An efficient approach for large scale prediction and feature selection

Miika Ahdesmäki
joint work with Verena Zuber and Korbinian Strimmer

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
University of Leipzig

July 9, 2009
useR
The R User Conference 2009

Motivation

Much of current analysis of *omics data focuses on biomarker discovery. Examples:

- Which genes are differentially expressed?
- Which features of a proteomic spectrum can be used to distinguish between cancer and healthy tissue?

There have been very many sugestions how to conduct statistical analyses of differential expression and classification.

Some very good (and well-known) choices:

- SAM or "moderated t" for gene ranking,
- PAM algorithm for classification and prediction

Motivation II

Our starting point: analysis of a proteomics data set (study of pancreas cancers)

Properties:

- very strong / pervasive correlation pattern among features
- dimension less extreme than in gene expression data

Question 1: is univariate feature selection appropriate *if features are correlated*?
Question 2: what role do gene / feature sets play in the analysis?
Question 3: is the FDR framework suitable for assigning signficance to features ?
Question 4: are there computationally efficient procedures?

Main themes: ranking and feature selection under dependence, application to classification

I. Differential Expression and Classification

Differential Expression - Setup

- Set-Up:
  - $K = 2$ groups (e.g. patient and control)
  - Large set of $i \in 1, ..., p$ genes
  - A small set of $n_1 + n_2 = n$ measurements
- Question: Which genes are differentially expressed (show a different expression profile)?
- Goal: Ranking the p genes according to their difference between the groups
- Tools: There exists an abundance of ranking statistics mostly modifications of the ordinary Student t-statistic:

$$t_{stud}(i) = \frac{\mu_1(i) - \mu_2(i)}{\sigma(i)\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Gene Ranking Statistics

Problem: gene expression and other omics data exhibit a a rich correlation-structure:

- Between measurements
- Between genes in certain clusters

How to incorporate gene-gene correlations ?

$\longrightarrow$ we revisit LDA for an idea!

Linear Discriminant Analysis I

= a simple yet very effective approach for classification

LDA assumes that each class k has a multivariate normal distribution $f(\boldsymbol{x}|k)$ with mean $\boldsymbol{\mu}_k$ and a common covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$ with correlations $\boldsymbol{P} = (\rho_{ij})$ and variances $\boldsymbol{V} = \text{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$.

A test sample is assigned to the class that maximizes the posterior probability $\Pr(k|\boldsymbol{x})$

The discriminant score is given by $d_k(\boldsymbol{x}) = \log\{\Pr(k|\boldsymbol{x})\}$.

For $K = 2$ a simple prediction rule is given by considering $\Delta(x) = d_1(x) - d_2(x)$

$$x \text{is assigned to group} \begin{cases} k = 1 & \text{if} \quad \Delta(x) > 0 \\ k = 2 & \text{if} \quad \Delta(x) < 0 \end{cases}$$

Linear Discriminant Analysis II

Some algebra simplifies the classification rule $\Delta(x)$:

$$\Delta(x) = \quad \boldsymbol{\omega}^t \boldsymbol{\delta}(x) + \log(\tfrac{n_1}{n_2})$$

with feature weights:

$$\boldsymbol{\omega} = \boldsymbol{P}^{-1/2} \boldsymbol{V}^{-1/2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{1}$$

and distance function:

$$\boldsymbol{\delta}(x) = \boldsymbol{P}^{-1/2} \boldsymbol{V}^{-1/2} (x - \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2})$$

Note that both $\boldsymbol{\omega}$ and $\boldsymbol{\delta}(x)$ are vectors.

Linear Discriminant Analysis III

If there is no correlation ($P = I$), LDA reduces to Diagonal Discriminant Analysis (DDA), with

$$\begin{aligned} \boldsymbol{\omega}^{\text{DDA}} &= \boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ \omega(i)^{\text{DDA}} &= \frac{\mu_1(i) - \mu_2(i)}{\sigma(i)} \\ &= (\frac{1}{n_1} + \frac{1}{n_2})^{1/2}\, t_{stud}(i) \end{aligned}$$

$\longrightarrow$ the feature weights $\boldsymbol{\omega}^{\text{DDA}}$ are proportional to t-score.

can we use the weights $\boldsymbol{\omega}^{\text{LDA}}$ of LDA for feature ranking and selection in the case of correlation?

II. The Correlation Adjusted t-Score (CAT-Score)

"Felix the Cat" by Pat Sullivan (1887–1933)

The CAT-Score

We define the correlation-adjusted t-Score (cat-score):

$$\begin{aligned} \boldsymbol{\tau}^{adj} &\equiv \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2} \boldsymbol{\omega} \\ &= \boldsymbol{P}^{-1/2} \times \left\{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{V}\right\}^{-1/2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{P}^{-1/2} \boldsymbol{\tau}. \end{aligned} \tag{2}$$

The vector $\boldsymbol{\tau}$ contains the gene-wise t-scores.

Interpretation of the cat score

- Weighted mean:
  The CAT-score is a weighted sum of all t-scores.
- Decorrelation:
  The CAT-score is the standardized and decorrelated mean difference between the two groups.
- Limiting case:
  If there exists no correlation, the CAT-score reduces to the ordinary $t-$score.

The cat score measures the individual contribution of each single feature to separate the two groups, after removing the effect of all other genes (note the similarity to partial correlation).

Evaluating Gene Sets: the Grouped Cat Score

Cat scores also offer a very simple means to evaluate the total effect on group separation of a set of features.

Connection with the Hotelling's T^2 statistic:

$$T^2 = \boldsymbol{t}^T \boldsymbol{R}^{-1} \boldsymbol{t} = (\boldsymbol{t}^{\text{adj}})^T \boldsymbol{t}^{\text{adj}},$$

i.e. the T^2 statistic is the sum of the squared cat scores.

Accordingly, we define the grouped cat score for gene i:

$$\tau_i^{\text{adj,grouped}} = \text{sign}(\tau_i^{\text{adj}}) \sqrt{\sum_{g \in \text{gene set}} (\tau_g^{\text{adj}})^2}\,.$$

Note that the gene sets considered need not be disjoint.

Gene Sets - Applications

There are two main cases when it is important to consider sets of genes rather than individual genes:

1. if gene sets specified a priori, if pathways or functional units are the interest of the study, but not individual genes $\longrightarrow$ gene set enrichment analysis.
2. if genes are very highly correlated and thus provide the same information on group separation.

In case 2 gene sets are given by correlation neighborhood around each gene (e.g Tibshirani and Wassermann 2006, Läuter et al 2009).

Estimating the CAT-Score

In a large p, small n setting, we use shrinkage procedures to estimate the cat score, by plugin of shrinkage estimates of:

1. The $t-$score;
   in particular, the variance $v(i)$ as mixture between the median variance v_{median} and the empirical variance estimator $\hat{v}(i)$:

$$\hat{v}_{\text{shrink}}(i) = \quad \lambda_1 v_{median} + (1-\lambda_1)\hat{v}(i)$$

2. The correlation matrix $\boldsymbol{P}$
   as mixture between the identity matrix $\boldsymbol{I}$ and the empirical correlation estimator $\boldsymbol{R}$:

$$\boldsymbol{R}_{\text{shrink}} = \quad \lambda_2 \boldsymbol{I} + (1-\lambda_2)\boldsymbol{R}$$

Computing $(\boldsymbol{R}_{\text{shrink}})^{-1/2}$

We use the following trick:

With $\boldsymbol{Z} = \boldsymbol{R}^{\text{shrink}}/\gamma$ we rewrite $\boldsymbol{Z} = \boldsymbol{I}_p + \frac{1-\gamma}{\gamma}\boldsymbol{R} = \boldsymbol{I}_p + \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^T$, where $\boldsymbol{M}$ is a symmetric positive definite matrix of size m times m and $\boldsymbol{U}$ an orthonormal basis. m is the rank of $\boldsymbol{R}$.

$$\boldsymbol{Z}^{\alpha} = \boldsymbol{I}_p - \boldsymbol{U}(\boldsymbol{I}_m - (\boldsymbol{I}_m + \boldsymbol{M})^{\alpha})\boldsymbol{U}^T, \tag{3}$$

This requires only the computation of the α-th power of the matrix $\boldsymbol{I}_m + \boldsymbol{M}$ which is of rank m.

Note the identiy is different from the Woodbury identity.

III. Efficient Classification

Pooled centroid formulation of LDA

So far we have only considered K=2. But the cat score can also defined for $K > 2$.

Recipe: Modify LDA discriminant score by adding a class-independent constant:

- Consider the pooled mean $\boldsymbol{\mu}_{\text{pool}} = \sum_{j=1}^{K} \frac{n_j}{n} \boldsymbol{\mu}_j$ and evaluate the pooled discriminant score $d_{\text{pool}}^{\text{LDA}}(\boldsymbol{x}) = \boldsymbol{\mu}_{\text{pool}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_{\text{pool}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\text{pool}}$
- The centered score $\Delta_k^{\text{LDA}}(\boldsymbol{x}) = d_k^{\text{LDA}}(\boldsymbol{x}) - d_{\text{pool}}^{\text{LDA}}(\boldsymbol{x})$ can be interpreted as log posterior ratio and further simplifies...

Pooled centroid formulation II

- ...into $\Delta_k^{\text{LDA}}(\boldsymbol{x}) = \boldsymbol{\omega}_k^T \boldsymbol{\delta}_k(\boldsymbol{x}) + \log(\pi_k)$ where we have a feature vector
  - $\boldsymbol{\omega}_k = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\text{pool}})$
- and a vector valued distance function
  - $\boldsymbol{\delta}_k(\boldsymbol{x}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_{\text{pool}}}{2})$.
- This formulation (*vector* dot product) allows the control of variable importance through $\boldsymbol{\omega}_k$, which is not dependent on the test data $\boldsymbol{x}$

Multiclass CAT Score

- Define the vector τ_k^{adj} of "correlation-adjusted t-scores" as a scaled version of ω_k:

$$\begin{aligned} \tau_k^{adj} &\equiv \left(\frac{1}{n_k} - \frac{1}{n}\right)^{-1/2} \omega_k \\ &= \boldsymbol{P}^{-1/2} \times \left\{\left(\frac{1}{n_k} - \frac{1}{n}\right)\boldsymbol{V}\right\}^{-1/2} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\text{pool}}) \\ &= \boldsymbol{P}^{-1/2} \tau_k \ . \end{aligned} \tag{4}$$

- Thus the score is a decorrelated version of the gene-wise gene-specific t-scores between the mean of group k and the pooled mean.

Multiclass correlation adjusted t-scores (continued)

- Summary score for measuring the total impact of feature $i \in \{1, \ldots, p\}$: $S_i = \sum_{j=1}^{K} (\tau_{i,j}^{adj})^2$
- For comparison, the nearest centroid classifier (a.k.a. PAM) uses $S_i' = \max_{j=1,\ldots,K}(|\tau_{i,j}|)$
- Pros of square-sum score:
  - approximately χ^2 distributed
  - takes more than one group into account

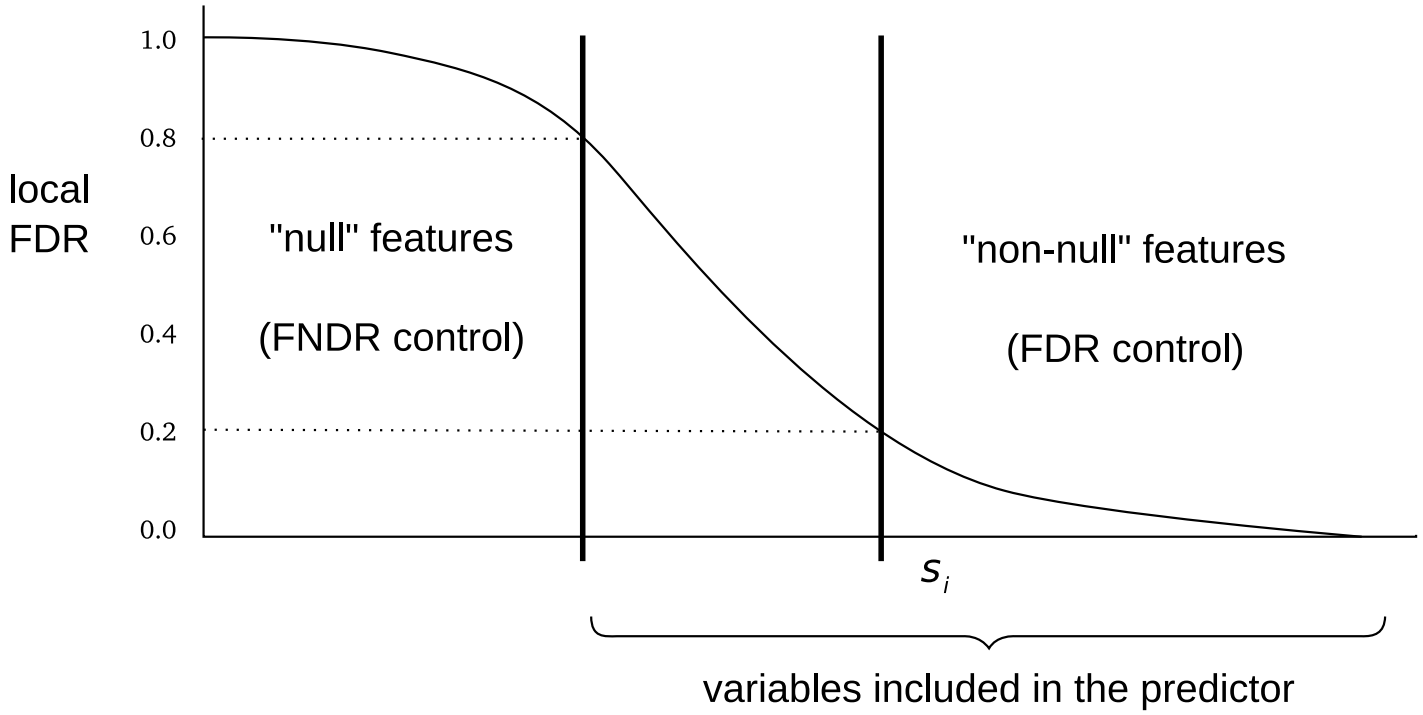
Local false non discovery rate for feature selection

- The feature-specific scores S_i are learned by plugging in the shrinkage estimators.
- Univariate thresholding is performed to select the important features.
- We advocate using the false discovery rate (FDR) framework or alternatively "Higher Criticism" to select features for classification.

Local false non discovery rate for feature selection (continued)

- When constructing classifiers the FDR approach can *not* be applied in the same way as in differential expression.
- This is because when training classifiers one aims at identifying with confidence the set of *null features* not informative about group separation.
- This is controlled by the false *non*-discovery rate.

Local false non discovery rate for feature selection (continued)



Results on real data: Singh Prostate cancer data

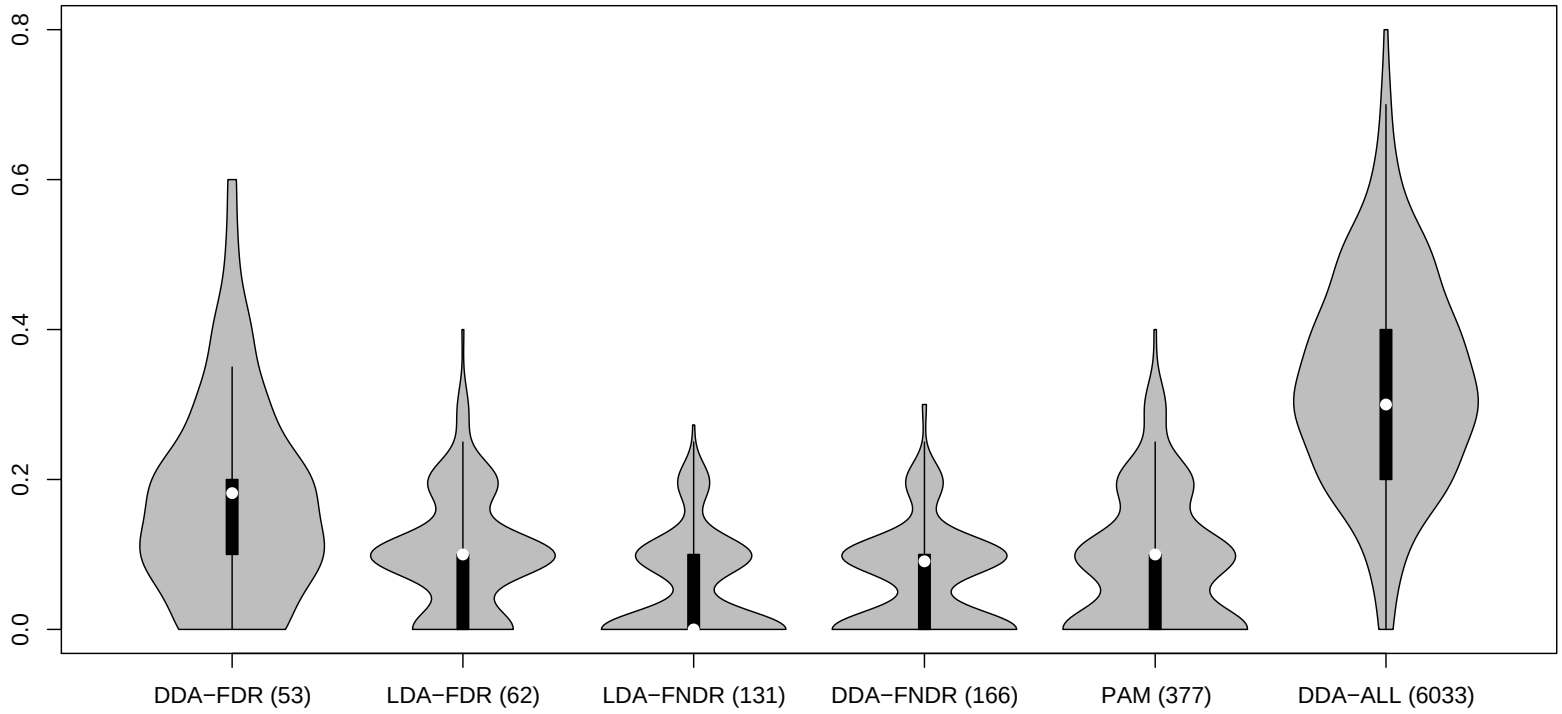
Table: Prediction errors and number of selected features for Singh *et al.* (2000) gene expression data. The number in the round brackets is the estimated standard error.

Method	Prediction Error	Features
Ebay	0.092	51
DDA-FDR	0.1682 (0.0093)	53
LDA-FDR	0.0989 (0.0056)	62
LDA-FNDR	**0.0550** (0.0048)	131
DDA-FNDR	0.0640 (0.0049)	166
PAM	0.0859 (0.0063)	172–482
DDA-ALL	0.3327 (0.0099)	6033

The prediction error of Ebay is taken from Efron (2008).

Comparison with other classifiers. Data from *Cancer Cell* 1:203–209.

Results on real data: Singh Prostate cancer data (continued)



Comparison with other classifiers. Data from *Cancer Cell* 1:203–209.

Results on real data: Lymphoma, SRBCT and Brain tumor data

Table: Estimated prediction errors for several multi-class reference data sets.

Data	Method	Prediction Error	Features	DE
Lymphoma	DDA-FNDR	0.0517 (0.0062)	162	0
($K = 3$, $n = 62$,	LDA-FNDR	**0.0036** (0.0018)	392	55
$p = 4026$)	PAM	0.0254 (0.0045)	2796–3201	
SRBCT	DDA-FNDR	0.0007 (0.0007)	90	62
($K = 4$, $n = 63$,	LDA-FNDR	**0.0000** (0.0000)	89	76
$p = 2308$)	PAM	0.0145 (0.0034)	39–87	
Brain	DDA-FNDR	0.1892 (0.0146)	33	8
($K = 5$, $n = 42$,	LDA-FNDR	**0.1525** (0.0120)	102	23
$p = 5597$)	PAM	0.1939 (0.0112)	197–5597	

The last column (DE) shows the number of differentially expressed genes, which equals the number of significant features if FDR rather than FNDR is used as criterion.

Comparison with other classifiers. Lymphoma data: Alizadeh et al. *Nature* 403:503–511. SRBCT data: Khan et al. *Nature Med.* 7:673–679. Brain tumor data: Pomeroy et al. *Nature* 415:436–442.

Higher criticism in comparison to FNDR

- Higher Criticism thresholding scores are z-scores computed from p-values.
- The rank of the highest ensuing value gives the number of important features.
- Similar performance to FNDR (result table skipped).
- NOTE: Both FNDR and HC need a fitted mixture-model (hence p must be moderately large).

IV. Conclusions

Conclusions

- the cat score is a natural univariate criterion that harmonizes ranking genes and feature selection and that takes account of correlation.
- We introduced a pooled centroid formulation of LDA (=LDA written the form of PAM).
- The formulation allows efficient feature selection without resampling.
- FNDR can be used efficiently for selecting the number of important features, (but not FDR!)
- Good performance and low computational time.

Limits: only moderately large dimensions possible, choice of optimal feature sets ambigous.

Software and further information

Software availability:

- Our R packages "st" and "sda", both available from CRAN (playing the same roles as "sam" and "pam" / "rda")

Preprints:

- V. Zuber and K. Strimmer. 2009. *Gene ranking and biomarker discovery under correlation.* `http://arxiv.org/abs/0902.0751`
- M. Ahdesmäki and K. Strimmer. 2009. *Feature selection in "omics" prediction problems using cat scores and false non-discovery rate control.* `http://arxiv.org/abs/0903.2003`

Thanks for your interest!

Any Questions (to be postponed)?

Thanks to Alexander von Humboldt Foundation for postdoc funding!