Working with a Response,

and supervised dimension reduction

Working with a response

Using microarray data to

- Explain/predict a classification

discriminate cancer status or types, and identify the global transcriptional signatures of cancer (which genes matter?)

- Explain/predict a measurable, continuous response

Model the intensity of a chemical in reaction to a stress, and identify the global transcriptional signatures of stress response (which genes matter?)

- also, explain/predict survival times (a special type of response)

Model the survival time following a treatment, and identify the global transcriptional signatures of treatment response (which genes matter?)

An Example: Using MA data to discriminate Leukemia types

Leukemia classification study, **Golub *et al*. (1999) *Science***.
(publicly available microarray data set, "model" for classification approaches)

- $i = 1 \ldots N\ (72)$ individuals diagnosed with one of 2(3) types of leukemia:
  Acute Lymphoblastic (ALL); ALL B-cell, ALL T-cell
  Acute Myeloid (AML)

Response (categorical)

- Samples from their bone marrow or blood

↓

Affymetrix chips (6817 human genes)

↓

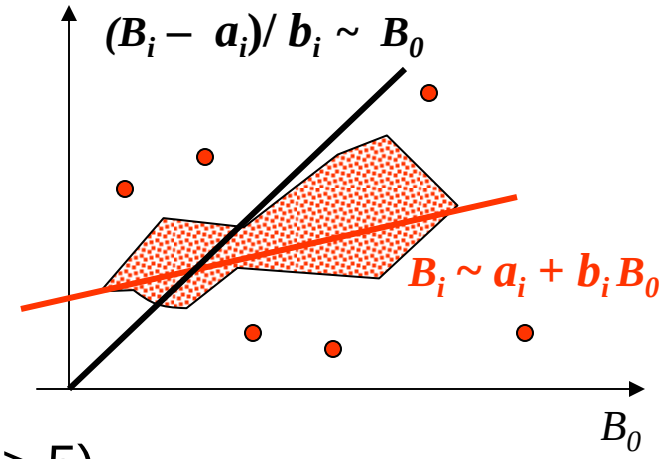
Expression readings

Predictors

Data Preprocessing

- Normalization: linear regression to an arbitrary sample for comparability (correct for chip-specific slope and intercept, assuming most genes do not show systematic expression variation)

$$X_{ji} = \log\left(\frac{B_{ji} - a_i}{b_i}\right)$$



- Filtering ($max - min > 500$ and $max / min > 5$)

$j = 1 \ldots p$ (3571) genes

- Centering by chip, again for comparability (remove any residual chip-specific level)

$$X_{ji} \leftarrow (X_{ji} - \overline{X}_i)$$

(same as in previous analyses of the same data)

$$\begin{pmatrix} X_{11} & \dots & X_{1i} & \dots & X_{1N} \\ \vdots & & & & \\ \vdots & & \vdots & & \vdots \\ \vdots & & & & \\ X_{j1} & \dots & X_{ji} & \dots & X_{jN} \\ \vdots & & \vdots & & \vdots \\ X_{p1} & \dots & X_{pi} & \dots & X_{pN} \end{pmatrix}$$

Expression data

$$\begin{bmatrix} Y_{bi,1} & \dots & Y_{bi,i} & \dots & Y_{bi,j} \\ Y_{tri,1} & \dots & Y_{tri,i} & \dots & Y_{tri,N} \end{bmatrix}$$

Classification response(s)

Bi: Acute Lymphoblastic (ALL)
Acute Myeloid (AML)

Tri: ALL B-cell, ALL T-cell, AML

Aims:

1. Construct a classifier to predict Y (Leukemia type) on the basis of the X's (gene expression levels)

2. Rank genes in terms of their importance, identify relevant X's: variable, i.e. gene, selection.

Serious Trouble

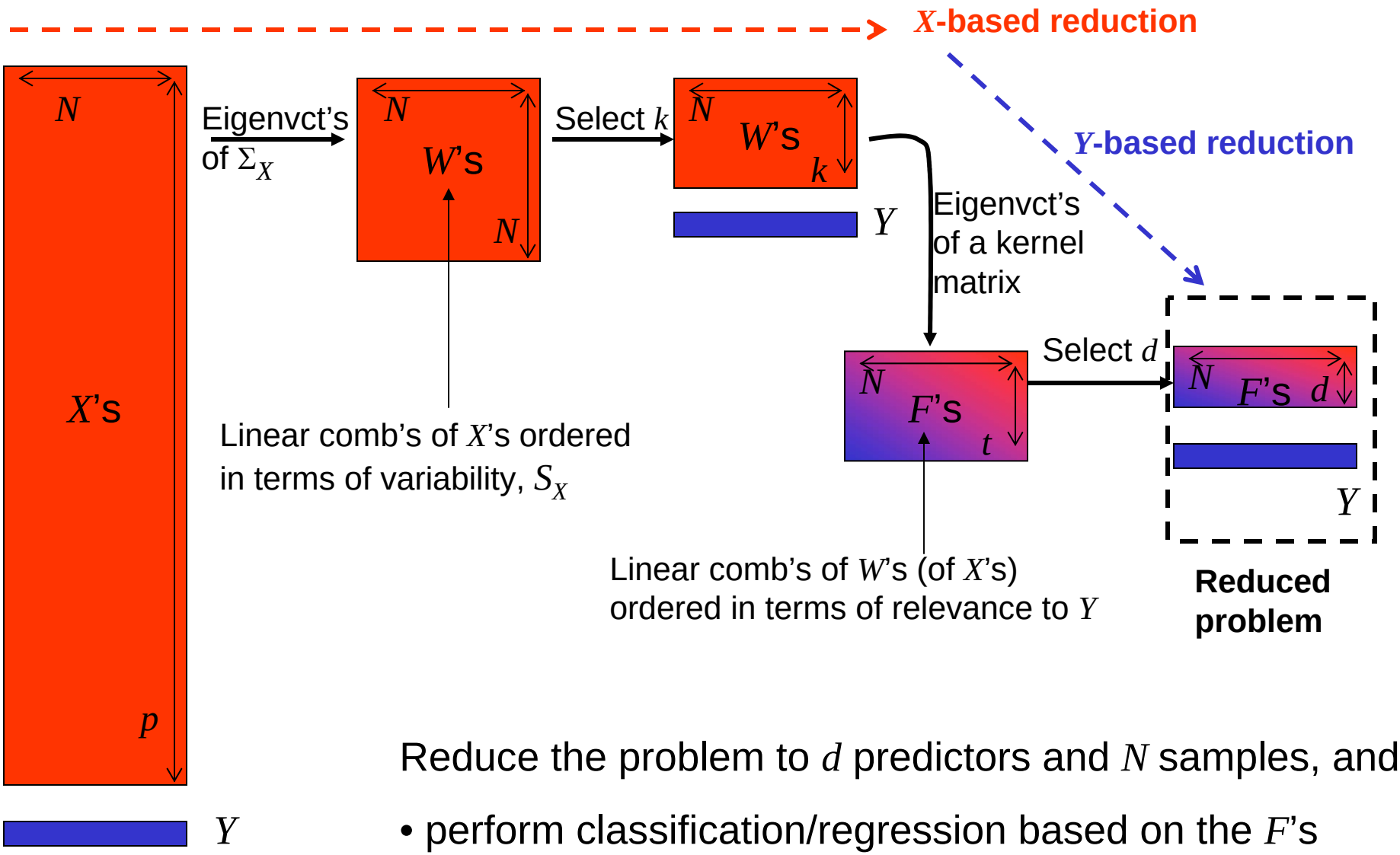
applying standard statistical tools:

large-p-small-N.

The number of predictors p (3571) exceeds by orders of magnitude the number of available samples N (72) .

Very high-dimensional "feature space" sparsely populated by data points.

One approach



Reduce the problem to d predictors and N samples, and

- perform classification/regression based on the F's
- *ex-post*, identify relevant X's (genes) using their "loadings" in the F's

X-based reduction: Principal Component Analysis

Spectral decomposition of the var/covariance matrix of the predictor data

$$\Sigma_X = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(X_i - \overline{X})' = \sum_{j=1}^{p}\lambda_j v_j v_j^T$$

$\{v_1 \dots v_p\}$ Eigenvectors, orthogonal directions ranked in terms of…

$\lambda_1 \geq \lambda_2 \dots \geq \lambda_p$ Eigenvalues, variability (if $N<p$, from N on = 0)

$W_1 = v_1^T X \quad \dots \quad W_k = v_k^T X$

$S_X = Span(v_1 \dots v_k)$

k-dimensional proj. best preserving the variability structure of the data cloud.

How many should be kept, *k* ?

… relevant for the variability structure of the X's

… use to retain "signal" and weed out "noise"

No targeting of the response variable Y .

We need to apply PCA to restrict the analysis to a space in which the data in non-singular… Why not go all the way down to d?

the target here is the "regression information", not the variability structure of the X's; want to retain the information X carries on Y.

$$F_1 = \beta_1^T X \quad \ldots \quad F_d = \beta_d^T X$$

$$S_{Y|X} = Span(\beta_1 \ldots \beta_d) \quad d \text{ small}$$

d-dimensional projection, within S_X , best preserving info on response

Perform a "supervised dimension reduction"

… through the eigenvectors of a kernel matrix

Kernel Matrices

Slice Y in H levels if continuous, or take the H classes, and consider

$$\Sigma_W = \frac{1}{N}\sum_{i=1}^{N}(W_i - \overline{W})(W_i - \overline{W})'$$ Overall var/cov matrix of W

$$\mathrm{B} = \sum_{h=1}^{H}\frac{N_h}{N}(\overline{W}_h - \overline{W})(\overline{W}_h - \overline{W})' \quad , \quad \overline{W}_h = \frac{1}{N_h}\sum_{Y_i \in h} W_i$$ "Between"

$$\mathrm{A} = \sum_{h=1}^{H}\frac{N_h}{N}\left(\frac{1}{N_h}\sum_{Y_i \in h}(W_i - \overline{W}_h)(W_i - \overline{W}_h)'\right)$$ "Within"

Two kernels: between matrix normalized by

$\Theta_{LDA} = \mathrm{A}^{-1/2}\mathrm{B}\mathrm{A}^{-1/2}$ ← within matrix (LINEAR DISCRIMINANT ANALYSIS)

$\Theta_{SIR} = \Sigma_W^{-1/2}\mathrm{B}\Sigma_W^{-1/2}$ ← overall matrix (SLICED INVERSE REGRESSION)

Well known "sufficient dimension reduction" technique
K.C. Li (1991) *JASA*; Cook and Weisberg, Wiley.

$$\Theta = \sum_{j=1}^{k}\theta_j t_j t_j'$$

spectral decomposition (all eigenvalues after t=H-1 are = 0)

take first d eigenvectors (largest eigenvalues)

How many should be kept, d ?
… relevant for the response Y

Under assumptions, for instance

• the predictor vector is Gaussian within each given Y class/slice (LDA)

$$(W \mid Y \in h) \sim N(\mu_h, \Delta_h) \quad h = 1 \ldots H$$

• the predictor vector satisfies (SIR)

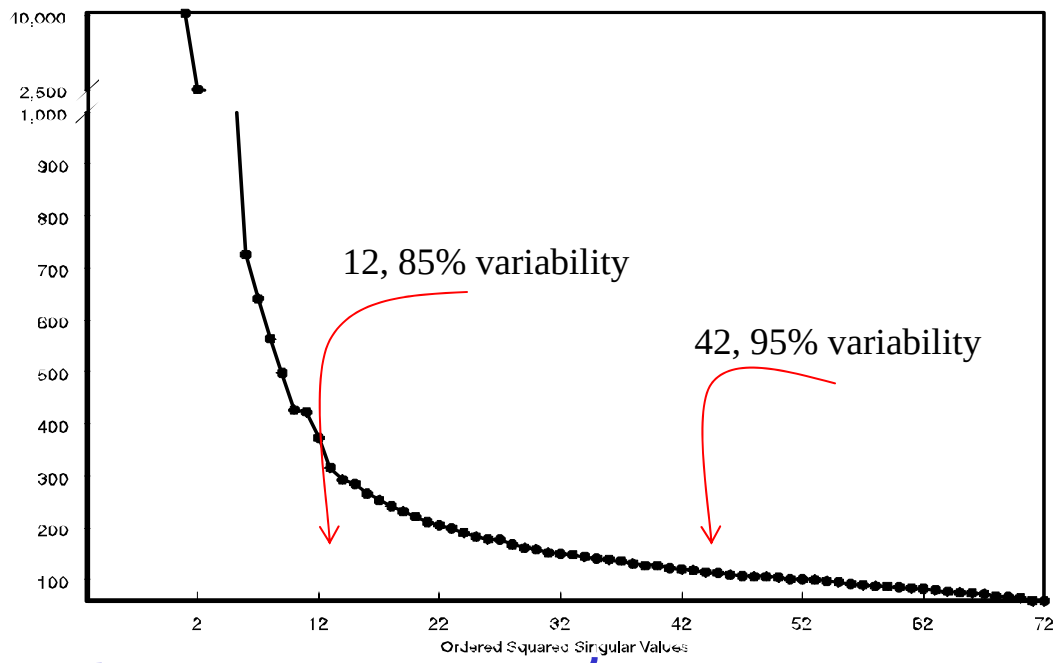
$$\begin{array}{ll} (\text{linearity}) & \mathrm{E}(W \mid F) = \text{linear in } F \\ (\text{const. cov}) & \mathrm{Var}(W \mid F) = \text{constant } C \end{array}$$

tests on the eigenvalues can be performed to make an inference on d.

But with these kernel we have necessarily $d \leq H$-1.

Back to the Leukemia example: (Chiaromonte and Martinelli, 2002)

How to pick *k*?



Only up to $j=N$, equal 0 after that

(ordered) W's, linear comb's of X's from PCA (SVD) $\longrightarrow$ S_X

(ordered) F's, linear comb's of W's from SIR $\longrightarrow$ $S_{Y|X} \subseteq S_X$

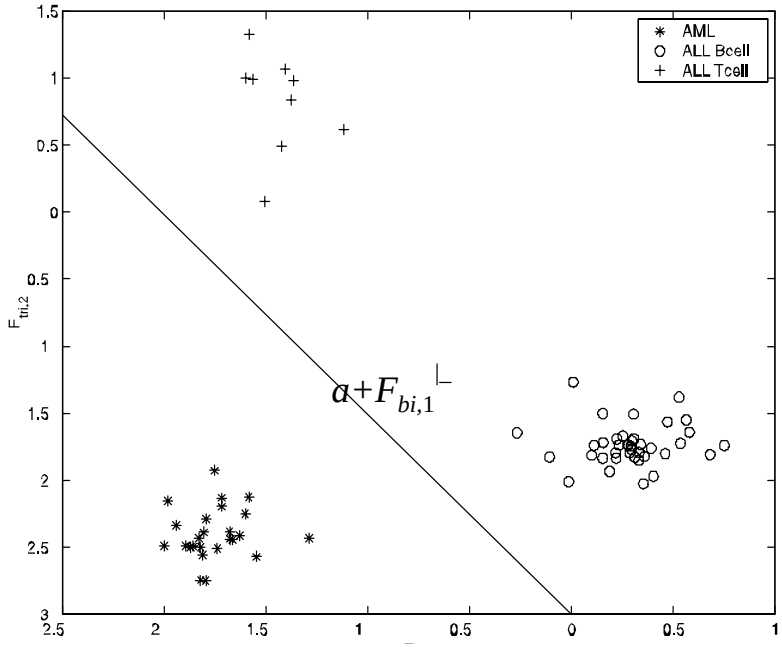
(–) Weed out "noise"

Retain "signal" **(+)**

More dof's (N-k) for SIR to work accurately.

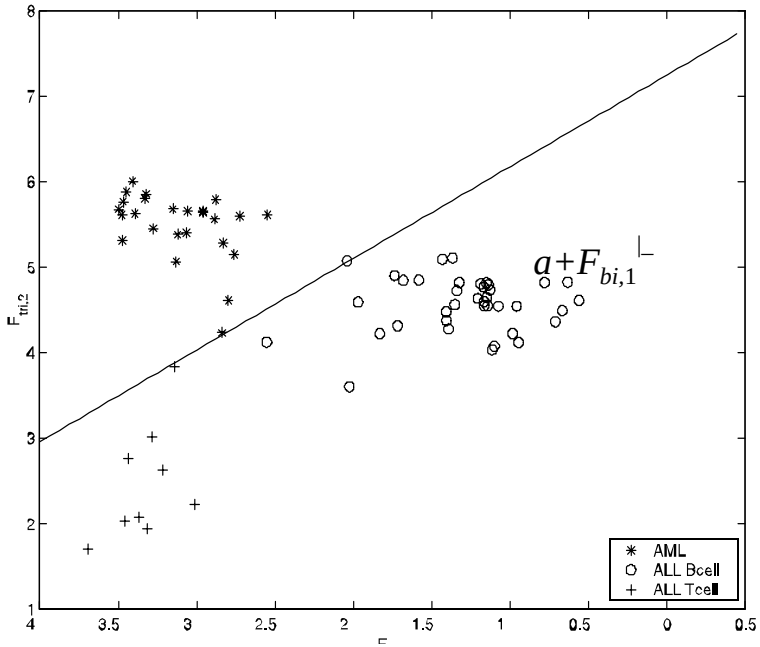
Less restrictions on location of $S_{Y|X}$. More basic patterns for SIR to work with (avoid "distorting" by undercutting input).

SIR on 42 PCA directions



Doing better by fitting “noise”?

SIR on 12 PCA directions



Catching only a projection of $S_{Y|X}$?

- $F_{bi,1}$ contained in *Span(*$F_{tri,1}$, $F_{tri,2}$*)*.
- SIR find at most H -1 directions. In both cases it detects one relevant direction for Y_{bi}, and two relevant directions for Y_{tri}.
- SIR seems very effective in processing what is fed into it, we can “dose” the feed.

Identifying relevant genes ex-post: one approach

Rank genes in terms of proximity to $S_{Y|X}$; individual contributions

$$R_j^2 = R^2(X_j \text{ on } F_{bi,1})$$

$$R_j^2 = R^2(X_j \text{ on } F_{tri,1}, F_{tri,2})$$

How many genes are closer to the *F*'s than could be expected "by chance"?

Along the ranking, "group" contribution of top-ranking genes (accounts for correl's)

$$R_{\le j}^2 = R^2\left(F_{bi,1} \text{ on } X_{(1)} \ldots X_{(j)}\right)$$

$$R_{\le j}^2 = R^2\left((\tau_1 F_{tri,1} + \tau_2 F_{tri,2}) \text{ on } X_{(1)} \ldots X_{(j)}\right)$$

How many among the closest genes are needed to produce a satisfactory reconstruction of the *F*'s?

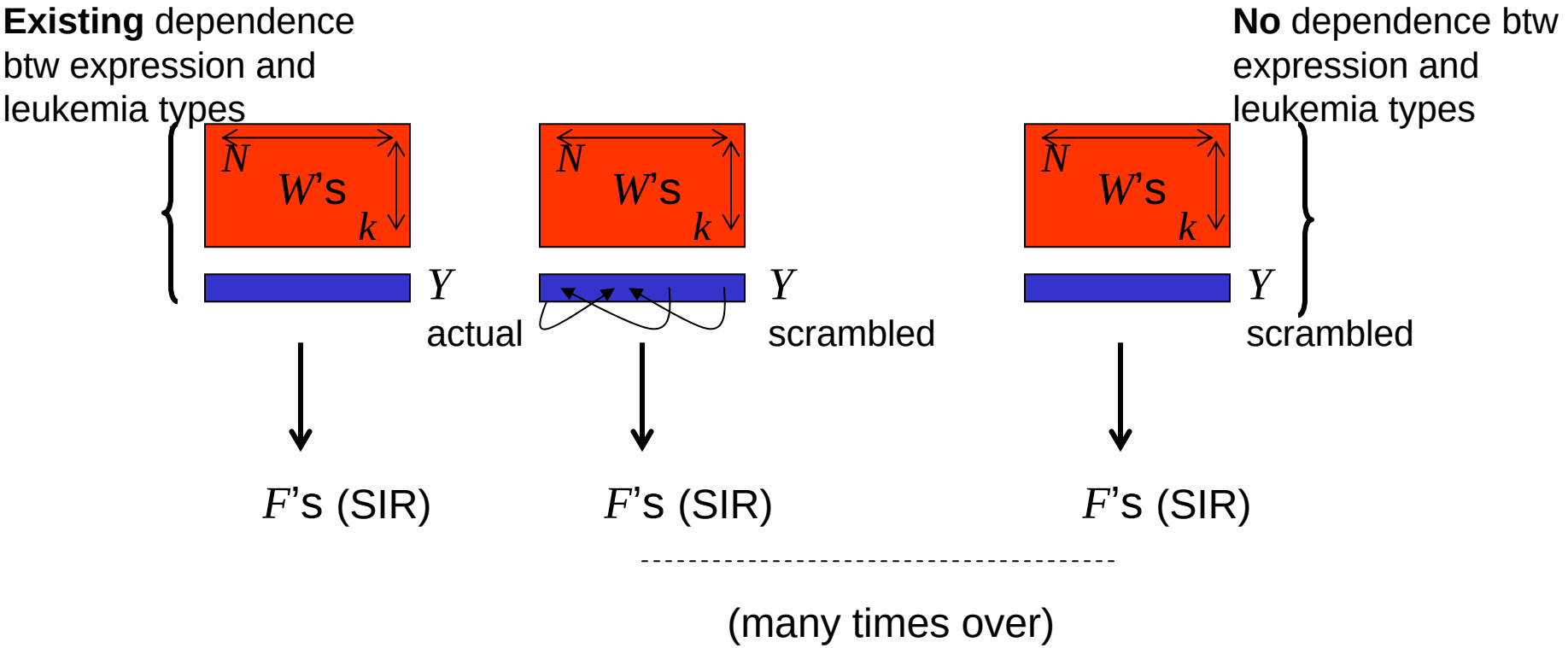
Only up to $j=N$, equal 1 after that

R^2 = Coefficient of determination, in [0,1].

Simulate a “chance scenario” by randomly permuting the samples’ Y’s, reapplying SIR in S_X and considering

- the first $F_{bi,1}$
- the first two $F_{tri,1}$, $F_{tri,1}$

(although d=0 score as relevant; no dependence)



Recomputing individual contributions, ranking, group contributions along it…

SIR on 42 PCA directions

(200 permutations)

Y binary

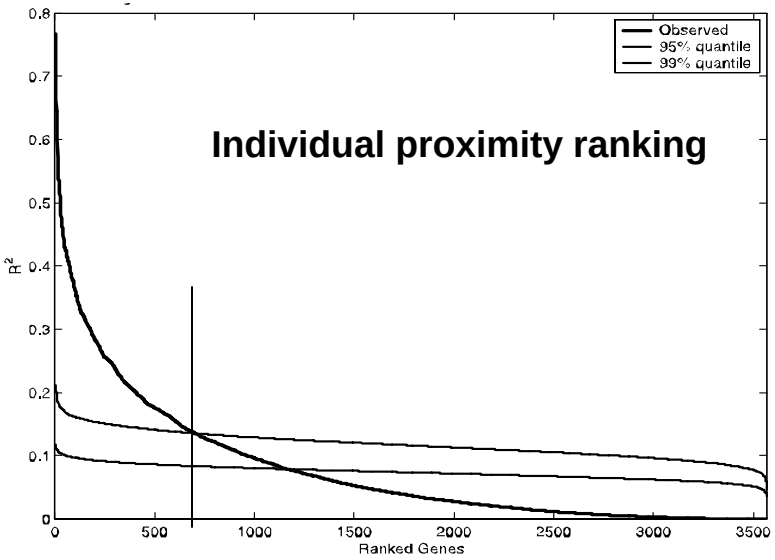
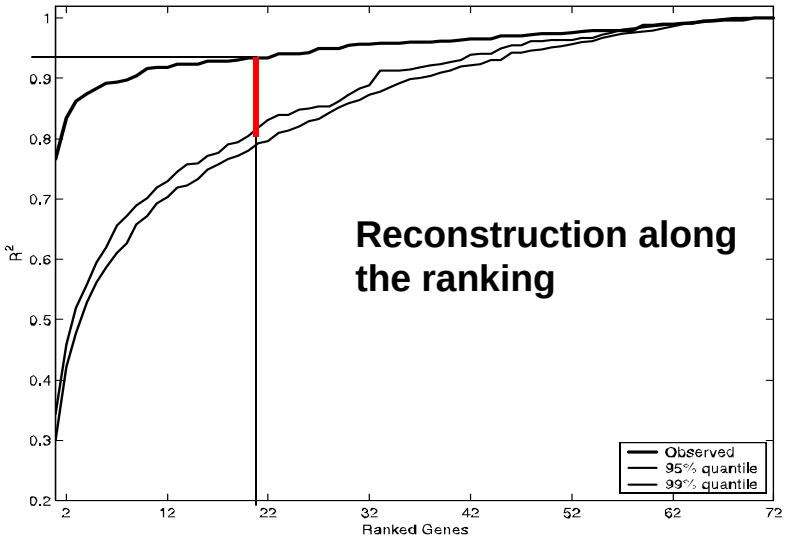


Figure 8: R^2_{sg} of the top genes along the ranking for Binary Response, with Quantile Curves (42 SVD directions)



Y trinary

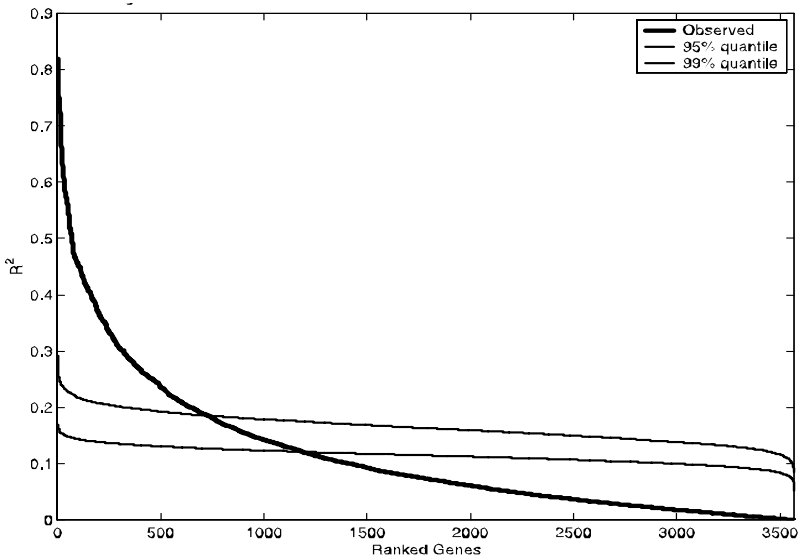
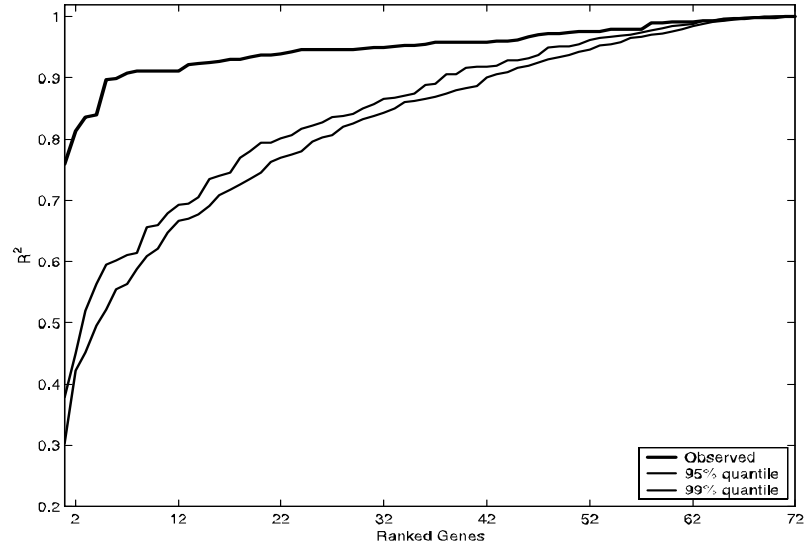
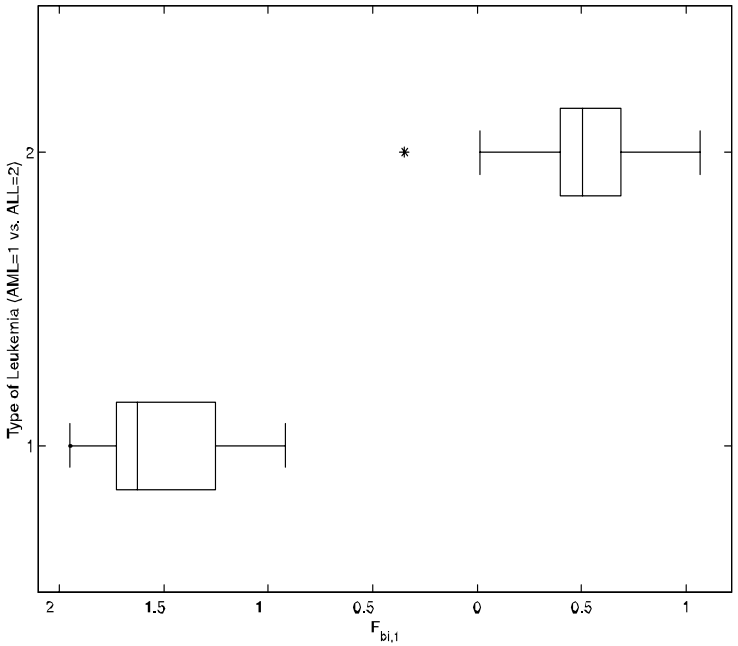


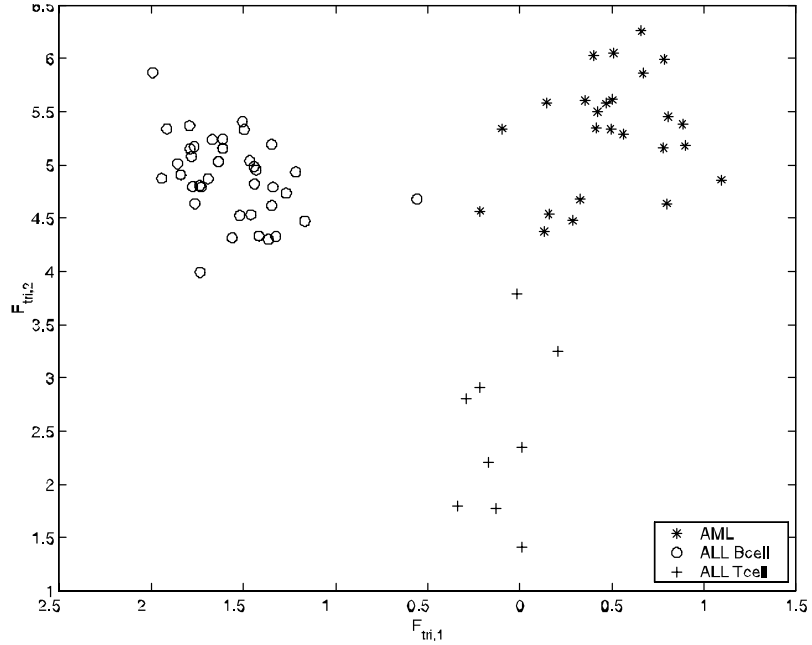
Figure 9: R^2_{sg} of the top genes along the ranking for Trinary Response, with Quantile Curves (42 SVD directions)



SIR on top 20 genes, Y binary



SIR on top 20 genes, Y trinary



- Not the same 20 genes; $F_{bi,1}$ not contained in *Span*($F_{tri,1}$, $F_{tri,2}$)
- SIR on top ranking genes detects one relevant direction for Y_{bi}, and two relevant directions for Y_{tri}.
- Increasing the number of top-ranking genes employed sharpens the separation… but are the additional genes really relevant?

Problem with this approach: there is no guarantee that $S_{Y|X} \subseteq S_X$

We need to restrict ourselves to a space S_X in which the data in non-singular… Σ_X is not invertible, while Σ_W is invertible.

Could we reach S_X without having to invert Σ_X ?

For continuous Y, and if we assume d =1, PARTIAL LEAST SQUARES

$$R_u = (\nu, \Sigma_X \nu \ldots \Sigma_X^{u-1} \nu) \quad , \quad \nu = Cov(X, Y)$$ **"seed"**

$$\beta = \Sigma_X^{-1} \nu$$

$$\beta_u = P_{R_u}^{(\Sigma)} \beta = R_u (R_u' \Sigma R_u)^{-1} R_u' \nu$$

OLS vector: reach it through a sequence of matrices that requires powers, not inversion, of Σ_X

$$\exists u^* : \quad \beta_u = \beta \quad u \geq u^*$$

The same approach works for all sorts of seeds!

(Cook, Li and Chiaromonte 2006).

Once the supervised dimension reduction has been achieved, one can apply a classifier or specify a regression model in terms of the Fs.

Useful references:

- LDA: multivariate analysis text books
- SIR:
  Li, K.C. (1991). Sliced inverse regression for dimension reduction *Journal of the American Statistical Association* **86**, 316–342.
  Cook R.D. and Weisberg S. *Applied Regression Including Computing and Graphics*. Wiley NY.
- Chiaromonte F., Martinelli J.A. (2002) Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, 176 (1), 123-144.
- Cook R.D., Li B. and Chiaromonte F. (2006) A method for sufficient dimension reduction in large-*p*-small-*n* regressions. (submitted)