Helsinki University of Technology
Dissertations in Computer and Information Science
Espoo 2007 Report D22

ADVANCES IN VARIABLE SELECTION AND VISUALIZATION METHODS FOR ANALYSIS OF MULTIVARIATE DATA

Timo Similä

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 19th of October, 2007, at 12 o'clock noon.

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science

Distribution:
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400
FI-02015 TKK
FINLAND
Tel. +358-9-451 3272
Fax +358-9-451 3277
http://www.cis.hut.fi

Available in PDF format at http://lib.tkk.fi/Diss/2007/isbn9789512289301

ISBN 978-951-22-8929-5 (printed version)
ISBN 978-951-22-8930-1 (electronic version)
ISSN 1459-7020

Multiprint Oy/Otamedia
Espoo 2007

Similä, T. (2007): **Advances in variable selection and visualization methods for analysis of multivariate data**. Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D22, Espoo, Finland.

Keywords: machine learning, dimensionality reduction, regression, information visualization, variable selection

ABSTRACT

This thesis concerns the analysis of multivariate data. The amount of data that is obtained from various sources and stored in digital media is growing at an exponential rate. The data sets tend to be too large in terms of the number of variables and the number of observations to be analyzed by hand. In order to facilitate the task, the data set must be summarized somehow. This work introduces machine learning methods that are capable of finding interesting patterns automatically from the data. The findings can be further used in decision making and prediction. The results of this thesis can be divided into three groups.

The first group of results is related to the problem of selecting a subset of input variables in order to build an accurate predictive model for several response variables simultaneously. Variable selection is a difficult combinatorial problem in essence, but the relaxations examined in this work transform it into a more tractable optimization problem of continuous-valued parameters. The main contribution here is extending several methods that are originally designed for a single response variable to be applicable with multiple response variables as well. Examples of such methods include the well known lasso estimate and the least angle regression algorithm.

The second group of results concerns unsupervised variable selection, where all variables are treated equally without making any difference between responses and inputs. The task is to detect the variables that contain, in some sense, as much information as possible. A related problem that is also examined is combining the two major categories of dimensionality reduction: variable selection and subspace projection. Simple modifications of the multiresponse regression techniques developed in this thesis offer a fresh approach to these unsupervised learning tasks. This is another contribution of the thesis.

The third group of results concerns extensions and applications of the self-organizing map (SOM). The SOM is a prominent tool in the initial exploratory phase of multivariate analysis. It provides a clustering and a visual low-dimensional representation of a set of high-dimensional observations. Firstly, an extension of the SOM algorithm is proposed in this thesis, which is applicable to strongly curvilinear but intrinsically low-dimensional data structures. Secondly, an application of the SOM is proposed to interpret nonlinear quantile regression models. Thirdly, a SOM-based method is introduced for analyzing the dependency of one multivariate data set on another.

Similä, T. (2007): **Menetelmiä muuttujien valintaan ja informaation visualisointiin moniulotteisen tietoaineiston analysoinnissa**. Tohtorin väitöskirja, Teknillinen korkeakoulu, Dissertations in Computer and Information Science, Raportti D22, Espoo, Suomi.

Avainsanat: koneoppiminen, ulotteisuuden pienentäminen, regressio, informaation visualisointi, muuttujien valinta

TIIVISTELMÄ

Tämä väitöskirja käsittelee moniulotteisen tietoaineiston analysointia. Lukuisista lähteistä peräisin olevien, digitaaliseen muotoon tallennettujen tietoaineistojen määrä kasvaa eksponentiaalisesti. Aineistot ovat usein hyvin isoja sekä havaintokertojen että mitattujen muuttujien lukumäärän suhteen. Jotta analysointi onnistuisi, aineistoa täytyy redusoida. Tässä työssä tutkitaan koneoppimisen menetelmiä, joilla voidaan automaattisesti löytää mielenkiintoisia piirteitä tietoaineistosta. Löydöksiä voidaan käyttää edelleen päätöksenteossa ja tilastollisessa ennustamisessa. Väitöskirjan tulokset voidaan jakaa kolmeen ryhmään.

Ensimmäinen tulosten ryhmä liittyy syötemuuttujien valintaan regressiotehtävässä, jossa useita vastemuuttujia pyritään ennustamaan samanaikaisesti. Muuttujien valinta on luonteeltaan hankala kombinatorinen ongelma, mutta väitöskirjassa tutkitut relaksaatiot muuntavat sen yksinkertaisemmaksi jatkuva-arvoisten parametrien optimointiongelmaksi. Tähän liittyvä väitöskirjan merkittävä kontribuutio on lukuisten yksivastemenetelmien laajentaminen siten, että niitä voidaan käyttää myös useiden vastemuuttujien ennustamiseen. Lasso-estimaatti ja least angle regression -algoritmi ovat esimerkkejä tällaisista yksivastemenetelmistä.

Toinen tulosten ryhmä koskee ohjaamatonta muuttujien valintaa, jossa kaikkia muuttujia käsitellään samalla tavalla tekemättä eroa syöte- ja vastemuuttujien välille. Tehtävänä on löytää muuttujat, jotka ovat tavalla tai toisella informatiivisia. Läheinen ongelma, jota väitöskirjassa myös tarkastellaan, on muuttujien valinnan ja aliavaruusprojektion yhdistäminen. Nämä ovat kaksi tärkeintä ulotteisuuden pienentämisen kategoriaa. Väitöskirjassa kehitetyt usean vastemuuttujan regressiomenetelmät tarjoavat pienin muunnoksin uudenlaisen lähestymistavan näihin ohjaamattoman oppimisen ongelmiin, mikä on tärkeä työn kontribuutio.

Kolmas tulosten ryhmä koostuu itseorganisoivan kartan (SOM) laajennuksista ja sovelluksista. SOM on käyttökelpoinen työkalu moniulotteisen tietoaineiston alustavassa, tutkiskelevassa analyysissä. Se tuottaa tietoaineistolle ryhmittelyn ja havainnollisen, matalaulotteisen esitysmuodon. Ensiksi väitöskirjassa esitetään SOM:n laajennus, joka soveltuu erityisesti voimakkaasti kaarevien tai mutkikkaiden mutta sisäisesti matalaulotteisten rakenteiden analysointiin tietoaineistossa. Toiseksi esitetään SOM:n sovellus, joka helpottaa epälineaaristen kvantiiliregressiomallien tulkintaa. Kolmanneksi esitetään SOM-pohjainen menetelmä, jolla voidaan tutkia moniulotteisen tietoaineiston riippuvuuksia jostakin toisesta moniulotteisesta aineistosta.

Preface

This work has been carried out at the Laboratory of Computer and Information Science (CIS) at Helsinki University of Technology. The work has been funded by the CIS laboratory and from the beginning of 2006 by the Graduate School in Computational Methods of Information Technology (ComMIT). In the early period, I also participated the Helsinki Graduate School in Computer Science and Engineering (HeCSE) without funding. The two graduate schools are acknowledged for their support. Some parts of the work have been funded by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. I am also grateful for the two personal grants received from the Technological Foundation of Finland (Tekniikan edistämissäätiö) and the one from the Nokia Foundation.

Prof. Olli Simula is the supervisor of this thesis. I am thankful for the trust and friendship that he has shared with me over the years. I have been privileged to choose my research topics and daily working practices without any formal responsibilities. This amount of freedom is rare even in the academic field, but it has suited me perfectly. I wish to thank Prof. Erkki Oja who, as the head of the CIS laboratory, has provided excellent facilities to do research.

I would like to thank the co-authors of the publications of this thesis Dr. Sampsa Laine and Mr. Jarkko Tikka. Sampsa deserves an extra thank for being the person who hired me to the CIS laboratory and instructed the early steps of my career. Some of the strongest publications of this thesis are joint works with Jarkko. I also thank my other closest co-workers Mr. Mika Sulkava and Mr. Pasi Lehtimäki for their enjoyable company.

I have received many comments on different versions of the manuscript of this thesis. Particularly, I thank Dr. Jaakko Hollmén, Dr. Francesco Corona, Mr. Jarkko Tikka, Mr. Mika Sulkava, and Mr. Elia Liitiäinen for their valuable help in improving the text. I wish to express my gratitude to the pre-examiners of the thesis Prof. Risto Ritala and Dr. Patrik Hoyer. Dr. Volker Tresp has kindly agreed to act as an opponent at the defense.

My parents have supported me in many ways during this research, which I truly appreciate. Most of all, I want to thank Piia for making my life a happy one.

Espoo, September 2007

Timo Similä

Contents

Publications of the thesis

I Timo Similä and Sampsa Laine (2005). Visual approach to supervised variable selection by self-organizing map, *International Journal of Neural Systems* **15**(1–2):101–110.

II Timo Similä (2005). Self-organizing map learning nonlinearly embedded manifolds, *Information Visualization* **4**(1):22–31.

III Timo Similä and Jarkko Tikka (2005). Multiresponse sparse regression with application to multidimensional scaling, *in* W. Duch, J. Kacprzyk, E. Oja and S. Zadrozny (eds), *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, 15th International Conference, Proceedings, Part II*, Vol. 3697 of *Lecture Notes in Computer Science*, Springer, pp. 97–102.

IV Timo Similä (2006). Self-organizing map visualizing conditional quantile functions with multidimensional covariates, *Computational Statistics & Data Analysis* **50**(8):2097–2110.

V Timo Similä and Jarkko Tikka (2006). Common subset selection of inputs in multiresponse regression, *Proceedings of the IEEE International Joint Conference on Neural Networks - IJCNN 2006*, pp. 1908–1915.

VI Timo Similä (2007). Majorize-minimize algorithm for multiresponse sparse regression, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2007*, Vol. II, pp. 553–556.

VII Timo Similä and Jarkko Tikka (2007). Input selection and shrinkage in multiresponse linear regression, *Computational Statistics & Data Analysis* **52**(1):406–422.

Abbreviations and notations

Lars	least angle regression
Lasso	least absolute shrinkage and selection operator
MDS	multidimensional scaling
MRSR	multiresponse sparse regression
OLS	ordinary least squares
SOM	self-organizing map
SVS	simultaneous variable selection

$\boldsymbol{X}$	$n \times m$ matrix whose columns $\boldsymbol{x}_i$ denote variables and rows $\underline{\boldsymbol{x}}_i$ observations
$\boldsymbol{Y}$	$n \times q$ matrix whose columns $\boldsymbol{y}_i$ denote variables and rows $\underline{\boldsymbol{y}}_i$ observations
$\boldsymbol{E}$	$n \times q$ matrix whose columns $\boldsymbol{e}_i$ denote noise vectors associated with $\boldsymbol{y}_i$
$\boldsymbol{Z}$	$n \times m$ matrix whose columns $\boldsymbol{z}_i$ are orthonormal
$\boldsymbol{W}$	$n \times m$ matrix of regression coefficients with the jth row $\boldsymbol{w}_j$
$\boldsymbol{I}$	identity matrix
$\boldsymbol{U}$	eigenvectors and/or left-singular vectors
$\boldsymbol{V}$	right-singular vectors
$\boldsymbol{D}$	diagonal matrix of eigenvalues
$\boldsymbol{S}$	matrix of singular values
t	nonnegative shrinking parameter
λ	nonnegative penalty parameter and/or Lagrange multiplier
$p(s)$	penalty function
$\mathcal{A}$	set of indices of active input variables
$\boldsymbol{\Omega}^{[k]}$	$m \times m$ diagonal matrix with the diagonal entries $p'(\Vert\boldsymbol{w}_j^{[k]}\Vert_2)/(\mu + \Vert\boldsymbol{w}_j^{[k]}\Vert_2)$
μ	positive parameter that measures the accuracy of an approximation
δ_{ij}	nonnegative dissimilarity value
$\boldsymbol{J}$	centering matrix
$\boldsymbol{\Delta}^{(2)}$	matrix of squared dissimilarities δ_{ij}^2
$\boldsymbol{m}_j$	prototype vector of the jth SOM node
$\boldsymbol{L}$	Laplacian matrix
θ	probability level between zero and one
$\rho_\theta(e)$	"check function" for the θth quantile
$Q_\theta(\underline{\boldsymbol{x}})$	model for the θth conditional quantile function
$\Vert\cdot\Vert_p$	p-norm of a vector
$\Vert\cdot\Vert_{\mathrm{F}}$	Frobenius norm of a matrix

Chapter 1

Introduction

1.1 Scope of the thesis

This thesis considers multivariate data analysis. Here the term data means any information that is already or can be transformed into numerical form. The data consist of observations, which are measured values of some variables. A data set is typically represented as a matrix whose rows correspond to the observations and columns to the variables. Data appear all around our society and a trend is that the sizes of the data sets are becoming larger all the time. For instance, a set of gene expression data can have tens of thousands of variables and some hundred observations. Evidently, such a data set cannot be reviewed by hand, but more sophisticated methods are needed for the analysis. Fortunately, another trend is the exponential growth of computing power, so there are excellent resources for designing and applying new methods.

The methods developed in this thesis are exploratory by nature. Exploratory data analysis is a branch of statistics, which aims to overview the data (Tukey, 1977). It differs from the classical statistics by generating novel hypotheses, instead of testing rigorously existing ones (Glymour et al., 1997). Little is known in the beginning, but new discoveries are made directly from the data. Information visualization is a field related to exploratory data analysis, which concerns representing some aspects of the data to the analyst. This is quite distinct from the traditional approach, where a domain expert knows in advance what is interesting in the data. The use of expert knowledge can be valuable, but it is not always available or it is too costly. Also, expert knowledge may be biased toward some established practice, which prevents discovering new phenomena. This thesis proposes novel machine learning methods, which aim at finding interesting patterns automatically.

Dimensionality reduction is the scope of the thesis. It means representing the data in a more compact form with as little loss of information as possible. The dimensionality of the data equals to the number of variables. Reducing the dimensionality means either selecting a subset of the variables or transforming (projecting)

the variables into a fewer number of new ones. These are the two main categories of dimensionality reduction: subset selection and subspace projection.

The objective of subset selection includes improving the prediction performance of the models, providing faster and more cost-effective models, and providing a better understanding of the underlying process that generated the data (Guyon and Elisseeff, 2003). In this thesis, variable selection in a regression problem is studied. Regression means predicting the value of a response variable given the values of input variables. In particular, this thesis presents methods for selecting a subset of the input variables, which is useful for predicting several response variables simultaneously. In addition to regression, unsupervised variable selection is considered. Unsupervised selection is distinguished from supervised selection by the fact that there is no a priori response, but all the variables are treated equally. Here the task is selecting principal variables that contain as much information as possible (McCabe, 1984).

Subspace projection shares the objective of subset selection with the difference that the results must be interpreted in terms of all the variables. The coordinate axes of the subspace are often called features and feature extraction denotes the process of finding the transformation that projects the data from the original space to the feature space. Unsupervised learning aims at discovering intrinsically low-dimensional structures that are embedded in high-dimensional data. This thesis proposes a method for the particular problem called manifold learning, where the embedding can be highly nonlinear. Projecting the data onto a two-dimensional feature space makes it possible to visualize the structure of the data. This thesis presents a visualization technique for interpreting nonlinear regression models.

1.2 Contributions of the thesis

The main scientific contributions of this thesis include the following.

- Many variable selection and shrinkage methods for single response regression are extended to be applicable with multiple response variables as well. Efficient algorithms for following the corresponding solution paths are proposed.
- Methods for unsupervised variable selection are proposed. In addition, methods that combine variable selection and subspace projection are introduced for unsupervised dimensionality reduction.
- Extensions and applications of the self-organizing map are presented.

The thesis consists of an introductory part and seven original publications. The publications, along with the contributions of the present author, are described more specifically in the next section.

The introductory part also contains some novel unpublished material from the present author. Firstly, the two theorems in Section 2.3.4 and their proofs in the Appendix establish the connection between the results of Publications V and VII. Secondly, variable selection for unsupervised dimensionality reduction is discussed in Publications III and V to some extent, but this topic is examined more extensively and it is presented in the unified framework of Publication VI in Section 2.4.

1.3 Contents of the publications and contributions of the present author

Publication I presents a nonparametric method for investigating the relationship between two sets of variables. The data are visualized using the self-organizing map (SOM) (Kohonen, 1982, 2001) in terms of response variables. The proposed method returns a subset of input variables that best corresponds to the description provided by the SOM. The method is based on nearest neighbor techniques and it is capable of depicting complex dependencies without assuming any particular model. This article is an extended version of an earlier paper (Laine and Similä, 2004), invited to a special issue of the journal from the ICONIP 2004 conference. The initial idea was developed jointly. While Dr. Laine was more active in preparing and writing the conference paper, the present author was almost solely responsible for the extended version. The extensions include a broader view in general, improvements to optimization, consideration of a smooth neighborhood kernel, and analysis of the kernel parameter.

Publication II introduces a modification of the SOM algorithm. The SOM has sometimes problems with data, which form a nonlinear manifold (e.g. a highly curvilinear surface) in a high-dimensional space. On the other hand, some more recently presented projection methods can handle these cases. The proposed technique is a hybrid of the SOM and one of these recent methods. It is shown quantitatively that the proposed method is capable of preserving local neighborhoods better than the SOM when there is a manifold geometry in the data. The article is the sole contribution of the present author.

Publication III extends the least angle regression algorithm by Efron et al. (2004) to multiple response variables. The proposed algorithm was named as the multi-response sparse regression (MRSR) algorithm. It is a forward selection method for linear regression, which updates the model with less greedy steps than traditional stepwise algorithms and is, thereby, less prone to overfitting. The present author suggested the idea originally, but the algorithm itself was developed jointly with Mr. Tikka. Mr. Tikka carried out the experiments, while the application to multidimensional scaling was outlined by the present author. The article was written together.

Publication IV considers the task of interpreting nonlinear quantile regression models. This is important in many fields of empirical research. It is proposed to use the SOM to visualize the parameters of the model, which helps understanding the dependency between the input variables and the response variable. The article is the sole contribution of the present author.

Publication V presents the MRSR algorithm in a more general framework, where the correlation criterion is defined by any vector norm. It is also shown that the outcome of the algorithm is unique, assuming implicitly that the selected inputs are linearly independent. Comparisons are extensive and they show the strengths of the MRSR against some other methods when the input variables are highly correlated. It is proposed to use the MRSR to select variables unsupervisedly in an application of image reconstruction. The present author was more responsible for the theoretical developments, including the proof of uniqueness, but this research

was continuously supported by Mr. Tikka. Image reconstruction was the present author's idea. Otherwise, the experiments were designed, carried out, and reported by Mr. Tikka. The present author wrote the other parts of the article.

Publication VI extends the majorize-minimize algorithm by Hunter and Li (2005) to multiple response linear regression models. The model fitting is penalized in a way that input selection and regularization occur simultaneously. The algorithm is applicable to any penalty function that is increasing, differentiable, and concave. In addition, an active set strategy is proposed for tracking input selection when the complete path of penalized solutions is computed. The active set strategy has computational advantages, because the optimization can be focused to a subset of parameters, while the solution remains the same. The article is the sole contribution of the present author.

Publication VII formulates the problem of input selection in multiple response linear regression as a convex optimization problem to minimize the error sum of squares subject to a sparsity constraint. The necessary and sufficient conditions for optimality and a condition that ensures the uniqueness of the solution are given. An interior point method is proposed for solving the problem. A predictor-corrector variant of the solver suits for following the solution path as a function of the constraint parameter. Extensive empirical comparisons are performed. The present author contributed the theory, implemented the algorithm, and wrote a large part of the article. Mr. Tikka provided insight into the theory and helped considerably in the way toward a workable implementation. Furthermore, Mr. Tikka was responsible for planning, conducting, and reporting the experiments.

1.4 Structure of the thesis

The rest of the introductory part of the thesis is organized as follows. Chapter 2 is devoted to variable selection. It starts with a review of necessary background knowledge on concepts like why variable selection is important, what variables should be selected, and what approaches exist to carry out the selection. After that, variable selection methods for single response linear regression are reviewed. Their counterparts for multiple response variables are introduced next. Finally, some approaches to unsupervised variable selection and other approaches that combine unsupervised variable selection and subspace projection are discussed. These techniques are strongly inspired by the regression methods that are studied in the previous sections of Chapter 2. Altogether, Chapter 2 positions Publications III and V–VII with respect to related work conducted by other researchers.

Chapter 3 provides a short tutorial on the visualization process based on the SOM. Three sections follow the tutorial, each of which introduces one application of the SOM according to Publications I, II, and IV. Firstly, an extension of the SOM to strongly curvilinear but intrinsically low-dimensional data structures is introduced. Secondly, a SOM-based framework is examined for interpreting nonlinear quantile regression models. Thirdly, a method is proposed for analyzing the dependency of one multivariate data set on another. One set is visualized using the SOM and the task is to find the variables of the other set that are related to the

visualization. This application bears similarity to the variable selection problem for multiresponse regression, discussed in Section 2.3, but differs in having more emphasis on data exploration than accurate prediction. Related articles from literature are reviewed in connection with the proposed methods in Chapter 3.

Chapter 4 concludes the introductory part of the thesis and provides some directions for future research. The Appendix consists of a pseudocode description of the MRSR algorithm and the proofs of the two theorems in Section 2.3.4.

Chapter 2

Variable selection for regression problems

2.1 General concepts

This section gives an overview of some general issues in relation to supervised variable selection. The issues include noting the difference between relevant and useful variables, and motivating variable selection as one way to tackle the problem of overfitting, which occurs in using a too complex model for the data. Some guidelines are examined to select the best combination of variables and the potential instability of this process is also discussed. Finally, three major approaches are introduced to implement the selection.

2.1.1 Relevant variables and useful variables

A number of notions exist for the definition of relevance of a variable. See, for example, the review articles by Blum and Langley (1997) and Kohavi and John (1997). Relevance depends on one's goals but, in general, relevant variables are those that carry meaningful information about the problem at hand. John et al. (1994) point out the weakness of existing definitions and propose two degrees of relevance: weak and strong. Strong relevance implies that the variable is indispensable in the sense that it cannot be removed without loss of prediction accuracy. Weak relevance implies that the variable can sometimes contribute to prediction accuracy. A variable is relevant if it is either weakly or strongly relevant, and is irrelevant otherwise.

A common factor of most definitions of relevance is that they are based on full distributions of data and a theoretical model that uses the distributions optimally. In practical problems, a data set is available but the distributions are not known. An implementable model can also have a restricted hypothesis space so that it cannot take full advantage of the subset of relevant variables. As a consequence,

the variables that are useful to build a good model can be different from the variables that are relevant. Here the goodness of a model means its generalization accuracy. There is no guarantee that a relevant variable will necessarily be useful to the model, particularly, if the variables are redundant. A variable is redundant if its value can be calculated from other variables. On the other hand, sometimes even irrelevant variables can be useful. This is rare in practice, but the following example illustrates the possibility (Example 3 by Kohavi and John, 1997). Consider a perceptron classifier, which separates observations into two classes based on the sign of a linear combination between the input variables and the model parameters. If the perceptron does not have the so-called bias term, a constant-valued input variable turns out to be useful to this particular classifier, although being irrelevant to the classification task. Differences between the concepts of relevance and usefulness are discussed, for example, by John et al. (1994), Blum and Langley (1997), Kohavi and John (1997), and Guyon and Elisseeff (2003).

2.1.2 Overfitting and the curse of dimensionality

Overfitting means that the model is very accurate on training data, but it has poor accuracy on previously unseen test data. It occurs when the model is too complex compared with the true underlying source of the data. The curse of dimensionality is a special case, which occurs in learning from few observations in a high-dimensional space. The term curse of dimensionality is introduced by Bellman (1961) to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to a vector space. More observations are needed to obtain the same density of data when the dimensionality increases. The mean squared prediction error of a model can be decomposed into two terms, known as bias and variance (Geman et al., 1992). Incorrect and too simple models lead to high bias, whereas overfitting produces excessive variance. A straightforward way to reduce the variance is to get more observations, but it is typically impossible. Often, however, the variance can also be reduced by deliberately introducing a small amount of bias so that the net effect is a reduction in mean squared error. This can be achieved by constraining the flexibility of the model somehow. Variable selection is one possibility to do this.

Irrelevant variables usually reduce the generalization accuracy of a model. For instance, by including an excess of irrelevant input variables in a linear regression model the response variable becomes predicted exactly in terms of training data. This is hardly a good model for test data if the training observations of the response variable contained any noise. Relevant but highly correlated variables are also problematic, for instance, by causing instability to the model parameters. The instability can lead to relatively large parameters in absolute values, which makes the model more sensitive to outliers (erroneous observations that are distant from the rest of the data). In general, focusing on the relevant variables does not save from overfitting. Kohavi and John (1997) give an example of a medical diagnosis test, where one of the variables is the patient's social security number. This variable is relevant in the sense that it alone solves any classification task applied to the measured patients. However, given only this variable, any practical model is expected to generalize poorly on measurements taken from a new patient.

2.1.3 Model selection and model assessment

Model selection means estimating the generalization accuracy of different models in order to choose the best one (Hastie et al., 2001). Variable selection is an integral part of model selection, since all combinations of variables correspond to different models. From this point of view, model selection aims at finding the subset of useful variables. Model selection is, however, a broader concept including the optimization of model parameters and fixing the values of hyperparameters. An example of a hyperparameter is a weighting factor that controls the tradeoff between the fit on the training set and the simplicity of the model. The hyperparameters are often difficult to tune together with the model parameters, so their values are kept constant when the model parameters are optimized. Different values of the hyperparameters are tried, followed by the optimization of the model parameters, until generalization accuracy becomes maximized.

Training error is usually a too optimistic measure of generalization accuracy, unless the training set is very large compared to the model complexity. In order to approximate the generalization accuracy, the data set can be divided into three parts: a training, a validation, and a test set. The training set is used for fitting the models, the validation set is used for estimating generalization accuracy in model selection, and the test set is used for evaluating the generalization accuracy of the selected model. In some cases the data set is too small to be split into three parts. Then the validation step can be approximated by various hold out techniques including cross-validation and the bootstrap (Efron and Tibshirani, 1993). Another approach is to use theoretical bounds that estimate the optimism in training error (Vapnik, 1998).

Model assessment, or performance prediction, means estimating the generalization accuracy of the selected model (Hastie et al., 2001). In most practical problems, it is not sufficient to provide a good model, it is important to predict how well it will perform on new unseen data (Guyon et al., 2006). Model assessment is essential to perform on an independent test set, which is not used in model selection. This way the estimate includes the uncertainty of model selection.

2.1.4 Variance in variable selection

Many variable selection techniques are sensitive to small perturbations of the data (Breiman, 1996). For instance, a different subset is found to be useful in different replicates of cross-validation. This causes instability to model selection. At worst, the chosen model can have much lower generalization accuracy than the optimal model would have. Even if the generalization accuracy is the same for models with different subsets of variables, this deviation is undesirable, because variance is often the symptom of a bad model, results are not reproducible, and one subset fails to capture the whole picture (Guyon and Elisseeff, 2003).

To stabilize model selection, several model averaging techniques have been proposed (Breiman, 1996; Brown et al., 2002; Kadane and Lazar, 2004). When many models are aggregated the risk of choosing a single bad model is avoided. Although model averaging usually improves generalization accuracy and stability, its draw-

back is that it does not necessarily lead to a reduced set of variables. The variable selection process itself can also be stabilized directly. The nonnegative garrote (Breiman, 1995) and the lasso (Tibshirani, 1996) are seminal methods that select variables in a non-discrete way as a function of a continuous-valued shrinking parameter. The smoothness of the process has the consequence that the methods do not suffer from high variability of subset selection.

2.1.5 Different approaches to variable selection

Many approaches have been proposed over the years for dealing with the problem of variable selection. One way to categorize the approaches is to consider the problem from the frequentist and Bayesian perpectives. This thesis belongs to the frequentist school. Readers interested in comparing the two perspectives are encouraged to see the review articles by George (2000) and Kadane and Lazar (2004). The frequentist perspective can be further divided into three categories: filter, wrapper, and embedded approaches (Blum and Langley, 1997; Guyon and Elisseeff, 2003).

The filter approach has two distinct steps. Firstly, irrelevant variables are discarded and, secondly, a model is fitted with the selected variables (John et al., 1994). Possible criteria for guiding the selection are, for example, mutual information (Rossi et al., 2006) and the Gamma statistic (Jones, 2004). The main disadvantage of the filter approach is that it searches for variables that are rather relevant than useful. Thus, it totally ignores the effects of the selected variable subset on the performance of the final predictive model. Filters are more like preprocessors, which can be used with any model to reduce the number of variables. Sometimes filters are used without the modeling step only to explore the data and detect relevant variables as shown in Publication I. Filters are typically computationally efficient, at least compared with wrappers.

The wrapper approach considers the model as a black box and uses only its interface. Different variable subsets are browsed using the estimated generalization accuracy of the model as the measure of usefulness for a particular variable subset (Kohavi and John, 1997). Wrappers can be coupled with any model and the approach usually leads to higher generalization accuracy than using filters. A drawback is that the computational load of the wrapper approach can be unbearable.

The embedded approach incorporates variable selection as a part of model fitting and the selection technique is specific to the model. The external search algorithms that are used in the filter and wrapper approaches cannot cover all possible variable combinations, excluding problems with only a few variables. Thereby, their solutions are likely to be suboptimal. On the contrary, global optimality is often easier to guarantee in the embedded approach. Embedded methods are typically computationally more efficient than wrappers. Publications III and V–VII consider embedded methods.

2.2 Single response linear regression

Subset selection in linear regression is a well studied area. The first wave of important developments occurred in the 1960s when computing was expensive. The focus on the linear model still continues, in part because its analytic tractability facilitates insight, but also because many tasks can be posed as linear variable selection problems (George, 2000). A comprehensive treatment of variable selection methods is provided by Miller (2002). Here the emphasis is on the most recent developments and such formulations of the problem, where a global solution can be computed without trying all possible combinations.

Suppose that $\boldsymbol{y}$, a response variable, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, a set of potential input variables, are vectors of n observations. For the sake of simplicity, we assume that all the variables are standardized to zero mean and unit variance. The columns of the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$ correspond to the input variables and rows to the observations. Linear regression means estimating $\boldsymbol{w}$, a vector of m parameters of the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{e}$ is an unknown noise vector of n elements. Model (2.1) does not include the bias term, because it is not needed under the assumption that the variables are standardized.

A traditional approach is to assume the noise independently normally distributed. In that case, minimizing the error sum of squares produces the maximum likelihood estimate. However, it is an ill-posed inverse problem whenever the columns of $\boldsymbol{X}$ are linearly dependent. This happens, for instance, when $n < m$. Even worse, if $\boldsymbol{w}$ is sparse in the sense that some parameters are exactly zero, then the estimation process has a combinatorial aspect. In practical problems, the data analyst may not have any prior knowledge about the sparsity of $\boldsymbol{w}$. Note that the parameters with zero value correspond to the input variables that do not contribute to the response variable at all, so identifying them is important. Subset selection means identifying the nonzero parameters and shrinking means placing some constraints on the parameters. Both techniques can be used to tackle the ill-posedness of the problem.

2.2.1 Best subset regression

Best subset regression is perhaps the most traditional formulation of variable selection. It means finding the best combination of k input variables in terms of the error sum of squares

$$\underset{\boldsymbol{w}}{\text{minimize}} \ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \ \text{ subject to } \ \|\boldsymbol{w}\|_0 = k. \tag{2.2}$$

Here $\|\cdot\|_2^2$ denotes the sum of squares of the elements in a vector and $\|\cdot\|_0$ denotes the number of nonzero entries in a vector. The number of different combinations of at least one variable given m variables is $2^m - 1$. The straightforward approach to compute all these combinations is impossible unless m is small, say, a few dozen.

The branch and bound strategy by Furnival and Wilson (1974) with some later improvements (Ni and Huo, 2006) is the current state-of-the-art for solving (2.2) for the complete sequence of k from 1 to m. It utilizes the fact that adding a new variable to the model can never increase the value of the objective, so some of the subsets can be discarded from the search. The computational burden may still be unbearable in many applications. Approximate solutions to (2.2) can be easily obtained using forward selection, backward elimination, other stepwise methods, or genetic algorithms (Miller, 2002; Oh et al., 2004). However, these techniques are notorious for failing to find the optimal subset when the input variables are highly correlated (Derksen and Keselman, 1992). Besides the difficulties in solving problem (2.2), another major deficiency of best subset regression is its weak stability (Breiman, 1996).

2.2.2 Regression shrinkage and selection

The following list briefly introduces combined shrinkage and selection methods. As a common denominator, the variable selection process is formulated as a continuous-valued optimization problem. When the problem is convex a globally optimal solution is easy to obtain. Shrinking stabilizes the estimate, so these methods offer a remedy to the weaknesses of best subset regression.

Nonnegative garrote. Breiman (1995) proposes the nonnegative garrote as a strategy for improving the ordinary least squares (OLS) solution

$$\boldsymbol{w}^{\text{OLS}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.3}$$

so that the strengths of both ridge regression (Hoerl and Kennard, 1970) and subset selection would be present. In ridge regression, the term $\lambda\|\boldsymbol{w}\|_2^2$ is added to the cost function and the value of $\lambda \geq 0$ is tuned in order to find the optimal penalization of the parameters. It is a very stable shrinking method, but it does not select variables (Breiman, 1996). In the garrote estimate, in turn, each parameter of the OLS solution has a multiplier and the multipliers are determined by the solution to the problem

$$\underset{\boldsymbol{c}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{w}^{\text{OLS}} \cdot \boldsymbol{c})\|_2^2 \ \text{ subject to } \ c_j \geq 0, \ \sum_{i=1}^{m} c_j \leq t. \tag{2.4}$$

The expression $\boldsymbol{w}^{\text{OLS}} \cdot \boldsymbol{c}$ denotes the entrywise product of the two vectors. By decreasing $t \geq 0$, more of the c_j become zero and those that are still positive shrink the OLS estimate. Importantly, the solutions form a continuous path as a function of t such that the task of subset selection simplifies to the task of fixing the value of this parameter. A drawback of the garrote is that the solution depends on both sign and magnitude of the OLS estimate. When the OLS estimate is poor, the garrote may suffer as a result. The OLS solution does not even exist when the input variables are linearly dependent.

Lasso. Tibshirani (1996) proposes the lasso (least absolute shrinkage and selection operator) estimate, which can be seen as an improvement of the nonnegative

garrote. The estimate solves the problem[1]

$$\underset{\boldsymbol{w}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 \ \text{ subject to } \ \|\boldsymbol{w}\|_1 \leq t. \tag{2.5}$$

Selection and shrinkage occur simultaneously, and contrary to (2.4), the estimate does not depend on the OLS solution. Osborne et al. (2000) and Efron et al. (2004) show that the solution path is piecewise linear as a function of t and give efficient algorithms for tracking this path. Efron et al. (2004) also derive the least angle regression (lars) algorithm, which nearly yields the lasso path but is a more simple forward selection procedure. A slight modification of the lars algorithm follows the lasso path exactly. Surprisingly, under certain conditions, the lasso estimate correctly identifies the variables that are effective in the best subset regression problem (2.2) (Fuchs, 2005; Donoho et al., 2006; Huo and Ni, 2007; Tropp, 2006b). As a consequence, the difficult combinatorial problem can be relaxed to a continuous-valued convex problem in some cases. The conditions are, however, too detailed to be considered in this work.

Extensions of the lasso. Zou and Hastie (2005) define the naïve elastic net estimate by the solution to the problem

$$\underset{\boldsymbol{w}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 \ \text{ subject to } \ (1-\alpha)\|\boldsymbol{w}\|_1 + \alpha\|\boldsymbol{w}\|_2^2 \leq t. \tag{2.6}$$

The parameter $\alpha \in (0,1)$ controls a convex combination of the lasso and ridge penalties and $t \geq 0$ controls the amount of shrinkage. The elastic net fixes two shortcomings of the lasso. Firstly, the lasso tends to select only one variable from a group of highly correlated variables, whereas the elastic net tends to give similar values (up to a change of sign if negatively correlated) for all the parameters within the group. This way the variables that are relevant to the underlying source of data are more likely the same as the variables that are useful to build the model. Secondly, the lasso selects at most n variables in the $n < m$ case, while the elastic net can select more than that. Zou and Hastie (2005) propose an efficient lars-en algorithm for computing the elastic net solution path, much like the algorithm lars does for the lasso.

Bakin (1999) and, more recently, Yuan and Lin (2006) consider an extension of the lasso to select variables at the group level. Contrary to the elastic net, the m input variables are divided into $\widetilde{m}$ non-overlapping groups beforehand. The group lasso estimate solves the following problem

$$\underset{\boldsymbol{w}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 \ \text{ subject to } \ \sum_{j=1}^{\widetilde{m}} \|\boldsymbol{w}_j\|_2 \leq t, \tag{2.7}$$

where $\boldsymbol{w}_j$ denotes the parameters of the jth group. As t decreases, more of the block norms $\|\boldsymbol{w}_j\|_2$ become zero and the corresponding groups of variables vanish from the model. Several approaches have been proposed to minimize a loss function

[1] The constrained minimization of $f(\boldsymbol{w})$ subject to $g(\boldsymbol{w}) \leq t$ is equivalent to the penalized minimization of $f(\boldsymbol{w}) + \lambda g(\boldsymbol{w})$ in the sense that for any $\lambda \geq 0$ there exists $t(\lambda) \geq 0$ such that $\boldsymbol{w}(\lambda)$ solves the both problems, and vice versa, as long as $f(\boldsymbol{w})$ and $g(\boldsymbol{w})$ are convex and $\exists \boldsymbol{w}_0$ such that $g(\boldsymbol{w}_0) < t$ (Hiriart-Urruty and Lemaréchal, 1996, chap. VII). The lasso problem and its various successors are often presented in the penalized form.

subject to the group lasso constraint, for instance, by Park and Hastie (2006), Meier et al. (2006), Kim et al. (2006), and Publications VI–VII.

Fan and Li (2001) consider a penalized least squares problem of the form

$$\underset{\boldsymbol{w}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda \sum_{j=1}^{m} p(|w_j|), \tag{2.8}$$

where the penalty function $p(s)$ is chosen such that the estimate satisfies the criteria of unbiasedness, sparsity, and continuity. For instance, the lasso penalty $p(s) = s$ does not satisfy the first criterion, because the estimate is biased when the true unknown value $|w_j|$ is large. In order to satisfy the criteria by Fan and Li (2001), $p(s)$ is necessarily strictly concave, which has the consequence that the objective in (2.8) may have local minima. Hunter and Li (2005) propose an algorithm, which is applicable to (2.8) with several choices of $p(s)$, and the algorithm converges to a stationary point of the objective function.

2.3 Multiresponse linear regression

The multiresponse linear regression model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{E}, \tag{2.9}$$

where the columns of the matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q]$ denote the response variables and the columns of the matrix $\boldsymbol{E} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_q]$ the corresponding noise vectors. The $m \times q$ matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m]^T$ denotes the regression coefficients. There are two approaches to multiresponse regression. Either a separate model is build for each response variable, or a single model is used to estimate all the responses simultaneously. The simultaneous estimation techniques have some advantages over the separate model building, especially when the responses are correlated (Breiman and Friedman, 1997; Srivastava and Solanky, 2003).

In the case of $\boldsymbol{w}_j = 0$, the jth input variable does not contribute to any of the response variables. In the following, we define the task of identifying and estimating the nonzero rows of $\boldsymbol{W}$ as simultaneous variable selection (SVS). Note that input selection in the separate model building approach does not necessarily lead to a reduced set of variables for predicting several response variables. This is uneconomical if the input variables have measurement costs, and it also complicates model interpretation. In the subsequent four sections, we investigate SVS techniques for the estimation of several response variables.

2.3.1 Regression shrinkage and selection

Various approaches have been proposed to extend the lasso estimate (2.5) to many response variables. Turlach et al. (2005) and Tropp (2006a) consider the L_∞-SVS problem

$$\underset{\boldsymbol{W}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 \ \text{ subject to } \ \sum_{j=1}^{m} \|\boldsymbol{w}_j\|_\infty \leq t \tag{2.10}$$

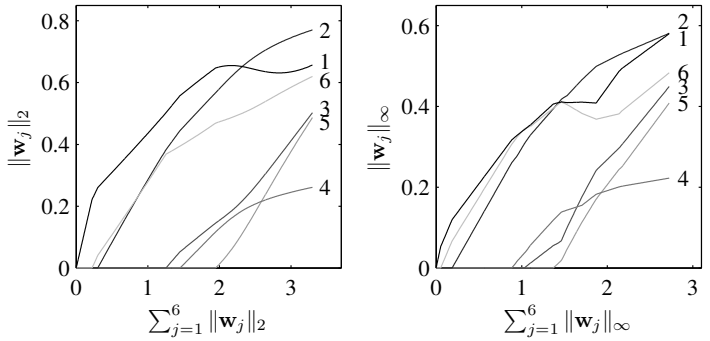

Figure 2.1: The SVS paths for Tobacco data whose variables are scaled to zero mean and unit variance: (*left*) L_2-SVS, (*right*) L_∞-SVS. The vertical axis shows the row norms of the coefficient matrix and the horizontal axis shows the aggregate of the row norms (left-hand sides of the constraints in (2.10) and (2.11)). The bigger the norm, the more important is the corresponding input variable.

and Malioutov et al. (2005), Cotter et al. (2005), and Publication VII consider the L_2-SVS problem

$$\underset{\boldsymbol{W}}{\text{minimize}} \ \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\text{F}}^2 \ \text{ subject to } \ \sum_{j=1}^{m} \|\boldsymbol{w}_j\|_2 \leq t. \tag{2.11}$$

In (2.10) and (2.11), $\|\cdot\|_{\text{F}}^2$ denotes the squared Frobenius norm of a matrix defined as the sum of squares of all entries of the matrix. As the parameter $t \geq 0$ decreases, more of the row norms become zero. This has the effect that some input variables are completely dropped from the SVS estimate and the regression coefficients associated with the rest of the inputs are shrunk toward zero. In the single response case, both SVS problems equal to (2.5).

To illustrate the SVS estimates, Fig. 2.1 shows the solution paths of (2.10) and (2.11) as a function of t for Tobacco data set (Anderson and Bancroft, 1952, p. 205), which has $m = 6$ inputs, $q = 3$ responses, and $n = 19$ observations. Both methods consider $\{1, 2, 6\}$ as the best subset of three input variables. The same subset is also recognized, for instance, by Sparks et al. (1985) and Bedrick and Tsai (1994) as the most important one. The results on simulated data in Publication VII show that the L_2-SVS estimate is better than the L_∞-SVS estimate both in terms of prediction accuracy and correctness of input selection. The situation might, of course, be different when the data points are generated in some other way. The results on real data sets in Publication VII do not bring out clear differences between the two estimates.

The L_2-SVS problem (2.11) is a convex cone programming problem. In Publication VII, it is shown that the solution is unique as long as the vectors $\boldsymbol{x}_j$, which correspond to the selected inputs, are linearly independent. This is less restricting than assuming linear independence of all the columns of $\boldsymbol{X}$. Furthermore, Publication VII proposes a predictor-corrector method for following the solution path of (2.11) as a function of t. A row-specific linear update is taken to predict the optimal $\boldsymbol{W}$ due to a change in t. The corrector step applies the barrier method

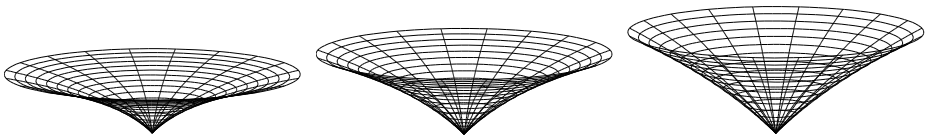

Figure 2.2: The mapping $p(\|\boldsymbol{w}_j\|_2)$ in the bivariate response case, where the penalty function is $p(s) = c\log(1+s/c)$. The value of c is doubled in each subfigure from left to right.

with Newton iterations to correct the prediction. The path following method is efficient due to two reasons. Firstly, the barrier method that is used in the corrector step takes advantage of the structure of the problem so that large matrix inversions are avoided. Secondly, the so-called active set of parameters is updated each time t changes so that only the rows of $\boldsymbol{W}$ that are likely to be nonzero are optimized. The nonzero rows are identified based on the property

$$\|(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T\boldsymbol{x}_j\|_2 \begin{cases} = \lambda, & j \in \{j : \boldsymbol{w}_j \neq 0\} \\ \leq \lambda, & j \in \{j : \boldsymbol{w}_j = 0\}, \end{cases} \tag{2.12}$$

which follows from the necessary and sufficient conditions for optimality as presented in Publication VII. The parameter λ denotes the Lagrange multiplier of the constraint in (2.11). It is readily available in the course of path following.

2.3.2 Bias reduction via concave penalization

Fan and Li (2001) propose the penalized least squares problem (2.8) to avoid unnecessary modeling bias in sparse estimation when the true unknown value $|w_j|$ is large. Publication VI extends this work to multiresponse regression by introducing the following novel penalized least squares problem

$$\underset{\boldsymbol{W}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\text{F}}^2 + \lambda\sum_{j=1}^{m} p(\|\boldsymbol{w}_j\|_2), \tag{2.13}$$

where $\lambda \geq 0$ is used to balance between model fitting and penalization. Any penalty function $p(s)$ that is increasing, differentiable, and concave on $s \geq 0$ can be used. It is shown in Publication VI that the penalty function encourages row sparsity only when its derivative is positive at the origin, that is $p'(0) > 0$ holds. The same condition has been derived for the single response case by Fan and Li (2001). Especially, the choice $p(s) = s$ makes (2.13) equivalent to the penalized formulation of the L_2-SVS problem (2.11). Fig. 2.2 depicts some possible penalty terms in (2.13). A strictly concave $p(s)$ may cause the objective in (2.13) to be nonconvex, or even worse, to have multiple local minima. This makes the search of a global minimum more difficult. On the other hand, such a penalty function has some advantages as demonstrated by the experiments in the end of this section.

Publication VI extends the majorize-minimize algorithm by Hunter and Li (2005) to solving (2.13). A general majorize-minimize algorithm operates on an auxiliary function, called the majorizing function. The majorizing function has to fulfill

two requirements. Firstly, it should touch the objective function at the supporting point and, secondly, it should never be below the objective function. Each iteration of the majorize-minimize algorithm updates the next supporting point to the minimizer of the current majorizing function.

The mapping $p(\|\boldsymbol{w}_j\|_2)$ is nondifferentiable at $\boldsymbol{w}_j = 0$ under the sparsity condition $p'(0) > 0$. Hence, the desired penalty function $p(s)$ is approximated by $p_\mu(s)$ so that the mapping $p_\mu(\|\boldsymbol{w}_j\|_2)$ becomes smooth. The smaller the value of the parameter $\mu > 0$, the more similar the approximation is to the desired function (see Hunter and Li (2005) and Publication VI for details). When $p_\mu(\|\boldsymbol{w}_j\|_2)$ is majorized by a quadratic function we have the following majorize-minimize algorithm

$$\widehat{\boldsymbol{W}}^{[k+1]} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{\Omega}^{[k]})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.14}$$

The matrix $\boldsymbol{\Omega}^{[k]}$ is computed at the current supporting point $\widehat{\boldsymbol{W}}^{[k]}$ as follows

$$\boldsymbol{\Omega}^{[k]} = \operatorname{diag}\left(\frac{p'(\|\widehat{\boldsymbol{w}}_1^{[k]}\|_2)}{\mu + \|\widehat{\boldsymbol{w}}_1^{[k]}\|_2}, \cdots, \frac{p'(\|\widehat{\boldsymbol{w}}_m^{[k]}\|_2)}{\mu + \|\widehat{\boldsymbol{w}}_m^{[k]}\|_2}\right). \tag{2.15}$$

For a fixed value of $\mu > 0$, algorithm (2.14) converges monotonically to a stationary point of the approximated version of the objective in (2.13). Taking $\mu \to 0$, algorithm (2.14) converges to a stationary point of the desired objective. These convergence results follow directly from general properties of the majorize-minimize algorithms (Lange, 1995; Hunter and Li, 2005). The parameter μ is essential for convergence, and it also has the practical benefit of avoiding singularities in (2.15). The regularized M-FOCUSS by Cotter et al. (2005) is a related algorithm, which equals to (2.14) under the choices $\mu = 0$ and $p(s) = \alpha^{-1}s^\alpha$ with $\alpha \in (0, 1]$.

It is shown in Publication VI that the following condition is necessary for optimality

$$\frac{\|(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T\boldsymbol{x}_j\|_2}{p'(\|\boldsymbol{w}_j\|_2)} \begin{cases} = \lambda, & j \in \{j : \boldsymbol{w}_j \neq 0\} \\ \leq \lambda, & j \in \{j : \boldsymbol{w}_j = 0\}. \end{cases} \tag{2.16}$$

Taking $p(s) = s$, condition (2.16) coincides with (2.12) under the interpretation that λ is the Lagrange multiplier. This confirms the connection between the L_2-SVS problem (2.11) and the more general penalized framework (2.13). Condition (2.16) can be used to track the nonzero rows of $\boldsymbol{W}$ when problem (2.13) is solved for a sequence of values of λ. Computational burden reduces, since only the rows that are likely to be nonzero need to be optimized. Similar strategies for tracking blockwise sparsity are used in Publication VII and by Park and Hastie (2006), but in both cases the penalty function is restricted to be linear, that is $p(s) = s$. Only differentiability of $p(s)$ and $p'(s) > 0$ for $s \geq 0$ is required in (2.16).

Fig. 2.3 depicts some results of applying the majorize-minimize algorithm to the whole range of relevant values of λ. Data are sampled as explained in Publication VI. In short, there are $n = 50$ observations, $m = 100$ inputs, and $q = 5$ responses. Only twenty inputs are effective in model (2.9) and the rest of the inputs are irrelevant. There are strong correlations among the inputs. The estimate mostly uses the relevant inputs for $\lambda \geq 1$, or at least the irrelevant ones have a small contribution, as shown by the two rightmost plots in Fig. 2.3. The plot on the left hand side in Fig. 2.3 illustrates condition (2.16). Note that $\|\boldsymbol{w}_j\|_2$ becomes nonzero at the moment the curve that represents the left hand side of (2.16) hits

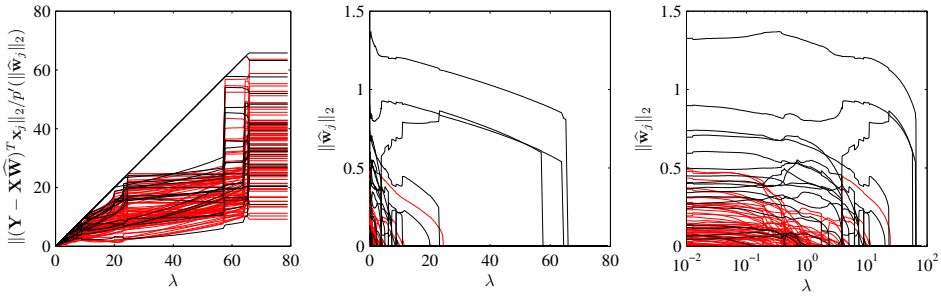

Figure 2.3: The path of penalized solutions for a single sampling of data when the penalty function $p(s) = c\log(1 + s/c)$ is used with $c = 0.4$: (*left*) illustration of condition (2.16), (*middle*) and (*right*) row norms of the coefficient matrix. Black curves correspond to relevant inputs and red curves to irrelevant ones.

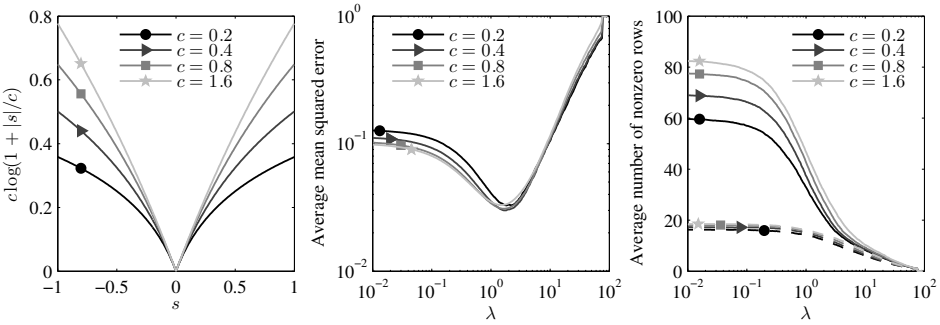

Figure 2.4: Results calculated from 500 replicates of sampling: (*left*) penalty function, (*middle*) average mean squared prediction error, (*right*) average number of selected inputs. The dashed lines show the average number of correct selections. The figure is taken from Publication VI.

the diagonal in the plot. Fig. 2.4 presents some further results, which are calculated from 500 replicates of sampling. The degree of concavity of the penalty function does not have a big influence on prediction error, which is the lowest at $\lambda = 2$ in all cases. However, high concavity favors parsimony. The number of relevant inputs in the selected subset is roughly the same in all cases, but the number of irrelevant inputs decreases as the degree of concavity increases.

2.3.3 The MRSR algorithm

Efron et al. (2004) propose the least angle regression (lars) algorithm for computing an entire path of regularized estimates, which nearly coincides with the solution path of the lasso problem (2.5). In the beginning, an input variable that correlates most with the response variable enters the model. The coefficient of the selected input is updated in the direction of the sign of its correlation until another input variable has the same absolute correlation with the current residuals as the selected one. From this point on, both coefficients are updated in the OLS direction of the

two inputs. The lars path is piecewise linear and a new input variable enters the model at each breakpoint. The path continues from a breakpoint in the direction of the OLS solution of the selected inputs. The absolute correlation between the residuals and each selected input is equal throughout the path and, moreover, larger than the correlation between any unselected input. The path is completed when the number of selected inputs reaches n, or when all m variables are selected and have attained the OLS solution. The entire sequence of breakpoints in the lars algorithm with $m < n$ inputs has a similar computational cost to the OLS estimation on m inputs.

Publications III and V propose the multiresponse sparse regression (MRSR) algorithm, which extends the lars algorithm to many response variables. In the MRSR algorithm, the correlation between an input variable and the residuals associated with all the responses is measured. The absolute correlation criterion of the lars algorithm is replaced with a norm of the vector of the q correlations. The correlation norms for the selected inputs are the same, and the largest, and this value is denoted by λ in the following presentation. The MRSR path $\widehat{\boldsymbol{W}}(\lambda)$ is piecewise linear as a function of λ and a new row becomes nonzero at each breakpoint $\lambda^{[k]}$. The MRSR algorithm generalizes the lars algorithm to the L_2-SVS problem (2.11) in the same way as the group lars algorithm by Yuan and Lin (2006) does to the group lasso problem (2.7). These connections are discussed more in Section 2.3.4.

The path begins at the point $\lambda^{[0]}$ for which we have $\widehat{\boldsymbol{W}}(\lambda^{[0]}) = 0$. This point and the set of active input variables are defined as follows,

$$\lambda^{[0]} = \max_{1 \le j \le m} \|\boldsymbol{Y}^T \boldsymbol{x}_j\|_\alpha \quad \text{and} \quad \mathcal{A}(\lambda^{[0]}) = \{j : \|\boldsymbol{Y}^T \boldsymbol{x}_j\|_\alpha = \lambda^{[0]}\}, \tag{2.17}$$

where $\alpha \ge 1$ denotes any vector norm. The case $\alpha = 1$ is proposed in Publication III and other choices are discussed in Publication V. From now on, we adopt the notation L_α-MRSR whenever it is necessary to specify the norm that is in use. The MRSR path is

$$\begin{aligned} \widehat{\boldsymbol{W}}_{\mathcal{A}^k}(\lambda) &= (\lambda/\lambda^{[k]})\widehat{\boldsymbol{W}}_{\mathcal{A}^k}(\lambda^{[k]}) + (1 - \lambda/\lambda^{[k]})\boldsymbol{W}_{\mathcal{A}^k}^{\text{OLS}} \\ \widehat{\boldsymbol{w}}_j(\lambda) &= 0, \quad j \notin \mathcal{A}(\lambda^{[k]}) \end{aligned} \tag{2.18}$$

in the segment $\lambda \in [\lambda^{[k+1]}, \lambda^{[k]}]$, where the subset OLS solution is

$$\boldsymbol{W}_{\mathcal{A}^k}^{\text{OLS}} = (\boldsymbol{X}_{\mathcal{A}^k}^T \boldsymbol{X}_{\mathcal{A}^k})^{-1} \boldsymbol{X}_{\mathcal{A}^k}^T \boldsymbol{Y} \tag{2.19}$$

and the subscript $\mathcal{A}^k$ defines the matrices

$$\widehat{\boldsymbol{W}}_{\mathcal{A}^k}(\lambda) = [\cdots \widehat{\boldsymbol{w}}_j(\lambda) \cdots]_{j \in \mathcal{A}(\lambda^{[k]})}^T \quad \text{and} \quad \boldsymbol{X}_{\mathcal{A}^k} = [\cdots \boldsymbol{x}_j \cdots]_{j \in \mathcal{A}(\lambda^{[k]})}. \tag{2.20}$$

The value of λ is decreased from $\lambda^{[k]}$ until a new index enters the active set

$$\mathcal{A}(\lambda) = \{j : \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{W}}(\lambda))^T \boldsymbol{x}_j\|_\alpha = \lambda\}. \tag{2.21}$$

This point is denoted by $\lambda^{[k+1]}$. It is shown in Publication V that there exists a unique value $\lambda_j \in [0, \lambda^{[k]}]$ such that $\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{W}}(\lambda_j))^T \boldsymbol{x}_j\|_\alpha = \lambda_j$ holds for any $j \notin \mathcal{A}(\lambda^{[k]})$. The largest value of λ_j for $j \notin \mathcal{A}(\lambda^{[k]})$ is the breakpoint $\lambda^{[k+1]}$ and the corresponding index joins the active set. As a consequence, the outcome of the

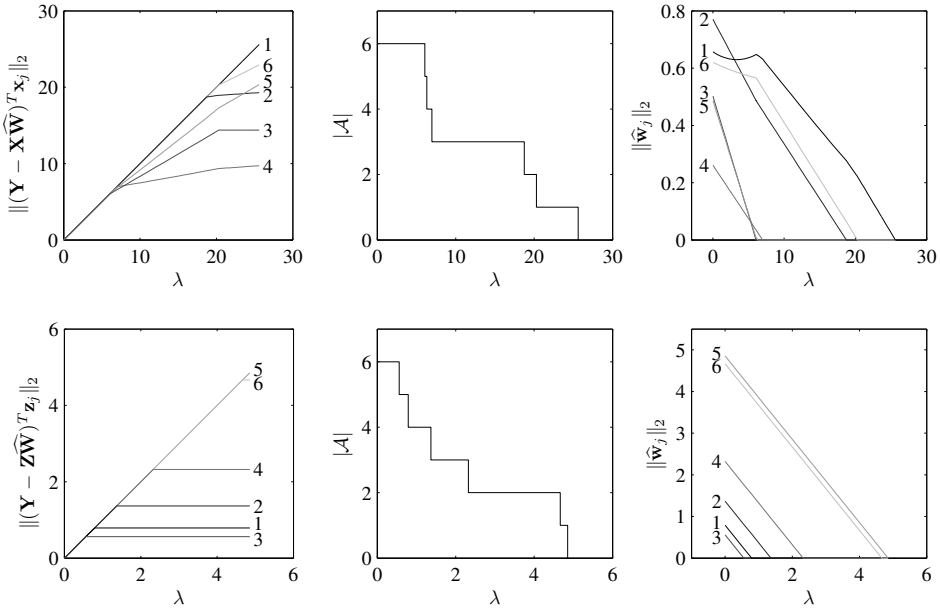

Figure 2.5: The L_2-MRSR path for Tobacco data: (*top*) all variables have zero mean and unit variance, (*bottom*) responses have zero mean and unit variance but inputs are also decorrelated, (*left*) correlation norms, (*middle*) number of selected inputs, (*right*) row norms of the coefficient matrix.

MRSR algorithm is unique, as long as the inverse in (2.19) is well defined, which actually happens when $\boldsymbol{X}_{\mathcal{A}^k}$ has a full column rank. It is possible, but unlikely, that λ_j equals $\lambda^{[k+1]}$ for several indices $j \notin \mathcal{A}(\lambda^{[k]})$, and then they are all added to $\mathcal{A}(\lambda)$ at $\lambda^{[k+1]}$. Pseudocode for the MRSR algorithm is presented in the Appendix.

To illustrate the algorithm, the upper panel in Fig. 2.5 shows the L_2-MRSR path for Tobacco data (Anderson and Bancroft, 1952, p. 205). A new index enters the set $\mathcal{A}(\lambda)$ each time the curve of a correlation norm hits the diagonal in the leftmost subfigure. At the same time, a new row of $\widehat{\boldsymbol{W}}(\lambda)$ becomes nonzero as shown in the rightmost subfigure. Comparisons in Publication V show that the MRSR algorithm is better than a greedy forward selection in terms of prediction accuracy and correctness of selection, especially when the input variables are highly correlated. On the other hand, the L_2-SVS estimate (2.11) is compared with the L_2-MRSR in Publication VII and the former is found to be better. The main strength of the MRSR algorithm is its computational efficiency compared with the SVS estimates that ensue from the minimization of some objective function. Using similar arguments as Efron et al. (2004) introduce for the lars algorithm, one can show that the computational load of the complete MRSR path is about the same order as an OLS fitting on m input variables. The computational load of the SVS paths is more difficult to quantify (see Publication VII for details). The capability of the MRSR to avoid overfitting is advantageous compared with the greedy algorithms.

2.3.4 The L_2-MRSR algorithm for orthonormal inputs

The lars algorithm can be modified such that it follows the solution path of the lasso problem (2.5) exactly (Efron et al., 2004). It is, therefore, natural to expect that the MRSR algorithm is tightly connected to some of the multiresponse shrinkage and selection estimates, but unfortunately, such a strong relationship does not seem to exist. However, if the columns of $\boldsymbol{X}$ are orthonormal, then the L_2-MRSR path (2.18) and the solution path of the L_2-SVS problem (2.11) coincide. In Publications V and VII, this connection is established by transforming (2.11) into a single response form, which is actually a particular form of the group lasso problem (2.7). Yuan and Lin (2006) state that their group lars algorithm follows the solution path of the group lasso problem under an orthonormality assumption on $\boldsymbol{X}$. It can be shown that the L_2-MRSR algorithm and the group lars algorithm are the same in this case.

For the sake of completeness, and because Yuan and Lin (2006) do not prove their statement rigorously, the following two theorems establish the connection between the L_2-MRSR algorithm and the L_2-SVS problem. Proofs can be found in the Appendix.

Theorem 2.3.1. *If $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ holds, then the L_2-MRSR path (2.18) can be written $\widehat{\boldsymbol{w}}_j(\lambda) = \max\{0, 1 - \lambda/\|\boldsymbol{Y}^T\boldsymbol{x}_j\|_2\}\boldsymbol{Y}^T\boldsymbol{x}_j$ for $\lambda \geq 0$ and $j = 1, \ldots, m$.*

Theorem 2.3.2. *If $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ holds, then the L_2-MRSR algorithm follows the solution path of the L_2-SVS problem (2.11) as a function of the constraint parameter under the mapping $t = \sum_{j=1}^{m} \max\{0, \|\boldsymbol{Y}^T\boldsymbol{x}_j\|_2 - \lambda\}$.*

The lower panel in Fig. 2.5 illustrates the L_2-MRSR path for Tobacco data whose variables are first normalized to zero mean and unit variance. Then the input variables are decorrelated (transformed into linearly uncorrelated variables) using the transformation $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{U}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{U}$ is the orthogonal matrix of eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{D}$ is the diagonal matrix of its eigenvalues. It is easy to check that $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}$ holds. The L_2-MRSR algorithm for the decorrelated input data does not make forward selection of the original input variables but increments a subset of principal components of the input data (Hotelling, 1933; Jolliffe, 1986) as λ decreases. Another, perhaps better justified, way to decorrelate the input data in regression problems is to extract the canonical components (Hotelling, 1936) instead of the principal components. Note that the path of $\|\widehat{\boldsymbol{w}}_j\|_2$ is linear for $\lambda \in [0, \|\boldsymbol{Y}^T\boldsymbol{z}_j\|_2]$ due to the decorrelation transformation, whereas correlation in the input data causes a crease at each breakpoint of the algorithm.

2.4 Other variable selection and feature extraction methods inspired by regression problems

Unsupervised learning examines multivariate data without explicit response variables or class labels to guide the analysis. In general, unsupervised learning seeks some internal structure of the data. Examples include cluster analysis (Jain and Dubes, 1988), which is the process of grouping similar data points together, and

visual analysis of high-dimensional data (de Oliveira and Levkowitz, 2003). Variable selection for unsupervised learning is important, because some of the variables may be redundant and some others may be irrelevant in terms of the cluster structure of the data. These variables hamper visual interpretation of the data and may cause misleading conclusions. Furthermore, useless variables cause degradation of accuracy and loss of running time efficiency in cluster analysis (Dy and Brodley, 2004; Wolf and Shashua, 2005), just to name few examples. Surprisingly, variable selection for unsupervised learning has been a bit overlooked in literature until recently (Dy and Brodley, 2004). Another interesting problem is finding such transformations of data that combine variable selection with further reduction of dimensionality. These issues are investigated in the next two sections.

2.4.1 Principal variable analysis

McCabe (1984) introduces the concept of principal variables of data. Principal variable analysis refers to the task of selecting a subset of the variables that contains, in some sense, as much information as possible. These principal variables are considered informative as themselves. Note the difference from principal component analysis, where the components are linear combinations of all the original variables (Hotelling, 1933; Jolliffe, 1986). Timely reviews of existing criteria for selecting the principal variables can be found from the articles by Al-Kandari and Jolliffe (2005) and Cumming and Wooff (2007). Just like in regression problems in Section 2.2.1, such criteria only rank combinations of variables and some stepwise algorithm is typically used to find promising combinations. Cadima et al. (2004) devote themselves to the combinatorial problem of identifying the optimal subset with respect to a given criterion. Heuristics seem to be the only possibility for many real problems unless the number of variables is small.

In contrast, the SVS framework to multiresponse regression offers another more tractable approach. We call the jth column of $\boldsymbol{X}$ as a principal variable if the $m \times m$ matrix $\boldsymbol{W}$ solves the problem

$$\underset{\boldsymbol{W}}{\text{minimize}}\ \tfrac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 + \lambda \sum_{j=1}^{m} p(\|\boldsymbol{w}_j\|_2) \tag{2.22}$$

and the jth row of $\boldsymbol{W}$ is nonzero. The first term of the objective function measures the error of a linear reconstruction of the data. Observe that principal component analysis can also be derived from this perspective[2]. Given a large enough λ, the second term in (2.22) has the effect that only some of the variables are used in the reconstruction. The identity matrix solves the problem at $\lambda = 0$ without any error. Principal variables, according to the above definition, are good at predicting many other variables.

Problem (2.22) is derived from (2.13) by substituting $\boldsymbol{Y} = \boldsymbol{X}$, so algorithm (2.14) and condition (2.16) are applicable. In the same way, any other SVS estimate could be extended to principal variable analysis. In Publication V, the MRSR algorithm

[2]Besides the maximum variance definition, the loadings of the principal components can be shown to minimize the reconstruction error $\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^T\|_{\mathrm{F}}^2$ under an orthonormality constraint on the columns of $\boldsymbol{W}$ (see, for example, Hastie et al., 2001).

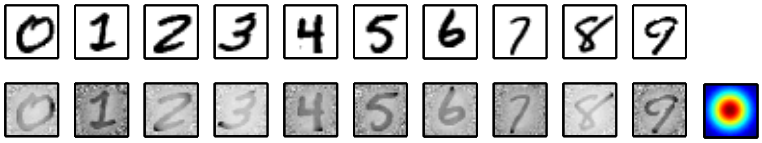

Figure 2.6: Images of handwritten digits: (*top*) sample images, (*bottom*) their normalized noisy versions, (*bottom right*) the color map of pixels that is used in Fig. 2.7.

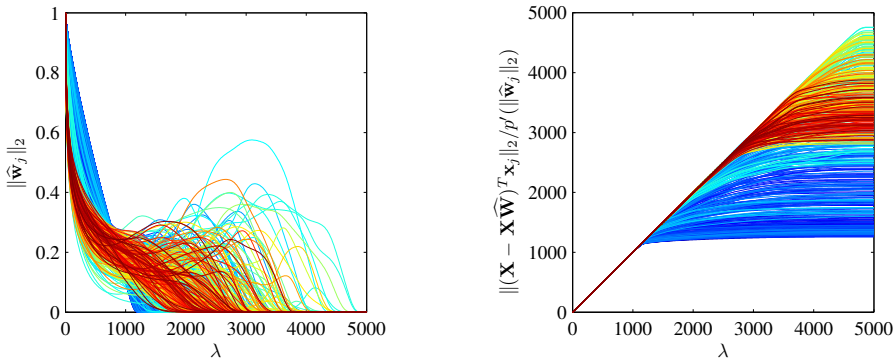

Figure 2.7: The path of penalized solutions in image reconstruction: (*left*) row norms of the coefficient matrix, (*right*) illustration of condition (2.16). The penalty function $p(s) = c\log(1 + s/c)$ is used with $c = 40$. See Fig. 2.6 for the connection between colors and pixels.

is used. If one still needs to reduce the dimensionality after extracting the principal variables, a low-rank approximation of $\boldsymbol{W}$ can be obtained by computing the singular value decomposition $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$. The matrix $\boldsymbol{U}$ has the same row sparse structure as $\boldsymbol{W}$. The columns of $\boldsymbol{U}$ associated with the largest singular values define an orthogonal projection to a low-dimensional subspace. The subspace only depends on the principal variables.

Image reconstruction is an application of principal variable analysis. Consider images of handwritten digits with 28×28 pixel grid[3]. Each row of $\boldsymbol{X}$ contains grayscale values of the $m = 784$ pixels of an image. The grayscale values are corrupted by Gaussian noise and they are normalized to zero mean and unit variance pixelwise. There are 100 images per digit and total $n = 1000$ images. Fig. 2.6 shows some sample images. The task is to reconstruct the complete images by a linear combination of some useful pixels. The normalization makes the task harder, since the useless pixels cannot simply be discarded by their low variance. The noise brings out the possibility of overfitting. Several variants of the MRSR algorithm are applied successfully to this task in Publication V. Instead, Fig. 2.7 shows the solution path of problem (2.22), which is computed by the majorize-minimize algorithm (2.14). The pixels in the middle enter to the model before the pixels close to the borders of an image as the value of λ is decreased. This

[3]The images are available from http://yann.lecun.com/exdb/mnist/.

is logical, since only the middle ones carry meaningful information. As shown in Publication V, greedy selection strategies are likely to fail in this task, because the nearby pixels are highly correlated.

2.4.2 Row sparsity in parametric multidimensional scaling

Multidimensional scaling (MDS) (Cox and Cox, 2001) methods aim at finding a configuration of data points in a vector space, usually Euclidean, such that the pairwise distances between the points match with given dissimilarities as well as possible. The dissimilarity values δ_{ij} can measure relations between objects of any type. This process is also known as embedding.

The necessary and sufficient condition for the values $\delta_{ij} = \delta_{ji} \geq 0$ to represent distances in a Euclidean space is that the matrix

$$\boldsymbol{B} = -\tfrac{1}{2}\boldsymbol{J}\boldsymbol{\Delta}^{(2)}\boldsymbol{J} \tag{2.23}$$

is positive semidefinite (Young and Householder, 1938; Torgerson, 1952). Here $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$ denotes the centering matrix and $\boldsymbol{\Delta}^{(2)}$ is the matrix of squared dissimilarities δ_{ij}^2 between all n objects. Assuming that $\boldsymbol{B}$ is positive semidefinite, the configuration $\boldsymbol{Y}$ that perfectly satisfies the dissimilarities can be recovered from the eigendecomposition $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ by setting $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}^{1/2}$. Furthermore, all eigenvalues are nonnegative and the number of nonzero values defines the lowest possible dimensionality, which is required to have an exact matching. In the classical MDS, the dimensionality of the perfect point configuration is reduced by using only a few, say q, columns of $\boldsymbol{Y}$ that correspond to the largest eigenvalues of $\boldsymbol{B}$. Gower (1966) shows that these dimensions are identical to the first q principal components of the perfect point configuration.

The connection to principal component analysis suggests that the dimensionality reduction step of the classical MDS is rather based on preservation of variance than on preservation of the dissimilarity representation. Additionally, the transformation is necessarily linear. A more flexible way to seek a q-dimensional, possibly nonlinear, embedding $\boldsymbol{Y}$ is to minimize the cost function[4]

$$E(\boldsymbol{Y}) = \sum_{i=1}^{n}\sum_{j>i}\alpha_{ij}(\|\underline{\boldsymbol{y}}_i - \underline{\boldsymbol{y}}_j\|_2 - \delta_{ij})^2, \tag{2.24}$$

which is sometimes called as the stress criterion. The perfect low-dimensional embedding is usually impossible, so the parameters $\alpha_{ij} \geq 0$ are introduced to weight some pairs more than the others. The sammon mapping (Sammon, 1969) has $\alpha_{ij} = \delta_{ij}^{-1}$ and, thereby, it focuses on preserving the relations between similar objects. Curvilinear component analysis (Demartines and Hérault, 1997) gives high values of α_{ij} to the pairs of objects whose embeddings $\underline{\boldsymbol{y}}_i$ and $\underline{\boldsymbol{y}}_j$ are close by. Venna and Kaski (2006) propose a user-tunable method for controlling the tradeoff between the two ends. The sammon mapping is used in Publication III.

[4]Throughout the text, $\underline{\boldsymbol{x}}_i$ denotes the ith observation of all m variables, whereas $\boldsymbol{x}_i$ is the vector of all n observations of the ith variable. The same thing differs $\underline{\boldsymbol{y}}_i$ from $\boldsymbol{y}_i$.

The MDS methods are often used in feature extraction by letting the dissimilarities denote distances in the m-dimensional input space according to some metric, that is $\delta_{ij} = d(\underline{\boldsymbol{x}}_i, \underline{\boldsymbol{x}}_j)$. Taking $q < m$, the input data become projected to a lower-dimensional feature space unsupervisedly. The metric can also depend on auxiliary data such as class labels of the observations. In that case, one can use metrics, where the between-class dissimilarity is considered larger than the within-class dissimilarity (Cox and Ferry, 1993; Zhang, 2003). Then the MDS method extracts discriminative features, which highlight class separation.

In order to use the feature representation of the input variables in any subsequent task on new observations, it must be possible to project these observations to the feature space. There exist certain out-of-sample techniques (Bengio et al., 2004) but, in general, the MDS framework described above does not give a parameterized transformation from the input space to the feature space in a consistent manner. In the remaining, we focus on the linear regression mapping $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}$ and present two algorithms for minimizing (2.24) while enforcing the matrix $\boldsymbol{W}$ to be row sparse. As pointed out in Publication III, this framework enables combining variable selection with feature extraction. Recently, Maniyar and Nabney (2006) have proposed a probabilistic approach to data visualization with simultaneous variable selection based on a generative topographic mapping, which offers an alternative to the MDS approach to be presented next.

The iterative majorization algorithm by Webb (1995) is applicable to minimizing the objective function $E(\boldsymbol{X}\boldsymbol{W})$, which denotes the stress criterion (2.24) under a linear mapping. Despite the different name, this is actually a majorize-minimize algorithm. The value of the objective function is reduced monotonically by minimizing a succession of quadratic approximations, each of which majorizes the objective. Consider the penalized MDS problem

$$\underset{\boldsymbol{W}}{\text{minimize}}\ \tfrac{1}{2}E(\boldsymbol{X}\boldsymbol{W}) + \lambda \sum_{j=1}^{m} p(\|\boldsymbol{w}_j\|_2). \tag{2.25}$$

It is straightforward to majorize the objective in (2.25) by applying the majorizing function by Webb (1995) to the first term and the majorizing function proposed in Publication VI to the second term. The next iterate

$$\widehat{\boldsymbol{W}}^{[k+1]} = (\boldsymbol{A} + \lambda \boldsymbol{\Omega}^{[k]})^{-1} \boldsymbol{D}^{[k]} \widehat{\boldsymbol{W}}^{[k]} \tag{2.26}$$

is taken to minimize the current majorizing function. The matrix $\boldsymbol{\Omega}^{[k]}$ is defined in (2.15). Webb (1995) defines $\boldsymbol{A}$ and $\boldsymbol{D}^{[k]}$, but they are not reproduced here in order to avoid introducing an excess of additional notation.

The shadow targets algorithm by Tipping and Lowe (1998) is a two-step iterative procedure to minimizing $E(\boldsymbol{X}\boldsymbol{W})$. The first step updates the configuration $\boldsymbol{Y}^{[k]}$ in the direction of the steepest descent $\boldsymbol{\Delta Y}^{[k]} \propto -\frac{\partial}{\partial \boldsymbol{Y}} E(\boldsymbol{Y}^{[k]})$. The second step makes an OLS fitting, where the updated configuration $\boldsymbol{T}^{[k+1]} = \boldsymbol{Y}^{[k]} + \boldsymbol{\Delta Y}^{[k]}$ serves as the response data and $\boldsymbol{X}$ denotes the input data. The output of the regression model gives the next configuration. Cox and Ferry (1993) have used this idea earlier, but in their method the complete MDS optimization and the consecutive model fitting are both performed only once and separately. The second step of the shadow targets algorithm can be replaced with solving a penalized problem

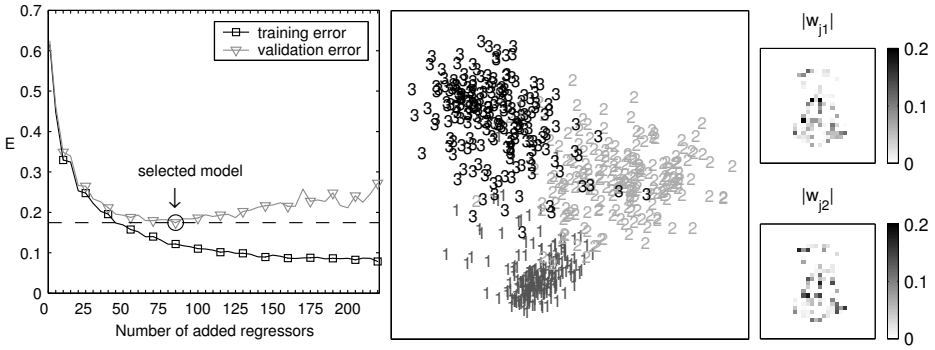

Figure 2.8: Discriminative projection of images representing handwritten digits 1, 2, and 3 to a two-dimensional feature space: (*left*) training and validation errors as a function of the number of selected pixels, (*middle*) the projection of independent test images to the feature space, (*right*) the contributions of individual pixels to the two features. The figure is taken from Publication III.

of the form (2.13) to enforce row sparsity, and so we have the modified algorithm

$$\boldsymbol{T}^{[k+1]} = \boldsymbol{X}\widehat{\boldsymbol{W}}^{[k]} + \boldsymbol{\Delta Y}^{[k]} \tag{2.27}$$

$$\widehat{\boldsymbol{W}}^{[k+1]} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{\Omega}^{[k]})^{-1}\boldsymbol{X}^T\boldsymbol{T}^{[k+1]}. \tag{2.28}$$

This is flexible to many changes. The first step (2.27) can be improved by using the update rule by Demartines and Hérault (1997), which is significantly faster to evaluate than the steepest descent update. If one agrees to decorrelate the input data, then the inverse operation becomes more simple in the second step (2.28). In Publication III, the MRSR algorithm is used to enforce row sparsity in the second step.

Fig. 2.8 shows some results that are taken from Publication III, where the modified shadow targets algorithm is applied to find a discriminative projection of images representing handwritten digits. The discrimination is implemented via the dissimilarity measure by Zhang (2003). The same set of images is used here as in Section 2.4.1, but including only the digits 1, 2, and 3. The selected group of about 11% of the pixels is apparently enough to form a successful projection to a two-dimensional feature space. While the image experiment is somewhat artificial, the combined variable selection and feature extraction approach of Publication III has also attracted interest from the field of genomic and proteomic data analysis (Li and Harezlak, 2007). Discovering genes or peptides, which dictate object class membership, are of ultimate interest to biologists.

Chapter 3

Information visualization using the self-organizing map

3.1 Self-organizing map

Clustering is the process of grouping similar data points together. It provides a summary, because the original set of data points $\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_n$ is represented by another set $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K$, where $K < n$. Vectors $\boldsymbol{m}_j$ are called prototypes and they lie in the same m-dimensional input space as the vectors $\underline{\boldsymbol{x}}_i$. Each data point is assigned to the cluster whose prototype is the nearest. The self-organizing map (SOM) (Kohonen, 1982, 2001) is a method for clustering and visualizing high-dimensional data. The visualization capabilities of the SOM are particularly useful in exploratory analysis of data (Vesanto, 2002). It is distinguished from most other prototype-based clustering methods, such as the well known K-means algorithm (MacQueen, 1967), due to these capabilities. There exist clustering methods that do not use prototypes (Jain and Dubes, 1988), but they are not discussed here.

The prototypes of the SOM constitute the nodes of a grid in the input space. The neighborhood connections are fixed and the topology is typically two-dimensional with either rectangular or hexagonal connections. The grid can be unfolded to form a uniform lattice, which is used as a visual display. The SOM reduces the dimensionality, because the display depicts how the prototypes reside in the input space. The prototypes are adapted according to the density of the data during an iterative learning process. More prototypes move to the dense areas and less to the areas of few data points. Finally, after successful learning, the nodes that are close in the lattice represent prototypes that are close in the input space as well.

Original incremental algorithm. The SOM algorithm iterates two steps that are the winner node selection and adaptation

$$c_i(k) = \operatorname*{argmin}_j \|\underline{\boldsymbol{x}}_i - \boldsymbol{m}_j(k)\|_2 \tag{3.1}$$

$$\boldsymbol{m}_j(k+1) = \boldsymbol{m}_j(k) + \alpha(k) h_{c_i,j}(k)[\underline{\boldsymbol{x}}_i - \boldsymbol{m}_j(k)], \tag{3.2}$$

where k is the index of the iteration. The winner node for a randomly picked data point $\underline{\boldsymbol{x}}_i$ is the one whose prototype is the nearest in the input space. The prototype of the winner node, as well as the prototypes of the neighboring nodes, is moved closer to the point $\underline{\boldsymbol{x}}_i$. The neighborhood function $h_{c_i,j}$ controls the adaptation. It is a decreasing function of the distance between the winner node c_i and another node j along the lattice. The Gaussian kernel is often used and the width of the kernel is decreased in the course of iteration. The parameter α denotes the learning rate, which is also decreased. See the book by Kohonen (2001) for further technical details and other possible forms of $h_{c_i,j}$.

Supervised incremental algorithm. Among many variants of the SOM the most relevant to this thesis is the one that is applicable to learning regression type of mappings. Suppose that we have observations of inputs $\underline{\boldsymbol{x}}_i$ and responses $\underline{\boldsymbol{y}}_i$, which are vectors of m and q variables, respectively. The following SOM algorithm learns the dependency of the response data on the input data

$$c_i(k) = \operatorname*{argmin}_j \|\underline{\boldsymbol{x}}_i - \boldsymbol{m}_j^{(1)}(k)\|_2 \tag{3.3}$$

$$\begin{bmatrix} \boldsymbol{m}_j^{(1)}(k+1) \\ \boldsymbol{m}_j^{(2)}(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{m}_j^{(1)}(k) \\ \boldsymbol{m}_j^{(2)}(k) \end{bmatrix} + \alpha(k) h_{c_i,j}(k) \begin{bmatrix} \underline{\boldsymbol{x}}_i - \boldsymbol{m}_j^{(1)}(k) \\ \underline{\boldsymbol{y}}_i - \boldsymbol{m}_j^{(2)}(k) \end{bmatrix}. \tag{3.4}$$

The supervised SOM can be used to predict unknown responses given a vector of inputs as follows. Firstly, the winner unit c for the input vector is computed by (3.3) and, secondly, the prediction of the responses is given by $\boldsymbol{m}_c^{(2)}$. However, the supervised SOM is hardly accurate enough for pure prediction purposes due to its discrete nature. It is more useful in visualizing the potentially nonlinear regression surface. This supervised variant is used, for instance, by Kiviluoto (1998) and in Publications II and IV.

3.1.1 Interpreting the SOM

The interpretation of the SOM is based on visualizations of the prototypes. It is assumed that properties seen from the visualizations will also hold for the original data. The visualization process is discussed thoroughly, for instance, in the review works by Vesanto (1999) and Himberg et al. (2001). Some basic methods are introduced here for the sake of completeness. The low-dimensional lattice is used as a visualization platform. A local area of nodes in the lattice represents a local area of prototypes in the input space. The lattice makes effective use of the visualization area. The density of the prototypes follows roughly the density of the data, but the lattice remains uniform. Thereby, the SOM offers an automatic adjustment of resolution without fear of overlaps. In contrast, methods that project data points directly may suffer from the fact that many points can end up very close to each other (overlap) and some of them very far away so that it is difficult to use the low-dimensional representation of the data for visualization purposes. The relational perspective map (Li, 2004) is one of the few other dimensionality reduction methods that is able to visualize data in a non-overlapping manner.

The coordinate axes of the lattice do not have clear interpretation in terms of the original variables. Instead, the variables are typically visualized by a component plane representation, where several lattices, one for each variable, are shown

side by side. The colors of the nodes in a lattice represent the locations of the prototypes in the input space in terms of the corresponding variable. A lattice supplemented with a variable-specific coloring is called a component plane. The component plane representation is useful in finding dependencies between variables. The dependencies are perceived as similar patterns in identical areas of different component planes. The dependency search can be eased by organizing the component planes such that similar planes are positioned near to each other (Vesanto and Ahola, 1999). Lampinen and Kostiainen (2000) discuss some potential problems of the SOM-based analysis including the risk that the prototypes overfit to the data and the tendency to overinterpret the dependencies seen from the component planes.

The SOM does not preserve distance information, but only topological information, that is the neighboring relationships between the prototypes. In many applications, however, it is important to analyze the density of the prototypes, because it reflects the density of the data. Recent contributions to this problem with a good survey of previous work can be found from the article by Pölzlbauer et al. (2006). A traditional way is to use the U-matrix (Ultsch and Siemon, 1990), which depicts local cluster boundaries of the prototypes. It means computing the distances from each prototype to its nearest topological neighbors and then visualizing the distances as grayscale values on the lattice. A rather different way is to disregard the topology, apply standard clustering methods to the prototypes, and then show the clusters on the lattice (Vesanto and Alhoniemi, 2000). One approach is to actually project the prototypes to a two- or three-dimensional space, where their mutual distances are visible. However, the risk of losing the orderliness of the grid is apparent when some standard dimensionality reduction method is used. Another concern is how to link the projection to the lattice such that the prototypes can be identified. The methods by Kaski et al. (1999) and Himberg (2000) aim at maintaining the orderliness of the grid in the projection and use a color coding of the prototypes in the linkage. In Publication II, a modification of the SOM is proposed, which consists of two grid layers with the same topology. One layer models a curvilinear data manifold in the high-dimensional observation coordinates in the same sense as the standard SOM does. The other layer lies in the internal coordinates, which pass through the data manifold. When the manifold is low-dimensional, the grid can be examined visually in the internal coordinates (see Fig. 3.1). This capability comes as a by-product of the algorithm of Publication II. However, there are also modifications of the SOM algorithm that are particularly intended for preserving distance information of the prototypes (Yin, 2002).

The methods discussed so far visualize properties of the prototypes with the goal of explaining the data set that is used in the learning process. Another important application of the SOM is to provide a groundwork for comparing external data. Individual observations can be compared by inserting markers to their winner units on top of the lattice. A data set can be linked to the SOM by counting the number of times a node is found to be the winner node for an observation and then displaying the histogram of the counts on the lattice. Several data sets can be compared by showing their histograms side by side. For more details, see (Vesanto, 1999). In the case of time-series data, a trajectory of adjacent winner units can be shown on top of the lattice to study the behavior of a process in time (Alhoniemi et al., 1999).

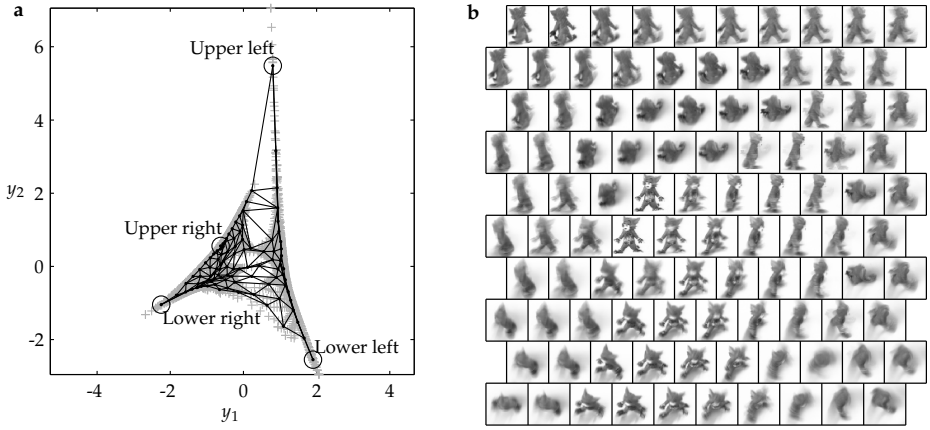

Figure 3.1: The grid of M-SOM (M denotes manifold, see Section 3.2.2) in the internal coordinates (*left*) and a lattice that visualizes the grid in the observation coordinates (*right*). Each node is characterized by pose parameters in the internal coordinates and by an image in the observation coordinates. The corner nodes have been marked to allow comparison between the coordinates. The data set consists of images of a toy object. The figure is taken from Publication II.

3.1.2 Assessing the SOM

All properties of high-dimensional data cannot be retained in low-dimensional representations in general, so visualization methods must make compromises regarding what kinds of relationships to represent. Given two visualizations of the same data set, it is crucial to know, which one is better than the other. The success of a visualization depends on the reliability and usefulness of the results to the user. Personal preferences have an influence, so ranking visualizations is a somewhat subjective task. The SOM is a particularly problematic method in its general form, because it neither has a proper cost function (Erwin et al., 1992) nor a clear probabilistic interpretation, which could be used to compare two different SOMs. Therefore, the quality of the SOM is usually determined in terms of two rather general aspects: quantization error and the quality of topology representation. The former aspect is common to all clustering algorithms and it is measured as the average distance from an observation to its nearest prototype. The latter aspect is common to all dimensionality reduction methods and it is more challenging to measure. The two aspects are competing, but they are not directly opposed.

One approach to assess the quality of the SOM is to consider a cost function that the SOM algorithm minimizes approximately. Vesanto et al. (2003) decompose such a cost function into three parts: quantization error, neighborhood variance, and neighborhood bias. The second term concerns topological quality and it measures the closeness of prototypes in the input space that are close to each other on the lattice. The third term links the quantization and ordering together. The decomposition offers a unified view of different aspects of quality.

Several approaches have been proposed that only focus on the topological quality

of the SOM. One of the earliest attempts is the topographic product (Bauer and Pawelzik, 1992), which measures the preservation of the neighborhood between the nodes in the lattice and their prototypes in the input space. However, as pointed out by Villmann et al. (1997), the topographic product is mostly useful when the data manifold is nearly linear, since it may not be capable of distinguishing between folding of the grid along true nonlinearities in the data and faulty folding. The topographic function (Villmann et al., 1997) is more appropriate when the SOM is fitted to a nonlinear data structure. It measures the frequency of having adjacent Voronoi polyhedrons[1] of the prototypes in the input space, while the corresponding nodes are further away on the lattice than a predefined threshold, or vice versa. The threshold is the argument of the topographic function. The topographic error (Kiviluoto, 1996) shifts the emphasis towards taking the density of the observations better into account. An error occurs when the nearest and the second nearest prototypes of an observation in the input space do not correspond to adjacent nodes in the lattice, and the frequency of these events is the topographic error. Kaski and Lagus (1996) propose a related criterion, where the extent of non-adjacency of the two nodes is measured as their shortest path distance along the lattice. Such a distance reflects the perceptual dissimilarity of a pair of nodes on the lattice display.

Venna and Kaski (2001) propose nonparametric measures for comparing topology preservation of nonlinear projections (see also Kaski et al., 2003). Two kinds of errors may occur. Either the projection introduces new observations to the neighborhood, or observations that are originally neighbors become projected further away. The former kind of error reduces trustworthiness of the visualization: observations found to be similar in the visual display (e.g. the two-dimensional SOM lattice) cannot be trusted to be proximate in the input space. The latter kind of error results from discontinuities in the projection: all proximities that exist in the input space are not present in the visual display. In most cases, both types of errors cannot be avoided. Then it is crucial to decide, which one of the errors is more harmful. The SOM is found to give trustworthy visualizations of data compared with many other projection methods (Venna and Kaski, 2001, 2006). The two measures are also used in Publications I and II.

3.2 Manifold learning

A manifold is by definition a topological space that is locally Euclidean. To illustrate the idea, consider the surface of the Earth, which looks on the small scales flat. However, unlike the ancient belief, we know that it constitutes globally a sphere. Manifolds arise in data when a set of high-dimensional data points can be modeled in a continuous way using only a few variables. A typical example is a set of images of an object taken under fixed lighting conditions with a moving camera. Each image represents a point in the high-dimensional space whose coordinate axes denote the pixels. As the camera moves, the images also move in the pixel space along a surface that is defined by the orientation parameters of the camera such as rotation and elevation (Weinberger and Saul, 2006).

[1] The Voronoi polyhedron of a prototype $\boldsymbol{m}_j$ is the set $\{\boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{m}_j\|_2 < \|\boldsymbol{z} - \boldsymbol{m}_i\|_2, \forall i \neq j\}$.

Manifold learning methods study intrinsically low-dimensional structures lying in a high-dimensional space aiming at discovering the low-dimensional representation. For linearly embedded manifolds, principal component analysis is guaranteed to succeed. MDS is another classical manifold learning method. The SOM produces nonlinear dimensionality reduction, so it also shares the objective of manifold learning. However, as pointed out by Tenenbaum (1998), it fails to model a visually obvious nonlinear structure in the data in some cases. Recently, several other methods have been developed that offer a more powerful framework of manifold learning than the SOM. As a common denominator, their error functions are convex, so there is no risk that the optimization yields poor local minima. Some of these methods are briefly reviewed in the next section. After that, it is shown how any manifold learning algorithm can be combined with the SOM to guide the learning process when there is a nonlinear manifold geometry in the data.

3.2.1 Nonlinear manifold learning methods

The isomap by Tenenbaum et al. (2000) is an extension of the classical MDS, where the dissimilarities denote geodesic distances. The geodesic distance between two points on a manifold is the length of the shortest curve that is on the manifold and connects the two points. In practice, the geodesic distances are approximated by graphical distances, which can be computed from the data. The low-dimensional configuration is recovered from the eigendecomposition of an $n \times n$ matrix just like in the classical MDS (see Section 2.4.2). However, any other MDS method can also be used in the recovery (Lee et al., 2004).

The locally linear embedding by Roweis and Saul (2000) consists of three steps. Firstly, compute the K nearest neighbors of each data point. Secondly, find the $n \times n$ matrix $\boldsymbol{W}$ that minimizes the reconstruction error $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{X}\|_{\mathrm{F}}^2$ subject to $\sum_j w_{ij} = 1$, and $w_{ij} = 0$ if $\underline{\boldsymbol{x}}_j$ does not belong to the K nearest neighbors of $\underline{\boldsymbol{x}}_i$. Thirdly, find the low-dimensional configuration $\boldsymbol{Y}$ that minimizes the embedding cost $\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{Y}\|_{\mathrm{F}}^2$ subject to $\sum_i \underline{\boldsymbol{y}}_i = \boldsymbol{0}$ and $\frac{1}{n}\sum_i \underline{\boldsymbol{y}}_i \underline{\boldsymbol{y}}_i^T = \boldsymbol{I}$. The solution to the second step is found by solving a least squares problem. The solution to the third step is found by computing the eigendecomposition of the $n \times n$ sparse matrix $(\boldsymbol{I} - \boldsymbol{W})^T(\boldsymbol{I} - \boldsymbol{W})$. The eigenvectors with the smallest positive eigenvalues provide the configuration $\boldsymbol{Y}$.

The laplacian eigenmap by Belkin and Niyogi (2003) is justified theoretically by its connection to the Laplace-Beltrami operator on manifolds, which can be used to construct an optimal embedding. The Laplace-Beltrami operator is approximated by the weighted Laplacian of the adjacency graph of the data points. For two connected points, the weight w_{ij} is a positive similarity value. For two disconnected points, the weight w_{ij} is zero. The low dimensional configuration is found by minimizing the objective $\sum_{ij} w_{ij}\|\underline{\boldsymbol{y}}_i - \underline{\boldsymbol{y}}_j\|_2^2$ subject to $\sum_i d_i \underline{\boldsymbol{y}}_i \underline{\boldsymbol{y}}_i^T = \boldsymbol{I}$, where $d_i = \sum_j w_{ij}$. This can be written as a generalized eigenvalue problem $\boldsymbol{L}\boldsymbol{u} = \lambda \mathrm{diag}(\boldsymbol{d})\boldsymbol{u}$, where $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{d}) - \boldsymbol{W}$ is the Laplacian matrix. The eigenvectors with the smallest positive eigenvalues give the configuration $\boldsymbol{Y}$. The hessian eigenmap by Donoho and Grimes (2003) is a related method, which relaxes the implicit requirement that the embedded manifold is sampled on a convex region.

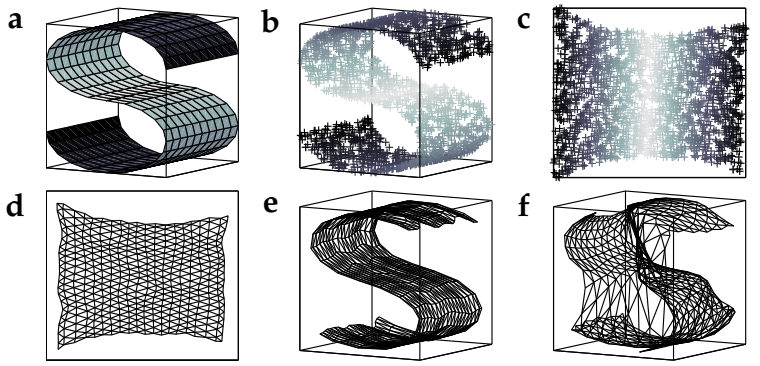

Figure 3.2: The problem of manifold learning for three-dimensional data (b) sampled from an intrinsically two-dimensional S-curve manifold (a). The locally linear embedding algorithm discovers the projection of the data to the internal coordinates on the manifold (c). The M-SOM learns similarities in the internal coordinates (d) and provides a successful representation of the manifold in the observation coordinates (e). The SOM algorithm fails to model the data properly (f). The figure is taken from Publication II.

Alignment methods (Brand, 2003; Zhang and Zha, 2005; Verbeek, 2006) fit several local linear models that give separate low-dimensional coordinate systems. Then these models are merged to get a global coordinate system.

The maximum variance unfolding by Weinberger and Saul (2006) attempts to project the data points apart while preserving the local distances. The objective $\sum_{ij} \|\underline{\boldsymbol{y}}_i - \underline{\boldsymbol{y}}_j\|_2^2$ is maximized subject to $\sum_i \underline{\boldsymbol{y}}_i = \boldsymbol{0}$, and $\|\underline{\boldsymbol{y}}_i - \underline{\boldsymbol{y}}_j\|_2^2 = \|\underline{\boldsymbol{x}}_i - \underline{\boldsymbol{x}}_j\|_2^2$ if there is an edge between $\underline{\boldsymbol{x}}_i$ and $\underline{\boldsymbol{x}}_j$ in the graph formed by pairwise connecting all K nearest neighbors. This can be reformulated as a semidefinite programming problem in terms of the elements of the inner product matrix. From the inner product matrix learned by semidefinite programming, the configuration $\boldsymbol{Y}$ is recovered by matrix diagonalization.

3.2.2 Extending the SOM to manifold learning

It is proposed in Publication II to use some nonlinear manifold learning algorithm to guide the learning process of the SOM. Fig. 3.2 shows an example, where the SOM fails to model the data properly, but the proposed M-SOM learning strategy succeeds with the help of locally linear embedding. The prefix M denotes a manifold. The M-SOM is actually a supervised SOM algorithm (see Eqs. (3.3)–(3.4)), where the winner node is computed in the manifold coordinates and the nodes are updated both in the manifold and observation coordinates. The M-SOM algorithm requires that the projection of the observations to the manifold coordinates is available before the learning process.

Visualization is the purpose of the M-SOM algorithm. The nonlinear manifold learning algorithms are able to discover the low-dimensional representation of the

data under suitable conditions, but the reduction of the dimensionality is not enough for the visualization purpose. For an efficient visualization, the representation must be summarized somehow. Besides the need to comprehend a possibly large data set, the projection to the manifold coordinates may be highly overlapping, which complicates the analysis. On the other hand, the M-SOM shares the same advantages that the SOM has in summarizing massive data sets (see Section 3.1.1).

Publication II provides several comparisons between the SOM and M-SOM algorithms. Given a representation of the data in the manifold coordinates, the learning process of the M-SOM algorithm is faster, because the winner node is sought in a lower-dimensional space and the extra cost in the adaptation step is minor. Experiments on image data with manifold geometry indicate that the M-SOM visualization is less trustworthy but preserves the original neighborhoods better than the SOM according to the criteria by Venna and Kaski (2001). If trustworthiness is considered important, the M-SOM can be used as an initialization algorithm and the learning process is continued with the SOM algorithm. The two-stage learning process may be useful, particularly, if there are doubts about the accuracy of the representation of the data in the manifold coordinates.

3.3 Interpreting quantile regression models

Regression analysis aims at estimating some statistical property of the response variable when the values of the input variables are known. The classical regression methods based on minimizing the sum of squared errors enable one to estimate models for the conditional mean function. In contrast, quantile regression (Koenker and Bassett, 1978; Koenker, 2005) offers a mechanism for estimating models for conditional quantile functions. The median is the best-known example of a quantile and the term quantile is synonymous with percentile. For example, the 25% and 75% quantiles can be defined as values that split ordered data into proportions of one- and three-quarters. Conditional quantile functions split the conditional distribution of the response variable accordingly given the values of input variables. By estimating several different conditional quantile functions a more complete picture of the conditional distribution of the response variable is obtained than using the classical regression methods.

Suppose that the model $Q_\theta(\underline{\boldsymbol{x}})$ is used for the θth conditional quantile function, where $\theta \in (0, 1)$ defines the probability level of interest. The loss function to be minimized in quantile regression is (Koenker and Bassett, 1978)

$$\sum_{i=1}^{n} \rho_\theta(\underline{y}_i - Q_\theta(\underline{\boldsymbol{x}}_i)), \tag{3.5}$$

where $\rho_\theta(\cdot)$ is the so-called "check function"

$$\rho_\theta(e) = \begin{cases} \theta e & , \ e \geq 0 \\ (\theta - 1)e & , \ e < 0. \end{cases} \tag{3.6}$$

If the model $Q_\theta(\underline{\boldsymbol{x}})$ is linear in parameters, then the minimization of (3.5) can be performed efficiently by linear programming methods. Koenker (2005) provides a

comprehensive review that includes both parametric and nonparametric models for conditional quantile functions. The overview can be further supplemented by the recent works by Takeuchi et al. (2006) and Li et al. (2007) on quantile regression in reproducing kernel Hilbert spaces, which appear to offer a tractable framework.

Model interpretation is important if the problem setting is extended from prediction toward explorative analysis, where the aim is to understand the underlying process that generated the data. For instance, economists are interested in estimating the effects of education on employment or earnings and the effects of a firm's inputs on its outputs. This type of inference is typically based on partial derivatives of the model $\frac{\partial}{\partial \underline{x}_j} Q_\theta(\underline{\boldsymbol{x}})$. For a linear model the partial derivatives are constant and they coincide with the regression coefficients, which are readily available. If the linear model does not fit to the data, then any inference that is based on it is incorrect. However, things become much more complicated when a nonlinear model is used. Accurate pointwise estimation of the derivatives is difficult due to the curse of dimensionality. The derivatives are no longer constant, so one has to interpret m partial derivative surfaces in the $(m+1)$-dimensional space given m input variables. As a remedy, Chaudhuri et al. (1997) propose to use the average derivative. This entails averaging nonparametric estimates of the derivative function over some appropriate region in the input space. Another approach is to use a nonlinear additive model of the form $Q_\theta(\underline{\boldsymbol{x}}) = w_0 + \sum_j Q_\theta^{(j)}(\underline{x}_j)$, which offers advantages at the model interpretation stage (Doksum and Koo, 2000; De Gooijer and Zerom, 2003; Yu and Lu, 2004). By displaying the curves $Q_\theta^{(j)}(\underline{x}_j)$ it is possible to examine the roles of the input variables in predicting the θth conditional quantile function.

A novel approach is proposed in Publication IV. The conditional quantile function and its partial derivatives are visualized using the SOM. It enables to examine the shapes of these potentially nonlinear functions of several input variables via the two-dimensional component plane representation of the SOM. Compared with the average derivatives, a more comprehensive picture is obtained with the possibility of identifying local properties of the partial derivative functions. Compared with visualizing an additive model, the SOM-based approach is more flexible, because any type of model can be used. In additive models, the effect of an input variable on the conditional quantile regression surface does not depend on the values of the other input variables, which may be too restrictive in practice. The next section serves as a word of caution by overviewing some pitfalls as regards to interpreting regression surfaces in general. After that, the SOM-based approach is presented.

3.3.1 Notions on interpreting regression surfaces

In practical situations, the choice of the model is limited by conditions like what input variables can be measured and what kinds of models can be estimated reliably, so the model is merely an approximation of the true underlying phenomenon. Even if the model happens to be an excellent predictor, the interpretation of the regression surface in terms of its partial derivatives often requires considerable care. Small fluctuations of the regression surface may not weaken the prediction accuracy much, but they can make the derivatives noisy. When the number of input variables increases the data cloud becomes sparser and more data points are close

to the boundaries. Sparsity makes the estimation problematic and the boundary effect has the consequence that most directions point away from the data cloud, so it is impossible to measure changes of the regression surface in those directions.

The partial derivatives are often interpreted to measure the change in the regression surface when an input variable is changed by one unit while all the other input variables are unchanged. This is strictly true only in well-designed experiments in which the input variables can be manipulated independently of each other. Given observations of some process that cannot be controlled, the justification to speak about changes is faint. It may not be possible even theoretically to change one input variable without changing the others. Furthermore, it is not allowed to state conclusions in causal terms without experimental control. Causality means that the change in the input variable causes the change in the regression surface. It is difficult to identify the dynamics of individual subjects even when causality can be inferred. If a particular subject happens to fall at the θth quantile initially, and then receives an increment $\Delta \underline{x}_j$, it does not necessarily fall on the θth conditional quantile function following the increment (Koenker, 2005).

The magnitudes of the partial derivatives with respect to different input variables can only be compared if the variables are measured in the same units. Often the input variables are normalized to have an equal standard deviation. It makes the comparison easier but is only justified if a change proportional to the standard deviation constitutes a meaningful quantity. Correlation between the input variables complicates the comparison further. A model with perfectly correlated input variables can be reparameterized in infinitely many ways in terms of these variables, so the corresponding partial derivatives are arbitrary as well. The partial derivatives may also be inaccurate and misleading, because they measure changes along the coordinate axes but correlated data points are mainly distributed along the diagonal direction.

3.3.2 Visualizing quantile regression surfaces using the SOM

The approach that is proposed in Publication IV is a direct application of the supervised SOM algorithm, shown in Eqs. (3.3)–(3.4). It is assumed that input data, models for the conditional quantile functions, and the partial derivatives of the models are available for some probability levels of interest $\theta_1, \ldots, \theta_\ell$. Local polynomial fitting is applied in Publication IV, but any other technique can also be used to estimate the models. The available information can be represented as follows

$$\{\underline{\boldsymbol{x}}_i, Q_{\theta_1}(\underline{\boldsymbol{x}}_i), \ldots, Q_{\theta_\ell}(\underline{\boldsymbol{x}}_i), \nabla Q_{\theta_1}(\underline{\boldsymbol{x}}_i), \ldots, \nabla Q_{\theta_\ell}(\underline{\boldsymbol{x}}_i)\}_{i=1}^n. \tag{3.7}$$

The proposed approach enables visualizing local properties of the above functions with the help of the SOM.

The winner node in the supervised SOM algorithm is computed using the first m terms (input variables) and the update concerns all the $m + \ell + \ell m$ terms in (3.7). The prototypes in the input space model the data in the same way as would be the case when using the original SOM algorithm. In the other dimensions, the prototypes represent sorts of conditional averages, which are computed over local areas of the input space. The averaging has the practical benefit of smoothing

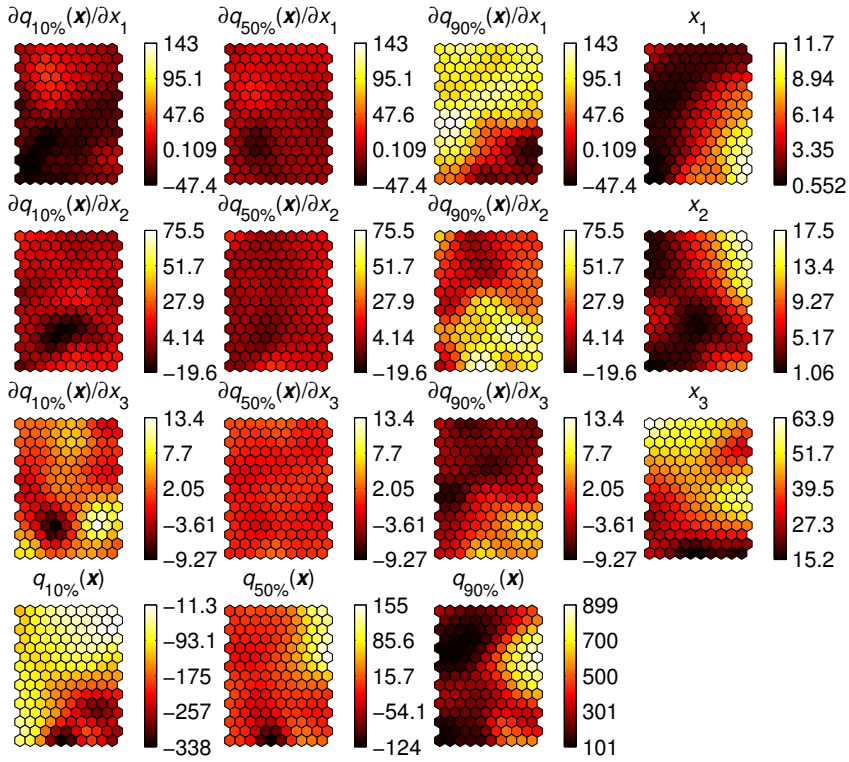

Figure 3.3: Component planes of the SOM that visualize the θth conditional quantile function and its partial derivative functions for the probability levels $\theta = 0.1,\ 0.5,\ 0.9$. Response variable: strike index. Input variables: x_1 unemployment rate (%), x_2 yearly inflation rate (%), x_3 social democratic parliamentary representation (%). The figure is taken from Publication IV.

the potentially noisy estimates of the quantile regression surfaces and their partial derivatives. The prototypes are likely to be in dense regions of the input space, so the method is somewhat resistant to the instability of model estimation near boundaries of the data.

The proposed approach is applied on real data in Publication IV. The data set consists of annual observations on the level of strike volume (days lost due to industrial disputes per 1000 wage salary earners) and their covariates in eighteen OECD countries from 1951–1985[2]. The years 1981–1985 are, however, missing for one country. Three observations are left out, because they have extremely high levels of strike volume. The response variable of the analysis is the so-called strike index, which is the difference between the observed level of strike volume and the country specific median level. Three input variables are used: yearly inflation rate (%), unemployment rate (%), and social democratic parliamentary representation (%). The conditional median, upper 90%, and lower 10% quantiles of strike index are modeled using nonparametric local linear quantile regression.

[2]The data set is available from http://lib.stat.cmu.edu/datasets/strikes.

Fig. 3.3 shows the component planes of the SOM that visualize the quantile regression models. The conditional median is the highest when inflation rate is also high. The 90% conditional quantile, in turn, is the highest when both inflation and unemployment rates are high. The extreme quantiles are close when unemployment rate is low, so strike index is rather invariant in that case. There is no clear dependency between social democratic representation and the distribution of strike index. An identical scaling of the axis is used in the component planes that represent the partial derivative functions of different quantiles with respect to a certain input variable. This helps to observe that the shape of the 90% conditional quantile function is more sensitive to the input variables than the two other functions. If unemployment rate is low, then the upper tail of the conditional distribution of strike index has a positive relationship between unemployment rate. If unemployment rate is high, then the relationship is negative. Inflation rate has a positive relationship between the upper tail in all cases and the connection is the strongest when unemployment rate is high and inflation rate is low. It is difficult to summarize the component planes, which represent partial derivatives with respect to social democratic representation.

3.4 Exploring the dependency of one data set on another

This section considers the analysis of two data sets of continuous-valued variables, say $\boldsymbol{X}$ and $\boldsymbol{Y}$. The goal of the analysis is somewhat similar to multiresponse regression, investigated in Section 2.3. However, the following approach focuses more on data exploration or data mining than accurate prediction. It serves the early step of a research problem, where the analyst is making the first conceptions of the data sets and their relations to each other. The relations can either be symmetric or asymmetric. Symmetric dependency modeling means finding things that are common to both data sets and the sets are treated equally. A classical example is canonical correlation analysis (Hotelling, 1936), which seeks linear combinations for both variable sets by maximizing their mutual correlation. Further pairs of maximally correlated linear combinations are chosen such that they are orthogonal to those already identified. Nonlinear extensions of canonical correlation analysis also exist, such as the kernel version by Bach and Jordan (2002). Nonparametric methods offer another flexible approach to symmetric dependency modeling. For instance, associative clustering by Kaski et al. (2005) groups similar observations within each data set such that the groups of different sets capture as much of the pairwise dependencies between the observations as possible.

From now on, only asymmetric dependency modeling is investigated. One set is considered as the response data, say $\boldsymbol{Y}$, and the goal is to find simplifying representations of $\boldsymbol{X}$ without loss of information on the conditional distribution of $\boldsymbol{Y}|\boldsymbol{X}$. No prespecified model is required, which differs from the techniques that select variables of $\boldsymbol{X}$ (see Section 2.3) or reduce the dimensionality of $\boldsymbol{X}$ (Abraham and Merola, 2005) for predictive purposes. Sliced inverse regression by Li (1991) is a seminal method that formalizes the model free framework by searching for a transformation that reduces the dimensionality of $\boldsymbol{X}$ while retaining the regres-

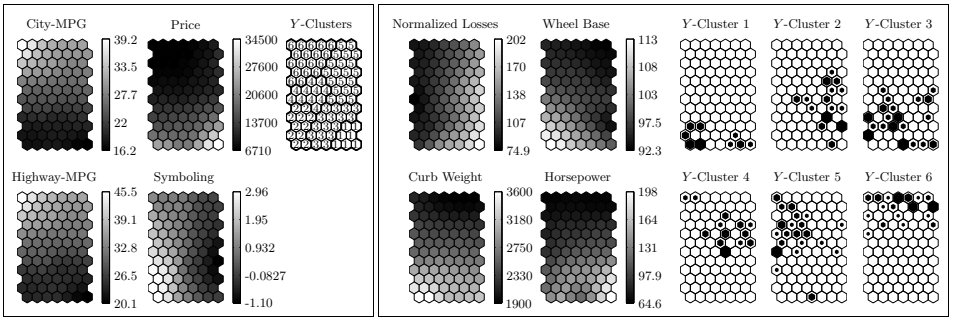

Figure 3.4: Analysis of car data: (*left*) component planes of the response variables that measure safety (Symboling describes insecurity) and economic aspects of the cars, (*right*) the optimal combination of four input variables and their SOM visualization. The clustering and the histograms illustrate the mapping of observations between the two SOMs. The figure is taken from Publication I.

sion information. The original sliced inverse regression assumes a single response variable, but there exist many extensions to multivariate responses (Setodji and Cook, 2004; Barreda et al., 2007).

In Publication I, a visualization approach to asymmetric dependency modeling is proposed. There would be an inherent need to apply visualization methods to the two data sets even if they were analyzed separately for explorative purposes, because both are multivariate. If the analyst has gained his/her understanding about the response data via some visualization, it is perceptually justified to derive such a representation of the input data that is related to the visualization, not the response data itself. The SOM is used to visualize the response data and the representation that is sought for is the subset of input variables that best supports the visualization. The next section describes the approach in more detail. Several extensions and related work are discussed in the subsequent section.

3.4.1 Variables that explain the SOM visualization

The response variables are fixed in the supervised learning framework, so it is quite natural to use the SOM to overview the response data. For example, a data set consisting of several measurements of various car models[3] is investigated in Publication I (see also Fig. 3.4). The four response variables measure safety and economic aspects of the cars. After the SOM-based visual exploration of the response data, the analyst may want to know, which other properties of the cars explain the notion of safety and economic efficiency provided by the SOM. The purpose could be to support market research of the car industry. In short, the problem is to find input variables that are visually salient to the response SOM.

The interpretation of the response SOM is based on some representation of the uniform lattice, such as the component planes. The dissimilarity of two nodes is perceived as their distance in the lattice. This is formalized, in the first place, by

[3]The data set is available from http://www.ics.uci.edu/∼mlearn/MLRepository.html.

the neighborhood function in the update rule of the SOM algorithm in Eq. (3.2). Each observation can be clustered to a certain node in the lattice by computing its winner node according to the response variables. Each observation also has a position in the input space. The task is to select relevant input variables without making parametric assumptions on the mapping from the continuous-valued input space to the uniform lattice of the response SOM. The criterion that is proposed in Publication I is a weighted sum of pairwise distances between observations in the lattice of the response SOM, where a weight depends on the closeness of the two observations in the input space. The simplest way to measure closeness is to give the unit weight for the K nearest neighbors of an observation and fix the other weights to zero. Different combinations of input variables are sought to minimize the criterion.

The minimization already gives additional information by discarding irrelevant input variables. If many relevant input variables are found, then it is useful to apply some visualization technique to them. The SOM is also used to visualize the reduced set of input variables in Publication I. The mapping between the two SOMs can be examined as follows. The observations are clustered using one SOM, and then, the SOM is itself clustered. The two-level approach to clustering is reviewed by Vesanto and Alhoniemi (2000). Finally, the cluster-specific histograms of observations are displayed on the top of the lattice of the other SOM. Fig. 3.4 illustrates the mapping of observations between the two SOMs in the car experiment.

3.4.2 Extensions and related work

A good comparison criterion is proposed in Publication I, but finding the combination of input variables that minimizes the criterion may be a difficult task when the number of candidates is large. Instead of going deeply into combinatorial optimization, alternative ways of combining the input variables or a subset of them to the response SOM are discussed next.

The MDS techniques introduced in Section 2.4.2 are readily applicable to the combining. To see this, let the dissimilarity δ_{ij} denote the distance between the winner nodes of a pair of observations in the lattice of the response SOM. Define the linear transformation $\boldsymbol{f}(\underline{\boldsymbol{x}}) = \boldsymbol{W}^T\underline{\boldsymbol{x}}$ as a solution to the penalized MDS problem (2.25). If $\boldsymbol{W}$ is set to have fewer columns than rows, then the input data become projected to a lower-dimensional feature space. With a sufficiently large value of the penalty parameter λ the feature space does not depend on all the input variables. This way the combinatorial subset selection problem can be relaxed to a continuous-valued optimization problem. The mapping from the feature space to the response SOM is handled nonparametrically by preserving local distances.

If one is willing to make parametric assumptions, any regression model to bivariate (the lattice is typically two-dimensional) ordinal response data (Agresti, 2002) can be used to model the dependency of the response SOM on the input variables. The regression model could be penalized or constrained such that input selection occurs. However, the parametric regression approach is already quite different from the one taken in Publication I, where the dependency is measured without requiring a prespecified predictive model.

A characteristic feature of the SOM is the topological ordering of the nodes. Ignoring this information, however, brings out several new possibilities to dependency modeling. A related method, K-means inverse regression by Setodji and Cook (2004), uses the K-means algorithm (MacQueen, 1967) to group the response data into K clusters. Sliced inverse regression (Li, 1991) is then applied, with the slices replaced by the clusters, to reduce the dimensionality of the input data. Sliced inverse regression could also follow the SOM-based clustering of the response data, but it is only one possibility. Several discriminative projection methods that have been developed for multiclass classification are applicable, starting from multiple discriminant analysis to the more recent methods (Peltonen and Kaski, 2005; Weinberger et al., 2006). In Publication I, the input data are visualized in terms of the selected variables using the SOM. Instead of this completely unsupervised visualization after supervised variable selection, it might be more reasonable to focus on visualizing the properties of the selected input variables that are relevant to the clustering of the response data. The learning metrics principle (Kaski et al., 2001) offers a mechanism for this purpose.

Chapter 4

Conclusions

4.1 Summary

The analysis of multivariate data is a frequent task in nearly any field of modern science. The problem of focusing on the most meaningful information has become increasingly important as the data sets have become larger. This thesis addressed the problem of finding simplifying representations automatically from the data. The proposed methods can be roughly divided into those that aim to discard irrelevant or useless variables and those that aim to overview the data in terms of a fixed set of variables. The thesis had three main contributions, which are summarized next.

The first half of the thesis considered methods for automatic model building in linear regression to enhance predictability and facilitate model interpretation. A traditional way is adding or deleting variables stepwise, or even trying all possible combinations. However, the stepwise approach is notorious for failing when the variables are strongly correlated and the exhaustive approach is limited to a small number of variables. Both approaches also suffer from weak stability. The first contribution of this thesis was extending several shrinkage and selection methods that offer a remedy to the aforementioned problems in single response regression to be applicable with multiple response variables as well. The key property of a selection and shrinkage method is that it can be formulated as a single optimization problem of continuous-valued parameters. The solution is more stable and it is more easily available than in the traditional combinatorial paradigm. Moreover, it is often possible to compute the whole path of solutions as a function of the shrinking parameter efficiently, which reduces the computational burden of model selection.

The MRSR algorithm was proposed as a multiresponse counterpart of the lars algorithm (Efron et al., 2004) in this thesis. Despite a close connection between the lars algorithm and the solution path of the lasso estimate (Tibshirani, 1996), it has been shown to be challenging to ensue the MRSR algorithm from some well defined objective function. However, Theorem 2.3.2 showed that the L_2-MRSR

algorithm follows the solution path of the L_2-SVS estimate under an orthonormality assumption on the input variables. The L_2-SVS estimate (Publication VII; Cotter et al., 2005; Malioutov et al., 2005), in turn, is a multiresponse counterpart of the lasso estimate. In a general case, the solution path of the L_2-SVS estimate is piecewise smooth but nonlinear, and the path was examined in this thesis for the first time. A predictor-corrector method and a related active set strategy were proposed for following the path efficiently. In addition, the necessary and sufficient conditions for optimality and a condition that ensures the uniqueness of the L_2-SVS estimate were derived. Yet another contribution to regression was extending the penalized least squares framework by Fan and Li (2001) and the majorize-minimize optimization algorithm by Hunter and Li (2005) to cover multiple response variables. In relation to that, an active set strategy for following the path of penalized solutions was proposed, which is not an extension but a novel result based on the necessary conditions for optimality developed in this work.

The second contribution concerned unsupervised variable selection, which was also discussed in the first half of the thesis. According to the definition by McCabe (1984), it is the task of finding principal variables that contain as much information as possible. The definition was formalized in this work as the optimal reconstruction of data by a linear combination of its most prominent variables. A practical implementation was obtained by regressing the data set against itself while forcing the coefficient matrix of the linear combination to be row sparse. Here, a surprising possibility opens up to the multiresponse regression techniques that were proposed in the thesis. Still supplementing the second contribution, parametric multidimensional scaling was examined with the goal of reducing the dimensionality of data via combined subset selection and subspace projection. It was shown how the iterative majorization algorithm by Webb (1995) and the shadow targets algorithm by Tipping and Lowe (1998) are applicable to the task as a result of an appropriate penalization of the parameters.

The third contribution was introduced in the second half of the thesis. It concerns extensions and applications of the SOM (Kohonen, 1982, 2001). The SOM was extended to a class of data sets in which it had previously lacked accuracy, namely strongly curvilinear but intrinsically low-dimensional data manifolds. With the assistance of an auxiliary projection of the data the proposed M-SOM algorithm is capable of preserving local neighborhoods of such data better than the SOM. An application of the SOM was proposed for interpreting nonlinear quantile regression models. Finally, it was examined how to find properties of one data set, which are related to a SOM-based visualization of another data set.

4.2 Directions for future work

The topics of the first half of the thesis offer several interesting possibilities for future research. One open theoretical problem is quantifying the deviation of the L_2-MRSR path from the L_2-SVS path when the input variables are correlated. Another open problem is finding sufficient conditions under which the MRSR algorithm and the L_2-SVS estimate identify the nonzero rows of the best subset regression estimate, which has already been done for the L_∞-SVS estimate by

Tropp (2006a). Most effort in this thesis was put in row sparsity of a coefficient matrix. However, some of the proposed methods are directly applicable to induce various other forms of structural sparsity. Requirements for such structures might potentially emerge in applications of nonlinear models or latent variable models. In relation to the latter, sparse principal component analysis by Zou et al. (2006) provides an interesting groundwork that could be extended toward structural sparsity.

Appendix

The L_α-MRSR algorithm.

Initialize $\widehat{\boldsymbol{W}}(\lambda) = 0$ for $\lambda \geq 0$, $k = 0$, and $\lambda^{[0]} = \max_{1 \leq j \leq m} \|\boldsymbol{Y}^T \boldsymbol{x}_j\|_\alpha$
while $\lambda^{[k]} > 0$ **do**
  $\mathcal{A} = \{j : \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{W}}(\lambda^{[k]}))^T \boldsymbol{x}_j\|_\alpha = \lambda^{[k]}\}$
  $\widehat{\boldsymbol{W}}_{\mathcal{A}}(\lambda) = \frac{\lambda}{\lambda^{[k]}} \widehat{\boldsymbol{W}}_{\mathcal{A}}(\lambda^{[k]}) + \big(1 - \frac{\lambda}{\lambda^{[k]}}\big)(\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}})^{-1} \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{Y}$ for $\lambda \in [0, \lambda^{[k]}]$
  if $|\mathcal{A}| < \min\{m, n\}$ **then**
    $\lambda^{[k+1]} = \max_{j \notin \mathcal{A}} \min_{\lambda \in [0, \lambda^{[k]}]} \{\lambda : \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{W}}(\lambda))^T \boldsymbol{x}_j\|_\alpha \leq \lambda\}$
  else
    $\lambda^{[k+1]} = 0$
  end if
  $k := k + 1$
end while

Proof of Theorem 2.3.1. Suppose, without loss of generality, that the index j enters the set $\mathcal{A}(\lambda)$ at $\lambda^{[k_j]}$. Since $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ holds, the MRSR path (2.18) is

$$\widehat{\boldsymbol{w}}_j(\lambda) = \begin{cases} (\lambda/\lambda^{[k]})\widehat{\boldsymbol{w}}_j(\lambda^{[k]}) + (1 - \lambda/\lambda^{[k]})\boldsymbol{Y}^T\boldsymbol{x}_j, & \lambda \in \{[\lambda^{[k+1]}, \lambda^{[k]}] : k \geq k_j\} \\ 0 & , \lambda \geq \lambda^{[k_j]}. \end{cases} \tag{A1}$$

Firstly, we formulate the following hypothesis about (A1)

$$\widehat{\boldsymbol{w}}_j(\lambda) = (1 - \lambda/\lambda^{[k_j]})\boldsymbol{Y}^T\boldsymbol{x}_j, \quad \lambda \in [0, \lambda^{[k_j]}], \tag{A2}$$

which is clearly true for $\lambda \in [\lambda^{[k+1]}, \lambda^{[k]}]$ with $k = k_j$. Next, we assume that (A2) holds for $\lambda \in [\lambda^{[k+1]}, \lambda^{[k]}]$ with some $k = k' \geq k_j$. But then, (A2) must also hold for $\lambda \in [\lambda^{[k+1]}, \lambda^{[k]}]$ with $k = k' + 1$, because we have

$$\begin{aligned} \widehat{\boldsymbol{w}}_j(\lambda) &= (\lambda/\lambda^{[k'+1]})\widehat{\boldsymbol{w}}_j(\lambda^{[k'+1]}) + (1 - \lambda/\lambda^{[k'+1]})\boldsymbol{Y}^T\boldsymbol{x}_j & \text{(A3)} \\ &= (\lambda/\lambda^{[k'+1]})(1 - \lambda^{[k'+1]}/\lambda^{[k_j]})\boldsymbol{Y}^T\boldsymbol{x}_j + (1 - \lambda/\lambda^{[k'+1]})\boldsymbol{Y}^T\boldsymbol{x}_j & \text{(A4)} \\ &= (1 - \lambda/\lambda^{[k_j]})\boldsymbol{Y}^T\boldsymbol{x}_j. & \text{(A5)} \end{aligned}$$

Hypothesis (A2) is, therefore, true by induction. The property $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ implies $(\boldsymbol{X}\widehat{\boldsymbol{W}}(\lambda))^T\boldsymbol{x}_j = \widehat{\boldsymbol{w}}_j(\lambda)$ and we have $\widehat{\boldsymbol{w}}_j(\lambda^{[k_j]}) = 0$, so the equivalence

$$\lambda^{[k_j]} = \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{W}}(\lambda^{[k_j]}))^T\boldsymbol{x}_j\|_2 = \|\boldsymbol{Y}^T\boldsymbol{x}_j\|_2 \tag{A6}$$

holds according to (2.21). Substituting (A6) into (A2) completes the proof. □

Proof of Theorem 2.3.2. Publication VII shows that the conditions

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T \boldsymbol{x}_j = \lambda \boldsymbol{w}_j / \|\boldsymbol{w}_j\|_2, \quad j \in \{j : \boldsymbol{w}_j \neq 0\} \tag{A7}$$

$$\|(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T \boldsymbol{x}_j\|_2 \leq \lambda, \quad j \in \{j : \boldsymbol{w}_j = 0\} \tag{A8}$$

$$\lambda\Big(\sum_{j=1}^{m} \|\boldsymbol{w}_j\|_2 - t\Big) = 0, \quad \lambda \geq 0 \tag{A9}$$

are necessary and sufficient for $\boldsymbol{W}$ to be the solution to (2.11) at t. It is easy to verify that the L_2-MRSR path $\widehat{\boldsymbol{W}}(\lambda)$, shown in Theorem 2.3.1, satisfies (A7) and (A8), where we have $(\boldsymbol{X}\widehat{\boldsymbol{W}}(\lambda))^T \boldsymbol{x}_j = \widehat{\boldsymbol{w}}_j(\lambda)$ due to the property $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$. In the case of $\lambda > 0$, (A9) holds when $t = \sum_{j=1}^{m} \|\widehat{\boldsymbol{w}}_j(\lambda)\|_2$ applies, and for $\lambda = 0$, we may choose any $t \geq \sum_{j=1}^{m} \|\widehat{\boldsymbol{w}}_j(0)\|_2$. By the definition of the L_2-MRSR path in Theorem 2.3.1, we have $\|\widehat{\boldsymbol{w}}_j(\lambda)\|_2 = \max\{0, \|\boldsymbol{Y}^T\boldsymbol{x}_j\|_2 - \lambda\}$. □

References

Abraham, B. and Merola, G. (2005). Dimensionality reduction approach to multivariate prediction, *Computational Statistics & Data Analysis* **48**(1): 5–16.

Agresti, A. (2002). *Categorical data analysis*, Wiley series in probability and statistics, 2 edn, Wiley-Interscience, New York.

Al-Kandari, N. M. and Jolliffe, I. T. (2005). Variable selection and interpretation in correlation principal components, *Environmetrics* **16**(6): 659–672.

Alhoniemi, E., Hollmén, J., Simula, O. and Vesanto, J. (1999). Process monitoring and modeling using the self-organizing map, *Integrated Computer-Aided Engineering* **6**(1): 3–14.

Anderson, R. L. and Bancroft, T. A. (1952). *Statistical Theory in Research*, New York: McGraw-Hill Book Company.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis, *Journal of Machine Learning Research* **3**: 1–48.

Bakin, S. (1999). *Adaptive Regression and Model Selection in Data Mining Problems*, PhD thesis, The Australian National University, School of Mathematical Sciences, Canberra, Australia.

Barreda, L., Gannoun, A. and Saracco, J. (2007). Some extensions of multivariate sliced inverse regression, *Journal of Statistical Computation and Simulation* **77**(1): 1–17.

Bauer, H.-U. and Pawelzik, K. R. (1992). Quantifying the neighborhood preservation of self-organizing feature maps, *IEEE Transactions on Neural Networks* **3**(4).

Bedrick, E. J. and Tsai, C.-L. (1994). Model selection for multivariate regression in small samples, *Biometrics* **50**: 226–231.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data compression, *Neural Computation* **15**(6): 1373–1396.

Bellman, R. E. (1961). *Adaptive Control Processes*, Princeton University Press.

Bengio, Y., Paiement, J., Vincent, P., Delalleau, O., Roux, N. L. and Ouimet, M. (2004). Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering, *in* S. Thrun, L. Saul and B. Schölkopf (eds), *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning, *Artificial Intelligence* **97**(1-2): 245–271.

Brand, M. (2003). Charting a manifold, *in* S. T. S. Becker and K. Obermayer (eds), *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, MA, pp. 961–968.

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

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics* **24**(6).

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression, *Journal of the Royal Statistical Society, Series B* **59**(1): 3–54.

Brown, P. J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors, *Journal of the Royal Statistical Society, Series B* **64**(3).

Cadima, J., Cerdeira, J. O. and Minhoto, M. (2004). Computational aspects of algorithms for variable selection in the context of principal components, *Computational Statistics & Data Analysis* **47**(2): 225–236.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression, *The Annals of Statistics* **25**(2): 715–744.

Cotter, S. F., Rao, B. D., Engan, K. and Kreutz-Delgado, K. (2005). Sparse solutions to linear inverse problems with multiple measurement vectors, *IEEE Transactions on Signal Processing* **53**(7): 2477–2488.

Cox, T. F. and Cox, M. A. A. (2001). *Multidimensional Scaling*, Chapman and Hall.

Cox, T. F. and Ferry, G. (1993). Discriminant analysis using non-metric multidimensional scaling, *Pattern Recognition* **26**(1): 145–153.

Cumming, J. A. and Wooff, D. A. (2007). Dimension reduction via principal variables, *Computational Statistics & Data Analysis* **52**(1): 550–565.

De Gooijer, J. G. and Zerom, D. (2003). On additive conditional quantiles with high-dimensional covariates, *Journal of the American Statistical Association* **98**(461): 135–146.

de Oliveira, M. C. F. and Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey, *IEEE Transactions on Visualization and Computer Graphics* **9**(3): 387–394.

Demartines, P. and Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets, *IEEE Transactions on Neural Networks* **8**(1): 148–154.

Derksen, S. and Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables, *British Journal of Mathematical and Statistical Psychology* **45**: 265–281.

Doksum, K. and Koo, J.-Y. (2000). On spline estimators and prediction intervals in nonparametric regression, *Computational Statistics & Data Analysis* **35**(1): 67–82.

Donoho, D. L., Elad, M. and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE Transactions on Information Theory* **52**(1): 6–18.

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Sciences* **100**(10): 5591–5596.

Dy, J. D. and Brodley, C. E. (2004). Feature selection for unsupervised learning, *Journal of Machine Learning Research* **5**: 845–889.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *The Annals of Statistics* **32**(2): 407–499.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall / CRC.

Erwin, E., Obermayer, K. and Schulten, K. (1992). Self-organizing maps: Ordering, convergence properties and energy functions, *Biological Cybernetics* **67**(1): 47–55.

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.

Fuchs, J. J. (2005). Recovery of exact sparse representations in the presence of bounded noise, *IEEE Transactions on Information Theory* **51**(10): 3601–3608.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds, *Technometrics* **16**(4): 499–511.

Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural networks and the bias/variance dilemma, *Neural Computation* **4**(1): 1–58.

George, E. I. (2000). The variable selection problem, *Journal of the American Statistical Association* **95**(452): 1304–1308.

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P. (1997). Statistical themes and lessons for data mining, *Data Mining and Knowledge Discovery* **1**: 11–28.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**(3/4): 325–338.

Guyon, I., Alamdari, A. R. S. A., Dror, G. and Buhmann, J. M. (2006). Performance prediction challenge, *Proceedings of the IEEE International Joint Conference on Neural Networks - IJCNN 2006*, pp. 2958–2965.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**: 1157–1182.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.

Himberg, J. (2000). A SOM based cluster visualization and its application for false coloring, *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks – IJCNN 2000*, Vol. 3, pp. 587–592.

Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J. and Simula, O. (2001). The self-organizing map as a tool in knowledge engineering, *in* N. R. Pal (ed.), *Pattern Recognition in Soft Computing Paradigm*, World Scientific Publishing, pp. 38–65.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (1996). *Convex Analysis and Minimization Algorithms I*, Springer-Verlag.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics* **12**(3): 55–67.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**: 417–520.

Hotelling, H. (1936). Relations between two sets of variates, *Biometrika* **28**(3/4): 321–377.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms, *The Annals of Statistics* **33**(4): 1617–1642.

Huo, X. and Ni, X. (2007). When do stepwise algorithms meet subset selection criteria?, *The Annals of Statistics* **35**(2): 870–887.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*, Prentice Hall.

John, G. H., Kohavi, R. and Pfleger, K. (1994). Irrelevant features and the subset selection problem, *Proceedings of the 11th International Conference on Machine Learning*, pp. 121–129.

Jolliffe, I. (1986). *Principal Component Analysis*, Springer Verlag.

Jones, A. J. (2004). New tools in non-linear modelling and prediction, *Computational Management Science* **1**(2): 109–149.

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection, *Journal of the American Statistical Association* **99**(465): 279–290.

Kaski, S. and Lagus, K. (1996). Comparing self-organizing maps, *Proceedings of the International Conference on Artificial Neural Networks – ICANN 1996*, pp. 809–814.

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P. and Castrén, E. (2003). Trustworthiness and metrics in visualizing similarity of gene expression, *BMC Bioinformatics* **4**(48).

Kaski, S., Nikkilä, J., Sinkkonen, J., Lahti, L., Knuuttila, J. E. A. and Roos, C. (2005). Associative clustering for exploring dependencies between functional genomics data sets, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**(3): 203–216.

Kaski, S., Sinkkonen, J. and Peltonen, J. (2001). Bankruptcy analysis with self-organizing maps in learning metrics, *IEEE Transactions on Neural Networks* **12**(4): 936–947.

Kaski, S., Venna, J. and Kohonen, T. (1999). Coloring that reveals high-dimensional structures in data, *Proceedings of the International Conference on Neural Information Processing – ICONIP 1999*, Vol. 2, pp. 729–734.

Kim, Y., Kim, J. and Kim, Y. (2006). Blockwise sparse regression, *Statistica Sinica* **16**(2): 375–390.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps, *Proceedings of the International Conference on Neural Networks – ICNN 1996*, pp. 294–299.

Kiviluoto, K. (1998). Predicting bankruptcies with the self-organizing map, *Neurocomputing* **21**(1–3): 191–201.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica* **46**(1): 33–50.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection, *Artificial Intelligence* **97**(1-2): 273–324.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics* **43**(1): 56–69.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd edn, Springer.

Laine, S. and Similä, T. (2004). Using SOM-based data binning to support supervised variable selection, *in* N. R. Pal, N. Kasabov, R. K. Mudi, S. Pal and S. K. Parui (eds), *11th International Conference on Neural Information Processing - ICONIP 2004, Proceedings*, Vol. 3316 of *Lecture Notes in Computer Science*, Springer, pp. 172–180.

Lampinen, J. and Kostiainen, T. (2000). Self-organizing map in data-analysis – Notes on overfitting and overinterpretation, *Proceedings of the European Symposium on Artificial Neural Networks – ESANN 2000*, pp. 239–244.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm, *Journal of the Royal Statistical Society, Series B* **57**(2): 425–437.

Lee, J. A., Lendasse, A. and Verleysen, M. (2004). Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis, *Neurocomputing* **57**: 49–76.

Li, J. X. (2004). Visualization of high-dimensional data with relational perspective

map, *Information Visualization* **3**(1): 49–59.

Li, K.-C. (1991). Sliced inverse regression for dimensionality reduction, *Journal of the American Statistical Association* **86**(414): 316–327.

Li, X. and Harezlak, J. (2007). Chapter 5: Visualization in genomics and proteomics, *in* W. Dubitzky, M. Granzow and D. P. Berrar (eds), *Fundamentals of Data Mining in Genomics and Proteomics*, Springer, pp. 103–122.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces, *Journal of the American Statistical Association* **102**(477): 255–268.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *in* L. M. Le Cam and J. Neyman (eds), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, pp. 281–297.

Malioutov, D., Çetin, M. and Willsky, A. S. (2005). A sparse signal reconstruction perspective for source localization with sensor arrays, *IEEE Transactions on Signal Processing* **53**(8): 3010–3022.

Maniyar, D. M. and Nabney, I. T. (2006). Data visualization with simultaneous feature selection, *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 156–163.

McCabe, G. P. (1984). Principal variables, *Technometrics* **26**(2): 137–144.

Meier, L., van de Geer, S. and Bühlmann, P. (2006). The group lasso for logistic regression, *Technical Report 131*, Eidgenössische Technische Hochschule, Seminar für Statistik, Zürich, Switzerland.

Miller, A. J. (2002). *Subset Selection in Regression*, 2nd edn, Chapman & Hall.

Ni, X. S. and Huo, X. (2006). Regressions by enhanced leaps-and-bounds via additional optimality tests (LBOT), *Technical report*, Georgia Institute of Technology.

Oh, I.-S., Lee, J.-S. and Moon, B.-R. (2004). Hybrid genetic algorithms for feature selection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(11): 1424–1437.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis* **20**: 389–404.

Park, M. Y. and Hastie, T. (2006). Regularization path algorithms for detecting gene interactions, *Technical Report 13*, Stanford University, Department of Statistics.

Peltonen, J. and Kaski, S. (2005). Discriminative components of data, *IEEE Transactions on Neural Networks* **16**(1): 68–83.

Pölzlbauer, G., Dittenbach, M. and Rauber, A. (2006). Advanced visualization of self-organizing maps with vector fields, *Neural Networks* **19**(6–7): 911–922.

Rossi, F., Lendasse, A., François, D., Wertz, V. and Verleysen, M. (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, *Chemometrics and Intelligent Laboratory Systems* **80**(2): 215–226.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**: 2323–2326.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers* **C-18**(5): 401–409.

Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression, *Technometrics* **46**(4): 421–429.

Sparks, R. S., Zucchini, W. and Coutsourides, D. (1985). On variable selection

in multivariate regression, *Communications in Statistics – Theory and Methods* **14**(7): 1569–1587.

Srivastava, M. S. and Solanky, T. K. S. (2003). Predicting multivariate response in linear regression model, *Communications in Statistics – Simulation and Computation* **32**(2): 389–409.

Takeuchi, I., Le, Q. V., Sears, T. D. and Smola, A. J. (2006). Nonparametric quantile estimation, *Journal of Machine Learning Research* **7**: 1231–1264.

Tenenbaum, J. B. (1998). Mapping a manifold of perceptual observations, *in* M. I. Jordan, M. J. Kearns and S. A. Solla (eds), *Advances in Neural Information Processing Systems 10*, MIT Press, Cambridge, MA, pp. 682–688.

Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science* **290**: 2319–2323.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**(1): 267–288.

Tipping, M. E. and Lowe, D. (1998). Shadow targets: A novel algorithm for topographic projections by radial basis functions, *Neurocomputing* **19**(1–3): 211–222.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method, *Psychometrika* **17**(4): 401–419.

Tropp, J. A. (2006a). Algorithms for simultaneous sparse approximation. Part II: Convex relaxation, *Signal Processing* **86**(3): 589–602.

Tropp, J. A. (2006b). Just relax: Convex programming methods for identifying sparse signals in noise, *IEEE Transactions on Information Theory* **52**(3): 1030–1051.

Tukey, J. (1977). *Exploratory Data Analysis*, Addison-Wesley.

Turlach, B. A., Venables, W. N. and Wright, S. J. (2005). Simultaneous variable selection, *Technometrics* **47**(3): 349–363.

Ultsch, A. and Siemon, H. P. (1990). Kohonen's self-organizing feature maps for exploratory data analysis, *Proceedings of the International Neural Network Conference – INNC 1990*, pp. 305–308.

Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley and Sons.

Venna, J. and Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: an experimental study, *in* G. Dorffner, H. Bischof and K. Hornik (eds), *Artificial Neural Networks - ICANN 2001, International Conference, Proceedings*, Vol. 2130 of *Lecture Notes in Computer Science*, Springer, pp. 485–491.

Venna, J. and Kaski, S. (2006). Local multidimensional scaling, *Neural Networks* **19**(6–7): 889–899.

Verbeek, J. (2006). Learning nonlinear image manifolds by global alignment of local linear models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(8).

Vesanto, J. (1999). SOM-based data visualization methods, *Intelligent Data Analysis* **3**(2): 111–126.

Vesanto, J. (2002). *Data Exploration Process Based on the Self-Organizing Map*, PhD thesis, Helsinki University of Technology.

Vesanto, J. and Ahola, J. (1999). Hunting for correlations in data using the self-organizing map, *Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications – CIMA 1999*, pp. 279–285.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map, *IEEE Transactions on Neural Networks* **11**(3): 586–600.

Vesanto, J., Sulkava, M. and Hollmén, J. (2003). On the decomposition of the self-organizing map distortion measure, *Proceedings of the Workshop on Self-Organizing Maps – WSOM 2003*, pp. 11–16.

Villmann, T., Der, R., Herrmann, M. and Martinetz, T. M. (1997). Topology preservation in self-organizing feature maps: Exact definition and measurement, *IEEE Transactions on Neural Networks* **8**(2): 256–266.

Webb, A. R. (1995). Multidimensional scaling by iterative majorization using radial basis functions, *Pattern Recognition* **28**(5): 753–759.

Weinberger, K. Q., Blitzer, J. and Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification, *in* Y. Weiss, B. Schölkopf and J. Platt (eds), *Advances in Neural Information Processing Systems 18*, MIT Press, Cambridge, MA, pp. 1473–1480.

Weinberger, K. Q. and Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming, *International Journal of Computer Vision* **70**(1): 77–90.

Wolf, L. and Shashua, A. (2005). Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach, *Journal of Machine Learning Research* **6**: 1855–1887.

Yin, H. (2002). ViSOM – A novel method for multivariate data projection and structure visualization, *IEEE Transactions on Neural Networks* **13**(1): 237–243.

Young, G. and Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances, *Psychometrika* **3**(1): 19–22.

Yu, K. and Lu, Z. (2004). Local linear additive quantile regression, *Scandinavian Journal of Statistics* **31**: 333–346.

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.

Zhang, Z. (2003). Learning metrics via discriminant kernels and multidimensional scaling: Toward expected Euclidean representation, *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 872–879.

Zhang, Z. and Zha, H. (2005). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, *SIAM Journal on Scientific Computing* **26**(1): 313–338.

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

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics* **15**(2): 265–286.