Classification models for high-dimensional data with sparsity patterns

Annika Tillander

Classification models for high-dimensional data with sparsity patterns

Annika Tillander

Abstract

Today's high-throughput data collection devices, e.g. spectrometers and gene chips, create information in abundance. However, this poses serious statistical challenges, as the number of features is usually much larger than the number of observed units. Further, in this high-dimensional setting, only a small fraction of the features are likely to be informative for any specific project. In this thesis, three different approaches to the two-class supervised classification in this high-dimensional, low sample setting are considered.

There are classifiers that are known to mitigate the issues of high-dimensionality, e.g. distance-based classifiers such as Naive Bayes. However, these classifiers are often computationally intensive and therefore less time-consuming for discrete data. Hence, continuous features are often transformed into discrete features. In the first paper, a discretization algorithm suitable for high-dimensional data is suggested and compared with other discretization approaches. Further, the effect of discretization on misclassification probability in high-dimensional setting is evaluated.

Linear classifiers are more stable which motivate adjusting the linear discriminant procedure to high-dimensional setting. In the second paper, a two-stage estimation procedure of the inverse covariance matrix, applying Lasso-based regularization and Cuthill-McKee ordering is suggested. The estimation gives a block-diagonal approximation of the covariance matrix which in turn leads to an additive classifier. In the third paper, an asymptotic framework that represents sparse and weak block models is derived and a technique for block-wise feature selection is proposed.

Probabilistic classifiers have the advantage of providing the probability of membership in each class for new observations rather than simply assigning to a class. In the fourth paper, a method is developed for constructing a Bayesian predictive classifier. Given the block-diagonal covariance matrix, the resulting Bayesian predictive and marginal classifier provides an efficient solution to the high-dimensional problem by splitting it into smaller tractable problems.

The relevance and benefits of the proposed methods are illustrated using both simulated and real data.

Keywords: High-dimensionality, supervised classification, classification accuracy, sparse, block-diagonal covariance structure, graphical Lasso, separation strength, discretization.

ISBN 978-91-7447-772-6

Printed in Sweden by US-AB, Stockholm 2013
Distributor: Department of Statistics, Stockholm University

To Micke

List of papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis

PAPER I: Tillander, A. (2012), "Effect of data discretization on the classification accuracy in a high-dimensional framework", *International Journal of Intelligent Systems* 27(4), 355-374.

PAPER II: Pavlenko, T., Björkström, A. and Tillander, A. (2012), "Covariance structure approximation via gLasso in high-dimensional supervised classification", *Journal of Applied Statistics* 39, 1643-1666.

PAPER III: Tillander, A. (2013), "Empirical evaluation of sparse classification boundaries and HC-feature thresholding in high-dimensional data", *Research report 2013:5*, Department of Statistics, Stockholm University.

PAPER IV: Corander, J., Koski, T., Pavlenko, T. and Tillader, A. (2013), "Bayesian Block-Diagonal Predictive Classifier for Gaussian Data ", *Synergies of Soft Computing and Statistics for Intelligent Data Analysis* 190, 543-551.

Acknowledgments

I would like to express my deepest appreciation to all those who provided me the possibility to complete this thesis.

Foremost, a special gratitude to my supervisor Docent Tatjana Pavlenko for the continuous support of my PhD study, for her commitment, accessibility and immense knowledge. I really value being introduced to such an interesting and modern research field.

I would also like to thank my assistant supervisor Prof. Daniel Thorburn for his helpfulness and effort to improve my writing. It is a true favor having the possibility to work with someone of such profound knowledge on literary all fields of statistics.

I am grateful to Prof. Timo Koski for valuable input and insightful comments on my licentiate thesis. I really appreciated the opportunity of writing a paper together and inspiring discussions.

I would like to offer my special thanks to my coauthor Dr. Anders Björkström for his willingness to give his time so generously and his kindness.

Advice and interest in my work given by Prof. Dan Hedlin has been a great help in writing the kappa.

Thanks to the Department of Statistics at Stockholm University for financial support during my period of study.

Gratitude goes to all my colleagues, former and present, at the Department. Especially the fellow doctoral students, your friendship, kindness and humour helped me get through this demanding task. I really appreciated the company during hot summer days and weekends. A special thanks to Sofia for the cover image suggestion.

I would like to thank my family and friends for their love and support. I would never have been able to manage this work without my wonderful husband, Mikael Tillander. Writing thesis while having small children requires the backing of a truly awesome man.

I am incredibly thankful for my two amazing sons; Matti who asked such thoughtful questions as what would happen if the thesis was not finished and promised that I would get Christmas presents even if I would not finish in time. Max who learned to stay up a couple of hours each night so we could spend some time together. Just wonder how long time it will take to unlearn. You two really are the best thing in my life.

Stockholm, October 2013

Annika

Contents

1 Introduction

Background

In several modern application fields, genomics and proteomics, for example, technical devices automatically generate measurements of thousands of features for each given sample unit. This type of feature glut is combined with the difficulty of obtaining good observational units; often the sample size is in the hundreds. These kinds of data represent high-dimensionality: a small sample setting where the number of measured features, p, can grow and exceed the total number of samples, n. Standard estimation methods are not designed to cope with this type of dimensionality and a number of modern statistical methods have started to address this challenging problem; see [4, 8, 32, 39].

The area of interest in this thesis is classification, which is a supervised learning technique. It arises frequently from bioinformatics like disease classification. For example, in gene expression microarrays, some genes demonstrate significant differences in expression levels, which can help distinguish between tumor and normal tissue. For a typical scenario, we have an outcome measure (e.g. tumor/normal tissue) that we want to predict based on a set of features (e.g. genes). The *training* data is then a set in which both the outcome and the features have been observed. Using this, a classification model is built which will enable us to predict of the outcome for new observations where only the features are known. For a recent overview of methods for high-dimensional classification we refer to [13, 21].

The mentioned distance-based classifiers in [13], such as Naive Bayes (NB) and k-nearest neighbor (k-NN), are known to be less sensitive to high-dimensionality, though the methods are usually designed for categorical features or are computationally intensive. These methods have been shown to perform better and have faster computational times for discrete features than continuous features for traditional dimensionality; see [46, 42, 33]. Several ways exist to transform continuous features into discrete features and many studies compare and evaluate different discretization methods for settings where $p < n$; see e.g. [20, 25]. In applications feature discretization for high-dimensional data have been used; see e.g. [36, 18].

When theoretically analyzing performance accuracy, the distributional properties of the suggested classifiers are needed. This, in part, stimulates the extensive research in adjusting linear and quadratic discriminant procedures to high-dimensional settings. The main challenge in this direction is the estimation of the inverse covariance matrix, Σ^{-1} in the case of $p > n$. Recently, a number of regularization techniques have been suggested for improving the estimation of inverse covariance in the classification framework; see, for instance, [19, 24, 45] where a number of modified classifiers were suggested and [22, 23, 40] where the performance characteristics of several regularized classifiers were examined. Other types of regularization are based on exploiting sparsity patterns in Σ^{-1}. Sparse inverse covariance matrices are widely studied as graphical models since

they imply a graph structure: under the Gaussian assumption, zeros in Σ^{-1} , i.e. conditional independences, correspond to absent edges in the graphical model. Hence, learning a sparse Gaussian graph corresponds precisely to recovering the structure of Σ^{-1} with many zeros. A popular technique in this direction is *graphical Lasso* (gLasso). It is based on applying an ℓ_1 penalty to the entries of the inverse covariance matrix; see [14]. A number of authors have proposed the estimation of sparse graphs by ℓ_1 regularization; see e.g. [30, 38, 7, 26] and the references therein.

Technical devices that make it possible to survey thousands of feature measurements at once seem to be attractive for applications. As an example, consider the gene expression microarray again: medical research teams seek for those genes that are highly informative for training a classifier, which, in turn, supposedly gives a reliable automatic diagnosis. However, it turns out that in many real problems there are simply too many useless features being produced by the automatic measurements, so that even if there are *really* discriminative features, they simply will be very difficult to be detect reliably. This type of setting is called *sparse and weak* (SW), meaning that in the underlying model the number of informative features is assumed to be small and the separation strength of each individual feature is low. (Observe that sparsity in the context of SW has different sense than in gLasso-based regularization technique considered above.) The detection of informative features, or feature thresholding, in SW settings while naturally improving classification accuracy is a challenging problem which has attracted a lot of attention in the recent literature [1, 12, 2, 3, 27]. An especially powerful technique is related to testing a very large number of hypotheses where the number of false-nulls is assumed to be very low, thereby representing the model sparsity. The crucial idea is based on analyzing the behavior of second-level significance testing for comparing the fraction of observed significances to the expected fraction under the global null; see Tukey's *Higher Criticism* (HC) ([9, 10, 11]).

Challenging problems when learning supervised classifiers with few observations in the training data, in the setting $p < n$, using the maximum likelihood approach, were noted in the early 90's by Seymour Geisser [16]. As a solution, a Bayesian strategy using various types of *a priori* classification uncertainty was suggested; see e.g [37, 29]. In general, the Bayesian approach in this context may be interpreted as a way of regularizing the problem through the information introduced by the prior distribution for the model parameters. A promising approach to handle high-dimensionality is the *predictive Bayesian classifier*, which explores a simultaneous prediction problem for all samples in the test data in contrast to the standard linear or quadratic classifiers where test samples are labeled one by one using the probabilistic model learned from the data. The inductive nature of Bayesian predictive inference (see [15, 16]) reveals how the uncertainty about both generating distributions of the classes (e.g. class parameter priors) and about the class memberships of all the test units can be simultaneously combined to define a predictive measure for classifying test data conditional on a training data set. Earlier studies in this direction include e.g.

classification of data from multiple finite alphabets [5].

Thesis contribution

In this thesis three different types of approaches to the challenge high-dimensionality poses to classification models are considered.

On the question of the discretization effect on classification performance when $p > n$ is evaluated, this thesis further suggests an effective algorithm for discretization with respect to classification accuracy. Since discretization is a data-transformation procedure, aspects of this step′s effect on the dependence structure between features is investigated.

The useful gLasso regularization technique is embedded in the classification framework and in this thesis we present a two-step approach that produces a block-wise sparse estimate of the inverse covariance matrix. We show that the block-diagonal approximation of the inverse covariance matrix leads to an additive classifier. Further, we show that our estimation approach allows for substantial improvement of the classification accuracy in high-dimensional situations.

In this thesis an asymptotic framework is derived that represents the sparse and weak block (SWB) model and suggests a technique for block-wise feature selection by thresholding. The procedure extends standard HC thresholding to the case where the dependence structure underlying the data can be taken into account and is shown to be optimally adaptive, i.e. performs well without knowledge of the sparsity and weakness parameters. The detection boundary for the extended HC procedure and the performance properties of some estimators of sparsity parameters are empirically investigated.

Due to the block-diagonal-structured covariance matrix and classes represented by Gaussian distribution, a closed form expression for the posterior predictive distribution of the data is established in this thesis. Given the factorization of the distribution, the resulting Bayesian predictive and marginal classifier provides an efficient solution to the high-dimensionality problem by splitting it into smaller tractable problems. Further, we show for synthetic data that our proposed method outperforms several alternative algorithms.

The remaining part of the thesis consists of a summary of the subject matter that is of relevance for the thesis followed by four appended papers (I-IV). In Section 2, a brief introduction to classification and high-dimensionality is presented. Section 3 explains some basics of the discretization procedure. In Section 4, some concepts used in Paper II are introduced. In Section 5, a short presentation of the SW model is given.

2 Classification in a high-dimensional framework

Classification

High-dimensional data collection devices such as microarrays produce enormous amount of information which cannot be overviewed without rearranging and summarizing it in sensible ways. This is where classification can assist. We give here a quick introduction to the elements of classification [13].

Let $\mathcal{X}$ be some input space and $\mathcal{Y}$ be some output space. We have a *training* data set $(\mathbf{X}_j, Y_j) \in \mathcal{X} \times \mathcal{Y}, j = 1, \ldots, n$ where $\mathbf{X}_j$ is the feature vector of the jth observation and Y_j is an associated outcome variable . Further, we assume that we have $\mathcal{C}$ categorical classes and $\mathcal{Y} = \{1, 2, \ldots, \mathcal{C}\}$. Given a new observation $\mathbf{X}$, classification aims to find a function $g : \mathcal{X} \to \mathcal{Y}$ which can predict the unknown class label Y for the new observation using the *training* data as accurately as possible. One way to measure the accuracy of the classifier is to introduce a loss function. Often used for classification is the *zero-one loss*:

$$L\left(y, g\left(\mathbf{x}\right)\right) = \begin{cases} 0 & \text{if } g\left(\mathbf{x}\right) = y \\ 1 & \text{if } g\left(\mathbf{x}\right) \neq y \end{cases}. \tag{2.1}$$

Then the expected misclassification for a new observation, i.e. the risk of the classification function g, takes the following form

$$\varepsilon = E\left[L(Y, g(\mathbf{X}))\right] = E\left[\sum_{c=1}^{\mathcal{C}} L(Y, g(\mathbf{X}))P(Y = c|\mathbf{X} = \mathbf{x})\right] = P(Y \neq g(x)|X = x), \tag{2.2}$$

where Y is the class label of $\mathbf{X}$. Hence, the optimal classifier in terms of minimizing the misclassification rate is

$$g^*(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} P(Y = c|\mathbf{X} = \mathbf{x}). \tag{2.3}$$

This is known as the *Bayes classifier*, which is a classifier that assigns a new observation to the most plausible class using the posterior probability of the response. Suppose that the observation $\mathbf{x}$ has the conditional density $f_c(\mathbf{x})$, being in class c, and let π_c be the prior probability of class c, with $\sum_{i=1}^{\mathcal{C}} \pi_{i=1}$. A simple application of Bayes′ theorem gives us

$$\mathbf{P}(Y = c|\mathbf{X} = \mathbf{x}) = \frac{f_c(\mathbf{x})\pi_c}{\sum_{i=1}^{\mathcal{C}} f_i(\mathbf{x})\pi_i}. \tag{2.4}$$

Then the Bayes classifier becomes

$$g^*(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} f_c(\mathbf{x})\pi_c. \tag{2.5}$$

In this thesis we focus on discriminating between two classes, i.e. $\mathcal{C} \in \{1, 2\}$, and assume that each class is modeled by the Gaussian distribution:

$$\mathbf{x}_c \sim N(\boldsymbol{\mu}_c, \Sigma), \tag{2.6}$$

where $\boldsymbol{\mu}_c$ is the class mean vector and Σ is the common covariance matrix. Next, we consider the well-known *Fisher linear discriminant* analysis. If an observation $\mathbf{x}$ belongs to class c, then its density is

$$f_c(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} e^{\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right\}}. \tag{2.7}$$

Given this assumption, the classifier assigns $\mathbf{x}$ to class 1 if

$$\pi_1 f_1(\mathbf{x}) \geq \pi_2 f_2(\mathbf{x}), \tag{2.8}$$

which is equivalent to

$$\log \frac{\pi_1}{\pi_2} + \left(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0. \tag{2.9}$$

This is the same as the Bayes classifier which can be seen in (2.5) and the classification rule defined in (2.8). The *Fisher discriminant function*, $D_F(\mathbf{x}) = \left(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, assigns $\mathbf{x}$ to class 1 if $D_F(\mathbf{x}) \geq \log \frac{\pi_2}{\pi_1}$ otherwise to class 2. Let $\varepsilon(D, \boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is a shift vector, be the misclassification rate of a classifier with discriminant function D. Then the discriminant function D_B of the Bayes classifier minimizes the misclassification. Further, when $\pi_1 = \pi_2 = \frac{1}{2}$ the misclassification rate for the Fisher discriminant function can be calculated as

$$\varepsilon(D_F, \boldsymbol{\mu}, \Sigma) = \Phi\left(-\frac{1}{2}\sqrt{\delta^2(\boldsymbol{\mu}, \Sigma)}\right), \tag{2.10}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function and $\delta^2(\boldsymbol{\mu}, \Sigma) = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$ is the Mahalanobis distance measuring the distance between the two classes. Under the normality assumption the Fisher discriminant analysis is the Bayes classifier and the misclassification rate in (2.10) is the Bayes risk. This is used throughout the thesis to have a controlled misclassification rate. Further, we focus on δ^2 to use it as measure of separation power for features, as ε is a function of δ^2. Since Φ is a monotone strictly decreasing function of δ^2, we can say that the separation power of features can be a measure of their contribution towards the distance between the classes.

High-dimensionality

As mentioned, it is the number of available features that defines dimensionality and it is the relation between p and the number of available observations that determines whether it is a high-dimensional problem or not. Standard statistical methods have been developed for situations having considerably more observations than features. However, in a case with more features than available observations the problem is said to be "high-dimensional". In the asymptotic analysis the number of features is no longer fixed, so for the situation with "large p, small n" the number of features $p = p_n$ can grow with n, possibly very fast, so that $p_n \gg n$ when $n \to \infty$[31].

Classical statistical methods are based on standard asymptotic behavior, where $n \to \infty$ while p remains fixed and the ratio $\frac{p}{n}$ is treated as $\frac{1}{n}$. For growing-dimension asymptotic, the number of variables can also go towards infinity, unlike the standard asymptotic ratio $\frac{p}{n} \to k$, where $k \in (0, \infty)$. To demonstrate a well-known reason why standard methods are not applicable we consider the inverse covariance matrix, Σ^{-1}, used in many methods such as discriminant analysis and regression analysis [43]. The expected value for the inverse of the standard maximum likelihood estimated covariance matrix, $\hat{\Sigma}$, using the properties of the Gaussian distribution [29] is

$$\begin{aligned} E\left[\hat{\Sigma}^{-1}\right] &= \psi(p,n)\,\Sigma^{-1} \\ \psi(p,n) &= \frac{n}{n-p-1} = \frac{1}{1-\frac{p-1}{n}} \end{aligned} \tag{2.11}$$

This shows why the standard asymptotical arguments break down for the high-dimensionality. For the standard asymptotic, $\psi(p,n) \to 1$ and $E\left[\hat{\Sigma}^{-1}\right] \to \Sigma^{-1}$; however, for the growing-dimension asymptotic the scenario is different depending on the ratio between p and n [34]. In order to see the effect, we simulated data with different ratios. We generated data as i.i.d. $\{\mathbf{x_j} \in \mathbf{N}(\mathbf{0}, \mathbf{I})$ with $j = 1 : 1000$ for $p = (10, 500, 1000, 2000)$. We estimated the covariance matrix for each data set and calculated the eigenvalues. The eigenvalues were ordered and plotted against the rank; see Figure 1.

For the special case when $k \in (0, 1)$ and $\Sigma = I_{p\times p}$ the empirical distribution of the eigenvalues of $\hat{\Sigma}$ follow the Marchenko-Pastur law $\left[\left(1-\sqrt{k}\right)^2; \left(1+\sqrt{k}\right)^2\right]$ [28]. If we consider the situation where p is the same order of magnitude as n and $k < 1$ but not negligible, then the covariance matrix is still invertible but inverting it amplifies estimation error dramatically. It can clearly be seen in Figure 1 that when the ratio between p and n is close to one, the estimation for the standard methods will be biased and when $p > n$ these methods are not applicable at all. A real data example can be seen in Figure 2, which shows the estimated misclassification of linear discriminant analysis when we let the number of features grow. The data in question is breast cancer microarray data; the set from [44] has 62 observations and the set from [35] has 159 observations. For the covariance matrix to be invertible the number of available observations is the limit for the number of features to be included. Both data sets contain several thousands of features but only a tiny fraction can be used; we ordered the features according to the absolute t-value and selected the most informative features. The estimated misclassification increases rapidly with the growing number of features.

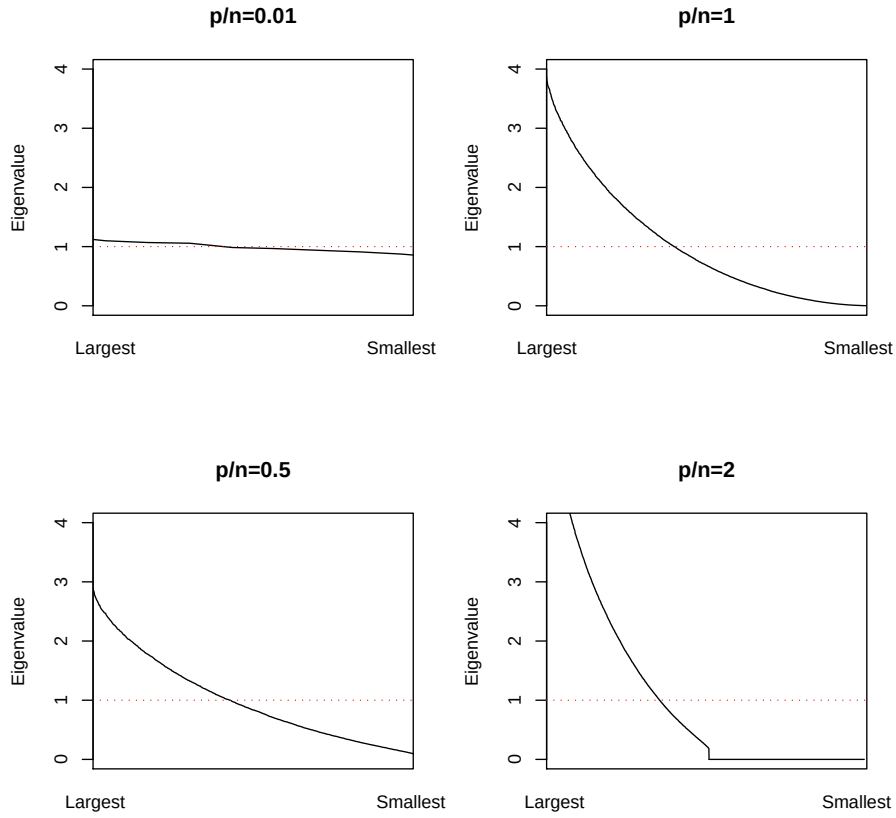

Figure 1: Sample and true eigenvalues. The solid line represents the distribution of $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ of $\hat{\Sigma}$, which are sorted from the largest to the smallest and plotted against their rank. For $\Sigma = I$, $\lambda_1 = \ldots = \lambda_p = \lambda = 1$ and the distribution λ_ith is plotted as a horizontal line at one [34].

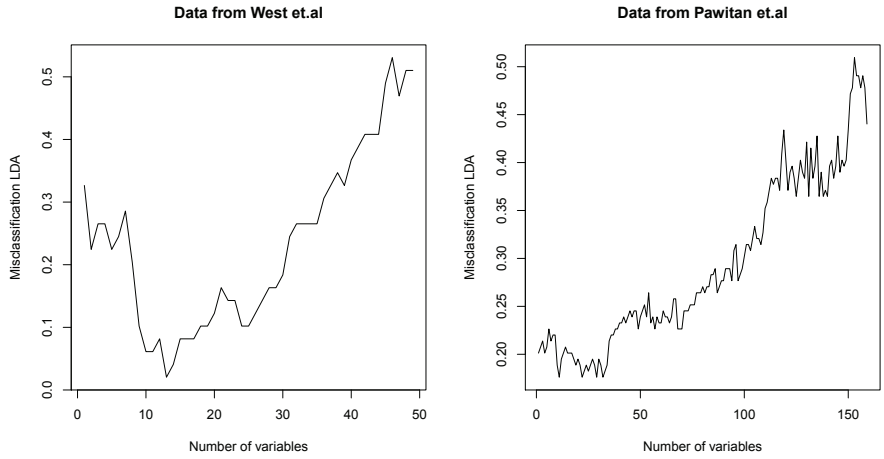

Figure 2: Estimated misclassification with linear discriminant analysis as a function of the number of features, for two real breast cancer microarray data sets

3 Discretization

The underlying distribution of the data is seldom known and classification methods that do not require distribution assumptions are important tools. The kinds of classifiers that require very mild assumptions and work even when $p \gg n$, e.g. k-NN, are often computationally intensive and therefore more effective for discrete data. The discretization procedure is the mapping of the continuous feature into discrete space, grouping together multiple values of the feature and partitioning the continuous domain into non-overlapping intervals. We have n observations and assume that we have n different values for the feature and partition into q number of intervals so that $q < n$. This of course leads to information loss but enables faster calculation times.

In this thesis we consider discretization methods that start by sorting the n values of each feature (x_i) separately in increasing order

$$x_{i1}, x_{i2}, \ldots, x_{in} \longrightarrow x_{i(1)}, x_{i(2)}, \ldots, x_{i(n)},$$

where i goes from 1 to p, $x_{i(1)} = \min\{x_j\}$ and $x_{i(n)} = \max\{x_i\}$. The next step is splitting or merging by sequence of thresholds, each denoted τ_k. Splitting is a top-down method that considers an interval containing all known n values of a feature, and then splits this interval into smaller and smaller subintervals until some predefined criterion for stopping is fulfilled. Merging is a bottom-up method that starts with a certain number of intervals, and these are merged during execution until the predefined stop criteria is fulfilled. For the splitting procedure τ_k is considered as a cut point where an interval of sorted values is split into two adjacent intervals, whereas for the merging process, τ_k is considered as a merging point where two adjacent intervals join. Therefore, to receive q intervals, $q - 1$ thresholds are needed for splitting and $n - q$ thresholds are needed for merging. The choices of thresholds are based on the discretization method and the intervals created are labeled by integers, $k_i \in \{k_1, \ldots, k_{q_i}\}$. Each possible value of the feature is assigned to one and only one interval. The assignments can be characterized by many-to-one mapping, or encoder $k_i = C(x_{i()})$, that assigns the ith value to the mth interval, where j goes from 1 to n and m goes from 1 to q_i.

4 Estimating the inverse of a sparse covariance matrix

In a situation when we have some knowledge about the distribution and the higher demands of assumptions can be fulfilled it is preferable to turn to the linear models due to their stability. Then we need to focus on the handling of the covariance matrix in $p > n$. For the estimation of the inverse covariance matrix we use gLasso [14] as a launching point and then apply the Cuthill-McKee ordering algorithm [6] to form a block-diagonal structure approximation of Σ^{-1}. gLasso is used to learn the sparsity patterns of the covariance matrix, where the algorithm applies a ℓ_1 penalty to the entries of the inverse covariance matrix [14]. gLassso uses the fact that we can learn about the dependence structure through multiple linear regression. With this algorithm we create the

skeleton ($\mathbf{S}$), which is described in more detail below, since when gLasso finds two variables to be conditionally dependent, the matrix entry (i, j) is non-zero. The Cuthill-McKee ordering algorithm aims at reducing the bandwidth, where the bandwidth of a matrix is the maximum value of $|i - j|$ for non-zero elements in the matrix. The bandwidth is reduced by moving the non-zero elements of the matrix closer to the main diagonal. How the non-zero elements should be moved is decided by relabeling the vertices in the graph associated with the matrix in consecutive order [6]. Both gLasso and Cuthill-McKee ordering are based on graph models, so the next section is a short introduction to graph terminology.

Graph terminology applied to skeletons

Let $\mathbf{S}$ denote a skeleton which is a symmetric positive definite Boolean matrix with i rows and j columns where $i, j \in \{1, \ldots, p\}$ and an element $s_{ij} = 0$ indicates that variables i and j are conditionally independent given all other variables. The undirected graph of $\mathbf{S}$ is denoted $G(\mathbf{S}) = (V, E)$, where V is a finite set of vertices together with a set of edges, E. In our context, we use $V = \{1, \ldots, p\}$ corresponding to some random variables $X_1, \ldots, X_p$ and $\{X_i, X_j\} \in E$ iff $s_{ij} = s_{ji} \neq 0, i \neq j$. Two vertices i and j are adjacent if there is an edge between them. The adjacent set of vertices i, denoted $adj(i, G(\mathbf{S}))$, is the set of all vertices that are adjacent to i in $G(\mathbf{S})$. $Deg(i) = |adj(i, G(\mathbf{S}))|$ is the degree of i, which is the number of vertices in $adj(i, G(\mathbf{S}))$. When the relabeling is done in the Cuthill-McKee ordering the vertices should be labeled in increasing order of degree. A path is a sequence of vertices $\{1, \ldots, k\}$ such that i is adjacent to $i+1$ for each $i = 1, \ldots, k-1$. The distance, $d(i, j)$ between two vertices i and j in the graph $G(\mathbf{S})$ is the length of the shortest path joining the two vertices. The eccentricity of vertex i is $e(i) = \max\{d(i, j)|j \in V\}$ and the diameter of $G(\mathbf{S})$ is given by $\delta(G(\mathbf{S})) = \max\{d(i, j)|i, j \in V\}$. A vertex is said to be peripheral if its eccentricity is equal to the diameter of the graph. This is an important concept for the Cuthill-McKee ordering algorithm since a peripheral vertex should be chosen as the starting point. The algorithm we use for finding a pseudo-peripheral node can be seen in [17].

Block diagonal structure

A graph is connected if every pair of distinct vertices is joined by at least one path; otherwise $G(\mathbf{S})$ is disconnected and consists of two or more connected components. It is clear that if $G(\mathbf{S})$ is disconnected and consists of r connected components and each component is labeled consecutively, the corresponding matrix will be block diagonal. Each connected component in the graph will correspond to a diagonal block in the matrix; see Figure 3.

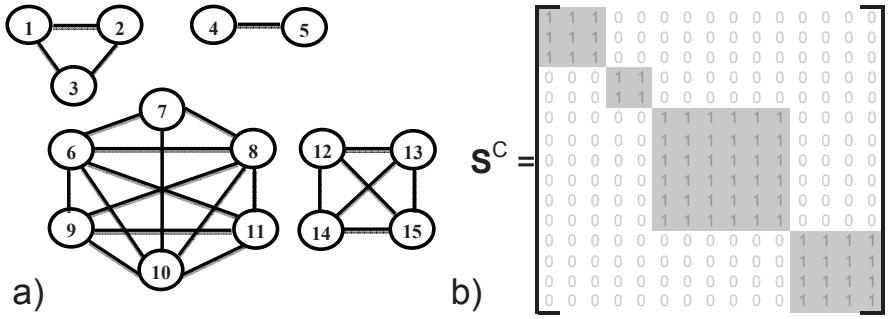

Figure 3: a) A disconnected graph with four connected components that have been labeled in consecutive order. b) A block diagonal matrix corresponding to the graph in a.

5 Feature thresholding for the sparse and weak model

As mentioned in the introduction, it is common for high-dimensional data that very few out of the thousands of measured features are informative for classification, i.e. the data is sparse. Further, those features that are informative often have low separation strength, i.e. the data is weak. This is known as the sparse and weak model and identifying the informative features is a challenging statistical problem. As an illustrative example, the distribution of observed separation scores from two different mixtures can be seen in Figure 4. The problem of identifying informative features is formulated as follows: Given p independent observations of separation strength $\mathcal{Z} = (Z_1, Z_2, \ldots, Z_p)$. For each $1 \leq i \leq p$, we suppose that Z_i has the probability β of being informative and $1-\beta$ of being non-informative. Here we consider the separation score for single features. Hence we model the non-informative features as samples from $N(0,1)$ and informative features as samples from $N(\theta,1)$. Then, β can be viewed as the proportion of informative features and $\theta \neq 0$ is the shift parameter. Then the goal is to test whether any informative features are present, i.e. we wish to test the hypothesis $\beta = 0$ or equivalently

$$\begin{aligned} H_0 : &\quad Z_i \sim N(0,1) \text{ i.i.d } 1 \leq i \leq p \\ H_1 : &\quad Z_i \sim (1-\beta)N(0,1) + \beta N(\theta,1) \text{ i.i.d } 1 \leq i \leq p. \end{aligned} \tag{5.1}$$

We adopt an asymptotic framework where β and θ are parameterized as functions of the driving variable p. For a fixed parameter $\frac{1}{2} < \gamma < 1$ we let

$$\beta = \beta_p = p^{-\gamma}.$$

In this sparse regime, $\beta_p \ll 1/\sqrt{p}$, the challenging situation is when the weakness parameter grows at a rate of $\sqrt{\log(p)}$ because outside this range it is either too easy to separate the two hypotheses or it is impossible. Hence we let

$$\theta = \theta_p(r;\gamma) = \sqrt{2r\log(p)},$$

where $0 < r < 1$.

Many of the commonly employed feature selection strategies are based on significance levels ("p-values"); see e.g. [1, 12, 2, 9]. Hence we obtain the

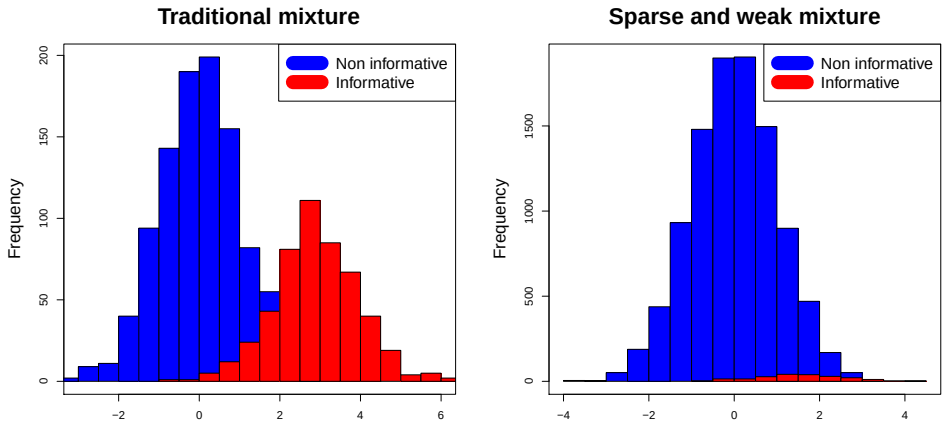

Figure 4: Observed separation strength from two different Gaussian mixtures.

"p-value" for the absolute value of the separation score with

$$\pi_i = P_{H_0}\left\{|Z_i| \geq |z_i|\right\} = 2(1 - \Phi(|z_i|)),$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function. As a next step the "p-values" are ranked in increasing order: $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(b)}$. Then the aim in this multiple-testing situation is identifying the false null. The feature thresholding procedure finds a cutoff which generates a subset with as many true informative features as possible while keeping the non-informative to a minimum, as illustrated in Figure 5.

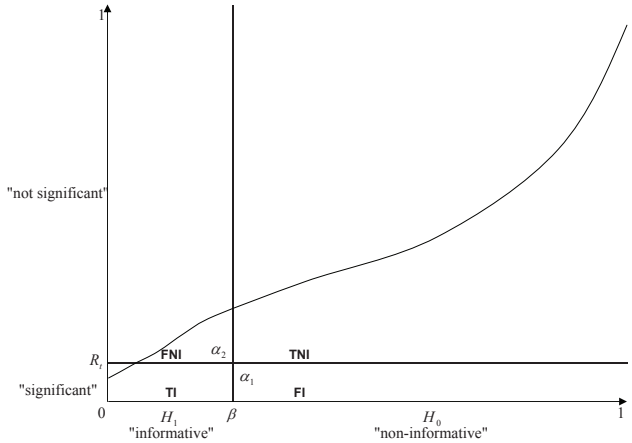

Figure 5: A two-component mixture model where the curve represents "p-values" and R_t is a cutoff point. Below this curve the null hypothesis is rejected. This implies a decision rule for type I (α_1) and type II (α_2) errors. The sum of the fractions false informative(FI), true informative (TI), false non-informative (FNI) and true non-informative (TNI), equals 1 [41].

6 Summary of papers

Paper I: Effect of data discretization on the classification accuracy in a high-dimensional framework

In Paper I classification methods that are less sensitive to high-dimensionality are considered. One of the disadvantages of these methods are that they are often computationally intensive. For continuous features and high-dimensional data the computation time becomes too demanding, hence the interest in discretization, i.e. transforming the continuous features into discrete features, with as few outcome categories as possible. For the standard setting $p < n$ it has been shown that discretization can improve classification.

The aim of this paper is to evaluate discretization and explore the effect of this procedure on the performance of high-dimensional classification. A second objective is, as discretization is a data-transformation procedure, to investigate how the structure of dependence between features is affected.

In the paper, a discretization algorithm is suggested that optimizes the discretization procedure using misclassification probability as a measure of classification accuracy. All the features are considered together in the algorithm and are discretized simultaneously instead of one at a time. This enables a fast discretization process which is suitable for high-dimensional data.

To examine the effect of discretization, classifiers that can handle both continuous and categorical data equivalently and are suitable for high-dimensional data are chosen. Further, for comparison several different discretization methods are applied. Then the classification accuracy and change in dependence structure are compared for both simulated and real data sets.

In classification accuracy the suggested method performs close to or better than classification for continuous features. In performance compared to classification based on other discretization methods the suggested method was always close to the best and never the worst. The imposed changes on the dependence structure varied greatly between different discretization methods. The suggested method caused major change in the dependence structure but the change was not significant compared to a baseline given by permutation.

In conclusion, the suggested method offers a highly effective discretization method.

Paper II: Covariance Structure Approximation via gLasso in High-Dimensional Supervised Classification

In Paper II we deal with constructing a sparse estimator of the inverse covariance matrix Σ^{-1} for supervised linear classification in high-dimensional settings. We propose a two-stage procedure for estimating an inverse covariance matrix. In the first step we identify the structural zeros of the inverse covariance through gLasso [14] in addition with bootstrapping to stabilize the non-zero elements. In the next step the non-zero elements are moved towards the main diagonal with Cuthill-McKee ordering [6]. These two steps enforce block sparsity and enable a block-diagonal approximation of the inverse covariance matrix.

The main advantage of the block-diagonal structure of Σ^{-1} is that it leads to a classifier that is a special case of the generalized additive model; see e.g. [21]. Further, the additive form of the classifier allows for block-wise feature selection.

To select a subset of blocks we need to know whether blocks are of importance for classification or not and for this we use block separation strength. In classification of real high-dimensional data, many of the blocks are likely to be "non-informative", i.e. only a small fraction of blocks actually contribute to the classification. Our goal is to select these blocks to include them into the classifier.

We simulate data with the given conditions in the lower and upper bounds to test classification accuracy for our classifier. We also compare our classification method to classification based on only gLasso. We can conclude that our proposed approach is relevant and beneficial in a high-dimensional setting, which we illustrate on both simulated and real data.

Paper III: Empirical evaluation of sparse classification boundaries and HC-feature thresholding on high-dimensional data

In Paper III the challenge of selecting a small subset of features that are likely to be informative for a specific project is further considered. This issue is crucial for success of supervised classification in very high-dimensional settings with sparsity patterns.

Here an asymptotic framework that represents the Sparse and Weak Block (SWB) model is derived and a technique for block-wise feature selection by thresholding is suggested, block Higher Criticism (bHC). The suggested procedure extends standard Higher Criticism (HC) thresholding [9] to the case where the dependence structure underlying the data can be taken into account.

The detection boundary of the bHC procedure and performance properties of some estimators of sparsity parameters are empirically investigated. Further, the bHC is shown to be optimally adaptive, i.e. it performs well without knowledge of the sparsity and weakness parameters. This property is of great importance, as the difficulties of obtaining reliable estimates of sparsity parameter are illustrated.

The relevance and benefits of the bHC approach are demonstrated in high-dimensional classification using both simulation and real data. As a conclusion, it can be stated that bHC outperforms other commonly employed selection methods. Taking the underlying dependence structure into account seems to improve classification accuracy for most of the real data sets. Though there are great differences in misclassification between set block-sizes, for further improvement in classification accuracy different block-sizes should be allowed, and the method can easily be extended to that.

Paper IV: Bayesian Block-Diagonal Predictive Classifier for Gaussian Data

In Paper IV we present a method for constructing a Bayesian predictive classifier in a high-dimensional setting. The predictive classifier provides an estimate of the probability of membership in each class for new observation, compared to many other classifiers that simply assign an observation to a class.

Also in this paper, we consider the block-structured covariance matrix, and then, given that classes are represented by Gaussian distributions, a closed form expression for the posterior predictive distribution of the data is established. Due to factorization of this distribution, the resulting Bayesian predictive and marginal classifier provides an efficient solution to the high-dimensional problem by splitting it into smaller tractable problems.

Further, in a simulation study we show that the suggested classifier outperforms several alternative algorithms such as linear discriminant analysis based on block-wise inverse covariance estimators and shrunken centroids regularize discriminant analysis.

7 Sammanfattning

Med dagens teknik, till exempel spektrometer och genchips, alstras data i stora mängder. Detta överflöd av data är inte bara till fördel utan orsakar även vissa problem, vanligtvis är antalet variabler (p) betydligt fler än antalet observation (n). Detta ger så kallat högdimensionella data vilket kräver nya statistiska metoder, då de traditionella metoderna är utvecklade för den omvända situationen ($p < n$). Dessutom är det vanligtvis väldigt få av alla dessa variabler som är relevanta för något givet projekt och styrkan på informationen hos de relevanta variablerna är ofta svag. Därav brukar denna typ av data benämnas som gles och svag (sparse and weak). Vanligtvis brukar identifiering av de relevanta variablerna liknas vid att hitta en nål i en höstack.

Denna avhandling tar upp tre olika sätt att klassificera i denna typ av högdimensionella data. Där klassificera innebär, att genom ha tillgång till ett dataset med både förklaringsvariabler och en utfallsvariabel, lära en funktion eller algoritm hur den skall kunna förutspå utfallsvariabeln baserat på endast förklaringsvariablerna. Den typ av riktiga data som används i avhandlingen är microarrays, det är cellprov som visar aktivitet hos generna i cellen. Målet med klassificeringen är att med hjälp av variationen i aktivitet hos de tusentals gener (förklaringsvariablerna) avgöra huruvida cellprovet kommer från cancervävnad eller normalvävnad (utfallsvariabeln).

Det finns klassificeringsmetoder som kan hantera högdimensionella data men dessa är ofta beräkningsintensiva, därav fungera de ofta bättre för diskreta data. Genom att transformera kontinuerliga variabler till diskreta (diskretisera) kan beräkningstiden reduceras och göra klassificeringen mer effektiv. I avhandlingen studeras huruvida av diskretisering påverkar klassificeringens prediceringsnoggrannhet och en mycket effektiv diskretiseringsmetod för högdimensionella data föreslås.

Linjära klassificeringsmetoder har fördelen att vara stabila. Nackdelen är att de kräver en inverterbar kovariansmatris och vilket kovariansmatrisen inte är för högdimensionella data. I avhandlingen föreslås ett sätt att skatta inversen för glesa kovariansmatriser med blockdiagonalmatris. Denna matris har dessutom fördelen att det leder till additiv klassificering vilket möjliggör att välja hela block av relevanta variabler. I avhandlingen presenteras även en metod för att identifiera och välja ut blocken.

Det finns också probabilistiska klassificeringsmetoder som har fördelen att ge sannolikheten att tillhöra vardera av de möjliga utfallen för en observation, inte som de flesta andra klassificeringsmetoder som bara predicerar utfallet. I avhandlingen förslås en sådan Bayesiansk metod, givet den blockdiagonala matrisen och normalfördelade utfallsklasser.

De i avhandlingen förslagna metodernas relevans och fördelar är visade genom att tillämpa dem på simulerade och riktiga högdimensionella data.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):pp. 289–300, 1995.

[2] E. Bradley. Local false discovery rates. Technical report, Department of Statistics, Stanford University, 2004.

[3] T.T. Cai, X.J. Jeng, and J. Jiashun. Optimal detection of heterogeneous and heteroscedastic mixtures. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(5):629–662, 2011.

[4] R. Coifman. Challenges in analysis. In N. Alon, J. Bourgain, A. Connes, M. Gromov, and V. Milman, editors, Visions in Mathematics, Modern Birkhauser Classics, pages 471–480. Birkhauser Basel, 2010.

[5] J. Corander, J. Xiong, Y. Cui, and T. Koski. Optimal viterbi bayesian predictive classification for data from finite alphabets. Journal of Statistical Planning and Inference, 143(2):261 – 275, 2013.

[6] E. Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. In Proceedings of the 1969 24th national conference, ACM '69, pages 157–172, New York, NY, USA, 1969. ACM.

[7] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. SIAM. J. Matrix Anal. & Appl., 30(56), 2008.

[8] D. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture delivered at the conference "Math Challenges of the 21st Century" held by the American Math. Society organised in Los Angeles, August 6-11, August 2000.

[9] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist, pages 962–994, 2004.

[10] D. Donoho and J. Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. Proceedings of the National Academy of Sciences, 105(39):14790–14795, 2008.

[11] D. Donoho and J. Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906):4449–4470, 2009.

[12] B. Efron, J.D. Storey, and R. Tibshirani. Microarrays, empirical bayes methods, and false discovery rates. Genet. Epidemiol, 23:70–86, 2001.

[13] J. Fan, Y. Fan, and Y. Wu. High dimensional classification. In T. T. Cai and X. Shen, editors, High-dimensional Data Analysis, pages 3–37. World Scientific, New Jersey, 2010.

[14] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441, 2008.

[15] S. Geisser. Predictive discrimination. In P.R. Krishnaiah, editor, Multivariate Analysis. Academic Press, New York, 1966.

[16] S. Geisser. Predicitive Infrence: An Introduction. Chapman and Hall, London, 1993.

[17] Alan George and Joseph W. Liu. Computer Solution of Large Sparse Positive Definite. Prentice Hall Professional Technical Reference, 1981.

[18] E. Georgii, L. Richter, U. Ruckert, and S. Kramer. Analyzing microarray data using quantitative association rules. Bioinformatics, 21 Suppl 2, September 2005.

[19] Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. Biostatistics, 8:86–100, 2007.

[20] A. Gupta, K.G. Mehrotra, and C. Mohan. A clustering-based discretization for supervised learning. Statistics and Probability Letters, 80:816–824, 2010.

[21] T. Hastie, R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning. Springer, second edition, July 2009.

[22] M. Hyodo, N. Shutoh, T Seo, and T. Pavlenko. Comparison of two high dimensional linear discrimination methods. Preprint, June 2011.

[23] M. Hyodo, N. Shutoh, T Seo, and T. Pavlenko. Modified estimator of the covariance matrix for high-dimensional data with monotone missing values. Preprint, June 2011.

[24] M. Hyodo and T. Yamada. Asymptotic properties of the epmc for modified linear discriminant analysis when sample size and dimension are both large. Journal of Statistical Planning and Inference, 140(9):2739 – 2748, 2010.

[25] D. Janssens, T. Brijs, K. Vanhoof, and G. Wets. Evaluating the performance of cost-based discretization versus entropy- and error-based discretization. Computers and Operations Research, 33(33):3107–3123, 2006.

[26] M. Kalisch and P. Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. Journal Of Computational And Graphical Statistics, 17:773–789, 2008.

[27] B. Klaus and K. Strimmer. Signal identification for rare and weak features: higher criticism or false discovery rates? Biostatistics, 14:129, 2013.

[28] V.A. Marchenko and L.A. Pastur. The distribution of eigenvalues in certain sets of random matrices. Math. USSR-Sbornik, 1:457–483, 1967.

[29] G.J. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. Wiley-Interscience, 2004.

[30] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. Annals Of Statistics, 34:1436–1462, 2006.

[31] N.F. Meinshausen. Analysis of High-Dimensional Data with Sparse Structure. PhD thesis, ETH Zurich, 2005.

[32] F. Murtagh, J-L. Starck, and M. W. Berry. Overcoming the curse of dimensionality in clustering by means of the wavelet transform. The Computer Journal, 43:107–120, 2000.

[33] T. Oates and D. Jensen. Large datasets lead to overly complex models: An explanation and a solution. In The fourth Internationel Conference on Knowledge Discovery and Data Mining, 1998.

[34] T. Pavlenko. Supervised classifications models in a high-dimensional framework. Department of Statistics, Stockholm University, November 2008.

[35] Y. Pawitan, J. Björhle, L. Amler, A.L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E.T. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P.M. Shaw, J. Smeds, L. Skoog, S. Wedren, and J. Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Research, 7(6):953–964, 2005.

[36] G. Potamias, L. Koumakis, and V. Moustakis. Gene selection via discretized gene-expression profiles and greedy feature-elimination. In George A. Vouros and Themistoklis Panayiotopoulos, editors, Methods and Applications of Artificial Intelligence, volume 3025 of Lecture Notes in Computer Science, pages 256–266. Springer Berlin / Heidelberg, 2004.

[37] B. D. Ripley. Pattern Recognition and Neural Networks. Cambridge, 1996.

[38] P. Rütimann and P. Bühlmann. High dimensional sparse covariance estimation via directed acyclic graphs. Electron. J. Statist., 3:1133–1160, 2009.

[39] D. Shenk. Data Smog: Surviving the Information Glut. HarperCollins, San Francisco, 1997.

[40] M. Srivastava and T. Kubokawa. Comparison of discrimination methods for high dimensional data. Journal of the Japan Statistical Society, 37:123Ũ–134, 2007.

[41] K. Strimmer. A unified approach to false discovery rate estimation. BMC Bioinformatics, 9(1):303, 2008.

[42] P. Utogoff. Incremental induction of decision trees. Machine Learning, 4:161–186, 1989.

[43] A.S. Wagaman and E. Levina. Discovering sparse covariance structures with the isomap. Journal of Computational and Graphical Statistics, 18(3):551–572, September 2009.

[44] M. West, C. Blanchette, H. Dressman, E. E Huang, S. Ishida, R. , Spang, H. Zuzan, J.A. Olson, J.R. Marks, and J.R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS, 98(20):11462–11467, September 2001.

[45] P. Xu, G.N. Brock, and R.S. Parrish. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. Computational Statistics & Data Analysis, 53(5):1674–1687, March 2009.

[46] Y. Yang and G.I. Webb. Discretization for naive-bayes learning: managing discretization bias and variance. Machine Learning, 74:39–74, 2009.