Statistical Methods for Bioinformatics

II-3: Variable Selection

Today

- Dimension reduction and factor analysis
- Principal Component Regression
- Partial Least Squares
- Considerations in High Dimensions

High Dimensionality is associated with elevated Variance

- Learning in High Dimensionality tends to a higher “variance” component in the error. Solutions up to now:
  - Subset selection
  - Ridge/Lasso regularization.
- From linear regression you may remember that highly correlated variables have high standard errors.
- Can we not combine correlated variables into a single variable?

The problem of co-linearity

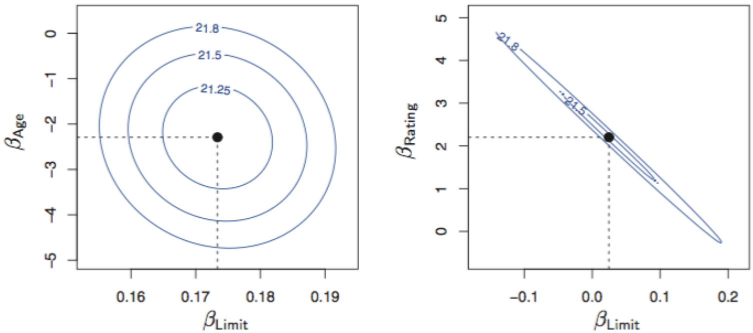


FIGURE 3.15. *Contour plots for the RSS values as a function of the parameters β for various regressions involving the* `Credit` *data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS.* Left: *A contour plot of RSS for the regression of* `balance` *onto* `age` *and* `limit`*. The minimum value is well defined.* Right: *A contour plot of RSS for the regression of* `balance` *onto* `rating` *and* `limit`*. Because of the collinearity, there are many pairs $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ with a similar value for RSS.*

Dimension Reduction Techniques and Factor Analysis

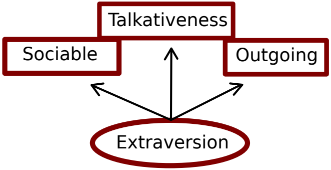
- Dimension Reduction or Factor Analysis try to describe *variability* in a dataset using a reduced set of dimensions
  - Mapping of (cor)related variables onto unobserved "factors"
- A multitude of approaches: Principal Component Analysis, Latent-variable models, Non-negative matrix factorization
- Important for many fields e.g. computer vision, text mining, psychology
  - Both exploratory and hypothesis driven analyses
- Active field of research, new methodologies under development.

Example: text mining

- Imagine there are 10000 documents about the protein p53
- You know the, say 10000, words that are used and the frequency in which they are used.
  - A sparse 10000 by 10000 matrix representing the literature about the gene
- Using Non-negative matrix factorization the word dimensionality is reduced by identifying words with similar occurrence patterns.
- The condensed variables can represent the topics discussed
  - summarize the literature
  - classify documents by topic
  - try for better document retrieval (by using the topics)
  - find genes active in similar processes

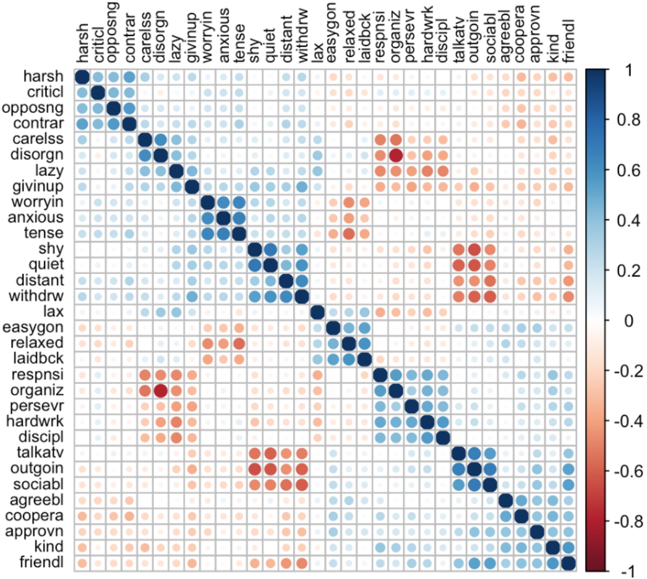
Factor Analysis in Psychology

- Factor: An underlying, but unobserved, construct or phenomena that 'shapes' or 'explains' the observed variables (at least partially).
- Used in Psychology for example to define hypothetical underlying components to explain observable traits
    - e.g. to describe personality traits



- The goal of factor analysis here is to explain the relations between variables using an underlying component
- On the following slide correlation matrix with data from Bernard Malle, figure from Psych253 Stanford online course, Statistical Theory, Models, and Statistical Methodology

Finding underlying Personality factors



Reducing Dimensionality for Statistical Learning

1. Find correlated variables and map them to a smaller set of new variables
2. Use the new variables in a regression
3. Use cross validation to find optimal number of variables.

Underlying ideas

- The dimension reduction may reduce variance component in the error
- The variability in the data is relevant for the response

Principal Component Analysis

- PCA reduces the dimensionality of a data-set of related variables, while retaining as much as possible of the variation.
- The set of variables is transformed to a new set of variables, principal components, which are uncorrelated and sorted by the variation they retain.
- The first principal component of a set of features is the normalized linear combination that has the largest variance

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

for this we work with the covariance matrix $C = {}^{X^T X}/_{n}$ with n observations of the p variable vector X_i. The optimization problem is to find $\max_\phi \phi^T C \phi$ with $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ (normalized)

Principal Component Analysis

Procedural description:

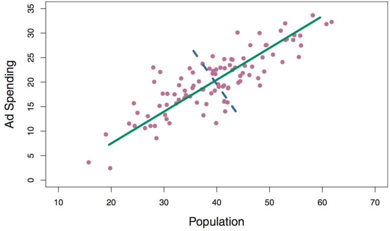
1. Find linear set of ϕ_{j1} so that:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

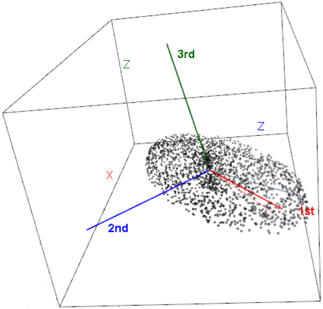
   (assuming all X centered around 0, so average of scaled X also 0, this formula represents variance)

2. Repeat till ϕ_{jp} ensuring no correlation between weighting sets

Principal Component Analysis



PCA applied to an ellipsoidically shaped point cloud

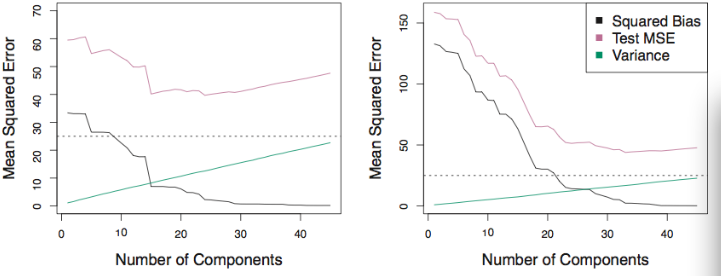


Finding the principal components

- You start with n observations of an p variable vector X_i with $X_i = \begin{bmatrix} X_{i,1} \\ \vdots \\ X_{i,p} \end{bmatrix}$
- Center all p variables around zero
- Construct a covariance matrix $C = \frac{1}{n} XX^T$. Entries in the matrix are $C_{l,k} = Cov(X_{.,l}, X_{.,k})$, which is estimated by $C_{l,k} = \frac{1}{n} \sum_i^n (x_{i,l} - \mu_l)(x_{i,k} - \mu_k)$. μ represents the mean (set to 0 here).
- Decompose C to find its eigenvectors. It can be shown that C can be decomposed as $C = VDV^T$ with V an orthonormal matrix, i.e. all columns have length 1 and are orthogonal to each other. V columns are eigenvectors of C, D is a diagonal matrix with eigenvalues.
- The eigenvectors define the mapping vectors ϕ of the principal components, the eigenvalues in D give the variance explained by every component

Principal Component Regression

After performing PCA, you choose a number of components M to make a regression. The fitted coefficients θ relate to the non reduced fit as: $\beta_j = \sum_{m=1}^{M} \theta_m \phi_{j,m}$. This puts a constraint on the coefficients. PCR is related in effect and form to ridge regression, but with a discrete form for the penalty.



Principal Component Regression

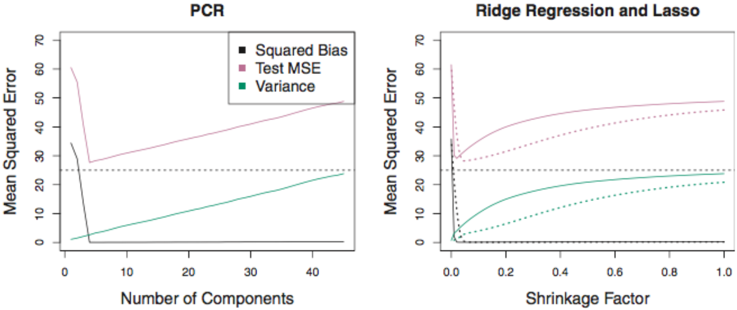


FIGURE 6.19. *PCR, ridge regression, and the lasso were applied to a simulated data set in which the first five principal components of X contain all the information about the response Y. In each panel, the irreducible error $Var(\epsilon)$ is shown as a horizontal dashed line.* Left: *Results for PCR.* Right: *Results for lasso (solid) and ridge regression (dotted). The x-axis displays the shrinkage factor of the coefficient estimates, defined as the ℓ_2 norm of the shrunken coefficient estimates divided by the ℓ_2 norm of the least squares estimate.*

Principal Component Regression: Considerations

Resume

Predictors mapped through a linear transformation to a reduced set of predictors:$Z_m = \sum_{j=1}^{p} \phi_{j,m} X_j$
Normal regression with this smaller set of predictors.
The fitted coefficients relate to the non reduced fit as
$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{j,m}$

1. PCR works best if the PCA transformation captures most of the variance in few dimensions AND this variance is associated with the response
2. It is a linear mapping approach, so strong non-linear relations will not be captured well.
3. Because PCA combines variables, the scale of each variable influences the outcome. If not directly comparable, standardize the variables.
4. PCA works best on normally distributed variables, strong departures will make PCA fail

The connection between Ridge and PCR

- If we take a $N \times p$ predictor matrix X , with zero-centered variables, we can apply the Singular Value Decomposition (SVD)

$$X = USV^T$$

U and V are $N \times p$ and $p \times p$ orthogonal matrices, S is a diagonal matrix with value s_i. Then the least squares fit for response y is:

$$\hat{y} = UU^T y$$

Ridge is given by:

$$\hat{y} = Udiag \left\{ \frac{s_i^2}{s_i^2 + \lambda} \right\} U^T y$$

PCR by:

$$\hat{y} = Udiag \left\{ 1, \ldots, 1, 0, \ldots, 0 \right\} U^T y$$

The connection between Ridge and PCR

$$\hat{y} = Udiag\left\{\frac{s_i^2}{s_i^2 + \lambda}\right\} U^T y$$

What is s? Remember the PCA: formulation $C = \frac{1}{n-1} X^T X = VDV^T$. Together with $X = USV^T$ we can derive that $D = \frac{S^2}{n-1}$. Hence the singular values s are related to the eigenvalues of the covariance matrix as follows:

$$d_i = \frac{s_i^2}{n-1}$$

Partial Least Squares Regression

- PLSR is another linear dimension reduction technique that fulfills

$$Z_m = \sum_{j=1}^{p} \phi_{j,m} X_j$$

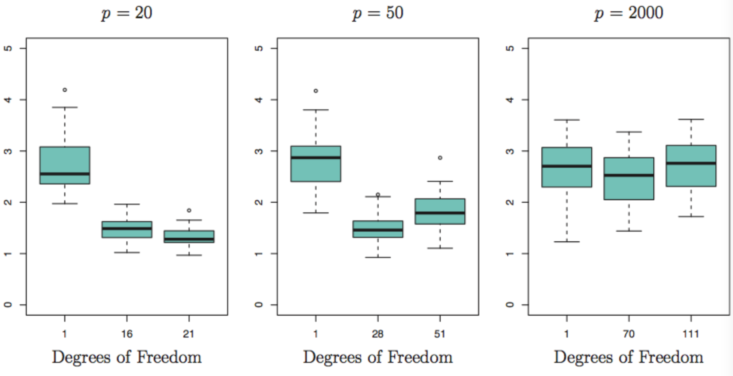
- It differs from PCR in that not just structure in the explanatory variables is captured, but also the relation between the explanatory variables and the response variables.
- The decomposition is such that most variation in Y is extracted and explained by a latent structure of X
- It works with a response *matrix*
- Resulting $Z_1 \ldots Z_m$ used with least squares to fit a linear model
- *vs PCR& Ridge:* Can reduce bias but increase variance

High dimensionality

- When p>n, a situation frequently encountered in modern science
- Least squares regression not appropriate (no remaining degrees of freedom)
- Large danger of over-fitting
- C_p, AIC and BIC are not appropriate (estimating error variance not possible)
- PLSR, PCR, forward step-wise regression, ridge and lasso are appropriate

Regressions in High Dimensions

The lasso with n = 100 observations and varying features (p). 20 features were associated with the response. Plots show test MSEs that result over the tuning parameter λ (degrees of freedom reported). Test MSE goes up with more features. This is related to the "curse of dimensionality".



Interpretation of regression in high dimensionality

- In high dimensional data sets many variables are highly co-linear
  - Selected variables may not be the unique or even best set of predictors for found prediction performance
- So even when we have a predictive model, we should not overstate the results
  - The found model is not unique, **one of many possible models**

The problem of co-linearity

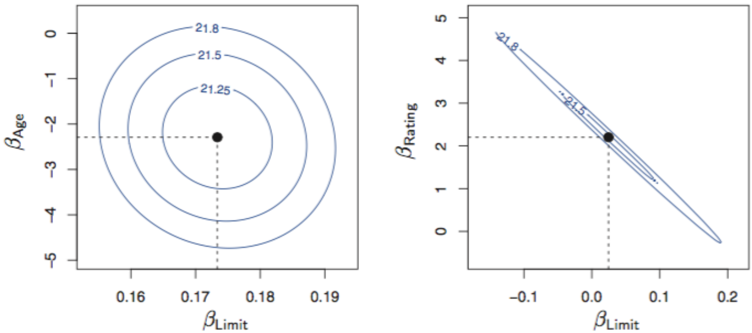
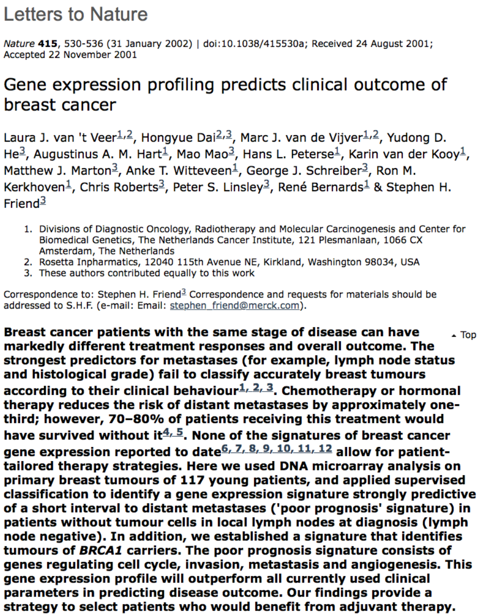


FIGURE 3.15. *Contour plots for the RSS values as a function of the parameters β for various regressions involving the* `Credit` *data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS.* Left: *A contour plot of RSS for the regression of* `balance` *onto* `age` *and* `limit`*. The minimum value is well defined.* Right: *A contour plot of RSS for the regression of* `balance` *onto* `rating` *and* `limit`*. Because of the collinearity, there are many pairs $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ with a similar value for RSS.*

Co-linearity is a motivation for regularization

- Even a small λ will stabilize coefficient estimates in ridge regression, also when p<<n
- When you have many co-linear variables, ridge and PCR will use them all in a sensible way.
  - You might want "group" selection: select the predictive set of correlated variables
- Lasso will tend to do feature selection and select the variable strongest related to the response
  - Perhaps arbitrarily.
  - Can be less robust.

High dimensionality in practice



Letters to Nature

Nature **415**, 530-536 (31 January 2002) | doi:10.1038/415530a; Received 24 August 2001; Accepted 22 November 2001

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer[1,2], Hongyue Dai[2,3], Marc J. van de Vijver[1,2], Yudong D. He[3], Augustinus A. M. Hart[1], Mao Mao[3], Hans L. Peterse[1], Karin van der Kooy[1], Matthew J. Marton[3], Anke T. Witteveen[1], George J. Schreiber[3], Ron M. Kerkhoven[1], Chris Roberts[3], Peter S. Linsley[3], René Bernards[1] & Stephen H. Friend[3]

1. Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands
2. Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA
3. These authors contributed equally to this work

Correspondence to: Stephen H. Friend[3] Correspondence and requests for materials should be addressed to S.H.F. (e-mail: Email: stephen_friend@merck.com).

▴ Top

Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour[1, 2, 3]. Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70–80% of patients receiving this treatment would have survived without it[4, 5]. None of the signatures of breast cancer gene expression reported to date[6, 7, 8, 9, 10, 11, 12] allow for patient-tailored therapy strategies. Here we used DNA microarray analysis on primary breast tumours of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature) in patients without tumour cells in local lymph nodes at diagnosis (lymph node negative). In addition, we established a signature that identifies tumours of *BRCA1* carriers. The poor prognosis signature consists of genes regulating cell cycle, invasion, metastasis and angiogenesis. This gene expression profile will outperform all currently used clinical parameters in predicting disease outcome. Our findings provide a strategy to select patients who would benefit from adjuvant therapy.

What you should learn from this class

1. What is dimension reduction and what is its use.
2. What is the procedure and motivation for a PCR (know what is PLSR)
3. How does PCR compare to the LASSO and ridge
4. Considerations in high dimensions

To do:

Preparation for next week

- Reading: chapter 7 through to 7.5
- Send in any questions day before class

Exercises

Finish labs of Chapter 6
and Exercise below.

Exercise

- In this exercise we will analyze the gene expression data set from Van de Vijver et al. (2002, N Engl J Med, 347). The study analyzed the gene expression in breast cancer tumors genome wide with DNA microarrays. The study compared the gene expression signature of the tumors with the presence or absence of distant metastasis (“DM” vs “NODM”). The idea was to use the gene expression signature as a clinical tool to decide if chemo- or hormone therapy would be beneficial.
- For the exercises load/install the following libraries: glmnet, with library(library) and install.packages(“library”).

Exercise

1. Load the file “VIJVER.Rdata”.
   Explore the dataset. How many variables? What do they represent? How many samples? What do these samples represent?
2. What challenges do you foresee in using gene expression for the stated goal (predict distant metastases).
3. For a couple of genes evaluate association with the phenotype. Do you see proof for some predictive potential? Test your intuition with a formal statistical test.
4. Demonstrate if co-linearity occurs between genes in this dataset. Do you think this represents a challenge in the analysis?
5. Use lasso, ridge and PCR methodology and make a predictor based on the gene expression values. How many genes are used for an optimal predictor? Evaluate the performance of the predictors, and comment on what you find.

Pointers: For lasso/ridge use library(”glmnet”), in the glmnet functions alpha=1