Prognostic Models for Cancer Progression using Gene Expression Signatures

5th Australian Microarray Conference, September 2005

Trevor Hastie

Stanford University

`http://www-stat.stanford.edu/~hastie`

Overview and Outline

Supervised learning with $p \gg N$ is a slippery game. Its easy to fool yourself that you are doing well. We need to be vigilant against the temptations of finding ghosts in the data.

- Pitfalls of supervised learning with $P \gg N$.
- Biological signatures.
- Supervised principal components.

Outline

- Pitfalls of supervised learning with $P \gg N$.
- Biological signatures.
- Supervised principal components.

Cross-Validation Misused

Ref: Christophe Ambroise and Geoff McLachlan, PNAS April 2002

Consider a simple classifier for microarrays:

1. Starting with 20,000 genes, find the 200 genes having the largest correlation with the class labels.
2. Compute the gene averages (centroids) in each class for these 200 genes.
3. Classify a new sample to the nearest-centroid using only these 200 genes

Cross-validation divides the data into a training set and a test set (many times) to evaluate the validity of a procedure.

Two ways to cross-validate this simple classifier

Wrong: Apply cross-validation to step 2, after the gene selection.

Right: Apply cross-validation to steps 1 and 2.

It is easy to simulate realistic data with the class labels independent of the gene expression so that the *true* (and *right*) test error is 50%, but where the *wrong* CV error estimate is zero!

We have seen this error made in several high-profile papers in the last couple of years.

A recent experience of colleague Rob Tibshirani

- Dave et al (NEJM Nov 2004) published a high-profile study in NEJM, reporting that they had found two sets of genes whose expression were highly predictive of survival in patients with Follicular Lymphoma.
- the paper got a lot of attention, because the genes in the clusters were largely expressed in non-tumor cells, suggesting that the host-response was the important factor
- One of our medical collaborators — Ron Levy, asked Rob to look over their paper — he wanted to apply their model to the Stanford FL patient population.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 18, 2004 VOL. 351 NO. 21

Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells

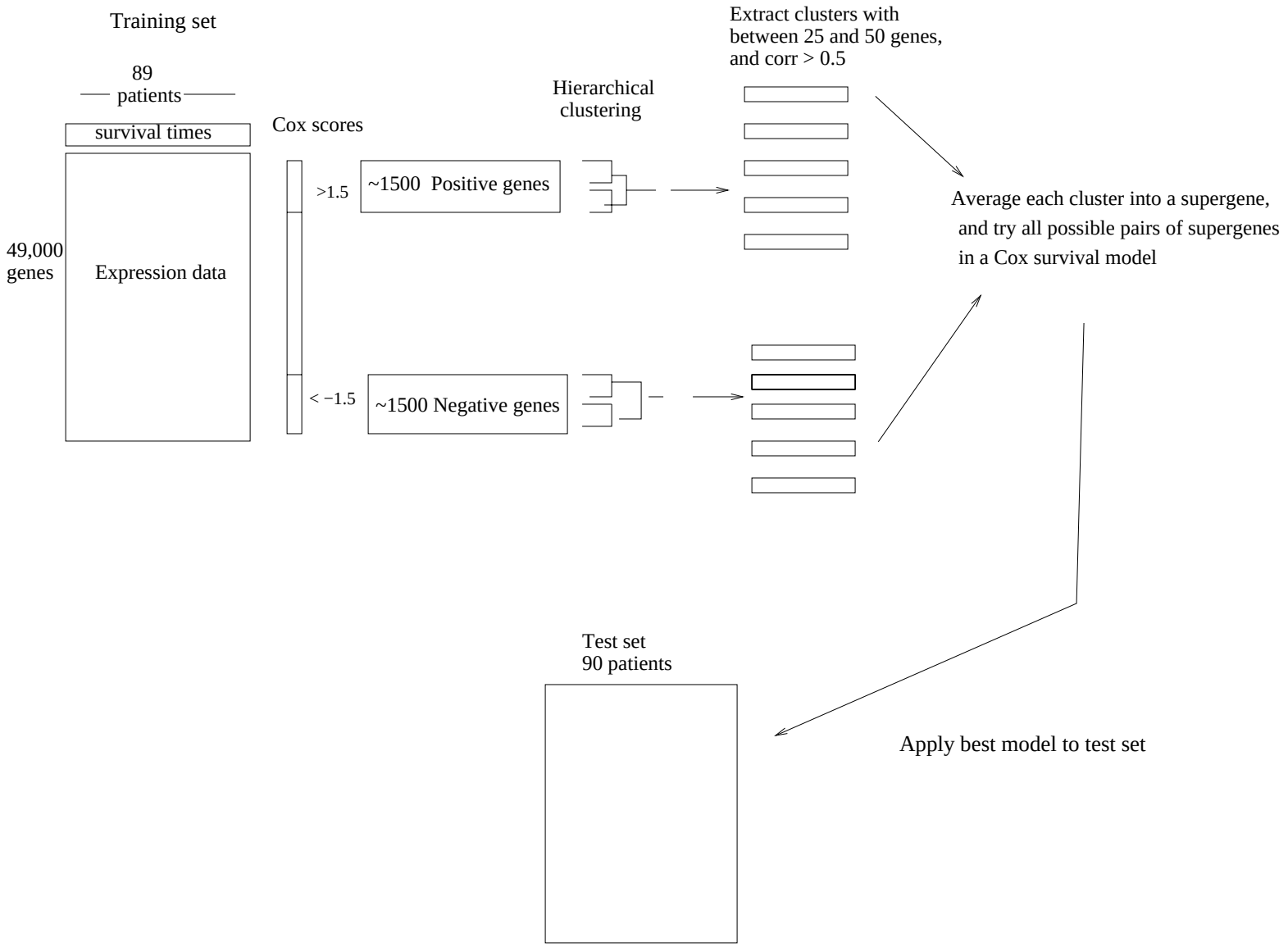
Sandeep S. Dave, M.D., George Wright, Ph.D., Bruce Tan, M.D., Andreas Rosenwald, M.D., Randy D. Gascoyne, M.D., Wing C. Chan, M.D., Richard I. Fisher, M.D., Rita M. Braziel, M.D., Lisa M. Rimsza, M.D., Thomas M. Grogan, M.D., Thomas P. Miller, M.D., Michael LeBlanc, Ph.D., Timothy C. Greiner, M.D., Dennis D. Weisenburger, M.D., James C. Lynch, Ph.D., Julie Vose, M.D., James O. Armitage, M.D., Erlend B. Smeland, M.D., Ph.D., Stein Kvaloy, M.D., Ph.D., Harald Holte, M.D., Ph.D., Jan Delabie, M.D., Ph.D., Joseph M. Connors, M.D., Peter M. Lansdorp, M.D., Ph.D., Qin Ouyang, Ph.D., T. Andrew Lister, M.D., Andrew J. Davies, M.D., Andrew J. Norton, M.D., H. Konrad Muller-Hermelink, M.D., German Ott, M.D., Elias Campo, M.D., Emilio Montserrat, M.D., Wyndham H. Wilson, M.D., Ph.D., Elaine S. Jaffe, M.D., Richard Simon, Ph.D., Liming Yang, Ph.D., John Powell, M.S., Hong Zhao, M.S., Neta Goldschmidt, M.D., Michael Chiorazzi, B.A., and Louis M. Staudt, M.D., Ph.D.

Summary of their findings

- They started with the expression of approximately 49,000 genes measured on 191 patient samples, derived from DNA microarrays. A survival time (possibly censored) was available for each patient
- they randomly split the data into a training set of 95 patients and a test set of 96 patients
- using a fairly complex multi-step procedure, they extracted two clusters of genes, called IR1 (immune response 1) and IR2 (immune response 2).
- They averaged the gene expression of the genes in each cluster, to create two "super-genes".

... continued

- They then fit these super-genes together in a Cox model for survival, and applied it to the training and test sets. The p-value in the training set was $< 10^{-7}$ and 0.003 in the test set. IR1 correlates with good prognosis; IR2 with poor prognosis
- In the remainder of the paper they interpret the genes in their model


Training set
89
patients
survival times
49,000
genes
Expression data
Cox scores
>1.5
~1500 Positive genes
< −1.5
~1500 Negative genes
Hierarchical
clustering
Extract clusters with
between 25 and 50 genes,
and corr > 0.5
Average each cluster into a supergene,
and try all possible pairs of supergenes
in a Cox survival model
Test set
90 patients
Apply best model to test set

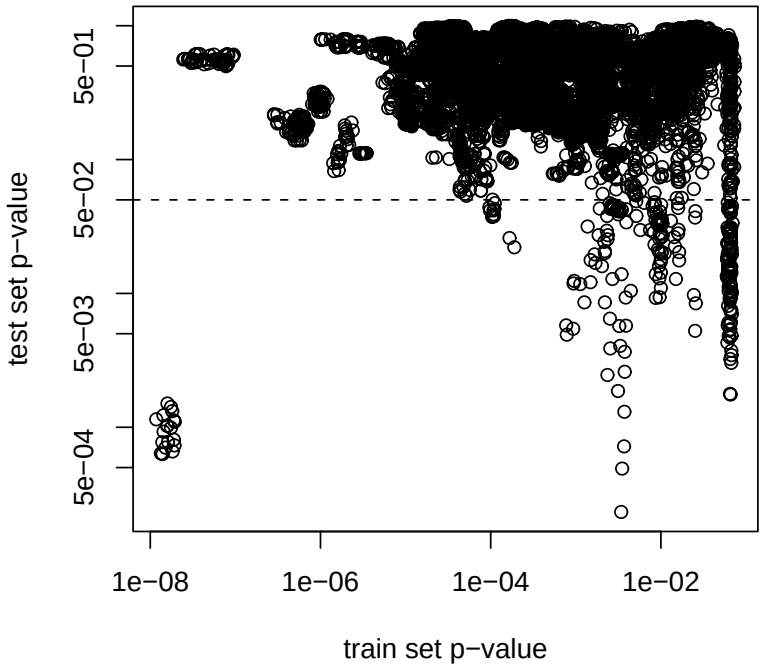
What happened next...

- Tibshirani downloaded the data
- Applied some familiar statistical tools — eg SAM (Significance Analysis of Microarrays), less familiar ones — supervised principal components. His initial finding — no significant correlation between gene expression and survival.
- He spent 2-3 weeks emailing back and forth with their statistician (George Wright) and programming in R, to recreate their analysis
- He tweaked their analysis (separately) in two simple ways:
  - Swapped training and test sets
  - Changed their cluster-size thresholds slightly

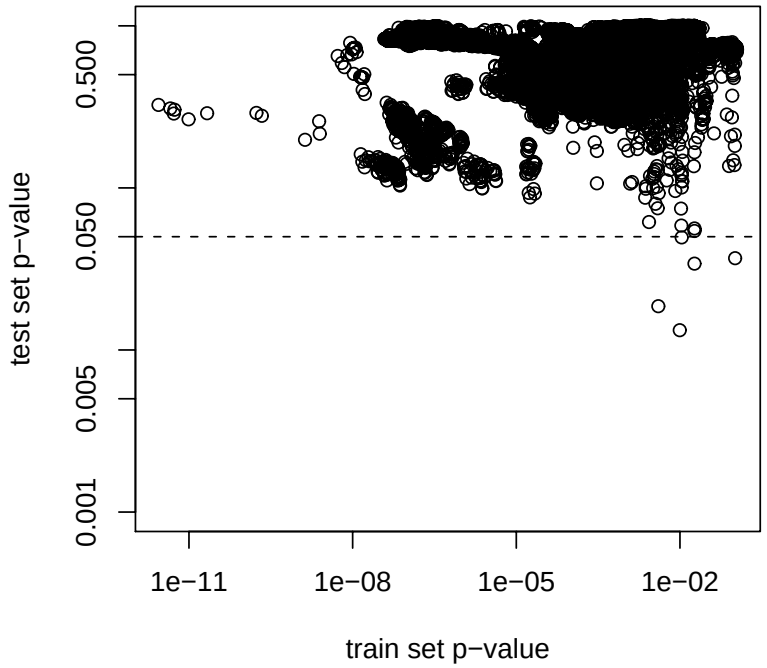
  In both cases their findings disappeared!

Swapping Train and Test Sets

Original Train-Test



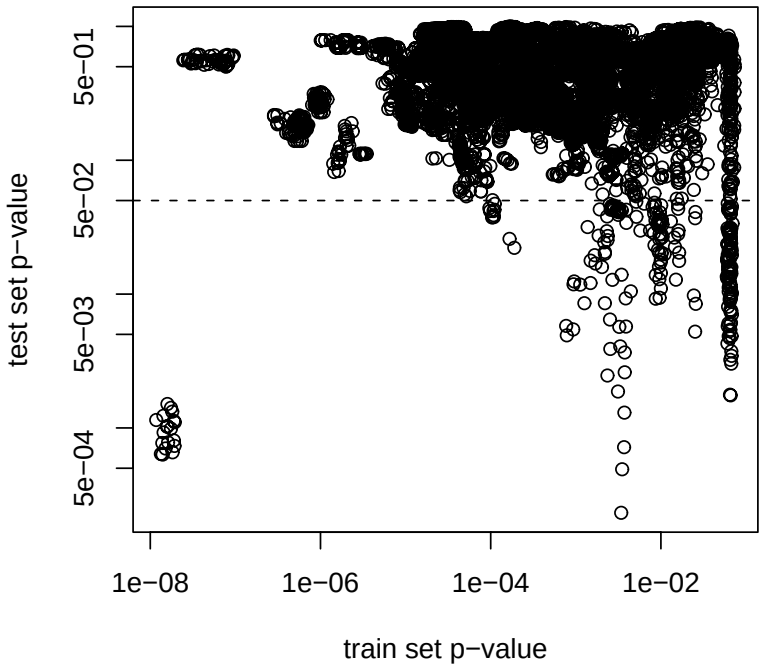
Swapped Train-Test



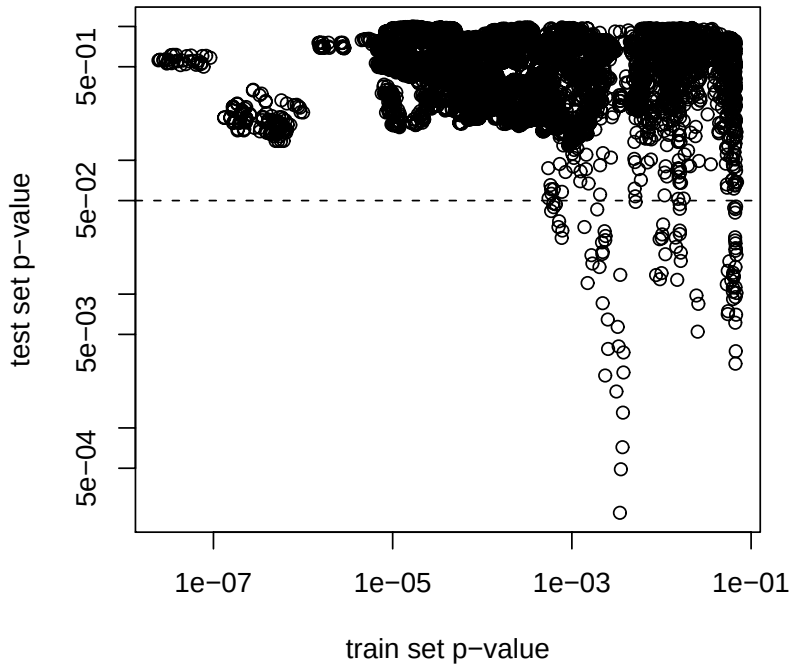
P-values in a multivariate Cox model. Their modeling procedure found nothing when the train-test set divisions were flipped.

Cluster size ranges (30,60) rather than (25,50)

Cluster size range (25,50)



Cluster size range (30,60)



P-values in a multivariate Cox model. Their modeling procedure found nothing when the cluster-size ranges were tweaked.

The Aftermath

- Tibshirani published a short letter to NEJM in March 2005; full details of his re-analysis appear on his website
- The authors published a rebuttal in the same issue. Their arguments:
  1. we followed standard statistical procedures, found a small p-value on the test set, therefore our finding is correct;
  2. our method found an interaction, which SAM can't find
  3. we get small p-values if we apply our original model coefficients to random halves of the data (????!!!!!!)

General comments

- Their finding is *fragile*. We don't believe that it is real or reproducible
- This experience uncovers a problem that is of general importance to our field:
  - with many predictors, it is too easy to overfit the data and find spurious results
  - we can inadvertently mislead the reader, and mislead ourselves. We have been guilty of this too
- Important to be very explicit about methodology used, provide scripts etc. *Reproducible research*

Outline

- Pitfalls of supervised learning with $P \gg N$.
- Biological signatures.
- Supervised principal components.

Biological Signatures

- Using in vitro biology, signatures of genes are hypothesized to play a role in cancer prognosis
- The signature is used to score each human cancer sample in a separate study
- These signatures (each one degree of freedom), are compared to traditional prognostic factors.

PNAS PNAS PNAS

Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival

Howard Y. Chang[a,b,c], Dimitry S. A. Nuyten[c,d,e], Julie B. Sneddon[b], Trevor Hastie[f], Robert Tibshirani[f], Therese Sørlie[b,g], Hongyue Dai[h,i], Yudong D. He[h,i], Laura J. van't Veer[d,i], Harry Bartelink[e], Matt van de Rijn[j], Patrick O. Brown[b,k,l], and Marc J. van de Vijver[d,l]

[a]Program in Epithelial Biology, Departments of [b]Biochemistry, [f]Health Research and Policy, and [j]Pathology, and [k]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of [d]Diagnostic Oncology and [e]Radiation Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands; [h]Rosetta Inpharmatics, Seattle, WA 98109; and [g]Norwegian Radium Hospital, 0310 Oslo, Norway

Contributed by Patrick O. Brown, January 5, 2005

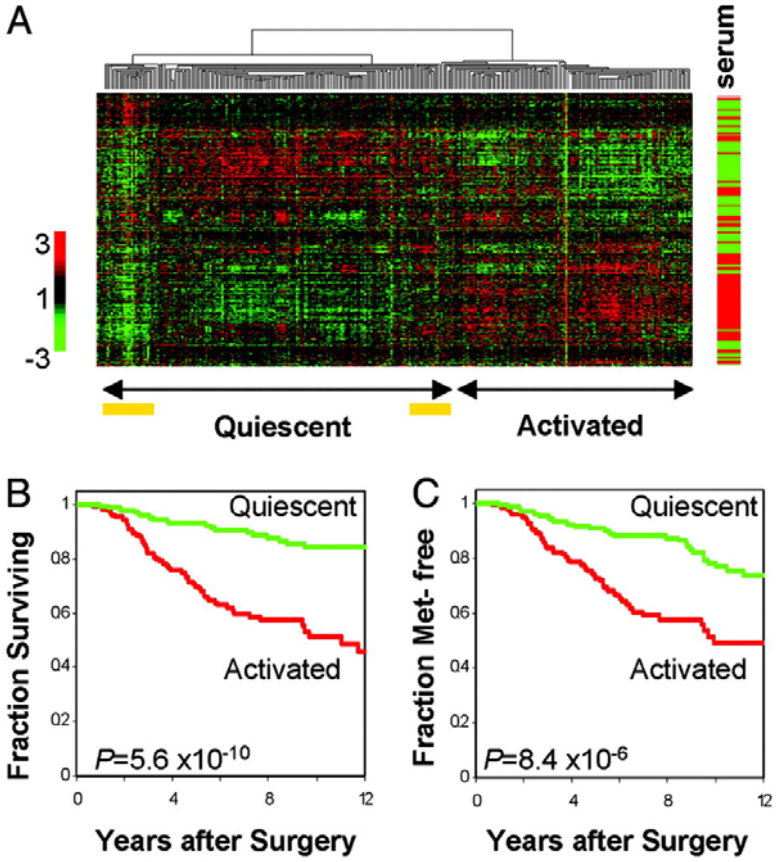
Based on the hypothesis that features of the molecular program of normal wound healing might play an important role in cancer metastasis, we previously identified consistent features in the transcriptional response of normal fibroblasts to serum, and used this ''wound-response signature'' to reveal links between wound healing and cancer progression in a variety of common epithelial tumors. Here, in a consecutive series of 295 early breast cancer patients, we show that both overall survival and distant metastasis-free survival are markedly diminished in patients whose tumors expressed this wound-response signature compared to tumors that did not express this signature. A gene expression centroid of the wound-response signature provides a basis for prospectively assigning a prognostic score that can be scaled to suit different clinical purposes. The wound-response signature improves risk stratification independently of known clinico-pathologic risk factors and previously established prognos-

response'' (CSR) genes and their canonical expression pattern in fibroblasts activated with serum, the soluble fraction of clotted blood and an important initiator of wound healing *in vivo*. The CSR genes were chosen to minimize overlap with cell cycle genes, but instead appeared to represent other important processes in wound healing, such as matrix remodeling, cell motility, and angiogenesis, processes that are likely also to contribute to cancer invasion and metastasis. In several common epithelial tumors such as breast, lung, and gastric cancers, expression of the wound-response signature predicted poor overall survival and increased risk of metastasis (10). These initial findings demonstrate the promise of using hypothesis-driven gene expression signatures to provide insights from existing gene expression profiles of cancers. However, as in other methodologies, reproducibility and scales for interpretation need to be evaluated before this strategy can be generally adopted for biologic dis-

Wound Signature

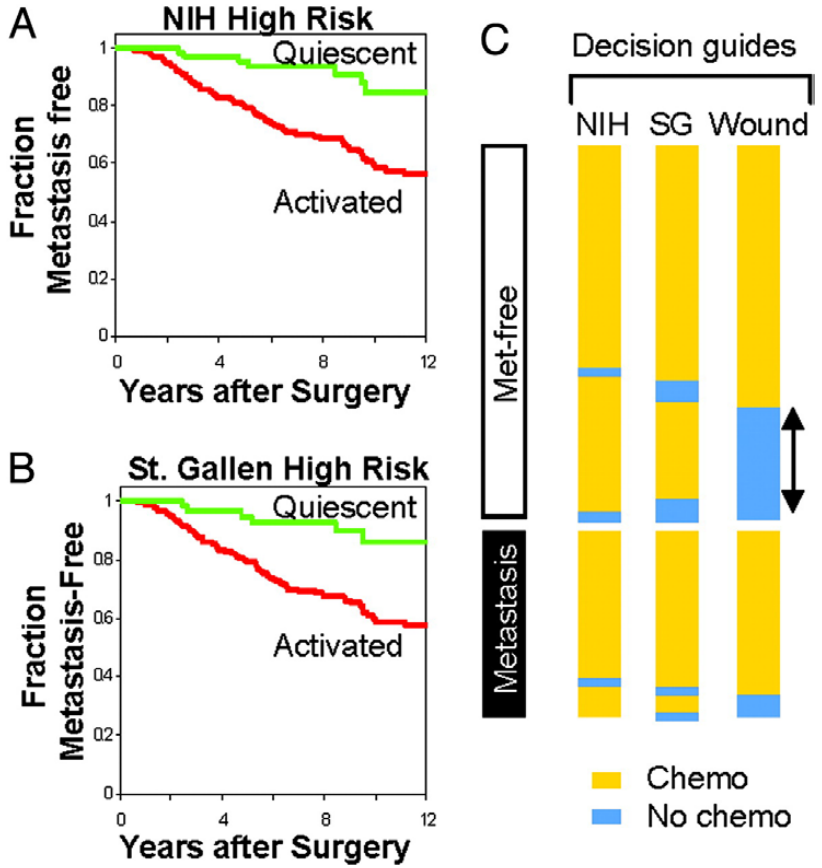
- Chang et al examined the transcriptional response of normal fibroblasts to serum *in vitro*.
- They identified a set of approximately 400 "Core Serum Response" genes that showed a wound response in a subset of the samples.
- The average of these genes in the subset gives a profile of up- and down-regulated genes.
- Any future sample (from a patient) can be scored for wound signature by computing the correlation of the expression of the corresponding genes with this profile.
- They evaluated this signature on an independent sample of 295 breast cancer samples (Netherlands Cancer Institute).

Fig. 1. Performance of a "wound response" gene expression signature in predicting breast cancer progression



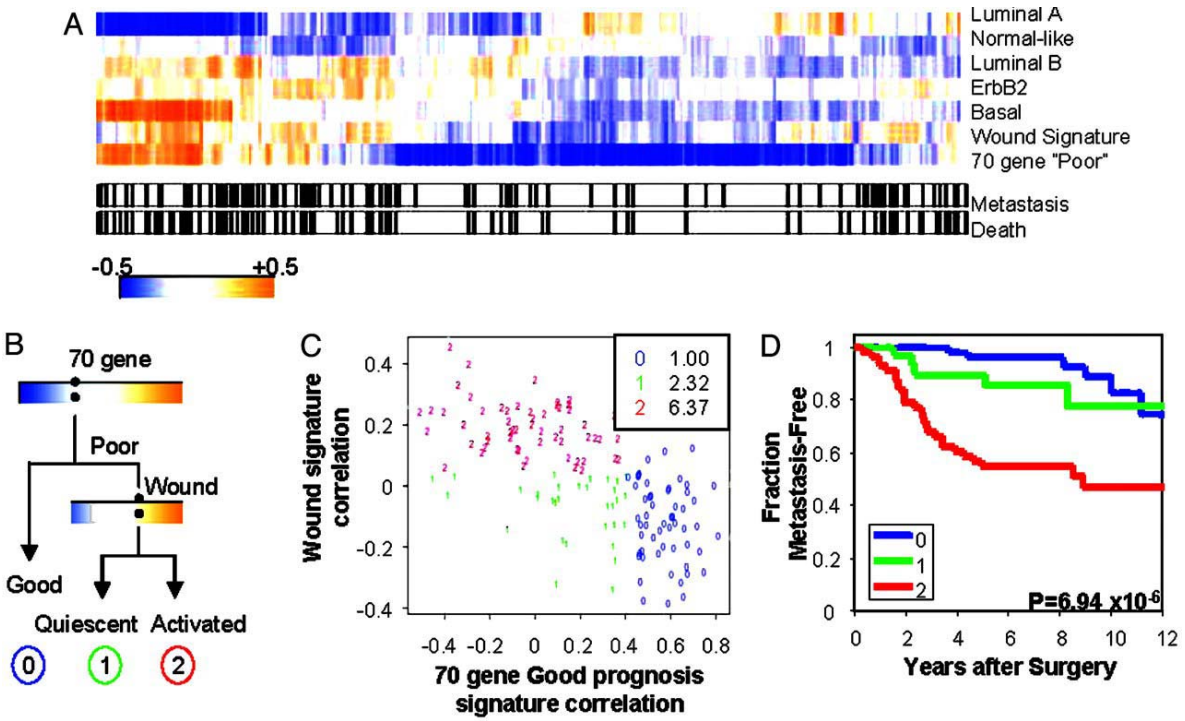
Chang, Howard Y. et al. (2005) Proc. Natl. Acad. Sci. USA 102, 3738-3743

Fig. 2. A scalable wound-response signature as a guide for chemotherapy



Chang, Howard Y. et al. (2005) Proc. Natl. Acad. Sci. USA 102, 3738-3743

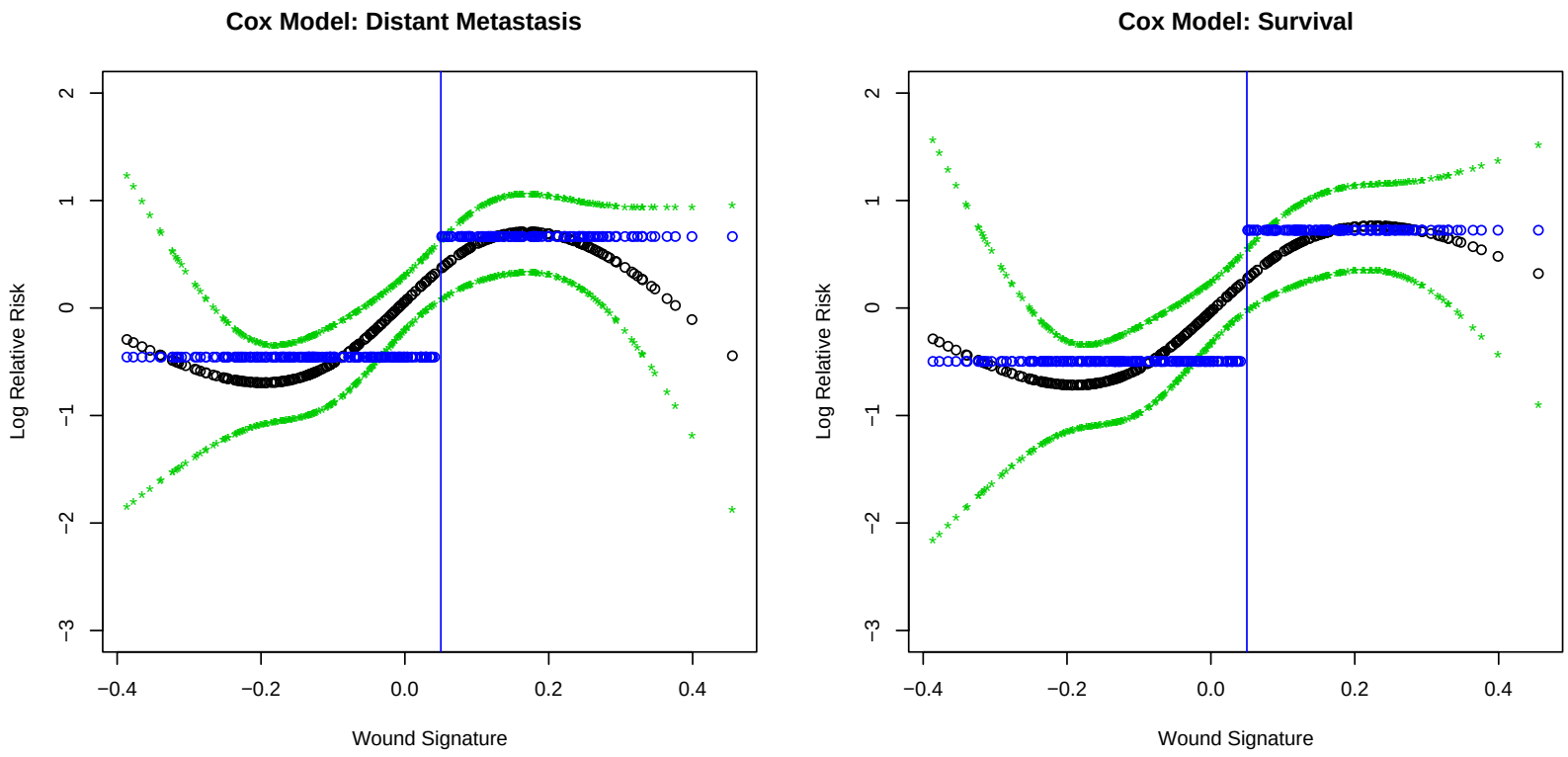
Fig. 3. Integration of diverse gene expression signatures for risk prediction



Chang, Howard Y. et al. (2005) Proc. Natl. Acad. Sci. USA 102, 3738-3743

Statistical Analysis

- Compared with traditional prognostic factors, using multivariate Cox model — tumor grade and size, lymph-node status, ER status, ... Summary later.
- Examine nature of wound signature score using semi-parametric methods [Hastie & Tibshirani, *Generalized Additive Models*, 1991]
- Using subgroups and scaling of the wound score, we showed that wound signature offers independent prognostic information, and could potentially spare 30% of women from chemotherapy.



Contribution of the wound score to the log-hazard in the proportional hazards model:

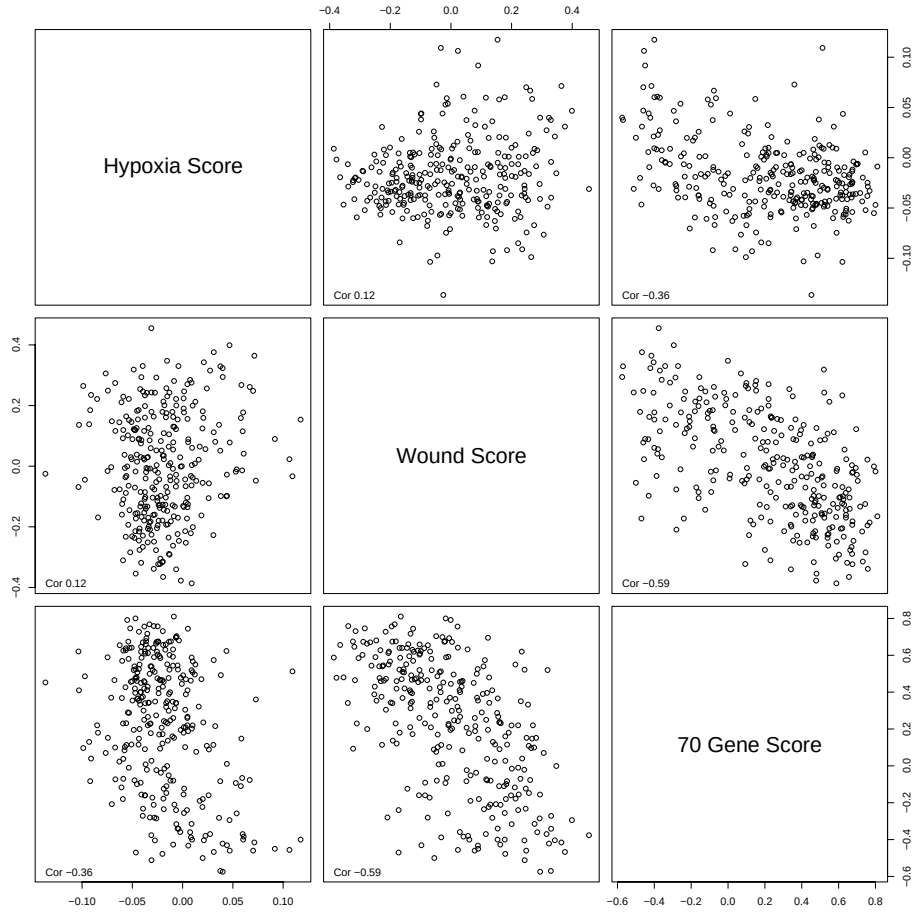
$$\log \lambda(T, X, W) = \log \lambda_0(T) + X^T \beta + f(W),$$

where $f(W)$ is modeled by cubic splines (black), or binary (blue).

Hypoxia Signature

lead author Jen-Tsan Ashley Chi, now at Duke MC

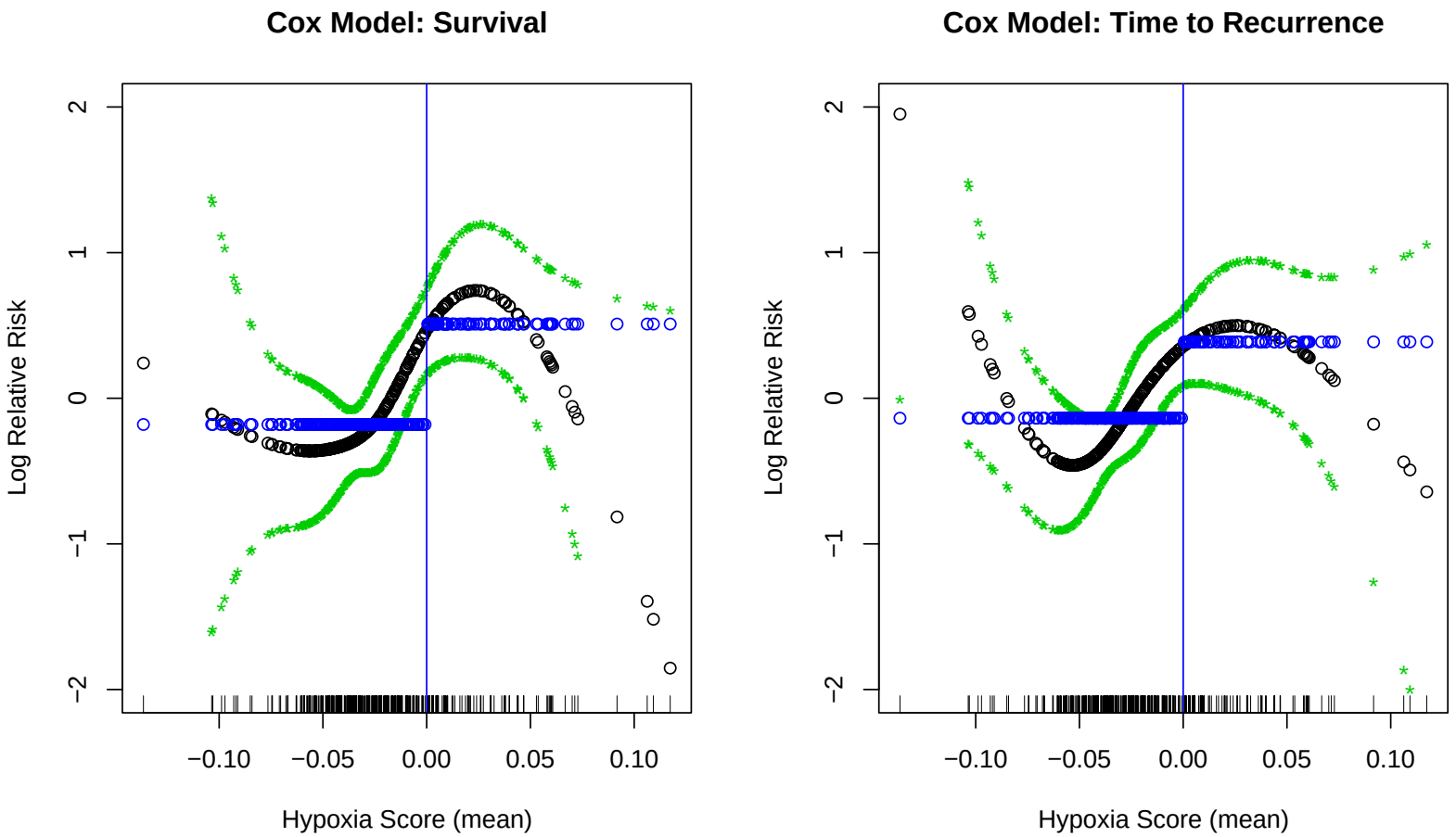
- About 250 genes showing response to hypoxia *in vitro* in cultured epithelial cells.
- They saw evidence on Stanford data that tumors with cells showing a strong response to hypoxia were associated with bad outcome.
- Here the signature is obtained by simply averaging the corresponding genes for each cancer patient.



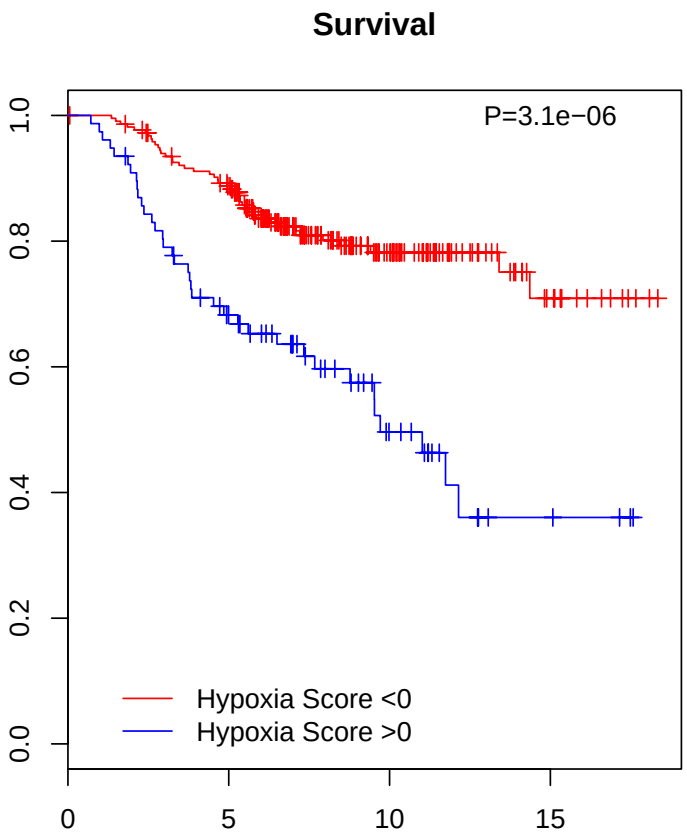
NKI breast cancer

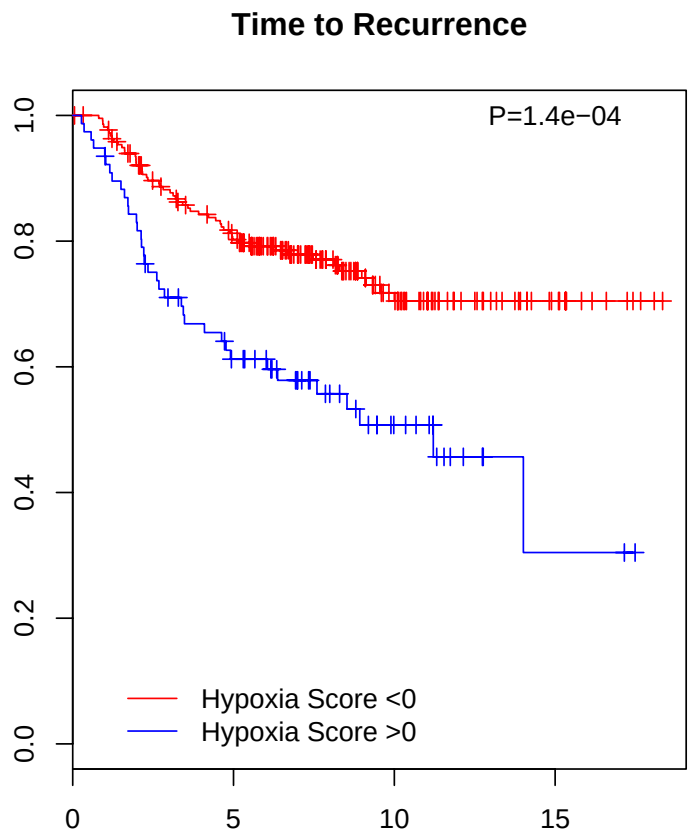
- Hypoxia and Wound have low correlation
- 70 Gene score is a supervised signature on these data, developed by original authors (van't Veer et al, Nature 2002).
- We focus on Wound and Hypoxia

Semi-parametric Cox Models



Kaplan-Meier Curves





Marginal and Partial Effects — Survival

	Df	Marginal	Partial	Pr(Chi)	
Angio Invasion	2	11.7	6.1	0.0660602	.
Age	1	8.8	11.4	0.0013773	**
Diameter	1	12.2	3.1	0.0963076	.
Node Status	2	6.9	1.7	0.4621051	
Grade	2	43.7	2.6	0.3142204	
ER Status	1	27.4	8.7	0.0053109	**
Mastectomy	1	0.7	0.3	0.5943224	
Chemotherapy	1	1.1	1.6	0.2341016	
Hormonal	1	1.8	0.1	0.7869375	
Hypoxia	1	22.6	8.6	0.0055803	**
Wound	1	39.9	15.8	0.0001708	***

As percent of the total deviance explained (89.5) in Cox model.

Signature Summary

- Wound signature provides an additional 18.7% in prognostic power in the Cox model (when added to all the other factors), and surpasses them all.
- Hypoxia signature adds 9.4%, similar to ER status and surpassed only by Age.

	Df	Marginal	Partial	Partial.add
Traditional	12	76.4	42.2	55.2
Signatures	2	57.8	23.6	40.8

- Wound and Hypoxia signature account for an additional 40.8% in prognostic power.
- External signatures avoid issues of overfitting associated with supervised signatures.

Outline

- Pitfalls of supervised learning with $P \gg N$.
- Biological signatures.
- Supervised principal components.

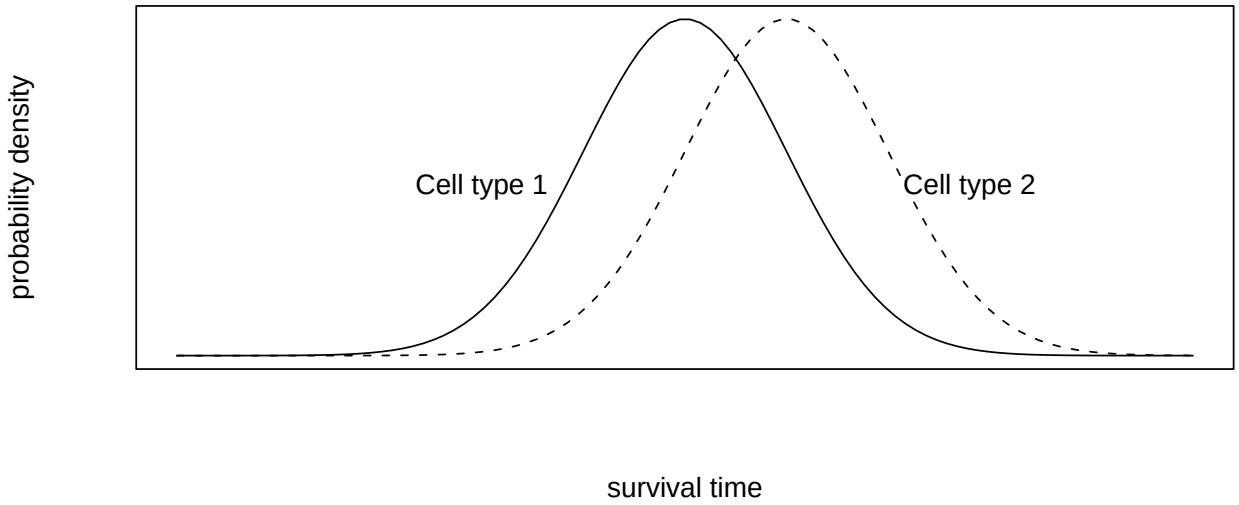
Supervised Principal Components

- method for regression or generalized regression (eg survival outcome), useful when number of predictors $p >> N$, the sample size
- Bair, Hastie, Paul, Tibshirani — to appear JASA 2005;
- Software available (Excel addin, R) on Tibshirani website

Usual approaches

- *Unsupervised approach* — cluster patients into groups, then hope that they differ in survival. Strategy widely used by Pat Brown, David Botstein and colleagues at Stanford. Idea is that biological subgroups may be reproducible but not specific gene lists that characterize these groups
- *Supervised approach* — find genes that correlate with survival. Some sort of regularization (eg ridge regression) is needed, since number of genes >> number of patients

A semi-supervised approach



Underlying conceptual model: survival time is a noisy surrogate for cell type, a real determinant of survival. Idea: rather than predict survival time directly, try to uncover the cell types and use these to predict survival time

Supervised principal components

Idea is to chose genes whose correlation with the outcome is largest, and using only those genes, extract the first (or first few) principal components.

Use these "supervised principal components" to predict the outcome

1. Compute (univariate) standard regression coefficients for each feature
2. Form a reduced data matrix consisting of only those features whose univariate coefficient exceeds a threshold θ in absolute value (θ is estimated by cross-validation)
3. Compute the first (or first few) principal components of the reduced data matrix
4. Use these principal component(s) in a regression model to predict the outcome

More details

- in our paper we develop a latent variable model, one that assumes the existence of latent variables (e.g cell types) shared by a subset of the features and the outcome.
- We show that the supervised PC approach estimates these latent variables consistently as $p, N \to \infty$ ($p = \#$ of features, $N = \#$ of samples)
- By contrast, standard principal components is not consistent in general— the large number of "noise" features corrupts the estimate

An underlying model

- Suppose we have a response variable Y which is related to an underlying *latent variable* U by a linear model

$$Y = \beta_0 + \beta_1 U + \varepsilon. \tag{1}$$

- In addition, we have expression measurements on a set of genes X_j indexed by $j \in \mathcal{P}$, for which

$$X_j = \alpha_{0j} + \alpha_{1j} U + \epsilon_j, \quad j \in \mathcal{P}. \tag{2}$$

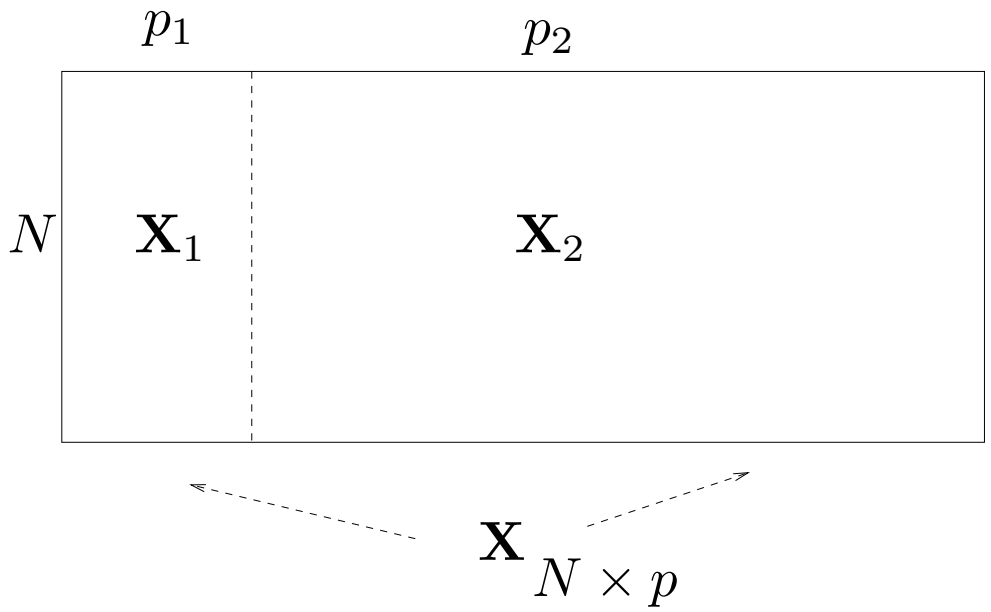
  We also have many additional genes X_k, $k \notin \mathcal{P}$ which are independent of U. We can think of U as a discrete or continuous aspect of a cell type, which we do not measure directly.

- The supervised principal component algorithm (SPCA) can be seen as an approximate method for fitting this model.

Natural since on average the score $||{X_j}^T Y||/||X_j||$ is non-zero only if α_{1j} is non-zero.

Consistency of supervised principal components

We consider a latent variable model of the form (1) and (2) for data with N samples and p features.

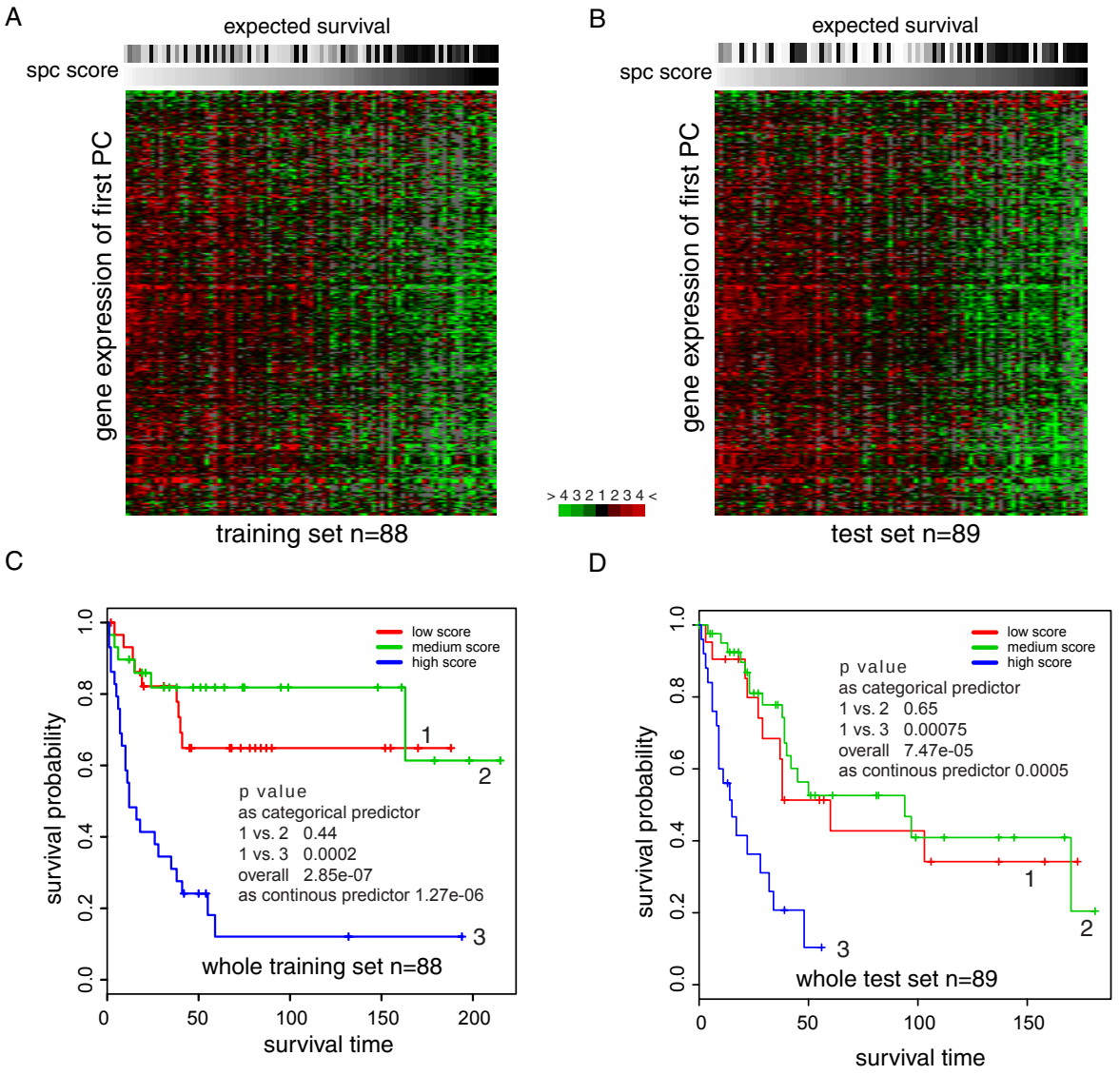


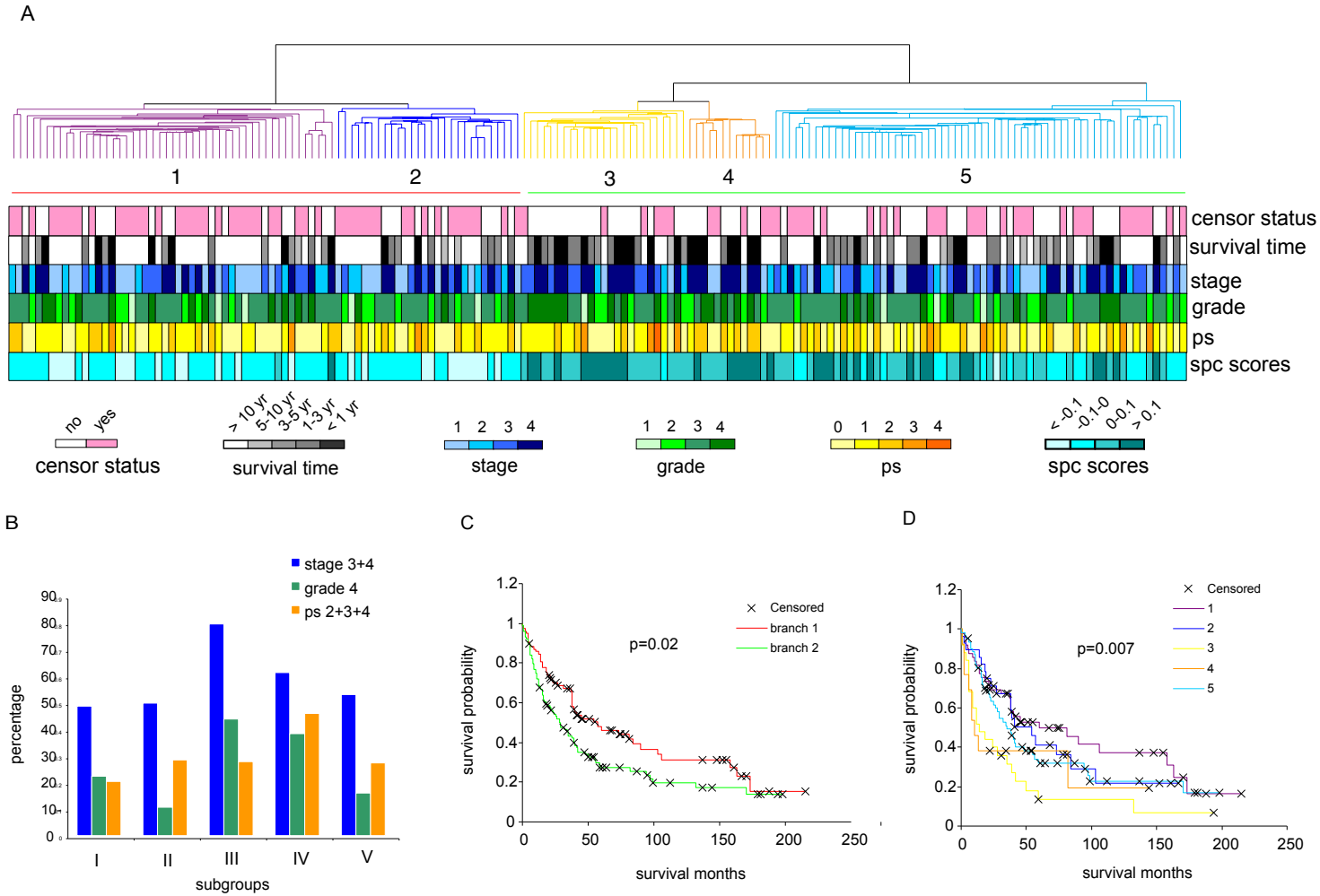
$$p/N \to \gamma \in (0, \infty)$$
$$p_1/N \to 0 \text{ fast}$$

Kidney cancer study

Jim Brooks, Hongjuan Zhao, Rob Tibshirani

14,000 genes; 180 samples — 90 in each of training and test


A
expected survival
spc score
gene expression of first PC
training set n=88
B
expected survival
spc score
gene expression of first PC
test set n=89
> 4 3 2 1 2 3 4 <
C
low score
medium score
high score
survival probability
1
2
3
p value
as categorical predictor
1 vs. 2 0.44
1 vs. 3 0.0002
overall 2.85e-07
as continous predictor 1.27e-06
whole training set n=88
survival time
D
low score
medium score
high score
survival probability
p value
as categorical predictor
1 vs. 2 0.65
1 vs. 3 0.00075
overall 7.47e-05
as continous predictor 0.0005
1
2
3
whole test set n=89
survival time


A
1
2
3
4
5
censor status
survival time
stage
grade
ps
spc scores
no
yes
censor status
> 10 yr
5-10 yr
3-5 yr
1-3 yr
< 1 yr
survival time
1 2 3 4
stage
1 2 3 4
grade
0 1 2 3 4
ps
< -0.1
-0.1-0
0-0.1
> 0.1
spc scores
B
stage 3+4
grade 4
ps 2+3+4
percentage
I
II
III
IV
V
subgroups
C
Censored
branch 1
branch 2
p=0.02
survival probability
survival months
D
Censored
1
2
3
4
5
p=0.007
survival probability
survival months