Supervised principal components

Robert Tibshirani
Stanford

MMDS conference June 2006

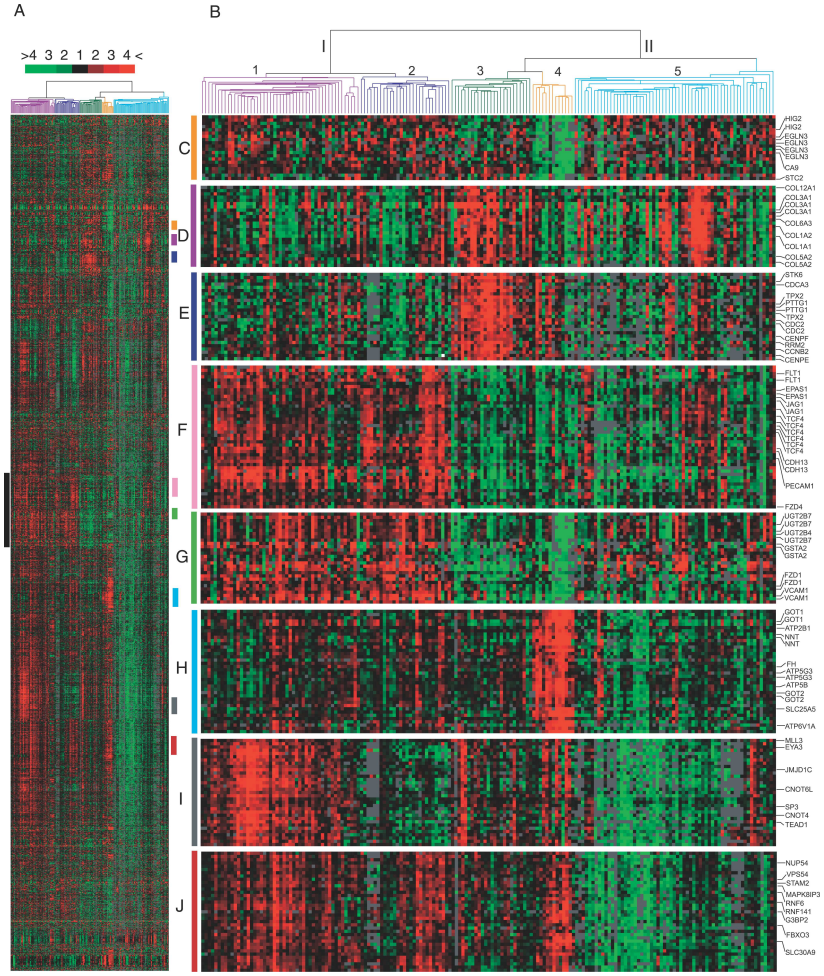
Co-authors- Eric Bair, Trevor Hastie, Debashis Paul- JASA 2005

(Google $->$ Tibshirani) for papers and software

Only 20 slides!

Supervised principal components

- method for predicting a quantitative measurement (especially survival time of a patient), useful when number of predictors $p >> N$, the sample size
- motivated by genomics applications, but could be useful in other problems
- Main application: gene expression measurements from microarrays. Each feature measurement x_{ij} is the expression of gene j for patient sample i.
- A possibly censored survival time y_i is also available for each patient. We wish to find genes j whose high or low expression is predictive of patient survival.
- In our main example: $N = 180$ patients with kidney cancer, with measurements on 14,000 genes.


A
B
>4 3 2 1 2 3 4 <
I
II
1
2
3
4
5
C
D
E
F
G
H
I
J
HIG2
HIG2
EGLN3
EGLN3
EGLN3
EGLN3
CA9
STC2
COL12A1
COL3A1
COL3A1
COL3A1
COL6A3
COL1A2
COL1A1
COL5A2
COL5A2
STK6
CDCA3
TPX2
PTTG1
PTTG1
TPX2
CDC2
CDC2
CENPF
RRM2
CCNB2
CENPE
FLT1
FLT1
EPAS1
EPAS1
JAG1
JAG1
TCF4
TCF4
TCF4
TCF4
TCF4
TCF4
CDH13
CDH13
PECAM1
FZD4
UGT2B7
UGT2B7
UGT2B4
UGT2B7
GSTA2
GSTA2
FZD1
FZD1
VCAM1
VCAM1
GOT1
GOT1
ATP2B1
NNT
NNT
FH
ATP5G3
ATP5G3
ATP5B
GOT2
GOT2
SLC25A5
ATP6V1A
MLL3
EYA3
JMJD1C
CNOT6L
SP3
CNOT4
TEAD1
NUP54
VPS54
STAM2
MAPKBIP3
RNF6
RNF141
G3BP2
FBXO3
SLC30A9

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. A bit of a "crapshoot".
- *Supervised approach*- build a prediction model for survival time as a function of gene expression. Some sort of regularization is needed (eg ridge regression or lasso (L_1)), since number of genes $\gg$ number of patients. Tends to overfit- gets confused by noisy genes.

The Lasso

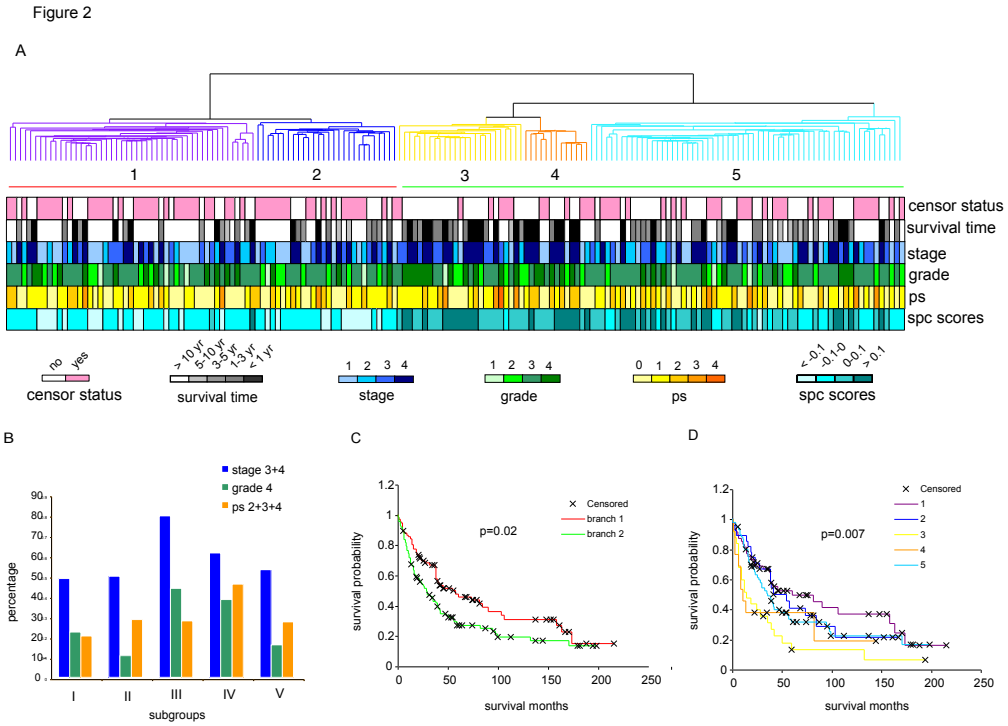
(Tibshirani, 1996)

- Least squares fitting with an L1 constraint
- Minimize over $\beta = (\beta_1, \beta_2, \ldots \beta_p)$:

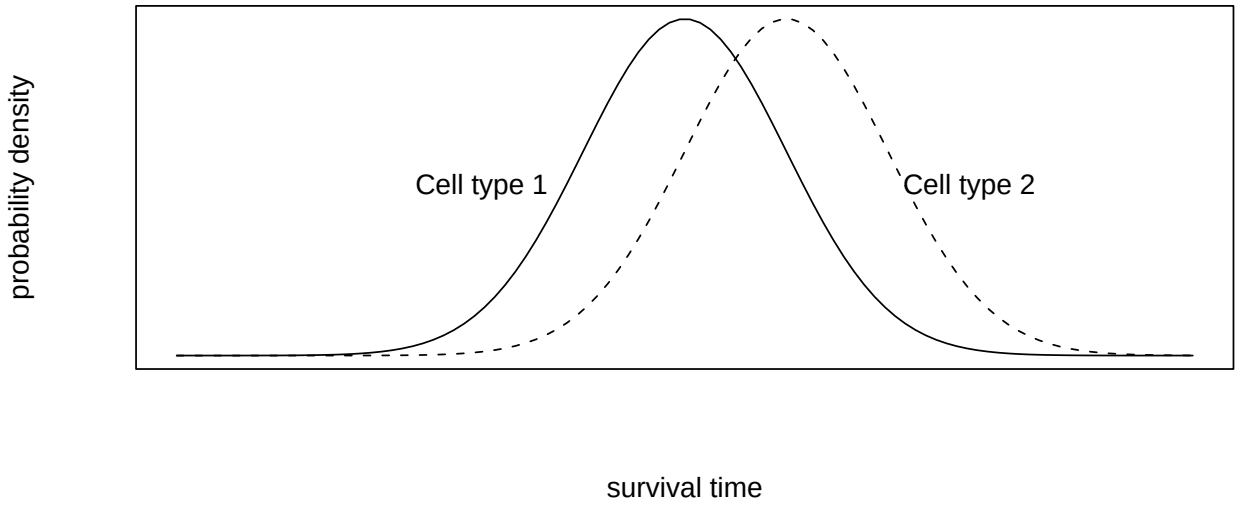
$$\sum_i (y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_j |\beta_j|$$

- Gives a sparse set of coefficients $\hat{\beta}_j$
- LAR algorithm of Efron, Hastie, Johnstone, Tibshirani (2002) gives a fast method for computing entire path of solutions. Has led to other path algorithms for SVMs, etc.

Kidney cancer study



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

Principal components

- Principal components are the linear combinations of the features showing the highest variation across the samples
- High variation is potentially interesting, but in our setting, only if it correlates with the outcome of interest
- Need to encourage principal components to find linear combinations that possess *both* high variance and significant correlation with the outcome

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.

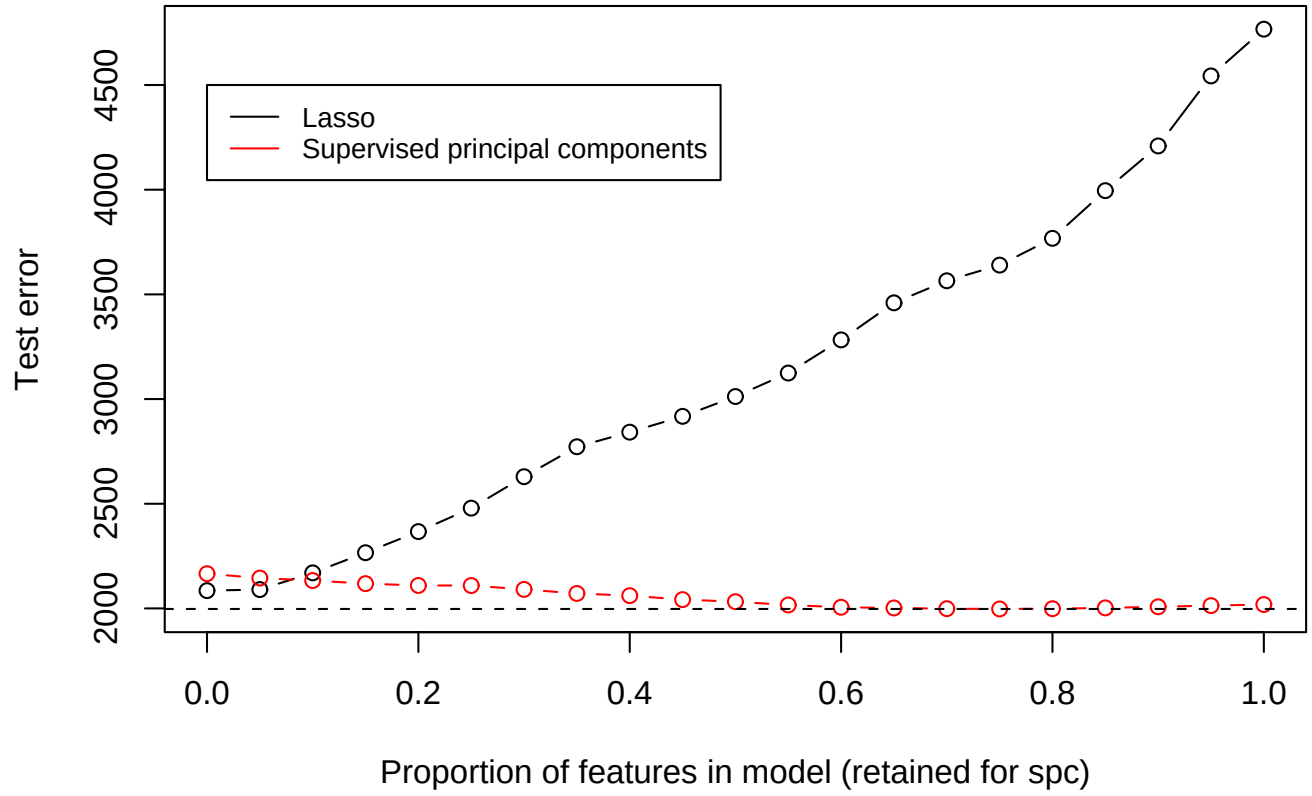
These "supervised principal components" are used to predict the outcome.

1. Compute correlation (Cox scores) between each feature (gene) and survival time
2. Form a reduced data matrix consisting of only those features whose correlation 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 survival prediction model

[SHOW MOVIE]

Kidney cancer example

For illustration, we treat all survival times as uncensored



Underlying latent variable 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

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

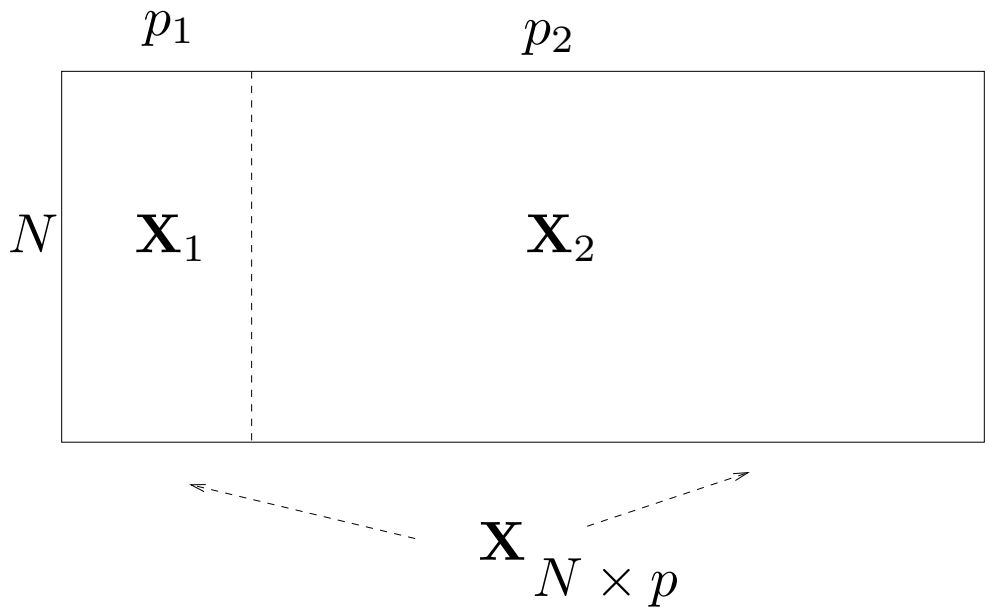
We also have many additional genes X_k, $k \notin \mathcal{P}$ which are independent of U. We can think of U as a discrete or continuous aspect of a cell type, which we do not measure directly.

Consistency

- In this model, standard principal components is not consistent in general— the large number of "noise" features corrupts the estimate
- In contrast, the supervised PC approach estimates the latent variables consistently as $p, N \to \infty$ ($p = \#$ of features, $N = \#$ of samples)

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.



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

Kidney cancer study

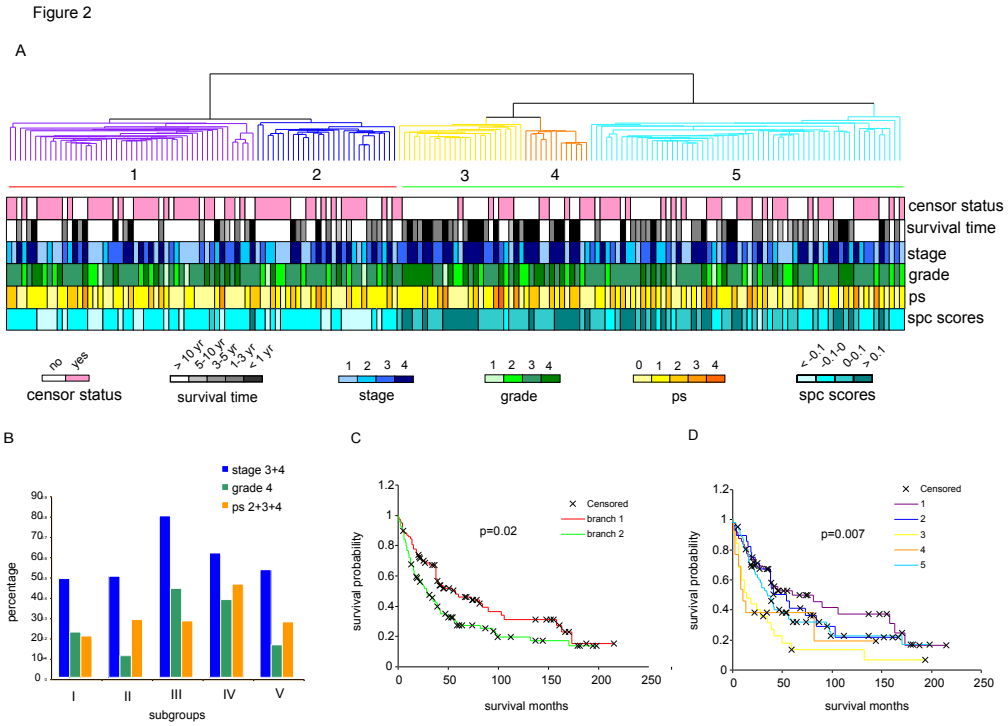
with Jim Brooks, Hongjuan Zhao

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

Results

- Supervised principal component score correlates with groups found by hierarchical clustering, but is a stronger predictor of survival
- Supervised principal component score has additional predictive power over and above traditional clinical measurements like tumor grade and stage

Kidney cancer study ctd...

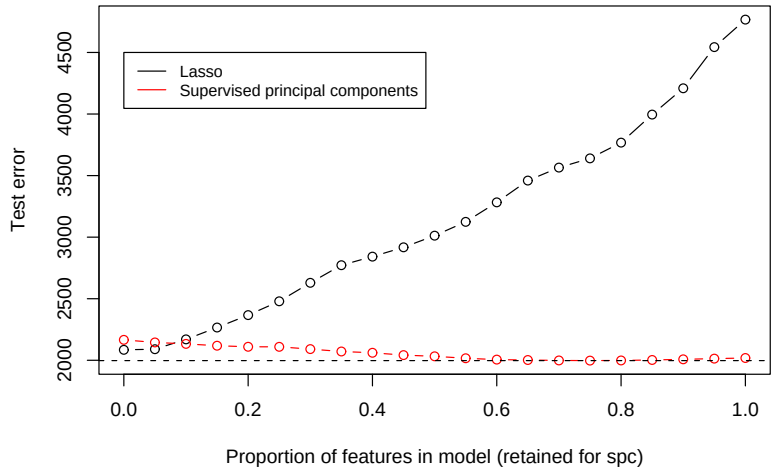


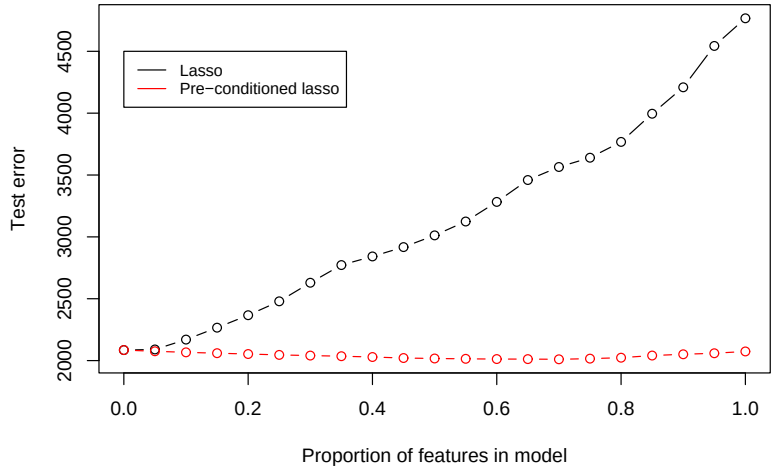
Further challenges

- given the SPC predictor $\hat{y}$, how can we find a reduced predictor that uses only a small number of genes?
- one approach: "pre-conditioning"– apply lasso to outcome $\hat{y}$. Works much better than usual lasso (applied to raw outcome) when $p \gg N$. (Paul, Bair, Hastie, Tibshirani)

Kidney cancer example- again

For illustration, we treat all survival times as uncensored


Lasso
Supervised principal components
Test error
2000
2500
3000
3500
4000
4500
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of features in model (retained for spc)


Lasso
Pre−conditioned lasso
Test error
2000
2500
3000
3500
4000
4500
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of features in model