Guiding the Lasso: Regression in High Dimensions

Linn Cecilie Bergersen

Dissertation presented for the degree of
Philosophiae Doctor (PhD)

Department of Mathematics
University of Oslo
2013

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1353

ISSN 1501-7710

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Acknowledgement

I started my work on this PhD thesis in September 2009 funded by Statistics for Innovation $(sfi)^2$. The Graduate School in Biostatistics provided four months additional funding during the completion of the thesis.

First and foremost, I would like to express my deepest gratitude to my supervisor Professor Ingrid K. Glad who has followed me through all of my studies and inspired me from the very beginning. During the years I have been her student I have learned to know her as the most genuine and caring person who has always provided me with invaluable guidance and support. I am extremely grateful for everything she has taught me, both related to statistics, about being a researcher and personally.

I am also truly grateful to Professor Arnoldo Frigessi who is one of the co-authors in the second paper. He has enthusiastically followed all of the projects in this thesis and given helpful advices and suggestions along the way. I would like to thank him for encouraging me to believe in myself, and for being enthusiastic and positive every time I was not.

My sincere gratitude also goes to Professor Sylvia Richardson who I admire both as a scientist and as a person. She kindly welcomed me to Paris for three months during October-December 2011 and I am truly grateful for her hospitality and for how she generously dedicated her valuable time to me and our discussions. I would also like to thank the students and staff at Inserm UMRS937 at Faculté de Médecine Pitié-Salpêtrière in Paris for being so welcoming and for making the three months I spent there unforgettable.

I have also enjoyed the collaboration with my three other co-authors: Heidi Lyng, Ismaïl Ahmed and Kukatharmini Tharmaratnam. They have all shared their knowledge and greatly contributed to the papers in this thesis. I am especially happy to have met and learned to know Kuha who always has a happy attitude and who I enjoyed very much collaborating with on the third paper. Many thanks to my co-supervisor Professor Nils Lid Hjort for valuable comments when reading the final manuscript and preliminary versions of some of the papers. I would also like to thank my colleagues and fellow PhD students at the eight floor at the Department of Mathematics for providing such a pleasant working environment. Also, a special thanks to Gro for our many lunch and breakfast meetings at Blindern over the years. You know they have been very much needed!

The final big THANKS goes to all my friends and family for their support and encouragement, but most of all for being able to get my mind off the thesis when I needed it. I owe my deepest gratitude to my parents for their endless care and dedication, and for always being there for me when I have needed it.

Last but not least, I would like to thank Andreas for all the joy and happiness we have had together during our studies and for the support and patience he has shown in this final period (especially the last six months, three weeks and two days as you just reminded me...).

Blindern, March 26th 2013
Linn Cecilie Bergersen

List of Papers

Paper I

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Statistical Applications in Genetics and Molecular Biology*, 10(1)

Paper II

Bergersen, L. C., Ahmed, I., Frigessi, A., Glad, I. K., and Richardson, S. (2013a). Preselection in lasso-type problems by cross-validation freezing. *Submitted manuscript*

Paper III

Bergersen, L. C., Tharmaratnam, K., and Glad, I. K. (2013b). Monotone splines lasso. *Submitted manuscript*

Contents

1 Introduction

Major technological developments and facilities have revolutionized our society during the last couple of decades. Today, nearly *everything* can be measured, stored and analyzed because of the technological advances that have changed our ability to generate and store vast amounts of data. Sophisticated statistical methods can be the golden key to turning the overwhelming amounts of information into useful knowledge. The goal is to gain insight by identifying patterns and understand hidden, and often very complex, relationships. The potential value is tremendous. Combined, the availability of data, efficient algorithms and clever statistical methodology can help solving yet unanswered real world problems in areas like medicine, business and climate research.

High-dimensional data are often referred to as one of the most challenging topics to deal with in modern statistics (Donoho, 2000; Bickel et al., 2009; Johnstone and Titterington, 2009; Ferraty, 2010; Fan and Lv, 2010). The field of high-dimensional statistics covers a wide range of models aiming at different aspects of learning from data of high dimension. This includes supervised methods in regression and classification models, as well as unsupervised approaches for clustering, multiple testing or even graphical models (Bühlmann and van de Geer, 2011). Methods should take care of the important, sometimes critical, effects high dimensionality has on the statistical analysis and handle the large data sets through computationally efficient algorithms, as well as answering relevant questions in the specific area of application.

One of the main statistical challenges with high-dimensional data analysis is in regression where the number of predictors by far exceeds the number of observations. In these situations standard estimation techniques, such as the method of ordinary least squares, cannot be applied. Therefore, huge efforts have been made to develop sufficient statistical approaches and today a wealth of methods and techniques handling the high-dimensional regression problem exists, typically employing some kind of regularization, dimension reduction and/or screening.

The so-called lasso, proposed by Tibshirani in 1996, is by far one of the most popular methods. By construction, the lasso does not only fit the regression model, it simultaneously performs variable selection by putting some of the regression coefficients exactly to zero. In this sense, it is suitable for prediction and by producing a sparse solution it also extracts the most important variables and constructs models that are easy to interpret. In many applications, however, the underlying patterns are more complex than what is possible to model by a standard lasso regression model. For example, the effects of the covariates might derail from linearity, they might interact with each other or even with other measurable quantities outside the data. In many situations, the data might also be of such a high dimension that even well implemented and efficient algorithms are insufficient.

Hence, even if solving the dimensionality problem, the standard lasso might not be adequate to answer the real questions in practice. As a consequence, the standard lasso has been extended and modified to deal with more complex data situations that appear in high-dimensional data applications resulting in numerous new lasso-type methods. We consider these modifications as a way of *guiding* the lasso, and by retaining many of the desirable properties and advantages of the standard lasso, such a guiding makes room for an extensive and flexible framework for

sparse high-dimensional regression.

Many of the methods within such a framework are developed to answer complex questions in the context of genomics. Although appearing in various other areas of application such as text classification or protein mass spectrometry (Hastie et al., 2009), genomics is somewhat the mother lode of high-dimensional data. The amount of such data in genomics is massive and the problems often involve predictive models with several thousands explanatory variables (e.g. gene expressions) though limited to a small number of observations (individuals/samples). Understanding the biological basis of disease may also require more information than provided by one type of data alone (Hamid et al., 2009), or need statistical methods guided by assumptions arising from relevant (biological) knowledge.

This thesis addresses different ways of guiding the lasso, aiming specifically at three problems where the standard lasso meets its limitations. Although thought as general methods for high-dimensional regression to be applied in any field, the proposed methods are indeed motivated in the light of applications from genomics. Incorporating external (biological) knowledge or assumptions in the regression analysis can easily be seen as a way of guiding the lasso, and is one of the main objectives of the thesis. This can be considered in the context of data integration, but can also refer to situations where certain assumptions, for example on the functional shape on the estimated effects, are imposed. Another important problem, especially in genomics, is connected to the eternal increase in the dimensionality of the data. To do some kind of preselection or screening of covariates prior to fitting the lasso has been shown to be necessary (Fan and Lv, 2008; El Ghaoui et al., 2011; Tibshirani et al., 2012), and useful in settings where the dimension of the problem becomes too large to be easily handled by standard statistical software. To preselect variables is, however, not without risk and special attention is needed to avoid overfitting and preselection bias. We address these issues in the lasso setting and suggest a more safe approach to preselection which can also be considered as a way of guiding the lasso in ultra high dimensions.

The thesis is organized as follows: In Section 2 we introduce the general problem of regression with high-dimensional data, as well as pointing to more specific applications and challenges in genomics. We review the standard lasso in Section 3, which is the building block for what follows in Section 4 about guided lasso methods. The aims of the thesis are given in Section 5, before presenting the methodology used to achieve these aims in Section 6. In Section 7 summaries of Papers I-III are given. A final discussion of the results obtained, as well as topics for further research, are given in Section 8. Papers I-III follow at the end of this introductory part of the thesis.

2 High-dimensional Regression Problems

The problem of regression is that of relating one or more covariates to a dependent variable which we call the response. The interest is in exploring how the response varies with changes in any one of the explanatory variables. That is, how the response is influenced by the covariates. Given observations $(y_i, \boldsymbol{x}_i), i = 1, ..., n$, the aim is to build a model describing this relationship

through a regression function $f(\cdot)$ by assuming a model

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \tag{1}$$

where y_i is the response value for observation i, $\boldsymbol{x}_i$ is the corresponding covariate vector and the ϵ_i's are i.i.d. error terms with $E(\epsilon_i) = 0$. Written in this general form, the model is highly flexible and the regression function can take on any shape. By using suitable techniques to estimate $f(\cdot)$, the regression model can be used to understand patterns, quantify relationships in the data and identify which explanatory variables are specifically relevant for describing the response. Regression models are often used for prediction, where the objective is to predict the outcome in future data from a set of relevant variables. Simple examples can be to predict the price of a stock from a set of economic variables or whether a cancer patient is likely to experience a relapse of his disease based on clinical variables such as tumor characteristics.

Typically, $f(\cdot)$ in (1) is assumed to be a linear combination of the covariates and we have a linear model. That is, using matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{y}$ is the vector of responses, the covariates are organized in the $n \times P$ design matrix $\boldsymbol{X}$, $\boldsymbol{\epsilon}$ is the vector of error terms with $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\boldsymbol{\beta}$ is the vector of unknown parameters to be estimated. The estimation problem is usually solved through ordinary least squares (OLS) where the parameters are estimated by the values minimizing the residual sum of squares $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$. Provided $\boldsymbol{X}$ is of full rank, such that $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular and can be inverted, this gives $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

High-dimensional regression is, in a similar manner, concerned with relating a potentially very large number of covariates to a response of interest. By large we typically think of problems where the number of covariates P exceeds the number of observations n, that is $P > n$ or even $P >> n$. These problems are common in genomic applications which are described in Section 2.1, as well as for example in text classification problems where a large number of words may act as covariates to classify a text to be of a certain subject. Regression with $P \leq n$, and the regression idea itself, is an old topic that has been subject to comprehensive amounts of research and applied in all kinds of disciplines. The high-dimensional problems, on the other hand, have evolved during the last (couple of) decade(s), with new challenges and interesting aspects still arising both in statistical theory and from an application point of view.

From a statistician's perspective, high-dimensional regression problems are interesting because they cannot be solved by classical estimation procedures like the method of ordinary least squares. The standard procedures rely on the assumption that $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular, otherwise $\boldsymbol{X}^T\boldsymbol{X}$ cannot be inverted and the parameters cannot be uniquely estimated. This obviously does not hold when $P > n$, as the covariate matrix does not have full column rank. There are no other differences in the model than the fact that $P > n$, but this highly influences the estimation problem. Thus to cope with regression when $P >> n$, some kind of preselection or regularization is needed. There are a number of both simple and more advanced methods available, successful to various extents. The most intuitive approach is maybe through preselection,

that is, to simply pick out a smaller subset of the covariates ($\leq n$) based on a certain relevant criterion and fit the (standard) model to these covariates only. This is, however, dangerous as it may exclude relevant variables and traditional ideas like best subset selection become computationally too expensive in high dimensions (Fan and Lv, 2010). Another approach is to use methods like principal components regression or partial least squares. These methods derive a small number of linear combinations of the original explanatory variables, and use these as covariates instead of the original variables. This may be reasonable for prediction purposes, but models are often difficult to interpret (Hastie et al., 2009).

The focus in this thesis is on a third regularization approach that has shown to be successful in handling high-dimensional data, that is, penalized regression methods. Penalized regression methods shrink the regression coefficients toward zero, introducing some bias to reduce variability. Shrinkage is done by imposing a size constraint on the parameters and the problem is often expressed by adding a penalty to the residual sum of squares,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sum_{j=1}^{P} J_\lambda(|\beta_j|) \right\}. \tag{3}$$

The penalty $J_\lambda(|\beta_j|)$ depends on a tuning parameter λ that controls the amount of shrinkage, and can take on various forms, typically involving $\lambda|\beta_j|^r$ and some proper value of r distinguishing different methods. Among the most famous is ridge regression (Hoerl and Kennard, 1970) with a penalty $J_\lambda(|\beta_j|) = \lambda|\beta_j|^r$ with $r = 2$. The lasso (Tibshirani, 1996), which is discussed in detail in Section 3 and which is the main building block of methods proposed in this thesis, is defined by putting $r = 1$.

All regularization methods depend on one or more tuning parameters controlling the model complexity, that is, the number of variables preselected in subset selection, the number of derived inputs to use in principal components regression or the amount of shrinkage in shrinkage methods. Choosing the tuning parameters is an important part of the model fitting. If aiming at prediction, the tuning parameters should find the right balance between bias and variance to minimize prediction error. Methods for choosing the tuning parameters are further described in Section 3.2 and Section 6.3.

2.1 High-dimensional Regression Problems in Genomics

Much of the research on high-dimensional regression has been related to applications in genomics and molecular biology. The objective of such studies is often to improve the understanding of human diseases such as cancer, and to identify suitable biomarkers. In medicine, biomarkers are used to indicate the severity or presence of a disease and can help to give an early diagnosis. They are also used to predict the effect of treatment or to choose the appropriate treatment for the patient (Simon, 2011). Genetic biomarkers can typically be a group of genes or sequences of DNA that are associated with a certain disease and relevant for prognosis. The discovery of genetic biomarkers can enable treatment that is tailored for the specific patient. For example, in cancer patients the target is often to quantify how aggressive the cancer is to be

able to assign the proper treatment.

With this objective in mind, one of the primary interests in analyzing genomic data, is to relate genomic measurements (e.g. gene expression) to a phenotype of interest, for example time to relapse of a disease or subtypes of a disease. When technological advances made it possible to make simultaneous measurements of thousands of genes, suitable statistical methods and tools which could cope with high-dimensional regression models became essential. The high-dimensionality of these problems is apparent; while the number of genomic measurements can be very large, typically tens of thousands or even more, the number of samples is often very limited.

Aiming at finding genomic features and relationships that can be used as prognostic indicators, regression models should capture the most important information in the current data, as well as being useful for prediction. The major concern is the problem of overfitting as the high dimensionality makes it possible to construct models that fit the current data perfectly, but are useless for prediction purposes (Bickel et al., 2009). Validation and proper tuning of the model is therefore crucial as findings can only be considered as potential biomarkers if supported in independent studies. To avoid overfitting to the data at hand, regression models used for discovery of important genetic components influencing disease should be tuned for prediction. It is also believed that only a small part of the genomic data plays a role in disease mechanisms (Bickel et al., 2009; Bühlmann and van de Geer, 2011). To take this into account and to ease interpretation, it makes sense to do some kind of variable selection to extract the genomic features that are the most relevant.

Not only have the amounts of genomic data increased during the last decades. The size of the data has grown and different types and structures of data have evolved (Hamid et al., 2009). As it becomes more common to have different kinds of genomic data available for the same study, the interest is no longer limited to understanding the relationships within one type of measurement and its association with a phenotype or response of interest, but also between the molecular entities that drive the biological processes. Incorporating biological knowledge and relationships in the statistical models may lead to deeper understanding and is believed to be of great importance and promise (Bickel et al., 2009). Gene expression data have traditionally constituted the covariates in high-dimensional regression analyses in genomics. The expression of a gene is the first process by which mRNA, and eventually protein, is synthesized from the DNA (Lee, 2004). Copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) are other types of data produced by high-throughput technologies in genomics. These distinct data types capture different and complementary information about the genome. To provide a unified view of the whole genome, data integration becomes an essential part of the modeling to capture more information than is provided by considering only one single type of data. Also, results are more likely to be reliable if they are confirmed in multiple sources of data (Hamid et al., 2009).

To have a concrete example in mind, consider for example gene expressions and copy number data. Genetic gains and losses regulate the expression level of genes and are considered as motive forces of disease development (Lando et al., 2009). Not all overexpressed genes are amplified, and not all amplified genes are highly expressed, but the genes that are both

highly expressed and amplified are interesting and considered as potential driving forces for disease development (Albertson, 2006). Studying correlation between expression and copy number is therefore often considered as relevant when combining the two types of data in statistical analyses. Various studies have aimed at integrating gene expression with copy number to identify disease causing genes (Pollack et al., 2002; Lando et al., 2009; Solvang et al., 2011; Fontanillo et al., 2012). For example, in a study of cervix cancer in Lando et al. (2009) we combined data on gene dosage alterations with expression profiles of the same tumors. The study revealed genes that are regulated primarily by the genetic events and hence to be considered as candidate driver genes that represent novel biomarkers in cervix cancer. Combined information from the two data sets strengthens the evidence for the new biomarkers really being regulated by recurrent and predictive gene dosage alterations. Relevant biological information may also enter the modeling in another form. For example Li et al. (2006); Tai and Pan (2007); Pan (2009) and Pan et al. (2010) consider information about known functional relations or pathways from biological literature or databases such as Gene Ontology (GO, `http://www.geneontology.org/`) and Kyoto Encyclopedia of Genes and Genomes (KEGG, `http://www.genome.jp/kegg/`). In general, if different sources of data point to the same findings, they are less likely to be false positives (Holzinger and Ritchie, 2012).

Rapid technological advances not only lead to many different types of data; the size of the data is also increasing. While gene expression data typically measure the expression level for about 20-30,000 genes simultaneously, the number of SNPs measured can be more than three million in genome wide association studies (Lai, 2001). Handling 20-30,000 covariates in a regression model is no longer a problem from a technical/computational perspective, but regression models with 1-3 millions of covariates obviously meet practical challenges. Also, there is an increased interest in gene-gene and gene-environment interactions as these are believed to play a crucial role in more complex diseases (Liang and Kelemen, 2008; Wu et al., 2009; Cantor et al., 2010). Even if reduced to pairwise interactions the potential number of covariates becomes rapidly prohibitive. There are $P(P-1)/2$ possible first-order interactions and with the large P occurring in these applications this will present an extensive computational challenge. When higher order interactions are considered, the problems become even more severe (Shah, 2012). Such ultra high-dimensional data sets call for preselection methods to reduce the number of covariates prior to the analysis to extend the applicability of high-dimensional regression models also to these settings.

3 Standard Lasso for Linear Models

The lasso was proposed by Tibshirani in 1996 as a new method for estimation in linear models. Inspired by the work of Breiman (1995) on the nonnegative garotte and wishing to improve upon unsatisfactory properties of the ordinary least squares (OLS) estimates, he introduced regression with a L_1 penalty. The method was not intended for high-dimensional problems, which at the time had not yet emerged as a hot topic in the statistical community (Tibshirani, 2011). It was, however, at the time when large data problems began to evolve, mostly in genomics, that the lasso started to receive more attention. The L_1 penalty appeared to have desirable properties that

could be exploited with great benefit in high-dimensional regression problems, and it is in the $P >> n$ problems that the lasso-type methods have really proven their superiority compared to other existing methods. Today, the methods of the lasso-type are by far the most popular group of methods solving regression problems when $P >> n$. In this section, we describe the lasso pointing especially to why it has become such an appreciated tool for regression in high-dimensional data.

Assuming the linear model in (2), the lasso estimator $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\operatorname{argmin}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\}, \tag{4}$$

where λ is a tuning parameter controlling the amount of shrinkage. We call the penalty of this form a L_1 penalty. In addition to shrinking the coefficients toward zero, the L_1 penalty has the advantageous property of doing variable selection. In this way the lasso performs a kind of continuous subset selection. Indeed the lasso was introduced to combine the favorable properties of both subset selection and ridge regression, and was not really intended on high-dimensional regression situations. Tibshirani pointed to the fact that the ordinary least squares estimates often had low bias, but could suffer from high variance. This could affect the prediction accuracy of the model. At the same time he wanted to construct more interpretable models by determining a smaller subset of the covariates that exhibited the strongest effects. While ridge regression improves upon possible inefficiencies in terms of prediction capability through shrinkage, subset selection provides interpretable models, though unstable. By using the L_1 penalty, Tibshirani was able to retain the good features of both ridge regression and subset selection (Tibshirani, 1996).

To understand in more detail how the lasso leads some regression coefficients to be exactly equal to zero and how the lasso and ridge penalties differ, note first that (4) is equivalent to minimizing the residual sum of squares with a size constraint of the form $\sum_{j=1}^{P} |\beta_j| \leq s$ on the parameters. Similarly for ridge regression, the residual sum of squares is minimized under a size constraint $\sum_{j=1}^{P} \beta_j^2 \leq s$. Here s is a tuning parameter that has a one-to-one correspondence with the penalty parameter λ.

For both the lasso and ridge regression, and in fact all penalized regression methods having similar size constraints, s controls the amount of shrinkage imposed on the estimates. By the form of the size constraint $\sum_{j=1}^{P} |\beta_j|^r \leq s$, smaller values of s correspond to more shrinkage, forcing the estimates toward zero. For the lasso, smaller values of s will shrink all coefficients, but in addition put some of them exactly equal to zero. This is a direct consequence of using the L_1 norm in the constraint. Since the lasso constraint is not differentiable at zero, the lasso has the ability of producing estimates that are exactly equal to zero. The ridge constraint, on the other hand, does not share this property as having $r > 1$ gives constraints that are differentiable at zero (Hastie et al., 2009). That is, the difference really lies in the shape of the constraint region. To illustrate this, we consider the simple situation with only two parameters in Figure 1. The figure shows the estimation picture for the lasso and ridge regression. The elliptical contour lines represent the residual sum of squares centered at the OLS estimate, while the shaded regions represent the constraint region for the lasso and ridge regression respectively. In

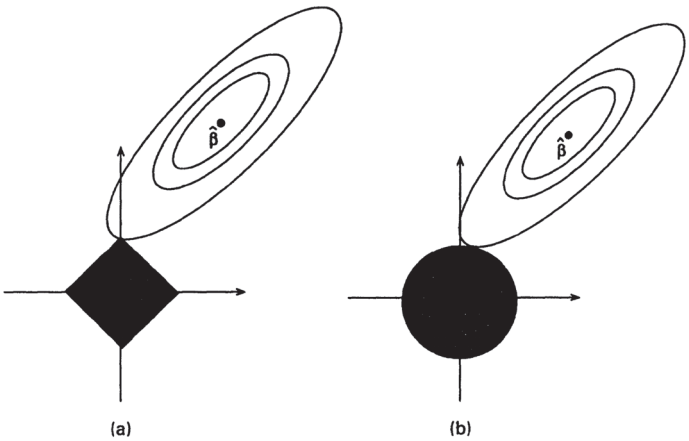


Figure 1: Illustration of estimation picture in the (a) lasso and (b) ridge regression. The figure is from the original paper of Tibshirani (1996).

both cases, the solution is at the first point where the elliptical contour lines of the residual sum of squares hit the constraint region. The important advantage of the lasso is that because of the diamond shape, it is more likely that the first time the elliptical contour lines hit the constraint region is in the corner, hence one of the parameters is estimated to be exactly zero. In higher dimensions the constraint region will have many corners and flat edges causing even more estimates to be zero (Hastie et al., 2009). Since the size of the constraint region is controlled by s, taking s small enough will force coefficients to be exactly zero. For ridge regression there are no sharp edges making it less likely for the contour lines to hit a corner. Hence estimated regression coefficients exactly equal to zero will rarely occur.

For a simple situation with only two estimated parameters, the example given in Figure 1 illustrates how the lasso constraint leads to variable selection. We may also gain further insights into the lasso if we consider the orthonormal case where (4) has an explicit solution in terms of the unrestricted estimators $\hat{\beta}_j^{OLS}$. That is, the lasso estimator of β_j corresponds to a soft-thresholded version of $\hat{\beta}_j^{OLS}$, whereas the ridge regression estimator is subject to proportional shrinkage (Hastie et al., 2009). Figure 2 shows how the threshold effect in the lasso results in estimates of β_j exactly equal to zero, compared to ridge regression and the unrestricted estimator.

Up to this point, we have only considered the linear model introduced in (2), but the lasso also extends naturally to generalized linear models (Tibshirani, 1996, 1997). For generalized linear models, we apply the same L_1 penalty, but the residual sum of squares is substituted by the relevant negative (partial) log-likelihood, such that the estimation is done by minimizing a penalized version of the negative log-likelihood. The properties of the lasso in the generalized linear models are very similar to those of the linear model (Bühlmann and van de Geer, 2011).

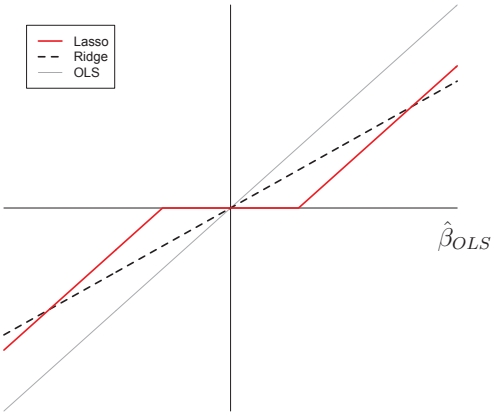


Figure 2: Estimators in the case of orthonormal design matrix $\boldsymbol{X}$. The grey line corresponding to an unrestricted estimate is added as a reference. The red line illustrates how the lasso puts small coefficients exactly to zero, while ridge regression performs proportional shrinkage.

When the lasso puts regression coefficients to zero, we say that it is producing a *sparse* solution. That is, only a few of the regression coefficients are estimated to be nonzero. This means that using the lasso there is an underlying assumption about sparsity; we assume that there are only a few of the covariates that are actually explaining the response. It is exactly this sparsity assumption that makes the lasso such a successful tool in high-dimensional regression analysis. Not only is sparsity a consequence of using the L_1 constraint and an important theoretical aspect to reduce the complexity and the number of effective parameters in the model, there are also intuitive as well as practical and computational reasons to assume sparsity in high-dimensional regression. The intention of producing more interpretable models is especially fruitful in the high-dimensional context. It is obviously easier and more convenient to interpret results from a lasso fit rather than a result involving estimated coefficients for all P covariates. In Section 2.1, we also discussed that in genomic applications we often assume from an application point of view that there is only a small set of the genes that are actually relevant for explaining the response. This is often the case in other types of problems as well, for example in text classification there is no reason to believe that all words in a text are important to classify it to be of a certain subject.

In standard regression models, the set of covariates is typically composed by a few variables that are well chosen and believed to be relevant and contributing to the model. The difference between the traditional setting and the high-dimensional problems is that the number of potential covariates is much larger, but more importantly, we do not know which of the covariates that might be relevant. In this sense, the fact that the lasso does variable selection makes it extremely attractive in determining the relevant covariates exhibiting the strongest effects. In fact, all constraints of the form $\sum_{j=1}^{P} |\beta_j|^r$ with $r \leq 1$ perform variable selection, but the lasso is the only constraint that has the advantage of producing a sparse solution while at the same time being convex. This also makes it an attractive method for computational reasons as non-convex

constraints make the optimization much more difficult (Hastie et al., 2009; Bühlmann and van de Geer, 2011).

3.1 Computational Algorithms

There is no closed form expression for the estimates in the lasso solution. The optimization problem becomes that of a convex problem with inequality constraints that are typically solved through quadratic programming (Friedman et al., 2007). Since the lasso is most frequently used in the presence of large data sets, computations can become extremely heavy if not efficiently implemented, hence much research has focused on computational efficiency.

Algorithms like the homotopy algorithm (Osborne et al., 2000) and the LARS algorithm (Efron et al., 2004) exploit the piecewise linearity of the coefficient paths yielding efficient algorithms that can solve the lasso problem for all values of λ. For generalized linear models the solution paths are in general not piecewise linear (Hastie et al., 2009). Hence Park and Hastie (2007) proposed another path algorithm solving the problem for generalized linear models which determine the entire coefficient path through a predictor-corrector method.

Another approach which is simple and well-suited for optimization in large convex problems, is the pathwise coordinate descent algorithm, which for the lasso problem has proven to be a strong competitor to the LARS algorithm (Friedman et al., 2010, 2007). Different from the exact path-following algorithms like LARS, the pathwise coordinate descent methods compute the regularized solution path for a fixed grid of λ values. For fixed λ, coordinate descent algorithms optimize successively over each parameter, that is, the optimization is done for one single parameter at a time. By considering the optimization problem as a sequence of single parameter problems that are easily solved by applying a soft-threshold operator, this is an attractive approach because each coordinate minimization can be done quickly and relevant updates are done by cycling through the variables until convergence (Friedman et al., 2007, 2010). To obtain the solution for the full grid of λ values, the procedure applies coordinate-wise descent for each value of the regularization parameter, varying the regularization parameter down a path.

3.2 Selection of Penalty Parameter

The lasso is, similarly to other penalized regression methods, depending on a tuning parameter λ controlling the model complexity. We know that different values of λ will influence how many variables that are selected by the lasso as well as the bias imposed on the estimated coefficients. It is therefore important to make a well deliberated choice of λ. There are several possible ways to choose the tuning parameter, all of them involving fitting the model for a range of λ-values. The final model is chosen from the set of candidate models based on some suitable criterion. Which criterion to use depends on the aim of the analysis.

Model selection in general is often done by estimating the performance of different models using criteria like the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Similar approaches can be used to choose the tuning parameter in the lasso if the focus

is primarily on recovering the true underlying set of active variables (Bühlmann and van de Geer, 2011). Another approach aiming at variable screening is to choose the λ corresponding to a predefined number of nonzero regression coefficients in the fitted model. This is relevant in situations such as in Wu et al. (2009) and El Ghaoui et al. (2011) where one has prior knowledge or strong reasons to anticipate how many variables that are really active. As an alternative, one can use the lasso as a screening method by considering the union of the variables selected for the entire range of λ values, that is, without selecting one specific value for λ at all (Bühlmann and van de Geer, 2011). Recently, stability selection based on subsampling was proposed to determine the right amount of regularization. In this case, the data are perturbed by subsampling many times before selecting variables that occur in a large fraction of the resulting selected sets (Meinshausen and Bühlmann, 2010).

None of these approaches are considering prediction performance. As discussed in Section 2.1, prediction is often a central part in the application of regression models. Therefore, maybe the most popular way to choose λ, is through K-fold cross-validation which involves minimizing an estimate of the prediction error. This is done by first splitting the data into K folds, typically $K = 10$. Leaving one fold out at a time, the remaining data are used to fit the model before computing the prediction error for the left out fold. The estimate $CV(\lambda)$ of prediction error is then obtained by aggregating over all folds and the model minimizing $CV(\lambda)$ is considered as the final model. In this case, the final model is tuned to be optimal for prediction, avoiding overfitting to the current data.

3.3 Theoretical Properties

There has been rigorous research devoted to understanding the theoretical properties of the lasso. The literature is extensive and it is not by any means possible to cover everything in detail in this thesis. A short summary of the key properties will be given in this section, reviewing the most central properties and relevant conditions for the standard lasso in the linear model. The results are obtained from Bühlmann and van de Geer (2011) where this is thoroughly presented, and which can be consulted for a comprehensive overview and further references. Specifically, the necessary assumptions and conditions referred to in this section can be found in Chapters 2, 6 and 7 of the book.

Consider a linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon},$$

with fixed design and with $\boldsymbol{\beta}^0$ being some true parameter vector. We also allow for the dimension $P = P_n >> n$ as $n \to \infty$. Let $S_0 = \{j : \beta_j^0 \neq 0, j = 1, ..., P\}$ be the active set of variables. Under no conditions on the design matrix or the non-zero coefficients, and rather mild conditions on the error,

$$||\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)||_2^2/n = O_P(||\boldsymbol{\beta}^0||_1 \sqrt{\log(P)/n}).$$

That is, the lasso is consistent for prediction if a sparsity assumption $||\boldsymbol{\beta}^0|| << \sqrt{n/\log(P)}$ is fulfilled. Optimal rates of convergence for prediction and estimation are obtained under certain

assumptions on the design. Under a compatibility or restricted eigenvalue condition, we achieve

$$||\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)||_2^2/n = O_P(s_0\phi^{-2}\log(P)/n),$$

where $s_0 = |S_0|$ and ϕ^2 is the compatibility constant or restricted eigenvalue depending on the compatibility between the design and the L_1-norm of the regression coefficients. Different from the prediction accuracy is the estimation accuracy of the parameter $\boldsymbol{\beta}$. Under the same compatibility assumptions on the design matrix $\boldsymbol{X}$ and on the sparsity s_0, it follows that

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0||_q = O_P(s_0^{1/q}\phi^{-2}\sqrt{\log(P)/n}), \quad q \in \{1, 2\}.$$

Since the lasso is also a method for variable selection, its ability to recover the true model is essential. For any method doing variable selection, the procedure should find an estimate $\hat{S} = \{j : \hat{\beta}_j \neq 0, j = 1, ...P\}$ such that $\hat{S} = S_0$ with high probability. There are, however, difficulties. For example, very small coefficients can be difficult to detect, and on the other hand, the lasso also tends to select too many variables, not being able to avoid false positives. Hence, consistency in variable selection requires the rather restrictive irrepresentable conditions on the design matrix as well as assumptions on the regression coefficients. First, note that the lasso has the variable screening property

$$\mathbf{P}(S_0 \subseteq \hat{S}) \to 1 \quad (P \geq n \to \infty),$$

in the sense that the selected variables from the lasso include the relevant variables with high probability, that is, we have $S_0 \subseteq \hat{S}$ with high probability. This follows under the restricted eigenvalue assumption and the so-called "beta-min" conditions

$$\inf_{j \in S_0^c} |\beta_j^0| >> \phi^{-2}\sqrt{s_0 \log(P)/n},$$

which require that the non-zero coefficients are not too small. Consistency for variable selection

$$\mathbf{P}(S_0 = \hat{S}) \to 1 \quad (P \geq n \to \infty)$$

on the other hand, requires in addition either a neighborhood stability condition for $\boldsymbol{X}$ or the equivalent irrepresentable condition. These are quite restrictive assumptions in practice, and can often fail to hold if the design matrix exhibits too strong correlations.

4 Guiding the Lasso

Since the cardinal paper of Tibshirani in 1996, the lasso penalty has been considered as a powerful and convenient method to handle the high-dimensional (generalized) linear regression problem. The method does, however, have drawbacks and cannot be considered as a universal approach. For example, it is limited to linear effects of the covariates and is not designed to cope with parameter vectors carrying certain (group) structure. In addition, the theoretical variable

selection properties require rather restrictive conditions on the design matrix, which are often not fulfilled in practice. Therefore substantial research and interest in these problems have led to new lasso-type methods being suggested and applied to a broad range of applications and problems with huge success. By retaining the desirable features of the L_1 penalty, these methods make up a flexible framework with extensive possibilities reaching far beyond the standard linear model.

Relevant but often simple modifications to the standard lasso model have expanded the area of use to problems of nonparametric regression, incorporation of information on the sparsity structure of the parameter vector, as well as improving theoretical properties of the standard method. We will call these kinds of modifications by the common term *guide* as they can all be viewed as a way of guiding the lasso toward more stable, relevant or tailored analyses and results.

A guide can either work directly on the model assumptions, enter as prior information in the model or influence the way the lasso estimates are obtained by limiting the lasso search to a smaller subset of the data. In the current section we will give an overview of some of the most relevant methods that have been proposed to solve problems that cannot be solved through the standard lasso model, hence working as lasso guides. We limit the section to methods of regression only, though substantial contributions have also been made using the L_1 penalty in graphical modeling (Yuan and Lin, 2007; Friedman et al., 2008; Banerjee et al., 2008). Many of the methods mentioned in this section also have equivalent counterparts appearing in the context of generlized linear models, but may not be listed here.

4.1 Improving Theoretical Properties

As pointed out in Section 3.3, the traditional lasso estimator may not be fully efficient in variable selection and may not be consistent. Several approaches have been suggested to guide the standard lasso toward consistency.

One of the main reasons for the lasso not to be consistent is the common amount of shrinkage that is imposed on the coefficients. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty which penalizes similarly to the lasso for small coefficients, but reduces the shrinkage for large coefficients. Hence the penalty produces less biased estimates and in fact SCAD possesses the oracle property which makes the method favorable in theory. However, the optimization criterion is not convex, which makes the computations more complicated and the method more difficult to apply in practice. The adaptive lasso (Zou, 2006) is a two-step procedure which was proposed to improve upon the theoretical properties of the standard lasso. In the adaptive lasso, the standard lasso penalty is replaced by a penalty that is weighted by the size of an initial estimator of the coefficients. When $P < n$, Zou (2006) suggested to use the ordinary least squares to obtain the initial estimator, such that when the initial estimator is a consistent estimator, the adaptive lasso is able to identify the true model consistently and the final estimator performs as well as the oracle (Zou, 2006). For the high-dimensional case, the lasso estimator itself can be used as an initial estimator (Bühlmann and van de Geer, 2011). The intention is to penalize large coefficients less, based on the initial estimator. Similar as for the

SCAD penalty, the result is less biased estimates and fewer false positives. A third approach to reduce the bias of the lasso is also a two-step procedure. The relaxed lasso was suggested by Meinshausen (2007) and involves fitting a standard lasso to obtain the set of nonzero coefficients, before fitting the lasso over again using only the variables in the nonzero set.

These modifications of the lasso are not changing the scope of the lasso per se, rather improving the properties of the standard method. All three methods can be viewed as ways to guide the lasso toward better theoretical properties by simple modifications in the penalty.

4.2 Linear Grouped Effects

In genomic applications there are often strong correlations among the covariates as they tend to operate in molecular pathways (Hastie et al., 2009). In this case it can be reasonable to consider them as being jointly relevant in the regression model, either by allowing for joint selection of correlated variables or of predefined groups that are assumed to act together.

The elastic net was proposed by Zou and Hastie (2005), and is especially designed to handle situations where there are strong correlations among the covariates. In blocks of correlated variables the lasso tends to pick one variable at random, discarding all other variables in that block. By adding a ridge penalty to the lasso penalty, the elastic net combines the advantages of both methods. The ridge penalty shrinks the coefficients of correlated features toward each other while the lasso penalty ensures sparsity among the averaged features. Another method proposed by Tibshirani et al. (2005) is the fused lasso which can take the ordering of the covariates into account by encouraging sparsity both of the coefficients and their differences.

Yuan and Lin (2006) developed the group lasso which is intended for situations where the predictors belong to predefined groups. The groups can consist of variables which for some reason are assumed to affect the response in a grouped manner, for example by belonging to the same pathways. Through the group lasso penalty, sparsity is encouraged at the group level such that the coefficients in a group should be either all zero or all nonzero. That is, covariates belonging to the same group are shrinked and selected together. The idea of introducing an adaptive step as discussed in Section 4.1 can also be applied for the group lasso to achieve better selection of groups (Bühlmann and van de Geer, 2011). Friedman et al. (2010) also suggest a sparse group lasso where sparsity can be achieved both at the group and at the individual feature level.

Stronger assumptions can also be imposed on the group structure. Chiquet et al. (2012) for example propose the cooperative lasso which does not only assume that groups should be selected jointly, but also that coefficients corresponding to variables of the same group are sign-coherent. That is, variables within a group influence the response in the same direction.

4.3 Nonlinear Effects

The lasso is designed to select linear effects, but meets limitations when the real effects derail from linearity. The methods discussed in Section 4.2 are utilizing relationships in the data or

incorporating information on known grouping structure of the covariates, but the selection is still limited to linear effects. The group lasso is, however, a highly flexible approach which can be used when the linearity assumption does not apply. Already in the initial paper, Yuan and Lin (2006) discussed how the group lasso can be applied when the explanatory variables are categorical. By redefining them as dummy variables and letting the dummy variables representing each covariate indicate the groups, a covariate is selected if its corresponding group of dummy variables is selected.

Much efforts have also been done to extend the high-dimensional regression methods to apply in high-dimensional nonparametric regression. Huang et al. (2010), Meier et al. (2009) and Ravikumar et al. (2009) all suggest to use the group lasso penalty in combination with splines to fit high-dimensional additive models. The covariates are represented through their spline basis expansions where the basis functions representing a covariate correspond to a group in the group lasso. In this way they are extending the linear model to allow for nonlinear effects in the individual components. By using B-splines and the adaptive group lasso, Huang et al. (2010) achieves consistency in both estimation and variable selection, while Meier et al. (2009) use a sparsity-smoothness penalty to control both sparsity and smoothness. Avalos et al. (2007) also proposed a method allowing for parsimonious solutions by the use of the L_1 penalty. Methods using lasso-type penalties or other penalization methods in partially linear models have also been subject to much research in recent years (Wang et al., 2011; Lian et al., 2012; Du et al., 2012). Typically this involves some spline representation for the nonparametric part and separate types of penalties for the parametric and nonparametric components.

A somewhat different approach in high-dimensional additive models is the Lasso Isotone (LISO) proposed by Fang and Meinshausen (2012). LISO fits an additive isotonic model where the component effects are assumed to be isotonic increasing. By introducing an adaptive step, the method also applies in more general situations where the direction of the functional components is not known. In this case the functions can be estimated to be either increasing or decreasing in the same model, but nevertheless the results are given as step functions.

4.4 Efficiency and Feasibility

Another type of problem that becomes hard and sometimes even impossible to solve by using the lasso and standard algorithms, is the problem where the data are of ultra high dimensionality. When the number of covariates becomes very large, standard algorithms become inefficient. Recent research by El Ghaoui et al. (2011) and Tibshirani et al. (2012) is devoted to this topic, proposing rules that discard variables which are not relevant for the regression for given values of λ. While the SAFE rule of El Ghaoui et al. (2011) is really safe, meaning that none of the variables that are active in the full lasso solution are eliminated by the rule, the STRONG rule of Tibshirani et al. (2012) is not guaranteed to be safe, but can achieve substantial improvements in terms of computational time. The rules operate by comparing the marginal correlation between the covariates and the response with certain criteria depending on λ. By construction the rules are able to discard a large proportion of the variables for large λ, but as λ decreases their standard rules are not as efficient and most variables are retained in the model fitting. El Ghaoui

et al. (2011) and Tibshirani et al. (2012) also provide sequential rules for which the elimination is done sequentially when moving down the λ scale. When implemented in combination with a lasso algorithm, especially the sequential STRONG rule is extremely beneficial and limits the lasso search to a small proportion of the data in each step.

Fan and Lv (2008) also consider methods for reducing the dimension from large to moderate based on correlation ranking and sure independence screening (SIS). Their aim is, however, somewhat different from the STRONG and SAFE rules which aim at finding the exact lasso solution. SIS does not guarantee that the screening does not exclude variables for which the coefficients would really be nonzero in the full solution with all covariates.

5 Aims of the Thesis

The lasso as presented in Section 3 solves the high-dimensional regression problem by selecting variables showing a linear effect on the response. Efficient algorithms exist, theoretical properties have been studied and are well deliberated, and the lasso has been widely appreciated in applied research (Kooperberg et al., 2010; Kohannim et al., 2012; Sveen et al., 2012). In more complicated problems the lasso has limitations and in Section 4 we addressed the fact that modifications and extensions of the standard lasso approach are necessary in order to expand the scope to include a broader range of applications. There exists a wide range of lasso-type methods, each of them guiding the lasso toward more stable or relevant results.

The main aim of this thesis is to take care of three specific problems that cannot be solved efficiently by the standard lasso itself. This includes problems where external (biological) knowledge or assumptions are reasonable to incorporate in the analysis. In Section 2.1 we discussed the need for methods taking this into account and that it might lead to deeper understanding, and elucidate potential casual mechanisms. With this in mind, we consider situations where external information enters the model by acting on the penalization scheme, either as a way of doing data integration or including prior information about the covariates in the model in order to tilt the analysis in a certain direction.

Another important problem is connected to the perpetual increase in the dimensionality of data. To cope with ultra high dimensionality, safer methods for preselection could facilitate computations and make it possible to analyze data that are so large that they exceed the feasibility limits in available algorithms and software. We address the issues that might arise when doing preselection by proposing an algorithm that focuses the lasso on a smaller and manageable set of relevant covariates. The aim is to make it possible to find the lasso solution in regression problems where the number of covariates is so large that we are not able to easily fit the full regression using all covariates.

Methods fitting nonparametric additive models through B-splines in combination with certain (group) lasso penalties were described in Section 4.3. These methods are very flexible, while in some situations it can be reasonable to assume certain shape restrictions on the functional components of each covariate. We propose a way of guiding the lasso where the aim is to retain

the monotonicity of the linear model, but allowing for nonlinearities in the estimated monotone effects.

Paper I is concerned with combining information from different types of genomic data in a weighted lasso. In Paper II we propose a strategy that enables analysis in ultra high-dimensional lasso problems that cannot easily be solved using standard procedures and software. Finally, Paper III deals with estimation and selection of nonlinear monotone effects. We approach the aim of solving these problems by methods that can be seen as ways of guiding the lasso.

6 Methodology

In this section, the methodology used to achieve the aims in Section 5 is described. First, in Section 6.1, the relevant penalization methods are described in the linear regression setting. This involves the standard lasso, the weighted lasso and the cooperative lasso. How penalized regression methods can be applied in generalized linear models is considered in Section 6.2. All these methods require a strategy for selecting the penalty parameter λ, thus the concept of cross-validation is described in Section 6.3. Finally, Section 6.4 considers monotone I-splines.

6.1 Penalization Methods

The current section considers the three lasso-type methods that are used to develop the proposed guided methods. Since the standard lasso is discussed in Section 3, it will only be repeated shortly in Section 6.1.1. Section 6.1.2 describes the weighted lasso with general weights, before the cooperative lasso is described in Section 6.1.3.

Suppose that we have data $\{y_i, \boldsymbol{x}_i\}$, $i = 1, ..., n$, where y_i is the response value and $\boldsymbol{x}_i = (x_{i1}, ..., x_{iP})$ is the vector of covariate measurements for observation i. Without loss of generality we assume that the intercept is zero and that all covariates $\boldsymbol{x}_j = (x_{1j}, ..., x_{nj})^T$, $j = 1, ..., P$, are centered and measured on the same scale. Let the covariates be organized in an $n \times P$ design matrix $\boldsymbol{X}$ and denote the response vector of length n by $\boldsymbol{y}$. We consider the linear model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is the P-dimensional parameter vector and the components ϵ_i of $\boldsymbol{\epsilon}$ are i.i.d. error terms with $E(\epsilon_i) = 0$.

6.1.1 Lasso

The lasso penalizes the regression coefficients by their L_1 norm. Hence the lasso estimates of the regression coefficients are given as

$$\hat{\boldsymbol{\beta}}^L = \underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\operatorname{argmin}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\}, \tag{5}$$

where $\lambda > 0$ is a penalty parameter controlling the amount of shrinkage. As mentioned in Section 3, there exist several algorithms to fit the lasso. In Paper II it is important that the lasso solutions are obtained for a fixed grid of λ and we use the coordinate descent algorithm as described in Section 3.1.

6.1.2 Weighted Lasso

In some settings, one might want to penalize the regression coefficients individually. This leads to the weighted lasso. That is, instead of a common penalty parameter λ, we consider a different penalty parameter $\lambda_j = \lambda w_j$ for each covariate such that each regression coefficient is penalized individually depending on the nonnegative generic weight w_j. The weighted lasso estimates can then be found by

$$\hat{\boldsymbol{\beta}}^{WL} = \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{P} w_j |\beta_j| \right\}. \tag{6}$$

We use the weighted lasso in Paper I, where weights are determined from external data, but the weighted lasso has previously been used in other contexts using weights chosen adaptively from the data (Zou, 2006; Bühlmann and van de Geer, 2011). The optimization can be done using any standard lasso algorithm, by using a simple reparametrization trick. That is, we rescale the covariates such that $\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j / w_j$ and $\tilde{\beta}_j = w_j \beta_j$, for $j = 1, ..., P$. Then we take $\tilde{\boldsymbol{x}}_j$ as covariates in the lasso algorithm to obtain estimates $\hat{\tilde{\beta}}_j$ and the weighted lasso estimates are found by transforming back, such that $\hat{\beta}_j^{WL} = \hat{\tilde{\beta}}_j / w_j$ for all j.

6.1.3 Cooperative Lasso

The cooperative lasso is a group penalty proposed by Chiquet et al. (2012), assuming that regression coefficients corresponding to variables in the same group are sign-coherent. Let $\{\mathcal{G}_k\}_{k=1}^K$ indicate the predefined groups which are mutually exclusive. The cooperative lasso penalty is then based on the group lasso norm

$$||\mathrm{v}||_{group} = \sum_{k=1}^{K} w_k ||\mathrm{v}_{\mathcal{G}_k}||,$$

where $||\cdot||$ is the Euclidean norm and $w_k > 0$ are fixed weights for each covariate that are used to adapt the amount of penalty in each group. Then the cooperative lasso estimates are defined as

$$\hat{\boldsymbol{\beta}}^{CL} = \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_{coop} \right\}, \tag{7}$$

where $\lambda \geq 0$ determines the amount of shrinkage and

$$||\mathrm{v}||_{coop} = ||\mathrm{v}^+||_{group} + ||\mathrm{v}^-||_{group} = \sum_{k=1}^{K} w_k \left(||\mathrm{v}^+_{\mathcal{G}_k}|| + ||\mathrm{v}^-_{\mathcal{G}_k}|| \right),$$

is the cooperative lasso norm with $\mathrm{v}^+ = (v_1^+, ..., v_p^+)^T$ and $\mathrm{v}^- = (v_1^-, ..., v_p^-)^T$ being the componentwise positive and negative part of v, that is, $v_j^+ = \max(0, v_j)$ and $v_j^- = \max(0, -v_j)$.

In Paper III, we use the cooperative lasso to ensure sign-coherence in the I-splines representation of the covariates. To fit the cooperative lasso in Paper III, we use the R package `scoop` available at `http://stat.genopole.cnrs.fr/logiciels/scoop`.

6.2 Penalization in Generalized Linear Models

Penalized regression also applies in the context of generalized linear models. It is not necessary to make an extensive review of generalized linear models here, but we stress that the solution can be obtained similarly as for the linear case by simply adding the desired penalty $J_\lambda(|\beta_j|)$ to the relevant negative log-likelihood;

$$\hat{\boldsymbol{\beta}}^{GLM} = \underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\operatorname{argmin}} \left\{ -l(\boldsymbol{\beta}) + \sum_{j=1}^{P} J_\lambda(|\beta_j|) \right\}. \tag{8}$$

In the experiments performed in this thesis, we have made use of the logistic regression model and the Cox proportional hazards model (Cox, 1972), in situations with binary and survival responses respectively.

6.2.1 Logistic Regression

For the special case of logistic regression, with binary response $y_i \in \{0, 1\}$, $i = 1, ..., n$, we have the logistic regression model

$$P(y = 1|\boldsymbol{x}) = p(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x})}.$$

If $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{iP})^T$ is the vector of covariates for the ith observation and $p_i = P(y_i = 1|\boldsymbol{x}_i)$, then the lasso estimate of the coefficient vector $\boldsymbol{\beta}$ is obtained by minimizing the penalized negative log-likelihood in (8) where $l(\boldsymbol{\beta})$ is replaced by the logistic regression log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\}. \tag{9}$$

Here the parameter vector $\boldsymbol{\beta}$ also contains an intercept.

6.2.2 Cox Regression

Suppose we have observations $(y_i, \boldsymbol{x}_i, \delta_i), i = 1, ..., n$, where $\delta_i \in \{0, 1\}$ is the censoring indicator and y_i is the survival time for observation i which is completely observed if the $\delta_i = 1$ and with corresponding covariate vector $\boldsymbol{x}_i = (x_{i1}, ..., x_{iP})^T$. Let $t_1 < ... < t_n$ denote the times

when events are observed assuming no tied events. In Cox proportional hazards models (Cox, 1972) the hazard function at time t given covariate vector $\boldsymbol{x}$ corresponds to

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}),$$

where $h_0(t)$ is the baseline hazard and $\boldsymbol{\beta}$ is the vector of unknown parameters. When $P \leq n$ the regression coefficients can be estimated by maximizing the partial log-likelihood given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}^T\boldsymbol{x}_i - \log\left[\sum_{k\in R(t_i)} \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i) \right] \right\},$$

where $R(t_i)$ is the risk set at time t_i. When $P > n$, penalized estimates of the regression coefficients in the Cox proportional hazards model can be obtained by substituting $l(\boldsymbol{\beta})$ in (8) (Verweij and Van Houwelingen, 1994; Tibshirani, 1997).

6.3 K-fold Cross-validation

K-fold cross-validation is often used to decide the amount of shrinkage in penalized regression methods. By dividing the observations into K folds, leaving one fold out at a time, the model is fitted to the remaining data for all values of λ in a grid. Typically we use $K = 10$. The prediction errors when predicting for the left out folds are aggregated to obtain the cross-validation score $CV(\lambda)$. That is, for the linear model, if f_k is the set of indices of the samples in fold k,

$$CV(\lambda) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in f_k}(y_i - \hat{y}_i^{-k}(\lambda))^2, \tag{10}$$

where $\hat{y}_i^{-k}(\lambda)$ is the fitted predicted value for observation i when fold k involving observation i is left out of the estimation. Typically one uses the prediction mean-squared error as in (10), but other measures of loss can also be considered.

For the generalized linear models we may express the cross-validation criterion in terms of the log-likelihood functions $l(\boldsymbol{\beta})$. For example for logistic regression, a typical choice is the mean deviance which corresponds to minus twice the log-likelihood on the left-out data (Friedman et al., 2010). For Cox regression the criterion used in Verweij and van Houwelingen (1993) and Bøvelstad et al. (2007) can be used to choose the value of λ. That is, maximizing

$$CV(\lambda) = \sum_{k=1}^{K}\{l(\hat{\boldsymbol{\beta}}_{(-k)}(\lambda)) - l_{(-k)}(\hat{\boldsymbol{\beta}}_{(-k)}(\lambda))\},$$

where $l_{(-k)}(\boldsymbol{\beta})$ is the log-likelihood function, and $\hat{\boldsymbol{\beta}}_{(-k)}$ is the estimate of $\boldsymbol{\beta}$, when the kth fold is left out.

6.4 Monotone I-Splines

I-splines are a type of splines introduced by Ramsay (1988) which can be used to fit monotone regression functions. By construction, the basis functions are all monotone, such that if they are combined with nonnegative spline coefficients, the fitted regression function will be monotone increasing. Similarly, if the spline coefficients are all nonpositive, the fitted regression function will be monotone decreasing.

For an explanatory variable x, suppose that its values are transformed to $[0, 1]$ and define the knot sequence t by $0 = t_1 = ... = t_l < ... < t_{K+l+1} = ... = t_{K+2l} = 1$ where K is the number of interior knots. The I-splines basis functions are defined as integrated versions of M-splines (Ramsay, 1988). M-splines of order l are defined by the recursive formula

$$M_k^{(l)}(x) = \begin{cases} \frac{l[(x-t_k)M_k^{(l-1)}(x)+(t_{k+l}-x)M_{k+1}^{(l-1)}(x)]}{(l-1)(t_{k+l}-t_k)}, & t_k \leq x \leq t_{k+l}, \\ 0 \quad \text{otherwise}, \end{cases}$$

for $k = 1, ..., K + 1$ and with

$$M_k^{(1)}(x) = \begin{cases} \frac{1}{t_{k+1}-t_k}, & t_k \leq x \leq t_{k+1}, \\ 0 \quad \text{otherwise}. \end{cases}$$

Then the I-splines basis functions are found by integrating

$$I_k^{(l)}(x) = \int_{t_1}^{x} M_k^{(l)}(u)\, du,$$

for $x \in [0, 1]$ and $k = 1, ..., K + l$. Hence with a suitable set of basis functions $I_k^{(l)}(x)$, $k = 1, ..., K + l$, a monotone piecewise polynomial of order k associated with knot sequence t can be represented as the linear combination

$$g(x) = \sum_{k=1}^{K+l} \beta_k I_k^{(l)}(x),$$

if $\beta_k, k = 1, ..., K + l$, are of the same sign. That is, either all nonnegative or all nonpositive.

7 Summary of the Papers

7.1 Paper I

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Statistical Applications in Genetics and Molecular Biology*, 10(1)

In the first paper we propose to use a weighted lasso for data integration in high-dimensional regression models. Our method is intended for high-dimensional regression models where more

than one measurement are available for each covariate. We suggest to let the additional data enter the model not directly, but by acting on the penalization scheme. Hence our approach is guided by the available external information and tuned to obtain the optimal solution for prediction through cross-validation.

General weights for data integration are introduced, as well as weighting schemes specifically designed and applied to two relevant data problems. The first example is an analysis of a data set where both gene expression and copy number data are available. The example illustrates how the method can be used to combine two types of genetic measurements available from the same study. Gene expression measurements are used as covariates in a Cox regression model with a lasso penalty weighted by information depending on copy number data for the same group of patients. The second example illustrates how the weighted lasso can be used to incorporate external information that is not a part of the specific data at hand. Literature annotations are used to define weights in a logistic regression analysis with gene expressions as covariates to describe metastasis/no metastasis in a data set of head and neck cancer patients. The results are validated either on a new independent data set or in the biological literature.

Performance in terms of both prediction and variable selection is also evaluated in simulation studies, and compared with the standard lasso and the adaptive lasso. When the external information is relevant, we find that the weighted lasso with data integration improves upon variable selection and prediction, as well as reducing the bias of the estimated coefficients.

7.2 Paper II

Bergersen, L. C., Ahmed, I., Frigessi, A., Glad, I. K., and Richardson, S. (2013a). Preselection in lasso-type problems by cross-validation freezing. *Submitted manuscript*

In ultra high-dimensional regression problems the full solution is often computationally difficult to obtain and it is necessary to use some initial preselection or screening strategy. Paper II addresses this issue by suggesting a way to find the full solution of a sparse regression problem by using a smaller subset of the covariates only. The proposed algorithm works in combination with cross-validation to find the solution optimal for prediction and is guided by a property we call freezing to ensure that all relevant covariates are included in the subset.

The concept of cross-validation freezing is introduced and illustrated in the context of the standard lasso regression model. By sequentially comparing the cross-validation curves of increasing subsets of covariates, the proposed algorithm recovers freezing patterns which is used to determine whether the relevant covariates are included in the preselected subset or not, and hence when the preselected set is sufficient to obtain the full solution.

In several simulation and real data experiments, we demonstrate that we are able to find the full cross-validated lasso solution based on a smaller subset of the data. One of the examples is a GWAS data example where we are not able to fit the full regression problem using standard software, but by using freezing patterns and the proposed algorithm we are able to find the full solution.

7.3 Paper III

Bergersen, L. C., Tharmaratnam, K., and Glad, I. K. (2013b). Monotone splines lasso. *Submitted manuscript*

In the third paper, we turn to regression problems where the effects are assumed to be monotone, but not necessarily linear. We consider an additive model where the component effects, that we want to estimate, are unknown functions of each variable. Our aim is to introduce a flexible alternative to the standard linear model, by allowing for the effects to take nonlinear shapes, but still preserving monotonicity.

We introduce the monotone splines lasso (MS-lasso), which by combining I-splines and the cooperative lasso penalty, selects important (nonlinear) monotone effects. Each variable is represented by its I-spline basis and the set of basis functions for each variable constitute a group in a cooperative lasso procedure. The combination of I-splines and the cooperative lasso is essential. For the I-splines to produce monotone functions, the spline coefficients need to be of the same sign. This is guaranteed by the cooperative lasso which provides sign-coherence for coefficients within a group. We also introduce the adaptive MS-lasso, which similarly to other adaptive lasso procedures, is shown to reduce the number of false positive selections.

Important differences and similarities with other existing methods for high-dimensional regression are pointed out. We also compare the MS-lasso and the adaptive MS-lasso with these methods in various simulation experiments. The performance is evaluated in terms of estimation and variable selection and the results indicate that if nonlinearities are present in the component effects, there can be a lot to gain by applying the (adaptive) MS-lasso instead of standard models. For illustration, we also apply the procedure in two data examples where the monotonicity assumption is relevant and discuss the findings.

8 Discussion

Problems of high-dimensional regression continue to challenge statisticians. The size of the data keeps growing and the scope of the analyses is no longer limited to finding significant linear relations or to analysis of a single data type alone. Standard (penalized) regression methods are not designed to cope with these kinds of complexities, and the need for adequate methodology solving the new challenges will keep evolving with the new types of data.

The aim of this thesis was to provide methods dealing with challenges where the standard penalized methods do not apply. We have focused on three specific problems of this type: combining data from different sources (Paper I), regression in ultra high dimensions (Paper II) and nonparametric monotone regression (Paper III). In Paper I we allow for incorporation of additional information on the covariates, while in Paper III we add flexibility to allow for the estimated effects to take on nonlinear shapes. In both cases we are changing the objective and assumptions of the model. The strategy introduced in Paper II is in some way different from those presented in Paper I and III. This is in the sense that the aim of Paper II is not to modify

the objective of the standard lasso, but rather to extend its applicability to ultra high dimensions where it is computationally difficult or infeasible to obtain the solution without reducing the number of covariates. Hence, the proposed methods aim at different issues and aspects of the lasso, but by making use of the flexibility of the L_1 penalty, our contributions are all to be considered as part of the *guided lasso* framework which we deliberated over in Section 4.

The main contribution of Paper I is to illustrate how the weighted lasso can be used to combine information from different sources. We have applied the procedure to data from genomics, with the main example being the one where both gene expression and copy number data are available together with survival data in a study of cervix cancer. We are aware that there are numerous possible approaches that can be considered to extract combined information from several data sets, and which might elucidate and uncover underlying dynamics of disease more accurately than the weighted lasso approach. The proposed procedure can, however, be seen as a convenient tool to include prior information in the regression analysis and even to tilt the analysis in a certain direction. As discussed in Section 4.1 about the adaptive lasso, the weighted lasso has been widely studied in the literature when the weights are decided from initial (consistent) estimators of the parameters. The weights are differentiating the penalization imposed on the coefficients, such that if the initial estimator is large in absolute value, the corresponding coefficient is penalized less and hence encouraged in the selection. We use the same reasoning, but allow for the weights to be determined from some prior information obtained from external data. A variable is given an advantage in the regression if the external information indicates that the variable is relevant. Based on biological knowledge, for example that genes with high correlation between their copy number and expression value are assumed to be driving forces in disease development, we are giving certain genes an advantage in the regression.

In future work it would be interesting to apply the procedure to other types of data. In Section 2.1 we pointed out that in genomic applications, it becomes more and more common to have data from multiple sources available for the same study (e.g. gene expression, copy number data, SNPs, methylation). It is also possible to construct weights based on several different types of external data. This should, however, be considered in combination with biological knowledge about the relations between the different biological processes.

In Paper I, we have used a fully frequentistic point of view, although including prior information in this way makes it natural to take on a Bayesian perspective. The standard lasso itself has a Bayesian interpretation; it is the mode of the posterior distribution of the regression coefficients when having a Gaussian likelihood and that each parameter is independently distributed with a double exponential prior (Hans, 2009). The variance of the prior distribution has a one-to-one correspondence to λ. Hence introducing individual penalty parameters $\lambda_j = \lambda w_j$ in the lasso, is the same as assuming double exponential priors with variances depending also on the external information. If the weights indicate that a variable is relevant, the prior will have more mass distributed away from zero making it more likely for the estimated coefficient to be nonzero. However, to take advantage of the multiple types of data available and do data integration one may want to consider a fully Bayesian approach. See for example Jensen et al. (2007) who perform variable selection and integrate several data sets through a Bayesian hierarchical model.

In Paper II we document the presence of preselection bias when variables are preselected based

on their marginal correlation with the response. Although it is well known that preselection may have an important impact on the results of the analysis, screening based on univariate association with the response variable is often done in practice, without recognizing the potential undesirable effect. The main focus of Paper II is the proposed algorithm which can be used to obtain the full solution by using only a subset of the covariates. The pattern in the cross-validation curves which we call *freezing* is used to provide the value of λ which would have been chosen by cross-validation when using the full set of covariates, but more importantly, it indicates when a preselection can be considered as *safe*. Hence freezing acts as a guide for preselection in lasso-type problems.

Apparently there are several similarities between the SAFE and STRONG rules for variable elimination and our approach. All three methods are specifically intended for lasso-type problems and can achieve substantial reductions in the number of covariates by using marginal correlation with the outcome to indicate the relevance of the covariates. It is, however, important to note that there are also some essential differences. While SAFE and STRONG operate at fixed values of λ, our method aims at finding the full lasso solution in the optimal value of λ for prediction. Also, the focus of SAFE and STRONG is mostly on computational efficiency and time, while our focus is on feasibility of problems that become too computationally demanding or even infeasible because of the large number of covariates. That is, our strategy might not necessarily save computational time. The computational time of the complete procedure will depend on how the subsets are chosen, the correlation structure in the design matrix and how soon the cross-validation curves freeze which depends on which λ that turns out to be optimal for the full data set. It is also important to note that the success of the SAFE and STRONG rules, in terms of number of variables they are able to discard, becomes limited for smaller values of λ. That is, when the lasso solution is not very sparse. For example in the GWAS example in Paper II, the STRONG rule will not be able to discard any variables when using the amount of penalization indicated by cross-validation while we will only use 2.5% of the covariates.

As for Paper III, the main contribution is the introduction of the new method MS-lasso which is designed to estimate and select relevant nonparametric monotone effects in additive models. Importantly, the proposed method is not only relevant in high dimensions, but may also be useful if one wants to do variable selection and at the same time ensure monotonicity in lower dimensions.

For the computations in Paper III, we use the algorithm available in the R-package `scoop` developed by Chiquet et al. (2012). The algorithm provides the necessary sign-coherence for the spline coefficients in the MS-lasso, but it does not scale to the necessary level to be able to fit the MS-lasso if the number of covariates becomes too large. Note that in the MS-lasso where each covariate is represented by m basis functions, the number of inputs in the algorithm is really $m \times P$, which can become extremely expensive in problems with $P = 20,000 - 30,000$ variables. One possible approach to deal with the computational issues is to use cross-validation freezing and the algorithm proposed in Paper II. Even if the strategy in Paper II is developed in the standard setting of the linear model, the concept has promising potential also in other settings. Especially in already high-dimensional settings where the covariates are to be represented by basis functions. We give an illustration for the specific case of the MS-lasso.

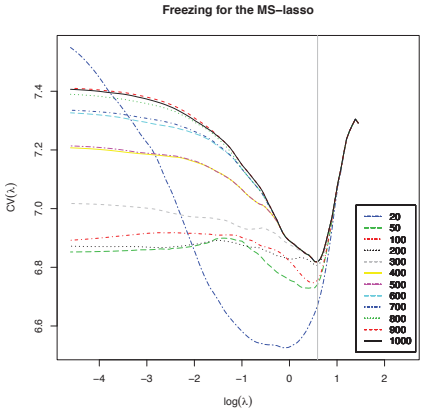


Figure 3: Illustrative example of cross-validation curves for the MS-lasso where the indicated subsets are determined from the initial ranking using Spearman correlation. The figure shows the results from a single simulation run from the experiments using Model A as described in Section 3.1 in Paper III.

Since the MS-lasso utilizes a group lasso type penalty, one should rank groups instead of single inputs in the algorithm. In our case, however, this is equivalent since each group represents one explanatory variable. Also, as the MS-lasso aims at finding nonlinear monotone effects, Spearman correlation (which is not limited to linear associations) between each covariate and the response could be a reasonable choice of initial ranking. An example is given in Figure 3 which indicates the characteristic freezing behavior, and that the procedure proposed in Paper II will extend naturally to the MS-lasso problem. Hence extending the algorithm in Paper II to nonparametric problems, as well as to other guided lasso methods, are interesting and relevant topics for future investigations.

As a final remark we point to another important problem that was briefly mentioned in the end of Section 2.1, and which has become a vibrant topic in genomics as well as in the statistical communities recently. That is, the problem of fitting models that include interaction effects between the covariates - in high dimensions. There is much recent work studying the interaction problem, typically involving selection of which interactions to include in the modeling. To mention a few examples, Bien et al. (2012) introduce a lasso for hierarchical interactions which imposes a hierarchy restriction such that an interaction is only included if its main effects are marginally important. Also, Shah (2012) propose a method that iteratively builds up sets of candidate interactions to include in the regression procedure. Hall and Xue (2012) discuss how to provide a proper ranking of the covariates and their pairwise interactions together. As a consequence of the increasing parameter space, it is common to all approaches to identify which interaction terms seem relevant such that only these are included in the model. Since the problem of fitting regression models with interactions is often considered challenging even in lower dimensions, limiting the number of interaction terms in the analysis through certain restrictions or preselection becomes an extremely essential part of the high-dimensional interaction problems. Extending the preselection strategy proposed in Paper II to interaction problems can possibly take the lasso to even higher dimensions in the future.

References

Albertson, D. (2006). Gene amplification in cancer. *Trends in genetics*, 22(8):447–455.

Avalos, M., Grandvalet, Y., and Ambroise, C. (2007). Parsimonious additive models. *Computational Statistics and Data Analysis*, 51:2851–2870.

Banerjee, O., El Ghaoui, L., and D'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Bergersen, L. C., Ahmed, I., Frigessi, A., Glad, I. K., and Richardson, S. (2013a). Preselection in lasso-type problems by cross-validation freezing. *Submitted manuscript*.

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Statistical Applications in Genetics and Molecular Biology*, 10(1).

Bergersen, L. C., Tharmaratnam, K., and Glad, I. K. (2013b). Monotone splines lasso. *Submitted manuscript*.

Bickel, P. J., Brown, J. B., Huang, H., and Li, Q. (2009). An overview of recent developments in genomics and associated statistical methods. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4313–4337.

Bien, J., Taylor, J., and Tibshirani, R. (2012). A lasso for hierarchical interactions. *ArXiv e-prints*. 1205.5050.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics*, 23(16):2080–2087.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):pp. 373–384.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6 – 22.

Chiquet, J., Grandvalet, Y., and Charbonnier, C. (2012). Sparsity with sign-coherent groups of variables via the cooperative-lasso. *Annals of Applied Statistics*, 6(2):795–830.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistics Society, Series B*, 34:187–220.

Donoho, D. L. (2000). Aide-Memoire. High-dimensional data analysis: The curses and blessings of dimensionality.

Du, P., Cheng, G., and Liang, H. (2012). Semiparametric regression models with additive nonparametric components and high dimensional parametric components. *Computational Statistics & Data Analysis*, 56(6):2006–2017.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

El Ghaoui, L., Viallon, V., and Rabbani, T. (2011). Safe feature elimination for the lasso and sparse supervised learning problems. *ArXiv e-prints*. 1009.4219.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148.

Fang, Z. and Meinshausen, N. (2012). LASSO Isotone for High-Dimensional Additive Isotonic Regression. *Journal of Computational and Graphical Statistics*, 21:72–91.

Ferraty, F. (2010). Editorial: High-dimensional data: a fascinating statistical challenge. *Journal of Multivariate Analysis*, 101(2):305–306.

Fontanillo, C., Aibar, S., Sanchez-Santos, J. M., and De Las Rivas, J. (2012). Combined analysis of genome-wide expression and copy number profiles to identify key altered genomic regions in cancer. *BMC Genomics*, 13(Suppl 5):S5.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *ArXiv e-prints*. 1001.0736.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Hall, P. and Xue, J.-H. (2012). On selecting interacting features from high-dimensional data. *Computational Statistics & Data Analysis*. http://dx.doi.org/10.1016/j.csda.2012.10.010.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. (2009). Data integration in genetics and genomics: Methods and challenges. *Human Genomics and Proteomics*, 1(1).

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition.

Hoerl, A. E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Holzinger, E. R. and Ritchie, M. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics*, 13(2):213–222.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38(4):2282–2313.

Jensen, S. T., Chen, G., and C. J. Stoeckert, J. (2007). Bayesian variable selection and data integration for biological regulatory networks. *The Annals of Applied Statistics*, 1(2):612–633.

Johnstone, I. M. and Titterington, D. M. (2009). *Philosophical Transactions. Series A: Mathematical, Physical, and Engineering Sciences*, 367(1906):4237–4253.

Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A., Jack Jr, C. R., Weiner, M. W., de Zubicaray, G. I., McMahon, K. L., Hansell, N. K., Martin, N. G., Wright, M. J., and Thompson, P. M. (2012). Discovery and replication of gene influences on brain structure using lasso regression. *Frontiers in Neuroscience*, 6(115).

Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology*, 34(7):643–52.

Lai, E. (2001). Application of SNP technologies in medicine: Lessons learned and future challenges. *Genome Research*, 11(6):927–929.

Lando, M., Holden, M., Bergersen, L. C., Svendsrud, D. H., Stokke, T., Sundfør, K., Glad, I. K., Kristensen, G. B., and Lyng, H. (2009). Gene dosage, expression, and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer. *PLOS Genetics*, 5(11):e1000719.

Lee, M.-L. T. (2004). *Analysis of microarray gene expression data*. Kluwer Academic Publishers.

Li, J., Li, X., Su, H., Chen, H., and Galbraith, D. W. (2006). A framework of integrating gene relations from heterogeneous data sources: an experiment on arabidopsis thaliana. *Bioinformatics*, 22(16):2037–2043.

Lian, H., Chen, X., and Yang, J.-Y. (2012). Identification of partially linear structure in additive models with an application to gene expression prediction from sequences. *Biometrics*, 68(2):437–445.

Liang, Y. and Kelemen, A. (2008). Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Survey*, 2:43–60.

Meier, L., Van De Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B*, 72:417–473.

Osborne, M., Presnell, B., and Turlach, B. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–403.

Pan, W. (2009). Network-based multiple locus linkage analysis of expression traits. *Bioinformatics*, 25(11):1390–1396.

Pan, W., Xie, B., and Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66(2):474–484.

Park, M. and Hastie, T. (2007). *glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model*. R package version 0.94.

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Børresen-Dale, A.-L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963–12968.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

Ravikumar, P., Lafferty, J., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71(5):1009–1030.

Shah, R. D. (2012). Modelling interactions in high-dimensional data with backtracking. *ArXiv e-prints*. 1208.1174v2.

Simon, R. (2011). Genomic biomarkers in predictive medicine. An interim analysis. *EMBO Molecular Medicine*, 3(8):429–435.

Solvang, H., Lingjaerde, O., Frigessi, A., Borresen-Dale, A.-L., and Kristensen, V. (2011). Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer. *BMC Bioinformatics*, 12(1):197.

Sveen, A., Agesen, T. H., Nesbakken, A., Meling, G. I., Rognum, T. O., Liestol, K., Skotheim, R. I., and Lothe, R. A. (2012). Cologuidepro: A prognostic 7-gene expression signature for stage III colorectal cancer patients. *Clinical Cancer Research*.

Tai, F. and Pan, W. (2007). Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics*, 23(14):1775–1782.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16:385–395.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society Series B*, 73(3):273–282.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108.

Verweij, P. J. M. and van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12:2305–2314.

Verweij, P. J. M. and Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, 13(23-24):2427–2436.

Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Annals of Statistics*, 39(4):1827–1851.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

I

Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 39

Weighted Lasso with Data Integration

Linn Cecilie Bergersen, *University of Oslo*
Ingrid K. Glad, *University of Oslo*
Heidi Lyng, *Norwegian Radium Hospital*

Recommended Citation:
Bergersen, Linn Cecilie; Glad, Ingrid K.; and Lyng, Heidi (2011) "Weighted Lasso with Data Integration," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 39.
DOI: 10.2202/1544-6115.1703

Weighted Lasso with Data Integration

Linn Cecilie Bergersen, Ingrid K. Glad, and Heidi Lyng

Abstract

The lasso is one of the most commonly used methods for high-dimensional regression, but can be unstable and lacks satisfactory asymptotic properties for variable selection. We propose to use weighted lasso with integrated relevant external information on the covariates to guide the selection towards more stable results. Weighting the penalties with external information gives each regression coefficient a covariate specific amount of penalization and can improve upon standard methods that do not use such information by borrowing knowledge from the external material. The method is applied to two cancer data sets, with gene expressions as covariates. We find interesting gene signatures, which we are able to validate. We discuss various ideas on how the weights should be defined and illustrate how different types of investigations can utilize our method exploiting different sources of external data. Through simulations, we show that our method outperforms the lasso and the adaptive lasso when the external information is from relevant to partly relevant, in terms of both variable selection and prediction.

KEYWORDS: adaptive lasso, cervix cancer, copy number alterations, data integration, gene expressions, head and neck cancer, Lasso, p>>n, penalized regression, prediction, variable selection, weighted lasso

Author Notes: Linn Cecilie Bergersen, Department of Mathematics, University of Oslo. Ingrid K. Glad, Department of Mathematics, University of Oslo. Heidi Lyng, Department of Radiation Biology, Norwegian Radium Hospital. This is a project of the centre Statistics for Innovation in Oslo. We thank F.C.P. Holstege and S. van Hooff for providing information on the head and neck cancer data.

1 Introduction

High throughput technologies in molecular biology allow to collect simultaneous information about thousands of individual characteristics, such as gene expressions, SNPs or proteins. Current genome wide association studies are easily based on a million SNPs per sample (Donnelly, 2008). The weak aspect of such studies is the insufficiently large sample size; studies today typically include around a hundred, sometimes a few thousand, individuals. The aims of genome wide studies can be several, for example to generate reliable classification or prediction rules for an outcome, say some time to event, based on a selection of genetic covariates. Such biomarkers can be used to predict outcome for future patients with the same medical conditions. Also, the selection per se of such covariates associated with the outcome, is of great interest, as it generates hypotheses for causal mechanisms. This situation leads to regression models (linear, logistic, Cox or others) where the number of covariates p by far exceeds the number of observations n, $p >> n$. Under such conditions, standard statistical theory breaks down.

The recent statistical literature is rich of exciting ideas and methods for handling such ultra high-dimensional models. Among the most popular are various penalization approaches, including the lasso (Tibshirani, 1996) and its many variations (Yuan and Lin, 2006, Zou, 2006, van de Geer et al., 2010, Meinshausen, 2007), the Dantzig (Candés and Tao, 2007), the SCAD (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005) and SIS (Fan and Lv, 2008). Efficient algorithms are now available for most of these methods, and theoretical studies have established that under various types of sparsity assumptions and regularity conditions on the design, many of these reach asymptotically reliable results, in terms of both prediction and variable selection. However, many problems remain in practice. Variable selection is highly unstable (Ein-Dor et al., 2005, Michiels et al., 2005, Meinshausen and Bühlmann, 2010), prediction rules are difficult to validate on new data (Chanock et al., 2007, McCarthy et al., 2008), known biological factors are not selected (Ioannidis, 2007). Low signal to noise ratios, technical noise in the covariates, inhomogeneous sample populations, and unmeasured confounders, are among the reasons for such difficulties.

In order to improve, strengthen and guide penalization based results, it is possible to incorporate more knowledge into the inferential exercise. In many situations, there is unexploited additional information available about which covariates are more likely to explain the outcome. For example, when the aim is to predict survival based on gene expressions, other genetic measurements for the same individuals might be available, like SNPs or copy number alterations. In cancer diseases, one can for instance expect that genes with increased copy number play a role in tumor progression and therefore are more likely to affect patient survival (Albert-

son, 2006). Moreover, lack of heritability of the baseline expression level of a gene (Morley et al., 2004) could be an indication that the gene is not involved in development of heritable diseases (Ferkingstad et al., 2008). In this paper we exploit the availability of several sources of relevant information to modulate the level of penalization, thus guiding the variable selection procedure. We propose a framework for genewise penalization, where the penalty parameter varies for each covariate and is modulated by external weights reflecting the expected relevance of the covariate (gene) for the outcome. Through two-dimensional K-fold cross-validation the method is data driven and the data decide the relative strength of the external weights, also allowing for no weights in case the additional source of information would turn out not to be really informative.

Other recent methods use relevant extra knowledge in the inference as well. In group lasso (Yuan and Lin, 2006, Ma et al., 2007), grouping of genes can be viewed as introducing additional information. Tai and Pan (2007) used prior grouping of genes in penalized classifiers and in penalized regression (Pan et al., 2010). Slawski et al. (2010) use prior structural information on genes to guide an elastic net regression procedure, while Percival et al. (2010) assume that nonzero covariates cluster together and incorporate this as an additional constraint. The work of Charbonnier et al. (2010) is related to our approach in that they use the weighted lasso formulation to incorporate structural information in inference for regulation networks from temporal data. Genes are assumed to belong to typically two classes of connectivity and penalized differently according to class membership. Also Xie et al. (2007) divide genes into two classes based on gene expression data, and then shrink the test-statistics for genomic location data only for the genes belonging to one of the two classes, enhancing power and reducing false discoveries. None of these approaches make however use of an additional dataset as we do.

Our method is one way to perform data integration, maintaining a hierarchical structure of the information: external data enters the model not directly but only by acting on the penalization scheme. This avoids a further increase in the number of covariates but allows to combine several measurements on the same genes in the analysis. The method also allows to use external information to bias the search in specific directions. For example, one might be interested in selecting genes whose effect on the outcome is complementary to, and thereby masked by, other known factors (Nowak and Tibshirani, 2008). The method can also be used to perform a joint analysis of independent datasets on the same disease, e.g. gene expressions measured on different patients in different labs. Ultimately, it is the appropriateness of the additional information that makes our approach advantageous. Bayesian interpretation of penalization schemes have been discussed (Park and Casella, 2008, Hans, 2009), as the penalization structure represents a prior model on the regression parameters. While we focus on a pure penalized likelihood setting, our method can

DOI: 10.2202/1544-6115.1703

easily be seen in this context. Data integration can be implemented into Bayesian inference through hierarchical models.

Section 2 gives a brief introduction to the general penalized likelihood methods with special emphasis on the lasso and the weighted lasso. The new methodology designed to incorporate external information is presented in Section 3. In Section 4 we illustrate our method on two cancer datasets with gene expressions as covariates in a Cox and a logistic regression setting using gene copy numbers and literature annotations as external information. We investigate the behaviour of our approach on simulated data in Section 5 and close with a discussion in Section 6.

2 Penalized Likelihood and the Lasso

Suppose we have data $(y_i, \mathbf{x}_i)$, where y_i is the response and $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ is the vector of p covariates, for $i = 1, ..., n$. Assume that the covariates are standardized. When $p > n$, classical statistical methods fail as there are infinitely many solutions to the estimating equations. It is necessary to add constraints to select the interesting solutions. Often, there are biological reasons to assume sparsity, that is, only few of the p covariates are actually associated to the outcome. Shrinkage and/or variable selection is then performed by maximizing a penalized version of the log-likelihood,

$$\frac{1}{n} l(\beta) - \sum_{j=1}^{p} J_\lambda(|\beta_j|)$$

with respect to the regression parameters β. Here $l(\beta)$ is the conditional (partial) log-likelihood of y_i given covariates $\mathbf{x}_i$ and $J_\lambda(|\beta_j|)$ is a penalty function controlling model complexity (Fan and Li, 2001, 2002). The penalty function is constructed so that the regression coefficients are shrunk towards zero or set equal to zero, resulting in a sparse solution. Various penalty functions $J_\lambda(|\beta_j|)$ have been proposed, see Fan and Lv (2010). Most common is to use a penalty of the form $\lambda \sum_{j=1}^{p} |\beta_j|^r$, where λ is the penalty parameter. The lasso corresponds to $r = 1$ and ridge regression (Hoerl and Kennard, 1970) to $r = 2$. Although the ideas presented in the following sections apply to other penalty functions as well, we concentrate on the lasso.

The lasso usually performs well for prediction, but for consistent variable selection the conditions on the design matrix are rather restrictive. These fail to be true for example in situations with strong correlations among covariates (van de Geer et al., 2010). Furthermore, the lasso does not possess oracle properties (Fan and Li, 2001). With oracle property we intend that the method can correctly select the nonzero coefficients with probability converging to one, and that the estimators for the nonzero coefficients are asymptotically unbiased and normally distributed.

The oracle property does not automatically imply optimal prediction performance, but for variable selection (finding the correct lists of genes) it is of course advantageous.

The lasso is sometimes presented as a special case of the weighted lasso (Zou, 2006, van de Geer et al., 2010, Grandvalet and Canu, 1998), where the penalty parameter λ is generalized to p values λ_j, such that each covariate is penalized individually, which could possibly improve the performance of the selection and estimation. We write $\lambda_j = \lambda w_j$ for generic non-negative weights w_j, so that

$$\frac{1}{n} l(\beta) - \lambda \sum_{j=1}^{p} w_j |\beta_j| \tag{1}$$

is maximized with respect to β. Note that by defining $\alpha_j = w_j \beta_j$, we can estimate the α_j's by standard lasso procedures substituting each entry in the data matrix with x_{ij}/w_j. Transforming back gives $\hat{\beta}_j = \hat{\alpha}_j / w_j$.

3 Weighted Lasso with Data Integration

In the weighted lasso setting, the weights are used to modulate the strength of the penalty of each covariate, based on what information we have from additional data. For a large value of w_j the regression coefficient for variable j is subject to a larger penalty and therefore is less likely of being included in the model, and vice versa.

Let the additional data be $\mathbf{Z}$, where $\mathbf{Z}$ is either a $m \times p$ matrix or a vector $\mathbf{z}$ of length p. We typically have $m = n$, but this is not necessary. In the most general form, we allow weights which capture specific relations between the additional data $\mathbf{Z}$, the response $\mathbf{y}$ and the covariates $\mathbf{X}$. Define these weights as $w_j(\mathbf{y}, \mathbf{X}, \mathbf{Z})$. These weights should be nonnegative and could take various forms depending on the data and question at hand. In our analyses we have found it useful to define weights as

$$w_j(\mathbf{y}, \mathbf{X}, \mathbf{Z}) = \frac{1}{|\eta_j(\mathbf{y}, \mathbf{X}, \mathbf{Z})|^q}, \tag{2}$$

for some function η_j, $j = 1, ..., p$. Here q is a parameter controlling the shape of the weight function. We assume that η_j increases in the expected relevance of the covariate. There are various possibilities for η_j. For example, η_j could depend only on elementwise external information z_j, or on the relation between $\mathbf{y}$ and $\mathbf{Z}$, or between $\mathbf{X}$ and $\mathbf{Z}$. To have a concrete example in mind, let $\mathbf{X}$ be a matrix of gene expressions, $\mathbf{Z}$ a matrix of gene copy numbers, and $\mathbf{y}$ a vector of right-censored survival times. For this example we will consider two important types of weights:

DOI: 10.2202/1544-6115.1703

W1 Spearman Correlation Coefficients We exploit the information in the correlation between gene expression and copy number: genes showing high correlation between expression and copy number are considered as possible driving forces for cancer progression (Albertson, 2006). We therefore use $\eta_j(\mathbf{x}_j, \mathbf{z}_j) = \hat{\rho}_j$ if $\hat{\rho}_j > 0$ and $\eta_j(\mathbf{x}_j, \mathbf{z}_j) = \{\min \hat{\rho} : \hat{\rho} > 0\}$ if $\hat{\rho}_j \leq 0$. Here $\hat{\rho}$ is the vector of Spearman correlation coefficients $\hat{\rho}_j$. Negative correlations are adjusted to the smallest positive observed correlation, since gene expressions which are negatively correlated with gene copy number express more complex dynamics. We also adjust when $\hat{\rho}_j = 0$ to avoid division by zero in w_j.

W2 Ridge Regression Coefficients Here we wish to exploit the association of gene copy number with survival. Genes whose changes in copy number explain survival, should be given less penalty than others. Since a copy number alteration influences survival by first affecting the expression level, the genes within these aberrated regions are more likely to explain survival through their expression as well. A quantity that captures the influence of copy number on survival, should be appropriate. We find such a quantity by fitting a Cox-ridge regression model to the copy number data to obtain estimates for the ridge regression coefficients γ_j. For gene j the weights are defined as $\eta_j(\mathbf{y}, \mathbf{Z}) = |\hat{\gamma}_j|$, $j = 1, ..., p$.

In (1), the penalty parameter λ controls the amount of shrinkage imposed on all the coefficients simultaneously, while the tuning parameter q in (2) controls the relative strength of the weights across all covariates. The q is real and positive and is not restricted to integers. In order to determine the pair of parameters (q, λ) most suited, we do full K-fold cross-validation with optimization on a two-dimensional grid.

We have used the cross-validation criterion of Verweij and van Houwelingen (1993) and Bøvelstad et al. (2007), but included the new parameter q so that the cross-validation criterion

$$CV(q, \lambda) = \sum_{k=1}^{K} \{l(\hat{\beta}_{(-k)}(q, \lambda)) - l_{(-k)}(\hat{\beta}_{(-k)}(q, \lambda))\}$$

is maximized leading to $(\hat{q}, \hat{\lambda})$. By including $q = 0$ in the grid, the procedure is allowed to choose the standard lasso in case the weights (or the additional data) are not informative. Note that if the weights depend on the relation between $\mathbf{y}$ and $\mathbf{X}$, the weights have to be recalculated inside each cross-validation step, as in Kramer et al. (2009).

Properties of the externally weighted lasso method will depend on the actual weights. Let β_k^{true}, $k = 1, ..., p$ be the true parameters in the regression model. The

general weighted lasso is shown to possess oracle properties if for the active set $A = \{k : \beta_k^{true} \neq 0\}$, all weights w_j, $j \in A$, are bounded, and all weights w_j, $j \notin A$, go to infinity as p and n grow (adapted from Huang et al. (2006)). Less strict conditions on the weights are possible, see Zou (2006), Huang et al. (2008) and van de Geer et al. (2010) for the precise theory. These conditions should be kept in mind when constructing the external weights, aiming at weights that give large enough penalty to the non active set and small enough penalty to the active set. As the bias (shrinkage towards zero) increases with the penalty parameter, smaller weights for the active set should also ensure less bias in the estimation. In Section 5 we present finite sample simulations when the conditions on the weights are fulfilled to various controlled degrees.

3.1 Relation to the Adaptive Lasso

Our approach has similarities to the adaptive lasso (Zou, 2006, Zhang and Lu, 2007, Huang et al., 2008, van de Geer et al., 2010) where they maximize

$$\frac{1}{n} l(\beta) - \lambda \sum_{j=1}^{p} \frac{1}{|\hat{\beta}_j|^{\gamma}} |\beta_j| \quad \text{for fixed } \gamma > 0,$$

where $\hat{\beta}_j$'s are OLS estimates when $p \leq n$ (Zou, 2006). When $p > n$, Huang et al. (2008) use estimated coefficients from p univariate regressions of each covariate on $\mathbf{y}$. A recent version of the adaptive lasso (van de Geer et al., 2010) apply the results from a standard lasso run as the $\hat{\beta}_j$'s. Hence the coefficients that were set to zero in a first lasso run, are left out in the second round, while the coefficients with largest $\hat{\beta}_j$ receive the smallest penalty and hence smaller bias. This two step procedure is shown to possess oracle properties under certain conditions on the design matrix (van de Geer et al., 2010) and is employed in several recent papers, for example in Kramer et al. (2009) and Benner et al. (2010). The adaptive lasso is an example of a weighted lasso. The difference with our approach is that in the adaptive lasso the weights are constructed from a preliminary analysis of the same data $\mathbf{y}$ and $\mathbf{X}$. In our approach there is data integration through the exploitation of additional data sets $\mathbf{Z}$.

4 Gene Signatures for Cervix and Head and Neck Cancer

We demonstrate our approach on two different data sets; cervix cancer and head and neck cancer. Gene expressions constitute the covariates $\mathbf{X}$ in both cases, while

DOI: 10.2202/1544-6115.1703

the response $\mathbf{y}$ is right-censored survival time or presence/absence of metastasis, respectively, calling for Cox proportional hazard and logistic versions of the weighted lasso. The additional data used for genewise weighting are: a matrix $\mathbf{Z}$ of array comparative genomic hybridization data (aCGH, gene copy numbers) for the cervix cancer analysis; a p-vector $\mathbf{z}$ of literature annotations from Pubgene (Jenssen et al., 2001) for the head and neck cancer analysis. Unless otherwise specified, we have applied the lasso implementation in the R package `glmpath` (Park and Hastie, 2007) with 10-fold cross-validation, for all methods, in all analyses and simulations.

4.1 Example 1: Cervix Cancer Data

We have two datasets containing survival data for patients diagnosed with cervix cancer. Clinical information and details regarding the aCGH and gene expression experiments are presented in Lando et al. (2009).

The first set of data contains 102 patients. We have survival data and cDNA microarray gene expression data for $n = 100$ of these patients and aCGH data for 97 of them. Both measurements are available for 95 of the patients. The genomic data contain measurements for $p = 7754$ genes with unique gene identification. The aCGH data measure genetic gains and losses, which may cause changes in the gene expression levels. These may disturb the primary function of the genes and lead to highly aggressive disease and poor clinical outcome (Albertson, 2006). With weighted lasso penalization we integrate the aCGH data as additional information on each gene, thus giving genes within aberrated regions a larger chance to be selected.

In addition we have a separate data set for validation of the prediction performance. The validation set is an independent data set containing survival data and gene expression measurements (but no aCGH measurements) of the same genes for 41 new patients. These gene expressions are obtained from Illumina gene expression beadarrays.

We considered two different weighting schemes of the form in Eq.(2). The quantity η_j in Eq.(2) will be either W1: the correlation between the gene copy number and gene expression (Spearman correlation coefficients), or W2: the effect on survival of the gene copy numbers (estimated by ridge regression coefficients), as described in Section 3. The correlation coefficients $\hat{\rho}$ in W1 were calculated from the vectors $\mathbf{x}_j$ and $\mathbf{z}_j$, $j = 1, ..., p$, for the 95 patients where both measurements were available. The ridge coefficients in W2 were calculated from the 97 patients with copy number and survival data available.

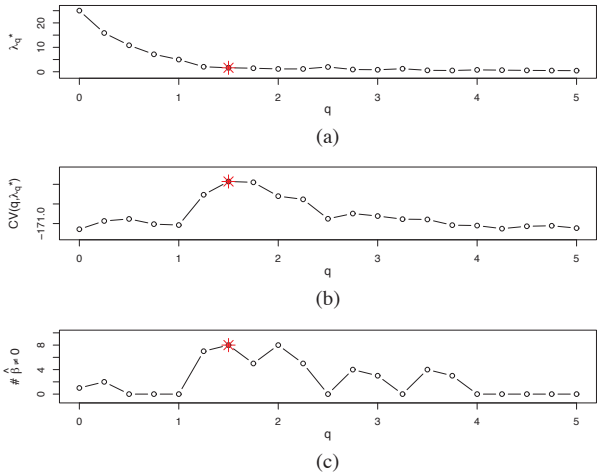


Figure 1: Cross-validation results for W1: Spearman correlation coefficients. The optimal value of q is marked by the red star. (a) Optimal λ values for various q. (b) Cross-validation curve as function of q, with optimal values λ_q^* inserted for λ. (c) Number of nonzero estimated regression coefficients for various q and corresponding optimal λ.

The weighted lasso estimates are then obtained by maximizing the penalized version of the Cox partial log-likelihood, as in Eq.(1), using two-dimensional 10-fold cross-validation.

4.1.1 Results

The results are summarized in Figure 1 and 2 for W1 and W2 respectively. In Figure 1(a) and 2(a) the optimal values of λ, $\lambda_q^* = \arg\max_\lambda CV(q,\lambda)$, are plotted versus q. As q increases the optimal value of λ decreases, leaving more of the penalization to the weights. The cross-validation curve as a function of q for given λ_q^* is plotted in Figure 1(b) and 2(b). Maximizing $CV(q,\lambda)$ over the two-dimensional grid gives the pair $(\hat{q},\hat{\lambda})$, where $\hat{\lambda} = \lambda_{\hat{q}}^*$, is used to fit the final model. We found $(\hat{q},\hat{\lambda}) = (1.500, 1.625)$ for W1, and $(\hat{q},\hat{\lambda}) = (2.750, 0.003)$ for W2. In both situations, the cross-validation prefers to include the external information through the weighted penalties instead of fitting a standard lasso model $(q = 0)$. Figure 1(c) and 2(c) show the number of nonzero estimated coefficients for each q (with $\lambda = \lambda_q^*$). In the final model 8 genes are selected for W1 and 21 genes for W2. The lasso surprisingly

DOI: 10.2202/1544-6115.1703

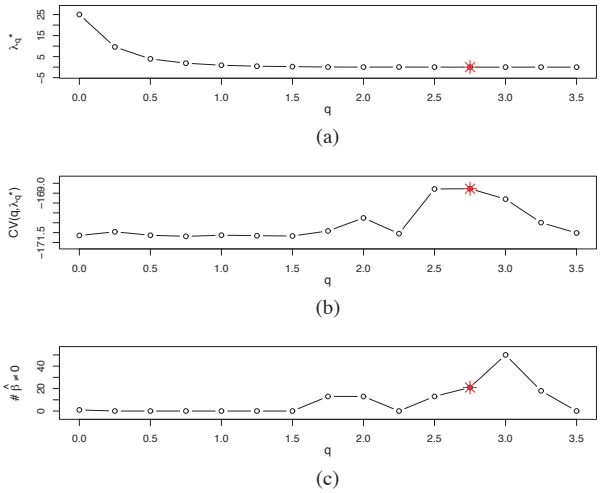


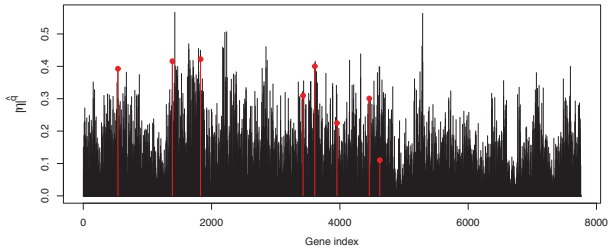
Figure 2: Cross-validation results for W2: Ridge regression coefficients. The optimal value of q is marked by the red star. (a) Optimal λ values for various q. (b) Cross-validation curve as function of q, with optimal values λ_q^* inserted for λ. (c) Number of nonzero estimated regression coefficients for various q and corresponding optimal λ.

finds only one gene (with $\hat{\lambda} = 25.030$). With the adaptive lasso also this gene is discarded in the second run. Here we used lasso in the first step and recalculated the weights within each cross-validation fold.

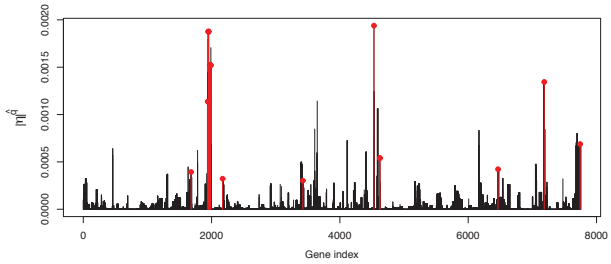
From Figure 3 it is obvious that genes corresponding to large values of $|\eta_j|^{\hat{q}}$ are promoted in the analysis. Note, however, that the selected genes not necessarily have the largest values of $|\eta_j|^{\hat{q}}$. As long as the weights do not insist too strongly they should be penalized out, genes with lower $|\eta_j|^{\hat{q}}$ can be selected if their expression shows a strong effect.

4.1.2 Validation

To evaluate survival predictions based on the selected set of genes, we use the independent test data set of 41 patients to calculate a prognostic index $PI = \mathbf{X}_{new}^T \hat{\beta}$ (Bøvelstad et al., 2007), using the estimated coefficients $\hat{\beta}$ calculated from the training data. Following Bøvelstad and Borgan (2011) we ranked the 41 patients according to their PI, divided them into two groups, and performed a simple log-rank test



(a)



(b)

Figure 3: External information: Plot of $|\eta_j|^{\hat{q}}$, where $\hat{q}$ is found by cross-validation. The selected genes are marked in red. (a) Spearman correlation coefficient between expression and copy number for each gene, (adjusted at zero). (b) Ridge regression coefficient indicating the copy number effect of each gene on survival (absolute value).

in order to test whether the hazard rates of the two groups were significantly different. In both training and test data, 1/3 of the patients show good prognosis and 2/3 bad prognosis, hence we maintain the same ratio in the division into two groups. For this test, the single lasso gene gave a P-value $P = 0.031$, the W2 gave $P = 0.025$ and the correlation weights W1 gave $P = 0.002$, indicating that the selected genes are able to discriminate between high-risk and low-risk patients. Following the arguments of Bøvelstad and Borgan (2011) we also computed a time-integrated version of the area under the ROC curve (*iAUC*) in addition to the log-rank test. As the results of the log-rank test might depend on how the patients are divided into the two groups, Bøvelstad and Borgan (2011) argue in favor of the *iAUC* measure since it examines all possible divisions into high-risk and low-risk groups. For the lasso we find $iAUC = 0.561$, for W2 $iAUC = 0.610$ and for W1 $iAUC = 0.753$, compared to the benchmark value of 0.5 where all covariate information is ignored. Thus for

DOI: 10.2202/1544-6115.1703

these data, both criteria found the simple correlation between gene expression and copy number to yield more robust results.

4.2 Example 2: Head and Neck Cancer Data

We re-analyze data concerning head and neck cancer from Roepman et al. (2006), using gene expressions as explanatory variables, while the response is binary (metastasis/metastasis free). A penalized logistic regression model is fitted. From the 3064 genes analyzed in Roepman et al. (2006) we extract only those for which there exists a unique gene symbol. The resulting data consist of gene expression measurements for 2060 genes in 65 samples.

The weighted penalties were determined through relevant literature annotations. Pubgene is a database providing associations between genes and other biological terms through text mining of the literature, see Jenssen et al. (2001) and `http://www.pubgene.com/`. Pubgene provides a list of gene symbols published together with a chosen keyword. It gives a score related to how often each gene is mentioned in connection with the keyword of interest. The score can be the number of articles in which both the gene and the biological keyword were found, which can be utilized to tilt the search in specific directions by weighting.

For illustration we used the biological keyword *anoxia*. Anoxia, or lack of oxygen, influences the expression of genes and has been shown to sometimes promote metastasis formation in cancer diseases (Gort et al., 2008). Many anoxia regulated genes have nothing to do with metastasis formation, and metastasis genes are regulated by a variety of other processes too (copy number alterations, mutations). However, in the search for genes associated with metastasis, it should be of help to know how strongly they are associated with anoxia. Hence we used $\eta_j(\mathbf{z}) = \log(z_j)$, where z_j is the number of articles associating gene j with anoxia reported by Pubgene. Genes never linked to anoxia in the literature were given a small positive value of z_j, smaller than the smallest positive count. The logarithm was used to reduce extreme effects of some very large values of z_j. We normalized all η_j to values $0 < \eta_j \leq 1$.

4.2.1 Results

The results are summarized in Figure 4. We found optimal values $(\hat{q}, \hat{\lambda}) = (4.750, 0.080)$, leading to 8 selected genes for the weighted lasso. For the lasso, $\hat{\lambda} = 2.484$ and 15 genes were selected. Again the adaptive lasso discarded all of these. In Figure 5 the transformed η_j values are plotted for each gene in the analysis with the two gene lists highlighted.

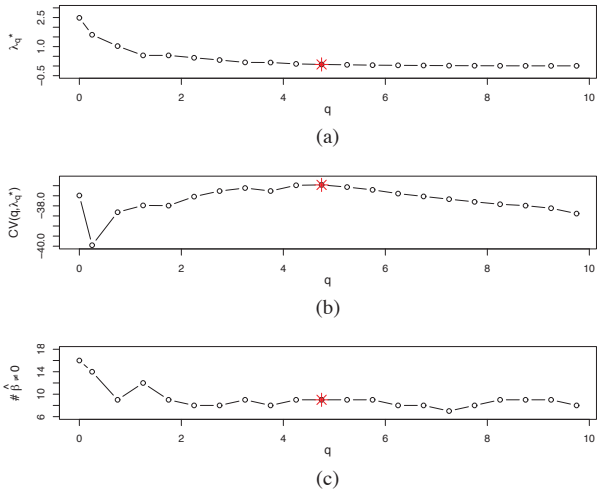


Figure 4: Cross-validation results for Example 2: Head and neck cancer data. The optimal value of q is marked by the red star. (a) Optimal λ values for various q. (b) Cross-validation curve as function of q, with optimal values λ_q^* inserted for λ. (c) Number of nonzero estimated regression coefficients for various q and corresponding optimal λ.

We did a biological validation of the resulting gene lists, comparing them with previous findings of association between genes and metastasis, as reported in the literature, see Tables 1 and 2. Both gene lists include genes that had previously been associated with metastasis and the signs of the estimated regression coefficients were in correspondence with previous findings. For the lasso analysis, however, more genes that have not been reported before as associated with metastasis were selected; 8 out of 15 for lasso, versus 2 out of 8 for weighted lasso. Although some of the unknown genes may play a role in metastasis development, the number of false positives might be higher in the lasso analysis. Lasso tends to select only one variable from groups of highly correlated variables (Zou and Hastie, 2005), while the external weighting is accommodating this by including relevant external information in the selection. The gene list we obtain with the weighted lasso should be viewed as a list of genes explaining the chance of metastasis, where the selection is guided by information on previously found associations with anoxia. Here all but two genes have previously been related to metastasis, which supports that the weights might have helped select more relevant genes.

DOI: 10.2202/1544-6115.1703

Table 1: List of the 8 genes selected by the weighted lasso. The last column indicates whether the selected genes have previously been associated with metastasis. Genes marked "-" are unknown in relation with metastasis.

	Gene symbol	Pubgene	$\hat{\beta}_j$	$\hat{\lambda}_j$	Biological validation
1	CD40	17	-0.208	2.086	-
2	FOS	60	0.049	0.527	Montell (2005)
3	IFNG	28	0.238	1.150	-
4	REN	143	0.927	0.249	Ino et al. (2006)
5	SERPINE1	23	0.004	1.441	Speleman et al. (2007)
6	MMP2	44	3.125	0.712	Danilewicz et al. (2003)
7	F3	30	1.332	1.065	Kasthuri et al. (2009)
8	HIF1A	598	0.190	0.090	Xueguan et al. (2008)

Table 2: List of the 15 genes selected by the standard lasso. The last column indicates whether the selected genes have previously been associated with metastasis. Genes marked "-" are unknown in relation with metastasis.

	Gene symbol	Pubgene	$\hat{\beta}_j$	Biological validation
1	TCAP	0	-0.106	-
2	CXCL10	0	-0.063	Jiang et al. (2009)
3	COL6A3	0	0.179	Sherman-Baust et al. (2003)
4	GFRA1	0	0.624	Iwahashi et al. (2002)
5	ZNF852	0	-0.462	-
6	HDAC5	1	0.065	Krivoruchko and Storey (2010)
7	RWDD2B	0	0.144	-
8	MYBPH	0	0.079	-
9	FN1	1	0.114	Wong et al. (2009)
10	LLGL2	0	-0.757	-
11	SPARC	1	0.690	Wong et al. (2009)
12	XKR8	0	0.238	-
13	BMPR1A	1	0.002	-
14	HNRNPL	1	0.140	-
15	TGFBI	0	0.810	Joshi and Cao (2010)

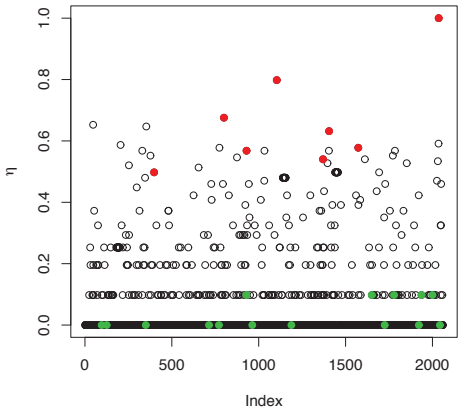


Figure 5: Plot of η_j for each gene j in Example 2. The genes selected by the lasso are marked in green, whereas the genes selected by the weighted lasso are marked in red.

5 Simulation Study

To assess the overall performance of our weighted lasso, and to compare it to the lasso and the adaptive lasso in settings where $p > n$, we present several simulation studies. The simulation experiments are designed to mimic real data situations similar to that of weighting scheme W1 in the cervix cancer example, where gene expressions correlated with their corresponding copy number were favored in the penalization. Covariates and external information $(\mathbf{x}_j, \mathbf{z}_j)$ were generated as detailed below. The response variables were simulated from $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$. The Spearman correlation coefficient was used to determine the weights and the vector of regression coefficients β was chosen in several ways, creating different degrees of sensible weighting to be spelled out below. We ran two-dimensional 10-fold cross-validation to find the best combination of penalty parameters λ and q for the weighted lasso, as described in Section 3. The penalty parameter λ' for the standard lasso was found through 10-fold cross-validation. To save computational costs we used the implementation of Kramer et al. (2009), which recalculates the adaptive weights within each cross-validation fold for the adaptive lasso.

DOI: 10.2202/1544-6115.1703

5.1 Simulation 1

Covariates and external information $(\mathbf{x}_j, \mathbf{z}_j)$ were simulated from a bivariate standard normal distribution with correlation $\rho = 0.8$ for $j = 1, ..., 10$, forming a group of 10 covariates $\mathbf{x}_j$ highly correlated with the external information $\mathbf{z}_j$. For $j = 11, ..., p$, $\mathbf{x}_j$ and $\mathbf{z}_j$ were drawn from uncorrelated standard normal distributions. The sample size $n = 50$ was kept fixed, whereas p was 100, 1000 and 10000, giving various degrees of sparsity. Two noise levels were considered; $\sigma = 1$ and $\sigma = 5$.

By choosing the true regression coefficients in various ways, five settings A-E were designed to account for scenarios where the weights are in correspondence with the true model to different degrees.

Setting A In the first scenario the true model has regression coefficients $(\beta_1, ..., \beta_{12}) = (-2, -2, -2, -2, -2, 2, 2, 2, 2, 2, 0, 0)$ and $(\beta_{13}, ..., \beta_p) = \mathbf{0}$. This corresponds to situations where our assumption is correct; the ten covariates $\mathbf{x}_j$ explaining the response are those which are highly correlated with $\mathbf{z}_j$. We expect the weighted lasso to do well when we use the Spearman correlation coefficients to determine the weights.

Setting B Next we consider the same scenario as in Setting A but exclude two variables correlated with $\mathbf{z}_j$ from the model, $(\beta_1, ..., \beta_{12}) = (-2, -2, -2, -2, 2, 2, 2, 2, 0, 0, 0, 0)$ and $(\beta_{13}, ...\beta_p) = \mathbf{0}$. Hence two variables that are not supposed to be in the model, have favorable weights. We show that variables which are not related to the response are not included just because of favorable weights.

Setting C It is of interest to select variables that are important for the response, without having as favorable weights as some of the other important variables. We let $(\beta_1, ..., \beta_{12}) = (-2, -2, -2, -2, -2, 2, 2, 2, 2, 2, 5, -5)$ and $(\beta_{13}, ...\beta_p) = \mathbf{0}$. Remember that only the ten first covariates are simulated to have favorable weights, thus two of the covariates which are not designed to have advantages through the weighting scheme are set to have nonzero regression coefficients.

Setting D We combine scenario B and C, and let $(\beta_1, ..., \beta_{12}) = (-2, -2, -2, -2, 2, 2, 2, 2, 0, 0, 5, -5)$ and $(\beta_{13}, ...\beta_p) = \mathbf{0}$. This reflects the situation where both variables not influencing the response are given a small penalty, and variables influencing the response are given a large penalty.

Setting E In the last scenario the variables given a favorable weight are not associated with the outcome. This scenario illustrates the effect of applying the weighted

lasso when the information we include is completely useless. $(\beta_1,...,\beta_{10}) = \mathbf{0}$, $(\beta_{11},...,\beta_{20}) = (-2,-2,-2,-2,-2,2,2,2,2,2)$ and $(\beta_{21},...,\beta_p) = \mathbf{0}$.

100 pairs of training and test sets were generated to evaluate the performance. Variable selection was assessed by sensitivity and specificity. The prediction mean squared error, $PMSE = n^{-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, where $\hat{y}_i$ is the fitted value of the training data, and y_i is the response value in the test data, was also evaluated.

5.2 Simulation 2

Covariates $\mathbf{x}_j$ and external information $\mathbf{z}_j$ of the cervix cancer data were used directly as covariates and for computation of weights to account for a more complex dependence structure between the covariates by conserving the dependencies among the variables from the data. The Spearman correlation coefficients r were used to define two groups;

- Group 1: Gene expression highly correlated with copy number ($r > 0.5$, 71 genes),
- Group 2: Gene expression less correlated with copy number, ($r \leq 0.5$, 7683 genes).

In each simulation we draw $p_{act} = 10$ variables $\mathbf{x}_1,...,\mathbf{x}_{10}$ from either Group 1 or Group 2 to constitute the variables with nonzero regression coefficients as follows.

$$(\beta_1,...,\beta_{10}) = \beta_{act} = (-2,-2,-2,-2,-2,2,2,2,2,2)$$
$$(\beta_{11},...,\beta_p) = \mathbf{0}$$

The rest of the covariates were included in different manners, leading to different scenarioes A', B' and E' comparable to A, B and E of Simulation 1. 100 simulated data sets were generated with $\sigma = 1$ and 5 as in Simulation 1.

Setting A' $\mathbf{x}_1,...,\mathbf{x}_{10}$ were randomly drawn from Group 1 to form the active set (the true model). All genes in Group 2 were included as covariates, whereas the rest of Group 1 was kept out of the analysis. This corresponds to a scenario where the weights are well designed and less penalization are given to the active set as in Setting A of Simulation 1. We fit a model with $p = 7693$ covariates.

DOI: 10.2202/1544-6115.1703

Setting B' $\mathbf{x}_1, ..., \mathbf{x}_{10}$ were randomly drawn from Group 1 to form the active set (the true model). The rest of Group 1 was also included as covariates, along with all genes in Group 2. This corresponds to a setting where the weights are partly informative, as in Setting B of Simulation 1. Some of the covariates included in the analysis will be penalized less, even if they are not a part of the true model. Here the total number of covariates is $p = 7754$.

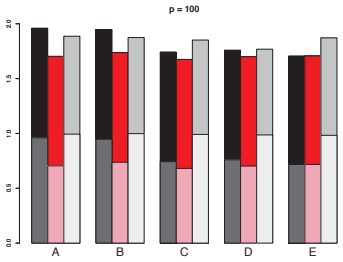
Setting E' $\mathbf{x}_1, ..., \mathbf{x}_{10}$ were randomly drawn from Group 2 to form the active set (the true model). The rest of Group 2 was included as covariates, as well as 10 random variables from Group 1. This corresponds to a setting where the weights are nonsense; the 10 covariates from Group 1 are subject to less penalization compared to the active set which are subject to a larger amount of penalization. We fit a model with $p = 7693$ covariates.

5.3 Results

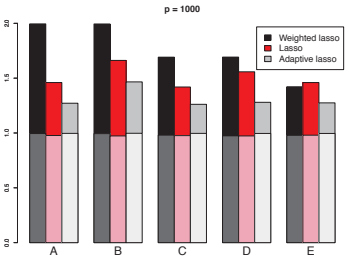
Variable Selection The results for variable selection are summarized in Figure 6, 7 and 8 and Supplementary Tables 1, 2 and 3 in the Supplementary material. The sensitivity and specificity are reported in the bar charts of Figure 6, 7 and 8. Methods having sensitivity and specificity close to 1 will have bars close to the ideal value of 2. Sensitivity measures the proportion of the true positive set, that is actually selected. Overall, the weighted lasso does much better in selecting the right variables, than both the lasso and the adaptive lasso. In situations B-D, where the weighting is not perfectly designed as in A, the weighted lasso still performs at least as well as the two other methods in terms of sensitivity. Standard deviations given in Supplementary Tables 1,2 and 3 are similar, almost always smaller for the weighted lasso than for the lasso. In situation E the results are comparable with the lasso with little price paid by introducing weights based on contradictory information.

In Simulation 1 we can study the effect when the number of covariates increases. Even if the lasso selects more covariates than the weighted lasso, sensitivity is decreasing for the lasso, while staying close to one for the weighted lasso. We observe a remarkable improvement for the weighted lasso compared to both the other methods for higher dimensions ($p = 1000$ and $p = 10000$), see Figure 6 and 7. It seems that the lasso overfits the training data, while the weighted lasso constructs more robust estimates and is able to select the right variables even when p is large. In the adaptive lasso, we see that the lasso used in the first step forces the right variables out, leaving no possibilities to adaption, see Benner et al. (2010). The same tendencies are seen in Simulation 2. Specificity measures the proportion of the true null set that is not selected, and is always high due to the sparse design of the simulations. Note that as p grows, the size of the active set remains unchanged.

	$p = 100$			
	$\#\beta \neq 0$	WL	Lasso	AL
A	10	13.6	36.6	10.5
		(6.4)	(7.2)	(1.1)
B	8	12.8	32.2	8.2
		(6.4)	(7.2)	(0.5)
C	12	34.6	40.0	12.0
		(7.5)	(5.9)	(2.1)
D	10	31.6	36.7	10.0
		(8.0)	(6.7)	(0.9)
E	10	35.4	35.4	10.4
		(7.2)	(7.2)	(1.3)



	$p = 1000$			
	$\#\beta \neq 0$	WL	Lasso	AL
A	10	15.1	27.6	6.6
		(8.2)	(16.6)	(6.3)
B	8	13.8	32.2	7.0
		(6.9)	(14.9)	(4.8)
C	12	27.8	29.0	6.9
		(13.2)	(14.1)	(5.2)
D	10	32.9	32.5	7.3
		(12.2)	(11.9)	(4.0)
E	10	25.5	24.8	7.1
		(15.0)	(15.2)	(6.5)



	$p = 10000$			
	$\#\beta \neq 0$	WL	Lasso	AL
A	10	15.3	12.6	1.7
		(8.6)	(15.8)	(3.6)
B	8	15.6	16.4	2.0
		(10.7)	(17.3)	(3.9)
C	12	23.3	18.9	4.5
		(13.6)	(15.3)	(4.5)
D	10	24.4	23.5	3.9
		(14.7)	(16.3)	(3.7)
E	10	18.1	13.9	2.0
		(15.5)	(15.6)	(3.8)

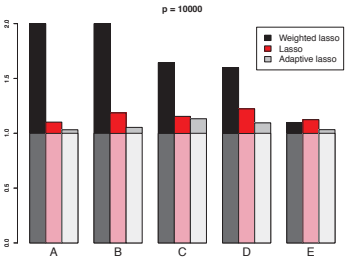
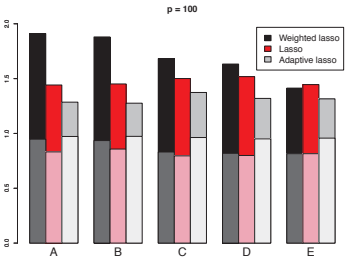


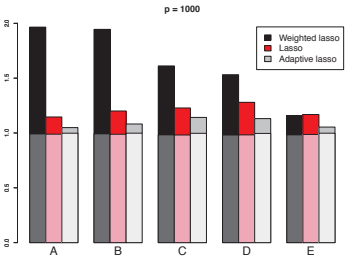
Figure 6: Simulation 1, $\sigma = 1$: Comparison of the weighted lasso, the standard lasso and the adaptive lasso. To the left, the average number of selected variables, $\hat{\beta} \neq 0$, is reported for each of the three methods, with standard deviations given in parentheses. To the right, sensitivity and specificity are reported as stacked bars for the different scenarioes A-E. The dark colors (on top) represent the sensitivity and the lighter colors (lower part of bars) the specificity. The reported measures are the means over the 100 simulated data sets.

DOI: 10.2202/1544-6115.1703

$p = 100$				
	$\#\beta \neq 0$	WL	Lasso	AL
A	10	14.3	21.2	5.9
		(7.5)	(13.1)	(5.4)
B	8	13.5	17.9	5.2
		(8.4)	(12.9)	(5.3)
C	12	25.1	26.4	8.5
		(10.0)	(12.3)	(5.6)
D	10	24.4	25.3	8.4
		(9.7)	(9.8)	(5.9)
E	10	22.6	22.9	7.4
		(14.4)	(14.2)	(6.4)



$p = 1000$				
	$\#\beta \neq 0$	WL	Lasso	AL
A	10	18.3	12.4	2.6
		(11.3)	(15.6)	(4.7)
B	8	15.2	13.5	2.5
		(10.7)	(14.7)	(4.0)
C	12	24.1	21.2	4.8
		(13.1)	(15.9)	(5.2)
D	10	23.6	20.2	5.4
		(14.4)	(13.8)	(4.8)
E	10	18.6	15.2	2.6
		(15.5)	(15.7)	(4.4)



$p = 10000$				
	$\#\beta \neq 0$	WL	Lasso	AL
A	10	16.4	13.3	2.1
		(9.4)	(16.9)	(4.6)
B	8	19.4	10.1	2.1
		(13.2)	(13.6)	(4.7)
C	12	20.9	17.8	3.2
		(14.1)	(17.5)	(4.0)
D	10	20.1	15.1	3.0
		(14.1)	(16.2)	(4.1)
E	10	19.3	12.0	1.4
		(17.2)	(16.0)	(3.2)

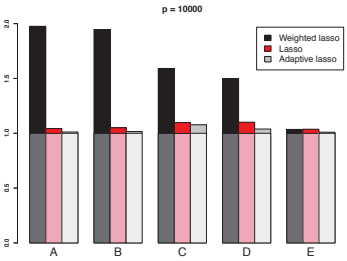
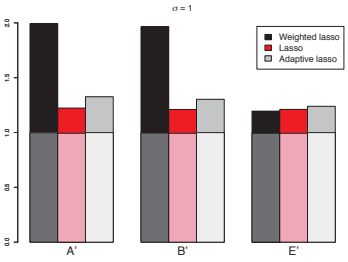


Figure 7: Simulation 1, $\sigma = 5$: Comparison of the weighted lasso, the standard lasso and the adaptive lasso. To the left, the average number of selected variables, $\hat{\beta} \neq 0$, is reported for each of the three methods, with standard deviations given in parentheses. To the right, sensitivity and specificity are reported as stacked bars for the different scenarioes A-E. The dark colors (on top) represent the sensitivity and the lighter colors (lower part of bars) the specificity. The reported measures are the means over the 100 simulated data sets.

$\sigma = 1$

	$\#\beta \neq 0$	WL	Lasso	AL
A'	10	25.3	57.2	8.4
		(10.8)	(21.2)	(3.4)
B'	10	40.0	53.2	8.3
		(12.8)	(25.9)	(3.7)
E'	10	55.1	52.9	7.2
		(20.6)	(22.7)	(3.0)



$\sigma = 5$

	$\#\beta \neq 0$	WL	Lasso	AL
A'	10	18.3	31.4	1.4
		(14.7)	(10.6)	(3.0)
B'	10	18.5	32.6	1.7
		(17.1)	(10.7)	(3.5)
E'	10	21.2	32.1	2.4
		(18.6)	(9.5)	(3.2)

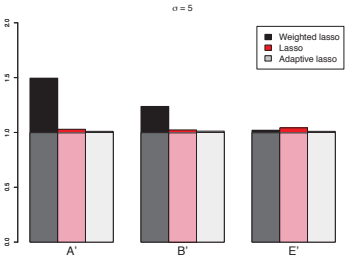


Figure 8: Simulation 2: Comparison of the weighted lasso, the standard lasso and the adaptive lasso. To the left, the average number of selected variables, $\hat{\beta} \neq 0$, is reported for each of the three methods, with standard deviations given in parentheses. To the right, sensitivity and specificity are reported as stacked bars for the different scenarioes A', B' and E'. The dark colors (on top) represent the sensitivity and the lighter colors (lower part of bars) the specificity. The reported measures are the means over the 100 simulated data sets.

DOI: 10.2202/1544-6115.1703

All three methods, in both simulations, have more problems finding the correct variables when noise increases. The weighted lasso, however, does much better than the lasso. Although the lasso now sometimes selects fewer variables than the weighted lasso, it tends to select the wrong ones. It is similar for the adaptive lasso.

Prediction Performance We investigated the prediction performance in Simulation 1 (for Simulation 2 we do not have test data); the results are given in terms of prediction mean squared error (PMSE) in Table 3. In situation A and B, where the weights are informative, the weighted lasso is clearly better than both the lasso and the adaptive lasso in terms of PMSE. For A and B the PMSE ratios are far below 1 in 23 out of 24 cases, as high-lighted in Table 3. We are here able to select the correct variables and estimate their coefficients more accurately leading to very good predictions. Note that when p increases, the effect of the external weighting improves even more upon the prediction performance and the weighted lasso performs remarkably much better than both other methods in situation A and B. For $\sigma = 1$ and $p = 100$ we see that also the adaptive lasso predicts the response quite well for situation A and B. As we saw for variable selection, the weighted lasso is able to select all the relevant variables, while the adaptive lasso exludes some of them. However, it seems that the adaptive lasso is able to explain the response based on its selected variables quite well, even if not all of the nonzero coefficients are found.

When the noise increases the prediction performance of the weighted lasso is always better or comparable with the two others, and fairly stable across different values of p.

In situations where the external information is not reasonable, the weighted lasso is more similar to both the lasso and the adaptive lasso. Actually $q = 0$ was selected for several of the replications in situation C, D and E; when the external weights are not informative, the standard lasso is selected.

Bias Reduction It is also interesting to compare the values of the estimated regression coefficients to comment on the bias. The lasso is known to overshrink the final coefficients (James and Radchenko, 2009). Several methods help reducing this bias, for example the elastic net (Zou and Hastie, 2005), the adaptive lasso (Zou, 2006), the relaxed lasso (Meinshausen, 2007) and SIS (Fan and Lv, 2008). Our approach also produce remarkably less biased estimates than the lasso. This is illustrated in Figure 9, where the first 15 regression coefficients are plotted for Simulation 1 with $p = 1000$: the standard lasso estimates become biased towards zero as a consequence of overshrinking. When different amounts of penalization are imposed on the coefficients in the weighted lasso, the estimates are less biased,

Table 3: Simulation 1: Comparison of the prediction performance for the weighted lasso, the standard lasso and the adaptive lasso.The reported measures are the means over the 100 simulated data sets, standard deviations given in parentheses. WL/L and WL/AL are the means of the ratio of PMSE for the weighted lasso and the two other methods respectively.

	$\sigma = 1, p = 100$					$\sigma = 5, p = 100$				
	$PMSE_{WL}$	$PMSE_L$	$PMSE_{AL}$	WL/L	WL/AL	$PMSE_{WL}$	$PMSE_L$	$PMSE_{AL}$	WL/L	WL/AL
A	1.48	3.93	2.33	**0.38**	**0.64**	38.12	56.40	58.64	**0.68**	**0.65**
	(0.41)	(3.14)	(3.66)			(15.33)	(14.70)	(14.73)		
B	1.39	2.44	1.38	**0.57**	1.01	36.32	52.03	54.16	**0.70**	**0.67**
	(0.37)	(0.94)	(0.36)			(11.93)	(13.94)	(13.56)		
C	3.18	5.56	8.33	**0.57**	**0.38**	63.65	68.18	68.22	**0.93**	**0.93**
	(1.69)	(4.56)	(11.44)			(28.37)	(20.25)	(20.95)		
D	2.76	4.38	3.64	**0.63**	**0.76**	59.35	59.72	61.47	**0.99**	**0.97**
	(1.40)	(4.49)	(5.33)			(38.01)	(15.75)	(17.64)		
E	4.31	4.29	2.64	1.04	1.63	59.00	56.78	59.11	1.04	1.00
	(4.78)	(4.78)	(4.45)			(17.31)	(14.94)	(15.01)		
	$\sigma = 1, p = 1000$					$\sigma = 5, p = 1000$				
	$PMSE_{WL}$	$PMSE_L$	$PMSE_{AL}$	WL/L	WL/AL	$PMSE_{WL}$	$PMSE_L$	$PMSE_{AL}$	WL/L	WL/AL
A	1.45	34.17	36.26	**0.04**	**0.04**	37.51	66.60	67.43	**0.56**	**0.56**
	(0.40)	(12.60)	(13.47)			(9.92)	(13.97)	(13.75)		
B	1.42	20.58	20.54	**0.07**	**0.07**	37.97	58.17	60.06	**0.65**	**0.63**
	(0.44)	(11.24)	(14.41)			(11.72)	(12.42)	(13.40)		
C	47.45	55.60	52.37	**0.85**	**0.91**	90.20	97.54	95.42	**0.93**	**0.95**
	(25.10)	(15.63)	(16.70)			(25.50)	(27.35)	(28.74)		
D	35.19	40.87	34.96	**0.86**	1.01	86.59	85.96	83.80	1.01	1.03
	(23.98)	(16.18)	(15.04)			(25.70)	(23.73)	(27.12)		
E	34.71	33.52	35.00	1.04	**0.99**	72.93	68.21	69.41	1.07	1.05
	(11.05)	(9.87)	(11.04)			(21.54)	(15.40)	(15.23)		
	$\sigma = 1, p = 10000$					$\sigma = 5, p = 10000$				
	$PMSE_{WL}$	$PMSE_L$	$PMSE_{AL}$	WL/L	WL/AL	$PMSE_{WL}$	$PMSE_L$	$PMSE_{AL}$	WL/L	WL/AL
A	1.47	41.93	42.53	**0.04**	**0.03**	36.64	66.68	68.78	**0.55**	**0.53**
	(0.44)	(8.09)	(8.70)			(11.22)	(14.18)	(16.36)		
B	1.43	32.98	34.05	**0.04**	**0.04**	38.83	57.97	59.11	**0.67**	**0.66**
	(0.45)	(6.92)	(7.54)			(11.41)	(11.89)	(13.94)		
C	72.66	76.32	70.40	**0.95**	1.03	112.74	113.72	111.58	**0.99**	1.01
	(19.36)	(19.68)	(24.63)			(28.24)	(23.22)	(25.88)		
D	61.84	65.87	56.18	**0.94**	1.10	97.44	99.52	101.02	**0.98**	**0.97**
	(28.30)	(22.00)	(22.91)			(25.75)	(22.94)	(28.37)		
E	45.21	41.87	43.78	1.08	1.03	72.82	67.15	67.99	1.08	1.07
	(11.70)	(7.87)	(9.56)			(19.21)	(15.63)	(17.93)		

since the weighted lasso penalizes less on the nonzero regression coefficients. In setting A and B, the estimates of the weighted lasso are perfectly centered around the true value, whereas the lasso estimates are strongly biased toward zero. In setting C and D the estimates of the weighted lasso and the lasso are more similar. This is probably because the lasso is a special case of our weighted lasso ($q = 0$) and is in fact selected in some replications if the weighting is not informative. This is also clearly apparent in Setting E corresponding to the situation where the weights

DOI: 10.2202/1544-6115.1703

are not informative. The bias of the estimates for the regression coefficients in the active set $(\hat{\beta}_{11}, ..., \hat{\beta}_{20})$ are very similar to the lasso.

We also see that if the weights in the weighted lasso are informative, as in Setting A and B, the estimates are clearly less biased than estimates of the adaptive lasso. The adaptive lasso is constructed to reduce the bias for large coefficients (Zou, 2006). We see this in setting C and D, where the adaptive lasso gives less biased estimates for the two large coefficients, than the other two methods. Overall in setting C and D, that is, for all of the variables in the active set, the weighted lasso does better in reducing the bias. As discussed, in setting E the weighted lasso performs similarly to the lasso. Hence the adaptive lasso is slightly better in estimating the nonzero true coefficients when our weights are nonsense.

6 Discussion

We have proposed a method for weighted penalization with data integration. Variables that are important due to external information are promoted in the analysis. We have focused on the lasso, but the idea of incorporating external information in the penalties can of course be used with any other type of penalty.

The proposed approach is general and does not require the specific choice of weight function (2). Problem specific weight functions with alternative shapes can be designed for example inspired by Green (1990). For more flexibility in the shape of weights, one could introduce a second tuning parameter in the weight function, at the cost of having one more dimension in the cross-validation.

From a Bayesian perspective, introducing λ_j instead of a common λ in a lasso regression model corresponds to a Bayesian regression approach where Laplace priors with unequal variances are assumed for the regression coefficients. The external information thus determines the variance of the prior distributions of the regression coefficients.

Depending on what kind of additional information is included in the analysis, there are different interpretations of the resulting selected variables. In the cervix cancer example we had both expression and copy number data for each gene for the same patients. We used two biologically justified weights based on gene copy numbers to guide the analysis. Not surprisingly the genes selected in each analysis differ, as different aspects of copy number alterations are exploited. In one case, the analysis favored genes with high correlation between expression and copy number, in the second case, genes whose copy number explains survival. Both lists were confirmed by validation, and can include possible driving forces for cancer progression. Integrating data through weighting can provide stronger predictors, in the search for new biomarkers.

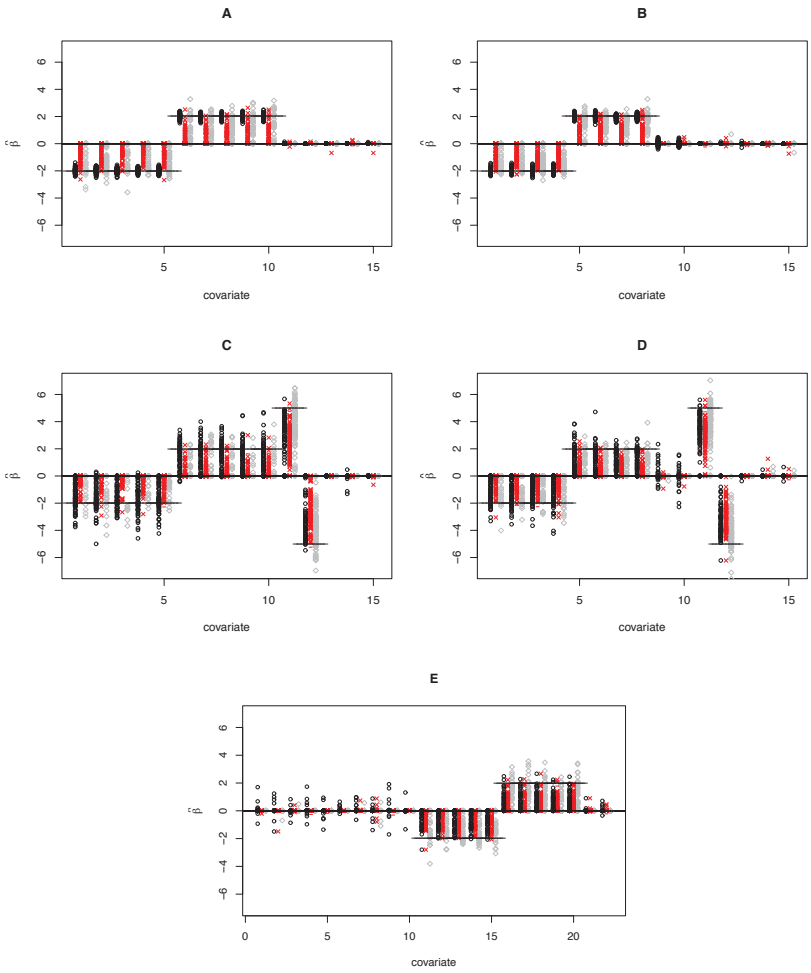


Figure 9: Plot of regression coefficients, weighted lasso (○), the lasso (×) and the adaptive lasso (◇) for the situation with $p = 1000$ and $\sigma = 1$. The true values are marked as horizontal lines. In Situation A and B the weighted lasso gives clearly less biased estimates than both the lasso and the adaptive lasso.

DOI: 10.2202/1544-6115.1703

In other situations, one might want to incorporate external information that is not a part of the specific data at hand. The head and neck cancer example is of this type, where literature information is included in the weights. Including information of this type could lead to more stable results because the penalization is based on information from previous studies, and thus suppress the effects of random artifacts in the data at hand. On the other side, new discoveries would be penalized. Using literature annotations as external information and defining weights appropriately can also be used to tilt the search for relevant associations away from known factors, thus encouraging new discoveries.

References

Albertson, D. (2006): "Gene amplification in cancer," *Trends in genetics*, 22, 447–455.

Benner, A., M. Zucknick, T. Hielscher, C. Ittrich, and U. Mansmann (2010): "High-dimensional cox models: The choice of penalty as part of the model building process," *Biometrical Journal*, 52, 50–69, URL http://dx.doi.org/10.1002/bimj.200900064

Bøvelstad, H. M. and Ø. Borgan (2011): "Assessment of evaluation criteria for survival prediction from genomic data," *Biometrical Journal*, 53, 202–216.

Bøvelstad, H. M., S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjærde (2007): "Predicting survival from microarray data - a comparative study," *Bioinformatics*, 23, 2080–2087.

Candés, E. and T. Tao (2007): "The dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, 35, 2313–2351.

Chanock, S. J., T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter, et al. (2007): "Replicating genotype-phenotype associations," *NATURE*, 447, 655–660.

Charbonnier, C., J. Chiquet, and C. Ambroise (2010): "Weighted-lasso for structured network inference from time course data," *Statistical Applications in Genetics and Molecular Biology*, 9.

Danilewicz, M., B. Sikorska, and M. Wagrowska-Danilewicz (2003): "Prognostic significance of the immunoexpression of matrix metalloproteinase MMP2 and its inhibitor TIMP2 in laryngeal cancer," *Medical Science Monitor*, 9, MT42–7.

Donnelly, P. (2008): "Progress and challenges in genome-wide association studies in humans," *Nature*, 456, 728–731.

Ein-Dor, L., I. Kela, G. Getz, D. Givol, and E. Domany (2005): "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics*, 21, 171–178.

Fan, J. and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and R. Li (2002): "Variable selection for Cox's proportional hazards model and frailty model," *The Annals of Statistics*, 30, 74–99.

Fan, J. and J. Lv (2008): "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911, URL http://dx.doi.org/10.1111/j.1467-9868.2008.00674.x.

Fan, J. and J. Lv (2010): "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, 20, 101–148.

Ferkingstad, E., A. Frigessi, H. Rue, G. Thorleifsson, and A. Kong (2008): "Unsupervised empirical Bayesian multiple testing with external covariates," *Annals of Applied Statistics*, 2, 714–735.

Gort, E., A. Groot, E. van der Wall, P. J. van Diest, and M. A. Vooijs (2008): "Hypoxic regulation of metastasis via hypoxia-inducible factors," *Current Molecular Medicine*, 8, 60–67.

Grandvalet, Y. and S. Canu (1998): "Outcomes of the equivalence of adaptive ridge with least absolute shrinkage," in M. S. Kearns, S. Solla, and D. Cohn, eds., *Advances in Neural Information Processing Systems 11*, MIT Press, 445–451.

Green, P. J. (1990): "Bayesian reconstructions from emission tomography data using a modified em algorithm," *Medical Imaging, IEEE Transactions on*, 9, 84–93.

Hans, C. (2009): "Bayesian lasso regression," *Biometrika*, 96, 835–845.

Hoerl, A. E. and R. Kennard (1970): "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Huang, J., S. Ma, and C.-H. Zhang (2006): "Adaptive lasso for sparse high-dimensional regression models," Technical Report 374, The department of Statistics and Actuarial Science, The University of Iowa.

Huang, J., S. Ma, and C.-H. Zhang (2008): "Adaptive lasso for sparse high-dimensional regression models," *Statistica Sinica*, 18, 1603–1618.

Ino, K., K. Shibata, H. Kajiyama, A. Nawa, S. Nomura, and F. Kikkawa (2006): "Manipulating the angiotensin system - new approaches to the treatment of solid tumours," *Expert Opinion on Biological Therapy*, 6, 243–255.

Ioannidis, J. P. (2007): "Non-replication and inconsistency in the genome-wide association setting," *Human Heredity*, 64, 203–213.

DOI: 10.2202/1544-6115.1703

Iwahashi, N., T. Nagasaka, G. Tezel, T. Iwashita, N. Asai, Y. Murakumo, K. Kiuchi, K. Sakata, Y. Nimura, and M. Takahashi (2002): "Expression of glial cell line-derived neurotrophic factor correlates with perineural invasion of bile duct carcinoma," *Cancer*, 94, 167–174.

James, G. M. and P. Radchenko (2009): "A generalized dantzig selector with shrinkage tuning," *Biometrika*, 96, 323–337.

Jenssen, T.-K., A. Laegreid, J. Komorowski, and E. Hovig (2001): "A literature network of human genes for high-throughput analysis of gene expression." *Nature Genetics*, 28, 21–28.

Jiang, Z., Y. Xu, and S. Cai (2009): "CXCL10 expression and prognostic significance in stage II and III colorectal cancer," *Molecular Biology Reports*, 37, 3029–3036.

Joshi, A. and D. Cao (2010): "TGF-beta signaling, tumor microenvironment and tumor progression: the butterfly effect," *Frontiers in Bioscience*, 15, 180–194.

Kasthuri, R. S., M. B. Taubman, and N. Mackman (2009): "Role of tissue factor in cancer," *Journal of clinical oncology*, 27, 4834–4838.

Kramer, N., J. Schafer, and A.-L. Boulesteix (2009): "Regularized estimation of large-scale gene association networks using graphical gaussian models," *BMC Bioinformatics*, 10, 384, URL `http://www.biomedcentral.com/1471-2105/10/384`

Krivoruchko, A. and K. B. Storey (2010): "Epigenetics in anoxia tolerance: a role for histone deacetylases," *Molecular and Cellular Biochemistry*, 342, 151–161.

Lando, M., M. Holden, L. C. Bergersen, D. H. Svendsrud, T. Stokke, K. Sundfør, I. K. Glad, G. B. Kristensen, and H. Lyng (2009): "Gene dosage, expression, and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer," *PLOS Genetics*, 5, e1000719, URL `http://dx.doi.org/10.1371%2Fjournal.pgen.1000719`.

Ma, S., X. Song, and J. Huang (2007): "Supervised group lasso with applications to microarray data analysis," *BMC Bioinformatics*, 8, doi:10.1186/1471–2105–8–60.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn (2008): "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature reviews*, 9, 356–369.

Meinshausen, N. (2007): "Relaxed lasso," *Computational Statistics & Data Analysis*, 52, 374–393.

Meinshausen, N. and P. Bühlmann (2010): "Stability selection (with discussion)," *Journal of the Royal Statistical Society: Series B*, 72, 417–473.

Michiels, S., S. Koscielny, and C. Hill (2005): "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *Lancet*, 365, 488–492.

Montell, D. J. (2005): "Anchors away! fos fosters anchor-cell invasion," *Cell*, 121, 816–817.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung (2004): "Genetic analysis of genome-wide variation in human gene expression," *Nature*, 430, 743–747.

Nowak, G. and R. Tibshirani (2008): "Complementary hierarchical clustering," *Biostatistics*, 9, 467–483.

Pan, W., B. Xie, and X. Shen (2010): "Incorporating predictor network in penalized regression with application to microarray data," *Biometrics*, 66, 474–484.

Park, M. and T. Hastie (2007): *glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model*, r package version 0.94.

Park, T. and G. Casella (2008): "The bayesian lasso," *Journal of the American Statistical Association*, 103, 672–680.

Percival, D., K. Roeder, R. Rosenfeld, and L. Wassermann (2010): "Structured sparse regression with application to hiv drug resistance", *Annals of Applied Statistics*, arXiv:1002.3128v2.

Roepman, P., P. Kemmeren, L. F. A. Wessels, P. Slootweg, and F. Holstege (2006): "Multiple robust signatures for detecting lymph node metastasis in head and neck cancer," *Cancer Research*, 66, 2361–2366.

Sherman-Baust, C. A., A. T. Weeraratna, L. B. Rangel, E. S. Pizer, K. R. Cho, D. R. Schwartz, T. Shock, and P. J. Morin (2003): "Remodeling of the extracellular matrix through overexpression of collagen VI contributes to cisplatin resistance in ovarian cancer cells," *Cancer cell*, 3, 377–386.

Slawski, M., W. zu Castell, and Tutz (2010): "Feature selection guided by structural information," *The Annals of Applied Statistics*, 4, 1056–1080.

Speleman, L., J. D. Kerrebijn, M. P. Look, C. A. Meeuwis, J. A. Foekens, and E. M. Berns (2007): "Prognostic value of plasminogen activator inhibitor-1 in head and neck squamous cell carcinoma," *Head & neck*, 29, 341–350.

Tai, F. and W. Pan (2007): "Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms," *Bioinformatics*, 23, 1775–1782.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, 58, 267–288.

van de Geer, S., P. Bühlmann, and S. Zhou (2010): "Prediction and variable selection with the adaptive lasso," ArXiv e-prints, 1001.5176v2

Verweij, P. J. M. and H. C. van Houwelingen (1993): "Cross-validation in survival analysis," *Statistics in Medicine*, 12, 2305–2314.

DOI: 10.2202/1544-6115.1703

Wong, F. H., C. Y. Huang, L. J. Su, Y. C. Wu, Y. S. Lin, J. Y. Hsia, H. T. Tsai, S. A. Lee, C. H. Lin, C. H. Tzeng, P. M. Chen, Y. J. Chen, S. C. Liang, J. M. Lai, and C. C. Yen (2009): "Combination of microarray profiling and protein-protein interaction databases delineates the minimal discriminators as a metastasis network for esophageal squamous cell carcinoma," *International journal of oncology*, 34, 117–128.

Xie, Y., W. Pan, K. S. Jeong, and A. Khodursky (2007): "Incorporating prior information via shrinkage: a combined analysis of genome-wide location data and gene expression data," *Statistics in Medicine*, 26, 2258–2275.

Xueguan, L., W. Xiaoshen, Z. Yongsheng, H. Chaosu, S. Chunying, and F. Yan (2008): "Hypoxia inducible factor-1 alpha and vascular endothelial growth factor expression are associated with a poor prognosis in patients with nasopharyngeal carcinoma receiving radiotherapy with carbogen and nicotinamide," *Clinical Oncology*, 20, 606–612.

Yuan, M. and Y. Lin (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, 68, 49–67.

Zhang, H. H. and W. Lu (2007): "Adaptive lasso for Cox's proportional hazards model," *Biometrika*, 94, 691–703.

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the America Statistical Association*, 101, 1418–1429.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, 67, 301–320.

Weighted Lasso with Data Integration

Supplementary Material

Linn Cecilie Bergersen, Department of Mathematics, University of Oslo
Ingrid K. Glad, Department of Mathematics, University of Oslo
Heidi Lyng, Department of Radiation Biology, Norwegian Radium Hospital

1 Results of Simulation Study

The details of the results of Simulation 1 are given in Supplementary Table S1 and S2 for $\sigma = 1$ and $\sigma = 5$ respectively. Supplementary Table S3 reports the results of Simulation 2.

Supplementary Table S1.

Table 1: Simulation 1: Comparison of the weighted lasso, the standard lasso and the adaptive lasso. Sensitivity and specificity are reported for variable selection. The reported measures are the means over the 100 simulated data sets, with standard deviations given in parentheses.

		$\sigma = 1, p = 100$								
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A	10	13.6	0.960	1.000	36.6	0.705	0.999	10.5	0.994	0.893
		(6.4)	(0.071)	(0.000)	(7.2)	(0.081)	(0.010)	(1.1)	(0.016)	(0.041)
B	8	12.8	0.947	1.000	32.2	0.737	1.000	8.2	0.998	0.876
		(6.4)	(0.070)	(0.000)	(7.2)	(0.079)	(0.000)	(0.5)	(0.005)	(0.013)
C	12	34.6	0.743	1.000	40.0	0.681	0.994	12.0	0.991	0.861
		(7.5)	(0.085)	(0.000)	(5.9)	(0.066)	(0.030)	(2.1)	(0.020)	(0.139)
D	10	31.6	0.759	1.000	36.7	0.703	0.997	10.0	0.986	0.782
		(8.0)	(0.089)	(0.000)	(6.7)	(0.074)	(0.022)	(0.9)	(0.007)	(0.074)
E	10	35.4	0.717	0.990	35.4	0.717	0.990	10.4	0.983	0.889
		(7.2)	(0.078)	(0.059)	(7.2)	(0.078)	(0.059)	(1.3)	(0.014)	(0.072)
		$\sigma = 1, p = 1000$								
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A	10	15.1	0.995	1.000	27.6	0.977	0.483	6.6	0.996	0.275
		(8.2)	(0.008)	(0.000)	(16.6)	(0.015)	(0.297)	(6.3)	(0.005)	(0.262)
B	8	13.8	0.994	1.000	32.2	0.973	0.689	7.0	0.997	0.469
		(6.9)	(0.007)	(0.000)	(14.9)	(0.013)	(0.292)	(4.8)	(0.004)	(0.328)
C	12	27.8	0.980	0.711	29.0	0.976	0.443	6.9	0.996	0.265
		(13.2)	(0.013)	(0.235)	(14.1)	(0.013)	(0.172)	(5.2)	(0.004)	(0.115)
D	10	32.9	0.974	0.718	32.5	0.973	0.585	7.3	0.996	0.284
		(12.2)	(0.012)	(0.219)	(11.9)	(0.011)	(0.190)	(4.0)	(0.003)	(0.176)
E	10	25.5	0.979	0.443	24.8	0.980	0.480	7.1	0.996	0.279
		(15.0)	(0.014)	(0.273)	(15.2)	(0.014)	(0.259)	(6.5)	(0.005)	(0.246)
		$\sigma = 1, p = 10000$								
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A	10	15.3	0.999	1.000	12.6	0.999	0.103	1.7	1.000	0.032
		(8.6)	(0.001)	(0.000)	(15.8)	(0.001)	(0.128)	(3.6)	(0.000)	(0.068)
B	8	15.6	0.999	1.000	16.4	0.999	0.188	2.0	1.000	0.054
		(10.7)	(0.001)	(0.000)	(17.3)	(0.002)	(0.192)	(3.9)	(0.000)	(0.102)
C	12	23.3	0.998	0.646	18.9	0.998	0.156	4.5	1.000	0.133
		(13.6)	(0.001)	(0.239)	(15.3)	(0.001)	(0.103)	(4.5)	(0.000)	(0.089)
D	10	24.4	0.998	0.599	23.5	0.998	0.225	3.9	1.000	0.095
		(14.7)	(0.001)	(0.255)	(16.3)	(0.002)	(0.129)	(3.7)	(0.000)	(0.074)
E	10	18.1	0.998	0.100	13.9	0.999	0.125	2.0	1.000	0.033
		(15.5)	(0.001)	(0.119)	(15.6)	(0.001)	(0.135)	(3.8)	(0.000)	(0.067)

Supplementary Table S2.

Table 2: Simulation 1: Comparison of the weighted lasso, the standard lasso and the adaptive lasso. Sensitivity and specificity are reported for variable selection. The reported measures are the means over the 100 simulated data sets, with standard deviations given in parentheses.

		$\sigma = 5$, $p = 100$								
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A	10	14.3	0.948	0.963	21.2	0.832	0.610	5.9	0.972	0.313
		(7.5)	(0.085)	(0.106)	(13.1)	(0.126)	(0.264)	(5.4)	(0.039)	(0.228)
B	8	13.5	0.935	0.945	17.9	0.858	0.594	5.2	0.973	0.302
		(8.4)	(0.091)	(0.137)	(12.9)	(0.120)	(0.314)	(5.3)	(0.040)	(0.246)
C	12	25.1	0.831	0.852	26.4	0.796	0.705	8.5	0.963	0.411
		(10.0)	(0.106)	(0.162)	(12.3)	(0.119)	(0.219)	(5.6)	(0.041)	(0.211)
D	10	24.4	0.819	0.814	25.3	0.799	0.719	8.4	0.950	0.369
		(9.7)	(0.103)	(0.206)	(9.8)	(0.097)	(0.200)	(5.9)	(0.047)	(0.196)
E	10	22.6	0.815	0.598	22.9	0.815	0.630	7.4	0.957	0.358
		(14.4)	(0.137)	(0.296)	(14.2)	(0.134)	(0.286)	(6.4)	(0.049)	(0.243)
		$\sigma = 5$, $p = 1000$								
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A	10	18.3	0.991	0.974	12.4	0.989	0.157	2.6	0.998	0.051
		(11.3)	(0.012)	(0.073)	(15.6)	(0.014)	(0.186)	(4.7)	(0.004)	(0.081)
B	8	15.2	0.992	0.953	13.5	0.988	0.212	2.5	0.998	0.084
		(10.7)	(0.011)	(0.106)	(14.7)	(0.013)	(0.214)	(4.0)	(0.003)	(0.137)
C	12	24.1	0.983	0.628	21.2	0.982	0.247	4.8	0.997	0.145
		(13.1)	(0.014)	(0.247)	(15.9)	(0.015)	(0.166)	(5.2)	(0.004)	(0.112)
D	10	23.6	0.982	0.550	20.2	0.983	0.297	5.4	0.996	0.135
		(14.4)	(0.014)	(0.244)	(13.8)	(0.013)	(0.164)	(4.8)	(0.004)	(0.094)
E	10	18.6	0.983	0.176	15.2	0.987	0.182	2.6	0.998	0.056
		(15.5)	(0.015)	(0.169)	(15.7)	(0.015)	(0.177)	(4.4)	(0.004)	(0.087)
		$\sigma = 5$, $p = 10000$								
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A	10	16.4	0.999	0.976	13.3	0.999	0.045	2.1	1.000	0.012
		(9.4)	(0.001)	(0.062)	(16.9)	(0.002)	(0.077)	(4.6)	(0.000)	(0.038)
B	8	19.4	0.999	0.946	10.1	0.999	0.051	2.1	1.000	0.016
		(13.2)	(0.001)	(0.110)	(13.6)	(0.001)	(0.097)	(4.7)	(0.000)	(0.049)
C	12	20.9	0.999	0.593	17.8	0.998	0.099	3.2	1.000	0.077
		(14.1)	(0.001)	(0.248)	(17.5)	(0.002)	(0.083)	(4.0)	(0.000)	(0.078)
D	10	20.1	0.998	0.501	15.1	0.999	0.102	3.0	1.000	0.039
		(14.1)	(0.001)	(0.266)	(16.2)	(0.002)	(0.095)	(4.1)	(0.000)	(0.053)
E	10	19.3	0.998	0.037	12.0	0.999	0.038	1.4	1.000	0.010
		(17.2)	(0.002)	(0.068)	(16.0)	(0.002)	(0.069)	(3.2)	(0.000)	(0.036)

Supplementary Table S3.

Table 3: Simulation 2: Comparison of the weighted lasso, the standard lasso and the adaptive lasso. The reported measures are the means over the 100 simulated data sets, with standard deviations given in parentheses.

					$\sigma = 1$					
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A'	10	25.3	0.998	0.995	57.2	0.993	0.230	8.4	0.999	0.327
		(10.8)	(0.001)	(0.022)	(21.2)	(0.003)	(0.153)	(3.4)	(0.000)	(0.176)
B'	10	40.0	0.996	0.968	53.2	0.993	0.217	8.3	0.999	0.304
		(12.8)	(0.002)	(0.057)	(25.9)	(0.003)	(0.175)	(3.7)	(0.000)	(0.187)
E'	10	55.1	0.993	0.202	52.9	0.993	0.218	7.2	0.999	0.240
		(20.6)	(0.003)	(0.139)	(22.7)	(0.003)	(0.146)	(3.0)	(0.000)	(0.139)
					$\sigma = 5$					
		The weighted lasso			The lasso			The adaptive lasso		
	$\#\beta \neq 0$	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.	$\#\hat{\beta} \neq 0$	Specif.	Sensit.
A'	10	18.3	0.998	0.497	31.4	0.996	0.033	1.4	1.000	0.010
		(14.7)	(0.002)	(0.331)	(10.6)	(0.001)	(0.057)	(3.0)	(0.000)	(0.030)
B'	10	18.5	0.998	0.239	32.6	0.996	0.027	1.7	1.000	0.012
		(17.1)	(0.002)	(0.185)	(10.7)	(0.001)	(0.047)	(3.5)	(0.000)	(0.041)
E'	10	21.2	0.997	0.023	32.1	0.996	0.046	2.4	1.000	0.010
		(18.6)	(0.002)	(0.051)	(9.5)	(0.001)	(0.063)	(3.2)	(0.000)	(0.030)

II

III