BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 14 2011, pages 1986–1994
doi:10.1093/bioinformatics/btr300

Data and text mining

Advance Access publication May 16, 2011

Classification with correlated features: unreliability of feature ranking and solutions

Laura Tolosi* and Thomas Lengauer
Department of Computational Biology and Applied Algorithmics, Max-Planck-Institute for Informatics, Saarbrücken, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Classification and feature selection of genomics or transcriptomics data is often hampered by the large number of features as compared with the small number of samples available. Moreover, features represented by probes that either have similar molecular functions (gene expression analysis) or genomic locations (DNA copy number analysis) are highly correlated. Classical model selection methods such as penalized logistic regression or random forest become unstable in the presence of high feature correlations. Sophisticated penalties such as group Lasso or fused Lasso can force the models to assign similar weights to correlated features and thus improve model stability and interpretability. In this article, we show that the measures of feature relevance corresponding to the above-mentioned methods are biased such that the weights of the features belonging to groups of correlated features decrease as the sizes of the groups increase, which leads to incorrect model interpretation and misleading feature ranking.

Results: With simulation experiments, we demonstrate that Lasso logistic regression, fused support vector machine, group Lasso and random forest models suffer from correlation bias. Using simulations, we show that two related methods for group selection based on feature clustering can be used for correcting the correlation bias. These techniques also improve the stability and the accuracy of the baseline models. We apply all methods investigated to a breast cancer and a bladder cancer arrayCGH dataset and in order to identify copy number aberrations predictive of tumor phenotype.

Availability: R code can be found at: http://www.mpi-inf.mpg.de/~laura/Clustering.r.

Contact: laura.tolosi@mpi-inf.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 20, 2010; revised on April 19, 2011; accepted on May 10, 2011

1 INTRODUCTION

The accelerated development of microarrays and, more recently, of high-throughput sequencing techniques affords genome-wide measurements of molecular changes in the cell that have an impact on cancer onset and progression. High-resolution experiments targeting gene expression, DNA copy number or DNA methylation in tumors can be the basis for discovering patterns predictive of diagnosis, prognosis and therapy selection (Hicks *et al*., 2006; Ma *et al*., 2007; Mikeska *et al*., 2007; van't Veer *et al*., 2001). Machine-learning techniques for classification and feature selection are often used for automated identification of variables associated with particular tumor phenotypes. In this article, we are concerned with two widely discussed aspects of microarray classification: handling high dimensionality and ill conditioning.

The high dimensionality of microarray-based experiments contrasting to the small number of samples easily leads to overfitting. Regularized linear models such as logistic regression with ridge (Hastie *et al*., 2001) or Lasso penalty (Tibshirani, 1996) are popular solutions to fitting sparse models in which only a small subset of features plays a role. More sophisticated penalties for sparse model selection are discussed by Zou and Li (2008).

The problem of ill conditioning refers to the existence of groups of highly correlated features. The high correlations often have a biological basis, for example if the correlated features relate to the same molecular pathway (coregulated genes in expression data), are in close proximity in the genome sequence (neighboring genes in copy number data) or share similar methylation profile (consecutive CpG dinucleotides in CpG islands). Methods using simple penalties like Lasso typically discard most of the correlated features: only one or a few arbitrary representatives from every group of correlated features enter the model, provided they are relevant for the outcome. As a consequence, the models become unstable: small changes in the training set result in dramatic changes in the selected subset of features. If the purpose of feature selection includes biological interpretation of the model, then stability must be ensured. A successful approach used in many recent articles is that of selection of groups of features. For example, the group Lasso model (Meier *et al*., 2008) consists of Lasso selection of predefined groups of features. The fused support vector machine (Rapaport *et al*., 2008) combines a Lasso and a fused penalty for enforcing similar weights on correlated features, this way performing group discovery and group selection simultaneously. Another approach to group selection adopted in a large class of methods uses clustering procedures to discover feature groups, compute super features to summarize every cluster and apply feature selection on the set of super features. For example, in Park *et al*. (2007), the features are grouped with a hierarchical clustering procedure and the cluster centroids are used for training linear models. The Metagene method (Huang *et al*., 2003a, b) consists of *k*-means clustering of the features, followed by computing the principal components of the clusters, called *metagenes*, which are used for model training. Jäger *et al*. (2003) use fuzzy clustering to determine groups of features and then select a limited number of representatives from each cluster for training SVM models. Yu *et al*. (2008) search for dense groups of

*To whom correspondence should be addressed.

features by kernel density estimation. The *pelora* method (Dettling and Bühlmann, 2004) performs supervised grouping of features, by iteratively updating the groups such that the accuracy of a penalized logistic regression model is increased.

A non-parametric model often used in microarray classification is the random forest (Breiman, 2001; Díaz-Uriarte and Alvarez de Andrés, 2006; Pang *et al*., 2008). In a recent study, Strobl *et al*. (2008) observe that correlated variables are used interchangeably in the decision trees of the random forest models. The authors analyze the consequence of this phenomenon by simulating artificial datasets containing few correlated variables with different predictive values. They notice that the less relevant variables often replace the predictive ones (due to correlation) and thus receive undeserved, boosted importance. Strobl *et al*. (2008) introduce a new variable importance measure that better reflects the predictive power of each feature within a correlated group. In contrast to the study by Strobl *et al*. (2008), we assume that the correlated features in a group share the same predictive value (due to a common underlying biological event) and we investigate how correlation affects the feature importance given by random forest.

The main contribution of this article is to raise awareness of a specific effect involving feature correlation in several of the methods mentioned above, that can misguide model interpretation. We observed that the Lasso penalized logistic regression, the group Lasso, the fused SVM and the random forest report feature weights which are affected by a type of bias which we call *correlation bias*. Specifically, the features which belong to larger groups of correlated features receive smaller weights, proportional to the group size, due to a shared responsibility in the model. Therefore, if the group is large enough, all features may appear irrelevant, even if they yield high correlations with the outcome. This effect is expected in the case of the sparse Lasso logistic regression, but is surprising in the case of group Lasso and fused SVM, which are specifically designed to afford selection at group level and improved model interpretation. Moreover, such bias has not been reported on random forest models previously.

We show using simulations that correlation bias exists and affects several widely used classification models for microarray data. We also show that group selection based on feature clustering such as provided by the method presented in Park *et al*. (2007) can be successfully used for removing the correlation bias. We test and compare the methods investigated on two biological datasets.

2 METHODS

We consider only binary classification problems. Let $(x_i, y_i), i = 1, ..., N$ be N i.i.d. observations of a p-dimensional vector $x_i \in \mathcal{R}^p$ and a response variable $y_i \in \{0, 1\}$. Denote by $X = (x_1, ..., x_N) \in \mathcal{R}^{N \times p}$ the input matrix and $y \in \{0, 1\}^N$ the binary outcome. In general, we will use small letters to refer to samples $x_1, ..., x_N$ and capital letters to refer to features $X_1, ..., X_p$ of the input matrix X. In the manuscript, we will use the notion *feature importance* to refer to the measures of feature relevance commonly used for model interpretation, such as feature weights in linear models or variable importance in random forest.

2.1 Classification methods

Logistic regression is a popular method for classification of biological data. It models the logarithm of the posterior probabilities of the classes as linear functions of the input features. The parameters $w \in \mathcal{R}^p$ of the model are estimated by maximizing the log-likelihood $L(w; X, y)$ over the observations in the training set. Model sparsity is obtained by adding a Lasso penalty λ [see Equation (1)], which can be optimized with cross-validation. Feature importance is given by the model weights w_{LLR}:

$$w_{\mathrm{LLR}} = \arg\max_w L(w; X, y) - \lambda \sum_{j=1}^{p} |w_j| \quad (1)$$

In what follows, we will call this model Lasso logistic regression (LLR). In our experiments, we used the **R** package *glmnet* (Friedman and Hastie and Tibshirani, 2010) for training LLR models.

Logistic group Lasso (Meier *et al*., 2008) uses the logistic regression model with a more specialized penalty, which takes into account some natural grouping of the features. Assume there are G groups of predictors and each group must be entirely included in the model by receiving non-zero weights or be discarded as irrelevant. The group penalty is a combination of a Lasso penalty acting at the group level and a ridge penalty on the predictors within each group. If I_g is the index of the features belonging to group g, then the weights of the logistic group Lasso (GL) model are given by:

$$w_{\mathrm{GL}} = \arg\max_w L(w; X, y) - \lambda \sum_{g=1}^{G} \|w_{I_g}\|_2 \quad (2)$$

We use the **R** package *grplasso* (by Lukas Meier) for training GL models and cross-validation for estimating the optimum penalty λ.

The *fused SVM* (Rapaport *et al*., 2008) has been proposed for the special case that the features can be ordered such that neighboring features are expected to be correlated. This is the case in data on copy number aberrations, where the features are genomic sites ordered by position in the genome. Fused SVM (FSVM) is a linear support vector machine model with two supplementary penalties: a Lasso penalty for model sparsity and a fused penalty, which acts as a smoother of the weights, in such a way that weights of neighboring features are forced to be similar. The weights of the model w_{FSVM} are obtained by minimizing a penalized hinge loss, as follows [see Rapaport *et al*. (2008) for details]:

$$w_{\mathrm{FSVM}} = \arg\min_w \sum_{i=1}^{N} [1 - y_i w^T x_i]_+ + \lambda \sum_{i=1}^{p} |w_i| + \mu \sum_{i=2}^{p} |w_i - w_{i-1}| \quad (3)$$

The optimization problem given by System 3 can be solved by a linear program. We implemented this method using Matlab and the CVX optimization toolbox (Grant and Boyd, 2008). Cross-validation for both penalty parameters λ and μ is necessary, which makes fitting an FSVM model slower than fitting the other methods.

Random forest (RF) models (Breiman, 2001) are non-parametric and non-linear models, attractive due to their interpretability. They are based on averaging over a large collection of decision trees, each trained on a separate bootstrap sample of the input set. The aggregate model has lower variance and is less susceptible to overfitting than a single decision tree. *Gini Importance* (GI) and *Variable Importance* (VI) are two measures of feature relevance that can be computed based on the RF model. We use the **R** package *randomForest* (Liaw and Wiener, 2002) for training RF models. There are two parameters that influence the performance of RF: the number *ntree* of trees in the collection and the number *mtry* of variables considered for each tree split. In our experiments, we use the recommended value $mtry = \sqrt{\text{number of features}}$ and we select the optimal value for *ntree* via cross-validation. Díaz-Uriarte and Alvarez de Andrés (2006) evaluate the performance of RF models for various parameter settings in 10 real-world learning instances. Their results suggest that the default value of *mtry* affords either optimal or close to optimal performance.

2.2 Correlation bias

In the classification of high-dimensional data containing (large) groups of correlated features, the requirements of model sparsity and of retrieving of all predictive features are in direct competition. In applications in which assessment of feature importance is the main objective, models that give

priority to the latter requirement should be preferred. In what follows, we formulate three key properties that we believe a classification model should meet in order to be a good instrument for assessment of feature importance.

Assume that two independent biological events P_1 and P_2 (e.g. the deletion of a chromosome arm can be an event) influence the binary phenotype Y (e.g. tumor stage) and let us denote the magnitude of their effects on Y with $E(P_1)$ and $E(P_2)$, respectively and assume that $E(P_1) > E(P_2)$. Assume that, by means of an experimental technology, variables associated with each of the two events are measured (e.g. all genes located within a deleted chromosome arm). Let us denote with $U_1, ..., U_q$ the variables associated with P_1 and with $V_1, ..., V_p$, the variables associated with P_2, $p, q \geq 1$. Consequently, $\{U_i\}_{1 \leq i \leq q}$ and $\{V_j\}_{1 \leq j \leq p}$ form two groups of correlated variables. Assume a classification model $\mathcal{M}$ is used to predict Y from a set of N observations on features $U_1, ..., U_q, V_1, ..., V_p$ and this model assigns importance values to features: $w_1^1, ..., w_q^1, w_1^2, ..., w_p^2$. Without losing generality, assume all the importance values are positive and a larger value indicates a more predictive feature. The following three properties should hold:

(1) The importance values of the correlated features are similar: $w_1^1 \approx w_2^1 \approx ... \approx w_q^1$ and $w_1^2 \approx w_2^2 \approx ... \approx w_p^2$.

(2) The importance of the variables reflect the magnitude of the effect of the corresponding process on the outcome: $w_i^1 \geq w_j^2, \forall i = 1..q, \forall j = 1..p$.

(3) The importance of the variables $\{w_i^1\}_{1 \leq i \leq q}$ and $\{w_i^2\}_{1 \leq i \leq p}$ does not depend on the corresponding group sizes, namely q and p, respectively.

We require that property (i) holds because, in absence of a true model, it is wise to give fair chances to all correlated variables for being considered as causative for the phenotype. In this case, supplementary evidence from other sources should be used for identifying the causative variable from a correlated group. Property (ii) is based on the assumption that $E(P_1) > E(P_2)$ and hence any of the features $\{U_i\}_{i \leq q}$ contributes more to the outcome than any of the features $\{V_j\}_{j \leq p}$. Thus, the property ensures a fair ranking of the variables, which is important in applications because often only a few top ranking groups are considered for further investigation. Property (iii) demands that the importance of the features does not change as more evidence (more variables) about the corresponding events is added to the data.

In this article, we show that in classification problems with groups of correlated features, the assignment of feature importance by LLR, RF, GL and FSVM does not meet requirements (ii) and (iii). Specifically, the reported feature importance varies with the sizes of the correlated groups of features and results in biased feature ranking. In the context of the example above, the feature weights $\{w_i^1\}_{1 \leq i \leq q}$ and $\{w_i^2\}_{1 \leq i \leq p}$ depend on the values of q and p, respectively, in a way that larger group size leads to smaller importance values. As a consequence, if q is much larger than p, variables $\{U_i\}_{1 \leq i \leq q}$ can falsely appear less predictive than variables $\{V_j\}_{1 \leq j \leq p}$ and P_2 is considered more relevant than P_1.

In sparse models like LLR, correlated features are generally discarded in favor of a single representative. Instability of feature importance is a known issue in such models (Jäger *et al*., 2003; Park *et al*., 2007), and it is easy to observe that the larger the group, the smaller the chance of each particular variable within the group is to be selected by the model. Therefore, under repeated perturbations of the training set, the average weights of the features decrease as the size of the group increases. In the case of FSVM, the weights of correlated features are forced to be equal (or similar). Consequently, if the group of correlated features becomes larger, the common weights need to be decreased, in order to accommodate all features in the model and not violate the Lasso penalty. This rescaling of the weights is possible without decreasing the accuracy of the model, since correlated features provide only redundant information. In the Supplementary Material, we show how the interaction between the two penalties of FSVM can cause correlation bias. (see Example of correlation bias, Supplementary Material.) A similar effect can be observed in GL models. In RF, the correlation bias is caused by the bootstrap sampling of the observations and by the sampling of the features at each node of the trees, which causes correlated features to be used interchangeably in the tree components.

In this article, we say that models that do not meet requirements (ii) and (iii) are affected by *correlation bias*.

2.3 Methods for reducing correlation bias based on FC

Intuitively, a good strategy for reducing the correlation bias is to group the correlated features prior to model fitting and derive corresponding *feature representatives* as a summary of each group. The importance of the original features can be defined as the importance of the corresponding representatives. A very simple and intuitive approach is described by Park *et al*. (2007): average-linkage hierarchical clustering with Euclidean distance is performed on the features. Cluster centroids are used afterwards for training linear regression models. The features are standardized to mean zero and standard deviation one before clustering, such that the Euclidean distance is equivalent with the correlation distance between features. The optimal number of clusters is selected in a supervised manner, by visiting each level of the clustering dendrogram and estimating the model accuracy by means of cross-validation. Park *et al*. (2007) make use of the monotonicity of the penalized loss and propose a path-following algorithm for efficient search for optimal parameters of the method. In this article, we apply the same approach for clustering features, in combination with LLR and RF models. Since RF are non-parametric models, an efficient path-following algorithm for simultaneous optimization of the number of feature clusters and number of trees in the RF is not available. Therefore, the cross-validation procedure is slower than the one presented in Park *et al*. (2007). For a dataset with a few thousands of features, it becomes inefficient to train a model for all levels of the clustering dendrogram. The running time of such a procedure is $\sum_{k=1}^{p} \mathcal{C}(k)$, where $\mathcal{C}(k)$ is the running time required by the classification algorithm to fit datasets with k features and and p is the total number of features. Using a univariate filtering for eliminating irrelevant features can help in applications, in which relatively small number of features are expected to be relevant. For example, the sure independence screening method (Fan and Lv, 2008) ensures that all predictive variables survive the filtering procedure. However, such filtering would not provide with substantial dimension reduction in applications like classification of copy number aberrations, because the number of predictive features can be very large (e.g. all probes covering several chromosomes). Therefore, in this study, we propose a cheaper, unsupervised alternative for estimating the number of clusters, based on silhouette values (Rousseeuw, 1987).

For simplicity, we will refer to the method by Park *et al*. (2007) as *supervised feature clustering* (FC-Sup). By feature clustering (FC), we mean the alternative approach using silhouette values for estimating the number of clusters. We use both methods in combination with LLR and RF and compare them with the baseline methods and with the group-penalty methods GL and FSVM.

2.4 Model evaluation

We compare the performance of classification methods on training sets with (large) groups of correlated features. In particular, we analyze the measures of feature relevance provided by the models investigated and seek for evidence of correlation bias. Additionally, we report and discuss the prediction accuracy of the models (estimated via 10-fold cross-validation) and the stability of the respective feature importance measures.

The stability of feature importance is defined as the variability of feature weights under perturbations of the training set. When the goal of classification is to select the most relevant features, small modifications in the training set should not lead to considerable changes in the set of important covariates. When the true distribution of the training set is not known, stability can be inferred via repeated sampling from the available training observations. In Kalousis *et al*. (2005), classical 10-fold cross-validation is used in order to create 10 overlapping training sets and model stability is estimated by comparing the 10 resulting feature weightings. For this purpose,

the authors propose several measures of similarity between two vectors of feature weights. For our purposes, the Pearson's correlation coefficient is most suitable. Overall model stability is given by the average of all pairwise Pearson correlations between feature weight vectors provided by the models fitted on the 10 variations of the training set. The stability score has a value between -1 and $+1$, with higher values for more stable models.

3 DATA

3.1 Simulated data

Simulation A: we generated datasets with $N=100$ samples and $p=250$ features. The features are divided into three groups: G_1, G_2 and R. Group R has 50 features and the cardinality of group G_2 is $200-|G_1|$, for different values of $|G_1| \in \{100,120,140,160,180\}$. The features in each group are mutually correlated and any two features belonging to different groups are independent. The features in group G_1 are generated from the prototype vector U, which is sampled from a mixture of two Gaussians with equal probabilities: $g_0=\mathcal{N}(0,0.2)$ and $g_1=\mathcal{N}(1,0.3)$. The particular choice for a Gaussian mixture stems from gene copy number data, where g_0 corresponds to those samples with normal copy number and g_1 indicates aberrations (copy number gains, in this case). We generate features $U_1,...,U_{|G_1|}$ with the following procedure: randomly select 20% of the components of U and alter them by adding Gaussian noise $\mathcal{N}(0,0.5)$, then repeat $|G_1|$ times. The features generated this way are correlated with U and with each other and resemble segmented copy number data, which are piecewise constant with occasional changes. Using the same algorithm, we generate a prototype vector V independent from U and corresponding features $V_1,...,V_{|G_2|}$, which form group G_2, and then repeat the procedure to simulate group R. Last, we generate a binary outcome y with the following linear classification rule:

$$y=\begin{cases} 1, \text{ if } 5U+4V-(\overline{5U+4V})+\varepsilon>0, \\ 0, \text{ otherwise.} \end{cases}$$

where $\varepsilon \sim \mathcal{N}(0,0.1)$ and $\overline{5U+4V}$ denotes the average of $5U+4V$. From the simulation parameters, it follows that features in group G_1 are most relevant to the outcome (being correlated to U), features in group G_2 are less predictive and features in group R are irrelevant. Table 1 shows the within group and between group average correlations, summarized over 100 simulated datasets.

Simulation B: we generate more complex artificial datasets by considering 10 groups of predictive features $G_1,...,G_{10}$ and 20 groups of irrelevant features, $R_1,...,R_{20}$. The number of samples is $N=100$. Each of the groups G_2 to G_{10} and R_1 to R_{20} contains 10 correlated features and the cardinality of group G_1 takes, in turn, one of the values $\{10,50,100,200\}$. The groups of correlated variables $G_1,...,G_{10},R_1,...,R_{20}$ are generated from the prototype variables $U_1,...,U_{10},V_1,...,V_{20}$, respectively. The simulation procedure is similar to that of Simulation A, with different parameters: we alter *all* components of the corresponding prototype vector by adding Gaussian noise $\mathcal{N}(0,0.6)$. The binary outcome y is given by the linear rule:

$$y=\begin{cases} 1, \text{ if } 10U_1+9U_2+\cdots+1U_{10}-(\overline{10U_1+9U_2+\cdots+1U_{10}})+\varepsilon>0, \\ 0, \text{ otherwise.} \end{cases}$$

The groups G_1 to G_{10} are ordered decreasingly by their relevance to the outcome. In Simulation B, the within-group correlations are smaller than in Simulation A and the number of features is larger, which makes the identification of the groups of features more difficult. In Supplementary Figure S1, we show the average correlations between features and the average correlations between features and the outcome variable, for each group, summarized over 100 simulations.

3.2 Real data

Bladder tumors: we tested our methods on a set of 98 CGH arrays measuring copy number aberrations in bladder tumors. The experimental settings and data have been described in Blaveri *et al*. (2005). DNA copy number has been measured for 2142 probes distributed over all autosomes. The correlation between adjacent probes is very high (median 0.82), see Supplementary Figure S2. We considered two binary classification problems, by tumor grade and by tumor stage. For grade classification, we used 19 samples with low grade (Grade 1) and 77 samples with high grade (Grade 2 or 3). For stage classification, 84 samples were grouped into two classes: stage Ta (29 samples) and stage T2+ (55 samples). We excluded the intermediary stage T1 [as in Rapaport *et al*. (2008)]. For each classification scenario, we train RF, LLR, FSVM and GL models. We also fit RF and LLR in combination with FC and FC-Sup. For each model, we report accuracy (using 10-fold cross validation), area under the curve (AUC) and feature importance.

Table 1. Pairwise Pearson's correlation between features, averaged within and between groups

	G_1	G_2	R	y
G_1	0.86	0	−0.02	0.63
G_2	0	0.86	0	0.42
R	−0.02	0	0.87	0

The last column shows the average correlation between each feature group and the outcome y.

Breast tumors: In Climent *et al*. (2007), 185 early stage breast tumors were analyzed using arrayCGH technology (UCSF Hum Array 2.0). Copy number aberrations are measured for 2369 BAC probes (chromosomes X and Y excluded). High correlations between neighboring probes are observed, with median value of 0.69 (Supplementary Fig. S3). The authors of the study use statistical tests and report significant associations between certain genetic alterations and ER status (oestrogen receptor) and PR status (progesterone receptor) of the tumors. Using the methodology introduced here, we identify genetic lesions which help discriminate between ER positive and ER negative tumors, and PR positive and PR negative tumors, respectively. In the cohort, there are 60 ER negative and 101 ER positive tumors, and 65 PR negative and 96 PR positive tumors. For all models considered, we report classification accuracy and AUC (using 10-fold cross-validation) and feature importance.

4 RESULTS

4.1 Simulated data

Simulation A: we evaluated the performance of RF, LLR, FSVM and GL with respect to the criteria described in Section 2.4. Figure 1 summarizes the importance values assigned to features from the three groups G_1, G_2 and R. The average feature weights over 100 simulations are shown for each chosen cardinality of group G_1, with indication of standard deviation added. The correlation bias is clearly demonstrated by the decreasing importance of group G_1 as its cardinality increases and conversely, the increasing relevance of G_2 as its cardinality decreases.

In the case of RF, when the number of features in group G_1 is larger than 140 ($|G_1|/|G_2|>2.3$), the ranking of the groups given by GI and VI is incorrect in that features in G_2 falsely appear most relevant for the model (Fig. 1a and b). On average, the same effect is observed in LLR models (depicted in Fig. 1c).

In the case of FSVM and GL, the correlation bias is noticeable even if $G_1=120$ ($|G_1|/|G2|=1.5$) (Fig. 1e). In the context of the formal example from the Supplementary Material (see Example of correlation bias in Supplementary Material), note that our

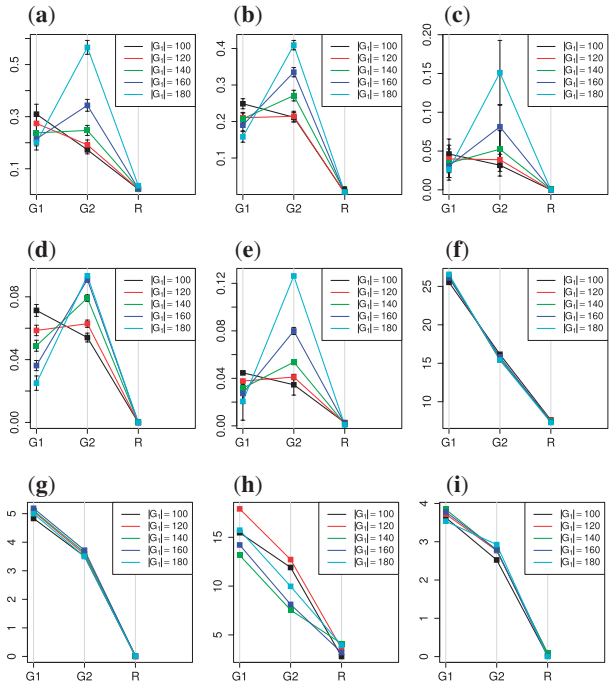

Fig. 1. Average importance of features for classification of data from Simulation A. The importance is averaged over groups G_1, G_2 and R. (**a**) RF (GI); (**b**) RF (VI); (**c**) LLR; (**d**); GL (**e**); FSVM (**f**); RF-FC (**g**); LLR-FC (**h**) RF-FC-Sup; (**i**) LLR-FC-Sup.

experimental results agree with the set of conditions (6) ($a=5, b=4$, $q=120$, $r=80$).

We used FC and FC-Sup as preprocessing step before training RF and LLR models. The FC procedure almost always discovered the correct number of groups of features (three). FC-Sup more often than not overestimates the number of groups. Moreover, the RF models select a larger number of groups than the LLR. Figure 2a shows a summary of the selected number of feature groups by FC and FC-Sup. Importantly, both FC and FC-Sup succeed in removing the correlation bias, which is evident from Figure 1f–i.

The prediction accuracy of all models is summarized in Table 2. All linear models outperform RF, which is expected because the simulations are based on a linear model. With FC, both baseline methods RF and LLR achieve higher accuracy. FC-Sup always outperforms FC, which is surprising, given that FC discovers the true number of feature groups and FC-Sup does not. Most probably, the classification is improved if each group is further split into several subgroups. This is possible because the feature groups are not spherical, but rather elongated, thus a single centroid is not the best representation of the group. As the cardinality ratio $|G_1|/|G_2|$ increases, the accuracy of FSVM and GL decreases and thus the LLR with FC becomes significantly better than FSVM and GL.

Table 3 shows the stability estimates for the various models. The RF are most unstable, probably due to their increased complexity and thus tendency to overfitting. As expected, the LLR are the most unstable among the linear models. FC improves dramatically the stability of RF and LLR. FC-Sup is always more stable than the baseline methods. Interestingly, FC is more stable than FC-Sup. This is the case because the grouping of the features by FC is driven only

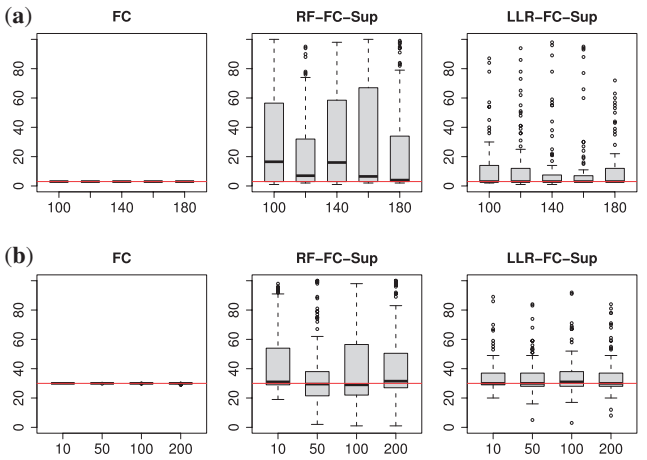

Fig. 2. Boxplot summarizing the number of feature groups selected by FC and FC-Sup (in combination with RF and LLR) for (**a**) simulation A and (**b**) simulation B. The red horizontal line shows the true number of groups. On the x-axis, the cardinality of G_1 is given.

Table 2. Accuracy of classification models on data from Simulation A

$\lvert G_1\rvert$	100	120	140	160	180
RF	0.913±0.03	0.906±0.02	0.911±0.02	0.912±0.02	0.901±0.03
RF-FC	0.915±0.03	0.921±0.03	0.916±0.03	0.921±0.02	0.915±0.02
RF-FC-Sup	0.928±0.03	0.930±0.02	0.925±0.02	0.924±0.02	0.922±0.02
LLR	0.940±0.03	0.941±0.02	0.939±0.02	0.940±0.02	0.938±0.02
LLR-FC	0.966±0.01	0.966±0.02	0.987±0.02	0.970±0.02	0.964±0.02
LLR-FC-Sup	0.972±0.01	0.972±0.02	0.973±0.02	0.973±0.01	0.969±0.01
FSVM	0.967±0.02	0.966±0.02	0.963±0.02	0.965±0.02	0.951±0.02
GL	0.970±0.01	0.965±0.02	0.959±0.02	0.941±0.02	0.884±0.03

The values are averaged over 100 simulations.

Table 3. Stability of feature importance of classification models on data from Simulation A

$\lvert G_1\rvert$	100	120	140	160	180
RF	0.56±0.14	0.52±0.17	0.55±0.16	0.55±0.16	0.55±0.18
RF-FC	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01	1.00±0.01
RF-FC-Sup	0.88±0.14	0.89±0.14	0.86±0.15	0.88±0.15	0.90±0.14
LLR	0.72±0.08	0.72±0.08	0.71±0.08	0.73±0.08	0.75±0.07
LLR-FC	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
LLR-FC-Sup	0.87±0.21	0.86±0.20	0.91±0.17	0.91±0.18	0.93±0.14
FSVM	0.96±0.03	0.98±0.02	0.96±0.07	0.95±0.04	0.95±0.04
GL	0.95±0.02	0.94±0.02	0.94±0.01	0.94±0.01	0.91±0.02

The scores are averaged over 100 simulations.

by the features themselves, while in the case of FC-Sup, the outcome also plays a role. The results show that FSVM and GL are also very stable models.

Simulation B: Figure 3 clearly demonstrates the correlation bias affecting RF (GI) (Fig. 3a), RF (VI) (Fig. 3b), LLR (Fig. 3c), GL (Fig. 3d) and FSVM (Fig. 3e) models. Most dramatically, in the case of FSVM, as the cardinality of group G_1 increases to 200

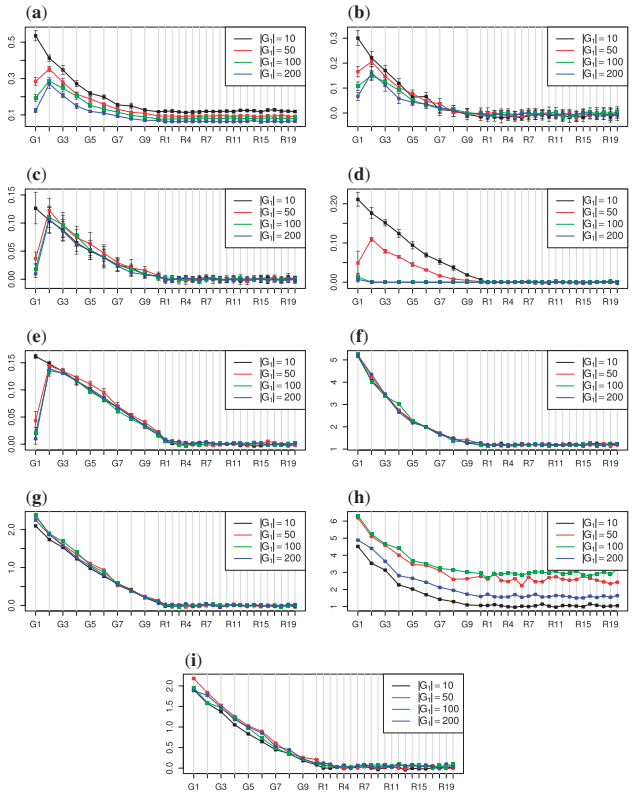

Fig. 3. Average importance of features for classification of data from Simulation B. The importance is averaged over groups $G_1, ..., G_{10}, R_1, ..., R_{20}$ (**a**) RF (GI); (**b**) RF (VI); (**c**) LLR; (**d**) GL; (**e**) FSVM; (**f**) RF-FC; (**g**) LLR-FC; (**h**) RF-FC-Sup; (**i**) LLR-FC-Sup.

features, the features in G_1 appear almost irrelevant. When the size of G_1 exceeds 100 features, the GL model selects only group G_1 and disregards all other predictive groups. As in the case of Simulation A, FC and FC-Sup succeed to remove the correlation bias (Fig. 3f–i). FC in general finds the true number of groups (30), but as the number of correlated features in G_1 increases, the number of groups is sometimes underestimated (Fig. 2b). FC-Sup often selects a larger number of feature groups.

The accuracy of the models is given in Table 4. The models show similar relative performance as seen in Simulation A: FC and FC-Sup always outperform the baseline models and FC-Sup slightly outperforms FC. LLR always outperforms RF, probably due to the underlying linear model. The FSVM and GL lose accuracy as the size of the group G_1 increases.

Supplementary Table S1 shows the stability scores of all models. FC and FC-Sup improve the stability of the baseline models, with FC being more stable than FC-Sup.

4.2 Real data

Bladder tumors: by applying FC to the bladder data, we obtained 113 ± 11 feature groups with grade labeling and 140 ± 31 groups with stage labeling. The varying number of groups corresponds to the 10 training sets of the cross-validation procedure. When applied with RF, FC-Sup method finds 107 groups (with grades labeling) and 20 groups (with stages labeling), respectively. When applied

Table 4. Accuracy of classification models on data from Simulation B

$\|G_1\|$	10	50	100	200
RF	0.741±0.05	0.744±0.03	0.724±0.04	0.714±0.04
RF-FC	0.754±0.05	0.758±0.05	0.758±0.05	0.753±0.05
RF-FC-Sup	0.756±0.05	0.758±0.05	0.758±0.05	0.761±0.05
LLR	0.777±0.06	0.787±0.05	0.783±0.05	0.788±0.05
LLR-FC	0.857±0.03	0.874±0.03	0.868±0.04	0.863±0.03
LLR-FC-Sup	0.870±0.03	0.890±0.03	0.883±0.03	0.880±0.03
FSVM	0.894±0.03	0.898±0.03	0.891±0.03	0.889±0.04
GL	0.828±0.05	0.739±0.04	0.693±0.05	0.690±0.05

The values are averaged over 100 simulations.

with LLR, FC-Sup partitions the set of features into smaller number of groups: 24 (with grade labeling) and 19 (with stage labeling).

In both classification scenarios (with grades and stages labeling), the LLR with FC-Sup yields best accuracy. RF with FC and FC-Sup improve the baseline model in the case of tumor grade prediction but decrease it slightly when tumor stage is predicted.

The stability scores of the prediction models (Table 6) lead to the same conclusions as the experiments on the simulated data: the RF and LLR models are most unstable; however, using FC and FC-Sup results in significant improvements. FC-Sup is more unstable than FC.

Supplementary Figures S4 and S5 show the feature importance reported by all methods investigated. For comparison purposes, to the set of prediction methods analyzed, we added a univariate measure of feature relevance, consisting of t-test p-values (log-transformed) (Supplementary Fig. S4i). A t-test was applied to each feature independently, in order to evaluate the significance of the difference between the means of the two classes. We do not perform multiple testing correction because we use the log-transformed p-values as scores, and not as indicators of relevance. Concerning interpretability, FC and FC-Sup with LLR and RF or FSVM are the better models, reporting clear groups of features with identical weights. GL is also suitable for finding relevant groups; however, there is high variance among the weights within groups.

In the absence of a true model, it is difficult to show how correlation bias affects classification models. However, in the case of classification with stage labeling, we speculate that correlation bias is observable in the feature importance given by FSVM. A large group of 175 correlated features on chromosome 7 is ranked fifth (w.r.t. absolute value of the weights) by FSVM (Fig. 4a). However, a large subgroup of this group of features located toward the short arm of the chromosome is indicated as most relevant by RF with FC, LLR with FC, LLR with FC-Sup, as well as by univariate t-tests (Fig. 4b–e). It is possible that the lower rank of this group of features in the FSVM model can be caused by the correlation bias. The FSVM includes the entire group of correlated features, at the price of lower average weights. In order to verify this hypothesis, we constructed a new dataset, by assigning one feature representative to each group with identical weights in the FSVM model. The representatives are computed by averaging over the corresponding group. All features with null weights were excluded. This procedure essentially uses FSVM for discovering groups of correlated features and computes the centroid of each group selected by the FSVM model. The resulting reduced dataset has 11 features. We trained and evaluated

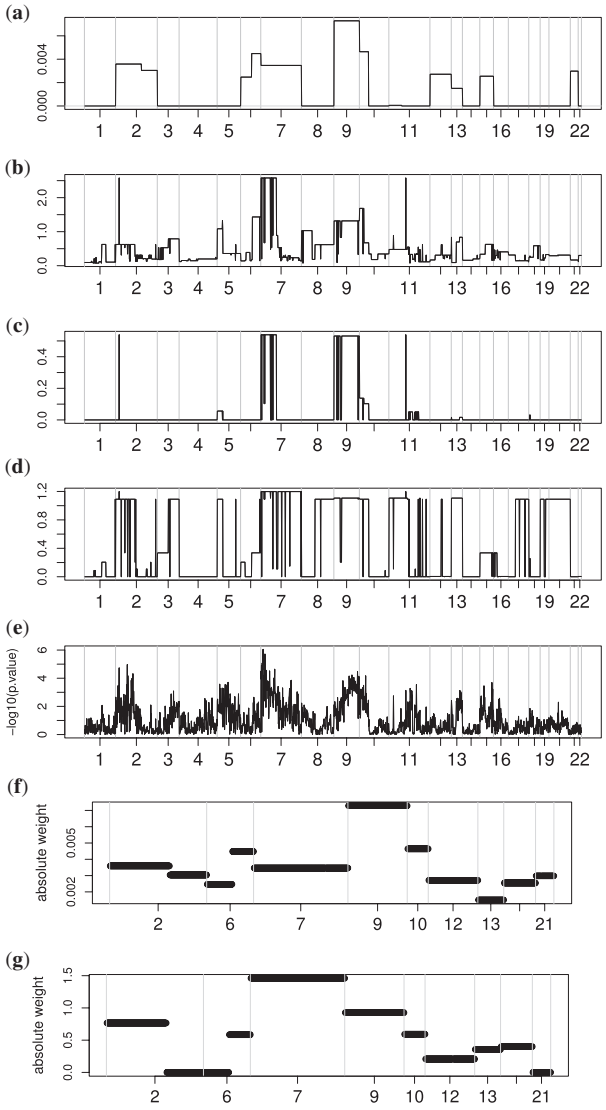

Fig. 4. Feature relevance (in absolute value) by different classification models on the bladder dataset with stage labeling. The features are sorted according to genomic position and the chromosomes are shown along the *x*-axis. (**a**) FSVM; (**b**) RF-FC; (**c**) LLR-FC; (**d**) LLR-FC-Sup; (**e**) *t*-test; (**f**) FSVM (only relevant groups); (**g**) reduced FSVM.

an FSVM model on the new dataset, this time without fused penalty [$\mu = 0$ in Equation (3)], since most probably there are no further groups to be discovered. We call the new model *reduced FSVM*. In Figure 4f, the weights of the features in the original FSVM model are represented (in absolute value and only the regions with non-zero weights). Figure 4g shows the weights of the reduced FSVM (in absolute values), extended for convenience so as to be aligned to the original weights. The representative feature corresponding to the group located on chromosome 7 receives highest absolute weight, which could indicate that the correlation bias has been removed.

Breast tumors: FC method identifies 130 ± 37 groups of features with ER labeling and 127 ± 32 groups of features, with PR labeling. FC-Sup in combination with RF selects an optimal partitioning into

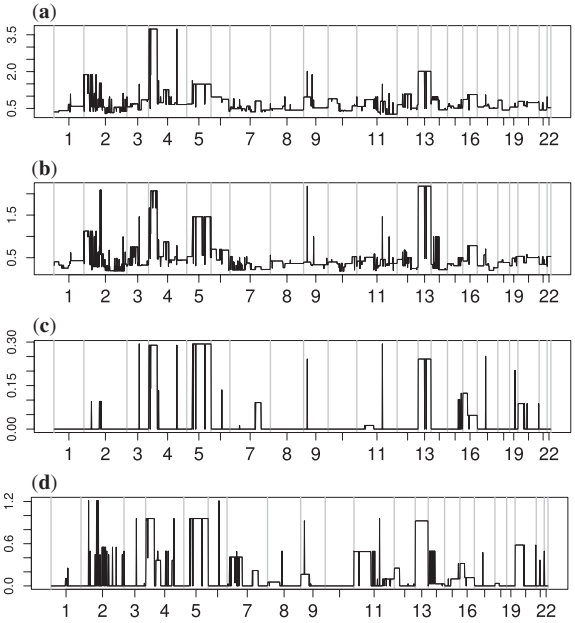

Fig. 5. Feature relevance (in absolute value) by different classification models on the breast dataset with PR labeling. The features are sorted according to genomic position and the chromosomes are shown along the *x*-axis. (**a**) RF-FC; (**b**) RF-FC-Sup; (**c**) LLR-FC; (**d**) LLR-FC-Sup.

195 groups (with ER labeling) and 163 groups (with PR labeling), respectively. Table 7 summarizes the accuracy of the different algorithms on ER and PR classification. FC improves the accuracy of the RF in both cases and of LLR in the case of PR labeling, but decreases slightly the accuracy of the LLR model when ER status is predicted. In the case of ER classification, all RF and LLR models outperform FSVM and GL, however, by a small margin. In the case of the PR classification, FSVM performs best.

The models investigated have similar stability scores as in the case of bladder tumors: RF and LLR are most unstable, FC with LLR and RF have increased stability, comparable to that of GL and FSVM models and FC-Sup is less stable than FC (Supplementary Table S2).

The feature importance reported by the various methods investigated in general confirms the findings reported in the original study (Climent *et al*., 2007) (Supplementary Figs S6 and S7). An interesting aspect is shown in Figure 5: FC and FC-Sup in combination with RF and LLR select a group of features in chromosome 13 as highly relevant for classification of PR status. In the original study (Climent *et al*., 2007), none of these features were reported significant, based on univariate association with the outcome (corrected *t*-test *p*-value). Allelic loss at chromosome 13 is known to be associated with poor prognosis in breast cancer, due to the loss of the tumor suppressor gene BRCA2, located in this region. Associations with low progesterone content have been reported previously in the literature (Eiriksdottir, 1998), which we confirm in our study.

5 DISCUSSION

We have shown that several widely used classification algorithms can generate misleading feature rankings when the training datasets

Table 5. Performance of different classifiers on the bladder data

	Grades		Stages	
	Acc	AUC	Acc	AUC
RF	0.792±0.02	0.827	0.833±0.03	0.882
RF-FC	0.833±0.01	0.878	0.810±0.02	0.882
RF-FC-Sup	0.833±0.01	0.885	0.810±0.03	0.884
LLR	0.823±0.01	0.800	0.798±0.01	0.821
LLR-FC	0.854±0.02	0.838	0.774±0.01	0.757
LLR-FC-Sup	0.865±0.01	0.771	0.845±0.02	0.873
FSVM	0.813±0.02	0.642	0.810±0.05	0.780
GL	0.833±0.02	0.775	0.833±0.04	0.780

Table 6. Stability of feature importance of classification models on the bladder data

	Grades	Stages
RF	0.55±0.03	0.60±0.03
RF-FC	0.80±0.04	0.83±0.04
RF-FC-Sup	0.78±0.06	0.69±0.12
LLR	0.61±0.12	0.66±0.11
LLR-FC	0.86±0.04	0.87±0.08
LLR-FC-Sup	0.72±0.11	0.66±0.18
FSVM	0.75±0.16	0.88±0.05
GL	0.72±0.13	0.87±0.09

Table 7. Performance of different classifiers on the breast data

	ER status		PR status	
	Acc	AUC	Acc	AUC
RF	0.658±0.01	0.664	0.677±0.02	0.673
RF-FC	0.670±0.06	0.635	0.682±0.08	0.660
RF-FC-Sup	0.665±0.02	0.663	0.671±0.02	0.667
LLR	0.683±0.03	0.692	0.683±0.03	0.733
LLR-FC	0.671±0.02	0.718	0.689±0.04	0.723
LLR-FC-Sup	0.696±0.02	0.676	0.714±0.01	0.691
FSVM	0.658±0.02	0.660	0.745±0.02	0.800
GL	0.658±0.01	0.692	0.702±0.02	0.698

contain large groups of correlated features. This can confound model interpretation, since large groups of predictive features can be masked and falsely appear irrelevant. Such an effect is likely to occur because variables relating to a biological process or genomic location of high interest (w.r.t. a phenotype) are overrepresented in the probes set of microarray-based experiments. In this article, we have described the correlation bias and have shown that it affects random forest, Lasso logistic regression, group Lasso and fused SVM models. We used two artificial datasets based on linear models to show that the expected importance of the features in a correlated group decreases as the size of the group increases. We also illustrated the correlation bias caused by the combination of fused and Lasso penalties by means of a theoretical example, which considers the particular case of two groups of correlated features. We showed that correlation bias can be reduced using a group-selection algorithm which combines feature clustering with any classification method. We tested two methods for estimating the number of clusters, based on a unsupervised (FC) and supervised approach (FC-Sup), respectively.

We showed using simulated data experiments that FC and FC-Sup successfully remove the correlation bias, improve the stability of feature importance and increase the accuracy of the baseline methods. FC-Sup outperforms FC in terms of accuracy, but FC is faster and has higher stability. The classification of the real data shows that FC dramatically increases the model interpretability and stability of feature importance. Moreover, in five out of eight classification tasks, FC improved the accuracy of the baseline models. FC-Sup improves the accuracy of the baseline models in six out of eight classification tasks. FC-Sup used in combination with Lasso logistic regression yields highest accuracy in three out of four cases.

Using hierarchical clustering of the features and then computing cluster centroids using the average is certainly not the only solution for identifying and summarizing groups of correlated features. Depending on the distribution of the features in the sample space, methods using principal component as cluster centroid [as in Huang *et al*. (2003a)] or even several representatives [as in Jäger *et al*. (2003)] may yield better performance.

ACKNOWLEDGEMENT

We would like to thank Jörg Rahnenführer, Adrian Alexa, Hiroto Saigo and Konstantin Halachev for their helpful discussions.

Funding: German National Genome Research Network (NGFNplus) (01GS08100 to L.T.).

Conflict of Interest: none declared.

REFERENCES

Blaveri,E. *et al*. (2005) Bladder cancer stage and outcome defined by array based comparative genomic hybridization. *Clin. Cancer Res.*, **11** (19 Part 1), 7012–7022.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45** , 5–32.

Climent,J. *et al*. (2007) Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. *Cancer Res.*, **67**, 818–826.

Dettling,M. and Bühlmann,P. (2004) Finding predictive gene groups from microarray data. *J. Multivar. Anal.*, **90**, 106–131.

Díaz-Uriarte,R. and Alvarez de Andrés,S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Eiriksdottir,G. (1998) Mapping loss of heterozygozity at chromosome 13q: loss at 13q12-q13 is associated with breast tumor progression and poor prognosis. *Eur. J. Cancer*, **34**, 2076–2081.

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **70**, 849–911.

Friedman,J. *et al*. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softwr.*, **33**, 1–22.

Grant,M. and Boyd,S. (2008) Graph implementations for nonsmooth convex programs. *Recent Advances in Learning and Control*, Vol. 371, Springer, pp. 95–110.

Hastie,M. (2001) *The Elements of Statistical Learning*. Springer, NY.

Hicks,J. (2006) Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.,* **16**, 1465–1479.

Huang,E. *et al*. (2003a) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.

Huang,E. *et al*. (2003b) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.*, **34**, 226–230.

Jäger,J. *et al*. (2003) Improved gene selection for classification of microarrays. *Pac. Sympos. Biocomput.*, **8**, 53–64.

Kalousis,A. *et al*. (2005) Stability of feature selection algorithms. In *ICDM '05 Proceedings*, pp. 218–225.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.

Ma,S. *et al*. (2007) Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, **8**, 60.

Meier,L. *et al*. (2008) The group lasso for logistic regression. *J. R. Stat. Soc. B*, **70**, 53–71.

Mikeska,T. *et al*. (2007) Optimization of quantitative MGMT promoter methylation analysis using pyrosequencing and combined bisulfite restriction analysis. *J. Mol. Diagn.*, **9**, 368–381.

Pang,H.*et al*. (2008) Building pathway clusters from Random Forests classification using class votes. *BMC Bioinformatics*, **9**, 87.

Park,M.Y. *et al*. (2007) Averaged gene expression for regression. *Biostatistics*, **8**, 212–227.

Rapaport,F. *et al*. (2008) Classification of arrayCGH data using fused SVM. *Bioinformatics*, **24**, i375–i382.

Rousseeuw,P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Strobl,C. *et al*. (2008) Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

van't Veer,L.J. (2001) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Yu,L. *et al*. (2008) Stable feature selection via dense feature groups. In *Proceedings of the 14th ACM KDD'08.*

Zou,H. and Li,R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.*, **36**, 1509–1533.