KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

LS-SVM Regression Modelling and its Applications

Promotoren:
Prof. dr. ir. J. Vandewalle
Prof. dr. ir. J. Suykens

Proefschrift voorgedragen tot het behalen van het doctoraat in de toegepaste wetenschappen

door

Jos De Brabanter

June 2004

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

LS-SVM Regression Modelling and its Applications

Jury:
Prof. dr. P. Verbaeten, voorzitter
Prof. dr. ir. J. Vandewalle, promotor
Prof. dr. ir. J. Suykens, promotor
Prof. dr. ir. S. VanHuffel
Prof. dr. ir. A. Barbé
Prof. dr. J. Beirlant
Prof. dr. D. Bollé
Prof. dr. N. Veraverbeke (LUC)
Prof. dr. L. Györfi (Budapest Univ.)

Proefschrift voorgedragen tot het behalen van het doctoraat in de toegepaste wetenschappen

door

Jos De Brabanter

U.D.C. 519.233.5

June 2004

ISBN 90-5682-521-6
UDC 519.233.5
D/2004/7515/58

Voorwoord

De jaren studie en onderzoek aan het departement elektrotechniek waren een interessante en leerzame periode, waarin tal van interessante onderzoeksuitdagingen geformuleerd en opgelost werden. Tijdens deze periode heb ik ook de gelegenheid gehad met vele mensen samen te werken aan publicaties. Bij het begin van dit proefschrift wil ik hen graag bedanken voor de constructieve bijdragen en de aangename werksfeer.

In de eerste plaats dank ik mijn beide promotoren, prof. dr. ir. Joos Vandewalle en prof. dr. ir. Johan Suykens. Prof. dr. ir. Joos Vandewalle wil ik bedanken voor de inleiding tot neurale netwerken. Tegelijk ben ik hem dankbaar voor de vrijheid om me te verdiepen in statistische toepassingen. Prof. dr. ir. Johan Suykens ben ik vooral dankbaar voor het voorstellen van talrijke boeiende onderzoeksopdrachten. Hij bracht me de basisprincipes van support vector machines bij. De interne technische discussies waren bijzonder leerrijk en productief. Beide promotoren samen waren een steun en stimulans voor mijn onderzoek.

De assessoren van het leescomité, Prof. dr. ir. Sabine Van Huffel, Prof. dr. ir. André Barbé, Prof. dr. Jan Beirlant en Prof. dr. Desiré Bollé, wil ik bedanken voor hun begeleiding gedurende de vier onderzoeksjaren en voor hun opbouwende kritiek in verband met het verbeteren van de tekst.

Prof. dr. Noël Veraverbeke (LUC) ben ik erkentelijk omdat hij onmiddellijk bereid was deel uit te maken van de jury. It is for our research group and in particular for myself a big honour that prof. dr. László Györfi wants to participate in the jury. Tenslotte wil ik prof. dr. P. Verbaeten bedanken voor het waarnemen van het voorzitterschap van de examencommissie.

Tevens wil ik Prof. dr. Dirk Timmerman, Prof. dr. Ignace Vergote en dr. Dirk Amant bedanken van de afdeling gynaecologie-verloskunde van het U.Z. Leuven waarmee op regelmatige basis is samengewerkt. In dit verband zou ik hier ook Prof. dr. ir. Sabine Van Huffel willen vermelden voor de aangename samenwerking.

Ook de collega's van de onderzoeksgroep wil ik bedanken voor de aangename werksfeer. Hierbij denk ik dan vooral aan de directe collega's Bart, Kristiaan en Luc. Zeker mag ik mijn collega's binnen de bio-informaticagroep en SCD niet vergeten, die altijd klaar stonden als ik hulp nodig had. Een speciale vermelding verdienen zeker Tony, Frank, Patrick, Lieveke, Andy en Lukas. Ida, Pela, Ilse en Bart wil ik bedanken omdat ze altijd klaar stonden om praktische vragen en problemen op te lossen. Tevens ben ik de Katholieke Universiteit Leuven erkentelijk voor de financiële steun.

Tenslotte wil ik benadrukken dat dit proefschrift er ook gekomen is dankzij de steun van mijn familie, waarbij ik bij deze gelegenheid vooral mijn echtgenote en zoon Kris wil bedanken.

Jos De Brabanter
Leuven, juni 2004

Abstract

The key method in this thesis is least squares support vector machines (LS-SVM), a class of kernel based learning methods that fits within the penalized modelling paradigm. Primary goals of the LS-SVM models are regression and classification. Although local methods (kernel methods) focus directly on estimating the function at a point, they face problems in high dimensions. Therefore, one can guarantee good estimation of a high-dimensional function only if the function is extremely smooth. We have incorporated additional assumptions (the regression function is an additive function of its components) to overcome the curse of dimensionality.

We have studied the properties of the LS-SVM regression when relaxing the Gauss-Markov conditions. It was recognized that outliers may have an unusually large influence on the resulting estimate. However, asymptotically the heteroscedasticity does not play any important role. We have developed a robust framework for LS-SVM regression. It allows to obtain a robust estimate based upon the previous LS-SVM regression solution, in a subsequent step. The weights are determined based upon the distribution of the error variables. We have shown, based on the empirical influence curve and the maxbias curve, that the weighted LS-SVM regression is a robust function estimation tool. We have used the same principle to obtain an LS-SVM regression estimate in the heteroscedastic case. However, the weights are then based upon a smooth error variance estimate.

Most efficient learning algorithms in neural networks, support vector machines and kernel based learning methods require the tuning of some extra tuning parameters. For practical use, it is often preferable to have a data-driven method to select these parameters. Based on location estimators (e.g., mean, median, M-estimators, L-estimators, R-estimators), we have introduced robust counterparts of model selection criteria (e.g., Cross-Validation, Final Prediction Error criterion).

Inference procedures for both linear and nonlinear parametric regression models in fact assume that the output variable follows a normal distribution. With nonparametric regression, the regression equation is determined from the data. In this case, we relax the normality assumption and standard inference procedures are no longer applicable in that case. We have developed a robust approach for obtaining robust prediction intervals by using robust external bootstrapping methods.

Finally, we apply LS-SVM regression modelling in the case of density estimation.

Korte Inhoud

Dit proefschrift handelt over de kleinste kwadraten support vector machines (LS-SVM), een klasse van kernel gebaseerde leermethoden die behoren tot het regularizeerd modellerings paradigma. Voornaamste doelen van LS-SVM modellen zijn regressie en classificatie. Hoewel lokale methodes zich onmiddellijk focussen op de schatting van de functie in een punt, ondervinden zij problemen in hoge dimensies. Daarom kan men enkel een goede schatting van een functie bekomen in hoge dimensies als de functie extreem glad is. We hebben bijkomende veronderstellingen toegevoegd (de regressie functie is een additieve functie in zijn componenten) om de vloek van de dimensionaliteit te overwinnen.

De eigenschappen van LS-SVM regressie werden bestudeerd in geval de Gauss-Markov voorwaarden niet vervuld zijn. Uitschieters kunnen een abnormaal grote invloed hebben op de resulterende schatting. Maar asymptotisch heeft de heteroscedasticiteit geen belangrijke invloed. Een kader voor de LS-SVM regressie werd ontwikkeld. Dit laat toe een robuuste schatting te bekomen gebaseerd op een voorgaande LS-SVM oplossing, in een volgende stap. De daartoe ingevoerde gewichten zijn gebaseerd op de kansverdeling van de foutvariabelen. Via de empirische invloeds curve en de maxbias curve hebben we aangetoond dat de gewogen LS-SVM regressie een robuuste schattingstechniek is. Hetzelfde principe werd toegepast om een LS-SVM schatting te bekomen in het heteroscedastisch geval, waarbij dan de gewichten gebaseerd zijn op een gladde foutvariantie schatting.

De meest efficiente leeralgoritmen in neurale netwerken, support vector machines en kernel gebaseerde leermethoden vereisen de bepaling van extra leerparameters. Bij praktisch gebruik wordt de voorkeur gegeven aan data-gedreven methodes om deze parameters te selecteren. Gebaseerd op lokatie schatters (vb. mediaan, M-schatters, L-schatters, R-schatters) hebben we robuuste equivalente modelselectie criteria (bvb. Cross-Validatie, Final Prediction Error Criterion) geïntroduceerd.

Inferentie procedures voor beide lineaire- en niet lineaire parametrische regressie modellen veronderstellen een normaal onderliggende kansverdeling voor de uitgangsvariabelen. Bij niet-parametrische regressie wordt de regressie vergelijking afgeleid van de data. In dit geval wordt de veronderstelling van normaliteit afgezwakt en de standaard inferentieprocedures kunnen niet meer worden toegepast in dat geval. Door gebruik te maken van robuuste External Bootstrapping methodes hebben we een robuuste manier ontwikkeld tot het bekomen van robuuste prediktie intervallen.

Ten slotte hebben we LS-SVM regressie gebruikt als kansdichtheid schatter.

List of Symbols

$a \in A$	a is an element of the set A
$A \subseteq B$	Set A is contained in the set B; i.e., A is a subset of B
$A \subset B$	$A \subseteq B$ and $A \neq B$; i.e., set A is a proper subset of B
$\Rightarrow$	Implies
$\lfloor x \rfloor$	Integer part of the real number x
O, o	Order of magnitude symbols
$\sim$	Asymptotically equal
$I_A(x) = I_{\{x \in A\}}$	Indicator function of a set A
$\{x : ...\}$	Set of all elements with property ...
$\log$	Natural logarithm (base e)
$[x]_+$	$\max\{x, 0\}$
$d(.,.)$	Distance function
d_1, d_2, d_∞	Particular distance functions
$\|\cdot\|_\infty$	Uniform norm
$\|\cdot\|_p$	p-norm
$\sup A$	Supremum or least upper bound of the set A
$\inf A$	Infimum or greatest lower bound of the set A
$\mathbb{N}$	Set of all natural numbers, $\{1, 2, ...\}$
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of nonnegative real numbers
$\mathbb{R}^d$	Set of d-dimensional real numbers
$\mathcal{F}$	Class of functions $f : \mathbb{R}^d \to \mathbb{R}$
$f : C \to D$	A function from C to D
$f(x)$	The value of the function at x
φ	Nonlinear mapping from input space to feature space
$C\left(\mathbb{R}^d\right)$	Set of all continuous functions $f : \mathbb{R}^d \to \mathbb{R}$
$C^v(\mathcal{X})$	Set of all v times continuously differentiable functions $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^d$
$C^\infty\left(\mathbb{R}^d\right)$	Set of all infinitly often continuously differentiable - functions $f : \mathbb{R}^d \to \mathbb{R}$
L^2	Space of square-integrable functions

F	Distribution function of a random variable
$\Pr(A)$	Probability of the event A
$\hat{F}_n$	Empirical distribution
$\mathcal{N}\left(\mu, \sigma^2\right)$	The one-dimensional normal distribution or random variable with mean μ and variance σ^2
$\mathcal{AN}\left(\mu, \sigma^2\right)$	Asymptotic normal
$T(F)$	Statistic
$T(\hat{F}_n)$	Estimation of the statistic
$E\left[X\right]$	Expectation value of X
$\mathrm{Bias}\left[T(\hat{F}_n), T(F)\right]$	Bias of the estimator $T(\hat{F}_n)$
$\mathrm{Var}[X]$	Variance of X
$\mathrm{Cov}[X, Y]$	Covariance of X and Y
$\mathrm{Corr}[X, Y]$	Correlation of X and Y
$\xrightarrow{wp1}$	Convergence with probability 1
$\xrightarrow{p}$	Convergence in probability
$\xrightarrow{d}$	Convergence in distribution
$\mathcal{D}_n = \left\{\left(x_1, y_1\right), ..., \left(x_n, y_n\right)\right\}$	Training data
$m = E\left[Y \mid X = x\right]$	Regression function
$\hat{m}_n\left(x\right)$	Regression estimate
$x^{(1)}, ..., x^{(d)}$	Components of the d-dimensional column vector
$\mathcal{R}$	Risk functional
$\mathcal{R}_{emp}$	Empirical risk functional
$u = \arg\min\limits_{x \in D} f(x)$	Abbreviation for $u \in D$ and $f(z) = \min\limits_{x \in D} f(x)$
$K : \mathbb{R}^d \to \mathbb{R}$	Kernel function
$h > 0$	Smoothing parameter for kernel function
$(V, \Vert . \Vert)$	Normed space

Acronyms

SVM	Support Vector Machine
LS-SVM	Least Squares Support Vector Machine
RSS	Residual Sum of Squares
CV	Cross-Validation
GCV	Generalized Cross-Validation
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
VC	Vapnik-Chervonenkis dimension
SRM	Structural Risk Minimization
MSE	Mean Squared Error
FPE	Final Prediction Error
i.i.d.	Independent and identically distributed
cdf	Cumulative distribution function
pdf	Probability density function
QQ	Quantile-Quantile
OLS	Ordinary least squares
LS	Least squares
LAD	Least Absolute Deviations
MAD	Minimum absolute deviations
MAE	Minimum absolute errors
LAR	Least absolute residuals
LAV	Least absolute values
IF	Influence Function
ERM	Empirical Risk Minimization
SRM	Structural Risk Minimization

Samenvatting

Modellering en toepassingen van LS-SVM regressie

Hoofdstuk 1: Inleiding

In 1896, publiceerde Pearson zijn eerste verhandeling i.v.m. correlatie en regressie in de Filosofische Transacties van de Koninklijke Maatschappij van Londen. In feite werden de belangrijkste ideeën van het parametrische paradigma ontwikkeld tussen 1920 en 1960 (zie Fischer, 1952). Tijdens deze periode, werd de methode van maximum waarschijnlijkheid voor het schatten van parameters geïntroduceerd. Nochtans, toonde Tukey aan dat echte problemen niet door klassieke statistische verdelingsfuncties kunnen worden beschreven. Bovendien construeerden James en Stein (1961) een geregulariseerde schatter van het gemiddelde (normaal verdeelde vectoren) dat voor om het even welk vast aantal observaties uniform beter is dan de raming door de steekproef. Deze moeilijkheden met het parametrische paradigma en verscheidene ontdekkingen (samengevat in de volgende 4 punten) die in de jaren '60 worden gemaakt, waren een keerpunt in de statistiek en leidden tot een nieuw paradigma: (*i*) Het bestaan van hoge snelheid, goedkope gegevensverwerking. (*ii*) De theorie van slecht-gestelde problemen. (*iii*) De generalisatie van het glivenko-cantelli-Kolmogorov theorema. (iv) De controle van de capaciteit.

Een nieuwe richting werd aangekondigd, de zogenaamde "gegevensanalyse". Aan het eind van de jaren '60, werd de theorie van de Empirische Minimalisering van het Risico (ERM) voor het classificatie probleem geconstrueerd (Vapnik en Chervonenkis, 1974). Binnen 10 jaar, werd de theorie van het ERM principe eveneens veralgemeend voor reeksen van functies (Vapnik, 1979). Het idee van het minimaliseren van de testfout door twee tegenstrijdige factoren te controleren werd geformaliseerd door een nieuw principe, het Minimaliseren van het Structureel Risico (SRM). De Support vector methode realiseert het SRM principe. De SVM voor het schatten van functies werd geintroduceerd door Vapnik (1995). Kleinste kwadraten support vector machines (LS-SVM) (Suykens en Vandewalle, 1999; Suykens et al, 2002) zijn herformuleringen van de standaard SVM die leiden tot het oplossen van lineaire systemen voor classificatietaken en regressie. Naast zijn lange geschiedenis, is het probleem van regressieschatting vandaag nog steeds aan de orde.

Structuur van de thesis

Deel I behandelt de methoden en technieken van niet-parametrische regressie modellering. *Hoofdstuk 2* introduceert het probleem van de regressiefunctie schatting en beschrijft belangrijke eigenschappen van de regressieramingen. In *hoofdstuk 3* verklaren wij support vector machines. In *Hoofdstuk 4* beschrijven wij methoden (bvb. cross-validation en Final Prediction Error criterium) voor prestatiebeoordeling. *Hoofdstuk 5* bespreekt de Jackknife en bootstrap technieken.

In **Deel II** beschouwen we het probleem van hoog-dimensionale data, het heteroscedastische geval en het probleem van de waarschijnlijkheidsdichtheid schatting. *Hoofdstuk 6* bespreekt belangrijke kenmerken van hogere dimensionale problemen. In *Hoofdstuk 7* beschrijven wij methoden voor het schatten van de foutvariantie. In *Hoofdstuk 8* gebruiken wij de LS-SVM regressie modellering voor kansdichtheid schatting.

Deel III verstrekt een inleiding tot methoden van robuuste statistiek. In *Hoofdstuk 9* bekijken wij diverse maten van robuustheid (bvb. invloedsfunctie, maxbias curve). Daarnaast introduceren wij een robuuste versie van de LS-SVM. In *Hoofdstuk 10* construeren wij een gegeven-gedreven lossfunctie voor regressie. *Hoofdstuk 11* beschrijft robuuste tegenhangers van modelselectie criteria (bvb. cross-validation en Final Prediction Error criterium). *Hoofdstuk 12* illustreert inferentie met niet-parametrische modellen. Wij bespreken een robuuste methode voor het verkrijgen van robuuste voorspellingsintervallen. In *Hoofdstuk 13* worden de belangrijkste resultaten van deze thesis samengevat en de onderwerpen voor verder onderzoek worden aangehaald.

Bijdragen

De belangrijkste methode in deze thesis is de LS-SVM, een voorbeeld van het geregulariseerde modellerings paradigma. Wij hebben een nieuwe methode, componentwise LS-SVM geïntroduceerd, voor het schatten van modellen die uit een som van niet-lineaire componenten bestaan (Pelckmans et al, 2004).

We hebben het idee van de ruisvariantie schatter geintroduceerd door Rice (1984) veralgemeend voor multivariate data. We hebben de eigenschappen van de LS-SVM regressie bestudeerd bij afgezwakte Gauss-Markov condities. Kwadratische residuen plots werden voorgesteld om de heteroscedasticiteit te karakteriseren.

In LS-SVM's worden de oplossing gegeven door een lineair systeem (gelijkheidsbeperkingen) i.p.v. een QP probleem (ongelijkheidsbeperkingen). De SVM aanpak (Mukherjee en Vapnik, 1999) vereisen ongelijkheidsbeperkingen voor kansdichtheid schattingen. Een manier om deze ongelijkheidsbeperkingen te omzeilen, is het gebruik van regressie gebaseerde kansdichtheid schattingen. We hebben de LS-SVM regressie gebruikt voor kansdichtheid schatting.

Wij hebben een robuust kader voor LS-SVM regressie ontwikkeld. Het kader laat toe om een robuuste raming te verkrijgen die op de vorige LS-SVM regressie oplossing wordt gebaseerd, in een opeenvolgende stap. De gewichten worden bepaald welke gebaseerd zijn op de verdeling van de foutvariabelen (Suykens et al, 2002). Wij hebben aangetoond, gebaseerd op de empirische invloedsfunctie en de maxbias curve, dat de gewogen LS-SVM regressie een robuuste functieschatting is. Wij hebben hetzelfde principe gebruikt om een LS-SVM regressieraming in het heteroscedastisch geval te verkrijgen. Nochtans zijn de gewichten nu gebaseerd op een gladde raming van de foutvariantie.

Thans bestaat er een variatie van loss functies (bvb., least squares, least absolute deviations, M-estimators, generalized M-estimators, L-estimators, R-estimators, S-estimators, least trimmed sum of absolute deviations, least median of squares, least trimmed squares). Anderzijds brengt dit de data analyst in een moeilijke situatie. Een idee voor deze situatie, voorgesteld in deze thesis, is als volgt. Gegeven de data, de methode kan gesplitst worden in twee hoofddelen: (*i*) opbouwen van een robuust niet parametrisch regressie model en berekenen van de residuen, en (*ii*) de foutverdeling via robuuste bootstrap bekomen en bepalen van de loss functie (in een maximum likelihood omgeving).

Meest efficiente leeralgoritmen in neurale networken, support vector machines en kernel based methoden (Bishop, 1995; Cherkassky *et al.*, 1998; Vapnik, 1999; Hastie *et al.*, 2001; Suykens *et al.*, 2002b) vereisen de bepaling van extra leerparameters. In praktijk wordt de voorkeur gegeven aan data-gedreven methoden voor het selecteren van de leerparameters. Gebaseerd op locatie schatters (bvb. mediaan, M-schatters, L-schatters, R-schatters), hebben we de robuuste tegenhangers geintroduceerd van modelselectiecriteria (bvb. Cross-Validation, Final Prediction Error criterion).

Bij niet-parametrische regressie wordt de regressie vergelijking bepaald via de data. In dit geval kunnen de standaard inferentie procedures niet toegepast worden. Daarom hebben we robuuste voorspellingsintervallen ontwikkeld gebaseerd op robuuste bootstrap technieken.

Hoofdstuk 2: Model Opbouw

De beschrijving betreffende de drie paradigma's in niet-parametrische regressie is gebaseerd op (Friedman, 1991).

Parametrische modellering

De klassieke benadering voor het schatten van een regressiefunctie is de parametrische regressieschatting. Men veronderstelt dat de structuur van de regressiefunctie gekend is en slechts afhankelijk is van enkele parameters. Het lineaire regressiemodel verstrekt een flexiebel kader. Nochtans, zijn de lineaire regressiemodellen niet aangewezen voor alle situaties. Er zijn vele situaties waar de afhankelijke veranderlijke en onafhankelijke variabelen door een bekende niet-lineaire functie verwant zijn.

Laat $\mathcal{F}$ de klasse zijn van lineaire combinaties van de componenten $x = \left(x^{(1)}, ..., x^{(d)}\right)^T \in \mathbb{R}^d$,

$$\mathcal{F} = \left\{ m : m\left(x\right) = \beta_0 + \sum_{l=1}^{d} \beta_l x^{(l)},\ \beta_0, ..., \beta_d \in \mathbb{R} \right\}.$$

Men gebruikt de data $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ om de onbekende parameters $\beta_0, ..., \beta_d \in \mathbb{R}$ te schatten door gebruik te maken van het kleinste kwadraten principe:

$$\left(\hat{\beta}_0, ..., \hat{\beta}_d\right) = \underset{\beta_0, ..., \beta_d \in \mathbb{R}}{\arg\min} \left[\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \beta_0 + \sum_{l=1}^{d} \beta_l x_k^{(l)} \right)^2 \right],$$

hierin is $x_k^{(l)}$ de lth component van $x_k \in \mathbb{R}^d$, $k = 1, ..., n$ en de schatting is gedefinieerd als

$$\hat{m}_n(x) = \hat{\beta}_0 + \sum_{l=1}^{d} \hat{\beta}_l x^{(l)}.$$

Nochtans, hebben de parametrische schattingen een nadeel. Ongeacht de data, kan een parametrische raming de regressiefunctie niet beter benaderen dan de beste functie met de veronderstelde parametrische structuur. Deze inflexibiliteit betreffende de structuur van de

regressiefunctie kan vermeden worden door niet-parametrische regressieschattingen.

Niet-parametrische modellering

Lokale averaging en lokale modellering

Een voorbeeld van *local averaging* schatting (kernel methoden) is de Nadaraya-Watson kernel schatting. Per definitie

$$\begin{aligned} m\left(x\right) = E\left[Y\left|X = x\right.\right] &= \int y f_{Y|X}\left(y\left|x\right.\right) dy \\ &= \int y \frac{f_{XY}(x,y)}{f_X(x)} dy, \end{aligned}$$

hierin zijn $f_X(x)$, $f_{XY}(x,y)$ en $f_{Y|X}\left(y\left|x\right.\right)$ de marginale kansdichtheid van X, de samengestelde kansdichtheid van X en Y, en de voorwaardelijke kansdichtheid van Y gegeven X, respectievelijk. Laat $K : \mathbb{R}^d \to \mathbb{R}$ de kernelfunctie zijn en laat $h > 0$ de bandbreedte zijn. De Nadaraya-Watson kernel schatter is gegeven door

$$\hat{m}_n\left(x\right) = \sum_{k=1}^{n} \frac{K\left(\frac{x-x_k}{h}\right) y_k}{\sum_{l=1}^{n} K\left(\frac{x-x_l}{h}\right)}.$$

Globale modellering

Men moet de set van functies beperken over de welke men de empirische L_2 risk functionaal minimaliseerd. De globale modellering schatting is dan gedefinieerd als

$$\hat{m}_n\left(\cdot\right) = \arg\min_{f\in\mathcal{F}_n} \left[\frac{1}{n}\sum_{k=1}^{n}\left(f\left(x_k\right) - y_k\right)^2\right]$$

en minimaliseert de empirische L_2 risk functionaal.

Gepenaliseerde modellering

In plaats van de set van functies te beperken, voegt de gepenaliseerde kleinste kwadraten schatting expliciet een term bij de functionaal dewelke moet geminimaliseerd worden. Laat $r \in \mathbb{N}$, $\lambda_n > 0$ en laat de univariate gepenaliseerde kleinste kwadraten schatting gedefinieerd worden door

$$\hat{m}_n(\cdot) = \arg\min_{f\in C^r(\mathbb{R})} \left[\frac{1}{n}\sum_{k=1}^{n}\left(f\left(x_k\right) - y_k\right)^2 + \lambda_n J_{n,v}\left(f\right)\right],$$

hierin is $J_{n,v}\left(f\right) = \int \left(f^v\left(u\right)\right)^2 du$ en $C^v(\mathbb{R})$ is de set van alle v keer differentieerbare functies $f : \mathbb{R}^d \to \mathbb{R}$. Voor de penalty term, $v = 2$, de minimum wordt bereikt door een cubic spline met knots in de x_k's.

Hoofdstuk 3: Kernel Geinduceerde Kenmerkenruimte en Support Vector Machines

In dit hoofdstuk geven wij een kort overzicht over de formuleringen van de standaard Vectormachines (SVM) zoals die door Vapnik werden geïntroduceerd. Wij bespreken niet-lineaire functieschatting door SVMs die gebaseerd zijn op de Vapnik -ϵ-insensitive kost. Daarna verklaren wij de basismethoden van kleinste kwadraten Vectormachines (LS-SVMs) voor niet-lineaire functieschatting.

LS-SVM regressie

Gegeven een training set gedefinieerd als $\mathcal{D}_n = \{(x_k, y_k) : x_k \in \mathcal{X}, y_k \in \mathcal{Y}; k = 1, ..., n\}$ met grootte n in overeenstemming met

$$y_k = f(x_k) + e_k, \qquad k = 1, ..., n,$$

waar $E[e_k |X = x_k] = 0$, $Var\,[e_k] = \sigma^2 < \infty$, $m\,(x)$ een ongekende gladde functie is en $E[y_k|x = x_k] = m(x_k)$. Het doel is de parameters w en b (primaire ruimte) te bepalen welke de emprische risk functionaal

$$\mathcal{R}_{emp}\,(w, b) = \frac{1}{n}\sum_{k=1}^{n}\left(\left(w^T\varphi\,(x_k) + b\right) - y_k\right)^2$$

minimaliseert met restrictie $\|w\|_2 \leq a$, $a \in \mathbb{R}_+$. Men kan het optimalisatie probleem voor het bepalen van de vector w en $b \in \mathbb{R}$ reduceren door het volgende optimilisatie probleem op te lossen

$$\min_{w,b,e} \mathcal{J}\,(w, e) = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{k=1}^{n} e_k^2,$$

zodanig dat

$$y_k = w^T\varphi\,(x_k) + b + e_k,\ k = 1, ..., n$$

Om het optimalisatieprobleem (in de duale ruimte) op te lossen definieert men de volgende Lagrangiaan functionaal

$$\mathcal{L}(w, b, e; \alpha) = \mathcal{J}(w, e) - \sum_{k=1}^{n}\alpha_k\left(w^T\varphi\,(x_k) + b + e_k - y_k\right),$$

met Lagrangiaan vermenigvuldigers $\alpha_k \in \mathbb{R}$ (support waarden). De condities voor optimaliteit zijn gegeven door

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial w} = 0 \rightarrow w = \sum\limits_{k=1}^{n}\alpha_k\varphi\,(x_k) \\ \dfrac{\partial\mathcal{L}}{\partial b} = 0 \rightarrow \sum\limits_{k=1}^{n}\alpha_k = 0 \\ \dfrac{\partial\mathcal{L}}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, & k = 1, ..., n \\ \dfrac{\partial\mathcal{L}}{\partial\alpha_k} = 0 \rightarrow w^T\varphi\,(x_k) + b + e_k = y_k,\ k = 1, ..., n \end{cases}$$

Na eliminatie van w, e bekomt men de oplossing

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\gamma} I_n \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right],$$

met $y = (y_1, ..., y_n)^T$, $1_n = (1, ..., 1)^T$, $\alpha = (\alpha_1; ...; \alpha_n)^T$ en $\Omega_{kl} = \varphi(x_k)^T \varphi(x_l)$ voor $k, l = 1, ..., n$. Overeenkomstig het Mercer's theorema, het resulterende LS-SVM model voor functie schatting wordt gegeven door

$$\hat{m}_n(x) = \sum_{k=1}^{n} \hat{\alpha}_k K(x, x_k) + \hat{b}.$$

Support Vector Machines

Gegeven de training data $(x_1, y_1), ..., (x_n, y_n)$, om een benadering van functies te vinden met volgende vorm $(x) = \sum_{k=1}^{n} \beta_k K(x, x_k) + b$, de empirische risk functionaal

$$\mathcal{R}_{emp}(w, b) = \frac{1}{n} \sum_{k=1}^{n} \left| \left(w^T \varphi(x_k) + b \right) - y_k \right|_\varepsilon$$

wordt geminimaliseerd rekening houdend met de restrictie $\|w\|_2 \leq a_n$, waarbij $|\cdot|_\varepsilon$de Vapnik ε-insensitive kostfunctie is, en gedefinieerd wordt als

$$|f(x) - y|_\varepsilon = \begin{cases} 0, & \text{als } |f(x) - y| \leq \varepsilon, \\ |f(x) - y| - \varepsilon, & \text{anders.} \end{cases}$$

Na constructie van de Lagrangiaan functionaal en de condities voor optimaliteit bekomt men het volgende duale probleem

$$\begin{aligned} \text{[D]} \min_{\alpha, \alpha^*} J_{\mathrm{D}}(\alpha, \alpha^*) = & - \frac{1}{2} \sum_{k,l=1}^{n} (\alpha_k - \alpha_k^*)(\alpha_l - \alpha_l^*) K(x_k, x_l) \\ & - \frac{1}{2} \sum_{k,l=1}^{n} (\alpha_k - \alpha_k^*)(\alpha_l - \alpha_l^*) K(x_k, x_l) \\ & - \varepsilon \sum_{k=1}^{N} (\alpha_k + \alpha_k^*) + \sum_{k=1}^{N} y_k (\alpha_k - \alpha_k^*) \end{aligned}$$

$$\text{such that} \quad \sum_{k,l=1}^{n} (\alpha_k - \alpha_k^*) = 0, \qquad \alpha_k, \alpha_k^* \in [0, c]$$

waar $\beta_k = (\alpha_k - \alpha_k^*)$, $k = 1, ..., n$.

Hoofdstuk 4: Model Beoordeling en Selectie

In dit hoofdstuk worden de belangrijkste methoden beschreven (cross-validation en complexity criteria) voor model selectie. We beginnen dit hoofdstuk met het bias-variantie evenwicht en model complexiteit. Tenslotte geven we een parameter selectie strategie.

Introductie

Het meest efficiënte leeralgoritme in neurale netwerken, support vector machines en kernel gebaseerde methoden (Bishop, 1995; Cherkassky *et al.*, 1998; Vapnik, 1999; Hastie *et al.*, 2001; Suykens *et al.*, 2002b) vereisen de bepaling van extra leerparameters, hier voorgesteld door θ. De leerparameter selectie methoden kunnen ingedeeld worden in drie klassen:

(*i*) Cross validation en bootstrap.

(*ii*) Plug-in methoden.

(*iii*) Complexiteit criteria. Mallows' C_p (Mallows, 1973), Akaike's information criterion (Akaike, 1973), Bayes Information Criterion (Schwartz 1979) en Vapnik-Chernovenkis dimensie (Vapnik, 1998).

Het typisch gedrag van de test en trainingsfout, wanneer de model complexiteit verandert, wordt weergegeven in Figuur 1. De trainingsfout vertoont een dalende karakteristiek wanneer de modelcomplexiteit stijgt (Bishop, 1995) en (Hastie *et al.*, 2001). Bij overfitting zal het model zichzelf zodanig aanpassen aan de traingsdata zodat het niet goed generaliseerd.

Bij een te lage modelcomplexiteit stijgt de bias en de generalisatie is slecht. Om dit welgekend probleem te vermijden verdeeld men de data set $\mathcal{D}_n = \{(x_k, y_k) : x_k \in \mathcal{X}, y_k \in \mathcal{Y}; k = 1, ..., n\}$ in drie delen: een training set voorgesteld door $\mathcal{D}_n$, een validatie set voorgesteld door $\mathcal{D}_v$, en een test set voorgesteld door $\mathcal{D}_{test}$. De training set wordt gebruikt om de modellen te fitten; de validatie set wordt gebruikt om de predictie fout voor de modelselectie te schatten; de test set om de generalisatie fout van het eindmodel toe te kennen. De complexiteit criteria en de cross-validatie methoden benaderen de validatiestap respectievelijk analytisch en bij hergebruik van de sample.

Cross-validatie

Leave-one-out cross-validatie score functie

De kleinste kwadraten cross-validatie keuze van θ voor de LS-SVM schatters gebaseerd op het gemiddelde van de gekwadatreerde predicitiefout is de minimizer van

$$\inf_{\theta} CV(\theta) = \frac{1}{n} \sum_{k=1}^{n} (y_k - \hat{m}_n^{(-k)}(x_k; \theta))^2.$$

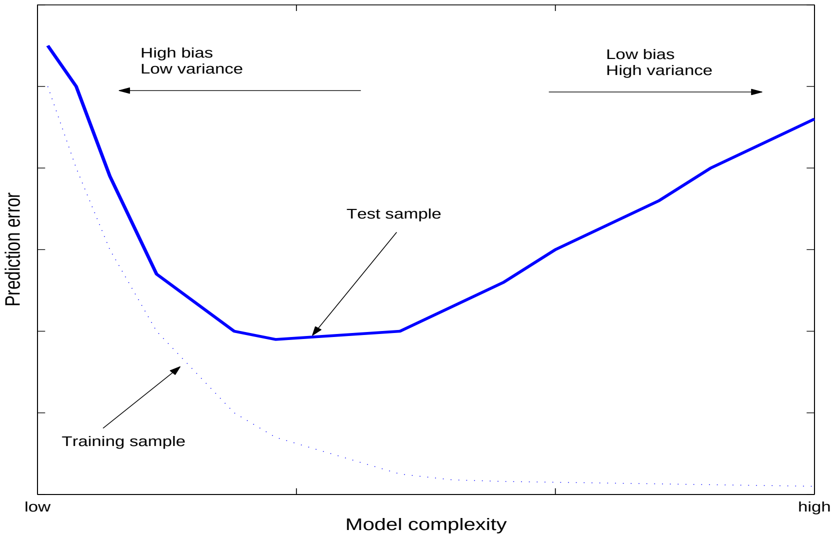

Figure 1: Gedrag van de test sample en training sample fout in functie van de model complexiteit.

Generalized cross-validatie score functie

De generalized cross-validatie score is gegeven door

$$GCV\left(\theta\right)=\frac{1}{n}\frac{\sum_{k=1}^{n}\left(y_k-\hat{m}_n\left(x_k;\theta\right)\right)^2}{\left(1-n^{-1}\mathrm{tr}\left[S(\theta)\right]\right)^2}.$$

Hierin is $S(\theta)$ de smoother matrix. Zoals bij de gewone cross-validatie, de GCV keuze van de leerparameters worden dan verkregen bij het minimaliseren van de functie $GCV\left(\theta\right)$ over θ.

V-fold cross-validatie score functie

We beginnen de data willekeurig te verdelen in V disjunct sets van ongeveer gelijke grootte. De grootte van de v^{de} groep wordt voorgesteld door m_v en verondersteld dat $\lfloor n/V \rfloor \leq m_v \leq \lfloor n/V \rfloor + 1$ voor alle v. Voor elke verdeling passen we Leave-one-out toe en maken het gemiddelde van deze schattingen. Het resultaat is de V-fold cross-validatie schatting van de predictie fout

$$CV_{V-fold}\left(\theta\right)=\sum_{v=1}^{V}\frac{m_v}{n}\sum_{k=1}^{m_v}\frac{1}{m_v}\left(y_k-\hat{m}_n^{(-m_v)}\left(x_k,\theta\right)\right)^2.$$

hierin stelt $\hat{f}^{(-m_v)}$ het verkregen model gebaseerd op de data welke niet behoren tot de groep v.

Complexiteit criteria

Final Prediction Error (FPE) criterium

Laat $\mathcal{P}$ een eindige set van parameters zijn. Voor $\alpha \in \mathcal{P}$, laat $\mathcal{F}_\beta$ een set van functies zijn

$$\mathcal{F}_\beta = \left\{ m : m\left(x, \beta\right) = \beta_0 + \sum_{l=1}^{d} \beta_l x^{(l)},\ x \in \mathbb{R}^d \text{ and } \beta \in \mathcal{P} \right\},$$

laat $Q_n(\beta) \in \mathbb{R}^+$ een complexiteitsterm voor $\mathcal{F}_\beta$ zijn en laat $\hat{m}_n$ een schatter zijn van m in $\mathcal{F}_\beta$. De leerparameters worden zodanig bepaald zodat de cost functie gedefinieerd als

$$J_\beta(\lambda) = \frac{1}{n} \sum_{k=1}^{n} L\left(y_k, \hat{m}_n\left(x_k; \beta\right)\right) + \lambda \left(Q_n(\beta)\right) \hat{\sigma}_e^2$$

zijn minimum bereikt. Hierin is $\sum_{k=1}^{n} L(y_k, \hat{m}_n(x_k; \beta))$ de som van de geschatte kwadratische fouten, $Q_n(\beta) \in \mathbb{R}^+$ is een complexiteitsterm, $\lambda > 0$ is een cost complexiteits parameter en de term $\hat{\sigma}_e^2$ is een schatting voor de error variantie. The Final Prediction Error criterium is enkel afhankelijk van $\hat{m}_n$ en de data.

Vapnik-Chervonenkis dimensie

De Vapnik-Chernovenkis theorie geeft een andere meting van de complexiteit dan het effectief aantal parameters en geeft de hierbij behorende begrenzingen. Veronderstel dat we een klasse van functies hebben

$$\mathcal{F}_{n,\beta} = \left\{ m : m(x, \beta),\ x \in \mathbb{R}^d \text{ en } \beta \in \Lambda \right\},$$

waarin Λ een parameter vector set is en beschouw de indicator klasse

$$\mathcal{I}_{\beta,\tau} = \left\{ I : I\left(m(x, \beta) - \tau\right),\ x \in \mathbb{R}^d,\ \beta \in \Lambda \text{ en } \tau \in \left(\inf_x m(x, \beta), \sup_x m(x, \beta) \right) \right\}.$$

De VC-dimensie (Vapnik, 1998) van reële waarde functies $\mathcal{F}_{n,\beta}$ is gedefinieerd als de VC-dimensie van de indicator klasse $\mathcal{I}_{\beta,\tau}$. De VC-dimensie van de klasse $\mathcal{F}_\beta$ is gedefinieerd als het grootste aantal punten welke kunnen gescheiden worden door elementen van $\mathcal{F}_{n,\beta}$.

Als $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ past, gebruik makende van een functie klasse $\mathcal{F}_{n,\beta}$ met VC-dimensie h, met probabiliteit $(1 - \alpha)$ over de training sets, zal de volgende ongelijkheid

$$R\left(f\right) \leq \frac{R_n\left(f\right)}{\left(1 - c\sqrt{\xi\left(n\right)}\right)_+}$$

gelden, waarin

$$\xi\left(n\right) = a_1 \frac{h\left(\log\left(\frac{a_2 n}{h}\right) + 1\right) - \log\left(\frac{\alpha}{4}\right)}{n},$$

en $a_1 = a_2 = c = 1$ (Cherkassky en Mulier, 1998). Deze begrenzingen zijn gelijktijdig van toepassing voor alle elementen van $\mathcal{F}_{n,\beta}$.

	σ^2	Smoother matrix	Opmerkingen
Leave-one-out	niet nodig	niet nodig	Grote variantie lage bias
V-fold-CV	niet nodig	niet nodig	lage variantie grote bias
GCV	niet nodig	nodig	(*)
AIC	nodig	nodig	
BIC	nodig	nodig	
SRM	niet nodig	niet nodig	

Table 1: De strategie voor het selecteren van een goede leer parameter vector. (*): Voor een gegeven data set, GCV selecteert altijd dezelfde leer parameter vector, ongeacht de grootte van de ruis.

Keuze van de leerparameters

De strategie, voor het selecteren van een goede leerparameter vector, is het gebruik maken van één of meerdere selectie criteria. De keuze van het gebruikte criteria is afhankelijk van de situatie. Tabel 1 geeft een samenvatting van verschillende situaties.
Als σ^2 onbekend is en geen aanvaardbare schatter is beschikbaar, kan GCV of cross-validatie gebruikt worden aangezien zij geen schatting van de error variantie vereisen. Het gebruik van de cross-validatie zal leiden tot meer rekenwerk dan GCV. In de praktijk is het mogelijk om twee of meer risk schattingen te berekenen.

Hoofdstuk 5: De Jackknife en de Bootstrap

We beginnen dit hoofdstuk met de Jacknife. Vervolgens bespreken we de bootstrap als een algemene tool voor het toekennen van statistische nauwkeurigheid.

De Jackknife

De Jackknife schatter werd voorgesteld door (Quenouille, 1949) en benoemd door (Tukey, 1958). Deze techniek verlaagt de bias van een schatter (de Jackknife schatter). De procedure is als volgt. Laat $X_1, ..., X_n$ een willekeurig sample zijn met grootte n van een onbekende waarschijnlijkheidsverdeling F. Gebruik makend van de geobserveerde waarden $x_1, ..., x_n$ is men geïnteresseerd in een bepaalde statistic $T(F)$. Laat $T(\hat{F}_n)$ een schatter zijn voor $T(F)$. Verdeel het willekeurig sample in r groepen met grootte $l = \frac{n}{r}$ observaties. Verwijder groep per groep, en schat $T(F)$ gebaseerd op de overblijvende $(r-1)\,l$ observaties, gebruik makend van dezelfde voorgaande schattings procedure met een sample grootte n. Stel de schatter van $T(F)$ verkregen met de i^{de} groep te verwijderen door $T(\hat{F}_{(i)})$. Voor $i = 1, ..., r$, van pseudowaarden

$$J_i = rT(\hat{F}_n) - (r-1)\,T(\hat{F}_{(i)}),$$

en beschouw de Jackknife schatter van $T(F)$ gedefineerd door

$$\begin{aligned} J\left(T(\hat{F}_n)\right) &= \frac{1}{r}\sum_{i=1}^{r}\left(rT(\hat{F}_n) - (r-1)\,T(\hat{F}_{(i)})\right) \\ &= T(\hat{F}_n) - (r-1)\,\bar{T}(\hat{F}_{(i)}) \end{aligned}$$

waar $\bar{T}(\hat{F}_{(i)}) = \frac{1}{r}\sum_{i=1}^{r} T(\hat{F}_{(i)})$.

De Bootstrap

De bootstrap is een methode voor het schatten van de parameterdistributie door herbemonstering van de data. Een zeer goede inleiding tot de bootstrap kan gevonden worden in het werk van (Efron en Tibshirani, 1993). In vele situaties zijn aanpassingen mogelijk, door het wijzigen van herbemonsteringsschema of door wijziging van andere aspecten van de methode. Het bootstrap principe is geïllustreerd in het volgende algoritme (bootstrap principe).

Algoritme 1 *(bootstrap principe).*

(i) *Van $X = (x_1, ...x_n)$, bepaal de schatter $T_n(\hat{F}_n)$.*

(ii) *Construeer de empirische verdeling, $\hat{F}_n$, welke gelijke probabiliteit $1/n$ aan iedere observatie toekent (gelijk verdeelde willekeurige bemonstering).*

(iii) *Van de geselecteerde $\hat{F}_n$, ,neem een sample $X^* = (x_1^*, ...x_n^*)$, genaamd het bootstrap sample.*

(iv) *Benader de verdeling van $\mathcal{J}_n(X, T(F))$ door de verdeling van $\mathcal{J}(X^*, T(\hat{F}_n))$*

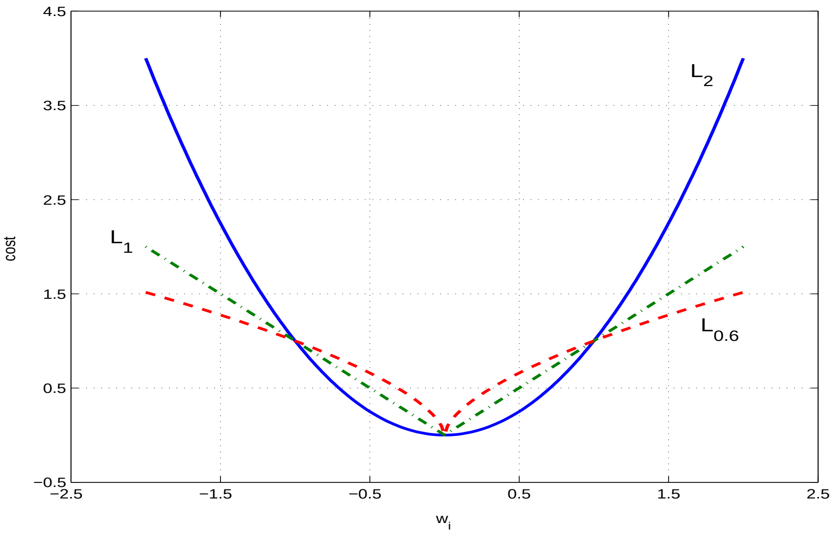

Figure 2: De L_p penalty familie voor $p = 2, 1$ en 0.6.

Hoofdstuk 6: LS-SVM voor Regressie Schatting

In dit hoofdsuk introduceren we een nieuwe methode, componentsgewijze LS-SVM, voor de schatting van additieve modellen (Pelckmans *et al.*, 2004).

Componentsgewijs LS-SVM regressie modellering

Beschouw de geregulariseerde kleinste kwadraten cost functie gedefineerd als

$$\mathcal{J}_\lambda \left(w^{(i)}, e \right) = \frac{\lambda}{2} \sum_{i=1}^{d} L \left(w^{(i)} \right) + \frac{1}{2} \sum_{k=1}^{n} e_k^2,$$

hierin is $L(w^{(i)})$ een penalty functie en $\lambda \in \mathbb{R}_0^+$ gedraagt zich als een regularisatie parameter. We stellen $\lambda L(\cdot)$ voor door $L_\lambda(\cdot)$, zodat het afhankelijk is van λ. Voorbeelden van penalty functies zijn:

(i) De L_p penalty functie $L_\lambda^p \left(w^{(i)} \right) = \lambda \left\| w^{(i)} \right\|_p^p$ leidt tot een bridge regressie (Frank en Friedman, 1993; Fu, 1998). Het is bekend dat de L_2 penalty functie resulteert in ridge regressie. Voor de L_1 penalty functie is de oplossing de soft thresholding regel (Donoho en Johnstone, 1994). (zie Figuur 2).

(ii) Wanneer de penalty functie gegeven is door

$$L_\lambda \left(w^{(i)} \right) = \lambda^2 - \left(\left\| w^{(i)} \right\|_1 - \lambda \right)^2 I_{\left\{ \left\| w^{(i)} \right\|_1 < \lambda \right\}}$$

(zie Figuur 3), de oplossing is een hard-thresholding regel (Antoniadis, 1997).

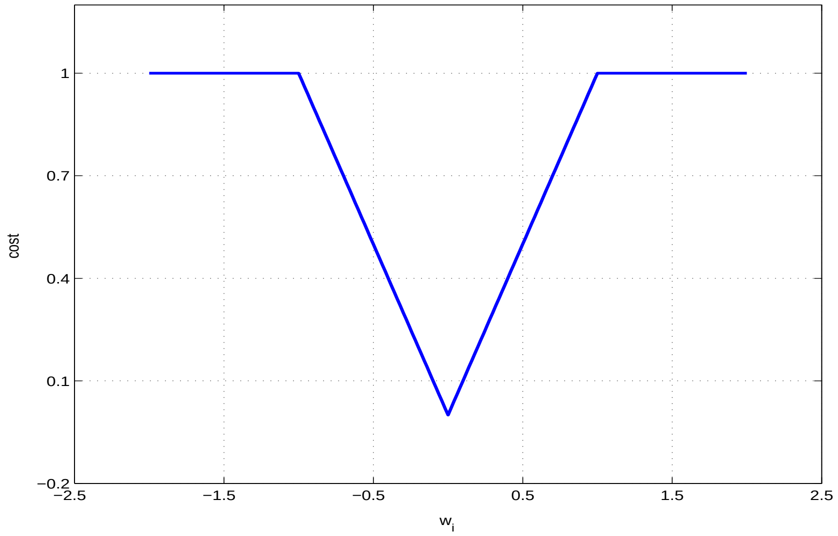

Figure 3: Hard thresholding penalty functie.

De L_p en de hard thresholding penalty functies voldoen tegelijkertijd niet aan de condities voor unbiasedness, sparsity en continuity (Fan and Li, 2001). De hard thresholding heeft een discontinue cost oppervlak. De enige continue cost oppervlak (gedefinieerd als de cost functie geassocieerd met de oplossingsruimte) met een thresholding regel in de L_p-familie is de L_1 penalty functie, maar de resulterende schatter is opgeschoven met een constante λ. Om deze ongemakken de vermijden, (Nikolova, 1999) definieert de penalty functie als volgt

$$L_{\lambda,a}\left(w^{(i)}\right) = \frac{a\lambda \left\|w^{(i)}\right\|_1}{1 + a\left\|w^{(i)}\right\|_1},$$

met $a \in \mathbb{R}$. Deze penalty functie gedraagt zich nogal gelijkaardig als de Smoothly Clipped Absolute Deviation (SCAD) penalty functie voorgesteld door (Fan, 1997). De Smoothly Thresholding Penalty (TTP) functie $L_{\lambda,a}\left(w^{(i)}\right)$ verbetert de eigenschappen van de L_1 penalty functie en de hard thresholding penalty functie (zie Figuur 4), zie (Antoniadis en Fan, 2001).

De onbekenden a en λ gedragen zich als regularisatie parameters. Een aanvaardbare waarde voor a werd afgeleid in (Nikolova, 1999; Antoniadis en Fan, 2001) als $a = 3.7$. Het componentsgewijze regularisatie schema wordt gebruikt voor de emulatie van de penalty functie $L_{\lambda,a}\left(w^{(i)}\right)$

$$\min_{w^{(i)},b,e_k} \mathcal{J}\left(w^{(i)}, e\right) = \frac{1}{2}\sum_{i=1}^{d} L_{\lambda,a}\left(w^{(i)}\right) + \frac{\gamma}{2}\sum_{k=1}^{n} e_k^2$$

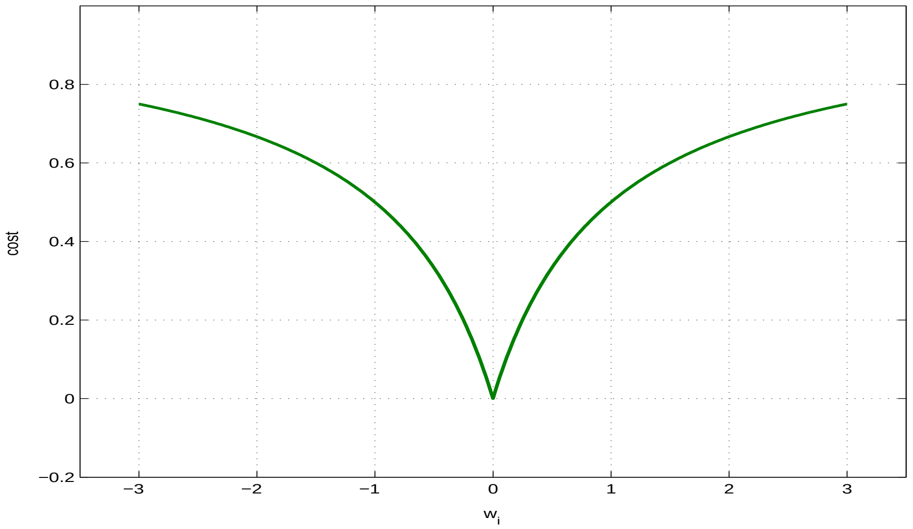

Figure 4: De getransformeeerde L_1 penalty functie.

zodat

$$y_k = \sum_{i=1}^{d} w^{(i)T} \varphi_i \left(x_k^{(i)} \right) + b + e_k,\ k = 1, ..., n.$$

welke niet convex wordt. Voor praktische toepassingen, wordt de iteratieve aanpak gebruikt voor het oplossen van niet convexe cost functies (Pelckmans *et al.*, 2004). De iteratieve aanpak is gebaseerd op de graduated non-convexity algoritme zoals voorgesteld in (Blake, 1989; Nikolova, 1999; Antoniadis en Fan, 2001) voor de optimisatie van niet convexe cost functies.

Hoofdstuk 7: Foutvariantie schatting

In dit hoofdstuk generaliseren we het idee van de niet-parametrische ruisvariantie schatter (Rice, 1984) voor multivariate data gebaseerd op U-statistics en differogram modellen (Pelckmans *et al.*, 2003). In het tweede deel van het hoofdstuk bestuderen we het gebruik van LS-SVM regressie in geval van heteroscedasticiteit.

Homoscedastische foutvariantie

Een voorbeeld van een variantie schatter σ^2 werd door Rice (1984) als volgt voorgesteld

$$\hat{\sigma}^2 = \frac{1}{2\,(n-1)} \sum_{k=1}^{n-1} \left(y_{k+1} - y_k\right)^2 .$$

Vervolgens zullen we het idee van Rice (1984) generaliseren voor multivariate data.

Definitie 2 *(U-statistic). Laat $g : \mathbb{R}^l \to \mathbb{R}$ een symmetrische functie zijn. De functie*

$$U_n = U\left(g; X_1, ..., X_n\right) = \frac{1}{\binom{n}{l}} \sum_{1 \le i_1 < ... < i_l \le n} g\left(X_{i_1}, ..., X_{i_l}\right), \quad l < n, \tag{1}$$

waar $\sum_{1 \le i_1 < ... < i_l \le n}$ de sum over $\binom{n}{l}$ combinaties van l verschillende elementen $\{i_1, ..., i_l\}$ van $\{1, ..., n\}$ is, wordt een U-statistic van orde l met kernel g genoemd.

Definitie 3 *(Differogram). De differogram $\Upsilon : \mathbb{R} \to \mathbb{R}$ wordt gedefinieerd door*

$$\Upsilon\left(\Delta x_{ij}\right) = \frac{1}{2} E\left[\Delta y_{ij} \left| \Delta x = \Delta x_{ij}\right.\right] \quad \textit{for } \Delta x \to 0, \tag{2}$$

waar $\Delta x_{ij} = \left\|x_i - x_j\right\|_2$, $\Delta y_{ij} = \left\|y_i - y_j\right\|_2 \in \mathbb{R}^+$ is. Gelijkaardig als in de variogram, geeft de intercept $\frac{1}{2} E\left[\Delta y_{ij} \left| \Delta x = \Delta x_{ij} = 0\right.\right]$ de ruisvariantie weer.

Differogram modellen gebaseerd op Taylor reeksontwikkeling

Beschouw de één-dimensionaal Taylor reeksontwikkeling van orde r in het center $x_i \in \mathbb{R}$

$$T_r\left(x_j - x_i\right) = m\left(x_i\right) + \sum_{l=1}^{r} \frac{1}{l!} \nabla^{(l)} m\left(x_j - x_i\right)^l + O\left(\left(x_j - x_i\right)^{r+1}\right),$$

waar $\nabla m\left(x\right) = \frac{\partial m}{\partial x}$, $\nabla^2 m\left(x\right) = \frac{\partial^2 m}{\partial x^2}$, enz. voor $l \le 2$. We beschouwen de r^{de} orde Taylor reeksbenadering van het differogram model met center $x_i = 0$ in het geval $\Delta x \to 0$. De differogram wordt gegeven door

$$\Upsilon\left(\Delta x, a\right) = a_0 + \sum_{l=1}^{r} a_l \Delta^l x, \quad a_0, ...a_r \in \mathbb{R}_+,$$

waar de parameter vector $a = \left(a_0, a_1, ..., a_r\right)^T \in \mathbb{R}_+^{r+1}$ wordt verondersteld uniek te zijn. De variantie functie ϑ van de schatter kan begrensd worden als volgt

$$\begin{aligned}\vartheta\left(\Delta x, a\right) = E\left[\left(\Delta y - \Upsilon\left(\Delta x, a\right)\left|\Delta x\right.\right)^2\right] &= \left[\left(\Delta y - a_0 - \sum_{l=1}^{r} a_l \Delta^l x \left|\Delta x\right.\right)^2\right] \\ &\leq E\left[\left(a_0 + \sum_{l=1}^{r} a_l \Delta^l x \left|\Delta x\right.\right)^2\right] + E\left[\left(\Delta y \left|\Delta x\right.\right)^2\right] \\ &= 2\left(a_0 + \sum_{l=1}^{r} a_l \Delta^l x\right)^2,\end{aligned}$$

steunend op de driehoeksongelijkheid en het differogram model. Volgende kleinste kwadraten methode kan worden gebruikt

$$\hat{a} = \arg\min_{a \in \mathbb{R}_+^{r+1}} \mathcal{J}\left(a\right) = \sum_{i \leq j}^{n} \frac{c}{\vartheta\left(\Delta x_{ij}, a\right)} \left(\Delta y_{ij} - \Upsilon\left(\Delta x_{ij}, a\right)\right)^2,$$

waar de constante $c \in \mathbb{R}_+^0$ de wegingsfunctie normaliseert zodanig dat $\sum_{i \leq j}^{n} \frac{c}{\vartheta(\Delta x_{ij}, a)} = 1$. De functie $\vartheta\left(\Delta x_{ij}, a\right) : \mathbb{R}_+ \to \mathbb{R}_+$ wordt als correctie gebruikt voor de heteroscedastische variantie structuur.

De differogram voor het schatten van de ruisvariantie

Gebaseerd op het differogram kunnen we de foutvariantie schatten. Bijvoorbeeld, laat $r = 0$, de 0^{de} orde Taylor polynomiaal van m in het punt x_i en geevalueerd in het punt x_j wordt gegeven door $T_0\left(x_j - x_i\right) = m\left(x_i\right)$ en de variantie schatter is

$$\begin{aligned}\hat{\sigma}_e^2 &= U\left(g; e_1, ..., e_n\right) \\ &= U\left(g; \left(y_1 - m\left(x_1\right), ..., \left(y_1 - m\left(x_1\right)\right)\right.\right. \\ &= \frac{1}{n\left(n-1\right)} \sum_{1 \leq i < j \leq n} \frac{1}{2}\left(y_i - y_j\right)^2.\end{aligned}$$

waar de benadering verbeterd als $x_i \to x_j$. Om dit te corrigeren kan men volgende kernel $g_1 : \mathbb{R}^2 \to \mathbb{R}$ gebruiken

$$g_1\left(y_i, y_j\right) = \frac{1}{2}\Delta y_{ij} \frac{c}{\vartheta\left(\Delta x_{ij}, a\right)}$$

waar de constante $c \in \mathbb{R}_+^0$ gekozen wordt zodanig dat de som van de gewogen termen constant zijn $2c\left(\sum_{i\leq j}^{n}\frac{1}{\vartheta(\Delta x_{ij})}\right) = n\,(n-1)\,.$ De variantie schatter wordt dan

$$\hat{\sigma}_e^2 = \frac{1}{n\,(n-1)}\sum_{1\leq i<j\leq n}\frac{1}{2}\left(\Delta y_{ij} - \sum_{l=1}^{r} a_l \Delta^l x\right)\frac{c}{\vartheta\,(\Delta x_{ij})}$$

Heteroscedastische foutvariantie

Kernel smoothing van lokale variantie schatters

Om de heteroscedasticiteit te schatten, maken we gebruik van kernel gebaseerde lokale variantie schatters. We veronderstellen dat: (i) De foutvariabelen e_k, $k = 1, ..., n$ zijn onafhankelijk, $E\left[e_k\right] = 0$, $E\left[e_k^2\right] = \sigma^2\,(z)$ waar $z = (x \text{ of } y)$ en $E\left[\left|e_k\right|^{2r}\right] \leq M < \infty$, $r > 1$. (ii) $m \in C^{\infty}\,(\mathbb{R})\,,$ en (iii) $\sigma^2\,(z) \in C^{\infty}\,(\mathbb{R})\,.$ Beschouw het regressie model

$$v_k = \sigma^2\,(z_k) + \varepsilon_k,\ \ k = 1, ..., n$$

waar v_k de initiele variantie schatters zijn. Om consistente schatters te bekomen (Müller en Stadtmüller, 1987), maken we gebruik van de Nadaraya-Watson schatter

$$\hat{\sigma}^2\,(z) = \sum_{k=1}^{n}\frac{K\left(\frac{z-z_k}{h}\right)v_k}{\sum_{l=1}^{n}K\left(\frac{z-z_l}{h}\right)},$$

waar K de kernel functie is en h de bandbreedte is zodanig dat $h \to 0$, $nh \to \infty$ als $n \to \infty$.

LS-SVM regressie schatting

Om een schatting te bekomen (heteroscedastisch geval) gebaseerd op de voorgaande LS-SVM oplossing, in een opeenvolgende stap, weegt men de foutvariabelen $e_k = \alpha_k/\gamma$ door wegingsfactoren ϑ_k . Dit leidt tot volgend optimalisatie probleem:

$$\min_{w^*,b^*,e^*}\mathcal{J}\,(w^*,e^*) = \frac{1}{2}w^{*T}w^* + \frac{1}{2}\gamma\sum_{k=1}^{n}\vartheta_k e_k^{*2} \qquad (3)$$

zodat $y_k = w^{*T}\varphi\,(x_k) + b^* + e_k^*, \quad k = 1, ..., n.$ De Lagrangiaan wordt geconstrueerd op een gelijkaardige manier als voordien. De ongekende variabelen voor dit gewogen LS-SVM probleem worden voorgesteld door het symbool $*$. Tengevolge van de condities voor optimaliteit en eliminatie van w^*, e^* bekomt men het Karush-Kuhn-Tucker systeem:

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \mathcal{V}_\gamma \end{array}\right]\left[\begin{array}{c} b^* \\ \hline \alpha^* \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right]$$

waar de diagonaal matrix $\mathcal{V}_\gamma$ wordt gegeven door $\mathcal{V}_\gamma = \text{diag}\left\{\frac{1}{\gamma\vartheta_1}, ..., \frac{1}{\gamma\vartheta_n}\right\}$. De gewichten

$$\vartheta_k = \frac{1}{\hat{\sigma}^2\left(z_k\right)},\ k = 1, ..., n, \tag{4}$$

worden bepaald gebaseerd op de lokale foutvariantie schatter. Gebruik makend van deze gewichten kan er gecorrigeerd worden in geval van heteroscedasticiteit.

Hoofdstuk 8: Kansdichtheid schatting

In dit hoofdstuk bespreken we de regressie kijk op de kansdichtheid schatting. Vervolgens gebruiken we de LS-SVM regressie modellering in het geval van kansdichtheid schatting.

Veronderstel dat $X_1, ..., X_n$ willekeurige variabelen zijn welke onafhankelijk en identiek verdeeld zijn volgens een welbepaalde probabiliteitsverdelingsfunctie F, waarin $F \in \mathcal{F}$, een familie van waarschijnlijkheidsverdelingsfuncties en waarschijnlijkheids kansdichtheidsfunctie f. De waarschijnlijkheidskansdichtheidsfunctie (pdf), welke volgende eigenschappen heeft $f(x) \geq 0$, f is stapsgewijs continue en $\int_{-\infty}^{\infty} f(x)dx = 1$, is gedefineerd als

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

Het probleem is een rij van schatters $\hat{f}_n(x)$ van $f(x)$ op te bouwen gebaseerd op de sample $x_1, ..., x_n$. Omdat niet vertekende schatters niet bestaan voor f (Rao, 1983), is men geïnteresseerd in asymptotisch niet vertekende schatters $\hat{f}_n(x)$ zodanig dat

$$\lim_{n\to\infty} E_{f\in\mathcal{F}_n}\left[\hat{f}_n(x)\right] = f(x), \qquad \forall x.$$

Support Vector Methode voor kansdichtheidschatting

De SVM aanpak (Mukherjee en Vapnik, 1999) beschouwd het probleem van pdf schatting als een probleem om $F(x) = \int_{-\infty}^{x} f(u)du$ op te lossen waar in plaats van $F(x)$ men een plug-in schatter $\hat{F}_n(x)$ gebruikt, de empirische verdelingsfunctie. Het oplossen van $Tf = F$ met benaderende $\hat{F}_n(x)$ is een slecht gesteld probleem. Methoden voor het oplossen van slecht gestelde problemen werden voorgesteld door (Tikhonov, 1963) en (Philips, 1962). Het oplossen van $F(x) = \int_{-\infty}^{x} f(u)du$ in een set van functies behorende tot een reproducerende kernel Hilbert ruimte, gebaseerd op de methoden voor het oplossen van slecht gestelde probelemen voor welke SVM technieken kunnen aangewend worden. Men minimaliseert

$$\begin{array}{ll} \min \sum_{i,j=1}^{n} \vartheta_i\vartheta_j K(x_i, x_j, h) & \\ \text{s.t.} & \left|\hat{F}_n(x) - \sum_{j=1}^{n} \vartheta_j \int_{-\infty}^{x} K(x_j, u, h)du\right|_{x=x_i} \leq \kappa_n, \qquad 1 \leq i \leq n, \\ & \vartheta_i \geq 0 \text{ en } \sum_{i=1}^{n} \vartheta_i = 1, \end{array}$$

waarin κ_n de bekende nauwkeurigheid is van de benadering van $F(x)$ door $\hat{F}_n(x)$ (Mukherjee en Vapnik, 1999). Om een oplossing te bekomen als een samenstelling van waarschijnlijkheidsdichtheidsfuncties moet de kernel een waarschijnlijke dichtheidsfunctie zijn en $\vartheta_i \geq 0$, $\sum_{i=1}^{n} \vartheta_i = 1$. Gewoonlijk zijn de meeste ϑ_i waarden in de SVM schatting gelijk aan nul en men bekomt een sparse schatter van een waarschijnlijkheidsdichtheidsfunctie. Een typische eigenschap van de SVM is dat de oplossing wordt gekarakteriseerd door een convex optimalisatie probleem, meer bepaald een kwadratisch programmeer (QP) probleem: in de LS-SVM wordt de oplossing gegeven door een lineair stelsel (gelijkheidsrestricties) in plaats van een QP probleem (ongelijkheidsresitricties). De SVM aanpak (Mukherjee en Vapnik, 1999) vereisen ongelijkheidsrestricties voor dichtheidschatting. Een mogelijkheid om deze ongelijkheidsresitricties te omzeilen is gebruik te maken van de regressie gebaseerde dichtheidsschatting aanpak. In deze aanpak kan men de LS-SVM regressie gebruiken voor kansdichtheidschatting.

Smoothing parameter selectie

Beschouw de Parzen kernel dichtheidsschatter. De vorm van de kernel is niet belangrijk (Rao, 1983). Een belangrijk probleem is het bepalen van de smoothing parameter. In de kernel dichtheidsschatting, heeft de bandbreedte een veel groter effect op de schatter dan op de kernel zelf. Er zijn vele methoden voor smoothing parameters selectie (bvb., least-squares cross-validation, least squares plug-in methods, the double kernel method, L_1 plug-in methods, etc.). In deze thesis gebruiken we een combinatie van cross-validatie en bootstrap voor het bepalen van de bandbreedte voor de Parzen kernel schatter.

Regressie kijk op de dichtheidsschatting

De kernel schatter heeft een nadeel wanneer gebruik gemaakt wordt van lange staart verdelingen. Een voorbeeld, gebaseerd op de data set aangehaald door (Copas en Fryer, 1980), van dit nadelig gedrag wordt voorgesteld in Figuur 5 en Figuur 6. De data set geeft de lengte van behandeling van controle patiënten in een zelfmoordstudie. De schatter weergegeven in Figuur 5 is ruisgevoelig in de rechter staart, terwijl de schatter weergeven in Figuur 6 gladder is. Noteer dat de data waarden positief zijn, de schatting weergegeven in Figuur 6 behandelt de data als observaties in het interval $(-\infty, \infty)$.

Om deze moeilijkheid te behandelen, werden verschillende adaptieve methoden voorgesteld (Breiman *et al.*, 1977). Logspline kansdichtheidsschatting, voorgesteld door (Stone en Koo, 1986) en (Kooperberg en Stone, 1990), volgt de staart vloeiend van de kansdichtheid, maar de implementatie van het algoritme is enorm moeilijk (Gu, 1993). In dit hoofdstuk ontwikkelen we een kansdichtheidsschatting gebruik makend van de LS-SVM regressie. De voorgestelde methode heeft bijzondere voordelen ten opzichte van de Parzen kernel schatters, wanneer schattingen zich in de staart bevinden.

Ontwerpen van regressie data

Veronderstel $z_1, ..., z_n$ is een willekeurig sample afkomstig van een continue waarschijnlijkheidsdichtheidsfunctie $f(z)$. Laat $A_k(z)$, $k = 1, ..., s$ het bin interval zijn, laat $h = (a_{k+1}(z) - a_k(z))$ de bin breedte zijn. Laat U_k het aantal sample punten zijn die in het bin interval A_k liggen. De histogram wordt dan

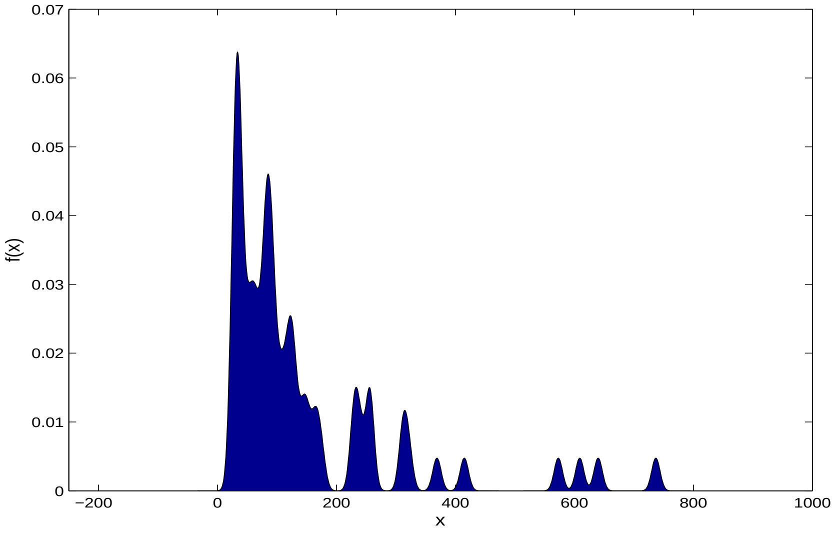

Figure 5: Kernel schatting voor zelfmoord data (Bandbreedte: h =10). De schatting is ruisgevoelig in de rechter staart.

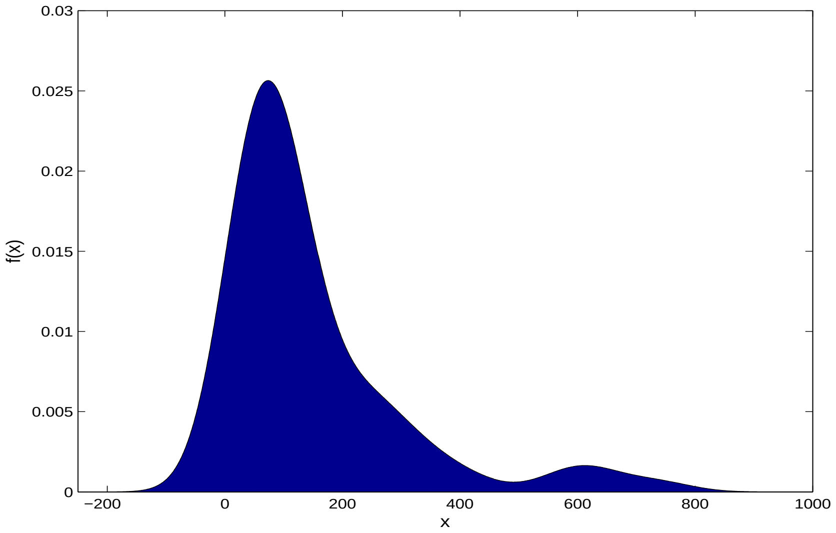

Figure 6: Kernel schatting voor zelfmoord data (Bandbreedte: h =80). De schatting is gladder dan in Figuur 5. De data waarden zijn positief, alhoewel de dichtheidsschatting de data als observaties in het interval $(-\infty, \infty)$ behandelt.

gedefineerd als

$$\hat{f}(z) = \frac{U_k}{nh} = \frac{1}{nh}\sum_{k=1}^{n} I_{[a_k,a_{k+1})}(z_k) \quad \text{voor } z \in A_k,$$

waarin U_k een binominale verdeling heeft, $U_k \backsim \text{Bin}(np_k(z), np_k(z)(1-p_k(z)))$ (Johnson *et al.*, 1997). De optimale keuze voor h vereist kennis van de onderliggende kansdichtheidsfunctie f, (Tukey, 1977) en (Scott, 1979). Praktisch is de smoothing parameter van de vorm $h^* = c3.5\hat{s}n^{-\frac{1}{3}}$ (Scott, 1979).

LS-SVM en dichtheidsschatting

Laat x_k, de onafhankelijke variabele, het center van A_k, $k = 1, ..., s$ zijn. Laat y_k, de afhankelijke variabele, de proportie van de data z_k liggend in het interval A_k gedeeld door de bin breedte h_n. Gebruik makend van Taylor's expansie, $f(\xi) = f(z) + (\xi - z) f'(z) + O\left(h^2\right)$, voor $\xi \in A_k$. Er kan worden berekend dat

$$E[y_k] = f(x_k) + O(h), \quad Var[y_k] = \frac{f(x_k)}{nh_n} + O\left(\frac{1}{n}\right).$$

De ruis inherent aan het histogram varieert in functie van zijn hoogte. Dus, kan men het kansdichtheidschattingsprobleem bekijken als een heteroscedastisch niet parametrisch regressie probleem, gedefinieerd als

$$y_k = m(x_k) + \varepsilon_k, \quad \varepsilon_k = e_k [\eta(m(x_k), x_k)]$$

waarin e_k onafhankelijk en identiek verdeeld zijn. De functie $\eta(g(x_k), x_k)$ drukt de mogelijke heteroscedasticiteit uit en $m : \mathbb{R}^d \to \mathbb{R}$ is een onbekende gladde functie welke we wensen te schatten. Noteer dat asymptotisch de heteroscedasticiteit geen enkele rol speelt aangezien de smoothing lokaal wordt en dusdanig de data in een klein venster bijna homoscedastisch wordt. De kensdichtheidschatter wordt gedefinieerd door

$$\hat{f}(x) = \mathcal{C}[\hat{m}_n(x)]_+,$$

waarin de constante $\mathcal{C}$ een normalisatie constant is zodanig dat $\hat{f}(x)$ integreerd naar 1 en $\hat{m}_n(x_k)$ is de LS-SVM regressie smoother.

Hoofdstuk 9: Robuustheid

In de voorgaande hoofdstukken werden basismethoden voor LS-SVM regressie modellen bestudeerd. Het gebruik van de kleinste kwadraten en gelijkheidsrestricties resulteren in een eenvoudige formulering, maar deze eenvoudige modellen hebben het nadeel dat ze niet robuust zijn. In dit hoofdstuk bespreken we het robuust maken van de LS-SVM modellen door gebruik te maken van methoden voorkomende in de robuuste statistiek. Gewogen LS-SVM versies worden geïntroduceerd om te kunnen omgaan met data waarin uitschieters in voorkomen (De Brabanter *et al.*, 2002). Om de robuustheid te meten van deze schatters maken we gebruik van de empirische invloedfuncties en maxbias curves.

Robuustheidmetingen

Empirische invloedfuncties

De meest belangrijke empirische versies van invloedfuncties zijn de sensitiviteitscurve (Tukey, 1970) en de Jackknife (Quenouille, 1956) en (Tukey, 1958).

De sensitiviteitscurve Er zijn twee versies, één met toevoeging en één met vervanging. In het geval van toevoeging van een observatie, start men met de sample $(x_1, ..., x_{n-1})$. Laat $T(F)$ een 'statistic' zijn en laat $T(\hat{F}_{n-1}) = T\left(x_1, ..., x_{n-1}\right)$ de schatter zijn. De verandering van de schatting wanneer de n^{de} observatie $x_n = x$ wordt toegevoegd is $T\left(x_1, ..., x_{n-1}, x\right) - T\left(x_1, ..., x_{n-1}\right)$. Men vermenigvuldigt de verandering met n en het resultaat is de sensitiviteitscurve.

Definitie 4 *(sensitiviteitscurve) Men bekomt de sensitiviteitscurve als men F vervangt door $\hat{F}_{n-1}$ en ϵ door $\frac{1}{n}$ in de invloedsfunctie:*

$$
\begin{aligned}
SC_{n-1}(x, T, \hat{F}_{n-1}) &= \frac{T\left[\left(\frac{n-1}{n}\right)\hat{F}_{n-1} + \frac{1}{n}\Delta_x\right] - T\left(\hat{F}_{n-1}\right)}{\frac{1}{n}} \\
&= (n-1)\,T\left(\hat{F}_{n-1}\right) + T(\Delta_x) - nT\left(\hat{F}_{n-1}\right) \\
&= n\left[T_n\left(x_1, ..., x_{n-1}, x\right) - T_{n-1}\left(x_1, ..., x_{n-1}\right)\right].
\end{aligned}
$$

Jackknife benadering Een andere aanpak voor het benaderen van de IF, maar enkel gebruik makend van de sample waarden $x_1, ..., x_n$, is de Jackknife.

Definitie 5 *(De Jackknife benadering). Men bekomt de sensitiviteitscurve als men F vervangt door $\hat{F}_n$ en -$\frac{1}{(n-1)}$ voor ϵ in de invloedsfunctie*

$$
\begin{aligned}
J_{IF}\left(x_i, T, F_n\right) &= \frac{T\left[\left(\frac{n}{n-1}\right)F_n - \frac{1}{n-1}\Delta_{x_i}\right] - T\left(F_n\right)}{-\frac{1}{n-1}} \\
&= -\left(n-1\right)\left[\left(T\left(F_n\right) - T(\Delta_{x_i})\right) - T\left(F_n\right)\right] \\
&= (n-1)\left[T_n\left(x_1, ..., x_n\right) - T_{n-1}\left(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\right)\right].
\end{aligned}
$$

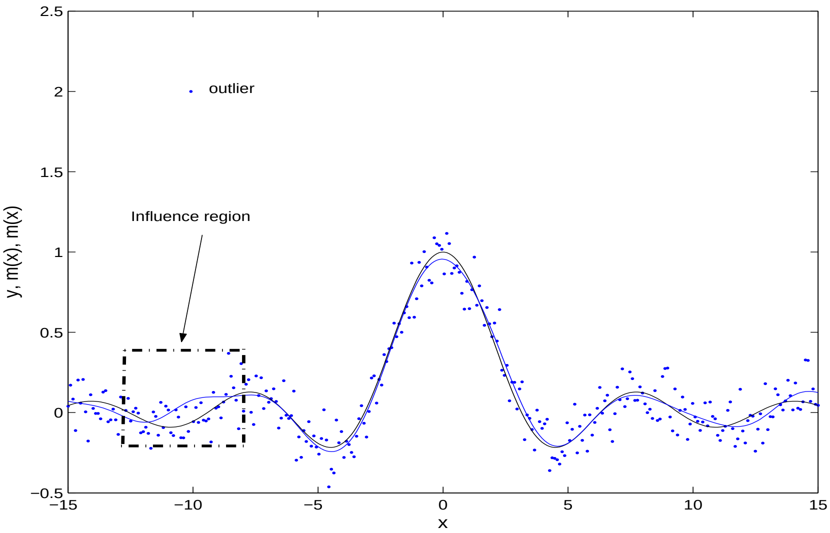

Figure 7: De effecten van een uitschieter (y-richting). Schatting van de sinc functie door LS-SVM regressie.

Residuals en uitschieters in Regressie

Kernel gebaseerde regressie

Herinner dat de LS-SVM regressie schatter wordt gegeven door

$$\hat{m}_n(x) = \sum_{k=1}^{n} \hat{\alpha}_k K\left(\frac{x - x_k}{h}\right) + \hat{b},$$

waarin $\hat{\alpha}_k \in \mathbb{R}$ en $b \in \mathbb{R}$. Figuur 7 laat de effecten zien van een uitschieter in de y-richting voor de LS-SVM regressie schatting.

De analyse van de robuustheidseigenschappen van kernel gebaseerde schatters worden in termen van de geschatte regressiefunctie uitgedrukt. Laat (x_i, y_i°) een uitschieter zijn (y-richting) en laat $\mathcal{A}$ de invloedsregio zijn. In dit geval heeft de uitschieter een kleine invloed op de schatter $\hat{m}_n(x_i)$ wanneer $(x_i, \hat{m}_n(x_i)) \in \mathcal{A}$ en heeft geen invloed als $(x_j, \hat{m}_n(x_j)) \notin \mathcal{A}$. De residuen van de LS-SVM regressie schatting zijn zeer nuttig als uitschieter detectors.

We tonen de sensitiviteitscurve (één met vervanging) voor $(x, \hat{m}_n(x)) \in \mathcal{A}$ en $(x_i, \hat{m}_n(x_i)) \notin \mathcal{A}$ in Figuur 8. Het meest belangrijkste aspect is dat de sensitiviteitscurve van de $\hat{m}_n(x)$ onbegrensd wordt ($x \in \mathcal{A}$) voor beide $y \to \infty$ en $y \to -\infty$, waarbij de $\hat{m}_n(x_i)$ constant blijft $(x_i \notin \mathcal{A})$.

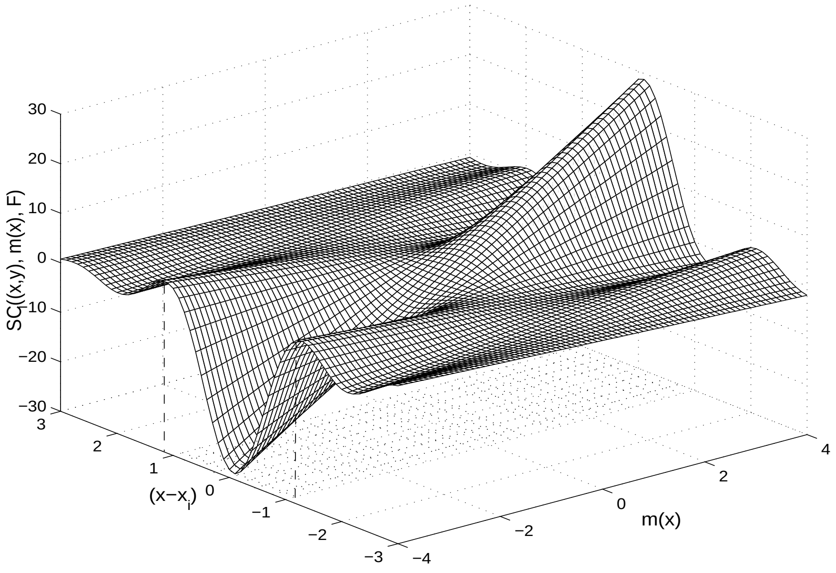

Figure 8: Empirische invloedsfunctie van $\hat{m}_n(x)$ als functie van $(x - x_i)$. De invloedscurve (in stippelijn) is onbegrensd in $\mathbb{R}$, waarbij in de andere regio's de invloedscurve begrensd blijft in $\mathbb{R}$.

Gewogen LS-SVM

Om een robuuste schatter gebaseerd op een voorgaande LS-SVM oplossing te bekomen, in een volgende stap, kan men de foutvariabelen $e_k = \alpha_k/\gamma$ wegen met wegingsfactoren v_k (Suykens *et al.*, 2002). Dit leidt tot volgend optimalisatie probleem:

$$\min_{w^\circ, b^\circ, e^\circ} \mathcal{J}\left(w^\circ, e^\circ\right) = \frac{1}{2} w^{\circ T} w^\circ + \frac{1}{2}\gamma \sum_{k=1}^{n} v_k e_k^{\circ 2}$$

zodat $y_k = w^{\circ T}\varphi\left(x_k\right) + b^\circ + e_k^\circ, \quad k = 1, ..., n$. De Lagrangiaan wordt geconstrueerd op een gelijkaardige manier als voordien. De ongekende variabelen voor het gewogen LS-SVM probleem worden aangeduid met het $\circ$ symbool. Vanuit de condities van optimaliteit en eliminatie van w°, e° bekomt met het Karush-Kuhn-Tucker systeem:

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + D_\gamma \end{array}\right] \left[\begin{array}{c} b^\circ \\ \hline \alpha^\circ \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right]$$

waarbij de diagonaal matrix D_γ wordt gegeven door $D_\gamma = \operatorname{diag}\left\{\frac{1}{\gamma v_1}, ..., \frac{1}{\gamma v_n}\right\}$. De keuze van de gewichten v_k worden bepaald gebaseerd op de foutvariabelen $e_k = \alpha_k/\gamma$ vanuit het (ongewogen) LS-SVM geval. Robuuste schatters worden dan verkregen (Rousseeuw en Leroy, 1986) bvb. door

$$v_k = \begin{cases} 1 & \text{als } \left|e_k/\hat{s}\right| \leq c_1 \\ \frac{c_2 - \left|e_k/\hat{s}\right|}{c_2 - c_1} & \text{als } \mathrm{c}_1 \leq \left|e_k/\hat{s}\right| \leq c_2 \\ 10^{-4} & \text{anders} \end{cases}$$

waar $\hat{s} = 1.483 \operatorname{MAD}\left(e_k\right)$ een robuuste schatting van de standaard afwijking van de LS-SVM foutvariabelen e_k is en MAD staat voor de median absolute deviation. De constanten c_1, c_2 worden typisch als $c_1 = 2.5$ en $c_2 = 3$ gekozen (Rousseeuw en Leroy, 1987). Gebruik makend van deze wegingen kan men corrigeren voor uitschieters (y-richting).

Ten eerste, tonen we de sensitiviteitscurve voor $(x, \hat{m}_n^\circ(x)) \in \mathcal{A}$ en $(x_i, \hat{m}_n^\circ(x_i)) \notin \mathcal{A}$ in Figuur 9. Het meest belangrijkste aspect is dat de sensitiviteitscurve voor $\hat{m}_n^\circ(x)$ onbegrensd wordt $(x \in \mathcal{A})$ voor beide $y \to \infty$ en $y \to -\infty$, waarbij de $\hat{m}_n^\circ(x_i)$ constant blijft $(x_i \notin \mathcal{A})$.

Ten tweede, berekenen we de maxbias curve voor beide LS-SVM en gewogen LS-SVM ten opzichte van een test punt. Gegeven 150 "good" observaties $\left\{\left(x_1, y_1\right), ..., \left(x_{150}, y_{150}\right)\right\}$ welke voldoen aan de relatie

$$y_k = m(x_k) + e_k, \quad k = 1, ..., 150,$$

waar $e_k \sim \mathcal{N}(0, 1^2)$. Laat $\mathcal{A}$ een bepaalde regio (43 data punten) zijn en laat x een test punt van die regio zijn (Figuur 10). Dan beginnen we met de data te contamineren in de regio $\mathcal{A}$. Bij elke stap verwijderen we één "good" punt in de regio $\mathcal{A}$ en vervangen we het door een "bad" punt $\left(x_i, y_i^\circ\right)$. We herhalen dit tot de schatting waardeloos wordt. Een maxbias curve wordt getoond in Figuur

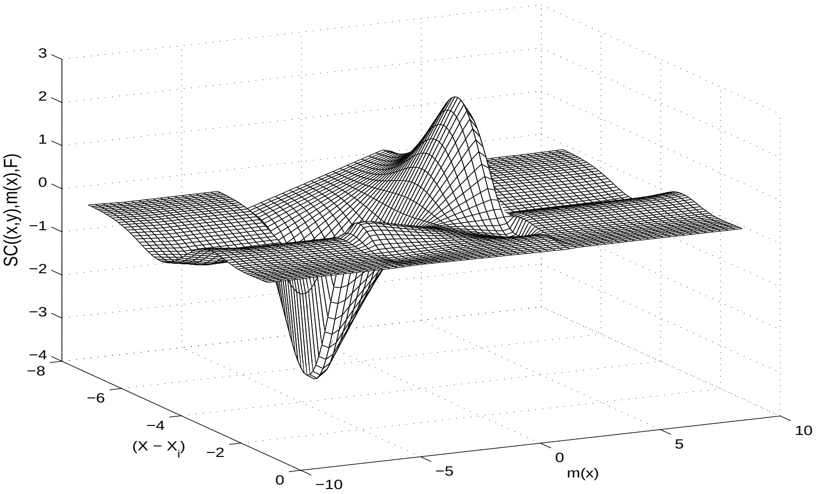

Figure 9: Empirische invloedsfunctie van $\hat{m}_n(x)$ als functie van $(x - x_i)$. De invloedscurve is begrensd in $\mathbb{R}$.

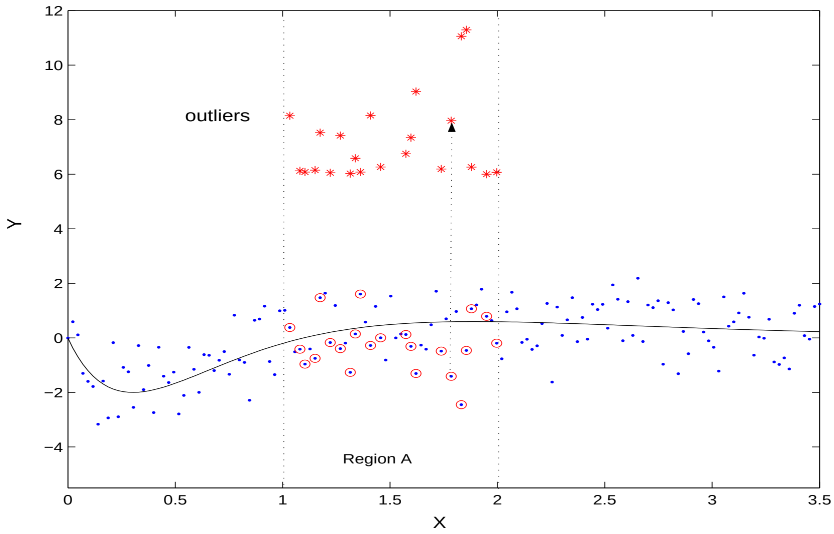

Figure 10: Gegeven 150 training data (Wahba, 1990). Beshouw de regio $\mathcal{A}$ tussen $x = 1$ en $x = 2$. In elke stap wordt de data in de regio $\mathcal{A}$ gecontamineerd door goede punten (aangeduid door "$\circ$") te vervangen door slechte punten (aangeduid door "$*$").

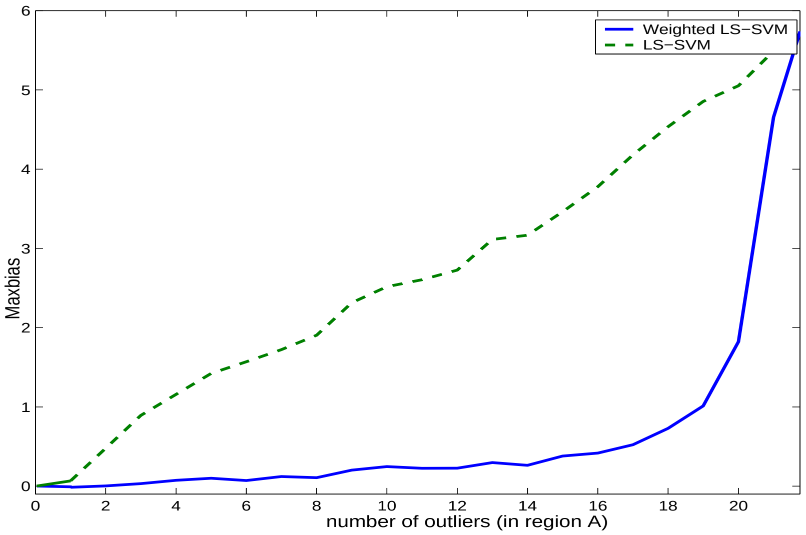

Figure 11: Maxbias curves voor de LS-SVM regressie schatter $\hat{m}_n\,(x)$ en de gewogen LS-SVM regressie schatter $\hat{m}_n^{\circ}\,(x)$.

11 waarbij de waarden van $\hat{m}_n(x)$ en $\hat{m}_n^{\circ}(x)$ getekend zijn als functie van het aantal uitschieters in de regio $\mathcal{A}$. De maxbias van $\hat{m}_n^{\circ}(x)$ stijgt zeer langzaam in functie van het aantal uitschieters in de regio $\mathcal{A}$ en blijft begrensd tot het rechtse breekpunt. Dit geldt niet voor $\hat{m}_n(x)$ met 0% als breekpunt.

Hoofdstuk 10: Data-gedreven Kostfuncties voor Regressie

Thans bestaat er een variatie van kostfuncties (bvb., least squares, least absolute deviations, M-estimators, generalized M-estimators, L-estimators, R-estimators, S-estimators, least trimmed sum of absolute deviations, least median of squares, least trimmed squares). Anderzijds brengt dit de data analyst in een moeilijke situatie.

Een idee voor deze situatie, voorgesteld in deze Sectie, is als volgt. Gegeven de data, de methode kan gesplitst worden in twee hoofddelen: (*i*) opbouwen van een robuust niet parametrisch regressie model en berekenen van de residuen, en (*ii*) de foutverdeling via robuuste bootstrap bekomen en bepalen van de kostfunctie (in een maximum likelihood omgeving).

Robuuste niet parametrische regressie modellen

De Nadaraya-Watson kernel schatter is niet robuust. Gebaseerd op het functionaal kader (Aït-Sahalia, 1995) zullen we de invloedsfunctie van de schatter bepalen om deze niet robuutsheid te verifiëren. Naar anologie met Hampel's invloedsfunctie (Hampel, 1994) en gebaseerd op het Generalized Delta theorem, de invloedsfunctie van de Nadaraya-Watson kernel schatter wordt gedefinieerd als

$$\begin{aligned} IF\left((x_k, y_k); T, F_{XY}\right) = & \frac{1}{f_X(x)h^d} \int yK\left(\frac{x-x_k}{h}\right) K\left(\frac{y-y_k}{h}\right) dy - \\ & \frac{1}{f_X(x)h^{d-1}} K\left(\frac{x-x_k}{h}\right) m\left(x\right) \\ = & \frac{K\left(\frac{x-x_k}{h}\right)}{f_X(x)h^{d-1}} \left(\frac{1}{h}\int yK\left(\frac{y-y_k}{h}\right) dy - m(x)\right). \end{aligned}$$

De invloedsfunctie is niet begrensd voor y in $\mathbb{R}$. Gebruik makende van dalende kernels, kernels zodat $K(u) \rightarrow 0$ als $u \rightarrow \infty$, de invloedsfunctie is begrensd voor x in $\mathbb{R}$. Gemeenschappelijke keuzes voor dalende kernels zijn: $K(u) = \max\left(\left(1-u^2\right), 0\right)$, $K(u) = \exp-\left(u^2\right)$ en $K(u) = \exp\left(-u\right)$.

Naar analogie (Boente en Fraiman, 1994), zijn we geïnteresseerd in de L-robuuste Nadaraya-Watson kernel schatter. De invloedsfunctie voor de schatter

$T_L\left(\hat{F}_{XY}\right)$ is gegeven door

$$\begin{aligned}
IF\left((x_k,y_k);T,F_{XY}\right) &= S_{F_{XY}}T.\left(\hat{F}_{XY}-F_{XY}\right)\\
&= \int \frac{\mathcal{J}(u)}{f\left(x,F^{-}_{Y|X}(u)\right)}u\left(\frac{1}{h^{d-1}}K\left(\frac{x-x_k}{h}\right)-f(x)\right)du\\
&-\frac{1}{h^d}K\left(\frac{x-x_k}{h}\right)\int\frac{\mathcal{J}(u)}{f\left(x,F^{-}_{Y|X}(u)\right)}K\left(\frac{F^{-}_{Y|X}(u)-y_k}{h}\right)du\\
&+\int\frac{\mathcal{J}(u)}{f\left(x,F^{-}_{Y|X}(u)\right)}\frac{\partial^{d-1}F}{\partial x^{(1)}...\partial x^{(d-1)}}\left(x,F^{-}_{Y|X}(u)\right)du
\end{aligned}$$

De invmoedsfunctie is begrensd voor y in $\mathbb{R}$ en $\hat{F}_{Y|X}$ is gedefinieerd als

$$\hat{F}_{Y|X}=\sum_{k=1}^{n}\frac{K\left(\frac{x-x_k}{h}\right)}{\sum_{l=1}^{n}K\left(\frac{x-x_l}{h}\right)}I_{[Y_k\leq y]},$$

waarin K de Gaussiaanse kernel is. De trimming parameter werd gelijk gesteld aan 2.5%.

Berekenen van de kostfunctie

Laat $f(y,m(x))$ het ruismodel zijn en laat $L(y,m(x))$ de kostfunctie zijn. In een maximum likelihood omgeving, voor symmetrische kansdichtheidsfunctie $f(y,m(x))$, een zekere kostfunctie is optimaal voor een gegeven ruismodel zodanig dat de kostfunctie gelijk is aan

$$L(y,m(x))=-\sum_{k=1}^{n}\log f\left(y_k-m(x_k)\right).$$

Nauwkeurigheid van de kostfunctie

De robuustificatie van de residual bootstrap is gebaseerd op een controle mechanisme in het herbemonsteringsplan, bestaande uit een verandering van de herbemonsteringswaarschijnlijkheden, door identificatie en weging van deze data punten die de functie schatter beïnvloeden (zie hoofdstuk robuust predictie intervallen).

Hoofdstuk 11: Robuuste leerparameter selectie

In dit hoofdstuk bestuderen we robuuste methoden voor het selecteren van leer parameters door cross-validatie en de final prediction error (FPE) criterium. Voor het robuust schatten van leerparameters worden robuuste locatieschatters zoals het getrimde gemiddelde gebruikt.

Robuuste V-fold Cross-validatie Score Functie

De algemene vorm van de V-fold cross-validatie score functie wordt gegeven door

$$CV_{V-fold}\left(\theta\right)=\sum_{v=1}^{V}\frac{m_v}{n}\int L\left(z,\hat{F}_{(n-m_v)}\left(z\right)\right)d\hat{F}_{m_v}\left(z\right).$$

Een nieuwe variant van de de klassieke cross-validatie score functie gebaseerd op het getrimde gemiddelde wordt geïntroduceerd. De robuuste V-fold cross-validatie score functie wordt dan geformuleerd als

$$CV_{V-fold}^{Robust}\left(\theta\right)=\sum_{v=1}^{V}\frac{m_v}{n}\int_{0}^{F^{-}(1-\beta_2)}L\left(z,F_{(n-m_v)}\left(z\right)\right)dF_{m_v}\left(z\right).$$

Laat $\hat{f}_{Robust}\left(x;\theta\right)$ een robuuste regressieschatting zijn, bijvoorbeeld de gewogen LS-SVM (Suykens *et al.*, 2002). De kleinste kwadraten robuuste V-fold cross-validatie schatting is gegeven door

$$\begin{aligned}CV_{V-fold}^{Robust}\left(\theta\right)=\sum_{v=1}^{V}\frac{m_v}{n}\sum_{k}^{m_v}\frac{1}{m_v-\lfloor m_v\beta_2\rfloor}\left(y_k-f_{Robust}^{(-m_v)}(x_k;\theta)\right)_{m_v(k)}^{2}\\ I_{[m_v(1),m_v(m_v-\lfloor m_v\beta_2\rfloor)]}((y_k-f_{Robust}^{(-m_v)}(x_k;\theta))^2),\end{aligned}$$

waar $(y_k-f_{Robust}^{(-m_v)}(x_k;\theta))_{m_v(k)}^{2}$ een geordende statistic is en de indicator functie $I_{[a,b]}(z)=1$ als $a<z<b$ en anders 0.

Robuuste Generalized Cross-validatie Score Functie

De GCV kan geschreven worden als

$$GCV(\theta)=\frac{1}{n}\sum_{k=1}^{n}L(\vartheta_k)=\frac{1}{n}\sum_{k=1}^{n}\vartheta_k^2,$$

waar ϑ_k gedefinieerd is als

$$\vartheta_k=\left(\frac{y_k-f^{*}(x_k;\theta)}{1-(1/\sum_k v_k)tr(S^{*})}\right),\quad k=1,\ldots,n$$

waar $f^*(x_k;\theta)$ de gewogen LS-SVM is, de weging van $f^*(x_k;\theta)$ overeenstemmend met $\{x_k, y_k\}$ wordt voorgesteld door v_k. Gebruik makend van de $(0,\beta_2)$ - getrimde gemiddelde, de robuuste GCV wordt gedefinieerd door

$$GCV_{robust}(\theta) = \frac{1}{n-\lfloor n\beta_2\rfloor}\sum_{k=1}^{n-\lfloor n\beta_2\rfloor} I_{[\vartheta_{n(1)},\vartheta_{n(n-\lfloor n\beta_2\rfloor)}]}(\vartheta^2)$$

waar $I_{[\cdot,\cdot]}(\cdot)$ een indicator functie is.

Robuust Final Prediction Error (FPE) criterium

De model parameters, θ worden zodanig bepaald dat de generalized Final Prediction Error (FPE) criterium gedefinieerd als

$$J_C(\theta) = \frac{1}{n}RSS + \left(1 + \frac{2\mathrm{tr}(S(\hat{\theta}))+2}{n-\mathrm{tr}(S(\hat{\theta}))-2}\right)\hat{\sigma}_e^2.$$

minimaal is. Een natuurlijke aanpak om het Final Prediction Error (FPE) criterium $J_C(\theta)$ te robuustifiëren is als volgt:

(i). Een robuuste schatter $\hat{m}_n^\circ(x,\theta)$ gebaseerd op (bvb. M-schatter (Huber, 1964) of gewogen LS-SVM (Suykens *et al.*, 2002)) vervangt de LS-SVM $\hat{m}_n(x,\theta)$.

(ii). De $RSS = \frac{1}{n}\sum_{k=1}^n (y_k - \hat{m}_n(x_k;\theta))^2$ vervangen door een robuuste tegenhanger RSS_{robust}. Laat $\xi = L(e)$ een functie van een willekeurige variabele e zijn. Een realisatie van de willekeurige variabele e wordt gegeven door $e_k = (y_k - \hat{m}_n(x_k;\theta))$, $k = 1,...,n$, en de $\frac{1}{n}RSS = J_1(\theta)$ kan geschreven worden als een locatie probleem

$$J_1(\theta) = \frac{1}{n}\sum_{k=1}^n L(e_k) = \frac{1}{n}\sum_{k=1}^n \xi_k,$$

waar $\xi_k = e_k^2$, $k = 1,...,n$. Gebruik makend van $(0,\beta_2)$ - getrimde gemiddelde, de robuuste $J_1(\theta)$ wordt gedefinieerd als

$$J_1^{robust}(\theta) = \frac{1}{n-\lfloor n\beta_2\rfloor}\sum_{k=1}^{n-\lfloor n\beta_2\rfloor}\xi_{n(k)},$$

waar $\xi_{n(1)},...,\xi_{n(n)}$, $e_k = (y_k - \hat{m}_n^\circ(x_k;\theta))$ en $\hat{m}_n^\circ(x_k,\theta)$ is een gewogen representatie van de functie schatter.

(iii). De variantie schatter $\hat{\sigma}_e^2$ wordt vervangen door de corresponderende rubuuste tegenhanger $\hat{\sigma}_{e,robust}^2$. Beschouw het NARX model (Ljung, 1987)

$$\hat{y}(t) = f(y(t-1),...,y(t-q),u(t-1),...,u(t-p)).$$

In praktijk, is het meestal het geval dat enkel de geordende data $y(k)$ met de discrete tijdsindex k, gekend is. De variantie schatter voorgesteld door (Gasser *et al.*, 1986) wordt gebruikt

$$\hat{\sigma}_e^2\left(y\left(t\right)\right)=\frac{1}{n-2}\sum_{t=2}^{n-1}\frac{(y(t-1)a+y(t+1)b-y(t))^2}{a^2+b^2+1}$$

waar $a=\frac{y(t+1)-y(t)}{y(t+1)-y(t-1)}$ en $b=\frac{y(t)-y(t-1)}{y(t+1)-y(t-1)}$. Laat $\zeta=L\left(\vartheta\right)$ een functie van een willekeurige variabele zijn, een realisatie van de willekeurige variabele ϑ wordt gegeven door

$$\vartheta_k=\frac{(y(t-1)a+y(t+1)b-y(t))}{\sqrt{a^2+b^2+1}}.$$

De variantie schatter kan nu geschreven worden als een gemiddelde van willekeurige samples $\vartheta_1^2,...,\vartheta_n^2$ (een locatie probleem):

$$\hat{\sigma}_e^2=\frac{1}{n-2}\sum_{k=2}^{n-1}\zeta_k,$$

waar $\zeta_k=\vartheta_k^2$, $k=2,...,n-1$. Gebruik makend van $(0,\beta_2)$ - getrimde gemiddelde, de robuuste $\hat{\sigma}_{e,robust}^2$ word gedefinieerd door

$$\hat{\sigma}_{e,robust}^2=\frac{1}{m-\lfloor m\beta_2\rfloor}\sum_{l=1}^{m-\lfloor m\beta_2\rfloor}\zeta_{n(l)},$$

waar $m=n-2$.

De robuuste FPE criterium wordt gegeven door

$$J_C(\theta)_{robust}=J_1(\theta)_{robust}+\left(1+\frac{2[\mathrm{tr}(S^*(v_k,\hat{\theta}))+1]}{n-\mathrm{tr}(S^*(v_k,\hat{\theta}))-2}\right)\hat{\sigma}_{e,robust}^2$$

waar de smoother matrix $S^*(v_k,\hat{\theta})$ nu gebaseerd is op de gewogen elementen v_k.

Hoofdstuk 12: Robuuste Predictie Intervallen

In dit hoofdstuk introduceren we robuuste predictie intervallen voor LS-SVM regressie gebaseerd op een robuuste externe bootstrap methoden.

Constructie van predictie intervallen

Waarschijnlijk één van de meest populaire methoden voor het construeren van predictiesets is gebruik te maken van pivots (Barnard, 1949, 1980), gedefinieerd als

Definitie 6 *Laat $X = (x_1, ...x_n)$ een willekeurige variabele met een onbekende samengestelde verdeling $F \in \mathcal{F}$, en laat $T(F)$ een reële waarde parameter zijn. Een willekeurige variabele $\mathcal{J}(X, T(F))$ is een pivot als de verdeling van $\mathcal{J}(X, T(F))$ onafhankelijk is van alle parameters.*

Hall (1992) bewees dat pivot methoden, voor het probleem van bootstrap predictie intervallen, moeten verkozen worden boven de niet-pivot methoden. Het belangrijkste probleem voor het construeren van predictie intervallen bij niet-parametrische regressie berust op het feit dat een consistente schatter van $m(x)$ noodzakelijk vertekend is (Neumann, 1995).

Robuuste Predictie Intervallen

Gewogen LS-SVM voor robuuste functie schatting

Smoother matrix voor predictie We vestigen de aandacht op de keuze van een RBF kernel $K(x_k, x_l; h)$ =exp$\left\{- \|x_k - x_l\|_2^2 / h^2\right\}$. In matrix vorm, laat $\theta = (h, \gamma)^T$ en voor alle nieuwe input data gedefinieerd als $\mathcal{D}_{x,test} = \{x : x_l^{test} \in \mathbb{R}^d, l = 1, ..., s\}$:

$$\begin{aligned}
\hat{m}_n\left(x^{test}; \theta\right) &= \Omega^{test}\hat{\alpha}^{train} + 1_n\hat{b}^{train} \\
&= \left[\Omega^{test}\left(Z^{-1} - Z^{-1}\frac{J_{nn}}{c}Z^{-1}\right) + \frac{J_{sn}}{c}Z^{-1}\right] y \\
&= S(x^{test}, x^{train}; \theta) y,
\end{aligned}$$

waar $c = 1_n^T\left(\Omega^{train} + \frac{1}{\gamma}I_n\right)^{-1} 1_n$, $Z = (\Omega^{train} + \frac{1}{\gamma}I_n)$, J_{nn} een vierkante matrix met alle elementen gelijk aan 1 is, J_{sn} is een $s \times n$ matrix met alle elementen gelijk aan 1, $y = (y_1, \ldots, y_n)^T$, $\hat{m}_n\left(x^{test}; \theta\right) = (\hat{m}_n(x_1^{test}; \theta), \ldots, \hat{m}_n(x_s^{test}; \theta))^T$, $\Omega_{k,l}^{test} = K\left(x_k^{train}, x_l^{test}\right)$ zijn de elementen van de $s \times n$ kernel matrix en $\Omega_{k,l}^{train} = K\left(x_k^{train}, x_l^{train}\right)$ zijn de elementen van de $n \times n$ kernel matrix.

Robuuste bootstrap

Gegeven een willekeurig sample $\{(x_1, y_1), ..., (x_n, y_n)\}$ met gemeenschappelijke verdeling F. Definieer voor elk paar (x_k, y_k) de residuen als $\hat{e}_k = y_k - \hat{m}_n(x_k)$. Gebaseerd op de residuen, gewichten werden als volgt gedefineerd

$$v_k = \vartheta\left(\frac{\hat{e}_k}{\hat{s}}\right)$$

waar $\vartheta(.)$ een functie is en $\hat{s}$ een robuuste schaalschatter is. Laat het bemonsteringsschema van de uniforme bootstrap voorgesteld worden door $p_{unif} = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$ en, laat $p = (p_1, ..., p_n)$ het herbemonteringsschema van de gewogen bootstrap zijn. Laat m het aantal data punten zijn met $(v_k \neq 1)$ en $\sum_{k=1}^{n} p_k = 1$. De hoeveelheid p_l, $l = 1, ...n - m$, wordt gegeven door

$$p_l = \frac{1}{n} + \frac{\sum_{i=1}^{m} \frac{1}{n}(1 - v_i)}{n - m}, \quad l = 1, ..., n - m \quad ; i = 1, ..., m$$

en de hoeveelheid p_j, $j = 1, ..., m$, wordt gegeven door

$$p_j = \left(1 - \sum_{l=1}^{n-m} p_l\right)\left(1 - \frac{v_j}{\sum_{j=1}^{m} v_l}\right), \quad j = 1, ..., m.$$

Bepalen van robuuste predictie intervallen

Gegeven een LS-SVM functie schatter $\hat{m}_{n,h}(x_0)$, waar x_0 een nieuw input data punt is, predictie intervallen worden geconstrueerd door gebruik te maken van een pivot statistic. Laat $\mathcal{J}(m(x_0), \hat{m}_{n,h}(x_0))$ een pivot statistic zijn, gedefineerd als

$$\mathcal{J}(m(x_0), \hat{m}_{n,h}(x_0)) = \frac{\hat{m}_{n,h}(x_0) - m(x_0) - B(x_0)}{(V(x_0))^{\frac{1}{2}}},$$

waar $B(x_0)$ de bias is en $V(x_0)$ de variantie is van de LS-SVM functie schatter $\hat{m}_{n,h}(x_0)$. De asymptotische pivot $\mathcal{J}(m(x_0), \hat{m}_{n,h}(x_0))$ kan niet worden gebruikt voor het bepalen van predictie intervallen omdat beiden $B(x_0)$ en $V(x_0)$ ongekend zijn. We beschouwen een alternatieve methode die de verdeling van de pivot schatten

$$\mathcal{T}(m(x_0), \hat{m}_{n,h}(x_0)) = \frac{\hat{m}_{n,h}(x_0)(x_0) - m(x_0)}{\left(\hat{V}(x_0)\right)^{\frac{1}{2}}}$$

door een externe bootstrap methode. Men benadert de verdeling van de pivotal statistics $\mathcal{T}(m(x_0), \hat{m}_{n,h}(x_0))$ door de corresponderende verdeling van de gebootstrapte statistics

$$\mathcal{V}(\hat{m}_{n,g}(x_0), \hat{m}^*_{n,h}(x_0)) = \frac{\hat{m}^*_{n,h}(x_0) - \hat{m}_{n,g}(x_0)}{\left(\hat{V}^*(x_0)\right)^{\frac{1}{2}}},$$

waar $*$ bootstrap tegenhangers zijn.

Een natuurlijke aanpak voor het robuustifiëren van de pivotal $\mathcal{V}(\hat{m}_{n,g}(x_0), \hat{m}^*_{n,h}\,(x_0))$ wordt bekomen door het vervangen van de LS-SVM functieschatter door een robuuste functieschatter (de gewogen LS-SVM) en het vervangen van de variantieschatter $\hat{V}^*\,(x_0)$ door zijn robuuste tegenhanger $\hat{V}^{*\diamond}\,(x_0)$

$$\mathcal{Z}(\hat{m}^{\diamond}_{n,g}(x_0), \hat{m}^{*\diamond}_{n,h}\,(x_0)) = \frac{\hat{m}^{*\diamond}_{n,h} - \hat{m}^{\diamond}_{n,g}(x_0)}{\left(\hat{V}^{*\diamond}\,(x_0)\right)^{\frac{1}{2}}}.$$

Gegeven nieuwe input data gedefinieerd als $\mathcal{D}_{x,test}$, robuuste predictie intervallen met $1-\alpha$ zijn gegeven door

$$I_{\mathcal{Z}} = \left[\hat{m}^{\diamond}_{n,h}(x_0) + \left(\hat{V}^{*\diamond}\,(x_0)\right)^{\frac{1}{2}} Q_{\alpha/2s},\ \hat{m}^{\diamond}_{n,h} + \left(\hat{V}^{*\diamond}\,(x_0)\right)^{\frac{1}{2}} Q_{(1-\alpha)/2s}\right],$$

waar Q_α de α-quantile van de bootstrap verdeling van de pivotal statistic $\mathcal{Z}(\hat{m}^{\diamond}_{n,g}(x_0), \hat{m}^{*\diamond}_{n,h}\,(x_0))$ is.

Hoofdstuk 13: Besluit en verder onderzoek

In this thesis, we have given an overview of basic techniques for non-parametric regression. In this chapter, we first give a chapter by chapter overview of our contributions and the conclusions. Topics for further research are pointed out in the second section of this chapter.

Besluit

De belangrijkste methode in deze thesis is de LS-SVM, een voorbeeld van het geregulariseerde modellerings paradigma. Wij hebben een nieuwe methode, componentwise LS-SVM geïntroduceerd, voor het schatten van modellen die uit een som van niet-lineaire componenten bestaan (Pelckmans et al, 2004).

We hebben het idee van de ruisvariantie schatter geintroduceerd door Rice (1984) veralgemeend voor multivariate data. We hebben de eigenschappen van de LS-SVM regressie bestudeerd bij afgezwakte Gauss-Markov condities. Kwadratische residuen plots werden voorgesteld om de heteroscedasticiteit te karakteriseren.

In LS-SVM's worden de oplossing gegeven door een lineair systeem (gelijkheidsbeperkingen) i.p.v. een QP probleem (ongelijkheidsbeperkingen). De SVM aanpak (Mukherjee en Vapnik, 1999) vereisen ongelijkheidsbeperkingen voor kansdichtheid schattingen. Een manier om deze ongelijkheidsbeperkingen te omzeilen, is het gebruik van regressie gebaseerde kansdichtheid schattingen. We hebben de LS-SVM regressie gebruikt voor kansdichtheid schatting.

Wij hebben een robuust kader voor LS-SVM regressie ontwikkeld. Het kader laat toe om een robuuste raming te verkrijgen die op de vorige LS-SVM regressie oplossing wordt gebaseerd, in een opeenvolgende stap. De gewichten worden bepaald welke gebaseerd zijn op de verdeling van de foutvariabelen (Suykens et al, 2002). Wij hebben aangetoond, gebaseerd op de empirische invloedsfunctie en de maxbias curve, dat de gewogen LS-SVM regressie een robuuste functieschatting is. Wij hebben hetzelfde principe gebruikt om een LS-SVM regressieraming in het heteroscedastisch geval te verkrijgen. Nochtans zijn de gewichten nu gebaseerd op een gladde raming van de foutvariantie.

Thans bestaat er een variatie van kostfuncties (bvb., least squares, least absolute deviations, M-estimators, generalized M-estimators, L-estimators, R-estimators, S-estimators, least trimmed sum of absolute deviations, least median of squares, least trimmed squares). Anderzijds brengt dit de data analyst in een moeilijke situatie. Een idee voor deze situatie, voorgesteld in deze thesis, is als volgt. Gegeven de data, de methode kan gesplitst worden in twee hoofddelen: (*i*) opbouwen van een robuust niet parametrisch regressie model en berekenen van de residuen, en (*ii*) de foutverdeling via robuuste bootstrap bekomen en bepalen van de kostfunctie (in een maximum likelihood omgeving).

Meest efficiente leeralgoritmen in neurale networken, support vector machines en kernel based methoden (Bishop, 1995; Cherkassky *et al.*, 1998; Vapnik, 1999; Hastie *et al.*, 2001; Suykens *et al.*, 2002b) vereisen de bepaling van extra leerparameters. In praktijk wordt de voorkeur gegeven aan data-gedreven methoden voor het selecteren van de leerparameters. Gebaseerd op locatie schatters (bvb. mediaan, M-schatters, L-schatters, R-schatters), hebben we de robuuste tegenhangers geintroduceerd van modelselectiecriteria (bvb. Cross-Validation, Final Prediction Error criterion).

Bij niet-parametrische regressie wordt de regressie vergelijking bepaald via de data. In dit geval kunnen de standaard inferentie procedures niet toegepast worden. Daarom hebben we robuuste voorspellingsintervallen ontwikkeld gebaseerd op robuuste bootstrap technieken.

Verder onderzoek

Verder onderzoek is nodig om de kernel methoden robuuster te makent. Mogelijk steunt dit verder onderzoek op twee peilers:

(1) bestaande robuuste methoden moeten worden bestudeerd voor het gebruik in kernel gebaseerde methoden.

(2) Robuustifieren van de cost functies.

Verder moeten de robuuste eigenschappen van deze methoden theoretisch worden bestudeerd. Hiervoor dienen de functionele benadering van Von Mises, welke gebruikt wordt in de parametrische statistiek en leidt tot de invloedsfunctie en het asymptotisch breekpunt, uitgebreid te worden tot niet-lineaire en niet-parametrische schatters.

Uiteindelijk zal men de ontwikkelde methoden toepassen op reele data sets. In het bijzonder denken we aan gegevens uit de chemometrie en de bio-informatica, omdat deze vaak vele variabelen bevatten en een klein aantal of juist een zeer groot aantal observaties.

Contents

Chapter 1

Introduction

In 1896, Pearson published his first rigorous treatment of correlation and regression in the Philosophical Transactions of the Royal Society of London. In this paper, Pearson credited Bravais, (1846) with ascertaining the initial mathematical formulae for correlation. In his four-volume biography of Galton, Pearson described the genesis of the discovery of the regression slope Pearson (1930). Subsequent efforts by Galton and Pearson brought about the more general techniques of multiple regression and the product-moment correlation coefficient. In fact, the main ideas of the parametric paradigm were developed between 1920 and 1960 (see Fischer, 1952). During this period, the method of maximum likelihood for estimating parameters was introduced. However, Tukey demonstrated that the statistical components of real-life problems cannot be described only by classical statistical distribution functions. In addition, James and Stein (1961) constructed a biased estimator of the mean of random normally distributed vectors that for any fixed number of observations is uniformly better than the estimate by the sample mean. These difficulties with the parametric paradigm and several discoveries (summarized in the next 4 items) made in the 1960s was a turning point in statistics and led to a new paradigm:

(*i*) The existence of high speed, inexpensive computing has made it easy to look at data in ways that were once impossible. Where once a data analyst was forced to make restrictive assumptions before beginning, the computer power now allows great freedom in deciding where an analyst should go. One area that has benefited greatly from this new freedom is that of nonparametric density and regression estimation, or what are generally called smoothing methods. Local regression modelling traces back to the 19th century. The work on local modelling starts in the 1950's with kernel methods introduced within the probability density estimation setting (Rosenblatt, 1956; Parzen, 1962) and within the regression setting (Nadaraya, 1964; Watson, 1964). The aim of nonparametric techniques is to relax the restrictive form of a regression function. It provides a useful tool for validating or suggesting a parametric form. However, nonparametric techniques have no intention of replacing parametric techniques. In fact, a combination of them can lead to discovering many interesting findings that are difficult to accomplish by any single method.

(*ii*) The theory of ill-posed problems. Tikhonov (1943), proving a lemma about an inverse operator, described the nature of well-posed problems and therefore discovered ways for regularization of ill-posed problems. 20 years later Phillips (1962), Ivanov (1962), Tikhonov (1963) and Lavrentev (1962) came to the same constructive regularization idea in a different form. The regularization technique in solving ill-posed problems was not only the first indication of the existence of non obvious solutions to the problems that are better than the obvious solutions, but it also gave an idea how to construct these non obvious solutions.

(*iii*) The generalization of the Glivenko-Cantelli-Kolmogorov theory was constructed in the late 1960s (Vapnik and Chervonenkis, 1968; 1971). The theory is based on new capacity concepts for a set of events (a set of indicator functions). Of particular importance is the VC dimension of the set of events which characterizes the variability of the set of events (indicator functions).

(*iv*) Capacity control makes it possible to take into account the amount of training data. This was discovered in the mid-1970s for the classification problem and by the beginning of 1980, all of the results obtained for sets of indicator functions were generalized for sets of real-valued functions (the problem of regression estimation). Capacity control in a structured set of functions became the main tool of the new paradigm.

A new direction was declared, the so-called "data analysis," where the goal was to perform inference from the data, rather than using purely statistical techniques. At the end of the 1960s, the theory of Empirical Risk Minimization (ERM) for the classification problem was constructed (Vapnik and Chervonenkis, 1974). Within 10 years, the theory of the ERM principle was generalized for sets of real-valued functions as well (Vapnik, 1979). It was found that both the necessary and sufficient conditions of consistency and the rate of convergence of the ERM principle depend on the capacity of the set of functions implemented by the learning machine. It was also found that distribution-free bounds on the rate of uniform convergence depend on the VC dimension (the capacity of the machine), the number of training errors and the number of observations. Therefore, to find the best guaranteed solution, one has to make a compromise between the accuracy of approximation of the training data and the capacity of the machine that one uses to minimize the number of errors. The idea of minimizing the test error by controlling two contradictory factors was formalized by introducing a new principle, the Structural Risk Minimization (SRM) principle. The support vector method realizes the SRM principle.

In 1963 the method of support vector machines (SVM) for constructing an optimal hyperplane in the separable case was under investigation (Vapnik and Lerner, 1963) and (Vapnik and Chervonenkis, 1964). In the mid-1960s, the expansion of the optimal hyperplane on support vectors and the constructing hyperplane in feature space using Mercer kernels were known. However, the combination of the two elements was done 30 years later in an article by Boser, Guyon and Vapnik (1992). After combining the support vector expansion with kernel representation of the inner product, the main idea of the SVM was realized. The extension of the SV technique for nonseparable cases was obtained in an article by Cortes and Vapnik (1995). The generalization of SVM for estimating real-valued functions was done by Vapnik (1995). Least Squares Support Vector Machines (LS-SVM) (Suykens and Vandewalle, 1999; Suykens *et al.*, 2002) are reformulations to standard SVMs which lead to solving linear KKT systems for classification tasks as well as regression. In (Suykens *et al.*, 2002) LS-SVMs have been proposed as a class of kernel machines with primal-dual formulations in relation to kernel Fisher Discriminant Analysis (FDA), Ridge Regression (RR), Partial Least Squares (PLS), Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), recurrent networks and control. The dual problems for the static regression without bias term are closely related to Gaussian processes (MacKay, 1992), regularization networks (Poggio and Girosi, 1990) and Kriging (Cressie, 1993), while LS-SVMs rather take an optimization approach with primal-dual formulations which have been exploited towards large scale problems and in developing robust versions.

Besides its long history, the problem of regression estimation is of increasing importance today. Stimulated by the growth of information technology in the past 20 years, there is a growing demand for procedures capable of automatically extracting information from massive high-dimensional databases. One of the fundamental approaches for dealing with this “data-mining problem” is regression estimation. Usually there is no prior knowledge about the data, leaving the analyst with no other choice but a nonparametric approach.

1.1 Practical applications

Scientific data must be clean and reliable. While this is the case in the majority of physical, chemical and engineering applications, *biomedical data* rarely possess such qualities. The very nature of biomedical objects is volatile and irregular, as are the results of biomedical assessments collected in large biomedical data sets. These data sets contain the results of tests which fluctuate with the patient’s state, and the long term trends are difficult to distinguish from the short term fluctuations, taking into account that these data sets rarely contain reliable longitudinal components. The other typical problem is the large number of incomplete records, for example, if certain tests are missing for some individuals, then deleting such records may essentially reduce the power of the ongoing calculations. Even mortality statistics, probably the most reliable type of biomedical data, are not free from error: while the date of death is usually known precisely, the date of birth can be biased.

There are three types of *economic/financial data*: Time series, cross-sectional and pooled data. Time series data may be collected at regular time intervals, such as daily (e.g. stock prices), monthly (e.g. the unemployment rate, the consumer price index). Although time series data are used in many econometric studies, they present some special problems. For example, most of the empirical work based on time series data assumes that the underlying time series is stationary. The problem of heteroscedasticity is more common in cross-sectional than in time series data. In cross-sectional data, one usually deals with members of a population at a given point in time, such as individual consumers, industries, country, etc. Moreover, these members may be of different sizes (e.g. small, medium, or high income). In time series data, on the other hand, the variables lend to be of similar order of magnitude because one generally collects the data for the same entity over a period of time. In the pooled data are elements of both time series and cross-sectional data.

Next we describe several applications in order to illustrate the practical relevance of regression estimation in both domains.

1.1.1 Biomedical data

Example 1 *(The Stanford heart transplant program). Various versions of data from the transplant study have been reported in a number of publications (Crowley and Hu, 1977), (Fan and Gijbels, 1994) and (Akritas, 1996). The sample consisted of 157 cardiac patients who where enrolled in the transplantation program between October 1967 and February 1980. Patients alive beyond February 1980 were considered to be censored (55 in total). One of the questions of interest was the effect of the age of a patient receiving a heart transplantation, on his survival time after transplantation.*

Example 2 *Prognostic factors in 1545 patients with stage I invasive epithelial ovarian carcinoma. A total of 1545 patients with invasive epithelial FIGO stage I ovarian cancer were included in this study. The patient records of 6 existing databases were retrospectively reanalysed according to predefined criteria. The Norwegian patient population consisted of 380 patients referred to the Norwegian Radium Hospital between January 1, 1980, and July 1, 1988. The 277 Danish patients were treated between September 1981 and September 1986 and registered in the Danish Ovarian Cancer Study Group (DACOVA) register. Canadian patients (n = 242) were treated at the Princess Margaret Hospital, Toronto, between April 1, 1971 and December 31, 1982. The patients from United Kingdom (n = 258) were referred to the Royal Mardsen NHS Trust London between January 1980 and December 1994. The 267 Swedish patients were referred to Radiumhemmet, Stockholm in the period 1974 –1986, and 121 Austrian patients were treated at the First Department of Obstetrics and Gynecology of the University of Vienna between December 1975 and June 1987. The aim of the study is to prove (statistically) the importance of degree of differentiation and cyst rupture in predicting relapse. For more details and results refer to "Case studies".*

Example 3 *(Endometria carcinome). The distinction between endometrial cancer patients with and without deep myometrial invasion is an important factor that guides clinical management. Between September 1994 and February 2000 we collected ultrasound and histopathological data from 97 women with endometrial carcinoma (called the training set) and divided them into two groups (group I: stage Ia and Ib – group II: stage Ic and higher). The transition between FIGO* *surgical stage Ib and Ic endometrial carcinoma is determined by the degree of myometrial invasion (less or more than 50%) and is important in determining the treatment schedule in many institutions. Accurate preoperative discrimination between group I and group II would allow to identify high-risk patients who might need pelvic and para-aortic lymphadenectomy. This might be important because in many countries patients who need lymphadenectomy are referred to a gynaecological oncologist while patients not needing lymphadenectomy are operated by the general gynaecologist or surgeon. For more details and results refer to "Case studies".*

1.1.2 Economic/financial data

Example 4 *(Boston housing data set). Harrison and Rubinfeld (1978) considered the effect of air pollution concentration on housing values in Boston. The data consisted of 506 samples of median home values in a neighborhood with attributes such as nitrogen oxide concentration, crime rate, average number of rooms, percentage of nonretail businesses, etc. A regression estimate was fitted tot the data and it was then used to determine the median value of homes as a function of air pollution measured by nitrogen oxide concentration. For more details refer tot Harrison and Rubinfeld (1978) an Breiman et al. (1984).*

Example 5 *(loan management). A bank is interested in predicting the return on a loan given to a customer. Available to the bank is the profile of the customer including his credit history, assets, profession, income, age, etc. The predicted return affects the decision as to whether to issue or refuse a loan, as well as the conditions of the loan. For more details refer to Krahl et al. (1998).*

Example 6 *(Interest rate data). Short-term risk-free interest rate play a fundamental role in financial markets. They are directly related to consumer spending, inflation and the overall economy. This data set concerns the yields of three-month, six-month and twelve-month treasury bills from the secondary market rates (on Fridays). The data consist of 2386 weekly observations. For more details refer to Anderson and Hund, 1997).*

1.2 Structure of the thesis

Part I addresses in a general manner the methods and techniques of nonparametric regression modelling. *Chapter 2* introduces the problem of regression function estimation and describes important properties of regression estimates. An overview of various paradigms to nonparametric regression is also provided. In *chapter 3* we explain a tool (identification of nonlinear structure in data) that uses a nonlinear mapping from the d-dimensional data space to an n_f-dimensional feature space. The feature space can have many more dimensions than the data space. This is essentially the approach of support vector machines. A simple type of mapping that is used in support vector machines is one defined by an inner product, called a kernel function. In addition, we discuss the least squares support vector machine and the fixed size least squares support vector machine. In *Chapter 4* we describe the methods (e.g., Final Prediction Error criterion, cross validation) for performance assessment. We begin the Chapter with a discussion of the bias-variance tradeoff. *Chapter 5* discuss the approach of resampling plans. The resampling methods replace theoretical derivations required in applying traditional methods in statistical analysis (nonparametric estimation of bias, variance and more general measures of error) by repeatedly resampling the original data and making inferences from the resamples. The most popular data-resampling methods used in statistical analysis are the bootstrap and jackknife.

In **Part II** we consider, the problem of high-dimensional data, the heteroscedastic case, the general problem of estimation of probability density functions, and the problem of modelling with censored data. *Chapter 6* discusses important characteristics of higher dimensional problems. For example, the asymptotic rate of convergence decreases with increasing input dimension when the characteristic of smoothness remains fixed (Vapnik, 1998). Therefore, one can guarantee good estimation of a high-dimensional function only if the function is extremely smooth. However, circumventing the curse of dimensionality can be done by impose additional assumptions (additivity) on the regression functions. There are several ways to approach estimation of additive models. The iterative backfitting algorithm (Hastie and Tibshirani, 1990) is used to fit an additive model. A great deal of effort has gone into developing estimators of the underlying regression function while the estimation of error variance has been relatively ignored. In *Chapter 7* we describe methods for error variance estimation. For example, an estimator based on the LS-SVM regression modelling and an estimator based on U-statistics. In *Chapter 8* a brief summary is given of the main methods for density estimation. We explain the connection between categorical data smoothing, nonparametric regression and density estimation. In addition we use the LS-SVM regression modelling for density estimation.

Part III provides an introduction and methods of robust statistics. Roughly speaking, robustness is concerned with the fact that many assumptions commonly made in statistics (e.g., normality) are at most approximations to reality. In *Chapter 9* we look at various measures of robustness (e.g., influence function, maxbias curve, breakdown point). The most important empirical versions of the influence function are illustrated with several examples. Based on Huber robust theory (Huber, 1964) we calculate a family of robust loss function for LS-SVM regression modelling. We discuss the weighted LS-SVM formulation. Empirical influence curves and maxbias curves are calculated for a comparison between LS-SVM regression and weighted LS-SVM. In addition we introduce a robust version of the fixed size LS-SVM. In *Chapter 10* we construct a data-driven loss function for regression. *Chapter 11* describes location estimators. Based on these location estimators, robust counterparts of model selection criteria (e.g., cross validation, generalized cross validation, Final Prediction Error criterion) are developed. *Chapter 12* illustrates inference for linear parametric models and nonparametric models. We discuss a robust method, based on external bootstrap technique, for obtaining robust prediction intervals.

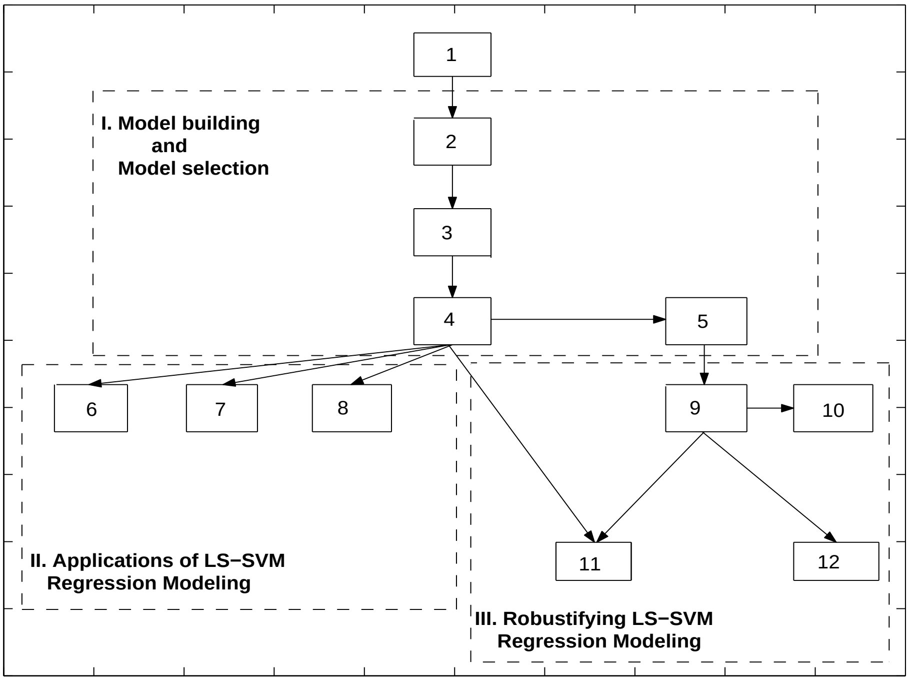

Figure 1.1: Structure of the thesis.

In *Chapter 13* the main results of this thesis are summarized and topics for further research are pointed out. The structure of the thesis is shown in Figure 1.1. It shows the sequence of chapters needed to be covered in order to understand a particular chapter.

1.3 Contributions

The key method in this thesis is least squares support vector machines, an example of the penalized modelling paradigm. The primal-dual representation is considered as an additional advantage. For large data sets it is advantageous if one solve the problem in the primal space (Suykens *et al.*, 2002). Figure 1.2 gives a general overview of our contributions. While the main goal of the first LS-SVM formulation was to solve an ordinary least squares problem (LS-SVM regression and LS-SVM classification), we have used the LS-SVM regression for density estimation.

Although local methods (kernel methods) focus directly on estimating the function at a point, they face problems in high dimensions. The asymptotic rate of convergence decreases with increasing input dimension when the characteristic of smoothness remains fixed (Vapnik, 1998). Therefore, one can guarantee

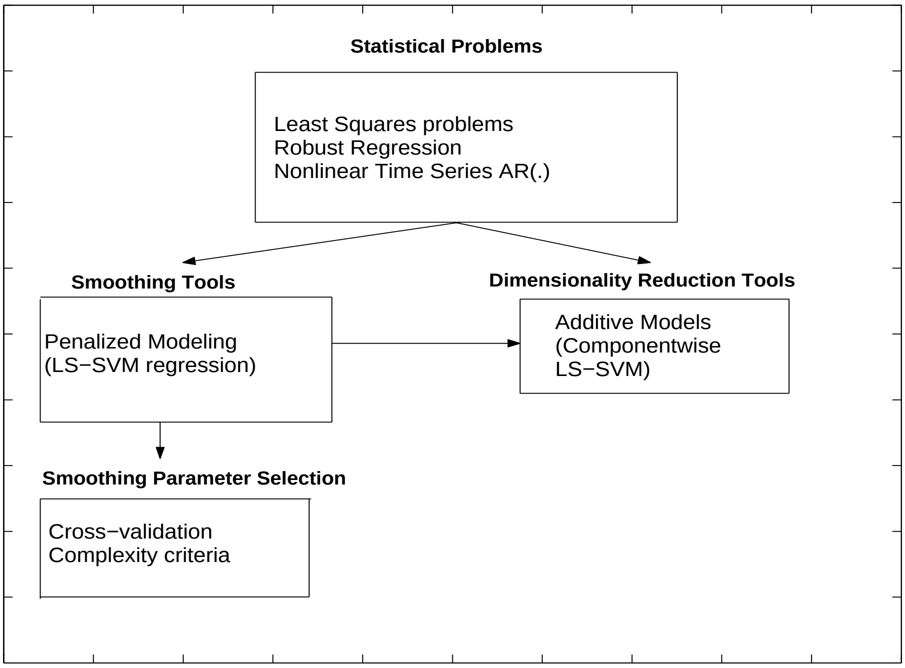

Figure 1.2: A general overview of our contributions.

good estimation of a high-dimensional function only if the function is extremely smooth. We have incorporated additional assumptions (the regression function is an additive function of its components) to overcome the curse of dimensionality. There are several ways to approach estimation of additive models. The iterative backfitting algorithm (Hastie and Tibshirani, 1990) was used to fit the additive model. We have introduced a new method, componentwise LS-SVM, for the estimation of additive models consisting of a sum of nonlinear components (Pelckmans *et al.*, 2004).

Model-free estimators of the noise variance are important for doing model selection and setting learning parameters. We have generalized the idea of the noise variance estimator introduced by Rice (1984) for multivariate data based on U-statistics and differogram models. While the method of least squares (under the Gauss-Markov conditions) enjoys well known properties, we have studied the properties of the LS-SVM regression when relaxing these conditions. It was recognized that outliers may have an unusually large influence on the resulting estimate. However, asymptotically the heteroscedasticity does not play any important role. Squared residual plots are proposed to assess heteroscedasticity in regression diagnostics.

A typical property of support vector machines (SVM) is that the solution is

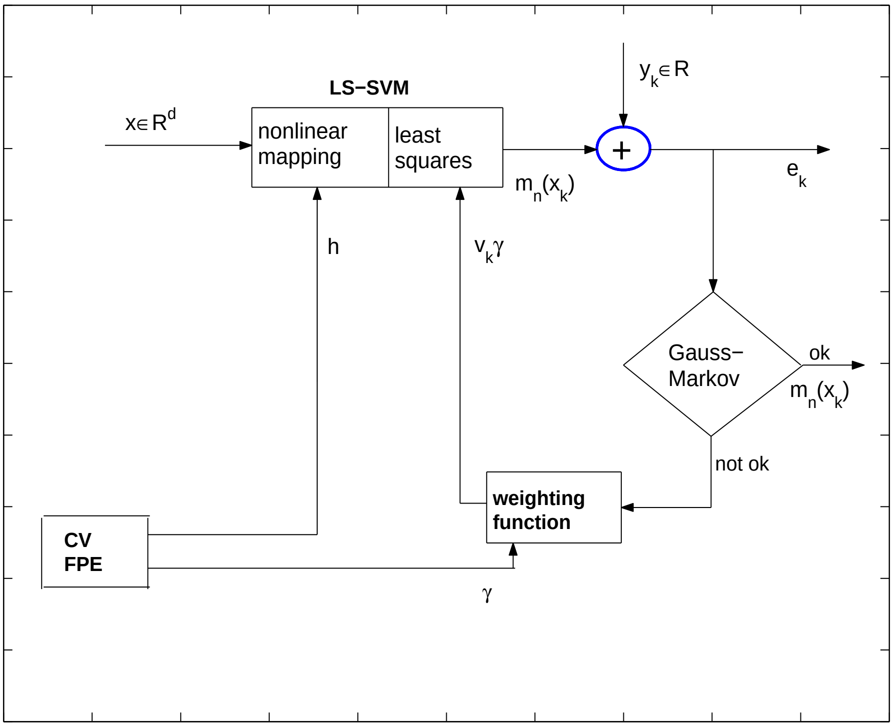

Figure 1.3: A robust framework for LS-SVM regression modelling.

characterized by a convex optimization problem, more specifically a quadratic programming (QP) problem. But in LS-SVM's the solution is given by a linear system (equality constraints) instead of a QP problem (inequality constraints). The SVM approach (Mukherjee and Vapnik, 1999) requires inequality constraints for density estimation. One way to circumvent these inequality constraints is to use the regression-based density estimation approach. We have used the LS-SVM regression for density estimation.

We have developed a robust framework (Figure 1.3) for LS-SVM regression. The framework allows to obtain a robust estimate based upon the previous LS-SVM regression solution, in a subsequent step. The weights are determined based upon the distribution of the error variables (Suykens *et al.*, 2002). We have shown, based on the empirical influence cure and the maxbias curve, that the weighted LS-SVM regression is a robust function estimation tool. We have used the same principle to obtain an LS-SVM regression estimate in the heteroscedastic case. However the weights are now based upon a smooth error variance estimate.

At present, there exists a variety of loss functions (e.g., least squares, least absolute deviations, M-estimators, generalized M-estimators, L-estimators, R-estimators, S-estimators, least trimmed sum of absolute deviations, least median of squares, least trimmed squares). On the other hand, this progress has put applied scientists into a difficult situation: if they need to fit their data with a regression function, they have trouble deciding which procedure to use. If more information was available, the estimation procedure could be chosen accordingly. We have proposed a method for such a situation: Given the data the method can basically be split up into two main parts: (*i*) constructing a robust nonparametric regression model and computing the residuals, and (*ii*) finding the distribution of the errors via a robust bootstrap and computing the loss function. Based on these distributions we can compute, in a maximum likelihood sense, the loss function.

Most efficient learning algorithms in neural networks, support vector machines and kernel based methods (Bishop, 1995; Cherkassky *et al.*, 1998; Vapnik, 1999; Hastie *et al.*, 2001; Suykens *et al.*, 2002b) require the tuning of some extra learning parameters, or *tuning parameters.* For practical use, it is often preferable to have a data-driven method to select the learning parameters. Based on location estimators (e.g., mean, median, M-estimators, L-estimators, R-estimators), we have introduced robust counterparts of model selection criteria (e.g., Cross-Validation, Final Prediction Error criterion).

Inference procedures for both linear and nonlinear parametric regression models in fact assume that the output variable follows a normal distribution. With nonparametric regression, the regression equation is determined from the data. In this case, we relax the normal assumption and standard inference procedures can not be strictly applicable. We have developed a robust approach for obtaining robust prediction intervals by using robust external bootstrapping methods.

Part I

MODEL BUILDING and MODEL SELECTION

Chapter 2

Model building

A model is just an abstraction of reality and it provides an approximation of some relatively more complex phenomenon. Models may be broadly classified as deterministic or probabilistic. Deterministic models abound in the sciences and engineering; examples include Ohm's law, the ideal gas law and the laws of thermodynamics. An important task in statistics is to find a probabilistic model, if any, that exist in a set of variables when at least one is random, being subject to random fluctuations and possibly measurement error. In regression problems typically one of the variables, often called the response, output or dependent variable, is of particular interest. The other variables, usually called explanatory, input, covariates, regressor or independent variables, are primarily used to explain the behavior of the response variable.

Consider the case of a quantitative output. Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denote a real valued random input vector, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ a real valued random output variable, with joint distribution F_{XY}. In regression analysis one is interested to find a measurable function $f : \mathcal{X} \to \mathcal{Y}$, such that $f(X)$ is a "good approximation of Y". Since X and Y are random vectors, $(f(X) - Y)$ is random as well. This requires a loss function $L(f(X), Y)$ for penalizing errors, and one can use the L_2 risk functional or mean squared error of f,

$$\mathcal{R}(f) = E\left[L(f(X), Y)\right] = E\left[(f(X) - Y)^2\right] \tag{2.1}$$

which is to be minimized. There are two reasons for considering the L_2 risk. First, this simplifies the mathematical treatment of the whole problem. Second, and more important, trying to minimize the L_2 risk leads naturally to estimates which can be computed rapidly. So, one is interested in a measurable function $m^* : \mathcal{X} \to \mathcal{Y}$, such that

$$m^*(X) = \arg \min_{f:\mathbb{R}^d \to \mathbb{R}} E\left[(f(X) - Y)^2\right]. \tag{2.2}$$

Such a function can be obtained explicitly as follows. Let

$$m(x) = E\left[Y \left| X = x\right.\right], \tag{2.3}$$

be the conditional expectation, also known as the regression function. Thus the best estimation of Y at any point $X = x$ is the conditional mean, when the best approximation is measured in mean squared error. Indeed, for an arbitrary $f : \mathbb{R}^d \to \mathbb{R}$, one has

$$\begin{aligned} E\left[(f(X) - Y)^2\right] &= E\left[(f(X) - m(X) + m(X) - Y)^2\right] \\ &= E\left[(f(X) - m(X))^2\right] + E\left[(m(X) - Y)^2\right] \end{aligned}$$

where we have used

$$\begin{aligned} &E\left[(f(X) - m(X))\,(m(X) - Y)\right] \\ &= E\left[E\left[(f(X) - m(X))\,(m(X) - Y)\,|X\right]\right] \\ &= E\left[(f(X) - m(X))\,E\left[(m(X) - Y)\,|X\right]\right] \\ &= E\left[(f(X) - m(X))\,(m(X) - m(X))\right] \\ &= 0 \end{aligned}$$

Hence,

$$E\left[(f(X) - Y)^2\right] = \int_{\mathbb{R}^d} (f(x) - m(X))^2\, dF(x) + E\left[(m(X) - Y)^2\right] \qquad (2.4)$$

The first term is always nonnegative and is zero if $f(x) = m(x)$. Therefore, $m^*(x) = m(x)$, i.e., the optimal approximation of Y by a function of X is given by $m(X)$.

In applications the distribution F_{XY} is usually unknown. Therefore it is impossible to estimate Y using $m(X)$. But it is often possible to observe data according to the distribution F_{XY} and to estimate the regression function from these data. In the regression function estimation problem one wants to use the data $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ in order to construct an estimate $\hat{m}_n : \mathcal{X} \to \mathcal{Y}$ of the regression function m. In general, estimates will not be equal to the regression function. Several distinct error criteria, which measure the difference between the regression function and an arbitrary estimate $\hat{m}_n$, are used: first, the pointwise error,

$$d(\hat{m}_n, m) = (\hat{m}_n(x) - m(x))^2 \quad \text{for some fixed } x \in \mathcal{X},$$

second, the supremum norm error,

$$d_\infty(\hat{m}_n, m) = \|\hat{m}_n - m\|_\infty = \sup_{x \in C} (\hat{m}_n(x) - m(x)) \quad \text{for some fixed set } C \subseteq \mathbb{R}^d,$$

and third, the L_1 (integrated absolute error) and L_2 (integrated squared error) error, respectively defined as

$$d_1(\hat{m}_n, m) = \int_C |\hat{m}_n(x) - m(x)|\, dx$$

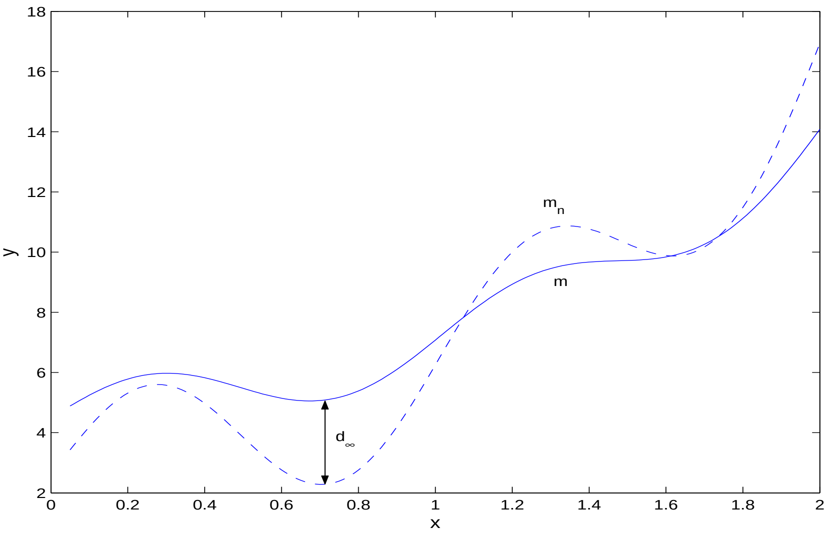

Figure 2.1: Greatest width $d_\infty (m, \hat{m}_n)$. The vertical line drawn at the place where the width is largest apart, the lenght of this arrow is $d_\infty (m, \hat{m}_n)$ and this gives the distance between m and $\hat{m}_n$ in $C[a,b]$.

$$d_2 (\hat{m}_n, m) = \int_C (\hat{m}_n(x) - m(x))^2 \, dx,$$

where the integration is with respect to the Lebesgue measure, C is a fixed subset of $\mathbb{R}^d$. Another measure of the difference in $\hat{m}_n$ and m over the full range of x is the the Hellinger distance

$$\left(\int_C \left(\hat{m}_n(x)^{\frac{1}{p}} - m(x)^{\frac{1}{p}} \right)^p dx \right)^{\frac{1}{p}}$$

and $p \geq 1$ is arbitrary. As an example, let $[a,b] \subset \mathbb{R}$ be a (nonempty) closed and bounded interval and let $\hat{m}_n$, $m \in C[a,b]$. Figure 2.1 illustrates the meaning of the distance function $d_\infty (\hat{m}_n, m) = \sup_{x \in [a,b]} |m(x) - \hat{m}_n(x)|$, $m, \hat{m}_n \in C[a,b]$ and Figure 2.2 explains $d_1 (\hat{m}_n, m) = \int\limits_a^b |m(x) - \hat{m}_n(x)| \, dt$ by the area between the two curves.

Recall that the main goal was to find a function f such that the L_2 risk $E\left[(f(X) - Y)^2 \right]$ is small. The minimal value of this L_2 risk is $E\left[(m(X) - Y)^2 \right]$, and is achieved by the regression function m. One can show, given the data $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$, that the L_2 risk $E\left[(\hat{m}_n(X) - Y)^2 \right]$ of an esti-

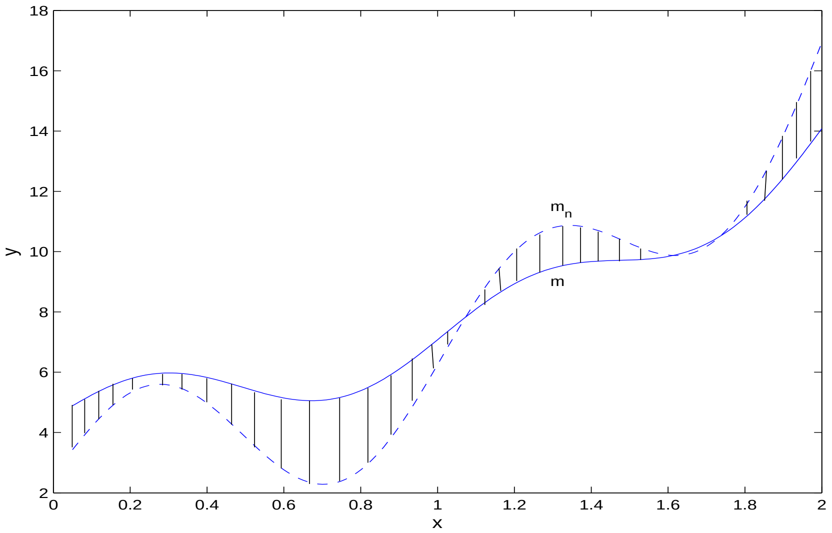

Figure 2.2: $d_1\left(m,\hat{m}_n\right)=$ area of the dashed portion.

mate $\hat{m}_n$ is close to the optimal value if and only if the L_2 error

$$\int_{\mathbb{R}^d}\left(\hat{m}_n(x)-m(x)\right)^2 dF(x) \tag{2.5}$$

is close to zero. Therefore we will use the L_2 error in order to measure the quality of an estimate.

2.1 Assumptions and restrictions

We know from Section 2.1 that the regression function m satisfies

$$E\left[\left(m\left(X\right)-Y\right)^2\right]=\inf_f E\left[\left(f\left(X\right)-Y\right)^2\right],$$

where the infimum is taken over all measurable functions $f:\mathcal{X}\to\mathcal{Y}$. This is impossible in the regression function estimation problem, because the risk functional to be optimized depends on F_{XY}.

Given empirical data $\mathcal{D}_n=\left\{\left(x_1,y_1\right),...,\left(x_n,y_n\right)\right\}$, minimizing the empirical L_2 risk functional defined as

$$\mathcal{R}_{emp}\left(f\right)=\frac{1}{n}\sum_{k=1}^{n}\left(f\left(x_k\right)-y_k\right)^2, \tag{2.6}$$

leads to infinitely many solutions: any function $\hat{f}_n$ passing through the training points $\mathcal{D}_n$ is a solution. In order to obtain useful results for finite n, one must "*restrict*" the solution to (2.6) to a smaller set of functions. Therefore one first chooses a suitable class of functions $\mathcal{F}$ and then selects a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, where $f \in \mathcal{F}_n$ which minimizes the empirical L_2 risk functional, i.e. one defines the estimate $\hat{m}_n$ by

$$\hat{m}_n \in \mathcal{F}_n \text{ and } \frac{1}{n}\sum_{k=1}^{n}\left(\hat{m}_n\left(x_k\right)-y_k\right)^2 = \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{k=1}^{n}\left(f\left(x_k\right)-y_k\right)^2. \qquad (2.7)$$

Recall that the empirical data $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ can be written as

$$y_k = m\left(x_k\right) + e_k. \qquad (2.8)$$

One "*assumes*" that the error term e in the model has zero mean and constant variance σ^2, that is, $E\left[e_k \left|X = x_k\right.\right] = 0$ and $E\left[e_k^2\right] = \sigma^2 < \infty$, and that the $\{e_k\}$ are uncorrelated random variables.

The design points $x_1, ..., x_n$ are random, typically far from beign uniformly distributed as in the fixed design, but we "*assume*" that $x_1, ...x_n$ could be measured accurately and $y_1, ..., y_n$ would not have the same accuracy. Otherwise, if the $x_1, ...x_n$ are measured with error the true values of $x_1, ...x_n$ are unknown and a random-regressor (errors-in-variables) model is needed.

For most systems, as represented in Figure 2.3, the input-output pairs (X, Y) will not have a deterministic relationship $y_k = m\left(x_k\right)$. Generally there will be other unmeasurable variables $q_1, ..., q_m$ that also contribute to Y, including measurement errors.

The additive error model (2.8) "*assumes*" that one can capture all these departures from a deterministic relationship via the error e.

The average L_2 risk functional $E\left[\left(\hat{m}_n - m\right)^2\right] = E\int\left(\hat{m}_n(x) - m(x)\right)^2 dF(x)$ is completely determined by the distribution of the pair (X, Y) and the regression function estimator $\hat{m}_n$. There exist universally consistent regression estimates (e.g., kernel estimates, neural networks estimates, radial basis function networks estimates,...), but it is impossible to obtain a nontrivial rate of convergence results without imposing strong "*restrictions*" on the distribution of (X, Y), by imposing some smoothness conditions on the regression function depending on a parameter τ (e.g., m is τ times continuously differentiable). For classes $\mathcal{F}_\tau$, where m is τ times continuously differentiable, the optimal rate of convergence will be $n^{-\frac{2\tau}{2\tau+d}}$.

The estimation of a regression function is very difficult if the dimension of the design variable X is large. The phenomenon is commonly referred to as the curse of dimensionality (Bellman, 1961). There are many manifestations of this problem, and we will examine a few here. This material is available in scattered references (see Kendall, 1961), for example.

Optimal rate of convergence for the estimation of a κ continuously differentiable regression function covergence to zero rather slowly if the dimension d of $X \in \mathbb{R}^d$ is large compared to κ. The only possibility is to impose additional "*assumptions*" on the regression function.

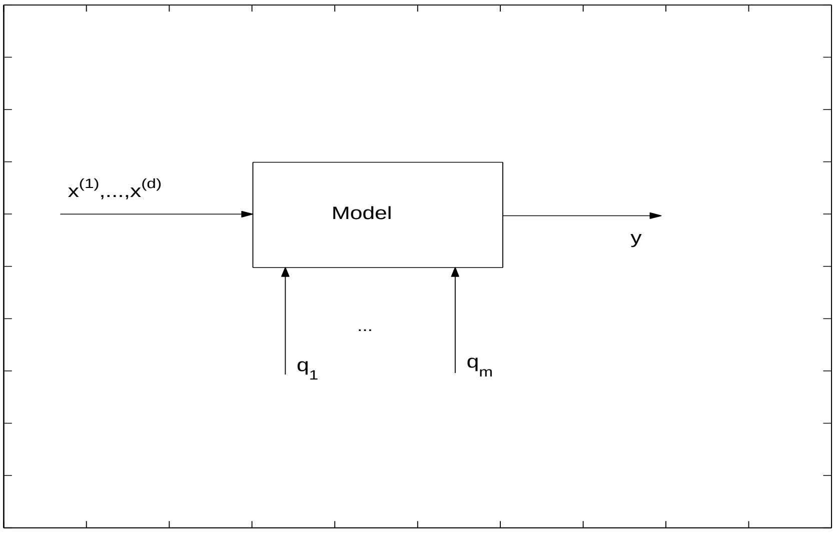

Figure 2.3: General model. Some of the variables $x^{(1)}, ..., x^{(d)}$ are controllable, whereas other variables $q^{(1)}, ..., q^{(p)}$ are uncontrollable.

2.2 Classes of restricted regression estimators

The description concerning the three paradigms in nonparametric regression is based on (Friedman, 1991) and (Györfi *et al.*, 2002). The kernel estimate is due to (Nadaraya, 1964; 1970) and (Watson, 1964). The principle of least squares (global modelling) is much older. For historical details we refer to (Hald, 1998; Farebrother, 1999; Stigler, 1999). The principle of penalized modelling, in particular, smoothing splines, goes back to (Whittaker, 1923; Schoenberg, 1964; Reinsch, 1967).

2.2.1 Parametric modelling

The classical approach for estimating a regression function is the parametric regression estimation. One assumes that the structure of the regression function is known and depends only on finitely many parameters. The linear regression model provide a flexible framework. However, linear regression models are not appropriate for all situations. There are many situations where the dependent variable and the independent variables are related through a known nonlinear function. It should be clear that in dealing with the linear and nonlinear regression models the normal distribution played a central role. There are a lot of practical situations where this assumption is not going to be even approximately satisfied. The generalized linear model was developed to allow us to fit regression models for dependent data ($y \in \mathbb{R}^+, y \in \mathbb{N}$ or $y \in \{0, 1\}$) that follows a general distribution called the exponential family.

As an example, consider the linear regression estimation. Let $\mathcal{F}$ be the class of linear combinations of the components of $x = \left(x^{(1)}, ..., x^{(d)}\right)^T \in \mathbb{R}^d$, i.e.,

$$\mathcal{F} = \left\{ m : m\left(x\right) = \beta_0 + \sum_{l=1}^{d} \beta_l x^{(l)},\ \beta_0, ..., \beta_d \in \mathbb{R} \right\}, \tag{2.9}$$

One use the data $\mathcal{D}_n = \left\{\left(x_1, y_1\right), ..., \left(x_n, y_n\right)\right\}$ to estimate the unknown parameters $\beta_0, ..., \beta_d \in \mathbb{R}$, e.g. by applying the principle of least squares:

$$\left(\hat{\beta}_0, ..., \hat{\beta}_d\right) = \underset{\beta_0,...,\beta_d \in \mathbb{R}}{\arg\min} \left[\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \beta_0 + \sum_{l=1}^{d} \beta_l x_k^{(l)} \right)^2 \right], \tag{2.10}$$

where $x_k^{(l)}$ denotes the lth component of $x_k \in \mathbb{R}^d$, $k = 1, ..., n$ and the estimate is defined as

$$\hat{m}_n(x) = \hat{\beta}_0 + \sum_{l=1}^{d} \hat{\beta}_l x^{(l)}. \tag{2.11}$$

However, parametric estimates have a drawback. Regardless of the data, a parametric estimate cannot approximate the regression function better than the best function with the assumed parametric structure. This inflexibility concerning the structure of the regression function is avoided by nonparametric regression estimates.

2.2.2 Local averaging and local modelling

An example of a local averaging estimate (kernel methods) is the Nadaraya-Watson kernel estimate. By definition

$$\begin{aligned} m\left(x\right) = E\left[Y \left| X = x\right.\right] &= \int y f_{Y|X}\left(y \left| x\right.\right) dy \\ &= \int y \frac{f_{XY}(x,y)}{f_X(x)} dy, \end{aligned} \tag{2.12}$$

where $f_X(x)$, $f_{XY}(x,y)$ and $f_{Y|X}\left(y \left| x\right.\right)$ are the marginal density of X, the joint density of X and Y, and the conditional density of Y given X, respectively. Let $K : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function called the kernel function and let $h > 0$ be the bandwidth or smoothing parameter. A product kernel estimate of $f_{XY}(x,y)$ is

$$\hat{f}_{XY}(x,y) = \frac{1}{nh_x h_y} \sum_{k=1}^{n} K_x\left(\frac{x - x_k}{h_x}\right) K_y\left(\frac{y - y_k}{h_y}\right),$$

while a kernel estimate of $f_X(x)$ is

$$\hat{f}_X(x) = \frac{1}{nh_x} \sum_{k=1}^{n} K_x\left(\frac{x - x_k}{h_x}\right).$$

Substituting into (2.12), and noting that $\int K_y(u)du = 1$, yields the Nadaraya-Watson kernel estimator

$$\hat{m}_n(x) = \sum_{k=1}^{n} \frac{K\left(\frac{x-x_k}{h}\right) y_k}{\sum_{l=1}^{n} K\left(\frac{x-x_l}{h}\right)}. \tag{2.13}$$

The Nadaraya-Watson kernel estimator is most natural for data using a random design, as in (2.12) (when the design is a random sample from some distribution having density f_X). If the design is not random, but rather a fixed set of ordered nonrandom numbers $x_1, ..., x_n$, the intuition of (2.12) is lost, and a different form of kernel estimator could be considered. An estimator intended for the fixed design case is the Gasser-Müller kernel estimator.

The second example of local averaging is the k-nearest neighbor estimate. For $X \in \mathbb{R}^d$, let $\left\{\left(x_{\pi(l)}, y_{\pi(l)}\right)\right\}_{l=1}^{n}$ be a permutation of $\{(x_l, y_l)\}_{l=1}^{n}$ such that $\left\|x - x_{\pi(1)}\right\| \leq ... \leq \left\|x - x_{\pi(n)}\right\|$. The k-nearest neighbor estimate is defined as

$$\hat{m}_n(x) = \frac{1}{k} \sum_{l=1}^{n} y_{\pi(l)}. \tag{2.14}$$

Here $\frac{1}{k}$ is a weight if x_l is among the k-nearest neighbors of x, and equals zero otherwise.

Basic calculus shows that the Nadaraya-Watson kernel estimator is the solution to a natural weighted least squares problem, being the minimizer $\hat{\beta}_0$ of

$$\sum_{l=1}^{n} (y_l - \beta_0)^2 K\left(\frac{x - x_l}{h}\right).$$

The Nadaraya-Watson kernel estimator corresponds to locally approximating $m(x)$ with a constant, weighting values of Y corresponding to x_l's closer to x more heavily. This suggests fitting higher order polynomials, since a local constant usually makes sense only over a small neighborhood. The most popular example of a local modelling estimate is the local polynomial kernel estimate. Let $g(x, \beta) : \mathbb{R} \to \mathbb{R}$ be a function depending on parameters $\beta \in \mathbb{R}^d$. For each $x \in \mathbb{R}$, choose values of these parameters by a local least squares criterion

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{l=1}^{n} (y_l - g(x_l, \beta))^2 K\left(\frac{x - x_l}{h}\right). \tag{2.15}$$

2.2.3 Global modelling

Least squares estimates are defined by minimizing the empirical L_2 risk functional over a general set of functions $\mathcal{F}_n$. This leads to a function which interpolates the data and hence is not a reasonable estimate. Thus one has to restrict the set of functions over which one minimizes the empirical L_2 risk functional. The global modelling estimate is defined as

$$\hat{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \left[\frac{1}{n} \sum_{k=1}^{n} (f(x_k) - y_k)^2\right] \tag{2.16}$$

and it minimizes the empirical L_2 risk functional.

As an example, consider neural networks (multilayer perceptrons with one hidden layer) regression function estimators. Given a training set $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ the parameters of the network are chosen to minimize the empirical L_2 risk functional. However, in order to obtain consistency, one restricts the range of some of the parameters. Thus one minimizes the empirical L_2 risk functional for the class of neural networks

$$\mathcal{F}_n = \left\{ \sum_{l=1}^{h} \beta_l g\left(w_l^T x + b_l\right) + \beta_0 : h \in \mathbb{N},\ w_l \in \mathbb{R}^d,\ b_l \in \mathbb{R},\ \sum_{l=0}^{h} |\beta_k| \leq a_n \right\}, \tag{2.17}$$

where $g : \mathbb{R} \to [0, 1]$ is a sigmoidal function (or often tanh) and $w_1, ..., w_h \in \mathbb{R}^d$, $b_1, ..., b_h \in \mathbb{R}$, $\beta_0, ..., \beta_h \in \mathbb{R}$ are the parameters that specify the network and obtain $\hat{m}_n \in \mathcal{F}_n$ satisfying

$$\frac{1}{n}\sum_{k=1}^{n} \left(\hat{m}_n\left(x_k\right) - y_k\right)^2 = \min_{f \in \mathcal{F}_n} \left[\frac{1}{n}\sum_{k=1}^{n} \left(f\left(x_k\right) - y_k\right)^2\right]. \tag{2.18}$$

If $h \to \infty$, $a_n \to \infty$ and $\frac{h a_n^4 \log\left(h a_n^2\right)}{n} \to 0$, then $E \int \left(\hat{m}_n\left(x\right) - m\left(x\right)\right)^2 dF(x) \to 0$ $(n \to \infty)$ for all distributions of (X, Y) with $E\left[Y^2\right] < \infty$ (Lugosi and Zeger, 1995).

2.2.4 Penalized modelling

Instead of restricting the class of functions, penalized least squares estimates explicitly adds a term to the functional to be minimized. Let $r \in \mathbb{N}$, $\lambda_n > 0$ and let the univariate penalized least squares estimate be defined as

$$\hat{m}_n(\cdot) = \arg \min_{f \in C^r(\mathbb{R})} \left[\frac{1}{n}\sum_{k=1}^{n} \left(f\left(x_k\right) - y_k\right)^2 + \lambda_n J_{n,v}\left(f\right)\right], \tag{2.19}$$

where $J_{n,v}\left(f\right) = \int \left(f^v\left(u\right)\right)^2 du$ and $C^v(\mathbb{R})$ is the set of all v times differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$. For the penalty term, $v = 2$, the minimum is achieved by a cubic spline with knots at the x_k's. In the multivariate case, the estimate is given by

$$\hat{m}_n(\cdot) = \arg \min_{f \in W^V(\mathbb{R}^d)} \left[\frac{1}{n}\sum_{k=1}^{n} \left(f\left(x_k\right) - y_k\right)^2 + \lambda_n \mathcal{J}_{n,v}\left(f\right)\right], \tag{2.20}$$

where $\mathcal{J}_{n,v}\left(f\right) = \sum_{i_1,...,i_v \in \{1,...,d\}} \int_{\mathbb{R}^d} \left(\frac{\partial^v f(u)}{\partial u_{i1} ... \partial u_{iv}}\right)^2 du$ is the penalty term for the roughness of the function $f : \mathbb{R}^d \to \mathbb{R}$, $\lambda_n > 0$ is the smoothing parameter of the estimate and $W^v(\mathbb{R}^d)$ is the Sobolev space consisting of all functions where weak derivatives of order v are contained in $L^2(\mathbb{R}^d)$ (Kohler and Krzyzak, 2001).

Chapter 3

Kernel Induced Feature Spaces and Support Vector Machines

In this Chapter we give a short overview on the formulations of standard Support Vector Machines as introduced by Vapnik. We discuss nonlinear function estimation by SVMs based on the Vapnik ϵ-insensitive loss function. Next we explain basic methods of Least Squares Support Vector Machines (LS-SVMs) for nonlinear function estimation. Finally we discus an approach in order to solve LS-SVM problems for function estimation in the case of large data sets. A technique of fixed size LS-SVM is presented.

3.1 Primal and dual representation

Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denote a real valued random input vector, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ a real valued random output variable and let $\Psi \subseteq \mathbb{R}^{n_f}$ denote a high-dimensional feature space. A key ingredient of the support vector machine is the following: It maps the random input vector into the high-dimensional feature space Ψ through some nonlinear mapping $\varphi : \mathcal{X} \rightarrow \Psi$. In this space, one consider the class of linear functions

$$\mathcal{F}_\Psi = \left\{ f : f(x) = w^T \varphi\left(x\right) + b : \varphi : \mathcal{X} \rightarrow \Psi,\ w \in \mathbb{R}^{n_f},\ b \in \mathbb{R} \right\}. \tag{3.1}$$

However, even if the linear function in the feature space (3.1) generalizes well and can theoretically be found, the problem of how to treat the high-dimensional feature space remains. Note that for constructing the linear function (3.1) in the feature space Ψ, one does not need to consider the feature space in explicit form. One only replaces the inner product in the feature space $\varphi\left(x_k\right)^T \varphi\left(x_l\right)$ with the corresponding kernel $K\left(x_k, x_l\right)$ satisfying Mercer's condition.

Theorem 7 *(Mercer, 1909). Let $K \in L^2(C)$, $g \in L^2(C)$ where C is a compact subset of $\mathbb{R}^d$ and $K(t, z)$ describes an inner product in some feature space. To guarantee that a continuous symmetric function K has an expansion*

$$K(t, z) = \sum_{k=1}^{\infty} a_k \phi_k(t) \phi_k(z)$$

with positive coefficients $a_k > 0$, (i.e., $K(t, z)$ describes an inner product in some feature space), it is necessary and sufficient that the condition

$$\int_C \int_C K(t, z) g(t) g(z) dt dz \geq 0$$

be valid for all $g \in L^2(C)$.

Assume $w = \sum_{k=1}^{n} \beta_k \varphi(x_k)$ and based on Mercer theorem, the class of linear functions in feature space (3.1) has the following equivalent representation in input space $\mathcal{X}$:

$$\mathcal{F}_{\mathcal{X}} = \left\{ f : f(x) = \sum_{k=1}^{n} \beta_k K(x, x_k) + b : \ b \in \mathbb{R}, \ \beta_k \in \mathbb{R} \right\}. \tag{3.2}$$

where x_k are vectors and $K(x, x_k)$ is a given function satisfying Mercer's condition.

3.2 LS-SVM regression

3.2.1 The unweighted case

Consider now the case where there is noise in the description of the functions. Given a training set defined as $\mathcal{D}_n = \{(x_k, y_k) : \ x_k \in \mathcal{X}, y_k \in \mathcal{Y}; \ k = 1, ..., n\}$ of size n drawn i.i.d. from an unknown distribution F_{XY} according to

$$y_k = m(x_k) + e_k, \qquad k = 1, ..., n, \tag{3.3}$$

where $e_k \in \mathbb{R}$ are assumed to be i.i.d. random errors with $E[e_k | X = x_k] = 0$, $Var[e_k] = \sigma^2 < \infty$, $m(x) \in \mathcal{F}_\Psi$ is an unknown real-valued smooth function and $E[y_k | x = x_k] = m(x_k)$. Our goal is to find the parameters w and b (primal space) that minimize the empirical risk functional

$$\mathcal{R}_{emp}(w, b) = \frac{1}{n} \sum_{k=1}^{n} \left(\left(w^T \varphi(x_k) + b \right) - y_k \right)^2 \tag{3.4}$$

under constraint $\|w\|_2 \leq a$, $a \in \mathbb{R}_+$. One can reduce the optimization problem of finding the vector w and $b \in \mathbb{R}$ to solve the following optimization problem

$$\min_{w,b,e} \mathcal{J}(w, e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{k=1}^{n} e_k^2, \tag{3.5}$$

such that

$$y_k = w^T \varphi\left(x_k\right) + b + e_k,\ \ k = 1, ..., n$$

Note that the cost function $\mathcal{J}$ consists of a RSS fitting error and a regularization term, which is also a standard procedure for the training of MLP's and is related to ridge regression (Golub and Van Loan, 1989). The relative importance of these terms is determined by the positive real constant γ. In the case of noisy data one avoids overfitting by taking a smaller γ value. SVM problem formulations of this form have been investigated independently in (Saunders *et al.*, 1998) (without bias term) and (Suykens and Vandewalle, 1999).

To solve the optimization problem (in the dual space) one defines the Lagrangian functional

$$\mathcal{L}(w, b, e; \alpha) = \mathcal{J}(w, e) - \sum_{k=1}^{n} \alpha_k \left(w^T \varphi\left(x_k\right) + b + e_k - y_k\right), \tag{3.6}$$

with Lagrangian multipliers $\alpha_k \in \mathbb{R}$ (called support values). The conditions for optimality are given by

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum\limits_{k=1}^{n} \alpha_k \varphi\left(x_k\right) \\ \dfrac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum\limits_{k=1}^{n} \alpha_k = 0 \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, \qquad\qquad\qquad k = 1, ..., n \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \rightarrow w^T \varphi\left(x_k\right) + b + e_k = y_k,\ \ k = 1, ..., n \end{cases} \tag{3.7}$$

After elimination of w, e one obtains the solution

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\gamma} I_n \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \tag{3.8}$$

with $y = \left(y_1, ..., y_n\right)^T$, $1_n = (1, ..., 1)^T$, $\alpha = \left(\alpha_1; ...; \alpha_n\right)^T$ and $\Omega_{kl} = \varphi\left(x_k\right)^T \varphi\left(x_l\right)$ for k, $l = 1, ..., n$. According to Mercer's theorem, the resulting LS-SVM model for function estimation becomes

$$\hat{m}_n\left(x\right) = \sum_{k=1}^{n} \hat{\alpha}_k K\left(x, x_k\right) + \hat{b}, \tag{3.9}$$

where $\hat{\alpha}$, $\hat{b}$ are the solution to (3.8)

$$\hat{b} = \frac{1_n^T \left(\Omega + \dfrac{1}{\gamma} I_n\right)^{-1} y}{1_n^T \left(\Omega + \dfrac{1}{\gamma} I_n\right)^{-1} 1_n} \tag{3.10}$$

$$\hat{\alpha} = \left(\Omega + \frac{1}{\gamma} I_n\right)^{-1} \left(y - 1_n \hat{b}\right). \tag{3.11}$$

3.2.2 Smoother matrix

In this thesis, we focus on the choice of an RBF kernel $K(x_k, x_l; h) = \exp\left\{-\|x_k - x_l\|_2^2 / h^2\right\}$. Let $\theta = (h, \gamma)^T$ and for all training data $\{x_k, y_k\}_{k=1}^n$, one has

$$\begin{aligned} \hat{m}_n &= \Omega\alpha + 1_n b \\ &= \left[\Omega\left(Z^{-1} - Z^{-1}\frac{J_n}{c}Z^{-1}\right) + \frac{J_n}{c}Z^{-1}\right] y \\ &= S(\theta)y, \end{aligned} \tag{3.12}$$

where $c = 1_n^T\left(\Omega + \frac{1}{\gamma}I_n\right)^{-1} 1_n$, $Z = (\Omega + \frac{1}{\gamma}I_n)$, J_n is a square matrix with all elements equal to 1, $y = (y_1, \ldots, y_n)^T$ and $\hat{m}_n = (\hat{m}_n(x_1), \ldots, \hat{m}_n(x_n))^T$. The LS-SVM for regression corresponds to the case with $\hat{f}_\theta$ defined by (3.12) and

$$S(\theta) = \Omega\left(Z^{-1} - Z^{-1}\frac{J_N}{c}Z^{-1}\right) + \frac{J_n}{c}Z^{-1}. \tag{3.13}$$

Therefore, the LS-SVM for regression is an example of a linear smoother. This is because the estimated function in (3.12) is a linear combination of the y. The linear operator $S(\theta)$ is known as the smoother matrix. Linear operators are familiar in linear regression (least squares fitting), where the fitted values $\hat{y}$ can be expressed as linear combinations of the output (dependent) variable y with the elements of the matrix that involves only the observations on the input (independent) variable u. Here the linear operator $H(u) = u(u^Tu)^{-1}u^T$ is a projection operator also known as the hat matrix in statistics. There are some important similarities and differences between the hat matrix $H(u)$ and the smoother matrix $S(\theta)$. Both matrices are symmetric, positive semidefinite and the hat matrix is idempotent ($S^2 = S$) while the smoother matrix $S(\theta)^T S(\theta) \leqslant S(\theta)$, (meaning that $S^TS - S \leqslant 0$ is negative semidefinite). This is a consequence of the shrinking nature of $S(\theta)$. The trace of $H(u)$ gives the dimension of the projection space, which is also the number of parameters involved in the fit. By analogy one defines the effective degrees of freedom of the LS-SVM for regression (effective number of parameters) to be

$$d_{eff}(\theta) = \operatorname{tr}\left[S(\theta)\right]. \tag{3.14}$$

Another important property of the smoother matrix, based on an RBF kernel, is that the $\operatorname{tr}[S(\theta)] < n$, except in the case ($h \to 0$, $\gamma \to \infty$) where $\operatorname{tr}[S(\theta)] \to n$.

3.3 Support Vector Machines

Given training data $(x_1, y_1), ..., (x_n, y_n)$, to find an approximation of functions of the form $f(x) = \sum_{k=1}^n \beta_k K(x, x_k) + b$ that are equivalent (in feature space) to the function $f(x) = w^T \varphi(x)$, one minimize the empirical risk functional in feature space

$$\mathcal{R}_{emp}(w,b) = \frac{1}{n}\sum_{k=1}^{n} \left| \left(w^T\varphi(x_k)+b\right) - y_k \right|_\varepsilon \tag{3.15}$$

subject to the constraint $\|w\|_2 \leq a_n$, where $|\cdot|_\varepsilon$is the Vapnik ε-insensitive loss function defined as

$$|f(x)-y|_\varepsilon = \begin{cases} 0, & \text{if } |f(x)-y| \leq \varepsilon, \\ |f(x)-y| - \varepsilon, & \text{otherwise.} \end{cases} \tag{3.16}$$

This optimization problem is equivalent to the problem of finding w, b that minimizes the quantity defined by slack variables ξ_k, ξ_k^*, $k = 1, ..., n$

$$\begin{array}{lll} [\mathrm{P}] & \min\limits_{w,b,\xi,\xi^*} J_{\mathrm{P}}(w,\xi,\xi^*) & = \frac{1}{2}w^Tw + c\sum_{k=1}^{n}(\xi_k+\xi_k^*) \\ & & y_k - w^T\varphi(x_k) - b \leq \varepsilon + \xi_k, \quad k = 1,...,n \\ & \text{such that} & w^T\varphi(x_k) + b - y_k \leq \varepsilon + \xi_k^*, \quad k = 1,...,n \\ & & \xi_k, \xi_k^* \geq 0, \qquad k = 1,...,n. \end{array} \tag{3.17}$$

After constructing the Lagrangian functional and conditions of optimality one obtains the following dual problem

$$\begin{aligned} [\mathrm{D}] \min_{\alpha,\alpha^*} J_{\mathrm{D}}(\alpha,\alpha^*) = & -\frac{1}{2}\sum_{k,l=1}^{n}(\alpha_k-\alpha_k^*)(\alpha_l-\alpha_l^*)K(x_k,x_l) \\ & -\frac{1}{2}\sum_{k,l=1}^{n}(\alpha_k-\alpha_k^*)(\alpha_l-\alpha_l^*)K(x_k,x_l) \\ & -\varepsilon\sum_{k=1}^{N}(\alpha_k+\alpha_k^*) + \sum_{k=1}^{N} y_k(\alpha_k-\alpha_k^*) \\ \text{such that} \quad & \sum_{k,l=1}^{n}(\alpha_k-\alpha_k^*) = 0, \qquad \alpha_k,\alpha_k^* \in [0,c] \end{aligned} \tag{3.18}$$

where $\beta_k = (\alpha_k - \alpha_k^*)$, $k = 1, ..., n$.

3.4 Fixed-size LS-SVM

3.4.1 Estimation in primal weight space

For large data sets it is often advantageous if one could solve the problem in the primal space (Suykens *et al.*, 2002). However, one would then need an explicit expression for some nonlinear mapping $\varphi : \mathcal{X} \rightarrow \Psi$. Let $x_k \in \mathbb{R}^d$, $k = 1, ..., n$ be a random sample from an unknown distribution $F_X(x)$. Let C be a compact subset of $\mathbb{R}^d$, let $V = L^2(C)$ and let $M(V,V)$ be a class of linear operators from

V into V. Consider the eigenfunction expansion of a kernel function

$$K\left(x,t\right)=\sum_{i=1}^{s}\lambda_{i}\phi_{i}\left(x\right)\phi_{i}\left(t\right), \tag{3.19}$$

where $s\leq\infty$, $K\left(x,u\right)\in V$, $\lambda_{i}\in\mathbb{C}$ and $\phi_{i}\in V$ are respectively the eigenvalues and the eigenfunctions, defined by Fredholm integral equation of the first kind

$$\begin{aligned}\left(T\phi_{i}\right)\left(t\right)&=\int_{C}K\left(x,t\right)\phi_{i}\left(x\right)dF_{X}\left(x\right)\\&=\lambda_{i}\phi_{i}\left(t\right),\end{aligned} \tag{3.20}$$

where $T\in M\left(V,V\right)$.

One can discretize (3.20) on a finite set of evaluation points $\{x_{1},...,x_{n}\}\in C$ with associated weights $v_{k}\in\mathbb{R}$, $k=1,...,n$. Define a quadrature method Q_{n}, $n\in\mathbb{N}$

$$Q_{n}=\sum_{k=1}^{n}v_{k}\psi\left(x_{k}\right). \tag{3.21}$$

Let $v_{k}=\frac{1}{n}$, $k=1,...,n$, then the Nyström method approximates the integral by means of Q_{n} and determines an approximation ϕ_{i} by

$$\lambda_{i}\phi_{i}\left(t\right)\approx\frac{1}{n}\sum_{k=1}^{n}K\left(x_{k},t\right)\phi_{i}\left(x_{k}\right),\ \forall t\in C. \tag{3.22}$$

Let $t=x_{j}$, in matrix notation one obtains then

$$\Omega U_{n\times n}=U_{n\times n}\Lambda_{n\times n}, \tag{3.23}$$

where $\Omega_{kj}=K\left(x_{k},x_{j}\right)$ are the elements of the kernel matrix, $U_{n\times n}=\left(u_{1},...,u_{n}\right)$ is a $n\times n$ matrix of eigenvectors of Ω and $\Lambda_{n\times n}$ is a $n\times n$ diagonal matrix of nonnegative eigenvalues in a decreasing order. Expression (3.22) delivers direct approximations of the eigenvalues and eigenfunctions for the $x_{k}\in\mathbb{R}^{d}$, $k=1,...,n$ points

$$\phi_{i}\left(x_{j}\right)\approx\sqrt{n}u_{li,n} \tag{3.24}$$

and

$$\lambda_{i}\approx\frac{1}{n}\lambda_{i,n}, \tag{3.25}$$

where $\lambda_{i,n}$ are the eigenvalues of (3.22) and λ_{i} are the eigenvalues of (3.20). Substituting (3.24) and (3.25) in (3.22) gives an approximation of an eigenfunction evaluation in point $t\in C$

$$\hat{\phi}_{i}\left(t\right)\approx\frac{\sqrt{n}}{\lambda_{i,n}}\sum_{k=1}^{n}K\left(x_{k},t\right)u_{ki,n}. \tag{3.26}$$

One obtains, based on the Nyström approximation, an explicit expression for the entries of the approximated nonlinear mapping $\varphi_i : \mathcal{X} \rightarrow \Psi$:

$$\begin{aligned}\hat{\varphi}_i(x) &= \sqrt{\lambda_i}\hat{\phi}_i\left(x\right)\\ &= \frac{1}{\sqrt{\lambda_{i,n}}}\sum_{k=1}^{n} u_{ki,n} K(x_k, x).\end{aligned} \tag{3.27}$$

In order to introduce parsimony, one chooses as fixed size n_{FS} $(n_{FS} \ll n)$ for a working subsample. A likewise n_{FS}-approximation can be made and the model takes the form

$$\begin{aligned}y(x) &= w^T\hat{\varphi}(x) + b\\ &= \sum_{i=1}^{n_{FS}} w_i \frac{1}{\sqrt{\lambda_{i,l}}} \sum_{k=1}^{n_{FS}} u_{ki,n_{FS}} K(x_k, x) + b.\end{aligned} \tag{3.28}$$

One can solve now the following ridge regression problem in the primal weight space with unknowns $w \in \mathbb{R}^{n_{FS}}, b \in \mathbb{R}$

$$\min_{w,b} \frac{1}{2}\sum_{i=1}^{n_{FS}} w_i^2 + \gamma\frac{1}{2}\sum_{k=1}^{n}\left(y_k - \sum_{i=1}^{n_{FS}}\left(w_i\hat{\varphi}_i\left(x_k\right) + b\right)\right)^2. \tag{3.29}$$

This approach gives explicit links between primal and the dual space representation.

3.4.2 Active selection of a subsample

In order to make a more suitable selection of the subsample instead of a random selection, one can relate the Nyström method to an entropy criterion. Let $x_k \in \mathbb{R}^d$, $k = 1, ..., n$ be a set of input samples from a random variable $X \in \mathbb{R}^d$. The success of a selection method depends on how much information about the original input sample $x_k \in \mathbb{R}^d$, $k = 1, ..., n$, is contained in a subsample $x_j \in \mathbb{R}^d$, $j = 1, ..., n_{FS}$ $(n_{FS} \ll n)$. Thus, the purpose of a subsample selection is to extract n_{FS} $(n_{FS} \ll n)$ samples from $\{x_1, ..., x_n\}$, such that $H_{n_{FS}}(X)$, the information or entropy of the subsample becomes as close to $H_n(X)$, the entropy of the original sample.

One estimates the density function $f\left(x\right)$ by the kernel density estimation

$$\hat{f}\left(x\right) = \frac{1}{n_{FS}h^d}\sum_{k=1}^{n_{FS}} K\left(\frac{x - x_k}{h}\right), \tag{3.30}$$

where h denotes the bandwidth and the kernel $K : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies $\int_{\mathbb{R}^d} K(u)du = 1$. When the differential entropy (Shannon, 1948) defined by

$$\begin{aligned}H_S\left(X\right) &= E\left[-\log f\left(x^{(1)}, ..., x^{(d)}\right)\right]\\ &= -\int ... \int f(x^{(1)}, ..., x^{(d)}) \log f(x^{(1)}, ..., x^{(d)}) dx^{(1)} ... dx^{(d)}\end{aligned} \tag{3.31}$$

is used along with the kernel density estimate $\hat{f}(x)$, the estimation of the entropy $H_{n_{FS},S}(X)$ becomes very complex. But Renyi's entropy of order $q = 2$ (also called quadratic entropy) leads to a simpler estimate of entropy $H_{n_{FS},R2}(X)$. Renyi's entropy of order q is defined as:

$$H_{Rq}(X) = \frac{1}{1-q} \log \int f(x^{(1)}, ..., x^{(d)})^q dx^{(1)}...dx^{(d)}, \quad q > 0, q \neq 1. \tag{3.32}$$

The differential entropy can be viewed as one member of the Renyi's entropy family, because $\lim_{q\to 1} H_{Rq}(X) = H_S(X)$. Although Shannon's entropy is the only one which possesses the properties (e.g., continuity, symmetry, extremal property, recursivity and additivity) for an information measure, the Renyi's entropy family is equivalent with regards to entropy maximization. In real problems, which information measure to use depends upon other requirements such as ease of implementation. Combining (3.30) and (3.32), Renyi's quadratic entropy estimator becomes

$$\begin{aligned} H_{n_{FS},R2}(X) &= -\log \frac{1}{n_{FS}^2 h^{2d}} \sum_{k=1}^{n_{FS}} \sum_{l=1}^{n_{FS}} K\left(\frac{x_k - x_l}{h}\right) \\ &= -\log \frac{1}{l^2 h^{2d}} 1_{n_{FS}}^T \Omega 1_{n_{FS}}. \end{aligned} \tag{3.33}$$

One chooses a fixed size n_{FS} ($n_{FS} \ll n$) for a working set of data points and actively selects points from the pool of training input samples as a candidate for the working set. In the working set a point is randomly selected and replaced by a randomly selected point from the training input sample if the new point improves Renyi's quadratic entropy criterion. This leads to the following fixed size LS-SVM algorithm as introduced in (Sukens *et al.*, 2002)

(i) Given a training set $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$, construct a standardized input training set $\mathcal{S}_n = \left\{\tilde{x} : \tilde{x} = \frac{x-E[x]}{Var[x]},\ \tilde{x}_k \in \mathcal{X},\ k = 1, ..., n\right\}$, choose a working set $\mathcal{W}_{n_{FS}} = \{\tilde{x} : \tilde{x}_j \in \mathcal{X}; j = 1, ..., n_{FS} \ll n\} \subset \mathcal{S}_n$.

(ii) Randomly select a sample point $\tilde{x}^* \in \mathcal{W}_l$, and $\tilde{x}^{**} \in \mathcal{S}_n$, swap$(\tilde{x}^*, \tilde{x}^{**})$.
If $H_{n_{FS},R2}(\tilde{x}_1, ...\tilde{x}_{n_{FS}-1}; \tilde{x}^{**}) > H_{n_{FS},R2}(\tilde{x}_1, ...\tilde{x}_i^*, ...\tilde{x}_{n_{FS}})$ then $\tilde{x}^{**} \in \mathcal{W}_{n_{FS}}$ and $\tilde{x}^* \notin \mathcal{W}_{n_{FS}}, \tilde{x}^* \in \mathcal{S}_n$.

(iii) Calculate $H_{n_{FS},R2}(\tilde{x})$ for the present $\mathcal{W}_{n_{FS}}$.

(iv) Stop if the change in entropy value (3.33) is small.

(v) Estimate w, b in the primal space after estimating the eigenfunctions from the Nyström approximation. according to (3.29).

Chapter 4

Model Assessment and Selection

In this Chapter we describe the key methods (cross-validation and complexity criteria) for performance assessment. We begin the chapter with a discussion of the bias-variance tradeoff and model complexity. Finally, we give a strategy for selecting a good learning parameter vector.

4.1 Introduction

Most efficient learning algorithms in neural networks, support vector machines and kernel based methods (Bishop, 1995; Cherkassky *et al.*, 1998; Vapnik, 1999; Hastie *et al.*, 2001; Suykens *et al.*, 2002b) require the tuning of some extra learning parameters, or *tuning parameters*, denoted here by θ. The tuning parameter selection methods can be divided into three broad classes:

(i) Cross validation and bootstrap.

(ii) Plug-in methods. The bias of an estimate of an unknown real-valued smooth function is usually approximated through Taylor series expansions. A pilot estimate of the unknown function is then "plugged in" to derive an estimate of the bias and hence an estimate of the mean integrated squared error. The optimal tuning parameters minimize this estimated measure of fit. More complete descriptions of these approaches are given in (Härdle, 1989).

(iii) Complexity criteria. Mallows' C_p (Mallows, 1973), Akaike's information criterion (Akaike, 1973), Bayes Information Criterion (Schwartz 1979) and Vapnik-Chernovenkis dimension (Vapnik, 1998).

Figure 4.1 shows the typical behavior of the test and training error, as model complexity is varied (Bishop, 1995) and (Hastie *et al.*, 2001). The training error tends to decrease whenever one increases the model complexity.

However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well. In contrast, if the model is not complex enough, it will underfit and may have large bias, again resulting in

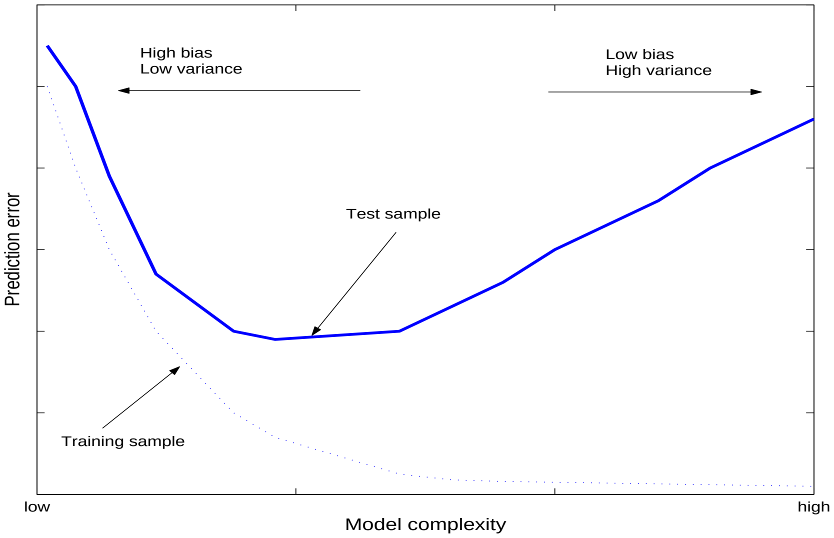

Figure 4.1: Behavior of test sample and training sample error as the model complexity is varied.

poor generalization. To avoid this well-known problem, one divides the data set $\mathcal{D}_n = \{(x_k, y_k) : \ x_k \in \mathcal{X}, y_k \in \mathcal{Y}; \ k = 1, ..., n\}$ into three parts: a training set denoted by $\mathcal{D}_n$, a validation set denoted by $\mathcal{D}_v$, and a test set denoted by $\mathcal{D}_{test}$. The training set is used to fit the models $\hat{m}_n$ (prediction model); the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final model. The test data are completely left untouched within the training and validation process. The complexity criteria and cross-validation methods approximate the validation step respectively analytically and by sample re-use methods.

As in Chapter 2, assume that the empirical training data $\mathcal{D}_n$ can be written as $y_k = m(x_k) + e_k$ where $E[e_k | X = x_k] = 0$ and $E\left[e_k^2\right] = \sigma^2 < \infty$. The expected generalization error, based on the L_2 risk functional, of $\hat{m}_n(x^{new})$ is given by

$$\begin{aligned}\mathcal{R}(x^{new}) &= E\left[(\hat{m}_n(x^{new}) - Y)^2\right] \\ &= \sigma_e^2 + (E[\hat{m}_n(x^{new}) - m(x^{new})])^2 \\ &+ E[\hat{m}_n(x^{new}) - E[\hat{m}_n(x^{new})]]^2 \\ &= \sigma_e^2 + bias^2[\hat{m}_n(x^{new})] + Var[\hat{m}_n(x^{new})] \qquad (4.1)\end{aligned}$$

where $x^{new} \in \mathcal{D}_{test}$. Note from (4.1), called the prediction error decomposition, that the expected loss or risk functional of using $\hat{m}_n(x^{new})$ to predict y is the sum of the variances of $\hat{m}_n(x^{new})$ and y plus the squared bias. The variance of y is beyond our control and is known as the irreducible error. However, the bias and variance of $\hat{m}_n\left(x^{new}\right)$ are functions of our estimator and can therefore potentially be reduced.

The complexity criteria estimate the generalization error via an estimate of the complexity term and then add it to training error. In contrast, the cross-validation and bootstrap methods, are direct estimates of the generalization error. The tuning parameter selection methods (e.g., Cross-validation, Complexity criteria) used in this thesis are described in the following sections.

4.2 Cross-validation

4.2.1 Leave-one-out cross-validation score function

Next, we will motivate the cross-validation procedure as the θ selection rule. Let the distance $d_{ISE}\left(m\left(x\right), \hat{m}_n\left(x;\theta\right)\right)$ denote the integrated squared error measure of accuracy for the estimator $\hat{m}_n\left(x;\theta\right)$. Write

$$\begin{aligned} d_{ISE}\left(m\left(x\right), \hat{m}_n\left(x;\theta\right)\right) &= \int \left(m\left(x\right) - \hat{m}_n\left(x;\theta\right)\right)^2 g\left(x\right) dx \\ &= \int m^2\left(x\right) g\left(x\right) dx + \int \hat{m}_n^2\left(x;\theta\right) g\left(x\right) dx \\ &- 2\int m\left(x\right) \hat{m}_n\left(x;\theta\right) g\left(x\right) dx. \end{aligned} \tag{4.2}$$

Since the first term is independent of θ, minimizing this loss is equivalent to minimizing

$$Q = \int \hat{m}_n^2\left(x;\theta\right) g\left(x\right) dx - 2\int m\left(x\right) \hat{m}_n\left(x;\theta\right) g\left(x\right) dx. \tag{4.3}$$

But this cannot be realized in practice because this quantity depends on the unknown real-valued function $m\left(x\right)$ and $g\left(x\right)$ the density function over the input space. The first term of (4.3) can be computed entirely from the data, and the second term of (4.3) may be written as

$$Q_2 = \int m\left(x\right) \hat{m}_n\left(x;\theta\right) g\left(x\right) dx = E_{(x,y)}\left[\hat{m}_n\left(x;\theta\right) y\right]. \tag{4.4}$$

If we estimate (4.4) by $n^{-1}\sum_{k=1}^{n} y_k \hat{m}_n\left(x_k;\theta\right)$, the selection rule will be a biased estimator of $d_{ISE}\left(m\left(x\right), \hat{m}_n\left(x;\theta\right)\right)$. The reason for the bias in the selection rule is that the observation y_k is used in $\hat{m}_n\left(x_k;\theta\right)$ to predict itself. This is equivalent to considering the apparent (resubstitution) estimate of the prediction error (Hastie *et al.*, 2001). There are several methods to find an unbiased estimate of $d_{ISE}\left(m\left(x\right), \hat{m}_n\left(x;\theta\right)\right)$, for example: a plug-in method, leave-one-out technique and a modification such that bias terms cancel asymptotically. We will use here the leave-one-out technique, in which one observation is left out. Therefore, a better estimator for (4.4) instead of $n^{-1}\sum_{k=1}^{n} y_k \hat{m}_n\left(x_k;\theta\right)$ will be

$$\hat{Q}_2 = \frac{1}{n}\sum_{k=1}^{n} y_k \hat{m}_n^{(-k)}\left(x_k;\theta\right), \tag{4.5}$$

where $\hat{m}_n^{(-k)}\left(x_k;\theta\right)$ denotes the leave-one-out estimator with point k left out from the training. Similarly, the first term of (4.3) may be approximated by

$$\hat{Q}_1 = \frac{1}{n}\sum_{k=1}^{n}\left(\hat{m}_n^{(-k)}\left(x_k;\theta\right)\right)^2. \tag{4.6}$$

From (4.5) and (4.6), the cross-validation function is

$$CV\left(\theta\right) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{m}_n^{(-k)}\left(x_k;\theta\right)\right)^2. \tag{4.7}$$

The above motivation is related to some ideas of (Rudemo, 1982) and (Bowman, 1984). In the context of kernel smoothing this score function for finding the bandwidth was proposed by (Clark, 1975). (Wahba and Wold, 1975) proposed a similar technique in the context of spline smoothing. The least squares cross-validated choice of θ for the LS-SVM estimates, based on the average squared prediction error, is the minimizer of

$$\inf_{\theta} CV(\theta) = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{m}_n^{(-k)}(x_k;\theta))^2. \tag{4.8}$$

4.2.2 Generalized cross-validation score function

The GCV criterion was first proposed by (Craven and Wahba, 1979) for the use in the context of nonparametric regression with a roughness penalty. However, (Golub, Heath and Wahba, 1979) showed that GCV can be used to solve a wide variety of problems involving estimation of minimizers for (4.3). In the leave-one-out cross-validation it is necessary to solve n separate LS-SVM's, in order to find the n models $\hat{m}_n^{(-k)}\left(x_k;\theta\right)$. From (3.12), the values of the LS-SVM $\hat{m}_n\left(x_k;\theta\right)$ depend linearly on the data y_k. We can write the deleted residuals $y_k - \hat{m}_n^{(-k)}\left(x_k;\theta\right)$ in terms of $y_k - \hat{m}_n\left(x_k;\theta\right)$ and the *k-th* diagonal element of the smoother matrix $S(\theta)$. The CV score function satisfies

$$CV\left(\theta\right) = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{y_k - \hat{m}_n\left(x_k;\theta\right)}{1 - s_{kk}(\theta)}\right)^2, \tag{4.9}$$

where $\hat{m}_n\left(x_k;\theta\right)$ is the LS-SVM calculated from the full data set $\{(x_k, y_k)\}_{k=1}^{n}$ and $s_{kk}(\theta)$ is the k-th diagonal element of the smoother matrix. The proof of (4.9) is identical to the one obtained in the development of the PRESS criterion for deciding about the complexity for parametric multivariate regression; see (Cook and Weisberg, 1982). Assuming that $\mathrm{tr}[S(\theta)] < n$ and $s_{ii} < 1$, $\forall i$, the basic idea of generalized cross-validation is to replace the factors $1 - s_{kk}(\theta)$ by their average value, $1 - n^{-1}\mathrm{tr}[S(\theta)]$. The generalized cross-validation score is then constructed, by analogy with ordinary cross-validation, by summing the squared residuals corrected by the square of $1 - n^{-1}\mathrm{tr}[S(\theta)]$. Since $1 - n^{-1}\mathrm{tr}[S(\theta)]$ is the same for all k, we obtain

$$GCV\left(\theta\right) = \frac{1}{n}\frac{\sum_{k=1}^{n}\left(y_k - \hat{m}_n\left(x_k;\theta\right)\right)^2}{\left(1 - n^{-1}\mathrm{tr}\left[S(\theta)\right]\right)^2}. \tag{4.10}$$

As in ordinary cross-validation, the GCV choice of the tuning parameters is then carried out by minimizing the function $GCV\left(\theta\right)$ over θ.

4.2.3 V-fold cross-validation score function

In general there is no reason that training sets should be of size $n-1$. There is the possibility that small perturbations, when single observations are left out, make $CV\left(\theta\right)$ too variable, if fitted values $\hat{m}_n\left(x;\theta\right)$ do not depend smoothly on the empirical distribution $\hat{F}_n$ or if the loss function $L\left(y, \hat{m}_n\left(x;\theta\right)\right)$ is not continuous. These potential problems can be avoided to a large extent by leaving out groups of observations, rather than single observations. We begin by splitting the data randomly into V disjoint sets of nearly equal size. Let the size of the vth group be m_v and assume that $\lfloor n/V \rfloor \leq m_v \leq \lfloor n/V \rfloor + 1$ for all v. For η real, $\lfloor \eta \rfloor$ denotes the greatest integer less or equal to η. For each such split we apply (4.7), and then average these estimates. The result is the V-fold cross-validation estimate of prediction error

$$CV_{V-fold}\left(\theta\right) = \sum_{v=1}^{V}\frac{m_v}{n}\sum_{k=1}^{m_v}\frac{1}{m_v}\left(y_k - \hat{m}_n^{(-m_v)}\left(x_k,\theta\right)\right)^2, \tag{4.11}$$

where $\hat{f}^{(-m_v)}$ represents the model obtained from the data outside group v. Practical experience suggests that a good strategy is to take $V = \min\left(\sqrt{n}, 10\right)$, because taking $V > 10$ may be computationally too expensive when the prediction rule is complicated, while taking groups of size at least $\sqrt{n}$ should perturb the data sufficiently to give small variance of the estimate (Davison and Hinkley, 1997). The use of groups will have the desired effect of reducing variance, but at the cost of increasing bias. According to (Beran, 1984), (Serfling, 1984) and (Burman, 1989), the bias of $CV_{V-fold}\left(\theta\right) \approx a_0\left[\left(V-1\right)^{-1}n^{-1}\right]$, for $V = n$ (leave-one-out) the bias is of order $O\left(n^{-2}\right)$, but when V is small, the bias term is not necessarily very small. The term a_0, depending on L and $\hat{F}_n$, is of order the number of parameters being estimated. For LS-SVM, a_0 becomes a constant multiplied with the number of effective parameters (see (3.14)). Therefore, if the number of effective parameters is not small, the $CV_{V-fold}\left(\theta\right)$ is a poor estimate of the prediction error. But the bias of $CV_{V-fold}\left(\theta\right)$ can be reduced by a simple adjustment (Burman, 1989). The adjusted V-fold cross-validation estimate of prediction error is

$$CV^{adj}_{V-fold}(\theta) = CV_{V-fold}(\theta) + \left[\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{m}_n\left(x_k;\theta\right)\right)^2 - \sum_{v=1}^{V}\frac{m_v}{n}\sum_{k=1}^{n}\frac{1}{n}\left(y_k - \hat{m}_n^{(-m_v)}\left(x_k;\theta\right)\right)^2\right]. \tag{4.12}$$

The bias of $CV^{adj}_{V-fold}(\theta) \approx a_1\left[(V-1)^{-1}n^{-2}\right]$, for some constant a_1 depending on L and $\hat{F}_n$. The $CV^{adj}_{V-fold}(\theta)$ has a smaller bias than $CV_{V-fold}(\theta)$ and works better asymptotically as n increases. The $CV^{adj}_{V-fold}(\theta)$ is almost as simple to calculate, because it requires no additional LS-SVM fits.

4.3 Complexity criteria

4.3.1 Final Prediction Error (FPE) criterion, Mallows' C_p, AIC and BIC

Let $\mathcal{P}$ be a finite set of parameters. For $\alpha \in \mathcal{P}$, let $\mathcal{F}_\beta$ be a set of functions

$$\mathcal{F}_\beta = \left\{ m : m\left(x,\beta\right) = \beta_0 + \sum_{l=1}^{d}\beta_l x^{(l)},\ x \in \mathbb{R}^d \text{ and } \beta \in \mathcal{P} \right\}, \tag{4.13}$$

let $Q_n(\beta) \in \mathbb{R}^+$ be a complexity term for $\mathcal{F}_\beta$ and let $\hat{m}_n$ be an estimator of m in $\mathcal{F}_\beta$. The learning parameters are chosen to be the minimizer of a cost function defined as

$$J_\beta(\lambda) = \frac{1}{n}\sum_{k=1}^{n} L\left(y_k, \hat{m}_n\left(x_k;\beta\right)\right) + \lambda\left(Q_n(\beta)\right)\hat{\sigma}_e^2 \tag{4.14}$$

where $\sum_{k=1}^{n} L(y_k, \hat{m}_n(x_k;\beta))$ is the residual sum of squares (RSS), $Q_n(\beta) \in \mathbb{R}^+$ is a complexity term, $\lambda > 0$ is a cost complexity parameter and the term $\hat{\sigma}_e^2$ is an estimate of the error variance. The Final Prediction Error criterion depends only on $\hat{m}_n$ and the data. If $\hat{m}_n$ is defined by minimizing the empirical L_2 risk over some linear vector space $\mathcal{F}_\beta$ of functions with dimension d_β, then $J_\beta(\lambda)$ will be of the form:

- Let $\lambda = 2$ and $Q_n(\alpha) = n^{-1}d_\beta$

$$C_p(\lambda) = \frac{1}{n}RSS + 2\left(\frac{d_\beta}{n}\right)\hat{\sigma}_e^2. \tag{4.15}$$

The Akaike information criterion (AIC) is a similar but more generally applicable estimate when a log-likelihood loss function is used. For the Gaussian model (with variance $\sigma_e^2 = \hat{\sigma}_e^2$ assumed known), the AIC statistic is equivalent to C_p.

- Let $\lambda = \log n$ and $Q_n(\beta) = n^{-1} d_\beta$

$$BIC(\lambda) = \frac{1}{n} RSS + (\log n) \left(\frac{d_\beta}{n} \right) \hat{\sigma}_e^2$$

- Let $\lambda = \log\log n$ and $Q_n(\beta) = n^{-1} d_\beta$

$$J_1(\lambda) = \frac{1}{n} RSS + (\log\log n) \left(\frac{d_\beta}{n} \right) \hat{\sigma}_e^2,$$

  which has been proposed by (Hannon and Quinn, 1979) in the context of autoregressive model order determination.

The AIC was originally designed for parametric models as an approximately unbiased estimate of the expected Kullback-Leibler information. For linear regression and time series models, (Hurvich and Tsai, 1989) demonstrated that in small samples the bias of the AIC can be quite large, especially as the dimension of the candidate model approaches the sample size (leading to overfitting of the model), and they proposed a corrected version, $AICC$, which was found to be less biased than the AIC. The $AICC$ for hyperparameter selection is given by

$$AICC_\beta\left(\lambda\right) = \frac{1}{n} RSS + \left(1 + \frac{2\left(d_\beta + 1\right)}{n - d_\beta - 2} \right) \hat{\sigma}_e^2 \tag{4.16}$$

where $\lambda = 1$ and $Q_n(\beta) = 1 + \frac{2(d_\beta+1)}{n-d_\beta-2}$.

For $\theta \in \mathcal{Q}$, let $\mathcal{F}_{n,\theta}$ be a set of functions

$$\mathcal{F}_{n,\theta} = \left\{ m : m\left(x, \theta\right),\ x \in \mathbb{R}^d,\ y \in \mathbb{R}^n,\ \theta \in \mathcal{Q} \text{ and } \hat{m}_n\left(\hat{\theta}\right) = S\left(\hat{\theta}\right) y \right\}, \tag{4.17}$$

let $Q_n(\theta) \in \mathbb{R}^+$ be a complexity term for $\mathcal{F}_{n,\theta}$ and let $\hat{m}_n$ be an estimator of m in $\mathcal{F}_{n,\theta}$. For example, regression spline estimators, wavelet and LS-SVM estimators are linear estimators , in the sense that $\hat{m}_n(\hat{\theta}) = S(\hat{\theta})y$, where the matrix $S(\hat{\theta})$ is called the smoother matrix and depends on $x \in \mathcal{D}_n$ but not on y. Based on (Moody, 1992) and by analogy, the learning parameters are chosen to be the minimizer of a more generalized cost function defined as

$$JC_\theta(\lambda) = \frac{1}{n} RSS + \left(1 + \frac{2\mathrm{tr}(S(\hat{\theta})) + 2}{n - \mathrm{tr}(S(\hat{\theta})) - 2} \right) \hat{\sigma}_e^2. \tag{4.18}$$

Each of these selectors depends on $S(\hat{\theta})$ through its trace ($\mathrm{tr}(S(\hat{\theta})) < n - 2$), which can be interpreted as the effective number of parameters used in the fit.

4.3.2 Vapnik-Chervonenkis dimension

A difficulty in using (4.18) is the need to specify the number of parameters (or complexity) used in the fit. The Vapnik-Chernovenkis theory provides an other measure of complexity than the effective number of parameters, and gives associated bounds. Suppose we have a class of functions

$$\mathcal{F}_{n,\beta} = \left\{m : m(x,\beta),\ x \in \mathbb{R}^d \text{ and } \beta \in \Lambda\right\}, \tag{4.19}$$

where Λ is some parameter vector set and consider the indicator class

$$\mathcal{I}_{\beta,\tau} = \left\{I : I\left(m(x,\beta) - \tau\right),\ x \in \mathbb{R}^d,\ \beta \in \Lambda \text{ and } \tau \in \left(\inf_x m(x,\beta), \sup_x m(x,\beta)\right)\right\}. \tag{4.20}$$

The VC-dimension (Vapnik, 1998) of real-valued functions $\mathcal{F}_{n,\beta}$ is defined to be the VC-dimension of the indicator class $\mathcal{I}_{\beta,\tau}$. The VC-dimension of the class $\mathcal{F}_\beta$ is defined to be the largest number of points that can be shattered by members of $\mathcal{F}_{n,\beta}$.

Example 8 *Let the class of functions be defined as*

$$f(x,\beta) = \beta_0 + \sum_{l=1}^{d} \beta_l x^{(l)}$$

and $I\left(\beta_0 + \beta_1 x - \tau\right)$ is the linear indicator function. The VC-dimension of the class $m(x,\beta)$ is equal to the number of parameters $(d+1)$ of the set of functions.

Example 9 *Let the class of functions be defined as*

$$f(x,\beta) = \sin\left(\beta x\right)$$

and $I\left(\sin\left(\beta x\right) - \tau\right)$ is the indicator function. This class of functions has only one parameter, but it has infinite VC-dimension .

If one fits $\mathcal{D}_n = \{(x_1,y_1),...,(x_n,y_n)\}$ using a class of functions $\mathcal{F}_{n,\beta}$ having VC-dimension h, with probability $(1-\alpha)$ over the training sets, the inequality

$$R\left(f\right) \leq \frac{R_n\left(f\right)}{\left(1 - c\sqrt{\xi\left(n\right)}\right)_+}$$

is valid, where

$$\xi\left(n\right) = a_1 \frac{h\left(\log\left(\frac{a_2 n}{h}\right) + 1\right) - \log\left(\frac{\alpha}{4}\right)}{n},$$

and $a_1 = a_2 = c = 1$ (Cherkassky and Mulier, 1998). These bounds hold simultaneously for all members of $\mathcal{F}_{n,\beta}$.

4.4 Choosing the learning parameters

Loader (1999) has studied, in the context of kernel density and kernel regression, a wide range of smoothing parameters on both real and simulated data. The plug-in approaches have fared rather poorly while cross-validation and AIC produce good estimators. For practical use, it is often preferable to have a data-driven method to select learning parameters. For this selection process, many data-driven procedures have been discussed in the literature. Commonly used are those based on the cross-validation criterion of Stone (Stone, 1974) and the generalized cross-validation criterion of Craven and Wahba (1979). One advantage of cross-validation and generalized cross-validation over some other selection criteria such as Mallows' C_p, Akaike's information criterion is that they do not require estimates of the error variance. This means that Mallows' C_p, Akaike's information criterion require a roughly correct working model to obtain the estimate of the error variance. Cross-validation does not require this. The motivation behind cross-validation is easily understood, see (Allen, 1974) and (Stone, 1974). Much work has been done on the ordinary or leave-one-out cross-validation (Bowman, 1984) and (Härdle and Marron, 1985). However, the difficulty with ordinary cross-validation is that it can become computationally very expensive in practical problems. Therefore, (Burman, 1989) has introduced V-fold cross-validation. For more references on smoothing parameter selection, see (Marron, 1987, 1989) and (Härdle and Chen, 1995).

For a comparison of cross-validation with other sample re-use techniques (e.g., bootstrap methods) see (Efron, 1982). The bootstrap procedures are nothing more than smoothed versions of cross-validation, with some adjustments made to correct for bias. The improvement of the bootstrap estimators over cross-validation, in the dichotomous situation where both y and the prediction rule are either 0 or 1, is due mainly to the effect of smoothing. In smoother prediction problems, when y and the prediction rule are continuous, there is little difference between cross-validation and bootstrap methods.

The strategy, for selecting a good learning parameter vector, is to choose one (or more) of the selection criteria. The choice of which criterion to use will depend on the situation. Table 4.1 summarizes some characteristics of a number of situations.

If σ^2 is unknown and no reasonable estimate is available, GCV or cross-validation can be used since they do not require estimation of the error variance. The use of cross-validation will involve more computational labor than GCV. In practice it may be feasible to compute two or more risk estimates. By computing two or more of the measures one obtains some basis for comparison.

	σ^2	Smoother matrix	Remarks
Leave-one-out	not required	not required	High variance low bias
V-fold-CV	not required	not required	low variance high bias
GCV	not required	required	(*)
AIC	required	required	
BIC	required	required	
SRM	not required	not required	

Table 4.1: The strategy for selecting a good learning parameter vector. (*): For a given data set, GCV always selects the same learning parameter vector, no matter whether the magnitude noise is 100 or is just 0.01.

Chapter 5

The Jackknife and the Bootstrap

Kernel based methods (e.g. Nadaraya-Watson estimator, Support Vector Machines,...) are effective methods for function estimation in a flexible nonparametric way (without making assumptions about its shape). For example, prediction intervals are an important approach to get an impression about the accuracy that can be expected for a particular estimator. A classical way of constructing prediction intervals for an unknown regression function consists of using the limit distribution of the properly normalized difference between the regression function and some estimator. However, the traditional approach to statistical inference in complicated independent identically distributed (i.i.d.) data problems (support vector machines, kernel based methods) is difficult to mathematically analyse. The approach of resampling plans (Bootstrap, Jackknife, Balanced repeated replications, subsampling, ...) is a computationally attractive alternative. A second example is the potential of the Jackknife in obtaining empirical influence functions. The resampling methods replace theoretical derivations required in applying traditional methods in statistical analysis (nonparametric estimation of bias, variance and more general measures of error) by repeatedly resampling the original data and making inferences from the resamples. The most popular data-resampling methods used in statistical analysis are the bootstrap (Efron, 1979) and jackknife (Quenouille, 1949; Tukey, 1958).

We begin the Chapter with a discussion of the Jacknife. Next we describe the bootstrap as a general tool for assessing statistical accuracy and then show how it can be used in the regression context. Bootstrap algorithms (paired bootstrap, residual bootstrap and external bootstrap) are given in the context of kernel regression.

5.1 The Jackknife

The Jackknife estimator was introduced by (Quenouille, 1949) and named by (Tukey, 1958). This technique's purpose is to decrease the bias of an estimator (the Jackknife estimator). The procedure operates as follows. Let $X_1, ..., X_n$ be a random sample of size n from an unknown probability distribution F. Having observed values $x_1, ..., x_n$ one is interesed in some statistic $T(F)$. Simple examples are the mean $\mu = \int x dF(x)$ and the variance $\sigma^2 = \int (x-\mu)^2 dF(x)$. Let $T(\hat{F}_n)$ be an estimator of $T(F)$. Divide the random sample into r groups of size $l = \frac{n}{r}$ observations each. Delete one group at a time, and estimate $T(F)$ based on the remaining $(r-1)\,l$ observations, using the same estimation procedure previously used with a sample of size n. Denote the estimator of $T(F)$ obtained with the ith group deleted by $T(\hat{F}_{(i)})$, called a Jackknife statistic $(i = 1, ..., r)$. For $i = 1, ..., r$, form pseudovalues

$$J_i = rT(\hat{F}_n) - (r-1)\,T(\hat{F}_{(i)}), \tag{5.1}$$

and consider the Jackknife estimator of $T(F)$ defined by

$$\begin{aligned} J\left(T(\hat{F}_n)\right) &= \frac{1}{r}\sum_{i=1}^{r}\left(rT(\hat{F}_n) - (r-1)\,T(\hat{F}_{(i)})\right) \\ &= T(\hat{F}_n) - (r-1)\,\bar{T}(\hat{F}_{(i)}) \end{aligned} \tag{5.2}$$

where $\bar{T}(\hat{F}_{(i)}) = \frac{1}{r}\sum_{i=1}^{r} T(\hat{F}_{(i)})$. Note that the Jackknife estimator can be written as

$$J\left(T(\hat{F}_n)\right) = T(\hat{F}_n) + (r-1)\left(T(\hat{F}_{(i)}) - \bar{T}(\hat{F}_{(i)})\right) \tag{5.3}$$

which shows the estimator $J\left(T(\hat{F}_n)\right)$ as an adjustment to $T(\hat{F}_n)$, with the amount of adjustment depending on the difference between $T(\hat{F}_n)$ and $\bar{T}(\hat{F}_{(i)})$. The special case $l = 1$ is the most commonly used Jackknife, in which case

$$J\left(T(\hat{F}_n)\right) = nT(\hat{F}_n) - (n-1)\,\bar{T}(\hat{F}_{(i)}). \tag{5.4}$$

Tukey (Tukey, 1958) suggested how the recomputed statistics $T(\hat{F}_{(i)})$ could also provide a nonparametric estimate of the variance.

5.2 The Bootstrap

The bootstrap is a method for estimating the distribution of an estimator or statistic by resampling the data. Under mild regularity conditions, the bootstrap yields an approximation to the distribution of an estimator or statistic that is at least as accurate as the approximation obtained from first-order asymptotic theory. Thus, the bootstrap provides a way to substitute computation for mathematical analysis if calculating the asymptotic distribution of an estimator or statistic is difficult. The bootstrap is often more accurate in finite samples than first-order asymptotic approximations but does not have the algebraic complexity of higher-order expansions. Thus, it can provide a practical method for improving upon first-order approximations. An excellent introduction to the bootstrap may be found in the work of (Efron and Tibshirani, 1993). More theoretical properties may be found in the work of (Hall, 1992) and (Davison & Hinkley, 1997). Over the past decade bootstrap methods have become widely used in statistical applications. But as with any statistical procedure, it is important to be clear about the assumptions that underlie the validity and accuracy of bootstrap calculation. There are several situations in which standard bootstrap calculations and methods (Algorithm 1) do not give reliable answers. In many situations correction actions is possible, by modifying the resampling scheme or by modifying some other aspect of the method. We briefly list some problem situations:

(i). *Inconsistency of bootstrap method.* The combination of model, statistic and resampling scheme may be such that bootstrap results fail to approximate the required properties. Cases of bootstrap inconsistency include the sample maximum (Politis and Romano, 1994), kernel estimates for densities and for regression curves (Härdle and Bowman, 1988).

(ii). *Effect of data outliers.* Outliers influence not only the estimator but also the resampling properties. Depending upon the resampling model used, an outlier may occur with variable frequency in bootstrap samples, and the effect of this may be hard to anticipate even if the estimator itself is not affected.

5.2.1 The bootstrap approximation

Let $X = (x_1, ...x_n)$ be a sample of n (*i.i.d.*) random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is a sample space (a set of observations), $\mathcal{A}$ is a σ-field of subsets of Ω and $\mathbb{P}$ is a probability distribution or measure defined on the elements on $\mathcal{A}$. $\mathbb{P}$ is assumed to belong to a certain collection $\mathcal{P}$ of distributions. For example, interest might focus on some parameter $T(\mathbb{P})$. An estimator $T_n(\hat{\mathbb{P}})$ is suggested and an estimated variance is desired, or possibly an estimator of bias of $T_n(\hat{\mathbb{P}})$ as an estimate of $T(P)$. Another goal is to construct a confidence region for $T(\mathbb{P})$.

The problem of estimating the entire cumulative distribution function leads to considering a random variable $\mathcal{J}(X, T(\mathbb{P}))$, which is some functional depending on both X and $T(\mathbb{P})$. The idea is that confidence intervals could be constructed if the distribution of $\mathcal{J}(X, T(\mathbb{P}))$ were known. For example, an estimator $T_n(\hat{\mathbb{P}})$ of a real-valued parameter $T(\mathbb{P})$ might be given so that a natural choice is $\mathcal{J}(X, T(\mathbb{P})) = \left[T_n(\hat{\mathbb{P}}) - T(\mathbb{P})\right]/S_n$, where S_n is an estimate of the standard deviation of $T_n(\hat{\mathbb{P}})$. For estimating the sampling distribution of $T_n(\hat{\mathbb{P}})$, set $\mathcal{J}(X, T(\mathbb{P})) = T_n(\hat{\mathbb{P}})$.

In general, the sampling distribution of $\mathcal{J}(X, T(\mathbb{P}))$ under $\mathbb{P}$ is required and is defined by

$$H\left(x, \mathbb{P}\right) = \text{Prob}\left\{\mathcal{J}(X, T(\mathbb{P})) \leq x\right\} \tag{5.5}$$

Based on the data X, $\mathbb{P}$ is estimated by some probability mechanism $\hat{\mathbb{P}}$. The bootstrap approximation to the distribution of $\mathcal{J}(X, T(\mathbb{P}))$ under $\mathbb{P}$ is the distribution (conditional on x) of $\mathcal{J}(X^*, T(\hat{\mathbb{P}}))$ where X^* are (*i.i.d.*) from $\hat{\mathbb{P}}$

$$H_{boot}(x) = H(x, \hat{\mathbb{P}}) = \text{Prob}^* \left\{ \mathcal{J}(X^*, T(\hat{\mathbb{P}})) \leq x \,|X \right\} \tag{5.6}$$

where $\text{Prob}^* \{\cdot \,|X\}$ denotes the conditional probability for given X.

Usually, the explicit form for $H_{boot}(x)$ is not available and a stochastic approximation to $H_{boot}(x)$ is necessary, a Monte Carlo approximation to $H_{boot}(x)$ is

$$H_{boot}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^{B} I_{\left\{\mathcal{J}(X_b^*, T(\hat{\mathbb{P}})) \leq x\right\}}, \tag{5.7}$$

where $I_{\{\cdot\}}$ is an indicator function which outputs the value 1 if the event $\left\{\mathcal{J}(X_b^*, T(\hat{\mathbb{P}})) \leq x\right\}$ occurs and X_b^*, $b = 1, ..., B$ are independent bootstrap samples from $\hat{\mathbb{P}}$. The bootstrap principle is illustrated in algorithm ......

Algorithm 10 *(bootstrap principle).*

(i) From $X = (x_1, ...x_n)$, calculate the estimate $T_n(\hat{\mathbb{P}})$.

(ii) Construct the empirical distribution, $\hat{\mathbb{P}}$, which puts equal mass $1/n$ at each observation (uniformly random sampling with replacement).

(iii) From the selected $\hat{\mathbb{P}}$, draw a sample $X^* = (x_1^*, ...x_n^*)$, called the bootstrap sample.

(iv) Approximate the distribution of $\mathcal{J}_n(X, T(\mathbb{P}))$ by the distribution of $\mathcal{J}(X^*, T(\hat{\mathbb{P}}))$

Simulation of independent bootstrap samples and their use is usually easily programmed and implemented (Algorithm ...). But sometimes the statistic is very costly to compute or the procedure will be repeated many times. More sophisticated Monte Carlo techniques exist that reduce the number of simulations needed to obtain a given precision. They are discussed in books on Monte Carlo methods, such as (Hammersley and Handscomb, 1964; Fox and Schrage, 1987; Ripley, 1987; Niederreiter, 1992). Balanced bootstrap simulation was introduced by (Davison, Hinkley and Schechtman, 1986). Linear approximations were used as control variates in bootstrap sampling by (Davison, Hinkley and Schechtman, 1986), a different approach was taken by (Efron, 1990). Importance resampling was suggested by (Johns, 1988) and (Davison, 1988), and was exploited by (Hinkley and Shi, 1989) in the context of iterated bootstrap confidence intervals. (Hall and Wood, 1993) describe algorithms for balanced importance resampling. The saddlepoint method, which eliminates the need for simulation, is used by (Davison and Hinkley, 1988; Daniels and Young, 1991; Wang, 1993; DiCiccio, Martin and Young, 1992, 1994). Other methods applied to bootstrap simulation include antithetic sampling (Hall, 1989) and Richardson extrapolation (Bickel and Yahav, 1988). Here we will use only the bootstrap method as introduced by (Efron, 1979).

Example 11 *Consider the following example in which it is desired to estimate the mean squared error (MSE) of a parameter. Let* $X = (x_1, ...x_n)$ *denote the data set of* N *(i.i.d.) observations from an unknown distribution* F *and let* $T_n(\hat{F})$ *be an estimator of an unknown parameter* $T(F)$*. The MSE of* $T_n(\hat{F})$ *as an estimator of* $T(F)$ *is*

$$MSE\left(T_n(\hat{F})\right) = E\left(T_n(\hat{F}) - T(F)\right)^2 = Var\left(T_n(\hat{F})\right) + \left(bias\left(T_n(\hat{F})\right)\right)^2. \tag{5.8}$$

The bias of $T_n(\hat{F})$ *is defined as*

$$\begin{aligned} bias\left(T_n(\hat{F})\right) &= E\left(T_n(\hat{F})\right) - T(F) \\ &= \int xdH\left(x, F\right) - T(F), \end{aligned} \tag{5.9}$$

where $H\left(x, F\right)$ *is given by (5.5) with* $\mathcal{J} = T_n(\hat{F})$*. We can substitute the unknown* F *and* $T(F)$ *in (5.9) by their* $\hat{F}$ *and* $T_n(\hat{F})$*, respectively, and obtain the bootstrap estimator*

$$bias^*\left(T_n(\hat{F})\right) = \int xdH\left(x, \hat{F}\right) - T_n(\hat{F}). \tag{5.10}$$

When the integral in (5.10) has no explicit form, we can use the Monte Carlo approximation

$$\begin{aligned} bias^*_{(B)}\left(T_n(\hat{F})\right) &= \int xdH_{(B)}\left(x, \hat{F}\right) - T_n(\hat{F}) \\ &= \frac{1}{B}\sum_{b=1}^{B} T^*_{n,b}(\hat{F}) - T_n(\hat{F}). \end{aligned} \tag{5.11}$$

The variance of $T_n(\hat{F})$ *is defined as*

$$Var\left(T_n(\hat{F})\right) = \int \left[T_n(x) - \int T_n(x)dF\left(x\right)\right]^2 dF\left(x\right). \tag{5.12}$$

Substituting $\hat{F}$ *for* F *in (5.12), we obtain the bootstrap variance*

$$Var^*\left(T_n(\hat{F})\right) = \int \left[T_n(x) - \int T_n(x)d\hat{F}\left(x\right)\right]^2 d\hat{F}\left(x\right). \tag{5.13}$$

$B:$	10	100	500	1000	2000	5000	10000
$\widehat{s.e}_B$	0.0959	0.1282	0.1291	0.1280	0.1277	0.1277	0.1275

Table 5.1: The bootstrap estimate of standard error for the sample correlation coefficient (0.1147). A run of 100000 bootstrap replications gave the tabled values as B increased from 10 to 10000.

When the right-hand side of (5.12) is not a simple analytic expression, we cannot evaluate $Var\left(T_n(\hat{F})\right)$ exactly, even if F is known. Monte Carlo techniques can be used to approximate $Var\left(T_n(\hat{F})\right)$ numerically

$$Var^*_{(B)}\left(T_n(\hat{F})\right) = \frac{1}{B}\sum_{b=1}^{B}\left[T^*_{n,b}(\hat{F}) - \frac{1}{B}\sum_{j=1}^{B}T^*_{n,j}(\hat{F})\right]^2 \tag{5.14}$$

The Monte Carlo approximation of the MSE

$$\begin{aligned} MSE^*_{(B)}\left(T_n(\hat{F})\right) &= Var^*_{(B)}\left(T_n(\hat{F})\right) + \left[bias^*_{(B)}\left(T_n(\hat{F})\right)\right]^2 \\ &= \frac{1}{B}\sum_{b=1}^{B}\left[T^*_{n,b}(\hat{F}) - T_n(\hat{F})\right]^2 . \end{aligned} \tag{5.15}$$

We can see in this theoretical example, that the bootstrap is a mixture of two techniques: The substitution principle and the numerical approximation.

Example 12 *As a second example consider the sample correlation coefficient between two groups g_1 and g_2 both of size $n = 15$. The sample correlation coefficient $\hat{\rho} = 0.7764$. The textbook formula for the standard error of the correlation coefficient is $\frac{(1-\hat{\rho}^2)}{\sqrt{n-3}}$. Substituting $\hat{\rho} = 0.7764$ gives a value of 0.1147. Table 5.1 shows the bootstrap estimate of standard error for B bootstrap replications ranging from 10 to 10000.*

We can look at the bootstrap data graphically, rather than relying entirely on a single summary statistic like the standard error denoted by $\widehat{s.e}_B$. Figure 5.1 shows the histogram of $\hat{\rho}$ for 2000 samples of size $n = 15$ drawn from the original sample.

5.2.2 Resampling schemes for regression models

Bootstrapping pairs and bootstrapping residuals

In general, there are two types of x_k: deterministic x_k (fixed design) and random x_k (random design). In the former case, the e_k are assumed (*i.i.d.*) with mean zero and variance σ^2, in the latter case, the (x_k, y_k) are assumed (*i.i.d.*) and $E[e_k|x_k] = 0$. Depending on whether or not the x_k are random there are different resampling schemes in this problem:

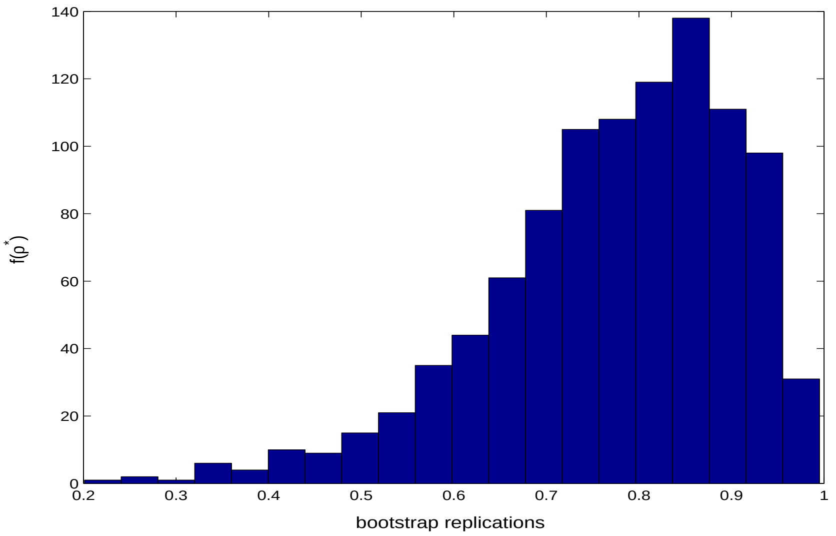

Figure 5.1: Histogram of 2000 bootstrap replications of $\hat{\rho}^{*}$, from the second example (two groups of size $n = 15$).

(a) . The *paired bootstrap* seems to be a natural procedure when the x_k are random and (x_k, y_k) are $(i.i.d.)$ from an unknown multivariate distribution F. In this case, $\mathbb{P} = F_{XY}$ and can be identified by the joint distribution of (x_k, y_k) and estimated by the empirical distribution function putting mass n^{-1} to (x_k, y_k). The bootstrap data are generated from this empirical distribution $\hat{F}_{XY}$.

Algorithm 13 *(The paired bootstrap).*

(i) The unknown probability model $\mathbb{P}$ was taken to be F_{XY}.

(ii) The bootstrap data are generated from this empirical distribution $\hat{F}_{XY}$: probability $\frac{1}{n}$ on (x_k, y_k).

(iii) Calculate the bootstrap estimates $\hat{m}_n^{*}(x_k)$ based on $\{(x_k^{*}, y_k^{*})\}_{k=1}^{n}$.

(iv) This whole process must be repeated B times.

(b) . The bootstrap based on *residuals* was proposed by Efron (Efron, 1979). The x_k are nonrandom and $e_1, ..., e_n$ are $(i.i.d.)$ from an unknown distribution F_e with zero mean. In this case, $\mathbb{P}$ can be identified as $(m(x), F_e)$. Let $\hat{m}_n(x_k)$ denote the estimation of $m(x_k)$, then F_e can be estimated by the empirical distribution $\hat{F}_e$ putting mass n^{-1} to $\hat{e}_k - n^{-1}\sum_{j=1}^{n}\hat{e}_j$, where $\hat{e}_k = y_k - \hat{m}_n(x_k)$

is the *k-th* residual. $\mathbb{P}$ is now estimated by $\hat{\mathbb{P}} = (\hat{m}_n(x), \hat{F}_e)$. To generate bootstrap data (x_k, y_k^*), we first generate (*i.i.d.*) data $e_1^*, ..., e_n^*$ from $\hat{F}_e$ and then define $y_k^* = \hat{m}_n(x_k) + e_k^*$.

Algorithm 14 *(The bootstrap based on residuals)*

(i) The unknown probability model $\mathbb{P}$ was taken to be $y_k = m(x_k) + e_k$, $k = 1, ..., n$ with $e_1, ..., e_n$ independent errors drawn from some unknown probability distribution F_e.

(ii) Calculate $\hat{m}_n(x_k)$, and the estimated errors (residuals) are $\hat{e}_k = y_k - \hat{m}_n(x_k)$, from which was obtained an estimated version of $\hat{F}_e$: probability $\frac{1}{n}$ on $\hat{e}_k$.

(iii) Bootstrap data $\{y_k^*\}_{k=1}^n$ were generated according to $y_k^* = \hat{m}_n(x_k) + e_k^*$, with $e_1^*, ..., e_n^*$ independent errors drawn from $\hat{F}_e$ by Monte Carlo.

(iv) Having generated $\{y_k^*\}_{k=1}^n$, calculate the bootstrap estimates $\hat{m}_n^*(x_k)$.

(v) This whole process must be repeated B times.

External bootstrap

However, the paired bootstrap generating (x_k^*, y_k^*) from the empirical distribution $\hat{F}_{XY}$ of the pairs $(x_1, y_1), ..., (x_n, y_n)$ works for linear models, nonlinear models, generalized linear models and Cox's regression model, but it does not work for nonparametric regression models (Härdle, 1989). The bootstrap distribution estimator based on $(\hat{m}_n^* - \hat{m}_n)$ is inconsistent. A bias correction is needed as (Härdle and Bowman, 1988) did in the bootstrap based on residuals. Härdle (Härdle, 1989) considered the application of the external or wild bootstrap. An example where bootstrap breaks down and where wild bootstrap works is given in (Härdle and Mammen, 1993). Further discussions of wild bootstrap can be found in (Liu, 1988), (Liu and Singh, 1992), (Zheng and Tu, 1988) and (Mammen, 1992a, 1992b). Here $x_1, ..., x_n$ are deterministic or random, $e_1, ..., e_n$ are independent with mean zero and it is not assumed that they have the same distribution.

Algorithm 15 *(The external (wild) bootstrap).*

(i) The unknown probability model $\mathbb{P}$ was taken to be $y_k = m(x_k) + e_k$, with $e_1, ..., e_n$ independent errors drawn from some unknown probability distribution F_e.

(ii) Calculate $\hat{m}_n(x_k)$, and the estimated errors (residuals) are $\hat{e}_k = y_k - \hat{m}_n(x_k)$.

(iii) Draw the bootstrap residuals $\hat{e}_k^*$ *from a two-point centered distribution in order that its second and third moment fit the square and the cubic power of the residual* $\hat{e}_k$*. For instance, the distribution of* $\hat{e}_k^*$ *could be* $\eta I_{[a\hat{e}_k]} + (1-\eta) I_{[b\hat{e}_k]}$ *with* $\eta = \frac{5+\sqrt{5}}{10}$, $a = \frac{1-\sqrt{5}}{2}$, $b = \frac{1+\sqrt{5}}{2}$ *and* $\delta_{[x]}$ *being the Dirac measure at* x*. Alternatively, one can choose* $\hat{e}_k^*$ *distributed as* $\hat{e}_k^* = \hat{e}_k \left(\frac{Z_1}{\sqrt{2}} + \frac{Z_2^2 - 1}{2} \right)$, *with* Z_1 *and* Z_2 *being two independent standard normal random variables, also independent of* $\hat{e}_k$.

(iv) Having generated $\{y_k^*\}_{k=1}^n$, *calculate the bootstrap estimates* $\hat{m}_n^*(x_k)$.

(v) This whole process must be repeated B *times.*

Part II

APPLICATIONS of LS-SVM REGRESSION MODELLING

Chapter 6

LS-SVM for Regression Estimation

Direct estimation of high dimensional nonlinear functions using a non-parametric technique without imposing restrictions faces the problem of the curse of dimensionality (Bellman, 1961). Several attempts were made to overcome this obstacle, including projection pursuit regression (Friedmann and Stuetzle, 1981) and additive regression modeling (Hastie and Tibshirani, 1990). These methods and their extensions have become one of the widely used nonparametric techniques as they offer a compromise between the somewhat conflicting requirements of flexibility, dimensionality and interpretability.

In this chapter we begin our discussion of some manifestations of the curse of dimensionality. The LS-SVM regression modeling, introduced in Chapter 3, is discussed in this context. We consider analysis-of-variance (ANOVA) decompositions and then introduce some structure by eliminating some of the higher-order terms. Additive models assume only main effect terms. We describe iterative backfitting algorithms for fitting LS-SVM regression models. We introduce a new method, componentwise LS-SVM, for the estimation of additive models consisting of a sum of nonlinear components (Pelckmans *et al.*, 2004). New contributions are made in Section 6.3.

6.1 Low dimensional examples

Given a training set $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$, the problem is to estimate the regression function on test data $\mathcal{D}_{test} = \{(x_{n+1}, y_{n+1}), ..., (x_r, y_r)\}$.

(i) *Estimation of nonlinear regression functions.* Consider the following model

$$y = \frac{\sin(\|x\|_2)}{\|x\|_2} + e \tag{6.1}$$

where $x \in \mathbb{R}^d$. The values y_k are corrupted by noise with Normal distribution. Figure 6.1 and Figure 6.2 show the LS-SVM regression estimation ($x \in \mathbb{R}$ uni-

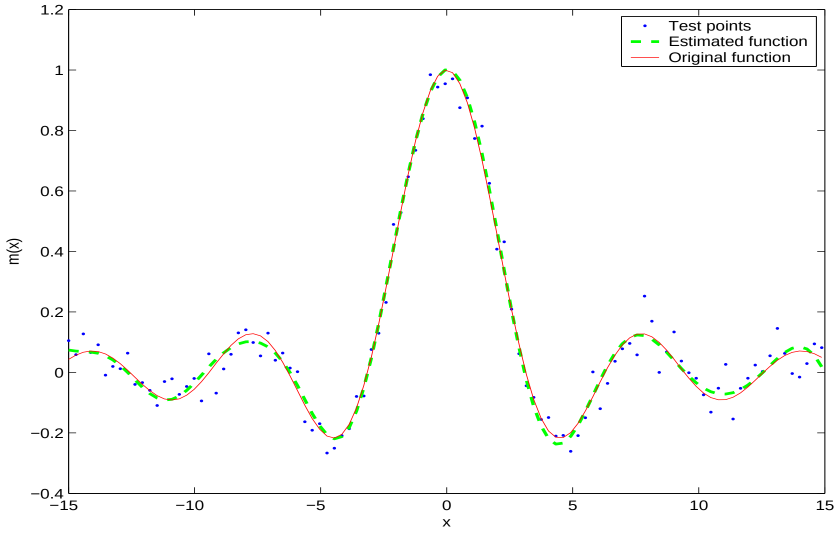

Figure 6.1: The regression function and its estimation obtained from the data (200 observations) with $\sigma = 0.1$.

formly distributed in the region $-15 \leq x \leq 15$) corrupted by different levels of noise. Figure 6.3 shows the LS-SVM regression estimation where $x \in \mathbb{R}^2$ is defined on a uniform lattice on the interval $[-15, 15] \times [-15, 15]$.

(ii) *Estimation of linear regression functions.* Consider the simple linear regression model

$$y = \beta_0 + \beta_1 + e, \tag{6.2}$$

where the values y_k are corrupted by noise with Normal distribution with parameters $N\left(0, 1^2\right)$. This model has one independent variable uniformly distributed in the region $0 \leq x \leq 1$. Figure 6.4 shows the LS-SVM regression estimation (linear kernel and RBF kernel) from 50 observations and the ordinary least square estimator.

6.2 Curse of dimensionality

Let X be i.i.d. distributed $\mathbb{R}^d$-valued random variable. If X takes values in a high dimensional space (e.g. d is large), estimation the regression function is difficult (Vapnik, 1998). The reason for this is that in the case of large d it is not possible to densely pack the space of X with finitely many sample points, even if the sample size n is very large. This fact is often referred to as the "curse of dimensionality" (Bellman, 1961). There are many manifestations of this problem, and we will examine a few here.

Consider a set of functions $\mathcal{F} \in C^s\left([0,1]^d\right)$. For any function $m \in \mathcal{F}$ the

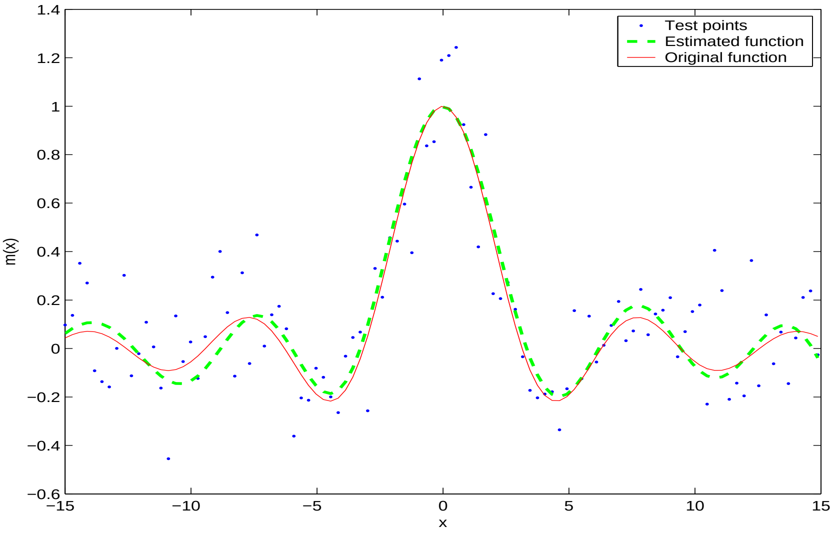

Figure 6.2: The regression function and its estimation obtained from the data (200 observations) with $\sigma = 0.2$.

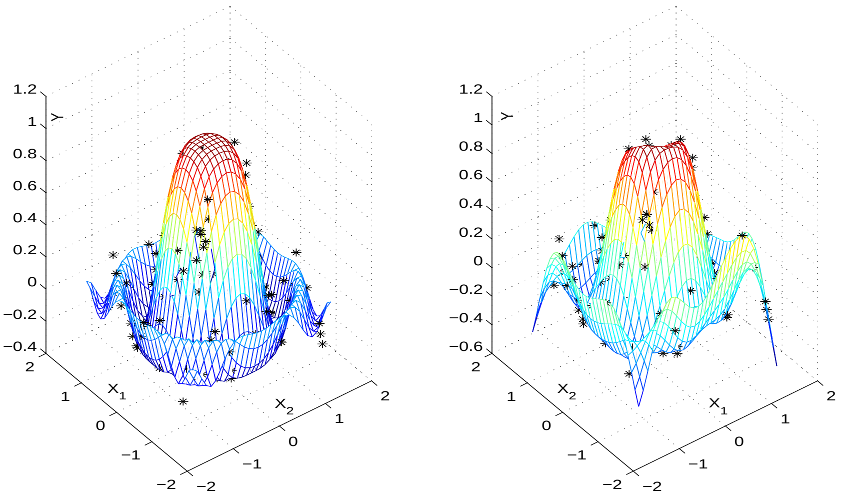

Figure 6.3: The regression function and its estimation obtained from the data (120 observations) with σ=0.1.

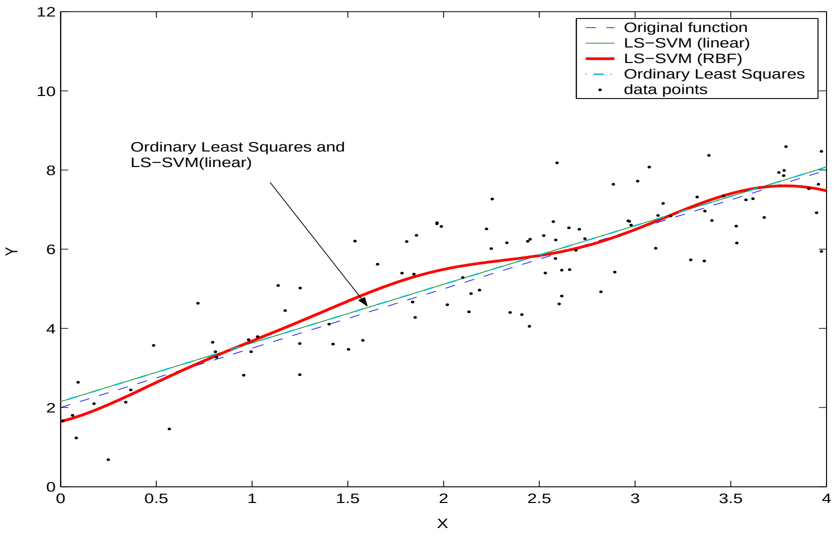

Figure 6.4: The regression function and its estimations (LS-SVM with linear kernel, LS-SVM with RBF kernel and ordinary least squares) obtained from the data with $\sigma = 1$.

optimal minimax rate of convergence, $n^{-\frac{2s}{2s+d}}$ (Vapnik, 1998; Györfi *et al.*, 2002), for the estimation $\hat{m}_n$ converges to zero rather slowly if the dimension d is large compared to s. The asymptotic rate of convergence decreases with increasing input dimension when the characteristic of smoothness remains fixed (Vapnik, 1998). Therefore, one can guarantee good estimation of a high-dimensional function only if the function $m(x) \in \mathcal{F}$ with $s \rightarrow \infty$ (extremely smooth).

6.2.1 Geometry of higher dimensions

The geometry of higher dimensions and statistical concepts is available in the book by (Kendall, 1961), *A course in the geometry of d dimensions*. The book gives numerous examples in which common statistical concepts are explained by geometrical constructs. Some interesting consequences are:

(i) *Tail probabilities of multivariate Normal.* Assume the data, $\{x_k\}_{k=1}^n$ and $x_k \in \mathbb{R}^d$, follow the standard d-dimensional normal distribution. The origin (mode) is the most likely point and the contours of equal probability are spheres,

$$\frac{f(x)}{f(0)} = \exp\left(-\frac{1}{2}x^T x\right) \qquad \text{and} \qquad -2\log\frac{f(x)}{f(0)} = \sum_{i=1}^{d} x_i^2 \sim \chi_d^2, \qquad (6.3)$$

where $f(x) = (2\pi)^{-\frac{p}{2}} \exp\left(-\frac{1}{2}x^T x\right)$. The probability that a point is within the

$a\backslash d$	1	3	5	7	9	10	15	20
0.01	0.0024	0.0266	0.1010	0.2379	0.4181	0.5123	0.8663	0.9803
0.02	0.0052	0.0498	0.1662	0.3484	0.5520	0.6460	0.9306	0.9930
0.05	0.0144	0.1120	0.3070	0.5407	0.7408	0.8160	0.9799	0.9989

Table 6.1: Around dimension 5 (a=0.01), the probability mass of a multivariate Normal density starts rapidly migrating to the tails.

$a\%$ spherical contour may be computed by

$$\Pr\left(\frac{f(x)}{f(0)} \leq a\right) = 1 - \Pr\left(\chi_d^2 \leq -2\log a\right). \tag{6.4}$$

Equation (6.4) gives the probability that a random point will fall within the tails. In Table 6.1, these probabilities are tabulated for several dimensions and a. The consequence is that an observation in higher dimensions is more likely to appear to be an outlier than one in lower dimensions. In a multivariate distribution whose density is the product of identical univariate densities, the relative probability content within extreme regions becomes larger as the dimension increases.

(ii) Given a uniformly distributed unit hypercube $[0,1]^d$ in d-dimensions. Let $x_0 \in \mathbb{R}^d$ be a point of the unit hypercube and let ζ denote a fraction of the observations. Consider a hypercubical neighborhood $\mathcal{A}_{x_0}$ about x_0, the fraction of the volume of the observations contained in the neighborhood is given by

$$\vartheta\left(d,\zeta\right) = \frac{\text{volume unit hypercube } \left[-0,1\right]^d}{\text{volume } \mathcal{A}_{x_0}} = \frac{2\left(1\right)^d}{2\zeta^d} = \zeta^{\frac{1}{d}}. \tag{6.5}$$

The fraction of the observations is shown in Figure 6.5. In dimension 5, to capture 1% of the observations, $\vartheta\left(5, 0.01\right) = 0.40$. To form a "local" neighborhood, we must cover 40% of the range of each input variable.

6.2.2 Dimension reduction techniques

The estimation of a regression function is difficult if the dimension d of the input variable $x \in \mathbb{R}^d$ is large. Circumventing the curse of dimensionality can be done by imposing additional assumptions on the regression functions. Consider the classical linear regression model

$$m\left(x\right) = \sum_{i=1}^{d} \beta_i x^{(i)} \tag{6.6}$$

where $\beta \in \mathbb{R}^d$. This restrictive parametric assumption can be generalized in several ways.

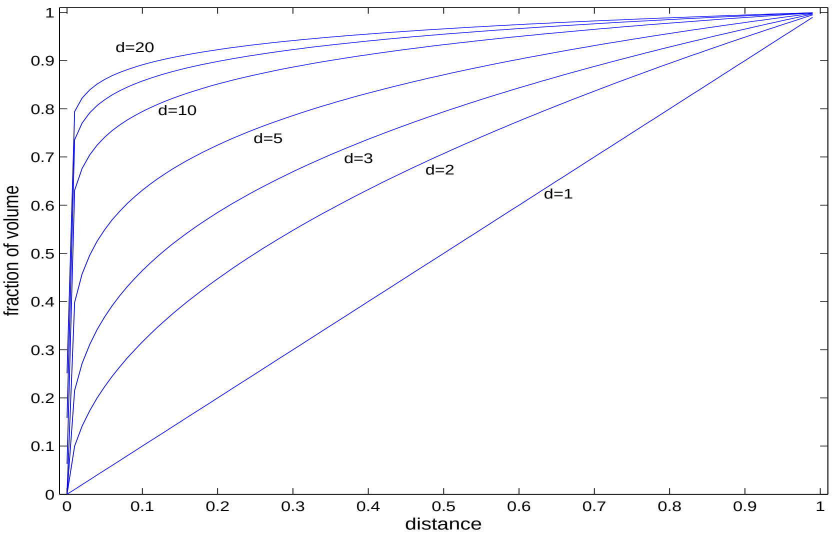

Figure 6.5: The edge length of the neighborhood hypercube needed to capture a fraction of the volume of the observations, for different dimensions.

(i) For *additive models*, one assumes that $m(x)$ is a sum of univariate functions $m_i : \mathbb{R} \rightarrow \mathbb{R}$ applied to the components of x, i.e.,

$$m\,(x) = \sum_{i=1}^{d} m_i(x^{(i)}). \tag{6.7}$$

The additive model and its generalization have been investigated by (Breiman and Friedman, 1981; Hastie and Tibshirani, 1990; Kohler, 1998). This assumption will be used to simplify the problem of regression estimation.

(ii) In *projection pursuit*, ones assumes that $m(x)$ is a sum of univariate functions $m_i : \mathbb{R} \rightarrow \mathbb{R}$ applied to projections of x onto various directions $a_i \in \mathbb{R}^d$

$$m\,(x) = \sum_{i=1}^{d} m_i(x, a_i). \tag{6.8}$$

Model (6.8) is an additive model, but in the derived features (x, a_i). Projection pursuit was proposed by (Friedman and Tukey, 1974) and specialized to regression estimation by (Friedman and Stuetzle, 1981). Note that a neural network model with one hidden layer is of the same form as the projection pursuit model (Hastie *et al.*, 2001).

(iii) *Tree-based* methods (Breiman *et al.*, 1984) partition the input space in regions and fit a simple model in each region. A modification of tree-based methods is the multivariate adaptive regression splines (Friedman, 1991) and a variant of tree-based methods is the hierarchical mixtures of experts (Jordan and Jacobs, 1994).

6.3 Additive LS-SVM regression modelling

6.3.1 Backfitting

There are several ways to approach estimation of additive models. The backfitting algorithm (Friedman and Stuetzle, 1981) and (Hastie and Tibshirani, 1990) is a general algorithm that enables to fit an additive model using any regression-type fitting mechanism. Consider the following additive model

$$y = a + \sum_{i=1}^{d} m_i \left(x^{(i)} \right) + e, \tag{6.9}$$

where the errors e are assumed to be independent of the $x^{(i)}, E[e] = 0$ and the $m_i, i = 1, ..., d$ are univariate functions, one for each independent variable. Implicit in (6.9) is the assumption that $E\left[m_i\left(x^{(i)}\right)\right] = 0$, since otherwise there will be free constants in each of the functions (Hastie and Tibshirani, 1990). Based on the conditional expectation

$$E\left[y - a - \sum_{i \neq j} m_i \left(x^{(i)} \right) \left| X = x^{(j)} \right. \right] = m_j \left(x^{(j)} \right), \quad \forall j = 1, ..., d, \tag{6.10}$$

the following iterative algorithm for computing all m_j is given by

Algorithm 16 *Backfitting algorithm. Let $\mathcal{T} : \mathbb{R} \rightarrow \mathbb{R}$ be a smooth operator and let $\mathcal{T}_l \left[y \left| X = x^{(l)} \right. \right]$ denotes a smooth estimation of y given $X = x^{(l)}$.*

(i) *Initialize*

$$\hat{a} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ and } \hat{m}_{n,i}^{[0]} \left(x^{(i)} \right), \ i = 1, ..., d \tag{6.11}$$

where $\hat{m}_{n,i}^{[0]} \left(x^{(i)} \right), \ i = 1, ..., d$ are fits resulting from a linear regression $E \left[y \left| X = x^{(i)} \right. \right]$.

(ii) *For each $i = 1, ..., d$, obtain*

$$\hat{m}_{n,i}^{[q]} \left(x^{(i)} \right) = \mathcal{T}_l \left[\left\{ y_k - \hat{a} - \sum_{i \neq l} \hat{m}_{n,i}^{[q-1]} \left(x_k^{(i)} \right), \ k = 1, ..., n \right\} \left| X = x_k^{(i)} \right. \right] \tag{6.12}$$

and

$$\hat{m}_{n,i}^{*[q]} \left(x^{(i)} \right) = \hat{m}_{n,i}^{(q)} \left(x^{(i)} \right) - \frac{1}{n} \sum_{k=1}^{n} \hat{m}_{n,i}^{(q)} \left(x_k^{(i)} \right). \tag{6.13}$$

(iii) *Repeat (ii) until convergence.*

Dimension	MSE	R^2
1	0.04322	0.91469
2	0.00506	0.90019
3	0.05280	0.89530
4	0.06481	0.87170
5	0.06571	0.87130
10	0.07880	0.84315
15	0.09417	$0,81891$
20	0.098867	$0,80629$

Table 6.2: Result of applying LS-SVM (RBF kernel) on 10.000 test data in function of the input dimension.

6.3.2 Simulation examples

Example 1

Consider the following nonlinear regression model defined as

$$y_k = \frac{\sin\left(2\pi \left\|x\right\|_2\right)}{2\pi \left\|x\right\|_2} + e_k,\ k = 1, ..., 100 \tag{6.14}$$

where the values y_k are corrupted by noise with Normal distribution $\mathcal{N}\left(0, 0.2^2\right)$ and independent variables $x^{(1)}, ..., x^{(10)}$ are uniformly distributed in the region $\left\|x\right\|_2 \leq 1$. Table 6.2 shows the MSE (on the test data) as a function of the dimension and $R^2 = 1 - \frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{\sum_{k=1}^n (y_k - \bar{y})^2}$ a measure of fit.

Example 2

This example describes experiments with LS-SVM in estimating linear regression functions. We compare the LS-SVM (linear kernel and RBF kernel) with ordinary least squares and additive LS-SVM (based on backfitting). Consider the linear regression estimation from the data set $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $x_k \in \mathbb{R}^{10}$ and $y_k \in \mathbb{R}$. The regression model depends only on two coordinates

$$y_k = 2x_k^{(1)} + x_k^{(2)} + 0\sum_{l=3}^{10} x_k^{(l)} + e_k \tag{6.15}$$

where the values y_k are corrupted by noise with Normal distribution with parameters $\mathcal{N}\left(0, 0.5^2\right)$ and independent variables $x^{(1)}, ..., x^{(10)}$ are uniformly distributed in the region $0 \leq x \leq 1$. Table 6.3 shows that the ordinary least squares, LS-SVM (linear kernel) and backfitting LS-SVM (RBF kernel) give the same results, which is 10% better than the LS-SVM (RBF kernel).

Ordinary Least Squares	LS-SVM (linear kernel)	LS-SVM (RBF kernel)	Backfitting LS-SVM (RBF kernel)
0.0042	0.0042	0.0047	0.0042

Table 6.3: Results of comparision ordinary least squares, LS-SVM (linear kernel and RBF kernel), and additive LS-SVM (based on backfitting).

	LS-SVM (RBF kernel)	Additive LS-SVM(RBF kernel) Backfitting
Model 1	0.6881	0.3368
Model 2	0.1554	0.1172

Table 6.4: Results on test data of numerical experiments on the data sets of example 3. Comparision of LS-SVM (RBF kernel) with additive LS-SVM (RBF kernel).

Example 3

This example describes experiments with LS-SVM in estimating nonlinear regression functions. We compare the LS-SVM (RBF kernel) to additive LS-SVM (based on backfitting). For these regression estimation experiments we chose the following regression functions (see Vapnik, 1998):

(i). "Model 1" (suggested by Friedman (1991)) considered the following nonlinear regression function of 10 variables

$$y_k = 10\sin(\pi x_k^{(1)}) + 20\left(x_k^{(2)} - 0.5\right)^2 + 10x_k^{(3)} + 5x_k^{(4)} + 0\sum_{l=5}^{10} x_k^{(l)} + e_k. \quad (6.16)$$

This function depends on only 5 variables. In this model the 10 variables are uniformly distributed in the region $0 \leq x \leq 1$ and the noise is normal with parameters $\mathcal{N}(0, 1^2)$.

(ii). "Model 2" considered the following nonlinear regression function of 10 variables

$$y_k = 10\sin(\pi x_k^{(1)} x_k^{(2)}) + 20\left(x_k^{(3)} - 0.5\right)^2 + 10x_k^{(4)} + 5x_k^{(5)} + 0\sum_{l=6}^{10} x_k^{(l)} + e_k. \quad (6.17)$$

This function depends on only 6 variables. In this model the 10 variables are uniformly distributed in the region $0 \leq x \leq 1$ and the noise is Normal with parameters $\mathcal{N}(0, 1^2)$. Note that 6.17 is not an additive model. Table 6.4 shows that backfitting LS-SVM (RBF kernel) outperforms the LS-SVM (RBF kernel).

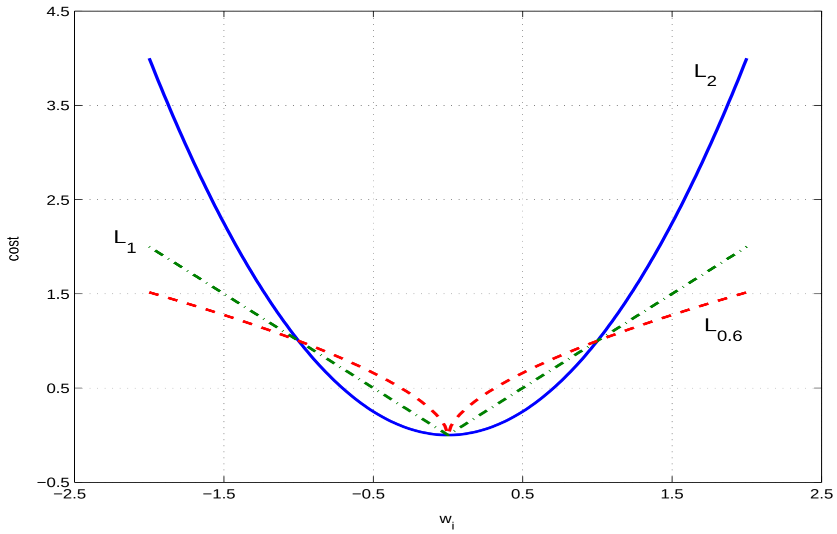

Figure 6.6: The L_p penalty family for $p = 2, 1$ and 0.6.

6.3.3 Componentwise LS-SVM regression modelling

Consider the regularized least squares cost function defined as

$$\mathcal{J}_\lambda\left(w^{(i)}, e\right) = \frac{\lambda}{2}\sum_{i=1}^{d} L\left(w^{(i)}\right) + \frac{1}{2}\sum_{k=1}^{n} e_k^2, \tag{6.18}$$

where $L(w^{(i)})$ is a penalty function and $\lambda \in \mathbb{R}_0^+$ acts as a regularization parameter. We denote $\lambda L(\cdot)$ by $L_\lambda(\cdot)$, so it may depend on λ. Examples of penalty functions include:

(i) The L_p penalty function $L_\lambda^p\left(w^{(i)}\right) = \lambda \left\|w^{(i)}\right\|_p^p$ leads to a bridge regression (Frank and Friedman, 1993; Fu, 1998). It is known that the L_2 penalty function results in the ridge regression. For the L_1 penalty function the solution is the soft thresholding rule (Donoho and Johnstone, 1994). LASSO, as proposed by (Tibshirani, 1996; Tibshirani, 1997), is the penalized least squares estimate using the L_1 penalty function (see Figure 6.6).

(ii) When the penalty function is given by

$$L_\lambda\left(w^{(i)}\right) = \lambda^2 - \left(\left\|w^{(i)}\right\|_1 - \lambda\right)^2 I_{\left\{\left\|w^{(i)}\right\|_1 < \lambda\right\}}$$

(see Figure 6.7), the solution is a hard-thresholding rule (Antoniadis, 1997).

As an example, the regularized least squares cost function with L_2 penalty

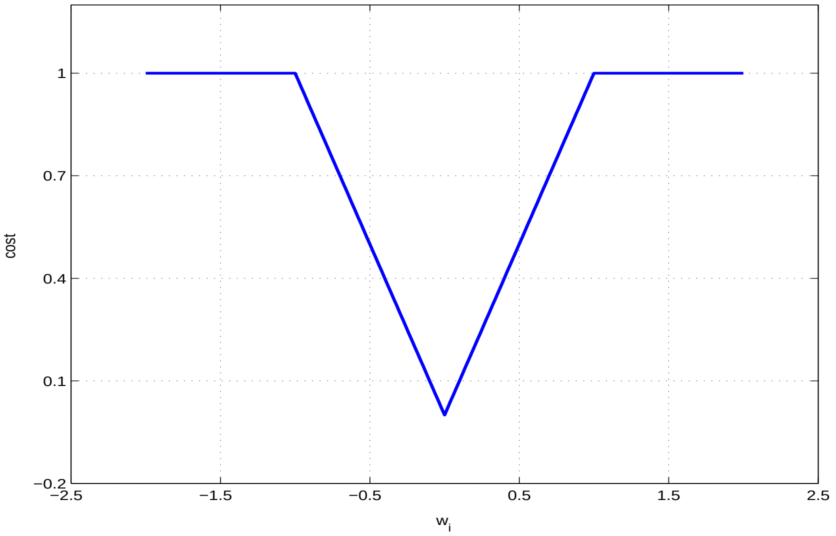

Figure 6.7: Hard thresholding penalty function.

function is given as (Suykens *et al.*, 2002)

$$\min_{w_i,b,e_k} \mathcal{J}\left(w_i,e\right)=\frac{1}{2}\sum_{i=1}^{d} w_i^T w_i+\frac{\gamma}{2}\sum_{k=1}^{n} e_k^2 \tag{6.19}$$

such that

$$y_k=\sum_{i=1}^{d} w^{(i)T}\varphi_i\left(x_k^{(i)}\right)+b+e_k,\ \ k=1,...,n.$$

To solve the optimization problem (in the dual space) one defines the Lagrangian functional

$$\mathcal{L}(w^{(i)},b,e;\alpha_k)=\mathcal{J}(w^{(i)},e)-\sum_{k=1}^{n}\alpha_k\left(\sum_{i=1}^{d} w^{(i)T}\varphi_i\left(x_k^{(i)}\right)+b+e_k-y_k\right). \tag{6.20}$$

By taking the conditions for optimality $\frac{\partial\mathcal{L}}{\partial w^{(i)}}=0,\ \frac{\partial\mathcal{L}}{\partial b}=0,\ \frac{\partial\mathcal{L}}{\partial e_k}=0,\ \frac{\partial\mathcal{L}}{\partial \alpha_k}=0$ and application of $K^i\left(x_k^{(i)},x_j^{(i)}\right)=\varphi_i\left(x_k^{(i)}\right)^T\varphi_i\left(x_j^{(i)}\right)$, the dual problem is summarized in matrix notation as

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega+\frac{1}{\gamma}I_n \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right]=\left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \tag{6.21}$$

with $y = (y_1, ..., y_n)^T$, $1_n = (1, ..., 1)^T$, $\alpha = (\alpha_1; ...; \alpha_n)^T$ and $\Omega \in \mathbb{R}^{n \times n}$ with $\Omega = \sum_{i=1}^{d} \Omega^i$ where $\Omega^i_{kj} = K^i \left(x_k^{(i)}, x_j^{(i)} \right)$ of all $k, j = 1, ..., n$. A new point $x \in \mathbb{R}^d$ can be evaluated as

$$\hat{m}_n (x) = \sum_{k=1}^{n} \hat{\alpha}_k \sum_{i=1}^{d} K^i \left(x^{(i)}, x_k^{(i)} \right) + \hat{b}, \tag{6.22}$$

where $\hat{\alpha}$ and $\hat{b}$ is the solution to (6.21).

The L_p and the hard thresholding penalty functions do not simultaneously satisfy the mathematical conditions for unbiasedness, sparsity and continuity (Fan and Li, 2001). The hard thresholding has a discontinuous cost surface. The only continuous cost surface (defined as the cost function associated with the solution space) with a thresholding rule in the L_p-family is the L_1 penalty function, but the resulting estimator is shifted by a constant λ. To avoid these drawbacks, (Nikolova, 1999) suggests the penalty function defined as

$$L_{\lambda,a} \left(w^{(i)} \right) = \frac{a\lambda \left\| w^{(i)} \right\|_1}{1 + a \left\| w^{(i)} \right\|_1}, \tag{6.23}$$

with $a \in \mathbb{R}$. This penalty function behaves quite similarly as the Smoothly Clipped Absolute Deviation (SCAD) penalty function as suggested by (Fan, 1997). The Smoothly Thresholding Penalty (TTP) function (6.23) improves the properties of the L_1 penalty function and the hard thresholding penalty function (see Figure 6.8), see (Antoniadis and Fan, 2001).

The unknowns a and λ act as regularization parameters. A plausible value for a was derived in (Nikolova, 1999; Antoniadis and Fan, 2001) as $a = 3.7$. The transformed L_1 penalty function satisfies the oracle inequalities (Donoho and Johnstone, 1994). One can plug-in the described semi-norm $L_{\lambda,a} (\cdot)$ to improve the component based regularization scheme (6.19). The componentwise regularization scheme is used for the emulation of this scheme

$$\min_{w^{(i)}, b, e_k} \mathcal{J} \left(w^{(i)}, e \right) = \frac{1}{2} \sum_{i=1}^{d} L_{\lambda,a} \left(w^{(i)} \right) + \frac{1}{2} \sum_{k=1}^{n} e_k^2 \tag{6.24}$$

such that

$$y_k = \sum_{i=1}^{d} w^{(i)T} \varphi_i \left(x_k^{(i)} \right) + b + e_k, \; k = 1, ..., n.$$

which becomes non-convex. For practical applications, the iterative approach is used for solving nonconvex cost-functions as (6.24) (Pelckmans *et al.*, 2004). The iterative approach is based on the graduated non-convexity algorithm as proposed in (Blake, 1989; Nikolova, 1999; Antoniadis and Fan, 2001) for the optimization of non-convex cost functions.

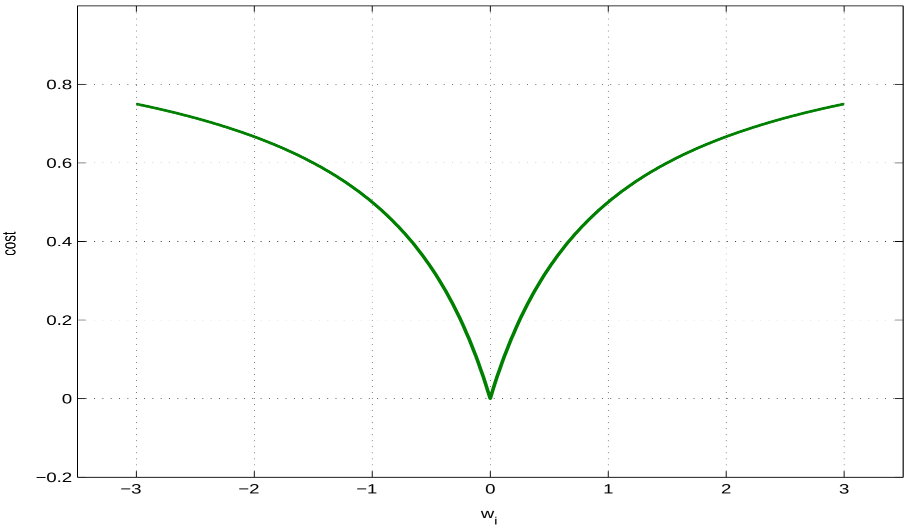

Figure 6.8: The transformed L_1 penalty function.

Method	Test Performance	Sparse components
LS-SVM (RBF kernel)	0.1110	0% recovered
Componentwise LS-SVM (6.19)	0.0603	0% recovered
STP and LS-SVM (6.24)	0.0608	100% recovered

Table 6.5: Results on test data of numerical experiments on the Friedman data set. The sparseness is expressed in the rate of components which is selected only if the input is relevant (100STP: Smootly thesholding penalized cost function.

6.3.4 Simulation examples

Simulation 1

To illustrate the additive model estimation method, a classical example was constructed as in (Friedman, 1991). The data were generated according to

$$y_k = 10\mathrm{sinc}(x_k^{(1)}) + 20(x_k^{(2)} - 0.5)^2 + 10x_k^{(3)} + 5x_k^{(4)} + 0\sum_{l=5}^{10} x_k^{(l)} + e_k, \quad (6.25)$$

were $e_k \sim \mathcal{N}(0,1)$, $n = 100$ and the input data X are randomly chosen from the interval $[0,1]^{10}$. The described techniques were applied on this dataset and tested on a test set. Furthermore, Table 6.5 reports whether the algorithm recovered the structure in the data (if so, the measure is 100%).

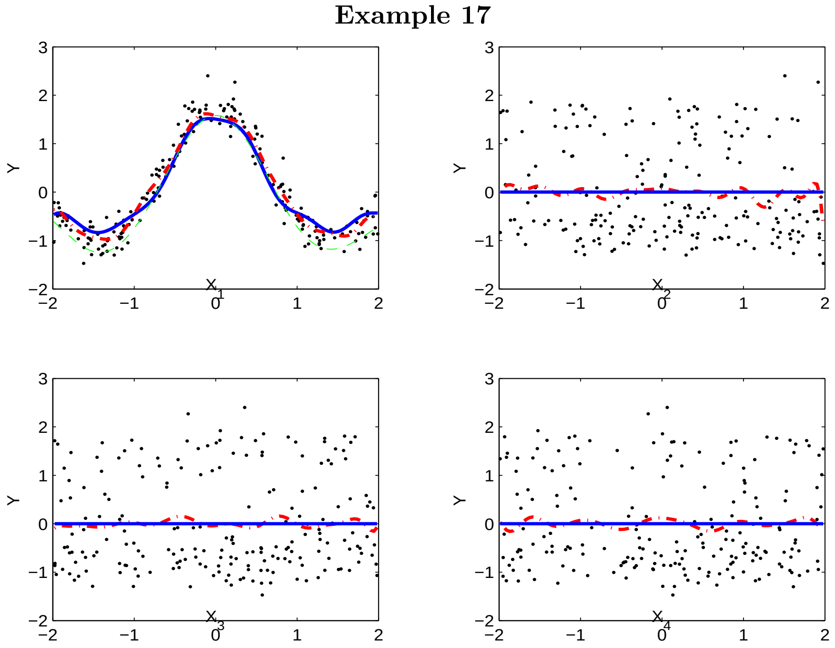

Figure 6.9: Example of a toy data set consisting of four input components $x^{(1)}, ..., x^4$ where only the first one is relevant to predict the output $y =$sinc$\left(x^{(1)}\right)$. A componentwise LS-SVM regressor (dashed line) has good prediction performance, while the L_1 penalized costfunction also recovers the structure in de data set as the estimated components corresponding with $x^{(2)}, x^{(3)}$ and $x^{(4)}$.

Simulation 2

The data were generated according to

$$y_k = \text{sinc}(x_k^{(1)}) + 0 \sum_{l=2}^{4} x_k^{(l)} + e_k,$$

were $e_k \sim \mathcal{N}(0,1)$, $n = 150$ and the input data $x^{(2)}, x^{(3)}$ and $x^{(4)}$ are white noise with $\sigma^2 = 0.1$. A componentwise LS-SVM regressor has good prediction performance, while the L_1 penalized costfunction also recovers the structure in de data set as the estimated components corresponding with $x^{(2)}, x^{(3)}$ and $x^{(4)}$ Figuur (6.9).

6.4 Conclusion

The iterative backfitting algorithm for fitting LS-SVM regression is simple, allowing one to choose a fitting method appropriate for each input variable. Important is that at any stage, one-dimensional kernel regression is all that is needed. Although consistency of the iterative backfitting algorithm is shown under certain conditions, an important practical problem (number of iteration steps) are still left. However the iterative backfitting algorithm (for large data problems) fits all input variables, which is not feasible or desirable when a large number are available. Table 6.4 shows that backfitting LS-SVM (RBF kernel) outperforms the LS-SVM (RBF kernel).

Recently we have developed a new method, componentwise LS-SVM, for the estimation of additive models consisting of a sum of nonlinear components (Pelckmans *et al.*, 2004). The method combines the estimation stage with structure detection. Advantages of using componentwise LS-SVMs include the efficient estimation of additive models with respect to classical practice, interpretability of the estimated model, opportunities towards structure detection and the connection with existing statistical techniques.

Chapter 7

Error Variance Estimation and its Application in Regression Modelling

Model-free estimators of the noise variance are important for doing model selection and setting learning parameters. In this chapter we generalize the the idea of the Rice estimator (Rice, 1984) for multivariate data based on U-statistics and differogram models (Pelckmans *et al.*, 2003). In the second part of this chapter we study the use of LS-SVM regression in the heteroscedastic case. Squared residual plots are proposed to assess heteroscedasticity in regression diagnostics. Contributions are made in Section 7.1 and Section 7.2.

7.1 Homoscedastic error variance

Consider the regression problem where we have observations $y_k \in \mathbb{R}$ at design points $x_k \in \mathbb{R}^d$ for $k = 1, ..., n$, and the observations are assumed to satisfy

$$y_k = m(x_k) + e_k, \qquad k = 1, ..., n. \tag{7.1}$$

The e_k values are assumed to be uncorrelated random variables with zero means and variance σ^2, and $m : \mathbb{R}^d \to \mathbb{R}$ a smooth function. A great deal of effort has gone into developing estimators of the underlying regression model m while the estimation of σ^2 has been relatively ignored. Estimation of σ^2 is also important, it has applications to interval estimation of m (inference) and to choose the amount of smoothing to be applied to the data. There are essentially two different approaches to the estimation of σ^2 : (i) Model based estimators and (ii) Another approach (model free estimators) to the problem of estimating σ^2 is to use the idea, common in time series analysis, of differencing the data to remove local trends effects. See, for example, (Rice, 1984; Gasser *et al.*, 1986; Pelckmans *et al.*, 2003).

7.1.1 Model based error variance estimation

Any estimator $\hat{m}_n$ of m can be used to estimate σ^2 by suitably normalizing its associated Residual Sums of Squares (RSS). See, for example, (Wahba, 1978, 1983) and (Cleveland, 1979). Consider a general class of variance estimators

$$\mathcal{V} = \left\{ \begin{array}{l} \sigma^2 : \hat{\sigma}^2 \text{ is quadratic in the data, } \hat{\sigma}^2 > 0 \text{ and} \\ E\left[\hat{\sigma}^2\right] = \sigma^2 \text{ if } m \text{ is a straight line} \end{array} \right\}. \tag{7.2}$$

The conditions restrict the class of estimators to the form

$$\hat{\sigma}^2 = \frac{y^T Q y}{\operatorname{tr}[Q]}, \tag{7.3}$$

with $y = (y_1, ..., y_n)^T$ and Q is a symmetric $n \times n$ positive semi-definite matrix (Buckley *et al.*, 1988)

Under model (7.1),

$$\begin{aligned} \hat{\sigma}^2 &= \frac{(m+e)^T Q (m+e)}{\operatorname{tr}[Q]} \\ &= \frac{\left(m^T Q m + 2m^T Q e + e^T Q e\right)}{\operatorname{tr}[Q]}, \end{aligned} \tag{7.4}$$

where $m = (m(x_1), ..., m(x_n))^T$ and $e = (e_1, ..., e_n)^T$. Hence quadratic estimates of the variance are made up of three parts: a positive bias, $\frac{m^T Q m}{\operatorname{tr}[Q]}$; a natural estimator of σ^2, $\frac{e^T Q e}{\operatorname{tr}[Q]}$; and a random perturbation, $\frac{2m^T Q e}{\operatorname{tr}[Q]}$.

An estimator of the above type can be obtained using the LS-SVM regression estimator. Recall that the LS-SVM regression estimator is of the class $\hat{m}_n(\theta) = S(\theta)\, y$, with $\theta \in \Theta$, representing some parameter vector set, and S is a smoother matrix. For the LS-SVM regression estimator the residual sum of squares can be written as

$$\begin{aligned} RSS(\theta) &= (y - \hat{m}_n)^T (y - \hat{m}_n) \\ &= y^T y - \hat{m}_n^T y - y^T \hat{m}_n + \hat{m}_n^T \hat{m}_n \\ &= y^T y - y^T S^T(\theta)\, y - y^T S(\theta)\, y + y^T S^T(\theta)\, S(\theta)\, y \\ &= y^T (I_n - S(\theta))^2\, y, \end{aligned} \tag{7.5}$$

where $\hat{m}_n = (\hat{m}(x_1), ..., \hat{m}(x_n))^T$. From (7.5) one can see that Q in (7.4) is equal to $(I_n - S(\theta))^2$. The variance estimator, based on LS-SVM regression, is now defined as

$$\hat{\sigma}^2 = \frac{y^T (I_n - S(\theta))^2\, y}{\operatorname{tr}\left[(I_n - S(\theta))^2\right]}. \tag{7.6}$$

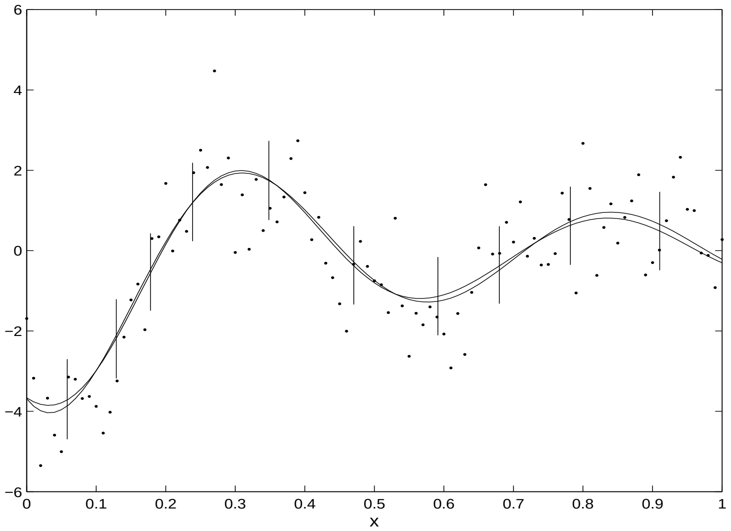

Figure 7.1: Error bars with 2 standard error bands.

7.1.2 Applications

(i) An application of *inference*, where $\hat{\sigma}^2$ can be used, will be given. Consider a simple example where $m(X) = \frac{\cos(12(X+0.2))}{X+0.2}$ with $X \sim U[0,1]$ and $e \sim \mathcal{N}\left(0, 0.5^2\right)$. The training sample consists of $n = 200$ pairs (x_k, y_k) drawn independently from this model. The error bars (given for some training points) in Figure 7.1 represent the pointwise standard error of $\hat{m}_n$, that is, the region between $\hat{m}_n(x_k) \pm 2\sqrt{Var[\hat{m}_n(x_k)]}$. Since $\hat{m}_n(x,\theta) = S(x,\theta)\, y$,

$$
\begin{aligned}
Cov[\hat{m}_n] &= Cov[S(\theta)\, y] \\
&= S(\theta)\, Cov[y]\, S(\theta)^T \\
&= S(\theta)\, \hat{\sigma}^2 I_n S(\theta)^T. \qquad (7.7)
\end{aligned}
$$

The diagonal of the variance-covariance matrix contains the pointwise variances at the training point x_k.

Note that the standard error only reflects the difference between the $\hat{m}_n$ and the mean of $\hat{m}_n$ and not the difference between $\hat{m}_n$ and the m. The latter requires knowledge of the training bias. To derive standard error for the prediction, the equation (7.7), has to be modified to reflect both the additive noise in the sample at the new point and the prediction error of the estimator.

(ii) *Tuning parameter selection.* Suppose, for example, that $y = (y_1, ..., y_n)^T$ is an $n \times 1$ vector of observations, $X_{\mathcal{D}}$ is an $n \times d$ matrix of independent variables. In the linear regression setting, the independent variables are related to the response by $y = Z\beta + e$, where $Z = (\mathbf{1}_n \; X_{\mathcal{D}})$ is the $n \times (d+1)$ design matrix including the constant term, β is a $(d+1)$ dimensional vector of parameters and e is an n dimensional vector of errors such that $E[e] = [0]$ and $Var[e] = \sigma^2$. In the standard application of ridge regression (Hoerl and Kennard, 1970), the constant term (intercept) of the fitted model is forced to be the mean of y by centering all columns of the $X_{\mathcal{D}}$ matrix about their mean values. In addition, the new independent variable columns are, typically, rescaled. The ridge regression estimator is then given by

$$\hat{\beta}_{ridge} = \left(\bar{y}, \left(A^T A + cI_d\right)^{-1} A^T y\right)^T, \tag{7.8}$$

where A is the appropriate $n \times d$ "$X_{\mathcal{D}}$-matrix" after the centering and rescaling procedure have been applied and $c \geq 0$ (often called the shrinkage parameter) Several properties of this estimator justify its consideration as an alternative to the least squares estimator. For example, Hoerl and Kennard (1970) show that there exists a range of c values for which the total mean squared error for the ridge estimator is smaller than the corresponding least squares quantity. However, c depends on the unknown parameters β_{ridge} and σ^2 and as such cannot be determined (Thistead, 1978). Hoerl, Kennard and Baldwin (1975) propose the use of

$$c = \frac{d\hat{s}^2}{\hat{\beta}_{LS}^T \hat{\beta}_{LS}}, \tag{7.9}$$

where $\hat{\beta}_{LS}$ is a least squares parameter vector, $\hat{s}^2$ is an estimate of σ^2 defined by

$$\hat{s}^2 = \frac{\left(y - Z\hat{\beta}_{LS}\right)^T \left(y - Z\hat{\beta}_{LS}\right)}{n-d-1}. \tag{7.10}$$

Monte Carlo evaluation of some ridge estimators are given by (Wichern and Churchill, 1977; McDonald and Galarneau, 1975; Beverley, 1980; Hoerl *et al.*, 1986). Setting tuning parameters in the nonparametric regression context depend also on the error variance. Note that model based error variance estimators can not be used because the estimators $\hat{s}^2$ depend on the tuning parameters, see for example (7.6).

(*iii*) *Model selection.* In the linear regression setting, a model selection problem is a subset selection problem. One of the goals of subset selection procedures is consistent selection i.e., picking the true underlying submodel with probability tending to 1 as the sample size gets large. Let $\mathcal{I}$ be any non-empty subset of the d independent variables. The Final Prediction Error (FPE) criterion (Akaike, 1970) can be used to estimate the prediction error and is defined as

$$J_{\mathcal{I}}(\lambda) = \frac{1}{n} RSS_{\mathcal{I}} + \lambda \left(\frac{d_{\mathcal{I}}}{n}\right) \sigma_e^2, \tag{7.11}$$

for a fixed positive value of a cost complexity parameter λ. The term σ_e^2 is estimated by $\hat{s}^2$ (see (7.10)), the unbiased estimate of σ_e^2 under the full model of degree d. The term $RSS_{\mathcal{I}}$ is the residual sum of squares for submodel $\mathcal{I}$ and $d_{\mathcal{I}}$ is the number of independent variables in $\mathcal{I}$.

In the nonparametric regression setting, a model selection problem consists of two parts: (a) tuning parameter selection, and (b) subset selection. Suppose we are only interested in tuning parameter selection. We can use a generalized Final Prediction Error (FPE) criterion given by

$$J_\theta(\lambda) = \frac{1}{n} RSS\,(\theta) + \lambda \left(\frac{\operatorname{tr}(S(\hat{\theta}))}{n} \right) \sigma_e^2, \tag{7.12}$$

where the matrix $S(\hat{\theta})$ is called the smoother matrix. Note that model based error variance estimators can not be used because the estimators $\hat{s}^2$ depend on the tuning parameters, see for example (7.6). Nonparametric estimates of the error variance are a solution of this problem.

7.1.3 Model free error variance estimators

Rice (1984) and Gasser, Sroka & Jennen-Steinmetz (1986) have proposed estimators of σ^2 based on first- and second-order differences of the y_k's, respectively. For example Rice suggested estimating σ^2 by

$$\hat{\sigma}^2 = \frac{1}{2\,(n-1)} \sum_{k=1}^{n-1} \left(y_{k+1} - y_k \right)^2 . \tag{7.13}$$

Gasser, Sroka & Jennen-Steinmetz has suggested a similar idea for removing local trend effects by using

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{k=2}^{n-1} c_k^2 \hat{\varepsilon}_k^2, \tag{7.14}$$

where $\hat{\varepsilon}_k$ is the difference between y_k and the value at x_k of the line joining (x_{k-1}, y_{k-1}) and (x_{k+1}, y_{k+1}) . The c_k are chosen to ensure that $E\left[c_k^2 \hat{\varepsilon}_k^2\right] = \sigma^2$ for all k when the function m of (7.1) is linear. Note that one assumes that $x_1 < ... < x_n$, $x_k \in \mathbb{R}$ in both methods. Next we will generalize the idea of (7.13) for multivariate data based on U-statistics.

U-statistic

Let $X_1, ..., X_n$ be independent observations on an unknown probability distribution function F_X. Let $T(F) = \int \left(x - E\left[X\right]\right)^2 dF(x)$ be the statistic of interest. For any unbiased estimator $T(\hat{F}_n)$ of $T(F)$ there exists a U-statistic U_n estimating $T(F)$ based on the same n observations such that

$$Var\left[U_n\right] \leq Var\left[T(\hat{F}_n)\right] \tag{7.15}$$

with equality if and only if $U_n = T(\hat{F}_n)$. That means U-statistics are best in variance under all unbiased estimators (Lee, 1990).

Definition 18 *(U-statistic). Let $g : \mathbb{R}^l \to \mathbb{R}$ be a measurable and (without loss of generality) symmetric function. The function*

$$U_n = U\left(g; X_1, ..., X_n\right) = \frac{1}{\binom{n}{l}} \sum_{1 \leq i_1 < ... < i_l \leq n} g\left(X_{i_1}, ..., X_{i_l}\right), \quad l < n, \tag{7.16}$$

where $\sum_{1 \leq i_1 < ... < i_l \leq n}$ denotes the summation over the $\binom{n}{l}$ combinations of l distinct elements $\{i_1, ..., i_l\}$ from $\{1, ..., n\}$, is called U-statistic of degree l with kernel g.

Example 19 *(i) $T(F) = \int x dF(x)$. For the kernel $g(x) = x$, the corresponding U-statistic is*

$$U_n = U\left(g; X_1, ..., X_n\right) = \frac{1}{n} \sum_{k=1}^{n} x_k \tag{7.17}$$

the sample mean.

(ii) $T(F) = \int \left(x - E\left[X\right]\right)^2 dF(x)$. For the kernel $g\left(x_1, x_2\right) = \frac{x_1^2 + x_2^2 - 2x_1 x_2}{2} = \frac{1}{2}\left(x_1 - x_2\right)^2$, the corresponding U-statistic is

$$\begin{aligned} U_n = U\left(g; X_1, ..., X_n\right) &= \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} g\left(x_i, x_j\right) \\ &= \frac{1}{n-1} \sum_{k=1}^{n} \left(x_k - \bar{x}\right)^2 \\ &= s^2, \end{aligned} \tag{7.18}$$

where $\bar{x} = \frac{1}{n} \sum_{l=1}^{n} x_l$ and s^2 is an unbiased estimator of $T(F) = \sigma^2$.

Based on (7.13) and motivated by (7.15) a weighted U-statistic can be written as

$$\begin{aligned} U_{n,v} = U_v\left(g; y_1, ..., y_n\right) &= \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} g\left(y_i, y_j\right) \mathcal{V}_{ij} \\ &= \frac{1}{n\left(n-1\right)} \sum_{1 \leq i < j \leq n} \frac{1}{2}\left(y_i - y_j\right)^2 \mathcal{V}_{ij}, \end{aligned} \tag{7.19}$$

where the random weights $\mathcal{V}_{ij}$ will be based on the independent variables $\left(x^{(1)}, ..., x^{(d)}\right)$ only.

Constructing weights based on density estimation

Consider the regression model as defined in (7.1). Assume that $e_1, ..., e_n$ are i.i.d. with common probability distribution function F belonging to the family

$$\mathcal{F} = \left\{F : \int x dF(x) = 0,\ 0 < \int \left|x\right|^r dF(x) < \infty\right\},\ r \in \mathbb{N}_0 \text{ and } 1 \leq r \leq 4. \tag{7.20}$$

The weights $\mathcal{V}_{ij}$ are non-negative, symmetric and average to 1. Müller, Schick and Wefelmeyer (2003) suggested an error variance estimator given by

$$\hat{\sigma}_e^2 = \frac{1}{n\left(n-1\right)h} \sum_{1 \leq i < j \leq n} \frac{1}{2}\left(y_i - y_j\right)^2 \frac{1}{2}\left(\frac{1}{\hat{f}_i} + \frac{1}{\hat{f}_j}\right) K\left(\frac{x_i - x_j}{h}\right), \tag{7.21}$$

where $\hat{f}_i$ is defined as

$$\hat{f}_i = \frac{1}{\left(n-1\right)h} \sum_{j=1, j \neq i} K\left(\frac{x_i - x_j}{h}\right), \; i = 1, ..., n. \tag{7.22}$$

Let $K : \mathbb{R}^d \to \mathbb{R}$ be a function called the kernel function and let $h > 0$ be a bandwidth or smoothing parameter. The cross-validation principle will be used to select the bandwidth h.

Constructing weights based on the estimated differogram

Definition 20 *(Semi-variogram). Let $\{Z_k, \; k \in \mathbb{N}\}$ be a random process with mean $\bar{z}$, $Var\left[Z_k\right] < \infty$ for all $k \in \mathbb{N}$ and correlation function which only depends on $\Delta x_{ij} = \left\|x_i - x_j\right\|_2$ for all $i, j \in \mathbb{N}$. It follows from the stationarity of the process $z_1, z_2, ...,$ that*

$$\begin{aligned} \frac{1}{2} E\left[\left(z_i - z_j\right)^2\right] &= \sigma^2 + \tau^2 \left(1 - h\left(\left\|x_i - x_j\right\|_2\right)\right) \\ &= \eta\left(\Delta x_{ij}\right), \; \forall i, j, \end{aligned} \tag{7.23}$$

where σ^2 is the variance of the measurement observations, τ^2 is the variance of the serial correlation component and $h(\cdot)$ is the correlation function (Diggle, 1990) and (Cressie, 1993). The function $\eta\left(\Delta x_{ij}\right)$ is called the semi-variogram, and it only depends on the points $x^{(i)}$ through $\Delta x_{ij} = \left\|x_i - x_j\right\|_2$.

In practice, the function $\eta\left(\Delta x\right)$ is estimated from the scatter plot of the half-squared differences $\frac{\left(z_i - z_j\right)^2}{2}$ versus the corresponding $\Delta x_{ij} = \left\|x_i - x_j\right\|_2$. This estimate is the sample semi-variogram $\hat{\eta}\left(\Delta x_{ij}\right)$.

This is illustrated in Figure (7.2) for the case of exponential correlation. Note that decreasing correlation function $h(\cdot)$ yield increasing semi-variograms $\eta\left(\Delta x\right)$, with $\eta\left(0\right) = \sigma^2$, which converge to $\sigma^2 + \tau^2$ as Δx grows to infinity.

Definition 21 *(Differogram). The differogram $\Upsilon : \mathbb{R} \to \mathbb{R}$ is defined as*

$$\Upsilon\left(\Delta x_{ij}\right) = \frac{1}{2} E\left[\Delta y_{ij} \left| \Delta x = \Delta x_{ij}\right.\right] \quad \textit{for } \Delta x \to 0, \tag{7.24}$$

where we define $\Delta x_{ij} = \left\|x_i - x_j\right\|_2$, $\Delta y_{ij} = \left\|y_i - y_j\right\|_2 \in \mathbb{R}^+$ which denote the differences of two input variables and of the corresponding output variables. Similar as in the variogram, the intercept $\frac{1}{2} E\left[\Delta y_{ij} \left| \Delta x = \Delta x_{ij} = 0\right.\right]$ gives the variance of the noise.

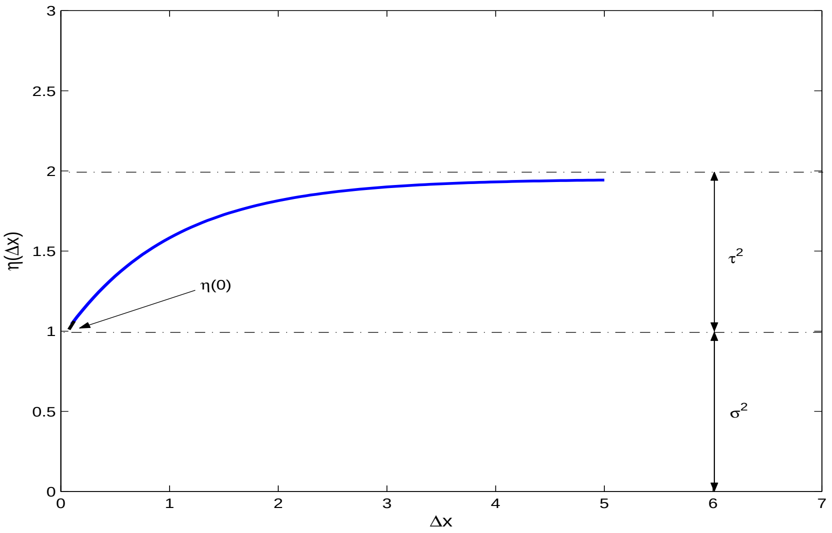

Figure 7.2: The semi-variogram with exponential correlation.

Differogram models based on Taylor series expansions Consider the one-dimensional Taylor series expansion of order r at center $x_i \in \mathbb{R}$

$$T_r\left(x_j - x_i\right) = m\left(x_i\right) + \sum_{l=1}^{r} \frac{1}{l!} \nabla^{(l)} m\left(x_j - x_i\right)^l + O\left(\left(x_j - x_i\right)^{r+1}\right), \qquad (7.25)$$

where $\nabla m\left(x\right) = \frac{\partial m}{\partial x}$, $\nabla^2 m\left(x\right) = \frac{\partial^2 m}{\partial x^2}$, etc. for $l \leq 2$.The rth order Taylor series approximation of the differogram model is considered with center $x_i = 0$ as one is interested only in the case $\Delta x \to 0$. The differogram is given by

$$\Upsilon\left(\Delta x, a\right) = a_0 + \sum_{l=1}^{r} a_l \Delta^l x, \quad a_0, ...a_r \in \mathbb{R}_+, \qquad (7.26)$$

where the parameter vector $a = \left(a_0, a_1, ..., a_r\right)^T \in \mathbb{R}_+^{r+1}$is assumed to exist uniquely. The parameter vector is enforced to be positive as the (expected) differences should always be strictly positive. The variance function ϑ of the estimator can be bounded as follows

$$\begin{aligned}\vartheta\left(\Delta x, a\right) &= E\left[\left(\Delta y - \Upsilon\left(\Delta x, a\right)\left|\Delta x\right.\right)^2\right] = \left[\left(\Delta y - a_0 - \sum_{l=1}^{r} a_l \Delta^l x \left|\Delta x\right.\right)^2\right] \\ &\leq E\left[\left(a_0 + \sum_{l=1}^{r} a_l \Delta^l x \left|\Delta x\right.\right)^2\right] + E\left[\left(\Delta y \left|\Delta x\right.\right)^2\right] \\ &= 2\left(a_0 + \sum_{l=1}^{r} a_l \Delta^l x\right)^2, \end{aligned} \tag{7.27}$$

where we apply the triangle inequality and the differogram model (7.24). This bound may be rather rough, but one has to keep in mind that the function $\Upsilon\left(\Delta x, a\right)$ only explains the data for $\Delta x \rightarrow 0$. Instead of deriving the parameter vector a from the (estimated) underlying function m, they are estimated immediately based on the observed differences Δx_{ij} and Δy_{ij} for $i < j = 1, ..., n$. The following weighted least squares method can be used

$$\hat{a} = \arg\min_{a \in \mathbb{R}_+^{r+1}} \mathcal{J}\left(a\right) = \sum_{i \leq j}^{n} \frac{c}{\vartheta\left(\Delta x_{ij}, a\right)} \left(\Delta y_{ij} - \Upsilon\left(\Delta x_{ij}, a\right)\right)^2, \tag{7.28}$$

where the constant $c \in \mathbb{R}_+^0$ normalizes the weighting function such that $\sum_{i \leq j}^{n} \frac{c}{\vartheta(\Delta x_{ij}, a)} = 1$. The function $\vartheta\left(\Delta x_{ij}, a\right) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ corrects for the heteroscedastic variance structure inherent to the differences. As the parameter vector a is positive, the weighting function is monotonically decreasing and as such represents always a local weighting function.

The differogram for noise variance estimation If the regression function m where known, the errors $e_k = y_k - m\left(x_k\right)$ were observable, the sample variance based on the errors can be written as

$$\begin{aligned}\hat{\sigma}_e^2 &= U\left(g; e_1, ..., e_n\right) \\ &= \frac{1}{n\left(n-1\right)} \sum_{1 \leq i < j \leq n} \frac{1}{2}\left(e_i - e_j\right)^2. \end{aligned} \tag{7.29}$$

But the regression function m is unknown and we have only the data $\mathcal{D}_n = \left\{\left(x_1, y_1\right), ..., \left(x_n, y_n\right)\right\}$. Based on the differogram, we can estimate the error variance. As an example, let $r = 0$, the 0th order Taylor polynomial of m centered at x_i and evaluated at x_j is given by $T_0\left(x_j - x_i\right) = m\left(x_i\right)$ and the variance estimate is

$$\begin{aligned}\hat{\sigma}_e^2 &= U\left(g; e_1, ..., e_n\right) \\ &= U\left(g; \left(y_1 - m\left(x_1\right), ..., \left(y_1 - m\left(x_1\right)\right)\right.\right. \\ &= \frac{1}{n\left(n-1\right)} \sum_{1 \leq i < j \leq n} \frac{1}{2}\left(y_i - y_j\right)^2. \end{aligned} \tag{7.30}$$

where the approximation improves as $x_i \to x_j$. To correct for this, one can use a kernel $g_1 : \mathbb{R}^2 \to \mathbb{R}$

$$g_1(y_i, y_j) = \frac{1}{2}\Delta y_{ij}\frac{c}{\vartheta(\Delta x_{ij})} \tag{7.31}$$

where the decreasing weighting function $\mathcal{V}_{ij} = \frac{1}{\vartheta(\Delta x_{ij})}$ is taken from (7.27). The constant $c \in \mathbb{R}_+^0$ is chosen such that the sum of the weighting terms are constant $2c\left(\sum_{i\leq j}^{n} \frac{1}{\vartheta(\Delta x_{ij})}\right) = n(n-1)$. The resulting variance estimator based on (7.31) and (7.24) becomes

$$\hat{\sigma}_e^2 = \frac{1}{n(n-1)} \sum_{1\leq i<j\leq n} \frac{1}{2}\left(\Delta y_{ij} - \sum_{l=1}^{r} a_l \Delta^l x\right)\frac{c}{\vartheta(\Delta x_{ij})} \tag{7.32}$$

7.1.4 Simulations

Simulation 1

Consider a small simulation to study the bias and the variance of the error variance estimators (for small to medium sample size). The following functional model was used:

$$m(x) = \frac{\cos(12(x+0.2))}{x+0.2},$$

where $x \sim U[0,1]$. The sample size was taken as $n = 50, 100, 300, 700, 900$ and the noise is normal with parameters $\mathcal{N}(0, \sigma^2)$. The results of the simulation are described in table 7.1, table 7.2 and table 7.3 based on 10 realizations. The third column gives the mean, the fourth the bias and the next the variance.

Figure 7.3 and 7.4 shows that the model based estimator and the model free estimator are consistent estimators ($n \to \infty$, $Var\left[\hat{\sigma}^2\right] \to 0$).

Simulation 2: Noise Variance estimation and Model Selection

To randomize the design of underlying functions in the experiment, we consider the following class of underlying functions

$$m(\cdot) = \sum_{k=1}^{n} \bar{\alpha} K(x_k, \cdot), \tag{7.33}$$

where $\bar{\alpha}$ is an i.i.d sequence of uniformly distributed terms. The kernel is fixed as $K(x_k, x_l) = \exp\left(-\|x_k - x_l\|_2^2\right)$ for any $k, l = 1, ..., n$. Datapoints were generated as $y_k = m(x_k) + e_k$ for $i = 1, ..., n$ where e_k are n i.i.d samples. The experiment compares results between using exact prior knowledge of the noise level, a model-free estimate using the differogram and using data-driven methods as V-fold cross-validation, leave-one-out, Mallows *Cp* statistic (Mallows, 1973; De Brabanter *et al.*, 2002). An important remark is that the method based on the differogram is orders of magnitudes faster than any data-driven method

σ^2	n	$mean(\hat{\sigma}^2)$	$bias\left(\sigma^2, \hat{\sigma}^2\right)$	$var\left(\hat{\sigma}^2\right)$
0.1	50	0.10284	0.00284	0.00045
	100	0.10183	0.00183	0.00024
	300	0.10087	0.00087	0.00007
	700	0.10101	0.00101	0.00003
	900	0.10005	0.00005	0.00002
0.25	50	0.24025	0.00975	0.00273
	100	0.25174	0.00174	0.00126
	300	0.24648	0.00352	0.00032
	700	0.25056	0.00056	0.00015
	900	0.25012	0.00012	0.00013
1.0	50	0.95742	0.04258	0.06735
	100	0.97867	0.02133	0.02275
	300	0.97999	0.02001	0.00778
	700	1.00343	0.00343	0.00376
	900	1.00124	0.00124	0.00197

Table 7.1: LS-SVM estimate of the error variance: mean, bias and variance for 100 replications

σ^2	n	$mean(\hat{\sigma}^2)$	$bias\left(\sigma^2, \hat{\sigma}^2\right)$	$var\left(\hat{\sigma}^2\right)$
0.1	50	0.10216	0.00216	0.00062
	100	0.09936	0.00064	0.00047
	300	0.10057	0.00057	0.00009
	700	0.10094	0.00094	0.00007
	900	0.09935	0.00065	0.00003
0.25	50	0.24398	0.00602	0.00499
	100	0.24756	0.00244	0.00211
	300	0.24824	0.00176	0.00072
	700	0.24971	0.00029	0.00035
	900	0.25221	0.00221	0.00031
1.0	50	0.97256	0.02744	0.10880
	100	0.98102	0.01898	0.03692
	300	0.96707	0.03293	0.01185
	700	$1,01817$	0.01817	0.00617
	900	0.99598	0.00402	0.00337

Table 7.2: Estimate (Gasser et al., 1986) of the error variance: mean, bias and variance for 100 replications

σ^2	n	$mean(\hat{\sigma}^2)$	$bias\left(\sigma^2, \hat{\sigma}^2\right)$	$var\left(\hat{\sigma}^2\right)$
0.1	50	0.10256	0.00744	0.00055
	100	0.10021	0.00021	0.00033
	300	0.10040	0.00040	0.00008
	700	0.10037	0.00037	0.00003
	900	0.10042	0.00042	0.00002
0.25	50	0.24516	0.00494	0.00382
	100	0.25438	0.00438	0.00158
	300	0.24892	0.00108	0.00036
	700	0.25235	0.00235	0.00016
	900	0.25041	0.00041	0.00014
1.0	50	1.04489	0.04489	0.15244
	100	1.03406	0.03406	0.03139
	300	1.03586	0.03586	0.00981
	700	1.02649	0.02649	0.00474
	900	1.02173	0.02173	0.00026

Table 7.3: Estimate (based on the variogram) of the error variance: mean, bias and variance for 100 replications

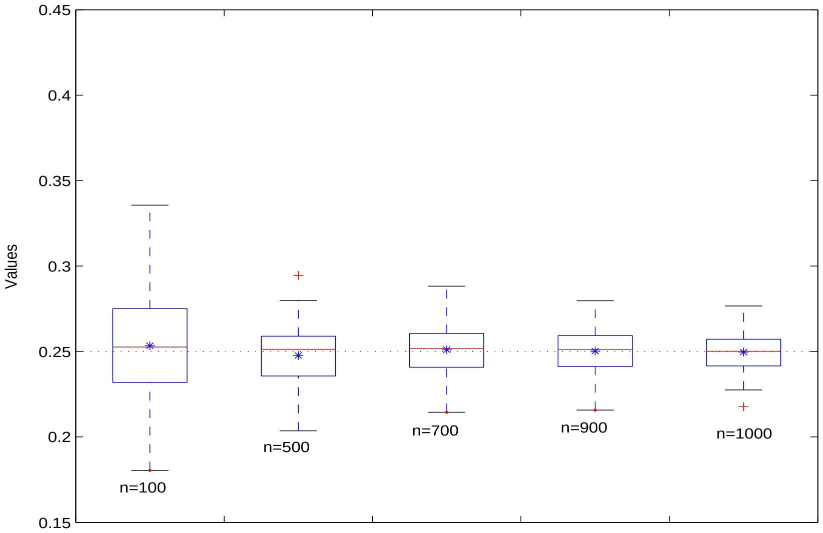

Figure 7.3: Results (model based estimator) of the experiment based on 50 realizations for different numbers of samples.

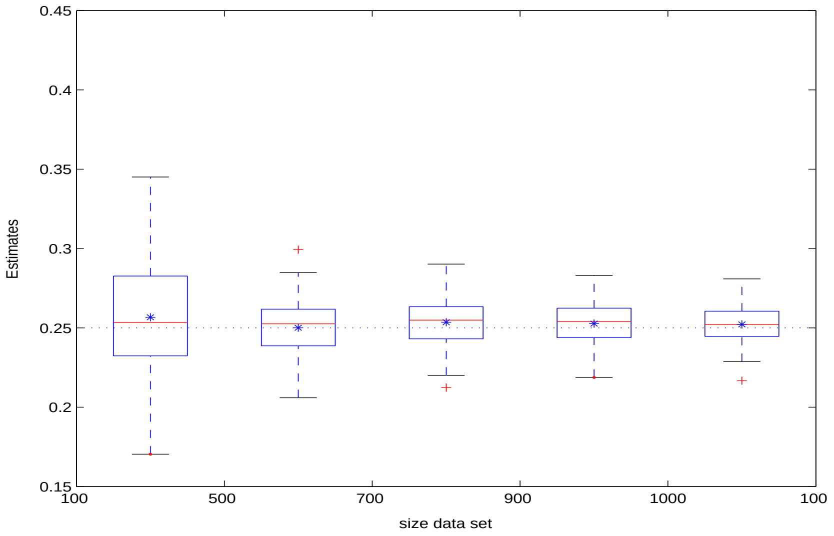

Figure 7.4: Results (model free estimator) of the experiment based on 50 realizations for different numbers of samples.

which makes it perfectly suited as a method for picking a good starting-value for a local search based on a more powerful and computationally intensive way to achieve good generalization. Experiments on the higher dimensional Boston housing data (with standardized inputs and outputs) even suggest that the proposed measure can be sufficiently good as a model selection criterion on its own. For this experiment, one third of the data was reserved for test purposes, while the remaining data were used for the training and selection of the regularization parameter. This procedure was repeated 500 times in a Monte-Carlo experiment. The kernel parameter was kept fixed in the experiments.

7.2 Heteroscedastic error variance

An excellent survey and discussion of the problems of heteroscedasticity (linear regression) is given by Judge (Judge *et al.*, 1980; Carroll and Ruppert, 1981; Horn, 1981; Cook and Weisberg, 1983). One of the important assumptions of the classical regression model is that the variance of each disturbance term e_k, $k = 1, ..., n$ conditional on the independent variables, is some constant $E\left[e_k^2 \,|x_k\right] = \sigma^2 < \infty, k = 1, ..., n$.

	Differogram	10-fold CV	leave-one-out	C_p	"true"
	Toy example: 25 data points				

mean(MSE)
s.e(MSE)
	0.4385
	1.9234
		0.3111
		0.3646
			0.3173
			1.5926
				0.3404
				0.3614
					0.2468
					0.1413

	Toy example: 200 data points				

mean(MSE)
s.e(MSE)
	0.2600
	0.5240
		0.0789
		0.0355
			0.0785
			0.0431
				0.0817
				0.0289
					0.0759
					0.0289

	Boston Housing dataset				

mean(MSE)
s.e(MSE)
	0.1503
	0.0199
		0.1538
		0.0166
			0.1518
			0.0217
				0.1522
				0.0152
					0.1491
					0.0184

Table 7.4: Results from experiments on regularization constant tuning. The experiment compares results when using the estimate based on the differogran and from classical datadriven techniques as 10-fold cross-validation, leave-one-out, Mallows Cp statistic of the hyper-parameters.

7.2.1 Error variance estimation

Detection of heteroscedasticity

Given a training set defined as $\mathcal{D}_n = \left\{(x_k, y_k) : \ x_k \in \mathbb{R}^d, y_k \in \mathbb{R}; \ k = 1, ..., n\right\}$, the regression model can be represented as

$$y_k = m(x_k) + e_k, \tag{7.34}$$

where the $e_k \in \mathbb{R}$ are assumed to be i.i.d. random errors $e_k \sim \mathcal{N}\left(0, \sigma^2\right)$ with $E[e_k\,|X = x_k] = 0$, $E[e_k^2\,|X = x_k] = Var\,[e_k] = \sigma^2 < \infty$ and $E\,[e_k e_l\,|x_k, x_l] = 0$ for all $k \neq l$. Based on model (7.34), the LS-SVM regression estimator of $m(x_k)$ is

$$\hat{m}_n(x_k) = \sum_{l=1}^{n} s_{kl} y_l, \tag{7.35}$$

where s_{kl} is the klth element of the smoother matrix S. Violation of $Var\,[e_k\,|X = x_k] = E[e_k^2\,|X = x_k] = \sigma^2$ is called heteroscedasticity and the true error covariance matrix can be defined as

$$E\left[ee^T\,|x\right] = \tau^2 V, \tag{7.36}$$

where $e = (e_1, ..., e_n)^T$, τ^2 is the true scale parameter and $V = diag\,(v_1, ..., v_n)$.

The variance is a function of the independent variables The type of the heteroscedastic regression model is given by

$$y_k = m(x_k) + e_k, \tag{7.37}$$

where $v_k = h(x_k; \beta)$ and the errors are assumed to be independent with $E[e_k\,|X = x_k] = 0$. The variance function $h\,(\cdot)$ express the heteroscedasticity,

τ^2 is the unknown scale parameter and β is an unknown parameter vector. For example, the variance may be modeled as $\tau^2 h(x_k;\beta) = \left(1+\beta_1 x_k + \beta_2 x_k^2\right)\tau^2$.

Based on $E[\hat{e}_k\,|X=x_k]=0$ where $\hat{e}_k = y_k - \hat{m}_n(x_k)$ and (7.36), the conditional variance of the residuals given x is defined as

$$\begin{aligned}
E[\hat{e}_k^2\,|X=x_k] &= Var\,[\hat{e}_k\,|X=x_k] \\
&= Var\,[(y_k-\hat{m}_n(x_k))\,|X=x_k] \\
&= Var\left[\left(y_k - \sum_{l=1}^{n} s_{kl}y_l\right)|X=x_k\right] \\
&= Var\left[\left(y_k - s_{kk}y_k - \sum_{l\neq k} s_{kl}y_l\right)|X=x_k\right] \\
&= (1-s_{kk})^2\,Var\,[y_k\,|X=x_k] - \sum_{l\neq k} s_{lk}^2 Var\,[y_l\,|X=x_l] \\
&= \left((1-s_{kk})^2\,v_k - \sum_{l\neq k} s_{kl}^2 v_l\right)\tau^2 \qquad (7.38)
\end{aligned}$$

and the conditional variance of the squared residuals given x is given by

$$Var\left[\hat{e}_k^2\,|x_k\right] = 2\left(Var\,[\hat{e}_k\,|X=x_k]\right)^2,\ k=1,...,n. \qquad (7.39)$$

¿From $E[\hat{e}_k\,|X=x_k]=0$ and (7.38), plotting $\hat{e}_k^{*2} = \frac{\hat{e}_k^2}{1-s_{kk}}$ versus x_k is determined by

$$Var\,[\hat{e}_k^*\,|X=x_k]\,E[\hat{e}_k^{*2}\,|X=x_k] = \left(v_k - \frac{\sum_{l\neq k} s_{kl}^2 v_l}{(1-s_{kk})^2}\right)\tau^2. \qquad (7.40)$$

The pattern of this plot is determined by $E[\hat{e}_k^{*2}\,|X=x_k]+\sqrt{2}Var\,[\hat{e}_k\,|X=x_k]$ and if the second term of (7.40) is ignored, then $Var\,[\hat{e}_k\,|X=x_k]=v_k$. The pattern of the squared residual plot can identify the weighting function v_k.

We present an example of a squared residual plot in assessing nonconstant variance. Consider the following model

$$y_k = \frac{\sin(x_k)}{x_k} + e_k, \quad k=1,...,n, \qquad (7.41)$$

a one-dimensional sinc function. Errors were simulated from the normal distribution, with mean zero, variance $v(x_k)\,\tau^2$ and $\tau^2=1$. The true weighting function $v(x_k)$ is $|x_k|$. The squared residual plot (Figure 7.5) shows a nonconstant variance pattern.

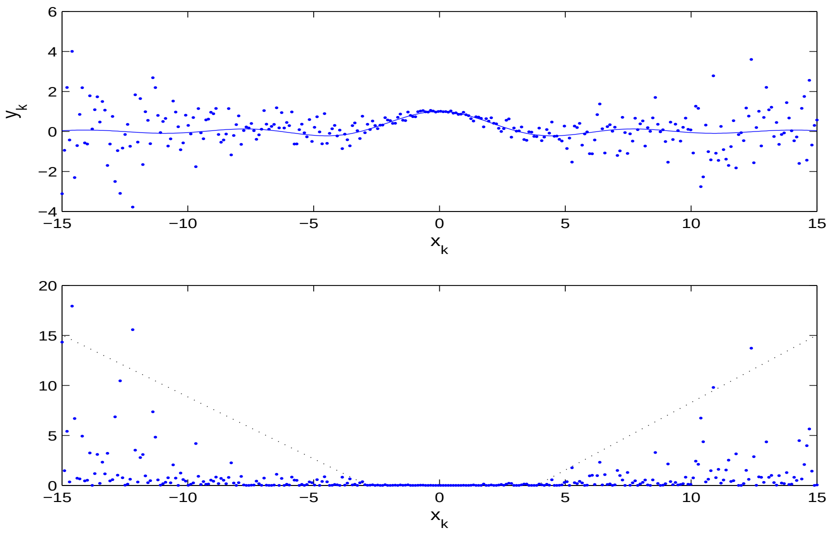

Figure 7.5: The data and the function (row 1). The squared (row 2) residual plot for assessing the heteroscedasticity as $v_k = \exp(x_k)$.

The variance is a function of the expected dependent variable The type of the heteroscedastic regression model is given by

$$y_k = m(x_k) + e_k, \tag{7.42}$$

where $v_k = h(E\left[y_k \left| X = x_k \right.\right]; \beta)$ and the errors are assumed to be independent with $E[e_k \left| X = x_k \right.] = 0$. The variance function $h(\cdot)$ express as the heteroscedasticity, τ^2 is the unknown scale parameter and β is an unknown parameter vector. In this case, $\hat{e}_k$ and $\hat{m}_n(x_k)$ are both random variables and the joint density function must be calculated. Based on the central limit theorem $\hat{m}_n(x_k) = \sum_{l=1}^{n} s_{kl} y_l \sim \mathcal{N}\left(m(x_k), \sigma^2_{m(x_k)}\right)$, and based on the assumption that $e_k \sim \mathcal{N}\left(0, \sigma^2_{e_k}\right)$, the joint density function of $(\hat{e}_k, \hat{m}_n(x_k))$ is given by

$$(\hat{e}_k, \hat{m}_n(x_k)) \sim N\left(\left(\begin{array}{c} 0 \\ m(x_k) \end{array}\right), \left(\begin{array}{cc} Var\left[\hat{e}_k\right] & cov\left[\hat{e}_k, \hat{m}_n(x_k)\right] \\ cov\left[\hat{m}_n(x_k), \hat{e}_k\right] & Var\left[\hat{m}_n(x_k)\right] \end{array}\right) \tau^2\right), \tag{7.43}$$

where

$$\begin{aligned} Var\left[\hat{e}_k\right] &= Var\left[(y_k - \hat{m}_n(x_k))\right] \\ &= \left((1 - s_{kk})^2 v_k - \sum_{l \neq k} s_{kl}^2 v_l\right) \tau^2, \end{aligned} \tag{7.44}$$

$$\begin{aligned} Var\left[\hat{m}_n(x_k)\right] &= Var\left[\sum_{l=1}^{n} s_{kl} y_l\right] \\ &= \sum_{l=1}^{n} s_{kl}^2 Var\left[y_l\right] \\ &= \sum_{l=1}^{n} s_{kl}^2 v_l \tau^2, \end{aligned} \tag{7.45}$$

and

$$\begin{aligned} cov\left[\hat{m}_n(x_k), \hat{e}_k\right] &= E\left[\hat{m}_n(x_k)\hat{e}_k\right] - E\left[\hat{m}_n(x_k)\right] E\left[\hat{e}_k\right] \\ &= E\left[\sum_{l=1}^{n} s_{kl} y_l \left(y_k - s_{kk} y_k - \sum_{l \neq k} s_{kl} y_l\right)\right] \\ &\quad - E\left[\sum_{l=1}^{n} s_{kl} y_l\right] E\left[y_k - s_{kk} y_k - \sum_{l \neq k} s_{kl} y_l\right] \\ &= \left(s_{kk} v_k - \sum_{l=1}^{n} s_{kl}^2 v_l\right) \tau^2. \end{aligned} \tag{7.46}$$

and

$$cov\left[\hat{e}_k, \hat{m}_n(x_k)\right] = cov\left[\hat{m}_n(x_k), \hat{e}_k\right]. \tag{7.47}$$

For calculating $E\left[\hat{e}_k \left|\hat{m}_n(x_k)\right.\right]$ and $Var\left[\hat{e}_k \left|\hat{m}_n(x_k)\right.\right]$ we need the conditional density of $\hat{e}$ given $\hat{y}$. This is a normal density with mean $E\left[\hat{e}\right] + \rho \frac{\tau_{\hat{e}}^2}{\tau_{\hat{m}_n(x)}^2}$ $\left(\hat{m}_n(x) - E\left[\hat{m}_n(x)\right]\right)$ and the variance $\tau_{\hat{e}}^2\left(1-\rho^2\right)$ with $-\infty < E\left[\hat{e}\right] < \infty$, $-\infty < E\left[\hat{m}_n(x)\right] < \infty$ and $-1 < \rho < 1$ the correlation coefficient.

Hence,

$$\begin{aligned} E\left[\hat{e}_k \left|\hat{m}_n(x_k)\right.\right] &= E\left[\hat{e}_k\right] + \frac{cov\left[\hat{m}_n(x_k), \hat{e}_k\right]}{\sqrt{Var\left[\hat{e}_k\right] Var\left[\hat{m}_n(x_k)\right]}} \left(\hat{m}_n(x_k) - m(x_k)\right) \\ &\qquad \frac{\sqrt{Var\left[\hat{e}_k\right]}}{\sqrt{Var\left[\hat{m}_n(x_k)\right]}} \\ &= \frac{cov\left[\hat{m}_n(x_k), \hat{e}_k\right]}{Var\left[\hat{m}_n(x_k)\right]} \left(\hat{m}_n(x_k) - m(x_k)\right) \\ &= \frac{\left(s_{kk} v_k - \sum_{l=1}^{n} s_{kl}^2 v_l\right) \tau^2}{\sum_{l=1}^{n} s_{kl}^2 v_l \tau^2} \left(\hat{m}_n(x_k) - m(x_k)\right) \\ &= \left(\frac{s_{kk} v_k}{\sum_{l=1}^{n} s_{kl}^2 v_l} - 1\right) \left(\hat{m}_n(x_k) - m(x_k)\right), \end{aligned} \tag{7.48}$$

and

$$Var\left[\hat{e}_k\left|\hat{m}_n(x_k)\right.\right]=Var\left[\hat{e}_k\right]\left(1-\frac{\left(cov\left[\hat{m}_n(x_k),\hat{e}_k\right]\right)^2}{Var\left[\hat{e}_k\right]Var\left[\hat{m}_n(x_k)\right]}\right)$$
$$=\left(1-\frac{s_{kk}^2v_k}{\sum_{l=1}^n s_{kl}^2v_l}\right)v_k\tau^2.$$

¿From $Var\left[u\right]=E\left[u\right]-\left(E\left[u\right]\right)^2$, the conditional expectation of the squared residuals given $\hat{m}_n(x_k)$ equals

$$E\left[\hat{e}_k^2\left|\hat{m}_n(x_k)\right.\right]=Var\left[\hat{e}_k\left|\hat{m}_n(x_k)\right.\right]+\left(E\left[\hat{e}_k\left|\hat{m}_n(x_k)\right.\right]\right)^2 \qquad (7.49)$$

and the conditional variance of the squared residuals given $\hat{m}_n(x_k)$ is defined as

$$Var\left[\hat{e}_k^2\left|\hat{m}_n(x_k)\right.\right]=2\left(Var\left[\hat{e}_k\left|X=x_k\right.\right]\right)^2,\ k=1,...,n. \qquad (7.50)$$

In this case we plot $\hat{e}_k^2$ versus $\hat{m}_n(x_k)$. The pattern of this plot is determined by $E[\hat{e}_k^{*2}\left|X=x_k\right.]+\sqrt{2}Var\left[\hat{e}_k\left|X=x_k\right.\right]$. The pattern of the squared residual plot can identify the weighting function v_k. As a second example, consider the following model

$$y_k=\frac{\cos(15(x_k+0.5))}{x_k+0.5}+e_k,\quad k=1,...,n, \qquad (7.51)$$

with $x\sim U\left[0,1\right]$. Errors were simulated from the normal distribution, with zero mean, variance $v\left(E\left[y_k\left|X=x_k\right.\right]\right)\tau^2$ and $\tau^2=1$. The weighting function $v\left(E\left[y_k\left|X=x_k\right.\right]\right)$ is $\exp\left(m\left(x_k\right)\right)$. The squared residual plot (Figure 7.6) shows a nonconstant variance pattern.

Kernel smoothing of local variance estimates

Estimation of the local variance has been considered in the context of linear regression with the aim of estimating optimal weights for weighted least squares by (Carroll and Ruppert, 1982; Davidian and Carrol, 1987; Carrol *et al.*, 1988; Hooper, 1993; Welsh *et al.*, 1994). Several parameter estimation methods for dealing with heteroscedasticity in nonlinear regression are described by (Beal and Sheiner, 1988). These include variations on ordinary, weighted, iteratively reweighted, extended, and generalized least squares. See, for example, (Rice, 1984; Gasser *et al.*, 1986; Pelckmans *et al.*, 2003) for a constant variance estimate in a nonparametric regression model.

For estimation heteroscedasticity in regression, we will use a kernel smoothed local variance estimator. We assume that: (i) The error variables e_k, $k=1,...,n$ are independent, $E\left[e_k\right]=0$, $E\left[e_k^2\right]=\sigma^2\left(z\right)$ where $z=(x \text{ or } y)$ and in addition $E\left[\left|e_k\right|^{2r}\right]\leq M<\infty$, $r>1$. (ii) $m\in C^\infty\left(\mathbb{R}\right)$, and (iii) $\sigma^2\left(z\right)\in C^\infty\left(\mathbb{R}\right)$. Consider the regression model

$$v_k=\sigma^2\left(z_k\right)+\varepsilon_k,\ k=1,...,n \qquad (7.52)$$

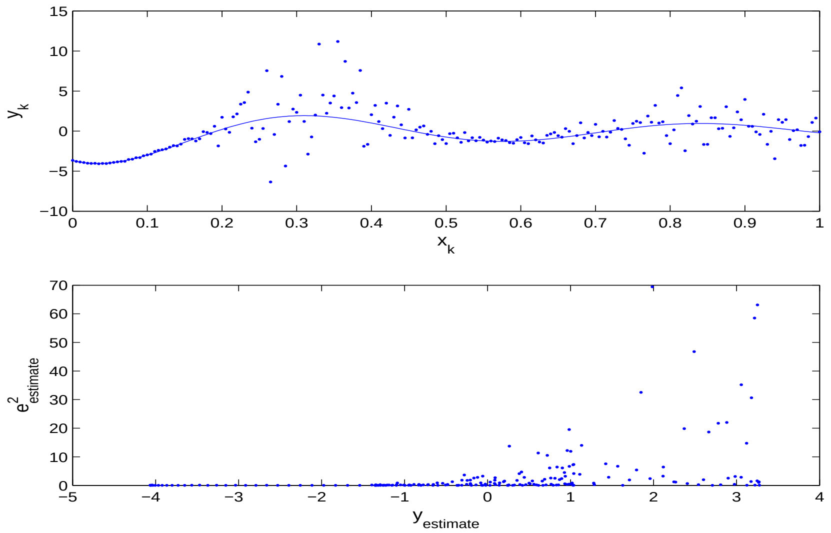

Figure 7.6: The data and the function (row 1). The squared (row 2) residual plot for assessing the heteroscedasticity as $v_k = \exp(m(x_k))$.

where v_k are the initial variance estimates. To obtain consistent estimators (Müller and Stadtmüller, 1987), we apply the Nadaraya-Watson estimator

$$\hat{\sigma}^2(z) = \sum_{k=1}^{n} \frac{K\left(\frac{z-z_k}{h}\right) v_k}{\sum_{l=1}^{n} K\left(\frac{z-z_l}{h}\right)}, \tag{7.53}$$

where K denotes the kernel function and the sequence of bandwidths h has to satisfy $h \to 0$, $nh \to \infty$ as $n \to \infty$.

7.2.2 Estimation of the regression function

To keep things simple, consider univariate regression. The regression model is defined as

$$y_k = \beta_0 + \beta_1 x_k + e_k, \tag{7.54}$$

where $E[e_k] = 0$, $E\left[(e_k)^2\right] = \sigma_k^2 < \infty$ and $E[e_k e_l] = 0$, $\forall k \neq l$. For example, let $\hat{\beta}_1$ be the ordinary least squares estimator of β_1. Heteroscedasticity does not destroy the unbiasedness and consistency properties of $\hat{\beta}_1$. But $\hat{\beta}_1$ is no longer minimum variance or efficient. A general approach to deriving accurate estimates of β_1 is weighted least squares. The idea behind weighted least squares is that least squares is still a good thing to do if the target and the independent variables are transformed to give a model with errors with constant variance. Let $\hat{\beta}_1^*$ be the weighted least squares estimator of β_1. On average, $\hat{\beta}_1^*$ will be closer to the true regression coefficient than are the ordinary least squares estimates. Hypothesis tests, such as t-test and F-tests, follow the assumed distributions and are thus proper tools for inference. Examples can be found in (Sen and Srivastava, 1997; Neter *et al.*, 1990).

LS-SVM regression estimate

In order to obtain an estimate (heteroscedastic case) based upon the previous LS-SVM solution, in a subsequent step, one can weight the error variables $e_k = \alpha_k/\gamma$ by weighting factors ϑ_k . This leads to the optimization problem:

$$\min_{w^*,b^*,e^*} \mathcal{J}\left(w^*, e^*\right) = \frac{1}{2} w^{*T} w^* + \frac{1}{2}\gamma \sum_{k=1}^{n} \vartheta_k e_k^{*2} \tag{7.55}$$

such that $y_k = w^{*T}\varphi\left(x_k\right) + b^* + e_k^*, \quad k = 1, ..., n.$ The Lagrangian is constructed in a similar way as before. The unknown variables for this weighted LS-SVM problem are denoted by the $*$ symbol. From the conditions for optimality and elimination of w^*, e^* one obtains the Karush-Kuhn-Tucker system:

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \mathcal{V}_\gamma \end{array}\right] \left[\begin{array}{c} b^* \\ \hline \alpha^* \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \tag{7.56}$$

where the diagonal matrix $\mathcal{V}_\gamma$ is given by $\mathcal{V}_\gamma = \operatorname{diag}\left\{\frac{1}{\gamma\vartheta_1}, ..., \frac{1}{\gamma\vartheta_n}\right\}$. The weights

$$\vartheta_k = \frac{1}{\hat{\sigma}^2\left(z_k\right)}, \; k = 1, ..., n, \tag{7.57}$$

are determined based upon the smoothing error variance estimator (7.53). Using these weightings one can correct for heteroscedasticity. This leads us to the following algorithm:

Algorithm 22 *(heteroscedastic LS-SVM).*

1. Given training data $\mathcal{D} = \left\{\left(x_1, y_1\right), ..., \left(x_n, y_n\right)\right\}$, find an optimal (h, γ) combination (e.g., by cross-validation, FPE criterion) by solving linear systems (Chapter 3, (3.8)).

2. Estimate the variance and determine the weights $\vartheta_k = \frac{1}{\hat{\sigma}^2(z_k)}$

3. Solve the weighted LS-SVM (7.56).

7.2.3 Simulations

In these examples we illustrate the method of weighted LS-SVM (heteroscedastic case).

Simulation 1

Consider the following model

$$y_k = \frac{\sin(x_k)}{x_k} + e_k, \quad k = 1, ..., n, \tag{7.58}$$

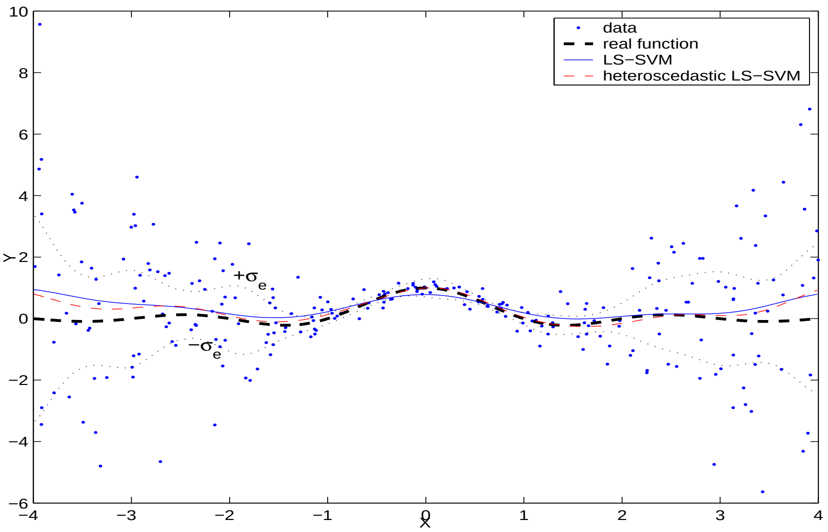

Figure 7.7: Both estimated regression functions (dashed line) and (doted line) were obtained from the heteroscedastic LS-SVM regression fit and from the LS-SVM regression fit respectively. Errors were simulated from the normal distribution, with mean zero, variance $w(x_k)\ \sigma_0^2$ and $\sigma_0^2 = 1$. The weighting function $v(x_k)$ is $0.5x_k^2 + 0.1$.

a one-dimensional sinc function. Errors were simulated from the normal distribution, with mean zero, variance $w(x_k)\ \sigma_0^2$ and $\sigma_0^2 = 1$. The weighting function $v(x_k)$ is $0.5x_k^2 + 0.1$. Figure 7.7 shows the results from unweighted and weighted LS-SVM.

The weighted LS-SVM resulted in a test set MSE of 0.0739, which was an improvement over the unweighted LS-SVM test set MSE of 0.1273.

Simulation 2

As a second example, consider the following model

$$y_k = \frac{\cos(15(x_k + 0.5))}{x_k + 0.5} + e_k, \quad k = 1, ..., n, \tag{7.59}$$

with $x \sim U[0,1]$. Errors were simulated from the normal distribution, with mean zero, variance $v(x_k)\ \sigma_0^2$ and $\sigma_0^2 = 1$. The weighting function $v(x_k)$ is $2x_k^3 + 0.4$. Figure 7.8 shows the results from unweighted and weighted LS-SVM.

The weighted LS-SVM resulted in a test set MSE of 0.0932, which was an improvement over the unweighted LS-SVM test set MSE of 0.3617.

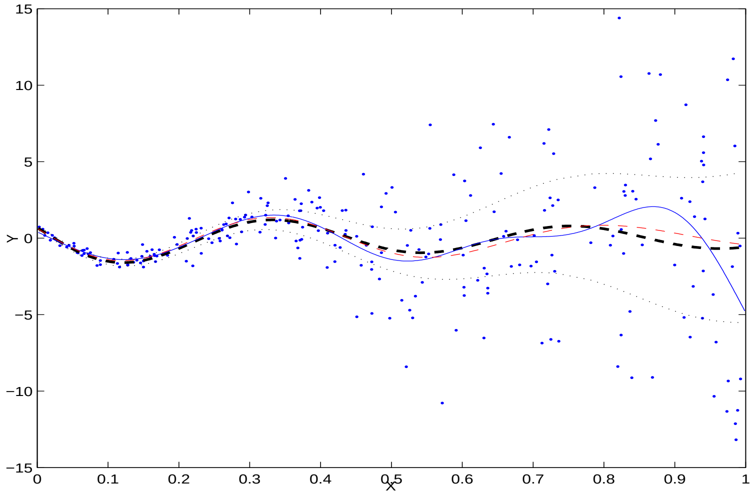

Figure 7.8: Both estimated regression functions (dashed line) and (doted line) were obtained from the heteroscedastic LS-SVM regression fit and from the LS-SVM regression fit respectively. Errors were simulated from the normal distribution, with mean zero, variance $v\left(x_k\right)\ \sigma_0^2$ and $\sigma_0^2 = 1$. The weighting function $v\left(x_k\right)$ is $2x_k^3 + 0.4$.

7.3 Conclusions

We proposed a non-parametric data analysis tool for noise variance estimation towards a machine learning context. By modelling the variation in the data for observations that are located close to each other, properties of the data can be extracted without relying on an explicit model of the data. These ideas are translated by considering the differences of the data instead of the data itself in the so-called differogram cloud. A model for the differogram can be inferred for sufficiently small differences. By deriving an upper bound on the variance of the differogram model, this locality can be formulated without having to rely explicitly on a hyper-parameter as the bandwidth. Furthermore, a number of applications of modelfree noise variance estimators for model selection and hyper-parameter tuning have been given.

While the method of least squares (under the Gauss-Markov conditions) enjoys well known properties, we have studied the properties of the LS-SVM regression when relaxing these conditions. It was recognized that outliers may have an unusually large influence on the resulting estimate. However, asymptotically the heteroscedasticity does not play any important role. Squared residual plots are proposed to assess heteroscedasticity in regression diagnostics.

Chapter 8

Density estimation

In this chapter we give a survey of parametric and nonparametric density estimation. We briefly discuss some methods for choosing the smoothing parameter in the kernel density estimator. Next, we discuss the regression view of density estimation. Finally, we apply LS-SVM regression modelling in the case of density estimation. Contributions are made in Section 8.3 and Section 8.5.

8.1 Introduction

In many cases, one wishes to make inferences only about some finite set of parameters, such as the mean and variance, that describe the population. In other cases, one wants to predict a future value of an observation. Sometimes, the objective is more difficult: one wants to estimate a function that characterizes the distribution of the population. The cumulative distribution function (cdf) or the probability density function (pdf) provides a complete description of the population, so one may wish to estimate these functions. Although the cdf in some ways is more fundamental in characterizing a probability distribution (it always exists and is defined the same for both continuous and discrete distributions), the probability density function is more familiar and important properties can be seen more readily from a plot of the pdf than from the plot of the cdf.

In the simpler cases of statistical inference, one assumes that the form of the pdf is known and that there is a finite dimensional parameter vector that characterizes the distribution within the assumed family. The normal distribution is a good model for symmetric, continuous data. For skewed data, the lognormal and gamma distributions often work very well. Discrete data are often modelled by the Poisson or binomial distributions. Distributions such as these are families of distributions that have various numbers of parameters to specify the distribution completely. A standard way of estimating a probability density function is to identify appropriate characteristics, such as symmetry, modes, range, etc., choose some well-known parametric distribution that has those characteristics, and then estimate the parameters of that distribution. For example, if the pdf of interest has a finite range, a beta distribution may be used to model it, and if it has an infinite range at both ends, a normal distribution, a Student´s t distribution, or a stable distribution may be a useful approximation. Rather than trying to fit an unknown pdf to a single parametric family of distributions, it may be better to fit it to a finite mixture of densities. (Priebe, 1994) describes an adaptive method of using mixtures to estimate a pdf. (Solka *et al.*, 1995) describe visualization methods for selection of the mixture estimator. (Everitt and Hand, 1981) provide a general discussion of the use of finite mixtures for representing distributions. (Roeder and Wasserman, 1997) describe the use of mixtures in Bayesian density estimation.

In a non-parametric approach, no assumptions or only weak assumptions, are made about the form of the distribution or density function. These assumptions may only address the shape of the distribution, such as an assumption of unimodality or an assumption of continuity or other degrees of smoothness of the density function. The estimation problem is much more difficult and the estimation problem may be computationally intensive. A very large sample is usually required in order to get a reliable estimate of the pdf. Starting from the definition of a *pdf*, a class of density estimators (e.g., histograms, Nadaraya-Watson kernel estimators, nearest neigbour methods, variable Nadaraya-Watson kernel estimators and orthogonal series estimator) are defined. Apart from the histogram, the Nadaraya-Watson kernel estimator is the most commonly used estimator and is the most studied mathematically (Rosenblatt, 1956) and (Parzen, 1962). The variable kernel method is related to the nearest neigbour class of estimators and is a method which adapts the amount of smoothing to the local density of the data. Orthogonal series density estimators, or projection estimators, introduced by (Čencov, 1962), are smoothers of the empirical density function. The local nature of wavelet functions promises superiority over projection estimators that use classical orthonormal bases (Fourier, Hermite, etc.). The wavelet estimators are simple and share a variety of optimality properties. The estimation procedures fall into the class of so-called projection estimators. For a critical discussion of the advantage and disadvantages of wavelets in density estimation see (Walter and Ghorai, 1992). A theoretical overview of wavelet density estimation can be found in (Hädle *et al.*, 1998). The methods discussed so far are all derived in an ad hoc way from the definition of a density. A survey of density estimation based on a standard statistical technique, the maximum likelihood, is given next.

In their pioneering article, (Good and Gaskin, 1971) introduced the idea of *maximum penalized likelihood* density estimation. The idea is to minimize a penalized minus log likelihood functional. The log likelihood dictates the estimate to adapt to the data, the roughness penalty counteracts by demanding less variation in the density function, and the smoothing parameter controls the tradeoff between the two conflicting goals. The use of penalty has a long history, which may trace back to (Whittaker, 1923) and (Tikhonov, 1963); see (Wahba, 1990) for a review.

The method of sieves was introduced by (Grenander, 1981) and studied by many authors including (Geman and Hwang, 1982; van de Geer, 1993, 1996; Shen and Wong, 1994; Wong and Shen, 1995; Birgé and Massart, 1998). The idea of sieves is to restrict the minimization in the maximum likelihood problem to classes of smooth densities. There are two kind of sieves. One kind is obtained by considering finite-dimentional subspaces of $L^1(\mathbb{R})$, and then the sieve consists of all pdfs in a particular subspace. The other kind of sieve is obtained by considering compact subsets of $L^1(\mathbb{R})$. Examples of sieves are: Gaussian mixture sieve (Genovese and Wasserman, 2000) and the class of beta mixtures defined by Bernstein polynomials (Ghosal, 2001).

There are various semi-parametric approaches in which, for example, parametric assumptions may be made only over a subset of the range of the distribution and non-parametric assumptions for others. Local likelihood was introduced by (Tibshirani and Hasti,) as a method of smoothing by local polynomials in non-Gaussian regression models. An extension of these *local likelihood* methods to density estimation can be found in (Loader, 1996) and (Hjort and Jones, 1996). The idea is, that around each given data point one define the local log likelihood. The local polynomial approximation assumes that the logarithm of the density function can be well approximated by a low-degree polynomial in a neighborhood of the fitting point.

A number of parametric methods, using neural networks, have been proposed in the literature. For example, (Traven, 1991; Bishop and Legleye, 1995; Miller and Horn, 1998). Neural network methods for density estimation based on approximating the distribution function can be found in (Magdon-Ismail and Atiya, 2002).

Another class of probability density function estimators based on Support vector methods (SVM) are proposed by (Mukherjee and Vapnik, 1999). While the structure of the Nadaraya-Watson kernel density estimator can be too complex, the support vector approach provides a sparse estimate of a density and therefore a reduced computational cost.

8.2 Survey of density estimation

Suppose that $X_1, ..., X_n$ are random variables that are independent and identically distributed according to some probability distribution function F, where $F \in \mathcal{F}$, a family of probability distribution functions and probability density function f. The probability density function (pdf), which has the properties that $f(x) \geq 0$, f is piecewise continuous and $\int_{-\infty}^{\infty} f(x)dx = 1$, is defined as

$$F(x) = \int_{-\infty}^{x} f(u)du, \qquad (8.1)$$

The problem is to construct a sequence of estimators $\hat{f}_n(x)$ of $f(x)$ based on the sample $x_1, ..., x_n$. Because unbiased estimators do not exist for f (Rao, 1983), one is interested in asymptotically unbiased estimators $\hat{f}_n(x)$ such that

$$\lim_{n\to\infty} E_{f\in\mathcal{F}_n}\left[\hat{f}_n(x)\right] = f(x), \qquad \forall x.$$

Starting from the definition of a *pdf*, a class of density estimators (e.g., histogram estimator, kernel estimator, orthogonal series estimator, wavelets) are defined. Let $W(z,x)$ be a function of two arguments, which will satisfy the conditions $\int_{-\infty}^{\infty} W(z,x)\,dx = 1$ and $W(z,x) \geq 0,\ \forall\ z,x$. An estimate of the density underlying the data can be obtained by putting

$$\hat{f}_n(x) = \frac{1}{n}\sum_{k=1}^{n} W(x_k,x)\,. \tag{8.2}$$

For example, the kernel estimate can be obtained by putting $W(z,x) = \frac{1}{h}K\left(\frac{x-z}{h}\right)$ and the Parzen kernel density estimator is defined as

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x-x_k}{h}\right), \tag{8.3}$$

where K is some chosen unimodal density, symmetric about zero.

The variable kernel method is related to the nearest neigbour class of estimators and is a method which adapts the amount of smoothing to the local density of the data. Define $\Delta_{k,l}$ to be the distance from x_k to the lth nearest point in the set $\mathcal{D}^{(-k)} = \{x_1,...,x_{k-1},x_{k+1},...,x_n\}$ and let $l \in \mathbb{N}_0$. The variable kernel estimate with smoothing parameter h is defined by

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{k=1}^{n} \frac{1}{\Delta_{k,l}} K\left(\frac{x-x_k}{h\Delta_{k,l}}\right). \tag{8.4}$$

The window width of the kernel placed on the point Z_k is proportional to $\Delta_{k,l}$ so that data points in regions where the data are sparse well have flatter kernels associated with them. For any fixed l, the overall degree of smoothing will depend on the parameter h.

The methods discussed so far are all derived in an ad hoc way from the definition of a density. A survey of density estimation based on a standard statistical technique, the maximum likelihood, is given in the next Subsections.

8.2.1 Maximum likelihood estimation

Parametric maximum likelihood estimation

When the density f on a domain $\mathcal{H}$ is known to belong to a finite dimensional parametric family $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ described by a (low dimensional) parameter belonging to the set of all possible parameters Θ. Then, there exists a $\theta_0 \in \Theta$ such that $f(x) = f(x;\theta_0)$, $-\infty < x < \infty$. The standard method for estimating θ_0 is by maximum likelihood estimation (Fisher, 1922). Note that a parametric approach puts rigid constraints on the estimator. Rather than maximizing the likelihood itself, the estimation problem (under reasonable conditions) is equivalent to

$$\begin{array}{l} \min -\int_{\mathbb{R}} \log f(x;\theta) d\hat{F}_n(x) \\ \text{s.t. } \theta \in \Theta \end{array} \tag{8.5}$$

where $\hat{F}_n(x)$ is the empirical distribution function. There are other methods for estimating θ_0, such as the method of moments, the method of maximal spacing and quantile regression. A choice of θ_0, based on least squares estimation, would minimize $\int_{\mathbb{R}} (f(x;\theta) - f(x))^2 \, dx$. Because $f(x)$ is unknown, the least squares estimator of the parameter θ is the solution to

$$\begin{array}{l} \min -2 \int_{\mathbb{R}} f(x;\theta) d\hat{F}_n(x) + \int_{\mathbb{R}} (f(x;\theta))^2 \, dx \\ \text{s.t. } \theta \in \Theta \end{array} \tag{8.6}$$

Non-parametric maximum likelihood estimation

When a parametric form is not available, a naive maximum likelihood estimator without any nonintrinsic constraint is a sum of delta function spikes at the sample points, which apparently is not an appealing estimator when the domain $\mathcal{H}$ is continuous. The maximum likelihood solution or least squares solution for f is then given by a histogram (Thompson and Tapia, 1990). An alternative is to spread out the point masses to obtain the Nadaraya-Watson kernel density estimator. To incorporate the information that the solution of

$$\begin{array}{l} \min - \int_{\mathbb{R}} \log f(x) d\hat{F}_n(x) \\ \text{s.t. } f \text{ is a continuous } \textit{pdf} \end{array}$$

should be a smooth *pdf* is done by the method of *maximum penalized likelihood* estimation, the method of *local parametric non-parametric maximum likelihood* estimation and the method of *sieves* (Grenander, 1981).

In their pioneering article, (Good and Gaskin, 1971) introduced the idea of *maximum penalized likelihood* density estimation. The idea is to minimize a penalized minus log likelihood functional

$$\begin{array}{l} \min - \int_{\mathbb{R}} \log f(x) d\hat{F}_n(x) + h\mathcal{J}(f) \\ \text{s.t. } f \text{ is a continuous } \textit{pdf}, \end{array} \tag{8.7}$$

where the $\mathcal{J}(f) = \int_{-\infty}^{\infty} \left(\frac{d}{dx} \sqrt{f(x)} \right)^2 dx$ is a roughness penalty and h is called a smoothing parameter. The log likelihood dictates the estimate to adapt to the data, the roughness penalty counteracts by demanding less variation in f, and the smoothing parameter controls the tradeoff between the two conflicting goals. The penalized version of least squares is defined by

$$\begin{array}{l} \min -2 \int_{\mathbb{R}} f(x) d\hat{F}_n(x) + \int_{\mathbb{R}} (f(x))^2 \, dx + h\mathcal{J}(f) \\ \text{s.t. } f \text{ is a continuous } \textit{pdf}, \end{array} \tag{8.8}$$

with the choice of $\mathcal{J}(f) = \int_{-\infty}^{\infty} \left(\frac{d}{dx} f(x) \right)^2 dx$, the solution of (8.8) satisfies the boundary value problem

$$\begin{array}{c} -h^2 \frac{d^2}{dx^2} (f(x)) + f(x) = d\hat{F}_n(x), \quad -\infty < x < \infty \\ f(x) \to 0 \quad \text{for } |x| \to \infty, \end{array}$$

and is given by the density function $f(x) = \int_{\mathbb{R}} K\left(\frac{x-z}{h}\right) d\hat{F}_n(z)$, where $K(x,z) = \frac{1}{2h}\exp\left(-\frac{|x-z|}{h}\right)$ is the scaled, two-sided exponential kernel (see Courant and Hilbert, 1953). The use of penalty has a long history, which may trace back to (Whittaker, 1923) and (Tikhonov, 1963); see (Wahba, 1990) for a review.

Local likelihood was introduced by (Tibshirani and Hasti,) as a method of smoothing by local polynomials in non-Gaussian regression models. This procedure is designed for non-parametric regression modelling such as logistic regression and proportional hazards models. An extension of these *local likelihood* methods to density estimation can be found in (Loader, 1996) and (Hjort and Jones, 1996). The idea is, that around each given x, one defines the local log likelihood to be

$$\begin{aligned}\log(L(x;\theta)) &= \int_{\mathbb{R}} K\left(\frac{x-z}{h}\right)\left[\log f(z;\theta)d\hat{F}_n(z) - f(z;\theta)dz\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} K\left(\frac{x-z_k}{h}\right)\log f(z_k;\theta) - \int_{\mathbb{R}} K\left(\frac{x-z}{h}\right) f(z;\theta)dz, \quad (8.9)\end{aligned}$$

where $\hat{F}_n(z)$ is the empirical distribution function, K is a suitable non-negative weighting function and h is the bandwidth. The local polynomial approximation assumes that $\log f(z)$ can be well approximated by a low-degree polynomial in a neighborhood of the fitting point x. With this approximation, the local log likelihood estimator of the parameter θ is the solution to

$$\begin{array}{l}\min -\int_{\mathbb{R}} K\left(\frac{x-z}{h}\right)\left[\log f(z;\theta)d\hat{F}_n(z) - f(z;\theta)dz\right] \\ \text{s.t. } f \text{ is a continuous } \textit{pdf}.\end{array} \quad (8.10)$$

For example, a local constant fitting , (8.10) gives $\hat{f}_n(x) = \frac{\sum_{k=1}^{n} K\left(\frac{x-z_k}{h}\right)}{n\int_{\mathbb{R}} K\left(\frac{x-z}{h}\right)dz}$, which is the kernel estimate introduced by (Rosenblatt, 1956) and (Parzen, 1962).

The sieves method was introduced by (Grenander, 1981) and studied by many authors including (Geman and Hwang, 1982; van de Geer, 1993, 1996; Shen and Wong, 1994; Wong and Shen, 1995; Birgé and Massart, 1998). Consider the following set of functions

$$L^p(\Omega) \overset{\text{def}}{=} \left\{\xi : \Omega \to \bar{\mathbb{R}} \text{ measurable } : \|\xi\|_p < \infty\right\},$$

for a given domain $\Omega \subseteq \mathbb{R}$ and a positive number $p \in [1,\infty]$. The idea of sieves is to restrict the minimization in the maximum likelihood problem to classes of smooth densities. There are two kind of sieves. One kind is obtained by considering finite-dimentional subspaces of $L^1(\mathbb{R})$, and then the sieve consists of all pdfs in a particular subspace. The other kind of sieve is obtained by considering compact subsets of $L^1(\mathbb{R})$. For example, let $g \in L^2(\mathbb{R})$ then $g^2 \in L^1(\mathbb{R})$, a simple example of sieves is when one have a nested sequence of finite-dimentional subspaces of subspaces of $L^2(\mathbb{R})$

$$\Upsilon_1 \subset \ldots \subset \Upsilon_m \subset L^2(\mathbb{R}),$$

which is dense in $L^2(\mathbb{R})$. An exponential family of densities is given by

$$f(x) = g_0(x)\exp(\varphi(x)), \qquad -\infty < x < \infty,$$

where g_0 is some fixed pdf and $\varphi \in \Upsilon_m$. The minimization problem may then be formulated by

$$\begin{array}{l} \min -\int_{\mathbb{R}} \varphi(x)\, dF_n(x) \\ \text{s.t. } \varphi \in \Upsilon_m, \text{ and } \int_{\mathbb{R}} g_0(x)\exp(\varphi(x))\, dx = 1. \end{array} \tag{8.11}$$

Remark that estimation by the method of sieves reveals that the dimension of the subspaces Υ_m plays the role of the smoothing parameter. Other sieves are for example, Gaussian mixture sieve (Genovese and Wasserman, 2000) and the class of beta mixtures defined by Bernstein polynomials (Ghosal, 2001).

8.2.2 Support Vector Method for density estimation

The SVM approach (Mukherjee and Vapnik, 1999) considers the problem of pdf estimation as a problem of solving (8.1) where instead of $F(x)$ one uses a plug-in estimator $\hat{F}_n(x)$, the empirical distribution function. Solving $Tf = F$ with approximation $\hat{F}_n(x)$ is an ill-posed problem. Methods for solving ill-posed problems where proposed by (Tikhonov, 1963) and (Philips, 1962). Solving (8.1) in a set of functions belonging to a reproducing kernel Hilbert space, based on methods for solving ill-posed problems for which SVM techniques can be applied, one minimizes

$$\begin{array}{ll} \min \sum_{i,j=1}^{n} \vartheta_i\vartheta_j K(x_i, x_j, h) & \\ \text{s.t.} & \left| \hat{F}_n(x) - \sum_{j=1}^{n} \vartheta_j \int_{-\infty}^{x} K(x_j, u, h)du \right|_{x=x_i} \leq \kappa_n, \qquad 1 \leq i \leq n, \\ & \vartheta_i \geq 0 \text{ and } \sum_{i=1}^{n} \vartheta_i = 1, \end{array} \tag{8.12}$$

where κ_n is the known accuracy of approximation of $F(x)$ by $\hat{F}_n(x)$ (Mukherjee and Vapnik, 1999). To obtain the solution as a mixture of probability density functions, the kernel must be a probability density function and $\vartheta_i \geq 0$, $\sum_{i=1}^{n} \vartheta_i = 1$. Usually most of the ϑ_i values in the SVM estimate will be zero and one obtains a sparse estimate of a probability density function.

A typical property of SVM's is that the solution is characterized by a convex optimization problem, more specifically a quadratic programming (QP) problem see (8.12). But in LS-SVM's the solution is given by a linear system (equality constraints) instead of a QP problem (inequality constraints). The SVM approach (Mukherjee and Vapnik, 1999) requires inequality constraints for density estimation. One way to circumvent these inequality constraints is to use the regression-based density estimation approach. In this approach one can use the LS-SVM for regression for density estimation.

8.3 Smoothing parameter selection

Smoothing methods provide a powerful methodology for gaining insights into data. Many examples of this may be found in monographs of (Eubank, 1988;

Härdle, 1990; Müller, 1988; Scott, 1992; Silverman, 1986; Wahba, 1990; Wand and Jones, 1994). But effective use of these methods requires: (*a*) choice of the kernel, and (*b*) choice of the smoothing parameter (bandwidth).

Consider the Parzen kernel density estimator. As it turns out, the kernel density estimator is not very sensitive to the form of the kernel (Rao, 1983). An important problem is to determine the smoothing parameter. In kernel density estimation, the bandwidth has a much greater effect on the estimator than the kernel itself does. When insufficient smoothing is done, the resulting density estimate is too rough and contains spurious features. When excessive smoothing is done, important features of the underlying structure are smoothed. There are many methods for smoothing parameter selection (e.g., least-squares cross-validation, least squares plug-in methods, the double kernel method, L_1 plug-in methods, etc.). However, only two methods are considered here:

(*i*) In the least squares cross-validation method, the smoothing parameter h is chosen by

$$\min \int \left(\hat{f}_n(x) - f(x)\right)^2 dx \text{ subject to } h > 0. \tag{8.13}$$

One first derives an unbiased estimator of $\int \left(\hat{f}_n(x) - f(x)\right)^2 dx$ by observing that

$$\int \left(\hat{f}_n(x) - f(x)\right)^2 dx = \int \left(\hat{f}_n(x)\right)^2 dx + \int (f(x))^2 dx \\ - 2\int \hat{f}_n(x) dF(x). \tag{8.14}$$

The second term on the right is independent of h, and since we wish to minimize over h, only the last term needs to be estimated. However, Devroye and Lugosi (2001) provides proofs that the banddwidth selection based on L_2 would not be universally useful.

(*ii*) From the L_1point of view, one choose the smoothing parameter as the solution to

$$\min \int \left|\hat{f}_n(x) - f(x)\right| dx \quad \text{subject to } h > 0. \tag{8.15}$$

The nice properties associated with the L_2 norm, especially (8.14), do not apply to the L_1 norm. Consider the double kernel method (Devroye, 1989) for choosing the smoothing parameter. One choose the smoothing parameter as the solution to

$$\min \int \left|\hat{f}_{n,h}(x) - \hat{g}_{n,h}(x)\right| dx \quad \text{subject to } h > 0, \tag{8.16}$$

where $\hat{g}_{n,h}$ is much more accurate estimator of f than $\hat{f}_{n,h}$. Following the Devroye (1989) the following double kernel methods are considered. With the Epanechnikov kernel, the second kernel is taken to be the Berlinet-Devroye kernel defined as

$$BD = \begin{cases} \frac{1}{4}\left(7 - 31x^2\right) & \text{if } |x| \le \frac{1}{2}, \\ \frac{1}{4}\left(x^2 - 1\right) & \text{if } \frac{1}{2} < |x| \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{8.17}$$

In this thesis we use a combination of cross-validation and bootstrap for choosing the bandwidth for the Parzen kernel estimator. The algorithm is as follows:

Algorithm 23 *(Smoothing parameter selection)*

(i) Cross-Validation step. From $x_1, ..., x_n$, construct an initial estimate of the probability density function

$$\hat{f}_n(x) = \frac{1}{nh_0} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h_0}\right),$$

where h_0 is chosen by minimizing $\int E\left(\hat{f}_n(x) - f(x)\right)^2 dx$ which can be estimated by the Jackknife principle

$$\begin{aligned} CV\left(h_0\right) &= \int \left(\hat{f}_n(x)\right)^2 dx - \frac{2}{n}\sum_{k=1}^{n} \hat{f}_n^{(-k)}(x_k) \\ &= \frac{1}{n^2 h_0} \sum_{l=1}^{n}\sum_{k=1}^{n} K\left(\frac{x_l - x_k}{h_0}\right) \circ K\left(\frac{x_l - x_k}{h_0}\right) \\ &+ \frac{1}{h_0}\sum_{l=1}^{n} \frac{1}{(n-1)} \sum_{k \neq l} \frac{1}{h_0} K\left(\frac{x_l - x_k}{h_0}\right). \end{aligned} \tag{8.18}$$

where $\hat{f}_n^{(-k)}$ is the density estimate based on all of the data except x_k and $K\left(u\right) \circ K\left(u\right)$ is the convolution of the kernel with itself.

(ii) Bootstrap step

(ii.1) Construct a smoothed bootstrap sample Construct the empirical distribution, $\hat{F}_n$, which puts equal mass, $1/n$, at each observation (uniform random sampling with replacement). From the selected $\hat{F}_n$, draw a sample $x_1^*, ..., x_n^*$, called the bootstrap sample. Adding a random amount $h_0\xi$ to each x_k^*, $k = 1, ..., n$ where ξ is distributed with density $K\left(\cdot\right)$. So $x_k^{**} = x_k^* + h_0\xi$.

(ii.2) Estimate the integrated mean absolute error by

$$IMAE_{boot}\left(h, h_0\right) = \frac{1}{B}\sum_{b=1}^{B} \int \left| \hat{f}_{n,b}^{**}(x; h) - \hat{f}_n(x; h_0) \right| dx,$$

where $\hat{f}_{n,b}^{**}(x; h) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x - x_k^{**}}{h}\right)$ for $b = 1, ..., B$ and B is the number of bootstrap samples to be taken.

(ii.3) Obtain the bootstrap choice of the bandwidth h_{boot} by minimizing $IMAE_{boot}\left(h, h_0\right)$ over h.

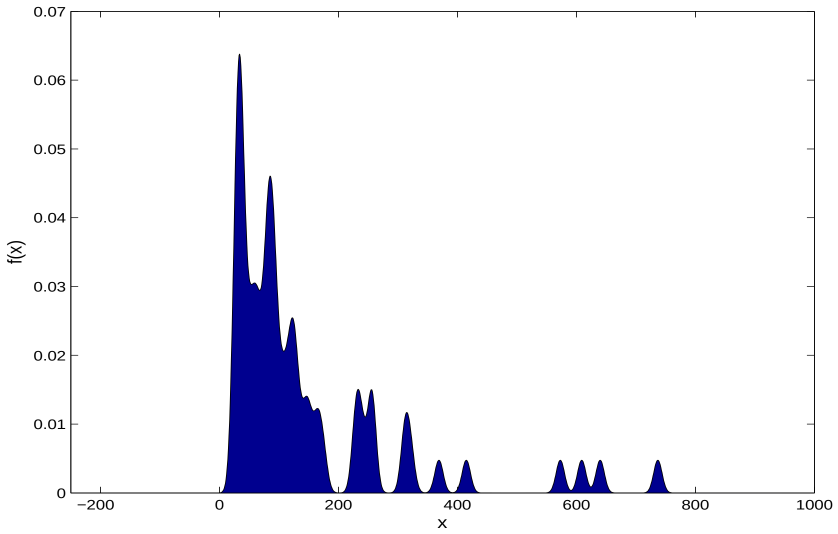

Figure 8.1: Kernel estimate for suicide data (Bandwidth: h =10). The estimate is noisy in the right-hand tail.

8.4 Regression view of density estimation

The kernel estimator suffers from a drawback when applied to data from long-tailed distributions. An example, based on the data set reported by (Copas and Fryer, 1980), of this behaviour is given by Figure 8.1 and Figure 8.2. The data set gives the lengths of treatments of control patients in a suicide study. The estimate shown in Figure 8.1 is noisy in the right-hand tail, while the estimate shown in Figure 8.2 is more smooth. Note that the data are positive, estimation of the density shown in Figure 8.2 treats the data as observations on $(-\infty, \infty)$. In order to deal with this difficulty, various adaptive methods have been proposed (Breiman *et al.*, 1977). Logspline density estimation, proposed by (Stone and Koo, 1986) and (Kooperberg and Stone, 1990), captures nicely the tail of a density but the implementation of the algorithm is extremely difficult (Gu, 1993). In this Chapter we develop a density estimation using LS-SVM regression. The proposed method has particularly advantages over Nadaraya-Watson kernel estimators, when estimates are in the tails. The data sample is prebinned and the estimator employs the bin center as the 'sample points'. This approach also provides a sparse estimate of a density. The multivariate form of the binned estimator is given in (Holmström, 2000). Consistency of multivariate data-driven histogram methods for density estimation are proved by (Lugosi and Nobel, 1996). The connection between probability density function estimation and non-parametric regression is made clear via smoothing ordered categorical data.

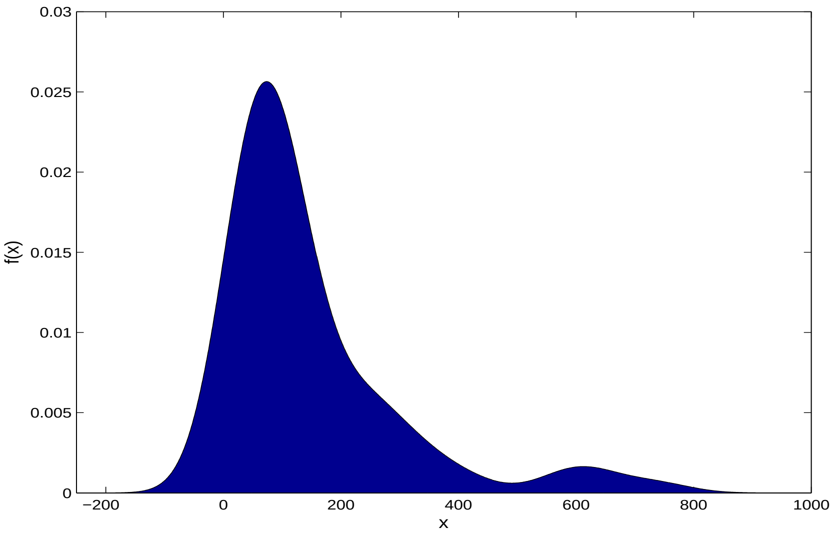

Figure 8.2: Kernel estimate for suicide data (Bandwidth: h =80). The estimate is more smooth than in Figure 8.1. The data are positive, however estimation of the density treats the data as observations on $(-\infty, \infty)$.

On the other hand, if the intention is not to look at the density function but instead to use it as an ingredient in some other statistical technique, a stategy is to first reduce the dimension (e.g., projection pursuit) of the data and then perform density estimation (e.g., Nadaraya-Watson variable kernel) in the smaller-dimensional subspace. Projection pursuit was proposed by (Friedman and Tukey, 1974), and specialized to regression by (Friedman and Stuetzle, 1981). Huber gives an overview (Huber, 1985).

8.4.1 Smoothing ordered categorical data, regression and density estimation

A categorical variable is one for which the measurement scale consists of a set of categories. There are two primary types of measurement scales for categorical variables, Many categorical scales have a natural ordering. Examples are, response to a medical treatment (excellent, good, fair, poor). Categorical variables having ordered scales are called ordinal variables. Categorical variables having unordered scales are called nominal variables. Examples are, favorite type of music (classical, country, jazz, rock, others) and mode of transportation to work (automobile, bicycle, bus, train, others). For these variables, the order of listing the categories is irrelevant, and the statistical analysis should not depend on that ordering.

Consider a series of n independent trails, in each of which just one of r mutually exclusive events $A_1, ..., A_r$ must be observed, and in which the probability of occurrence of event A_k in any trail is equal to p_k. Let $U_1, ..., U_r$ be ordinal random variables denoting the numbers of occurrences of the events $A_1, ..., A_r$ respectively, in these n trails, with $\sum_{k=1}^{r} U_k = n$. The probability model for U_k is a multinomial distribution, $U_k \backsim Bin(n, p_k)$, with $E[U_k] = np_k$ and $E[\check{p}_k] = p_k$. The elements of the vector $p = (p_1, ..., p_r)$ are called cell probabilities and n the sample size. The joint distribution of $U_1, ..., U_r$ is given (Johnson *et al.*, 1997) by

$$P\left[\bigcap_{k=1}^{r} (U_k = u_k)\right] = n! \prod_{k=1}^{r} \left(\frac{(p_k)^{u_k}}{u_k!}\right)$$

and can be expressed as

$$P[u_1, ..., u_r] = \left(\begin{array}{c} n \\ u_1, ..., u_r \end{array}\right) \prod_{k=1}^{r} (p_k)^{u_k}. \tag{8.19}$$

Ignoring constants, the relative frequencies $p_1, ..., p_r$ are the solution to

$$\begin{array}{l} \min - \sum_{k=1}^{r} U_k \log p_k \\ \text{s.t. } \sum_{k=1}^{r} p_k = 1. \end{array} \tag{8.20}$$

Most of the examples in the literature, estimation of a smooth density function f, are referred to continuous data. Since smoothness, continuity and differentiability would seem to be naturally linked to each other. For a nominal variable smoothing is not very helpful, since it is very difficult to define how "close" two categories are. An ordinal variable, where the categories do have a natural ordering, can arise as a discretization of an underlying continuous variable (e.g., $0 < x \leq 5$, $5 < x \leq 10$, $10 < x \leq 15$, etc.). For such a variable, smoothing makes sense, as it is likely that the number of observations that fall in a particular cell provide information about the probability of falling in nearby cells as well. For example, if the variable represents a discretization of a continuous variable with smooth density f, the probability vector $p = (p_1, ..., p_r)$ also reflect that smoothness, with $p_k$being close to p_l for k close to l. In many situations, the number of cells is close to the number of observations, resulting in many small (or zero) cell counts. Such a table of counts is called a sparse table. For such tables, $\hat{p}_k$ is not a good estimator of p_k, as the asymptotic approximation does not apply. Smoothing methods provide a way around this problem. Information in nearby cells can be borrowed to help provide more accurate estimation in any given cell. Let the vector $p = (p_1, ..., p_r)$ be generated from an underlying continuous variable with a smooth density function f on $[0, 1]$ through the relation $p_k = \int_{\frac{k-1}{r}}^{\frac{k}{r}} f(v)dv$. The Mean Value Theorem implies that $p_k = \frac{f(x)}{r}$ for some $x_k \in \left\{\frac{k-1}{r}, \frac{k}{r}\right\}$.

A natural way to define a smooth estimator $\hat{p} = \{\hat{p}_k\}_{k=1}^{r}$ is by analogy with a regression of outcome values $\check{p}_k = \frac{u_k}{n}$ on the equispaced design points $\frac{k}{r}, k = 1, ..., r$. The aim of a regression is to estimate $E[\check{p}_k | X = x_k] = \hat{p}_k$. So a

Nadaraya-Watson kernel estimator of $\hat{p}_k$ would be

$$\hat{p}_k = \frac{\sum_{l=1}^{r} K\left(\frac{\frac{l}{r}-\frac{k}{r}}{h}\right)\check{p}_k}{\sum_{l=1}^{r} K\left(\frac{\frac{l}{r}-\frac{k}{r}}{h}\right)}, \tag{8.21}$$

where $K(.)$ is the kernel function. Thus, in a sense all these smoothing problems can be treated as a special case of a general regression problem.

8.4.2 Design of the regression data

Suppose $z_1, ..., z_n$ is a random sample from a continuous probability density function $f(z)$. Let $A_k(z)$, $k = 1, ..., s$ be the bin interval, let $a_k(z)$ denote the left-hand endpoint of $A_k(z)$ and let $h = (a_{k+1}(z) - a_k(z))$ denote the bin width. Let U_k denote the bin count, that is, the number of the sample points falling in the bin A_k. Then the histogram is defined has

$$\hat{f}(z) = \frac{U_k}{nh} = \frac{1}{nh}\sum_{k=1}^{n} I_{[a_k, a_{k+1})}(z_k) \quad \text{for } z \in A_k, \tag{8.22}$$

where U_k has a binomial distribution, $U_k \backsim \text{Bin}(np_k(z), np_k(z)(1-p_k(z)))$ (Johnson *et al.*, 1997) with

$$p_k(z) = \int_{a_k}^{a_k+h} f(\xi)\, d\xi \ , \xi \in A_k. \tag{8.23}$$

Minimizing the mean integrated squared error, denoted by MISE$\left(\hat{f}(z), f(z,h)\right)$, one obtains

$$h = \left(\frac{6}{\int_{-\infty}^{\infty}(f'(z))^2\, dz}\right)^{\frac{1}{3}} n^{-\frac{1}{3}} \tag{8.24}$$

which, asymptotically, is the optimal choice for h. The optimal choice for h requires knowledge of the unknown underlying density f, (Tukey, 1977) and (Scott, 1979) have suggested using the Gaussian density as a reference standard and modify the normal rule when the data is skewed or heavy-tailed. Hence from (8.22)

$$\hat{h} = 3.5\hat{s}n^{-\frac{1}{3}}\kappa_1\kappa_2, \tag{8.25}$$

where $\hat{s}$ is a robust scale parameter (MAD estimator), κ_1 is a skewness correction factor and κ_2 is a kurtosis correction factor. In practice, the smoothing parameter is of the from $h^* = c\hat{h}$. (Scott, 1979) has investigated the sensitivity of the MISE to local deviations of c from 1 (for example, $c = \frac{1}{2}$ vs. $c = 2$). Based on that sensitivity analysis, one prefers the bin width $h^* = \frac{1}{2}\hat{h}$, even though it contains several spurious small bumps.

8.5 LS-SVM and density estimation

Let x_k, the independent variable, be the center of A_k, $k = 1, ..., s$. Let y_k, the dependent variable, be the proportion of the data $\{z_k\}_{k=1}^n$ falling in the interval A_k divided by the bin width h_n. Using Taylor's expansion, $f(\xi) = f(z) + (\xi - z) f^{'}(z) + O\left(h^2\right)$, for $\xi \in A_k$ in (8.23) it can be calculated that

$$E[y_k] = f(x_k) + O(h), \quad Var[y_k] = \frac{f(x_k)}{nh_n} + O\left(\frac{1}{n}\right). \tag{8.26}$$

The noise inherent in the histogram varies directly with its height see (8.26). Thus, one can regard the density estimation problem as a heteroscedastic nonparametric regression problem defined as

$$y_k = m(x_k) + \varepsilon_k, \quad \varepsilon_k = \ e_k\left[\eta\left(m(x_k), x_k\right)\right] \tag{8.27}$$

where e_k are independent and identically distributed. The function $\eta(g(x_k), x_k)$ expresses the possible heteroscedasticity and $m : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued smooth function that we wish to estimate. Recall from chapter ..., that asymptotically the heteroscedasticity does not play any important role since smoothing is conducted locally and hence the data in a small window are nearly homoscedastic. The density estimator is defined by

$$\hat{f}(x) = \mathcal{C}\left[\hat{m}_n(x)\right]_+, \tag{8.28}$$

where the constant $\mathcal{C}$ is a normalization constant such that $\hat{f}(x)$ integrates to 1 and $\hat{m}_n(x_k)$ is the LS-SVM regression smoother.

8.6 Experiments

We include two experiments. First we select some densities on a benchmark data set (Berlinet and Devroye, 1994). The group of densities contains three smooth bell-shaped ones such as the normal, lognormal and the t-distribution with 4 degrees of freedom. These have varying tail sizes and asymmetries. A continuous density with discontinuous first derivative is included: the beta$(2, 2)$. The discontinuity occurs near the extrema of the support. Next we take the simplest density that is non-smooth: the uniform density (it has no tails). All these densities are unimodal. The last two densities are mixtures of normal densities and are multimodal. The marrionite density is included to test the robustness with respect to well-separeted modes of varying scales. Let $\phi(\mu, \sigma)$ denote the normal density with mean μ and standard deviation σ. The selected densities are defined as follows:

(1) The standard normal density: $f(x) = \phi(0, 1)$.

(2) The standard lognormal density: $f(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\log x)^2}{2}\right)$ on $[0, \infty)$.

Density	Parzen density estimation	LS-SVM density estimation
	average L_1 error	average L_1 error
Standard Normal	0.0059	0.0055
Standard lognormal	0.0046	0.0041
$t\,(4)$	0.0033	0.0033
Beta$(2,2)$	0.0828	0.0494
$U\,[0,1]$	0.0222	0.0272
Skewed bimodal	0.0088	0.0085
marronite density	0.0212	0.0142

Table 8.1: LS-SVM estimate of the error variance: mean, bias and variance for 100 replications.

(3) The central t-distribution: $f\,(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}$, $v > 0$, $-\infty < x < \infty$.

(4) The beta$(2,2)$ density: $f\,(x) = 6x\,(1-x)\,,\ 0 \leq x \leq 1$.

(5) The uniform density on $[0,1]$.

(6) The skewed bimodal density: a normal mixture: $f\,(x) = \frac{3}{4}\phi\,(0,1) + \frac{1}{4}\phi\left(1.5,\frac{1}{3}\right)$.

(7) The marronite density: another normal mixture: $f\,(x) = \frac{1}{3}\phi\left(-20,\frac{1}{4}\right) + \frac{2}{3}\phi\,(0,1)$.

For each of the densities, we generated 50 samples of size 500, and used the Parzen density estimator and the LS-SVM regression estimator. We use a combination of cross-validation and bootstrap for choosing the bandwidth for the Parzen kernel estimator. The average L_1 errors are estimated for each density (Table 8.1). Both methods gives similar results.

In the last experiment we apply both methods to the suicide data (Copas and Fryer, 1980). Note that the data are positive, the estimates shown in Figure 8.3 that the Parzen estimator treating the data as observations on $(-\infty,\infty)$, while the LS-SVM (RBF kernel) estimate deals with this difficulty.

8.7 Conclusions

The SVM approach (Mukherjee and Vapnik, 1999) requires inequality constraints for density estimation. One way to circumvent these inequality constraints is to use the regression-based density estimation approach. In this

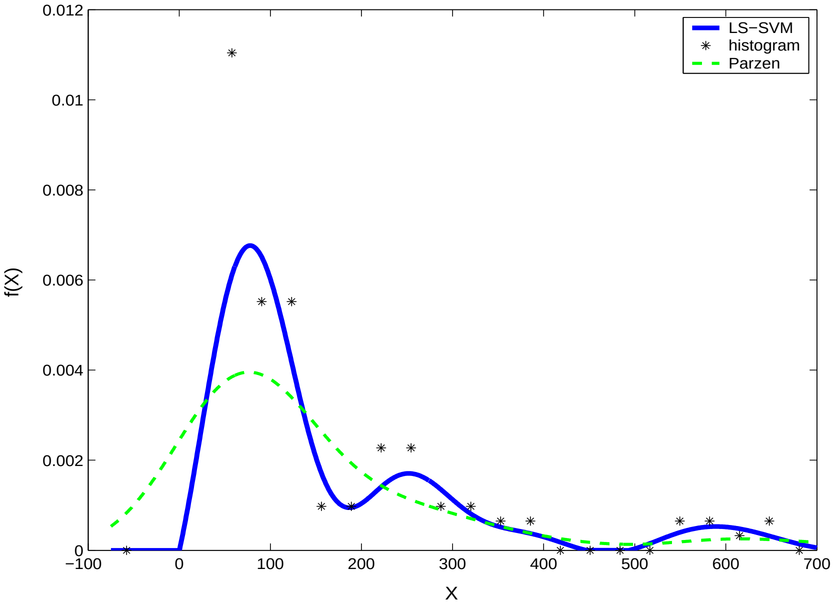

Figure 8.3: Estimates for the suicide data study. The * points denotes the centers of the histogram bins.

approach one can use the LS-SVM for regression for density estimation. The proposed method (density estimation using LS-SVM regression) has particularly advantages over Nadaraya-Watson kernel estimators, when estimates are in the tails. The data sample is pre-binned and the estimator employs the bin center as the 'sample points'. This approach also provides a sparse estimate of a density. The multivariate form of the binned estimator is given in (Holmström, 2000). Consistency of multivariate data-driven histogram methods for density estimation are proved by (Lugosi and Nobel, 1996).

In the first experiment we used the Parzen density estimator and the LS-SVM regression estimator. We used a combination of cross-validation and bootstrap for choosing the bandwidth for the Parzen kernel estimator. The average L_1 errors are estimated for each density. Both methods gives similar results (Table 8.1).

In the last experiment we applied both methods to the suicide data (Copas and Fryer, 1980). Note that the data are positive, the estimates shown in Figure 8.3 that the Parzen estimator treating the data as observations on $(-\infty, \infty)$, while the LS-SVM (RBF kernel) estimate deals with this difficulty. In order to deal with this difficulty, various adaptive methods have been proposed (Breiman *et al.*, 1977). Logspline density estimation, proposed by (Stone and Koo, 1986) and (Kooperberg and Stone, 1990), captures nicely the tail of a density but the implementation of the algorithm is extremely difficult (Gu, 1993).

Part III

ROBUSTIFYING LS-SVM REGRESSION MODELLING

Chapter 9

Robustness

In the previous chapters basic methods for LS-SVM regression models were discussed. The use of least squares and equality constraints for the models results into simpler formulations but on the other hand has potential drawback such as the lack of robustness. In this chapter we explain some measures of robustness. Next, we will discuss ways to enhance the robustness of LS-SVM models for nonlinear function estimation by incorporating methods from robust statistics. Weighted LS-SVM versions are explained in order to cope with outliers in the data (Suykens *et al.*, 2002). In order to understand the robustness of these estimators against outliers, we use the empirical influence function and maxbias curves. We also discuss how these robust techniques can be applied to the primal as well as dual representations of LS-SVMs. Contributions are made in Section 9.3.

9.1 Introduction

A common view on robustness is to provide alternatives to least squares methods and Gaussian theory. In fact, a statistical procedure based on L_2 works well in situations where many assumptions (such as normality, no outliers) are valid. These assumptions are commonly made, but are usually at best approximations to reality. For example, non-Gaussian noise and outliers are common in datasets and are dangerous for many statistical procedures (Hampel *et al.*, 1986). The importance of using robust statistical methods was recognized by eminent statisticians like (Pearson, 1902; Student, 1927; Jeffreys, 1939; Box, 1953; Tukey, 1960, 1962). It was convincingly demonstrated by Pearson for tests and by Tukey for estimators. Pearson (Pearson, 1929, 1931) showed the nonrobustness even of the level of chisquare- and F-tests for variances; in the context, Box (Box, 1953) and Box & Andersen (12) introduced the term "robust". Tukey (Tukey, 1960), (summarizing earlier work) showed the nonrobustness of the arithmetic mean even under slight deviations from normality. However, the modern theory of robust statistics emerged more recently, led by Huber's (1964) classic minimax approach and Hampel's (1974) introduction of influence functions.

Huber (Huber, 1964) gave the first theory of robustness. He considered the general gross-error model or ϵ-contamination model

$$\mathcal{U}\left(F_0, \epsilon\right) = \left\{F : F\left(x\right) = \left(1-\epsilon\right) F_0(x) + \epsilon G(x), \quad 0 \leq \epsilon \leq 1\right\}, \tag{9.1}$$

where $F_0(x)$ is some given distribution (the ideal nominal model), $G(x)$ is an arbitrary continuous distribution and ϵ is the first parameter of contamination. The contamination scheme describes the case where, with large probability $(1-\epsilon)$ the data occurs with distribution $F_0(x)$ and with small probability ϵ outliers occur according to the distribution $G(x)$. Examples of the ϵ-contamination model are:

Example 24 *ϵ-contamination model with symmetric contamination*

$$F\left(x\right) = \left(1-\epsilon\right)\mathcal{N}(0,\sigma^2) + \epsilon\mathcal{N}(0,\kappa^2\sigma^2), \quad 0 \leq \epsilon \leq 1, \quad \kappa > 1. \tag{9.2}$$

Example 25 *ϵ-contamination model for the mixture of the normal and Laplace or double exponential distribution*

$$F\left(x\right) = \left(1-\epsilon\right)\mathcal{N}(0,\sigma^2) + \epsilon\,\mathrm{Lap}(0,\lambda), \quad 0 \leq \epsilon \leq 1, \tag{9.3}$$

where respectively κ and λ are the second parameters of contamination describing the rate of variance of $G(x)$ over the variance of $F_0(x)$ ($\kappa \gg 1$). He considered also the class of M-estimators of location (also called estimating equation, generalized maximum likelihood estimators) described by some suitable function. The Huber estimator is a minimax solution: it minimizes the maximum asymptotic variance over all F in the gross-error model. The gross-error model can be interpreted as yielding exactly normal data with probability $(1-\epsilon)$, and gross errors (or some other, "contaminating" distribution) with small probability ϵ (usually between 0% and 5%).

Huber developed a second theory (Huber, 1965, 1968) and (Huber and Strassen, 1973) for censored likelihood ratio tests and exact finite-sample confidence intervals, using more general neighborhoods of the normal model. This approach may be mathematically deepest, but seems very hard to generalize and therefore plays hardly any role in applications.

Hampel developed a third (Hampel, 1968, 1971, 1974) and (Hampel *et al.*, 1986) also very closely related robustness theory which is more generally applicable than Huber's first and second theory. Three main concepts are introduced: qualitative robustness, which is essentially continuity of the estimator viewed as functional in the weak topology; the influence curve (IC) or influence function (IF), which describes the first derivative of the estimator, as far as existing; and the breakdown point (BP), a global robustness measure which describes how many percent gross errors are still tolerated. Small perturbations should have small effects; a first order Taylor expansion describes the small effects quantitatively (often in very good approximation); and the breakdown point tells under which load the bridge breaks down (or the estimator is totally unreliable). Thus, we can call robustness theory the stability statistical procedures.

In historical summary, robustness has provided at least two major insights into statistical theory and practice: (i) Relatively small perturbations form nominal models can have very substantial deleterious effects on many commonly used statistical procedures and methods (as in Tukey's example at the beginning of this article). (ii) Robust methods are needed for detecting or accommodating outliers in the data (Rousseeuw and Hubert, 1999).

Besides the books by (Huber, 1981) and (Hampel et al. 1986), some other books on robust statistics are (Rousseeuw and Leroy, 1987; Staudte and Shaether, 1990; Stahel and Weisberg, 1991; Morgenthaler *et al.*, 1993); more on the mathematical side (Rieder, 1994; Jureckova and Sen, 1996); applied books with relevance for robustness (Mosteller and Tukey, 1977; Box *et al.*, 1978; Hoaglin *et al.*, 1983; Gnanadesikan, 1977; Box *et al.*, 1983); on special related topics (Müller, 1997; Morgenthaler and Tukey, 1991); on computer programs for robust statistics (Marazzi, 1993).

9.2 Measures of Robustness

In order to understand why certain estimators behave the way they do, it is necessary to look at the various measures of robustness. There exist a large variety of approaches towards the robustness problem. The approach based on influence functions (Hampel, 1968, 1974) will be used here. The effect of one outlier on the estimator can be described by the influence function (IF) which (roughly speaking) formalizes the bias caused by one outlier. Another measure of robustness of an estimator is the maxbias curve. The maxbias curve gives the maximal bias that an estimator can suffer from when a fraction of the data come from a contaminated distribution. By letting the fraction vary between zero and the breakdown value a curve is obtained. The breakdown value is how much contaminated data an estimator can tolerate before it becomes useless.

9.2.1 Influence functions and breakdown points

Let F be a fixed distribution and $T\left(F\right)$ a statistical functional defined on a set $\mathcal{U}$ of distributions satisfying some regularity conditions (Hampel *et al.*, 1986). Let the estimator $T_n = T(\hat{F}_n)$ of $T(F)$ be the functional of the sample distribution F_n.

Definition 26 *(Influence function). The influence function $IF(x;T,F)$ is defined as*

$$IF\left(x;T,F\right) = \lim_{\epsilon\downarrow 0} \frac{T\left[\left(1-\epsilon\right)F + \epsilon\Delta_x\right] - T\left(F\right)}{\epsilon}. \tag{9.4}$$

Here Δ_x denotes the pointmass distribution at the point x.

The IF reflects the bias caused by adding a few outliers at the point x, standardized by the amount ϵ of contamination. Note that this kind of differentiation of statistical functionals is a differentiation in the sense of von Mises (Fernholz,

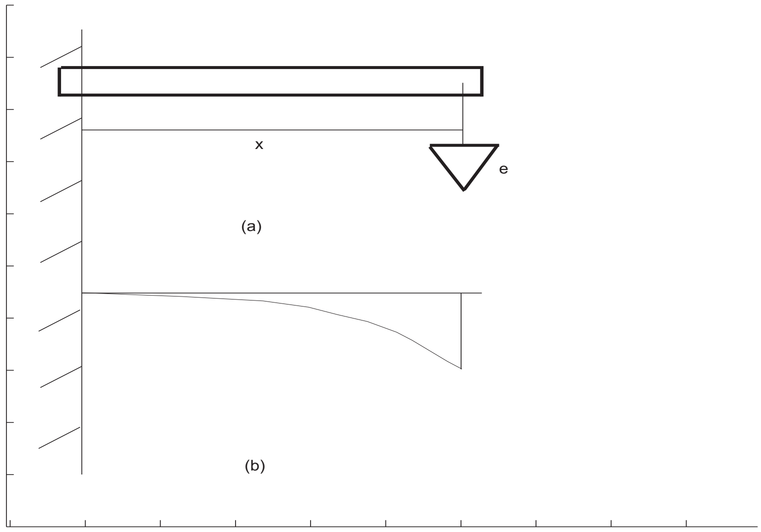

Figure 9.1: Illustration for showing the difference between the concept of influence function and breakdown point. The breakdown point is proportional to the width of the beam multiplied by the square of its height. The Influence function is just a plot of the elasticity as a function of x, the postion of the external force e.

1983). ¿From the influence function, several robustness measures can be defined: the gross error sensitivity, the local shift sensitivity and the rejection point (Hampel, 1968, 1974). Mathematically speaking, the influence function is the set of all partial derivatives of the functional T in the direction of the point masses. For functionals, there exist several concepts of differentiation; Gâteaux, Hadamard of compact, and Fréchet derivative have been used in statistics, the Fréchet derivative being the strongest concept and formely considered to be very rarely applicable; but the main reason for this belief seems to be the non-robustness of most classical estimators, while at least some (if not most) smooth M-estimators are indeed Fréchet-differentiable (Clarke 1983, 1986), (Bedmarski, 1993). The IF describes the derivative of a functional in whatever sense it exists.

A mechanical analogy of the concept of influence function is shown in Figure 9.1. Given a beam which is fixed at one end and a stone with weight ϵ is attached on the other end.

The influence function is a plot of the elasticity (based on the differential equation of the elastic line) as a function of x, the position of the weight. The elastic line is given in Figure 9.1.b.

Definition 27 *(Maxbias curve). Let $T(F)$ denote a statistical functional and let the contamination neighborhood of F be defined by (9.1) for a fraction of* *contamination ϵ. The maxbias curve is then defined by*

$$B\left(\epsilon, T(F)\,, F\right) = \sup_{F \in \mathcal{U}(F_0,\epsilon)} \left| \frac{T(F) - T(F_0)}{\sigma_0} \right|$$

Definition 28 *(Breakdown point). The breakdown point ϵ^* of the estimator $T_n = T(\hat{F}_n)$ for the functional $T(F)$ at F is defined by*

$$\epsilon^*(T, F) = \inf\left\{\epsilon > 0 \,|B\left(\epsilon, T(F)\,, F\right) = \infty\right\}. \tag{9.5}$$

This notion defines the largest fraction of gross errors that still keeps the bias bounded.

A mechanical analogy of the concept of breakdown point is very simple, how heavy does the weight ϵ has to be made such that the beam breaks? The breakdown point is proportional to the width of the beam multiplied by the square of its height. Next, we will give some examples of influence functions and breakdown points for location estimators and scale estimators.

Example 29 *(sample mean). The corresponding functional $T\left(F\right) = \int x dF\left(x\right)$ of the mean is defined for all probability measures with existing first moment. From (9.4), it follows that*

$$\begin{aligned} IF\left(x; T, F\right) &= \lim_{\epsilon \downarrow 0} \frac{\int xd\left[(1-\epsilon)F + \epsilon\Delta_x\right](x) - \int xdF(x)}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{\epsilon \int xd\Delta_x(x) - \epsilon \int xdF(x)}{\epsilon} \\ &= x - T(F). \end{aligned} \tag{9.6}$$

The IF of the sample mean is sketched in Figure 9.2. We see that the IF is unbounded in $\mathbb{R}$.

This means that an added observation at a large distance from $T(F)$ gives a large value in absolute sense for the IF. The finite sample breakdown point of the sample mean has $\epsilon_n^* = 1/n$, often the limiting value $\lim_{n\to\infty} \epsilon_n^* = 0$ will be used as a measure of the global stability of the estimator. One of the more robust location estimators is the median. Although the median is much more robust (breakdown point is 0.5) than the mean, its asymptotic efficiency is low. But in the asymmetric distribution case the mean and the median does not estimate the same quantity.

Example 30 *($(0, \beta_2)$-trimmed mean). The corresponding statistical functional for the $(0, \beta_2)$-trimmed mean is given in terms of a quantile function and is defined as*

$$\mu_{(0,\beta_2)} = T_{(0,\beta_2)}(F) = \frac{1}{1-\beta_2}\int_0^{F^-(1-\beta_2)} xdF(x) = \frac{1}{1-\beta_2}\int_0^{(1-\beta_2)} F^-\left(q\right) d(q). \tag{9.7}$$

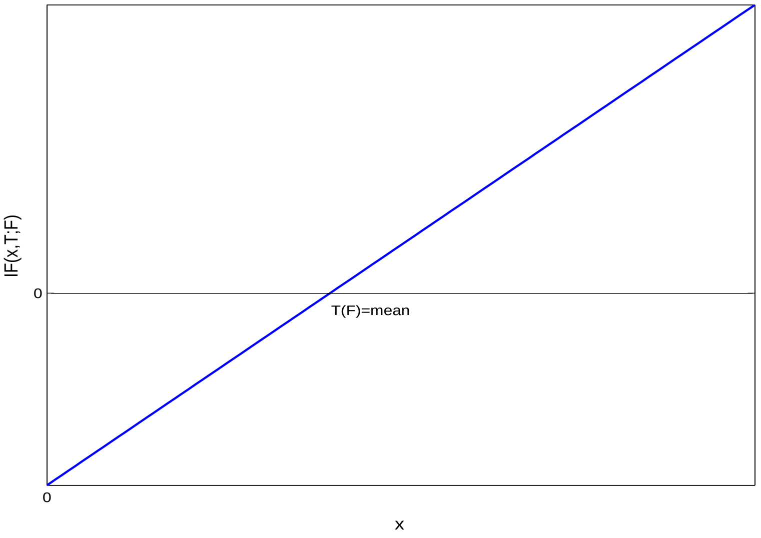

Figure 9.2: Influence function of the mean. The influence function is unbounded in $\mathbb{R}$.

The quantile function of a cumulative distribution function F is the generalized inverse $F^{-}:(0,1)\rightarrow\mathbb{R}$ given by

$$F^{-}(q)=\inf\left\{x:F(x)\geq q\right\}. \tag{9.8}$$

In the absence of information concerning the underlying distribution function F of the sample, the empirical distribution function F_N and the empirical quantile function F_n^{-1} are reasonable estimates for F and F^{-}, respectively. The empirical quantile function is related to the order statistics $x_{n(1)}\leq\ldots\leq x_{n(n)}$ of the sample through

$$F^{-}(q)=x_{n(i)},\quad \textit{for } q\in\left(\frac{i-1}{n},\frac{i}{n}\right). \tag{9.9}$$

To derive the influence function $IF(x;F^{-}(q),F)$ for the qth quantile functional $F^{-}(q)$, assume that F has a density f which is continuous and positive at $x_q=F^{-}(q)$. Let $F_\epsilon=F+\epsilon\left(\Delta_x-F\right)$ and apply (9.8)

$$\begin{aligned} T\left[F+\epsilon\left(\Delta_{x_0}-F\right)\right] &= \inf\left\{x:F(x)+\epsilon\left(\Delta_{x_0}(x)-F(x)\right)\geq q\right\} \\ &= \inf\left\{x:F(x)+\epsilon\left[I\left(x\geq x_0\right)-F(x)\right]\geq q\right\} \\ &= \inf\left\{x:F(x)\geq\frac{q-\epsilon\left[I\left(x\geq x_0\right)\right]}{(1-\epsilon)}\right\}. \end{aligned} \tag{9.10}$$

One finds $IF(x;F^{-}(q),F)=(\partial/\partial\epsilon)\left[F_\epsilon^{-1}(q)\right]_{\epsilon=0}$ indirectly by first calculating $(d/d\epsilon)\left[F_\epsilon^{-1}(q)\right]$ for $\epsilon>0$ and then taking $\lim_{\epsilon\downarrow 0}(d/d\epsilon)\left[F_\epsilon^{-1}(q)\right]$. From (9.10) *we have* $F_\epsilon^{-1}(q) = F^-\left(\frac{q-\epsilon[I(x\geq x_0)]}{(1-\epsilon)}\right)$. *Thus*

$$\begin{aligned}\frac{d}{d\epsilon}\left[F^-\left(\frac{q-\epsilon\left[I\left(x\geq x_0\right)\right]}{(1-\epsilon)}\right)\right] &= \frac{\frac{d}{d\epsilon}\left(\frac{q-\epsilon[I(x\geq x_0)]}{(1-\epsilon)}\right)}{f\left(F^-\left(\frac{q-\epsilon[I(x\geq x_0)]}{(1-\epsilon)}\right)\right)} \\ &= \frac{q-I\left(x_o\leq F^-\left(q\right)\right)}{f\left(F^-\left(\frac{q-\epsilon[I(x\geq x_0)]}{(1-\epsilon)}\right)\right)},\end{aligned}$$

so that

$$\lim_{\epsilon\downarrow 0}\frac{d}{d\epsilon}\left[F^-\left(\frac{q-\epsilon\left[I\left(x\geq x_0\right)\right]}{(1-\epsilon)}\right)\right] = \frac{q-I\left(x_0\leqslant F^-\left(q\right)\right)}{f\left(F^-\left(q\right)\right)}. \tag{9.11}$$

The $IF(x;F^-(q),F)$ *is*

$$IF(x;F^-(q),F) = \begin{cases}\frac{(q-1)}{f(F^-(q))} & x_0 < F^-\left(q\right) \\ 0 & x_0 = F^-\left(q\right) \\ \frac{q}{f(F^-(q))} & x_0 > F^-\left(q\right).\end{cases} \tag{9.12}$$

Now we can calculate the influence function of the $(0,\beta_2)$*-trimmed means. Define*

$$\begin{aligned}T_{(0,\beta_2)}(F_\epsilon) &= \frac{1}{1-\beta_2}\int_0^{F_\epsilon^{-1}(1-\beta_2)} ydF_\epsilon(y) \\ &= \frac{1}{1-\beta_2}\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)} ydF(y) + \epsilon\int_0^{F_\epsilon^{-1}(1-\beta_2)} yd\left(\Delta_x - F\right)(y)\right].\end{aligned} \tag{9.13}$$

We will find $IF(x;\mu_{(0,\beta_2)},F) = (d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]_{\epsilon=0}$ *indirectly by first calculating* $(d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]$ *for* $\epsilon > 0$ *and then taking* $\lim_{\epsilon\downarrow 0}(d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]$. *From (9.13)*

$$\begin{aligned}\frac{d}{d\epsilon}\left[T_{(0,\beta_2)}(F_\epsilon)\right] &= \frac{F_\epsilon^{-1}\left(1-\beta_2\right)}{(1-\beta_2)}f\left(F_\epsilon^{-1}\left(1-\beta_2\right)\right)\frac{d}{d\epsilon}\left[F_\epsilon^{-1}\left(1-\beta_2\right)\right] \\ &\quad + (\frac{1}{1-\beta_2})\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)} yd\left(\Delta_x - F\right)(y)\right] \\ &\quad + \epsilon(\frac{1}{1-\beta_2})\frac{d}{d\epsilon}\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)} yd\left(\Delta_x - F\right)(y)\right],\end{aligned} \tag{9.14}$$

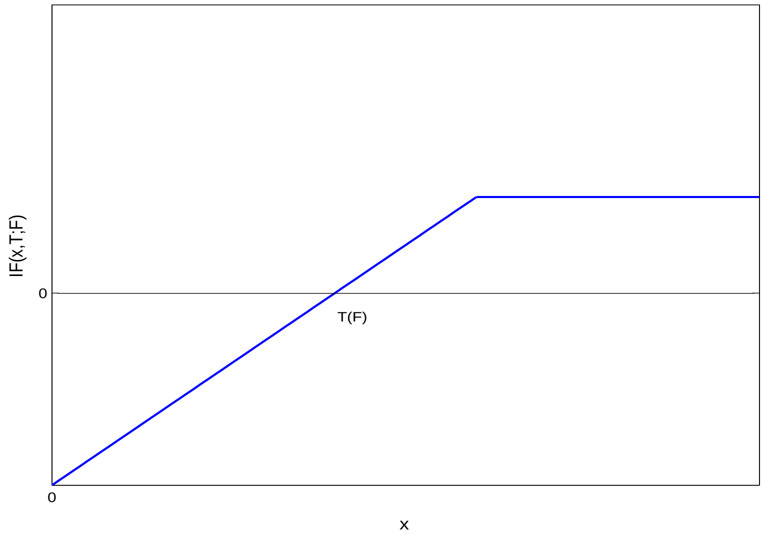

Figure 9.3: The influence function of the $(0, \beta_2)$-trimmed mean. The influenece function is both continuous and bounded in $\mathbb{R}$.

so that

$$\begin{aligned}
IF(x; \mu_{(0,\beta_2)}, F) &= \lim_{\epsilon \downarrow 0} \frac{d}{d\epsilon} \left[T_{(0,\beta_2)}(F_\epsilon) \right] \\
&= \frac{F^-(1-\beta_2)}{(1-\beta_2)} f\left(F^-(1-\beta_2)\right) IF(x; F^-(q), F) \\
&+ \frac{1}{1-\beta_2} \left[\int_0^{F^-(1-\beta_2)} y d\Delta_x(y) - \int_0^{F^-(1-\beta_2)} y dF(y) \right] \\
&= \frac{F^-(1-\beta_2)}{(1-\beta_2)} f\left(F^-(1-\beta_2)\right) IF(x; F^-(q), F) - \mu_{(0,\beta_2)} \\
&+ \frac{x}{(1-\beta_2)} I\left(x \le F^-(1-\beta_2)\right). \qquad (9.15)
\end{aligned}$$

Substituting the influence function $IF(x; F^-(q), F)$, with $q = (1-\beta_2)$, given in (9.12) into (9.15) yields:

$$IF(x; \mu_{(0,\beta_2)}, F) = \begin{cases} \frac{x - \beta_2 F^-(1-\beta_2)}{1-\beta_2} - \mu_{(0,\beta_2)} & 0 \le x \le F^-(1-\beta_2) \\ F^-(1-\beta_2) - \mu_{(0,\beta_2)} & F^-(1-\beta_2) < x. \end{cases} \qquad (9.16)$$

The IF of the $(0, \beta_2)$-trimmed mean is sketched in Figure 9.3. Note that it is both continuous and bounded in $\mathbb{R}$.
The finite sample breakdown point of the $(0, \beta_2)$-trimmed mean has $\epsilon_n^* = (\lfloor n\beta_2 \rfloor + 1)/n$ and the limiting value $\lim_{n\to\infty} \epsilon_n^* = \beta_2$.

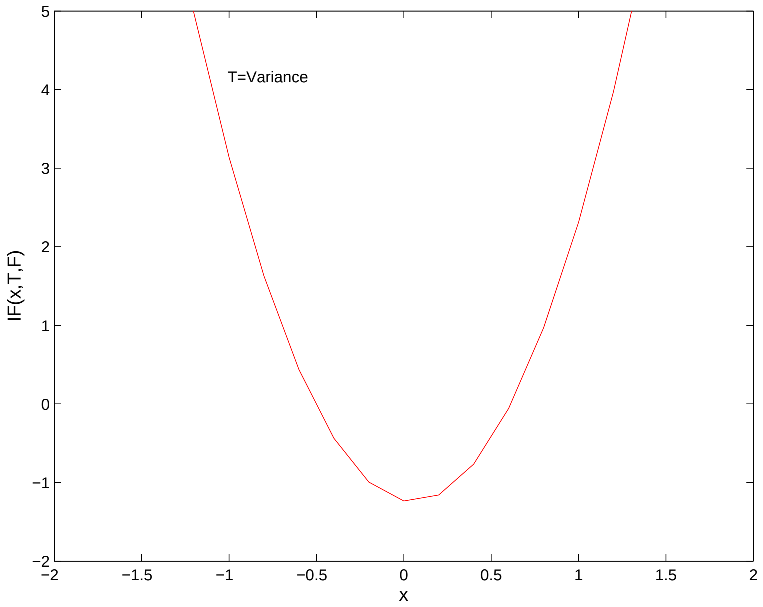

Figure 9.4: The influence function of the sample variance. The influence function is unbounded in $\mathbb{R}$.

Example 31 *(variance). Given the corresponding functional* $T(F) = \int (x-\mu)^2 \, dF(x)$ *and from (9.4), it follows that*

$$\begin{aligned} IF(x;T,F) &= \lim_{\epsilon \downarrow 0} \frac{\int xd\left[(1-\epsilon)F + \epsilon\Delta_x\right](x) - \int xdF(x)}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{\epsilon \int (x-\mu)^2 \, d\Delta_x(x) - \epsilon \int (x-\mu)^2 \, dF(x)}{\epsilon} \\ &= (x-\mu)^2 - T(F). \end{aligned} \tag{9.17}$$

The IF of the sample variance is sketched in Figure 9.4. We see that the IF is unbounded in $\mathbb{R}$.

This means that an added observation at a large distance from $T(F)$ gives a large value in absolute sense for the IF. The finite sample breakdown point of the sample variance has $\epsilon_n^* = 1/n$, often the limiting value $\lim_{n\to\infty} \epsilon_n^* = 0$ will be used as a measure of the global stability of the estimator.

9.2.2 Empirical influence functions

The most important empirical versions of (9.4) are the sensitivity curve (Tukey, 1970) and the Jackknife (Quenouille, 1956) and (Tukey, 1958).

The sensitivity curve

There are two versions, one with addition and one with replacement. In the case of an additional observation, one starts with a sample $(x_1, ..., x_{n-1})$ of size $n-1$. Let $T(F)$ be an estimator. Let $T(F)$ be an estimator and let $T(\hat{F}_{n-1}) = T(x_1, ..., x_{n-1})$ be denote the estimate. The change in the estimate when an nth observation $x_n = x$ is included is $T(x_1, ..., x_{n-1}, x) - T(x_1, ..., x_{n-1})$. One multiply the change by n and the result is the sensitivity curve.

Definition 32 *(sensitivity curve) One obtains the sensitivity curve if one replaces F by $\hat{F}_{n-1}$ and ϵ by $\frac{1}{n}$ in (9.4):*

$$\begin{aligned} SC_{n-1}(x, T, \hat{F}_{n-1}) &= \frac{T\left[\left(\frac{n-1}{n}\right)\hat{F}_{n-1} + \frac{1}{n}\Delta_x\right] - T\left(\hat{F}_{n-1}\right)}{\frac{1}{n}} \\ &= (n-1)\,T\left(\hat{F}_{n-1}\right) + T(\Delta_x) - nT\left(\hat{F}_{n-1}\right) \\ &= n\left[T_n(x_1, ..., x_{n-1}, x) - T_{n-1}(x_1, ..., x_{n-1})\right]. \qquad (9.18) \end{aligned}$$

Example 33 *(mean) Consider the one-sample symmetric Gaussian location model defined by*

$$X_k = E[X] + e_k,\ \ k = 1, ..., n, \qquad (9.19)$$

where the errors are i.i.d., and symmetric about 0 with common density f and F. If the error distribution is normal, $\bar{X}$ is the best estimate in a variety of senses. Let $T(F) = \mu = E[X]$ denote the mean in a population and let $x_1, ..., x_{n-1}$ denote a sample from that population. The sensitivity curve of the mean is then

$$\begin{aligned} SC_{n-1}(x, \bar{x}, \hat{F}_{n-1}) &= n\left(\hat{\mu}(x_1, ..., x_{n-1}, x) - \hat{\mu}(x_1, ..., x_{n-1})\right) \\ &= n\left(\frac{1}{n}\sum x_i + \frac{1}{n}x - \hat{\mu}(x_1, ..., x_{n-1})\right) \\ &= (n-1)\,\hat{\mu}(x_1, ..., x_{n-1}) + x - n\hat{\mu}(x_1, ..., x_{n-1}) \\ &= x - \hat{\mu}(x_1, ..., x_{n-1}). \end{aligned}$$

Example 34 *(median) The sample median is defined by*

$$med = \begin{cases} x_{n(k+1)} & \text{if } n = 2k+1 \\ \left(x_{n(k)} + x_{n(k+1)}\right)\frac{1}{2} & \text{if } n = 2k \end{cases} \qquad (9.20)$$

where $x_{n(1)} \leq ... \leq x_{n(n)}$ are the order statistics. The sensitivity curve of the mean is then

$$SC_{n-1}(x, med, \hat{F}_{n-1}) = n\left(med(x_1, ..., x_{n-1}, x) - med(x_1, ..., x_{n-1})\right). \quad (9.21)$$

Depending on the rank of x, the sensitivity curve of the median is given by

$$SC_{n-1}(x, med, \hat{F}_{n-1}) = \begin{cases} n\left(x_{(k)} - med(x_1, ..., x_{n-1})\right) & \text{for } x < x_{(k)} \\ x & \text{for } x_{(k)} \leq x \leq x_{(k+1)} \\ n\left(x_{(k+1)} - med(x_1, ..., x_{n-1})\right) & \text{for } x > x_{(k+1)} \end{cases} \qquad (9.22)$$

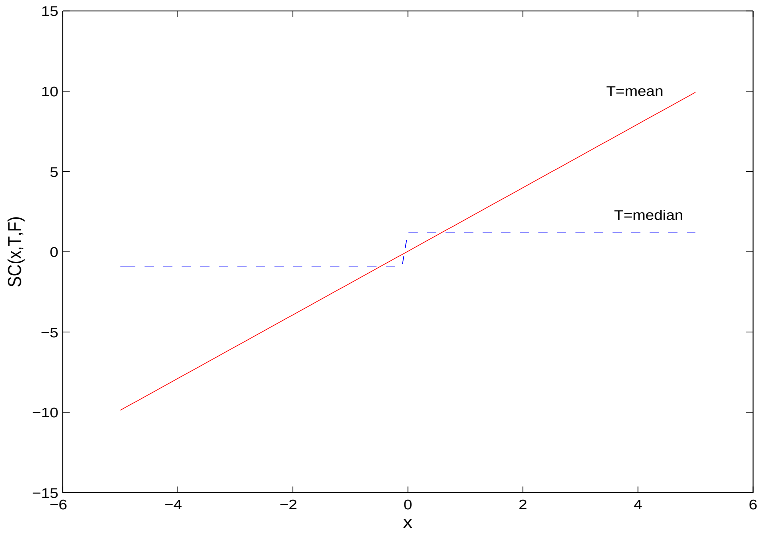

Figure 9.5: Sensitivity curve of the sample mean and the sample median.

Given an univariate data set $x = (x_1, ..., x_{100})^T$ where $x \sim \mathcal{N}\left(0, 1^2\right)$. Location estimators applied to this sample are the sample mean and the sample median. We show the sensitivity curve for the location estimators in Figure (9.5). The most important aspect is that the sensitivity curve of the mean becomes unbounded for both $x \to \infty$ and $x \to -\infty$, whereas the median remains constant.

Example 35 *(spread). Let $T(F) = \sigma^2$ denote the variance in a population and let $x_1, ..., x_n$ denote a sample from that population. Then $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ is the plug-in estimate of σ^2. Shift the horizontal axis so that $\sum_{k=1}^{n-1} x_k = 0$. The sensitivity curve of the variance is then*

$$\begin{aligned}
SC_{n-1}(x, \hat{\sigma}^2, \hat{F}_{n-1}) &= n\left(\hat{\sigma}_n^2 - \hat{\sigma}_{n-1}^2\right) \\
&= n\left(\frac{1}{n}\sum_{i=1}^{n-1}(x_i - \bar{x}_n)^2 + (x - \bar{x}_n)^2 - \hat{\sigma}_{n-1}^2\right) \\
&= \left(\sum_{i=1}^{n-1}(x_i - \bar{x}_n)^2 + \left(\frac{n-1}{n}\right)x^2 - n\hat{\sigma}_{n-1}^2\right) \\
&= (n-1)\,\hat{\sigma}_{n-1}^2 + \left(\left(\frac{n-1}{n}\right) + \frac{1}{n^2}\right)x^2 - n\hat{\sigma}_{n-1}^2 \\
&= \left(\left(\frac{n-1}{n}\right) + \frac{1}{n^2}\right)x^2 - \hat{\sigma}_{n-1}^2 \simeq x^2 - \hat{\sigma}_{n-1}^2.
\end{aligned}$$

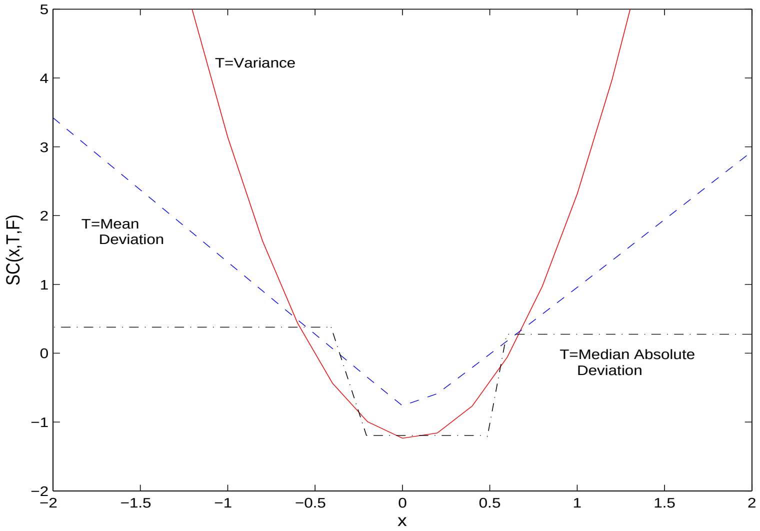

Figure 9.6: Sensitivity curves of mean deviation, median absolute deviation and the sample variance

Scale estimators applied to the sample $x = (x_1, ..., x_{100})^T$where $x \sim \mathcal{N}\left(0, 1^2\right)$, are the sample variance, the Mean Deviation and the Median Absolute Deviation. The Mean Deviation is defined as

$$T\left(\hat{F}_n\right) = \frac{1}{n}\sum_{k=1}^{n} |x_k - \bar{x}| \,. \tag{9.23}$$

This estimator is nonrobust to outliers and has a breakdown point $\epsilon^* = 0$. The Median Absolute Deviation (MAD), one of the more robust scale estimators, is defined as

$$T\left(\hat{F}_n\right) = med\,|x_k - med\,(x_i)| \,. \tag{9.24}$$

This estimator has a high breakdown point $\epsilon^* = 0.5$. We show the sensitivity curve for the scale estimators in Figure 9.6.

The most important aspect is that the sensitivity curves of the variance and the Mean Deviation become unbounded for both $x \to \infty$ and $x \to -\infty$, whereas the sensitivity curve of the Median Absolute Deviation is bounded.

Jackknife approximation

An other approach to approximating the IF, but only at the sample values $x_1, ..., x_n$ themselves, is the Jackknife.

Definition 36 *(The Jackknife approximation). If one substitutes $\hat{F}_n$ for F and* -$\frac{1}{(n-1)}$ *for ϵ in (9.4), one obtains*

$$\begin{aligned} J_{IF}\left(x_i, T, F_n\right) &= \frac{T\left[\left(\frac{n}{n-1}\right) F_n - \frac{1}{n-1}\Delta_{x_i}\right] - T\left(F_n\right)}{-\frac{1}{n-1}} \\ &= -\left(n-1\right)\left[\left(T\left(F_n\right) - T(\Delta_{x_i})\right) - T\left(F_n\right)\right] \\ &= \left(n-1\right)\left[T_n\left(x_1, ..., x_n\right) - T_{n-1}\left(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\right)\right]. \end{aligned} \tag{9.25}$$

In some cases, namely when the influence function does not depend smootly on F, the Jackknife is in trouble.

9.3 Residuals and Outliers in Regression

Residuals are used in many procedures designed to detect various types of disagreement between data and an assumed model. In this section, we consider observations that do not belong to the model and often exhibit numerically large residuals and, in case they do, they are called outliers. This type of situation is a special case of heteroscedasticity and is prevented by imposing the condition $E\left[e_k^2\right] = \sigma^2$.

Although the detection of outliers in a univariate sample has been investigated extensively in the statistical literature (see Barnett and Lewis, 1984), the word outlier has never been given a precise definition. For example, we use the one of (Barnett and Lewis, 1984). A quantitative definition has been given by (Davis and Gather, 1993).

Definition 37 *(Barnett and Lewis, 1984). An outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.*

A good elementary introduction to residuals and outliers is given by Fox (Fox, 1991). More advanced treatments are given by Cook and Weisberg (Cook and Weisberg, 1982) and by Atkinson (Atkinson, 1985).

9.3.1 Linear regression

As an example, let the simple linear model assumes a relation of the type

$$y_k = \beta_0 + \beta_1 x_k + e_k, \quad k = 1, ..., n \tag{9.26}$$

in which the slope β_1 and the intercept β_0 are to be estimated. Figure 9.7 illustrates the effects of an outlier in the y-direction. The outlier has a rather large influence on the least squares (LS) regression line.

Unlike LS, the L_1 regression line protects us against outlying y_k and is robust with respect to such an outlier. The residuals from an LS fit are not very useful

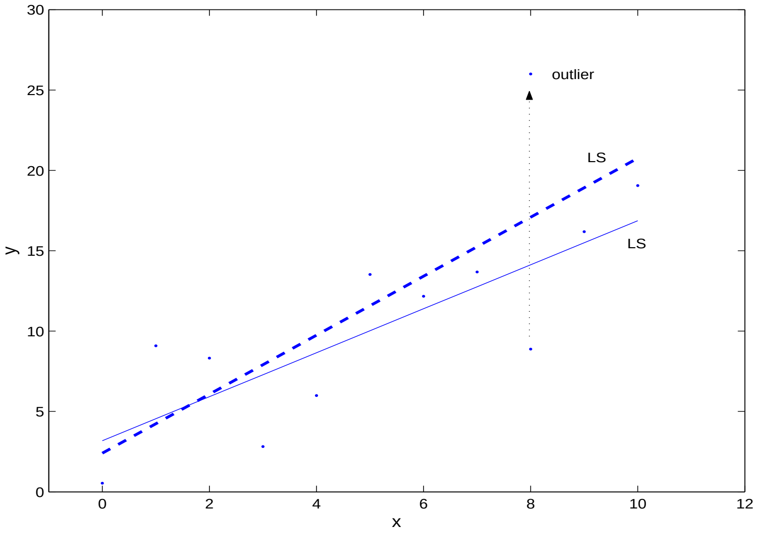

Figure 9.7: The original data and one outlier in the y-direction. The solid line corresponds to the LS fit without the outlier. The dashed line corresponds to the LS fit with the outlier.

as outlier diagnostics, on the other hand, the residuals computed from a robust estimator (e.g. L_1 regression, least median of squares, least trimmed squares) embody powerful information for detecting all the outliers present in the data. For LS, we have seen that one outlier can totally bias the LS estimator. On the other hand, the L_1 regression can handle the outlier. Figure 9.8 gives a schematic summary of the effect of one outlier on LAD regression, in the same situation as Figure 9.7.

In the following experiment we will check whether the estimators can deal with several outliers in the data set. Given 50 "good" observations $\{(x_1, y_1), ..., (x_{50}, y_{50})\}$ according to the linear relation

$$y_k = 5.0 + 1.5x_k + e_k, \quad k = 1, ..., 50, \tag{9.27}$$

where $e_k \sim N(0, 0.5^2)$ and $x_k \sim U[0, 5]$. To these data, we applied LS and LAD techniques. The estimators yield values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which are close to the original β_0, β_1. Then we started to contaminate the data. At each step we deleted one "good" point and replace it by a "bad" point generated (x_l, y_l°) according to a normal distribution $y_k^\circ \sim \mathcal{N}(2, 5^2)$. We repeated this until only 25 "good" points remained. A breakdown plot is shown in Figure 9.9 where the value of $\hat{\beta}_1$ is drawn as a function of the percentage outliers.
The LS was immediately affected by these outliers and breaks down, whereas the LAD holds on. An disadvantage of some robust methods (e.g. L_1 regression, least median of squares) is its lack of efficiency when the errors would really be

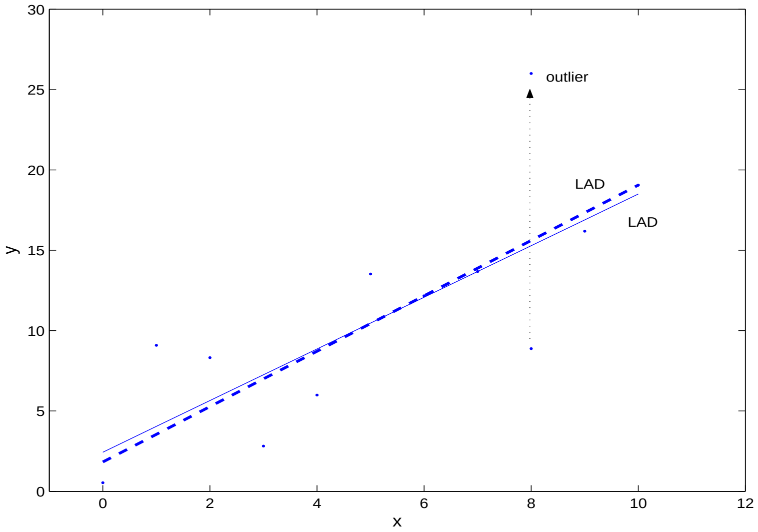

Figure 9.8: The original data and one outlier in the y-direction. The solid line corresponds to the LAD fit without the outlier. The dashed line corresponds to the LAD fit with the outlier.

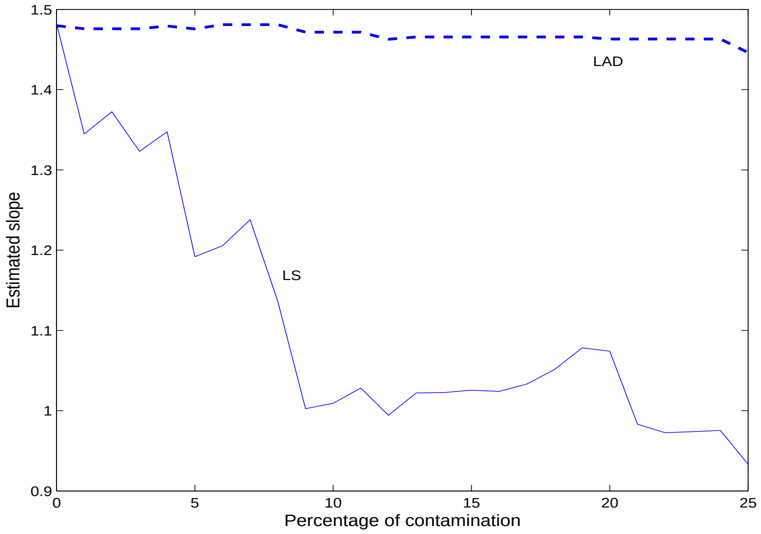

Figure 9.9: Breakdown plot, showing the estimated slope as a function of the percentage of contamination.

normally distributed. One solution to improve the efficiency of these robust methods is to use reweighted least squares. This leads us to the following algorithm:

Algorithm 38 *(Weighted Least squares).*

(i) Choose the initial L_1 estimator for β

$$\hat{\beta} = \arg\min_{\beta} \sum_{k=1}^{n} \left| y_k - \sum_{j=1}^{p} \beta_j x_k^{(j)} \right| . \tag{9.28}$$

(ii) Evaluate the error estimators

$$\hat{e}_k = y_k - \sum_{j=1}^{p} \hat{\beta}_j x_k^{(j)} \tag{9.29}$$

and calculate a robust variance estimator

$$\hat{\sigma}^2 = 1.483\ MAD \tag{9.30}$$

(iii) Find a weighting function, for example

$$v_k = \begin{cases} 1 & \text{if } \left|\frac{\hat{e}_k}{\hat{\sigma}}\right| \leq c \\ 0 & \text{otherwise.} \end{cases} \tag{9.31}$$

(iv) Once a weighting function is selected, one replaces all observations (x_k, y_k) by $\left(\sqrt{v_k} x_k, \sqrt{v_k} y_k\right)$.

(v) On these weighted observations, a standard LS may be used to obtain the final estimate.

9.3.2 Kernel based regression

Recall that (i) the Nadaraya-Watson kernel estimate of a regression function takes the form

$$\hat{m}_n(x) = \sum_{k=1}^{n} \frac{K\left(\frac{x-x_k}{h}\right) y_k}{\sum_{l=1}^{n} K\left(\frac{x-x_l}{h}\right)}, \tag{9.32}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ and $h > 0$, and (ii) the LS-SVM regression estimate is given by

$$\hat{m}_n(x) = \sum_{k=1}^{n} \hat{\alpha}_k K\left(\frac{x - x_k}{h}\right) + \hat{b}, \tag{9.33}$$

where $\hat{\alpha}_k \in \mathbb{R}$ and $b \in \mathbb{R}$. Figure 9.10 and 9.11 plot the effects of an outlier in the y-direction respectively for the Nadaraya-Watson kernel estimate and the LS-SVM regression estimate.

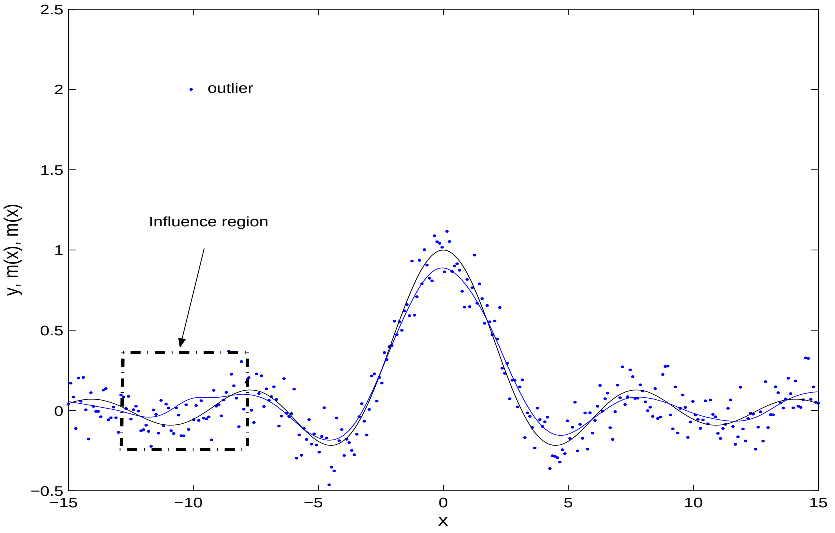

Figure 9.10: The effects of an outlier (y-direction). Estimation of the sinc function by Nadaraya-Watson kernel regression.

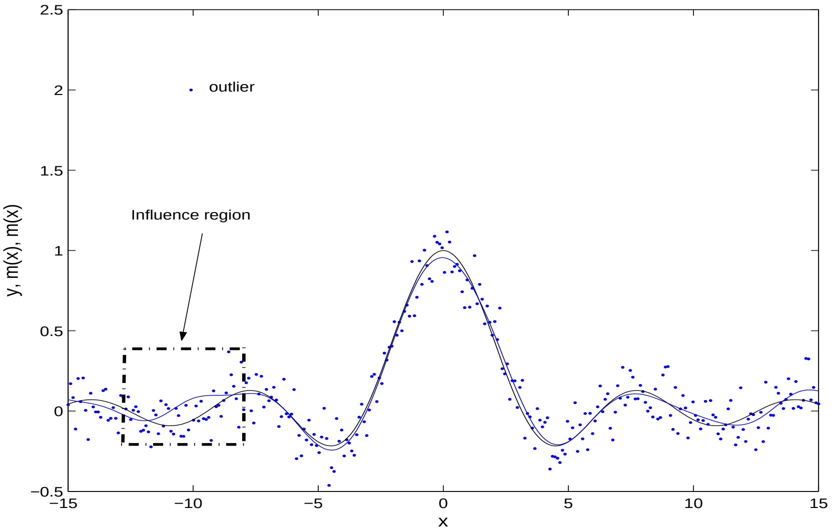

Figure 9.11: The effects of an outlier (y-direction). Estimation of the sinc function by LS-SVM regression.

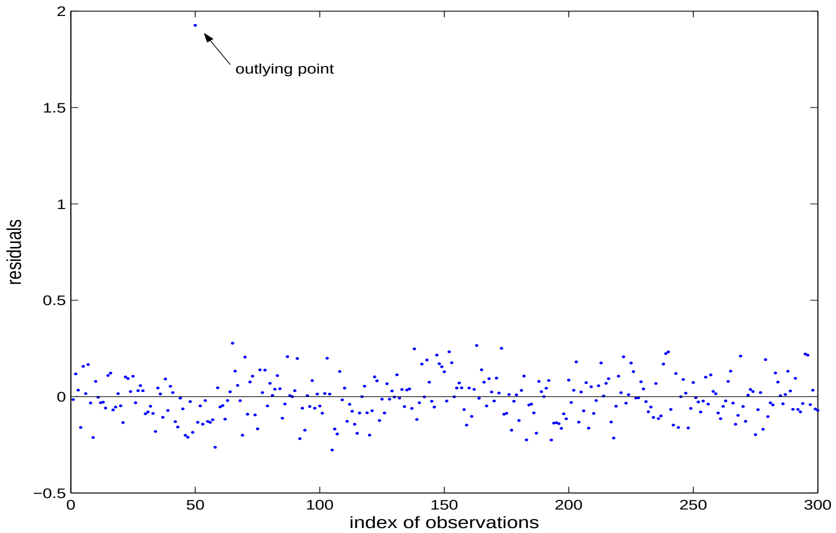

Figure 9.12: Index plot associated with LS-SVM regression. From this plot we can conclude that the data set contains one outlier.

Unlike in the linear parametric regression case, analysis of the robustness properties of kernel based estimators are in term of the estimated regression function. Let (x_i, y_i°) be an outlier (y-direction) and let $\mathcal{A}$ be the influence region. In both cases the outlier has a small influence on the estimate $\hat{m}_n(x_i)$ when $(x_i, \hat{m}_n(x_i)) \in \mathcal{A}$ and has no influence when $(x_j, \hat{m}_n(x_j)) \notin \mathcal{A}$. The residuals from both (Nadaraya-Watson kernel estimate, LS-SVM regression estimate) are very useful as outlier diagnostics. Figure 9.12 gives evidence of the presence of an outlying observation. The residual plot is given for the LS-SVM regression.

Using decreasing kernels, kernels such that $K(u) \to 0$ as $u \to \infty$, the influence for both $x \to \infty$ and $x \to -\infty$, is bounded in $\mathbb{R}$. Common choices for decreasing kernels are: $K(u) = \max\left(\left(1 - u^2\right), 0\right)$, $K(u) = \exp\left(-u^2\right)$ and $K(u) = \exp\left(-u\right)$.

We show the sensitivity curve (one with replacement) for $(x, \hat{m}_n(x)) \in \mathcal{A}$ and $(x_i, \hat{m}_n(x_i)) \notin \mathcal{A}$ in Figure 9.13. The most important aspect is that the sensitivity curve of the $\hat{m}_n(x)$ becomes unbounded $(x \in \mathcal{A})$ for both $y \to \infty$ and $y \to -\infty$, whereas the $\hat{m}_n(x_i)$ remains constant $(x_i \notin \mathcal{A})$.

In the following experiment we will check of the estimators can deal with several outliers in a particular region $\mathcal{A}$. Given 300 "good" observations $\{(x_1, y_1), ..., (x_{300}, y_{300})\}$ according to the relation

$$y_k = m(x_k) + e_k, \quad k = 1, ..., 300, \tag{9.34}$$

where $e_k \sim N(0, 0.1^2)$ and $x_k \sim U[-15, 15]$. To these data, we applied kernel regression techniques. For example the estimators (Nadaraya-Watson kernel

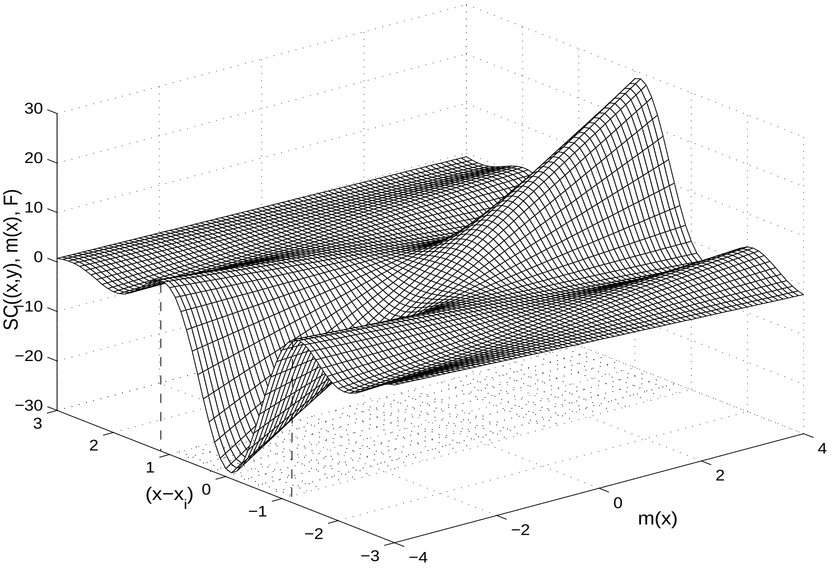

Figure 9.13: Empirical influence function of $\hat{m}_n(x)$ as a function of $(x - x_i)$. The influence curve (dotted region) is unbounded in $\mathbb{R}$, whereas in the other region the influence curve remains bounded in $\mathbb{R}$.

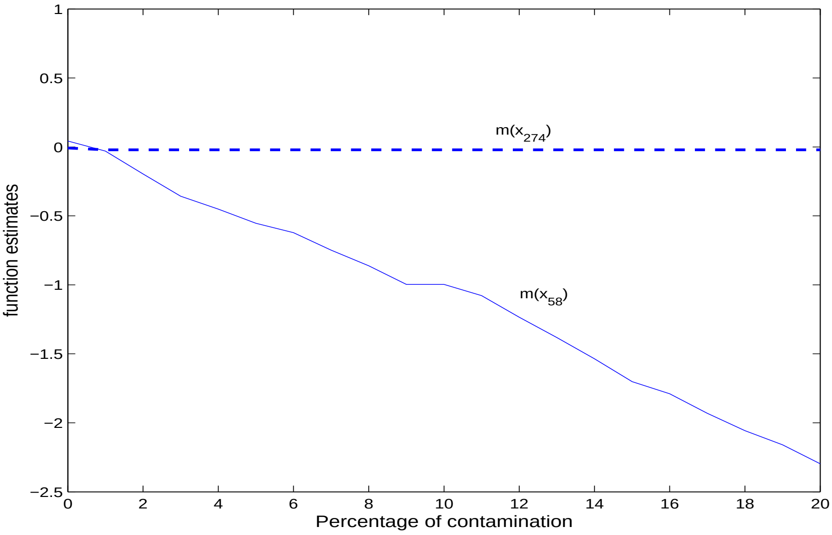

Figure 9.14: Breakdown plot, showing the estimated function values ($\hat{m}_n(x_{58})$ and $\hat{m}_n(x_{274})$) as a function of the percentage of contamination.

estimate and the LS-SVM regression estimate) yielded values of $\hat{m}_n(x_{58})$ and $\hat{m}_n(x_{274})$ which were close to the $m(x_{58})$ and $m(x_{58})$. Then we started to contaminate the data. At each step we deleted one "good" point in the region $\mathcal{A}\,(i)$ where $i = 48, ..., 68$ and replace it by a "bad" point (x_i, y_i°) . We repeated this until no "good" points remained in $\mathcal{A}$. A breakdown plot is shown in Figure 9.14 where the values of $\hat{m}_n(x_{58})$ and $\hat{m}_n(x_{274})$ are drawn as a function of the percentage outliers.

9.3.3 Robust LS-SVM

Under conditions where the dependent variable y is the result of measuring a regression function with normal additive noise e, the empirical risk minimization principle provides for the loss function $L\,(f(x), y) = (f(x) - y)^2$ an efficient estimator of the regression $m(x)$. Minimizing the empirical risk with respect to this loss function is known as the least squares (LS) method. The classical approach to the regression problem uses this method. The origin of the least squares method goes back to Legendre (Legendre, 1805). While the method of least squares enjoys well known properties within Gaussian parametric models, it is recognized that outliers which arise from heavy-tailed distributions have an unusually large influence on the resulting estimates.

If one haves only general information about the noise model, e.g., the density of the noise is a symmetric smooth function, then the best minimax strategy for regression estimation (Huber, 1964) provides the loss function $L\,(f(x), y) =$

$|f(x) - y|$. Minimizing the empirical risk with respect to this loss function is called the Least Absolute Deviations (LAD) method. The LAD estimation appears under a variety of names in the literature: the minimum absolute deviations method (MAD), minimum absolute errors (MAE), least absolute residuals (LAR), L_1-norm and least absolute values (LAV). Although it was known and studied as early as 1757 by Boscovich (Boscovich, 1757), due to mathematical inconveniences it never received full attention. However, it came into practical use with the development of mathematical programming. It is of interest to note that the LAD approach is equivalent to the method of maximum likelihood when the noise e has a double exponential distribution, and when the noise e has a uniform distribution with unknown range, the maximum likelihood criterion consists of minimizing the maximum $|e_k|$. Robust choices of the loss function can be motivated from both Huber's and Hampel's approaches for the location-scale models. In the regression setting (He, 1991) showed that obtaining local robustness requires using information beyond the residuals.

LS-SVM and robust cost functions

Based on Huber's robust theory (Huber, 1964), one can calculate a family of robust loss functions depending on how much information about the noise is available.

Theorem 39 *(Huber, 1964). Consider the class $\mathcal{A}$ of densities formed by mixtures $p(u) = (1-\varepsilon)\,g(u) + \varepsilon h(u)$ where $u = (f(x) - y)$. Let $-\log g(u) \in C^2[a,b]$ where a and b are endpoints. The class $\mathcal{A}$ possesses the following robust density*

$$p_{robust}(u) = \begin{cases} (1-\varepsilon)\,g(a)\exp(-c(a-u)) & \text{for } u < a \\ (1-\varepsilon)\,g(u) & \text{for } a \leq u \leq b \\ (1-\varepsilon)\,g(b)\exp(-c(u-b)) & \text{for } u \geq b. \end{cases} \tag{9.35}$$

The monotonic function

$$-\frac{d\log g(u)}{du} = -\frac{g^{'}(u)}{g(u)}$$

is bounded in absolute value by a constant c determined by the normalization condition

$$(1-\varepsilon)\left(\int_a^b g(u)du + \frac{g(a)+g(b)}{c}\right) = 1$$

Using (9.35), one can construct a robust regression estimator. The robust regression estimator is the one that minimizes the empirical risk functional

$$\mathcal{R}_{emp}(f) = -\sum_{k=1}^{n} \log p_{robust}(u_k).$$

As an example, given the normal density the class $\mathcal{A}$ of densities are defined as

$$p\left(u\right)=\frac{\left(1-\varepsilon\right)}{\sigma\sqrt{2\pi}}\exp\left(-\frac{u^{2}}{2\sigma^{2}}\right)+\varepsilon h(u).$$

According to the theorem the density

$$p_{robust}(u)=\left\{\begin{array}{ll}\frac{(1-\varepsilon)}{\sigma\sqrt{2\pi}}\exp\left(\frac{c^{2}}{2\sigma^{2}}-\frac{c}{\sigma}\left|u\right|\right) & \text{for }\left|u\right|>c\sigma\\ \frac{(1-\varepsilon)}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\xi^{2}}{2\sigma^{2}}\right) & \text{for }\left|u\right|\leq c\sigma\end{array}\right. \tag{9.36}$$

will be robust in the class, where c is determined from the normalization condition

$$\frac{\left(1-\varepsilon\right)}{\sigma\sqrt{2\pi}}\left(\int_{-c\sigma}^{c\sigma}\exp\left(-\frac{u^{2}}{2\sigma^{2}}\right)du+\frac{2\exp\exp\left(-\frac{c^{2}}{2}\right)}{c}\right)=1.$$

The loss function derived from this robust density is the Huber loss function (right panel of Figure 9.15)

$$L_{H}\left(u\right)=-\log p_{robust}(u)=\left\{\begin{array}{ll}c\left|u\right|-\frac{c^{2}}{2} & \text{for }\left|u\right|>c,\\ \frac{u^{2}}{2} & \text{for }\left|u\right|\leq c.\end{array}\right. \tag{9.37}$$

It is interesting to contrast this with the ϵ-insensitive error measure (support vector machines error measure), ignoring errors of size less than ϵ. The ϵ-insensitive loss function has the form

$$L_{\epsilon}\left(u\right)=\left\{\begin{array}{ll}0, & \text{if }\left|u\right|\leq\epsilon,\\ \left|u\right|-\epsilon, & \text{otherwise,}\end{array}\right. . \tag{9.38}$$

shown in the left panel of Figure 9.15.

The loss function $L_{\epsilon}\left(u\right)$ has the same structure as the loss function $L_{H}\left(u\right).$ The ϵ-insensitive error measure also has linear tails, but in addition it flattens the contributions of those cases with small residuals. ϵ is a parameter of the loss function $L_{\epsilon}\left(u\right)$, just like c is for $L_{H}\left(u\right).$

Consider the class of support vector machines. Let $\mathcal{F}$ denote a set of linear functions defined as

$$\mathcal{F}=\left\{f:f(x)=w^{T}\varphi\left(x\right)+b:\ w\in\mathbb{R}^{n_{f}},\ b\in\mathbb{R},\ \varphi:\mathbb{R}^{d}\to\mathbb{R}^{n_{f}}\right\}, \tag{9.39}$$

where $\mathbb{R}^{n_{f}}$ denote a high-dimensional feature space, $w\in\mathbb{R}^{n_{f}}$, $b\in\mathbb{R}$ are the parameters and $\varphi:\mathbb{R}^{d}\to\mathbb{R}^{n_{f}}$ is the feature map. With $f\in\mathcal{F}_{n}$, one can define the optimization problem in the primal space as

$$\left(\hat{w},\hat{b}\right)=\arg\min_{w\in\mathbb{R}^{\mathcal{M}},b\in\mathbb{R}}\left[\gamma\sum_{k=1}^{n}L\left(\left(w^{T}\varphi\left(x_{k}\right)+b\right)-y_{k}\right)+\frac{1}{2}\left\Vert w\right\Vert_{2}^{2}\right]. \tag{9.40}$$

where $L\left(.\right)\in C^{v}\left(\mathbb{R}\right),\ v\geq1$. Next we examine of the family of robust cost functions that fit within the LS-SVM formulation.

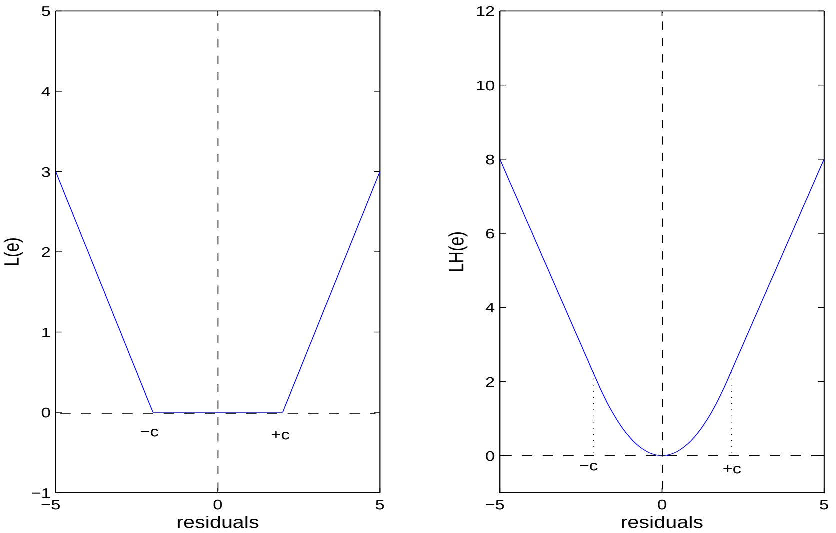

Figure 9.15: The left figure shows the ϵ-insensitive loss function used by the SVM regression. The right figure shows the loss function used in Huber's robust regression.

(i) . The LAD estimator is recommended when the shape of the noise distribution is unknown and one only assumes that it belongs to a broad family of distributions. In this case $L\left(e\right) = |e| \in C^0\left(\mathbb{R}\right).$ To overcome this difficulty, one can take an approximation to $|e|$ like $\sqrt{e^2 + a_n} \in C^\infty\left(\mathbb{R}\right)$, where a_n is a small positive parameter. Remark that the sequence of functions $f_n\left(e\right) = \sqrt{e^2 + \frac{1}{n}}$, converge uniformly to $|e|$, a nondifferentiable function. If one puts $L\left(e\right) = \sqrt{e^2 + a_n}$ in (9.40), one obtains a Quadratic Programming (QP) problem. Figure 9.16 plots the smooth approximation to the absolute value function. The solid curve represents the absolute value function; the dashed curve represents the approximation $\sqrt{e^2 + a_n}$ with parameter $a_n = 0.01$.

(ii) . Huber M-estimates use a function $L\left(e\right)$ that is a compromise between e^2 and $|e|$. Huber has combined the advantages of both methods by defining $L\left(e\right)$ to be equal to e^2 when e is near 0 and equal to $|e|$ when e is far from 0. The Huber loss function is defined in (9.37), where $c = 1.5\hat{s}$ (Huber, 1981). To estimate σ we use $\hat{s} = 1.483$ MAD, where MAD is the median of the absolute deviations $|\hat{e}_k|$. The multiplier 1.483 is chosen to ensure that $\hat{s}$ would be a good estimate of σ if it were the case that the distribution of the random errors were normal (Birkes and Dodge, 1993). The Huber loss function is convex and $L_H\left(e\right) \in C^1\left(\mathbb{R}\right)$, but has discontinuous second derivatives at points where $|\hat{e}_k| = c$. The mathematical structure of the Huber M-estimator seems first to have been considered in detail by Clark (Clark, 1985). If one puts $L_H\left(e\right)$ in (9.40), one obtains a Quadratic Programming (QP) problem.

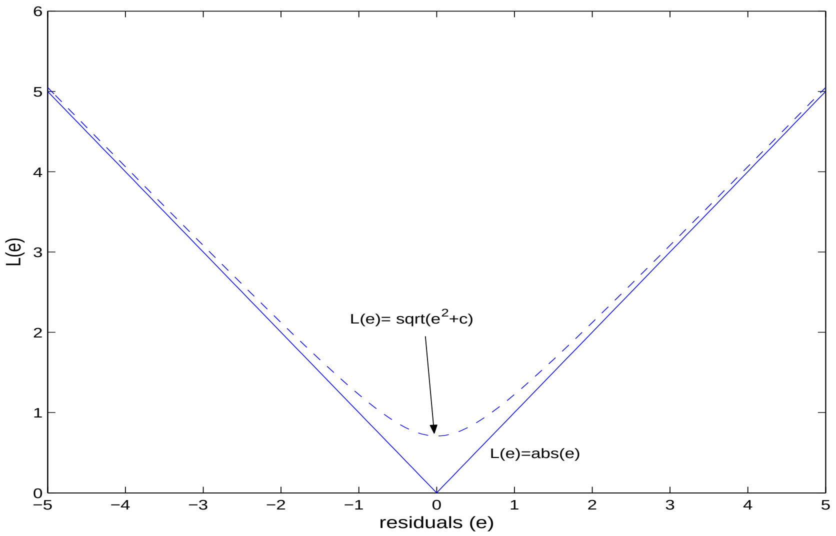

Figure 9.16: Smooth approximation to the absolute value function. The solid curve represents the absolute function $L(e) = |e|$; the dased curve represents the approximation $L(e) = \sqrt{e^2 + a_n^2}$ with the parameter $a_n = 0.1$.

(iii). Only for the case $L(e) = a_2 e^2 + a_1 e + a_0 \in C^{\infty}(\mathbb{R})$, $a_2, a_1, a_0 \in \mathbb{R}$ one has a linear system.

Iteratively LS-SVM

Huber M-estimates We can apply the LS-SVM iteratively to obtain Huber M-estimates. Given training data $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ and an optimal (h, γ) combination. The computational procedure (iteratively LS-SVM) is described below.

Algorithm 40 *(Iteratively LS-SVM).*

(1) Initialize: Let $y_k^{[0]} = y_k, k = 1, ..., n$

(2) Begin

(i) Apply the LS-SVM to the training data $\mathcal{D}_n = \left\{\left(x_1, y_1^{[0]}\right), ..., \left(x_n, y_n^{[0]}\right)\right\}$ *to obtain:*

(i.1) The support vectors $\hat{\alpha}_k^{[0]}, k = 1, ..., n$ *and the bias term* $\hat{b}^{[0]}$ *(by solving linear systems (3.8))*

(i.2) The vector $\hat{m}^{[0]} = \left(\hat{m}_n^{[0]}(x_1), ..., \hat{m}_n^{[0]}(x_n)\right)$ *of the estimated regression values are given by*

$$\hat{m}_n^{[0]}(x_i) = \sum_{k=1}^{n} \hat{\alpha}_k^{[0]} K(x_i, x_k; h) + \hat{b}^{[0]}, k = 1, ..., n; i = 1, ..., n$$

(ii) Let $\hat{e}^{[0]} = \left(\hat{e}_1^{[0]}, ..., \hat{e}_n^{[0]}\right)^T$ *where* $\hat{e}_k^{[0]} = \frac{\hat{\alpha}_k^{[0]}}{\gamma}$ *and:*

(ii.1) Calculate $\hat{s}^{[0]}$

$$\hat{s}^{[0]} = 1.483 \ MAD\left(\left|\hat{e}_1^{[0]}\right|, ..., \left|\hat{e}_n^{[0]}\right|\right)$$

(ii.2) Truncate the residuals by defining

$$\Delta_k^{[0]} = \max\left(-1.5\hat{s}^{[0]}, \min\left(\hat{e}_k^{[0]}, 1.5\hat{s}^{[0]}\right)\right), k = 1, ..., n.$$

(iii) Let $y_k^{[0]} = \hat{m}_n^{[0]}(x_k) + \Delta_k^{[0]}; \ \forall k, \ k = 1, ..., n.$

(3) Repeat the calculations with $y_k^{[1]}, k = 1, ..., n$ *in place of* $y_k^{[0]} = y_k, k = 1, ..., n$. *Until consecutive estimates* $\hat{\alpha}_k^{[l-1]}$ *and* $\hat{\alpha}_k^{[l]}$ *are sufficiently close to one other* $\forall k, \ k = 1, ..., n$.

Weighted LS-SVM. In order to obtain a robust estimate based upon the previous LS-SVM solution, in a subsequent step, one can weight the error variables $e_k = \alpha_k/\gamma$ by weighting factors v_k (Suykens *et al.*, 2002). This leads to the optimization problem:

$$\min_{w^\circ, b^\circ, e^\circ} \mathcal{J}(w^\circ, e^\circ) = \frac{1}{2} w^{\circ T} w^\circ + \frac{1}{2}\gamma \sum_{k=1}^{n} v_k e_k^{\circ 2} \tag{9.41}$$

such that $y_k = w^{\circ T}\varphi(x_k) + b^\circ + e_k^\circ, \quad k = 1, ..., n$. The Lagrangian is constructed in a similar way as before. The unknown variables for this weighted LS-SVM problem are denoted by the $\circ$ symbol. From the conditions for optimality and elimination of w°, e° one obtains the Karush-Kuhn-Tucker system:

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + D_\gamma \end{array}\right] \left[\begin{array}{c} b^\circ \\ \hline \alpha^\circ \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \tag{9.42}$$

where the diagonal matrix D_γ is given by $D_\gamma = \text{diag}\left\{\frac{1}{\gamma v_1}, ..., \frac{1}{\gamma v_n}\right\}$. The choice of the weights v_k is determined based upon the error variables $e_k = \alpha_k/\gamma$ from the (unweighted) LS-SVM case. Robust estimates are obtained then (Rousseeuw and Leroy, 1986) e.g. by taking

$$v_k = \begin{cases} 1 & \text{if } \ |e_k/\hat{s}| \leq c_1 \\ \frac{c_2 - \ |e_k/\hat{s}|}{c_2 - c_1} & \text{if } \ c_1 \leq |e_k/\hat{s}| \leq c_2 \\ 10^{-4} & \text{otherwise} \end{cases} \tag{9.43}$$

where $\hat{s} = 1.483$ MAD (e_k) is a robust estimate of the standard deviation of the LS-SVM error variables e_k and MAD stands for the median absolute deviation. One assumes that e_k has a symmetric distribution which is usually the case when (h, γ) are well-determined by an appropriate model selection method. The constants c_1, c_2 are typically chosen as $c_1 = 2.5$ and $c_2 = 3$ (Rousseeuw and Leroy, 1987). Using these weightings one can correct for outliers (y-direction). This leads us to the following algorithm:

Algorithm 41 *(Weighted LS-SVM).*

(i) Given training data $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$, find an optimal (h, γ) combination (e.g. by cross-validation, FPE criterion) by solving linear systems (Chapter 3, (3.8)). For the optimal (h, γ) combination one computes $e_k = \alpha_k / \gamma$ from (Chapter 3, (3.8)).

(ii) Compute $\hat{s} = 1.483$ MAD(e_k) from the e_k distribution.

(iii) Determine the weights v_k based upon $e_k, \hat{s}$ and the constants c_1, c_2.

(iv) Solve the weighted LS-SVM (9.42), giving the model $\hat{m}_n^{\circ}(x) = \sum_{k=1}^{n} \alpha_k^{\circ} K(x, x_k) + b^{\circ}$.

First, we graph the sensitivity curve (one with replacement) for $(x, \hat{m}_n^{\circ}(x)) \in \mathcal{A}$ and $(x_i, \hat{m}_n^{\circ}(x_i)) \notin \mathcal{A}$ in Figure 9.17. The most important aspect is that the sensitivity curve of the $\hat{m}_n^{\circ}(x)$ becomes bounded $(x \in \mathcal{A})$ for both $y \to \infty$ and $y \to -\infty$, whereas the $\hat{m}_n^{\circ}(x_i)$ remains constant $(x_i \notin \mathcal{A})$.

Second, we compute the maxbias curve for both LS-SVM and weighted LS-SVM on a test point. Given 150 "good" observations $\{(x_1, y_1), ..., (x_{150}, y_{150})\}$ according to the relation

$$y_k = m(x_k) + e_k, \quad k = 1, ..., 150, \tag{9.44}$$

where $e_k \sim \mathcal{N}(0, 1^2)$. Let $\mathcal{A}$ be a particular region (43 data points) and let x be a test point from that region (Figure 9.18). Then we started to contaminate the data in region $\mathcal{A}$. At each step we deleted one "good" point in the region $\mathcal{A}$ and replace it by a "bad" point (x_i, y_i°). We repeated this until the estimation becomes useless. A maxbias plot is shown in Figure 9.19 where the values of $\hat{m}_n(x)$ and $\hat{m}_n^{\circ}(x)$ are draw as a function of the number of outliers in region $\mathcal{A}$. The maxbias of $\hat{m}_n^{\circ}(x)$ increases only very slightly with the number of outliers in region $\mathcal{A}$ and stays bounded right up to the breakdown point. This doesn't hold for $\hat{m}_n(x)$ with 0% breakdown point.

Finally, illustrative examples are given. In this example we illustrate the method of weighted LS-SVM. First, we show two examples of estimating a sinc function from noisy data: (a) strong outliers are superimposed on zero mean Gaussian noise distribution (Figure 9.20-9.21) (b) non-Gaussian noise distribution (central t-distribution with 4 degrees of freedom, i.e. heavy tails (Johnson and Kotz, 1970)) (Figure 9.22).

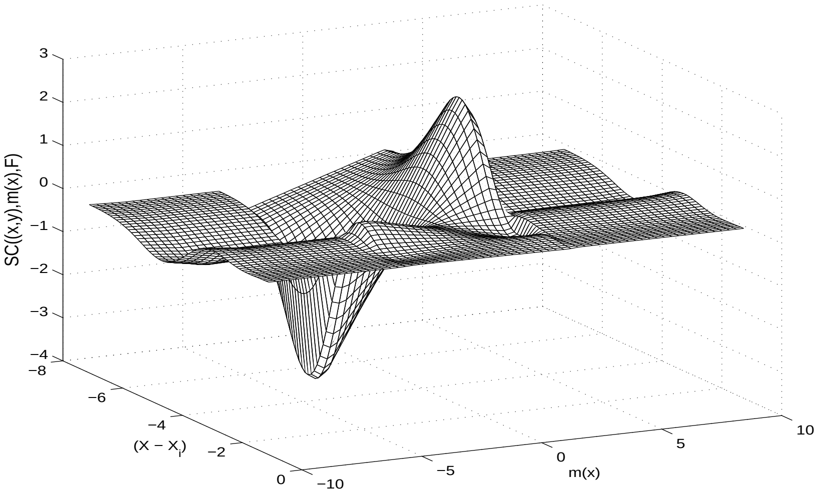

Figure 9.17: Empirical influence function of $\hat{m}_n(x)$ as a function of $(x - x_i)$. The data set used for this experiment is y_k =sinc$(x_k) + e_k$ where the errors $e \sim \mathcal{N}\left(0, 0.2^2\right)$ for all $k = 1, ..., 150$. The influence curve is bounded in $\mathbb{R}$.

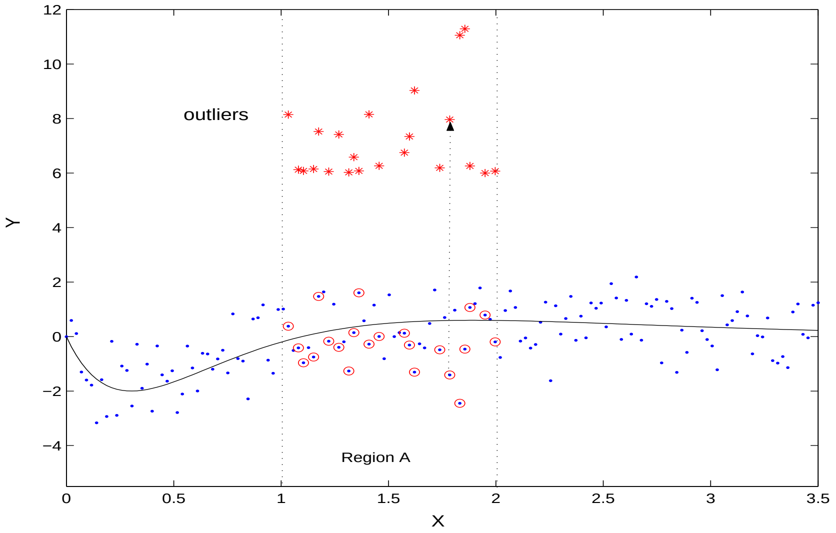

Figure 9.18: Given 150 training data (Wahba, 1990) corrupted with $e \sim \mathcal{N}\left(0, 1^2\right)$. Consider the region $\mathcal{A}$ between $x = 1$ and $x = 2$. In each step the data in the region $\mathcal{A}$ is contaminated by replacing a good point (given as "∘") by a bad point (denoted as "*") until the estimation becomes useless.

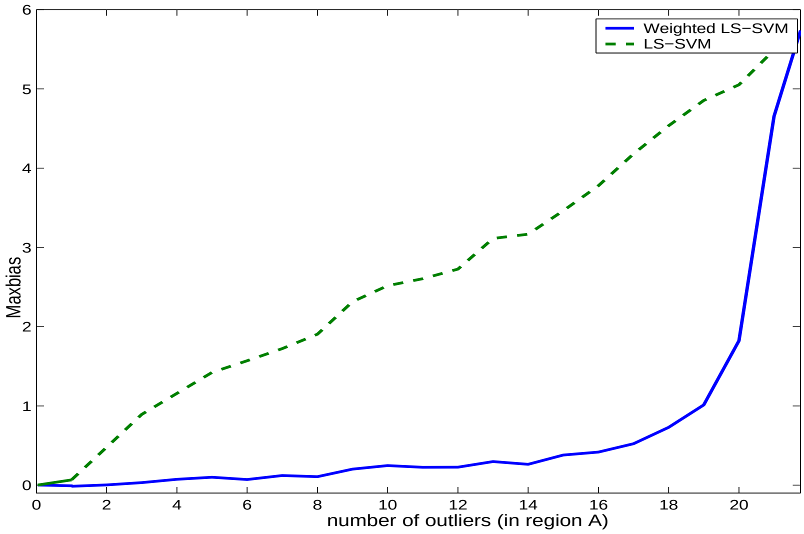

Figure 9.19: Maxbias curves of the LS-SVM regression estimator $\hat{m}_n\,(x)$ and the weighted LS-SVM regression estimator $\hat{m}_n^{\circ}\,(x)\,.$

(b) non-Gaussian noise distribution (central t-distribution with 4 degrees of freedom, i.e. heavy tails (Johnson and Kotz, 1970)) (Figure 9.22).

Given is a training set of $N = 300$ data points. From the simulation results it is clear that the unweighted LS-SVM is quite robust and does not breakdown (Figure 9.20). The generalization performance is further improved by applying weighted LS-SVM (Algorithm: *Weighted LS-SVM*), shown in Figure 9.21 and Figure 9.22, respectively. The good generalization performance on fresh test data is shown for all cases.

An additional comparison with a standard SVM with Vapnik ϵ-insensitive loss function is made. The Matlab SVM Toolbox by Steve Gunn was used to generate the SVM results. Here $\epsilon = 0$ was taken and as upper bound on support values $C = Inf$. An optimal σ value was selected. Other ϵ, C combinations resulted in worse results. These comparative results are shown in (Figure 9.23). In these examples the weighted LS-SVM results show the best results. The unweighted LS-SVM is also quite robust. Due to the choice of a 2-norm this may sound surprising. However, one should be aware that only the output weights (support values α) follow from the solution to the linear system while (γ, σ) are to be determined at another level.

Figure 9.24 shows comparative results on the motorcycle data, a well-known benchmark data set in statistics (Eubank, 1999). The x values are time measurements in milliseconds after simulated impact and the y values are measurements of head acceleration. The x values are not equidistant and in some cases multiple y observations are present for certain x values. The data are heteroscedastic.

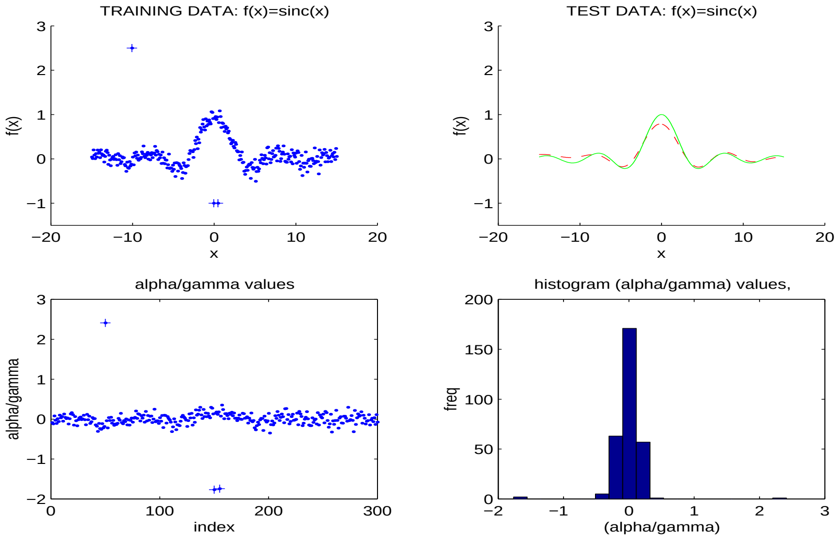

Figure 9.20: Estimation of a sinc function by LS-SVM with RBF kernel, given 300 training data points, corrupted by zero mean Gaussian noise and 3 outliers(denoted by +). (Top-left) Training data set; (Top-right) resulting LS-SVM model evaluated on an independent test set: (solid line) true function, (dashed line) LS-SVM estimate; (Bottom-left) e_k values; (Bottom-right) histogram of e_k values.

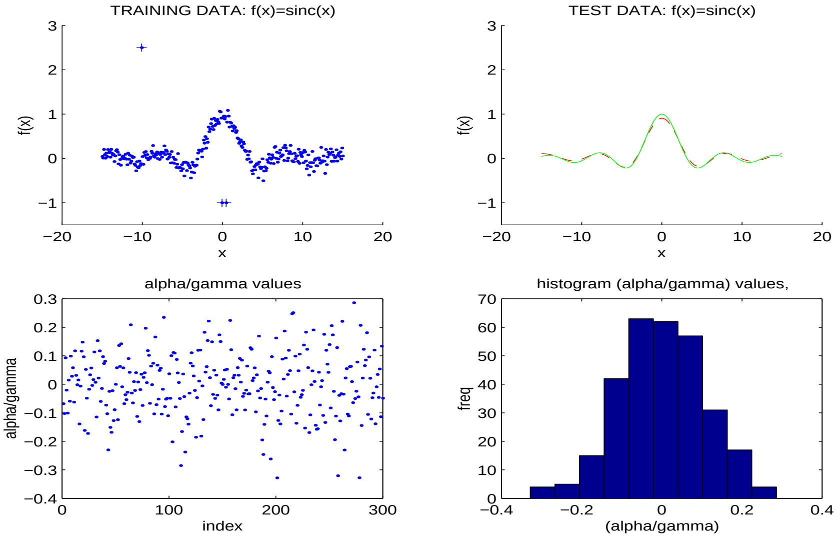

Figure 9.21: Weighted LS-SVM applied to the results of Figure 9.21. The e_k distribution becomes Gaussian and the generalization performance on the test data improves.

In this sense it forms a challenging test case. Figure 9.24 show the results from unweighted and weighted LS-SVM in comparison with standard SVM. In this example standard SVM suffers more from boundary effects. The tuning parameters are: $\gamma = 2$, $\sigma = 6.6$ (LS-SVM) and $\sigma = 11$, $\epsilon = 0$, $C = Inf$ (Vapnik SVM).

Figure 9.25 shows the improvements of weighted LS-SVM on the Boston housing data in comparison with unweighted LS-SVM. The weighted LS-SVM achieves an improved test set performance after determination of (γ, σ) by 10-fold CV on a randomly selected training set of 406 points. The remaining test set consisted of 100 points. The data were normalized except the binary variables. Optimal values of (γ, σ) were determined by 10-fold CV on the training set. The weighted LS-SVM resulted in a test set MSE error of 0.1638, which was an improvement over the unweighted LS-SVM test set MSE error of 0.1880. The improved performance is achieved by suppressing the outliers in the histogram shown in Figure 9.25.

Robust fixed size LS-SVM regression

The training data $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ are assumed to be zero mean. We known from chapter 3.4 that one can constructs a new training set defined as $\mathcal{D}_n^{(feature)} = \{(\hat{\varphi}(x_k), y_k) : \hat{\varphi}(x_k) \in \mathbb{R}^{n_{FS}}, y_k \in \mathcal{Y};\ k = 1, ..., n,\ n_{FS} \leq n\}$ based on the Nyström approximation. The following fixed size LS-SVM model

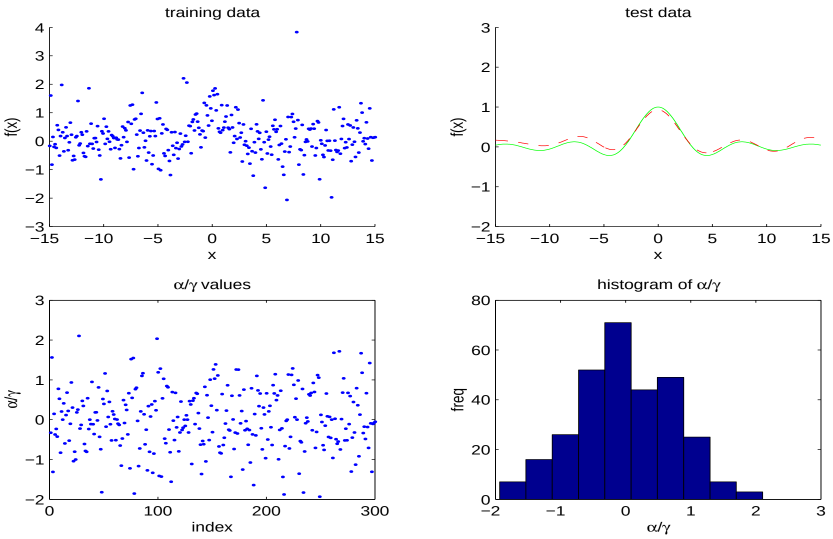

Figure 9.22: Estimation of a sinc function by LS-SVM with RBF kernel, given 300 data points, corrupted by a central t-distribution widt degrees of freedom equal to 4. (Top-Left) Training data set; (Top-Right) resulting weighted LS-SVM model evaluated on an independent test set: (solid line) true function, (dashed line) weighted LS-SVM estimate; (Bottom-Left) residuals obtained with weighted LS-SVM; (Bottom-Right) histogram of the residuals.

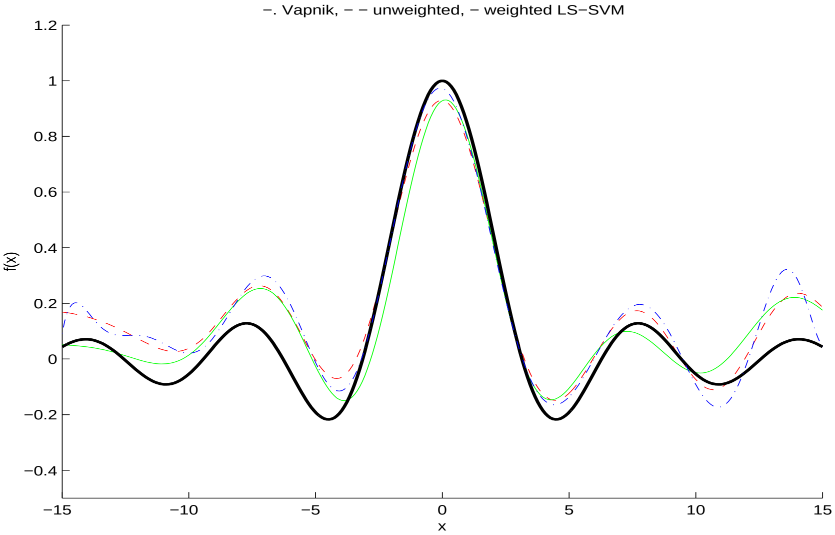

Figure 9.23: Comparison between standard SVM, unweighted LS-SVM and weighted LS-SVM. Weighted LS-SVM gives the best estimate for this example.

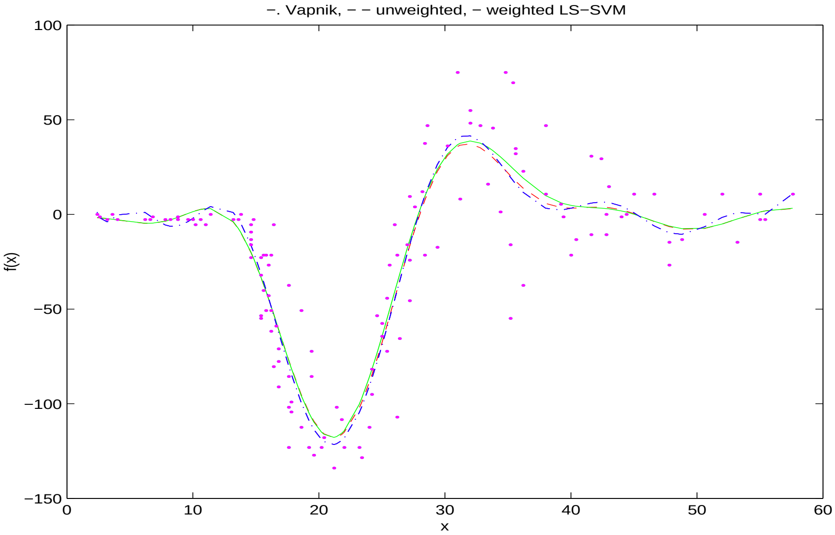

Figure 9.24: Motorcycle data set: comparison between standard SVM, unweighted LS-SVM and wighted LS-SVM.

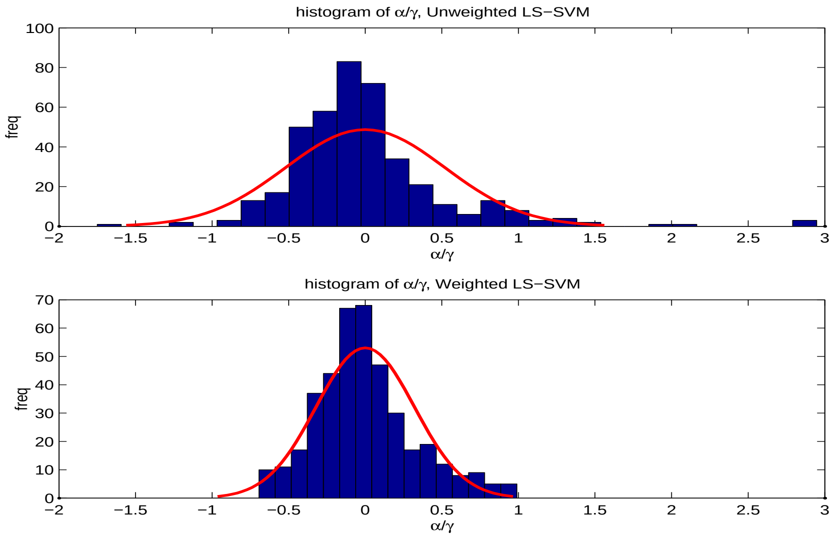

Figure 9.25: Boston housing data set: (Top) histogram of e_k for unweighted LS-SVM with RBF kernel. The outliers are clearly visible; (Bottom) histogram resulting from weighted LS-SVM with improved test set performance.

is assumed

$$y = Aw + e, \tag{9.45}$$

where $y = (y_1, ..., y_n)^T$, $e = (e_1, ..., e_n)^T$ and the $n \times n_{FS}$ matrix A is defined as

$$A = \begin{pmatrix} a_1^T \\ \cdot \\ \cdot \\ \cdot \\ a_n^T \end{pmatrix} \begin{pmatrix} \hat{\varphi}_1(x_1), & ..., & \hat{\varphi}_l(x_1) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \hat{\varphi}_1(x_n), & ..., & \hat{\varphi}_l(x_n) \end{pmatrix}. \tag{9.46}$$

Assume that the rows of A are i.i.d. observations from a n_{FS}-dimensional distribution F_a. Let F be the joint distribution for the $(n_{FS}+1)$-dimensional distribution, a^T is a random row of A and y is the associated dependent variable. Let $\hat{F}_n$ be the empirical distribution which put mass $\frac{1}{n}$ on each of the n rows $\left(a_k^T, y_k\right)$, $k = 1, ..., n$, of the matrix $[A\,|y]$. Define the real-valued functional $B(F) = E[ya]$. It maps the class of all distributions on $\mathbb{R}^{n_{FS}+1}$ onto $\mathbb{R}^{n_{FS}}$ (for which the expectation $E[ya]$ exists), $y \in \mathbb{R}$ and $a \in \mathbb{R}^{n_{FS}}$. Evaluated at $\hat{F}_n$

$$B\left(\hat{F}_n\right) = \frac{1}{n}\sum_{k=1}^{n} y_k a_k = \frac{1}{n}A^T Y. \tag{9.47}$$

Define $Q(F) = E\left[aa^T\right]$, evaluated at $\hat{F}_n$

$$Q\left(\hat{F}_n\right) = \frac{1}{n}\sum_{k=1}^{n} a_k a_k^T = \frac{1}{n}A^T A. \tag{9.48}$$

The ridge regression functional $w_{ridge} = T(F) = \left(Q(F) + \frac{1}{\gamma}I\right)^{-1} B(F)$ yields

$$\hat{w}_{ridge} = T(\hat{F}_n) = \left(A^T A + \frac{1}{\gamma}I_n\right)^{-1} A^T Y. \tag{9.49}$$

Lemma 42 *The influence function of $T(F)$ at $\left(a^T, \tilde{y}\right)$ with $\tilde{y} \in \mathbb{R}$ is*

$$IF\left(\left(a^T, \tilde{y}\right); T, F\right) = \left(Q(F) + \frac{1}{\gamma}I\right)^{-1} \left(B(H) - Q(F)T(F)\right) \tag{9.50}$$

Proof: By definition, the influence function gives for each $\left(a_k^T, y_k\right)$ (where $a_k^T \in \mathbb{R}^{n_{FS}}, y_k \in \mathbb{R}, k = 1, ..., n$) the directional derivative (Gâteau derivative) of T at F in the direction of $H = \Delta - F$

$$\begin{aligned}
IF\left(\left(a^T, \tilde{y}\right); T, F\right) &= \lim_{\epsilon \downarrow 0} \frac{T\left[(1-\epsilon)F + \epsilon\Delta_{[a^T\tilde{y}]}\right] - T(F)}{\epsilon} \\
&= \frac{d}{d\epsilon}\left[T(F + \epsilon H)\right]_{\epsilon=0} \\
&= \left[\frac{d}{d\epsilon}\left(Q(F) + \epsilon Q(H) + \frac{1}{\gamma}I\right)^{-1}(B(F) + \epsilon B(H))\right]_{\epsilon=0} \\
&+ \left[\left(Q(F) + \epsilon Q(H) + \frac{1}{\gamma}I\right)^{-1} B(H)\right]_{\epsilon=0} \\
&= \left(Q(F) + \frac{1}{\gamma}I\right)^{-1} B(H) \\
&- \left(\left(Q(F) + \frac{1}{\gamma}I\right)^{-1} Q(H)\left(Q(F) + \frac{1}{\gamma}I\right)^{-1}\right) B(F) \\
&= \left(Q(F) + \frac{1}{\gamma}I\right)^{-1} \left((B(H) - Q(F)T(F))\right) \\
&= \left(Q(F) + \frac{1}{\gamma}I\right)^{-1} \left(a\left(\tilde{y} - a^T T(F)\right)\right) \qquad (9.51)
\end{aligned}$$

□

This influence factors into an influence of position $\left(Q(F) + \frac{1}{\gamma}I\right)^{-1} a$ and the influence of the residual $\left(\tilde{y} - a^T T(F)\right)$. The influence function is unbounded in $\mathbb{R}$, the observations with large residuals have a much larger contribution to the influence function.

Consider the fixed-size LS-SVM regression where $\gamma \to \infty$. Let $T\left(\hat{F}_n\right)$ be denote a robust fixed-size LS-SVM estimator based on a M-estimator.

Lemma 43 *The influence function of the robust $T(F)$ at $\left(a^T, \tilde{y}\right) \in \mathbb{R}^{n_{FS}+1}$ is*

$$IF\left(\left(a^T, \tilde{y}\right); T, F\right) = \frac{\rho\left(a, \tilde{y} - a^T T(F)\right)}{\int D\left(a, \tilde{y} - a^T T(F)\right) a a^T dF(a, \tilde{y})}. \tag{9.52}$$

The functional $T(F)$ corresponding to a robust fixed-size LS-SVM is the solution of

$$\int \rho\left(a, \tilde{y} - a^T T(F)\right) a dF(a, \tilde{y}) = 0. \tag{9.53}$$

Then the influence function of T at a distribution F is given by

$$\begin{aligned} IF\left(\left(a^T, \tilde{y}\right); T, F\right) &= \lim_{\epsilon \downarrow 0} \frac{T\left[(1-\epsilon) F + \epsilon \Delta_{[a^T \tilde{y}]}\right] - T(F)}{\epsilon} \\ &= \frac{d}{d\epsilon}\left[T\left(F + \epsilon H\right)\right]_{\epsilon=0} \\ &= \frac{\rho\left(a, \tilde{y} - a^T T(F)\right)}{\int D\left(a, \tilde{y} - a^T T(F)\right) a a^T dF(a, \tilde{y})} \end{aligned} \tag{9.54}$$

where $D(t) = \frac{d\rho(t)}{dt}$. This influence function is bounded in $\mathbb{R}$. For a complete proof of the M-estimator (linear regression context), see (Hampel et al. 1986).

Example 1

In this example we illustrate the method of fixed size LS-SVM. Given the training data ($n = 500$) where 10 outliers are superimposed on a $\mathcal{N}\left(0, \sigma^2\right)$ noise distribution. From the simulation results (Figure 9.26) it is clear that the generalization performance for both robust methods (the L_1-estimate and the Huber-estimate) is improved with respect to the L_2-estimate. The L_1-estimate and the Huber-estimate resulted in a test set MSE of respectively 0.0044 and 0.0042, which was an improvement over the L_2-estimate test set MSE of 0.1347.

9.4 Conclusions

Unlike in the linear parametric regression case, analysis of the robustness properties of kernel based estimators are in term of the estimated regression function. The residuals from LS-SVM regression estimate is very useful as outlier diagnostics. While standard SVM's approaches starts from choosing a given convex cost function and obtain a robust estimate in a top-down fashion, this procedure has the disadvantage that one should know in fact beforhand which cost function is statistically optimal. We have successfully demonstrated and alternative bottom-up procedure which starts from an unweighted LS-SVM and then robustifies the solution bij defining weightings based upon the error distribution.

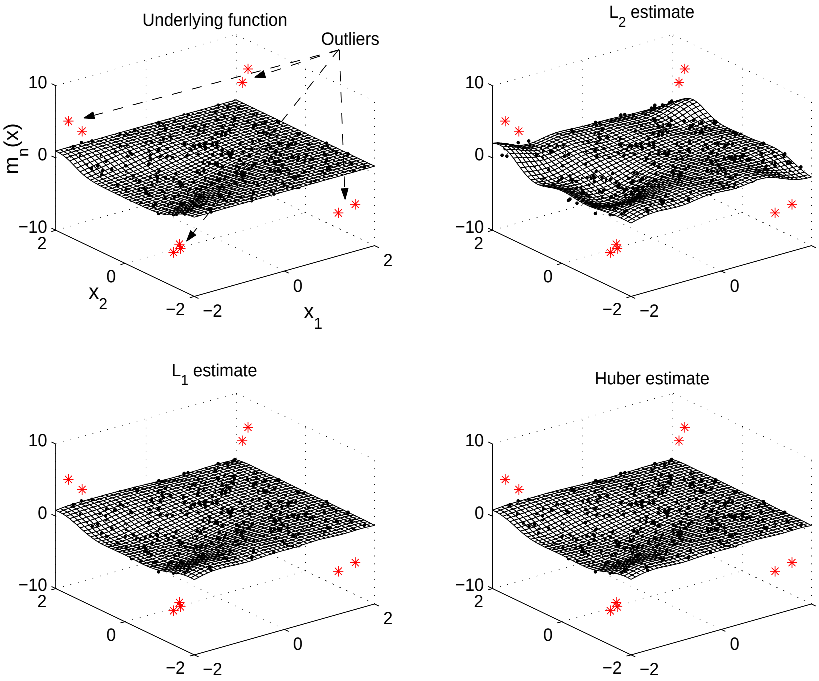

Figure 9.26: Estimation of a 2-dimensional function by fixed-size LS-SVM with RBF kernel, given 500 training data points, corrupted by zero mean, variance=0.1 Gaussian noise and 10 outliers (denoted by '*')(Top-left) Trainind data set;(Top-Right) resulting Fixed size LS-SVM evaluated on an independent test set; (Bottom-Left) L_1 estimate and (Bottom-Right) Huber estimate.

Chapter 10

Data-driven Loss Functions for Regression

In this chapter we study the Nadaraya-Watson estimator and show its nonrobustness in the sence of the influence function. We show that the L- robustifed Nadaraya-Watson kernel estimator has a boundend influence function. In a maximum likelihood sense, we calculate a loss function which is optimal for a given noise model. Contributions are made in Section 10.2, 10.3, and 10.4.

10.1 Introduction

Currently, there exists a variety of loss functions (e.g., least squares, least absolute deviations, M-estimators, generalized M-estimators, L-estimators, R-estimators, S-estimators, least trimmed sum of absolute deviations, least median of squares, least trimmed squares). On the other hand, this progress has put applied scientist into a difficult situation: if they need to fit their data with a regression function, they have trouble deciding which procedure to use. If more information was available, the estimation procedure could be chosen accordingly.

An idea for such a situation is to combine two convenient methods. Arthanari and Dodge (1981) introduced an estimation method in the linear model based on a direct convex combination of LAD and LS estimators with a fixed weight. Adaptive combination of LS and LAD estimators was first introduced by Dodge and Jurečkova (1987). Dodge (1984) introduced a convex combination of M- and LAD estimates; this convex combination was considered as adaptive by Dodge and Jurečkova (1988), who constructed an adaptive combination of LAD with Huber's M-estimate. The adaptive combination of LAD and the trimmed LS estimators was first studied by Dodge and Jurečkova (1992).

Another idea, proposed in this Section, is as follows. Given the data the method can basically be split up into two main parts: (i) constructing a robust nonparametric regression model and computing the residuals, and (ii) finding the distribution of the errors via a robust bootstrap and computing the loss function. Given the data, a robust nonparametric regression method will be used to obtain the residuals. Based on these residuals we can compute, in a maximum likelihood sense, the loss function. The evaluation of the accuracy of that loss function will be based on bootstrap accuracy intervals. We exhaustively describe the different parts in the next subsections.

10.2 Robust nonparametric regression models

The theory of robustness considers deviations from the various assumptions of parametric models. Robust statistics is thus an extension of classical parametric statistics. It studies the behavior of statistical procedures, not only under strict parametric models, but also both in smaller and in larger neighborhoods of such parametric models. Nonparametric statistics allows "all" possible probability distributions. When robustness to outliers is concerned, studying the behavior of nonparametric estimators is important. Davis and Gather (1993) give a quantitative definition of an outlier:

Definition 44 *(Davis and Gather, 1993). Let $x = (x_1, ..., x_n)$ denote a random sample. For any β, $0 < \beta < 1$, the β outlier region of a null distribution (distribution to be tested) F with mean μ and variance σ^2 is defined by*

$$\Xi\left(\beta, \mu, \sigma^2\right) = \left\{x : |x - \mu| > q_{1-\frac{\beta}{2}}\sigma\right\}, \tag{10.1}$$

where q is the $1 - \frac{\beta}{2}$ quantile of the null distribution F. A number x is called an β outlier with respect to F if $x \in \Xi\left(\beta, \mu, \sigma^2\right)$.

Based on this definition outliers may also affect nonparametric estimators.

10.2.1 Robust Nadaraya-Watson kernel estimator

The Nadaraya-Watson kernel estimator is nonrobust. Based on a functional framework (Aït-Sahalia, 1995) we will calculate the influence function of the estimator to quantify this nonrobustness. First we will define the concept of Frechet differentiability and the uniform Sobolev norm.

Definition 45 *Let $(V; \|\cdot\|_v)$, $(U; \|\cdot\|_u)$ be Banach spaces, let T be a functional $T : V \to U$ and let $B(V, U)$ be the class of all bounded linear operators from V into U. The functional T is Frechet differentiable at $x \in V$ with respect to $\|\cdot\|_v$ if there exists an operator $S_x(T) \in B(V, U) : V \to U$ such that*

$$T(x + h) = T(x) + S_x(T)h + o(h). \tag{10.2}$$

Definition 46 *For any function $g \in C^r$, mth ($m \leq r$) order uniform Sobolev norm is defined as*

$$\|g\| = \sup_{0 \leq c \leq m} \sup_{|a|=c} \sup_{t \in \mathbb{R}^d} \left| \frac{\partial^{(a_1 + ... + a_d)} g(t)}{\partial t_1 ... \partial t_d} \right|, \tag{10.3}$$

where $|a| = a_1 + ... + a_d$. Consider now the case where $V \subset C^r$ and $U = \mathbb{R}$. Let the statistical functional be denoted by $T(F_{XY})$ where $F_{XY} \in V$ is the cumulative joint distribution of the data and the natural estimate of $T(F_{XY})$ is $T(\hat{F}_{XY})$ where $\hat{F}_{XY}$ is the cumulative sample joint distribution function. Let $F_{Y|X}$ denotes the conditional probability distribution function of Y given X.

Theorem 47 *Generalized Delta theorem (Aït-Sahalia, 1995). Let $F_{XY} \in V$ and let $G_{XY} = \hat{F}_{XY} - F_{XY}$.*

(i) The functional T is Frechet differentiable at the point F_{XY} for the norm $\|\cdot\|$ and its differential is given by

$$S_{F_{XY}} T.G_{XY} = \frac{1}{f_X(x)} \left(\int y g_{XY}(x,y) dy - g_X(x) m(x) \right). \tag{10.4}$$

(ii) The functional $T(F_{XY})$ is consistently estimated by $T(\hat{F}_{XY})$ and

$$\sqrt{n} h^{\frac{d-1}{2}} \left(T(\hat{F}_{XY}) - T(F_{XY}) \right) \to \mathcal{N}\left(0, V\left(T(F_{XY})\right)\right). \tag{10.5}$$

By analogy with Hampel's influence function (Hampel, 1994) and based on the Generalized Delta theorem, the influence function of the Nadaraya-Watson kernel estimator is defined as

$$IF\left((x_k, y_k); T, F_{XY}\right) = S_{F_{XY}} T \left(\hat{F}_{XY} - F_{XY} \right) \tag{10.6}$$

with $S_{F_{XY}} T F_{XY} = 0$, we obtain

$$\begin{aligned} IF\left((x_k, y_k); T, F_{XY}\right) &= \frac{1}{f_X(x) h^d} \int y K\left(\frac{x - x_k}{h}\right) K\left(\frac{y - y_k}{h}\right) dy - \\ &\quad \frac{1}{f_X(x) h^{d-1}} K\left(\frac{x - x_k}{h}\right) m\left(x\right) \\ &= \frac{K\left(\frac{x - x_k}{h}\right)}{f_X(x) h^{d-1}} \left(\frac{1}{h} \int y K\left(\frac{y - y_k}{h}\right) dy - m(x) \right). \end{aligned} \tag{10.7}$$

The influence function is unbounded for y in $\mathbb{R}$. Using decreasing kernels, kernels such that $K(u) \to 0$ as $u \to \infty$, the influence function is bounded for x in $\mathbb{R}$. Common choices for decreasing kernels are: $K(u) = \max\left(\left(1 - u^2\right), 0\right)$, $K(u) = \exp - \left(u^2\right)$ and $K(u) = \exp\left(-u\right)$.

By analogy (Boente and Fraiman, 1994), we are interested in the L- robustifed Nadaraya-Watson kernel estimator. A convenient subclass of L-estimators is given by

$$T_L\left(F_{XY}\right) = \int \mathcal{J}\left(F_{Y|X}\left(v\right)\right) F^{-}_{Y|X}\left(F_{Y|X}\left(v\right)\right) dF_{Y|X}\left(v\right) + \sum_{j=1}^{m} a_j F^{-}_{Y|X}\left(q_j\right). \tag{10.8}$$

It is assumed that $0 < q_1 < ... < q_m < 1$ and that $a_1, ..., a_m$ are nonzero constants. This requires that $\mathcal{J}$ must be integrable, but lends itself to formulation of L-estimates as statistical functionals. Thus L-estimates of form (10.8) are sums of two special types of L-estimate, one type weighting all the observations according to a smooth function, the other type consisting of a weighted sum of a fixed number of quantiles. Let $u = F_{Y|X}(v)$ and let $m = 0$, we obtain the L functional

$$T_L(F_{XY}) = \int u\mathcal{J}(u)\, d(u). \tag{10.9}$$

Following Aït-Sahalia, Bickel and Stoker, (2000) the functional T_L is Frechet differentiable at the point F_{XY} for the norm $\|\cdot\|$ and its differential is given by

$$S_{F_{XY}} T_L G_{XY} = \int \frac{\mathcal{J}(u)}{f\left(x, F^{-}_{Y|X}(u)\right)} \left(ug(x) - \frac{\partial^{d-1} g}{\partial x^{(1)} ... \partial x^{(d-1)}}\left(x, F^{-}_{Y|X}(u)\right)\right) du. \tag{10.10}$$

The influence function for the estimator $T_L\left(\hat{F}_{XY}\right)$ is given by

$$\begin{aligned} IF((x_k, y_k); T, F_{XY}) &= S_{F_{XY}} T.\left(\hat{F}_{XY} - F_{XY}\right) \\ &= \int \frac{\mathcal{J}(u)}{f\left(x, F^{-}_{Y|X}(u)\right)} u\left(\frac{1}{h^{d-1}} K\left(\frac{x - x_k}{h}\right) - f(x)\right) du \\ &- \frac{1}{h^d} K\left(\frac{x - x_k}{h}\right) \int \frac{\mathcal{J}(u)}{f\left(x, F^{-}_{Y|X}(u)\right)} K\left(\frac{F^{-}_{Y|X}(u) - y_k}{h}\right) du \\ &+ \int \frac{\mathcal{J}(u)}{f\left(x, F^{-}_{Y|X}(u)\right)} \frac{\partial^{d-1} F}{\partial x^{(1)} ... \partial x^{(d-1)}}\left(x, F^{-}_{Y|X}(u)\right) du \end{aligned} \tag{10.11}$$

The influence function is bounded for y in $\mathbb{R}$. Alternative derivations of the influence function can be found in (Van Der Vaart, 1998) and (Serfling, 1980).

Let $\mathcal{J}(u) = \frac{1}{1-2\beta} I_{[\beta, 1-\beta]}(u)$ which correspond to the β conditional trimmed expectation (10.9), the estimator of $T(F_{XY})$ is given by

$$T\left(\hat{F}_{XY}\right) = \frac{1}{1 - 2\beta} \int_{\beta}^{1-\beta} \hat{F}^{-}_{Y|X}(z)\, dz \tag{10.12}$$

and $\hat{F}_{Y|X}$ is defined as

$$\hat{F}_{Y|X} = \sum_{k=1}^{n} \frac{K\left(\frac{x - x_k}{h}\right)}{\sum_{l=1}^{n} K\left(\frac{x - x_l}{h}\right)} I_{[Y_k \leq y]}, \tag{10.13}$$

where K is the Gaussian kernel. The trimming parameter was set equal to 2.5%.

10.2.2 Smoothing parameter selection for regression

The Nadaraya-Watson kernel estimate requires the tuning of an extra learning parameter, or *tuning parameter*, denoted here by h. We will use the cross-validation as tuning parameter selection method. Essentially, cross-validation uses the data set at hand to verify how well a particular choice of smoothing parameter (bandwidth) does in terms of the residuals. More precisely, the cross-validation score function is defined as

$$CV_{(L_2)}(h) = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{m}_n^{(-k)}(x_k;h))^2, \tag{10.14}$$

where $\hat{m}_n^{(-k)}$ is the so called "leave-one-out" version of $\hat{m}_n$. That is, $\hat{m}_n^{(-k)}$ is constructed with $n-1$ data points by leaving out the data point (x_k, y_k). Notice that the cross-validation criterion is essentially the same as the residual sum of squares, using the L_2 measure. This measure is not a robust one and what is needed is a robust version of the cross-validation. We define the absolute value cross-validation score function by

$$CV_{(L_1)}(h) = \frac{1}{n}\sum_{k=1}^{n}\left|y_k - \hat{m}_n^{(-k)}(x_k;h)\right|. \tag{10.15}$$

This criterion score function should be resistant against outliers.

10.3 Computing the loss function

Let $f(y, m(x))$ denote the model of noise and let $L(y, m(x))$ denote the loss function (the statistic of interest here). In a maximum likelihood sense, for the symmetric density function $f(y, m(x))$, a certain loss function is optimal for a given noise model such that the loss function equals

$$L(y, m(x)) = -\sum_{k=1}^{n}\log f(y_k - m(x_k)). \tag{10.16}$$

10.3.1 Kernel density estimation

Smoothing methods provide a powerful methodology for gaining insights into data. Many examples of this may be found in monographs of (Eubank, 1988; Härdle, 1990; Müller, 1988; Scott, 1992; Silverman, 1986; Wahba, 1990; Wand and Jones, 1994). But effective use of these methods requires: (a) choice of the kernel, and (b) choice of the smoothing parameter (bandwidth). Let $K : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function called the kernel function and let $h > 0$ be a bandwidth or smoothing parameter. The Parzen kernel density estimator is defined as

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{k=1}^{n}K\left(\frac{x - x_k}{h}\right). \tag{10.17}$$

10.3.2 Smoothing parameter selection for density estimation

As it turns out, the kernel density estimator is not very sensitive to the form of the kernel (Rao, 1983). An important problem is to determine the smoothing parameter. In kernel density estimation, the bandwidth has a much greater effect on the estimator than the kernel itself does. When insufficient smoothing is done, the resulting density estimate is too rough and contains spurious features. When excessive smoothing is done, important features of the underlying structure are smoothed (see chapter 8). An objective, for the Parzen kernel estimator, is to choose the smoothing parameter that minimizes the $\mathrm{MISE}\left(\hat{f}_n, f\right)$. Devroye and Lugosi (2001) provides proofs that the banddwidth selection based on L_2 would not be universally useful. In this thesis we use a combination of cross-validation and bootstrap for choosing the bandwidth for the Parzen kernel estimator. The algorithm is as follows:

Algorithm 48 *(Smoothing parameter selection).*

(i) Cross-Validation step. From $x_1, ..., x_n$, construct an initial estimate of the probability density function

$$\hat{f}_n(x) = \frac{1}{nh_0} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h_0}\right),$$

where h_0 is chosen by minimizing the integrated mean squared error (IMSE),

$$\int E\left(\hat{f}_n(x) - f(x)\right)^2 dx \tag{10.18}$$

which can be estimated by the Jackknife principle by

$$\begin{aligned} CV\left(h_0\right) &= \int \left(\hat{f}_n(x)\right)^2 dx - \frac{2}{n}\sum_{k=1}^{n} \hat{f}_n^{(-k)}(x_k) \\ &= \frac{1}{n^2 h_0}\sum_{l=1}^{n}\sum_{k=1}^{n} K\left(\frac{x_l - x_k}{h_0}\right) \circ K\left(\frac{x_l - x_k}{h_0}\right) \\ &+ \frac{1}{h_0}\sum_{l=1}^{n}\frac{1}{(n-1)}\sum_{k\neq l}\frac{1}{h_0}K\left(\frac{x_l - x_k}{h_0}\right). \end{aligned} \tag{10.19}$$

where $\hat{f}_n^{(-k)}$ is the density estimate based on all of the data except x_k and $K(u) \circ K(u)$ is the convolution of the kernel with itself.

(ii) Bootstrap step

(ii.1) Construct a smoothed bootstrap sample Construct the empirical distribution, $\hat{F}_n$, which puts equal mass, $1/n$, at each observation (uniform random sampling with replacement). From the selected $\hat{F}_n$, draw

a sample $x_1^*, ..., x_n^*$, called the bootstrap sample. Adding a random amount $h_0\xi$ to each x_k^*, $k = 1, ..., n$ where ξ is distributed with density $K(\cdot)$. So $x_k^{**} = x_k^* + h_0\xi$.

(ii.2) Estimate the integrated mean absolute error by

$$IMAE_{boot}(h, h_0) = \frac{1}{B}\sum_{b=1}^{B}\int\left|\hat{f}_{n,b}^{**}(x;h) - \hat{f}_n(x;h_0)\right|dx,$$

where $\hat{f}_{n,b}^{**}(x;h) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x-x_k^{**}}{h}\right)$ for $b = 1, ..., B$ and B is the number of bootstrap samples to be taken.

(ii.3) Obtain the bootstrap choice of the bandwidth h_{boot} by minimizing $IMAE_{boot}(h, h_0)$ over h.

10.4 Accuracy of the loss function

10.4.1 Bootstrap method

Figure 10.1 is a schematic diagram of the bootstrap method as it applies to general data structures. On the left an unknown probability mechanism $\mathbb{P}$ has given the observed data $z = (z_1, ..., z_n)$ by random sampling. Having observed data z, we calculate a statistic of interest $S(z)$, and wish to know something about the statistical behavior.

The advantage of the bootstrap is that we can calculate as many replications of $S(z^*)$ as we want. This allows us to probabilistic calculations.

Given a training data set of n points $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ with output data $y_k \in \mathbb{R}$ and input data $x_k \in \mathbb{R}^d$ according to

$$y_k = m(x_k) + e_k, \qquad k = 1, ..., n, \tag{10.20}$$

where $m : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued function. The e_k are assumed to be independent random errors $E[e_k] = 0$ and $E[y_k|X = x_k] = m(x_k)$. The $e_1, ..., e_n$ are (*i.i.d.*) from an unknown distribution F_e with mean zero. In this case, the unknown probability model $\mathbb{P}$ can be identified as $(m(x), F_e)$. Let $\hat{m}_{robust}(x_k)$ be denoted the robust estimation of $m(x_k)$, then F_e can be estimated by the empirical distribution $\hat{F}_e$ putting mass n^{-1} to $\hat{e}_k^\circ - n^{-1}\sum_{j=1}^{n}\hat{e}_j^\circ$, where $\hat{e}_k^\circ = y_k - \hat{m}_{robust}(x_k)$ is the *k-th* residual. $\mathbb{P}$ is now estimated by $\hat{\mathbb{P}} = (\hat{m}_{robust}(x), \hat{F}_e)$. To generate bootstrap data $z_k^* = (x_k, y_k^*)$, we first generate (*i.i.d.*) data $e_1^{\circ *}, ..., e_n^{\circ *}$ from $\hat{F}_{e^\circ}$ and then define $y_k^* = \hat{m}_{robust}(x_k) + e_k^{\circ *}$.

Algorithm 49 *(The bootstrap based on residuals)*

(i) The unknown probability model $\mathbb{P}$ was taken to be $y_k = m(x_k) + e_k$, $k = 1, ..., n$ with $e_1, ..., e_n$ independent errors drawn from some unknown probability distribution F_e

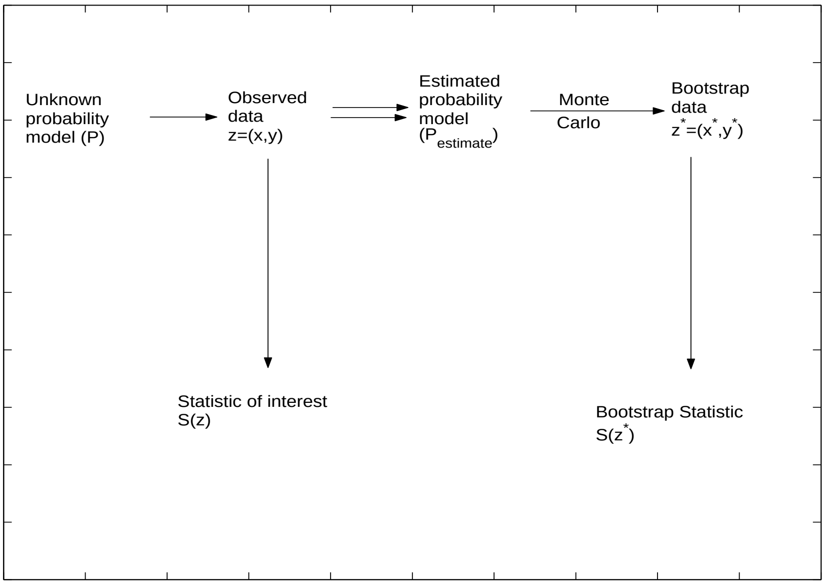
Unknown probability model (P)
Observed data z=(x,y)
Estimated probability model (P_estimate)
Monte Carlo
Bootstrap data z*=(x*,y*)
Statistic of interest S(z)
Bootstrap Statistic S(z*)

Figure 10.1: A general diagram of the bootstrap method for assessing statistical accuracy. On the right hand side of the diagram, the empirical probability mechanism $\hat{\mathbb{P}}$ gives bootstrap samples $z^* = (z_1^*, ..., z_n^*)$ by random sampling, from which we calculate bootstrap replications of the statistic of interest $S(z^*)$.

(ii) *Calculate* $\hat{m}_{robust}\left(x_k\right)$*, and the residuals are* $\hat{e}_k^{\circ} = y_k - \hat{m}_{robust}\left(x_k\right)$*, from which was obtained:*

(ii.1) *An estimated version of* $\hat{F}_{e^{\circ}}$ *: probability* $\frac{1}{n}$ *on* $\hat{e}_k^{\circ}$*.*

(ii.2) $L(y_k, \hat{m}_{robust}\left(x_k\right)) = -\log f(y_k - \hat{m}_{robust}\left(x_k\right))$

(iii) *Bootstrap data* $\mathcal{D}_n^* = \{(x_1, y_1^*), ..., (x_n, y_n^*)\}$ *were generated according to* $y_k^* = \hat{m}_{robust}\left(x_k\right) + \hat{e}_k^{\circ *}$*,* $\hat{e}_1^{\circ *}, ..., \hat{e}_n^{\circ *}$ *independent errors drawn from* $\hat{F}_{e^{\circ}}$ *by Monte Carlo.*

(iv) *Having generated* $\mathcal{D}_n^*$ *, the estimated errors (residuals) are* $\tilde{e}_k^{\circ *} = y_k^* - \hat{m}_{robust}^*\left(x_k\right)$*. Calculate* $L^*(y_k^*, \hat{m}_{robust}^*\left(x_k\right)) = -\log f^*(y_k^* - \hat{m}_{robust}^*\left(x_k\right))$*.*

(v) *This whole process must be repeated B*

where $\hat{m}_{robust}^*\left(x_k\right) = \sum_{robust,i} \frac{K\left(\frac{x_k - x_i}{h}\right)y_i^*}{\sum_{robust,j} K(\frac{x_k - xj}{h})}$ and
$\hat{m}_{robust}\left(x_k\right) = \sum_{robust,i} \frac{K\left(\frac{x_k - x_i}{h}\right)y_i}{\sum_{robust,j} K(\frac{x_k - xj}{h})}$.

10.4.2 Robust bootstrap

The robustification of the bootstrap in based on a control mechanism in the resampling plan, consisting of an alteration of the resampling probabilities, by identifying and downweighting those data points that influence the function estimator (see Chapter robust prediction interval).

10.5 Simulations

This example describes experiments with kernel based regression in estimating the loss function. For these experiments we chose the following regression function

$$y_k = \frac{\sin\left(x\right)}{x} + e_k,\ k = 1, ..., 250 \tag{10.21}$$

where the values y_k are corrupted by (i) $e_k \sim \mathcal{N}\left(0, 0.2^2\right)$ and (ii)
$e_k \sim \mathcal{L}\text{ap}\left(0, 0.2^2\right)$. The results are shown in Figure (10.2) and (10.3). Figure (10.2) gives the results for experiment with $e_k \sim \mathcal{N}\left(0, 0.2^2\right)$ and Figure (10.3) gives the results for experiment with $e_k \sim \mathcal{L}ap\left(0, 0.2^2\right)$. Both Figures give the empirical probability distribution function, the empirical probability density function and the empirical loss function. Note that respectively the L_2 norm loss function and the L_1 loss function can be recognized.

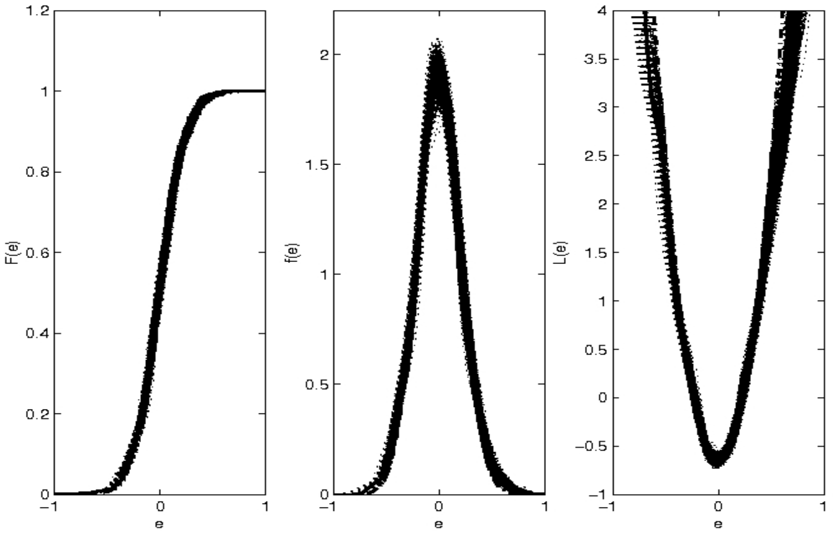

Figure 10.2: Numerical results for the bootstrap experiment with $e_k \sim \mathcal{N}\left(0, 0.2^2\right)$ for recovering the true maximum likelihood loss function.

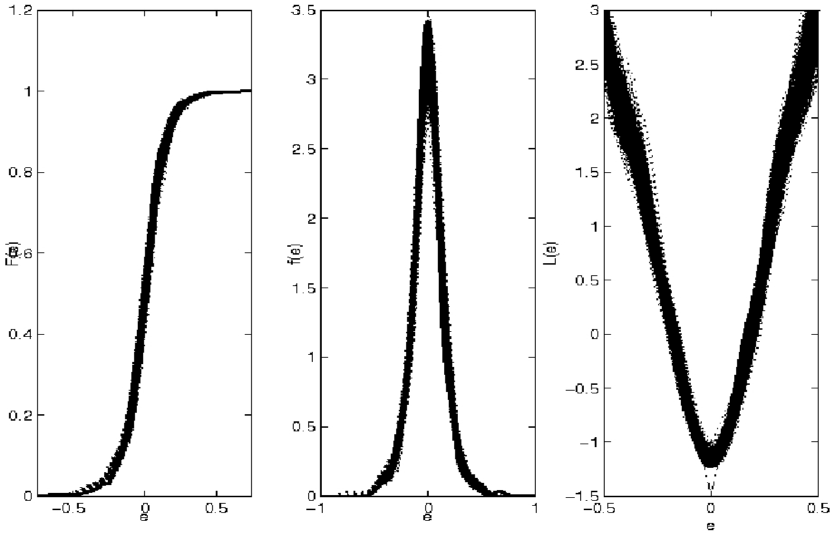

Figure 10.3: Numerical results for the bootstrap experiment with $e_k \sim \mathcal{L}ap\left(0, 0.2^2\right)$ for recovering the true maximum likelihood loss function.

10.6 Conclusions

We have shown that Nadaraya-Watson estimator is nonrobust in the sence of the influence function and that L- regression achieved robustness. Based on the estimated noise model we have calculated the empirical loss function. In an experiment, we have recognized respectively the L_2 norm loss function and the L_1 loss function.

Chapter 11

Robust tuning parameter selection

In this chapter we study the use of robust statistics towards learning parameter selection by cross-validation and the final prediction error (FPE) criterion. For robust learning parameter selection methods robust location estimators such as trimmed mean are applied. Together with robust versions of LS-SVMs, robust counterparts for cross-validation (De Brabanter *et al.*,) and FPE criterion (De Brabanter *et al.*,) are proposed. Finally, simulation results for weighted LS-SVM function estimation are given to illustrate that the proposed robust methods outperforms other cross-validation procedures and methods based on a number of other complexity criteria. Contributions are made in Section 11.3, 11.4, 11.5 and 11.7.

11.1 Introduction

As explained in Chapter 4, most efficient learning algorithms in neural networks, support vector machines and kernel based methods (Bishop, 1995; Cherkassky *et al.*, 1998; Vapnik, 1999; Hastie *et al.*, 2001; Suykens *et al.*, 2002b) require the tuning of some extra learning parameters, or *tuning parameters*, denoted here by θ. For practical use, it is often preferable to have a data-driven method to select θ. For this selection process, many data-driven procedures have been discussed in the literature. Commonly used are those based on the cross-validation criterion of Stone (Stone, 1974) and the generalized cross-validation criterion of Craven and Wahba (Craven and Wahba, 1979). One advantage of cross-validation and generalized cross-validation over selection criteria such as Mallows' C_p and the Final Prediction Error (FPE) criterion (Akaike, 1970) is that they do not require estimates of the error variance. This means that Mallows' C_p, Akaike's (FPE) criterion require a roughly correct working model to obtain the estimate of the error variance. Cross-validation does not require this. But for general dependent data, the cross-validation fails to capture the dependence structure of the data and require nontrivial modifications. An advantage with the Final Prediction Error (FPE) criterion is that the minimization can be performed with respect to different model structures, thus allowing for dependent dat. Based on location estimators (e.g. mean, median, M-estimators, L-estimators, R-estimators), one can find robust counterparts of model selection criteria (e.g. Cross-Validation, Final Prediction Error criterion).

Given a random sample $e = (e_1, ..., e_n)^T$ from a distribution $F(e)$. Let Γ_θ be a model selection criterion and let $\xi = L(e)$ be a function of the random variable e. One can transform the cost function of Γ_θ, based on $\xi = L(e)$, into a simple location problem. The robust counterpart of the model selection criterion is now based on a robust regression estimation and a robust location estimator (e.g. median, M-estimators, L-estimators, R-estimators). The choice of a robust location estimator depends on the distribution $F(\xi)$ and his robustness and efficiency properties.

Definition 50 *(Statistical location model). Letting $\xi = L(e) = (\xi_1, ..., \xi_n)^T$ and $\delta = (\delta_1, ..., \delta_n)^T$ we then write the statistical location model as*

$$\xi_k = \eta + \delta_k, \quad k = 1, ..., n, \tag{11.1}$$

where η is an unknown one-dimensional parameter and δ_k is normally distributed with mean zero and standard deviation σ.

Definition 51 *(Location estimator). Given a location model and a norm $\|\cdot\|$, an estimator $\hat{\eta} = T\left(\hat{F}_n\right)$ of η induced by the norm is*

$$\hat{\eta} = \arg\min_{\eta} \|\xi - 1_n \eta\|, \tag{11.2}$$

where 1_n denotes the vector whose components are 1. The estimator $T\left(\hat{F}_n\right)$ is called a univariate location estimator.

11.2 Location estimators

We will now discuss the location estimators. The least squares (LS) estimator minimizes

$$T\left(\hat{F}_n\right) = \arg\min_{\eta} \sum_{k=1}^{n} (\xi_k - \eta)^2, \tag{11.3}$$

which leads to the arithmetic mean. This estimator has a poor performance in the presence of contamination. Therefore Huber (Huber, 1964) has lowered the sensitivity of the LS loss function by replacing the square by a suitable function ρ. This leads to the location M-estimator.

M-estimators

In this subsection we briefly review the statistics which are obtained as solutions of equations. Often the equations result in an optimization procedure, e.g. in the case of maximum likelihood estimation (MLE), least squares estimation etc.. Such statistics are called M-estimates. An important subclass of M-estimates is introduced by Huber (Huber, 1964). A related class of statistics, L-estimates is treated in the next subsection.

Let $\xi_1, ..., \xi_n$ be a random sample from a distribution F with density $f(\xi - \eta)$, where η is the location parameter. Assume that F is a symmetric unimodal distribution, then η is the center of symmetry to be estimated. The M-estimator $\hat{\eta} = T\left(\hat{F}_n\right)$ of the location parameter is defined then as some solution of the following minimization problem

$$T\left(\hat{F}_n\right) = \arg\min_{\eta} \sum_{k=1}^{n} \rho\left(\xi_k - \eta\right), \tag{11.4}$$

where $\rho\left(t\right)$ is an even non-negative function called the contrast function (Phanzagl, 1969); $\rho(\xi_k - \eta)$ is the measure of discrepancy between the observation ξ_k and the estimated center. For a given density f the choice $\rho\left(t\right) = -\log f\left(t\right)$ yields the MLE. It is convenient to formulate the M-estimators in terms of the derivative of the contrast function $D\left(t\right) = d\rho(t)/dt$ called the score function. In this case, the M-estimator is defined as a solution to the following implicit equation

$$\sum_{k=1}^{n} D\left(\xi_k - \hat{\eta}_n\right) = 0. \tag{11.5}$$

Well-known examples of location parameter estimators are:

- *Example 1*: For $\rho\left(t\right) = t^2$, one obtains the least squares solution by minimization of $\sum_{k=1}^{n}\left(\xi_k - \eta\right)^2$. The corresponding score function is $D\left(t\right) = t$, $-\infty < t < \infty$. For this D, the M-estimate is the sample mean. The contrast function and the respectively score function are sketched in Figure 11.1.

- *Example 2*: For $\rho\left(t\right) = |t|$, one obtains the least absolute values by minimization of $\sum_{k=1}^{n}|\xi_k - \eta|$. The corresponding score function is

$$D\left(t\right) = \begin{cases} -1, & t < 0 \\ 0 & t = 0 \\ 1 & t > 0. \end{cases} \tag{11.6}$$

  The corresponding M-estimate is the sample median. The contrast function and the respectively score function are sketched in Figure 11.2.

- *Example 3*: Huber considers minimization of $\sum_{k=1}^{n}\rho\left(\xi_k - \eta\right)$, where

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & |t| \leq c \\ c\left|t\right| - \frac{1}{2}c^2 & |t| > c. \end{cases} \tag{11.7}$$

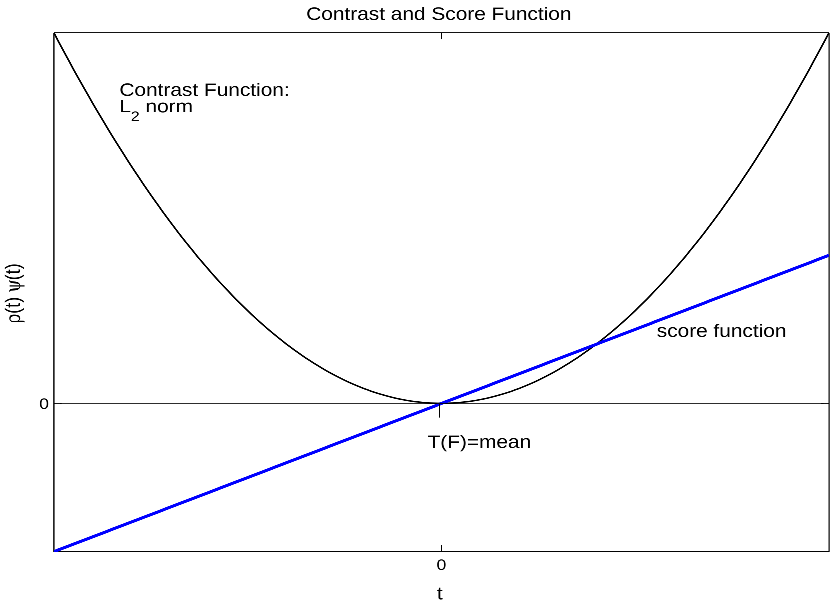

Figure 11.1: The contrast function and the score function of the L_2- norm.

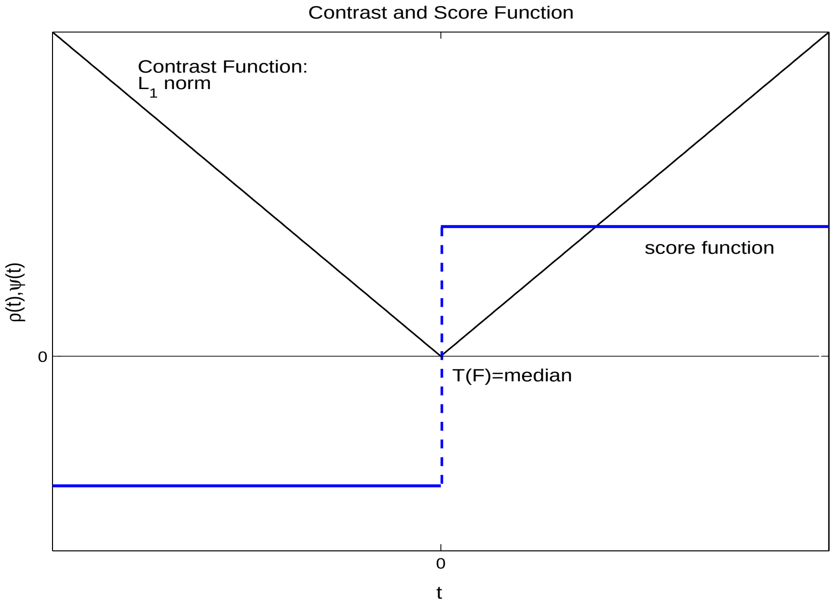

Figure 11.2: The contrast function and the score function of the L_1-norm.

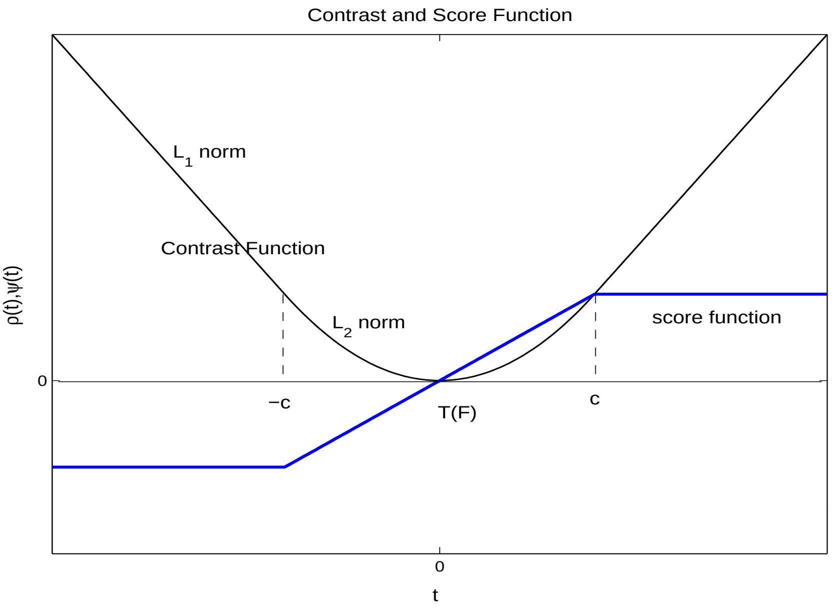

Figure 11.3: The contrast function and the score function of the Huber type of M-estimators.

The score function is

$$D(t) = \begin{cases} -c, & t < -c \\ t & |t| \leq c \\ c & t > c. \end{cases} \tag{11.8}$$

The corresponding M-estimate is a type of Winsorized mean (explained in further detail in next subsection). It turns out to be the sample mean of the modified ξ_k's, where ξ_k becomes replaced by $\hat{\eta} \pm c$, whichever is nearer, if $|\xi_k - \hat{\eta}| > c$. The contrast function and score function are sketched in Figure 11.3.

- *Example 4*: Hampel (1968, 1974) suggested a modification to the Huber estimator:

$$\Psi(t) = \begin{cases} t & 0 \leq |t| \leq a \\ a \text{ sign}(t) & a \leq |t| \leq b \\ a\left(\frac{c-|t|}{c-b}\right) \text{sign}(t) & b \leq |t| \leq c \\ 0 & |t| > c, \end{cases} \tag{11.9}$$

making $\Psi(t)$ zero for $|t|$ sufficiently large. This M-estimator has the property of completely rejecting outliers. The contrast function and score function are sketched in Figure 11.4.

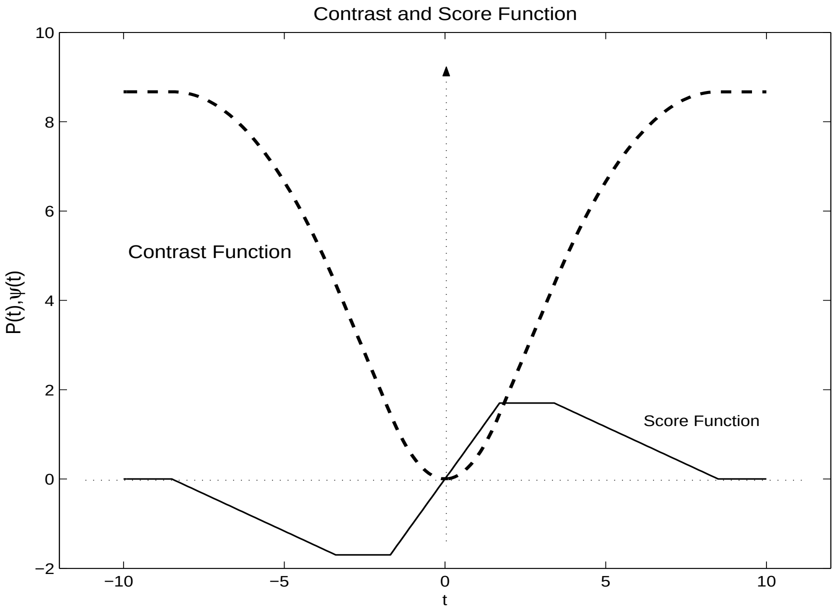

Figure 11.4: Hampel's modification to the Huber estimator. This M-estimator has the property of completely rejecting outliers.

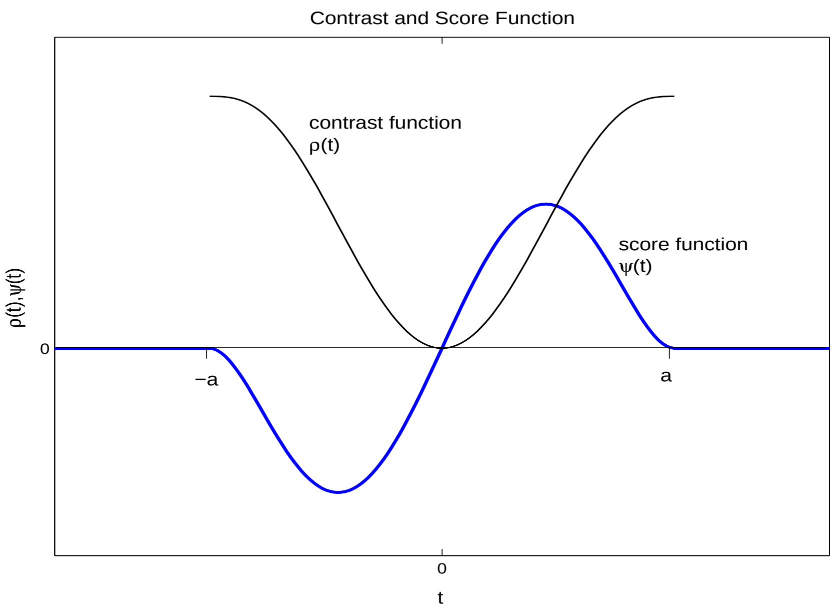

Figure 11.5: Tukey's biweight contrast and score function

- *Example* 5: A very smooth score function, the biweight was proposed by Tukey (1974) and has become increasingly popular. The score function is given by

$$\Psi\left(t\right)=t\left(a^{2}-t^{2}\right)^{2}\delta_{\left[-a,a\right]}(t), \tag{11.10}$$

  where

$$\delta_{\left[-a,a\right]}=\begin{cases}1 & \text{if } t\in\left[-a,a\right]\\ 0 & \text{otherwise.}\end{cases} \tag{11.11}$$

  The contrast function and the respectively score function are sketched in Figure 11.5.

L-estimators

L-estimators were originally proposed by Daniel (1920) and since then have been forgotten for many years, with a revival now in robustness studies. The description of L-estimators can be formalized as follows.

Let $\xi_1,...,\xi_n$ be a random sample on a distribution F, the ordered sample values $\xi_{n(1)}\leq...\leq\xi_{n(n)}$ are called the order statistics. A linear combination of (transformed) order statistics, or L-statistic, is a statistic of the form

$$T\left(\hat{F}_{n}\right)=\sum_{j=1}^{n}C_{n(j)}a\left(x_{n(j)}\right), \tag{11.12}$$

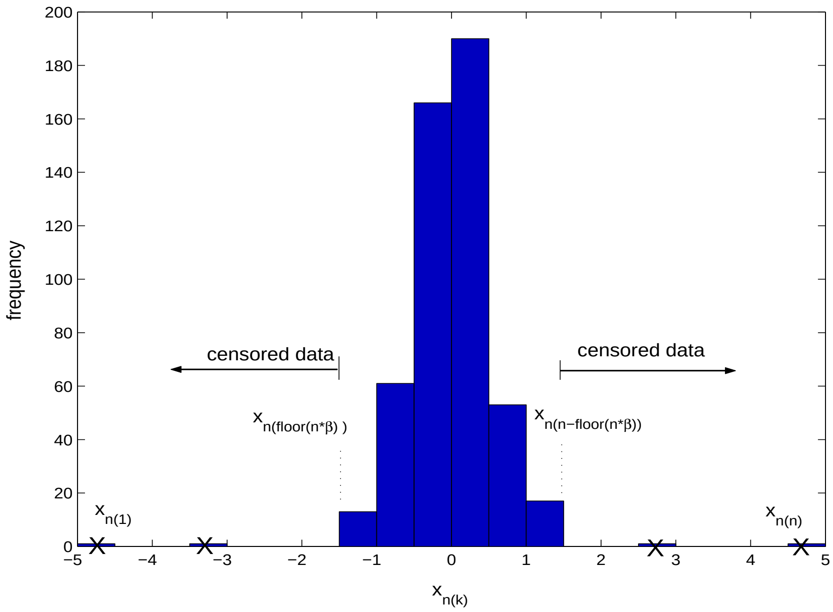

Figure 11.6: Schematic representation of the trimmed mean on a symmetric distribution.

for some choice of constants $C_{n(1)}, ..., C_{n(n)}$ where $\sum_{j=1}^{n} C_{n(j)} = 1$ and $a(\cdot)$ is some fixed function. The simplest example of an L-statistic is the sample mean. More interesting, a compromise between mean and median (trade-off between robustness and asymptotic efficiency), is the β_2-trimmed mean (Figure 11.6) defined as

$$\hat{\mu}_{(\beta_2)} = \frac{1}{n-2g} \sum_{j=g+1}^{n-g} \xi_{n(j)}, \tag{11.13}$$

where the trimming proportion β is selected so that $g = \lfloor n\beta_2 \rfloor$ and $a(\xi_{n(j)}) = \xi_{n(j)}$ is the identity function. The β-trimmed mean is a linear combination of the order statistics given zero weight to a number g of extreme observations at each end. It gives equal weight $1/(n-2g)$ to the number of $(n-2g)$ central observations. When F is no longer symmetric, it may sometimes be preferable to trim asymmetrically if the tail is expected to be heavier in one direction than the other. If the trimming proportions are β_1 on the left and β_2 on the right, the (β_1, β_2)-trimmed mean is defined as

$$\hat{\mu}_{(\beta_1,\beta_2)} = \frac{1}{n-(g_1+g_2)} \sum_{j=g_1+1}^{n-g_2} \xi_{n(j)}, \tag{11.14}$$

where β_1 and β_2 are selected so that $g_1 = \lfloor n\beta_1 \rfloor$ and $g_2 = \lfloor n\beta_2 \rfloor$. The (β_1, β_2)-trimmed mean is a linear combination of the order statistics giving zero weight

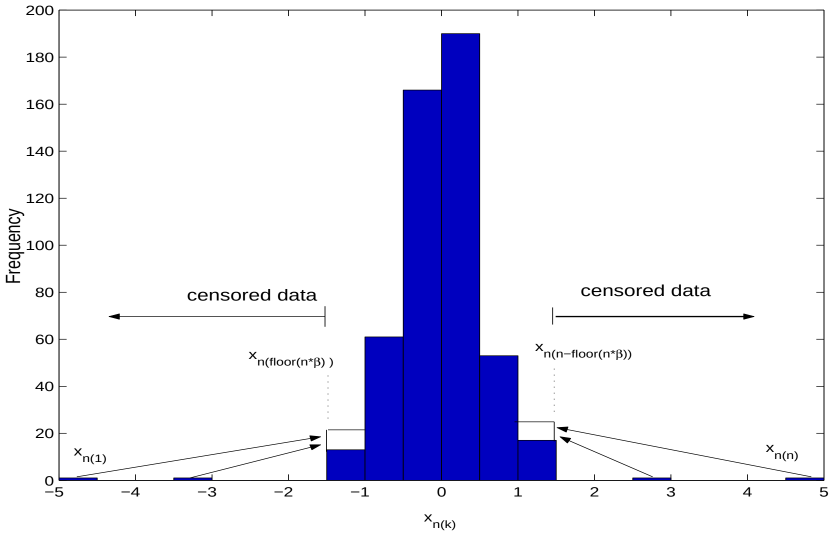

Figure 11.7: Schematic representation of the Winsorized mean on a symmetric distribution.

to g_1 and g_2 extreme observations at each end and equal weight $1/(n - g_1 - g_2)$ to the $(n - g_1 - g_2)$ central observations.

Another L-estimator is the β-Winsorized mean (Figure 11.7). Let $0 < \beta < 0.5$, then the β-Winsorized means (in the symmetric case) is defined as

$$\hat{\mu}_{W(\beta)} = \frac{1}{n}\left(g x_{n(g+1)} + \sum_{j=g+1}^{n-g} \xi_{n(j)} + g \xi_{n(n-g)} \right). \tag{11.15}$$

While the β-trimmed mean censors the smallest and largest $g = \lfloor n\beta \rfloor$ observations, the β-Winsorized means replaces each of them by the values of the smallest and the largest uncensored ones.

11.3 Robust V-fold Cross-Validation

The motivation behind cross-validation is easily understood, see (Allen, 1974) and (Stone, 1974). Much work has been done on the ordinary or leave-one-out cross-validation (Bowman, 1984) and (Härdle and Marron, 1985). However, the difficulty with ordinary cross-validation is that it can become computationally very expensive in practical problems. Therefore, (Burman, 1989) has introduced V-fold cross-validation. For more references on smoothing parameter selection, see (Marron, 1987, 1989) and (Härdle and Chen, 1995).

In recent years, results on L_2 and L_1 cross-validation statistical properties have become available (Yang and Zheng, 1992). However, the condition

$E\left[e_k^2\right] < \infty$ (respectively, $E\left[|e_k|\right] < \infty$) is necessary for establishing weak and strong consistency for L_2 (respectively, L_1) cross-validated estimators. On the other hand, when there are outliers in the output observations (or if the distribution of the random errors has a heavy tail so that $E\left[|e_k|\right] = \infty$), then it becomes very difficult to obtain good asymptotic results for the $L_2\,(L_1)$ cross-validation criterion. In order to overcome such problems, a robust cross-validation score function is proposed in this paper. This is done by first treating the values of the cross-validation score function as a realization of a random variable. In a second stage, the location parameter (e.g. the mean) of this realization is estimated by a robust method. The results of this paper illustrate that the robust methods can be very effective, especially with non-Gaussian noise distributions and outliers in the data.

The cross-validation procedure can basically be split up into two main parts: (a) constructing and computing the cross-validation score function, and (b) finding the tuning parameters by $\theta^* = \mathrm{argmin}_\theta\left[CV_{V-fold}\left(\theta\right)\right]$. In this thesis we focus on (a). Let $\{z_k = (x_k, y_k)\}_{k=1}^n$ be an i.i.d. random sample from some population with distribution function $F\left(z\right)$. Let $\hat{F}_n\left(z\right)$ be the empirical estimate of $F\left(z\right)$. Our goal is to estimate a quantity of the form

$$T(\hat{F}_n) = \int L\left(z, \hat{F}_n(z)\right) dF(z), \tag{11.16}$$

with $L(\cdot)$ the loss function (e.g. the L_2 or L_1 norm) and where $E\left[T(\hat{F}_n)\right]$ could be estimated by cross-validation. We begin by splitting the data randomly into V disjoint sets of nearly equal size. Let the size of the v-th group be m_v and assume that $\lfloor n/V \rfloor \leq m_v \leq \lfloor n/V \rfloor + 1$ for all v. Let $\hat{F}_{(n-m_v)}\left(z\right)$ be the empirical estimate of $F\left(z\right)$ based on $(n - m_v)$ observations outside group v and let $\hat{F}_{m_v}\left(z\right)$ be the empirical estimate of $F\left(z\right)$ based on m_v observations inside group v. Then a general form of the V-fold cross-validated estimate of $T(\hat{F}_n)$ is given by

$$CV_{V-fold}\left(\theta\right) = \sum_{v=1}^{V} \frac{m_v}{n} \int L\left(z, \hat{F}_{(n-m_v)}\left(z\right)\right) d\hat{F}_{m_v}\left(z\right). \tag{11.17}$$

Let $\hat{f}^{(-m_v)}\left(x;\theta\right)$ be the regression estimate based on the $(n - m_v)$ observations outside the group v. Then the least squares V-fold cross-validated estimate of $T(\hat{F}_n)$ is given by

$$CV_{V-fold}\left(\theta\right) = \sum_{v=1}^{V} \frac{m_v}{n} \sum_{k=1}^{m_v} \frac{1}{m_v} \left(y_k - \hat{f}^{(-m_v)}\left(x_k;\theta\right)\right)^2. \tag{11.18}$$

Let $\xi = L(e)$ be a function of a random variable e. In the V-fold cross-validation case (11.18), a realization of the random variable e is given by $e_k = \left(y_k - \hat{f}^{(-m_v)}\left(x_k;\theta\right)\right)$, $k = 1, ..., m_v\ \forall v$, and the cross-validation score function

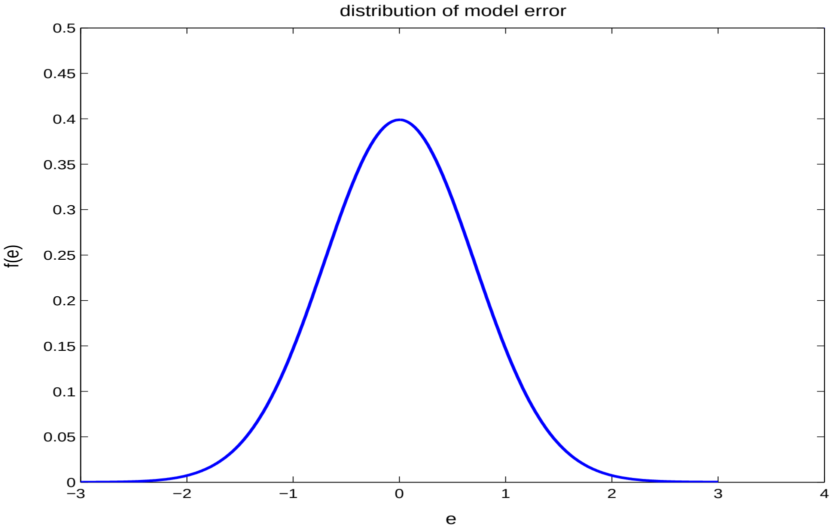

Figure 11.8: Noise distribution. $F(e)$ is unknown and is assumed to have zero mean.

can be written as function of the number of $V+1$ location problems. It estimates a location parameter of the corresponding v-samples.

$$CV_{V-fold}(\theta) = \sum_{v=1}^{V} \frac{m_v}{n} \left(\frac{1}{m_v} \sum_{k=1}^{m_v} L(e_k) \right) = \sum_{v=1}^{V} \frac{m_v}{n} \left(\frac{1}{m_v} \sum_{k=1}^{m_v} \xi_k \right)$$
$$= \hat{\mu}\left(\hat{\mu}_1(\xi_{11}, ..., \xi_{1m_1}), ..., \hat{\mu}_V(\xi_{V1}, ..., \xi_{Vm_v})\right), \qquad (11.19)$$

where ξ_{vj} denotes the j-th element of the v-th group, $\hat{\mu}_v(\xi_{v1}, ..., \xi_{vm_1})$ denotes the sample mean of the v-th group and $\hat{\mu}$ is the mean of all sample group means. Consider only the random sample of the v-th group and let $\hat{F}_{m_v}(\xi)$ be the empirical distribution function. Then $\hat{F}_{m_v}(\xi)$ depends in a complicated way on the noise distribution $F(e)$, the θ values and the loss function $L(\cdot)$. In practice $F(e)$ is unknown except for the assumption of symmetry around 0 (see Figure 11.8). Whatever the loss function would be (L_2 or L_1), the distribution $\hat{F}_{m_v}(\xi)$ is always concentrated on the positive axis with an asymmetric distribution (see Figure 11.9). The asymmetric distribution of $\hat{\mu}_1, ..., \hat{\mu}_V$, denoted by $F(\hat{\mu}_V)$ is sketched in Figure 11.10. There is a lot of variability in the V-fold cross validated estimate, because the number of ways that n random values can be grouped into V classes with m_v in the vth class, $i = 1, ..., V$, and $\sum_{i=1}^{V} m_v = n$ equals $\frac{n!}{m_1! m_2! ... m_V!}$.

We propose now the following procedure. Permute and split repeatedly the data - e.g. r times - into V groups as discussed. then the V-fold cross-validation score function is calculated for each split and finally take the average of the r

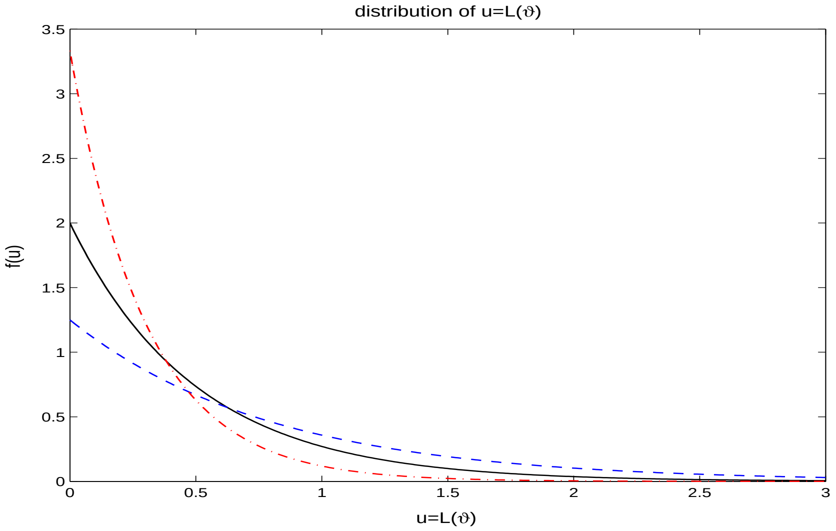

Figure 11.9: Squared residual distribution, $F_{m_v}(u)$ concentrated on the positive axis with an asymmetric distribution.

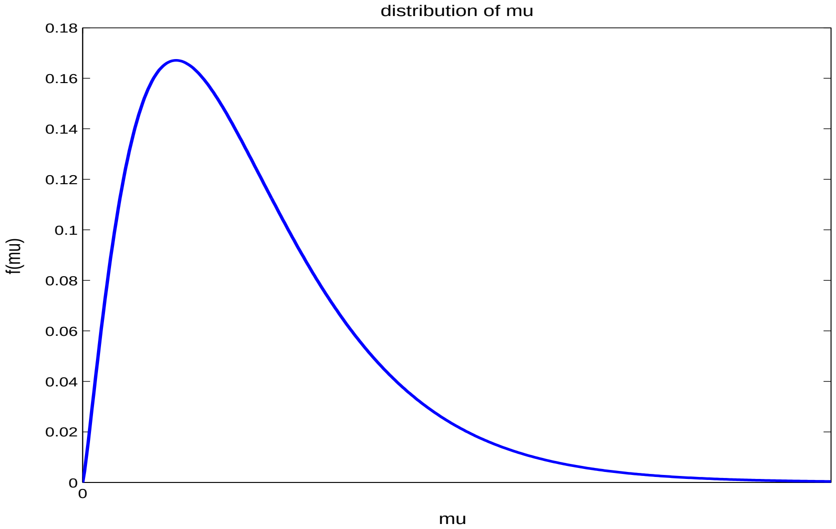

Figure 11.10: Asymmetric distribution $F(\mu_v)$.

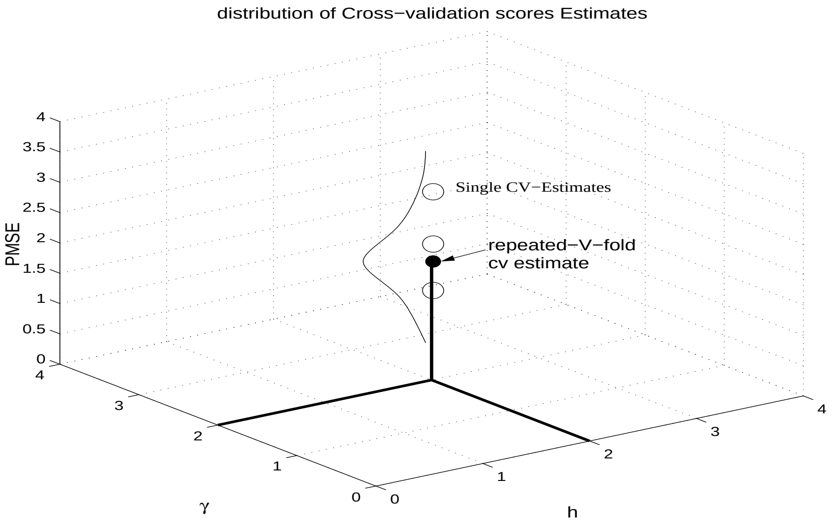

Figure 11.11: Schematic representation of the sampling distribution corresponding with one point in the tuning parameter space.

estimates

$$\text{Repeated\_}CV_{V-fold}\left(\theta\right)=\frac{1}{r}\sum_{j=1}^{r}CV_{V-fold,j}\left(\theta\right). \qquad (11.20)$$

The distribution of the Repeated_$CV_{V-fold}(\theta)$ is asymptotically normally distributed. The repeated V-fold cross-validation score function has about the same bias as the V-fold cross-validation score function, but the average of r estimates is less variable than for one estimate. This is illustrated in Figure 11.11.

11.4 Repeated Robust and Efficient V-fold Cross-validation Score Function

A classical cross-validation score function with L_2 or L_1 works well in situations where many assumptions (such as $e_k \sim \mathcal{N}\left(0,\sigma^2\right)$, $E\left[e_k^2\right] < \infty$ and no outliers) are valid. These assumptions are commonly made, but are usually at best approximations to reality. For example, non-Gaussian noise and outliers are common in data-sets and are dangerous for many statistical procedures and also for the cross validation score function. Given the previous derivations of robustness and efficiency, a new variant of the classical cross-validation score function is introduced based on the trimmed mean. There are several practical reasons to use this type of robust estimator, which is the least squares solution after discarding (in our case) the $g_2 = \lfloor n\beta_2 \rfloor$ largest observations:

- The trimmed mean can be applied when the sample distribution is symmetric or asymmetric.
- It is easy to compute. It is a reasonable descriptive statistic, which can be used as an estimator of the mean of the corresponding truncated distribution.
- For large n, the trimmed mean has an approximate normal distribution (Bickel and Peter, 1965). The standard deviation can be estimated based on the Winsorized sum of squares (Huber, 1970).
- It can be used as an adaptive statistic.

The general form of the V-fold cross-validation score function based on the sample mean is given in (11.17). The robust V-fold cross-validation score function based on the trimmed mean is formulated as

$$CV_{V-fold}^{Robust}\left(\theta\right)=\sum_{v=1}^{V}\frac{m_v}{n}\int_{0}^{F^{-}(1-\beta_2)}L\left(z,F_{(n-m_v)}\left(z\right)\right)dF_{m_v}\left(z\right). \qquad (11.21)$$

Let $\hat{f}_{Robust}\left(x;\theta\right)$ be a regression estimate constructed via a robust method, for example the weighted LS-SVM (Suykens *et al.*, 2002). Then the least squares robust V-fold cross-validation estimate is given by

$$CV_{V-fold}^{Robust}\left(\theta\right)=\sum_{v=1}^{V}\frac{m_v}{n}\sum_{k}^{m_v}\frac{1}{m_v-\lfloor m_v\beta_2\rfloor}\left(y_k-f_{Robust}^{(-m_v)}(x_k;\theta)\right)_{m_v(k)}^2$$
$$I_{[m_v(1),m_v(m_v-\lfloor m_v\beta_2\rfloor)]}((y_k-f_{Robust}^{(-m_v)}(x_k;\theta))^2), \qquad (11.22)$$

where $(y_k-f_{Robust}^{(-m_v)}(x_k;\theta))_{m_v(k)}^2$ is an order statistic and the indicator function $I_{[a,b]}(z)=1$ if $a<z<b$ and otherwise 0.

The robust V-fold cross-validation score function can also be written as

$$CV_{V-fold}^{Robust}\left(\theta\right)=\hat{\mu},...,\xi_{m_1(m_1)} \qquad (11.23)$$

It estimates a location parameter of the v-samples, where $\hat{\mu}_{(0,\beta_{2,v})}$ $\left(\xi_{m_v(1)},...,\xi_{m_v(m_v)}\right)$ is the sample $(0,\beta_2)$-trimmed mean of the v-th group, and $\hat{\mu}$ is the mean of all the sample group $(0,\beta_2)$-trimmed mean. To use a $(0,\beta_2)$-trimmed mean, one must decide on a value of β_2. Guidelines for selection of this value can be found in (Hogg, 1974). If one is particularly concerned with good protection against outliers and if from past experience one has an idea about the frequency of occurrence of such outliers (5 to 10% is typical for many types of data) one would choose a value β_2 somewhat above the expected proportion of outliers.

Similar as presented in Section 13.2 for the V-fold CV score function, the data is permuted and splitted repeatedly - e.g. r times - into V groups. For each split, the robust V-fold cross-validation score function is calculated. The final result is the average of the r estimates. This procedure reduces the variance of the score function

$$\text{Repeated\_}CV_{V-fold}^{Robust}\left(\theta\right)=\frac{1}{r}\sum_{j=1}^{r}CV_{V-fold,j}^{Robust}\left(\theta\right). \tag{11.24}$$

Remark that the robust cross-validation score function inherents all nice properties of the trimmed mean and his IF has the same form.

11.5 Robust Generalized Cross-validation Score Function

A natural approach to robustify the GCV is by replacing the linear procedure of averaging by the corresponding robust counterparts. Let $\xi = L(\vartheta)$ be a function of a random variable ϑ. In the GCV case, a realization of the random variable $\vartheta = g(e)$ is given by

$$\vartheta_k=\left(\frac{y_k-f^*(x_k;\theta)}{1-(1/\sum_k v_k)tr(S^*)}\right),\quad k=1,\ldots,n \tag{11.25}$$

where $f^*(x_k;\theta)$ is the weighted LS-SVM as described in Section 11.3.2, the weighting of $f^*(x_k;\theta)$ corresponding with $\{x_k, y_k\}$ is denoted by v_k and the smoother matrix based on these weightings is defined as in Eq.(3.12) where Z is replaced by $Z^* = (\Omega + V_\gamma)$ with $V_\gamma = \text{diag}\left\{\frac{1}{\gamma v_1},...,\frac{1}{\gamma vn}\right\}$. The GCV can now be written as

$$GCV(\theta)=\frac{1}{n}\sum_{k=1}^{n}L(\vartheta_k)=\frac{1}{n}\sum_{k=1}^{n}\vartheta_k^2. \tag{11.26}$$

Using a robust analog of the sum ($(0,\beta_2)$ - trimmed mean), the robust GCV is defined by

$$GCV_{robust}(\theta)=\frac{1}{n-\lfloor n\beta_2\rfloor}\sum_{k=1}^{n-\lfloor n\beta_2\rfloor}I_{[\vartheta_{n(1)},\vartheta_{n(n-\lfloor n\beta_2\rfloor)}]}(\vartheta^2) \tag{11.27}$$

where $I_{[\cdot,\cdot]}(\cdot)$ is an indicator function.

11.6 Illustrative examples

11.6.1 Artificial data set

In this example we compare eight criteria: leave-one-out CV, $CV_{V-fold}^{L_2}$, $CV_{V-fold}^{L_1}$, AIC, BIC, the repeated CV_{V-fold}^{robust}, GCV and robust GCV for use in tuning parameter selection of function estimation. First, we show three examples of estimating a sinc function where the noise model is described by:

Method	L_2	L_1	L_∞
L_2 Loo-CV+LS-SVM	0.000587	0.020209	0.083482
L_2 V–fold CV+LS-SVM	0.000621	0.020686	0.093063
robust V–fold CV+weighted LS-SVM	0.000586	0.020399	0.076741
L_1 V–fold CV+LS-SVM	0.000644	0.020979	0.097678
AIC	0.000645	0.021227	0.091463
BIC	0.000687	0.022292	0.085469
GCV+LS-SVM	0.000645	0.021227	0.091463
robust GCV+weighted LS-SVM	0.000645	0.021227	0.091463

Table 11.1: Numerical performance measured on fresh test data for the results of the sinc function without outliers. The results compare the performance of an LS-SVM on data with a Gaussian noise model tuned by different performance criteria. The robust procedures performs equally well as the classical methods in the non-contamination case.

(a) noise defined as $F_\epsilon(x) = \mathcal{N}(0,\sigma^2)$ (Table 11.1), (b) contamination noise defined as $F_\epsilon(x) = (1-\epsilon)\mathcal{N}(0,\sigma^2) + \epsilon\mathcal{N}(0,\kappa^2\sigma^2)$, $\epsilon = 0.15$, $\kappa = 1$ (Figures 11.12, 11.13 and 11.14) and (c) contamination noise defined as $F_\epsilon(x) = (1-\epsilon)\mathcal{N}(0,\sigma^2) + \epsilon\,\text{Lap}(0,\lambda)$, $\epsilon = 0.15$, $\lambda = 1$ (Figures 11.15, 11.16 and 11.17)

Given is a training set with $n = 150$ data points. From the simulation results it is clear that in all contaminated cases the LS-SVM tuned by the classical methods are outperformed by the robust methods for tuning the weighted LS-SVM. With the proposed robust procedures, the contamination has practically no influence on the tuning parameter selection. An important property of these robust procedures is that in the non-contamination case (c), it performs equally well as the classical methods (table 11.1). A Monte Carlo simulation (this experiment is repeated 150 times) was a carried out to compare the different criteria. The LS-SVM estimates are presented with tuning parameters selected by different criteria. Figures 11.12, 11.13 and 11.14 gives the boxplots of the simulations for case (a). Figures 11.15, 11.16 and 11.17 gives boxplots of the simulations for case (b).

11.6.2 Real data sets

Body fat data

In the body fat data set (Penrose *et al.*, 1985) the response variable "body fat" and the 18 independent variables are recorded for 252 men. The last third part of permuted observations is used as independent test set to compare the obtained results as given in Table 11.2. After examination of the data, the trimming proportion of the robust cross-validation procedure was set to 5%. The results show the improved performance of the proposed robust procedures in different norms (L_1, L_2 and L_∞).

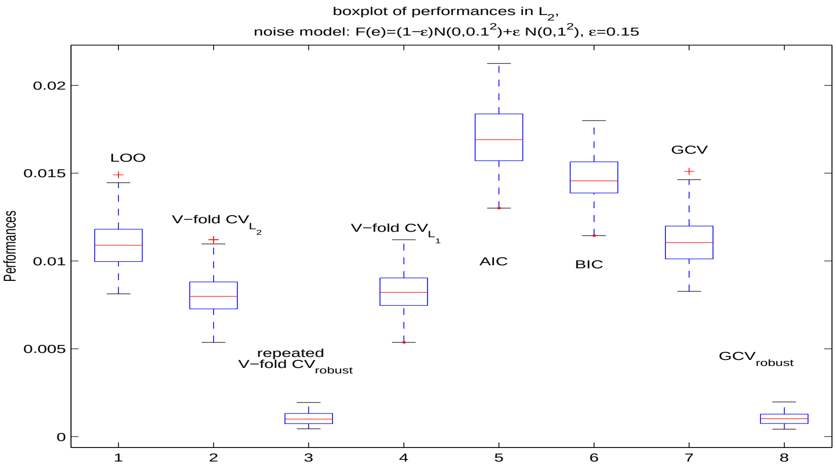

Figure 11.12: The boxplots of the Monte Carlo simulations on artificial data (sinc function) for the contamination noise $\epsilon\left(\mathcal{N}\left(0,1^2\right)\right)$, $\epsilon = 0.15$. Each box in the figure gives the median and the standard deviation of the sample. In the L_2 norm the best results (mean and variance) are obtained by robust V-fold crossvalidation and robust generalized crossvalidation combined with the weighted LS-SVM.

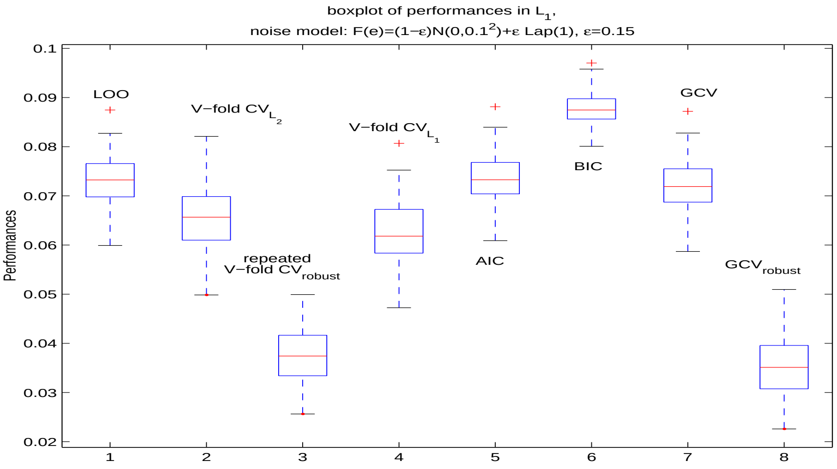

Figure 11.13: The boxplots of the Monte Carlo simulations on artificial data (sinc function) for the contamination noise $\epsilon\left(\mathcal{N}\left(0,1^2\right)\right)$, $\epsilon = 0.15$. Each box in the figure gives the median and the standard deviation of the sample. In the L_1 norm the best results (mean and variance) are obtained by robust V-fold crossvalidation and robust generalized crossvalidation combined with the weighted LS-SVM.

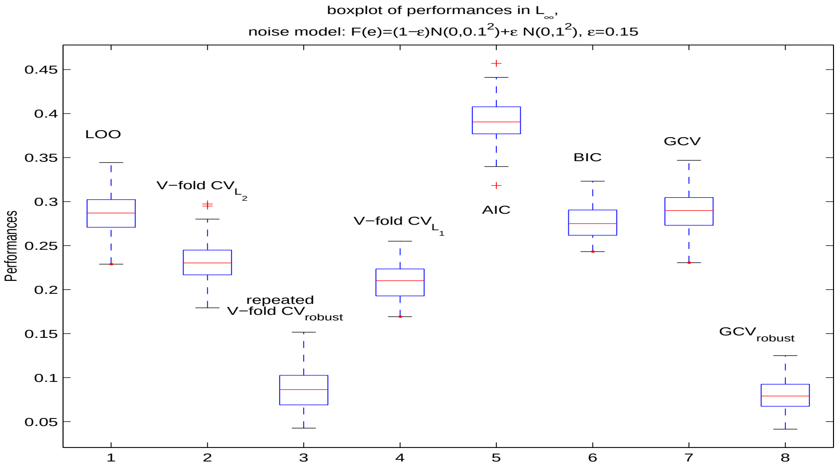

Figure 11.14: The boxplots of the Monte Carlo simulations on artificial data (sinc function) for the contamination noise $\epsilon\left(\mathcal{N}\left(0,1^2\right)\right)$, $\epsilon = 0.15$. Each box in the figure gives the median and the standard deviation of the sample. In the L_∞ norm the best results (mean and variance) are obtained by robust V-fold crossvalidation and robust generalized crossvalidation combined with the weighted LS-SVM.

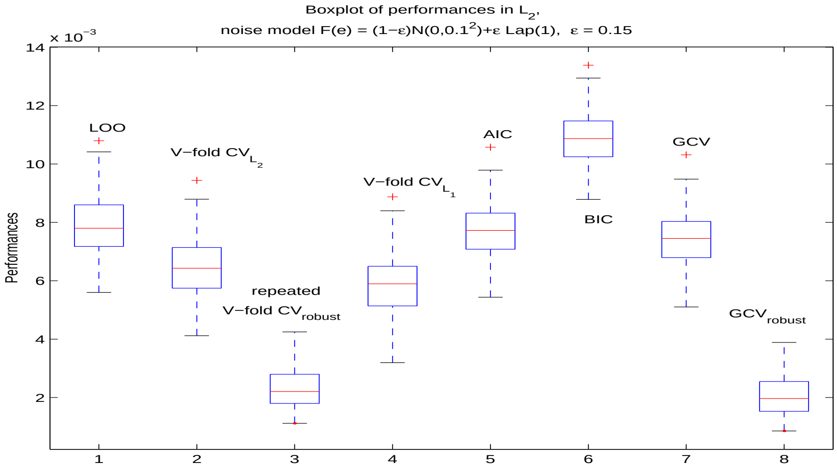

Figure 11.15: The boxplots of the Monte Carlo simulations on artificial data (sinc function) for the contamination noise $\epsilon\left(Lap\left(0,1^2\right)\right)$, $\epsilon = 0.15$. Each box in the figure gives the median and the standard deviation of the sample. In the L_2 norm the best results (mean and variance) are obtained by robust V-fold crossvalidation and robust generalized crossvalidation combined with the weighted LS-SVM.

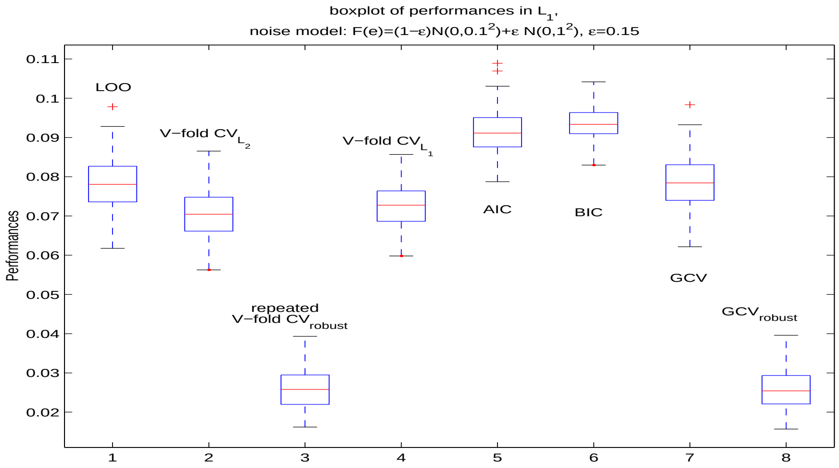

Figure 11.16: The boxplots of the Monte Carlo simulations on artificial data (sinc function) for the contamination noise $\epsilon\left(Lap\left(0,1^2\right)\right)$, $\epsilon = 0.15$. Each box in the figure gives the median and the standard deviation of the sample. In the L_1 norm the best results (mean and variance) are obtained by robust V-fold crossvalidation and robust generalized crossvalidation combined with the weighted LS-SVM.

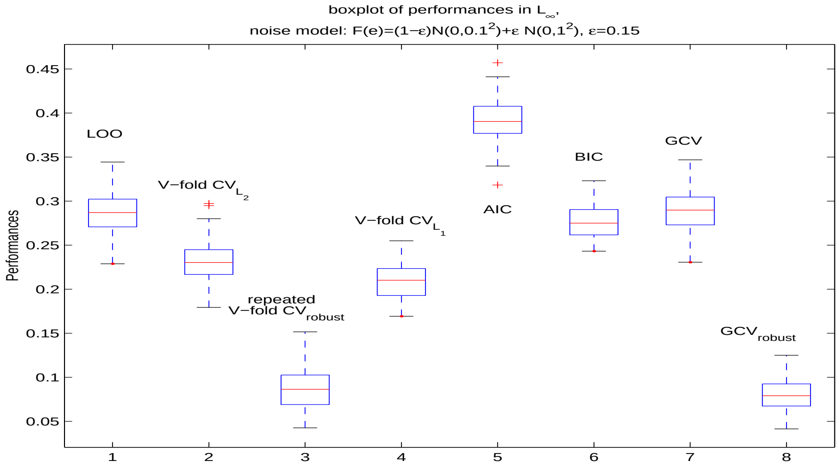

Figure 11.17: The boxplots of the Monte Carlo simulations on artificial data (sinc function) for the contamination noise $\epsilon\left(Lap\left(0,1^2\right)\right)$, $\epsilon = 0.15$. Each box in the figure gives the median and the standard deviation of the sample. In the L_∞ norm the best results (mean and variance) are obtained by robust V-fold crossvalidation and robust generalized crossvalidation combined with the weighted LS-SVM.

Method	L_2	L_1	L_∞
L_2 Loo-CV+LS-SVM	0.0000209	0.00363	0.0136
L_1 V–fold CV+LS-SVM	0.0000166	0.00306	0.0156
AIC	0.0000996	0.00819	0.0256
BIC	0.0000996	0.00819	0.0256
GCV+LS-SVM	0.0000193	0.00323	0.0138
robust V–fold CV+weighted LS-SVM	0.0000014	0.00078	0.0046
robust GCV+weighted LS-SVM	0.0000084	0.00226	0.0101

Table 11.2: Numerical performance measured on fresh test data for the body fat data set. The results compare the performance of an LS-SVM on this real data tuned by different performance criteria. The robust procedures outperform the classical methods in this case.

Method	L_2	L_1	L_∞
L_2 Loo-CV+LS-SVM	3.9974	1.5925	4.9841
L_1 V–fold CV+LS-SVM	3.9956	1.5918	4.9804
AIC	6.6044	1.4824	18.6141
BIC	11.6372	1.3864	29.0393
GCV+LS-SVM	4.7557	1.7083	5.4784
robust V–fold CV+weighted LS-SVM	3.9158	1.5846	5.0697
robust GCV+weighted LS-SVM	3.9316	1.5813	5.0104

Table 11.3: Numerical performance measured on fresh test data for the Boston housing data set. The results compare the performance of an LS-SVM on this real data tuned by different performance criteria. The robust procedures are slightly better (AIC, BIC and generalized cross-validation perform significantly worse than the others.

Boston housing data

The Boston housing data set (Harrison *et al.*, 1978) is composed of 506 objects. There are 13 continuous variables (including the response variable "MEDV") and one binary valued variable. The last third part of permuted observations is used as independent test set to compare the obtained results as given in Table 11.3

11.7 Robust complexity criteria

11.7.1 Robust Final Prediction Error (FPE) criterion

Let $\mathcal{Q}_n$ be a finite set of effective number of parameters. For $q \in \mathcal{Q}_n$, let $\mathcal{F}_{n,q}$ be a set of functions f, let $Q_n(q) \in \mathbb{R}^+$ be a complexity term for $\mathcal{F}_{n,q}$ and let

$\hat{f}$ be an estimator of f in $\mathcal{F}_{n,q}$. The model parameters, $\theta \in \Theta$, are chosen to be the minimizer a generalized Final Prediction Error (FPE) criterion defined as

$$J_C(\theta) = \frac{1}{n} RSS + \left(1 + \frac{2\mathrm{tr}(S(\hat{\theta})) + 2}{n - \mathrm{tr}(S(\hat{\theta})) - 2}\right) \hat{\sigma}_e^2. \tag{11.28}$$

Each of these selectors depends on $S(\hat{\theta})$ through its trace ($\mathrm{tr}(S(\hat{\theta})) < n - 2$), which can be interpreted as the effective number of parameters used in the fit. Because $J_C(\theta)$ is based on least squares estimation (via $L\left(y_k, \hat{f}\left(x_k; \theta\right)\right), \hat{\sigma}_e^2$), it is very sensitive to outliers and other deviations from the normality assumption on the error distribution. A natural approach to robustify the Final Prediction Error (FPE) criterion $J_C(\theta)$ is as follows:

(i). A robust estimator $\hat{m}_n^{\circ}\left(x, \theta\right)$ based on (e.g. M-estimator (Huber, 1964) or weighted LS-SVM (Suykens *et al.*, 2002)) replaces the LS-SVM $\hat{m}_n\left(x, \theta\right)$. The corresponding smoother matrix $S^*\left(v_1, ..., v_n; \hat{\theta}\right)$ is now based on the weighting elements v_k, $k = 1, ..., n$ of $\hat{m}_n^{\circ}\left(x, \theta\right)$.

(ii). The $RSS = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{m}_n\left(x_k; \theta\right))^2$ by the corresponding robust counterpart RSS_{robust}. Let $\xi = L(e)$ be a function of a random variable e. A realization of the random variable e is given by $e_k = (y_k - \hat{m}_n\left(x_k; \theta\right))$, $k = 1, ..., n$, and the $\frac{1}{n}RSS = J_1(\theta)$ can be written as a location problem

$$J_1(\theta) = \frac{1}{n}\sum_{k=1}^{n} L(e_k) = \frac{1}{n}\sum_{k=1}^{n} \xi_k, \tag{11.29}$$

where $\xi_k = e_k^2$, $k = 1, ..., n$. Using a robust analog of the sum ($(0, \beta_2)$ - trimmed mean), the robust $J_1(\theta)$ is defined by

$$J_1^{robust}(\theta) = \frac{1}{n - \lfloor n\beta_2 \rfloor} \sum_{k=1}^{n - \lfloor n\beta_2 \rfloor} \xi_{n(k)}, \tag{11.30}$$

where $\xi_{n(1)}, ..., \xi_{n(n)}$, $e_k = (y_k - \hat{m}_n^{\circ}\left(x_k; \theta\right))$ and $\hat{m}_n^{\circ}(x_k, \theta)$ is a weighted representation of the function estimate.

(iii). The variance estimator $\hat{\sigma}_e^2$ by the corresponding robust counterpart $\hat{\sigma}_{e,robust}^2$. Consider the NARX model (Ljung, 1987)

$$\hat{y}\left(t\right) = f\left(y\left(t-1\right), ..., y(t-q), u(t-1), ..., u(t-p)\right). \tag{11.31}$$

In practice, it is usually the case that only the ordered observed data $y(k)$ according to the discrete time index k, are known. The variance estimator suggested by (Gasser *et al.*, 1986) is used

$$\hat{\sigma}_e^2\left(y\left(t\right)\right) = \frac{1}{n-2}\sum_{t=2}^{n-1} \frac{(y(t-1)a + y(t+1)b - y(t))^2}{a^2 + b^2 + 1} \tag{11.32}$$

where $a = \frac{y(t+1)-y(t)}{y(t+1)-y(t-1)}$ and $b = \frac{y(t)-y(t-1)}{y(t+1)-y(t-1)}$. Let $\zeta = L(\vartheta)$ be a function of a random variable, a realization of the random variable ϑ is given by

$$\vartheta_k = \frac{(y(t-1)a + y(t+1)b - y(t))}{\sqrt{a^2+b^2+1}}. \tag{11.33}$$

The variance estimator (11.32) can now be written as an average of the random sample $\vartheta_1^2, ..., \vartheta_n^2$ (a location problem):

$$\hat{\sigma}_e^2 = \frac{1}{n-2} \sum_{k=2}^{n-1} \zeta_k, \tag{11.34}$$

where $\zeta_k = \vartheta_k^2$, $k = 2, ..., n-1$. Using a robust analog of the sum ($(0, \beta_2)$ - trimmed mean), the robust $\hat{\sigma}^2_{e,robust}$ is defined by

$$\hat{\sigma}^2_{e,robust} = \frac{1}{m - \lfloor m\beta_2 \rfloor} \sum_{l=1}^{m-\lfloor m\beta_2 \rfloor} \zeta_{n(l)}, \tag{11.35}$$

where $m = n-2$.

The final robust FPE criterion is given by

$$J_C(\theta)_{robust} = J_1(\theta)_{robust} + \left(1 + \frac{2[\mathrm{tr}(S^*(v_k,\hat{\theta}))+1]}{n - \mathrm{tr}(S^*(v_k,\hat{\theta})) - 2}\right) \hat{\sigma}^2_{e,robust} \tag{11.36}$$

where the smoother matrix $S^*(v_k, \hat{\theta})$ is now based on the weighting elements v_k.

11.7.2 Influence function of the Robust Final Prediction Error (FPE) criterion

Because $J_C(\theta)$ is based on least squares estimation (via $L\left(y_k, \hat{f}(x_k;\theta)\right), \hat{\sigma}_e^2$), it is very sensitive to outliers and other deviations from the normality assumption on the error distribution. The influence function of the Final Prediction Error (FPE) criterion is unbounded in $\mathbb{R}$.

The corresponding statistical functional for the robust FPE criterion is given by

$$T(F) = \frac{1}{1-\beta_2} \int_0^{F^-(1-\beta_2)} \xi dF(\xi) + M^{robust} \left(\frac{1}{1-\beta_2} \int_0^{F^-(1-\beta_2)} \zeta dF(\zeta) \right), \tag{11.37}$$

where $M^{robust} = \left(1 + \frac{2[\mathrm{tr}(S^*(v_k,\hat{\theta}))+1]}{n-\mathrm{tr}(S^*(v_k,\hat{\theta}))-2}\right)$. ¿From the definition of the influence function and (11.37) , it follows that

$$IF(\xi, \zeta; T, F) = IF(\xi; T_1, F) + m^{robust} \left(IF(\zeta; T_2, F) \right), \tag{11.38}$$

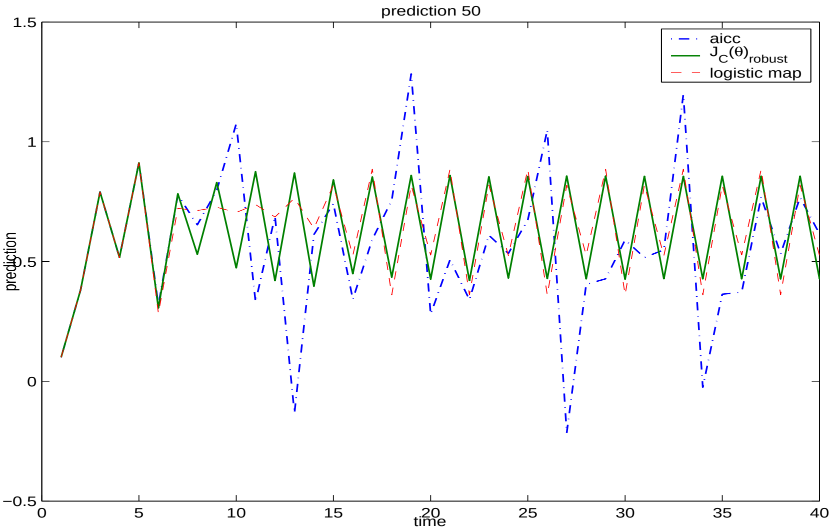

Figure 11.18: The logistic map with contaminated equation noise. Time plot of the validation data and its iterative predictions using mentioned LS-SVM estimators.

where

$$IF\left(\xi;T_1,F\right)=\begin{cases}\frac{\xi-\beta_2F(1-\beta_2)}{(1-\beta_2)}-T_1(F) & 0\leq\xi\leq F^-\left(1-\beta_2\right)\\ F^-\left(1-\beta_2\right)-T_1(F) & F^-\left(1-\beta_2\right)<\xi\end{cases}\qquad(11.39)$$

and the influence function $IF\left(\zeta;T_2,F\right)$ can be calculated. We can see that the influence functions are bounded in $\mathbb{R}$. This means that an added observation at a large distance from $T(\hat{F}_n)$ gives a bounded value in absolute sense for the influence functions.

11.7.3 Illustrative examples

Example 1: Artificial example

An example illustrates the advantage of the proposed criteria. It considers a stochastic version of the logistic map $y(t+1)=cy(t)\left(y(t)-1\right)+e_t$ with $c=3.55$ and contaminating process noise e_t. The recurrent prediction illustrates the difference of the NAR model based on the LS-SVM tuned by AICC and the weighted LS-SVM tuned by $J_C(\theta)_{robust}$ (see Fig 11.18).

The numerical test set performances of both are shown in Table 11.4 with improved results (in L_2, L_1, L_∞ norm) by applying the robust model selection criteria.

Method	L_2	L_1	L_∞
AICC+LS-SVM	0.1088	0.2251	1.0338
$J_C(\theta)_{robust}$+WLS-SVM	0.0091	0.0708	0.2659

Table 11.4: Performance of kernel based NARX model and its robust counterpart on time-series generated by logistic map with contaminated equation noise.

Example 2: Process data

The process is a liquid-satured steam heat exchanger, where water is heated by pressurized saturated steam through a copper tube. The output variable is the outlet liquid temperature. The input variables are the liquid flow rate, the steam temperature, and the inlet liquid temperature. In this experiment the steam temperature and the inlet liquid temperature are kept constant to their nominal values (See dataset 97-002 of DaISy, (De Moor 1998)). A number of different models are tried on the dataset (for$t = 1, ..., 3000$). A final comparison of the estimated models is done and measured on an independent testset (for $t = 3001, ..., 4000$)

1. At first, classical linear tools were used to explore properties of the data. An appropriate ARX model was selected using AICC and GCV. Although its one-step ahead predictor is excellent, iterative predictions are bad because of the overestimation of the orders of the data.

2. A way to overcome this, is to use robust methods for the parameter estimation in ARX modeling (based on Huber's cost function) and the robust counterparts of AICC and GCV. Although the performance in the one-step-ahead prediction on the test set is slightly worse compared to the previous non-robust models, the iterative prediction is the iterative prediction based on robust procedures outperforms the non-robust methods. Note that small orders are selected by the robust procedures.

3. Thirdly, the fixed-size LS-SVM (RBF kernel) was used for model identification of a NARX model (Table 11.5). The performance in the one-step-ahead prediction on test data is slightly worse compared to the linear models, while in the iterative prediction the fixed-size LS-SVM outperforms (in the L_1, L_1 and L_∞ norm) the linear models. The best results (L_1, L_1 and L_∞ norm) are obtained by robust selection criteria combined with a robust fixed-size LS-SVM.

11.8 Conclusions

Cross-validation methods are frequently applied for selecting tuning parameters in neural network methods, usually based on L_2 or L_1 norms. However, due to the asymmetric and non-Gaussian nature of the score function, better location parameters can be used to estimate the performance. In this thesis we have introduced a repeated robust cross-validation score function method by applying

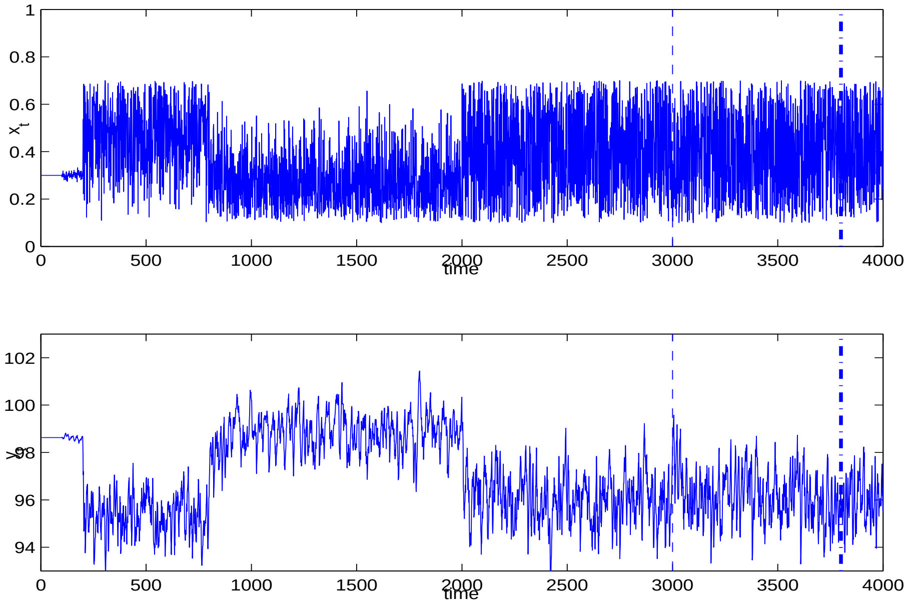

Figure 11.19: The process is a liquid-saturated steam heat exchanger, where water is heated by pressurized saturated steam through a copper tube.

		One-step-ahead			**Iterative**		
Model	**Selection**	L_2	L_1	L_∞	L_2	L_1	L_∞
Linear Models							
$J_C = 0.052$, $GCV = 0.059$							
ARX(40,40,8)	AICC	0.0823	0.226	1.10	3.04	1.42	4.48
ARX(40,18,8)	GCV	0.0819	0.226	1.10	0.722	0.708	2.28
Fixed-size LS-SVM (RBF) based Models							
$J_C = 0.048$, $GCV = 0.054$							
NARX(3,4,8)	AICC	0.0967	0.250	1.12	0.560	0.609	2.16
NARX(3,4,8)	GCV	0.0967	0.250	1.12	0.560	0.609	2.16
Robust fixed-size LS-SVM (RBF) based models							
$J_C^{robust} = 0.024$, $GCV^{robust} = 0.030$							
rNARX(3,3,8)	robust AICC	0.0958	0.246	1.11	0.501	0.589	2.09
rNARX(2,1,8)	robust GCV	0.0914	0.245	1.08	0.496	0.566	1.99

Table 11.5: Numerical performance measure on test data for the results of the process dataset. The results compare the performance of linear ARX models with NARX using fixed-size LS-SVM (RBF kernel) tuned by different model selection criteria (AICC, GCV) and its robust counterparts based on robust fixed-size LS-SVM and robust complexity criteria. Again, the robust procedures outperform the classical methods in the iterative prediction.

concepts from robust statistics to the cross-validation methodology. We have applied a similar technique to generalized cross-validation. Simulation results illustrate that these methods can be very effective, especially with outliers on data where the L_2 methods usually fails. The proposed methods have a good robustness / efficiency trade-off such that they perform equally well in cases where L_2 would perform optimally.

We have proposed robust estimation and robust model selection techniques for the use of least squares support vector machines with nonlinear ARX models. Robust techniques have been proposed for fixed-size LS-SVMs in the primal space as well as for the dual problem. Several examples illustrate that these methods can further improve standard non-robust techniques in the case of outliers and non-Gaussian noise distributions.

Chapter 12

Robust Prediction Intervals

In this chapter we give some definitions concerning confidence intervals and prediction intervals (Casella and Berger, 1990; Shao, 1999). Next, we discuss methods of constructing prediction intervals. Finally, we introduce robust prediction intervals for LS-SVM based on a robust external bootstrap method. Contributions are made in Section 12.3.

12.1 Definitions

Let $X = (x_1, ...x_n)$ be random variables with unknown joint distribution $F \in \mathcal{F}$ depending on a real-valued parameter $T(F)$ and let $C(X)$ denote a confidence set for $T(F)$. If

$$\inf_{F \in \mathcal{F}} \text{Prob}\left(T(F) \in C(X)\right) \geq q \tag{12.1}$$

where $q = 1 - \alpha$ and α is a fixed constant in $(0, 1)$. Then $C(X)$ is called a *confidence set* for $T(F)$ with coverage probability q. The concept of confidence sets can be extended to the case where $T(F)$ is a vector of m real-valued parameters and $C(X)$ is called a *confidence region.* If $C(X) = [L(X), U(X)]$ for a pair of statistics L and U, then $C(X)$ is called a *confidence interval.* If $C(X) = (-\infty, U(X)]$ or $[L(X), \infty)$, then L (or U) is called an upper (respectively a lower) *confidence bound* for $T(F)$. Let $T_s(F)$ denote a real-valued parameter with $s \in \mathcal{S}$ where $\mathcal{S}$ is an index set that may contain infinitely many elements, and let $C_s(X)$ be a class of confidence intervals. If

$$\inf_{F \in \mathcal{F}} \text{Prob}\left(T_s(F) \in C_s(X) \text{ for all } s \in \mathcal{S}\right) \geq q \tag{12.2}$$

then $C_s(X)$, $s \in \mathcal{S}$, are level $q = 1 - \alpha$ *simultaneous confidence intervals* or *confidence bands* for $T_s(F)$.

To fix the ideas, consider the following example in which it is desired to estimate a confidence interval, a tolerance interval and a standard error bar of a parameter. Let $x_1, ...x_n$ be a random sample from a the normal distribution $\mathcal{N}\left(\mu, \sigma^2\right)$ with unknown mean μ and unknown variance σ^2, and let $\bar{X}$ and s^2 be the sample mean and sample variance. A $(1-\alpha)\,100$ percent confidence interval on μ is given by

$$\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}},$$

where $t_{\frac{\alpha}{2},n-1}$ is the upper $\frac{\alpha}{2}$ percentage point of the t distribution with $n-1$ degrees of freedom. This means that in using $\bar{X}$ to estimate μ, the error $e = \left|\bar{X}-\mu\right|$ is less or equal to $t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$ with confidence $(1-\alpha)\,100$. Unlike the confidence interval, which estimates the range in which a population parameter falls, the tolerance interval estimates the range which should contain a certain percentage of each individual observation within the population. In practical situations, a bound that covers 95% of the population is given by

$$\bar{X} \pm cs\left(\bar{X}\right),$$

where c is determined such that one can state with confidence $(1-\alpha)\,100$ percent that the limits contain at least a proportion p of the population.
Finally, the standard error bar defined by

$$\bar{X} \pm s\left(\bar{X}\right),$$

gives some idea about the precision of $\bar{X}$.

Definition 52 *Let ξ be a random variable and suppose that the distribution of ξ is related to the distribution of a sample X from which prediction will be made. For instance, $X = (x_1, ...x_n)$ is the observed sample and $\xi = X_{n+1}$ is to be predicted, where $x_1, ...x_n, x_{n+1}$ are (i.i.d.) random variables. A set $C(X)$ is said to be a level $q = 1-\alpha$ prediction set for ξ if*

$$\inf_{F\in\mathcal{F}} Prob\left(\xi \in C(X)\right) \geq q \tag{12.3}$$

where F is the joint distribution of ξ and X. A prediction interval for an unobserved random variable ξ based on the observed data X is a random interval $C(X) = [L(X), U(X)]$.

Note the similarity in the definitions of a prediction interval and a confidence interval. The difference is that a prediction interval is an interval on a random variable, rather than a parameter. Methods for constructing confidence sets are for example: pivotal quantities, inverting acceptance regions of tests, the statistical method (Mood, Graybill and Boes, 1974), Bayesian approach (credible sets), confidence sets based on likelihood, invariant intervals (Berger, 1985), (lehmann, 1986) and bootstrap confidence sets.

12.2 Construction of prediction intervals

Perhaps one of the most popular methods of constructing prediction sets is the use of pivotal quantities (Barnard, 1949, 1980), defined as

Definition 53 *Let $X = (x_1, ...x_n)$ be random variables with unknown joint distribution $F \in \mathcal{F}$, and let $T(F)$ denote a real-valued parameter. A random variable $\mathcal{J}(X, T(F))$ is a pivotal quantity (or pivot) if the distribution of $\mathcal{J}(X, T(F))$ is independent of all parameters.*

Suppose that $x_1, ..., x_n$ are *i.i.d.* normal random variables with unknown mean and variance μ and σ^2, respectively. The random variable

$$\mathcal{J}(x, \mu) = \frac{\sqrt{n}\,(\bar{x} - \mu)}{s} \sim t_{n-1},$$

where s is an estimate of the square root of $Var\,[e_k] = \sigma^2$, has a t distribution with $(n-1)$ degrees of freedom. This distribution is independent of both μ, σ^2 and $\frac{\sqrt{n}(\bar{x}-\mu)}{\hat{\sigma}}$ is a pivot for μ. The random variable

$$\mathcal{J}(x, \sigma^2) = \frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{n-1},$$

where s^2 is an estimate of $Var\,[e_k] = \sigma^2$, has a χ^2 distribution with $(n-1)$ degrees of freedom. This distribution is independent of both μ, σ^2 and $\frac{(n-1)s^2}{\sigma^2}$ is a pivot for σ^2. To obtain confidence intervals for μ and σ^2, one can use the respectively pivots. In many problems, it is difficult to find exact pivots or to determine the distribution of an exact pivot if it does exist. However, in these cases, it is often possible to find an approximate pivot (approximate pivots are justified via asymptotic arguments).

Next, the given examples illustrate inference for linear parametric models and nonparametric models. The development of procedures for obtaining confidence intervals (for regression parameters) and prediction intervals (for new outputs of the regression model) requires that one assumes the errors to be normally and independently distributed with mean zero and variance σ^2. For nonparametric models one can obtain statistical inference based on bootstrap procedures.

Example 54 *Consider the multiple (several independent variables) linear regression with d input variables and assume that the Gauss-Markov conditions ($E\,[e_k] = 0$, $E\left[(e_k)^2\right] = \sigma^2 < \infty$ and $E\,[e_j, e_i] = 0$, $\forall i \neq j$), hold. One has a relationship of the form*

$$y_k = w_0 + \sum_{j=1}^{d} w_j x_k^{(j)} + e_k, \quad k = 1, ..., n, \; j = 1, ..., d,$$

where the unknown parameter vector $w = (w_0, w_1, ..., w_d)^T$ is assumed to be fixed and e_k are i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$. The least squares estimate of $w = (w_0, w_1, ..., w_d)^T$ is defined to be those values $\hat{w}_0, \hat{w}_1, ..., \hat{w}_d$ such that $\hat{w}_0 + \sum_{j=1}^{d} \hat{w}_j x_k^{(j)}$ minimizes *the residual sum of squares. Inferences regarding the regression parameters, under the assumption of normality, using a pivotal quantity*

$$\mathcal{J}(y, w_i) = \frac{\hat{w}_i - w_i}{s.e(\hat{w}_i)} \sim t_{n-(d+1)}, \quad i = 0, 1, ..., d,$$

for with one obtains the following confidence interval estimators with confidence coefficient $(1-\alpha)$*:*

$$\hat{w}_i \pm s.e(\hat{w}_i) t_{(n-d+1),\alpha/2}, \quad i = 0, 1, ..., d,$$

where $s.e(\hat{w}_i)$ *is the square root estimator of* $Var\left[\hat{w}_i\right]$*,* $i = 0, 1, ...d$ *and* $t_{(n-d+1),\alpha/2}$ *is the* $(1-\alpha/2)th$ *quantile of the* t*- distribution. A regression model can be used to predict future observations on the output variable* y *corresponding to particular values of the* d *input variables, for example* $x^{(1)}, ..., x^{(d)}$*. The point estimate for the future observation* y_0 *at the point* $x_0 = \left(x_0^{(1)}, ..., x_0^{(d)}\right)^T$ *is computed from equation* $\hat{y}(x_0) = \hat{w}^T x_0$*. A pivotal quantity is given by*

$$\mathcal{J}(y, y_0) = \frac{y_0 - \hat{w}^T x_0}{\hat{\sigma}\sqrt{1 + x_0^T \left(X^T X\right)^{-1} x_0}} \sim t_{n-d+1},$$

where $\hat{\sigma}$ *is an estimate of the square root of* $Var\left[e_k\right] = \sigma^2$*. The random variable* $\mathcal{J}(y, y_0)$ *has a t distribution* t_{n-d+1}*, this is because* y_0 *and* $\hat{w}^T x_0$ *are independently normal,* $(n-d+1)\,\hat{\sigma}^2$ *has a chi-square distribution* χ^2_{n-d+1} *and* y_0*,* $\hat{w}^T x_0$ *and* $\hat{\sigma}^2$ *are independent. A level* $1-\alpha$ *prediction interval for* y_0 *is then*

$$\hat{y}(x_0) \pm t_{\alpha/2, n-d+1} \hat{\sigma} \sqrt{1 + x_0^T \left(X^T X\right)^{-1} x_0}.$$

Suppose one is interested in several prediction intervals constructed from the same data. Such intervals are called simultaneous prediction intervals or prediction bands. If one sets up m *intervals as above, each at level* $1-\alpha$*, the overall inference will not be at the* $1-\alpha$ *level. A good solution is to use Bonferroni inequality (Sen and Srivastava, 1990).*

Example 55 *The linear regression model of example 61 provides a flexible framework. However, linear regression models are not appropriate for all situations. There are many situations where the output variable and the input variables are related through a known nonlinear function. Suppose one has a nonlinear relationship of the form*

$$y_k = m\left(x_k; w\right) + e_k, \quad k = 1, ..., n,$$

where the e_k *are i.i.d.* $\mathcal{N}\left(0, \sigma^2\right)$*,* x_k *is a* $(n \times 1)$ *vector and* $w \in \mathbb{R}^d$*. Assume that the Gauss-Markov conditions (*$E\left[e_k\right] = 0$*,* $E\left[e_k^2\right] = \sigma^2 < \infty$ *and* $E\left[e_j, e_i\right] = 0$*,* $\forall i \neq j$*), hold. To find the least-squares estimates, one must differentiate* $\sum_{k=1}^N \left(y_k - m\left(x_k, w\right)\right)^2$ *with respect to* w*. This will provide a set of*

d normal equations for the nonlinear situation. In a linear regression model, when the errors are normally and independently distributed, exact confidence intervals and prediction intervals based on t and F distributions are available. The least-squares parameter estimates have attractive statistical properties (unbiasedness, minimum variance and normal sampling distributions). However, this is not the case in nonlinear regression, even when the errors are i.i.d. statistical inference in nonlinear regression depends on large-sample or asymptotic results. The large-sample theory generally applies both for normally and non-normally distributed output variables. Using the asymptotic linearization of $y_k = m(x_k, w) + e_k$, $k = 1, ..., n$, one can apply existing linear methods to finding a prediction interval for y at $x = x_0$. The point estimate for the future observation y_0 at the point $x_0 = \left(x_0^{(1)}, ..., x_0^{(d)}\right)^T$ is computed from equation $\hat{y}(x_0) = m(x_0, w)$. An asymptotic pivotal quantity is given by

$$\mathcal{J}(y, y_0) = \frac{y_0 - m(x_0; \hat{w})}{\hat{\sigma}\sqrt{1 + g^T(x_0; \hat{w})\left(G^T G\right)^{-1} g(x_0; \hat{w})}} \sim t_{n-p},$$

where $\hat{\sigma}$ is an estimate of the square root of $Var[e_k] = \sigma^2$, $g^T(x_0; w) = \left(\frac{\partial m(x_0;w)}{\partial w_1}, ..., \frac{\partial m(x_0;w)}{\partial w_q}\right)$ and $G = \left[\left(\frac{\partial m(x_k;w)}{\partial w_j}\right)\right]$, $k = 1, ..., n$, $j = 1, ..., d$. The random variable $\mathcal{J}(y, y_0)$ has a t distribution t_{n-d}, under appropriate regularity conditions and for large n. An approximate $1 - \alpha$ level prediction interval for y_0 is then

$$\hat{y}(x_0) \pm t_{\alpha/2, n-d}\hat{\sigma}\sqrt{1 + g^T(x_0; \hat{w})\left(G^T G\right)^{-1} g(x_0; \hat{w})}.$$

The construction of prediction bands is discussed by (Khorasani and Milliken, 1982). Prediction intervals for multilayer perceptrons, a class of nonlinear models, can be obtained based on asymptotic results. Examples are find in (Hwang and Ding, 1997) and (De Veaux et al.*, 1998).*

Example 56 *It should be clear that in dealing with the linear and nonlinear regression models of the two previous examples, the normal distribution played a central role. Inference procedures for both linear and nonlinear regression models in fact assume that the output variable follows the normal distribution. There are a lot of practical situations where this assumption is not going to be even approximately satisfied. The generalized linear model (GLM) was developed to allow us to fit regression models for output data ($y \in \mathbb{R}^+, y \in \mathbb{N}$ or $y \in \{0, 1\}$) that follows a general distribution called the exponential family. The GLM is given by*

$$\mu_k = g(E[y_k | X = x_k]) = x_k^T w, \ k = 1, ..., n,$$

where x_k is a vector of input variables for the kth observation and w is a vector of parameters. Every GLM has three components: (a) an output variable distribution (error structure), (b) input variables, and (c) a link function (e.g.

logistic link, log link, probit link). For further details on the structure of GLM (see McCullagh and Nelder, 1987). An important member of the family of GLM is the logistic regression defined as

$$\pi(x_k) = \frac{1}{1+\exp\left(-x_k^T w\right)},\ k = 1,...,n,$$

where the term $x_k^T w = w_0 + \sum_{j=1}^{d} w_j x_k^{(j)}$*. Note that* $0 \leq \pi(x_k) \leq 1$*. One can associate a* $1-\alpha$ *level prediction interval on* $\pi(x_0)$ *through a prediction interval (Wald inference) on* $x_0^T w$*,* $x_0^T w \pm z_{\alpha/2}\sqrt{1 + x_0^T \left(X^T V X\right)^{-1} x_0}$ *with* $Var[\hat{w}] = \left(X^T V X\right)^{-1}$*. The point estimate for the future observation* y_0 *at the point* $x_{01}, ..., x_{0d}$ *is computed from equation* $\hat{\pi}(x_0) = \frac{1}{1+\exp\left(-x_0^T w\right)}$*. An asymptotic pivotal quantity is given by*

$$\mathcal{J}(y, \pi(x_0)) = \frac{y_0 - \hat{\pi}(x_0)}{(\pi(x_0))(1-\pi(x_0))\sqrt{1 + x_0^T \left(X^T V X\right)^{-1} x_0}} \sim \mathcal{AN}(0,1),$$

where an asymptotic standard normal distribution is denoted by $\mathcal{AN}(0,1)$*. An approximate* $1-\alpha$ *level prediction interval for* y_0 *is then*

$$\hat{\pi}(x_0) \pm z_{\alpha/2}\left[(\pi(x_0))(1-\pi(x_0))\right]\sqrt{1 + x_0^T \left(X^T V X\right)^{-1} x_0}.$$

Example 57 *An alternative approach to estimation is to use a biased estimation method. These methods of estimation are based on trading off bias for variance. Principal Component regression, Ridge regression (Hoerl and Kennard, 1970) and the shrinkage estimator (Stein, 1956) and (James and Stein, 1961) are three of the several methods belonging to the class of biased estimators. For example, Ridge regression modifies the method of least squares to allow biased estimators of the regression coefficients. Given the data* $(x_1, y_1), ..., (x_n, y_n)$*, for ordinary least squares the solution vector for the normal equations is given by* $\hat{w}_{ols} = \left(X^T X\right)^{-1} X^T Y$ *while the Ridge regression estimator is given by*

$$\hat{w}_{ridge} = \left(X^T X + cI\right)^{-1} X^T Y.$$

The constant c reflects the amount of bias in the estimators. When $c = 0$*,* $\hat{w}_{ridge}$ *reduces to the ordinary least squares* $\hat{w}_{ols}$*. When* $c > 0$*, the ridge regression coefficients are biased but tend to be more stable than ordinary least squares estimators. A limitation of Ridge regression is that ordinary inference procedures are not applicable and exact distributional properties are not known. One can obtain confidence intervals and prediction intervals by using bootstrapping, which is discussed in Section 5.*

Example 58 *The previous examples are parametrized by an Euclidean parameter vector. With nonparametric regression, the simplest type of semiparametric models, the regression equation is determined from the data. In this case,* *one relaxes the assumption* $e_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ *and inferential procedures are not strictly applicable, since those are based on the assumption of normal errors. One possible approach is to obtain prediction intervals by using bootstrapping.*

12.3 Robust Prediction Intervals

Methods to establish confidence intervals or prediction intervals are based on the principle of first estimating $m(x)$ by an initial estimator $\hat{m}_n(x)$ and then estimating the distribution of $m(x) - \hat{m}_n(x)$. In the statistical literature a distinction is made between pivotal and nonpivotal methods. Hall (1992) pointed out that pivotal methods, for the problem of bootstrap prediction intervals, should be preferred to nonpivotal methods. The main problem with prediction intervals in nonparametric regression rests on the fact that a consistent estimator of $m(x)$ is necessarily biased (Neumann, 1995). Regarding bias correction, there are two commonly used methods to deal with the bias of the initial estimator $\hat{m}_n(x)$, undersmoothing and explicit bias correction on the basis of yet another kernel estimator. In Hall (1991, 1992) it is shown that the undersmoothing methods leads to better results.

12.3.1 Weighted LS-SVM for robust function estimation

Smoother matrix for prediction

We focus on the choice of an RBF kernel $K(x_k, x_l; h) = \exp\left\{- \|x_k - x_l\|_2^2 / h^2\right\}$. In matrix form, let $\theta = (h, \gamma)^T$ and for all new input data defined as $\mathcal{D}_{x,test} = \{x : x_l^{test} \in \mathbb{R}^d, l = 1, ..., s\}$:

$$\begin{aligned}\hat{m}_n\left(x^{test}; \theta\right) &= \Omega^{test}\hat{\alpha}^{train} + 1_n \hat{b}^{train} \\ &= \left[\Omega^{test}\left(Z^{-1} - Z^{-1}\frac{J_{nn}}{c}Z^{-1}\right) + \frac{J_{sn}}{c}Z^{-1}\right] y \\ &= S(x^{test}, x^{train}; \theta) y, \end{aligned} \tag{12.4}$$

where $c = 1_n^T \left(\Omega^{train} + \frac{1}{\gamma} I_n\right)^{-1} 1_n$, $Z = (\Omega^{train} + \frac{1}{\gamma} I_n)$, J_{nn} is a square matrix with all elements equal to 1, J_{sn} is a $s \times n$ matrix with all elements equal to 1, $y = (y_1, \ldots, y_n)^T$ and $\hat{m}_n\left(x^{test}; \theta\right) = (\hat{m}_n(x_1^{test}; \theta), \ldots, \hat{m}_n(x_s^{test}; \theta))^T$. The LS-SVM for regression corresponds to the case with $\hat{m}_n\left(x^{test}; \theta\right)$ defined by (12.4) and

$$S(x^{test}, x^{train}; \theta) = \Omega^{test}\left(Z^{-1} - Z^{-1}\frac{J_{nn}}{c}Z^{-1}\right) + \frac{J_{sn}}{c}Z^{-1}. \tag{12.5}$$

where $\Omega_{k,l}^{test} = K\left(x_k^{train}, x_l^{test}\right)$ are the elements of the $s \times n$ kernel matrix and $\Omega_{k,l}^{train} = K\left(x_k^{train}, x_l^{train}\right)$ are the elements of the $n \times n$ kernel matrix.

12.3.2 Robust bootstrap

Contamination of a sample is an problem which can become worse when the bootstrap is applied, because some resamples may have a higher contamination level then the initial sample. Bootstrapping using robust function estimators, $\hat{m}_n^{\diamond}(x_k)$, may be a solution to this problem but it can lead to several complications (Stromberg, 1997) and (Singh, 1998). Stromberg (1997) studies alternative bootstrap estimates of variability for robust estimators. Singh (1998) suggests a robustification of bootstrap via winsorization. Hall and Presnell (1999) suggest a general approach, using a version of the weighted bootstrap, The method depends on measures of dispersion and on the distance between distributions. Salibian-Barrera and Zamar (2000) introduce a robust bootstrap based on a weighted representation for MM-regression estimates. The robustification of the bootstrap in this thesis is to introduce a control mechanism in the resampling plan, consisting of an alteration of the resampling probabilities, by identifying and downweighting those data points that influence the function estimator.

Given a random sample $\{(x_1, y_1), ..., (x_n, y_n)\}$ with common distribution F. To allow for the occurrence of outliers and other departures from the classical model we will assume that the actual distribution F of the data belongs to the contamination neighborhoud

$$F_\epsilon = \{F : F = (1-\epsilon)\, G + \epsilon H, \quad H \text{ arbitrary}\} \tag{12.6}$$

where $0 \leq \epsilon < 0.5$. For each pair (x_k, y_k) in the sample define the residuals as $\hat{e}_k = y_k - \hat{m}_n(x_k)$. Based on the residuals define the weights as

$$v_k = \vartheta\left(\frac{\hat{e}_k}{\hat{s}}\right) \tag{12.7}$$

where $\vartheta(.)$ is some function and $\hat{s}$ is a robust scale estimator. For instance, one could apply hard rejection or smooth rejection of outliers (see Suykens *et al.*, 2002). Let the resampling plan of the uniform bootstrap be denoted by $p_{unif} = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$ and let $p = (p_1, ..., p_n)$ be the resampling plan of the weighted bootstrap with mass p_k on $\hat{e}_k$. Let m be the number of data points with $(v_k \neq 1)$ and $\sum_{k=1}^{n} p_k = 1$. The mass p_l, $l = 1, ...n - m$, is given by

$$p_l = \frac{1}{n} + \frac{\sum_{i=1}^{m} \frac{1}{n}(1 - v_i)}{n - m}, \quad l = 1, ..., n - m \quad ; i = 1, ..., m \tag{12.8}$$

and the mass p_j, $j = 1, ..., m$, is given by

$$p_j = \left(1 - \sum_{l=1}^{n-m} p_l\right)\left(1 - \frac{v_j}{\sum_{j=1}^{m} v_l}\right), \quad j = 1, ..., m. \tag{12.9}$$

Algorithm 59 *(Robust external (wild) bootstrap)*

(i) *The unknown probability model P was taken to be $y_k = m(x_k) + e_k$, with $e_1, ..., e_n$ independent errors drawn from some unknown probability distribution F_e.*

(ii) *Calculate $\hat{m}_n(x_k)$, and the estimated errors (residuals) are$\hat{e}_k = y_k - \hat{m}_n(x_k)$, from which was obtained an estimated version of $\hat{F}_e : p = (p_1, ..., p_n)$ and p_k is given by (12.8 and 12.9).*

(iii) *Draw the bootstrap residuals $\hat{e}_k^*$ from a two-point centered distribution in order that its second and third moment fit the square and the cubic power of the residual $\hat{e}_k$. For instance, the distribution of $\hat{e}_k^*$ could be $\eta\delta_{[a\hat{e}_k]} + (1-\eta)\,\delta_{[b\hat{e}_k]}$ with $\eta = \frac{5+\sqrt{5}}{10}$, $a = \frac{1-\sqrt{5}}{2}$, $b = \frac{1+\sqrt{5}}{2}$ and $\delta_{[x]}$ being the Dirac measure at x Alternatively, one can choose $\hat{e}_k^*$distributed as $\hat{e}_k^* = \hat{e}_k\left(\frac{Z_1}{\sqrt{2}} + \frac{Z_2^2-1}{2}\right)$, with Z_1 and Z_2 being two independent standard Normal random variables, also independent of $\hat{e}_k$.*

(iv) *Having generated $\{y_k^*\}_{k=1}^n$, calculate the bootstrap estimates $\hat{f}^*(x_k)$.*

(v) *This whole process must be repeated B times.*

12.3.3 Computing robust prediction intervals

The prediction problem for nonparametric function estimation falls in two parts, the first being construction of a prediction interval (based on a pivotal method) for the expected value of the estimator and the second involving bias correction (undersmoothing). Given a LS-SVM function estimator $\hat{m}_{n,h}(x_0)$, where x_0 is a new input data point, prediction intervals are constructed by using the asymptotic distribution of a pivotal statistic. Let $\mathcal{J}(m(x_0), \hat{m}_{n,h}(x_0))$ be a pivotal statistic defined as

$$\mathcal{J}(m(x_0), \hat{m}_{n,h}(x_0)) = \frac{\hat{m}_{n,h}(x_0) - m(x_0) - B(x_0)}{(V(x_0))^{\frac{1}{2}}}, \tag{12.10}$$

where $B(x_0)$ is the bias and $V(x_0)$ is the variance of the LS-SVM function estimator $\hat{m}_{n,h}(x_0)$. The asymptotic pivot $\mathcal{J}(m(x_0), \hat{m}_{n,h}(x_0))$ can not be used for practical computations of prediction intervals because both $B(x_0)$ and $V(x_0)$ are unknown. We consider an alternative method that consists in estimating the distribution of the pivot

$$\mathcal{T}(m(x_0), \hat{m}_{n,h}(x_0)) = \frac{\hat{m}_{n,h}(x_0)(x_0) - m(x_0)}{\left(\hat{V}(x_0)\right)^{\frac{1}{2}}} \tag{12.11}$$

by an external bootstrap method. One approximate the distribution of the pivotal statistics $\mathcal{T}(m(x_0), \hat{m}_{n,h}(x_0))$ by the corresponding distribution of the bootstrapped statistics

$$\mathcal{V}(\hat{m}_{n,g}(x_0), \hat{m}_{n,h}^*(x_0)) = \frac{\hat{m}_{n,h}^*(x_0) - \hat{m}_{n,g}(x_0)}{\left(\hat{V}^*(x_0)\right)^{\frac{1}{2}}}, \tag{12.12}$$

where $*$ denotes bootstrap counterparts.

A natural approach to robustifying the pivotal $\mathcal{V}(\hat{m}_{n,g}(x_0), \hat{m}^*_{n,h}(x_0))$ is to replace the LS-SVM function estimator by a robust function estimator (the weighted LS-SVM) and replace the variance estimator $\hat{V}^*(x_0)$ by its robust counterpart $\hat{V}^{*\diamond}(x_0)$

$$\mathcal{Z}(\hat{m}^{\diamond}_{n,g}(x_0), \hat{m}^{*\diamond}_{n,h}(x_0)) = \frac{\hat{m}^{*\diamond}_{n,h} - \hat{m}^{\diamond}_{n,g}(x_0)}{\left(\hat{V}^{*\diamond}(x_0)\right)^{\frac{1}{2}}}. \tag{12.13}$$

Given new input data defined as $\mathcal{D}_{x,test}$, robust s simultaneous prediction intervals (applying the Bonferroni method) with asymptotic level $1-\alpha$ are given by

$$I_{\mathcal{Z}} = \left[\hat{m}^{\diamond}_{n,h}(x_0) + \left(\hat{V}^{*\diamond}(x_0)\right)^{\frac{1}{2}} Q_{\alpha/2s},\ \hat{m}^{\diamond}_{n,h} + \left(\hat{V}^{*\diamond}(x_0)\right)^{\frac{1}{2}} Q_{(1-\alpha)/2s}\right], \tag{12.14}$$

where Q_α denote the α-quantile of the bootstrap distribution of the pivotal statistic $\mathcal{Z}(\hat{m}^{\diamond}_{n,g}(x_0), \hat{m}^{*\diamond}_{n,h}(x_0))$.

12.4 Simulation

To illustrate the behavior of the robust prediction intervals proposed in the foregoing sections, we present an example using the following data set. Consider the following nonlinear regression model defined as

$$y_k = \frac{\sin(x)}{x} + e_k,\ k = 1, ..., 250 \tag{12.15}$$

where the values y_k are corrupted by noise with a ϵ-contamination model

$$\mathcal{U}(F_0, \epsilon) = \{F : F(x) = (1-\epsilon) F_0(x) + \epsilon G(x), \quad 0 \leq \epsilon \leq 1\}, \tag{12.16}$$

where $F_0(x)$ is a Normal distribution with parameters $\mathcal{N}(0, 0.2^2)$, $G(x) \sim \mathcal{N}(3, 2^2)$ and $\epsilon = 0.05$. The prediction intervals are shown in Figure (12.1) and Figure (12.2). Figure 12.2 shows the improvements of robust prediction intervals based on the weighted LS-SVM and robust bootstrap techniques in comparison with prediction intervals based on the unweighted LS-SVM and nonrobust bootstrap techniques.

12.5 Conclusions

We have robustified the bootstrap based on a control mechanism in the resampling plan, consisting of an alteration of the resampling probabilities, by identifying and downweighting those data points that influence the function estimator. We have demonstrated the improvements of robust prediction in comparison with prediction intervals.

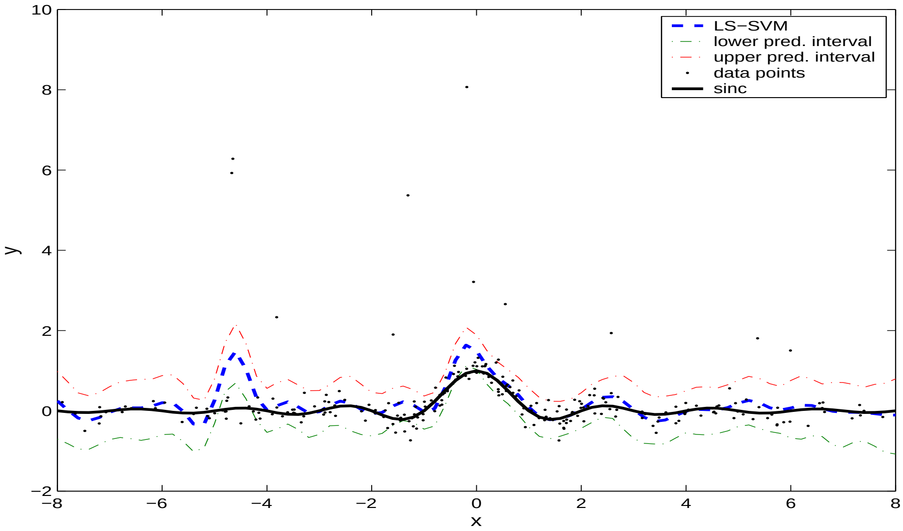

Figure 12.1: Both the estimated regression function (dashed line) and its associated 95% prediction intervals (dashdot lines) were obtained from the LS-SVM regression fit. The intervals are based on bootstrap techniques.

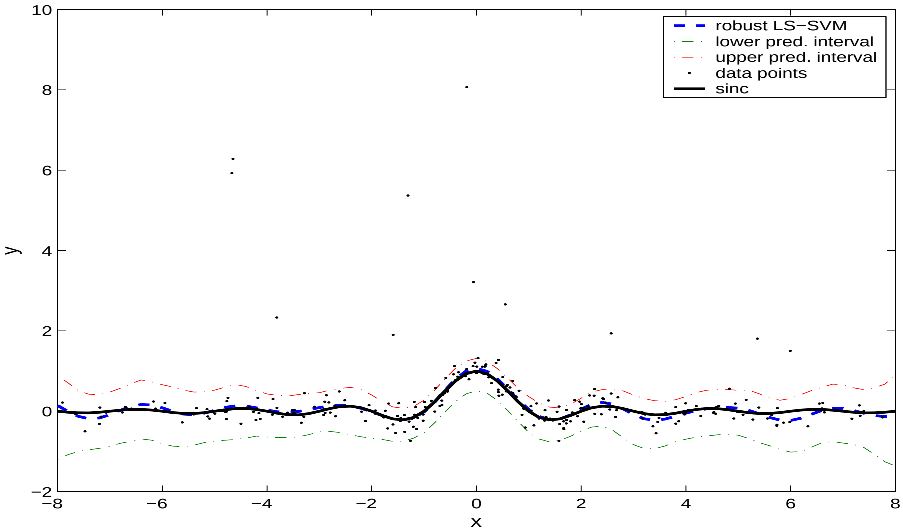

Figure 12.2: Both the estimated regression function (dashed line) and its associated 95% prediction intervals (dashdot lines) were obtained from weighted LS-SVM regression fit. The intervals are based on robust bootstrap techniques.

Chapter 13

Conclusions

In this thesis, we have given an overview of basic techniques for non-parametric regression. In this chapter, we first give a chapter by chapter overview of our contributions and the conclusions. Topics for further research are pointed out in the second section of this chapter.

13.1 Conclusion Summary

The key method in this thesis is least squares support vector machines (LS-SVM), a class of kernel based learning methods that fits within the penalized modelling paradigm. Primary goals of the LS-SVM models are regression and classification. Although local methods (kernel methods) focus directly on estimating the function at a point, they face problems in high dimensions. Therefore, one can guarantee good estimation of a high-dimensional function only if the function is extremely smooth. Additional assumptions (the regression function is an additive function of its components) overcome the curse of dimensionality.

The iterative backfitting algorithm for fitting LS-SVM regression is simple, allowing one to choose a fitting method appropriate for each input variable. Important is that at any stage, one-dimensional kernel regression is all that is needed. Although consistency of the iterative backfitting algorithm is shown under certain conditions, an important practical problem (number of iteration steps) are still left. However the iterative backfitting algorithm (for large data problems) fits all input variables, which is not feasible or desirable when a large number are available. Results show that backfitting LS-SVM (RBF kernel) outperforms the LS-SVM (RBF kernel). Recently we have developed a new method, componentwise LS-SVM, for the estimation of additive models consisting of a sum of nonlinear components (Pelckmans *et al.*, 2004). The method combines the estimation stage with structure detection. Advantages of using componentwise LS-SVMs include the efficient estimation of additive models with respect to classical practice, interpretability of the estimated model, opportunities towards structure detection and the connection with existing statistical techniques.

Model-free estimators of the noise variance are important for doing model selection and setting learning parameters. We have generalized the the idea of the Rice estimator (Rice, 1984) for multivariate data based on U-statistics and differogram models (Pelckmans *et al.*, 2003). We have studied the properties of the LS-SVM regression when relaxing the Gauss-Markov conditions. It was recognized that outliers may have an unusually large influence on the resulting estimate. However, asymptotically the heteroscedasticity does not play any important role.

We proposed a non-parametric data analysis tool for noise variance estimation towards a machine learning context. By modelling the variation in the data for observations that are located close to each other, properties of the data can be extracted without relying on an explicit model of the data. These ideas are translated by considering the differences of the data instead of the data itself in the so-called differogram cloud. A model for the differogram can be inferred for sufficiently small differences. By deriving an upper bound on the variance of the differogram model, this locality can be formulated without having to rely explicitly on a hyper-parameter as the bandwidth. Furthermore, a number of applications of modelfree noise variance estimators for model selection and hyper-parameter tuning have been given. While the method of least squares (under the Gauss-Markov conditions) enjoys well known properties, we have studied the properties of the LS-SVM regression when relaxing these conditions. It was recognized that outliers may have an unusually large influence on the resulting estimate. However, asymptotically the heteroscedasticity does not play any important role. Squared residual plots are proposed to assess heteroscedasticity in regression diagnostics.

A brief summary is given of the main methods for density estimation. We explain the connection between categorical data smoothing, nonparametric regression and density estimation. In addition we use the LS-SVM regression modelling for density estimation. The SVM approach (Mukherjee and Vapnik, 1999) requires inequality constraints for density estimation. One way to circumvent these inequality constraints is to use the regression-based density estimation approach. In this approach one can use the LS-SVM for regression for density estimation. The proposed method (density estimation using LS-SVM regression) has particularly advantages over Nadaraya-Watson kernel estimators, when estimates are in the tails. The data sample is pre-binned and the estimator employs the bin center as the 'sample points'. This approach also provides a sparse estimate of a density. The multivariate form of the binned estimator is given in (Holmström, 2000). Consistency of multivariate data-driven histogram methods for density estimation are proved by (Lugosi and Nobel, 1996).

In the first experiment we used the Parzen density estimator and the LS-SVM regression estimator. We used a combination of cross-validation and bootstrap for choosing the bandwidth for the Parzen kernel estimator. The average L_1 errors are estimated for each density. Both methods gives similar results

(Table 8.1). In the last experiment we applied both methods to the suicide data (Copas and Fryer, 1980). Note that the data are positive, the estimates shown in Figure 8.3 that the Parzen estimator treating the data as observations on $(-\infty, \infty)$, while the LS-SVM (RBF kernel) estimate deals with this difficulty. In order to deal with this difficulty, various adaptive methods have been proposed (Breiman *et al.*, 1977). Logspline density estimation, proposed by (Stone and Koo, 1986) and (Kooperberg and Stone, 1990), captures nicely the tail of a density but the implementation of the algorithm is extremely difficult (Gu, 1993).

We have developed a robust framework for LS-SVM regression. It allows to obtain a robust estimate based upon the previous LS-SVM regression solution, in a subsequent step. The weights are determined based upon the distribution of the error variables. We have shown, based on the empirical influence curve and the maxbias curve, that the weighted LS-SVM regression is a robust function estimation tool.
While standard SVM's approaches starts from choosing a given convex cost function and obtain a robust estimate in a top-down fashion, this procedure has the disadvantage that one should know in fact beforhand which cost function is statistically optimal. We have successfully demonstrated and alternative bottom-up procedure which starts from an unweighted LS-SVM and then robustifies the solution bij defining weightings based upon the error distribution.

We have shown that the Nadaraya-Watson estimator is nonrobust in the sence of the influence function and that L- regression achieved robustness. Based on the estimated noise model we have calculated the empirical loss function. In an experiment, we have recognized respectively the L_2 norm loss function and the L_1 loss function.

Most efficient learning algorithms in neural networks, support vector machines and kernel based learning methods require the tuning of some extra tuning parameters. For practical use, it is often preferable to have a data-driven method to select these parameters. Based on location estimators (e.g., mean, median, M-estimators, L-estimators, R-estimators), we have introduced robust counterparts of model selection criteria (e.g., Cross-Validation, Final Prediction Error criterion).

Cross-validation methods are frequently applied for selecting tuning parameters in neural network methods, usually based on L_2 or L_1 norms. However, due to the asymmetric and non-Gaussian nature of the score function, better location parameters can be used to estimate the performance. In this thesis we have introduced a repeated robust cross-validation score function method by applying concepts from robust statistics to the cross-validation methodology. We have applied a similar technique to generalized cross-validation. Simulation results illustrate that these methods can be very effective, especially with outliers on data where the L_2 methods usually fails. The proposed methods have a good robustness / efficiency trade-off such that they perform equally well in cases where L_2 would perform optimally. We have proposed robust estimation and robust model selection techniques for the use of least squares support vector machines with nonlinear ARX models. Robust techniques have been proposed for fixed-size LS-SVMs in the primal space as well as for the dual problem. Several examples illustrate that these methods can further improve standard non-robust techniques in the case of outliers and non-Gaussian noise distributions.

Inference procedures for both linear and nonlinear parametric regression models in fact assume that the output variable follows a normal distribution. With nonparametric regression, the regression equation is determined from the data. In this case, we relax the normality assumption and standard inference procedures are no longer applicable in that case. We have developed a robust approach for obtaining robust prediction intervals by using robust external bootstrapping methods.

13.2 Further research

Further research is necessary to make the kernel methods more robust. Possibly this research will support on two pillars:

(1) existing robust methods must be studied concerning the kernel based models with conversion of dual to the primal spaces, both for function estimation and kernel PCR, kernel PLS and kernel CCA.

(2) To robustify the costs function. Concretely we want replace the least squares treatment by trimmed least squares criterion, weighted least squares, L_1- criterion or ρ-function like an M-estimator. Each costs function is only an optimum for a particular error distribution. For this reason it would be desirable to make the choice of the cost function on the basis of the data adaptive. A possible working method exists to start with initial costs function and then to study the tail behaviour of the resulting errors. Then an improved cost function can be obtained. Further the robust properties of these methods must be theoretically examined. For this we base on the functional approach of Von Mises, which is used in parametric statistics and leads to the influence function and the asymptotic breakpoint, which have to be extended to nonparametric estimators. Finally the developed procedures will be applied on real data sets. In particular we are thinking of data from the chemometrics and the bio-informatics, because these contain frequently a lot of variables and a small number or a very large number of observations.

Appendix A

Preliminary Tools and Foundations

This Chapter outlines tools and foundations basic to asymptotic theory in statistics and functional analysis as treated in this thesis. The description concerning the asymptotic theory in statistics is based on (Serfling, 1980; Billingsley, 1986; Van Der Vaart, 1998). An excellent introduction of functional analysis is given by (Michel and Herget, 1981; Griffel, 1981; Aubin, 2000; Ponnusamy, 20002).

A.1 Definitions

A.1.1 The o, O and $\backsim$ notation

These symbols provide a convenient way to describe the limiting behavior of a function $f(x)$ as x tends to a certain limit a (not necessarily finite). These symbols are called "little oh," "big oh" and "twiddle" respectively. Let $f(x)$ and $g(x)$ be two functions defined on $D \subseteq \mathbb{R}$.

(i). Let the relationship between $f(x)$ and $g(x)$ is such that $\lim_{x \to a} \frac{f(x)}{g(x)} = 0$, then we say that $f(x)$ is of a smaller order of magnitude than $g(x)$ in a neigborhood of a. This fact is denoted by writing

$$f(x) = o\left(g(x)\right), \quad \text{as } x \to a,$$

which is equivalent to saying that $f(x)$ tends to zero more rapidly than $g(x)$ as $x \to a$. For example, $\sqrt{x} = o\left(x\right)$ as $x \to \infty$

(ii). Suppose that there exists a positive number M such that $|f(x)| \leq Mg(x)$ for all $x \in D$. Then $f(x)$ is said to be of an order of magnitude not exceeding that of $g(x)$. This fact is denoted by writing

$$f(x) = O\left(g(x)\right)$$

for all $x \in D$. For example, $x^2 + 2x = O(x^2)$ for large values of x.

(iii) . If $f(x)$ and $g(x)$ are any two functions such that $\lim_{x\to a}\frac{f(x)}{g(x)}=1$, then $f(x)$ and $g(x)$ are said to be asymptotically equal, written symbolically

$$f(x)\backsim g(x),\quad \text{as } x\to a.$$

For example, $\sin(x)\backsim x$ as $x\to 0$.

A.1.2 Indicator function

For any set A, the associated indicator function is

$$I_A(x)=\begin{cases}1, & x\in A,\\ 0, & x\notin A.\end{cases}\tag{A.1}$$

A.1.3 Probability space and random variables

Let Ω be an arbitrary nonempty space or set of points ω. Let $\mathcal{F}$ be a σ-field of subsets of Ω, that is, a nonempty class of subsets of Ω which contains Ω and is closed under countable union and complementation. Let P, a set function, be a probability measure defined on $\mathcal{F}$ satisfying $0\leq P(A)\leq 1$ for $A\in\mathcal{F}$, $P(0)=0$ and $P(\Omega)=1$. In probability theory Ω consists of all the possible results or outcomes ω of an experiment or observation. A subset of Ω is an event and an element $\omega\in\Omega$ is a sample point.

Definition 1 *An ordered triple* $(\Omega,\mathcal{F},P)$ *where*
(a) Ω *is a set of points* ω,
(b) $\mathcal{F}$ *is a* σ*-algebra of subsets of* Ω,
(c) P *is a probability on* $\mathcal{F}$,
is called a probabilistic model or a probability measure space.

Definition 2 *A real function* $X=X(\omega)$ *defined on* $(\Omega,\mathcal{F})$ *is a* $\mathcal{F}$*-measurable function, or a random variable, if*

$$X^{-1}(A)=\{\omega\in\Omega : X(\omega)\in A\}\in\mathcal{F},\qquad \textit{for every } A\in\mathcal{B}(\mathbb{R}),\tag{A.2}$$

where $\mathcal{B}(\mathbb{R})$ *is the* σ*-field of Borel sets in* $\mathbb{R}$. *That is, a random variable* X *is a measurable transformation of* $(\Omega,\mathcal{F},P)$ *into* $(\mathbb{R},\mathcal{B}(\mathbb{R}))$.

A.1.4 Distributions, expectations and quantiles

Note that a random variable X defined on $(\Omega,\mathcal{F},P)$ induces a measure P_X on $\mathcal{B}$ defined by the relation

$$P_X(A)=P\left(X^{-1}(A)\right),\quad A\in\mathcal{B}.$$

P_X is a probability measure on $\mathcal{B}$ and is called a probability distribution.

Definition 3 *For every $x \in \mathbb{R}$ set*

$$F_X(x) = P_X(-\infty, x] = P\{\omega \in \Omega : X(\omega) \leq x\}. \tag{A.3}$$

we call $F_X = F$ the distribution function of the random variable X.

In the following we write $\{X \leq x\}$ for the event $\{\omega \in \Omega : X(\omega) \leq x\}$. The distribution function F of a random variable X is a nondecreasing, right-continuous function on $\mathbb{R}$ which satisfies

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

and

$$F(+\infty) = \lim_{x \to \infty} F(x) = 1.$$

Definition 4 *Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution function F. Let $\hat{F}_n$ be the step function defined by*

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} I_{(-\infty,x]}(X_k), \quad x \in \mathbb{R}. \tag{A.4}$$

Then $\hat{F}_n$ is called the empirical (or sample) distribution function based on the sample $X_1, X_2, ..., X_n$.

Definition 5 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and X be a random variable defined on it. Let g be a real-valued Borel-measurable function on $\mathbb{R}$. The expectation of $g(X)$ exists if $g(X)$ is integrable over Ω with respect to P. In this case the expectation $E[g(X)]$ of the random variable $g(X)$ is defined by*

$$E[g(X)] = \int_\Omega g(X) dP. \tag{A.5}$$

Suppose that $E[g(X)]$ exists, the its follows that g is also integrable over $\mathbb{R}$ with respect to P_X (Halmos, 1950), and the relation

$$\int_\Omega g(X) dP = \int_\mathbb{R} g(u) dP_X(u) \tag{A.6}$$

holds. In particular, if g is continuous on $\mathbb{R}$ and $E[g(X)]$ exists, one can write

$$\int_\Omega g(X) dP = \int_\mathbb{R} g dP_X = \int_{-\infty}^{\infty} g(x) dF(x), \tag{A.7}$$

where F is the distribution function corresponding to P_X, and the last integral is a Riemann-Stieltjes integral. Two important special cases of (A.7) are as follows: Let F be discrete with the set of discontinuity points $\{x_n, n = 1, 2, ...\}$. Let $p(x_n)$ be the jump of F at $x_n, n = 1, 2, ...$ Then $E[g(X)]$ exists if and only if $\sum_{n=1}^{\infty} |g(x_n)| \, p(x_n) < \infty$, and in that case

$$E[g(X)] = \sum_{n=1}^{\infty} g(x_n) p(x_n) \tag{A.8}$$

Definition 6 *Let F be absolutely continuous on $\mathbb{R}$ with probability density function $f(x) = \frac{d}{dx}F(x)$. Then $E\left[g(X)\right]$ exists if and only if $\int_{-\infty}^{\infty} |g(x)| f(x)dx < \infty$, and in that case*

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx. \tag{A.9}$$

Definition 7 *The quantile function of a cumulative distribution function F is the generalized inverse $F^{-} : (0,1) \to \mathbb{R}$ given by*

$$F^{-}(q) = \inf\left\{x : F(x) \geq q\right\}. \tag{A.10}$$

In the absence of information concerning the underlying distribution function F of the sample, the empirical distribution function $\hat{F}_n$ and the empirical quantile function $\hat{F}_n^{-}$ are reasonable estimates for F and F^{-}, respectively. The empirical quantile function is related to the order statistics $x_{n(1)} \leq ... \leq x_{n(n)}$ of the sample through

$$F^{-}(q) = x_{n(i)}, \quad \textit{for } q \in \left(\frac{i-1}{n}, \frac{i}{n}\right). \tag{A.11}$$

A.2 Elementary properties of random variables with finite expectations

Denote by $\mathcal{L} = \mathcal{L}\left(\Omega, \mathcal{F}, P\right)$ the set of all random variables with finite expextations. Let $X, Y \in \mathcal{L}$, let $a, b, c \in \mathbb{R}$ and let $g(X)$ be integrable over Ω with respect to P.
(a) $aX + bY \in \mathcal{L}$ and $E\left[aX + bY\right] = aE\left[X\right] + bE\left[Y\right]$
(b) $E\left[X\right] \leq E\left[Y\right]$ if $X \leq Y$ on a set of probability 1.
(c) Let $g(X) = X^i$, $i \in \mathbb{N}_0$.
Then $E\left[X^i\right]$, if it exists, is called the moment of order i of the random variable X.
(d) Let $g(X) = (X - m)^i$, where $m \in \mathbb{R}$ and $i \in \mathbb{N}_0$.
Then $E\left[(X - m)^i\right]$, if it exists, is known as the moment of order i about the point m. In particular, if $m = E\left[X\right]$, then $E\left[(X - E\left[X\right])^i\right]$ is called the central moment of order n and is denoted by μ_i. For $i = 2$,

$$\mu_2 = E\left[(X - E\left[X\right])^2\right] = E\left[X^2\right] - (E\left[X\right])^2 \tag{A.12}$$

is called the variance of X and is denoted by $Var\left[X\right]$. The positive square root of $Var\left[X\right]$ is called the standard deviation of X. Note that $Var\left[X\right] \geq 0$ and $Var\left[X\right] = 0 \Longleftrightarrow X = c$, ($c$ constant).
(e) More generally, the covariance of two random variables X and Y denoted by $Cov\left[X, Y\right]$ is defined as

$$Cov\left[X, Y\right] = E\left[(X - E\left[X\right])(Y - E\left[Y\right])\right] \tag{A.13}$$

In the following examples we briefly introduce some standard distributions.

(i) Uniform(a, b) . The pdf of the uniform distribution is defined as

$$f\left(x\left|a, b\right.\right) = \frac{1}{b-a}, a \leq x \leq b, \tag{A.14}$$

with mean $E\left[X\right] = \frac{b+a}{2}$ and variance $Var\left[X\right] = \frac{(b-a)^2}{12}$.

(ii) Normal$\big(\mu, \sigma^2\big)$. The pdf of the normal distribution is defined as

$$f\left(x\left|\mu, \sigma^2\right.\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right), -\infty < x < \infty, \tag{A.15}$$

with mean $E\left[X\right] = \mu$ and variance $Var\left[X\right] = \sigma^2$ and $\sigma > 0$ $and - \infty < \mu < \infty$.

(iii) Double exponential(μ, σ) or Laplace distribution. The pdf of the Laplace distribution is defined as

$$f\left(x\left|\mu, \sigma\right.\right) = \frac{1}{2\sigma} \exp\left(-\left|\frac{(x-\mu)}{\sigma}\right|\right), -\infty < x < \infty, \tag{A.16}$$

with mean $E\left[X\right] = \mu$ and variance $Var\left[X\right] = 2\sigma^2$ and $\sigma > 0$ $and - \infty < \mu < \infty$.

(iv) Lognormal$\big(\mu, \sigma^2\big)$. The pdf of the lognormal distribution is defined as

$$f\left(x\left|\mu, \sigma^2\right.\right) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\log x-\mu\right)^2}{2\sigma^2}\right), 0 \leq x < \infty, \tag{A.17}$$

with mean $E\left[X\right] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ and variance $Var\left[X\right] = \exp\left(2\left(\mu + \sigma^2\right)\right) - \exp\left(2\mu + \sigma^2\right)$ and $\sigma > 0$ $and - \infty < \mu < \infty$.

(v) Exponential(λ) . The pdf of the exponential distribution is defined as

$$f\left(x\left|\lambda\right.\right) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right), 0 \leq x < \infty, \lambda > 0, \tag{A.18}$$

with mean $E\left[X\right] = \lambda$ and variance $Var\left[X\right] = \lambda^2$.

(vi) Chi-squared. The pdf of the Chi-squared distribution is defined as

$$f\left(x\left|v\right.\right) = \frac{1}{\Gamma\left(\frac{v}{2}\right)2^{v/2}} x^{(v/2)-1} \exp\left(-\frac{x}{\lambda}\right), 0 \leq x < \infty, v = 1, 2, ...$$

with mean $E\left[X\right] = v$ and variance $Var\left[X\right] = 2v$.
Figure A.1 shows the relationships among the common distributions.

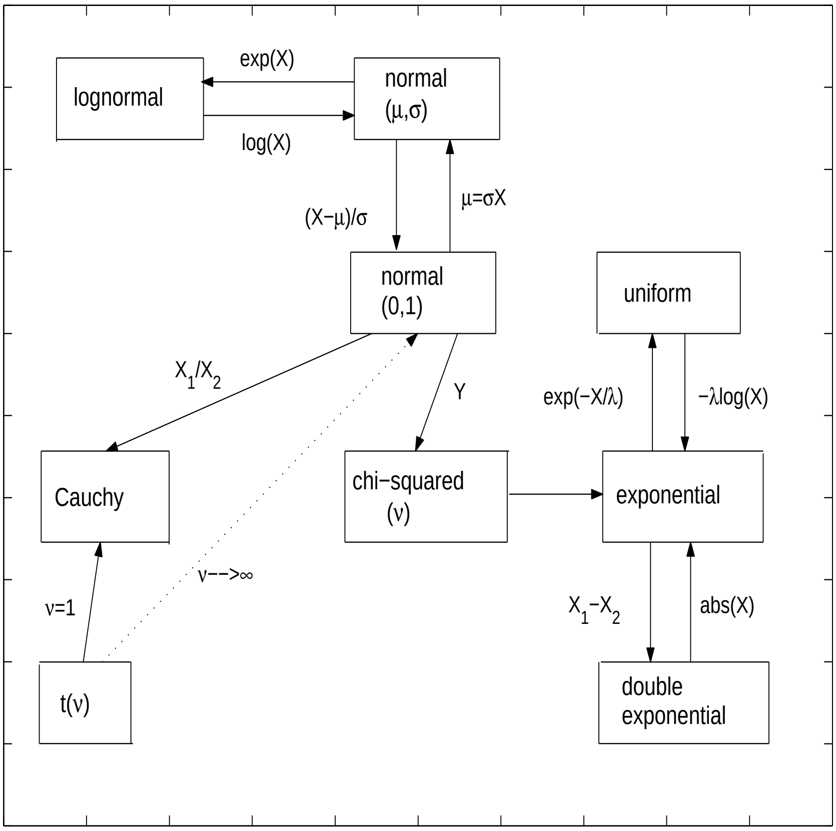

Figure A.1: Relationships among the common distributions.

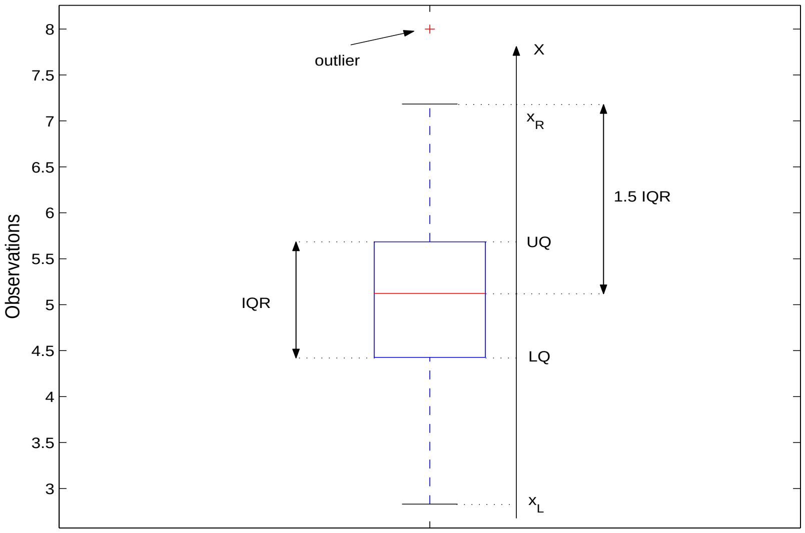

Figure A.2: The boxplot based on the sample interquartile range.

A.3 Statistical graphics

A.3.1 Boxplot for the univariate data

The univariate boxplot (Tukey, 1977) is a graphical tool for summarizing the distribution of a single random variable. Being a simple data analysis technique, it yields information about the location (the median), scale (the interquartile range), asymmetry, tails and outliers of a data distribution. A boxplot is the rectangle with the base equal to the sample interquartile range IQR, separated into two parts by the sample median (see Figure A.2).

¿From each side of the box, the two straight line segments are drawn describing the distribution tails, and finally, the observations lying aside these domains are marked and plotted being the candidates for outliers. The left and right boundaries for the distribution tails are given by

$$x_L = \max\left(x_{n(1)}, LQ - \frac{3}{2}IQR\right), \tag{A.19}$$

and

$$x_R = \min\left(x_{n(n)}, UQ + \frac{3}{2}IQR\right). \tag{A.20}$$

here LQ and UQ are the lower and upper sample quartiles, respectively. In general, they can be defined by $LQ = x_{n(j)}$ and $UQ = x_{n(n-j+1)}$ with $j = 0.25n$.

A.3.2 Quantile-quantile plots

Construction of a q-q Plot

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. The sample advantages of the q-q plot are: (i) The sample sizes do not need to be equal. (ii) Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

Interpreting quantile-quantile plots

If the data distribution matches the theoretical distribution, the points on the plot form a linear pattern. Thus, you can use a q-q plot to determine how well a theoretical distribution models a set of measurements. The following properties of these plots make them useful diagnostics to test how well a specified theoretical distribution fits a set of measurements: (i) If the quantiles of the theoretical and data distributions agree, the plotted points fall on or near the line $y = x$. (ii) If the theoretical and data distributions differ only in their location or scale, the points on the plot fall on or near the line $y = a + bx$. The slope a and intercept b are visual estimates of the scale and location parameters of the theoretical distribution. The interpretations of commonly encountered departures from linearity are summarized in the following Table A.1.

A.4 Modes of convergence of a sequence of random variables

In this subsection we consider some concepts of convergence of the sequence $\{X_n\}$.

A.4.1 Convergence in probability

Let $X_1, X_2, ...$and X be random variables on a probability space $(\Omega, \mathcal{F}, P)$. Then $\{X_n, n \geq 1\}$ converges in probability to X as $n \to \infty$ if for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0 \tag{A.21}$$

Description of Point Pattern	**Possible Interpretation**
All but a few points fall on a line	Outliers in the data
Left end of pattern is below the line, right end of pattern is above the line	Long tails at both ends of the data distribution
Left end of pattern is above the line, right end of pattern is below the line	Short tails at both ends of the distribution
Curved pattern with slope increasing from left to right	Data distribution is skewed to the right
Curved pattern with slope decreasing from left to right	Data distribution is skewed to the right
Staircase pattern (plateaus and gaps)	Data have been rounded or are discrete

Table A.1: Quantile-quantile plot diagnostics.

This is written $X_n \xrightarrow{p} X$ or $p\text{-}\lim_{n\to\infty} X_n = X$. Extensions to the case random elements of a metric space is straightforward, by replacing $|X_n - X|$ by the relevant metric (see Billingsley, 1968).

A.4.2 Convergence with probability 1

Consider random variables $X_1, X_2, \ldots$and X on $(\Omega, \mathcal{F}, P)$. Then $\{X_n\}$ converges with probability 1 (or strongly, almost surely, almost everywhere, etc.) to X if

$$\lim_{n\to\infty} \mathbb{P}\left(|X_m - X| < \varepsilon, \quad \forall\, m \geq n\right) = 1 \tag{A.22}$$

This is written $X_n \xrightarrow{wp1} X$ or $p1\text{-}\lim_{n\to\infty} X_n = X$. Extensions to the case random elements of a metric space is straightforward.

A.4.3 Convergence in distribution

Consider distribution functions $F_1(\cdot)$, $F_2(\cdot)$, ... and $F(\cdot)$. Let $X_1, X_2, \ldots$and X be random variables (not necessarily on a common probability space) having these distributions, respectively. Then $\{X_n\}$ converges in distribution (or in law) to X if

$$\lim_{n\to\infty} F_n(t) = F(t), \quad \text{each continuity point } t \text{ of } F. \tag{A.23}$$

This is written $X_n \xrightarrow{d} X$ or $d\text{-}\lim_{n\to\infty} X_n = X$.

Remark 8 *The convergences $\xrightarrow{p}$ and $\xrightarrow{wp1}$ each represent a sense in which, for n sufficiently large, $X_n(\omega)$ and $X(\omega)$ approximate each other as functions of ω, $\omega \in \Omega$. This means that the distribution of X_n and X cannot be too dissimilar, whereby approximation in distribution should follow. On the other hand, the convergence $\xrightarrow{d}$ depends only on the distribution functions involved and does not necessitate that the relevant X_n and X approximate each other as functions of ω.*

Relationship among the modes of convergence ($\overset{p}{\rightarrow}$, $\overset{wp1}{\rightarrow}$ and $\overset{d}{\rightarrow}$) can be summarized by the following scheme:

$$X_n \overset{wp1}{\rightarrow} X \Rightarrow X_n \overset{p}{\rightarrow} X \Rightarrow X_n \overset{d}{\rightarrow} X. \tag{A.24}$$

None of these implications can be reversed.

A.5 Foundations of functional analysis

Definition 9 *(Linear/vector space). A linear space, or vector space, over the field $\mathbb{R}$ of real numbers is a set V of elements called points or vectors, endowed with the operations of addition and scalar multiplication having the following properties:*

(a) $\forall u, v \in V$ *and* $\forall a, b \in \mathbb{R}$:

(1) $u + v \in V$,

(2) $au \in V$,

(3) $1u = u$,

(4) $a\,(bu) = (ab)\,u$,

(5) $(a + b)\,u = au + bu$,

(6) $a\,(u + v)\,au + av$.

(b) $(V, +)$ *is a commutative group; that is ,* $\forall u, v, w \in V$:

(1) $\exists 0 \in V$ *such that* $0 + u = u$,

(2) $\exists\,(-u) \in V$ *such that* $u + (-u) = 0$, $u + v = v + u$, $u + (v + w) = (u + v) + w$.

Definition 10 *(Topological space). A topological space is a set endowed with a topology, which is a family of subsets called open sets with the properties:*

(1) the intersection of any two open sets is an open set,

(2) the union of any collection of open sets is an open set,

(3) the empty set and the whole space are open sets.

Definition 11 (Metric spaces). *Let V be a nonempty set and let $d(.,.)$ be a mapping/function from $V \times V$ to $\mathbb{R}$, $d : V \times V \to \mathbb{R}$, satisfying the following conditions for all u, v and $w \in V$:*

(1) $d(u, v) = 0 \Longleftrightarrow u = v$

(2) $d(u, v) = d(v, u)$

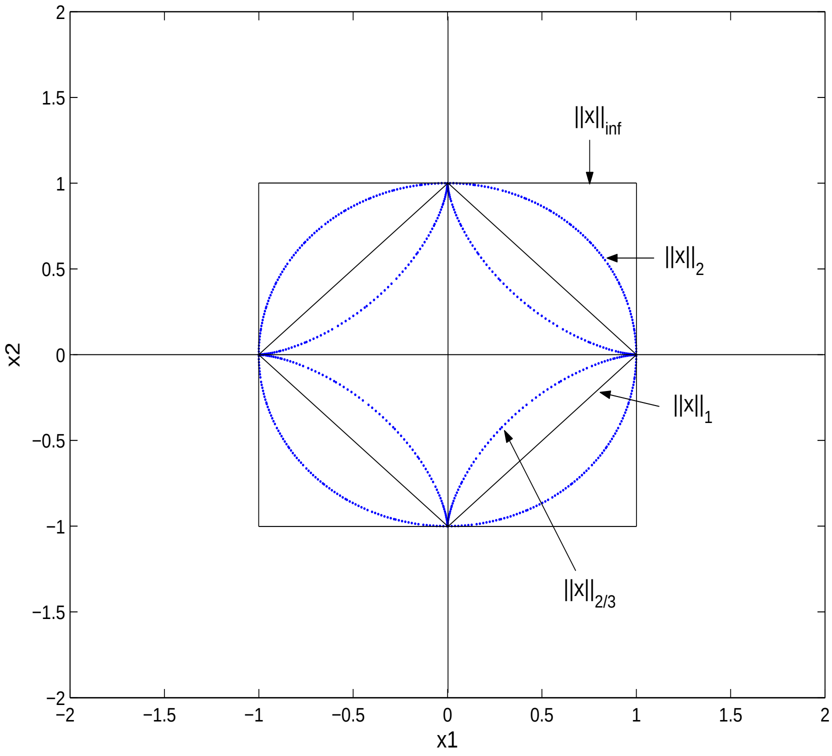

Figure A.3: Description (unit spheres) of balls with respect to the d_p metric.

(3) $d(u,v) \leq d(u,w) + d(w,v)$

Then d *is called a metric or a distance function on* V*. A metric space* V *is also a topological space in which the topology is given by a metric.*

Definition 12 *Let a metric space be denoted by* (V,d)*. The set* $B(x_0,r)$ *of all points* $x \in (V,d)$ *satisfying the inequality* $d(x_0,x) < r$ *is called an open ball with centre* $x_0 \in (V,d)$ *and radius* $r > 0$*. If the inequality is replaced by* $d(x_0,x) \leq r$ *one speaks of a closed ball.*

Example 13 (a) *Figure A.3 shows the unit closed ball* $d_p(x,0) \leq 1$, $0 < p \leq \infty$ *for the vector space* $\mathbb{R}^2$.

(b) *The open ball with center at* g_0 *and radius* δ *with respect to the supremum metric*

$$d_\infty(f,h) = \sup_{x\in[0,10]} |f(x) - h(x)| \tag{A.25}$$

on the space $C[0,10]$ *is given by*

$$B(g_0,\delta) = \{g \in C[0,10] : d_\infty(g,g_0) < \delta\}. \tag{A.26}$$

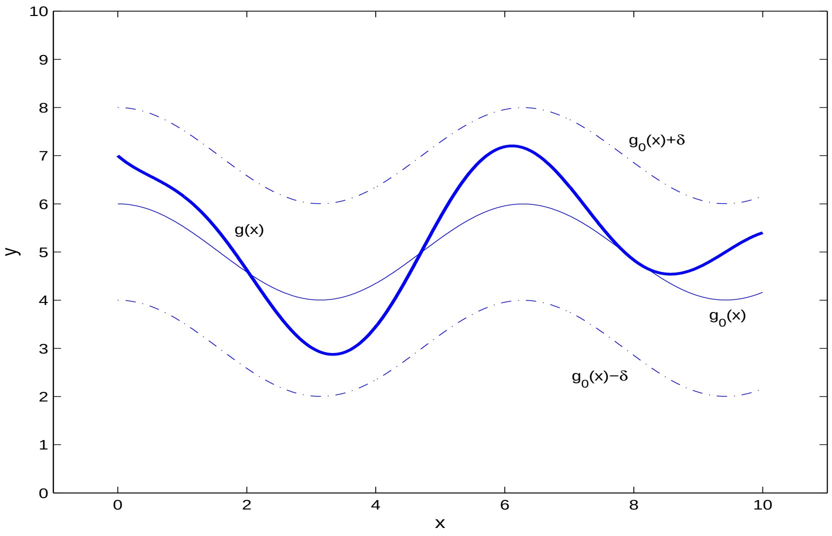

Figure A.4: Balls in $(C\,[0,10]\,,\ d_\infty)$. The ball around g_0 consists of all functions g such that the graph of g lies within a band around g_0 of width δ.

This means that $|g\,(x) - g_0\,(x)| < \delta$ for each $x \in [0,10]$ and therefore, the ball around g_0 consists of all functions g such that the graph of g lies within a band about g_0 of width δ. The region in which the graph of g must lie is shown in Figure A.4.

A.5.1 Normed linear spaces and Banach spaces

Examples of normed spaces may be divided into three kinds; namely coordinate spaces, sequence spaces and function spaces. A Banach space is simply a complete normed space. Thus it contains the limit of all its Cauchy sequences.

Definition 14 *(Normed spaces). Let V be a linear/vector space over the field $\mathbb{F}$ ($\mathbb{C}$ or $\mathbb{R}$). A norm on V is a mapping/function $\|.\|$ from V to $\mathbb{R}_0^+$, $\|.\| : V \to \mathbb{R}_0^+$ satisfying the following three axioms*

(i) $\|u\| = 0 \Rightarrow u = 0$

(ii) $\|\lambda u\| = |\lambda|\,\|u\|\ \forall u \in V$ *and* $\lambda \in \mathbb{F}$

(iii) $\|u + v\| \leq \|u\| + \|v\|\ \forall u, v \in V$

The pair $(V, \|.\|)$ is called a normed space.

Proposition 15 *Every normed space $(V, \|.\|)$ is a metric space with respect to the distance function $d(u, v) = \|u - v\|\,, \forall u, v \in V$.*

A.5.2 Inner product spaces, Hilbert spaces and reproducing kernel Hilbert spaces

Definition 16 *(Inner product spaces). An inner product on a vector space V is a scalar-valued function $\langle u, v\rangle$, defined for all ordered pairs of vectors $u, v \in V$ and which satisfies the following axioms:*

(i) $\langle u, v\rangle = \langle \overline{v, u}\rangle$ $\forall u, v \in V$, (the bar denotes complex conjugate)

(ii) $\langle au + bv, w\rangle = a\langle u, w\rangle + b\langle v, w\rangle$ $\forall u, v, w \in V$ and scalars a, b,

(iii) $\langle u, u\rangle > 0$, $\langle u, u\rangle = 0$ if, and only if, $u = 0$.

A real inner product space is often called a Euclidean space.

Proposition 17 *An inner product space which is complete with respect to the norm induced by the inner product is called a Hilbert space.*

Definition 18 *(reproducing kernel Hilbert spaces). A Hilbert space H of a real-valued function on a set D is called a reproducing kernel Hilbert space (RKHS) if, and only if, all the evaluation functionals on H are continuous (bounded).*

A.5.3 Function spaces

Let f be a function and $f(x)$ refer to the value of the function at the point x. If f is differentiable and its derivative function $f^{'}(x)$ is a continuous function of x, then f is continuously differentiable and $f \in C^1$. If f is vth order differentiable and $f^{(v)}(x)$ is a continuous function of x, then f is continuously vth order differentiable and $f \in C^v$. If f is smooth, each derivative of a smooth function is smooth, $f \in C^\infty$. A continuous function is of the class C^0. Summarized, one has the following regularity hierarchy

$$C_b \supset C^0 \supset C^1 \supset ... \supset \bigcap_{v\in\mathbb{N}} C^v = C^\infty. \tag{A.27}$$

where C_b denotes the set of all bounded functions. For example $f(x) = |x|$ is of class C^0 and $f(x) = x$ is of class C^∞.

A.6 Statistical functionals and differentiability

A.6.1 Statistical functionals

Let $X_1, ..., X_n$ be a sample from a population with probability distribution F and let $T_n = T_n(X_1, ..., X_n)$ be a statistic. When T_n can be written as a functional T of the empirical distribution $\hat{F}_n$, $T_n = T(\hat{F}_n)$, then we call T a statistical functional. A statistical functional $T(F)$ is any function of F. Examples are the mean $\mu = \int xdF(x)$, the variance $\sigma^2 = \int (x-\mu)^2 dF(x)$ and the median $med = F^{-}(1/2)$. Another example of a functional is $\int \Upsilon(x) dF(x)$ where $\Upsilon(x)$ is any function of x. Many important properties of statistics may be expressed in terms of analytic properties of statistical functionals. A statistical functional

$T(\hat{F}_n)$ is robust at F according to Hampel (Hampel, 1971) if $\mathcal{Z}\left(T(\hat{F}_n)\right)$, a function of the distribution F of a single observation, is a continuous function at F when the Prohorov metric is used in the spaces for both F and $\mathcal{Z}(T)$. The Influence Function is a form of derivative of a functional. The use of Taylor expansions of statistical functionals to prove asymptotic normality is known as Von Mises' method. An other advantage of statistical functionals is that there is often a natural extension to spaces that contain more than just distribution functions.

Consider the set of all distribution functions on $\mathbb{R}$ denoted by $\mathcal{U}$,

$$\mathcal{U} = \{ F | \, F : \mathbb{R} \to [0,1] \} . \tag{A.28}$$

A statistical functional is a mapping defined on a space of distribution functions. Usually the image space is $\mathbb{R}$ but it could also be a set of categories or a higher dimensional Euclidean space. The domain usually includes all empirical distribution functions and the hypothetical true distribution. Let statistical functionals be denoted by $T(F)$ where $F \in \mathcal{U}$ is the distribution of the data and the natural estimate of $T(F)$ is $T(\hat{F}_n)$ where $\hat{F}_n$ is the sample distribution function.

Definition 19 *A function $T : D_{\mathcal{U}} \to \mathbb{R}$, where $D_{\mathcal{U}} \subset \mathcal{U}$, is said to be a statistical functional if it satisfies the following two conditions:*
(i) $\hat{F}_n \in D_{\mathcal{U}}$ for all finite sequences $x_1, ..., x_n$.
(ii) The map $(x_1, ..., x_n) \mapsto T(\hat{F}_n)$ is for all fixed n a Borel function on $\mathbb{R}^n$.

Example 20 *Let $D_{\mathcal{U}} \subset \mathcal{U}$ be the set of all distribution functions with existing first moment, that is $D_{\mathcal{U}} = \left\{F \in \mathcal{U} : \int |x| \, dF(x) < +\infty\right\}$. Define $T : D_{\mathcal{U}} \to \mathbb{R}$ by $T(F) = \int x dF(x)$. Now $T(F)$ presents the mean of a population with distribution function F.*

Example 21 *Let $D_{\mathcal{U}}$ be the set defined by $D_{\mathcal{U}} = \left\{F \in \mathcal{U} : \int x^2 dF(x) < +\infty\right\}$. Define $T : D_{\mathcal{U}} \to \mathbb{R}$ by $T(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$. Now $T(F)$ presents the variance of a population with distribution function F.*

A.6.2 Differentiability

It is convenient first to establish a general form of differentiation and then restrict this to the form we wish to use. Let V and U be topological vector spaces and let $L(V,U)$ be the set of continuous linear transformations from V to U. Let S be a class of subsets of V such that every subset consisting of a single point belongs to S, and let A be an open subset of V.

Definition 22 *A function $T : A \to U$ is S-differentiable at $F \in A$ if there exists $T_F^{'} \in L(V,U)$ such that for any $B \in S$*

$$\lim_{\epsilon \to 0} \frac{T(F + \epsilon H) - T(F) - T_F^{'}(\epsilon H)}{\epsilon} = 0 \tag{A.29}$$

uniformly for $H \in B$. The linear function $T_F^{'}$ is called the S-derivative of T at F. Particular types of differentiation are:

(i) $S =$ bounded subset of V; this corresponds to Frechet differentiation.

(ii) $S =$ compact subset of V; this corresponds to Hadamard differentiation.

(iii) $S =$ finite subset of V; this corresponds to Gateaux differentiation.

A.7 Aspects of statistical inference

Following the usual terminology, we use the term "estimator" to denote a random variable, and "estimate" to denote a realization of the random variable. The statistical properties of an estimator of a function at a given point are analogous to the statistical properties of an estimator of a scalar parameter. Let f be the true function, $f(x)$ refer to the value of the function at the point x and $\hat{f}$ or $\hat{f}(x)$ denote the estimator of f or of $f(x)$.

A.7.1 Pointwise properties of function estimators

Bias

The bias of the estimator of a function value at the point x is

$$Bias\left(\hat{f}\left(x\right), f\left(x\right)\right) = E\left[\hat{f}\left(x\right)\right] - f\left(x\right). \tag{A.30}$$

If the bias is zero, the estimator is unbiased at the point x. If the estimator is unbiased at every point x in the domain of f, the estimator is pointwise unbiased.

Variance

The variance of the estimator at the point x is

$$Var\left[\hat{f}\left(x\right)\right] = E\left[\left(\hat{f}\left(x\right) - E\left[\hat{f}\left(x\right)\right]\right)^2\right]. \tag{A.31}$$

Mean Squared Error

The mean squared error, MSE, at the point x is

$$MSE\left(\hat{f}\left(x\right)\right) = E\left[\left(\hat{f}\left(x\right) - f\left(x\right)\right)^2\right]. \tag{A.32}$$

The mean squared error is the sum of the variance and the square of the bias

$$\begin{aligned} MSE\left(\hat{f}\left(x\right)\right) &= E\left[\left(\hat{f}\left(x\right)\right)^2 - 2\hat{f}\left(x\right) f\left(x\right) + \left(f\left(x\right)\right)^2\right] \\ &= Var\left[\hat{f}\left(x\right)\right] + \left(E\left[\hat{f}\left(x\right)\right] - f\left(x\right)\right)^2. \end{aligned} \tag{A.33}$$

Mean Absolute Error

The mean absolute error, MAE, at the point x is

$$MAE\left(\hat{f}\left(x\right)\right) = E\left[\left|\hat{f}\left(x\right) - f\left(x\right)\right|\right]. \tag{A.34}$$

It is more difficult to do mathematical analysis of the MAE than for the MSE. Furthermore, the MAE does not have a simple decomposition into other meaningful quantities similar to the MSE.

A.7.2 Global properties of function estimators

Often, one is interested in some measure of the statistical properties of an estimator of a function over the full domain D of the function. The obvious way of defining statistical properties of an estimator of a function is to integrate the pointwise properties. Three useful measures are the L_1-norm, also called the integrated absolute error, or IAE,

$$IAE\left(\hat{f}\right) = \int_D \left|\hat{f}\left(x\right) - f\left(x\right)\right| dx, \tag{A.35}$$

the square of the L_2 norm, also called the integrated squared error, or ISE,

$$ISE\left(\hat{f}\right) = \int_D \left(\hat{f}\left(x\right) - f\left(x\right)\right)^2 dx, \tag{A.36}$$

and the L_∞ norm, the sup absolute error, or SAE,

$$SAE\left(\hat{f}\right) = \sup_x \left|\hat{f}\left(x\right) - f\left(x\right)\right|. \tag{A.37}$$

Integrated Mean Squared Error, Mean Absolute Error and Mean sup Absolute Error

The integrated mean squared error is

$$IMSE\left(\hat{f}\right) = \int_D E\left[\left(\hat{f}\left(x\right) - f\left(x\right)\right)^2\right] dx. \tag{A.38}$$

The integrated mean absolute error is

$$IMAE\left(\hat{f}\right) = \int_D E\left[\left|\hat{f}\left(x\right) - f\left(x\right)\right|\right] dx. \tag{A.39}$$

The mean sup absolute error is

$$MSAE\left(\hat{f}\right) = \int_D E\left[\sup_x \left|\hat{f}\left(x\right) - f\left(x\right)\right|\right] dx. \tag{A.40}$$

Consistency

Let $\int \left(\hat{f}\left(x\right)-f\left(x\right)\right)^2 dF(x)$ be the L_2 error. A sequence of function estimates $\left\{\hat{f}_n\right\}$ is called weakly universally consistent if

$$\lim_{n\rightarrow\infty} E\left[\int \left(\hat{f}\left(x\right)-f\left(x\right)\right)^2 dF(x)\right]=0 \qquad \text{(A.41)}$$

for all distribution of (X,Y) with $E\left[Y^2\right]<\infty$.

Rate of convergence

Given data $\mathcal{D}_n=\left\{\left(x_1,y_1\right),...,\left(x_n,y_n\right)\right\}$ a function estimate is denoted by $\hat{f}_n$. Let $\mathcal{U}_\kappa$ be a class of distribution of (X,Y) where the corresponding regression function satisfies some smoothness condition depending on a parameter κ. In the minimax approach one tries to minimize the maximal expectation of the L_2 error within a class of distributions, e.g,

$$\inf_{\hat{f}_n} \sup_{(X,Y)\in\mathcal{U}_\varkappa} E\left[\int \left(\hat{f}\left(x\right)-f\left(x\right)\right)^2 dF(x)\right], \qquad \text{(A.42)}$$

where the infimum is taken over all estimates $\hat{f}_n$.

Bibliography

[1] Aït-Sahalia, Y. (1995). *The delta method for nonlinear kernel functionals.* University of Chicago, USA.

[2] Aït-Sahalia, Y., Bickel, P.J., Stoker, T.M. (2001). Goodness-of-t tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics* 105, 363–412

[3] Akaike, H. (1973). Statistical predictor identification. *Ann. Inst. Statist. Math.* 22, 203-217.

[4] Allen, D.M. (1974). The relationship between variables selection and data augmentation and a method of prediction. *Technometrics* 16, 125-127.

[5] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J. Rogers, W.H., Tukey, J.W. (1972). *Robust estimation of location.* Princeton, New Jersey: Princeton University Press.

[6] Antoniadis, A. (1997). Wavelets statistics: A review. *Journal of the Italian Statistical Association* (6), 97–144.

[7] Antoniadis, A. and J. Fan (2001). Regularized wavelet approximations (with discussion). *Journal of the American Statistical Association* 96, 939–967.

[8] Arthanari, T.S., Dodge, Y. (1981). *Mathematical programming in statistics.* John Wiley & Sons, INC.

[9] Atkinson, A.C. (1985). *Plots, Transformations and Regression.* Oxford: Oxford University Press.

[10] Aubin, J. (2000). *Applied functional analysis.* John Wiley & Sons, INC.

[11] Bednarski, T. (1993). Fréchet differentiability of statistical functionals and implications to robust statistics. *In New Directions in Statistical Data Analysis and Robustness*, (Morgenthaler, S., Ronchetti, E., Stahel, W.A. Eds.), Birkhäuser Verlag, Basel, pp. 25–34.

[12] Barnett, V., Lewis, T. (1984). *Outliers in statistical data.* John Wiley & Sons, INC.

[13] Barnard, G.A. (1949). Statistical inference. *Journal of the Royal Statistical Society*, Series B 11,115-139.

[14] Barnard, G.A. (1980). Pivotal inference and Bayesian controversy. Bayesian Statistics (Bernardo, J.M.,

[15] DeGroot, M.H., Lindley, D.V., and Smith, A.F.M., eds.) Valencia: University Press.

[16] Beaton, A.E., Tukey, J.W. (1974). The fitting of power series, meaning polynomials illustrated on band-spectroscopic data. *Technometrics* 16, 147-185.

[17] Beal, S.L., Sheiner, L.B. (1988). Heteroscedastic nonlinear regression. *Technometrics*, 30, 327-338.

[18] Bellman, R.E. (1961). *Adaptive control processes.* Princeton University Press.

[19] Beran, R.J. (1984). Jackknife approximation to bootstrap estimates. *Ann. Statist.* 12, 101-118.

[20] Berger, J.O. (1975). *Statistical decision theory and Bayesian analysis.* 2nd edition, Springer-Verlag. New York.

[21] Berlinet, A, Devroye, L. (1994). A comparision of kernel density estimates. *Publ. Inst. Stat. Univ. Paris* 38, 3-59.

[22] Berger,J.O. (1985). Statistical decision theory and Bayesian analysis, 2nd edition. New York: Springer-Verlag.

[23] Bickel, P.J., Lehman, E.L. (1975). Descriptive statistics for nonparametric models, *Ann. Statist.* 3, 1038-1045.

[24] Bickel, P.J. (1965). On some robust estimates of location, *Ann. Math. Statist.* 36, 847-858.

[25] Bickel, P.J., Yahav, J.A. (1988). Richardson extrapolation and the bootstrap. *J. Amer. Statist. Assoc.* 83, 387-393.

[26] Billingsley, P. (1986), *Convergence of probability Measures*, Wiley, New York.

[27] Birgé, L., Massart, P. (1998). Minimum contract estimators on sieves: exponential bounds and rates of convergence. *Bernoulli.* 4, 329-375.

[28] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

[29] Bishop, C.M., Legleye, C. (1995). Estimating conditional probability densities for periodic variables. *Advances Neural Inform. Prossesing Syst.*, 7.

[30] Blake, A. (1989). Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Transactions on Image Processing* 11, 2–12.

[31] Boente, G., Fraiman, R. (1994). Local L-estimators for nonparametric regression under dependence. *J. Nonparametric Statist.* 4, 91-101.

[32] Booth, J.G., Hall, P., Wood, A.T.A. (1993). Balanced importance resampling for the bootstrap. *Ann. Statist.* 21, 286-298.

[33] Boser, B., Guyon, I., Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, AMC, Pittsburgh, 144-152.

[34] Boscovich, R.J. (1757). De literaria expeditione per pontificiam ditionem et synopsis amplioris operis *Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii*, 4, 353-396.

[35] Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* 71, 353-360.

[36] Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* 40, 318–335.

[37] Box, G. E. P., Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. Royal Statist. Soc. B* 17 , 1–34.

[38] Box, G. E. P., Hunter, W. G., Hunter, J. S. (1998). *Statistics for Experimenters.* Wiley, N. Y., 1978.

[39] Box, G. E. P., Leonard, T., Wu, C. F., (1983). Eds. Scientific Inference, *Data Analysis, and Robustness.* Academic Press, New York, 24

[40] Breiman, L., Meisel, W., Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* 19, 135-144.

[41] Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and regression trees.* Wadsworth.

[42] Bunke, O., Droge, B. (1984). Bootstrap and cross-validation estimates of the prediction error for linear regression models, *Ann. Statist.* 12, 1400-1424.

[43] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76, 3, pp.503-514.

[44] Carroll, R.J., Ruppert, D. (1981). On robust tests for heteroscedasticity. *Ann. Statist.* 9, 205-209.

[45] Carroll, R.J., Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* 10, 429-441.

[46] Carroll, R.J.,Wu, C.F., Ruppert, D. (1988). The effect of estimating weights in weighted least squares. *J. Amer. Statist. Assoc.* 83, 1045-1054.

[47] Causey, B.D. (1980). A frequentist analysis of a class of ridge regression estimators. *J. Amer. Statist. Assoc.*, 75, 736-738.

[48] Čencov, N.N., (1962). Evaluation of an unknown distribution density from observations. *Doklady*, 3, 1559-1562.

[49] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 74,829-836.

[50] Choi, E., Hall, P., Rousson, V. (2000). Data Sharpening methods for bias reduction in nonparametric regression. *Ann. Statist.* 28, 1339-1955.

[51] Cherkassky, V., Mulier, F. (1998). *Learning from Data.* Wiley, New York.

[52] Clark, R.M. (1975). A calibration curve for radio carbon dates. *Antiquity* 49 251-266.

[53] Clark, D.I. (1985). The mathematical structure of Huber's M-estimator. SIAM *J. Sci. Statist. Comput.* 6, 209-219.

[54] Clarke, B. R. Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.* 11 (1983), 1196–1205.

[55] Clarke, B. R. (1986). Nonsmooth analysis and Fréchet differentiability of M-functionals. *Prob. Th. Rel. Field* 73, 197–209.

[56] Copas, J.B., Fryer, M.J. (1980). Density estimation and suicide risks in psychiatric treatment. *J. Roy. Statist. Soc. A*, 143, 167-176.

[57] Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman & Hall.

[58] Cook, R.D., Weisberg, S. (1983). Diagnostic for heteroscedasticity in regression. *Biometrika*, 70, 1-10.

[59] Cortes, C., Vapnik, V.N. (1995). Support vector networks, *Mach. Learn.* 20, 1-25.

[60] Courant, R., Hilbert, D. (1953). *Methods of mathematical physics.* Volume I, Wiley-Interscience, New York.

[61] Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.*, 31, 377-390.

[62] Cressie, N.A.C. (1993). *Statistics for spatial data.* John Wiley & Sons, INC.

[63] Daniel, C. (1920). Observations weighted according to order. *Amer.J. Math.* 42, 222-236.

[64] Daniels, H.E., Young, G.A. (1991). Saddlepoint approximations for the studentized mean, with an application to the bootstrap.

[65] Davidian, M., Carroll, R.J. (1987). Variance function estimation. *J. Amer. Statist. Assoc.* 82, 1079-1091.

[66] Davison, A.C. (1988). Discussion of the Royal Statistical Society meeting on the bootstrap. *Journal of the Royal Statistical Society*, Series B 50, 356-357.

[67] Davison, A.C., Hinkley, D.V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* 75, 417-431.

[68] Davison, A.C., Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

[69] Davison, A.C., Hinkley, D.V., Schechtman, E. (1986). Efficient bootstrap simulations. *Biometrika* 73, 555-566.

[70] Davis, R.A., Gather, U. (1993). The identification of multiple outliers. *J. Amer. Statist. Assoc.* 88, 782-801.

[71] De Brabanter, J., Pelckmans, K., Suykens, J.A.K., Vandewalle, J., De Moor, B.(2002). Robust cross-validation score function for LS-SVM non-linear function estimation, Internal Report 02-94, ESAT-SISTA, K.U.Leuven.

[72] De Brabanter, J., Pelckmans, K., Suykens, J.A.K., De Moor, B., Vandewalle, J. (2004). Robust statistics for Kernel Based NARX Modeling, Internal Report 04-38, ESAT-SISTA, K.U.Leuven.

[73] De Veaux, R.D., Schumi, J., Schweinsberg, J., Ungar, L.H. (1998). Prediction intervals for neural networks via nonlinear regression. *Technometrics. 40, 273-282.*

[74] DiCiccio, Martin and Young, (1992). Fast and accurate approximate double bootstrap approximations. *Biometrika*, 79, 285-295.

[75] DiCiccio, Martin and Young, (1994). Analytic approximations to bootstrap distribution functions using saddlepoint methods. *Statist. Sin.* 4, 281- 295.

[76] Diggle, P.J. (1990). *Time series: A biostatistical introduction.* Oxford: Oxford University Press.

[77] Dodge, G. (1984). Robust estimation of regression coefficient by minimizing a convex combination of least squares and least absolute deviations. *Comp. Statist.* 1, 139-153

[78] Dodge, G., Jureckova, J. (1987). *Adaptive combination of least squares and least absolute deviations estimators.* Statistical data analysis based on L_1-norm and related methods. (Dodge, Y. ed.), 275-284. North-Holland, Amsterdam.

[79] Dodge, G., Jureckova, J. (1988). *Adaptive combination of M-estimator and L_1-estimator in the linear model.* Optimal design and analysis of experiments (Dodge, Y., Fedorov, V.V., Wynn, H.P. eds.), 167-176. North-Holland, Amsterdam

[80] Dodge, G., Jureckova, J. (1992). *A class of estimators based on adaptive convex combinations of two estimation procedures.* L_1-statistical analysis and related methods (Dodge, Y. ed.), 31-44. North-Holland, Amsterdam.

[81] Donoho, D.L. and I.M. Johnstone (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika* 81, 425–455.

[82] Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Ann. Statist.* 7. pp. 1-26.

[83] Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. SIAM, Philadelphia.

[84] Efron, B. (1990). More efficient bootstrap computations. *J. Amer. Statist. Assoc.* 55, 79-89.

[85] Efron, B., Stein, C. (1981). The Jackknife estimate of variance. *The Annals of Statistics*, vol. 9, no 3, 586-596.

[86] Efron, B., Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman and Hall, New York.

[87] Eubank, R.L. (1988). *Spline smoothing and nonparametric regression.* Marcel Dekker, New York.

[88] Eubank, R.L. (1999). *Nonparameteric Regression and Spline Smoothing.* Marcel Dekker, Inc. New York.

[89] Everitt, B.S., Hand, D.J. (1981). Finite mixture distributions. Chapman & Hall. New York.

[90] Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* 87(420), 998-1004.

[91] Fan, J. (1997). Comments on wavelets in statistics: A review. *Journal of the Italian Statistical Association* (6), 131–138.

[92] Fan, J. and R. Li (2001). Variable selection via nonconvex penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.

[93] Fox, J. (1991). *Regression diagnostics.* Newberry Park, CA: Sage.

[94] Faraway, J.J. (1990). Bootstrap selecting of bandwidth and confidence bands for nonparametric regression. *Journal of Statistical Computations and Simulation*, 97, 97-44.

[95] Fernholz, L.T. (1983). *von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics, Springer-Verlag.

[96] Feller, W. (1966). *An Introduction to Probability theory and its Applications*, Vol. I (2nd ed.) and Vol. II, Wiley, New York.

[97] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. Philos. Trans. Royal. Soc. London, A 222, 309-368.

[98] Flury, B., Riedwyl, H. (1985). T^2 tests, the linear two-group discriminant function, and their computation by linear regression. *The American Statistician*, 39, 1, 20-25.

[99] Frank, L.E. and J.H. Friedman (1993). A statistical view of some chemometric regression tools. *Technometrics* (35), 109–148.

[100] Friedman, J., Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers Ser. C* 23, 881- 889.

[101] Friedman, J., Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, 79, 599-608.

[102] Fu, W.J. (1998). Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* (7), 397–416.

[103] Gasser, T., Müller, H.G. (1979). *Kernel estimation of regression functions, in Smoothing Techniques for* . Springer- Verlag, New York.

[104] Gasser, T., Sroka, L., Jenner, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73, 625-633.

[105] Gemen, S., Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* 10, 401-414.

[106] Genovese, C.R., Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* 28, 1105-1127.

[107] Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* 29, 1264-1280.

[108] Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations.* Wiley, N. Y.

[109] Golub, G.H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, 21, 215-223.

[110] Golub, G.H., Van Loan, C.F. (1989). *Matrix computations.* John Hopkins University Press.

[111] Good, I.J., Gaskins, R.A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, 58, 255-277.

[112] Györfi, L., Kohler, M., Krzyżak, A., Walk, H. (2002). *A distribution-free theory of nonparametric regression.* Springer-Verlag, New York.

[113] Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation* 10(6) 1455-1480.

[114] Grenander, U. (1981). *Abstract inference.* John Wiley & Sons, New York.

[115] Griffel, D.H. (1981). *Applied functional analysis.* Dover Publications, INC. New York.

[116] Gu, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* 88, 495-503.

[117] Hall., P. (1989). Antithetic resampling for the bootstrap. *Biometrika*, 76, 713-724.

[118] Hall, P. (1990). Using the bootstrap to estimate Mean Squared Error and Select Smoothing Parameters in nonparametric problems. *Journal of Multivariate Analysis*, 92, 177-203.

[119] Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* 22, 215-232.

[120] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[121] Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.* 20, 695-711.

[122] Hall, P., Presnell, B. (1999). Intentionally biased bootstrap methods. *J. Royal Statist. Soc. Ser. B* 61 143-158.

[123] Hampel, F.R. (1968). *Contributions to the Theory of Robust Estimation*, Ph.D. Thesis, University of California, Berkeley.

[124] Hampel, F. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* 42, 1887–1896.

[125] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *J. Amer. Statist. Assoc.* 69, 383-393.

[126] Hampel, F.R., Ronchetti, E.M. Rousseeuw, P.J., Stahel, W.A. (1986). *Robust Statistics, the approach based on Influence Functions*, New York, Wiley.

[127] Hampel, F.R. (1994). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* 62, 1179-1218.

[128] Halmos, P.R. (1950). *Measure theory*, D. Van Nostrad, New York.

[129] Hammersley, J.M., Handscomb, D.C. (1964). *Monte Carlo methods.* London: Methuen.

[130] Hannon, E.J., Quinn, B.G. (1979). The determination of the order of autoregression. *Journal of the Royal Statistical Society*, Series B 41,190-195.

[131] Hardle, W. (1989). Resampling for inference from curves. *Proceedings of the 47th Session of International Statistical Institute*, 59-69, Paris.

[132] Härdle, W. (1990). *Applied Nonparametric Regression, Econometric Society Monographs*, Cambridge University Press.

[133] HardIe, W., Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. J. Amer. Statist. Assoc. 89, 102-110.

[134] Härdle, W., Chen, R. (1995). Nonparametric time series analysis, a selective review with examples. *In Proceedings of the 50th ISI Session* (Beijing, 1995) vol. 1, 375-394.

[135] Hardle, W., Kerkyacharian, G., Pickard, D., Tsybakov, A. (1998). *Wavelets, approximations and statistical applications.* Lecture notes in Statistics 129. Springer-Verlag, New York.

[136] Hardle, W., Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* 21(4), 1926-1947.

[137] Härdle, W., Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, 13, 1465-1481.

[138] Hardle, W., Marron, S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics* 19, 778- 796.

[139] Hardly, G., Kemp, C.M. (1971). *Variational Methods in Economics.* American Elsvier, New York.

[140] Hart, D., Wehrly, T .E. (1992). Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models. J. Amer. Statist. Assoc. 87, 1018-1024.

[141] Hastie, T., Tibshirani, R. (1990). *Generalized additive models.* Chapman and Hall. London.

[142] Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*, Springer-Verlag, Heidelberg.

[143] He, X. (1991). A local breakdown property of robust tests in linear regression. *J. of Multivariate Analysis.* 38, 294-305.

[144] Hinkley, D.V., Shi, S. (1989). Importance sampling and the nested bootstrap. *Biometrika* 76, 435-446.

[145] Hjort, N.L., Jones, M.C. (1996). Locally parametric density estimation. *Ann. Statist.*, 24, 1619-1647.

[146] Hoaglin, D. C., Mosteller, F., Tukey, J. W. (1983). (Eds.) *Understanding Robust and Exploratory Data Analysis.* Wiley, N. Y.

[147] Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55-67.

[148] Hoerl, A.E., Kennard, R.W., Baldwin, K.F. (1975). Ridge regression: Some simulations. *Communications in Statistics*, 4, 105-123.

[149] Hoerl, R.W., Schuenemeyer, J.H., Hoerl, A.E., (1986). A simulation of biased estimation and subset selection regression techniques. *Technometrics*, 28, 369-380.

[150] Hogg, R.V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.*, 348, 909-927.

[151] Hooper, P.M., (1993). Iterative weighted least squares estimation in heteroscedastic linear models. *J. Amer. Statist. Assoc.* 88, 179-184.

[152] Horn, P. (1981). Heteroscedasticity of residuals: A nonparametric alternative to the Goldfeld-Quandt test. *Communications in Statistics A*, 10, 795-808.

[153] Huber, P.J. (1964). Robust Estimation of a Location Parameter, *Ann. Math. Statist.*, 35, 73-101.

[154] Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* 36, 1753–1758.

[155] Huber, P. J. (1968). Robust confidence limits. *Z. Wahrsch. verw. Geb.* 10, 269–278.

[156] Huber, P. J., Strassen, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* 1, 251–263. Corr: 2, 223–224.

[157] Huber, P.J. (1981). *Robust statistics.* John Wiley & Sons, INC.

[158] Hurvich, C.M., Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.

[159] Hwang, J.T., Ding, A.A. (1997). Prediction intervals for artificial neural networks. *J. Amer. Statist. Assoc.* 92, 748-757.

[160] Ivanov, V.V. (1962). On linear problems which are not well-posed. *Soviet Math. Dokl.* 3, 4, 981-983.

[161] Jaeckel, L. (1971). Robust estimation of location: symmetry and asymmetric contamination, *Annals of Mathematical Statistics*, 42, 1020-1034.

[162] Jeffreys, H. (1948). *Theory of Probability.* Clarendon Press, Oxford., 1939. Later editions: 1948, 1961, 1983.

[163] Johns, M.V. (1988). Importance sampling for bootstrap confidence intervals. *J. Amer. Statist. Assoc.* 83, 709-714.

[164] Johnson, N., Kotz, S., Balakrishnan, N. (1997). *Discrete multivariate distributions.* John Wiley & Sons, INC.

[165] James, W., Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the fourth Berkeley Symposium on Mathematical and Probability.* Berkeley: University of California Press. 361-379.

[166] Judge, G.G., Griffiths, W.E., Carter, H.R. (1980). *The theory and practice of econometrics.* John Wiley & Sons, INC.

[167] Jurečková, J., Sen, P. K. (1996). *Robust Statistical Procedures, Asymptotics and Interrelations.* John Wiley & Sons, INC.

[168] Kaplan, E.l., Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53, 457-481

[169] Kendall, M.G. (1961). *A course in the geometry of n dimensions.* Griffin's Statistical Monographs and Courses.

[170] Khorasani, F., Milliken, G.A. (1982), Simultaneous confidence bands for nonlinear regression models. *Comm. in Statist.* - Theory and Method, 11(11), 1241-1253

[171] Kohler, M., Krzyzak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transaction on Information Theory*, 47, 3054-3058.

[172] Lavrentiev, M.M. (1962). On ill-posed problems of mathematical physics. *Novosibirsk*, (in Russian).

[173] Lee, A.J. (1990). *U-Statistics, Theory and Practice.* Marcel Dekker, New York.

[174] Legendre, A.M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes. *Mme. Courcier*, Paris.

[175] Lehmann, E.L. (1986). *Testing statistical hypotheses.* 2nd edition, Wiley & Sons, INC.

[176] Liu, R. (1988). Bootstrap procedure under some non i.i.d. models. *Annals of Statistics* 16, 1696-1708.

[177] Liu, R., Singh, K. (1992). Efficiency and robustness in resampling. *Annals of Statistics* 20, 970-984.

[178] Ljung, L. (1987). System Identification, Theory for the User. Prentice Hall.

[179] Loader, C. (1996). Local likelihood density estimation. *Ann. Statist.*, 24, 1602-1618.

[180] Loader, C. (1999). *Local regression and likelihood.* Springer-Verlag, New York.

[181] Lugosi, G., Nobel, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.* 24, 687-706..

[182] Mallows, C.L. (1973). Some comments on C_p, *Technometrics* 15, 661-675.

[183] Magdon-Ismail, M., Atiya, A. (2002). Density estimation and random variate generation using multilayer networks. *IEEE Trans. Neural Networks*, 13,

[184] Mammen, E. (1992a). *When does bootstrap work: asymtotic results and simulations.* Lecture Notes in Statistics 77. Springer New York Heidelberg Berlin.

[185] Mammen, E. (1992b). Bootstrap, wild bootstrap, and asymptotic normality. Prob. Theory and Rei. Fields 99, 499-455.

[186] Marazzi, A. (1993). *Algorithms, routines, and S functions for robust statistics.* Wadsworth, Inc., Belmont, California.

[187] Marazzi, A., Ruffieux, C. (1996). *Implementing M-estimators of the Gamma Distribution*, Lecture Notes in Statistics, vol. 109. Springer-Verlag, Heidelberg.

[188] Marron, J.S. (1987). What does optimal bandwidth selection means for nonparametric regression estimation *In Statistical Data Analysis Based on the L_1-Norm and Related Methods* (Ed. Dodge, Y.), 379-391. North Holland, Amsterdam.

[189] Marron, J.S. (1989). Automatic smoothing parameter selection: a survey, *Empirical Econom.*, 13, 187-208.

[190] McCullagh, P., Nelder, J.A. (1989). *Generalized linear models.* 2nd ed. London: Chapman & Hall.

[191] McDonald, G.C., Galarneau, D.I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *J. Amer. Statist. Assoc.*, 70, 407-416.

[192] Michel, A.N., Herget, C.J. (1981). *Applied algebra and functional analysis.* Dover Publications, INC. New York.

[193] Miller, G., Horn, D. (1998). Probability density estimation using entropy maximization. *Neural Comput.*, 10, 1925-1938.

[194] Montgomery, D.C., Peck, E.A., and Vinning, G.G. (2001). *Introduction to linear regression analysis*, 3rd ed., John Wiley, New York.

[195] Mood, A.M., Graybill, F.A., and Boes, D.C. (1974). *Introduction to the theory of statistics*, 3rd edition. New York: McGraw-Hill.

[196] Morgenthaler, S., Ronchetti, E., Stahel, W. A. (Eds.) (1993). *New directions in statistical data analysis and robustness.* Birkhäuser Verlag, Basel.

[197] Morgenthaler, S., Tukey, J. W. (Eds.) (1991). *Configural polysampling: A route to practical robustness.* Wiley, N. Y.

[198] Mosteller, F., Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics.* Addison-Wesley, Reading, Mass.

[199] Müller, H.G. (1988). *Nonparametric regression analysis of longitudinal data.* Springer-Verlag. New York.

[200] Müller, H.G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika*, 78, 521-590.

[201] Müller, C.G. (1997). *Robust planning and analysis of experiments.* Springer-Verlag. New York.

[202] M ller, U.U., A. Schick and W. Wefelmeyer (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics* 37(3), 179–188.

[203] Myers, R.H. (1990). *Classical and modern regression with applications*, 2nd ed., Duxbury Press, Boston.

[204] Nadaraya, E.A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability* 10, 186-190.

[205] Neumann, M.H. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* 23, 1937-1959.

[206] Niederreiter, H. (1992). Random number generation and quasi-Monte Carlo methods. *CBMS-NSF Regional Conference Series in Applied Mathematics.* SIAM, Philadelphia.

[207] Nikolova, M. (1999). Local strong homogeneity of a regularized estimator. *SIAM Journal on Applied Mathematics* 61, 633–658.

[208] Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33, 1065-1076.

[209] Pelckmans, K., De Brabanter, J., Suykens, J.A;K., De Moor, B. (2003). Variogram based noise variance estimation and its use in Kernel Based Regression. *Proc. of the IEEE Workshop on Neural Networks for Signal Processing*, pp. 199-208.

[210] Pelckmans, K., De Brabanter, J., Suykens, J.A.K., De Moor, B. (2004). The differogram : Nonparametric noise variance estimation and its use for model selection, Internal Report 04-41, ESAT-SISTA, K.U.Leuven.

[211] Pelckmans, K., Goethals, I., De Brabanter, J., Suykens, J.A.K., De Moor, B. (2004). Componentwise Least Squares Support Vector Machines", Internal Report 04-75, ESAT-SISTA, K.U.Leuven.

[212] Penrose, K., Nelson, A., Fisher, A. (1985). Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques, *Medicine and Science in Sports and Excercise*, 17(2), 189.

[213] Pearson, K. (1902). On the mathematical theory of errors of judgement, with special reference to the personal equation. *Philos. Trans. Roy. Soc. Ser. A* 198, 235–299.

[214] Pearson, E. S. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika* 21, 259–286.

[215] Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika* 23, 114–133.

[216] Pfanzagl, J. (1969). On measurability and consistency of minimum contrast estimates, *Metrika*, 14, 248-278.

[217] Philips, D.L. (1962). A technique for the numerical solution of integral equations of the first kind. *J. Assoc. Comput. Machinery.* 9, 84-97.

[218] Poggio, T. and Girosi, F. (1990). Networks for approximation and learning, *Proceedings of the IEEE*, 78, 1481-1497.

[219] Politis, D.N., Romano, J.P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* 22, 2031-2050.

[220] Ponnusamy, S. (2002). *Foundations of functional analysis.* Alfa Science International Ltd. England.

[221] Priebe, C.E. (1994). Adaptive mixtures. *Journal of the American Statistical Association.* 89, 796-806.

[222] Quenouille, M. (1949). Approximate tests of correlation in times series. *J. Roy. Statist. Soc.* Ser. B, 11, pp.18-84.

[223] Rao, B.L.S.P. (1983). *Nonparametric functional estimation.* Academic Press.

[224] Rieder, H. (1994). *Robust asymptotic statistics.* Springer, N. Y., 1994.

[225] Rice, J.A. (1984). Boundary modification for kernel regression. *Communication in Statistics* 12, 899-900.

[226] Ripley, B.D. (1987). *Stochastic simulations.* John Wiley & Sons, INC.

[227] Roeder, K., Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association. 92, 894-902.

[228] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* 27, 832-835.

[229] Rousseeuw, P.J., Leroy, A. (1987). *Robust regression and outlier detection.* John Wiley & Sons, INC.

[230] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimation, *Scand. J. Statist.* 9 65-78.

[231] Salibian-Barrera, M., Zamar, R.H. (2000). Contributions to the Theory of Robust Inference. Unpublished Ph.D. Thesis. University of British Columbia, Department of Statistics, Vancouver, BC. Available on-line at http://mathstat.carleton.ca/˜matias/pubs.html

[232] Sargant, D.J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* 54, 1486-1497.

[233] Saunders, C., Gammerman, A., Vovk, V. (1998). Ridge regression learning algorithm in dual varables. Proc. of the 15th Int. Conf. on Machine learning (ICML '98), Morgan Kaufmann, 515-521.

[234] Schwartz, G. (1979). Estimating the dimension of a model, *Ann. of Statist.* 6, 461-464.

[235] Scott, D.W. (1992). *Multivariate density estimation. Theory, practice and visualization.* John Wiley & Sons, INC.

[236] Sen, A., Srivastava, M. (1997). *Regression analysis: Theory, methods and applications.* Springer-Verlag.

[237] Serfling, R.J. (1980). *Approximation theorems of mathematical statistics.* John Wiley & Sons, INC.

[238] Serfling, R.J. (1984). Generalized L-, M-, and R-statistics, *Ann. Statist.* 12, 76-86.

[239] Shen, X., Wong, W.H., (1994). Convergence rate of sieve estimates. *Ann. Statist.* 22, 580-615.

[240] Silverman, B.W., (1986). *Density estimation.* Chapman & Hall.

[241] Singh, K. (1998). Breakdown theory for bootstrap quantiles. *Ann. Statist.* 26, 1719-1732.

[242] Smith, P.J. (1996). Renovating interval-censored responses. *Lifetime data Analysis, 2, 1-11.*

[243] Solka, J.L., Poston, W. L., Wegeman, E.J. (1995). A visualization technique for studying the iterative estimation of mixture densities. Journal of Computational and Graphical Statistics 4, 180-198.

[244] Stahel, W. A., Weisberg, S. (1991). Eds. *Directions in robust statistics and diagnostics.* vol. 1, 2. Springer, N. Y.

[245] Staudte, R. G., Sheather, S. J. (1990). *Robust estimation and testing.* Wiley, N. Y.

[246] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the third Berkeley Symposium on Mathematical and Probability.* Berkeley: University of California Press. 197-206.

[247] Stigler, S.M. (1973). The asymptotic distribution of the trimmed mean, *Ann. Statist.* 1, 472-477.

[248] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statist. Soc. Ser. B* 36 111-147.

[249] Stone, C.J., Koo, C.Y. (1986). Log-spline density estimation, function estimates. *Contemp. Math.* 59, AMS, Providence, pp. 1-15.

[250] Stromberg, A.J. (1997). Robust covariance estimates based on resampling. *J. of Statist. Planning and Inference.* 57, 321-334.

[251] "Student"(W. S. Gosset). (1927). Errors of routine analysis. *Biometrika* 19, 151–164.

[252] Suykens, J.A.K., Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters.* Vol. 9(3), 293-300.

[253] Suykens, J.A.K., De Brabanter, J., Lukas, L., Vandewalle, J. (2002a). Weighted least squares support vector machines : robustness and sparse approximation, *Neurocomputing*, Special issue on fundamental and information processing aspects of neurocomputing, 48, 1-4, 85-105.

[254] Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J. (2002b). *Least Squares Support Vector Machines*, World Scientific, Singapore.

[255] Tibshirani, R., Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.*, 82, 559-567.

[256] Tibshirani, R.J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* (58), 267–288.

[257] Tibshirani, R.J. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine* (16), 385–395.

[258] Tikhonov, A. (1943). On the stability of inverse problem. *Dokl. Acad. Nauk.* USSR 39, 5

[259] Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regulariation method. *Soviet. Math. Dokl.* 5, 1035-1038.

[260] Thisted, R.A. (1978). On generalized ridge regression. *University of Chicago. Departement of Statistics Technical report 57, May.*

[261] Thompson, J.R., Tapia, G.F. (1990). *Nonparametric function estimation, modeling and simulation.* SIAM, Philadelphia.

[262] Traven, H. (1991). A neural network approach to statistical pattern classification by semiparametric estimation of probability density functions. *IEEE Trans. Neural Networks,* Vol 2, 366-377.

[263] Tsiatis, A.A. (1975). A nonidentifiability aspect of the problem of computing risks. *Proc. Nat. Acad. Sci.*, 72, 20-22.

[264] Tukey, J. (1958). Bias and confidence in not quite large samples, Abstract. *Ann. Math. Statist.* 29, p614.

[265] Tukey, J. W. A (1960). *Survey of sampling from contaminated distributions.* In Contributions to Probability and Statistics., (Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B. Eds.) Stanford University Press, Stanford, Calif., p. 448–485.

[266] Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* 33, 1–67.

[267] van der Vaart, A.W. (1998). *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

[268] van de Geer, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *Nonparametric Statistics.* 6, 293-310.

[269] Vapnik, V.N., Lerner, A. (1963). Pattern recognition using generalized portrait method. *Autom. Remote Control.* 24.

[270] Vapnik, V.N., Chervonenkis, A. Ya. (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Soviet Math. Dokl.* 9, 915-918.

[271] Vapnik, V.N., Chervonenkis, A.Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Prob. Appl.* 16, 264-280.

[272] Vapnik, V.N., Chervonenkis, A.Ya. (1974). *Theory of pattern recognition.* Nauka, Moscow.

[273] Vapnik, V.N. (1979). *Estimation of dependencies based on empirical data.* Nauka, Moscow. (English translation: Vapnik, V.N. (1982). *Estimation of dependencies based on empirical data.* Springer-Verlag, New York*).*

[274] Vapnik, V.N. (1995). *The nature of statistical learning theory.* Springer-Verlag, New York.

[275] Vapnik, V.N. (1999). *Statistical Learning Theory*, John Wiley & Sons, INC.

[276] Vapnik, V.N., Mukherjee, S. (1999). Support vector method for multivariate density estimation. *Neural Information Processing Letters.* 12, 659-665.

[277] Vergote I., De Brabanter J., Fyles A., Bertelsen K., Einhorn N., Sevelda P., Gore M.E., Karn J., Verrelst H., Sjovall K., Timmerman D., Vandewalle J., Van Gramberen M., Trope C.G. (2001). Prognostic factors in 1545 patients with stage I invasive epithelial ovarium carcinoma : Importance of degree of differential and cyst rupture in predicting relapse. *The Lancet*, vol. 357, pp. 176-182.

[278] Wand, M.P., Jones, M.C. (1995). *Kernel smoothing.* Chapman & Hall.

[279] Wahba, G., (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc.* Ser. B, 40, 364-372.

[280] Wahba, G., (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. Roy. Statist. Soc.* Ser. B, 45, 133-150.

[281] Wahba, G., Wold, S. (1975). A completely automatic french curve: fitting spline functions by cross-validation. *Comm. Statist.* 4 1-17.

[282] Wahba, G. (1990). *Spline models for observational data.* IMS, Hayward, CA.

[283] Yang, Y., Zheng, Z. (1992). Asymptotic properties for cross-validated nearest neighbour median estimates in non-parametric regression: the L_1-view. *Probability and Statistics*, 242-257.

[284] Wang, S.J. (1993) Saddlepoint expansions in finite populations problems. *Biometrika* 80, 583-590.

[285] Walter, G.G., Ghorai, J. (1992). Advantages and disadvantages of density estimation with wavelets. *In Computing Science and Statistics. Proceedings of the 24rd Symposium on the interface*, 234-243.

[286] Watson, G.S. (1964). Smooth regression analysis. *Sankhya* 26:15, 175-184.

[287] Welsh, A.H., Carroll, R.J., Ruppert, D. (1994). Fitting heteroscedastic regression models. *J. Amer. Statist. Assoc.* 89, 100-116.

[288] Wichern, D.W., Churchill, G.A. (1978). A comparision of ridge estimators. *Technometrics*, 20 301-311.

[289] Whittaker, E. (1923). On new method of graduation. *Proc. Edingburgh Math. Soc.* 2, 41.

[290] Wong, W.H., Shen, X. (1995). A probability inequality for the likelihood surface and convergence rate of the maximum likelihood estimate. *Ann. Statist.*, 23, 339-362.

[291] Zheng, L., Tu, D. (1988). Random weighting method in regression models. *Scientia Sinica A*, 91, 1442-1459.

Publication List

Book

(1) Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.(2002). *Least Squares Support Vector Machines*, World Scientific Publishing Co., Pte, Ltd. (Singapore).

Contribution to book

(1) Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J. (2003). Support Vector Machines : Least Squares Approaches and Extensions, in Chapter 8 of *Advances in Learning Theory : Methods, Models and Applications*, (Suykens J.A.K., Horvath G., Basu S., Micchelli C. and Vandewalle J., eds.), IOS Press (NATO-ASI Series in Computer and Systems Sciences) (Amsterdam, The Netherlands), 155-178.

International Journal Paper

(1) Suykens, J.A.K., De Brabanter, J., Lukas, L., Vandewalle, J. (2002). Weighted least squares support vector machines : robustness and sparse approximation, *Neurocomputing*, Special issue on fundamental and information processing aspects of neurocomputing, vol. 48, no. 1-4, 85-105.

(2) Vergote, I., De Brabanter, J., Fyles, A., Bertelsen, K., Einhorn, N., Sevelda, P., Gore, M.E., Karn, J., Verrelst, H., Sjovall, K., Timmerman, D., Vandewalle, J., Van Gramberen, M., Trope, C.G., (2001). Prognostic factors in 1545 patients with stage I invasive epithelial ovarium carcinoma : Importance of degree of differential and cyst rupture in predicting relapse, *The Lancet*, vol. 357, 176-182.

(3) Timmerman, D., Deprest, J., Verbesselt, R., Moerman, P., De Brabanter, J., Vergote, I. (2000). Absence of correlation between risk factors for endometrial cancer and the presence of tamoxifen-associated endometrial polyps in postmenopausal patients with breast cancer. *European Journal of Cancer*, vol. 36, S39-S41.

(4) Vergote, I., Timmerman, D., De Brabanter, J., Fyles, A. Trope, C.G. (2001). Cyst rupture during surgery. *Lancet*, no. 358, 72-73.

(5) Van Den Bosch, T., Van Schoubroeck, D., Ameye, L., De Brabanter, J., Van Huffel, S., Timmerman, D. (2003). Ultrasound assessment of endometrial thickness and endometrial polyps in women on hormonal replacement therapy. *American Journal of Obstetrics and Gynecology*, vol. 188, no. 5, 1249-1253.

(6) Van den Bosch, T., Van Schoubroeck, D., Lu, C., De Brabanter, J., Van Huffel, S., Timmerman, D. (2002). Color Doppler and Gray-Scale Ultrasound Evaluation of the Postpartum uterus. *Ultrasound Obstet. Gynecol.*, vol. 20, 586-591.

(7) Van Den Bosch, T., Van Schoubroeck, D., Ameye, L., De Brabante,r J., Van Huffel, S., Timmerman, D. (2002). The influence of hormone replacement therapy on the ultrasound features of the endometrium : a prospective study. *European Journal of Cancer*, vol. 38, no. 6, S78-S79.

(8) Van den Bosch, T., Donders, G., Riphagen, I., Debois, P., Ameye, L., De Brabanter, J., Van Huffel, S., Van Schoubroeck, D., Timmerman, D. (2002). Ultrasonographic features of the endometrium and the ovaries in women on etonogestrel implant. *Ultrasound Obstet. Gynecol.*, vol. 20, 377-380.

(9) Marchal, K., Engelen, K., De Brabanter, J., De Moor, B.(2002). A guideline for the analysis of two sample microarray data. *Journal of Biological Systems*, vol. 10, no. 4, 409-430.

(10) Timmerman, D., De Smet, F., De Brabanter, J., Van Holsbeke, C., Jermy, K., Moreau, Y., Bourne, T., Vergote, I. (2003). OC118 : Mathematical models to evaluate ovarian masses - can they beat an expert operator ?. *Ultrasound in Obstetrics and Gynecology*, vol. 22, no. S1, 33.

(11) De Brabanter, J., Vergote, I., Verrelst, H., Timmerman, D., Van Huffel, S., Vandewalle, J. (1999). Univariate and multivariate regression analyses : some basic statistical principles. *CME J. Gynecol. Oncol.*, no. 4, 262-270.

(12) Amant, F., De Brabanter, J. (2004). A possible role of the cytochrome P450c17alfa gene (CYP17) polymorphism in the pathobiology of uterine leiomyomas from black South African women: a pilot study. *Acta Obst. et Gynecol. Scand.*

International Conference Paper

(1) Van Gestel, T., Suykens, J., De Brabanter, J., De Moor, B., Vandewalle, J. (2001). Least squares support vector machine regression for discriminant analysis, *in Proc. of the International Joint Conference on Neural Networks (IJCNN'01)*, Washington DC, USA, 2445-2450.

(2) Ameye, L., Lu, C., Lukas, L., De Brabanter, J., Suykens, J.A.K., Van Huffel, S., Daniels, H., Naulaers, G., Devlieger, H. (2002). Prediction of mental development of preterm newborns at birth time using LS-SVM, *in Proc. of the European Symposium on Artificial Neural Networks (ESANN'2002)*, Bruges, Belgium, 167-172.

(3) Van Gestel, T., Suykens, J., De Brabanter, J., De Moor, B., Vandewalle, J. (2001). Kernel Canonical Correlation Analysis and Least Squares Support Vector Machines, *in Proc. of the International Conference on Artificial Neureal Networks (ICANN 2001)*, Vienna, Austria, 381-386.

(4) Lu, C., De Brabanter, J., Van Huffel, S., Vergote, I., Timmerman, D. (2001). Using Artificial Neural Networks to Predict Malignancy of Ovarian Cancers, *in Proc. of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2001)*, Istanbull, Turkey.

(5) Devos, A., Bergans, N., De Brabanter, J., Vanhamme, L., Vanstape,l F., Van Hecke, P., Van Huffel, S. (2002). Model selection for quantification of a multi-exponential MRS signal, *in Proc. of the 2nd European Medical and Biological Engineering Conference (EMBEC'02)*, Vienna, Austria, 1066-1067.

(6) Ameye, L., Van Huffel, S., De Brabanter, J., Suykens, J., Spitz, B., Cadron, I., Devlieger, R., Timmerman, D. (2002). Study of rupture of membranes before 26 weeks of gestation, *in Proc. of the 2nd European Medical and Biological Engineering Conference (EMBEC'02)*, Vienna, Austria, 760-761.

(7) De Brabanter, J., Pelckmans, K., Suykens, J.A.K., De Moor, B., Vandewalle, J. (2003). Robust complexity criteria for nonlinear regression in NARX models, *in Proc. of the 13th System Identification Symposium (SYSID2003)*, Rotterdam, Nederland, 79-84.

(8) De Brabanter, J., Pelckmans, K., Suykens, J.A.K., Vandewalle, J. (2002). Robust cross-validation score function for non-linear function estimation, in *Proc. of the International Conference on Artificial Neural Networks (ICANN 2002)*, Madrid, Spain, 713-719.

(9) Pelckmans, K., De Brabanter, J., Suykens, J.A.K., De Moor, B. (2003). Variogram based noise variance estimation and its use in Kernel Based Regression, in *Proc. of the IEEE Workshop on Neural Networks for Signal Processing*, 199-208.