Kernel Tricks, Means and Ends

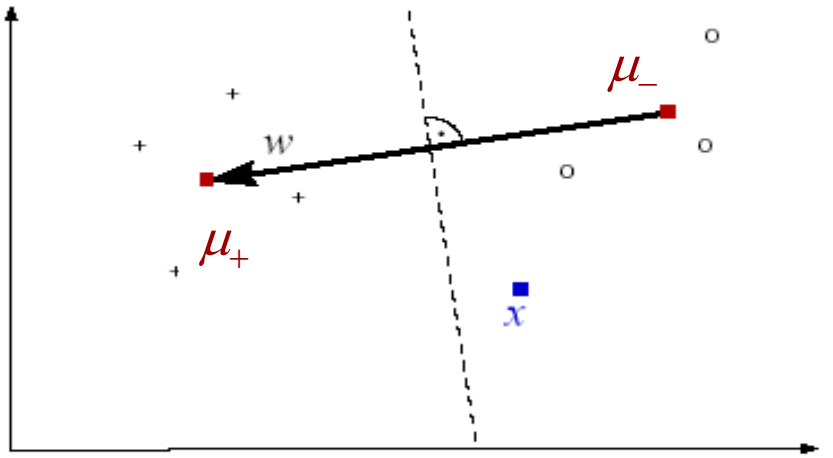
Bernhard Schölkopf
Max Planck Institute for Biological Cybernetics
Tübingen, Germany

Empirical Inference Department
http://www.kyb.tuebingen.mpg.de/bs

Example of a Pattern Recognition Algorithm

Idea: classify points x according to which of the two class means is closer.

$$\mu_+ := \frac{1}{m_+} \sum_{y_i=1} x_i, \quad \mu_- := \frac{1}{m_-} \sum_{y_i=-1} x_i$$



- Decision function: hyperplane with normal vector $w := \mu_+ - \mu_-$
- How about problems that are not linearly separable?

Kernel Feature Spaces

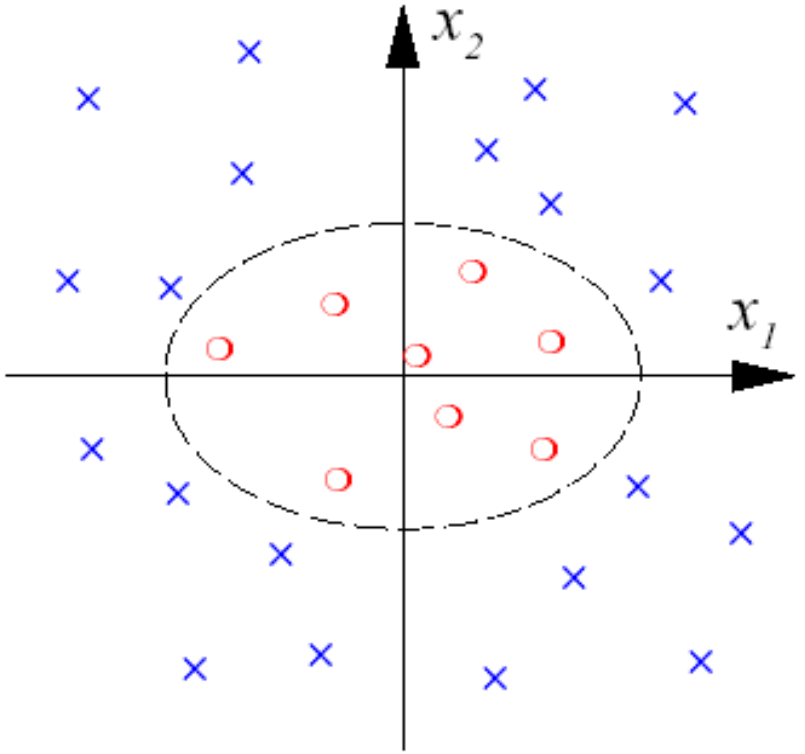
Preprocess the inputs with

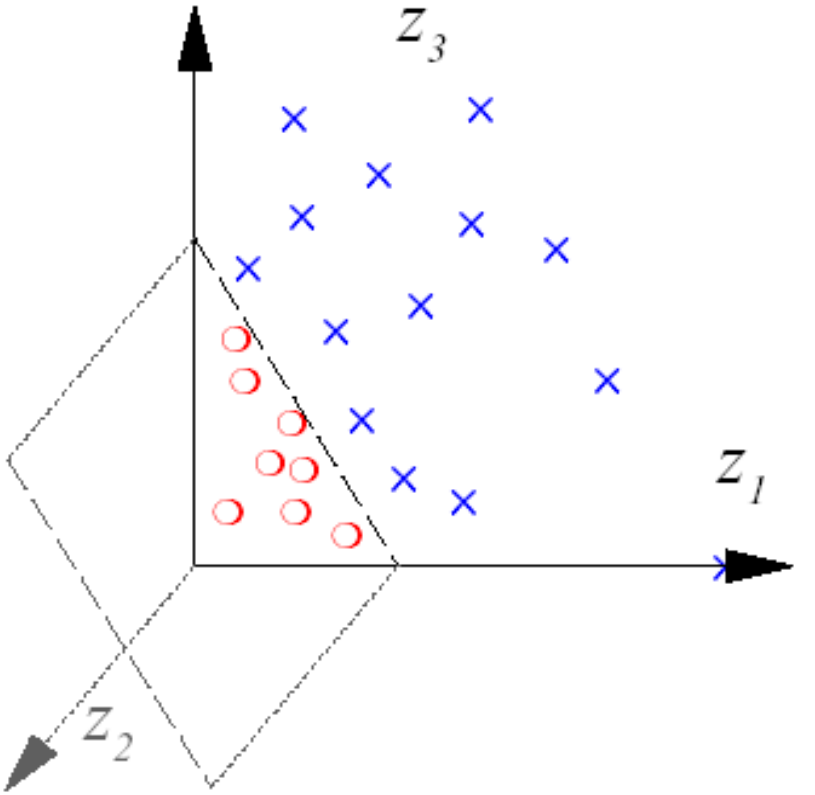
$$\Phi : \mathcal{X} \to \mathcal{H}$$
$$x \mapsto \Phi(x),$$

where $\mathcal{H}$ is a dot product space, and learn the mapping from $\Phi(x)$ to y.

Example: All Degree 2 Monomials

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$





The Kernel Trick

$$\begin{aligned}\langle \Phi(x), \Phi(x') \rangle &= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)(x'^2_1, \sqrt{2}\, x'_1 x'_2, x'^2_2)^\top \\ &= (x_1 x'_1 + x_2 x'_2)^2 \\ &= \langle x, x' \rangle^2 \\ &=: k(x, x')\end{aligned}$$

$\longrightarrow$ the dot product in $\mathcal{H}$ can be computed from the dot product in $\mathbb{R}^2$

More generally: for $x, x' \in \mathbb{R}^N$, $d \in \mathbb{N}$,

$$\langle x, x' \rangle^d = \left(\sum_{j=1}^{N} x_j \cdot x'_j \right)^d = \sum_{j_1, \ldots, j_d = 1}^{N} x_{j_1} \cdot \cdots \cdot x_{j_d} \cdot x'_{j_1} \cdot \cdots \cdot x'_{j_d} = \langle \Phi(x), \Phi(x') \rangle$$

More generally: works for *positive definite kernels*

Positive Definite Kernels

Let $\mathcal{X}$ be a nonempty set. The following two are equivalent:

- k is *positive definite (pd)*, i.e., k is symmetric, and for
  - any set of training points $x_1, \ldots, x_m \in \mathcal{X}$ and
  - any $a_1, \ldots, a_m \in \mathbb{R}$

  we have

$$\sum_{i,j} a_i a_j K_{ij} \geq 0, \quad \text{where } K_{ij} := k(x_i, x_j)$$

- there exists a map Φ into a dot product space $\mathcal{H}$ such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

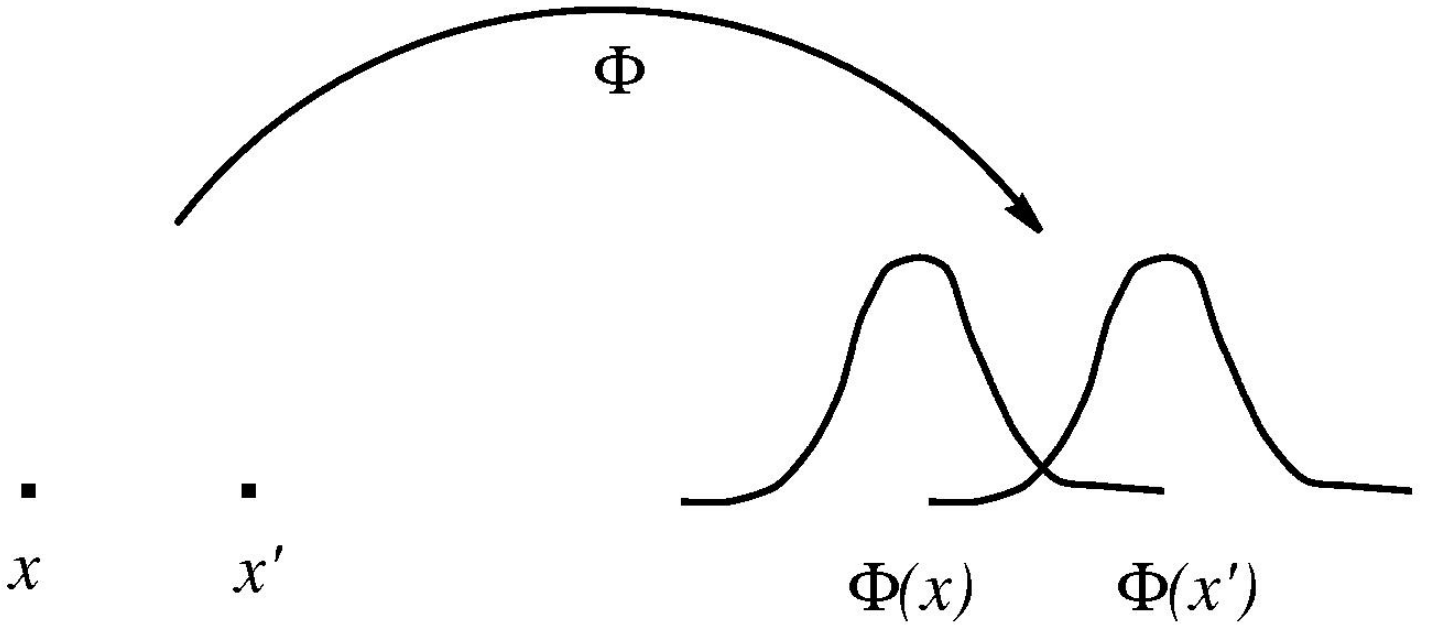
(RKHS)

$\mathcal{H}$ is a so-called *reproducing kernel Hilbert space*.

If for pairwise distinct points, Σ=0 iff all a_i = 0, call *k strictly p.d.*

- Example of a kernel: Gaussian
- Any algorithm that only depends on dot products can be kernelized
- Kernels can be defined on nonvectorial data

The Map into the Reproducing Kernel Hilbert Space



$\Phi(x) = k(x,.)$ *(Aronszajn 1950)*

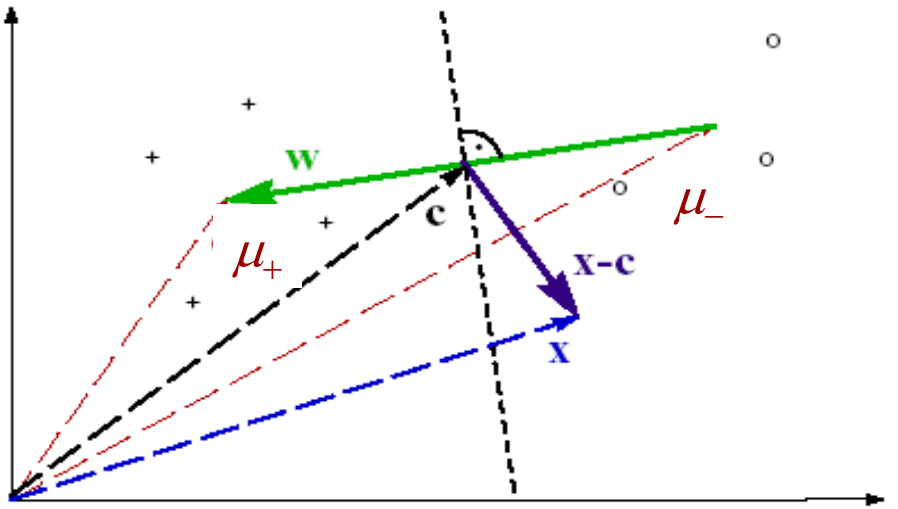
$< \Phi(x), \Phi(x') > = k(x,x')$

Point evaluation: $f(x) = < f, k(x,.) >$.

An Example of a Kernel Algorithm *(Schölkopf & Smola 2002)*

Classify points $\mathbf{x} := \Phi(x)$ *in feature space* according to which of the two class means is closer.

$$\mu_+ := \frac{1}{m_+} \sum_{\{i:y_i=1\}} \Phi(x_i), \quad \mu_- := \frac{1}{m_-} \sum_{\{i:y_i=-1\}} \Phi(x_i)$$



Compute the sign of the dot product between $\mathbf{w} := \mu_+ - \mu_-$ and $\mathbf{x} - \mathbf{c}$.

ctd.

$$f(x) = \operatorname{sgn}\left(\frac{1}{m_+}\sum_{\{i:y_i=1\}}\langle\Phi(x),\Phi(x_i)\rangle - \frac{1}{m_-}\sum_{\{i:y_i=-1\}}\langle\Phi(x),\Phi(x_i)\rangle + b\right)$$

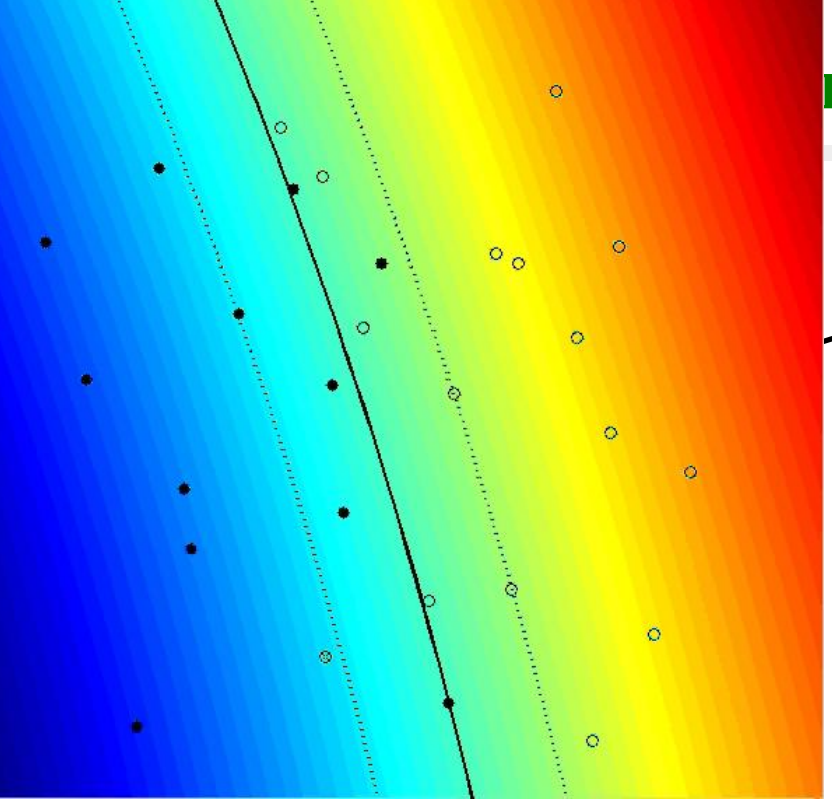
$$= \operatorname{sgn}\left(\frac{1}{m_+}\sum_{\{i:y_i=1\}}k(x,x_i) - \frac{1}{m_-}\sum_{\{i:y_i=-1\}}k(x,x_i) + b\right)$$

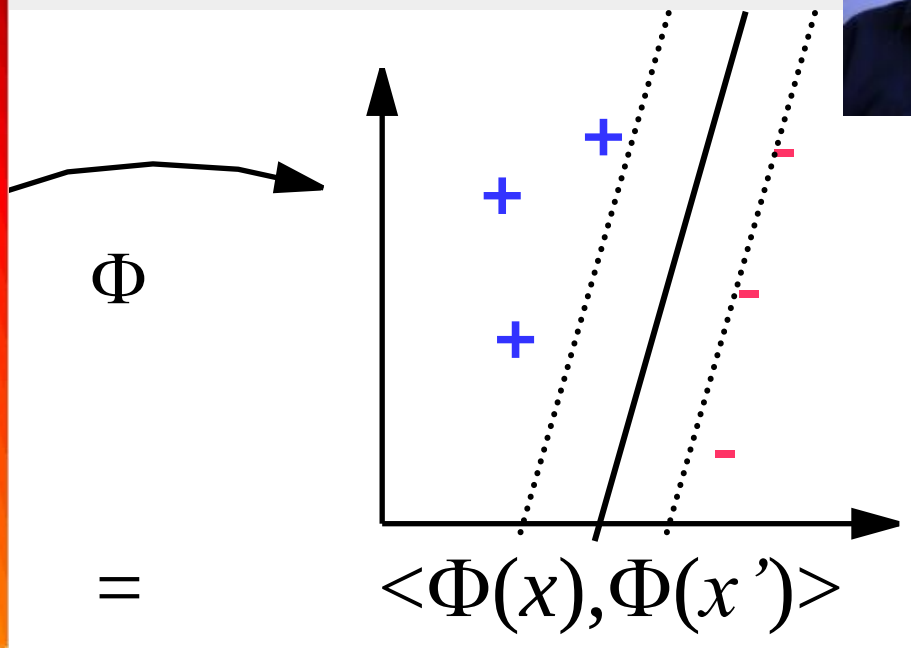
with the constant offset

$$b = \frac{1}{2}\left(\frac{1}{m_-^2}\sum_{\{(i,j):y_i=y_j=-1\}}k(x_i,x_j) - \frac{1}{m_+^2}\sum_{\{(i,j):y_i=y_j=1\}}k(x_i,x_j)\right).$$

If *k* is a density, this is a classifier based on *Parzen windows* plug-in estimates of the two classes.

r Machines in one Slide





- sparse expansion of solution in terms of SVs *(Boser, Guyon, Vapnik 1992):*

$$f(x) = \mathrm{sgn}(\sum_i \lambda_i k(x_i, x) + b)$$

representer theorem *(Kimeldorf & Wahba 1971, Schölkopf et al. 2000)*

- unique solution found by convex QP

Kernel Ends

Implicit Surface Fitting

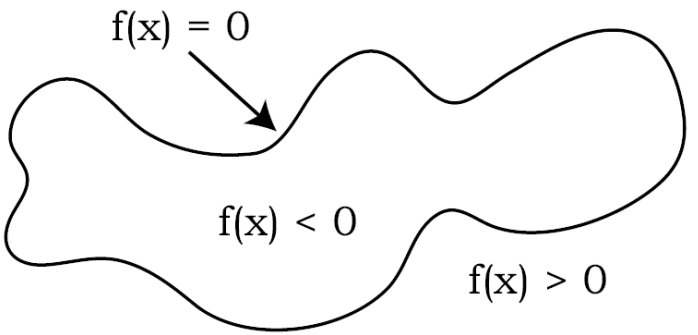
Given a sampling of a surface

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_m \subset \mathbb{R}^d$$

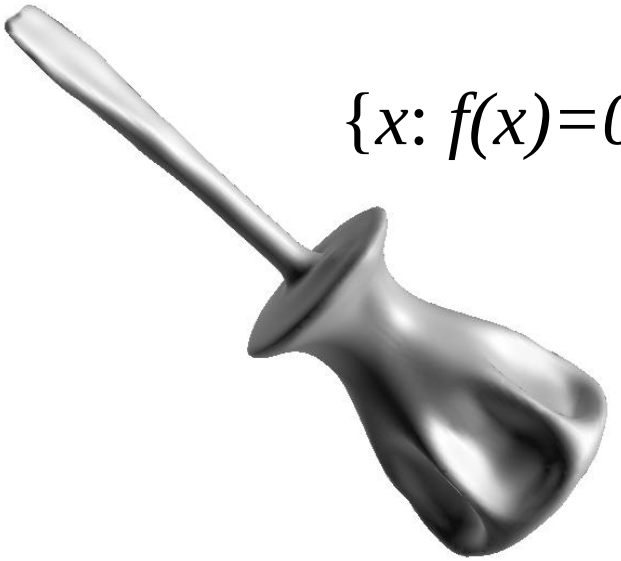
possibly with corresponding surface normals

$$\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots \boldsymbol{n}_m \subset \mathbb{R}^d$$

Construct a function f whose zero level approximates the surface

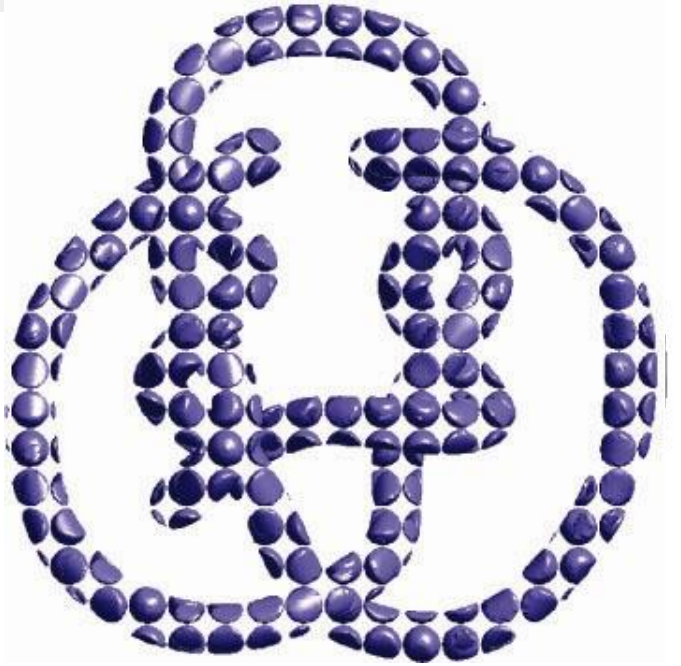


SVM Implicit Surface Approximation



$\{x: f(x)=0\}$

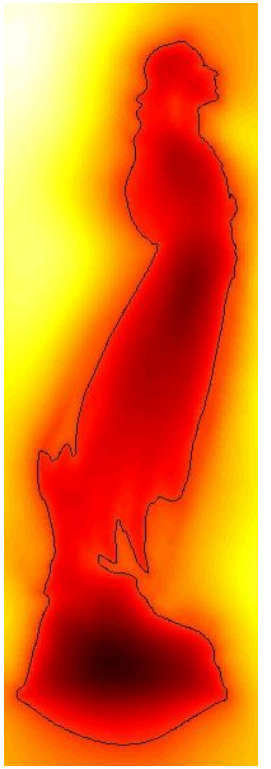
$\min(f_1,f_2)$



Schölkopf, 2004
Schölkopf, Giesen, & Spalinger, 2005
Walder, Chapelle, & Schölkopf, 2005

Steinke, Schölkopf, & Blanz, 2005
Signed distance functions f:
$|f(x)|$ = distance of x to the surface
$\text{sign}(f(x))=1$ iff x is outside the object

Large Scale Example *(Walder et al. 2006)*



***Left**: Rendered model of Lucy, constructed from 14 million points with normals.*
***Middle**: Each of the 364,982 basis function centres*
***Right**: A planar slice that cuts the nose.*

More Examples

Dragon 1: 440K points – decreasing regularisation

Dragon 2: 3.6M points

Thai Statue: 5M points

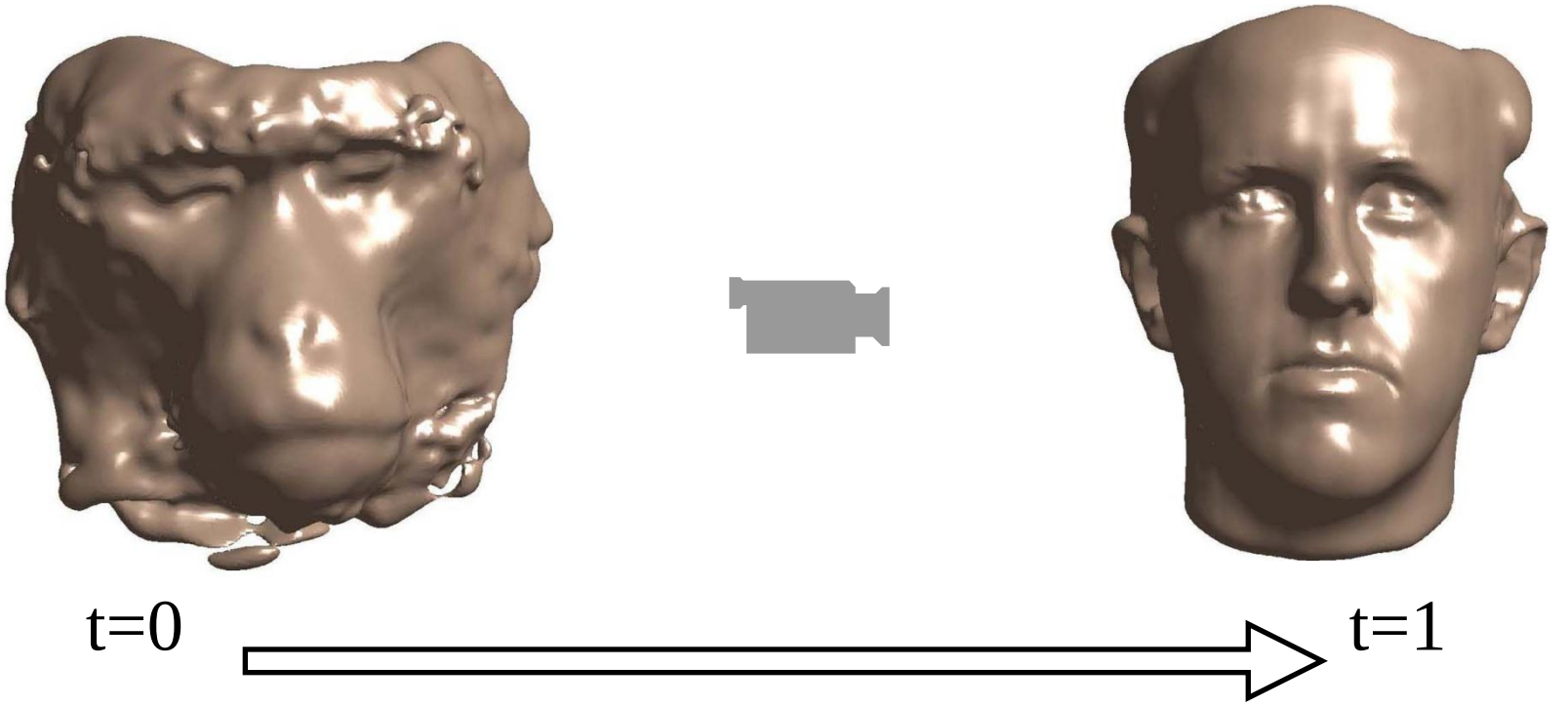
Temporal Scans

- fitted frame-wise in 3D

- fitted in 4D

Shape Interpolation

- We can interpolate shapes by embedding them in four dimensions:



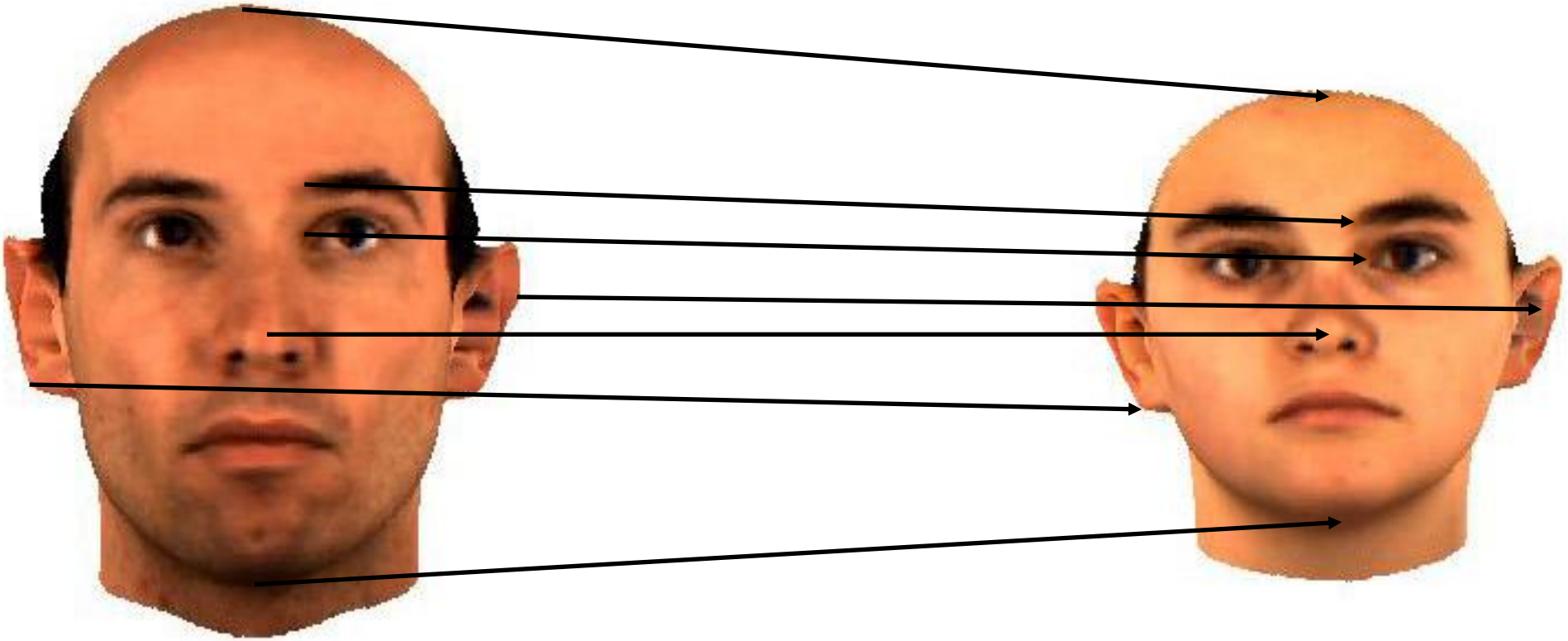
The Morphing Problem

I_1

$\frac{1}{2}(I_1 + I_2)$

I_2

Correspondence



Given a dense *correspondence field* (or *warp*), we can interpolate (and extrapolate) images, almost as in a linear space

(cf. Blanz & Vetter, 1999)

Correspondence via Machine Learning (*Schölkopf, Steinke, Blanz, 2005)*

- Assume the objects O_1 and O_2 live in a domain X. Then the warp is a mapping

  $\tau : X \to X.$

- Assume we are given surface points x_i of the object O_1 and z_i of O_2. *If* they are in correspondence, we have a training set $(x_1,z_1),\ldots,(x_m,z_m)$ and can do regression ("*landmark points*").
- What if they are not in correspondence?
- Main idea: τ should be such that

  O_1 relative to x looks like O_2 relative to $\tau(x)$

- Formalize this as a *locational cost*

  $c\,(O_1, x, O_2, \tau\,(x))$

Locational Cost Functions

feature functions $f_1, f_2 : \quad \mathcal{X} \to \mathbb{R}$

think of f_1, f_2 as the signed distance functions of O_1, O_2.

1. $d(f_1(x), f_2(\tau(x)))^2$

2. $\sum_{i=0}^{\infty} \alpha_i d\left(\nabla^i f_1(x), \nabla^i f_2(\tau(x))\right)^2$

3. If Ψ is the feature map associated with a p.d. kernel on $(\mathcal{O} \times \mathcal{X}) \times (\mathcal{O} \times \mathcal{X})$.

$$\|\Psi(O_1, x) - \Psi(O_2, \tau(x))\|^2$$

Optimization Problem

- Component functions: for $d=1,\ldots,D$,

$$\tau_d(x) = x_d + \langle \mathbf{w}_d, \Phi(x) \rangle$$

- Minimize

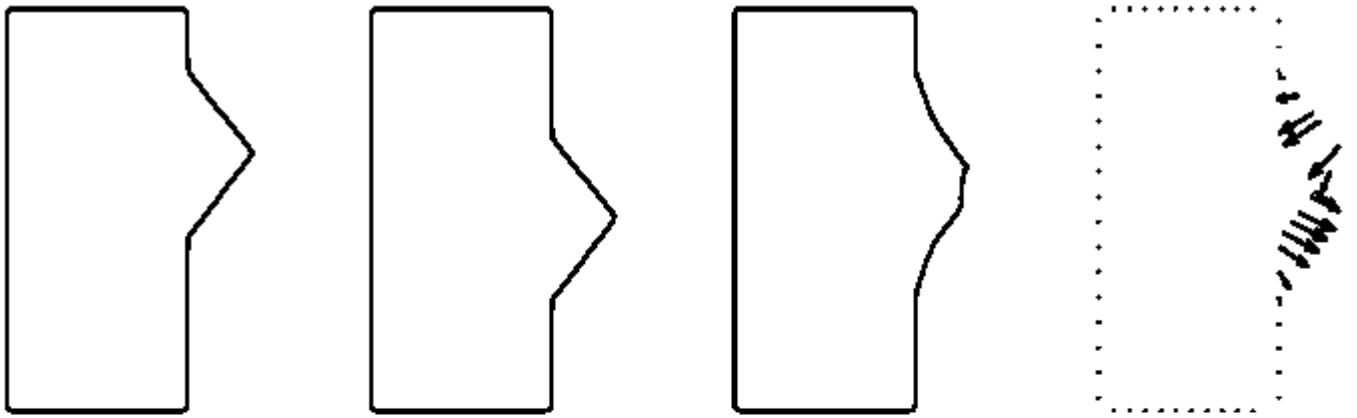
$$\frac{1}{2}\sum_{d=1}^{D} \|\mathbf{w}_d\|^2 + \lambda_p \sum_{i=1}^{m} \|\tau(x_i) - z_i\|^2$$

$$+\lambda_{loc} \int_{\mathcal{X}} c_{loc}(O_1, x, O_2, \tau(x))\, d\mu(x)$$

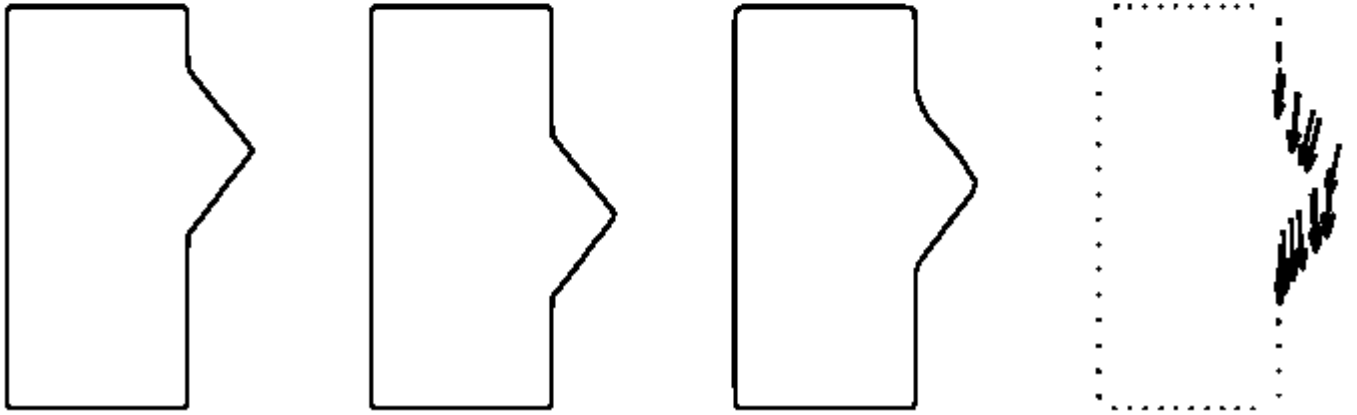
- For $\lambda_{loc} = 0$: D SVR problems with quadratic loss
- in the generic case, nonconvex

Toy Example

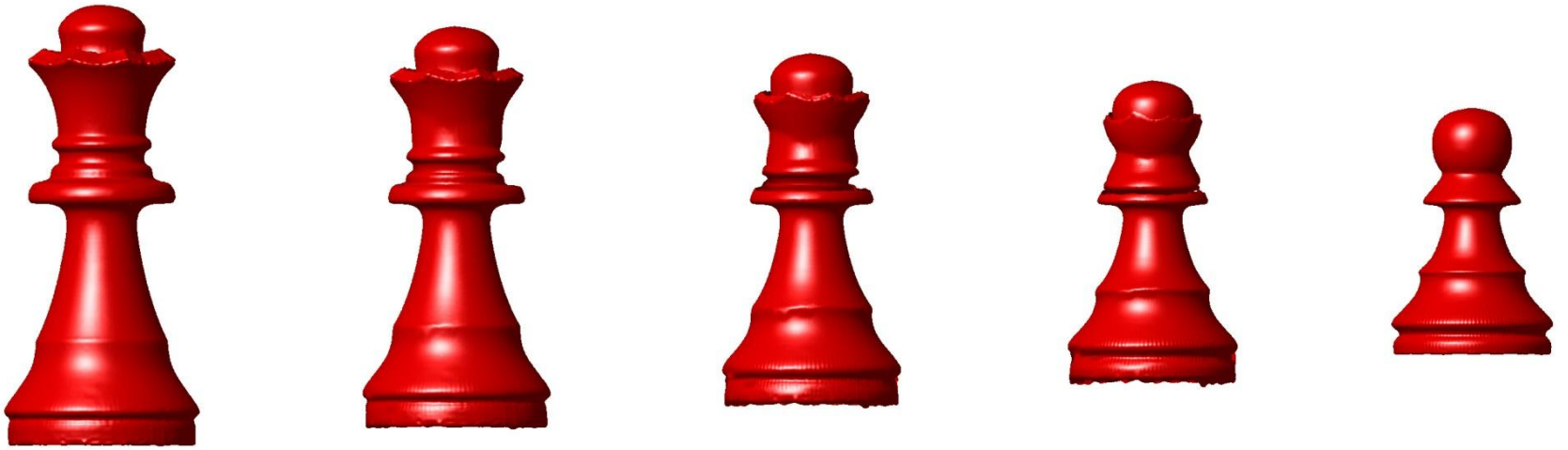
Signed distance



Signed distance + normals



Object Morphing



(signed distance and normals, no landmark points, no color information)

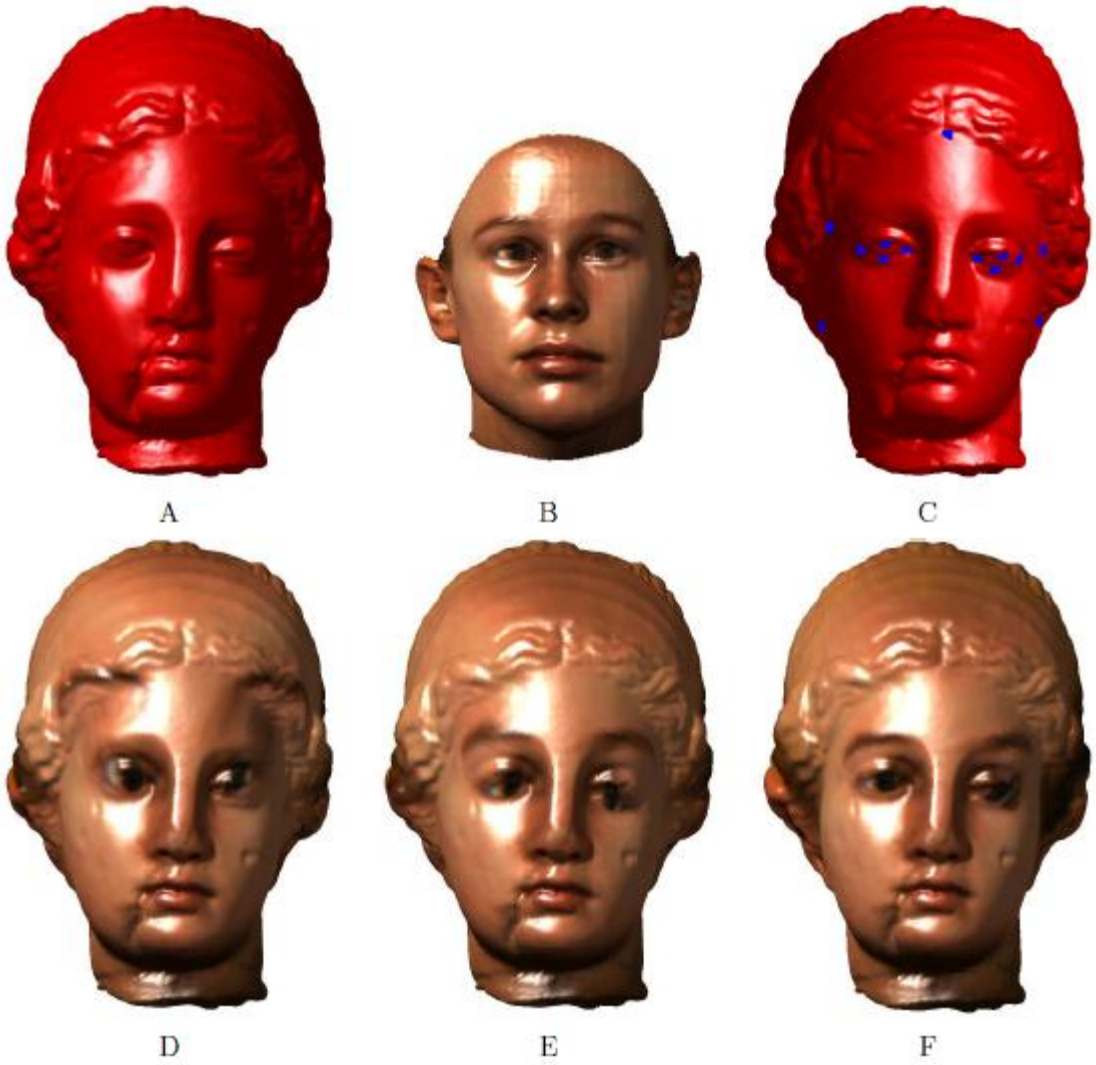
Head Morphing

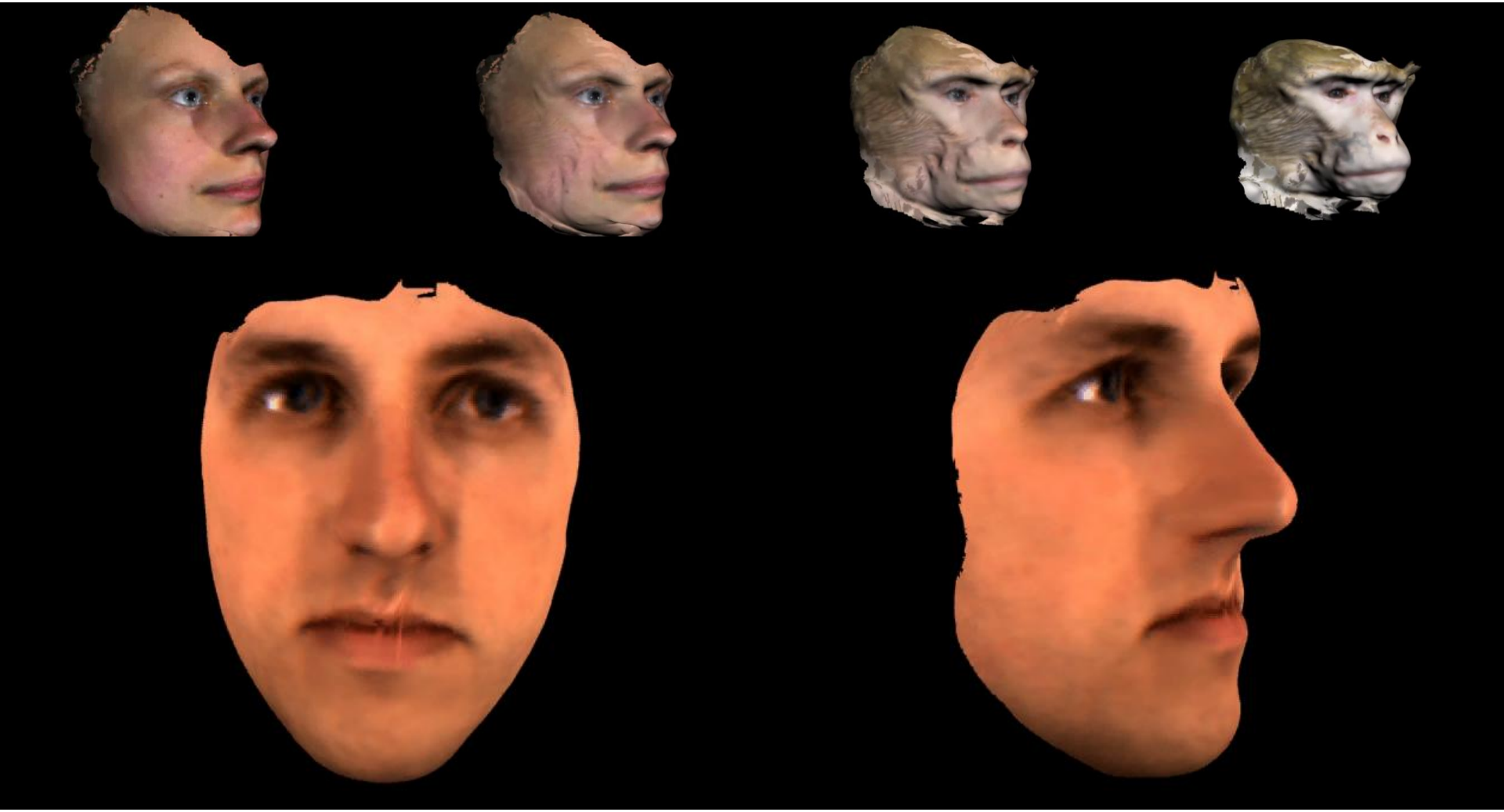
Start

Target

(signed distan olor information)

Texture Mapping





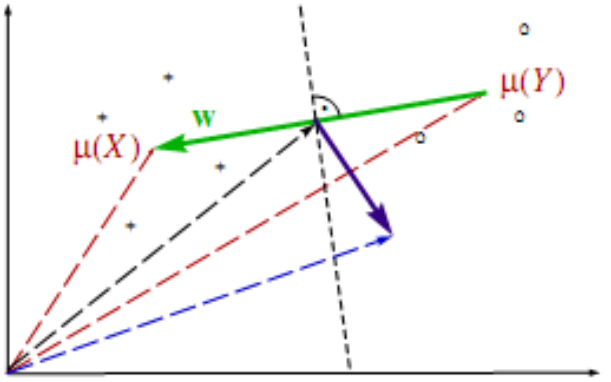
with Dept. of Physiology, MPI for Biological Cybernetics

Kernel Means

joint work with

Karsten Borgwardt, Kenji Fukumizu, Arthur Gretton, Jiayuan Huang, Quoc Le, Malte Rasch, Alex Smola, Le Song, Xiaohai Sun

An example of a kernel algorithm, revisited



$\mathcal{X}$ compact subset of a separable metric space, $m, n \in \mathbb{N}$.

Positive class $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$
Negative class $Y := \{y_1, \ldots, y_n\} \subset \mathcal{X}$

RKHS means $\mu(X) = \frac{1}{m}\sum_{i=1}^{m} k(x_i, \cdot)$, $\mu(Y) = \frac{1}{n}\sum_{i=1}^{n} k(y_i, \cdot)$.

Get a problem if $\mu(X) = \mu(Y)$.

Schölkopf & Smola, 2002

When do the means coincide?

$k(x, x') = \langle x, x' \rangle$: the means coincide

$k(x, x') = (\langle x, x' \rangle + 1)^d$: all empirical moments up to order d coincide

k strictly pd: $X = Y$.

The mean "remembers" each point that contributed to it.

Proposition 1 *Assume that k is strictly pd, and for all i, j, $x_i \neq x_j$, and $y_i \neq y_j$. If for some $\alpha_i, \beta_j \in \mathbb{R} - \{0\}$, we have*

$$\sum_{i=1}^{m} \alpha_i k(x_i, .) = \sum_{j=1}^{n} \beta_j k(y_j, .), \tag{1}$$

then $X = Y$.

Proof (by contradiction): W.l.o.g., assume that $x_1 \notin Y$. Subtract $\sum_{j=1}^{n} \beta_j k(y_j, .)$ from (1), and make it a sum over distinct points, to get

$$0 = \sum_i \gamma_i k(z_i, .),$$

where $z_1 = x_1, \gamma_1 = \alpha_1 \neq 0$, and $z_2, \cdots \in X \cup Y - \{x_1\}$, $\gamma_2, \cdots \in \mathbb{R}$.
Take the dot product with $\sum_j \gamma_j k(z_j, .)$, using $\langle k(z_i, .), k(z_j, .) \rangle = k(z_i, z_j)$, to get

$$0 = \sum_{ij} \gamma_i \gamma_j k(z_i, z_j),$$

with $\gamma \neq 0$, hence k cannot be strictly pd.

The mean map

$$\mu\colon X = (x_1, \ldots, x_m) \mapsto \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot)$$

satisfies

$$\langle \mu(X), f \rangle = \left\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot), f \right\rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i)$$

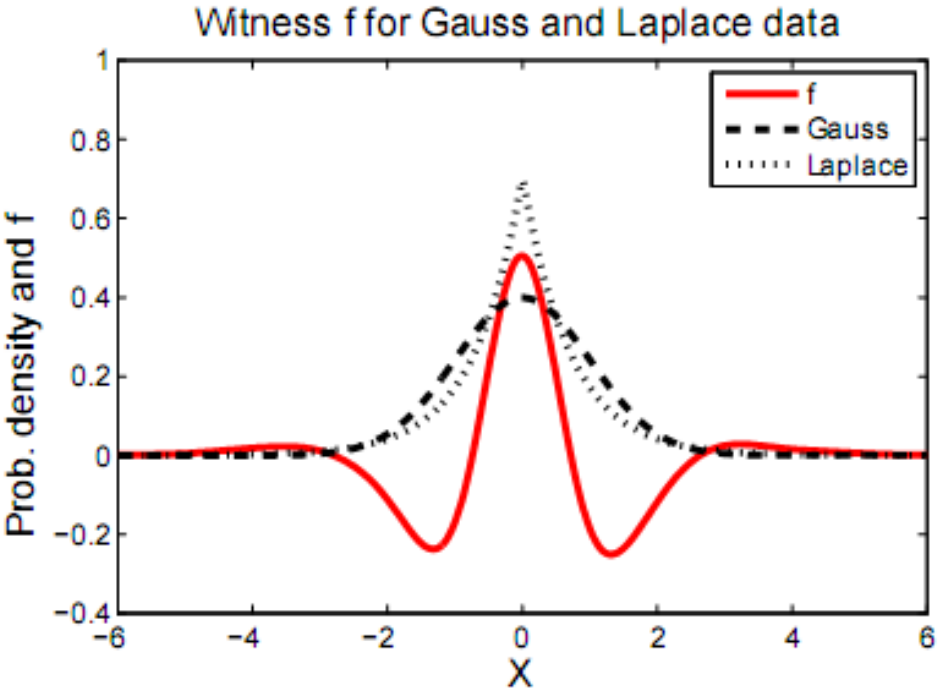
and

$$\|\mu(X) - \mu(Y)\| = \sup_{\|f\| \leq 1} |\langle \mu(X) - \mu(Y), f \rangle| = \sup_{\|f\| \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right| .$$

Note: distance in the RKHS = solution of a high-dimensional optimization problem.

Witness function

$f = \frac{\mu(X)-\mu(Y)}{\|\mu(X)-\mu(Y)\|}$, thus $f(x) \propto \langle \mu(X) - \mu(Y), k(x, .) \rangle$):



This function is in the RKHS of a Gaussian kernel, but not in the RKHS of the linear kernel.

The mean map for measures

p, q Borel probability measures,

$\mathbf{E}_{x,x' \sim p}[k(x,x')],\ \mathbf{E}_{x,x' \sim q}[k(x,x')] < \infty$ ($\|k(x,.)\| \leq M < \infty$ is sufficient)

Define

$$\mu\colon p \mapsto \mathbf{E}_{x \sim p}[k(x,\cdot)].$$

Note

$$\langle \mu(p), f \rangle = \mathbf{E}_{x \sim p}[f(x)]$$

and

$$\|\mu(p) - \mu(q)\| = \sup_{\|f\| \leq 1} \left| \mathbf{E}_{x \sim p}[f(x)] - \mathbf{E}_{x \sim q}[f(x)] \right|.$$

Recall that in the finite sample case, for strictly p.d. kernels, μ was injective — how about now?

Smola et al., ALT'07, Fukumizu et al., NIPS'07

Theorem 2 *[Fortet and Mourier (1953); Dudley (2002)]*

$$p = q \iff \sup_{f \in C(\mathcal{X})} \left| \mathbf{E}_{x \sim p}(f(x)) - \mathbf{E}_{x \sim q}(f(x)) \right| = 0,$$

where $C(\mathcal{X})$ is the space of continuous bounded functions on $\mathcal{X}$.

Replace $C(\mathcal{X})$ by the unit ball in an RKHS that is dense in $C(\mathcal{X})$ — universal kernel *(Steinwart, 2001)*, e.g., Gaussian.

Theorem 3 *[Gretton et al. (2007)] If k is universal, then*

$$p = q \iff \|\mu(p) - \mu(q)\| = 0.$$

- μ is invertible on its image
  $\mathcal{M} = \{\mu(p) \mid p \text{ is a probability distribution}\}$
  (the "marginal polytope", *Wainwright and Jordan (2003)*)
- generalization of the *moment generating function* of a RV x with distribution p:

$$M_p(.) = \mathbf{E}_{x \sim p}\left[e^{\langle x, \cdot \rangle}\right].$$

Uniform convergence bounds

Let X be an i.i.d. m-sample from p. The discrepancy

$$\|\mu(p) - \mu(X)\| = \sup_{\|f\| \leq 1} \left| \mathbf{E}_{x \sim p}[f(x)] - \frac{1}{m} \sum_{i=1}^{m} f(x_i) \right|$$

can be bounded using uniform convergence methods *(Smola et al., 2007)*.

Application 1: Two-sample problem *(Gretton et al., 2007)*

X, Y i.i.d. m-samples from p, q, respectively.

$$\begin{aligned}\|\mu(p) - \mu(q)\|^2 =& \mathbf{E}_{x,x' \sim p}\left[k(x,x')\right] - 2\mathbf{E}_{x\sim p, y\sim q}\left[k(x,y)\right] + \mathbf{E}_{y,y'\sim q}\left[k(y,y')\right] \\ =& \mathbf{E}_{x,x'\sim p, y,y'\sim q}\left[h((x,y),(x',y'))\right]\end{aligned}$$

with

$$h((x,y),(x',y')) := k(x,x') - k(x,y') - k(y,x') + k(y,y').$$

Define

$$\begin{aligned}D(p,q)^2 &:= \mathbf{E}_{x,x'\sim p, y,y'\sim q} h((x,y),(x',y')) \\ \hat{D}(X,Y)^2 &:= \frac{1}{m(m-1)} \sum_{i\neq j} h((x_i,y_i),(x_j,y_j)).\end{aligned}$$

$\hat{D}(X,Y)^2$ is an unbiased estimator of $D(p,q)^2$.
It's easy to compute, and works on structured data.

Theorem 4 *Assume k is bounded.*
$\hat{D}(X,Y)^2$ converges to $D(p,q)^2$ in probability with rate $\mathcal{O}(m^{-\frac{1}{2}})$.

This *could* be used as a basis for a test, but uniform convergence bounds are often loose..

Theorem 5 *We assume $\mathbf{E}\left(h^2\right) < \infty$. When $p \neq q$, then $\sqrt{m}(\hat{D}(X,Y)^2 - D(p,q)^2)$ converges in distribution to a zero mean Gaussian with variance*

$$\sigma_u^2 = 4\left(\mathbf{E}_z\left[(\mathbf{E}_{z'}h(z,z'))^2\right] - \left[\mathbf{E}_{z,z'}(h(z,z'))\right]^2\right).$$

When $p = q$, then $m(\hat{D}(X,Y)^2 - D(p,q)^2) = m\hat{D}(X,Y)^2$ converges in distribution to

$$\sum_{l=1}^{\infty} \lambda_l \left[q_l^2 - 2\right], \tag{2}$$

where $q_l \sim \mathcal{N}(0,2)$ i.i.d., λ_i are the solutions to the eigenvalue equation

$$\int_X \tilde{k}(x,x')\psi_i(x)dp(x) = \lambda_i\psi_i(x'),$$

and $\tilde{k}(x_i,x_j) := k(x_i,x_j) - \mathbf{E}_x k(x_i,x) - \mathbf{E}_x k(x,x_j) + \mathbf{E}_{x,x'}k(x,x')$ is the centred RKHS kernel.

Application 2: Dependence Measures

Assume that (x, y) are drawn from p_{xy}, with marginals p_x, p_y.
Want to know whether p_{xy} factorizes into its marginals.

Bach and Jordan (2002); Fukumizu et al. (2004): kernel generalized variance

Gretton et al. (2005a,b): kernel constrained covariance, HSIC

Main idea *(Rényi, 1959; Jacod and Protter, 2000)*:

x and y independent $\Longleftrightarrow$

$$\sup_{f,g \text{ bounded \& continuous}} \mathrm{Cov}(f(x), g(y)) = 0$$

Kernel version:

$$\sup_{f,g \in \text{unit balls in RKHS}} \mathrm{Cov}(f(x), g(y)) = 0$$

$$\mathrm{cov}(\mathsf{x}, \mathsf{y}) := \mathbf{E}_{\mathsf{x},\mathsf{y}}[\mathsf{xy}] - \mathbf{E}_{\mathsf{x}}[\mathsf{x}]\mathbf{E}_{\mathsf{y}}[\mathsf{y}]$$

k kernel on $\mathcal{X} \times \mathcal{Y}$.

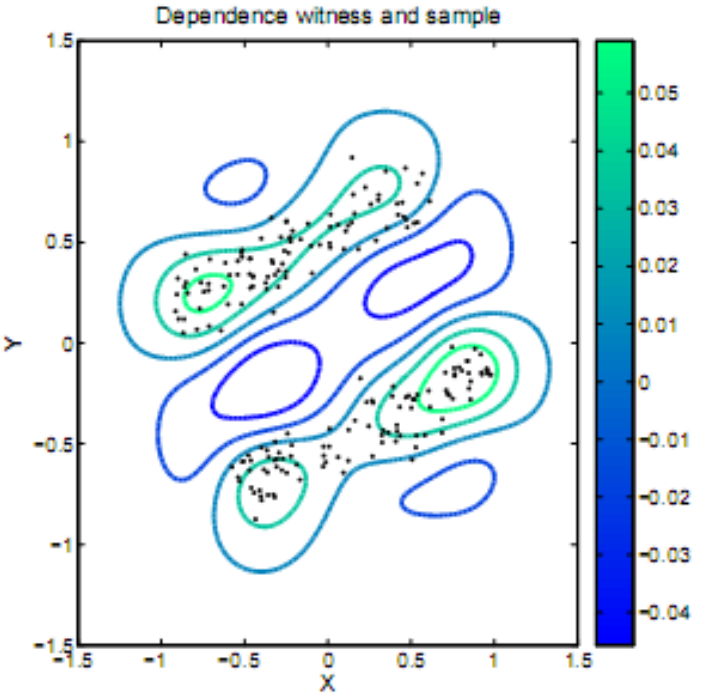
$$\mu(p_{xy}) := \mathbf{E}_{(x,y)\sim p_{xy}} [k((x,y),\cdot)]$$
$$\mu(p_x \times p_y) := \mathbf{E}_{x\sim p_x, y\sim p_y} [k((x,y),\cdot)] .$$

Use $\Delta := \left\| \mu(p_{xy}) - \mu(p_x \times p_y) \right\|$ as a measure of dependence.

For $k((x,y),(x',y')) = k_x(x,x')k_y(y,y')$:
Δ^2 equals the Hilbert-Schmidt norm of the covariance operator between the two RKHSs (HSIC), with empirical estimate $m^{-2} \operatorname{tr} HK_xHK_y$, where $H = I - \mathbf{1}/m$

Gretton et al. (2005a); Smola et al. (2007).

Witness function of the equivalent optimisation problem:



Application: learning causal structures *(Sun, Janzing, Schölkopf, Fukumizu, ICML 2007; Fukumizu, Gretton, Sun, Schölkopf, NIPS 2007))*

Application 3: Covariate Shift Correction and Local Learning

training set $X = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ drawn from p,
test set $X' = \{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}$ from $p' \neq p$.

Assume $p_{y|x} = p'_{y|x}$.

Shimodaira (2000): reweight training set

Minimize

$$\left\| \sum_{i=1}^{m} \beta_i k(x_i, \cdot) - \mu(X') \right\|^2 + \lambda \left\| \beta \right\|_2^2 \quad \text{subject to} \quad \beta_i \geq 0, \ \sum_i \beta_i = 1.$$

Equivalent QP:

$$\underset{\beta}{\text{minimize}} \ \frac{1}{2} \beta^\top (K + \lambda \mathbf{1}) \beta - \beta^\top l$$

$$\text{subject to } \beta_i \geq 0 \text{ and } \sum_i \beta_i = 1,$$

where $K_{ij} := k(x_i, x_j)$, $l_i = \left\langle k(x_i, \cdot), \mu(X') \right\rangle$.

Experiments show that in underspecified situations (e.g., large kernel widths), this helps *(Huang et al., 2007b)*.

$X' = \{x'\}$ leads to a local sample weighting scheme.

Application 4: Measure estimation and dataset squashing

(Dudík et al., 2004; Smola et al., 2007)

Given a sample X, minimize

$$\|\mu(X) - \mu(p)\|^2$$

over a convex combination of measures p_i,

$$p = \sum_i \alpha_i p_i, \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1.$$

This can be written as a convex QP with objective function

$$\|\mu(X) - \mu(p)\|^2 = \alpha^\top Q \alpha + 1_m^\top K 1_m - 2\alpha^\top L 1_m,$$

where

$$\begin{aligned} L_{ij} &:= \mathbf{E}_{x \sim p_i}\left[k(x, x_j)\right] \\ Q_{ij} &:= \mathbf{E}_{x \sim p_i, x' \sim p_j}\left[k(x, x')\right] \\ K_{ij} &= k(x_i, x_j) \\ 1_m &:= (1/m, \ldots, 1/m)^\top \in \mathbb{R}^m. \end{aligned}$$

In practice, use

$$\alpha^\top [Q + \lambda I]\alpha - 2\alpha^\top L 1_m$$

Some cases where Q and L can be computed in closed form *(Smola et al., 2007)*:

- Gaussian p_i and k (cf. *Balakrishnan and Schonfeld (2006); Walder et al. (2007)*)
- X training set, Dirac measures $p_i = \delta_{x_i}$: dataset squashing, *DuMouchel et al. (1999)*
- X test set, Dirac measures $p_i = \delta_{y_i}$ centered on the training points Y: covariate shift correction *Huang et al. (2007a)*

Kernel Tricks

Ancient Times

- p.d. kernels first used by *Hilbert (1904)*
- used to prove convergence of the potential function method *(Aizerman, Braverman, & Rozonoer, 1964)*
- Generalized Portrait method *(Vapnik & Chervonenkis, 1974)*
- *Duda & Hart (1973):* "The familiar functions of mathematical physics are eigenfunctions of symmetric kernels, and their use is often suggested for the construction of potential functions. However, these suggestions are more appealing for their mathematical beauty than their practical usefulness."
- *Grace Wahba (since 1970s)*

Renaissance

• used in Optimal Margin Classifiers *(Boser, Guyon & Vapnik 1992)*, Soft Margin Classifiers / Support Vector Networks *(Cortes & Vapnik 1995)*, Support Vector Machines *(Schölkopf, Burges & Vapnik 1995)*

• kernelization works for arbitrary dot product algorithms, e.g. Kernel PCA *(Schölkopf, Smola & Müller, ICANN 1997; Burges 1998)* --- ***"kernel trick"***

Industrial Revolution

- *1997 - 2000:* wide dissemination: NIPS kernel workshop, *Thorsten Joachims'* SVM light, "Kernel Machines" website *(2000)*

- X need not be a vector space *(Schölkopf, 1997, Haussler, Watkins, 1998, Zien et al. 1999)*

Modernity

- kernel ICA *(Bach & Jordan 2002, Fukumizu et al. 2005, Gretton et al. 2005)*
- use kernels to solve optimization problems over a large class of nonlinear functions: e.g.
- $\sup_{f \in \mathcal{F}} \mathbf{E}_p [f(x)] - \mathbf{E}_q [f(y)]$ Maximum Mean Discrepancy
- $\sup_{f \in F, g \in G} [\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y}))]$ Kernel Constrained Covariance

- [take home message]

  if you like to work with Expectations (rather than higher order statistics), map your data into an RKHS

Postmodernism


thank you for your attention


thank you for your attention

References

Y. Altun and A.J. Smola. Unifying divergence minimization and statistical inference via convex duality. In H.U. Simon and G. Lugosi, editors, *Proc. Annual Conf. Computational Learning Theory*, LNCS, pages 139–153. Springer, 2006.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.

N. Balakrishnan and D. Schonfeld. A maximum entropy kernel density estimator with applications to function interpolation and texture segmentation. In *SPIE Proceedings of Electronic Imaging: Science and Technology. Conference on Computational Imaging IV*, San Jose, CA, 2006.

M. Dudík, S. Phillips, and R.E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Proc. Annual Conf. Computational Learning Theory*. Springer Verlag, 2004.

R. M. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, UK, 2002.

W. DuMouchel, C. Volinsky, C. Cortes, D. Pregibon, and T. Johnson. Squashing flat files flatter. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 1999.

R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, 70:266–285, 1953.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, 2004.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19. The MIT Press, Cambridge, MA, 2007.

A. Gretton, O. Bousquet, A.J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Proceedings Algorithmic Learning Theory*, pages 63–77, Berlin, Germany, 2005a. Springer-Verlag.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129, 2005b.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007a. MIT Press.

J. Huang, A.J. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19. The MIT Press, Cambridge, MA, 2007b.

J. Jacod and P. Protter. *Probability Essentials*. Springer, New York, 2000.

A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10:441–451, 1959.

H. Shimodaira. Improving predictive inference under convariance shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 2000.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proc. Intl. Conf. Algorithmic Learning Theory*, volume 4754 of *LNAI*. Springer, 2007.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, 2001.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Department of Statistics, 2003.

C. Walder, K. Kim, and B. Schölkopf. Sparse multiscale gaussian process regression. Technical Report 162, Max-Planck-Institut für biologische Kybernetik, 2007.