Working with a response:

Beside expression data for N genes on T "conditions" (the X's), we have one or more other quantities recorded on conditions, say Y's.

Thus the data looks like

$$\begin{pmatrix} X_{11} & \cdots & X_{1T} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{NT} \end{pmatrix}$$

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1T} \\ \vdots & \ddots & \vdots \\ Y_{K1} & \cdots & Y_{KT} \end{pmatrix}$$

In our previous discussion, we thought of:

- N observations (the genes' expression levels), for each of
- T variables (the "conditions")

Now, the perspective is reversed. We think of:

- T observations (the "conditions"; usually cell samples from different individuals/organisms), for each of
- N explanatory variables (the gene's expression levels), plus
- K response variables

This is a **regression setting** : we try to explain/predict a response

- *through* (focus on creating a good prediction mechanism), and
- *in reference to* (focus on evaluating and understanding the roles of the variables; which ones are important? How?

a collection of explanatory variables; the genes' expression levels.

The most common responses are **categorical** variables expressing a classification (binary or multiple) of the samples – e.g. classification of a given cancer.

Note: Many approaches to classification exist, but here we focus on this regression-oriented logic.

Continuous responses are possible, too. For example:

- Survival time
- Continuous measurements of response to a treatment
- Continuous phenotypic trait

What distinguishes our setting from standard regression settings is that the # of explanatory variables is extremely high, and orders of magnitude higher than the # of observations: N > > T.

We have a **"variable under-resolution" problem** that makes standard regression techniques ill-posed.

Regression setting (somewhat unorthodox indexing!)

Y *response*

$X_1 \ldots X_N$ collection of explanatory variables or *predictors*

have T independent observations from $(Y, X_1 \ldots X_N)$ … or T fixed levels of $X_1 \ldots X_N$ and independent observations of Y on each of them.

Within the context of **(generalized) linear models**, one focuses the dependence analysis on $\mathrm{E}(Y| X_1 \ldots X_N)$ and introduces:

1. A *distributional assumption*, where the distribution is completely specified by the mean

   $Y| X_1 \ldots X_N \sim \mathrm{Dist}(\mathrm{E}(Y| X_1 \ldots X_N))$

2. A *link function*, connecting the mean and a linear expression in a certain number of functions of the predictors; *terms*

$$E(Y \mid X_1 \ldots X_N) = g(\gamma' F) \quad , \quad \gamma' F = g^{-1}(E(Y \mid X_1 \ldots X_N))$$

$$F = \begin{pmatrix} F_1 \\ \vdots \\ F_p \end{pmatrix} = \begin{pmatrix} f_1(X_1 \ldots X_N) \\ \vdots \\ f_p(X_1 \ldots X_N) \end{pmatrix} \quad , \quad \gamma \in R^p$$

g invertible (and differentiable, for ML purposes)
one of the f's is a const., corresponding parameter is the *intercept*
other parameters express slopes of the *terms*.

Example 1: *Y* continuous

distribution, normal: $Y|\, X_1 \ldots X_N \sim N(\, \mathrm{E}(Y|\, X_1 \ldots X_N),\, \sigma^2)$

possibly σ^2 = fct of $\mathrm{E}(Y|\, X_1 \ldots X_N)$

link, identity: $E(Y \mid X_1 ... X_N) = \gamma' F$

Example 2: *Y* binary

Distribution, Bernoulli: $Y|\, X_1 \ldots X_N \sim Ber(\, \mathrm{E}(Y|\, X_1 \ldots X_N)\,)$

Recall $\mathrm{E}(Y|\, X_1 \ldots X_N) = \Pr(Y=1)$
$\mathrm{var}(Y|\, X_1 \ldots X_N) = \mathrm{E}(Y|\, X_1 \ldots X_N)\,(\, 1 - \mathrm{E}(Y|\, X_1 \ldots X_N)\,)$

link 2.a, logit:

$$E(Y \mid X_1 ... X_N) = \frac{\exp(\gamma' F)}{1 + \exp(\gamma' F)}, \quad \gamma' F = \ln \frac{E(Y \mid X_1 ... X_N)}{1 - E(Y \mid X_1 ... X_N)}$$

link 2.b, probit:

$$E(Y \mid X_1 ... X_N) = \int_{-\infty}^{\gamma' F} \varphi_{N(0,1)}(y) dy, \quad \gamma' F = quantile_{N(0,1)}(E(Y \mid X_1 ... X_N))$$

Parameter estimates give sign and size of each term's contribution, in the context of all other terms in the model.

Also, can do inference (evaluate standard errors, produce confidence intervals for parameters and prediction intervals for the response, test for single parameters being = 0, or for groups of them being = 0 – comparison of nested models)

For example 1, ML equivalent to least squares (or appropriately weighted least squares). For example 2, ML.

ML allows Bayesian extension, putting priors on the parameter vector.

But in order for estimation and inference to be well posed, we need the F-data to span p dimensions or more.

So the issues are:

How to derive the terms to be used, dimension reduction:
variable selection among original predictors
selection of linear combinations of the original predictors

Model building, forming terms:
manageable when dealing with a very limited number of variables

Going back, from term parameter estimates and inferences to the original predictors: which ones matter and how (read: *selecting genes*).

Traditional approach:

Dimension reduction: variable selection in the context of a starting model including all original predictors without transformation

$$E(Y \mid X_1 ... X_N) = g(\beta' X)$$

Some of the slopes may be negligible because

- The *X*'s are linearly related, so they "duplicate" a certain amount of information on one another, and thus on Y, and/or
- Some *X*'s are irrelevant for the response under the chosen model

… go down to a few *X*'s, and build terms on them.

e.g. $X_1 \ldots X_N \longrightarrow X_1, X_2$

$$\downarrow$$

$$F_1 = X_1,\ F_2 = X_1^2,\ F_3 = X_1 X_2$$

Update: $E(Y \mid X_1 ... X_N) = g(\gamma' F)$

Linear dependencies among the *X*'s complicate variable selection

- Order matters when trying to apply sequential methods
- Loosely speaking, the reliability of the slope estimates is lower the stronger the linear dependencies
- In the extreme case, if some *X*'s are exact linear functions of other *X*'s, the slope estimates are not uniquely defined

The latter is exactly the case we face… we cannot just follow this approach.

We have a very bad case of unavoidable *multicollinearity* due to the fact that $N >> T$.

"Dimension-redundancy" i.e "variable under-resolution" problem.

We have just T points in the N-dimensional space of the X's.

No matter how high N is, these points will live at most in a
$(T+1)$-dimensional subspace
T-dimensional affine subspace

So our information is only at most $(T+1)$-dimensional, or at most T-dimensional if centering doesn't matter.

We have to eliminate this problem before we can do anything else.

Stage 1: X-based dimension reduction

We need to eliminate dimension redundancy, and reduce the data to what it "really spans" … or less.

Singular value decomposition of X
Extraction of principal components (idem, modulo centering of the data)

$N_o \leq T$ linear combinations, new variables, "building block" patterns.

N_o can be the exact rank, but also the "significant" rank, if we decide to neglect directions corresponding to small but non-zero eigenvalues.

We now pass to the new data matrix (N_o x T)

$$W = \begin{pmatrix} W_1' \\ \vdots \\ W_{N_o}' \end{pmatrix} = VX \qquad N_o \times T$$

$$V;\ N_o \times N \quad , \quad X;\ N \times T$$

(the row of Y's at the bottom stays the same)

At this point, we could apply in principle standard methods for **variable selection**. But N_o may still be quite large; i.e. not enough smaller than T to confidently use these methods. We may thus decide to go another way, and **select linear combinations** of the W's, instead of a subset of them.

Stage 2: dimension reduction in relation to *Y*

Suppose Y is categorical in $h=1\ldots H$ categories.

If Y is continuous, can always discretize it through ***slicing***.

Let's partition the data, and the covariance structure of the W's, according to Y (tildae indicate column vectors of the W matrix)

$$\tilde{\bar{W}}_h = \frac{1}{T_h}\sum_{t:h}\tilde{W}_t \quad , \quad \tilde{S}_{Wh} = \frac{1}{T_h}\sum_{t:h}(\tilde{W}_t - \tilde{\bar{W}}_h)(\tilde{W}_t - \tilde{\bar{W}}_h)' \quad N_o \times N_o \quad h = 1..H$$

$$M = \frac{1}{T}\sum_{h=1}^{H} T_h(\tilde{\bar{W}}_h - \tilde{\bar{W}})(\tilde{\bar{W}}_h - \tilde{\bar{W}})' \quad N_o \times N_o$$

$$E = \frac{1}{T}\sum_{h=1}^{H} T_h \tilde{S}_{Wh} \quad N_o \times N_o$$

$$\tilde{S}_W = M + E$$

Within covariances, assume invertible

Between covariance, can be singular (dependence of Y on W's, "inverted")

Decomposition

(Average) within covariance

"Standardize" the between covariance structure with the within:

Discriminant Analysis:

$$E^{-1/2}ME^{-1/2} = \Theta_D \Delta_D \Theta_D'$$

Discriminant matrix, spectral decomposition

$$\delta_{D1} \geq \ldots \geq \delta_{DN_o}$$

Some of the eigenvalues may be (approx) 0

$$\theta_{D1} \ldots \theta_{DN_o}$$

$$a_{Dj} = E^{-1/2}\theta_{Dj} \quad , \quad j = 1 \ldots N_o$$

Transform the eigenvectors through *E* to obtain the **discriminant directions**

$$\frac{a_{Dj}'Ma_{Dj}}{a_{Dj}'Ea_{Dj}} = \delta_{Dj} \quad , \quad j = 1 \ldots N_o$$

Between covariance structure along *a* standardized by within covariance structure along *a* …

$$\frac{\delta_{Dj}}{\sum_{l=1}^{N_o} \delta_{Dl}} \quad , \quad j = 1 \ldots N_o$$

Share of overall explanatory power for *Y* embodied in the discriminant matrix, that is captured by *a* .

…say one restricts attention to the first *d* discriminant directions.

This corresponds to the rank of M , if neglecting only dimensions corresponding to 0 eigenvalues, or to the "significant" rank of the standardized $E^{-1/2}ME^{-1/2}$, if neglecting also directions corresponding to small but non-zero eigenvalues.

"Standardize" the between covariance structure with the overall:

SIR (sliced inverse regression) Analysis:

$$\tilde{S}_W^{-1/2} M \tilde{S}_W^{-1/2} = \Theta_{sir} \Delta_{sir} \Theta_{sir}'$$

Sir matrix, spectral decomposition

$$\delta_{sir1} \geq \ldots \geq \delta_{sirN_o}$$

Some of the eigenvalues may be (approx) 0

$$\theta_{sir1} \ldots \theta_{sirN_o}$$

$$a_{sirj} = \tilde{S}_W^{-1/2} \theta_{sirj} \quad , \quad j = 1 \ldots N_o$$

Transform the eigenvectors to obtain **sir directions**

$$\frac{a_{sirj}' M a_{sirj}}{a_{sirj}' \tilde{S}_W a_{sirj}} = \delta_{sirj} \quad , \quad j = 1 \ldots N_o$$

Between covariance structure along *a* standardized by overall covariance structure along *a* …

$$\frac{\delta_{sirj}}{\sum_{l=1}^{N_o} \delta_{sirl}} \quad , \quad j = 1 \ldots N_o$$

Share of overall explanatory power for *Y* embodied in the sir matrix, that is captured by *a* .

…say one restricts attention to the first d sir directions.

This corresponds again to the rank of M , or to the "significant" rank of the standardized $\tilde{S}_W^{-1/2} M \tilde{S}_W^{-1/2}$.

Sir matrix:

$$\tilde{S}_W^{-1/2} M \tilde{S}_W^{-1/2} = \frac{1}{T} \sum_{h=1}^{H} T_h [\tilde{S}_W^{-1/2} (\tilde{\overline{W}}_h - \tilde{\overline{W}})][\tilde{S}_W^{-1/2} (\tilde{\overline{W}}_h - \tilde{\overline{W}})]' = \frac{1}{T} \sum_{h=1}^{H} \overline{Z}_h \overline{Z}_h'$$

Special case, H=2 (binary response)

It is easy to show that

$$M \propto (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)(\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)' \quad , \quad rank(M) = 1$$

Thus, for both the discriminant and SIR analysis only one eigenvalue can be non-zero. We have:

$$\delta_{D1} \propto (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)' E^{-1} (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$$
$$\vartheta_{D1} \propto E^{-1/2} (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$$
$$a_{D1} \propto E^{-1} (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$$

The overall explanatory power for *Y* embodied in the discriminant matrix is a multiple of the square distance between the group means, in the E^{-1} metric, and is captured by the single discriminant direction of $E^{-1}(\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$.

$$\delta_{SIR1} \propto (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)' \tilde{S}_W^{-1} (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$$
$$\vartheta_{SIR1} \propto \tilde{S}_W^{-1/2} (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$$
$$a_{SIR1} \propto \tilde{S}_W^{-1} (\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$$

The overall explanatory power for *Y* embodied in the SIR matrix is a multiple of the square distance between the group means, in the $\tilde{S}_W^{-1}$ metric, and is captured by the single SIR direction of $\tilde{S}_W^{-1}(\tilde{\overline{W}}_1 - \tilde{\overline{W}}_2)$.

Special case, "aligned" between and within covariance structures:

What happens if *M* and *E* admit the same spectral directions?

$$M = \Theta D(m)\Theta' \quad , \quad E = \Theta D(e)\Theta'$$
$$\tilde{S}_W = \Theta D(m+e)\Theta'$$
$$E^{-1/2} M E^{-1/2} = \Theta D(m/e)\Theta'$$
$$\tilde{S}_W{}^{-1/2} M \tilde{S}_W{}^{-1/2} = \Theta D(m/(m+e))\Theta'$$

Spectral directions of overall covariance, discriminant and SIR matrix are all the same.

The directions to surely neglect (corresponding to *m*'s = 0; exact rank of *M*) are the same in discriminant and SIR analysis, but eigenvalues ranking, and therefore directions to possibly neglect (corresponding to tail eigenvalues; "significant" rank of *M* after the two standardizations) may differ.

Remarks:

1. The discriminant approach is usually employed when the partitioned (by *Y*) *W*-data are all elliptical-looking, with approximately the same covariance structure (then *E* is a pooled estimate of it).

2. The SIR approach is employed when the overall *W*-data is elliptical-looking, and it has a strong theoretical foundation (*Theory of sufficient dimension reduction*, more next time).

3. Under appropriate (distributional) assumptions on *W* within *Y*-subpopulation, or on *W* overall, tests can be developed to establish the "significant" rank of *M* after the two standardizations, i.e. to determine how many of the tail eigenvalues of the discriminant or SIR matrix are approximately 0, and thus how many discriminant or SIR directions to retain. But these tests require large samples (asymptotic results for SIR, more next time).

4. CAN ALWAYS USE RANDOM PERMUTATIONS OF THE Y-LABELS TO COMPUTE EMPIRICAL P-VALUES FOR AN APPROPRIATE STATISTIC (e.g. sum of tail eigenvalues).

Another approach based on linear covariation, when *Y* is continuous:

OLS (Ordinary Least Squares); produces only one direction.

Consider the vector of covariances between *Y* and the *W*'s, and "standardize" the rank 1 cross-covariance structure it represents by the overall covariance structure of the *W*'s

$$S_{YW} = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \bar{Y})(\tilde{W}_t - \bar{\tilde{W}}) \quad N_o \times 1$$

$$S_{YW}S_{YW}' \quad rank(S_{YW}S_{YW}') = 1 \quad \rightarrow \quad \tilde{S}_W^{-1/2} S_{YW} S_{YW}' \tilde{S}_W^{-1/2}$$

$$\delta_{OLS1} \propto S_{YW}' \tilde{S}_W^{-1} S_{YW}$$

$$\vartheta_{OLS1} \propto \tilde{S}_W^{-1/2} S_{YW}$$

$$a_{OLS1} \propto \tilde{S}_W^{-1} S_{YW}$$

OLS matrix

The overall explanatory power for *Y* embodied in the OLS matrix is a multiple of the square norm of the covariance vector, in the $\tilde{S}_W^{-1}$ metric, and is captured by the single OLS direction of $\tilde{S}_W^{-1} S_{YW}$.

See also *partial least squares*, to get more than one direction.

SIR, and other sufficient dimension reduction methods work for both categorical and continuous variables.

Whatever the approach, if we retain, say d directions, we end up with the new-new data matrix:

$$F = \begin{pmatrix} F_1' \\ \vdots \\ F_d' \end{pmatrix} = AW \qquad d \times T$$

$$A;\ d \times N_o \quad , \quad W;\ N_o \times T$$

(the row of Y's at the bottom stays the same)

These are linear combinations of linear combinations of the X's.

Either the F's themselves, or possibly terms derived from them, will then be used in our prediction mechanism (regression model, or other classification algorithm).

Remarks:

1. In our microarray applications, we cannot stick with the original variables; they are simply too many for what our data actually spans. Even after X-based dimension reduction, an exact N_o may be still too large (too close to the number of observations, T) – too few degrees of freedom to apply standard procedures confidently. In any event, we have already broken direct interpretability passing to linear combinations (the W's), so we might as well reduce further, if possible.
2. Although we must eliminate dimension redundancy in the X's, for our aims what's important is not the variability structure of the X's, but the dependence structure of Y on the X's. Thus, staying with a fairly exact N_o ("conservative" X-based stage), and applying a Y-based second stage of dimension reduction, is preferable.
3. On the other hand, a small N_o may accomplish noise reduction.

(more on these issues next time)

Generalized linear model framework:

$$E(Y \mid X) \approx E(Y \mid F) = \gamma' F$$ Y continuous

$$E(Y \mid X) = \Pr(Y = 1 \mid X) \approx \Pr(Y = 1 \mid F) = g(\gamma' F)$$ Y binary, with link (d=1)

$$\hat{Y}_t = \hat{\gamma}' \tilde{F}_t \quad t = 1...T$$

$$\hat{p}_t = g(\hat{\gamma}' \tilde{F}_t) \quad \rightarrow \quad \hat{Y}_t = \begin{cases} 0 & \hat{p}_t < 0.5 \\ 1 & \hat{p}_t > 0.5 \end{cases} \quad t = 1...T$$

Fitted values/estimated prob's, after estimating the parameters

Prediction, for a new observation on which we know *X*'s, and therefore *F*'s, but not *Y*:

$$\hat{Y}^* = \hat{\gamma}' \tilde{F}$$ Y continuous

$$\hat{p}^* = g(\hat{\gamma}' \tilde{F}) \quad \rightarrow \quad \hat{Y}^* = \begin{cases} 0 & \hat{p}^* < 0.5 \\ 1 & \hat{p}^* > 0.5 \end{cases}$$ Y binary (flip a coin if exactly 0.5)

Note for the binary case we can "nuance" the confidence in the response prediction according to the estimated probability (the closer to 0 or 1 the estimated probability, the more confident we are in predicting class 0 or 1 for the new observation).

There are various ways to assess the performance of the model after its fit/estimation. Goodness of fit, diagnostics. In the case of a binary response, we can count misclassifications in the data we used to fit the model (for these we know *Y*)

$$\#(t : \hat{Y}_t \neq Y_t)$$

Cross-validation generalizes this idea, trying to work more conservatively ("against the model"; raising the bar). Split the original data, on which response is known, in two sets; use one to fit the model, and then predict response and count misclassifications on the other. Possibly, repeat on various splits.

Common ways of creating the splits:

- Split the data in two groups just once (at random, but sometimes according to a criterion)
- take out 1at a time: use T-1 observations to fit, 1 to validate, repeat T times
- take out 10% at a time: split the data in 10 groups of equal size (at random), use observations in 9 groups (90%) to fit, and observations in the remaining group (10%) to validate, repeat 10 times – of course could use a percentage other than 10%.

There is nothing wrong in principle in basing our prediction mechanism on the F's, but one of our aims (predict/explain in reference to the X's) was exactly to identify genes that are relevant for Y:

At least ex-post, we need to ***implement a variable/gene selection procedure***!

Idea: create a ***ranking*** of the variables/genes, and a reasonable criterion to decide ***how many*** of the top-ranking ones to label as relevant.

Our two-stage dimension reduction gives us a straightforward way to rank genes. Recall that the new-new variables we identify (the F's) do correspond to a d-dimensional space in the original N-dimensional one. Thus, we will consider how close each of the original X's is to this subspace, which is the one we have selected as relevant for our response.

(more on this next time)