

ZB MED - Informationszentrum Lebenswissenschaften

12th Day of the Doctoral Student

Seyfert, Hans-Martin

Dummerstorf, 2011

Pooled samples for classification - What about PPLS-DA? Anna Telaar,
Kristian Hovde Liland, Dirk Reipsilber and Gerd Nürnberg

urn:nbn:de:hbz:38m:1-30093

Pooled samples for classification – What about PPLS-DA?

Anna Telaar¹, Kristian Hovde Liland², Dirk Reipsilber¹ and Gerd Nürnberg¹

¹Leibniz Institute for Farm Animal Biology, Dummerstorf, ²Norwegian University of Life Sciences, Ås

1 Introduction

Sample pooling is necessary for microarray experiments if not enough amount of RNA exists for the hybridization step or because of financial reasons. For classification studies like biomarker search, pooling should be avoided, because data is required at the "individual level" (Kerr 2003). Nevertheless pooled samples are used for biomarker discovery (Searfoss et al. 2003). Therefore Telaar et al. (2010) initiated a simulation study to analyse the dependence of the pooling effect on the structure of informative features, pool size and statistical learning method used. As measures of the pooling effect we use the prediction error and the number of important features simulated as informative for the classification. We investigate the statistical learning method power partial least squares discriminant analysis (PPLS-DA). PPLS-DA is a specialized version of the method partial least squares discriminant analysis (PLS-DA) which is a well known and powerful approach. We present a further optimization of the parameter choice of PPLS-DA for the single sample design and the pooling designs.

2 Materials and methods

Pooling: The gene expression of sample i for gene g , (with $g = 1, \dots, G$ and G depicting the number of genes) can be modelled on the log scale as $\log_2(u_{g,i}) = \mu_g + \varepsilon_{b_i} + \varepsilon_{t_i}$, $i = 1, \dots, n$ (with n number of samples) with mean gene expression level μ_g , biological and technical variation ε_{b_i} and ε_{t_i} which are independently, identically and normally distributed with mean zero and variances σ_b^2 and σ_t^2 . Pooling, the combination of several RNA samples, takes place before measuring the expression of the genes of these RNA samples. The gene expression of a pool p_j , $1 \leq j \leq n_p$ (with n_p number of pools), is the average of the samples which form the pool: $u_{g,p_j} = \log_2\left(\sum_{i=l+1}^{m+l} \frac{1}{m} 2^{(\mu_g + \varepsilon_{b_i})}\right) + \varepsilon_{t_j}$ where $1 \leq m \leq n$ is the number of mixture components within a pool, $l = s \cdot m$, $s \in \{0, \dots, n_p - 1\}$ is the starting index of a sample. The size of a pool is determined by m .

Design: Studying a classification task with two classes, our pools are built inside the classes.

Gene expression data simulation: Comparing a two-class classification study with a single sample design and pooling designs with pool size $m = 2, 3, 5$, we simulate 60 single samples per class and 1000 genes partitioned in an informative part and non-informative part for the classification. For the test set, 30 single samples per class are randomly chosen, which are used to calculate the prediction error (PE). The remaining 30 single samples per class are the basis for the different training sets.

The non-informative part of the data matrix which shows no differences between the two classes, consists of normally distributed random variables. Moreover, noise is added as technical variance in different proportions of the biological variance: $\{\frac{1}{8}\sigma_b^2, \frac{1}{4}\sigma_b^2, \frac{1}{2}\sigma_b^2, \frac{3}{4}\sigma_b^2\}$. We simulate 10 differentially expressed genes with a mean class difference randomly chosen between $[0.1, 0.5]$, these we name informative simulated features (ISF). We make 500 repetitions of the classification task.

PPLS-DA: For the classification we investigate the method PPLS-DA. We implement an inner 10-fold cross-validation to choose the optimal number of PPLS-DA components, with these components we apply a linear discriminant analysis. In contrast to the usual PLS-DA an optimization step is done during the calculation of the loading weights. For the gene expression data matrix $\mathbf{X} \in \mathbb{R}^{(30 \times G)}$ and the dummy matrix $\mathbf{D} \in \{0, 1\}^{(30 \times 2)}$ (indicating the class-membership of the classes $t = 1, 2$) the transformation matrix $\mathbf{t}_0(\gamma)$, $\gamma \in U$ can be factorized into: $\mathbf{t}_0(\gamma) = \mathbf{D}(\gamma) \mathbf{D}(\gamma)$.

Here $\mathbf{D}(\gamma)$ is a diagonal matrix with the standard deviation of the columns of $\mathbf{X}$ as non-zero entries exponentiated by $\frac{1-\gamma}{\gamma}$, $\mathbf{D}(\gamma)$ is a matrix which consists of the correlation of with exponentiated by $\frac{\gamma}{1-\gamma}$, and $\mathbf{D} = \text{diag}(\{\rho_t \mid t = 1, 2\})$, where ρ_t depends on the prior probability and relative group frequency of group t . Then the correlation between the transformed matrix $\mathbf{X} \mathbf{t}_0(\gamma)$ and $\mathbf{D}$ is maximized by using canonical correlation analysis. The power parameter γ is optimized normally over the interval $U = (0, 1)$, the endpoints 1 and 0 are checked explicitly. Considering the usual PLS-DA method, standard deviation and correlation part are coequal (same exponent 1) for the calculation of the loading weights. Further details of optimization and calculation of the loading weights vector are described in Liland et al. (2009).

Important features: As measure of importance for the classification we take the mean value of the loading weights over the components and order them according to their absolute values.

For the comparison of the single sample design with pooling designs, we consider the intersections of the set of ISF with the important features sets for the pooling designs using the set of important features for the single sample design as reference set.

3 Results

In Figure 1 the mean prediction error over 500 runs is shown for PPLS-DA for different proportions of technical variance and $\gamma \in [0,1]$ (dashed line) and $\gamma \in [0.9,1]$ (solid line).

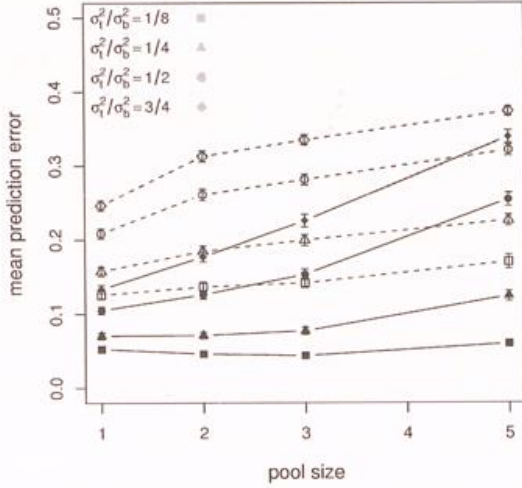


Figure 1: Mean PE with 95% confidence intervals for PPLS-DA, dashed lines and transparent symbols: $\gamma \in [0,1]$, solid line and filled symbols: $\gamma \in [0.9,1]$

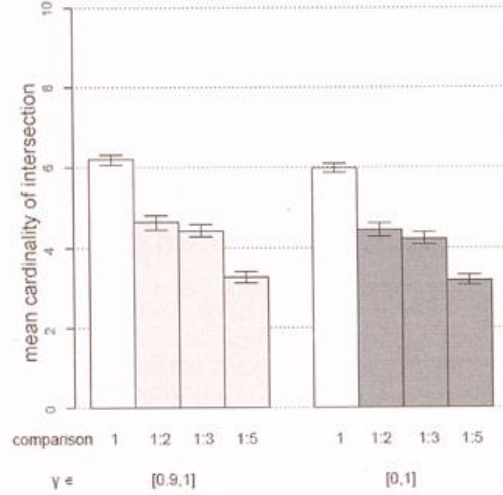


Figure 2: Cardinality of the intersection between important feature sets and the set of ISF

The results display a considerable lower prediction error (PE) for all designs for γ within the upper interval. Considering the results for $\sigma_t^2 = \frac{1}{4}\sigma_b^2$ as an example, the PE for $\gamma \in [0.9,1]$ ($\blacktriangle$) is nearly two times lower than for $\gamma \in [0,1]$ ($\triangle$) for the single sample design. Also for the pooling design, the PE is much lower for $\gamma \in [0.9,1]$. The PE is nearly equal, for designs with a pool size smaller than 3. Nevertheless for $\sigma_t^2 \geq \frac{1}{2}\sigma_b^2$, the PE increases stronger with increasing pool size for $\gamma \in [0.9,1]$.

Comparing the important feature sets for the classification, which is illustrated in Figure 2, no great differences are shown between the two power settings for γ for $\sigma_t^2 = \frac{1}{4}\sigma_b^2$. For the single sample design in average 6 of 10 ISF are important and of these sets between 4.5 and 3.2 are equal with the important features for the pooling design with $m = 2,3,5$.

The reason for the lower PE of PPLS-DA with $\gamma \in [0.9,1]$ is demonstrated in Figure 3 which

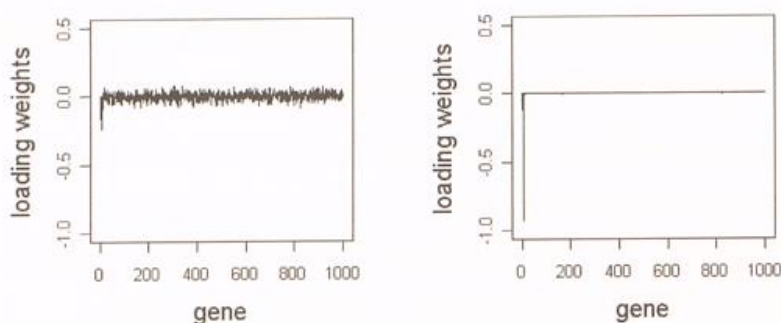


Figure 3: PPLS-DA loading weights for the first component (left) for $\gamma \in [0,1]$ and (right) for $\gamma \in [0.9,1]$ for the single sample design with $\sigma_t^2 = \frac{1}{4}\sigma_b^2$.

shows the loading weights for the first run for the single sample design with $\sigma_t^2 = \frac{1}{4}\sigma_b^2$ as an example. On the left side of Figure 3 the results are based on PPLS-DA $\gamma \in [0,1]$ where the optimization leads to $\gamma = 0.5$. Here also most of the non-informative features have loading weights equal to zero or near zero. On the right side of Figure 3 the result is shown for $\gamma \in [0.9,1]$ with optimal $\gamma = 0.9$. Almost all non-informative simulated features receive a loading weight near to zero, even smaller than for the left side, and differentially expressed genes with high mean class difference show large loading weights.

4 Discussion

The prediction error results for PPLS-DA are considerably improved by constraining the optimization interval of γ to the highest values ($[0.9,1]$) for the single sample design and all pooling designs. Therewith PPLS-DA, $\gamma \in [0.9,1]$ clearly outperforms all other methods used in Telaar et al. (2011) for the classification with differentially expressed genes. Using this setting, the non-informative genes, which could be considered as noise, get a loading weight near to zero and have therewith nearly no influence on the calculation of the PPLS-DA components which are used for a linear discriminant analysis. On the contrary, the values of the genes which have a high correlation to the vector of the group memberships receive larger loading weights.

5 Literature

- Liland, K. H. & Indhal, U. (2009). Powered partial least squares discriminant analysis. *Chemometrics* 23
- Kerr, M. K. (2003). Design considerations for efficient and effective microarray studies. *Biometrics* 59(4)
- Searfoss, G. H., Jordan, W. H., Calligaro, D. O., Galbreath, E. J., Schirtzinger, L. M., Berridge, B. R., Gao, H., Higgins, M. A., May, P. C. & Ryan, T. P. (2003). Adipsin, a biomarker of gastrointestinal toxicity mediated by a functional gamma-secretase inhibitor. *J Biol Chem* 278(46)
- Telaar, A., Nuernberg, G. & Repsilber, D. (2010). Finding biomarker signatures in pooled sample designs: A simulation framework for methodological comparisons. *Advances in Bioinformatics* 2010, 8.
- Telaar, A., Repsilber, D. & Nuernberg, G. (2011) Biomarker Discovery: Classification using pooled samples - A simulation study. *Computational Statistics*. (submitted)

Including metabolomic profiles to improve genetic value prediction: an integrated bioinformatics approach using weighted genome-wide marker information

Nina Melzer, Dörte Wittenburg and Dirk Repsilber

¹Leibniz Institute for Farm Animal Biology, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany

1 Introduction

We develop an integrative bioinformatics approach to improve the prediction of milk phenotypes from genome-wide marker data using metabolite profiles together with a weighting heuristics, based on a realistic simulation study. Towards this objective, we adopt an approach from the field of genomic selection (e.g. Meuwissen [1]). In this field, simulated genome-wide marker data and phenotypes are used to develop methods for predicting genetic values, but to date there does not exist a strategy to include additional information from functional levels of gene expression, such as for example metabolite levels. In our study, we integrate the metabolome level and employ a weighting for the SNP data. The metabolome level is simulated using the metabolism SBML model 70 from Holzhütter [2]. Metabolite profiles can be used to predict a milk phenotype trait using methods of the field of machine learning. Here, several methods can be used together with OMICs profiles, for example Random Forest [3]. From the trained learning machine it is possible to extract the important variables for phenotype prediction. These measures of importance can be used after locating the belonging enzymes, which catalyze the metabolite reactions of the important metabolites. From the identified enzymes it is possible to go further back to the genome level and weight SNPs in the corresponding gene region of the enzyme. The weights and the genome-wide marker data are jointly used to predict the genetic value. For our approach, we applied the extended fastBayesB method [4, 5]. In this study, it is exemplarily tested if and to which degree the weighting approach leads to an improvement of the genetic value prediction.

2 Materials and methods

Data sets: We use ten simulated data sets as described in Melzer et al. [6]. We build our analysis on the genome-wide marker data and on metabolite profiles. We simulate the phenotypes depending on three different values of broad-sense heritability. Further, the training generation and test generation have a