Microarray-based prediction with emphasis on clinical parameters: a PLS approach

Anne-Laure Boulesteix, Christine Porzelius, Martin Daumer

Sylvia Lawry Centre for Multiple Sclerosis Research
Munich, Germany

PLS'07 Conference
Aas, September 2007

Microarray gene expression data

The term 'microarray data' is almost always used to name 'gene expression' data, although

- Gene expression can be measured using other technologies (SAGE, RT-PCR).
- Microarray technology can be used for other purposes than gene expression measurement (array CGH, ChIP-on-chip).

Microarrays are very expensive (several hundreds dollars / array). Hence, microarray data sets *never* include more than a few hundreds of observations (arrays). But each array includes several (tens of) thousands of genes, thus yielding high dimensional data.

Microarray gene expression data (ctd.)

After normalization, one obtains a 'flat' matrix with n rows and p columns:

	Gene 1	...	...	Gene p
Array 1	.	.	.	.
...	.	.	.	.
...	.	.	.	.
Array n	.	.	.	.

where each value represents the amount of mRNA of the considered gene in the considered array and can be seen as a *gene expression level*. From a statistical point of view, genes are continuous variables.

Microarray data analysis: an overview

1. Data processing, normalization
2. Differential gene expression (univariate supervised analysis), where groups are defined by
   - different experimental conditions (stimulus/no stimulus, wildtype/knock-out)
   - different phenotypes (healthy/diseased, cancer subtypes)
   - different clinical outcomes (responder/non responder)
   - different time points during the cell-cycle
3. Microarray-based prediction (multivariate supervised analysis), where the response to be guessed may be, e.g.
   - a class ('classification' or 'class prediction')
   - a continuous variable ('regression')
   - a survival time ('survival analysis')
4. Clustering, genetic networks, applications to system biology

PLS for microarray data: an overview

- Regression (Huang et al, 2004)
- **Class prediction**
  (e.g. Nguyen and Rocke, 2002; Boulesteix, 2004)
- Survival analysis (e.g. Park et al, 2002; Li and Gui, 2004)
- Imputation of missing values (Bras and Menezes, 2006)
- Genetic networks
  - Prediction of transcription factor activities
    (Boulesteix and Strimmer, 2005)
  - **Estimation of the partial correlation matrix for the reconstruction of gene networks**

For an overview, see

- Boulesteix and Strimmer, 2007. Partial Least Squares: A versatile tool for the analysis of high dimensional genomic data. *Briefings in Bioinformatics* 8:24-32.

Class prediction

A. PLS-DA approach

B. PLS dimension reduction, then apply a standard class prediction method using the PLS components as predictors: Linear Discriminant Analysis (LDA), logistic regression, etc (Nguyen and Rocke, 2002; Boulesteix, 2004)

C. Generalizations of PLS

   c1. Marx' algorithm (1996): a generalization to Partial Least Squares of the IRLS algorithm used in GLMs. PLS is embedded into the IRLS algorithm. Problems for $n \ll p$ $\longrightarrow$ alternative methods by Ding and Gentleman (2005) and Fort and Lambert-Lacroix (2005).

   c2. Modification of the weight vectors to account for the specificity of the response. The structure of the PLS algorithm remains the same. Ex: Bastien et al (2005).

Data situation and addressed question

From now on, we consider the following data situation:

- A binary response variable Y, e.g. responder/non-responder
- Thousands of gene expression predictors $X_1, \ldots, X_p$
- A moderate number of clinical predictors $Z_1, \ldots, Z_q$, e.g. age, sex, disease duration, disease subtype, grade, etc.

Goal: Determine whether the gene expression predictors can improve the prediction accuracy yielded by the clinical parameters (which are much more easy to collect).

An example: MS and MRI (Daumer et al, 2006)

- **Multiple sclerosis**
  Multiple sclerosis is an inflammatory disease of the nervous system, which can typically take a large number of (very) different courses.
- **Magnetic Resonance Imaging (MRI)**
  MRI parameters such as the presence of enhancement or the number of newly enhancing lesions have been suggested as surrogate markers for relapses, a commonly used outcome in clinical trials.
- **Conclusion of Daumer et al (2006)**
  Evidence that if information on previous relapses is available, MRI parameters have no additional value for the prediction of future relapses.

The same for microarray data?

- Hundreds of articles on new classification methods reporting accuracy improvement (sometimes "fishing for significance").
- Most of them do not consider the *additional* improvement of classification accuracy yielded by microarray data when they are used in addition to classical parameters.
- "Adjustment for other classical predictors of the disease outcome are essential" (Ntzani and Ioannidis, 2003). However, this adjustment is usually not given much attention from a methodological point of view.

Solutions?

A. Subgroup analysis
   - Sample size problems
   - Even if the sample size is large, not all classic predictors can be taken into account.

B. Consider both genes and clinical parameters as covariates and use a classification method for high-dimensional data
   - Clinical parameters are "lost" within a huge amount of gene expression variables

C. Distinguish between mandatory predictors (clinical parameters) and optional predictors (gene expression), e.g., Tutz and Binder (2007, CSDA): "Boosting ridge regression".
   - Performs well, but computationally intensive in high dimension

Solutions?

D. Construct two distinct classification rules: one using clinical parameters only, one using gene expression data only
   - It is difficult to see whether clinical parameters do the same as gene expression data or rather explore other aspects of the classification problem.
   - One does not obtain a single classification rule combining both types of information.

E. Build a (cross-validated) score out of gene expression and use it as predictor in logistic regression (Tibshirani and Efron,2002).
   - A unique summarizing score may be too simplifying for optimal microarray-based prediction.

Proposal

- "Summarize" gene expression data via dimension reduction.
  - PCA (not recommended because unsupervised)
  - **PLS (recommended)**
- Use both clinical parameters and extracted components to build a classifier.
  - LDA (Nguyen and Rocke, 2002; Boulesteix, 2004). Problem: categorical clinical parameters.
  - Logistic regression. Problem: convergence when classes are well-separated.
  - **Random forests**.

This yields the "PLS+RF" method.

In a nutshell

Random forest with predictors, e.g.

- Sex
- Age
- Age at onset of disease
- ...

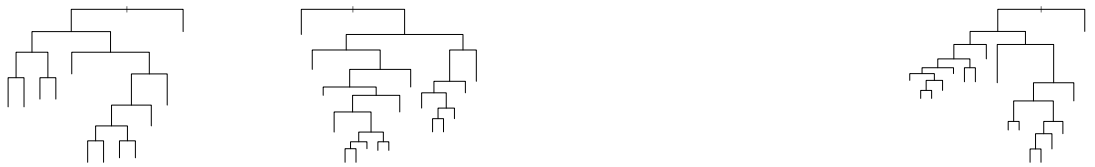
and

- 1st PLS component
- 2nd PLS component

Overview of random forests

- Suggested by Breiman (2001)
- Machine learning method based on the aggregation of (e.g., $B = 100$) classification or regression trees
- Each tree is constructed based on a different bootstrap sample of size n (usually) drawn with replacement from the original data set.
- Each split in each tree is selected from a subset of predictors (default is $\sqrt{p}$).
- Implemented in the `R` packages `randomForest` (original version) and `party` (with conditional inference trees by Hothorn et al, 2006).

Overview of random forests



$B = 1$ $B = 2$ ... ... $B = 100$

$\Rightarrow$ Robust prediction

- To avoid variable selection bias, sampling without replacement is preferable (e.g., Strobl, Boulesteix, Zeileis and Hothorn, 2007, BMC Bioinformatics).

Summary of the procedure (PLS+RFa)

- $X_1, \ldots, X_p$ = gene expression predictors
- $Z_1, \ldots, Z_q$ = clinical predictors (age, sex, etc)
- Y = binary outcome (disease type, responder status, etc)
- Suppose we have a learning data set $\mathcal{L}$ and a test data set $\mathcal{T}$.

1. Carry out variable selection based on $\mathcal{L}$ only, yielding $X_{s_1}, \ldots, X_{s_{p^*}}$.
2. Summarize $X_{s_1}, \ldots, X_{s_{p^*}}$ in form of a small number new linear components $T_1, \ldots, T_c$ (typically, $c = 2, 3$), based on $\mathcal{L}$ only.
3. Construct a random forest based on predictors $T_1, \ldots, T_c$ and $Z_1, \ldots, Z_q$.
4. Compute $T_1, \ldots, T_c$ for the test data set $\mathcal{T}$.
5. Apply this random forest to the prediction of the test observations from $\mathcal{T}$.

Other related procedures

- **Random forests without PLS**
  - Construct a random forest based on predictors $X_{s_1}, \ldots, X_{s_p*}, Z_1, \ldots, Z_q$.

- **PLS+RF, without distinguishing between clinical and gene expression data (PLS+RFb)**
  - Summarize $X_{s_1}, \ldots, X_{s_p*}, Z_1, \ldots, Z_q$ in form of a small number new linear components $T_1, \ldots, T_c$,
  - Construct a random forest based on predictors $T_1, \ldots, T_c$.

Data-driven simulation

- Different simulation settings corresponding to different prediction powers.
- Variable selection based on the Wilcoxon rank sum test. Number of retained genes: $\tilde{p} = 1000$ (similar results are obtained with other values of $\tilde{p}$).
- The error rate is estimated via LOOCV. Monte-Carlo cross-validation (random splitting into learning and test data set) with 5 observations in the test data set yields similar results.
- In each setting, the whole procedure is repeated 10 times. Averaged results are given.

Results for A vs B comparison (good prediction)

With 1000 genes

		normal Z_1, Z_2	binary Z_1, Z_2	$Z_1, \ldots, Z_{10}$
PLS+RFa	**X**, **Z**	0.11/0.03/0	0.10/0.08/0	0.13/0.01/0
PLS+RFb	**X**, **Z**	0.13/0.05/0.03	0.11/0.10/0.09	0.14/0.10/0.04
RF	**X**, **Z**	0.16/0.11/0.08	0.15/0.14/0.08	0.15/0.14/0.02
PLS+RF	**X**	0.10	0.10	0.10
RF	**X**	0.16	0.16	0.14
RF	**Z**	0.51/0.03/0	0.44/0.12/0	0.48/0.01/0

Results for B vs C comparison (poor prediction)

With 1000 genes

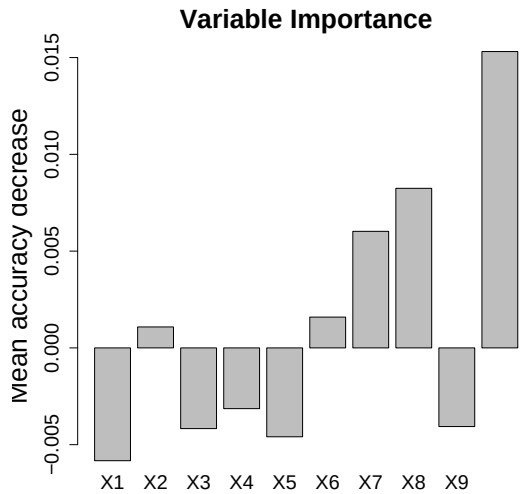
		normal Z_1, Z_2	binary Z_1, Z_2	$Z_1, \ldots, Z_{10}$
PLS+RFa	**X**, **Z**	0.40/0.04/0	0.39/0.10/0	0.40/0.01/0
PLS+RFb	**X**, **Z**	0.40/0.08/0.04	0.39/0.34/0.28	0.40/0.24/0.06
RF	**X**, **Z**	0.35/0.26/0.23	0.36/0.30/0.23	0.34/0.27/0.01
PLS+RF	**X**	0.38	0.38	0.34
RF	**X**	0.34	0.32	0.34
RF	**Z**	0.44/0.03/0	0.36/0.08/0	0.41/0.01/0

Alternatives to random forests

- logistic regression: problem with well-separated data
- logistic regression with Firth's correction: problem with highly collinear data
- penalized logistic regression: choice of λ?
- linear discriminant analysis: problem with categorical data
- all methods: since the PLS components separate the classes very well (overfitting), they may be artificially preferred by the applied classifier. With random forests, this problem is partly avoided by randomly eliminating predictors from the set of the candidate predictors at each split.

Outlook: model selection

- Model selection based on, e.g., unbiased variable importance measures (e.g., Strobl, Boulesteix, Zeileis and Hothorn, 2007) could be applied to eliminate irrelevant clinical parameters.
- It is then important to build the forest based on unbiased trees (see Hothorn et al, 2006).



Conclusions PLS+RF

- Clinical parameters should not be put together with gene expression variables!
- The PLS+RF method allows direct evaluation of the accuracy improvement induced by gene expression data.
- PLS+RF performs as well as the "best" methods (PAM, SVM, etc) in standard settings.
- The procedure may be generalized to other outcomes (e.g., survival analysis).

Correlation and partial correlation

- Joint work with Nicole Krämer and Juliane Schäfer

Graphical gaussian models (GGMs) are based on the partial correlation matrix. If

$$\mathbf{\Omega} = \mathbf{C}^{-1} = (\omega_{jk})$$

is the inverse of the Bravais-Pearson correlation matrix, the partial correlation matrix $\mathbf{\Pi} = (\pi_{jk})$ is obtained by

$$\pi_{jk} = -\omega_{jk} / \sqrt{\omega_{jj}\omega_{kk}}$$

Problem: When $n \ll p$, $\mathbf{C}$ cannot be inverted!

Correlation and partial correlation

The partial correlation of X_j and X_k given $Z_1, \ldots, Z_m$ (here all variables except X_j and X_k) can also be computed based on the linear regression model

$$x_{ij} = \beta_{kj} x_{ik} + (z_{i1}, \ldots, z_{im})\boldsymbol{\beta}_j + \epsilon_{ij} \quad (1)$$

$$x_{ik} = \beta_{jk} x_{ij} + (z_{i1}, \ldots, z_{im})\boldsymbol{\beta}_k + \epsilon_{ik} \quad (2)$$

as

$$\pi_{jk} = sign(\hat{\beta}_{kj})\sqrt{\hat{\beta}_{kj}\hat{\beta}_{jk}}$$

Problem: OLS can not be applied in the case $n \ll p$!
Proposal: Use PLS instead of OLS.

I thank my co-authors:

- Christine Porzelius
- Martin Daumer
- Nicole Krämer
- Juliane Schäfer

Thanks for your attention!
Any questions?