Korbinian Strimmer
Konstantinos Fokianos

Seminar Summer 2005:
Literature List

„Strategies for Analyzing Highdimensional Data“

Data with a large number of variables (p) and a relatively small number of samples (n) abound in many of today's large-scale modeling problems. This requires one to reconsider standard strategies in order to develop new and more appropriate tools.

Many of the methods discussed here were developed in application to high-throughput technologies in genomics and medicine(e.g. microarray, spectroscopy, proteomics, flow cytometry, functional magnetic resoncance imaging etc.). However, similar problems also arise, e.g. in the analysis of massive econometric or social science data.

Out focus here lies on methods that are connected with the „empirical Bayes“ framework and with „shrinkage“- these are promising because they are not only capable to deal with the „small n, large p“ problem but also are computationally very efficient.

PAPERS

A: General theory („old“ papers on biased estimation, regularization, empirical Bayes, shrinkage etc):

Hoerl, A, E, and Kennard, R.W. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**:55-67

Hoerl, A, E, and Kennard, R.W. 1970. Ridge regression: application to nonorthogonal problems. Technometrics **12**:69-82

Efron, B. and Morris, C. 1975. Data analysis using Stein's estimator and its generalizations. JASA **70**:311-319.

Morris, C.N. 1983. Parametric empirical Bayes inference: theory and applications. JASA **78**:47-55.

B: Small sample „t“ tests:

Smyth, G.K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. SAGMB **3**:3

Wu. B. 2005 Differential gene expression detection using penalized linear regression models: the improved SAM statistic. Bioinformatics **21**:1565-1571

Cui, X., Hwang, J.T.G., Qiu, J., Blades, N.J., and Churchill G.A. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics **6**:59-75.

Lönnstedt, I. and Britton, T. 2005. Hierachical Bayes models for cDNA microarray gene expression. Biostatistics: **6**:279-291.

Yang, Y.H., Xiao, Y., and Segal, M.R. 2005. Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics **21**:1084-1083.

C: Large-scale multiple testing (local fdr and tail area FDR):

Benjamini, Y, and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. JRSS B **57**:289-300.

Storey, J. D. 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Stat. **31**:2013-2035.

Storey, J. D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. PNAS **100**:9440-9445

Efron, B. 2004. Large-scale simultaneous hypothesis testing: the choice of a null-hypothesis. JASA **99**:96-104.

Efron, B. 2005. Local false discovery rates. Preprint (for Statistical Science)

D: Covariance estimation from small samples:

Ledoit. O. and . Wolf, M. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Emp. Finance **10**:603-621

Ledoit. O. and . Wolf, M. 2004. A well conditioned estimator for large-dimensional covariance matrices. J. Multiv. Anal. **88**:365-411-

E: Large-scale graphical models:

Wong, F., Carter C.K., and Kohn, R.. 2003. Efficient estimation of covariance selection models. Biometrika **90**:809-830

Dobra, A., Jones, B., Hans, C., Nevins, J.R., and West, M. Sparse graphical models for exploring gene expression data, 2004, J. Mult. Analysis, **90:**196-212.

F: Variable/model selection/regularization:

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. JRSS B **58**:267-288.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. 2004. Least angle regression. Ann. Statist. **32**:407-499.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. JRSS B **67**:301-320.

Efron, B. 2004. The estimation of prediction error: covariance penalties and cross-validation. JASA **99**:619-663

Hastie, T, and Tibshirani, R. 2004. Efficient quadratic regularization for expression arrays. Biostatistics **5**:329-340.

G: Classification, dimension reduction etc:

There are many recent papers about „small n, large p“ classification and dimension reduction, but we ignore them in our seminar on purpose - this is dealt with in the seminar by Gerhard Tutz and Florian Leitensdorfer.

F: Time series analysis:

This is quite interesting for „small n, large p“ , but there aren't too many papers yet!