Prediction by supervised principal components

IMS Medallion lecture 2007

Joint work with Eric Bair, Trevor Hastie, Debashis Paul
Stanford University

Based on *Prediction by supervised principal components*, Bair et al JASA 2006

Pre-conditioning for feature selection and regression in high-dimensional problems, Paul et. al., submitted.

Papers/Software available at

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

The Problem: $p >> N$

- Linear regression and Cox (survival) regression when p (number of features) is $>>$ N (number of observations)
- Motivation: gene expression studies. Objective is to correlate a survival time with gene expression. Typically $N \approx 100$ patients, $p = 10,000$ genes.

Why the problem is hard

- With a large number of features, there is a real danger of overfitting the data
- See for example the controversy in the New England Journal of Medicine on Non-Hodgkins Lymphoma (my homepage has full details)
- need statistical methods that are simple and can be internally validated

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.

ABSTRACT

BACKGROUND

Patients with follicular lymphoma may survive for periods of less than 1 year to more than 20 years after diagnosis. We used gene-expression profiles of tumor-biopsy specimens obtained at diagnosis to develop a molecular predictor of the length of survival.

METHODS

Gene-expression profiling was performed on 191 biopsy specimens obtained from patients with untreated follicular lymphoma. Supervised methods were used to discover expression patterns associated with the length of survival in a training set of 95 specimens. A molecular predictor of survival was constructed from these genes and validated in an independent test set of 96 specimens.

RESULTS

Individual genes that predicted the length of survival were grouped into gene-expression signatures on the basis of their expression in the training set, and two such signatures were used to construct a survival predictor. The two signatures allowed patients with specimens in the test set to be divided into four quartiles with widely disparate median lengths of survival (13.6, 11.1, 10.8, and 3.9 years), independently of clinical prognostic variables. Flow cytometry showed that these signatures reflected gene expression by nonmalignant tumor-infiltrating immune cells.

CONCLUSIONS

The length of survival among patients with follicular lymphoma correlates with the molecular features of nonmalignant immune cells present in the tumor at diagnosis.

From National Cancer Institute (S.S.D., G.W., B.T., A.R., W.H.W., E.S.J., R.S., H.Z., N.G., M.C., L.M.S.); Center for Information Technology (L.Y., J.P.); and National Heart, Lung, and Blood Institute (S.S.D.) — all in Bethesda, Md.; British Columbia Cancer Center, Vancouver, Canada (R.D.G., J.M.C., P.M.L., Q.O.); University of Nebraska Medical Center, Omaha (W.C.C., T.C.G., D.D.W., J.C.L., J.V., J.O.A.); Southwest Oncology Group, San Antonio, Tex. (R.I.F., T.M.G., T.P.M., M.L.); University of Rochester School of Medicine, Rochester, N.Y. (R.I.F.); Oregon Health and Science University, Portland (R.M.B.); University of Arizona Cancer Center, Tucson (L.M.R., T.M.G., T.P.M.); Fred Hutchinson Cancer Research Center, Seattle (M.L.); Norwegian Radium Hospital, Oslo (E.B.S., S.K., H.H., J.D.); Cancer Research UK, St. Bartholomew's Hospital, London (T.A.L., A.J.D., A.J.N.); University of Würzburg, Würzburg, Germany (A.R., H.K.M.-H., G.O.); and University of Barcelona, Barcelona, Spain (E.C., E.M.). Address reprint requests to Dr. Staudt at the National Cancer Institute, Bldg. 10, Rm. 4N114, NIH, Bethesda, MD 20892, or at lstaudt@mail.nih.gov.

N Engl J Med 2004;351:2159-69.

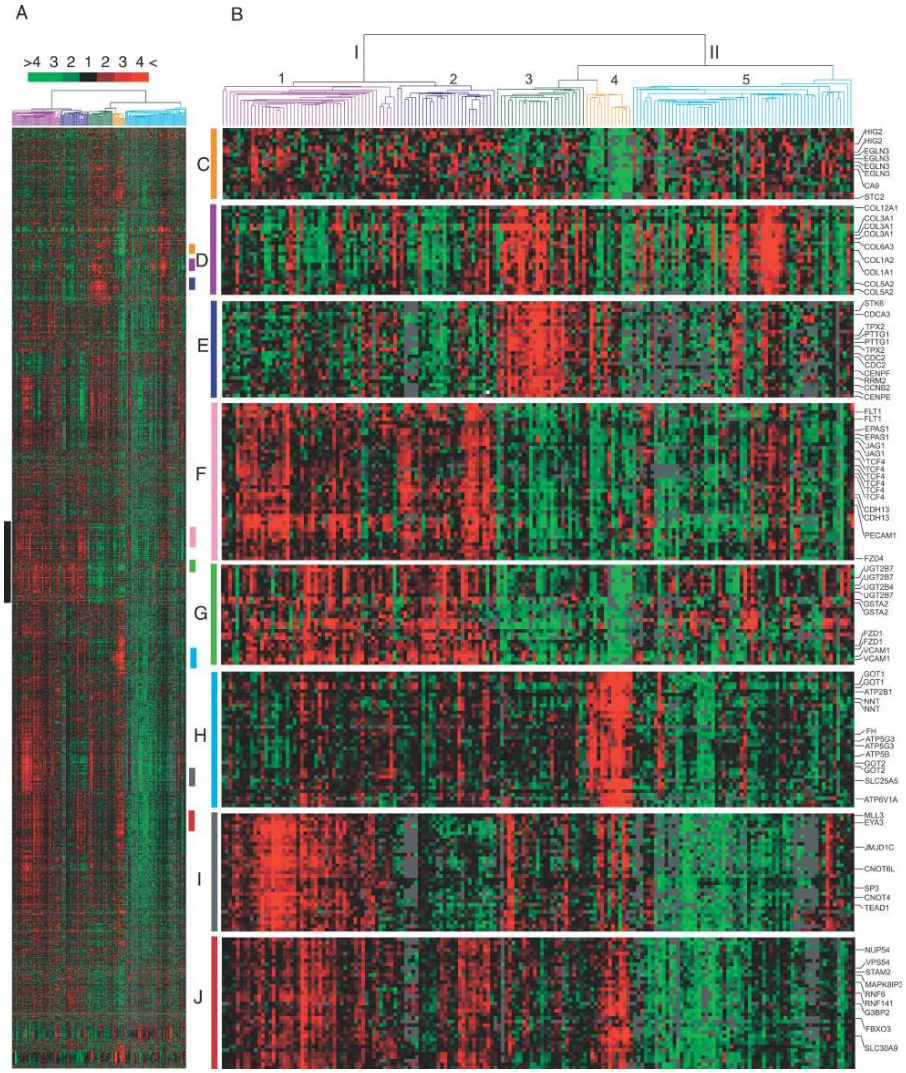
Copyright © 2004 Massachusetts Medical Society.

Downloaded from www.nejm.org at Stanford University on May 10, 2005 .
Copyright © 2004 Massachusetts Medical Society. All rights reserved.

Example

- Kidney cancer study, with Jim Brooks, Hongjuan Zhao: PLOS Medicine 2006
- Gene expression measurements for 14,814 genes on 177 patients- 88 in training set and 89 in test set
- Outcome is survival time. Would like a predictor of survival, for planning treatments, and also would like to understand which genes are involved in the disease

Kidney cancer data



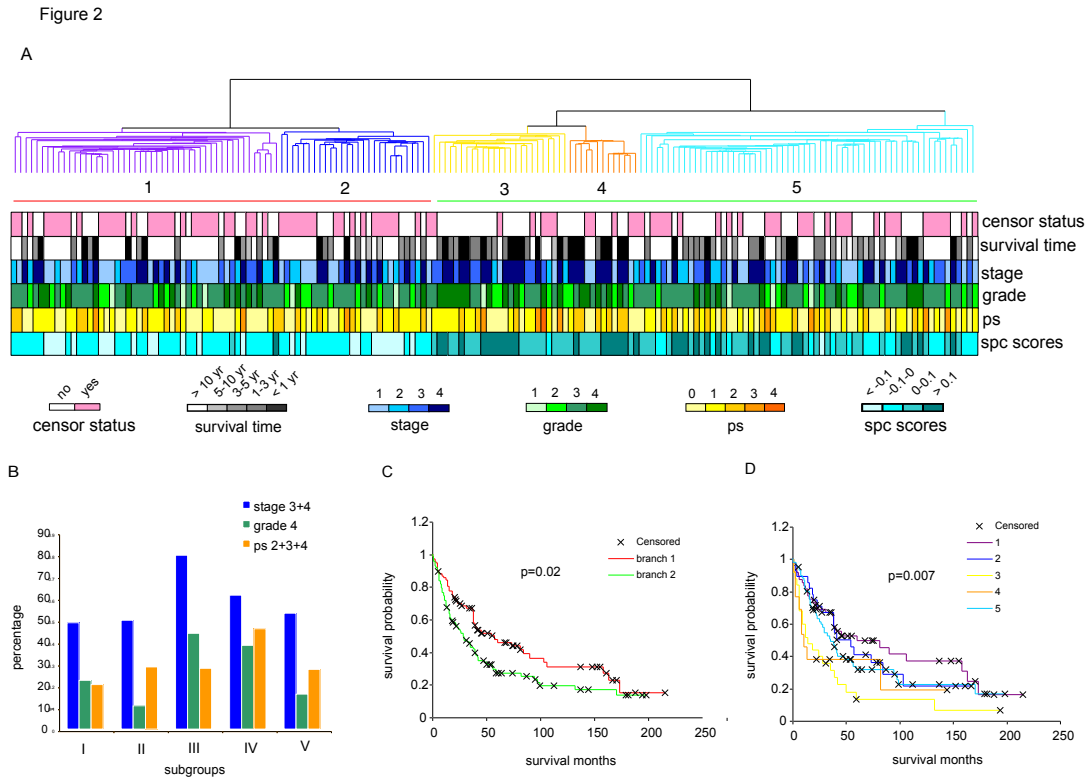
Two approaches

- *Supervised learning:* Some kind of (regularized) regression: eg ridge regression, lasso, partial least squares, SCAD (Fan and Li), elastic net (Zou and Hastie).
- *Unsupervised learning:* cluster the samples into say 2 groups and hope that they differ in terms of survival.

  Not as crazy as it sounds. Used in many microarray studies of cancer from Stanford labs (David Botstein, Pat Brown).

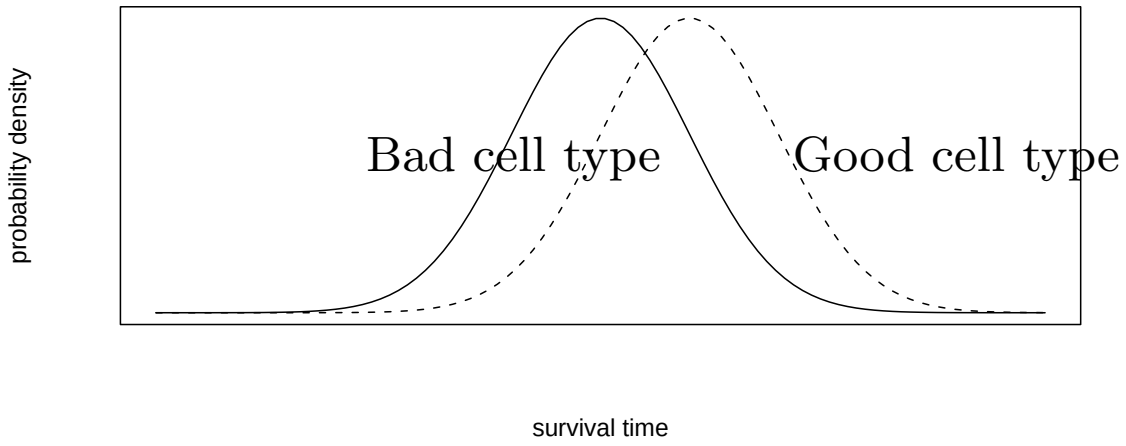
  Idea is to discover biologically distinct and meaningful groups. These groups will tend to be more reproducible than the genes that characterize them (*listen to your collaborators!)*

Unsupervised approach



Semi-supervised approach

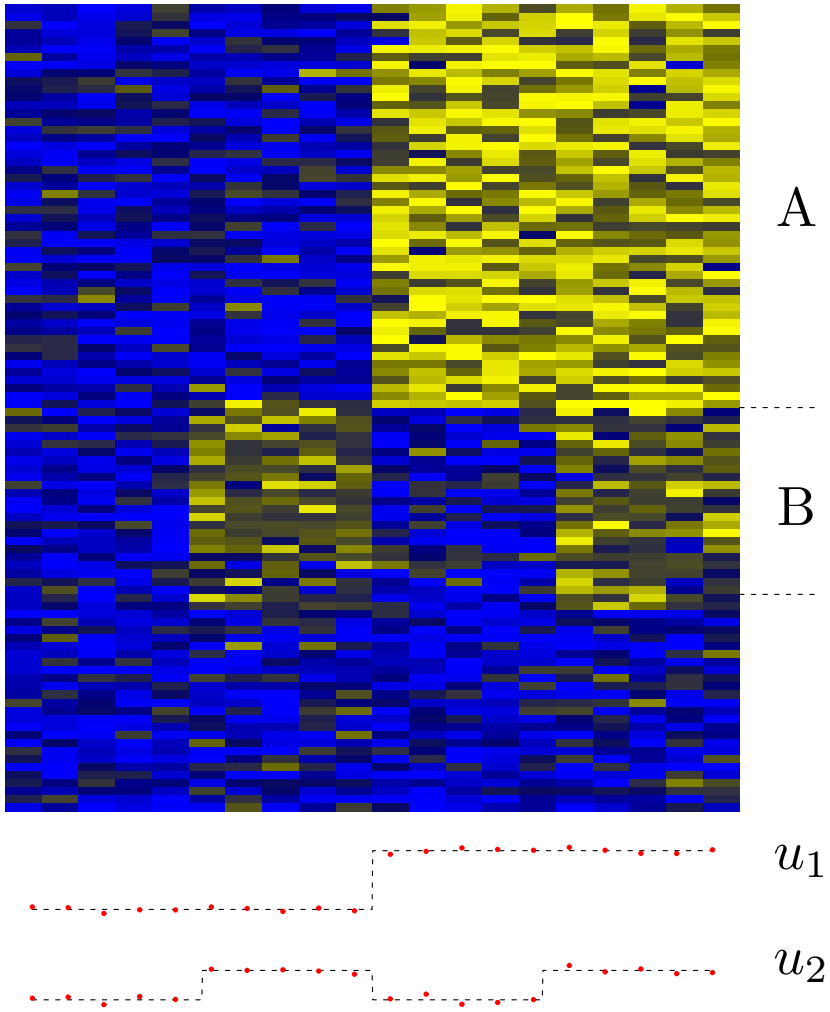
Underlying conceptual model



Supervised Principal components

- Idea is to chose genes whose correlation with the outcome (Cox score) is largest, and using only those genes, extract the first (or first few) principal components.
- Then we use these "supervised principal components" to predict the outcome, in a standard regression or Cox regression model

A toy example



[SHOW MOVIE]

Outline of talk

1. The idea in detail, for (normal) regression and generalized regression models like survival models
2. Underlying latent variable model
3. Summary of some asymptotic results
4. Kidney cancer example
5. Simulation studies, comparison to ridge, lasso, PLS etc
6. "Pre-conditioning" - selecting a smaller set of features for prediction

Supervised principal components

- We assume there are p features measured on N observations (e.g. patients). Let $\mathbf{X}$ be an N times p matrix of feature measurements (e.g. genes), and y the N-vector of outcome measurements.
- We assume that the outcome is a quantitative variable; below we discuss other types of outcomes such as censored survival times.

Supervised principal components

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

Details

- Assume that the columns of $\mathbf{X}$ (variables) have been centered to have mean zero.
- Write the singular value decomposition of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{1}$$

where $\mathbf{U}, \mathbf{D}, \mathbf{V}$ are $N \times m$, $m \times m$ and $m \times p$ respectively, and $m = \min(N-1, p)$ is the rank of $\mathbf{X}$. $\mathbf{D}$ is a diagonal matrix containing the singular values d_j; the columns of $\mathbf{U}$ are the principal components $u_1, u_2, \ldots u_m$; these are assumed to be ordered so that $d_1 \geq d_2 \geq \ldots d_m \geq 0$.

- Let s be the p-vector of standardized regression coefficients for measuring the univariate effect of each gene separately on y:

$$s_j = \frac{x_j^T y}{||x_j||}, \quad \text{(scale omitted)} \tag{2}$$

- Let C_θ be the collection of indices such that $|s_j| > \theta$. We denote by $\mathbf{X}_\theta$ the matrix consisting of the columns of $\mathbf{X}$ corresponding to C_θ. The SVD of $\mathbf{X}_\theta$ is

$$\mathbf{X}_\theta = \mathbf{U}_\theta \mathbf{D}_\theta \mathbf{V}_\theta^T \tag{3}$$

- Letting $\mathbf{U}_\theta = (u_{\theta,1}, u_{\theta,2}, \ldots u_{\theta,m})$, we call $u_{\theta,1}$ the first supervised principal component of $\mathbf{X}$, and so on.

- Now fit a univariate linear regression model with response y and predictor $u_{\theta,1}$,

$$\hat{y}^{\text{spc},\theta} = \bar{y} + \hat{\gamma} \cdot u_{\theta,1}. \tag{4}$$

- Use cross-validation to estimate the best value of θ.

Test set prediction

Given a test feature vector x^*, we can make predictions from our regression model as follows:

1. We center each component of x^* using the means we derived on the training data: $x_j^* \leftarrow x_j^* - \bar{x}_j$.
2. $\hat{y}^* = \bar{y} + \hat{\gamma} \cdot {x_\theta^*}^T w_{\theta,1}$,

where x_θ^* is the appropriate sub-vector of x^*, and $w_{\theta,1}$ is the first column of $\mathbf{V}_\theta \mathbf{D}_\theta^{-1}$.

Easy generalization to non-normal data

- Use a score statistic to assess each gene, and fit a generalized regression model at the end
- Unlike like ridge and lasso, no sophisticated special software is needed

An underlying model

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

$$Y = \beta_0 + \beta_1 U + \varepsilon. \tag{5}$$

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

$$X_j = \alpha_{0j} + \alpha_{1j} U + \epsilon_j, \quad j \in \mathcal{P}. \tag{6}$$

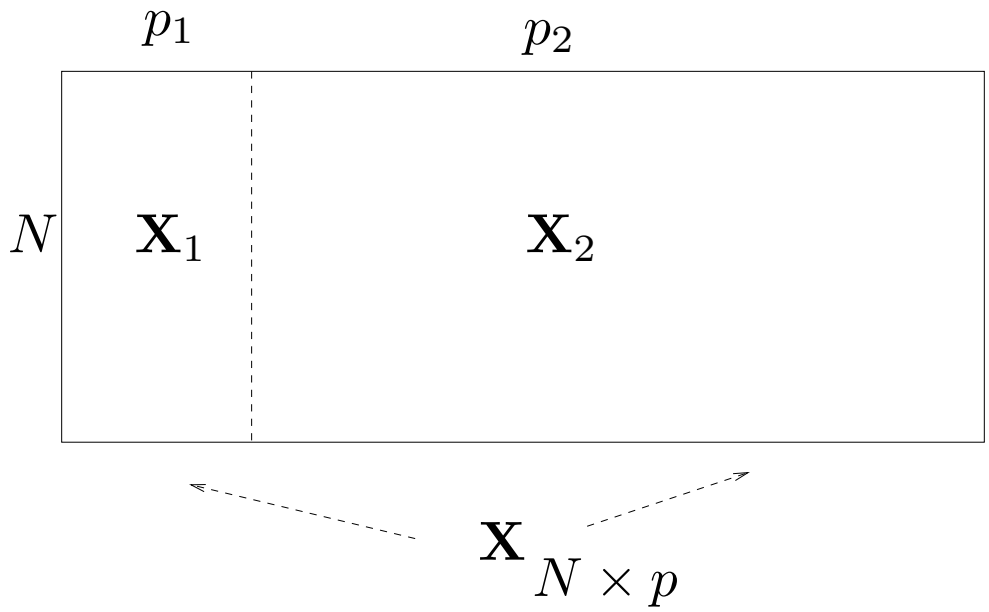
  We also have many additional genes X_k, $k \notin \mathcal{P}$ which are independent of U. We can think of U as a discrete or continuous aspect of a cell type, which we do not measure directly.

- 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 (5) and (6) for data with N samples and p features.



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

We prove:

- Let $\tilde{U}$ be the leading principal component of $\mathbf{X}$ and $\tilde{\beta}$ be the regression coefficient of Y on $\tilde{U}$. Then $\tilde{U}$ is not generally consistent for U and likewise $\tilde{\beta}$ is not generally consistent for β.
- Assume that we are given $\mathbf{X}_1$. Then if $\hat{U}$ is the leading principal component of $\mathbf{X}_1$ and $\hat{\beta}$ be the regression coefficient of Y on $\hat{U}$, these are both consistent.
- If $\mathbf{X}_1$ is not given but estimated by thresholding univariate features scores (as in the supervised principal component procedure), the corresponding $\hat{U}$ and $\hat{\beta}$ are consistent for $K = 1$ component. For $K > 1$, it's a longer story...

Importance scores and reduced models

- Having derived the predictor $u_{\theta,1}$, how do we assess the contributions of the p individual features? It is not true that the features that passed the screen $|s_j| > \theta$ are necessarily important or that they are only important features.
- Instead, we compute the *importance score* as the correlation between each feature and $u_{\theta,1}$: $\text{imp}_j = \text{cor}(x_j, u_{\theta,1})$

Kidney Cancer ctd.

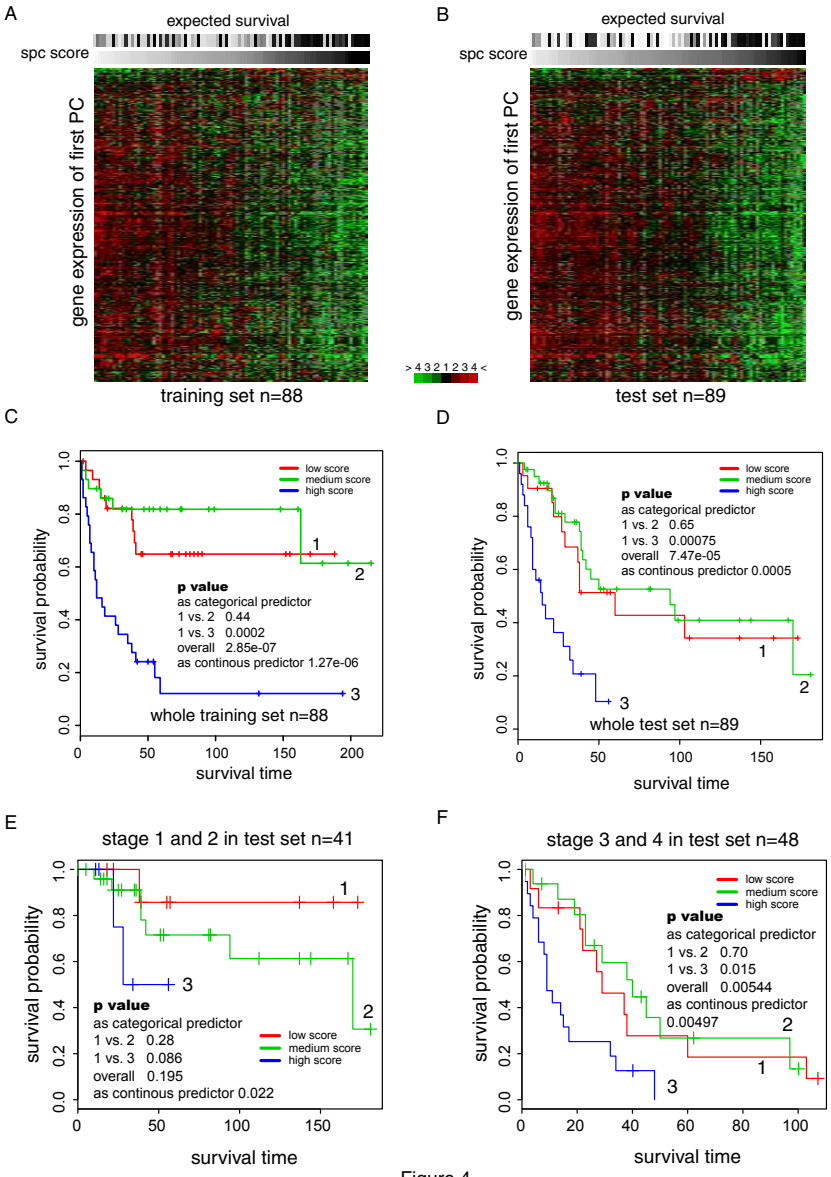
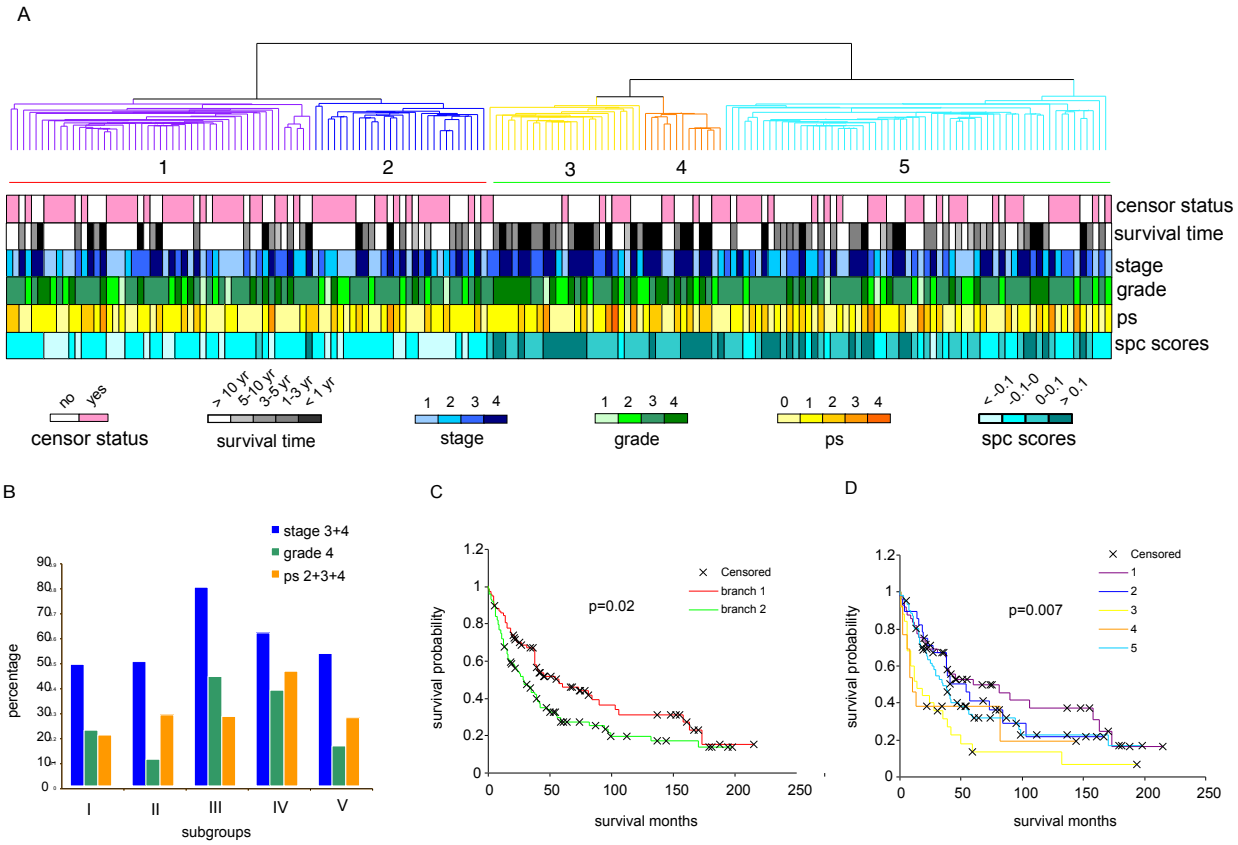


Figure 4

Some results- 200 selected genes

Figure 2


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

Five groups vs SPC

	coef	se(coef)	z	p
gr2	-0.414	0.588	-0.705	0.4800
gr3	0.505	0.580	0.870	0.3800
gr4	-0.977	0.738	-1.323	0.1900
gr5	-0.793	0.507	-1.563	0.1200
spc.pred	8.298	2.588	3.206	0.0013

dropping gr1--gr5: LR test =1.1, 4 degrees of freedom

Some alternative approaches

- *Ridge regression*:

$$\min_{\beta} ||y - \beta_0 - \mathbf{X}\beta||^2 + \lambda||\beta||^2, \tag{7}$$

- *Lasso*:

$$\min_{\beta} ||y - \beta_0 - \mathbf{X}\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{8}$$

- *Partial least squares*:

  Standardize each of the variables to have zero mean and unit norm, and compute the univariate regression coefficients $w = \mathbf{X}^T y$.

- define $u_{\text{PLS}} = \mathbf{X}w$, and use it in a linear regression model with y.

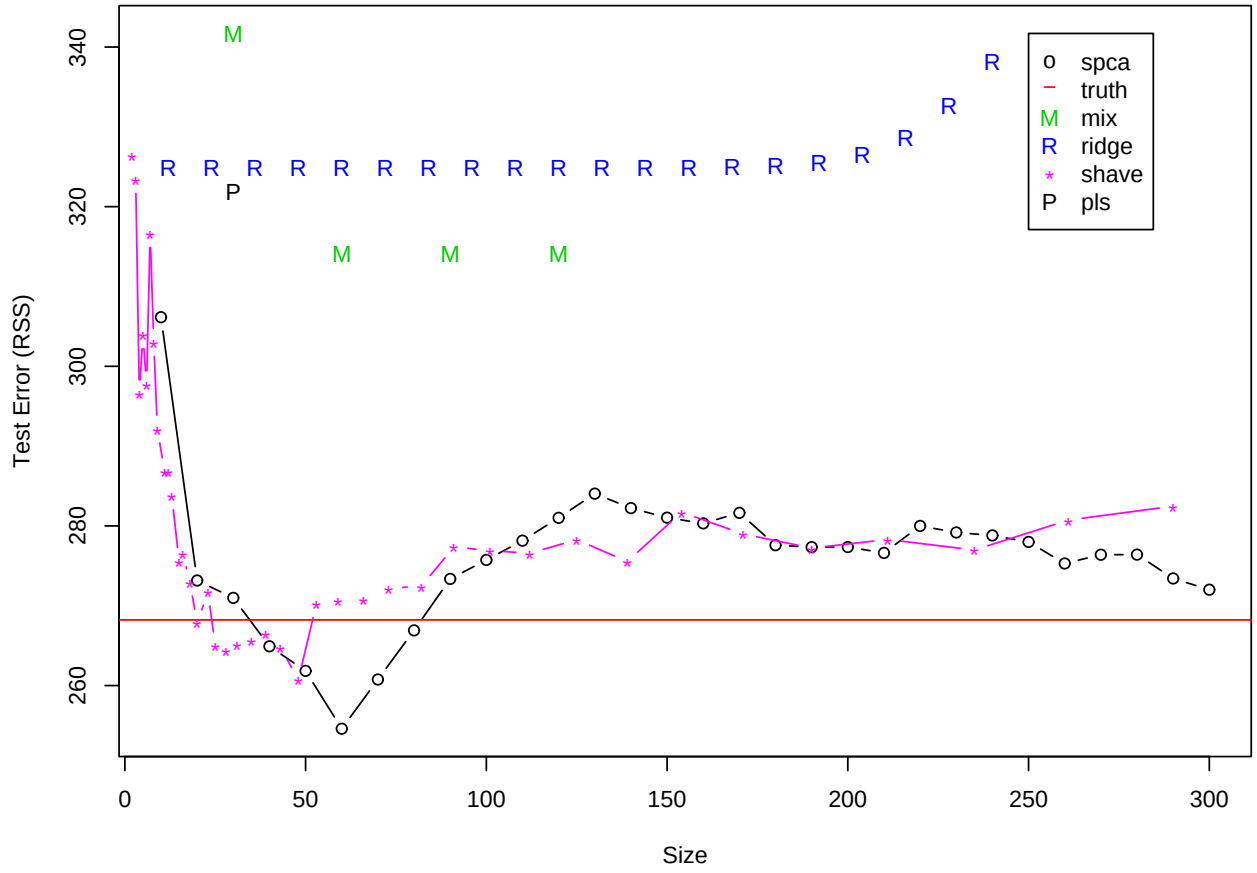
- *Supervised gene shaving*: Find $z = \mathbf{X}v$ to solve:

$$\max_{||v||=1} (1-\alpha)\text{Var}(z) + \alpha\text{Cov}(z,y)^2 \quad \text{s.t. } z = \mathbf{X}v. \tag{9}$$

We also call this a "mixed covariance" method.

Simulation studies

Data generated from a latent-variable model; first 50 features are important


Test Error (RSS)
Size
260
280
300
320
340
0
50
100
150
200
250
300
o spca
truth
M mix
R ridge
* shave
P pls

Simulation study

Gaussian prior for true coefficients

Method	CV Error	Test Error
PCR	293.4 (17.21)	217.6 (10.87)
PCR-1	316.8 (20.52)	239.4 (11.94)
PLS	291.6 (13.11)	218.2 (12.03)
Ridge regression	298.0 (14.72)	224.2 (12.35)
Lasso	264.0 (13.06)	221.9 (12.72)
Supervised PC	233.2 (11.23)	176.4 (10.14)
Mixed var-cov.	316.7 (19.52)	238.7 (10.24)
Gene shaving	223.0 (8.48)	172.5 (9.25)

More survival studies

	(a) DLBCL			(b) Breast Cancer		
Method	R^2	p-val	NC	R^2	p-val	NC
(1) SPCA	0.11	0.003	2	0.27	2.1×10^{-5}	1
(2) PC Regression	0.01	0.024	2	0.22	0.0003	3
(3) PLS	0.10	0.004	3	0.18	0.0003	1
(4) Lasso	0.16	0.0002	NA	0.14	0.001	NA
	(c) Lung Cancer			(d) AML		
Method	R^2	p-val	NC	R^2	p-val	NC
(1) SPCA	0.36	1.5×10^{-7}	3	0.16	0.0013	3
(2) PC Regression	0.11	0.0156	1	0.08	0.0376	1
(3) PLS	0.18	0.0044	1	0.07	0.0489	1
(4) Lasso	0.26	0.0001	NA	0.05	0.0899	NA

SPC vs Partial least squares

Can apply PLS after hard-thresholding of features.

Now PLS uses

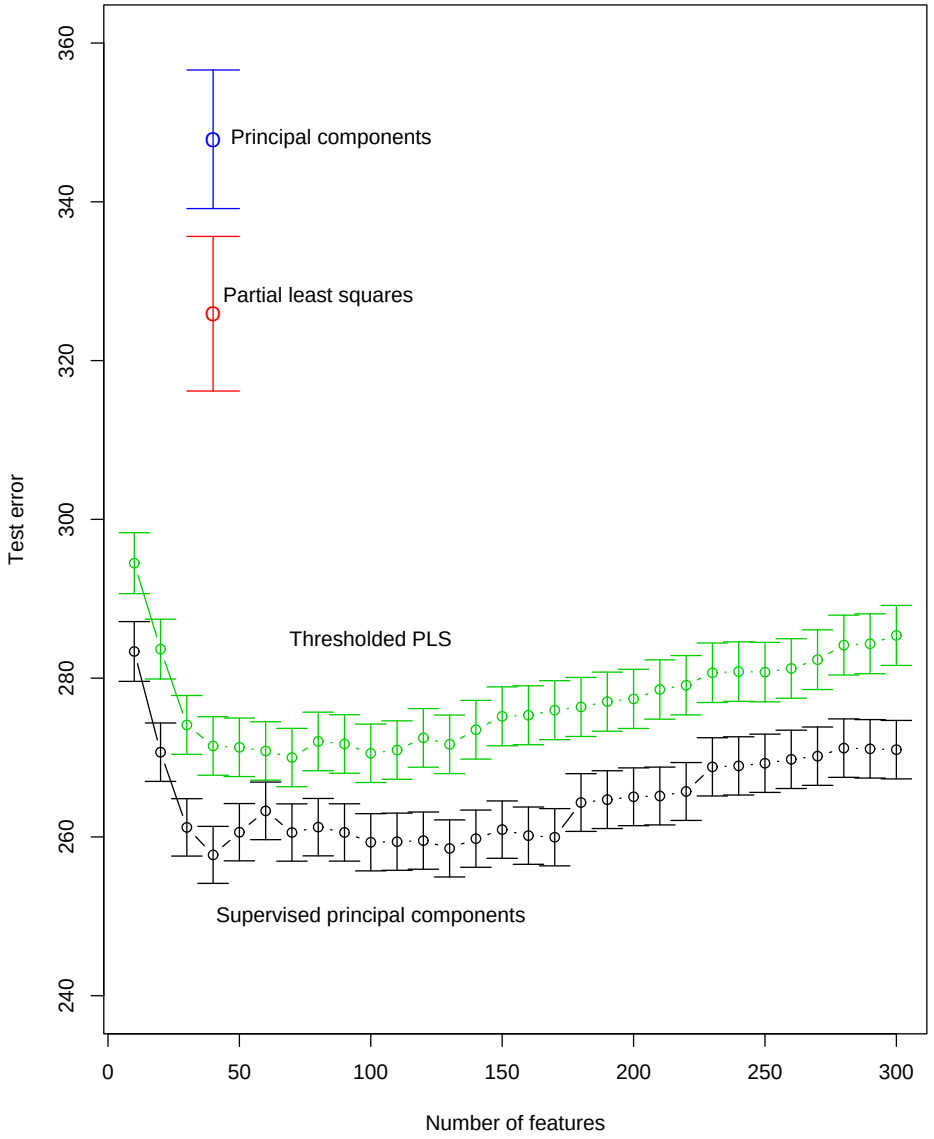
$$\mathbf{z} = \sum_{j \in \mathcal{P}} \langle \mathbf{y}, \mathbf{x}_j \rangle \mathbf{x}_j \tag{10}$$

where $\langle \mathbf{y}, \mathbf{x}_j \rangle = \sum_i y_i x_{ij}$, the inner product between the jth feature and the outcome vector $\mathbf{y}$.

In contrast, supervised principal components direction $\hat{\mathbf{u}}$ satisfies

$$\hat{\mathbf{u}} = \sum_{j \in \mathcal{P}} \langle \hat{\mathbf{u}}, \mathbf{x}_j \rangle \mathbf{x}_j \tag{11}$$

SPC vs Partial least squares ctd


Principal components
Partial least squares
Thresholded PLS
Supervised principal components
Test error
Number of features
240
260
280
300
320
340
360
0
50
100
150
200
250
300

Take home messages

- One key to the success of Supervised PC is the hard-thresholding (discarding) of noisy features– giving them low weight (as in ridge regression) is not harsh enough
- Given the chosen features, SPC makes more efficient use of the information than does partial least squares.

Pre-conditioning to find a reduced model

Paul, Bair, Hastie, Tibshirani (2007) submitted

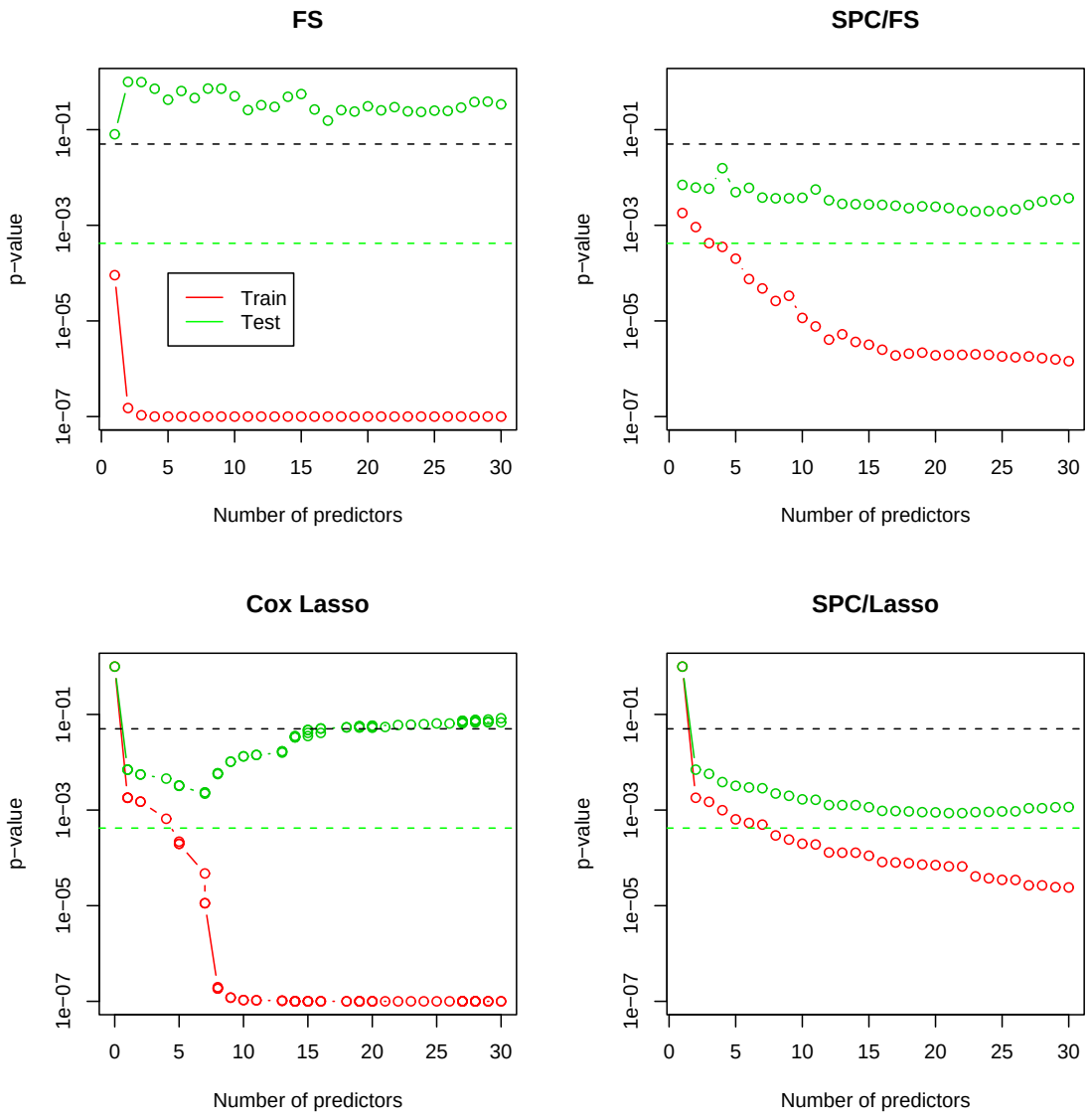
- Supervised principal components finds a good predictive model, but not necessarily a very parsimonious one.
- Features that pass the initial filter might not be the ones that are most correlated with the supervised principal component
- Highly correlated features will all tend to be included together
- need to do some sort of model selection, using eg forward stepwise regression or the lasso

Pre-conditioning continued

- *Usual approach*: apply forward stepwise regression or the lasso to the outcome y. There has been lots of recent work of the virtues of the lasso for model selection- Donoho, Meinhausen and Buhlmann, Meinhausen and Yu;
- *Pre-conditioning idea*: 1) compute supervised principal components predictions $\hat{y}$, then 2) apply forward stepwise regression or the lasso to $\hat{y}$
- *Why should this work?* The denoising of the outcome should help reduce the variance in the model selection process.

Kidney cancer again

Pre-conditioning pares the number of genes down from 200 to 20.


FS
SPC/FS
Cox Lasso
SPC/Lasso
p−value
1e−01
1e−03
1e−05
1e−07
0
5
10
15
20
25
30
Number of predictors
Train
Test

Asymptotics

- we show that the pre-conditioning procedure, combining supervised principal components with the lasso, under suitable regularity conditions leads to asymptotically consistent variable selection in the Gaussian linear model setting.
- We also show that the errors in the pre-conditioned response have a lower order than those in the original outcome variable.

Conclusions

- supervised principal components is a promising tool for regression when $p >> N$.
- computationally simple, interpretable. A useful competitor to ridge regression, lasso etc.
- papers/software available at

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