Penalized Estimation in High-Dimensional Data Analysis

Johan de Rooi

Printed by: Ipskamp drukkers, Enschede
Cover design: Johan de Rooi

ISBN: 978-94-6191-898-7

Penalized Estimation
in High-Dimensional Data Analysis

Schatten met penalties
in hoog-dimensionale data-analyse

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
donderdag 24 oktober 2013 om 11:30 uur

door

Jacob Jan de Rooi

geboren te Woerden

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie

Promotoren:	Prof.dr.ing. P.H.C. Eilers
	Prof.dr. P. van der Spek
Overige leden:	Dr. J.P.P. Meijerink
	Prof.dr. L.M.C. Buydens
	Prof.dr. A. Smilde

Contents

Introduction 1

A common task in statistical practice is the estimation of unknown parameters from available data. When proposing a model one could rely on unbiased estimators, meaning that the expected value of an estimator equals the true parameter being estimated. Historically, the absence of bias is often considered an attractive property of an estimator, because it allows intuitive interpretation of the results. However, a fixation on bias reduction increases the variance, which deteriorates the predictive performance of a model. It also occurs that the data do not provide enough information to produce well-behaved unbiased estimates. In these so called ill-conditioned problems, unbiased estimators will over-fit the data, or will be impossible to obtain.

This work focusses on ill-conditioned problems in high-dimensional data analysis. Typical examples of high-dimensional data are signals and spectra in analytical chemistry, image data in computer vision or microarray data in biology. In all chapters penalized estimators, or often called shrinkage estimators, are used to obtain estimates. This means that the usual loss function is augmented with some type of penalty function that constrains or shrinks the coefficients in the model.

The concept of ill-posedness is further introduced in the next section. A short introduction into penalized estimation is given in section 1.2. Section 1.3 provides a chapter by chapter introduction to the different topics within this thesis. In addition it provides some more details concerning the data and its technological aspects, not treated in the particular chapters.

1.1 Problem setting

A problem is well-posed if it meets the following three conditions (Vapnik, 1999):

- A solution exists
- The solution is unique
- The solution is stable

If one or more conditions are violated one speaks of an ill-posed problem. In this thesis we are faced with situations where an (approximate) solution exists. But because there are too many degrees of freedom the solution is not unique. This is for instance the case when estimating a smooth curve from observed data; the intended estimates contain much more detail compared to the 'coarse' data. A second example occurs when the number of variables is larger than the number of observations, which causes the estimated covariance matrix to be singular. Unstable estimates are often the result of overfitting, meaning that a too complex model is fitted. It leads to large variances of the estimates and it makes the model sensitive for changes in the data. A specific case occurs when the data matrix is full rank, but some columns are nearly linearly dependent, causing multicollinearity between the variables. The performance of these models is generally unsatisfactory.

One solution is to simplify the model. Feature selection and extraction are the classical options. In this thesis we rely on penalized estimation. It enables model estimation, prevents overfitting and the variance will be reduced, leading to better predictions. This however comes at the cost of an increased bias; the difference between the expectation of the estimator and the true value of the parameter. Bias and variance of a model can be expressed in terms of the mean squared error (MSE):

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2, \tag{1.1}$$

with the variance depending on the sample size. A graphical representation of the bias-variance trade-off, as a function of model complexity, is given in Figure 1.1. Penalization or shrinkage allows model tuning on a continuous scale, in contrast to the discrete behaviour of variable selection and extraction. The model complexity is regulated by the weight of the penalty. Varying the weight affects both the bias and variance, an optimal trade-off can be found using cross-validation or some type of information criterion. A thorough discussion of bias and variance is found in Friedman (1997).

1.2 Penalized estimation

Penalized methods have evolved from different directions. A a rough (but slightly artificial) distinction can be made between size penalties and roughness penalties, often called smoothers. The introduction of shrinkage estimators in the multivariate setting is largely due to Stein (1956). He showed that in many cases it is possible to find a biased estimator with a better bias variance trade-off compared to an unbiased one. The concept of penalization is however much older. Penalizing the differences of adjacent values in the fit to a series of datapoints was introduced by Whittaker (1923).

Size penalties

Size penalties merely shrink coefficients towards zero or to given values. A prominent shrinkage estimator is ridge regression, introduced in statistics by Hoerl in 1962 and popularized by Hoerl and Kennard (1970). The goal of this procedure is to circumvent multicollinearity problems in regression. In other communities ridge regression is known

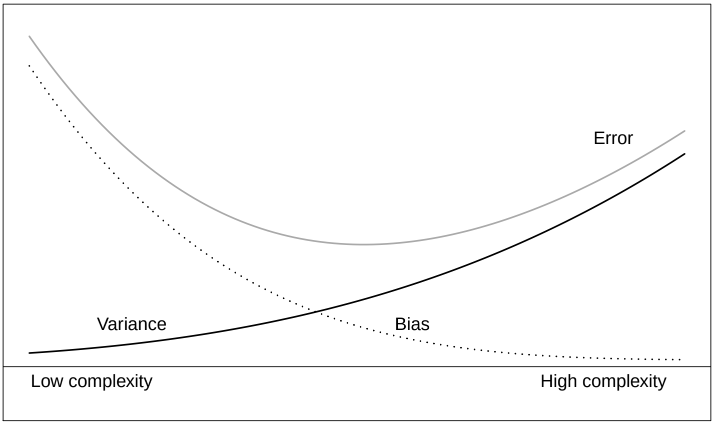


Figure 1.1: Bias variance trade-off as a function of the model complexity

as Tikhonov regularization (Tikhonov, 1963), as weight decay (see e.g. Bishop, 2006) or simply as L_2 regularization.

Ridge regression adds a quadratic penalty function to the usual ordinary least squares cost function. Before estimation, the data are centered so we can leave out the intercept. We minimize:

$$Q_2 = ||y - X\beta||_2^2 + \lambda||\beta||_2^2, \tag{1.2}$$

with the squared L_2 norm

$$||\beta||_2^2 = \sum_j \beta_j^2. \tag{1.3}$$

The solution can be obtained in closed form and is given by:

$$\beta = (X^T X + \lambda I)^{-1} X^T y. \tag{1.4}$$

The parameter λ is the tuning parameter. Sometimes it is called the bias parameter, related to the increased bias in favour of a reduction of the variance. The ridge regression alleviates problems due to collinearity, but because the penalty function takes the square of the model parameters, it does not yield a parsimonious model.

A penalty that exhibits the advantages mentioned above and also performs variable selection, is the L_1 norm penalty (in the following also abbreviated to L_1 penalty). It was introduced by Tibshirani (1996) as the least absolute shrinkage estimator (Lasso) and, independently, by Chen et al. (1998) as basis pursuit. Instead of taking the the square of the coefficients as in the case of ridge regression, the L_1 penalty takes the absolute values. The objective function is defined:

$$Q_1 = ||y - X\beta||_2^2 + \lambda||\beta||_1, \tag{1.5}$$

where $||b||_1$ denotes the L_1 norm

$$||\beta||_1 = \sum_j |\beta_j|. \tag{1.6}$$

Different from L_2 regularization, the L_1 penalty cannot be solved in closed form. However, the objective function is convex and thus a solution can be calculated relative efficiently. A large number of optimization schemes are available for the Lasso, some of them are discussed in Hastie et al. (2009) or Osborne et al. (2000).

Inspired by the success of the Lasso and in order to meet requirements in specific applications, researchers have proposed various other types of shrinkage estimators. Two examples are the grouped Lasso (Yuan and Lin, 2006) and the elastic net (Zou and Hastie, 2005). Both target at shrinking groups of variables instead of single features. In high-dimensional data the Lasso can generate a large amount of bias (see e.g. Zhang, 2011). Zou (2006) proposed the adaptive lasso as an alternative, which includes data-dependent weights and can obtain sparser results. A second option is to use the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), designed to penalize large parameters less heavily.

Similar to classical variable subset selection, it is also possible to penalize the non-zero elements in the model which results in the L_0 norm penalty:

$$Q_0 = ||y - X\beta||_2^2 + \lambda||\beta||_0 \tag{1.7}$$

with the L_0 norm

$$||\beta||_0 = \sum_j I_{\beta_j \neq 0} \tag{1.8}$$

The disadvantage of the L_0 norm penalty is that it results in a non-convex optimization problem. Nonetheless a number of successful optimization strategies have been proposed (see e.g. Blumensath and Davies, 2008; Candes et al., 2008; Bruckstein et al., 2009). Good results in signal recovery are reported by (Xiang et al., 2012). In chapters 3 and 4 of this thesis we demonstrate the value of the L_0 norm penalty in the context of sparse deconvolution of signal data. In chapter 7 it is used for sparse network estimation. A review of various types of penalties related to the Lasso is provided by Hesterberg et al. (2008); an additional set of penalties, used in imaging, is presented by Starck and Murtagh (2006).

Roughness penalties

Roughness penalties utilize the idea that observed data should consist of a smooth signal that is corrupted by noise. This notion of smoothness is useful in many areas of statistics. In this thesis, smoothers are used for density estimation, noise removal and functional modelling tasks. Other types of applications can be found in for instance Simonoff (1996) on smoothing of categorical data, Ramsay and Silverman (2005), discussing functional data analysis, and Hastie and Tibshirani (1990), introducing the generalized additive model (GAM).

The smoothers we use are the Whittaker smoother and P-splines. References to other methods are given at the end of the section. The Whittaker smoother is defined as (Eilers, 2003):

$$Q_W = \sum_{i=1}^{n}(y_i - z_i)^2 + \lambda \sum_{i=2}^{n}(\Delta z_i)^2. \tag{1.9}$$

Roughness is defined using first order differences, with Δ the differencing operator: $\Delta z_i = z_i - z_{i-1}$. The contribution of the penalty to Q is regulated by the tuning parameter λ. Increasing λ will result in a smoother curve. We can translate (1.9) into matrix form:

$$Q_W = ||y - z||_2^2 + \lambda ||Dz||_2^2, \tag{1.10}$$

with D is the difference matrix such that $Dz = \Delta z$. In the case with $n = 5$, the matrix is:

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \tag{1.11}$$

The penalty uses the squared L_2 norm; the sum of the squared differences of adjacent values. Applying (1.10) results is a smoothly varying curve, if more sudden jumps are required one could use the L_1 norm (Eilers and de Menezes, 2005). Results close to a step function can be obtained using the L_0 norm, as recently demonstrated by Rippe et al. (2012).

For various application the Whittaker smoother has turned out to be a very useful tool. However, when the data are irregularly sampled (i.e. not at equal distances), it cannot be applied. In addition, no reduction of the parameter vector takes place, meaning that for large datasets the systems to solve can become very large. In these cases P-splines can be applied, which are defined as B-splines with a penalty on the differences. B-splines are bell-shaped curves, composed of q+1 polynomial pieces, each of degree q. The polynomial pieces join at q points, called the knots, and together form a smooth bell-shaped curve. Figure 1.2 shows the individual polynomial pieces, the resulting smooth spline function and the positions of the knots, for a single B-spline. Standard references to the B-spline theory are Dierckx (1995) and de Boor (2001).

Using a series of equally spaced B-splines, a smooth curve $\hat{y}$ can be fitted to the observed data (x, y). If we let the data be represented by B-spline coefficients we get the following:

$$\hat{y} = \sum_{j=1}^{k} \alpha_j B_j(x, q), \tag{1.12}$$

with $B_j(x, q)$ being the value at x of the j-th B-spline of degree q. Fitting B-splines to our data results in the following minimization function:

$$Q_B = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} B_j(x_i)\alpha_j)^2 = ||y - B\alpha||_2^2. \tag{1.13}$$

The B-spline framework requires optimization of the number and positions of the knots. The P-spline approach, proposed by Eilers and Marx (1996), avoids these issues by relying on equally spaced knots and the penalty. The penalty on the differences (hence ”P-splines”), forces adjacent spline coefficients to be more similar:

$$P = \lambda \sum_{j=1}^{k} (\Delta_d \alpha_j)^2. \tag{1.14}$$

Parameter λ tunes the relative weight of the penalty. Combining (6.2) with the penalty results in the following objective function:

$$Q_P = ||y - B\alpha||_2^2 + \lambda ||D_d \alpha||_2^2, \tag{1.15}$$

with D_d the d-th order differences matrix, with most often $d = 1$, $d = 2$ or $d = 3$. When λ is chosen very large a constant fit is obtained when $d = 1$, a linear trend when $d = 2$ and a quadratic fit when $d = 3$. Minimization of (1.15) leads to

$$(B'B + \lambda D_d' D_d)\hat{\alpha} = B'y. \tag{1.16}$$

Interesting is that the Whittaker smoother as seen above is in fact a zero-degree B-spline model (matrix B becomes an identity matrix) with knots half-way between the data points. In addition to the order of the penalty, we can make other changes to the difference matrix D. In chapter 6 of this thesis, the difference matrix is adjusted in order to preserve the circular character of the data. Other options are discussed in Eilers and Marx (2010).

A prominent alternative smoother is the LOESS algorithm (Cleveland, 1979) which relies on local polynomials. A popular choice within signal processing literature is the polynomial least-squares filter (Wentzell and Brown, 2000), and is also known as the Savitzky and Golay filter (Savitzky and Golay, 1964). Model flexibility is tuned by varying the degree of the polynomial and the window width. For an overview of several methods we refer to Hastie et al. (2009). A short account on historical developments can be found in Loader (1999) and Shenoi et al. (2006).

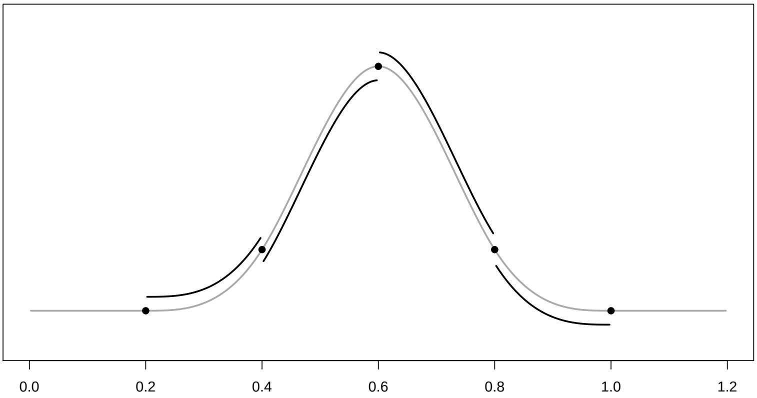


Figure 1.2: One cubic B-spline function, the knots and cubic (shifted) segments.

1.3 Thesis outline

This section provides an introduction to the separate chapters within this thesis. Per chapter the main problem and proposed solutions are shortly described. Often followed by some theoretical background concerning the used datasets.

Chapter 2. Mixture models for baseline estimation

Many signals from chemical instruments show a baseline drift. In this chapter a new baseline estimation procedure is proposed using weighted regression on P-splines. The weights follow from a mixture model with two components, one for the noise around the baseline and the second for the peaks. The method relies on a realistic statistical model, in contrast to many existing baseline correction procedures.

The applications originate from spectroscopy and chromatography. Spectroscopy is based on absorption, emission or scattering of electromagnetic radiation by atoms or molecules (Hollas, 2004). The substance under study may be gas, liquid or solid. The basic idea is to pass an electromagnetic beam through a sample and record the amount of energy returned at different wavelengths. From the recordings a transmittance or absorbance spectrum can be produced, which reveals details about the molecular structure of the sample. Because atoms and molecules have unique spectra, spectroscopy can be used to identify and quantify chemicals within mixtures.

Chromatography is a class of methods for separation of mixtures into the components of the substance (see e.g. McNair and Miller, 1998; Webster and Clark, 2010) . The result of the analysis is presented in a chromatogram: a graph showing the detector response as a curve. It essentially plots the concentration of analyte in the effluent (the mobile phase leaving the column) versus effluent time or volume. A typical chromatogram consists of a number of peaks representing the different separated components of the original mixture. In the clinical laboratory this class of techniques is used to detect complex substances such as drugs and hormones.

Chapter 3. Mixture models for two-dimensional baseline estimation

In chapter 2 we introduced a model to estimate the baseline in one-dimensional signals. In this chapter the concept is extended to cases with two dimensions, meaning we need to model a surface. To model a smooth baseline in the case of one dimensional data, we relied on a P-spline basis. When modelling a surface, a similar methodology can be used: tensor product P-splines. The advantage of the proposed model is that it incorporates the correlation of both dimensions in the data. In addition, it allows different degrees of smoothing in both directions. Although the algorithm is applicable to a wide range of baseline estimation problems, here we only focus on artefact elimination in time-resolved spectroscopy. Other possible applications are baseline estimation for microarrays and electrophoretic data.

The application that is presented comes from femtosecond spectroscopy and more specific pump-probe experiments (Dantus and Gross, 2003). The aim of these experiments is to study the vibration of (excited) molecules on a femtosecond (10^{-15} s) time resolution. The experiment basically consists of two laser pulses. The first, the pump pulse, excites molecules within the studied sample. After a certain delay time, the pump pulse is followed by the probe pulse, and will pass trough the same volume. The intensities of the probe pulse are registered, resulting in an absorption spectrum, similar to a steady state spectrometer. By repeating the experiment, with increasing time between the pump and pulse, one can monitor the dynamics of the system as a function of time. Combining the consecutive spectra results in a surface, with in one dimension information in the spectral domain, and in the second information in the temporal domain. On a more practical level, ultrafast spectroscopy can add to the understanding of e.g. semiconductors, or photosynthetic systems.

Chapter 4. Deconvolution of pulse trains

Chapter 2 and 3 focused on the baseline components of signals. In this chapter the peaks are studied. The peaks are modelled as a convolution of the true input signal, and a blurring function, caused by the instrument and additional sources of noise. This blurring function is called the impulse response function (IRF) or point spread function (PSF). To estimate the input spikes, we use deconvolution; the reverse process. Deconvolution is a broadly applied methodology, applications can be found in for example, astronomy (Starck et al., 2002), seismology (Silvia and Robinson, 1979) and engineering (Wei et al., 2009).

Deconvolution is a typical example of an ill-conditioned regression problem. Knowing that the input signal is sparse, we employ a penalty function with an L_0 norm. It makes the model estimable and to generates sparseness. If not only the input of the system, but also the impulse response is unknown, we speak of blind deconvolution. In these cases one iterates between estimating the input given the response function and updating the IRF given the current input.

One application presented in this chapter comes from chromatography, similar to data discussed in chapter two. A second example deals with DNA sequencing data, derived with electrophoresis. Normally these signals consist of four parallel channels (one for each DNA-base), here we only focus on one of these. A peak in the signal corresponds to a base call. The third data set consists of a series of hormone concentrations in human blood (Vis et al., 2010), measured over time. The aim is to reduce the signal to a series of spikes indicating the exact time and amounts of the hormone releases.

Chapter 5. Sparse deconvolution in one and two dimensions

This chapter is a follow-up of chapter 4. The deconvolution method as laid out in the previous chapter is applied to two-dimensional data. In addition, the blind deconvolution algorithm is adjusted, which makes it applicable in combination with virtually any unimodal response function. A small simulation shows the value of the proposed methodology. Performance of the blind deconvolution algorithm in case of an approximately exponential response function, is illustrated using the hormone release data, as discussed in chapter 4.

To illustrate two-dimensional deconvolution, we use fluorescence microscopy data. The data are a series of images showing actin structures, highlighted using fluorophores. The goal is to give a more detailed description of the actin filaments, but is difficult given the densely packed fluorophore molecules. Because the fluorophores blink in a random fashion, closely positioned molecules can be resolved in time. These properties are ultimately used to obtained super-resolution images.

Chapter 6. Classification of outlines using penalized signal regression

Many data analysis problems involve a classification problem. When the number of variables is small, the classification procedure can be kept relative simple. If their number is large, or when they are highly correlated, things are more complicated. Two-dimensional shape classification is a typical example. A solution to these problems is to apply some kind of summarization or variable extraction. A frequently used tool in this respect is principal component analysis. The results derived from these methods are generally difficult to interpret. This chapter presents an approach for classification of two dimensional shapes, without the need of summarizing the data.

In a first step the rectangular coordinates are converted into polar coordinates, which results in a set of 'signals'. Next, the data are summarized on a large (circular) P-spline basis. This facilitates preprocessing like rotation and noise removal, while essentially no detail is lost. For classification we rely on logistic penalized signal regression based upon the algorithm proposed by Marx and Eilers (1999), developed to analyse signals in spectroscopy. The method uses the spline coefficients as explanatory variables. The proposed methodology has several advantages. By analysing the complete outlines instead of a restricted number of features, no information is lost in advance of the classification. Preprocessing, like rotation or smoothing, is easily performed on the curves. The PSR algorithm is a variant of the generalized linear model (GLM) and as such complete outlines can be analysed using well established statistical tools. In addition the model is easily extended to three dimensions, other types of outcomes, or additional covariates can be included using additive modelling.

Three applications are presented. The motivating example is about abnormal head shapes, caused by premature fusion of one or more cranial sutures (Craniosynostosis). Depending on which sutures fuse, different deformities occur. Classifying individual skulls, into the right group is important for decisions on possible treatment. The actual data are a set of x, y coordinates, corresponding to one skull outline taken at a fixed height and coming from a CT-scan. Because of the small sample size, this dataset is used only for description. In a second application the aim is to distinguish two types of (fossil) rodents, using the outline of a molar. The dataset comes from Claude (2008) and contains 60 samples in total. A third example comes from biology and deals with the classification of diatoms; algae with a large variety of different shapes and of a size ranging between 2 and 200 micrometer. This dataset is described in Jalba et al. (2005).

Chapter 7. Exploring network structure using penalties and priors

Gene expression data have been a major research area within biology in the last years. By monitoring the activity levels of a large number of genes (measured by the amount of RNA they produce), researchers try to find links between genes and diseases. In addition, it is assumed that genes act in groups of interacting features. The aim is to establish direct relations between genes and represent these in a network. The estimated networks are supposed to be sparse, meaning that the number of relations is small compared to the total number of possible pairwise connections. Network estimation is often difficult due to the large number of variables compared to the number of samples. In this chapter we test the ability of three shrinkage estimators to reproduce a series of simulated networks. In a next step we try to utilize the vast amount of information stored online in the process of network estimation. We use an adaptation of the weighted lasso to combine the proposed prior with the data. Depending on the type of information available, the prior can be more or less specified with respect to the loadings of the individual edges. Two applications show the flexibility of the method; in one the prior is built using text mining. In the second, pathways that are found in a specific publication are used as prior information.

Chapter 8. Pseudo-Bayes smoothing of tables with very low counts

Chapter 8 applies penalized estimation to zeros or low counts in contingency tables. Low counts can occur due to the rare nature of the observed event or when many categories are recorded. These low counts are problematic for statistical inference and correcting these is advised. We propose a linear shrinkage procedure, based upon the available data in the study. The optimal balance between the data and the prior is determined using an adjusted Akaike information criterion. The procedure is general applicable, but here we focus on meta analysis of clinical trials, in which the observed data consists of series of two by two tables.

Chapter 9. Conclusions

In the last chapter of this thesis we repeat the major conclusions from the different chapters. In addition, a short more general discussion focusses on the interpretation of penalties within the statistical framework. Proposals for future research are made where possible.

Mixture models for baseline estimation

2

Abstract. Various instruments produce data consisting of a series of more or less isolated peaks, superimposed on a drifting baseline. The positions and heights of the peaks are of interest and the baseline is a nuisance. We model a smooth baseline by weighted regression on P-splines, a combination of B-splines and a discrete penalty to tune smoothness. The weights are computed from a mixture model with two component distributions, relative to the baseline, one for noise, the other for the peaks. The algorithm is fast and it shows excellent performance on simulated and experimental data.

2.1 Introduction

A variety of instruments deliver signals that consist of a series of more or less isolated peaks. The physical or chemical information is in the positions and heights of the peaks. Ideally the baseline should be flat, but this is seldom the case. In practice slow, but strong, fluctuations are seen, which are known as background, drift, or baseline. The same scenario holds for images or scans of various types. The image presents valuable information against an interfering background. Next to signal and background, random noise is generally present.

The aim of this chapter is to estimate a baseline from a single data series, exploiting two characteristic properties: the baseline changes smoothly, while all peaks have the same sign, either positive or negative. After estimation, the baseline is subtracted, and peak analysis can follow.

In the literature various strategies to estimate the baseline or background can be found, covering disciplines like bioinformatics, chemometrics and computer vision. We do not give an extensive review here, as this was recently done by Liland et al. (2010), primarily

This chapter is published as: de Rooi J.J. and Eilers P.H.C. (2012). Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, 117:56-60.

for spectroscopic data and by Komsta (2011) in the area of chromatography. We do mention some of the more popular algorithms. A simple approach is smoothing in a moving window method as applied by e.g. Ressom et al. (2007) or Coombes et al. (2003). The well-known Savitzky-Golay smoothing algorithm (Savitzky and Golay, 1964) has been adapted to baseline estimation (Wang et al., 2003). More advanced approaches rely on curve fitting with asymmetric weights. Examples are asymmetric least squares, as presented and applied by the authors (Eilers, 2004; de Rooi and Eilers, 2011), or quantile regression, as proposed by Komsta (2011). The idea is that observations located on peaks get a very small weight. The baseline can be described by a polynomial (Gan et al., 2006), this works well in simple cases. In complicated cases, of which we show an example in Figure 2.4, the polynomial needs to have very large degrees, leading to (numerical) instabilities.

Here we propose to use a mixture model. This kind of model is often used for cluster analysis. The actual size and composition of the clusters are the unknown components of the mixture and have to be estimated from the data. Signal data can be considered as a mixture with two components: a baseline including noise and peaks. Any data point of the signal has a probability to be on a peak, or on the baseline. Various books offer a quick introduction into mixture models, often in the context of clustering, an example is Bishop (2006). A book dedicated to the topic is McLachlan and Peel (2000).

The points on the baseline follow a smooth curve plus noise (assumed to be normally distributed, with unknown variance). For all the points on peaks we assume that they have been drawn from an unknown distribution on the positive (or negative, for down-pointing peaks) axis. The baseline is modelled as a smooth curve, using P-splines, a combination of B-splines and a discrete roughness penalty (Eilers and Marx, 1996). After subtraction of the baseline we get a signal with a two-component probability density. One component is related to the peaks and it is approximated by a uniform distribution on the positive axis. The other component is for the observation noise around the baseline; it is assume to be normal. Combining all this with the EM algorithm (Dempster et al., 1977) for parameter estimation, an effective algorithm is constructed. The method is illustrated on experimental data.

2.2 The model

The model with a flat baseline

We have a single data series y composed of n individual observations: $\{y_i, \ldots, y_n\}$. To simplify the presentation, we first consider the case of a constant but unknown background level μ. We assume normally distributed noise with an unknown standard deviation σ. We neglect the order of the observations, and we assume that they are drawn from a probability density $f(y)$. If an observation is on the baseline, it is drawn from the normal density $g(y|\mu, \sigma)$ while if it is on a peak, it is drawn from an unknown probability density $h(y-\mu)$, which is only supported on the positive real axis. Of course, we do not know if an observation is on the baseline or not, so we have to consider the mixture of densities:

$$f(y) = \pi g(y|\mu,\sigma) + (1-\pi)h(y-\mu), \tag{2.1}$$

where the unknown mixing ratio is denoted by π.

We use the established Expectation-Maximization algorithm to estimate the components of the mixture (McLachlan and Peel, 2000). It repeats the following two steps until convergence is reached. In the $expectation$ or E step, the current values of the parameters are used to calculate the posterior probabilities for the baseline and peaks of all data points. During the $maximization$ or M step, the parameters p, μ, σ, and the estimate for $h(.)$ are updated, given the calculated probabilities.

The details of our EM algorithm are as follows. Suppose we knew approximations to all parts of (2.1). Then we could compute the posterior probability p_i that observation y_i comes from density $g(y|\mu,\sigma)$ as:

$$p_i = \frac{\pi g(y|\mu,\sigma)}{\pi g(y|\mu,\sigma) + (1-\pi)h(y-\mu)}. \tag{2.2}$$

Then we can use p and y to compute weighted estimates of μ and σ (with weights p). And we can feed 1-p to a density smoother to estimate $h(.)$ and take the average of p to estimate π. From there we can start a next round. Proper starting values are needed. We skip the details here, as they will be discussed for the general case of a drifting baseline.

We mentioned density smoothing for estimating $h(.)$, the density of the observations on the peaks. Actually, we can simplify this component of the model to a uniform density (between zero and the maximum of $y-\mu$) without any problem. We are not interested in a precise density, but only need good estimates of the posterior probabilities p. For y_i near μ, p_i is very near to 1, and at a distance larger than 3σ from μ it will be essentially 0, whatever the shape of $h(.)$. The simplification will have some influence on weights not near 0 or 1, but they are a minority.

Figure 2.1 shows three panels which explain the basics of the proposed method. The simulated data consists of a constant baseline and a few positive peaks. The example serves as an illustration of the major concepts of the model. In the left upper panel of Figure 2.1 the raw data (colored dots) and estimated baseline (the black line) are shown. Each data point in the figure is either green or blue and depicts the assignment to respectively the baseline (if $p > 0.5$) or the peaks. In the lower left panel the posterior probabilities for the baseline, estimated using the EM algorithm, are shown. This figure can be viewed in parallel to the upper panel and shows strong discrimination between data on the baseline and data points which are part of the peaks.

In the right panel the raw data are presented using a histogram. The x-axis of the histogram forms the actual range of the observed data (as plotted in the upper left panel). The two colors in the plot represent the different components of the mixture: green for the

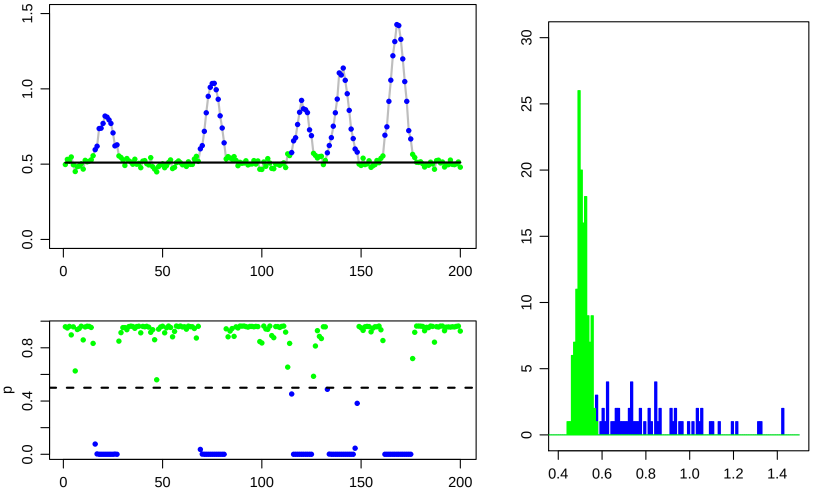


Figure 2.1: Estimation of a constant baseline with the mixture model. The upper left panel shows the data and the estimated baseline in black. In the lower left panel the posterior probabilities for the two mixture components are presented. In the right panel the histograms of both mixture components are shown, with in green the baseline and in blue the peaks.

baseline and blue for the peaks (depending on $p > 0.5$, or not). For this example, most of the mass in the histogram is situated at the left which is due to the prominent baseline and only few peaks.

The model with a drifting baseline

Now it is easy to see how more complicated baseline models can be constructed: specify μ as a curve, and apply the mixture model after subtraction of this curve. The model then looks as follows: $y_i = \mu_i + v_i$, with density,

$$f(y_i) = \pi g(v|\mu_i, \sigma) + (1 - \pi)h(y_i - \mu_i), \tag{2.3}$$

with g being a normal distribution and σ unknown. Again $h(.)$ is not specified; we do not try to estimate it reliably, but instead approximate it by a uniform distribution. The only concern with respect to this component is to properly estimate the posterior probabilities of the data points. Removing the baseline μ gives us the mixture model for $v = y - \mu$:

$$f(v) = \pi g(v|0, \sigma) + (1 - \pi)h(v). \tag{2.4}$$

The baseline is modelled using P-splines, which are B-splines with a penalty on the differences of the coefficients, to tune smoothness (Eilers and Marx, 1996). The B-splines part is given by:

$$\mu_i = \mu(x_i) = \sum B_j(x_i)\alpha_j, \tag{2.5}$$

with B being the $n \times m$ cubic B-spline basis with m the number of B-splines. Consider for the moment fitting B-splines to our data, to compute a trend, not a baseline. Then we would minimize

$$S = \sum_i (y_i - \sum B_j(x_i)\alpha_j)^2 = ||y - B\alpha||^2, \tag{2.6}$$

leading to the system of equations

$$B'B\hat{\alpha} = B'y \quad \text{or} \quad \hat{\alpha} = (B'B)^{-1}B'y. \tag{2.7}$$

P-splines extend the objective function (3.2) with a penalty on differences of α:

$$S = \sum_i (y_i - \sum B_j(x_i)\alpha_j)^2 + \lambda \sum (\Delta^d \alpha_j)^2 = ||y - B\alpha||^2 + \lambda||D\alpha||^2, \tag{2.8}$$

where Δ is the differencing operator: it turned out that $d = 3$ gave the best results for all examples. Then $\Delta^d\alpha_j = \Delta^3\alpha_j = \alpha_j - 3\alpha_{j-1} + 3\alpha_{j-2} - \alpha_{j-3}$. Say we summarized a signal using seven B-splines, the corresponding matrix D is given by:

$$D = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 & 0 \\ 0 & 0 & 0 & -1 & 3 & -3 & 1 \end{bmatrix}. \tag{2.9}$$

By forcing third order differences of the coefficients to be small, the variation in coefficients is reduced, leading to a smoother curve compared to when using only B-splines. With the number of B-splines being relatively large, the penalty allows continuous control over smoothness, from a rather wiggly curve to a quadratic line, by increasing λ. The spirit of P-splines is illustrated by Figure 2.2. The same (simulated) data are smoothed with a relatively large B-spline basis for two values of λ.

Of course, we are not interested here in computing a trend. If the posterior probabilities are available, we insert them as weights into (3.4), giving the new objective function

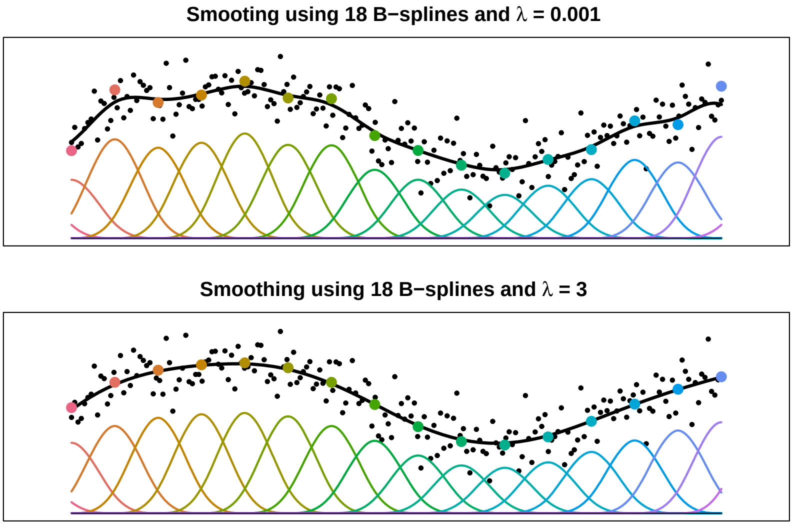
$$S^* = \sum_i p_i(y_i - \sum B_j(x_i)\alpha_j)^2 + \lambda \sum (\Delta^d \alpha_j)^2 = (y - B\alpha)'P(y - B\alpha) + \lambda||D\alpha||^2, \tag{2.10}$$

with $P = \text{diag}(p)$. Again we find an explicit solution:

$$\hat{\alpha} = (B'PB) + \lambda D_d' D_d)^{-1} B'Py, \tag{2.11}$$

A low posterior probability p, means no or little influence of an observation on the estimated baseline, as required.

To make the EM algorithm work, we need proper starting values. We use asymmetric least squares smoothing (Eilers, 2004), with 0.01 or 0.02 as value for the asymmetry parameter, a. Interestingly, the computational structure is almost identical to that of the mixture



Smooting using 18 B−splines and λ = 0.001

Smoothing using 18 B−splines and λ = 3

Figure 2.2: Illustration of P-spline smoothing for small λ (upper panel) and for large λ (lower panel). The basis consist of 18 B-splines. The grey dots are the (simulated) data. The colored curves show the individual B-splines, scaled by their coefficients (which are represented by the large colored dots). The black curve shows the P-spline fit.

model. Instead of the weights in $P = \text{diag}(p)$ in (3.7), one uses $W = \text{diag}(w)$, with $w_i = a$ if $y_i < \mu_i$ and $w_i = 1 - a$ otherwise. Weighted smoothing and re-computation of the weights are repeated until convergence, which usually occurs after a handful of iterations. One can prove that asymmetric least squares smoothing always converges to a unique solution, so for a chosen a and λ this is an automatic starting procedure.

2.3 Applications

In this section we show four applications, with data derived from different types of instruments. The first example is a chromatogram. The example is relatively simple with a slightly drifting baseline and a few strong peaks. In the upper panel of Figure 2.3 we show the data (blue) and the baseline (green). Notice that the estimated baseline is plotted 0.003 units lower in order to show both the baseline and the signal, which are otherwise overlapping in the sections without peaks. We used 50 B-splines and $\lambda = 0.1$. Actually, 50 B-splines are far more than needed for such a smooth baseline, but it was our deliberate choice, to illustrate that the size of the B-spline basis plays no role. The lower panel of Figure 2.3 shows the result after subtraction of the baseline.

For a more challenging example we use data from an FTIR experiment (Phillips and Hamilton, 1996), the results are shown in Figure 2.4. In the upper panel the raw data (blue)

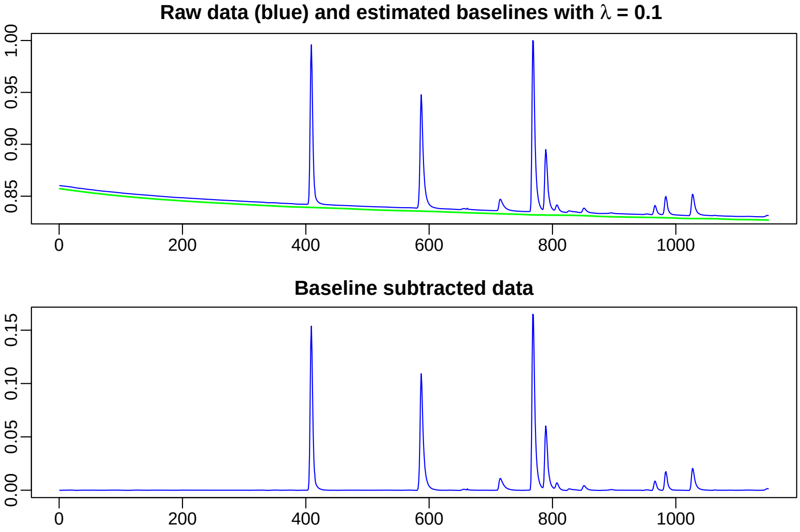


Figure 2.3: Background estimation with the mixture model applied to a chromatogram. The upper panel shows the data (blue) and the estimated baseline (green). In the figure the estimated baseline is plotted 0.003 lower in order to show both baseline and the signal, which are otherwise largely overlapping. The lower panel shows the result of subtracting the baseline.

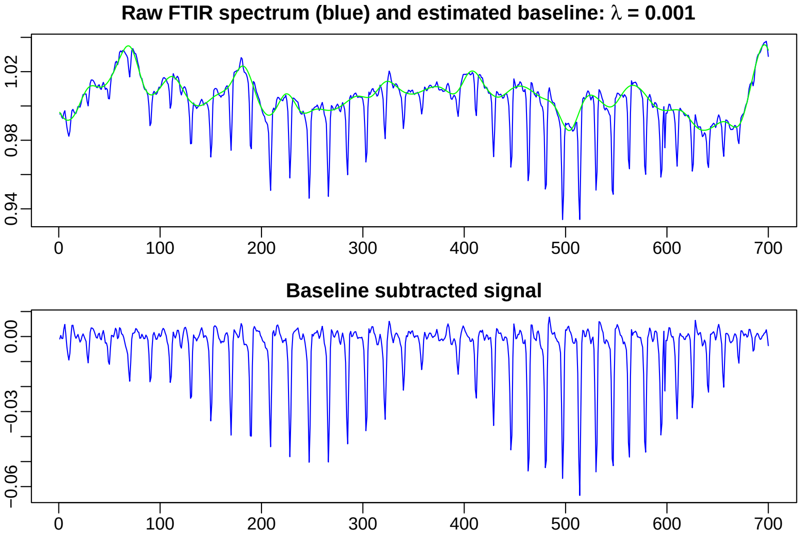


Figure 2.4: Background estimation with the mixture model applied to FTIR data. The upper panel shows the data and estimated baseline. The baseline is estimated on a basis of 50 B-splines. The lower panel shows the data after subtracting the baseline.

and estimated baseline are plotted. The baseline is estimated using 50 B-splines and $\lambda = 0.001$. Subtracting this baseline from the data results in the picture presented in the lower panel of Figure 2.4.

A third example is shown in Figure 2.5; it is a MALDI-TOF mass spectrum of blood serum (de Noo et al., 2005). A baseline is fitted using 200 B-splines and setting $\lambda = 50000$, the data and baseline are plotted in the upper panel of Figure 2.5. In the figure we observe that the amount of baseline noise is relatively large compared to the height of the peaks. Moreover we notice that the error is larger for lower mass-to-charge ratios. Although this is not assumed in the model, the baseline is estimated well.

As a last application we took a segment from a Raman spectrum. The data shows a relatively subtle baseline drift. More difficult in this example is however the presence of many overlapping peaks. The baseline was modelled using 200 B-splines. We present two possible outcomes, in the upper panel of Figure 2.6 we show the data and the estimated baseline, for λ equals 50. According to our opinion this penalty is too mild. A more satisfactory result, with a larger λ, is provided in the lower panel.

2.4 Discussion

We have presented a new approach to baseline estimation and have shown that it performs excellently on real data. The baseline itself is modelled by P-splines allowing a very flexible curve. The assignment of observation to either baseline or peaks is based on probabilities computed from a mixture model that incorporates observational noise.

Our model is similar to asymmetric curve fitting approaches, either by least squares or by quantile regression. However, it avoids the problem of choosing the amount of asymmetry, because that follows implicitly from the parameters π, σ and the model for distribution $h(.)$.

We did not discuss an automatic choice of the value of the penalty parameter λ. Instead we leave this to the user, relying on visual inspection. In principle cross-validation could be used. One fits the model to a subset of the data, chosen randomly or systematically, and computes the log-likelihood of the left-out part of the data. Changing the penalty parameter λ on a grid and computing the cross-validation likelihood for each value will give a curve that hopefully shows a global maximum. Experiments with simulated data seem to indicate that this can work well. But the simulations use independent noise. In real data the noise is frequently correlated, leading to complications. More research is needed here.

We also do not use the order of the observations. In many data sets, stretches of baseline alternate with peaks. Intuitively one feels that this fact should be exploited in some way. But until now we have not found a working procedure to implement it.

Presently we are working on extending our model to series of count data, like X-ray

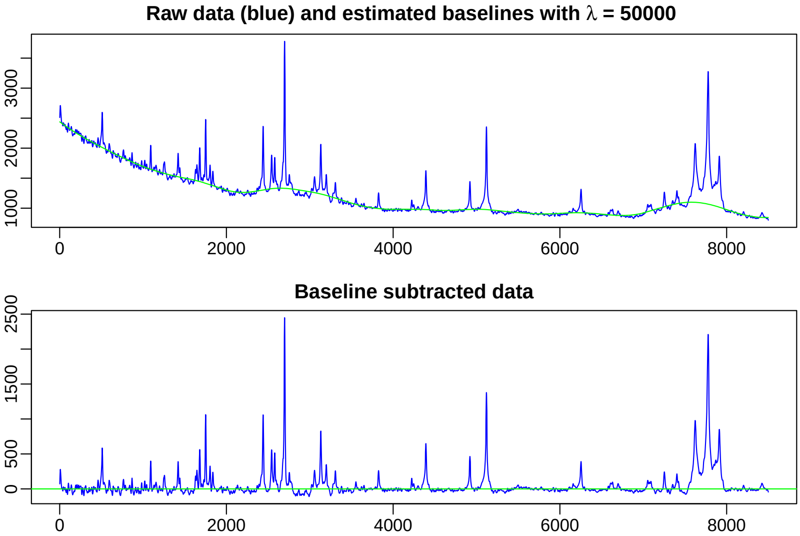


Figure 2.5: Background estimation with the mixture model applied to a MALDI-TOF mass spectrum. The upper panel shows the data and baseline, estimated using 200 B-splines. The lower panel shows the data after subtracting the baseline.

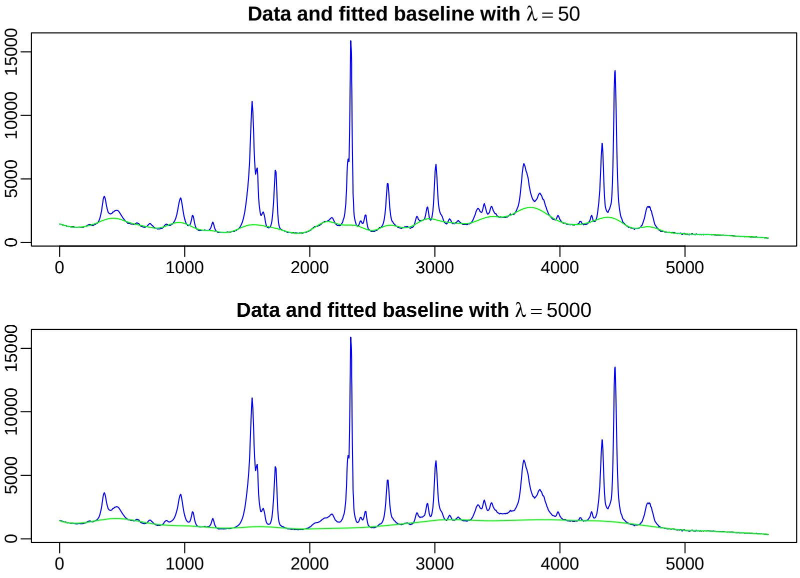


Figure 2.6: Background estimation with the mixture model applied to a Raman spectrum. The upper panel shows the data (blue) and the estimated baseline (green) using a moderate penalty. In the lower panel a larger penalty is applied to the same data. In both cases the baseline is modelled using 200 B-splines.

diffraction scans. There a baseline plus noise of constant variance is not appropriate. Instead, one has to introduce a curve that gives the expected value of the baseline counts, with variance equal to the expected value.

In the example using a MALDI-TOF mass spectrum, we saw the error of the baseline changing over the range of m/z values. In this case it turned out not to be problematic. Nevertheless, incorporating a flexible variance estimator is a possible improvement of the method, but needs further investigation.

Matlab code to estimate the baseline using the method discussed in this paper is available upon request. In the future we will also release an R-package.

In this chapter we focused on baseline estimation in one-dimensional data. The next chapter presents the straightforward extension to two-dimensional data, using tensor products of P-splines (Eilers et al., 2006).

Mixture models for baseline estimation in two dimensions 3

Abstract. Baseline correction and artefact removal are important preprocessing steps in analytical chemistry. We propose a correction algorithm using a mixture model in combination with penalized regression. The model is an extension of the algorithm introduced in the previous chapter, for baseline estimation in the case of one-dimensional data. The data are modelled as a smooth surface using tensor product P-splines. The weights of the P-splines regression model are computed from a mixture model where a datapoint is either allocated to the noise around the baseline, or to the artefact component. The method is broadly applicable for anisotropic smoothing of two-way data such as two-dimensional gel electrophoresis and two-dimensional chromatography data. We focus here on the application of the approach in femtosecond time-resolved spectroscopy, to eliminate strong artefact signals from the solvent.

3.1 Introduction

Chemical and physical interferences resulting in an unwanted baseline or artefacts, is a relevant issue in analytical spectroscopy (Rinnan et al., 2005; Dai and Eads, 2010; Bowie et al., 2006; Rinnan et al., 2009; JiJi and Booksh, 2000). Many pre-processing methods have been proposed for baseline removal. These methods can be divided into several categories, including band-pass filtering (Mosierboss et al., 1995; Alsberg et al., 1997; Galloway et al., 2009), baseline modeling with a function (polynomial, spline, etc.) (Gans and Gill, 1984; Eilers and Marx, 1996; Lieber and Mahadevan-Jansen, 2003; Vickers et al., 2001), data smoothing Eilers (2003) and statistical methods (Phillips and Hamilton, 1996; Ruckstuhl et al., 2001). Generally most of these methods require the tuning of some parameters, such as the weights in weighted least-squares procedures. The use of mixture models for

This chapter is based on: de Rooi J.J., Devos O., Sliwa M., Ruckebusch C. and Eilers P.H.C. (2013). Mixture models for two-dimensional baseline correction, applied to artefact elimination in time-resolved spectroscopy. *Analytica Chimica Acta*, 771:7-13.

baseline estimation based on P-spline regression was reported recently by de Rooi and Eilers (2012).

Mixture models are statistical models which assume that the probability density function consists of different components; they are popular for clustering and classification purposes (Jacques et al., 2010). For baseline modelling, the data can be decomposed in a two-component mixture: peaks and a smooth baseline plus noise. The proportions of both components are unknown and have to be estimated. One of the main advantages of the procedure is that the weights that are used in the penalized regression procedure to model the data with P-splines are computed directly from the statistical model.Very good results were obtained when dealing with one-dimensional data as, e.g., Raman spectra (de Rooi and Eilers, 2012). In this chapter, we illustrate how this approach can be usefully extended to two-way data by using two-dimensional tensor products of P-splines. We focus on time-resolved spectroscopy data, but other types of applications are also possible and are discussed in the last section.

Femtosecond transient absorption spectroscopy is an analytical technique to investigate short-lived chemical intermediates and transient states created during ultrafast chemical reactions, such as excited-states intramolecular proton transfer (Sliwa et al., 2010). Getting information about unknown overlapping transient species often requires multivariate analysis of the series of two-way time-resolved spectra. However, the use of ultrafast optical laser pulses, to trigger the photoreaction, results in the observation of coherent optical processes. An example is stimulated Raman amplification scattering signals of the solvent which show very specific spectro-kinetic features. These signals are observed during a few hundreds of femtoseconds, as their kinetic behavior is similar to the instantaneous response function, and their typical spectral signature consists of a series of intense and sharp negative signals. Regarding data analysis, these signals are artefacts, distorting broad UV-visible transient absorption spectra. In addition, these contributions alter the ideal bilinear low-rank data structure of the spectrokinetic data as the bilinear structure is usually required to apply classical chemometric methods such as multivariate curve resolution (MCR) (Aloise et al., 2008; Mouton et al., 2010; Ruckebusch et al., 2012). As a result, data decomposition may be biased.

Several ways of handling coherent spectrokinetic artefacts have been proposed in the literature. Due to the nature of the signals considered, simple solvent subtraction turns out to be very challenging in practical situations (Lorenc et al., 2002). Ernsting et al. (2001) proposed the introduction of additional contributions in the data decomposition to deal with coherent low-frequency oscillations. Multiset MCR analysis was also shown to be an alternative to unravel the information related to the photochemical processes from the perturbation of the unwanted solvent signals (Ruckebusch et al., 2009). Another alternative to remove artefacts from femtosecond transient spectra is provided by pre-processing techniques such as smoothing procedures (Devos et al., 2011). The aim is to remove narrow artefact peaks from smooth data, a problem which can be considered the 'mirror image' of a classical baseline correction problem. In Devos et al. (2011) preliminary results were obtained using asymmetric least squares smoothing as proposed by Eilers and Boelens (2005). However, this technique suffers from a few drawbacks. In particular, spectra have to be corrected sequentially, one by one, without considering the strong correlation between successive spectra in time-resolved spectroscopy. Additionally, two parameters need to be tuned for each spectrum, one for asymmetry and the other for smoothing, which can be problematic for series of spectra.

The method we propose here relies on penalized regression with P-splines to estimate a smooth two-dimensional baseline. Full series of time-resolved spectra can be corrected in one run, as for images, and advantage can be taken from the fact that the correlation between successive spectra is not ignored. In addition, the x and y direction can be tuned independently, allowing anisotropic smoothing with tensor products of P-splines. Contrary to more classical weighted least-squares procedures, only the penalty parameters require optimization.

The next section starts with introducing the method for the one-dimensional case, and subsequently extend it to two-way data. In the third section, simulated and experimental data are introduced. The results obtained after analysing these data are presented in section four, and we close with some conclusions in section five.

3.2 Theory

The baseline is modelled as a smooth surface and is estimated using P-splines. We assume a two-component probability density for the deviations from the baseline: one for the noise around the baseline, and the second for the peaks (or artefacts). The mixture components are estimated using an expectation-maximization (EM) algorithm (McLachlan and Peel, 2000). Below, we first discuss the theory for the case of one-dimensional data, as presented in de Rooi and Eilers (2012). In the next step the model is extended for two-dimensional data, the focus of the current work.

Baseline estimation in 1d

A B-spline is a smooth bell-shaped curve, constructed from $q + 1$ polynomial pieces of degree q. Using a set of equally spaced and identically shaped B-splines we can estimate a smooth trend to fit the data. Let y be the signal with $i = 1, ..., n$ observations. The matrix B is the spline basis, consisting of n rows, and K columns. Figure 3.1 shows a perspective plot of a basis with fifteen cubic B-splines.

The model for the baseline is,

$$\mu = B\alpha, \tag{3.1}$$

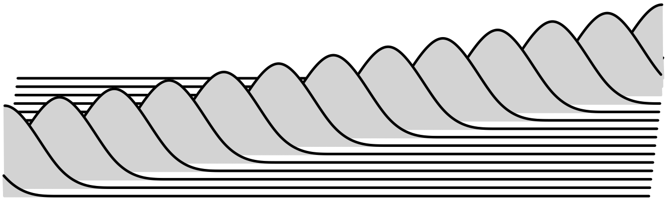


Figure 3.1: The basis B, here consisting of fifteen splines. Notice that for the presentation the matrix is rotated about 90° counter clockwise.

where α is the vector of spline coefficients. It can be estimated by linear regression, minimizing

$$S = \sum_{i=1}^{n}[y_i - \sum_{k=1}^{K} B_j(x_i)\alpha_k]^2 = ||y - B\alpha||^2, \tag{3.2}$$

giving the explicit solution

$$B'B\hat{\alpha} = B'y \quad \text{or} \quad \hat{\alpha} = (B'B)^{-1}B'y. \tag{3.3}$$

There are a number of recipes to determine the number of splines and the placement of knots. Here we follow the P-spline approach of Eilers and Marx (1996). They avoid the knot selection problem by using a generous number of B-splines and equally spaced knots. Smoothness is tuned by penalizing the differences of adjacent B-spline coefficients, based on the idea of a discrete roughness penalty introduced by Whittaker (1923), more recently discussed in Eilers (2003). Including the difference penalty, the objective functions can be written as

$$S = \sum_{i=1}^{n}[y_i - \sum_{k=1}^{K} B_k(x_i)\alpha_k]^2 + \lambda \sum_{k=1}^{K}(\Delta^d \alpha_k)^2 = ||y - B\alpha||^2 + \lambda||D\alpha||^2, \tag{3.4}$$

where Δ^d is the differencing operator of order d, as in de Rooi and Eilers (2012) we rely on third order differences. For seven B-splines this gives us the following matrix D:

$$D = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 & 0 \\ 0 & 0 & 0 & -1 & 3 & -3 & 1 \end{bmatrix}. \tag{3.5}$$

The weight of the penalty, and thus the smoothness of the curve, is regulated by the tuning parameter λ. Including the penalty, equation (3.3) becomes

$$\hat{\alpha} = (B'B + \lambda D'D)^{-1}B'y. \tag{3.6}$$

Instead of drawing a curve through the data cloud, the goal is to estimate a baseline μ. All observations in the signal are part of either the peaks, or the baseline including noise.

If we know to which component an observation belongs, we can include this information in the objective function using weights w. However, for observed data it is unknown to which component an observation belongs. We solve this by assuming the data are coming from a mixture of two distributions, and estimate all parameters, including w, using an EM algorithm (McLachlan and Peel, 2000). Weight parameter w is now defined as the probability to belong to the baseline component. When w_i is close to one, the data point is assigned to the baseline. We minimize the following:

$$S^* = \sum_{i=1}^{n} w_i [y_i - \sum_{k=1}^{K} B_k(x_i)\alpha_k]^2 + \lambda \sum_{k=1}^{K} (\Delta^d \alpha_k)^2 = (y - B\alpha)' W (y - B\alpha) + \lambda ||D\alpha||^2, \quad (3.7)$$

with $W = \text{diag}(w)$. The solution is:

$$\hat{\alpha} = (B'WB + \lambda D_d' D_d)^{-1} B'Wy. \quad (3.8)$$

We assume $y = \mu + v$, with μ the baseline and v a mixture drawn from the density $f(v)$:

$$f(v) = \pi g(v|0, \sigma) + (1 - \pi) h(v). \quad (3.9)$$

Here π is the unknown mixing ratio, and g a normal distribution with unknown σ representing the noise surrounding the baseline μ. The peaks or artefacts are part of the second component which is approximated by a uniform distribution $h(v)$. The weight or posterior probability that observation y_i comes from the baseline component $g(v|0, \sigma)$ is estimated:

$$w_i = \frac{\pi g(v_i|0, \sigma)}{\pi g(v_i|0, \sigma) + (1 - \pi) h(v_i)}. \quad (3.10)$$

In turn we estimate w given the parameter values in (3.10) (the *expectation* step) and the values for π, μ, σ and $h(.)$ given w (the *maximization* step). This routine is repeated until convergence. Proper starting values can be obtained by first estimating a simpler model, like using asymmetrical least squares (Eilers and Boelens, 2005).

Baseline estimation in 2d

In two dimensions the observation forms a matrix Y with $i = 1, ..., n$ rows and $j = 1, ..., m$ columns. The baseline is a similar matrix representing a smooth surface M and is estimated with tensor products of P-splines (Eilers et al., 2006), naturally following from the one-dimensional splines discussed in the previous section. The tensor products are formed from two bases: B ($n \times K$ matrix) in the row direction and $\breve{B}$ ($m \times L$ matrix) in the column direction, an impression of cubic B-spline tensor products is given in Figure 3.2. Notice that it is a sparse representation, a full basis would make the image too dense and hinder interpretation.

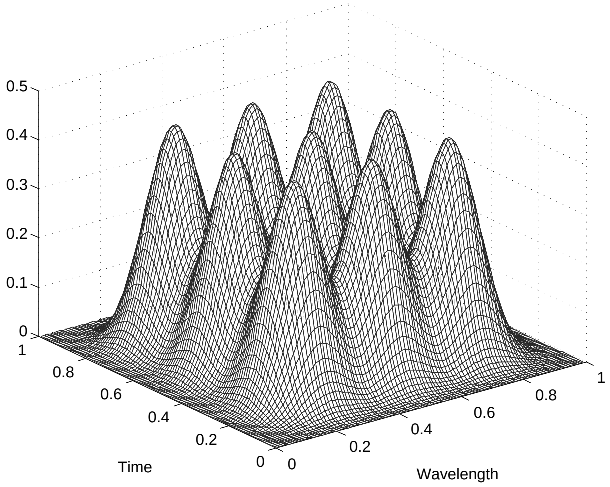


Figure 3.2: A sparse basis of cubic B-spline tensor products.

The fitted value for each point of the surface is expressed as

$$\mu_{ij} = \sum_{k=1}^{K} \sum_{l=1}^{L} \alpha_{kl} b_{ik} \check{b}_{jl}, \tag{3.11}$$

where α_{kl} is the k, l-th element of the matrix of coefficients A ($K \times L$ matrix). The expression (3.11) can be reformulated in matrix form:

$$M = BA\check{B}'. \tag{3.12}$$

Similar as in the one dimensional case, we introduce a penalty on the differences:

$$Pen = \lambda ||DA||_F + \check{\lambda} ||\check{D}' A||_F, \tag{3.13}$$

with $||.||_F$, the Frobenius norm, the sum of the squares of all elements. The tuning parameters for both directions λ and $\check{\lambda}$ can be tuned independently, allowing anisotropic smoothing.

The classical solution to estimate the model is to vectorize the data matrix: $y = \text{vec}(Y)$, the coefficient matrix $\alpha = \text{vec}(A)$, the smooth surface $\mu = \text{vec}(M)$ and the matrix with weights $w = \text{vec}(W)$. The penalty is written

$$P = \lambda(I_L \otimes D_d' D_d) + \check{\lambda}(\check{D}'_d \check{D}_d \otimes \check{I}_K), \tag{3.14}$$

with $\otimes$ being the Kronecker product and I the identity matrix. After vectorization, we are back at the weighted least squares as in (3.8), including the penalty P the equations become

$$((\check{B} \otimes B)' W^* (\check{B} \otimes B) + P)\hat{\alpha} = (Q + P)\hat{\alpha} = (\check{B} \otimes B)' W^* y, \tag{3.15}$$

with $W^* = \text{diag}(w)$.

Estimating the coefficients in A using vectorization and Kronecker products is simple but also very inefficient, especially when Y is large. If $\breve{B}$ is of dimensions $m \times L$ and B is $n \times K$ than $\breve{B} \otimes B$ results in an $mn \times KL$ matrix. Instead we rely on an algorithm introduced by (Eilers et al., 2006), later coined GLAM (generalized linear array model) (Currie et al., 2006). This algorithm is much more efficient, leading to large savings in time and memory use. The elements in Q in (3.15) are calculated in an other product,

$$G^* = (B\Box B)' W^* (\breve{B}\Box\breve{B}), \tag{3.16}$$

a matrix with dimensions $K^2 \times L^2$, and $B\Box B = (e'_K \otimes B) \odot (B \otimes e'_K)$ the row-wise Kronecker product of B with itself, and $\breve{B}\Box\breve{B} = (e'_L \otimes \breve{B}) \odot (\breve{B} \otimes e'_L)$. Here $\odot$ stands for element by element multiplication, and e_K (e_L) being a vector of ones of length K (L). The system in (3.15) can now be rewritten as

$$(G + P)\hat{\alpha} = r. \tag{3.17}$$

To get r we first create the matrix $R = B'(W \odot Y)\breve{B}$, and subsequently vectorize this column-wise. To obtain G from the matrix G^*, three steps are needed. First, G^* is reordered in a four dimensional $K \times K \times L \times L$ array. Second, the dimensions are reordered from 1,2,3,4 to 1,3,2,4. Third, the array is transformed into G, a $KL \times KL$ matrix. The estimated vector $\hat{\alpha}$ consist the columns of the matrix $\hat{A}$.

The mixture model looks similar as in (3.9) only now referring to the two dimensions of the surface, equation (3.10) is also modified:

$$w_{ij} = \frac{\pi g(v_{ij}|0,\sigma)}{\pi g(v_{ij}|0,\sigma) + (1-\pi)h(v_{ij})}, \tag{3.18}$$

the weights are iteratively updated using the discussed EM algorithm.

Optimal values for the penalty parameters are often determined using cross validation, the Akaike information criteria (AIC) (Akaike, 1974), or the Bayesian information criteria (BIC) (Schwarz, 1978). However, these approaches are not robust to highly correlated noise, often leading to poor performance in the case of artefact removal. The choice of the tuning parameters λ and $\breve{\lambda}$ is usually performed by visual inspection of the preprocessed data, as for the one-dimensional approach. In some applications, detailed information about the characteristics of the experimental data is available and, as a result, can be accurately resembled by simulated data. Here we illustrate this approach for the elimination of artefacts in time-resolved spectroscopy data, where the spectro-kinetic features of the signal are known. The simulated data will enable estimation of proper parameter values, or serve as good initial estimates for further refinement. The use of simulated data is not a requirement for the method to work, but can be helpful for parameter optimization. For the simulated data, the performance of the method can be evaluated using the sum of the squared errors (SQE):

$$SQE = \sum_{i=1}^{n}\sum_{j=1}^{m}(f_{ij} - t_{ij})^2, \tag{3.19}$$

the smooth surface modelled, F, is compared to the initial simulated two-way data (before adding artefacts) T. The optimal set of parameter values is subsequently used for the modelling of the real data.

In the studied application we use additional knowledge in the estimation procedure. Most of the time, the position of artefacts is partially known. In transient absorption spectroscopy stimulated Raman amplification signals can only be observed in a time range matching the time resolution of the experiment and at some spectral positions depending on the solvent used and the excitation wavelength. This information can be translated into a mask matrix, which has the same dimension as the data matrix. The mask matrix contains ones, in the region where artefacts are known to occur, and zeros elsewhere. During the mixture model optimization procedure, the weights of only data points corresponding to mask matrix values of one will be updated. The advantage of using a mask matrix is that the risk of deformation in areas where no artefact is present is limited, while all data is used for estimating the noise around the baseline. For many applications there will be no information regarding the location of artefacts and all weights are set to one.

3.3 Experimental section

Simulated data

Simulated data were constructed, mimicking the spectro-kinetic features of femtosecond time-resolved spectroscopy (see Figure S-1 in supplementary material). Kinetic profiles were chosen to describe a process involving three transient species following a first-order sequential model. The pure kinetic profiles were convoluted with an apparatus function assumed to be a normalized centered Gaussian function with a 50 fs standard deviation. A strong spectrokinetic artefact signal was added to the smooth two-way data, its dynamics following the one of the apparatus function. The transient spectrum corresponding to the artefact is composed of 2 negative Gaussian contributions (width 6.5 nm), centered at time zero at 450 nm and 465 nm, respectively. A blue shift evolving linearly with the time (-0.02 nm.fs-1) was applied to these Gaussian contributions to simulate some non-ideal behaviours usually observed in practice, when dealing with stimulated Raman amplification scattering signals. A homoscedastic noise corresponding to 1% of the maximum signal amplitude was added. The resulting data are shown in Figure 3.4a.

Experimental transient absorption spectra

In order to illustrate the method, two different UV-Vis transient absorption spectroscopy data sets were investigated in this study (see Figure S-2 in supplementary material). Both were obtained with the same experimental setting and similar experimental conditions, in particular, pump excitation wavelength was set around 390 nm (see Mouton et al., 2010, and references therein for experimental details). The first data set contains femtosecond transient absorption spectra of salicylidene aniline (from here called SA data) in acetonitrile. Measurements obtained in pure acetonitrile are also available. Stimulated Raman amplification scattering signals of acetonitrile consist of three strong negative artefact peaks at wavelength positions around 425, 435 and 450 nm, respectively. The artefact contributions are shifting toward lower wavelength with increasing delay time. It is clear that both artefact and transient absorption signals are appearing simultaneously at zero delay time and that these two signals are growing with the same dynamics. artefact signals vanish after a few hundreds of femtoseconds, in agreement with the time resolution (apparatus function) of the experiment ($\sim$150 fs) (Mouton et al., 2010). The second dataset (from now called SB data) corresponds to the femtosecond transient absorption spectra of 2-Pyridin-1-yl-1H-benzimidazole (Aloise et al., 2012). The same solvent (acetonitrile) was used as for the SA data. For this reason, comparable spectrokinetic artefact signature (three strong negative artefact contributions at roughly the same position) can be observed. However, solvent response was not measured independently for the SB data.

Software

Routines were written in MATLAB version 7.1.0 (The Mathworks, Natick, MA) and the R programming language (R Development Core Team, 2011), and are available upon request.

3.4 Results and discussion

Parameter settings and performance of the method.

Cubic B-splines were used in the two dimensions of the signal, the order of difference (d) was set to 3, as recommended for other applications (Devos et al., 2011). Eilers and Marx (1996) dictate the use of a generous number of B-splines, while the model complexity is tuned by the penalty. Fifty splines were considered in order to properly fit the data in the spectral dimension (351 wavelength values). In the time dimension, the profiles to model are often smoother, particularly at short time scale, as a result of the convolution with a relative large apparatus function. As 30 to 100 time points are usually available, the number of splines in this direction was fixed to 30. Using too few splines can be

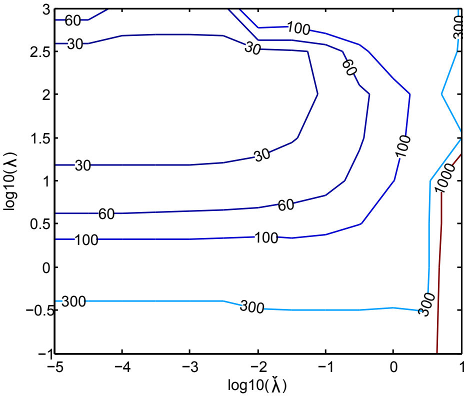


Figure 3.3: Grid search approach for the optimization of the regularization parameters λ and $\check{\lambda}$ by calculation of SQE. A 50 × 30 tensor product of cubic B-splines and third order differences was used.

problematic in this application due to the characteristics of the signal at short time delays (close to zero), where sampling steps are small and convolution with apparatus function is strong. As mentioned above, a large number of splines can be used as the smoothness constraint will penalize adjacent differences between spline coefficients.

For the estimation of the posterior probability (Eq. 3.18), the initial values for the standard deviation of the noise and the amplitude of the negative artefact signals are estimated from the data. For the simulated data, values of 0.1 and 5 were chosen to estimate first the densities $g(v)$ and $h(v)$, respectively. It should be noted that the standard deviation of noise will be updated during the procedure.

Important is the optimization of the tuning parameters λ and $\check{\lambda}$, for the spectral and time dimension, respectively. For underestimated λ values, some artefact contributions will remain, whereas deformation of the transient spectra will be observed for too large values. In this work, a grid search approach was applied, using the simulated data, for simultaneous optimization of λ and $\check{\lambda}$. The results obtained are presented in Figure 3.3. The sensitivity of the error response when varying the two parameters can be assessed visually from the figure. The best results (SQE ≤ 30) were obtained for values of $\lambda \approx 102$ and $\check{\lambda} \leq 0.01$ and the smallest SQE was found be equal to 28.2. This value corresponds to 100 and 0.01 for λ and $\check{\lambda}$, respectively. For the simulated data, the results are shown in Figure 3.4.

The noise standard deviation is estimated at 0.051 and the mixing ratio at 0.95. As observed in Figure 3.4b, the corrected data are smooth and artefact contributions were efficiently removed. This is confirmed by the observation of the removed spectrokinetic signals presented in Figure 3.4c. It can be noticed that in addition to the two shifting negative peaks properly extracted, noise is taken into account in the smoothing procedure.

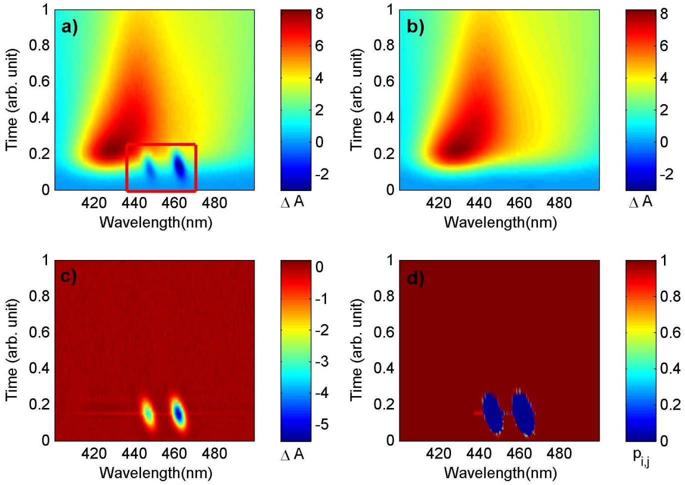


Figure 3.4: Raw a) and preprocessed simulated data b); the removed spectrokinetic signals and posterior probability values w_{ij} are presented in c) and d), respectively. Outside the zone marked by the rectangle (in red) the weights are forced to one.

Excellent agreement is observed between preprocessed data and smooth, artefact free, simulated data. Focusing on the values of the posterior probability w_{ij} obtained after convergence of the procedure and provided in Figure 3.4d, it is clearly observed that in the region where spectrokinetics artefacts are present, posterior probability values are close to zero. Elsewhere w_{ij} values are close to 1.

The benefits of the use of this smoothing procedure with respect to soft-modelling are discussed in the support information. It is shown from Singular Values Decomposition (SVD) results that the smoothing procedure enables restoring concordance between the rank observed for the simulated data and the one corresponding to the known simulated process scheme (see Figure S-3).

Application to real data

More challenging situations are provided by both experimental datasets. For these two real datasets we used parameter values, previously optimized on the simulated data. Only the initial values of the noise standard deviation and artefact amplitude have to be estimated for each dataset. As we will show in the following, when the real two-way data investigated show similar structure and behaviour, or smoothness, as the simulated ones, the consistency of the results is good. However, it was also checked that no clear improvement of the correction could be obtained after fine tuning of the regularization parameters.

For SA data, it turned out that direct solvent subtraction is inefficient to remove arte-

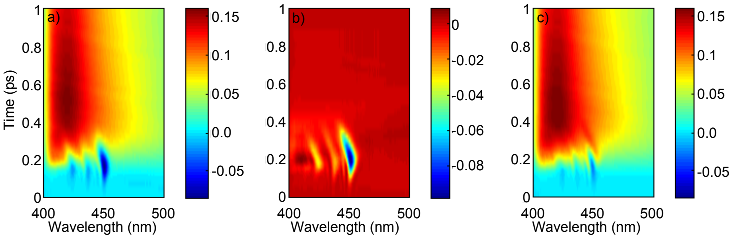


Figure 3.5: Transient absorption data observed for a) SA data b) acetonitrile solvent and c) results obtained after solvent subtraction.

facts, as shown Figure 3.5. The observed SA data is presented in Figure 3.5a, Figure 3.5b provides data obtained in pure solvent (acetonitrile) in the same experimental conditions. Comparing the features of artefact signals, it can be observed that not only the intensity but also the spectro-kinetic characteristics are not exactly similar for SA and solvent data. In order to get reproducible solvent signals, experiments in pure solvent and solution should be performed in perfectly identical optical conditions, which is hardly achievable as the absorption coefficient of the pure solvent changes with the pump energy. Figure 3.5c corresponds to the results obtained performing solvent subtraction. Obviously, an artefact pattern remains, emphasizing the need for alternative procedures to remove stimulated Raman amplification scattering signals in femtosecond transient absorption spectroscopy.

The results obtained after pre-processing the two-way data (Figure 3.6b) are very satisfying when compared to raw data (see Figure 3.6a). The removed artefact signals are shown in Figure 3.6c. From the posterior probabilities w_{ij} (Figure S-4), it is observed that after convergence the weights are close to 1, where no artefact is present and close to zero elsewhere. When looking at the weights magnitude, clearly three peaks are modelled in very good agreement with solvent data, as expected.

The SB data provides another example of the efficiency of the proposed approach. The situation is somewhat different, because the artefact superimpose on a part of the signal which is negative (Figure 3.6d). The corrected data and removed signals obtained with the mixture model are presented in Figure 3.6e and Figure 3.6f. As in the case of the SA dataset, three artefact peaks with important shift are extracted by pre-processing. Comparison of Figures 3.6c and 3.6f, shows good correspondence between the removed signals. This is in agreement with the fact that the origin of the spectrokinetic artefacts is the same (same solvent and similar excitation wavelength).

3.5 Conclusions

In this paper we presented a baseline correction algorithm for two-dimensional data. The method is a natural extension of earlier work, restricted to one-dimensional signals, by two

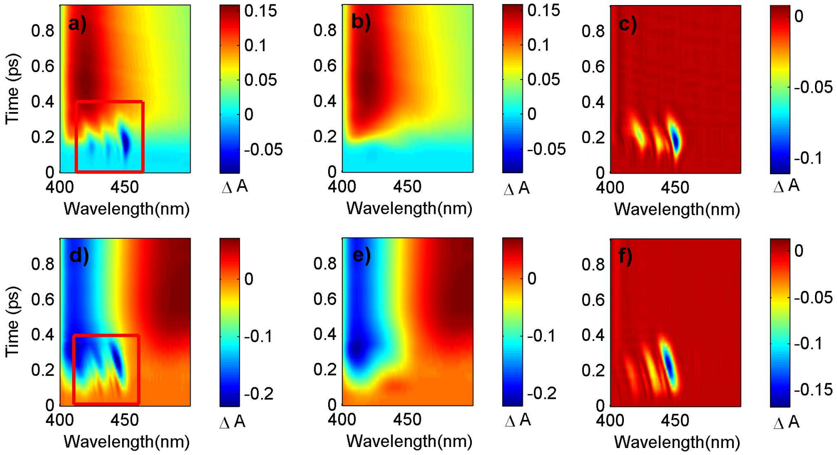


Figure 3.6: A) SA raw data and d) SB raw data; outside the red rectangle the weights are forced to one. b) and e) preprocessed time-resolved data using the mixture model. The removed signals are presented in c) and f).

of the authors. The performance of the method is demonstrated on femtosecond transient spectroscopy where the elimination of strong spectro-kinetic artefacts is considered.

The method uses a mixture model to distinguish the baseline (including noise) from the peaks or artefacts. Any data point y_{ij} thus gets a probability to be associated to peaks or to a contribution from the smooth surface, corresponding to the unknown proportion of the two-component mixture. The baseline surface is estimated using tensor product P-splines. This enables the method to take advantage of the correlation between successive spectra in series of time-resolved spectra.

In the demonstrated applications an optimal model was obtained using simulated data. If not available, or when it is impossible to simulate comparable data, the model can be tuned by the operator. Currently no automatic procedure for model selection is available. For future work it is valuable to investigate automated tuning of the model, using some type of information criterion. Problematic with these criteria is that they assume uncorrelated noise, which is not realistic in the present case. The method can be used in a wide range of applications. In this chapter we only showed negative artefacts, but performance is equally well in situations with positive peaks. Some further applications are background estimation for two-dimensional gel electrophoretic data and correcting microarray gene expression data. The GLAM framework allows fast computations on large matrices, and enables the extension to three dimensions, where we have two spatial directions as well as a time dimension.

Acknowledgement

Dr. Aloïse and Dr. Pawlowska are acknowledged for making the experimental data available.

Appendix

Singular value decomposition of the simulated dataset

The rank of the data is an important aspect in soft-modelling of the resolution of the time-resolved spectroscopy data, where one aims at recovering kinetics and pure spectra of the different species observed. The simulated data were obtained considering a sequential three-species first-order reactional scheme affected by an apparatus function in the time direction. Each of the three species was associated with a pure spectrum with quite large spectra signature and significant recovery in the wavelength range considered. The results of the singular value decomposition are provided in Figure S-6. Three series of values are reported corresponding respectively to the analysis of the two-way data affected by artefact contributions, artefact-free two-way data and preprocessed data. Focusing on the comparison between the two first series, it can be clearly observed the consequence of the addition of spectrokinetic artefact signals, which blur the ideal situation where three contributions are clearly detected. Comparing now the results for preprocessed data with the simulated smooth artefact-free two-way data, very good concordance is found for the three first singular values which indicates that the low rank bilinear data structure of the artefact-free data has been recovered.

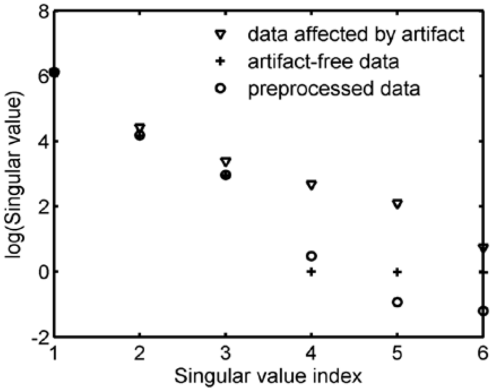


Figure A1: Singular values decomposition of the simulated data (two-way data affected by artefact contributions), of the corresponding artefact-free data and of the preprocessed data.

Datasets description

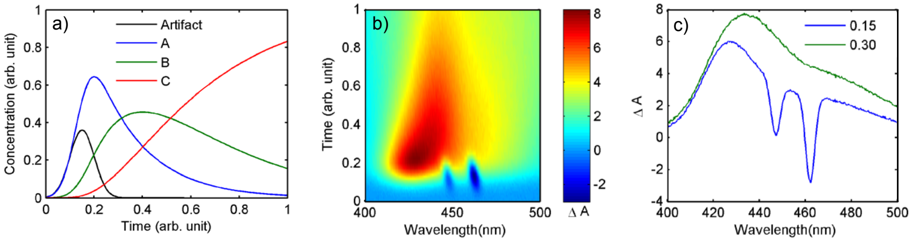


Figure A2: Simulated data.a) Kinetic Profile.b) Fifty transient spectra (time from 0 to1 in arbitrary unit) of 351 difference absorbance values in the range 400-500 nm.

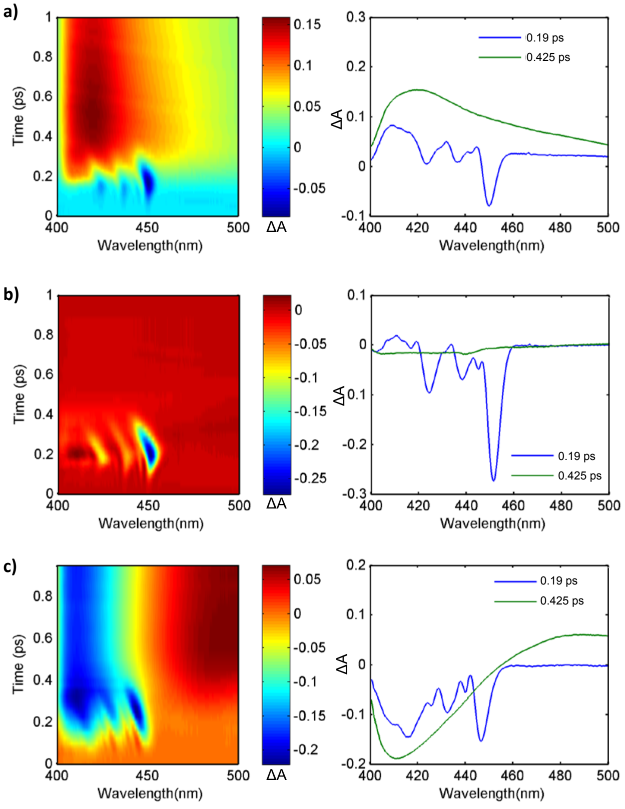


Figure A3: Transient absorption spectra in the visible region (400-500 nm) for a) SA data (50 spectra from 0 to 1 ps) b) Acetonitrile data (50 spectra from 0 to 1 ps) and c) SB data (29 spectra from 0 to 1 ps).

Weights optimized for the SA dataset

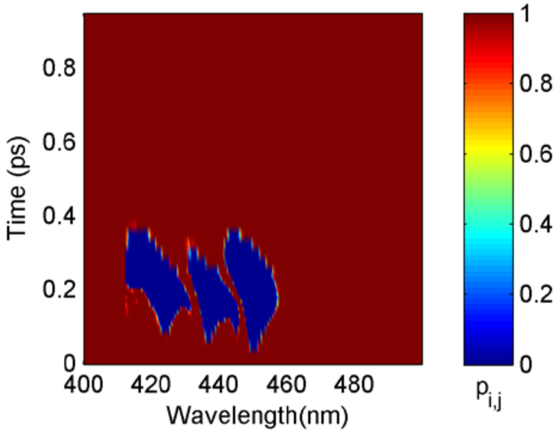


Figure A4: Posterior probability p_{ij} for the SA dataset.

Deconvolution of pulse trains 4

Abstract. The output of many instruments can be modelled as a convolution of an impulse response and a series of sharp spikes. Deconvolution considers the inverse problem: estimate the input spike train from an observed (noisy) output signal. We approach this task as a linear inverse problem, solved using penalized regression. We propose the use of an L_0 penalty and compare it with the more common L_2 and L_1 penalties. In all cases a simple and iterative weighted regression procedure can be used. The model is extended with a smooth component to handle drifting baselines. Application to three different datasets shows excellent results.

4.1 Introduction

Many instruments produce signals that consist of a series of pulses. Examples are electrophoretic DNA sequencers, chromatographs, and spectrometers. Some biological signals have the same characteristics; an example is hormone release in the human body. The pulses have (more or less) equal shapes but different heights, and they may overlap. These output signals are the convolution of a series of true (input) spikes or diracs and the impulse response function. The task is to deduce the heights and positions of the spikes from the output signal.

Essentially there are two ways to approach this issue. The first is to search for local maxima to find peak positions, followed by summarizing the signal in their neighbourhoods, to estimate peak heights. Examples of this approach are found in many places in the literature. We mention only a small selection. Yasui et al. (2003) search for zeros of the first derivative, while Mariscotti (1967) uses the second derivative. When computing derivatives, it is essential that proper noise filtering is first applied. Wavelets have been proposed as a tool for filtering by Coombes et al. (2005) in this setting, but other filters are also possible. Du et al. (2006) use a wavelet spectrum to locate peaks. It is also possible

This chapter is published as: de Rooi J. and Eilers P. (2011). Deconvolution of pulse trains with the L_0 penalty. *Analytica Chimica Acta*, 705(1-2):218-26.

to apply a discrete Markov chain as done by Silagadze (1996) and Morháč (2007), these approaches result in a probability distribution targeting the location of peaks.

The second approach is to model signals as a convolution of a series of sharp input spikes and a constant impulse response. The task then is to estimate the input from the observed output signal. This is the deconvolution problem that has been studied in many fields of science. It is a so-called inverse problem, and it is generally very badly conditioned, which means that small changes in the observed signal or the impulse response lead to large changes in the estimated input. Conversely, many very different inputs are compatible with the observed output.

To address the bad condition various deconvolution algorithms have been proposed. An early solution is the van Cittert algorithm (see e.g. Jansson, 1996) that was later improved in the form of the Gold algorithm (see e.g. Bandzǔch et al., 1997). Other iterative approaches are often based on the Richardson-Lucy algorithm or using the general class of expectation maximisation (EM) algorithms (see e.g. Vardi and Lee, 1993; Li and Speed, 2004). The EM algorithm iteratively redistributes the observed output, proportionally to the current estimate of the input. Averaging gives an improved estimate of the input, to be used in the next iteration.

The class of deconvolution algorithms also contains a branch of boosting algorithms. Cardot et al. (2004) propose to use boosting to find an optimal set of input spikes. As a first step the locations of the peaks are estimated and subsequently renewed in an updating step, in the third stage peaks that are too close to each other are merged into one. Recently Morháč and Matoušek (2011) proposed a boosted version of the Gold and Richardson-Lucy algorithms.

A general approach to ill-conditioned problems is the use of regularization: some form of penalty is imposed on the parameters of the model. We already referred to Li and Speed (2004). A familiar example in the chemometric literature is ridge regression (Hoerl and Kennard, 1970), where the penalty is on the sum of the squares of the regression coefficients. This is called the L_2 norm; generally the sum of absolute values (of the elements of a vector) to the power p is called the L_p norm. In recent years the Lasso (Tibshirani, 1996), a penalty based on the L_1 norm, the sum of absolute values, has become popular in many applications. References can be found in the next section.

Penalties with a norm based on $p < 1$ have received little attention. A main theoretical obstacle has been the fact that they lead to a non-convex optimization problem, in contrast to penalties with $p \geq 1$. Hence one cannot be sure of having found a global minimum. Another drawback is the lack of good practical algorithms. In this chapter we propose regularized deconvolution using the L_0 penalty, and we show very good results using an algorithm based on repeated weighted regression. Apparently, in the limited context of pulse train deconvolution, a non-convex objective function is not a real problem.

In the next section we introduce the deconvolution framework, and we show the effects of regularization with different norms. There we assume that the impulse response is known. In practice only an approximation will be available, so we also consider "blind deconvolution": the estimation of both input and impulse response from one signal. Drifting baselines are quite common; we present two ways to handle them.

In section 3 we present three applications to experimental data. Two of them are instrumental (electrophoretic DNA sequencing and gas chromatography), the third is a series of high-frequency measurements of concentrations of luteinizing hormone in human blood, which show strong pulsative behaviour.

In the final section we discuss possible extensions and refinements.

4.2 The model

Convolution and deconvolution

Consider a (causal) discrete linear system with an input signal x, and an impulse response (or spread function) c, which incorporates blurring and filtering effects. Actually observed is the output signal y of length m, plus noise e. Using the superposition principle y can be described by

$$y_i = \sum_{j=0}^{n} c_j x_{i-j} + e_i. \tag{4.1}$$

Details can be found in many books on linear system theory; see e.g. Brown (2006). Interpreting i as indexing time, this shows that y_i is a weighted sum of the present and all previous observations of x. The weights are given by the impulse response. In practice the length n of the spread function c, is small compared to the length of x and y.

In matrix-vector form we can write $y = Sx + e$. The convolution matrix S consists of $m + (n-1)$ rows and m columns. Row i of S contains c reversed and right-shifted by $i-1$. This means that the columns of S are identical but shifted, which results in a band matrix. An example of a convolution matrix is presented in Figure 5.1.

Given S and x, computation of (a noise-free) y is straightforward, but here we consider the inverse problem: y and S are given and we have to find x. The simple approach is to use least squares: minimize $||y - Sx||^2$, the objective function of linear regression. The closed form solution is:

$$x = (S'S)^{-1}S'y. \tag{4.2}$$

Although the problem now looks like a simple case of regression, this does not help us much. By construction, the columns of S are strongly correlated, leading to a a severely ill-conditioned problem, making the results of straightforward regression useless. We will not go into the theoretical details; one example should suffice. The middle panel of Figure 4.2 illustrates the problem for a small simulated dataset. Even though the amount of noise is relatively small, and the fit of the model to the data is very good, the estimated

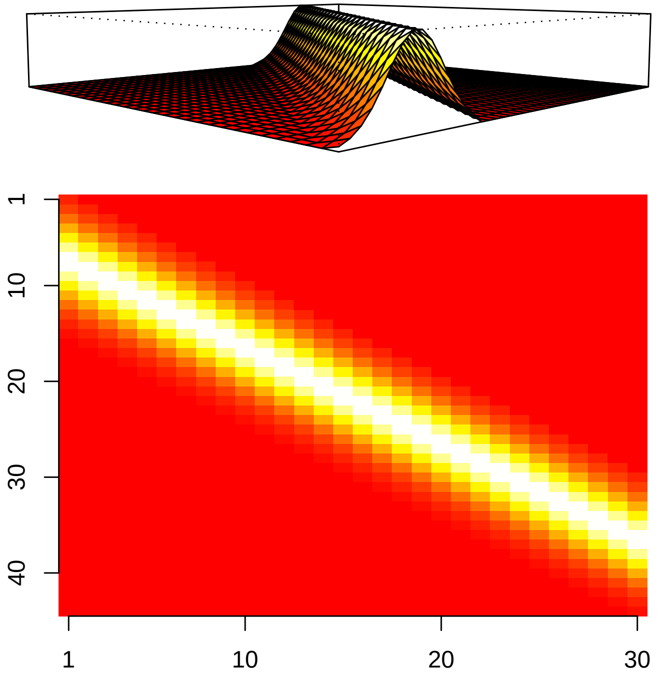


Figure 4.1: A convolution matrix S with $m + (n - 1)$ rows and m columns, in 3-d in the upper panel and in the lower panel in 2-d.

input is useless. Notice the size of the estimated input signal: it is a million times larger than the output. Also the elements of $\hat{x}$ systematically alternate signs.

Penalized regression

A general solution for ill-posed problems is regularization. A classic procedure in this vein is Tikhonov regularization, also known as ridge regression (Hoerl and Kennard, 1970). One adds a quadratic penalty $Q = \kappa \sum_j (x_j^2) = \kappa ||x||^2$ to the usual ordinary least squares objective function. Technically this works well: the calculation is stable and the fit to the data generally is quite good. The lower panel of Figure 4.2 shows the results when applying the L_2 penalty on the simulated data. In contrast to the unrestricted regression model, we observe estimated pulses that are relatively close to the truth.

The ridge penalty however does not result in a sparse solution: $\hat{x}$ is not very useful, because it does not look like the series of isolated sharp spikes that we expect. A penalty that applies shrinkage and creates sparseness is the L_1 penalty: $Q = \kappa \Sigma_j |x_j|$. The L_1 penalty was independently introduced by Chen et al. (1998) as basis pursuit and as the least absolute shrinkage estimator (Lasso) by Tibshirani (1996). The L_1 penalty has be-

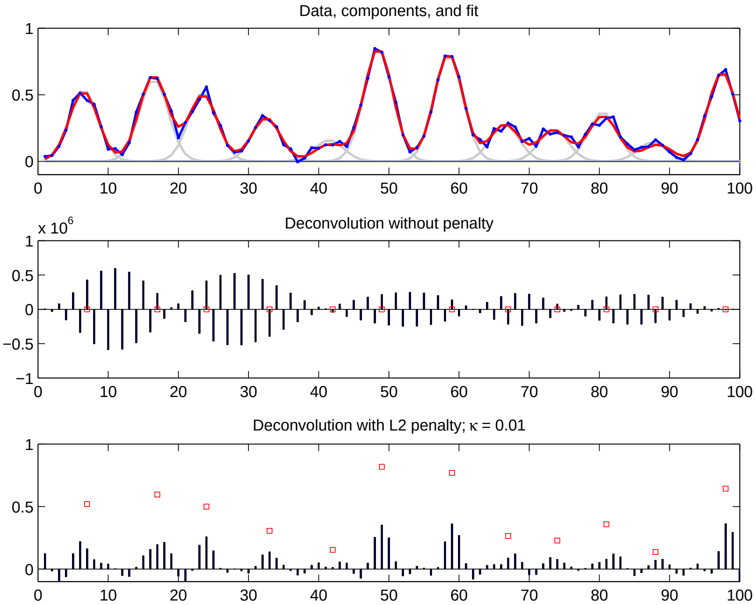


Figure 4.2: Simulated data. Top panel: data (full blue line with dots), fit (full red line) and individual pulses (thick gray lines). Middle panel: input as estimated without a penalty. Bottom panel: input as estimated with an L_2 penalty. The small squares give the positions and the heights of the nonzero elements of the input used for the simulation.

come very popular in various fields of statistics, a few examples are graphical modelling (Meinshausen and Buhlmann, 2006), time to event data (Goeman, 2009) and image denoising (Wang and Zhu, 2010). Applying the L_1 penalty gives a dramatic improvement, the positions and magnitudes of the spikes are often close to their simulated values, as Figure 4.3 shows. Although the L_1 penalty improves the results considerably, we do not find isolated single peaks and the peak height is often underestimated. As a third alternative we propose the L_0 penalty: the sum of nonzero elements of x, scaled by κ. In this way heavy shrinkage of the main peaks is prevented, while minor adjacent peaks are set to zero. The L_0 penalty can be written as: $Q = \kappa \sum_j I(x_j \neq 0) \equiv \sum_j \kappa |x_j|^0$. Results on the same simulated data using the L_0 penalty are shown in the lower panel of Figure 4.3. In most cases we do find isolated peaks that are strongly matching the position and height of the simulated spikes. The small mismatches with respect to position and height between truth and estimated are due to the amount of noise added to the system.

All three discussed penalties can be considered as members of a family of L_p penalties, with $0 \leq p \leq 2$. If $p = 2$, the ridge penalty, there is a closed form solution for x and looks:

$$x = (S'S + \kappa W)^{-1} S'y, \tag{4.3}$$

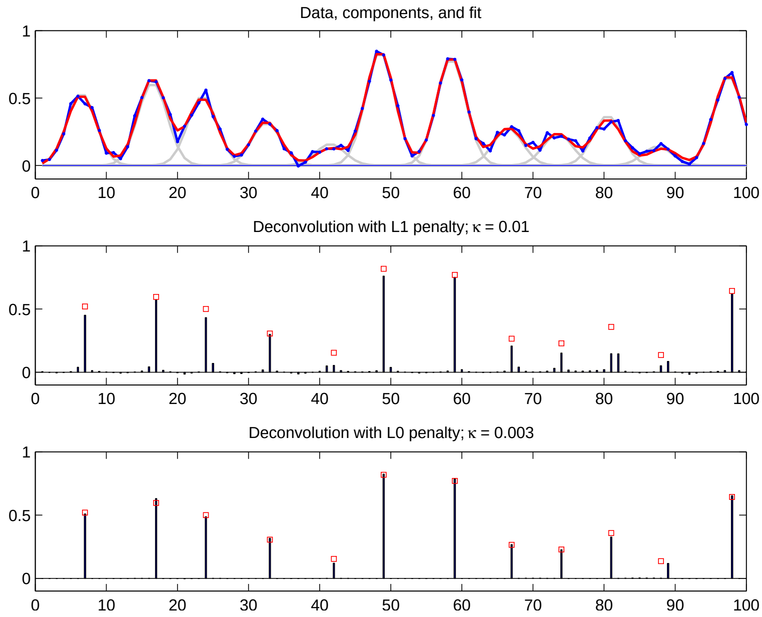


Figure 4.3: Simulated data. Top panel: data (full blue line with dots), fit (full red line) and individual pulses (thick gray lines). Middle panel: input as estimated with an L_1 penalty. Bottom panel: input as estimated with an L_0 penalty. The small squares give the positions and the heights of the nonzero elements of the input used for the simulation.

with $W = I$, an identity matrix and κ the tuning parameter for the penalty. When $1 \leq p < 2$ the objective function is still convex and several optimization methods are available in the literature. When $p = 0$ the objective function is non-convex, making the optimization problem harder. To optimize the objective function of the L_1 and L_0 penalty we use an iterative procedure, previously presented in Osborne et al. (2000) to optimize the L_1 penalty. It generally holds that:

$$|x_j|^p = x_j^2/(\tilde{x}_j^2 + \beta^2)^{(2-p)/2} = w_j x_j^2, \tag{4.4}$$

a weighted square, if $\tilde{x}_j = x_j$ and constant β equals zero. The idea is to iterate with this formula, using for $\tilde{x}_j$ the current approximation to the solution x_j. The constant β is added to reduce the number of iterations: β is a small number, say 1000 times smaller than the expected maximum of x. For both penalties the basis is equation (5.7), with for each a specific matrix W. In the case of the L_1 penalty W is a diagonal matrix with $w_{jj} = 1/\sqrt{\tilde{x}^2 + \beta^2}$ and for the L_0 penalty: $w_{jj} = 1/(\tilde{x}_j^2 + \beta^2)$.

In Figure 4.4 we see the penalty function plotted for all three norms. Compared to the other two norms we see that the L_2 norm imposes a large penalty on coefficients far away from zero. This is the shrinking effect. Although the penalty approaches zero when the

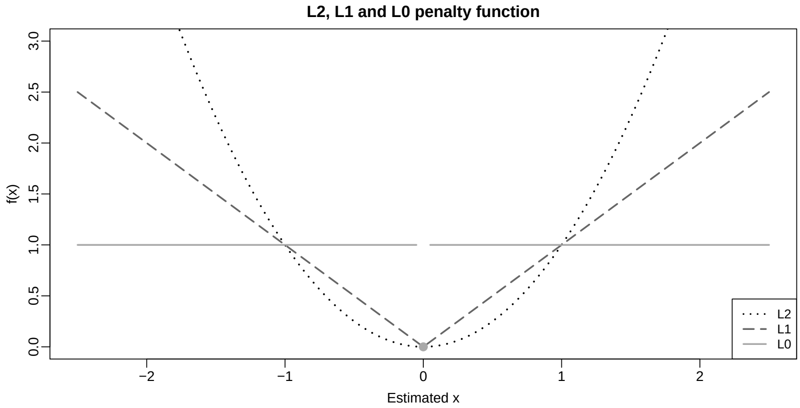


Figure 4.4: The shape of the penalty function for the L_2, L_1 and the L_0 penalty. The degree of shrinkage depends on the shape of the penalty function, the size of the tuning parameter and magnitude of the estimates.

coefficients are small it will never be zero. The L_1 penalty function is linear and compared to L_2 penalty function puts a bigger penalty on coefficients near zero and can also set coefficients to zero. Using the L_0 norm, the penalty is one if the estimate is nonzero and zero when a coefficient is zero.

Blind deconvolution

In the previous paragraph we assumed that the impulse response is available, however in practice this is almost never the case. Fortunately, we do get warnings when we use a wrong impulse response, because the estimated input signal compensates the error and we will not get the desired train of isolated spikes. If we use an impulse response that is chosen too broad this leads to additional negative spikes in the input between the real one, as shown in Figure 4.5. When the impulse response is chosen too narrow, we will see close double impulses, as Figure 4.6 shows.

If we know x, estimating the impulse response c leads to an other regression problem. So, by alternating between estimating x and c we get an algorithm for blind deconvolution. Using equation 4.2 we estimate x, to estimate matrix S we reverse the equation and treat x as given. If X is a matrix with, in its columns, shifted copies of $\hat{x}$, $\hat{c}$ is found from $X'X\hat{c} = X'y$. In general this is a well conditioned regression problem since the length of c is generally much smaller than that of x and y. A reasonable starting estimate for the impulse response c is needed, but this is not very critical and one can pick one isolated peak as a template.

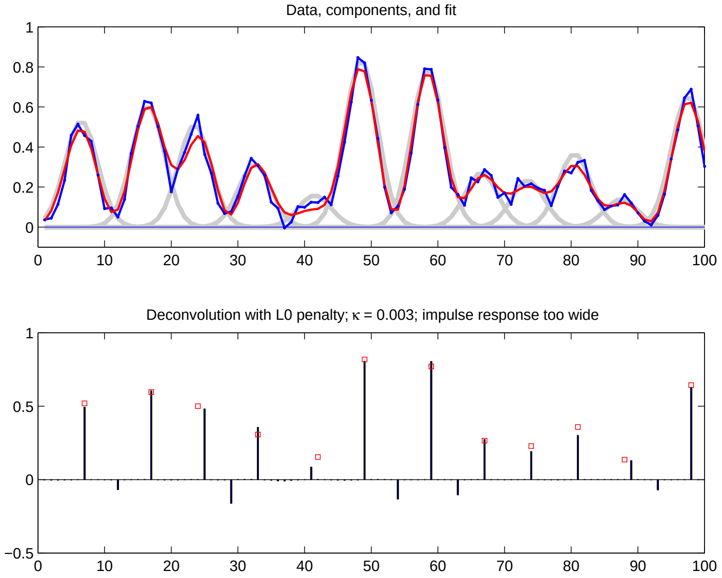


Figure 4.5: Simulated data. Upper panel: data (full blue line with dots), fit (full red line) and individual pulses (thick gray lines). Lower panel: deconvolution with an impulse response that is 20% too wide.

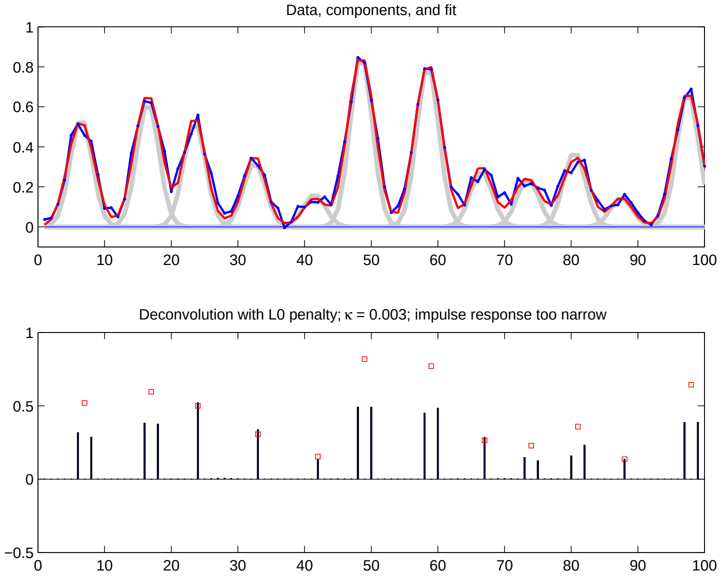


Figure 4.6: Simulated data. Upper panel: data (full blue line with dots), fit (full red line) and individual pulses (thick gray lines). Lower panel: deconvolution with an impulse response that is 20% too narrow.

Baseline correction

Many spectra and traces show a drifting baseline, which should be corrected to achieve proper results. We approach this problem in two different ways. One is to estimate a baseline from the observed data and subtract it. The other is to introduce it in the model and to estimate both input and baseline simultaneously.

In the literature we can find various ways of estimating a baseline component in traces and spectra. Examples are estimating local minima and fitting the baseline with interpolation (Coombes et al., 2003), peak clipping algorithms (Ryan et al., 1988; Morháč, 2009), or the application of robust local regression (Ruckstuhl et al., 2001). We use asymmetric least squares smoothing (Eilers, 2004). The baseline z is estimated by minimizing the following penalized least squares function:

$$R = \sum_i v_i(y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2. \tag{4.5}$$

The first term estimates the fit to the data, the second part $\lambda \sum_i (\Delta^2 z_i)^2$ is a roughness penalty based on the second order differences. Δ is the difference operator: $\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}$, λ is the tuning parameter. The weights v are asymmetric: when $y_i > z_i$, $v_i = \alpha$, a small number between 0 and 1. Otherwise $v_i = 1 - \alpha$. The weights are computed iteratively, starting from $v_i \equiv 0.5$. Two steps are repeated: smoothing with the current weights and updating the weights. Convergence is generally obtained after a handful of iterations. The parameter α determines the asymmetry; usually a value between 0.001 and 0.01 works well. This assumes positive peaks. If the peaks are negative, α should be replaced by $1 - \alpha$.

Increasing λ will result in a smoother trend, currently λ is tuned manually (as is α). The system of equations to be solved is: $z = (V + \lambda D'D)^{-1} V y$. Here, V is a diagonal matrix with v on the diagonal. $Dz = \Delta^2 z_i$ the (second order) difference matrix; an example with $m = 6$ looks like:

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}. \tag{4.6}$$

Alternatively we can add the baseline to the model, as follows:

$$y = Sx + z + e. \tag{4.7}$$

The roughness penalty $\lambda ||Dz||$ is applied as before, and the L_0 penalty is applied to x. In the iterative least squares approach the penalty can be written as $||Wx||$, with the current weights in W. This leads to the following system of equations:

$$\begin{bmatrix} I + \lambda D'D & S \\ S' & S'S + \kappa W \end{bmatrix} \times \begin{bmatrix} z \\ x \end{bmatrix} = \begin{bmatrix} y \\ S'y \end{bmatrix}. \tag{4.8}$$

In our experience, it is generally not possible to work with this system directly, unless the baseline is small. The weights in the matrix W have to be close enough to their right values to make this work.

The following procedure gave good results. First deconvolution is performed, with the extended model, using a ridge penalty and an asymmetric penalty on x. The latter penalty is included to force x to be positive. It has the form $Q^* = \kappa^* \sum u_i x_i^2$, where κ is a large number, say 10^6, and $u_i = 1$ if $x_i < 0$ and $u_i = 0$ otherwise. Like the weights v in asymmetric smoothing, u is updated iteratively.

Figure 4.7 illustrates results and intermediate results for the proposed procedure on simulated data. The input spikes are so closely spaced that the baseline is never reached, except at both ends. The intermediate step finds a positive $\hat{x}$ which is not sparse, but close enough to let the L_0 penalty do its work properly in the final step. In this example $\kappa = 0.01$, $\kappa^* = 3$ and $\lambda = 10^4$. We expect that one has to experiment with these parameter values to get a good result in specific situations. Unfortunately we cannot give general advice on parameter values.

4.3 Applications

In this section the algorithm is applied to three different datasets. The first example comes from DNA sequencing by electrophoresis. The original data consists of four channels; one for each of the four nucleotides. Here we analyse only one channel. In the upper panel of Figure 4.8 the signal is plotted. The middle panel shows a number of peaks that all coincide with the largest peaks present in the data, only a few smaller bumps are not considered a peak. We used blind deconvolution. In the lower panel the initial and final impulse response function are plotted.

The second dataset shows human hormone concentrations in human blood (Vis et al., 2010). A constant baseline is assumed, and subtracted beforehand. An exponentially decaying impulse response has been assumed. Both the value of the baseline and the time constant of the decay have been obtained by playing with parameter values and visually inspecting the results. We do not consider this a full analysis of the data; it only serves as an illustration of principles.

The third dataset is a gas chromatogram, with baseline drift. Prior to blind deconvolution we estimated this baseline with asymmetric smoothing and subtracted it. A Gaussian impulse response is used as a start. Both the data and final results are shown in Figure 4.10.

All three examples show excellent performance of deconvolution with the L_0 penalty.

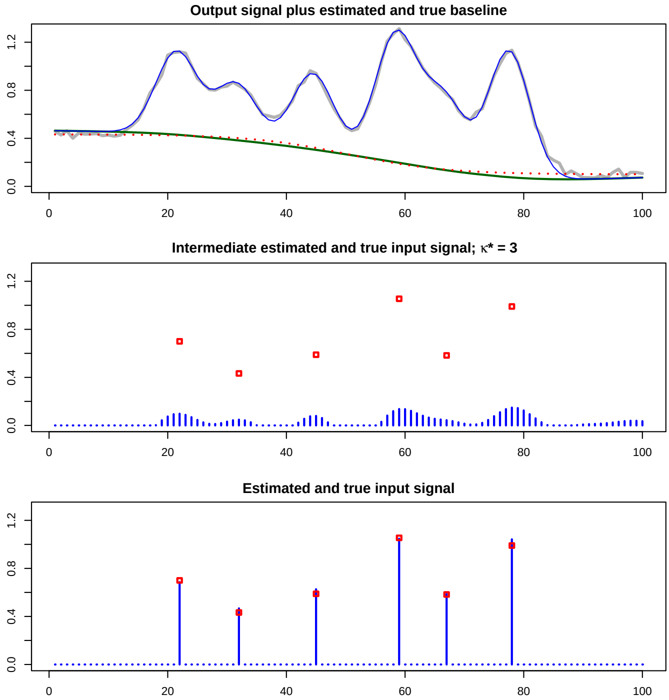


Figure 4.7: Simulation. Baseline estimation and peak detection with the extended model. Upper panel: the simulated data (in blue), the estimated (green) and true baseline (red, dotted). Middle panel: the intermediate solution using the L_2 penalty. Lower panel: final estimates using the L_0 penalty. The small squares give the positions and the heights of the nonzero elements of the input used for the simulation.

4.4 Discussion

Penalized regression is a popular choice to resolve identification issues in the case of ill-posed inverse problems. We have shown that for deconvolution of sparse input signals the L_2 penalty is not useful. The L_1 penalty gives decent results, but not as sparse as we would like to see, and deviations from the true peak height occur. The L_0 penalty shows superior performance: a very sparse result, close to the true input values.

The iterative weighted least squares algorithm works well, on simulated and on experimental data. This is remarkable, because the objective function is not convex, as is the case for the L_1 and L_2 penalties. Apparently there is enough structure in the problem to steer the computations in the right direction. It will be an interesting venture to develop more theoretical insight in why this is the case.

We proposed two ways to handle a drifting baseline: by prior elimination, after estimating it by asymmetric least squares smoothing, or as an explicit smooth component

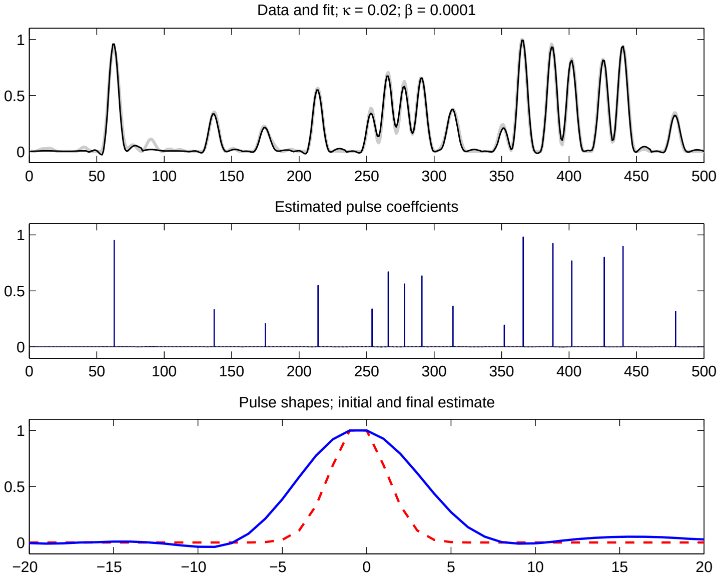


Figure 4.8: Peak detection on the DNA electrophoresis data. Upper panel: output signal and fitted model. Middle panel: estimated peaks. Lower panel: initial (in red) and final impulse response function.

in the model. The first approach has the advantage that the user can judge the reasonableness of the intermediate result. But if peaks are closely spaced, it might be hard to estimate a baseline by asymmetric smoothing. The positivity constraints used in the second approach could be applied even without the presence of a baseline. However, to test the full potential of these constraints in combination with the discussed penalties is left for future research.

Currently we tune the weight of the penalty by visual inspection. Of course, an automated procedure would be attractive, but we did not yet discover clear criteria. The essence of ill-posed problems is that the fit of a model to the data does not give much information: it will always be very good. The desired information should come from properties of the estimated input signal.

Throughout this chapter we modelled the impulse response in such a way that its maximum is at the same position as the impulse response. For real, causal, systems this cannot be true. It is a psychological issue: when judging the results of deconvolution, it is pleasant to see the positions of the estimated input spikes at the peaks of the corresponding pulses in the output.

We have assumed that the impulse response does not change over time. In our applications this turned out not to be a problem. But for other data, and especially for longer data series, this might no longer hold. If it is known how the impulse response changes

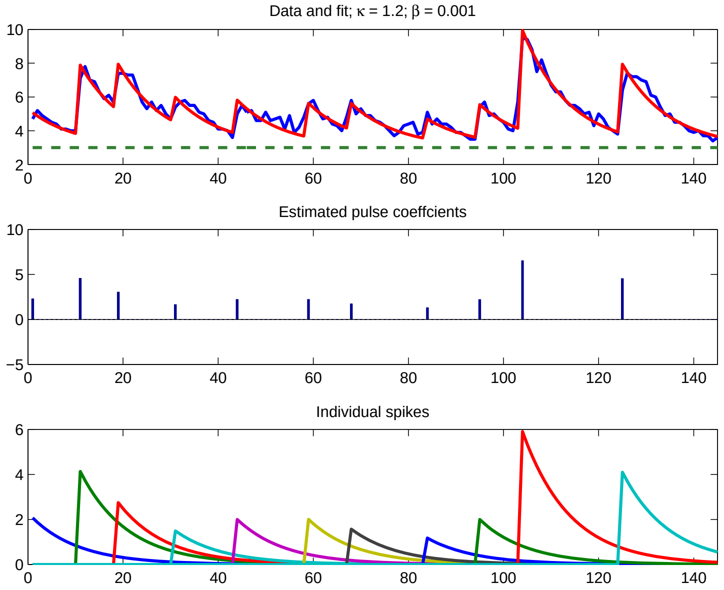


Figure 4.9: Peak detection on the hormone release data, using a constant baseline and exponential decay. Upper panel: output signal (in blue), fitted model (red) and baseline. Middle panel: estimated peaks. Lower panel: individual fitted components.

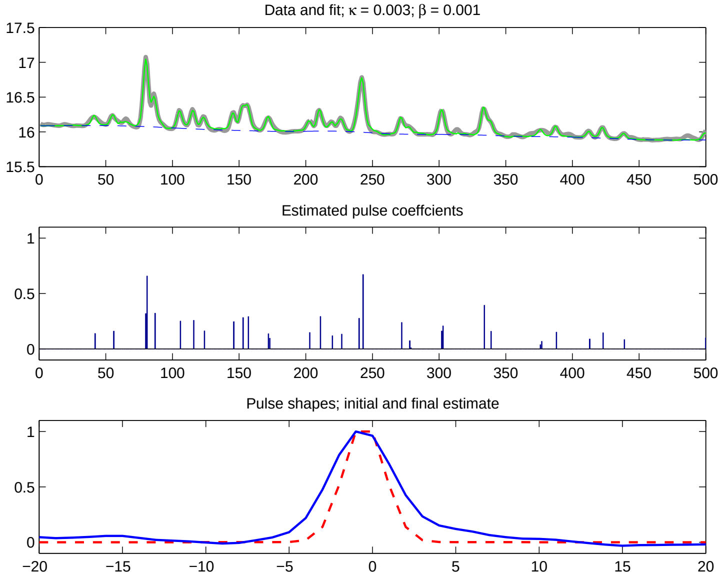


Figure 4.10: Peak detection on the chromatogram data. Upper panel: output signal (in blue) and fitted model. Middle panel: estimated peaks. Lower panel: initial (in red) and final impulse response function.

with time, it is easy to adapt the matrix S: column j contains the impulse response at input time j. Nothing else has to be changed in the algorithms.

A future challenge is to estimate a time-varying impulse response from a data series. We are considering two options for future research: 1) transform time in such a way that on the new scale the impulse response is constant, and 2) estimate the contents of the matrix S as a smooth two-dimensional surface.

To keep the presentation simple, we assumed the same sampling distance for input and output, but we are not limited to this choice. Especially when we model the impulse response with a flexible function, like a spline, to allow arbitrary interpolation, we can use any grid we like. In specialized applications, the output might be sampled on an arbitrary, non-uniform, grid. A uniform grid for the input might be more detailed than that of the output, suggesting opportunities for increased resolution.

The DNA sequencing data as discussed, consists of four traces in parallel, from which we only analysed one. These traces typically show some amount of crosstalk, resulting in one large peak in the correct channel and three smaller ones in the others, for each input impulse. Li and Speed (2000) proposed a model in which a four-by-four crosstalk matrix is estimated together with the distributions of the input signals. It seems natural and feasible to extend L_0-penalized regression to this setting, modelling crosstalk and input signals together.

All computations shown in this chapter were performed in R and Matlab. On request, the software is available from the first author.

Sparse deconvolution in one and two dimensions 5

Abstract. Deconvolution of noisy signals and images is an important task in various areas, examples are: chemometrics, biology and imaging. When the solution is required to be sparse, desirable results are obtained using penalized estimation techniques. Sparseness is realized by shrinking coefficients to zero. We use penalized regression with a penalty based on the L_0 norm, as presented in earlier work. Several extensions to this approach are presented. The model now applies to both one-dimensional and two-dimensional data. In case of blind deconvolution, a smoother is applied to improve the estimated impulse response, which is applicable to any unimodal response function.

5.1 Introduction

In many measurements of physical and biological systems some kind of convolution occurs. This is the phenomenon that an input signal gets 'blurred', in space or time (or both), through delays or spatial interactions. In microscopy, convolution is visible if we approach the optical limit in resolution. In a chromatogram, it is the spreading of peaks due to dispersion. Inside the human body it is the gradual decrease of hormone concentrations after a very short release burst. Convolution is also used in statistics, for instance in time series analysis, to model the effect of time on the output variable.

In many cases convolution is linear and the blurring effect can be described by a so-called impulse response function (IRF). If that is the case, one can try to estimate the input signal from an observed input signal by linear regression. This is called deconvolution. Generally this leads to an ill-posed statistical problem: the solution is extremely sensitive to small changes in the data or the assumed impulse response function. Often the results do not make sense at all against the background of the problem. A further complication is that in many cases only an approximation of the IRF will be available, meaning that

This chapter is based on: de Rooi J.J. and Eilers P.H.C. Sparse deconvolution in one and two dimensions. *Submitted*.

both the input and the spread function have to be estimated; this is often called blind deconvolution.

Stable estimates can be obtained by regularization of the problem. Constraints can be imposed according to the a priori information about the system under study. We know the input consists of a sparse series of positive spikes. All other channels are zero. The response function should be smooth and positive everywhere. We use a shrinkage estimator, employing a L_0 norm penalty. A smooth positive IRF is realized using a unimodal smoother.

The remainder of this chapter is structured as follows. In the next section we first introduce the general model for one-dimensional data. Thereafter, blind deconvolution and optimization of the tuning parameters are discussed. The third paragraph is devoted to sparse deconvolution of two-dimensional data, followed by a section showing some applications. The chapter closes with a discussion on the main findings and possible improvements.

5.2 The deconvolution problem

The signal y of length n is assumed to be a convolution of a series of delta functions x and a fixed impulse response function c of length m, summarizing the blurring or distortion in the instrument. The output signal y can be described by the model:

$$y(t) = \int_0^t x(\tau)c(t-\tau)d\tau. \tag{5.1}$$

Observed signals are almost always discrete. The relationship between input and output can be expressed by the discrete convolution equation:

$$y_j = \sum_{i=-\infty}^{\infty} x_i c_{j-i}, \tag{5.2}$$

resulting in y_j; a weighted sums of the present and all previous values of x, weighted by the impulse response function c. The problem can be rewritten into matrix-vector form:

$$y = Cx + \epsilon \tag{5.3}$$

with residuals ϵ and C being the convolution matrix with $m+(n-1)$ rows and m identical but shifted columns, resulting in a band matrix. An example of a convolution matrix is presented in Figure 5.1.

A solution to Eq. 5.3 can be obtained using least squares: $\min||y - Cx||^2$. Without any constraints this result is useless, as a consequence of the bad condition of the problem. Small errors in y generate large oscillations in the solution $\hat{x}$, and there are an infinite number of ways to fit the data.

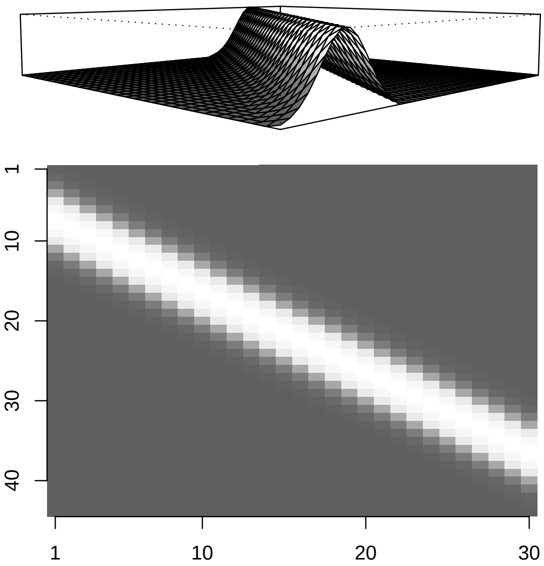


Figure 5.1: A convolution matrix C with $n + (m - 1)$ rows and n columns. The upper panel shows the (rotated) matrix in 3-d, the same matrix in 2-d is visualized in the lower panel.

In the literature we find various deconvolution algorithms. Popular are iterative approaches which are often related to the van Cittert method (see e.g. Jansson, 1996):

$$x^{k+1} = x^k + \gamma(y - Cx^k), \tag{5.4}$$

with γ a relaxation parameter. Iteration starts with an initial guess of the input, x^0. The differences between y and the convolution of x^0 with C are weighted by γ and added to x^0 to finish the first iteration. The van Cittert method has successfully been used for deblurring in imaging applications, but it also exhibits some problems. First, in case of sparse deconvolution, the solution has too many non-zero elements. Second, it allows negative estimates, which does not make sense in the context of the data studied here.

A similar iterative algorithm was proposed by Gold (see e.g. Jansson, 1996). To avoid negative estimates it is based upon multiplicative corrections, instead of additive ones. The algorithm is defined,

$$x_i^{k+1} = x_i^k \frac{y_i}{\sum_j C_{ij} x_j^k}. \tag{5.5}$$

Given that the observed data y and the response function c are positive everywhere, the estimates x will be positive as well. Iterations for both the van Cittert method and Gold deconvolution are repeated until a predefined number of cycles is reached or when a convergence criterion is met. Often a large number of iterations is required which makes the procedures relatively slow. A comparison between these methods can be found in

Coote (1997), very recently Morháč also proposed a boosted version of the Gold algorithm (Morháč and Matoušek, 2011).

An alternative is penalized regression. The least squares cost function is augmented with a penalty, $||y - Cx||^2 + P$. The penalty is a function of x and puts a constraint on the size of the input. By reducing the model complexity, the problem becomes estimable. A prominent example within this family is the ridge penalty by Hoerl and Kennard (1970), which combines the loss function with a quadratic term.

$$P = \lambda \sum_j x_j^2, \tag{5.6}$$

with λ a tuning parameter. Ridge regression is attractive because the solution can be obtained in closed form:

$$\hat{x} = (C'C + \lambda W)^{-1} C'y, \tag{5.7}$$

with $W = I$, the identity matrix. Ridge regression shrinks the estimated coefficients. By doing so, stable estimates are obtained and large variances in the predictions are prevented. A downside is that it does not result in a sparse vector $\hat{x}$. When penalizing the absolute values of the solution, both shrinkage and sparsity are obtained. The L_1 norm penalty (hereafter shortened to L_1 penalty), is well known as the Lasso (Tibshirani, 1996) in statistics or as Basis pursuit denoising (Chen et al., 1998) in signal processing:

$$P = \lambda \sum_j |x_j|. \tag{5.8}$$

It takes the sum of the absolute values of the elements of x. The L_1 norm penalty cannot be solved in closed form, but the goal function is still convex and many optimizers are available to obtain a stable solution. The L_1 penalty is very successful in various areas. In relation to signals and spectra it is used for classification, as demonstrated by Li and Speed (2004) and Du and Angeletti (2006), and for peak identification in for instance mass spectrometry data (Renard et al., 2008).

In applications with high-dimensional data, the L_1 penalty often returns relatively rich solutions. A further increase of the tuning parameter will lead to a sparser result, but at the same time more shrinkage of the estimates. These problems have been well recognized in the literature and have resulted in a number of alternatives. One example is the adaptive lasso (Zou, 2006), which introduces an additional weight for each component. In essence it is a weighted l_1 penalty:

$$P = \sum_j \lambda_j |x_j|, \tag{5.9}$$

with λ_j based on for instance the magnitude of initial estimates, obtained by using an L_1 penalty (see eg. Krämer et al., 2009)).

In the present case too much shrinkage is undesirable because we try to obtain an accurate estimate of the peak height. In fact, our problem is very close to traditional variable subset selection and naturally leads to the L_0 norm penalized cost function:

$$P = \lambda \sum_j I(x_j \neq 0) \equiv \lambda \sum_j |x_j|^0, \tag{5.10}$$

the L_0 norm penalty (hereafter shortened to L_0 penalty) is the weighted sum of non-zero elements of x,

$$|x_j|^0 = \begin{cases} 0, & \text{if } x_j = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{5.11}$$

Because of its discontinuity, proper optimization is challenging. Good results are reported however, by several authors such as Blumensath and Davies (2008), Elad et al. (2007) and Herrity et al. (2006). Attractive results in the context of sparse deconvolution of one-dimensional signals, can be found in de Rooi and Eilers (2011). In many cases either a greedy algorithm or relaxation techniques are used. Two examples of the first approach are forward stepwise regression and orthogonal matching pursuit (Bruckstein et al., 2009). Both algorithms add variables sequentially based on the significance or correlation with residuals. The downside of these methods is that they become slow when there are many variables.

Relaxation techniques convert the L_0 norm optimization into something that is more easy to solve. Here we use an algorithm based on iteratively reweighted least squares (IRLS). This method is earlier explained in connection to the L_1 norm by e.g. Osborne et al. (2000), but is also used for L_0 norm optimization (Bruckstein et al., 2009). It basically means that the problem is iteratively converted into a weighted L_2 norm.

It is clear that $|x_j|^p = x_j^2/(x_j^2)^{(2-p)/2}$. When x is unknown, but approximated by $\tilde{x}$ we have: $|x_j|^p \approx x_j^2/(\tilde{x}_j^2)^{(2-p)/2}$. For numerical stability, we also introduce β, a small constant about a thousand times smaller than the expected maximum of x, which results in:

$$|x_j|^p \approx x_j^2/(\tilde{x}_j^2 + \beta^2)^{(2-p)/2} = w_j x_j^2. \tag{5.12}$$

This weighting scheme can be used in Equation (5.7). In the case of the L_1 norm penalty W is a diagonal matrix with $w_{jj} = 1/\sqrt{\tilde{x}_j{}^2 + \beta^2}$ and for the L_0 penalty: $w_{jj} = 1/(\tilde{x}_j^2 + \beta^2)$. The idea is to iterate with this formula, using for $\tilde{x}_j$ the current approximation to the solution x_j. Starting values for x can be obtained from a ridge regression, or one can start from a constant x.

Blind deconvolution

In the preceding sections we assumed that the impulse response function is known. In practice, the true impulse response function is hard to obtain and one has to use an approximation, or estimate it from the data. In cases where the IRF can be well approximated by a parametric function, one could fit a series of alternative functions, and choose the one that fits the data best (see e.g. Reiss et al., 2008; Vis et al., 2010). Model fit can be determined by inspection of the residuals. Estimating both the input and the response function only by using the data, is often called blind deconvolution.

It is clear that blind deconvolution is more difficult compared to the case where the response function is known. However, when the (convoluted) peaks in the signals are relatively separated from each other, it becomes more easy to deduce information concerning the shape of the IRF. When estimating the response function we are also warned when it is too far of from reality. When the function is too narrow, compared to the truth, too many spikes will appear in the estimated x. In cases where the response function is taken too wide, negative peaks will be the result.

To estimate both the IRF and the input, we use an alternating minimization scheme. Assuming that we know the impulse response, we can make an estimate of the series of input spikes. Reversely, if we know x, we can estimate the IRF. This procedure is repeated until convergence. In Equation 5.12 we saw how to estimate the series of input spikes, given the impulse response function. The unconstrained solution for the IRF is given

$$\hat{c} = (X'X)^{-1}X'y, \tag{5.13}$$

with X being a band matrix with in the columns the (shifted) input x. In principle estimating c does not require regularization, but we can get better results with penalization. Good results are obtained by applying a penalty on the differences: $\hat{c} = (X'X + \lambda^* D'D)^{-1}X'y$. The matrix D is a dth order difference matrix, imposing smoothness of the vector c. λ^* is the tuning parameter and regulates the smoothness of the function. Using this set-up, overfitting is a risk if the penalty is not optimized.

The applications discussed within this chapter all have a response function that is unimodal and positive everywhere. Both constraints are implemented in the unimodal smoother as discussed in Eilers (2005). Using this smoother, blind deconvolution can be applied to any type of unimodal response function. One example is an exponential decay function, presented in the application section. Because the limit of the smoother is a bell curve, overfitting is no issue in case of a normal response function. The unimodal smoother is defined:

$$Q = \sum_i (y_i - \sum_j x_{i-j} e^{u_j})^2 + \lambda^* \sum_j (\Delta^3 u_j)^2 + \gamma \sum_j v_j (\Delta^2 u_j)^2, \tag{5.14}$$

with $c_j = e^{u_j}$, Δ being the differencing operator and γ a large value, say 1000 or larger. The weights of v depend on the sign of $\Delta^2 u_j$,

$$v_j = \begin{cases} 1, & \text{if } \Delta^2 u_j > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5.15}$$

As a result, the second penalty only shrinks where the differences in u are negative. For a detailed discussion we refer to Eilers (2005).

Model selection

In the previous sections we laid out the model estimation framework. Sparseness is obtained using an L_0 penalty and in cases where the impulse response is unknown it can be estimated in an iterative manner. Here we discuss three types of model selection criteria:

- Manual tuning
- Restriction of model complexity
- Using information criteria

In many cases the user knows the data well, meaning the model can be tuned by visual inspection satisfactorily. If one knows the limits of the studied system with respect to the pulse rate, it is also possible to restrict the maximum number of estimated peaks. The problem can be rewritten as a constrained minimization problem:

$$\min_x ||y - Cx||^2 \qquad \text{subject to} \qquad ||x||_0 = K, \tag{5.16}$$

with K being the number of spikes in x. The solution can be obtained by starting with a large λ and subsequently reduce it until $||x||_0 = K$. An example can be found in Vis et al. (2010), where the authors restrict the allowable number of estimated hormone secretion bursts using knowledge from endocrinology. Base calling in DNA sequencing data is a second example. Knowing the length of the signal one can calculate the maximum number of possible bases on the string.

The two most prominent statistical criteria are the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian in formation criterion (BIC) (Schwarz, 1978). Here we use the AIC which is defined as

$$\text{AIC} = n\text{log}(\hat{\sigma}^2) + 2K, \tag{5.17}$$

with

$$\hat{\sigma}^2 = \frac{||\hat{\epsilon}||_2^2}{n}. \tag{5.18}$$

Typical for penalized regression models is a smooth AIC curve. Using an L_0 norm penalty, this is not the case, this can be explained by the discrete steps of the effective dimension, when increasing the weight of the penalty.

Simulations

Simulations with known IRF

In this section we present a small simulation, to make a comparison between three deconvolution algorithms. A signal is generated, including six peaks and using a Gaussian shaped response function. The signal is short, so that performance can be inspected visually. The results for Gold deconvolution, penalized estimation using an L_1 norm penalty and an L_0 penalty are provided in Figure 5.2. The optimal penalty parameter for the L_1

and L_0 penalty is obtained using the AIC. The response function is known to all algorithms. The observed signal including noise is printed in red. The positions and heights of the true input spikes are depicted with the blue circles. The estimated pulses are presented in green.

The upper panel shows the results using the Gold algorithm, which converged after about 500 iterations. Observing the result, we see that too many peaks are estimated and it seems that the algorithm is rather sensitive to noise. Using an L_1 penalty yields better results, which can be seen in the middle panel of Figure 5.2. This method performs better in finding the right location of the peaks. However the shrinkage property of this penalty introduces bias in the form of small peaks. It could also be the case that the AIC is unable to detect the best configuration. However, manual tuning does not result in better estimates. Increasing the penalty term reduces the height of the estimated peaks and can even lead to the appearance of more peaks. The most satisfactory results are obtained using the L_0 norm penalty. These results are shown in the bottom panel of Figure 5.2. Both peak positions and heights are estimated with great precision.

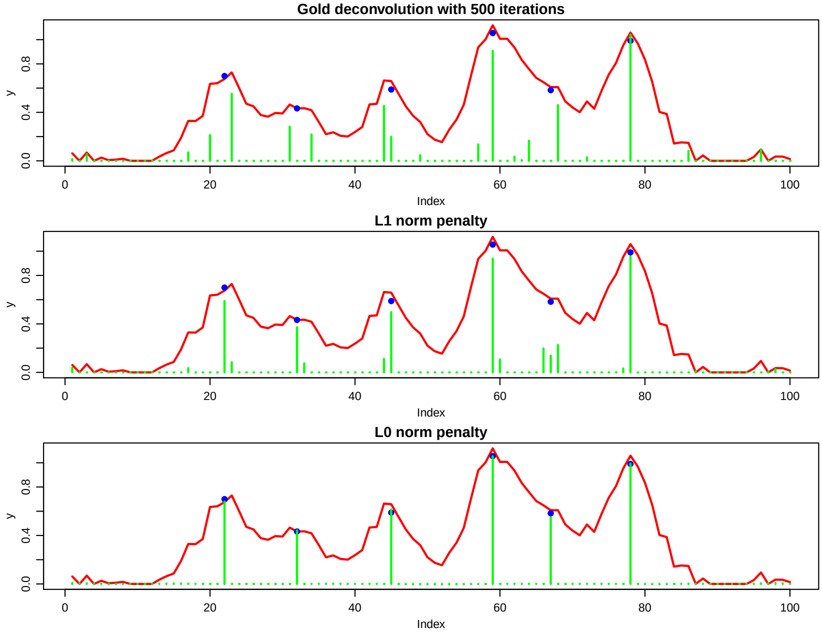


Figure 5.2: A simulated signal including noise. The observed data are printed in red, the estimated pulses in green. The position and height of the true input spikes are printed as blue dots. The upper panel shows the results for the Gold deconvolution, the middle panel the results using the L_1 penalty and the bottom panel using the L_0 penalty.

Simulations with unknown IRF

To construct an example to illustrate blind deconvolution, we use the same input signal as in Figure 5.2 and concentrate on the L_1 and L_0 norm penalized methods. To make the example somewhat harder, the true response function is changed to skewed normal. We choose to optimize the penalty manually, because the AIC is not suitable in this case. The effective dimension in Equation 5.17 only incorporates the size of the input x and does not includes restrictions on the estimated response function.

The results for the L_1 penalty are presented in Figure 5.3. The upper panel shows the raw signal (in red) the position and height of the true input spikes (blue dots) and the estimated input (in green). When using this penalty, we do not retrieve the right number of spikes, the estimated IRF is much too narrow and its right tail is missing completely. The initial IRF is presented in red, the truth in grey.

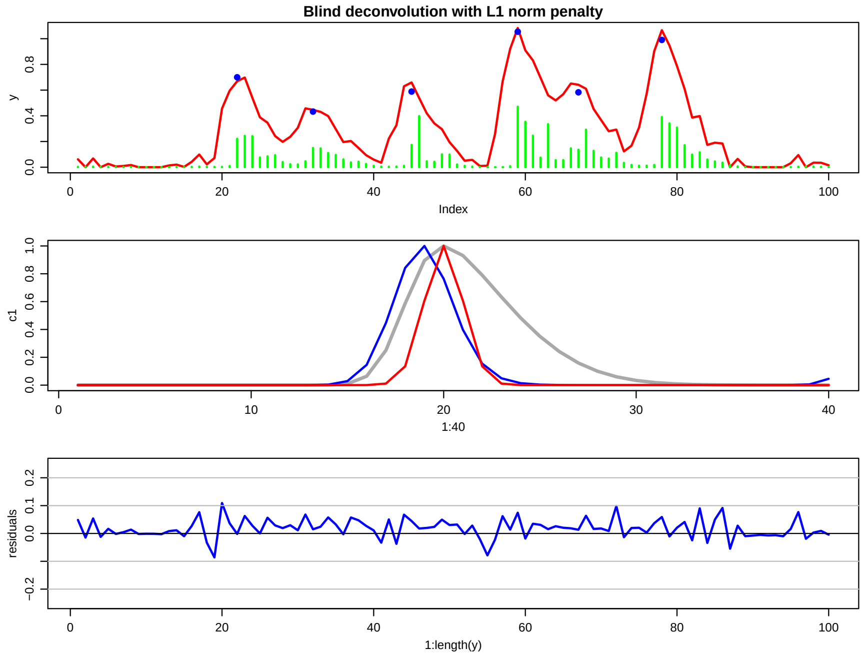


Figure 5.3: The results of blind deconvolution, using an L_1 norm penalty. The upper panel shows the observed signal (red), the estimated pulses (blue) and the true height and positions of the spikes (blue dots). The middle panel shows initial (red), estimated (blue) and true IRF (grey). Residuals are presented in the bottom panel.

The results for the L_0 penalty are very satisfactory, as can be seen in Figure 5.4. Six spikes are estimated, with right positions and intensities. The estimated response function is very close to the true one. It appears that even when starting with a much to narrow response function, there is enough information in the data to converge to a very good solution.

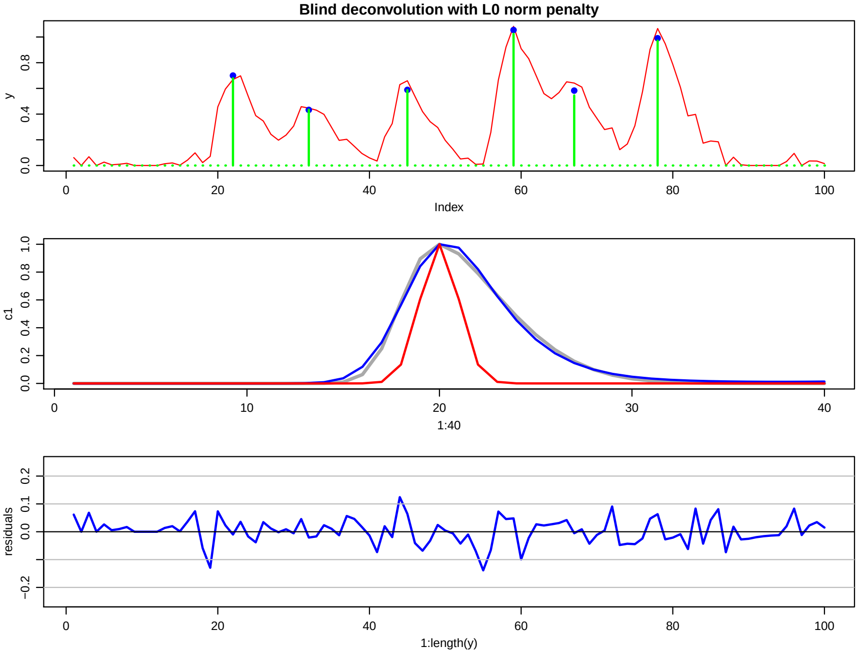


Figure 5.4: The results of blind deconvolution, using an L_0 norm penalty. The upper panel shows the observed signal (red), the estimated pulses (blue) and the true height and positions of the spikes (blue dots). The middle panel shows initial (red), estimated (blue) and true IRF (grey). Residuals are presented in the bottom panel.

5.3 2d deconvolution

In addition to one-dimensional deconvolution, there is a large volume of publications discussing applications in two dimensions. Many of those focus on some application in imaging, trying to remove optical distortions (see e.g. Starck and Murtagh, 2006, for an overview), but they do not require sparseness of the solution. Two-dimensional sparse deconvolution is discussed by Belghith et al. (2011) in relation to 2d NMR data, or by Morháč et al. (1997) where it is applied to γ-ray spectrometer data. Further applications can be found in fluorescence microscopy and will be treated in the next section.

Two-dimensional deconvolution is challenging, because of the large number of variables in the regression problem, one for each pixel. The data are summarized in a matrix Y, with dimensions $I \times J$. The model becomes:

$$y_{ij} = \sum_k \sum_l s_{i-k,j-l} x_{kl} + \epsilon_{ij}, \tag{5.19}$$

with the impulse response function s, a two dimensional (discretized) density. Alternatively, s can be translated into a four-dimensional array C:

$$y_{ij} = \sum_k \sum_l c_{ijkl} x_{kl} + \epsilon_{ij}. \tag{5.20}$$

To solve the system, we vectorize the data: $y = \text{vec}(Y)$, and translate C to the matrix C^* with dimensions $IJ \times IJ$. The special structure of this matrix can be seen in Figure 5.5. This matrix is still relatively sparse, but depending on the response function the bandwidth is considerable. If the images become large, estimation is computationally demanding and segmentation is required.

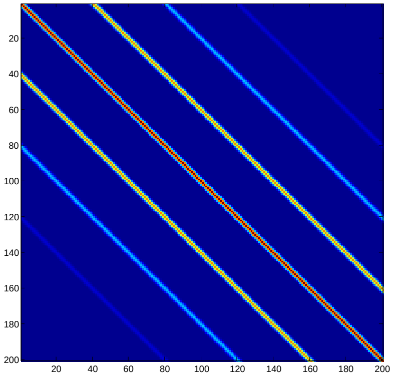


Figure 5.5: Vectorizing the two dimensional data requires an adjustment of the convolution matrix, resulting in the banded matrix C^*.

Figure 5.6 shows the results of a simulation. The size of the image is 40 × 40 pixels and contains five partly overlapping dots. The data is similar to the single-molecule fluorescence imaging data discussed in the application section. The goal is to estimate the positions of the dots, knowing the response function. The left panel of Figure 5.6 shows the observed data, the estimated image is depicted in the middle panel. The estimated input spikes are presented in the right panel. Interesting about this simulation is the convergence rate and the obtained degree of sparseness.

The convergence is depicted in Figure 5.7. The image shows the reduction in the number of non-zero coefficients for each iteration. The first step reduces the number of non-zero coefficients by more than 50 percent. After ten iterations, only six coefficients remain.

5.4 Applications

In this section two applications are discussed. We start with a discussion of hormone release data. As a second application we discuss sparse deconvolution of fluorescence images.

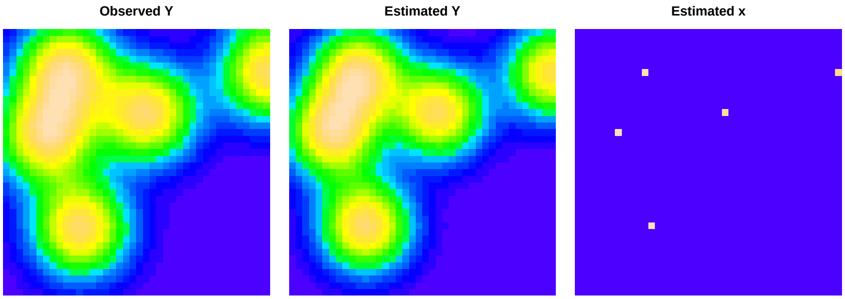


Figure 5.6: Simulated single-molecule fluorescence imaging data (left), the reconstructed image (middle) and the estimated input (right).

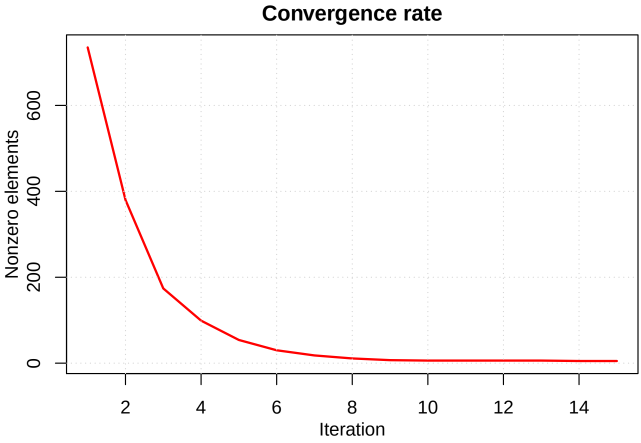


Figure 5.7: The convergence of the algorithm, expressed by plotting the iterations against the number non-zero spikes.

Pulse identification in endocrine time series data

The data are luteinizing hormone levels, measured during 24 hours at intervals of ten minutes, see Vis et al. (2010, 2012). The goal in this application is to model the secretion pattern, which is assumed to be a sparse series of spikes. The impulse response function is known to follow a decay close to an exponential function. We model it non-parametrically. In Figure 5.8 we see the result of the (blind) deconvolution. The upper panel shows the observed data in grey. The red signal is the estimated one; a convolution of the estimed IRF and series of spikes. The middle panel shows the sparse series of estimated input peaks. The 11 secretion bursts corresponds with the major peaks in the estimated data, all other coefficients are set to zero. The third panel shows the impulse response, estimated using blind deconvolution, in red. The blue IRF is used as an initial estimate, its decay

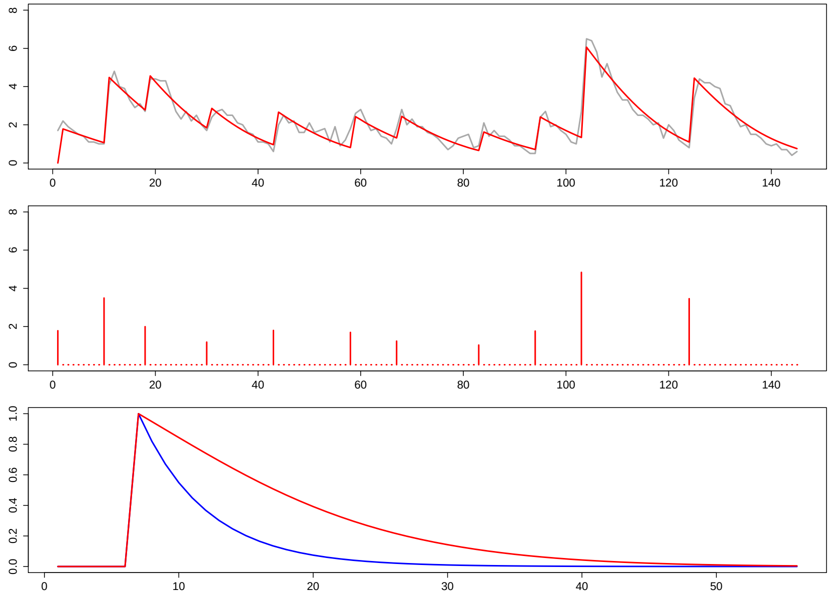


Figure 5.8: Estimating input pulses for the hormone release data. In the upper panel we show the data (grey) and estimated output. The middle panel shows the series of estimated peaks. The lower panel shows the initial IRF (blue) and the estimated one (red).

rate is clearly too rapid for the observed data. It shows the robustness of the procedure; even with a large discrepancy between the initial and true IRF, the algorithm converges to a very good approximation.

Single-molecule fluorescence imaging

In the second application we focus on two-dimensional data. Fluorescence microscopy is used to study the complex spatio-temporal interplay of biomolecules and is widely used in molecular and cell biology. The data are a series of images showing actin structures in which the structure is highlighted using quantum dots fluorophores. Due to their photoswitching property, the fluorophores blink randomly over time, and make closely positioned molecules resolvable in time. These properties are used to obtained super-resolution images. Here we focus on localisation; estimating the exact positions of single molecules. Analysing one image is an example of sparse deconvolution in two dimensions. A more elaborate treatment of these data will be given elsewhere.

In addition to the signal we assume a baseline which need to be removed before deconvolution. Here we rely on a mixture model approach, discussed in de Rooi et al. (2013). Figure 5.9 shows the deconvolution of the actin data. The panel on the left in the first row

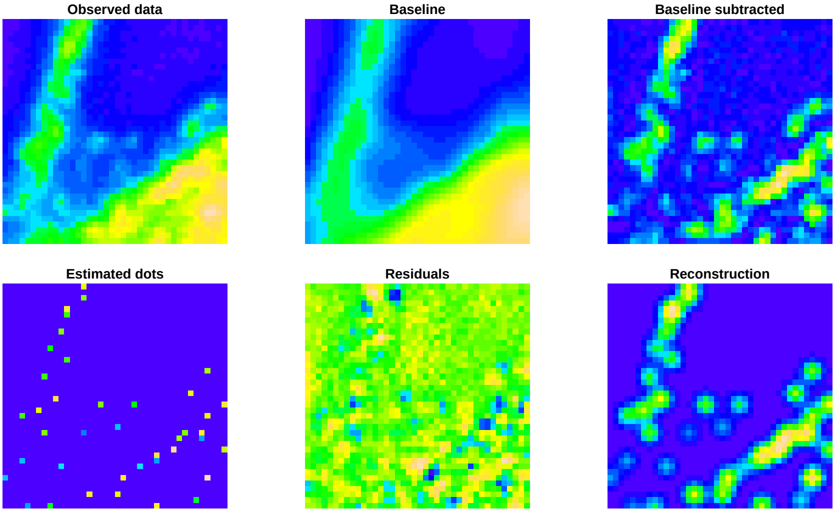


Figure 5.9: Deconvolution of the actin data, after baseline subtraction. The sparse reconstructed image is showed in the bottom right image.

show a part of the observed image. The estimated baseline in presented in the middle panel. Subtracting the baseline from the observed data results in the picture presented on the right. The left image on the second row shows the estimated positions of the dots. The residuals and the reconstructed image are presented in the two remaining figures.

5.5 Discussion

Sparse deconvolution is a prominent problem in various disciplines. In the previous chapter we proposed sparse deconvolution using an L_0 penalty, which is able to estimate very sparse models in a reliable way. In the current chapter some additions are proposed.

The algorithm is successfully applied to two-dimensional data. Vectorization of the two-dimensional data results in large systems, which require considerable computation time. A solution is to split images in (overlapping) segments and analyse these separate. In addition, the model now allows blind deconvolution with in theory any shape of unimodal impulse response function. The IRF is estimated from the data, given the (updated) vector of input pulses. Smoothness is imposed and results in a unimodal IRF. The blind deconvolution algorithm can be further extended to the two-dimensional case.

The deconvolution algorithm is based on penalized regression, and allows various further extensions. Some applications show band-broadening effects, meaning that the observed peaks gets broader over the observed range. To capture this effect the model can be extended with a time dependent impulse response function. In some situations with multiple signals, like DNA base calling and EEG signals, the data shows crosstalk. Meaning that one signal 'spills over' into other signals. A possible extension of the presented model corrects for this crosstalk by introducing a mixing matrix, aiming to separate the different sources.

The models as presented are often tuned visually, with satisfactory results. In a broader setting it is customary to apply model selection using for instance cross-validation or some type of information criteria. We have shown that the AIC is useful in some cases, but needs adjustment to be applicable to blind deconvolution.

Acknowledgements

We thank J.I. Hotta and J. Hofkens (K.U. Leuven) for providing the actin data, as well as D. Vis (University of Amsterdam) and H. Pijl (Leiden University Medical Center) for providing the hormone data.

Classification of outlines using penalized signal regression

6

Abstract. Various medical and biological applications require the classification of two-dimensional rounded outlines. In addition to classification, proper preprocessing is needed. We propose a scheme with several steps: 1) rectangular coordinates are converted to polar coordinates; 2) scaling and rotation is applied; 3) the radius is lightly smoothed using (circular) P-splines as a function of the angle; 4) the spline coefficients are used as explanatory variables in logistic penalized signal regression. Three applications to real-life datasets are presented, showing excellent classification performance.

6.1 Introduction

Shape analysis concerns the use of geometrical properties in order to distinguish different objects. Observed differences can be used for description or classification purposes. Applications can be found in fields like medicine, biology and engineering. Acharya and Ray (2005) and Dryden and Mardia (1998) show examples in several areas.

Shapes can be described using landmarks, locations or points of correspondence on measured samples that are used to calculate distances, ratios or angles to characterize groups. Dryden and Mardia (1998) distinguish three types of landmarks: anatomical, mathematical and pseudo-landmarks, all recorded in (x, y) space. The first two types of landmarks are concisely defined as distinguishable points or markers, with biological meaning or based on mathematical properties, and should be easily reproducible on independent samples. The advantage of these types of landmarks is that complex shapes are represented by a small set of variables, and subsequently can be analyzed using familiar tools like discriminant analysis. Sometimes clear markers are lacking, or whole shapes unveil more details. In these cases one can decide to analyze complete outlines. By using a moderate number of what Dryden and Marida call pseudo-landmarks, an accurate

This chapter is based on: de Rooi J.J. and Eilers P.H.C. Classification of almost convex outlines with penalized signal regression. *Submitted.*

description of shapes can be realized. In contrast to both other types, there is no longer a decision rule for the placements of the landmarks, the only requirement is that they are placed on such an interval that yields a satisfactory approximation of the original outline. With modern technologies the outlines can be easily tracked (semi-)automatically, and a large number of (x, y) coordinate pairs can be obtained.

The motivating example for this chapter is a set of skull outlines from children. The outlines show abnormal patterns, caused by premature fusion of one or more cranial sutures. Making the right diagnosis is crucial to decide on possible treatments. Our aim is to classify the samples using complete outlines in order to classify samples in the right group. There are two important issues that have to be addressed when analyzing outlines:

- Preprocessing of the data.
- The ill-conditioned nature of the problem, caused by the large number of coordinate pairs.

This chapter presents a framework in which both problems are solved in an elegant way. For the preprocessing, the pairs of (x, y) coordinates are translated into polar coordinates (ϕ, ρ) to facilitate both preprocessing and classification. The resulting curves are then summarized on a P-spline basis, which makes them of equal length, smooths possible noise and reduces the dimensionality of the problem. If required, the curves can be rotated and normalized, for proper comparison. In addition, it is possible to smooth areas of low support, or small parts of missing data.

For the classification part, measures have to be taken in order to resolve the condition problem. An often applied solution is to summarize the data on a lower dimension, e.g. using principal components analysis as pioneered by Cootes et al. (1995). A second option is to apply some regularization. One example, analyzing open outlines, can be found in Ramsay and Silverman (2002), who use linear discriminant analysis with a ridge penalty (Hoerl and Kennard, 1970). We use penalized signal regression (PSR), based on the algorithm as presented in Marx and Eilers (1999). In the regression, each P-spline coefficient is treated as a variable. In order to make the model estimable and prevent overfitting of the data, we force the coefficient vector to be smooth by using a difference penalty. Because the data are not transformed, the results of the regression are well interpretable.

We propose the following workflow:

- Determine the origin of the image.
- Convert cartesian to polar coordinates.
- Fit P-splines.
- Rotate to align with the reference.
- Normalize the radius.
- Classify using penalized signal regression.

In the next section all preprocessing steps are discussed. P-splines take a central role in the proposed methodology, and are therefore explained into detail. The classification algorithm is explained in the third section. In the fourth section, three applications are demonstrated. For comparison, we apply two additional classifiers, one using mean scores and a second based on principal components. A general overview and a discussion concerning possible improvements are given in the last section.

6.2 Methods

Data conversion

The studied data are digitized outlines, translated from images into series of rectangular coordinates, generally a different number for each sample. The cartesian system is probably the most common way of representing images, alternative systems are available and might be preferable in some applications. The methodology presented in this chapter applies to outlines that are star-convex with respect to their center of gravity o. This means that for any straight line drawn from this point to the outline, the whole line segment lies inside the figure. In the last section of this chapter we pay some attention to more complex shapes and ways to model these. In the upper panel of Figure 6.1 we see one head circumference plotted using rectangular coordinates, which is a clear example of a star convex shape. Translating these rectangular coordinates into the polar coordinate system, results in non-crossing curves. Preprocessing like rotation and transformation can be easily performed on the curves while centering of the original images is not needed.

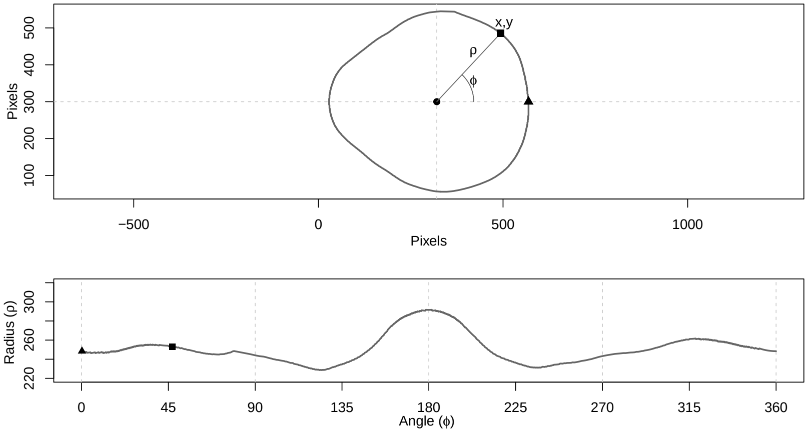


Figure 6.1: The upper panel shows one skull circumference using the cartesian system, in the lower panel the same data is plotted, now using polar coordinates.

Polar coordinates are a natural choice for relatively rounded shapes, and result in a simple representation of the data: a constant in the case of a circle, or a gradually changing radius as a function of the angle. They are regularly used to model outlines; applications can be found in e.g. Claude (2008) and Prossinger (2005). A similar methodology is presented by Seul et al. (2000). They calculate the difference between the radii of a sample and the radii of a reference curve.

In the lower panel of Figure 6.1 we see the same data as shown in the upper panel, here plotted using polar coordinates. The x and y coordinates are translated into the radius ρ and the polar angle ϕ. The pole is calculated by taking the mean of the (x, y) coordinates, assuming that the coordinates are sampled on a regular interval. If this assumption does not hold, we propose a two stage procedure, explained in the discussion section. The development of the radius ρ as a function of the angle ϕ is characteristic for the shape. The starting position of the curve coincides with the position of the square in the upper panel. The outline is subsequently described counter clockwise. Notice that the beginning and end of the curve should join.

P-spline fit

A single B-spline is composed of q+1 polynomial pieces, each of degree q. The polynomial pieces join at q points, called the knots and together form a smooth bell-shaped curve. Using a series of equally spaced B-splines, a smooth curve $f(\phi)$ can be fitted to the observed ρ (see Dierckx, 1995; de Boor, 2001, for a thorough discussion). Here we rely on the scheme as set out by Eilers and Marx (1996). They introduced P-splines: a basis of B-splines with in addition a penalty on the differences, forcing adjacent spline coefficients to be more similar.

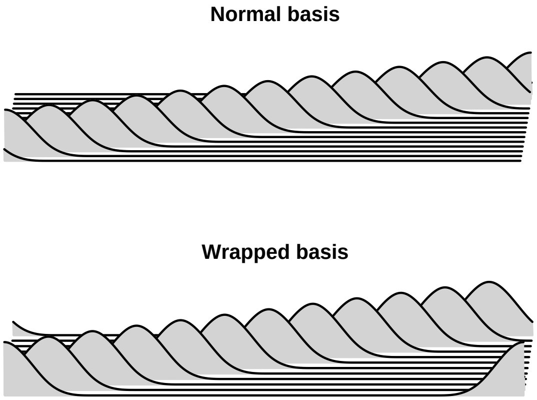
The model without a penalty can be written as $f(\phi) = B\alpha$. If we let the data be represented by B-spline coefficients we get the following:

$$\rho = \sum_{j=1}^{k} \alpha_j B_j(\phi, q), \tag{6.1}$$

with $B_j(\phi, q)$ being the value at ϕ of the j-th B-spline of degree q. In the following we only use cubic B-splines. Fitting B-splines to our data gives us the minimization function:

$$S = \sum_{i=1}^{n} (\rho_i - \sum_{j=1}^{k} B_j(\phi_i)\alpha_j)^2 = ||\rho - B\alpha||^2. \tag{6.2}$$

A regular basis B consists of $k + q$ basis functions, with k chosen by the user. We incorporate the cyclical component of the studied contours into the smoother, causing the beginning and end of the signal to join together (Eilers and Marx, 2010). The wrapped around basis consists of k columns, and starts in the first column with half a B-spline. The second half is placed on the bottom of this column, for the second column and the last column the same pattern appears: the missing piece for a full B-spline is placed on the opposite end. In Figure 6.2 the original (upper panel) and the cyclical (lower panel) basis are shown.



Normal basis

Wrapped basis

Figure 6.2: The upper panel shows the standard basis B, in the lower panel a wrapped basis is presented. For a better presentation, both matrices are rotated about 90° counter clockwise with respect to their normal orientation.

The penalty on the differences is defined:

$$P = \lambda \sum_{j=1}^{k} (\Delta_c \alpha), \tag{6.3}$$

with parameter λ tuning the relative weight of the penalty. The Δ_c denotes the first order circular difference operator, and is defined: $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$. Like matrix B, the difference matrix D has to be adjusted to the cyclical nature of the data. An example for a first order penalty using five B-splines:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix}, \tag{6.4}$$

notice the last line wraps around, forcing the beginning and end of the signal to join smoothly. Including the penalty term in the objective function (6.2) gives:

$$S^* = S + P = ||\rho - B\alpha||^2 + \lambda \sum_{j=1}^{k} (\Delta_c \alpha). \tag{6.5}$$

The system of equations, leading to the minimization of S^*:

$$\hat{\alpha} = (B'B + \lambda D'D)^{-1} B'y. \tag{6.6}$$

The optimal trade off between a smooth curve and fidelity to the data, can be calculated by using cross-validation, or the Akaike information criterion (AIC). Signals, consisting of hundreds of data points, can be easily summarized on a basis of about 50 B-splines. In our situation there is no need to optimize the smoothness at this stage, here we only add a very small penalty for noise removal. Optimization of the penalty with respect to the estimated regression coefficients is discussed in the next section. Because we are only interested in differences in shape, the data are normalized by making the average radius equal to each other. This corresponds to lowering or raising individual curves.

Rotational adjustment

Proper alignment is important to bring all curves in phase with each other, in order to be able to make a fair comparison. After the curves are summarized on a P-spline basis, they can be presented on a grid. For alignment we take one signal as a reference and match all others individually with this sample. If a sample is out of phase, a circular shift algorithm can be applied, and rotates the sample along the ϕ-axis. For every step, the association with the reference is calculated. In a final step the sample is shifted to the point of maximum association, and the P-spline coefficients are updated. The procedure is visualized in Figure 6.3. In the upper panel we see one sample shifted ten times 4° to the right along the ϕ-axis. The middle panel shows the correlation with the reference for a rotation over 100° of the curve. In the lower panel the reference curve (solid line), the rotated curve (dotted line) according to the highest correlation, and the sample in its original position are printed.

Classification with logistic regression

The result of the pre-processing steps is a matrix $A = [\alpha_{ij}]$, containing the coefficients of the P-splines that are fitted to the signals. Each row of A corresponds to an object, and the number of columns of A equals the number of B-splines. To classify the objects we can use logistic regression, with the columns of A as explanatory variables. The model is defined:

$$\eta_i = \log \frac{p_i}{1 - p_i} = \sum_{j=1}^{k} \alpha_{ij} \beta_j, \tag{6.7}$$

with the linear predictor η_i, and is the logit of the probability p_i that object i belongs to the reference class. In principle this is just a Generalized linear model (GLM) (McCullagh and Nelder, 2000). But because the number of explanatory variables is relatively large, leading to a singular or ill-conditioned regression problem. Therefore we borrow ideas from penalized signal regression (Marx and Eilers, 1999), and use a difference penalty as introduced in (6.3), but now applied to the estimated regression coefficients. The penalty: $P = \lambda \sum_{j=1}^{k} (\Delta_c \beta)$, forces adjacent estimates to be more similar. The relative weight of

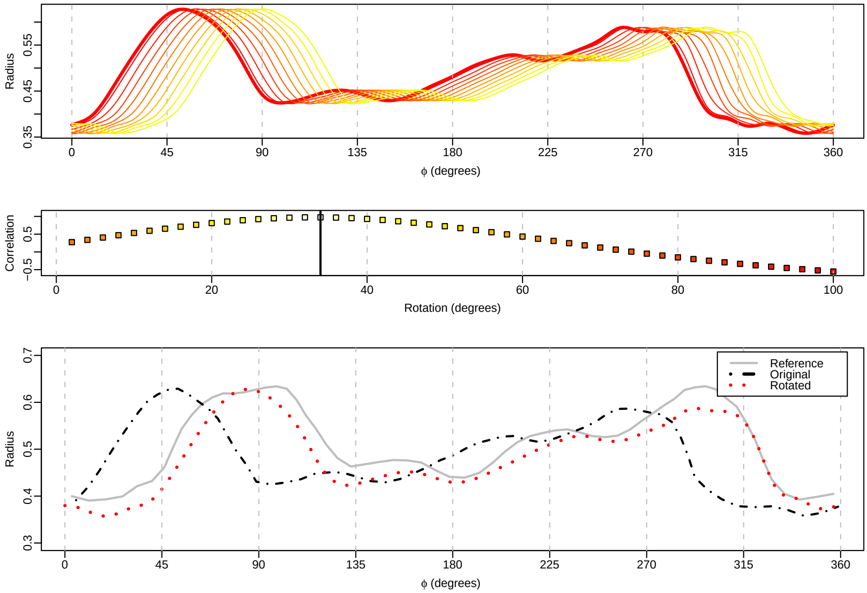


Figure 6.3: Rotation of the signals is needed when the original images are not properly aligned. In the upper panel we see one sample shifted ten times 4° along the ϕ-axis. The middle panel shows the correlation of the sample with the reference over 50 shifts of 2°. The vertical line indicates the best match between both samples. In the lower panel the reference sample (solid line), the original sample (dash-dot line) and the rotated sample according to the best match.

the penalty compared to the loss function (and hence smoothness of the coefficient vector) is regulated with λ. The vector β can be interpreted as normally in logistic regression. Plotting the estimates in cartesian coordinates gives a nice visualization of the most discriminating areas along the outlines. The log-likelihood for the normal model is defined:

$$l = \sum_{i=1}^{n} y_i \log p_i + (1 - y_i)\log(1 - p_i). \tag{6.8}$$

Adding the difference penalty to this function provides us with the penalized log-likelihood:

$$l^* = l - \frac{1}{2}\lambda \sum (\Delta_c \beta)^2. \tag{6.9}$$

There is no analytical solution for (6.8), but various optimizers are available to solve this problem. We use the following iterative algorithm, producing an updated vector of β coefficients with each iteration:

$$\beta = (A'\tilde{W}A + \lambda D'D + \kappa I)^{-1} A'(y - \tilde{p}), \tag{6.10}$$

including the difference penalty, and with $W = \text{diag}(p(1-p))$. In practice, convergence is usually reached in a handful of iterations. Notice that in addition to the difference penalty a ridge penalty is added to the equation, with κ being the tuning parameter and I an identity matrix (Hoerl and Kennard, 1970). The ridge penalty will make the model estimable over the whole range of λ, and results in more stable estimates as discussed in Marx and Eilers (2002). In practice we do not optimize κ but fix it at a small value.

The model is fitted using a training and a test set. The training set is used to optimize the weight of the penalty by varying parameter λ. The size of this set preferably comprises about 2/3 of the total sample, but depends on the number of available samples. To optimize the tuning parameter, we use the Akaike information criterion (Akaike, 1974) (see Burnham and Anderson, 2002, for a recent treatment):

$$\text{AIC} = \text{deviance} + 2\text{ED}, \tag{6.11}$$

the deviance equals $-2l$ (McCullagh and Nelder, 2000), and the ED is the effective dimension of the model, which is defined as the trace of the smoother matrix H, as in Hastie and Tibshirani (1990):

$$H = A(A'WA + \lambda D'D + \kappa I)^{-1}A'W, \tag{6.12}$$

with W and D as defined before.

The performance of the trained model is determined by predicting the samples in the test set. There are various recipes for training and testing a classifier (see e.g. Witten and Frank, 2005, for an overview). Here we rely on a repeated holdout procedure in order to asses the (average) performance of the model. Repeatedly all samples are divided into a training and test set, the model is trained and performance is determined using the test set. The error is expressed as the percentage of wrong predictions in the test set, calculated over all runs.

6.3 Results

In this section three applications are discussed. The first one is concerned with the outline of skulls. This data is the motivating example for this chapter, but because of unexpected delays, it currently consists of only eight samples, and is used only for description. In a second application the aim is to distinguish between two types of rodents, based upon the outlines of their molars; the data comes from Claude (2008). The third example is concerned with the classification of diatoms, and is described in Jalba et al. (2005).

In the two latter applications, we apply two alternative classifiers in addition to the PSR. As a first alternative we apply penalized logistic regression (PLR) on the cartesian coordinates. A ridge penalty is used to make the problem estimable, and is optimized using AIC. The second method is similar; again we use logistic regression only now using principal components (PCLR) as input for the classifier. The loadings of the principal components are derived from the training set, and projected onto the data of the test set. The number of principal components is optimized with the AIC as defined before. Although

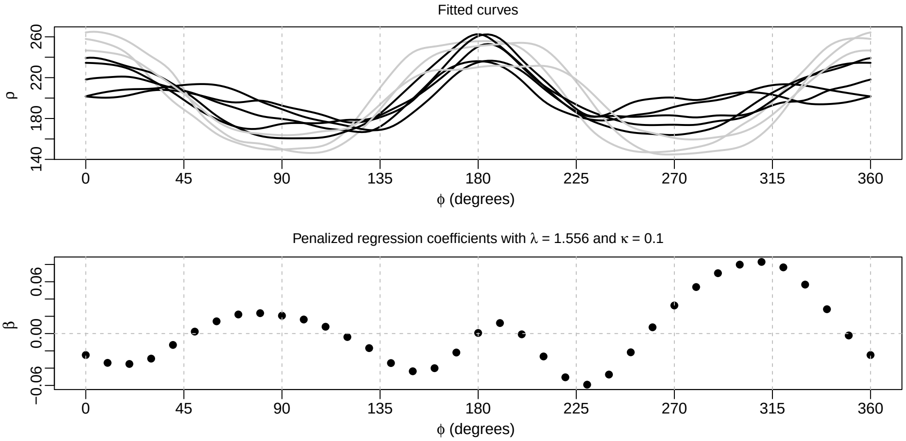


Figure 6.4: The upper panel shows all fitted curves of the head shape data. The lower panel presents the estimated regression coefficients.

more sophisticated algorithms are available, these two methods are easily programmed and fitted into the analysis. We do not pursue an elaborate comparison amongst methods, but rather show two additional methods as a first indication of relative performance of PSR.

Skull data

A horizontal cross-section of the skull of a normally developing child has a more or less elliptic shape, not too eccentric. Abnormal head shapes can occur due to premature fusion of one or more cranial sutures and is called Craniosynostosis. Depending on which sutures fuse, different deformities occur. Making the right diagnosis is crucial to decide on possible treatments. Treatment can have cosmetic purposes in order to improve the head shape, or it is done to reduce intracranial pressure.

Here we analyze a set of skull outlines to discriminate two subtypes of deformities. In the present literature the deformation is mostly described by relatively crude ratios, often defined using measures like the width and height of the skull. The outlines are one slice, coming from a CT-scan of the skull. The slices are taken at a fixed height and angle, and are determined using three landmarks (at the nose and both ears). Although there are various issues related to the analysis of these data, here we only focus on prediction of class membership, given the set of (x, y) coordinates. The images are digitized to about 1500 coordinate pairs, varying from sample to sample. With only eight samples available we only train the model, and evaluate the differences. In the upper panel of Figure 6.4, the smoothed fitted curves of all eight samples are plotted. Due to the large differences, a clear distinction between the two groups can be made. The lower panel of the same figure shows the estimated regression coefficients, and can be related to the regions where the two groups differ most. The general interpretation of the coefficients is the same as for a usual logistic regression.

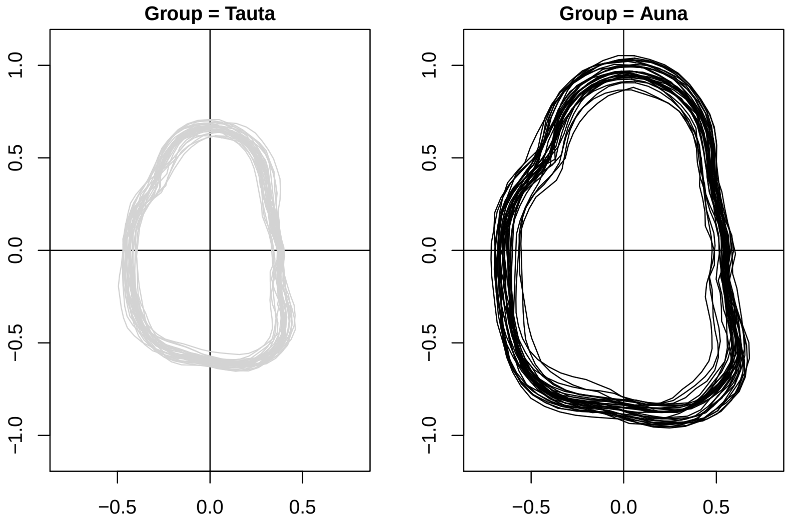


Figure 6.5: Two panels showing the molar data divided into the two classes. Because the focus lies on discrimination based on shape, the differences in size between the two groups are removed before classification.

Fossil rodent molar data

A second example comes from Claude (2008), where the outlines of fossil rodent molars are used to classify two species. The data consists of 60 samples, divided into a class of 31 ”Auna” samples, and a class of 29 ”Tauta” samples. All shapes are represented by 64 pairs of cartesian coordinates, both classes are depicted in the panels of Figure 6.5. Aside from the differences in size, it seems difficult to observe large differences between both classes. As a next step the data is summarized using a basis of 36 B-splines, including a very small penalty.

We repeatedly train the algorithm on 40 samples, randomly drawn from the data set. Figure 6.6 shows a typical AIC profile. The performance of the model is subsequently assessed using the remaining 20 samples. The average percentage of errors is calculated after 25 hold outs. The same setup is used for the PLR and PCLR classifiers, see Table 6.1 for the results. It turns out that the differences between both classes are substantial, since for all methods the error percentage is low.

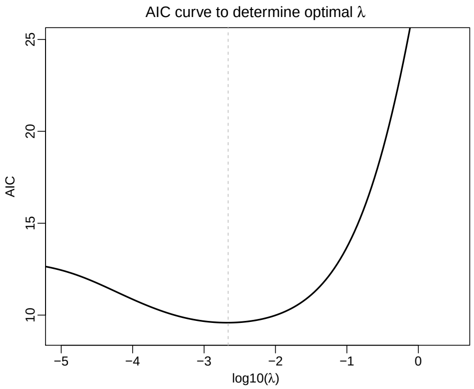


Figure 6.6: The AIC curve for the molar data, as a function of the tuning parameter λ.

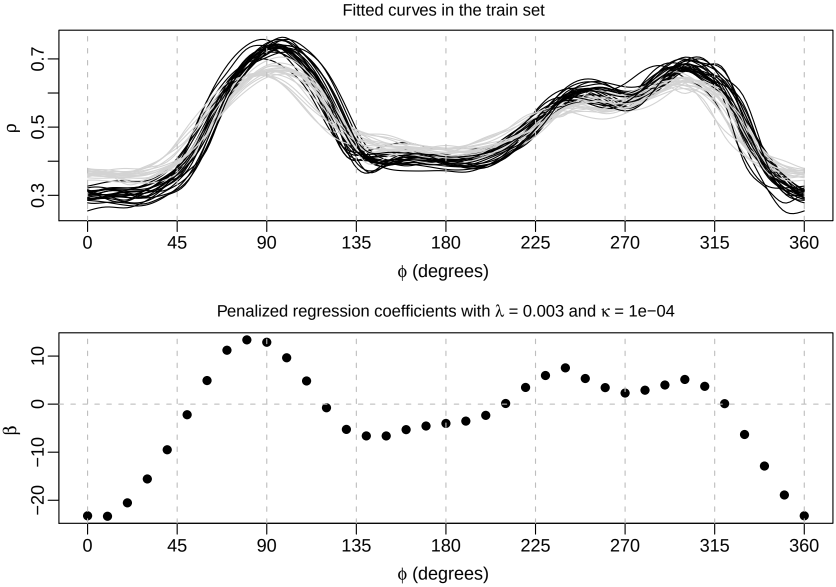


Figure 6.7: In the upper panel we see the preprocessed training samples of the molar data. The estimated coefficients from the logistic regression model are presented in the lower panel. The coefficients are calculated wrt. the Auna samples.

Figure 6.7 shows the results for one run of training and testing the data. In the upper panel of the figure we can see the fitted curves of the training samples. The estimated coefficients from the logistic regression model are presented in the lower panel. The coefficients are calculated with respect to the Auna samples.

From Figure 6.7 one can deduce the most discriminating regions along the outline. It is also useful to plot these coefficients in cartesian coordinates, as shown in Figure 6.8. In the inner circle of the plot the mean shapes of both classes are plotted. On the outside the coefficients are presented. To make the plot better interpretable, the coefficients are connected to the null line. From this figure we observe that the largest differences between the molars are around 360°, where the Tauta samples seem to be more flat.

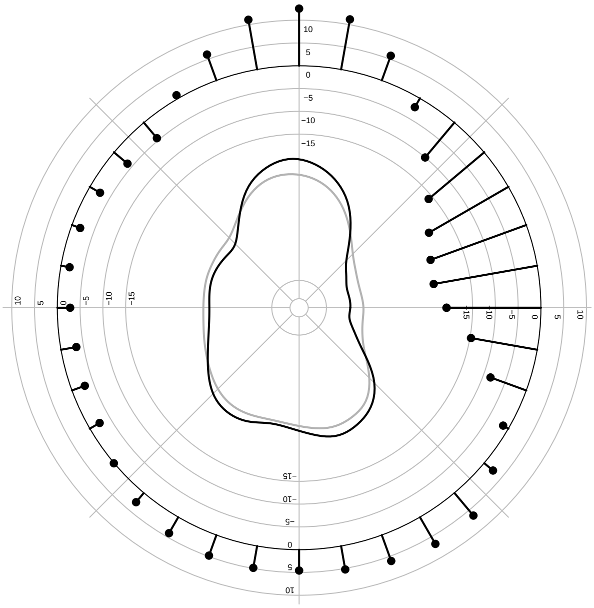


Figure 6.8: The regression coefficients plotted together with the mean curve per class.

Diatom data

As a third example we classify outlines of diatoms, very small single-celled algae. In this case we tested one group against a second group consisting of two different classes. The data comes from a larger dataset of in total 781 samples from 37 species, described in Jalba et al. (2005). We selected three relatively similar shaped species; in Figure 6.9 one sample from each of the three different groups are displayed. The Navicula reinhardtii are denoted as group one, the Navicula menisculus and Parlibellus delognei are combined in the second group. All the shapes are standardized before classification. After summarizing all data using 36 B-splines, the model is trained on 40 samples and tested on the 29 remaining samples. Here, we also used a repeated hold-out scheme with 25 repetitions, the results are displayed in Table 6.1. The performance of the three classifiers differs more

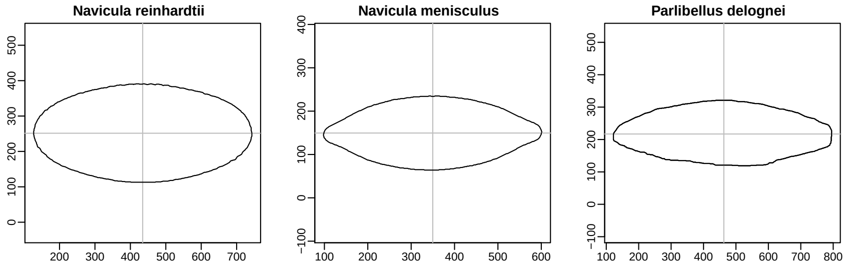


Figure 6.9: One example of each of the three groups of diatoms, plotted in rectangular coordinates. The two classes on the right are merged into one group.

Table 6.1: Predictive performance of the three methods for the molar and diatom data.

	Method		
Data	PLR	PCLR	PSR
Molars	1.6 %	3.0 %	0.8 %
Diatoms	17.3 %	7.8 %	1.4 %

in comparison to the previous example. Especially the PLR algorithm seems unable to detect the differences between the samples.

6.4 Discussion

We presented a framework for preprocessing and classification of outlines using polar coordinates. Preprocessing steps like rotation, normalization and noise removal are easily performed on the set of curves summarized on a P-spline basis. Moreover, using P-splines makes the signals of equal length while unsupported regions can be corrected using a difference penalty. Classification is performed using penalized signal regression. One of the benefits of PSR is that it fits into the framework of generalized linear models (GLM), and makes it a familiar tool for many possible users. Crucial is the role of the penalty, as it prevents overfitting of the data and makes the problem well-posed. In addition, the coefficients are easily interpretable and can be presented in combination with the original shapes.

One of the assumptions we made in the preprocessing stage, is that the data points are sampled on an approximately equally spaced grid, which makes it easy to calculate the origin. In cases with unsupported regions or irregularly sampled pseudo-landmarks, the center of mass will not be suitable as an indicator for the origin. This problem can be solved by, first using the initial origin to calculate polar coordinates. And subsequently summarizing the data on a B-spline basis, and translate the data back to rectangular coordinates using a regular grid. Now a proper origin can be calculated, which is used to determine a definite set of polar coordinates.

The model as presented here leaves room for a wide range of extensions. Above we only discussed a binary outcome variable. Because the model is a member of the class of generalized linear models, we are not limited to binomial data. Currently we do not have data for alternative outcomes, but one could think of applications in industry like the wear of machine parts related to the number of fault events. In addition, the model can be augmented with additional variables, next to a description of the shape.

Further extensions can be sought in the type of modelled shapes. A first example is the analysis of 3d shapes using the cylindrical or spherical coordinate system. Furthermore, the shapes as discussed in this article have a simple general structure, and thus can be analyzed as curves. Although various problems can be approached using the described methodology, many applications show more complicated shapes. They are not star-convex or the defined origin is not a star center (or even lies outside the shape). In these cases it is more difficult to use standard polar coordinates to describe the data. In other situations not a single, but a set of shapes per sample are of concern. In these cases it is an option to use a (normalized) distance along the outline, instead of the angle ϕ.

7 Exploring network structure using penalties and priors

Abstract. Estimating a high-dimensional network of features, has become a popular topic in various disciplines. In genetics, the aim is to establish functional modules of genetic features. Network estimation with a large number of variables is however a non-trivial task. Estimation procedures can benefit from the use of external resources, available in databases connected to the web. In this chapter we investigate the performance of three shrinkage estimators and propose a method to incorporate prior knowledge into the estimation process in order to retrieve an underpinned network of features. The method is based upon the weighted Lasso. Use of the method is illustrated with two applications, one relying on text mining and the second on a prior derived from literature.

7.1 Introduction

Estimating a high-dimensional network of features has become a popular topic in various disciplines. Examples include brain imaging in neurology (Smith et al., 2011; Lee et al., 2011), and portfolio selection in economics (Ledoit and Wolf, 2003). Networks also play a prominent role in genetics, where features like genes and proteins are considered as parts of functional modules or networks (see e.g. Azuaje, 2010; Kitano, 2002; Junker and Schreiber, 2008).

Determining associations in a high-dimensional dataset is however a non-trivial task, and becomes even more difficult when the number of features exceeds the number of observations. A few examples of solutions in literature are support vector machines (Bleakley et al., 2007), Bayesian networks (Friedman et al., 2000) and methods based on information theory (Margolin et al., 2006). Here we consider network estimation as a covariance selection problem (Dempster, 1972). By setting elements in the inverse covariance matrix to zero we search for a network structure that is sparse and at the same time fits the data well. Because of the large number of variables compared to the number of observations, the covariance matrix will be often ill-conditioned or singular. Shrinkage estimators are a popular remedy in these cases; they make the model estimable by setting relations in the covariance matrix to zero.

In most cases the estimation procedure only relies on the data at hand. In the genetical setting there is the advantage that, on top of the observed data, often additional knowledge is available. These external resources can be used to guide the model estimation. The prior is often a collection of covariances, probabilities or weights which expresses the likelihood of a direct relation between two features. This prior can be used to enhance network estimation, or test alternative network structures against each other. To build a prior one can use online sources like KEGG (Kanehisa and Goto, 2000) and GO (Ashburner et al., 2000), additional datasets, or curated publications.

In the next section we start with a general introduction to the problem of network estimation. Thereafter we investigate network estimation in two directions. First we look at three shrinkage estimators and their ability to reproduce simulated networks. Next we discuss a method to combine prior knowledge with the available the data. In addition, two applications are discussed.

7.2 Network estimation

Let $X = (x_1, x_2, ..., x_n)^T$ be the $n \times p$ data matrix having a multivariate normal distribution with mean μ and covariance matrix Σ. The standard estimate of the covariance matrix is

$$\hat{\Sigma} = \frac{1}{n-1} X'X. \tag{7.1}$$

If we are interested in conditional relations we have to identify the non-zero elements in the precision matrix $\hat{\Phi} = \hat{\Sigma}^{-1}$. Given that $p < n$ and assuming that we have a positive definite covariance matrix $\hat{\Sigma}$, we can simply take its inverse and from this to calculate the partial correlations:

$$\hat{\rho}_{ij} = \frac{-\hat{\phi}_{ij}}{\sqrt{\hat{\phi}_{ii}\hat{\phi}_{jj}}}. \tag{7.2}$$

The set of partial correlations can be translated into an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = 1, ..., p$ being the set of nodes and $\mathcal{E} = \{e_{ij}\}_{1 \leq i < j \leq p}$ the set of edges. A zero in the inverse covariance matrix corresponds to conditional independence between two nodes and is subsequently described in the edge set:

$$\phi_{ij} = 0 \Leftrightarrow x_i \perp x_j | x_{-j} \Leftrightarrow e_{ij} = 0. \tag{7.3}$$

The graph or network $\mathcal{G}$ represents the structure of the set of genes in a graphical way (see Figure 7.1 in section 8.3 for an example). Estimating the graph structure $\mathcal{E}$ is called a covariance selection problem after Dempster (1972).

The data is most often high-dimensional with $p \gg n$. As a result the observed covariance matrix will be singular and its inverse cannot be calculated. A quick solution is to use the generalized inverse, but that will most often return unsatisfactory results (Schäfer and Strimmer, 2005). Recently, various alternatives have been proposed. A relatively simple solution is to use so-called local models, which will be discussed in the next section, followed by a discussion on shrinkage estimators.

Local solutions

Using marginal instead of conditional associations avoids the inversion problem. For instance Stuart et al. (2003) and Steuer et al. (2003) use the Pearson correlation coefficient between subsets of variables in order to derive a network structure. To induce sparseness one can set a threshold and remove all relations below the required level. Simple correlations cannot distinguish between direct and indirect effects. To overcome this problem higher order correlations can be used (see e.g. Fuente et al., 2004) . The modelling now takes place on small sub-networks or 'neighbourhoods' consisting of three genes. In this case the relationship between gene i and gene j is investigated conditioning on a third gene z, which gives $x_i \perp x_j | x_z$. In the next step all sub networks are combined to form a large network. The advantage of this procedure is that it can estimate conditional relations, while singularity is not a problem due to the fact that only three variables are estimated each time. This makes it suitable for applications where there are more variables than observations. A downside of this method is that it only conditions on one (or a few if bigger neighbourhoods are considered) variable(s), whereas it could be very well the case that the relations are more complex than this. Reverter and Chan (2008) and Watson-Haigh et al. (2010) combine the ideas of higher order partial correlation with an information theory approach.

Shrinkage estimation

Full conditional models are computationally more complex, but can be obtained using a shrinkage estimator. Various penalized procedures are inspired by the Lasso, introduced by Tibshirani (1996). The Lasso adds a linear penalty function to the ordinary least squares cost function and is defined:

$$\min_{\beta} ||y - X\beta||_2^2 + \kappa \sum_{i=1}^{p} |\beta_i|, \tag{7.4}$$

the first part of the equation being the usual least squares part, while the second part expresses the penalty on the estimates. The amount of shrinkage is regulated using the tuning parameter κ. In the standard Lasso the coefficients are shrunken towards each other and to zero, yielding a sparse model. Adaptations to $p \gg n$ data are presented by e.g. Ambroise et al. (2009), Yuan and Lin (2007) and Friedman et al. (2008).

Here we use the 'neighbourhood selection' approach by Meinshausen and Buhlmann (2006). To determine the set of adjacent nodes for all nodes in the graph, we fit p regression models with in each model one node being the dependent variable and all others being the predictors. This gives us p linear regression models:

$$x_i = \sum_{j \neq i} \beta_j x_j + \varepsilon, \qquad \text{for i = 1, ..., } p, \tag{7.5}$$

with estimates $\beta^{(i)} = (\beta_1^{(i)}, ..., \beta_{i-1}^{(i)}, \beta_{i+1}^{(i)}, ..., \beta_p^{(i)})$. From the estimated regression coefficients we can calculate the partial correlations as:

$$\hat{\rho}_{ij} = \text{sign}(\hat{\beta}_j^{(i)}) \sqrt{\hat{\beta}_j^{(i)} \hat{\beta}_i^{(j)}}. \tag{7.6}$$

Because in the setting where $p \gg n$ the sign of $\hat{\beta}_j^{(i)}$ can differ from the sign of $\hat{\beta}_i^{(j)}$, we define that (7.6) only holds if $\text{sign}(\hat{\beta}_j^{(i)}) = \text{sign}(\hat{\beta}_i^{(j)})$, and is otherwise set to zero (Krämer et al., 2009). For sparseness Meinshausen and Buhlmann (2006) exclusively rely on the Lasso penalty. However, it is often recognized that the L_1 penalty leaves too many edges in the network, resulting in many false positives. A penalty claimed to be more successful in the present setting is the adaptive Lasso (Zou, 2006). For the neighbourhood selection problem this results in

$$\min_{\beta} ||x^i - X^{-i}\beta||_2^2 + \kappa \sum_{j \neq i} w_j |\beta_j|, \tag{7.7}$$

with X^{-i} the data matrix, exclusive the i-th column. The weights w adjust the influence of the penalty per coefficient. Weights can be based upon e.g. the magnitude of initial (shrunken) estimates (Krämer et al., 2009) or the variance of the estimate (Shimamura et al., 2007). Very sparse solutions are generally obtained by penalizing the number of non-zero coefficients, effectively implementing an L_0 norm penalty,

$$\min_{\beta} ||x^i - X^{-i}\beta||_2^2 + \kappa \sum_{j \neq i} I(\beta_j \neq 0). \tag{7.8}$$

One problem arising from this penalty is the non-convexity of the objective function. To optimize the objective function we use a smooth approximation, discussed in case of the L_1 penalty by e.g. Osborne et al. (2000): $|\beta_j| = \beta_j^2/|\beta_j|$. Using $\beta_j \approx \sqrt{\beta_j^2 + c^2}$, with c being a small constant, results in:

$$\beta_{new} = (X'X + \lambda V^{-1})^{-1} X'y, \tag{7.9}$$

with $V = \text{diag}(\sqrt{\beta^2 + c^2})$ and iterating gives us the optimal β. In case of the L_0 penalty V is replaced with: $U = \text{diag}(\beta^2 + c^2)$.

7.3 Shrinkage using additional knowledge

There are a number of methods to combine shrinkage estimators with additional information. A smoothing approach in the form of the fused Lasso is proposed by Tibshirani et al.

(2005), it encourages (hypothesised) adjacent coefficients to be similar to each other. The objective function, modelling an outcome y, includes both the L_1 norm and smoother,

$$\min_{\beta} ||y - X\beta||_2^2 + \kappa_1 \sum_{i=1}^{p} |\beta_i| + \kappa_2 \sum_{i=2} |\beta_i - \beta_{i-1}|. \tag{7.10}$$

A similar model inspired on the fused Lasso, is proposed by Li and Li (2008). They introduce so-called network-constrained regularization. The model includes an L_1 penalty, and in addition smooths coefficients of adjacent nodes in a network to each other. Adjacency of nodes is stored in a standardized Laplacian matrix.

In this work the only concern is to estimate the network. Possible additional knowledge can be used to compose a prior structure and is subsequently imposed on the data. Thus by increasing the penalty parameter, one shrinks towards the hypothesized structure. This idea is similar to the adaptive Lasso as shown in equation (7.7). For the neighbourhood selection problem w corresponds to one line taken from matrix W, indicating prior relations in the network. We adopt the following model:

$$\min_{\beta} ||x^i - X^{-i}\beta||_2^2 + \kappa_1 \sum_{j \neq i} \beta_j^2 + \kappa_2 \sum_{j \neq i} (1 - w_j)|\beta_j|. \tag{7.11}$$

In addition to the network-constrained penalty regulated by κ_2, an L_2 penalty is included. It is not optimized but added to the model in order to make it estimable, irrespective of the (amount of) weight assigned to the prior.

Structuring the weight matrix

Both the structure and the values of the elements of the matrix W follow from the prior information available. Depending on the source, one could compose a fully specified matrix with edge-specific weights, or create a more global structure. Examples of prior information are:

- Pathway information (e.g. MsigDB)
- Curated networks and functional data (e.g. KEGG, Reactome, BioCarta)
- Additional (similar) datasets
- Text mining

Pathway databases tell us information on which genes are present in certain pathways. Information about specific gene-gene interactions can be found in functional databases like KEGG (Kanehisa and Goto, 2000). Additional datasets or publications may serve as a first source, providing a list of interesting genes or even interactions among them. Text mining tools will provide co-occurrences of certain features in the literature. Especially in the last case, the weights in W will follow naturally from the prior. In other cases weights might be more subjective. A special case is the binary weight matrix with entries

$$[W]_{ij} = \begin{cases} 1, & \text{if node } i \text{ is adjacent to node } j, \\ 0, & \text{otherwise.} \end{cases} \tag{7.12}$$

Using a binary weight matrix and choosing a very large value for κ_1, the prior will be imposed entirely on the data.

In addition to the weights, the structure of W has to be defined. Two examples are:

- A block diagonal matrix. In this case we augment the standard model using gene set or pathway information. No specific functional information on the gene level is available. However, assuming that relations within the pathways are more likely than between them, this results in a block diagonal weight matrix, with each block representing a pathway or set of genes.

$$\begin{bmatrix} \Box & \Box & \Box & & \\ \Box & \Box & \Box & & \\ \Box & \Box & \Box & & \\ & & & \Box & \Box \\ & & & \Box & \Box \end{bmatrix} \tag{7.13}$$

- A fully specified network structure derived from e.g. functional databases, the literature (using text mining tools), or additional data. The example shows a matrix in which one gene serves as a hub to other genes and could be extracted from a functional database.

$$\begin{bmatrix} \Box & \Box & \Box & \Box & \Box \\ \Box & \Box & & & \\ \Box & & \Box & & \\ \Box & & & \Box & \\ \Box & & & & \Box \end{bmatrix} \tag{7.14}$$

7.4 Simulations

In this section a series of simulations are performed, assessing performance of different penalties with respect to network recovery. Tested are the L_1 norm penalty and the L_0 norm penalty, implemented using the smooth approximation discussed above. In addition, the adaptive Lasso is included, for which we rely on the *parcor* package available in *R* (R Development Core Team, 2011), and discussed in Krämer et al. (2009).

As a first step, a scale-free network (see e.g. Barabasi and Bonabeau, 2003) is simulated, of which an example is plotted in Figure 7.1. A sparse precision matrix Ω is created using the connection matrix based upon the simulated network. Non-zero coefficients in Ω correspond with ones in the connection matrix. The values of the non-zero off-diagonal relations in Ω are sampled from a uniform distribution, with boundaries as specified in Table 7.1. The actual data are sampled from $\mathcal{N}(0, \Omega^{-1})$.

Table 7.1: Characteristics of the different simulated models

	p	correlation
Sim I	50	[-0.7,-0.4]∪[0.4,0.7]
Sim II	50	[-1,-0.5]∪[0.5,1]
Sim III	50	[-1,-0.7]∪[0.7,1]

Performance is measured using discrete loss and mean squared error (MSE). In the case of discrete loss, the estimated coefficients are translated into a zero or a one. A zero corresponds with conditional independence and a one with a non-zero relation in the precision matrix. This estimated connection matrix can be compared with the truth using measures of precision and recall. Precision is defined as

$$\mathrm{P} = \frac{TP}{TP + FP}, \tag{7.15}$$

with TP and FP being the true positives and false positives, respectively. The precision is the proportion of true positives among the total number of relations that were found. This ignores however, the ability of the algorithm to detect relevant edges, which is instead measured by the recall score:

$$\mathrm{R} = \frac{TP}{TP + FN}, \tag{7.16}$$

in which FN is defined as the number of false negatives. The F-score summarizes both measures in one estimate,

$$F = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \tag{7.17}$$

To determine the performance of the penalties with respect to the true precision matrix the mean squared error is used. In Table 7.2 the results of the simulations are shown.

Changing the correlation structure does not result in large changes in the the performance of the methods. In general we see that the L_1 penalty returns networks that are too dense. The L_0 and the adaptive Lasso have about the same performance with respect to the precision. However, the adaptive Lasso performs better on the recall score, resulting in a better F-score.

7.5 Applications

In this section we show two applications. In the first example, we generate a prior using text mining algorithms. In the second case the prior information is less specific. From

Table 7.2: Average performance of the estimators per scenario (mean and standard deviation).

		Precision	Recall	F-score	MSE
Sim I					
	L_0	0.80 (0.10)	0.36 (0.05)	0.49 (0.06)	1.1×10^{-3}
	L_1	0.40 (0.05)	0.64 (0.07)	0.49 (0.05)	9.5×10^{-4}
	aL_1	0.77 (0.08)	0.44 (0.07)	0.56 (0.07)	9.4×10^{-4}
Sim II					
	L_0	0.79 (0.09)	0.35 (0.06)	0.48 (0.06)	1.3×10^{-3}
	L_1	0.44 (0.05)	0.67 (0.06)	0.53 (0.04)	9.1×10^{-4}
	aL_1	0.78 (0.08)	0.46 (0.05)	0.57 (0.05)	1.0×10^{-3}
Sim III					
	L_0	0.79 (0.09)	0.32 (0.05)	0.45 (0.06)	1.1×10^{-3}
	L_1	0.41 (0.07)	0.58 (0.05)	0.48 (0.05)	8.5×10^{-4}
	aL_1	0.80 (0.09)	0.40 (0.05)	0.53 (0.04)	8.4×10^{-4}

a previous publication we derive genesets which are supposed to form functional modules. The prior assumes edges within the two genesets, but no relations between the sets. Evidence for either support or rejection of the prior by the data is gathered using a randomization test.

External data prior

In this section we test whether functional networks found in human gliomas (Bredel et al., 2005) are reproducible in a similar type of dataset. The data we use concerns 292 microarray expression samples analysed in Gravendeel et al. (2009). Bredel et al. (2005) investigated the presence of networks and found four distinctive coexpression clusters. Here we investigate two of them, consisting of 65 and 49 genes. Because both studies report a different type of gene identifier, some translations need to take place. To translate the clone identifiers used in Bredel et al. (2005) into Entrez identifiers we use the MADGene tool (Baron et al., 2011). Not all identifiers were found and after conversion we have a dataset consisting of 81 genes (one pathway consisting of 42 genes and the other 39). Because we do not have edge specific information, a block diagonal prior is constructed. The values in the weight matrix are set to 0.5, if an edge is assumed and zero otherwise.

The estimated binary connection matrix is depicted in Figure 7.2. To determine whether the prior structure is likely, given the data, we use a permutation test. In addition to the estimated network, using data and a prior, we compute a number of permuted priors. Both columns and rows and permuted, meaning that a random prior structure is created. Each prior is used to estimate a network and the tuning parameter is optimised using cross-validation. Subsequently, for each randomization an average tuning parameter $\bar{\lambda}$ is calcu-

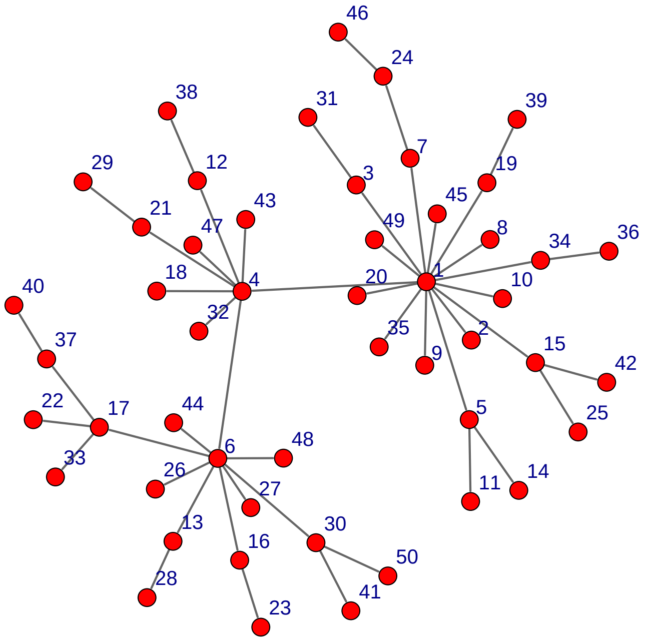


Figure 7.1: A graph, showing the scale-free topology of the simulated network.

lated. We assume that if the prior based upon the findings in Bredel et al. (2005) fits the data well, much shrinkage is allowed and if not, no or only little shrinkage is performed. The amount of shrinkage is compared to the random alternatives and the distribution of average shrinkage parameters is plotted Figure 7.3. The result connected to the observed prior is depicted as a grey, square, all randomizations as black dots. Note that the proposed prior does allow more shrinkage compared to a random alternative, suggesting the prior does fit much better.

Text mining

Various strategies are available for using text mining tools to compose a prior network. A very prominent one uses the vector space model and is for instance used in Faro et al. (2012). In this approach, genes are represented by a set of relevant documents, the vector space model is used to establish a measure of distance between the genes involved (see e.g. Glenisson et al., 2004; Liu, 2007). A second, less often used option is to derive a direct co-occurrence prior (Jenssen et al., 2001). PubMed is queried for publications relevant to our data, the retrieved documents are scanned for the presence of gene names. The

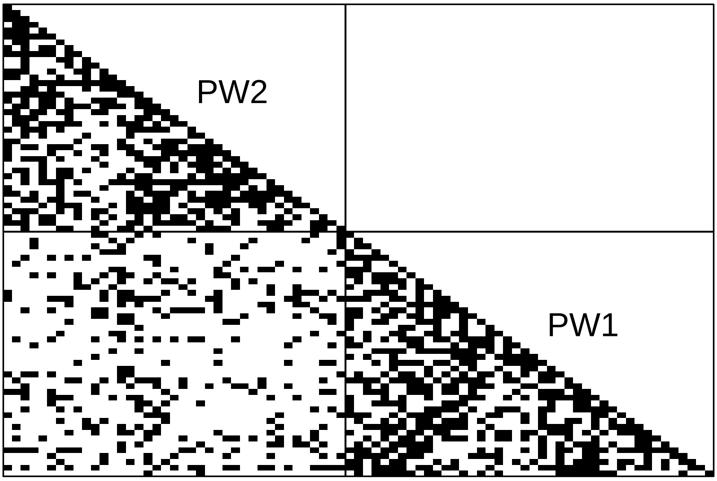


Figure 7.2: The posterior (binary) model presented in a graph. From the image we observe that, according to the model, associations within the pathways are more likely than between the pathways.

assumption underlying the method is that genes that are mentioned in the same article have a biological relationship of some type. The co-occurrence of the genes within the documents of the total document set are represented in a term-document incidence matrix Γ. This is a binary matrix, with for every element a one if a certain gene t is cited at least once in document d, and a zero otherwise. To determine the closeness of any pair of genes involved we take the average score for gene j of the document vectors in which gene i appears:

$$w_{ij}^* = \frac{1}{N_i} \sum_{d=1}^{N_i} \gamma_{di} \gamma_{dj} \mathrm{idf}_j, \tag{7.18}$$

with N being the size of the total number of documents and N_i the size of the set in which gene i is present. The idf denotes the inverse document frequency (see e.g. Manning et al., 2008), a weighting factor for the number of documents in which gene j appears:

$$\mathrm{idf} = \log_2 \frac{N}{N_j}. \tag{7.19}$$

Subsequently the weights are normalized and determines the weight matrix:

$$W = \left\{ \frac{w_{ij}^*}{\max(w_{ij}^*)} \right\}. \tag{7.20}$$

The method is applied to the 292 microarray gene expression samples, as introduced above. A network is built using penalized regression in combination with a prior based

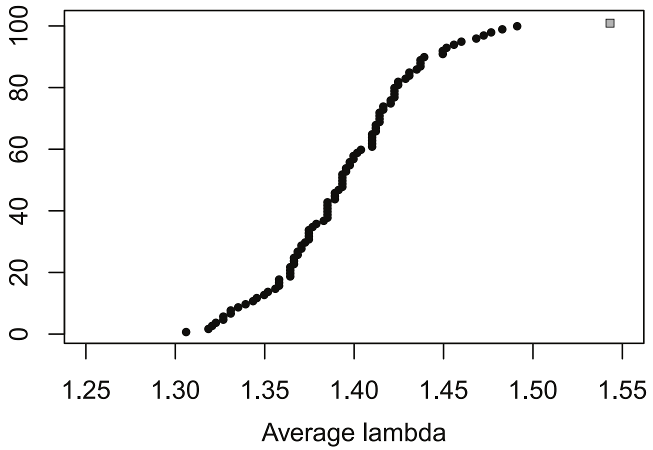


Figure 7.3: The result of the randomization test, with in red the modelled prior. The proposed prior does allow more shrinkage compared to a random alternative.

on a literature search through PubMed. In order to generate a (very) small but insightful example we restricted the number of documents to the first 2000 hits. 212 unique genes were derived from the document set, resulting in 204 hypothesized edges. A model is estimated using the framework discussed in the previous section and model selection is performed using cross-validation. The posterior network is still very dense, while only a limited number of edges are confirmed. Forty of the hypothesized edges are confirmed by the estimated model, 164 are not found, the number of new connections (with $\phi > 0.05$) equals 1802. When we use an adaptive Lasso on the same dataset, without any prior information, a similar model appears. The adaptive lasso confirms 25 edges, rejects 179 and finds 1239 new edges. The surprisingly large number of edges, even in the case of the adaptive Lasso, raises questions of various kinds. Setting a threshold on the minimum strength of associations could unveil clusters, but at the same time ignores available information. It might be that inclusion of more genes creates a better picture of the conditional relations in the network and even removing a number of the now present paths.

7.6 Conclusions

Estimating a network from a high-dimensional feature space has become an important task in various disciplines. In this study we focused on network estimation in genetics. First studied are present alternatives to estimate a network. Methods tested are the L_1 norm penalty, the adaptive Lasso and the L_0 norm penalty. The best performance is obtained using the adaptive Lasso, whereby the resulting graphs are sparser and the penalty controls better the number of false positives, and the model returns a low squared error.

In the next step we investigated the use of prior information in the process of network modelling. Weights, derived from external sources, form a prior connection matrix which is used in the estimation step in a manner similar to the weighted Lasso. Two different applications are provided, showing the potential of the methodology. In both settings the conditions are rather mild, meaning that the number of samples is larger than the number of features and $\hat{\Sigma}$ can be inverted. A sparse solution is, not obtained however by taking the inverse and thus requires additional constrains. The first application seems promising since the prior seems to fit to the data while random alternatives are worse. For a good judgement with respect to the results more expert knowledge from the field is required. In the case of the text mining application the situation is different. Because a relatively small number of edges are hypothesized, we are searching for a very sparse result. But the estimated network is very dense, even if no prior information is used. It would be very interesting to see what happens with the estimated connections when more genes are added to the dataset. In addition we should mention that for text mining much more sophisticated tools are available, likely leading towards a more informed prior.

Acknowledgements

The authors would like to acknowledge support from CTMM.

Pseudo-Bayes smoothing of tables with very low counts 8

Abstract. Sampling zeros or small counts are a nuisance in the analysis of contingency tables. Several strategies exist to make them less harmful, by filling in non-zero numbers, a process commonly called smoothing. We propose smoothing by a pseudo-Bayes procedure with appropriate priors. The procedure can be applied to any contingency table, but here we limit ourselves to meta-analysis of clinical trials. To optimize the weight of the prior we use a modification of AIC. Performance is evaluated with simulations and two real datasets are analyzed.

8.1 Introduction

Sampling zeros or small counts are a nuisance in the analysis of contingency tables. Several strategies exist to make them less harmful, by filling in non-zero numbers, a process commonly called smoothing. When the data have a natural order, one can borrow information from neighbours. In the setting of multinomial data Fienberg and Holland (1973), and later Simonoff (1996) (and references therein) and Burman (2004), proposed solutions of various kinds. In this chapter we propose a quite general pseudo-Bayes procedure, based on the ideas from Fienberg and Holland (1973) and earlier work by one of the authors (Eilers, 1996). It can be applied to any contingency table, but here we limit ourselves to meta-analysis of clinical trials. Examples of sparse data in clinical trials are numerous and can be found in e.g. Bradburn et al. (2007); Friedrich et al. (2007); Kuss and Gromann (2007) and Sweeting et al. (2004). With this type of data, zeros and small counts result in estimates of log-odds which are unreliable or do not exist, and thus are problematic for generating L'Abbé plots, funnel plots and forest plots. In the literature, various solutions to this problem have been proposed:

Fixed corrections. Add a fixed number to each count. A popular choice is 0.5, which is based on the continuity correction of Yates (1934). Laplace's "rule of succession",

This chapter is based on: de Rooi J.J., Frank L.E. and Eilers P.H.C. Pseudo-Bayes smoothing of tables with very low counts, applied to meta-analysis of clinical trials. *Submitted.*

which is based on a Bayesian argument and adds 1 to the observed numbers of failures and successes, can also be seen as a smoothing device. The effect of this correction on the estimated odds ratios per trial depends on the ratio of the group sizes. An elaborate study on the effects of applying these continuity corrections, in combination with various pooling methods for meta-analysis, is presented by Bradburn et al. (2007). A thorough discussion, more specifically on continuity corrections, can be found in Sweeting et al. (2004).

Empirical corrections. Adding constants does not use any information contained in the data. Sweeting et al. (2004) proposed two more advanced corrections, specifically for meta-analysis of clinical trials. The first method determines a specific correction for both arms of the study: $c_t = \frac{N_t}{N}$ and $c_c = \frac{N_c}{N}$, with c_t and c_c being the correction for respectively the treatment and control arm. The sum of increments over all cells in a trial is $2(N_t + N_c)/N = 2$. Notice that this notation is described in Rücker et al. (2009). The second method is the empirical continuity correction, and is based on a pooled estimate $\hat{\Omega}$ calculated over all trials without observed zeros. The derived odds ratio is used to calculate a correction for the studies with observed zeros. If we set the restriction that $c_t + c_c = 1$, the correction can be calculated as follows: $c_t \approx \frac{\hat{\Omega}}{R+\hat{\Omega}}$ and $c_c \approx \frac{R}{R+\hat{\Omega}}$, with $R = \frac{N_c}{N_t}$ the group size ratio. This correction is also applied on all cells of a trial which brings the sum of all corrections at 2.

Bayesian models. Various types of Bayesian approaches have been proposed. Empirical Bayesian estimators are discussed in for example Hedges and Olkin (1985), several references to similar methods can be found in Sutton and Abrams (2001). A fully Bayesian analysis is also possible, but is far from trivial. We do not review the literature, but point to the paper by Paul et al. (2010).

Generalized linear mixed models. A generalized linear mixed model with a logistic link is an advanced possibility. With many low counts or zeros this is not trivial. We will not review the literature, but refer to the recent paper by Stijnen et al. (2010)

Pooled estimates. It is also possible to do nothing, and just calculate a joint estimator using a procedure that is able to deal with zeros. One example is the Peto method, which requires at least one event per trial to estimate study-specific odds ratios and at least one event in the meta-study to estimate the pooled OR. A second option is the Mantel-Haenszel method which needs at least one event in each cell to estimate an OR on trial level, and at least one event in each arm overall, to estimate the pooled OR. A third option is to use the arcsine transform, as advocated by Rücker et al. (2009).

As said, in this chapter we explore pseudo-Bayes estimation. It has been described and applied in a paper by Fienberg and Holland (1973), and later in the classic book by Bishop et al. (1975, 2007). The pseudo-Bayes approach chooses a prior that consists of a set of numbers, one for each cell of a table of counts. The posterior estimates are weighted averages of the prior and the empirical probabilities.

Most often the prior is based upon the data, and can be derived using different methods. A simple prior is obtained by computing probabilities from the aggregated counts over all trials, either separately for the treatment and control arm, or combined. However, simply aggregating the individual studies can cause problems which are formulated in Simpson's paradox. To avoid this problem, alternatives are discussed as well.

The success of the pseudo-Bayes procedure depends on the proper choice of the weight of the prior. Bishop *et al.* presented a rather abstract recipe, based on a general asymptotic risk argument. They did not use the observed data to optimize the prior weight. We propose to use AIC (Akaike's Information Criterion) (Akaike, 1974), which combines the log-likelihood and a measure of model complexity, called the effective dimension (ED). In our experience, the influence of the latter was too small. When we used a more or less standard definition of ED. We propose a simple heuristic modification, inspired by the work of Hurvich et al. (1998).

To evaluate our procedure, and compare it to other recipes, we performed simulations, and illustrate its use for two datasets from the literature.

Pseudo-Bayes estimation for meta-analysis leads to results that fall between the fixed model (one probability per arm) and the mixed model or a fully Bayesian analysis. The latter approaches introduce a "population distribution" of event probabilities, from which posterior point estimates are computed. In the pseudo-Bayes approach there are only posterior point estimates. In fact it is a linear shrinking procedure, because it computes weighted averages of prior and posterior probabilities.

We see our proposal as a data smoothing procedure, that results in corrected counts, which can be plugged into a meta-analysis. We do not advocate a fixed or mixed model. Our goal is simply to remove zeros and correct very low counts. The use of priors as a correction method in observational studies is discussed by Greenland (2006).

In the next section we present pseudo-Bayes estimation and the modified AIC. Our simulations are presented in Section three and application to real data in Section four. We conclude with a short discussion.

8.2 Pseudo-Bayes smoothing

The pseudo-Bayes principle

Consider an m by n matrix $X = [x_{ij}]$ of observed counts. Let $x_{i+} = \sum_j x_{ij}$ be the row sums, $x_{+j} = \sum_i x_{ij}$ be the column sums, and x_{++} be the sum of all counts in the matrix X. Let $\Lambda = [\lambda_{ij}]$ be an m by n matrix, the prior, with $\sum_i \sum_j \lambda_{ij} = 1$. Then, according to

Bishop et al. (1975) the pseudo-Bayes estimates x^*_{ij} are given by

$$x^*_{ij} = w\lambda_{ij}x_{++} + (1-w)x_{ij}, \tag{8.1}$$

where $0 \leq w \leq 1$. This is a linear shrinkage procedure where the smoothed estimate x^*_{ij} is a weighted combination of the observed data x_{ij} and the prior λ_{ij}. The amount of shrinkage is determined by w.

Choosing a prior

Our goal is to correct zeros or very low counts, so we expect the prior to have a strong influence. We have several choices:

Independent of the data. The most simple choice is a two-by-two table with equal probabilities ($\lambda_{ij} = 0.25$). Using this prior comes near to adding the constant $wx_{++}/4$ to each cell. It is not exactly the same because of the correction $(1-w)x_{ij}$, to get the same total counts.

Based on marginal counts. All individual tables are added and the probabilities are computed from the result. Two choices are available: do this separately for treatment and control arms, or do this for the two-by-two tables. The former case we call a split prior, the latter a joint prior.

Constant log-odds prior. It is well known that when the size (of the arms) of studies varies, the marginal table can lead to a wrong estimate of the odds ratio. This is known as non-collapsibility and it is related to Simpson's paradox. A discussion of this problem in the context of clinical trials can be found in e.g. Cates (2002), Altman and Deeks (2002) and Greenland (2010). A graphical explanation of the problem can be found in Rücker and Schumacher (2008). For this reason one should use the Mantel-Hänszel or Peto estimator. This suggests that we construct a separate prior for each study, based on its marginal counts and the overall odds ratio.

Let the marginal counts of table X be given, and let $\tilde{\Omega}$ be the desired odds ratio of a modified table $\tilde{X}$. After some algebra one finds that $\tilde{x}_{ij}$ should be equal to the positive root of the quadratic equation $ax^2 + bx + c = 0$, with $a = 1 - 1/\tilde{\Omega}$, $b = -(x_{1+} + x_{+1} + (x_{+2} - x_{1+})/\tilde{\Omega})$ and $c = x_{1+}x_{+1}$. Once $\tilde{x}_{ij}$ is known, the other elements follow directly from the margins of X. The prior counts can be translated into prior expectencies: $\Lambda = \tilde{X}/x_{++}$

A serious disadvantage of this prior is that it cannot be computed if the number of events in both arms is zero.

Hybrid priors. A possible solution for cases with zero events in both arms is a hybrid approach. First, for each study the numbers in treatment and control arms are added, to give a table with two cells. The pseudo-Bayes procedure is applied to these tables, leading to non-zero totals for the number of events per study. In as second round the constant log-odds prior can be computed for each study.

In this chapter we only study priors based on marginal counts and on an overall odds ratio. All proposed priors can be applied to all studies within a meta-analysis, irrespective of possible zero event cells. A second option is to smooth only the trials with zero events.

Optimizing the weights

To determine the optimal combination of prior and data, different methods have been proposed. Fienberg and Holland (1973) and later Bishop et al. (1975) relied on the asymptotic properties of a risk-function that minimizes the expected mean squared error. Eilers (1996) proposed a fast leave-one-out cross-validation procedure. Unfortunately, it was based on reducing each cell count in turn by 1 which will not work with zero counts. Instead, we propose to use AIC. We have to modify the effective model dimension to make this work.

AIC is defined as

$$\mathrm{AIC}_C = D(X;X^*) + 2ED, \tag{8.2}$$

where the $D(X;X^*)$ and ED the effective model dimension. In the case of counts and a 1 by n table (for simplicity we suppress the index for the trial, replace w_i by w, and replace x_{i+} by x_+) the deviance is:

$$D(X;X^*) = 2\sum_j x_j \ln\Big(\frac{x_j}{x_j^*}\Big). \tag{8.3}$$

Ye (1998) defines the effective dimension of a model that estimates $\hat{y}_j$ for observation y_j quite generally as the following sum of partial derivatives

$$ED = \sum_j \frac{\partial \hat{y}_i}{\partial y_j}. \tag{8.4}$$

Applied to our problem, this becomes

$$ED = \sum_j \frac{\partial x_j^*}{\partial x_j}, \tag{8.5}$$

and we find

$$ED = \sum_j \frac{\partial[w\lambda_j x_+ + (1-w)x_j]}{\partial x_j} = n(1 - w + wx_+/x_{++}), \tag{8.6}$$

where we have taken into account that $\partial\lambda_j/\partial x_j = 1/x_{++}$. In the case of clinical trials, we have $n = 2$. The optimal weighted combination minimizes the AIC. This can simply be done by searching on a grid for w, or by a more sophisticated optimization routine.

In our experience this definition of AIC was not very successful. By summing the deviance and (two times) the effective dimension, it balances closeness to the data and complexity of a model, preventing over-fitting. In practice the influence of the effective dimension (which is between 0 and 2, if we work on a 1 by 2 table) was too small, resulting in a small weight for the prior.

Hurvich et al. (1998) studied AIC in the case where the number of parameters in a model is close to the number of observations. They found that corrections to AIC are necessary in this case. Their correction is very close to using $ED^* = ED * n/(n - ED)$ instead of ED itself. This modification works well in the present context. When ED approaches 2, ED^* will approach infinity, so there will always be given some weight to the prior, although it might be small.

This choice for a modified effective dimension is heuristic. We could not (yet) find a very convincing theoretical analysis to support it. But our simulations show good performance of this choice, especially on the corrections of individual trials.

8.3 Simulations

In this section we study the performance of pseudo-Bayes smoothing by simulation. We compare the following procedures.

- Pseudo-Bayes smoothing of all trials (global), with a joint prior based on marginal counts for the two arms. The individual trials are simply stacked to form one large table, and subsequently prior cell probabilities can be calculated.
- Pseudo-Bayes smoothing of only the trials with one or more observed zeros (local), using the joint prior as defined above.
- Constant continuity correction with two different values: $c = 0.5$ and $c = 0.1$, which are added to all cells of a trial with one or more observed zeros.
- The empirical continuity correction as presented by Sweeting *et al.*. The total sum of the corrections over all four cells in a trial equals two.
- The treatment arm correction of Sweeting et al. (2004), with the condition that the sum of the corrections over all cells equals two.

For all methods the total zero events studies are kept in the analysis, as advocated by Friedrich et al. (2007). The design of the simulation is inspired by Sweeting et al. (2004), but it differs in the selection of fixed and variable parameters. We focus on three parameters: heterogeneity in the odds ratio Ω between the trials within a meta-study, heterogeneity in the baseline event probability P_c and the magnitude of the odds ratio Ω. For calculating the pooled estimate we chose the Mantel-Haenszel method (Mantel and Haenszel, 1959) and its variance estimator presented by Robins et al. (1986), because of its widespread use and positive evaluation in connection with sparse data (Sweeting et al., 2004). Our

Table 8.1: Parameters and assigned values used in the simulations

Parameter	Assigned values
Odds ratio Ω ($= e^{\omega}$)	0.35, 0.45, 0.5, 0.55, 0.65, 0.75
Between trial heterogeneity in effect $\sigma_{log(\Omega)}$	0, 0.05, 0.10, 0.15, 0.18
Control group event probability P_c ($= e^{p}$)	0.01
Between trial heterogeneity in baseline σ_p	0, 0.08, 0.11, 0.14, 0.17, 0.20
Number of studies in each meta-analysis	10
Number of simulated meta-studies s	2000
Number of patients in treatment group	100
Ratio of group sizes R (C:T)	1:1

performance criteria are: coverage, bias and the root mean squared error (RMSE). In the subsections that follow, we present details of the design and performance criteria and evaluate the results of the simulations.

Simulation design

Sparse data is generated using the following design. The log of the odds ratio $\omega = \log \Omega$ is drawn from a normal distribution with mean = μ_ω and standard deviation σ_ω. The magnitude of Ω and σ_ω are varied over the different scenarios. The log of the event probability in the control group, $p = \log P_c$ (i.e. the baseline effect) is drawn from a normal distribution with the mean equals $\log 0.01$ and standard deviation according to the scenario. Having the odds ratio and the baseline event probability, the treatment group event probability P_t can be calculated:

$$P_t = \frac{(\frac{P_c}{1-P_c})\Omega}{1 + (\frac{P_c}{1-P_c})\Omega} \tag{8.7}$$

After determining P_t for all simulated meta-studies the actual data are generated by drawing from binomial distributions. Each meta-analysis consists of ten individual studies. Because of the low event probabilities, observed zeros in either one or both arms of a trial will be common. It might occur that all ten trials in one meta-study do not contain any event in the treatment or control group. In these situations the prior probability is zero as well and thus no pseudo-Bayes corrections are possible. To prevent this situation, data with only zero events are discarded. Table 8.1 presents the parameters involved in the simulation, the actual computations were done in R (version 2.10).

In total, fifteen simulation runs are performed, with in each run 2000 meta-studies, all consisting of ten trials. These fifteen runs can be divided into three series; first, there are five simulations where the standard deviation in the sampled log odds ratio σ_ω varies from 0 till 0.18. The second series consist of five simulations with σ_ω fixed at 0 and Ω ranging from 0.35 up to 0.75. In the last series of simulations the control group event probability

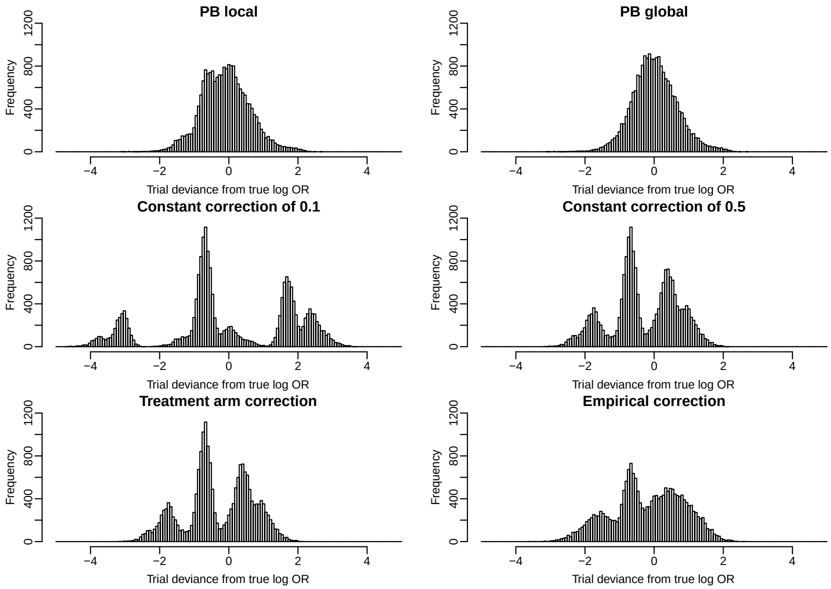


Figure 8.1: Distributions of the differences between estimated and simulated log odds ratio (DLO), per trial, for six correction methods, as indicated by the titles of the panels. The simulation settings are $\Omega = 0.5$, σ_ω = 0.18 and σ_p= 0.

becomes a flexible parameter, following a log-normal distribution with the mean equals $\log(0.01)$ and standard deviation on the log scale ranging from 0.08 to 0.20 (see Table 8.1). This time σ_ω is fixed at zero, and Ω is set to 0.5. Most of the sampled P_c values are in the range from 0.004 to 0.022.

Evaluation

We evaluate the performance of the correction procedures on several criteria and on two levels. For the individual trials we compute differences between the simulated and estimated log odds ratio, which we abbreviate to DLO. The histograms include all trials, corrected and uncorrected.

Typical results for DLO,are shown in Figure 8.1. The simulation settings are $\Omega = 0.5$, σ_ω = 0.18 and σ_p= 0. The histograms show the distribution of DLO for the six correction methods we studied. For both pseudo-Bayes procedures we get an almost normal distribution. The difference between the two distributions are due to the presence of uncorrected studies in the first one. The other four methods return more complicated, multimodal distributions with longer tails.

The multimodality has two reasons. First it is a consequence of the specific set-up of the simulations. The size of the treatment and the control group are fixed at 100, this in

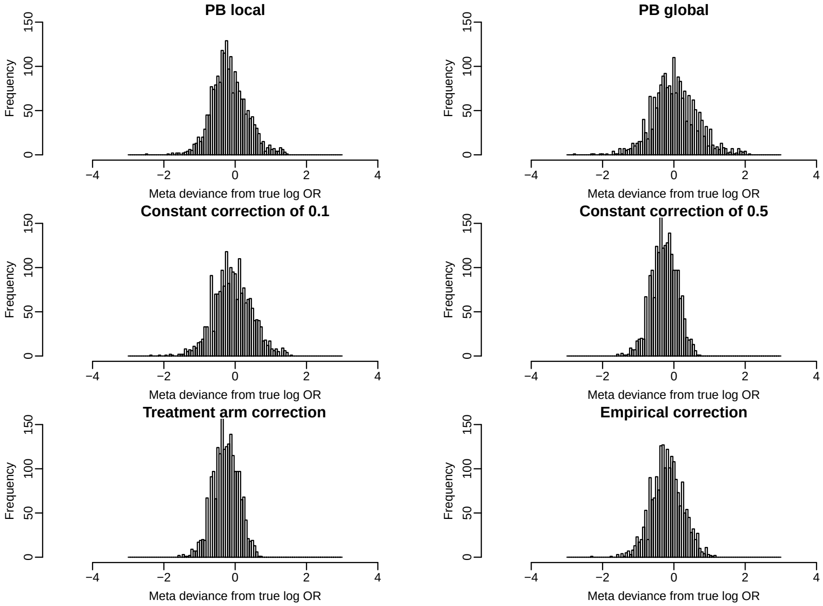


Figure 8.2: Distributions of the differences between estimated and simulated log odds ratio per meta-study, for six correction methods, as indicated by the titles of the panels. The simulation settings are $\Omega = 0.5$, σ_ω = 0.18 and σ_p= 0.

combination with a small event rate, will result in many studies with similar distributions of counts in the four cells. Second, it also also shows some of the characteristics of the different corrections. Small changes in the data can cause drastic changes in the estimated effect. This is especially the case for the constant continuity correction of 0.1. Applying this correction to trials with only one observed case in one of both arms results in a large estimated effect.

The histograms of the two constant corrections are basically the same, although when adding a constant of 0.5 the effects on the estimates are milder. However, both corrections set many coefficients to zero. Adding a constant of 0.1 should clearly be avoided. The output for the treatment arm correction is equal to the constant correction of 0.5. The empirical correction returns a distribution that is less dispersed.

On the level of the meta-analysis we first compare the estimated log of the odds ratio with the simulated one using the Mantel-Haenszel estimator. The results for the same simulation as discussed in Figure 8.1 are shown in Figure 8.2. The histograms show the somewhat conservative nature of most of the corrections, except for the PB overall procedures which seems to be a bit optimistic. Looking at the histograms of both constant continuity corrections we observe a large peak close to zero, indicating very small effects on the level of a meta-study caused by many zero effects on the trial level.

We also evaluate output of the simulations given the three scenarios as presented above. Results connected to the first scenario, in which the odds ratio is varied, are given in Figure 8.3.

In the left top corner we see the RMSE on the meta-level. The general picture is that for all methods the error decreases when the true effect is nearer to zero. When smoothing all trials using the pseudo-Bayes procedure the largest error is produced. The treatment arm correction and the constant continuity correction of 0.5 generate the least amount of error.

The coverage of all procedures is affected by a change in the true odds ratio. The empirical correction the treatment arm correction and constant correction of 0.5 all benefit from a smaller odds ratio. The same holds for the pseudo-Bayes procedure which only corrects studies with empty cells. Lowering the odds ratio has a negative effect on the coverage of the pseudo-Bayes procedure which smoothes all trials. The bias for this method is small and positive over the whole range, but hardly affected by a change in the OR. The positive bias indicates that the effect is often slightly overestimated. Al other corrections return negative bias, which decreases when the odds ratio becomes smaller. It is likely that they benefit from effects closer to zero because of their conservative nature.

For the RMSE at the trial level the picture is very stable. Both pseudo-Bayes procedures generate the least amount of bias, while the constant correction of 0.1 generates the most. Actually, this is a confirmation of the pictures we saw in Figure 8.1. Comparing both pseudo-Bayes procedures we can conclude that shrinkage helps with respect to individual trials. As said, shrinking to an overall estimate reduces the bias. There is, however, a price to pay: is seems that the prior is often too extreme and strong smoothing on the trial level will generate more bias at the meta-level. This problem can be alleviated by smoothing only the studies with zero cells.

For the second scenario heterogeneity of the odds ratio is introduced, of which results are shown in Figure 8.4. For all performance criteria we see a rather stable pattern over the different amounts of heterogeneity. The constant of 0.5 and the treatment arm correction are identical in their performance. The broader picture is comparable to the previous scenario. The pseudo-Bayes method smoothing all studies generates the largest RMSE, slightly overestimates the effect, but on the trial level comes closest to the true log odds ratio. If only studies with zeros are corrected, the trial level RMSE is slightly larger. The correction of 0.1 again creates a lot of error on this level. For all six methods the coverage starts above 95 per cent and drops slightly when increasing the variance.

Figure 8.5 shows the output of scenario three, where we introduce heterogeneity in the control group event probability. For all performance criteria we again see a similar ordering in the corrections as we saw in the previous scenarios. For the coverage and bias we observe a sharp dip at $\sigma_{log(p)} = 0.14$, which is due to many estimated zero effects at the meta-level, caused by the sampling procedure. With introducing heterogeneity in the baseline event probability we also see that in some instances the bias of the pseudo-Bayes

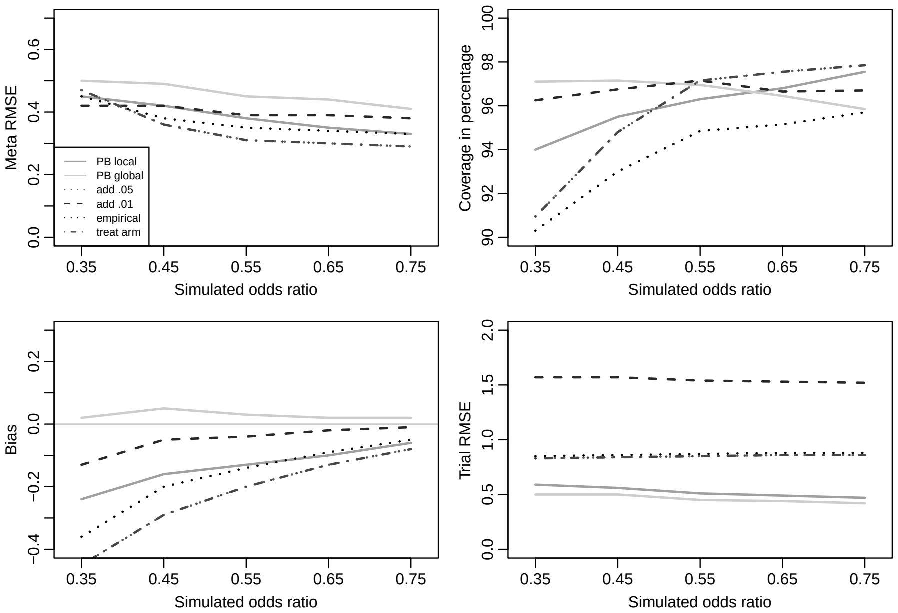


Figure 8.3: Results showing the influence of the odds ratio on performance for six correction methods (see legend). In this scenario Ω ranges from 0.35 up to 0.75, while σ_{ω} and σ_{logP_c} are fixed at zero. Counter-clockwise from upper left: RMSE and bias of pooled log odds ratio, RMSE of log odds ratio per trial, and coverage.

smoother for all trials becomes negative, instead of the positive bias observed earlier. The RMSE at the trial level remains stable over the whole range for all corrections.

In general we observe that the smoothed effect of the global pseudo-Bayes procedure is often a bit too optimistic. If we only smooth trials with zero we notice two effects. On the meta-level we observe a drop in the RMSE, which makes it competitive with other corrections. The bias indicates a more conservative estimate close to the empirical correction. However, on the level of individual trials we see a slight increase in RMSE. Nonetheless this error is considerably smaller than the error of the other four corrections. This combination makes it an attractive alternative when both the meta-study and individual trials are of concern.

8.4 Applications

In this section two applications are presented. The first example contains only a few trials with observed zeros, while the second shows extremely sparse data. In both cases we calculate pooled estimates after applying all six corrections that were studied in the previous section. In case of the EFM data we used the pseudo-Bayes procedure in combination

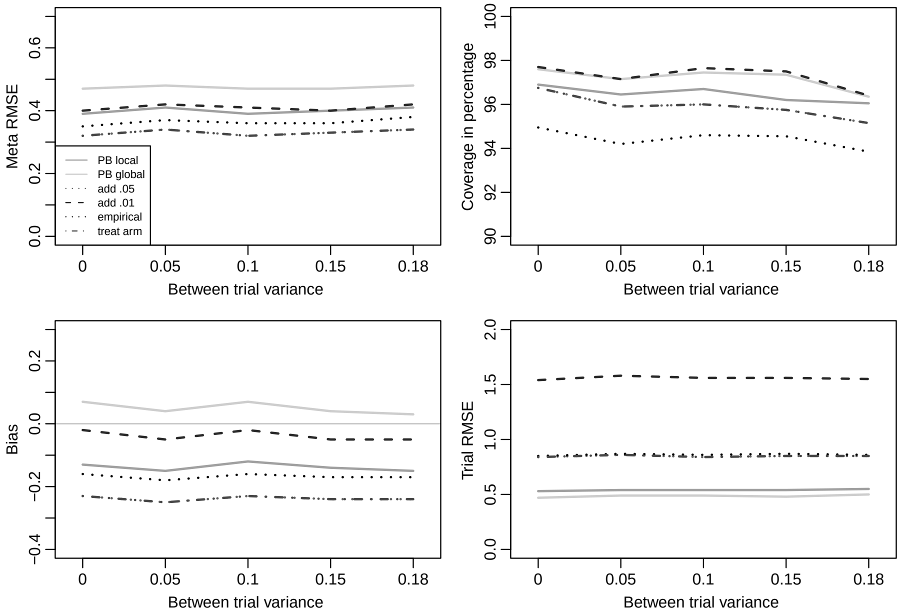


Figure 8.4: The influence of the heterogeneity of trials (σ_ω) on performance for six correction methods (see legend). In this scenario σ_ω ranges from 0 till 0.18, Ω is fixed at 0.5 σ_p equals zero. Counter-clockwise from upper left: RMSE and bias of pooled log odds ratio, RMSE of log odds ratio per trial, and coverage.

with the prior based on marginal counts (the joint prior) and the constant log-odds prior as discussed in section 2. In both cases the smoother is applied on all individual trials. After correcting the data, a pooled odds ratio is calculated with the Mantel-Haenszel (MH) method.

Table 8.2 shows data from seven different studies investigating whether electronic fetal heart rate monitoring (EFM) causes a drop in the perinatal mortality rate (Sweeting et al., 2004). A large part of the trials is not balanced. First, we created a simple prior by collapsing the data into one table. Second, we generated a prior based on the MH odds ratio over the seven studies. The smoothed counts using both methods are given in Table 8.3. For both corrections the assigned weights are shown in the last two columns of the table.

Table 8.4 shows pooled estimates for the log odds ratio for different corrections. The largest value is found using the pseudo-Bayes procedure with the prior based on marginal counts. The most conservative estimate is found applying the constant correction of 0.5. In general, the estimated effects and variances are relatively close to each other. This is to be expected, since this dataset is not very sparse and thus does not need a large correction.

A second example, in which the data are very sparse, is provided by Cheng et al. (2005)

Table 8.2: The raw data from seven different studies investigating whether EFM causes a drop in the perinatal mortality rate.

	EFM		No EFM	
	No Event	Event	No Event	Event
1	1160	2	5410	17
2	150	0	6821	15
3	607	1	6142	37
4	4209	1	2914	9
5	553	1	689	3
6	4978	0	8632	2
7	45870	10	66163	45

Table 8.3: The corrected counts for the EFM study using a prior based on the marginal table, and a prior based on the Mantel-Haenszel (italic). The last two columns present the weight of the prior.

	EFM		No EFM		Weight of prior	
	No Event	Event	No Event	Event	mar. pr.	OR pr.
1	1335.9 *1161.3*	1.82 *0.68*	5235.9 *5415.9*	15.43 *11.01*	0.14	1
2	290.5 *149.6*	0.04 *0.04*	6681.0 *6825.9*	14.47 *10.07*	0.06	1
3	782.0 *607.7*	0.97 *0.25*	5969.9 *6153.4*	34.14 *25.64*	0.09	1
4	3979.3 *4208.9*	0.96 *1.10*	3144.2 *2917.8*	8.54 *5.23*	0.15	1
5	507.4 *553.8*	0.55 *0.25*	736.0 *690.7*	1.99 *1.26*	0.51	1
6	5038.6 *4976.7*	0.87 *1.30*	8564.4 *8626.9*	1.71 *7.12*	0.66	1
7	44721.2 *45868.0*	10.25 *11.96*	67298.1 *66145.6*	58.36 *62.45*	0.28	1

Table 8.4: Pooled log odds ratio for the EFM data after different corrections.

Correction	$\log(\Omega_{MH})$	SE $\log(\Omega_{MH})$
PB$_{marginal}$	-1.40	0.27
PB$_{\tilde{\Omega}}$	-1.24	0.28
Constant: c = 0.5	-1.16	0.27
Constant: c = 0.1	-1.22	0.28
Treatment arm	-1.22	0.28
Empirical	-1.24	0.28

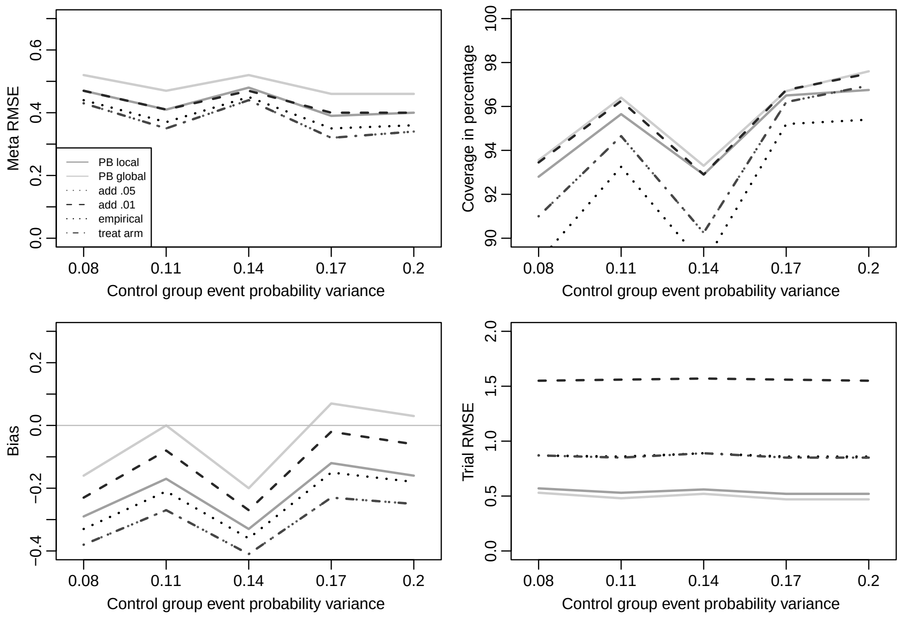


Figure 8.5: The influence of variability (σ_p) of the control group event probability on performance for six correction methods (see legend). Here σ_{logP_c} ranges from 0.08 till 0.2, simulated Ω is fixed at 0.5 and σ_ω is fixed at zero. Counter-clockwise from upper left: RMSE and bias of pooled log odds ratio, RMSE of log odds ratio per trial, and coverage.

and is also studied by Rücker et al. (2009) in relation to correcting sparse data. The meta-study investigates the effect of off-pump surgery in coronary artery bypass grafting on postoperative stroke. In this example we only use the pseudo-Bayes procedure in combination with a marginal counts prior calculated separate for the treatment arm and control arm. The procedure based on a prior OR cannot be used here because there are several studies with zero total events.

As we can see for many trials the weight assigned to the prior is rather large. For most trials it is 1, even in both arms. This makes sense: if counts are very low or zero, a rather strong correction is necessary.

Also for this study corrected pooled estimates have been calculated and compared with other corrections: they are presented in Table 8.6. The pooled estimates are calculated including the total zero event studies.

In Table 8.6 we can observe that the differences in the estimated effect size and variances between the different corrections are more pronounced in comparison with the previous example. This is due to the very sparse data of this example. The pseudo-Bayes correction again shows the largest effect. The estimated variance of the prior is the largest as well. Adding a constant of 0.5 returns the most conservative estimate.

Table 8.5: Meta-analysis of off-pump surgery in coronary artery bypass grafting on postoperative stroke. Observed counts and corrected counts (italic) using the split prior are presented. The last two columns shows the weight of the prior. The last row shows the prior probabilities.

	Treatment group				Control group				Weight of Λ	
Study	No Event		Event		No Event		Event		TG	CG
1	100	*99.58*	0	*0.42*	100	*99.02*	0	*0.98*	1.00	1.00
2	100	*99.58*	0	*0.42*	101	*100.01*	0	*0.99*	1.00	1.00
3	28	*27.88*	0	*0.12*	36	*36.64*	1	*0.36*	1.00	1.00
4	14	*13.94*	0	*0.06*	16	*15.84*	0	*0.16*	1.00	1.00
5	19	*19.76*	1	*0.24*	19	*19.74*	1	*0.26*	0.83	0.92
6	80	*79.66*	0	*0.34*	80	*79.22*	0	*0.78*	1.00	1.00
7	54	*53.77*	0	*0.23*	49	*48.52*	0	*0.48*	1.00	1.00
8	148	*149.29*	2	*0.71*	150	*148.54*	0	*1.46*	0.95	1.00
9	60	*59.75*	0	*0.25*	58	*59.28*	2	*0.72*	1.00	0.90
10	15	*14.94*	0	*0.06*	16	*15.84*	0	*0.16*	1.00	1.00
11	10	*9.96*	0	*0.04*	10	*9.90*	0	*0.10*	1.00	1.00
12	15	*14.94*	0	*0.06*	19	*19.74*	1	*0.26*	1.00	0.92
13	88	*87.63*	0	*0.37*	86	*87.14*	2	*0.86*	1.00	1.00
14	141	*141.40*	1	*0.60*	137	*137.64*	2	*1.36*	1.00	1.00
15	136	*135.43*	0	*0.57*	131	*129.72*	0	*1.28*	1.00	1.00
16	204	*203.14*	0	*0.86*	182	*182.20*	2	*1.80*	1.00	1.00
17	97	*97.59*	1	*0.41*	97	*98.03*	2	*0.97*	1.00	1.00
18	23	*23.76*	1	*0.24*	25	*25.74*	1	*0.26*	0.85	0.99
19	25	*24.90*	0	*0.11*	25	*24.76*	0	*0.24*	1.00	1.00
20	41	*40.83*	0	*0.17*	67	*66.35*	0	*0.65*	1.00	1.00
21	21	*20.912*	0	*0.09*	16	*15.84*	0	*0.16*	1.00	1.00
Λ		.9902		.0098		.9958		.0042		

Table 8.6: Pooled log odds ratio for the stroke data after different corrections.

Correction	$\log(\Omega_{MH})$	SE $\log(\Omega_{MH})$
$\text{PB}_{marginal}$	-0.80	0.43
Constant: $c = 0.5$	-0.42	0.35
Constant: $c = 0.1$	-0.70	0.45
Treatment arm	-0.44	0.33
Empirical	-0.63	0.35

8.5 Discussion

To resolve problems caused by observed zeros and low counts simple continuity corrections have been proposed. Also more involved mixed model and Bayesian approaches are available. We propose the pseudo-Bayes procedure, which forms a weighted average of the observed counts and a prior. In contrast to a fully Bayesian analysis, the combination of prior and posterior distributions and a likelihood function is replaced by a weighted average of two numbers. Pseudo-Bayes is a linear shrinkage procedure.

In this chapter we discussed different strategies to derive a prior. A simple recipe is to pool the counts of all trials and calculate the cell probabilities from them. This can be done separately for treatment and control arm, or for both arms combined. If the trials are unbalanced, Simpson's paradox can strike. To avoid that, we also proposed a prior using the meta-level odds ratio.

We use an adjusted Akaike Information Criterion to determine the optimal weight of the prior. AIC combines the deviance and an effective model dimension. We have to deal with a situation in which the effective dimension, as defined by Ye (1998) can get very close to the number of observations. By analogy of the work in Hurvich et al. (1998), we introduce a factor that increases the effective dimension when it approaches the boundary. This is a heuristic device, but it seems to work well, as shown by our simulation results. Although it appears to work well, we are not completely happy with our heuristic solution for the modified effective dimension. AIC was derived on the basis of a quite general analysis of prediction performance. We are working on a similar analysis, in the specific setting of contingency tables.

Judging from the precision of the log odds ratios estimated for the individual trials, pseudo-Bayes clearly outperforms its competitors. It is also strong on bias and coverage of the estimate of the pooled, Mantel-Haenszel-based, log odds ratio. However, for the latter it shows larger variability than the other corrections.

Simulations showed that pseudo-Bayes is a useful tool. However, its value is to a large extent determined by the chosen prior. In some situations the prior can suffer from Simpson's paradox and lead to incorrect outcomes. Still, if a proper prior is derived, the correcting of trial counts can be recommended using pseudo-Bayes estimation. Discrete outcomes are replaced by continuous numbers, which allows for the computation of differences of logarithms of event probabilities. This is a clear advantage for graphical exploration, e.g. by the L'Abbé plot (L'Abbé et al., 1987).

Conclusions 9

This chapter is divided into two sections. The first section reflects, on a broad level, on the general idea of this thesis. The second section summarizes the main findings reported in the different chapters. In both sections possible improvements and initiatives for future research are proposed.

9.1 General conclusions

Shrinkage estimators have become a popular tool for a wide range of analysis problems. Examples are the widespread use of penalized regression and the generalized additive models (GAM). In this thesis, we use penalties to overcome the ill-conditioned nature of data in various settings. A concise overview of major developments for size penalties or roughness penalties is provided in Chapter 1.

Roughness penalties occur in Chapters 2, 3, 4, 5 and 6. In all cases smoothness is obtained using discrete penalties. They are easy to construct and quickly adapted to specific needs, have attractive numerical properties and are well grounded in statistical theory. For long signals the systems to solve can get very large, but using sparse matrix calculations alleviates this burden. Further gain in speed is obtained by efficiently composing the B-spline matrix.

Size penalties are discussed in Chapters 4, 5, 7 and 8. In the literature they are prominently based on the L_2 and L_1 norm. In some applications these penalties are unsuccessful and alternatives are required. We concentrate on the L_0 norm penalty. The problem with all L_p penalties with $p \leq 1$ is the non-convex objective function, making optimization of the coefficient vector hard. Nonetheless a number of this type of penalties and optimizers have been proposed (see Gasso et al., 2009, for an overview). Although the algorithms cannot claim to reach a global minimum, satisfactory results are obtained in many settings. In this thesis the L_0 norm penalty is successfully applied to sparse deconvolution of signals and images. Our results and simulations suggest that the obtained local minimum is at least close to the global one. Estimates are obtained using an iteratively reweighted least squares (IRWLS) based algorithm. It essentially replaces the L_p norm by a weighted L_2 norm.

The great advantage of penalization techniques is that they allow the estimation of high dimensional models, without the necessity to apply some type of feature extraction. The penalty makes the model estimable and proper tuning prevents overfitting of the data. The typical way to achieve this is to use some type of information criterion or to apply cross-validation. In reference to especially Chapters 3, 4 and 5 we note that further investigations are required on model selection criteria in the presence of complex (serial) correlations within the errors. The familiar information criteria, like the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), often fail in these cases. A possible alternative is the L-curve, as positive results are reported in the case of signal calibration (Kalivas, 2012) and smoothing of time series (Frasso and Eilers, 2012).

Not discussed within this thesis, but important to mention are alternative interpretations of shrinkage estimators. Penalized estimation is formally equivalent to a mixed model, as discussed by Wand (2003). In this case, the penalty is written as an a priori normal distribution on the (spline) coefficients. Attractive about this presentation is that the amount of smoothing is automatically controlled by the variance components in the mixed model, calculated during the estimation of the model.

By fitting penalties in a mixed model view, we are half way towards the Bayesian interpretation of penalized estimation. Bayesian P-splines are proposed by Lang and Brezger (2004), and covers various types of models. Smoothness is obtained by introducing a random walk prior, which is dependent on neighbouring data points. The degree of smoothness is controlled by the variance parameter of the prior and that of the observation noise. The connection between Bayesian estimation and shrinkage was pointed out by Tibshirani (1996) and further investigated by e.g. Park and Casella (2008) and Hans (2009). Inference is carried out using MCMC algorithms, which often makes the procedure slow especially for long signals as discussed in earlier chapters.

A downside of both the mixed model view and the Bayesian interpretation, is the lack of flexibility in the penalty. In addition to the general model, implied by the use of a particular penalty, it is sometimes required to impose further (local) constrains. By writing the penalty in a more explicit way it is easily adjusted in order to incorporate for instance, circular behaviour, periodicity, interpolation and extrapolation and the limiting behaviour of the penalty. Ruppert et al. (2003) made a comparison of the performance of the mixed model approach of penalized estimation against methods relying on cross-validation or information criteria. They notice clear differences, but call for more research to make a judgement.

9.2 Chapter-specific conclusions

Central to Chapters 2 and 3 is the proper estimation of a baseline from a noisy signal. The signal is often composed of a series of peaks, superimposed on a drifting baseline. The physical or chemical information is in the positions and the heights of the peaks and the baseline is a nuisance. The proposed baseline estimation method relies on a two-component mixture model: one for the baseline including noise and a second for the peaks. The baseline is estimated using P-splines. In Chapter 2, the focus is on one-dimensional data. Application on various types of signals shows its excellent performance. The model is similar to asymmetric curve fitting approaches, but avoids the problem of choosing the amount of asymmetry, because this follows implicitly from the mixture model. The large gain of this approach is that the baseline is approached as a statistical model.

In Chapter 3 the method is extended to two-dimensional data, meaning that a baseline surface has to be estimated. The same mixture principle applies, only tensor product P-splines are used to estimate the surface. To demonstrate the value of the procedure we focus on artefact removal in time-resolved spectroscopy. The method can be applied in other situations as well. Examples are, correction of two-dimensional electrophoretic data or microarray images used for gene expression analysis. A logical extension of the approach is to model three dimensional data. This means we have two directions in space and one in time. Practical examples are weather data with a spatial dimension and measured in time. A second example comes from tribology and concerns the wear of mechanical parts as a function of mileage. Further extensions in terms of applications are model series of count data, like X-ray diffraction scans, and sinusoidal data series where both a lower bound and upper bound need to be estimated. Topics requiring more research are automatic model selection and baseline estimation in the presence of heteroscedastic noise. The second problem could be handled by incorporating a flexible variance parameter.

In Chapter 4, the focus is on the estimation of a small set of input spikes from a noisy convolved signal. The sparsity of the input requires a penalty that sets many parameters to zero, but does not create too much bias in the estimated spikes. We find that regression with an L_0 norm penalty performs very good, for simulated data and real data. In many cases not only the input, but also the impulse response function (IRF) is unknown. Estimating both from the data is often called blind deconvolution, which is implemented using an alternating minimization scheme. Assuming we know the impulse response we can make an estimate of the input. Reversely, if we know the input, we can make an estimate of the response function. This procedure is repeated until convergence. For optimal performance, the estimated impulse response is smoothed using a Whittaker smoother.

Chapter 5 is a continuation of Chapter 4. Deconvolution using penalized regression with an L_0 norm penalty is successfully applied to two-dimensional data. In addition, the blind deconvolution algorithm now applied to virtually any unimodal impulse response function. Vectorizing two-dimensional data quickly results in large systems of equations, which leads to an increase in computation time. A possible solution is to split images in (overlapping) segments and analyse these separate.

Chapter 4 and chapter 5 both demonstrate that penalized regression with an L_0 penalty is a very useful tool for deconvolution. To model crosstalk between parallel signals we could extend the framework using the composite link model (CLM). The CLM considers the data to be composed of a linear combination of generalized linear models.

Chapter 6 discusses the binary classification of two-dimensional shapes. By transforming the x and y coordinates into polar coordinates, the shapes are represented as signals. Consequently, the shapes can be summarized by P-splines, and the signals can be analysed using penalized signal regression. From a data modelling perspective this is attractive, because the samples can now be analysed by penalized logistic regression, which is a member of the generalized linear model (GLM). Besides attractive statistical properties of the GLM, other outcomes than the binary can be modelled. Treating the data as signals has the additional advantage that alignment and rotation of the shapes becomes easy. Applications show excellent classification performance. Discriminative regions can be visualized by plotting the beta coefficients in a circle emphasizing which parts of the shape are most influential. The model can be extended to three-dimensional shapes by relying on spherical or cylindrical coordinates. In a next step the shapes can be summarized using tensor product P-splines. The circular character of the data is preserved by adjusting the basis and difference penalty. When covariates (clinical parameters) are available, they can be easily included into the regression model.

Estimating gene networks, using expression data, is the central theme of Chapter 7. Obtaining a conditional network is difficult because of the relative small number of samples, while the number of genes easily runs into thousands. To get an estimable model and induce sparseness, size penalties are used. Different types of penalties are discussed, and a small simulation is performed to assess the performance of three of them. To restrict the search space and possibly improve the quality of the network, knowledge from external sources is included in the estimation procedure. A simple framework to combine the prior and data is proposed and two applications are presented. In the first application a prior is build using text mining. In the second, one specific publication and the connections found therein are used as a basis for a prior. With text mining there is the advantage that the prior matrix is fully specified by the algorithm, meaning that gene to gene weights are available. In other situations this is often not the case, which makes it difficult to generate a weighting scheme for hypothesized edges.

In Chapter 8 we propose a pseudo-Bayes procedure to resolve problems caused by observed zeros and low counts in contingency tables. The cell counts are corrected by a linear shrinkage procedure and are calculated as weighted average of the observed counts and a prior. Different priors are possible depending on the application. The optimal combination of the data and prior is determined with an adjusted AIC. Both simulations and applications show that correcting of cell counts using pseudo-Bayes estimation can be recommended.

Summary

Central to this thesis are ill-conditioned problems, in a high-dimensional setting. In many cases this problem is caused by the fact that the number of features is (much) larger than the number of observations. As a result, proper model estimation is difficult and additional measures have to be taken. A number of solutions are available, often roughly divided into variable selection and extraction strategies. Imposing constrains is a third option. In the preceding chapters we relied on penalties on the size and smoothness of a solution. An optimal balance between a parsimonious model and fidelity to the data can be found using for instance information criteria or cross-validation. Chapter 1 offers a concise overview of the conditioning problem and penalized estimation. The chapters 2-8 showing the versatility of penalized estimators and their ability to reach a satisfactory solution in a broad range of applications.

Chapter 2 and Chapter 3 focus on the correction of baselines and artefacts, as present in instrumental signals, studied for instance in analytical chemistry. The physical or chemical information is in the positions and the heights of the peaks and the baseline is a nuisance. The proposed solution relies on a mixture model of two components. One component for the peaks, and a second for the baseline including noise. The model is optimised using an EM algorithm, the smooth baseline is estimated using P-splines. One advantage of the procedure is that the baseline can be considered as a statistical model, rather than a fixed element. In Chapter 2 the method is presented and applied to one-dimensional signals. An extension into two dimensions is proposed and evaluated in Chapter 3. The smooth surface is modelled using tensor product P-splines.

Baseline removal is important for the analysis of signals in which the shape and height of the peaks that are present need to be determined. Assuming that the peaks are a convolution of an input signal with an impulse response function (IRF), peak characterization can be performed using deconvolution. The task is to estimate the input signal, given the observed data and the point spread function. One-dimensional sparse deconvolution is discussed in Chapter 4. Sparseness is imposed using an L_0-norm penalty. In many cases both the input signal and the point spread function are unknown. We estimate both in an iterative manner, and smooth the IRF for optimal results. Chapter 5 present some extensions to the proposed procedure. First, the method is applied to two-dimensional data, as found in for instance microscopy. Second, by adjusting the updating function of the IRF, the algorithm can now perform blind deconvolution for a much broader class of functions. An adjusted smoother forces the estimated response function to be unimodal

and positive everywhere. An application is shown where an IRF with an approximate exponential decay is estimated.

A seemingly totally different topic is treated in chapter 6, namely the classification of two-dimensional shapes. However, by transforming the x, y coordinates into polar coordinates, the data can be analysed as signals. After transformation, the data are summarised on a P-spline basis. This facilitates preprocessing steps like sample alignment and noise removal. By adjusting the basis and the penalty, the circular character of the data is preserved, irrespective of the amount of smoothing. Classification is performed using penalized signal regression (PSR). The elegance of this model is that it is a generalized linear model, and thus connects shape analysis to very familiar statistical tools. The estimated coefficients are interpretable, like any other GLM. By representing them on a (mean) shape, discriminating areas are easily located. Three applications are discussed, all showing excellent results.

Chapter 7 elaborates on network estimation and, more in particular, identification of gene modules. Because of the large number of features and the limited number of samples, estimation is not trivial. A number of approaches are present in the literature, in which shrinkage estimators play a prominent role. We investigate the value of three different penalties using simulations. In addition to the penalty, one could utilize the large amount of information on gene interaction on the web. We use a simple framework in order to combine prior knowledge and the observed data in one network estimation procedure. To demonstrate the procedure, two applications are provided. In the first application, a prior set of connections is built using one particular publication, reporting the discovery of a number of gene modules. The second application uses text mining algorithms in order to retrieve gene-to-gene information.

Chapter 8 discusses a correction for low and zero counts in contingency tables. Adjustments are required because only relying on the observed data will result in unreliable estimates, or estimation will be impossible. The cell counts are corrected by a linear shrinkage procedure. The corrected counts are calculated as weighted averages of the observed counts and a prior. An optimal value for the penalty parameter is obtained using an adjusted Akaike information criterion (AIC). The procedure is demonstrated by way of series of 2 × 2 tables, that are part of a meta study. If a suitable prior is applied, correcting of trial counts can be recommended using pseudo-Bayes estimation. Further evaluation of the method is done by a series of simulations.

The last chapter of this thesis provides an overview of the main findings and points of discussion on a chapter level. In addition, shrinkage estimators are discussed from a broader perspective. New questions and possible approaches for future research are provided.

Samenvatting

Centraal in deze thesis staan slecht geconditioneerde problemen, in een hoogdimensionale context. Het betreft hier doorgaans situaties waarin het aantal variabelen het aantal observaties (ver) overstijgt. Om toch uitspraken te kunnen doen op basis van de beschikbare data, zijn extra maatregelen vereist. Hiervoor zijn verschillende oplossingen voorhanden. Te denken valt aan variabelenselectie op basis van een van de meer klassieke methoden, of variabelenextractie, bijvoorbeeld met behulp van principale componenten analyse. In dit werk vertrouwen we echter volledig op een derde optie, de toepassing van een penalty. De gebruikelijke verliesfunctie wordt aangevuld met een penalty. Deze zorgt voor een tegenbeweging bij het optimaliseren van de parameters, zodat nu de balans wordt gezocht tussen de fitaanpassing aan de data (de fit) en de bias die geïntroduceerd wordt door de penalty. Omdat de penalty de modelcomplexiteit omlaag brengt, is het resultaat een beter geconditioneerd probleem, wat het stabiel schatten van de parameters van een model mogelijk maakt.

Een globaal onderscheid is gemaakt naar twee type penalty′s, de *size penalty* en de *roughness penalty*. De eerste penaliseert de grootte van een vector van parameters en krimpt deze richting nul of een andere vooraf vastgestelde waarde. Deze penalty wordt veel toegepast binnen regressie. De *roughness penalty* is vooral terug te vinden in de niet-parametrische setting en kan gebruikt worden voor het schatten van een curve door een serie datapunten. Een gladde curve (*smoothness*) wordt bereikt door de verschillen tussen naburige coëfficienten te bestraffen. De invloed van de penalty, en dus de *smoothness*, is te variëren met behulp van een extra parameter. Het optimale gewicht van de penalty, of in andere woorden de optimale balans tussen de fit en modelcomplexiteit, kan gezocht worden met behulp van verschillende hulpmiddelen. Meest gebruikt hiervoor zijn informatiecriteria of kruisvalidatie.

Hoofdstuk 1 biedt een beknopte introductie in slecht geconditioneerde problemen en gepenaliseerd schatten. In de daarop volgende hoofdstukken komen verschillende toepassingen aan bod waarbij in alle gevallen een penalty toegepast wordt om tot een goed resultaat te komen.

In de hoofdstukken 2 en 3 wordt gekeken naar instrumentele data. Belangrijke informatie in dergelijke data vormt doorgaans de positie en hoogte van de aanwezige pieken. De in veel gevallen aanwezige basislijn is hierbij storend en dient gecorrigeerd te worden. In hoofdstuk 2 wordt een methode voor basislijn correctie van eendimensionale data gepresenteerd. De voorgestelde methode gaat uit van een mengselmodel met twee componenten: een voor de basislijn inclusief ruis en een voor de pieken. Door uit te gaan van een normale verdeling voor eerste component kan een gladde lijn geschat worden. De schatter wordt verkregen met behulp van P-splines. Typische toepassingen zijn te vinden in de chromatografie en de spectroscopie. Beide technieken worden veelvuldig gebruikt voor de identificatie van stoffen en het onderscheiden van individuele bestanddelen binnen een mengsel.

Hoofdstuk 3 bespreekt een uitbereiding van de methode zoals gepresenteerd in hoofdstuk 2, naar twee dimensies. Dit komt neer op het schatten van een vlak in plaats van een curve. De basisconfiguratie is als in hoofdstuk 2: een mengselmodel met twee componenten en P-splines voor de basislijn. Om het schatten in twee dimensies mogelijk te maken worden P-splines met tensorproducten gebruikt, zodat associaties in beide richtingen opgenomen worden. De besproken toepassing komt uit de femtochemie, waarbij gekeken wordt naar trillingen van moleculen op een tijdschaal van femtoseconden (10^{-15} s). Inherent aan dit experiment zijn artefacten in de data en dienen verwijdert te worden voor een optimale interpretatie van de experimentele resultaten.

In de literatuur is veel onderzoek te vinden omtrent het goed schatten van een basislijn. Met name voor de in hoofdstuk 2 besproken eendimensionale data zijn dan ook verscheidene alternatieve methoden beschikbaar. Echter Het voordeel van de hier voorgestelde procedure is dat de basislijn als statistisch model wordt beschouwd. Waarvan ruis en onzekerheden expliciet onderdeel van zijn. Hiernaast wordt de curve geschat met behulp van P-splines, wat zorgt voor een aanmerkelijke reductie van de omvang van het te schatten model.

Voor de signaaldata besproken in de hoofdstukken 2 en 3 is het schatten van een basislijn een mogelijk noodzakelijke voorbehandeling. Piek analyse is een volgende stap en deconvolutie is hiervoor een veelgebruikt instrument. De veronderstelling is dat de geobserveerde data het resultaat is van een convolutie van een ingangssignaal met een impulsresponsfunctie. Het ingangssignaal is het werkelijke onderliggende effect, maar dat vervolgens verstoord is door de impulsrespons. Het doel is om het ingangssignaal zo accuraat mogelijk te schatten. Wanneer de data slechts een beperkt aantal pieken vertoont, dan zal het te schatten ingangssignaal ook slechts enkele spikes tonen en spreken we van ijle deconvolutie. Hoofdstuk 4 bespreekt de ijle deconvolutie van eendimensionale signalen met behulp van een zogenaamde L_0-norm penalty. Omdat in veel gevallen zowel het ingangssignaal als de impulsrespons onbekend zijn, moeten deze beide geschat worden. Uit de getoonde toepassingen blijkt dat het voorgestelde recept zeer goed werkt. Het geschatte signaal is ijl, terwijl de aanwezige pieken, in combinatie met de geschatte impulsrespons, het signaal accuraat kunnen reproduceren.

Hoofdstuk 5 is wederom een uitbreiding van het voorgaande hoofdstuk en ook hier betreft het de gang van een dimensie naar twee dimensies. De toepassing komt uit de microscopie. Tevens wordt het schatten van de impulsrespons mogelijk voor alle unimodale vormen. Deconvolutie in twee dimensies wordt aangepakt door de data te vectorizeren.

In hoofdstuk 6 wordt een schijnbaar volledig ander thema besproken, namelijk het classificeren van reeksen tweedimensionale, min of meer ronde, vormen. Echter, doordat de vormen getransformeerd worden van het Cartesiaans stelsel naar polaire coördinaten, zijn we terug bij de analyse van signalen. Een drietal toepassingen wordt besproken. Het onderscheiden van een aandoening bij kinderen die zich uit in afwijkingen in de schedelvorm, het correct indelen van verschillende soorten diatomeeën en een paleontologishe toepassing waarbij het gaat om het onderscheiden van fossielen van knaagdieren.

Voor het classificeren kan gebruik gemaakt worden van signaalregressie met een binaire uitkomst. Na de transformatie wordt de data samengevat met gebruik van P-splines. Dit heeft het voordeel dat datapunten met onderling ongelijke afstanden samengevat worden op een laagdimensionaal raster van splines. Door de penalty worden ruis en kleine gebreken eenvoudig gladgestreken. De splinebasis en de penalty zijn aangepast aan het circulaire karakter van de data. Eventuele verschillen in de oriëntatie van de vormen kunnen simpel gecorrigeerd worden door middel van rotatie. Elegant aan het voorgestelde model is dat de analyse van vormen onderbrengt binnen het gegeneraliseerd lineair model (GLM). De geschatte coëfficienten zijn eenvoudig te interpreteren en door ze te presenteren langs de originele vorm, zijn de sterkst discriminerende gebieden eenvoudig aan te wijzen. Een logische uitbreiding van het model is het analyseren van driedimensionale vormen. Afhankelijk van de te analyseren vorm kan hierbij gebruik gemaakt worden van bijvoorbeeld cilinder- of bolcoördinaten.

Centraal in hoofdstuk 7 staat het schatten van een netwerk van genen. Gebleken is dat de activiteit van genen te koppelen is aan ziektebeelden. Ook wordt verondersteld dat genen niet individueel opereren, maar clusters vormen, waarbinnen zij elkaar beïnvloeden. De clusters kunnen gepresenteerd worden als een netwerk. Voor het schatten van de netwerkstructuur kan gebruik gemaakt worden van partiële correlaties. Echter, omdat het aantal genen vaak vele malen groter is dan de omvang van de getrokken groep, is een dergelijke dataset een duidelijk voorbeeld van een slecht geconditioneerd probleem. Het gebruik van een penalty is hierbij een van de mogelijkheden om toch een netwerk structuur te kunnen schatten. Een aantal mogelijke penalty′s worden besproken en een drietal wordt nader bekeken met behulp van simulatie. Hierbij blijkt dat de adaptieve lasso het beste presteert. Veel informatie uit eerder onderzoek op het gebied van de genetica wordt opgeslagen in online databanken. Met deze gegevens kunnen bestaande netwerk structuren getoetst worden, of kunnen op basis van deze voorkennis bepaalde relaties meer of minder gepenaliseerd worden.

In het tweede deel van hoofdstuk 7 presenteren we een eenvoudige methode om de a priori-kennis te combineren met de data. De methode wordt geïllustreerd aan de hand van een tweetal voorbeelden. In de eerste toepassing wordt een prior gemaakt met behulp van een text mining algoritme. In het tweede voorbeeld wordt uitgegaan van bevindingen van een eerdere publicatie.

Hoofdstuk 8 behandelt geen hoogdimensionale data, maar reeksen kruistabellen waarbij in sommige cellen zeer weinig of geen observaties aanwezig zijn. In dergelijke gevallen is het lastig zo niet onmogelijk om uitspraken te doen op basis van de data. De gepresenteerde oplossing corrigeert de lage tellingen op het niveau van de individuele cellen. De voorgestelde correcties zijn verkregen op basis van een reeks tabellen, onderdeel van bijvoorbeeld een meta analyse.

De verwachtte tellingen per cel, gegeven de meta studie, kunnen op verschillende wijze berekend worden. Het krimpen richtingde voorgestelde correctie is een lineaire procedure, waarbij de optimale balans tussen data en correctie vastgesteld wordt met behulp van een aangepast informatie criterium van Akaike. Simulaties tonen aan dat op het niveau van individuele studies het toepassen van de correctie tot bevredigende resultaten leidt.

In het laatste hoofdstuk, wordt nogmaals stilgestaan bij de voornaamste bevindingen uit voorgaande hoofdstukken. Tevens worden penalty's beschouwd vanuit verschillende gezichtspunten. Ook worden nieuwe vragen opgeworpen, die een reeks handvatten vormen voor nader onderzoek en mogelijk nieuwe toepassingen.

List of publications

Refereed journal papers

Gravendeel, L. A. M., Kouwenhoven, M. C. M., Gevaert, O., de Rooi, J. J., Stubbs, A. P., Duijm, J. E., Daemen, A., Bleeker, F. E., Bralten, L. B. C., Kloosterhof, N. K., De Moor, B., Eilers, P. H. C., van der Spek, P. J., Kros, J. M., Sillevis Smitt, P. A. E., van den Bent, M. J. and French, P. J. (2009). Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Research*, 69(23):9065–9072.

Homminga, I., Pieters, R., Langerak, A. W., de Rooi, J. J., Stubbs, A., Verstegen, M., Vuerhard, M., Buijs-Gladdines, J., Kooi, C., Klous, P., van Vlierberghe, P., Ferrando, A. A., Cayuela, J. M., Verhaaf, B., Beverloo, H. B., Horstmann, M., de Haas, V., Wiekmeijer, A., Pike-Overzet, K., Staal, F. J. T., de Laat, W., Soulier, J., Sigaux, F., and Meijerink, J. P. P. (2011). Integrated transcript and genome analyses reveal nkx2-1 and mef2c as potential oncogenes in t cell acute lymphoblastic leukemia. *Cancer Cell*, 19(4):484–497.

de Rooi, J. and Eilers, P. (2011). Deconvolution of pulse trains with the L_0 penalty. *Analytica Chimica Acta*, 705(1–2):218–226.

Gravendeel, L. A. M., de Rooi, J. J., Eilers, P. H. C., van den Bent, M. J., Sillevis Smitt, P. A. E., and French, P. J. (2012). Gene expression profiles of gliomas in formalin-fixed paraffin-embedded material. *Br J Cancer*, 106:538–545.

de Rooi J. J., and Eilers, P. H. C. (2012). Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, 117:56–60.

Erdem-Eraslan, L., Gravendeel, L. A. M., de Rooi, J., Eilers, P. H. C., Idbaih, A., Spliet, W. G. M., den Dunnen, W. F. A., Teepen, J. L., Wesseling, P., Sillevis Smitt, P. A. E., Kros, J. M., Gorlia, T., van den Bent, M. J., and French, P. J. (2013). Intrinsic Molecular Subtypes of Glioma Are Prognostic and Predict Benefit From Adjuvant Procarbazine, Lomustine, and Vincristine Chemotherapy in Combination With Other Prognostic Factors in Anaplastic Oligodendroglial Brain Tumors: A Report From EORTC Study 26951. *Journal of Clinical Oncology*, 31(3):328–336.

de Rooi, J. J., Devos, O., Sliwa, M., Ruckebusch, C., and Eilers, P. H. C. (2013). Mixture models for two-dimensional baseline correction, applied to artifact elimination in timeresolved spectroscopy. *Analytica Chimica Acta*, 771:7-13.

de Rooi, J. (2013). Book review: Chemometrics with R. *Journal of Chemometrics*, 27:141–142.

Rijken, B. F. M., Lequin, M. H., de Rooi, J. J., van Veelen, M-L., and Mathijssen, I. M. J. Foramen magnum size and involvement of its intra-occipital synchondroses in Crouzon syndrome. *Plastic and Reconstructive Surgery, accepted*,

van den Bent, M. J., Erdem-Eraslan, L., Idbaih, A., de Rooi, J., Eilers, P. H. C., Spliet, W. G. M., den Dunnen, W. F. A., Tijssen, C., Wesseling, P., Sillevis Smitt, P. A. E., Kros, J. M. Gorlia, T., and French.P. J. MGMT-STP27 methylation status as predictive marker for response to PCV in anaplastic oligodendrogliomas. A report from EORTC study 26951. *Clinical Cancer Research, accepted.*

Zuurbier, L., Gutierrez, Mulighan, C. G., de Rooi, J. J., Smits, W. K., Sonneveld, E., Horstmann, M., Look, A. T., Pieters, R., and Meijerink, J. P. P. Characterization of immature t-cell acute lymphoblastic leukemia: Etp-all, immature cluster cases and patients lacking bi-allelic trg@ deletions. *Haematologica, accepted.*

de Rooi J. J., and Eilers, P. H. C. Classification of almost convex outlines with penalized signal regression. *submitted.*

de Rooi J. J., and Eilers, P. H. C. Pseudo-Bayes smoothing in tables with very low counts. *submitted.*

de Rooi J. J., and Eilers, P. H. C. Constrained estimation for sparse deconvolution in one and two-dimensions. *submitted.*

de Rooi, J. J., Böttger, A. J., Delhez, R., van der Pers, N. M., Hendrikx, R. W. A., and Eilers, P. H. C. Smoothing of X-ray diffraction scans and $K_{\alpha 2}$ elimination using penalized likelihood and the composite link model. *submitted.*

de Rooi J. J., and Eilers, P. H. C., Stubbs, A., and van der Spek, P. Exploring prior structures in gene expression data. *in preparation.*

Conference papers

de Rooi J. J., and Eilers, P. H. C. (2008). Smoothing zeros and small counts in meta analysis of clinical trials. In *Proceedings of the 23th international workshop on statistical modelling.*

de Rooi J. J., and Eilers, P. H. C. (2010). Recovering gene-networks using l_1 and l_0 penalties. In *Proceedings of the 25th international workshop on statistical modelling.*

de Rooi J. J., and Eilers, P. H. C. (2011). Using text mining tools to compose structure priors for inferring gene networks. In *Proceedings of the 26th international workshop on statistical modelling.*

Phd portfolio summary

Presentations

- Smoothing zeros in meta analysis of clinical trials, *23th International Workshop Statistical Modeling*, Utrecht, The Netherlands, July 7, 2008.
- Sparse network estimation with microarray expression data, *26th Annual Symposium on Chemometrics*, Utrecht, The Netherlands, May 20, 2010.
- Recovering gene regulatory networks using L_1 and L_0 penalties, *The 7th International Symposium on Networks in Bioinformatics*, Amsterdam, The Netherlands, April 22, 2010.
- Discovering gene regulatory networks using L_1 and L_0 penalties, *31st Annual Conference of the International Society for Clinical Biostatistics*, Montpelier, August 30, 2010.
- Priors from text mining for gene network inference, *26th International Workshop Statistical Modeling*, Valencia, Spain, July 14, 2011.
- Classifying shape outlines using polar coordinates and signal regression, *International Biometric Society Channel Network 3rd conference*, Bordeaux, France, April 12, 2011.
- Deconvolution as variable selection, *Spring Symposium in Biostatistics*, Rotterdam, The Netherlands, March 9, 2012.
- Classification of rounded shapes with penalized signal regression, *The 1st Annual Conference of the Dutch/Flemish Classification Society*, Tilburg, The Netherlands, May 25, 2012.
- Sparse deconvolution in 1d and 2d, using penalized regression, *XIII Chemometrics in Analytical Chemistry*, Budapest, Hungary, June 27, 2012.
- Penalized decomposition of Kα_1, Kα_2 X-ray diffraction profiles, *13th Scandinavian Symposium on Chemometrics*, Stockholm archipelago, Sweden, June 20, 2013.

Conferences

- 23th International Workshop on Statistical Modelling, Utrecht, The Netherlands, 2008.
- 2nd International Biometric Society Channel Network, Gent, Belgium, 2009.
- NDNS+ workshop: High dimensional inference and complex data, Groningen, The Netherlands, 2009.
- 25th International Workshop on Statistical Modelling, Glasgow, Scotland, 2010.
- 7th International Symposium on Networks in Bioinformatics, Amsterdam, The Netherlands, 2010.
- 31rd Annual Conference of the International Society for Clinical Biostatistics, Montpellier, France, 2010.
- 26th International Workshop on Statistical Modelling, Valencia, Spain, 2011.
- 3rd International Biometric Society Channel Network, Bordeaux, France, 2011.
- 5th International Chemometrics Research Meeting, Nijmegen, The Netherlands, 2011.
- Conference of the Dutch/Flemish Classification Society (VOC), Tilburg, The Netherlands, 2012.
- XIII Chemometrics in Analytical Chemistry, Budapest, Hungary, 2012.
- Chemometrics in time-resolved and imaging spectroscopy, Lille, France, 2012.
- 13th Scandinavian Symposium on Chemometrics, Stockholm archipelago, Sweden, 2013.

PhD training

- SNP's and human diseases, MolMed, Erasmus Medical Center, 2008.
- Pattern recognition, Netherlands Bioinformatics Centre (NBIC), 2009.
- Repeated measures analysis, Nihes, Erasmus Medical Center, 2010.
- Optimization techniques, Netherlands Bioinformatics Centre (NBIC), 2010.
- The craft of smoothing, Nihes, Erasmus Medical Center, 2010.
- Frailty models, Nihes, Erasmus Medical Center, 2009.
- Multistate models, Nihes, Erasmus Medical Center, 2009.
- Bayesian Methods and Bias Analysis, Department of Biostatistics, Erasmus Medical Center, 2010.

Teaching

- Classical Methods (practicals), Nihes, 2008/2009-2012.
- Modern Methods (practicals), Nihes, 2008.
- Basic analysis of gene expression (lecture), MolMed, 2009/2010.
- Introduction to R, MolMed, 2010-2012.
- Vaardighedenonderwijs, onderdeel statistiek (tutorials), Nihes, 2010/2011/2013.
- Analysis of microarray gene expression data (practicals), MolMed, 2011.

Over de auteur

De auteur van dit proefschrift werd geboren op 7 juni 1981 te Woerden. Hij groeide op in Nieuwveen en volgde middelbaar onderwijs aan het Groene Hart Lyceum te Alphen aan den Rijn. Vervolgens bezocht hij het ROC Utrecht waar hij, gedreven door zijn belangstelling voor techniek, koos voor de richting motorvoertuigentechniek. Na het behalen van dit diploma werkte hij als monteur, maar besloot al snel om terug te keren naar de schoolbanken.

Een jaar hoger beroepsonderwijs werd opgevolgd door de opleiding Sociologie aan de Universiteit Utrecht. Tijdens deze studie onstond de interesse in statistiek. Met het behalen van het bachelordiploma, werd hij in 2006 aangenomen bij de master Methods and Statistics of the Social and Behavioural Sciences. Na het succesvol afronden van deze studie start de auteur in 2008 een promotietraject bij de afdelingen Bioinformatica en Biostatistiek van het Erasmus Medisch Centrum te Rotterdam. Ondanks omzwervingen is de belangstelling voor techniek gebleven, wat onder andere tot uitdrukking komt in het onderzoek omtrent signaaldata en statistische vraagstukken binnen de analytische chemie.

Momenteel werkt de auteur aan analysemethoden voor röntgendiffractiedata, binnen een gezamenlijk project van de afdeling Biostatistiek van het Erasmus Medisch Centrum Rotterdam en het departement Technische Materiaalwetenschappen van de Technische Universiteit Delft.

Dankwoord

Bij het afronden van dit proefschrift wil ik graag een aantal mensen bedanken. Peter van der Spek en Andrew Stubbs van de afdeling Bioinformatica en Emmanuel Lesaffre van de afdeling Biostatistiek, voor de geboden kans om binnen deze departementen een proefschrift te schrijven.

Paul Eilers, beste Paul, jij staat aan de basis van dit proefschrift. Onze samenwerking is gestart tijdens mijn masteropleiding in Utrecht en we hebben deze vervolgens voortgezet in Rotterdam. Een gedeelde interesse in techniek heeft tot interssant onderzoek en boeiende discussies geleid. Naast de inhoudelijke raakvlakken hebben we ook zeer plezierig samengewerkt. Je hebt dikwijls gesteld dat promoveren ook gewoon leuk moet zijn en ik denk dat dit aardig gelukt is. Omdat je na je pensioen toch niet gaat tuinieren, kunnen we misschien nog wat projecten afronden.

Mijn collega's van Biostatistiek, waarbij in het bijzonder mijn oud-kamergenoten Karolina, Baoyue en Siti. En mijn huidige kamergenoten Kazem en Nicole. Hierbij valt ook Ralph te noemen, collega op menig congres en tafelgenoot in diverse restaurants welke moeiteloos passen in het oeuvre van Hopper.

In de periode bij het Erasmus Medisch Centrum zijn er een aantal samenwerkingen ontstaan die ik hier graag wil vermelden. Pim French en de collega's van het neurooncologie laboratorium. Pim, Lonneke, Nanne en Lale, naast een productieve samenwerking was het altijd leuk om elkaar even te spreken. Te denken valt ook aan alle keren waarbij ik gewoon op zoek was naar iets anders dan automatenkoffie. Beste Jules, bedankt voor je rol binnen mijn promotiecommissie. In onze gesprekken kwam jij regelmatig met een bijna bedwelmende hoeveelheid biologie, ik heb daar veel van opgestoken, maar ik vrees toch dat ik ook veel ben vergeten. Dear Cyril and Olivier, thanks for the fruitful collaboration, stimulating discussions and warm welcome in Lille. Cyril, I am very pleased to have you in the committee. Beste Terence, hoofdstuk zes van mijn thesis is terug te voeren op jouw vragen. We moeten hier in de komende tijd nog een vervolg aan geven. Hakim, jouw initiatief heeft geleid tot een leuk gezamenlijk project, waar bovendien nog vele interessante zaken aan onderzocht kunnen worden.

Mijn dank gaat ook uit naar Amarante Böttger, Niek van der Pers, Rob Delhez en Ruud Hendrikx van het departement Technische Materiaalwetenschappen van de TU Delft, voor het getoonde vertrouwen en geboden vrijheid binnen ons gezamenlijk project.

Ik maak ook graag van de gelegenheid gebruik om een aantal mensen te bedanken die niet zozeer rechtstreeks bij mijn promotietraject betrokken zijn geweest, maar elders op de route belangrijk zijn geweest of nog zijn. Vrienden uit de studietijd in Utrecht: Theo, Ellen, Ditza, Ismini en Liesbeth, jullie hebben deze tijd gemaakt tot iets waar ik met plezier op terug kijk. Nijs, jij bent, via mijn studentenbaan bij jou op de afdeling Methoden en Statistiek, tenslotte mede 'schuldig' aan het feit ik in de statistiek ben beland. Marco Fransen, het is plots lang geleden dat ik regelmatig bij jou over de vloer kwam, maar je hebt zeker invloed gehad op de keuzes die ik destijds gemaakt heb.

Mijn familie, het is altijd fijn om thuis in Nieuwveen te zijn. Om iedereen te zien en de drukte van de stad in te wisselen voor de rust van het platteland. Te genieten van de tuin en de goede zorgen. In het bijzonder 'de broertjes': typetjes, imitaties of gewoon een slechte actiefilm, met jullie is het altijd lol.

Tenslotte Merel, bedankt voor het aanhoren van diverse betogen, voor de discussies, gesprekken en voor alle steun in de jaren dat we samen zijn. Hopelijk kunnen we nog vele jaren door met het plezier dat we hebben en de interesses die we delen.

Johan de Rooi
Rotterdam, september 2013

Bibliography

Acharya, T. and Ray, A. (2005). *Image processing: principles and applications*. John Wiley, Hoboken.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on automatic control*, 19:716–723.

Aloise, S., Pawlowska, Z., Ruckebusch, C., Sliwa, M., Dubois, J., Poizat, O., Buntinx, G., Perrier, A., Maurel, F., Jacques, P., Malval, J. P., Poisson, L., Piani, G., and Abe, J. (2012). A two-step ICT process for solvatochromic betaine pyridinium revealed by ultrafast spectroscopy, multivariate curve resolution, and TDDFT calculations. *Physical Chemistry Chemical Physics*, 14:1945–1956.

Aloise, S., Ruckebusch, C., Blanchet, L., Rehault, J., Buntinx, G., and Huvenne, J.-P. (2008). The benzophenone S_1 (n,π^*) $\rightarrow$ T_1 (n,π^*) states intersystem crossing reinvestigated by ultrafast absorption spectroscopy and multivariate curve resolution. *The Journal of Physical Chemistry A*, 112(2):224–231.

Alsberg, B. K., Woodward, A. M., and Kell, D. B. (1997). An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics and Intelligent Laboratory Systems*, 37:215–239.

Altman, D. and Deeks, J. (2002). Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Medical Research Methodology*, 2(1):3.

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Azuaje, F. (2010). What does systems biology mean for biomarker discovery? *Expert Opinion on Medical Diagnostics*, 4(1):1–10.

Bandzůch, P., Morháč, M., and Krištiak, J. (1997). Study of the van cittert and gold iterative methods of deconvolution and their application in the deconvolution of experimental spectra of positron annihilation. *Nuclear Instruments and Methods in Physics Research A*, 384:506–515.

Barabasi, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288:60–69.

Baron, D., Bihouée, A., Teusan, R., Dubois, E., Savagner, F., Steenman, M., Houlgatte, R., and Ramstein, G. (2011). Madgene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets. *Bioinformatics*, 27(5):725–726.

Belghith, A., Collet, C., Rumbach, L., and Armspach, J.-P. (2011). A unified framework for peak detection and alignment: application to hr-mas 2d nmr spectroscopy. *Signal, Image and Video Processing*, pages 1–10.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis*. The MIT Press, Cambridge, Massachusetts.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis*. Springer, New York.

Bleakley, K., Biau, G., and Vert, J. (2007). Supervised reconstruction of biological networks with local models. *Bioinformatics*, 23:57–65.

Blumensath, T. and Davies, M. E. (2008). Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14:629–654.

Bowie, B. T., Chase, D. B., Lewis, I. R., and Griffiths, P. R. (2006). Anomalies and artifacts in raman spectroscopy. In *Handbook of Vibrational Spectroscopy*. John Wiley & Sons, Ltd.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Localio, A. R. (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, 26(1):53–77.

Bredel, M., Bredel, C., Juric, D., Harsh, G., Vogel, H., Recht, L., and Sikic, B. (2005). Functional network analysis reveals extended gliomagenesis pathway maps and three novel myc-interacting genes in human gliomas. *Cancer research*, 65(19):8679–8689.

Brown, F. T. (2006). *Engineering System Dynamics*. CRC Press, New York.

Bruckstein, A. M., Donoho, D. L., and Elad, M. (2009). From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81.

Burman, P. (2004). On some testing problems for sparse contingency tables. *Journal of multivariate analysis*, 88:1–18.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference*. Springer, New York.

Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905.

Cardot, H., Koo, J.-Y., Park, H. J., and Trubuil, A. (2004). Boosting diracs for eletrophoresis. *Journal of Computational and Graphical Statistics*, 13(3):659–673.

Cates, C. (2002). Simpson's paradox and the calculation of number needed to treat from meta-analysis. *BMC Medical Research Methodology*, 2:1.

Chen, S., Donoho, D., and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computing*, 20(1):33–61.

Cheng, D., Bainbridge, D., Martin, J., and Novick, R. (2005). The evidence-based perioperative clinical outcomes research group. does off-pump coronary artery bypass reduce mortality, morbidity, and resource utilization when compared with conventional coronary artery bypass? A meta-analysis of randomized trials. *Anesthesiology*, 102(1):188–203.

Claude, J. (2008). *Morphometrics with R*. Springer, New York.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Coombes, K. R., Fritsche, H. A., Clarke, C., neng Chen, J., Baggerly, K. A., Morris, J. S., chun Xiao, L., Hung, M.-C., and Kuerer, H. M. (2003). Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clinical Chemistry*, 49(10):1615–1623.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C., and Kuerer, H. M. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5:4107–4117.

Coote, G. E. (1997). Iterative smoothing and deconvolution of one- and two-dimensional elemental distribution data. *Nuclear Instruments and Methods in Physics Research B*, 130:118–122.

Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59.

Currie, I. D., Durban, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):259–280.

Dai, B. and Eads, C. D. (2010). Efficient removal of unwanted signals in nmr spectra using the filter diagonalization method. *Magnetic Resonance in Chemistry*, 48:230–234.

Dantus, M. and Gross, P. (2003). *Encyclopedia of Applied Physics*, chapter Ultrafast Spectroscopy. Wiley-VCH, Berlin.

de Boor, C. (2001). *A practical guide to splines*. Applied mathematical sciences. Springer, New York.

de Noo, M. E., Tollenaar, R. A. E. M., Özalp, A., Kuppen, P. J. K., Bladergroen, M. R., Eilers, P. H. C., and Deelder, A. M. (2005). Reliability of human serum protein profiles generated with c8 magnetic beads assisted maldi-tof mass spectrometry. *Analytical Chemistry*, 77(22):7232–7241.

de Rooi, J., Devos, O., Sliwa, M., Ruckebusch, C., and Eilers, P. (2013). Mixture models for two-dimensional baseline correction, applied to artifact elimination in timeresolved spectroscopy. *Analytica Chimica Acta*, 771:7–13.

de Rooi, J. and Eilers, P. (2011). Deconvolution of pulse trains with the L_0 penalty. *Analytica Chimica Acta*, 705:218–226.

de Rooi, J. J. and Eilers, P. H. C. (2012). Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, 117:56–60.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Devos, O., Mouton, N., Sliwa, M., and Ruckebusch, C. (2011). Baseline correction methods to deal with artifacts in femtosecond transient absorption spectroscopy. *Analytica Chimica Acta*, 705:64–71.

Dierckx, P. (1995). *Curve and surface fitting with splines*. Monographs on numerical analysis. Oxford University Press, Oxford.

Dryden, I. and Mardia, K. (1998). *Statistical shape analysis*. John Wiley, Hoboken.

Du, P. and Angeletti, R. (2006). Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Analytical chemistry*, 78(10):3385–3392.

Du, P., Kibbe, W. A., and Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continious wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065.

Eilers, P. (2004). Parametric time warping. *Analytical Chemistry*, 76(2):404–411.

Eilers, P. and de Menezes, R. X. (2005). Quantile smoothing of array cgh data. *Bioinformatics*, 21(7):1146–1153.

Eilers, P. and Marx, B. (2010). Splines, knots and penalties. *WIREs Computational Statistics*, 2:637–653.

Eilers, P. H. C. (1996). Sparse contingency tables, pseudo-Bayes estimates and cross-validation. In *Proceedings of the 11th international workshop on statistical modelling*, pages 402–405.

Eilers, P. H. C. (2003). A perfect smoother. *Analytical Chemistry*, 75:3631–3636.

Eilers, P. H. C. (2005). Unimodal smoothing. *Journal of chemometrics*, 19(5-7):317–328.

Eilers, P. H. C. and Boelens, H. (2005). Baseline correction with asymmetric least squares smoothing. *Unpublished manuscript*.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50:61–76.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102.

Elad, M., Matalon, B., Shtok, J., and Zibulevsky, M. (2007). A wide-angle view at iterated shrinkage algorithms. In *SPIE (Wavelet XII)*.

Ernsting, N. P., Kovalenko, S. A., Senyushkina, T., Saam, J., and Farztdinov, V. (2001). Wave-packet-assisted decomposition of femtosecond transient ultraviolet-visible absorption spectra: Application to excited-state intramolecular proton transfer in solution. *Journal of Physical Chemistry A*, 105:3443–3453.

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

Faro, A., Giordano, D., and Spampinato, C. (2012). Combining literature text mining with microarray data: advances for system biology modeling. *Briefings in bioinformatics*, 13(1):61–82.

Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American statistical association*, 68(343):683–691.

Frasso, G. and Eilers, P. H. C. (2012). Smoothing parameter selection using the l-curve. In *Proceedings of the 27th international workshop on statistical modelling*, pages 402–405.

Friedman, J. (1997). On bias, variance, 0/1–loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77.

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

Friedman, N. I. R., Linial, M., and Nachman, I. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620.

Friedrich, J. O., Adhikari, N. K. J., and Beyene, J. (2007). Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Medical Research Methodology*, 7(1):5.

Fuente, A. D. L., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.

Galloway, C. M., Le Ru, E. C., and Etchegoin, P. G. (2009). An iterative algorithm for background removal in spectroscopy by wavelet transforms. *Applied Spectroscopy*, 63:1370–1376.

Gan, F., Ruan, G., and Mo, J. (2006). Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82:59–65.

Gans, P. and Gill, J. B. (1984). Smoothing and differentiation of spectroscopic curves using spline functions. *Applied Spectroscopy*, 38:370–376.

Gasso, G., Rakotomamonjy, A., and Canu, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698.

Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y., and De Moor, B. (2004). Txtgate: profiling gene groups with text-based information. *Genome biology*, 5(6):R43.

Goeman, J. J. (2009). L_1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 51:1–15.

Gravendeel, L., Kouwenhoven, M., Gevaert, O., de Rooi, J., Stubbs, A., Duijm, J., Daemen, A., Bleeker, F., Bralten, L., Kloosterhof, N., de Moor, B., Eilers, P., van der Spek, P., Kros, J., Sillevis Smitt, P., van den Bent, M., and French, P. (2009). Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer research*, 69(23):9065–9072.

Greenland, S. (2006). Smoothing observational data: A philosophy and implementation for the health sciences. *International Statistical Review*, 74(1):31–46.

Greenland, S. (2010). Simpsons paradox from adding constants in contingency tables as an example of bayesian noncollapsibility. *The American Statistician*, 64(4):340–344.

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

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Taylor and Francis, Boca Raton.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York.

Herrity, K., Gilbert, A., and Tropp, J. (2006). Sparse approximation via iterative thresholding. In *Proceedings of international conference on acoustic, speech and signal processing*, volume 3, pages 624–627.

Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and l_1 penalized regression: A review. *Statistics Surveys*, 2:61–93.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58(3):54–59.

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

Hollas, J. (2004). *Modern spectroscopy*. Wiley, New York.

Hurvich, C., Simonoff, J., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, 60:271–293.

Jacques, J., Bouveyron, C., Girard, S., Devos, O., Duponchel, L., and Ruckebusch, C. (2010). Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Journal of Chemometrics*, 24:719–727.

Jalba, A., Wilkinson, M., Roerdink, J., Bayer, M., and Juggins, S. (2005). Automatic diatom identification using contour analysis by morphological curvature scale spaces. *Machine Vision and Applications*, 16(4):217–228.

Jansson, P. A., editor (1996). *Deconvolution of Images and Spectra.* Academic Press, San Diego.

Jenssen, T. K., Laegreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1):21–28.

JiJi, R. D. and Booksh, K. S. (2000). Mitigation of rayleigh and raman spectral interferences in multiway calibration of excitation-emission matrix fluorescence spectra. *Analytical Chemistry*, 72:718–725.

Junker, B. and Schreiber, F. (2008). *Analysis of Biological Networks*. Wiley, New York.

Kalivas, J. (2012). Overview of two-norm (l_2) and one-norm (l_1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *Journal of Chemometrics*, 26:218–230.

Kanehisa, M. and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560):1662–1664.

Komsta, Ł. (2011). Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia*, 73:721–731.

Krämer, N., Schäfer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(1):384.

Kuss, O. and Gromann, C. (2007). An exact test for meta-analysis with binary endpoints. *Methods of Information in Medicine*, 46:662–668.

L'Abbé, K., Detsky, A., and O'Rourke, K. (1987). Meta-analysis in clinical research. *Ann. Intern. Medicine*, 107:224–233.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Lee, H., Lee, D., Kang, H., Kim, B., and Chung, M. (2011). Sparse brain network recovery under compressed sensing. *IEEE Transactions on Medical Imaging*, 30(5):1154–1165.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182.

Li, L. and Speed, T. (2000). Parametric deconvolution of positive spike trains. *The Annals of Statistics*, 28(5):1279–1301.

Li, L. and Speed, T. (2004). Deconvolution of sparse positive spikes. *Journal of Computational and Graphical Statistics*, 13(4):853–870.

Lieber, C. A. and Mahadevan-Jansen, A. (2003). Automated method for subtraction of fluorescence from biological raman spectra. *Applied Spectroscopy*, 57:1363–1367.

Liland, K. H., Almøy, T., and Mevik, B. (2010). Optimal choice of baseline correction for multivariate calibration of spectra. *Applied Spectroscopy*, 64(9):1007–1016.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer, New York.

Loader, C. (1999). *Local regression and likelihood*. Statistics and computing. Springer, New York.

Lorenc, M., Ziolek, M., Naskrecki, R., Karolczak, J., Kubicki, J., and Maciejewski, A. (2002). Artifacts in femtosecond transient absorption spectroscopy. *Applied Physics B: Lasers and Optics*, 74(1):19–27.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press, Cambridge.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 15:1–15.

Mariscotti, M. A. (1967). A method for automatic identification of peaks in the presence of background and its application to spectrum analysis. *Nuclear instruments and methods*, 50:309–320.

Marx, B. D. and Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics*, 41(1):1–13.

Marx, B. D. and Eilers, P. H. C. (2002). Multivariate calibration stability: a comparison of methods. *Journal of Chemometrics*, 16(3):129–140.

McCullagh, P. and Nelder, J. (2000). *Generalized linear models*. Champman and Hall/CRC, Boca Raton.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

McNair, H. and Miller, J. (1998). *Basic gas chromatography*. Wiley, New York.

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Morháč, M., Kliman, J., Matoušek, V., Veselskỳ, M., and Turzo, I. (1997). Efficient one-and two-dimensional gold deconvolution and its application to γ-ray spectra decomposition. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 401(2):385–408.

Morháč, M. (2007). Multidimensional peak searching algorithm for low-statistics nuclear spectra. *Nuclear Instruments and Methods in Physics Research A*, 581:821–830.

Morháč, M. (2009). An algorithm for determination of peak regions and baseline elimination in spectroscopic data. *Nuclear Instruments and Methods in Physics Research A*, 600:478–487.

Morháč, M. and Matoušek, V. (2011). High-resolution boosted deconvolution of spectroscopic data. *Journal of Computational and Applied Mathematics*, 235:1629–1640.

Mosierboss, P. A., Lieberman, S. H., and Newbery, R. (1995). Fluorescence rejection in raman-spectroscopy by shifted-spectra, edge-detection, and FFT filtering techniques. *Applied Spectroscopy*, 49:630–638.

Mouton, N., Sliwa, M., Buntinx, G., and Ruckebusch, C. (2010). Deconvolution of femtosecond time-resolved spectroscopy data in multivariate curve resolution. application to the characterization of ultrafast photo-induced intramolecular proton transfer. *Journal of Chemometrics*, 24:424–433.

Osborne, M. R., Presnell, B., and Turlach, B. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Paul, M., Riebler, A., Bachmann, L., Rue, H., and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested laplace approximations. *Statistics in Medicine*, 29(12):1325–1339.

Phillips, A. J. and Hamilton, P. A. (1996). Improved detection limits in fourier transform spectroscopy from a maximum entropy approach to baseline estimation. *Analytical Chemistry*, 68(22):4020–4025.

Prossinger, H. (2005). *Modern morphometrics in physical anthropology*, chapter Problems with landmark-based morphometrics for fractal outlines: the case of frontal sinus ontogeny, pages 167–185. Kluwer Academic.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramsay, J. and Silverman, B. (2002). *Applied functional data analysis*. Springer, New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer, New York.

Reiss, D. J., Facciotti, M. T., and Baliga, N. S. (2008). Model-based deconvolution of genome-wide dna binding. *Bioinformatics*, 24:396–403.

Renard, B., Kirchner, M., Steen, H., Steen, J., and Hamprecht, F. (2008). Nitpick: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9(1):355.

Ressom, H. W., Varghese, R. S., Drake, S. K., Hortin, G. L., Abdel-Hamid, M., Loffredo, C. A., and Goldman, R. (2007). Peak selection from maldi-tof mass spectra using ant colony optimization. *Bioinformatics*, 23(5):619–626.

Reverter, A. and Chan, E. K. F. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, 24(21):2491–2497.

Rinnan, Å., Berg, F. v. d., and Engelsen, S. B. (2009). Review of the most common preprocessing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28:1201–1222.

Rinnan, Å., Booksh, K. S., and Bro, R. (2005). First order rayleigh scatter as a separate component in the decomposition of fluorescence landscapes. *Analytica Chimica Acta*, 537:349–358.

Rippe, R., Meulman, J., and Eilers, P. (2012). Visualization of genomic changes by segmented smoothing using an l_0 penalty. *PloS one*, 7(6):e38230.

Robins, J., Greenland, S., and Breslow, N. E. (1986). A general estimator for the variance of the Mantel-Haenszel odds ratio. *American Journal of Epidemiology*, 124:719–723.

Ruckebusch, C., Sliwa, M., Pernot, P., de Juan, A., and Tauler, R. (2012). Comprehensive data analysis of femtosecond transient absorption spectra: A review. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, 13:1–27.

Ruckebusch, C., Sliwa, M., Rehault, J., Naumov, P., Huvenne, J. P., and Buntinx, G. (2009). Hybrid hard- and soft-modelling applied to analyze ultrafast processes by femtosecond transient absorption spectroscopy: Study of the photochromism of salicylidene anilines. *Analytica Chimica Acta*, 642:228–234.

Rücker, G. and Schumacher, M. (2008). Simpson's paradox visualized: The example of the rosiglitazone meta-analysis. *BMC Medical Research Methodology*, 8(1):34.

Rücker, G., Schwarzer, G., Carpenter, J., and Olkin, I. (2009). Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, 28(5):721–738.

Ruckstuhl, A., Jacobson, M., Field, R., and Dodd, J. (2001). Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric regression*, volume 12. Cambridge University Press, New York.

Ryan, C., Clayton, E., Griffin, W., Sie, S., and Cousens, D. (1988). Snip, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 34:396–402.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Seul, M., O'Gorman, L., and Sammon, M. (2000). *Practical algorithms for image analysis*. Cambridge University Press, New York.

Shenoi, B., Wilkins, C., and Lay, J. (2006). *Introduction to digital signal processing and filter design*. Wiley, New York.

Shimamura, T., Imoto, S., Yamaguchi, R., and Miyano, S. (2007). Weighted lasso in graphical gaussian modeling for large gene network estimation based on microarray data. *Genome Inform*, 19:142–153.

Silagadze, Z. K. (1996). A new algorithm for automatic photopeak searches. *Nuclear Instruments and Methods in Physics Research A*, 376:451–454.

Silvia, M. and Robinson, E. (1979). *Deconvolution of geophysical time series in the exploration for oil and natural gas*. Developments in petroleum science. Elsevier, Amsterdam.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer, New York.

Sliwa, M., Mouton, N., Ruckebusch, C., Poisson, L., Idrissi, A., Aloise, S., Potier, L., Dubois, J., Poizata, O., and Buntinx, G. (2010). Investigation of ultrafast photoinduced processes for salicylidene aniline in solution and gas phase: toward a general photo-dynamical scheme. *Photochemical & Photobiological Sciences*, 9:661–669.

Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., Ramsey, J., and Woolrich, M. (2011). Network modelling methods for fmri. *Neuroimage*, 54(2):875–891.

Starck, J. and Murtagh, F. (2006). *Astronomical image and data analysis*. Springer, New York.

Starck, J.-L., Pantin, E., and Murtagh, F. (2002). Deconvolution in astronomy: a review. *Publication of the Astronomical Society of the Pacific*, 114:1051–1069.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proc. Third Berkeley Symp. Math. Statist. Prob.*

Steuer, R., Kurths, J., Fiehn, O., and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026.

Stijnen, T., Hamza, T., and Özdemir, P. (2010). Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*, 29(29):3046–3067.

Stuart, J., Segal, E., Koller, D., and Kim, S. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10:277–303.

Sweeting, M. J., Sutton, A. J., and Lambert, P. C. (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9):1351–1375.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108.

Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Dokl.*, volume 4, pages 1035–1038.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer, New York.

Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investements - maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society: Series B*, 55(3):569–612.

Vickers, T. J., Wambles, R. E., and Mann, C. K. (2001). Curve fitting and linearity: Data processing in raman spectroscopy. *Applied Spectroscopy*, 55:389–393.

Vis, D., Westerhuis, J., Hoefsloot, H., Roelfsema, F., Hendriks, M., and Smilde, A. (2012). Detecting regulatory mechanisms in endocrine time series measurements. *PloS one*, 7(3):e32985.

Vis, D. J., Westerhuis, J. A., Hoefsloot, H. C. J., Pijl, H., Roelfsema, F., van der Greef, J., and Smilde, A. K. (2010). Endocrine pulse identification using penalized methods and a minimum set of assumptions. *Am J Physiol Endocrinol Metab*, 298(2):E146–E155.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18(2):223–250.

Wang, L. and Zhu, J. (2010). Image denoising via solution paths. *Annals of Operations Research*, 174(1):3–17.

Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry*, 75:4818–4826.

Watson-Haigh, N. S., Kadarmideen, H. N., and Reverter, A. (2010). Pcit: an r package for weighted gene co-expression networks based on partial correlation and information theory approaches. *Bioinformatics*, 26(3):411–413.

Webster, J. and Clark, J. (2010). *Medical instrumentation: application and design*. Wiley, New York.

Wei, L., Huang, Z., and Que, P. (2009). Sparse deconvolution method for improving the time-resolution of ultrasonic NDE signals. *NDT & E International*, 42(5):430–434.

Wentzell, P. D. and Brown, C. D. (2000). *Encyclopedia of Analytical Chemistry*, chapter Signal Processing in Analytical Chemistry, pages 9764–9800. Wiley, New York.

Whittaker, E. (1923). On a new method of graduation. In *Proceedings of the Edinburgh Mathematical Scociety*, volume 41, pages 63–75.

Witten, I. and Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Morgan Kaufman, San Francisco.

Xiang, L., Xunbo, L., Wei, L., and Liang, C. (2012). l_0- norm regularized minimum entropy deconvolution for ultrasonic ndt&e. *NDT & E International*, 47:80–87.

Yasui, Y., Pepe, M., Thompson, M. L., Adam, B.-L., Wright, G. L., Qu, Y., Potter, J. D., Winget, M., Thornquist, M., and Feng, Z. (2003). A data analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4(3):449–463.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the journal of the Royal statistical society*, 1(2):217–235.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131.

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

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

Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57:4689–4708.

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

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

Subject index

Author index