Fusing prior knowledge with microbial metabolomics

Verouden, M.P.H.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Weighted Smooth Principal Component Analysis (WSPCA): validation and application to missing value estimation

†This chapter has been submitted for publication: Maikel P.H. Verouden, Johan A. Westerhuis and Age K. Smilde Copyright © 2011
Abstract

In the various fields of functional genomics longitudinal data plays an important role to improve understanding and knowledge of the dynamics within biological systems. The time-resolved data are expected to contain underlying dynamic profiles that are smooth. However, estimating these underlying smooth dynamic phenomena from such data is complicated due to the high complexity of the data and the limited number of techniques that can deal with this type of data. Traditional multivariate data analytical techniques, such as Principal Component Analysis, ignore the underlying dynamics in the data and give solutions that tend towards explaining variance rather than explaining dynamics and understanding biology. In this paper we present WSPCA, a method that incorporates smoothness into the scores of PCA by using a roughness penalty. WSPCA can be used for data with consecutive samples that are linked by time (time-resolved) or position (spatially resolved). By means of a synthetic data set we show that applying this restriction leads to a better estimation of the dynamic phenomena underlying the data. For determination of the model meta parameters (the number of components and the smoothness parameter) we present a leave elements out cross-validation procedure, that for our synthetic data set is capable of estimating the underlying noiseless data. The WSPCA method and leave elements out cross-validation are then applied to a real-life metabolomics data set from a *E. coli* batch fermentation, that has been sampled over time, to estimate the missing elements in the data set.

**KEYWORDS:** PCA, smooth PCA, weighting, cross-validation, missing values.
5.1 Introduction

Time-resolved experiments play an increasingly important role within the various fields of functional genomics research to improve understanding of the dynamics within biological systems. The longitudinal data obtained from such experiments are expected to contain smooth dynamic profiles of interest. However, due to the high complexity of the data and the limited number of techniques for analyzing this type of data [132], it is hard to discern the underlying dynamic phenomena. Traditional multivariate data analytical techniques such as Principal Component Analysis (PCA) ignore the underlying dynamics in the data. PCA focusses on maximizing the explained variation, without taking into account the relation that exists between samples obtained at consecutive time points. Addition of prior information in the form of restrictions to these multivariate data analytical techniques can improve the biological interpretability of the created models and increase the predictive power of those models. Similar to the use of much used restrictions as non-negativity or unimodality in Multivariate Curve Resolution (MCR) we aim that a roughness penalty also leads the analysis of the data towards more “biological” and less “variance explained” solutions.

In this paper the goal is to show that incorporation of smoothness into the scores of PCA can lead to a better estimation of the dynamic phenomena underlying the data, when consecutive objects or samples are not independent of each other but are connected for example by time, i.e. time-resolved, or sample position, i.e. spatially resolved. The better estimation of the underlying longitudinal phenomena can e.g. be used to improve the estimation of missing values. Adding roughness penalties to data analysis methods is often used for smoothing data [133] and extensions of PCA with roughness penalties such as smoothed PCA [134][135] have been proposed in the context of functional data analysis [111]. However, these methods are designed for data where the variables have a functional relationship whereas our method focusses on multivariate data with the functional relationship between the samples. Note that simply taking the transpose of the data before using the FDA approach does not solve the problem of analyzing multivariate data with the functional relationship between the samples, as we would like to be able to predict new smooth scores for a new set of samples using the loadings of the model. Recently Yamamoto et al. introduced smooth PCA with smooth scores by using a generalized eigenvalue approach [136]. Their approach calculates the smooth score as a linear combination of the variables in the data whereas our WSPCA method does not use this restriction.

For the introduction of the weighted smooth PCA method and describing its properties we use a synthetic dataset, based on biologically relevant time profiles. In the simulation study we aim to show that, although the fit with respect to the analysed data becomes lower, the true underlying structure in the data is better captured. We aim to show that by restricting the PCA model with a roughness penalty, the well-known overfit of the method is reduced. By using cross-validation the optimal roughness penalty and number of components to be used in the model will be determined. The cross-validation is based on leaving multiple elements out.
5. Weighted Smooth Principal Component Analysis (WSPCA)

of the data and estimating them back with the created model. This approach creates the necessity for the smooth principal component analysis to be weighted. Next we will apply smoothness in combination with the same cross-validation procedure to microbial metabolomics data obtained from a *Escherichia coli* batch fermentation with missing values to estimate those missing values.

In Section 5.2 weighted smooth principal component analysis and the cross-validation procedure will be explained and discussed in detail. Section 5.3 describes how the simulated data has been created and how the *E. coli* batch fermentation metabolomics data has been obtained experimentally. The results are presented and discussed in Section 5.4. Finally in Section 5.5 the conclusions of our study will be presented.

5.2 Methods

5.2.1 Weighted Smooth Principal Component Analysis

Weighted Smooth Principal Component Analysis (WSPCA) is a penalized form of weighted principal component analysis (PCA W) [99, 103, 105], that can be used for multivariate data with objects or samples that are not independent of each other but are linked via a clear relationship with expected smoothness in the underlying phenomena. The consecutive objects or samples are for example time-resolved or spatially resolved forming longitudinal profiles. The variables in the data used for WSPCA can have relationships among each other but there is no expected smoothness between them. Metabolomics data from a microbial batch fermentation experiment sampled over time is an example of this type of data. In WSPCA the scores, calculated for successive samples, are penalized for their roughness, by means of a roughness penalty, while still providing a good description of the data. By implying smoothness on the scores in WSPCA we assume that the underlying latent components are smooth, not necessarily the manifest variables. The smoothed score profiles relate to the major dynamic events in the data set.

We use a weighted version of the smooth PCA model. Weighting can be used to incorporate prior information about the experimental error into the data analytical method [103, 105, 137]. Here the weighted approach is used to indicate the missing elements in the data to be estimated and for the selection of test and validation samples in the cross-validation of the model meta parameters. The weight matrix $W$ is constructed as a binary matrix containing ones (1) for elements that are present and zeros (0) for elements that are missing or have been excluded for cross-validation.

Weighted smooth PCA (WSPCA) is implemented using the solver for nonlinear least-squares problems (lsqnonlin) from the Optimization Toolbox (version 5.1) for Matlab® where the calculated scores $Z [I \times R]$ are penalized for their roughness. The sum of squares $Q$ (Equation 5.1) is minimized with respect to the smoothed scores and the corresponding loadings $P [J \times R]$ for a data set $X [I \times J]$, a given weight matrix $W [I \times J]$ and a given smoothing parameter $\lambda$ with $R$ representing the number of

---

*R2010b Version 7.11.0.584, The Mathworks®, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098, U.S.A.*
5.2. Methods

In Equation 5.1 the character \( \circ \) denotes the Hadamard or element-wise product. Matrix \( D_n \) represents the \( n \)th order difference matrix, which is used for the finite differences approximation of derivatives. Commonly a first or second order difference matrix is used, but difference matrices of higher order can also be applied. When \( D_n \) is set to 1st order (\( D_1 \)) the consecutive score values are penalized for their slope (the first derivative) and are forced to become equal, leading in the extreme (large \( \lambda \)) to straight flat lines with zero slope. With a 2nd order difference matrix (\( D_2 \)) the consecutive scores are penalized for the change in slope (the second derivative) and are forced to become straight lines with arbitrary slope in the extreme. The use of a first or second order difference matrix depends on the problem at hand, generally for removing noise a first order difference probably will suffice. However, if the expected smoothness in the underlying profiles has parabolic or higher order polynomial form a difference matrix of second or higher order might be more appropriate as a first order derivative has a tendency to quickly deteriorate the smoothness profiles by pushing them to straight flat lines with zero slope.

In our presented WSPCA method the only constraint is the roughness penalty on the scores. There is no active constraint on the loadings (\( P \)). We only normalize the loadings to length 1 and adapt the smoothed scores (\( Z \)) accordingly. This is necessary as else the scores would become very small and the roughness penalty would not have an effect. When \( \lambda \) is set to 0, the penalty will not be active and the obtained scores and loadings from weighted smooth PCA will be equal to or a rotated version of regular weighted PCA. If \( \lambda \) is increased, the obtained score vectors will become more smooth, however this will result in a reduced description of the variation in the data \( X \). The smoothing parameter \( \lambda \) determines the amount of emphasis that is put on the roughness penalty.

The smoothed scores (\( Z \)) obtained from minimization of Equation 5.1 are not orthogonal to each other, \( Z^T Z \) is a symmetric but not a diagonal matrix. The loadings are neither orthogonal to each other, \( P^T P \) is a symmetric matrix with ones on its diagonal but not an identity matrix of order \( R \). We did not want to enforce orthogonality onto the scores and loading, for reasons which will be clarified in the following section and, therefore, have implemented WSPCA as a method in which all components are estimated simultaneously.

**On the freedom for smoothed scores \( Z \) and loadings \( P \).**

Contrary to the common assumption of keeping the scores in the column space of \( X \) and loadings in the row space of \( X \), either as a feature of or as a restriction in the applied data analysis method, we do not impose this as a restriction in our method. This means that in WSPCA both can be varied freely. When the smoothing parameter (\( \lambda \)) is unequal to zero, the smoothed scores \( Z \) are not restricted to the range of \( X \). The space spanned by \( X \) depends highly on the noise realization and remeasured data usually spans a different space. In our opinion it is, therefore,
unwanted to constrain modelled information about the samples or objects (e.g. the score profiles) to being linear combinations of the columns space of $X$. In general smooth score profiles can not be found in the range of $X$, the optimal smooth solution lies outside that range.

In Burnham et al. [138] a statistical framework for Latent Variable Multivariate Regression (LVMR) is explored with various degrees of error variance in $X$ and $Y$. In this framework the estimates $\hat{T}$, whose columns provide a basis for the latent variable space, are restricted to $\hat{T} = X\hat{W}$, where $\hat{W}$ is a function of the training data. This restriction is also used in more commonly applied methods (e.g. PLS, PCR, RRR and CCR) and depending on the degree of error variance in $X$ and $Y$ the framework behaves like some of these methods. When the restriction is not applied the framework is referred to as maximum likelihood latent root regression (MLRR). In cases where the errors in $X$ are large relative to those in $Y$ the unrestricted version of the framework (MLRR) performs uniformly better. Some algorithms provide sub- and even final solutions (concentration and spectral profiles) that can be outside the range of the data matrix, without the user being aware of it, as was recently shown for the multivariate curve resolution-alternating least squares (MCR-ALS) algorithm [139][141].

5.2.2 Cross-validation procedure

Cross-validation is a resampling technique which is used to simplify the selection of model parameters, such as the number of components to use, and provides a basis for residual and influence analysis [142]. Here cross-validation will be applied to determine the model meta parameters, being the number of components ($R$) to use and the smoothing parameter ($\lambda$).

Most cross-validation methods leave out one or more complete objects or samples and predict them back to assess the model meta parameters. This approach can not be applied in case of WSPCA because leaving out one or more complete objects or samples would disrupt the relationship between the objects or samples, which are not independent of each other. Therefore, we have developed a cross-validation procedure in which randomly selected elements $x_{ij}$ are left out of the data matrix $X$. Our procedure has a similar approach as the cross-validation method for principal component analysis as proposed by Wold [143]. The user has to choose the number of elements to leave out, e.g. one could perform a 20 fold cross-validation by leaving out 5% of the elements at each round. A WSPCA model consisting of smoothed scores ($Z$) and loadings ($P$), with a pre-set number of components ($R$) and $\lambda$, is fitted to the remaining data and the values of the left out elements are predicted from the calculated model. The values for the left out elements are obtained by first reconstructing the data matrix from the calculated $Z$ and $P$ as $\hat{X} = ZP^T$ and selecting the left out elements. The step of leaving out elements and fitting a WSPCA model is repeated until all elements $x_{ij}$, with the exception of missing elements, have been left out once and predicted by the various constructed WSPCA models. The resulting data matrix of predicted elements ($\hat{X}_{leo} \{lx\}$), which is obtained via the described leave elements out procedure (leo), can be used to calculate a predicted
residual error sum of squares (PRESS) as shown in Equation 5.2:

\[
\text{PRESS} = \| W \odot (X - \hat{X}_{\text{leo}}) \|^2
\] (5.2)

As in Equation 5.1, the character \( \odot \) denotes the Hadamard product. Here the weight matrix \( W \) is, as mentioned before in Section 5.2.1, also a binary matrix containing ones (1) for elements that are present and zeros (0) for elements that are missing. During construction of \( \hat{X}_{\text{leo}} \) in the leave elements out cross-validation procedure the weight matrix is adjusted such that not only missing elements but also the left out elements are not used in the calculation of the various WSPCA models.

Each calculated PRESS value depends on which elements are randomly selected for the calculation of \( \hat{X}_{\text{leo}} \). In order to correct for the random selection of elements 10 PRESS values are calculated, each one originating from different random selections of elements for estimating \( \hat{X}_{\text{leo}} \). By varying \( R \) and \( \lambda \) and repeating the cross-validation procedure the optimal number of components and smoothing parameter can be assessed by finding the minimum of the average PRESS of the 10 calculated PRESS values. We used 10 PRESS values because this proved to have a strong stabilizing effect on the average PRESS value.

5.2.3 Missing value estimation

When a multivariate data set of related samples with expected smoothness in the underlying phenomena (e.g. microbial metabolomics fermentation data), contains values that are missing completely at random, WSPCA in combination with the leave elements out cross-validation can be used to estimate the missing values. At first the missing values have to be located and marked in the weight matrix \( W \). After that the cross-validation procedure, described above in Section 5.2.2, can be used to estimate the model meta parameters for WSPCA. Keep in mind that the true missing values do not participate in this cross-validation and that various sets of additional elements are sequentially temporarily marked for the construction of the matrix \( \hat{X}_{\text{leo}} \). Once the optimal model meta parameters \( R \) and \( \lambda \), for the roughness penalty, have been determined a WSPCA model can be build using the weight matrix \( W \) in which the missing values have been marked. By reconstructing the data matrix \( \hat{\mathbf{X}} \) from the obtained smoothed scores \( \mathbf{Z} \) and accompanying loadings \( \mathbf{P} \) as \( \hat{\mathbf{X}} = \mathbf{ZP}^T \), estimates for the missing elements can be obtained by selecting the missing values in \( \hat{\mathbf{X}} \). The estimates calculated in this way are adjusted for the dynamic aspects in the data.

5.3 Materials

5.3.1 Simulated data

To show that incorporation of smoothness into the scores of PCA can lead to a better estimation of the smooth phenomena underlying the time-resolved data and to show that the proposed leave elements out cross-validation procedure can be used for the determination of model parameters, a simulated data set has been
5. Weighted Smooth Principal Component Analysis (WSPCA)

created. The simulated data set consists of a noiseless data set and white noise. The noiseless data set ($X_{\text{true}}$), consisting of 25 evenly spaced timepoints for 45 variables, has been generated by multiplying three score vectors with three loading vectors (as depicted in Figure 5.1).

The first and second score vector ($s_1$ and $s_2$) were generated using time-shifted logistic functions with $t$ representing time, to represent well known dynamic behavior of metabolites that are formed or disappearing during the batch fermentation, as described by Equation 5.3 and 5.4. The first score vector is apart from being time-shifted also inverted and the second score vector contains an offset value.

$$s_1(t) = 1 - \frac{1}{1 + e^{-t+8}}$$  \hspace{1cm} (5.3)

$$s_2(t) = 0.3 + \frac{1}{2 + e^{-t+13}}$$  \hspace{1cm} (5.4)

The third score vector ($s_3$) represents a metabolite that is first being formed and later consumed during the batch fermentation. The following Gaussian function (Equation 5.5) was used:

$$s_3(t) = \frac{3}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$  \hspace{1cm} (5.5)

with $\mu = 8$ and $\sigma = 2$.

The three loading vectors, displayed in Figure 5.1(b) as loading profiles $l_1$–$l_3$, are used to make linear combinations of the three dynamic effects. They have been constructed such that the first three variables exactly describe the three dynamics effects, whereas the remainder 42 variables create random linear combinations of
the three dynamic effects. Each loading vector has been normalized to unit length and the angle between the loading vectors is given in Table 5.1.

<table>
<thead>
<tr>
<th>loading vectors</th>
<th>Angle $\angle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_1-l_2$</td>
<td>54.7°</td>
</tr>
<tr>
<td>$l_1-l_3$</td>
<td>56.0°</td>
</tr>
<tr>
<td>$l_2-l_3$</td>
<td>55.8°</td>
</tr>
</tbody>
</table>

Table 5.1: Angle between loading vectors used to generate noiseless data.

The noiseless data set ($X_{true}$) is obtained by multiplication of the three score and loading vectors,

$$X_{true} = [s_1 \, s_2 \, s_3][l_1 \, l_2 \, l_3]^T$$  \hspace{1cm} (5.6)

Addition of random normally distributed noise ($E$) with mean 0 and standard deviation 0.045 to the noiseless data ($X_{true}$) gives the data set ($X$), as represented by Equation 5.7.

$$X = X_{true} + E$$  \hspace{1cm} (5.7)

The resulting data set ($X$) contains approximately 8.2% noise, calculated as:

$$\frac{\|E\|^2}{\|X\|^2} \times 100\%$$  \hspace{1cm} (5.8)

5.3.2 *E. coli* batch fermentation metabolomics data

*Escherichia coli* batch fermentation experiments have been performed under different conditions to maximize the variation in the production of the amino acid phenylalanine using a full factorial design with three factors on two levels. During the time course of these batch fermentations 12 samples were taken from each bioreactor and analysed for intracellular metabolites by means of GC-MS and LC-MS. The experimental design, the experiments and the data collection are extensively described in Rubingh et al. [129]. To show the practical application of the leave elements out cross-validation procedure for model parameter estimation and the use of the WSPCA method for missing data estimation we used GC-MS data of one of the *E. coli* batch fermentations performed with Glucose at pH 6 and low phosphate concentration. The data consists of intensity values that represent the non-calibrated concentration levels of 21 putative intracellular metabolites. Prior to analyzing the data we range scaled each putative metabolite, making all metabolites equally important and comparable relative to their biological response range [58]. The range scaled measured intensities have been stored in a matrix consisting of 12 rows, representing the samples taken during the batch fermentation, and 21 columns, representing the putative metabolites (variables).
5. Weighted Smooth Principal Component Analysis (WSPCA)

Figure 5.2: Sum of Squares penalty, Sum of Squares residuals and Q plotted versus $\lambda$. Although $\lambda = 0$ cannot be shown on a common (base 10) logarithmic scale, it is shown here to indicate the non-regularized solution.

Missing values

The described *E. coli* batch fermentation metabolomics data set has 6 missing elements out of a total of $12 \times 21 = 252$ elements, making the percentage of missing values 2.38%. The missing values are present in the putative metabolite profiles of variables 018, 019 and 220. In variable 018 timepoints 2 and 4 are missing. Variable 019 is missing timepoints 10 and 11 and variable 220 misses timepoints 4 and 7.

5.4 Results and discussion

5.4.1 WSPCA method

To test WSPCA we have applied the implemented method with a second order difference matrix ($D_2$) on the created simulated data set, described in Section 5.3.1, using 3 components and 25 different values (ranging from $10^{-4}$ to $10^{1.5}$) for the smoothing parameter $\lambda$. We chose 3 components, because the simulated data was generated using 3 score and 3 loading profiles (Figure 5.1). Figure 5.2 graphically displays the obtained sum of squared penalty ($\|D_2Z\|^2$), the obtained sum of squared residuals ($\|W \circ (X - ZP^T)\|^2$) and $Q$ (see Equation 5.1) as a function of $\lambda$. The graph shows that over 2 orders of magnitude, from $\lambda = 0.001$ to 0.1, the sum of squared penalty rapidly decreases, whereas the sum of squared residuals and
Table 5.2: Values of Sum of Squares residuals, Sum of Squares penalty, Q and Fit for various Log$_{10} \lambda$ values, as shown in Figure 5.2

<table>
<thead>
<tr>
<th>Log$_{10} \lambda$</th>
<th>SS residuals</th>
<th>SS penalty</th>
<th>Q</th>
<th>Fit (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.000</td>
<td>1.85</td>
<td>5.72</td>
<td>1.85</td>
<td>93.33</td>
</tr>
<tr>
<td>-3.000</td>
<td>1.85</td>
<td>5.16</td>
<td>1.86</td>
<td>93.33</td>
</tr>
<tr>
<td>-2.000</td>
<td>1.87</td>
<td>2.31</td>
<td>1.89</td>
<td>93.28</td>
</tr>
<tr>
<td>-1.500</td>
<td>1.89</td>
<td>0.92</td>
<td>1.92</td>
<td>93.20</td>
</tr>
<tr>
<td>-1.000</td>
<td>1.92</td>
<td>0.33</td>
<td>1.96</td>
<td>93.09</td>
</tr>
<tr>
<td>-0.500</td>
<td>1.95</td>
<td>0.15</td>
<td>2.00</td>
<td>92.98</td>
</tr>
<tr>
<td>-0.250</td>
<td>1.97</td>
<td>0.11</td>
<td>2.03</td>
<td>92.92</td>
</tr>
<tr>
<td>+0.000</td>
<td>1.99</td>
<td>0.08</td>
<td>2.07</td>
<td>92.85</td>
</tr>
<tr>
<td>+0.125</td>
<td>2.00</td>
<td>0.07</td>
<td>2.09</td>
<td>92.81</td>
</tr>
<tr>
<td>+0.250</td>
<td>2.01</td>
<td>0.06</td>
<td>2.12</td>
<td>92.75</td>
</tr>
<tr>
<td>+0.375</td>
<td>2.03</td>
<td>0.05</td>
<td>2.16</td>
<td>92.69</td>
</tr>
<tr>
<td>+0.500</td>
<td>2.05</td>
<td>0.04</td>
<td>2.19</td>
<td>92.62</td>
</tr>
<tr>
<td>+0.750</td>
<td>2.09</td>
<td>0.03</td>
<td>2.25</td>
<td>92.48</td>
</tr>
<tr>
<td>+0.875</td>
<td>2.12</td>
<td>0.02</td>
<td>2.28</td>
<td>92.37</td>
</tr>
<tr>
<td>+1.000</td>
<td>2.15</td>
<td>0.02</td>
<td>2.33</td>
<td>92.26</td>
</tr>
<tr>
<td>+1.125</td>
<td>2.19</td>
<td>0.01</td>
<td>2.38</td>
<td>92.13</td>
</tr>
<tr>
<td>+1.250</td>
<td>2.23</td>
<td>0.01</td>
<td>2.44</td>
<td>91.99</td>
</tr>
<tr>
<td>+1.375</td>
<td>2.27</td>
<td>0.01</td>
<td>2.50</td>
<td>91.82</td>
</tr>
<tr>
<td>+1.500</td>
<td>2.32</td>
<td>0.01</td>
<td>2.57</td>
<td>91.64</td>
</tr>
</tbody>
</table>

Q only slightly increase. In other words over those two orders of magnitude in $\lambda$ the scores get smoothed quite strongly, hence the rapid drop in the sum of squared penalty. However, the created model fits the data still quite well, since the sum of squared residuals has not increased much. Determining an optimal value for the smoothing parameter $\lambda$ based on the graph in Figure 5.2 is, however, impossible. Qualitatively it is expected that the optimal $\lambda$ will be found, where the sum of squared penalty has reduced considerably and the sum of squared residuals has not increased much, meaning that the created model still fits the data well. Table 5.2 contains some of the values used for making Figure 5.2 and additionally contains the fit given in percentage, which is calculated as

$$
(1 - \frac{\|X - ZP^T\|^2}{\|X\|^2}) \times 100\% \quad (5.9)
$$

From this table can be seen that with the increase of $\lambda$ the fit of the calculated model $(ZP^T)$ with respect to the measured data $(X)$ becomes smaller, which is logical as the sum of squared residuals increases with increasing $\lambda$. The table also shows, that even when the sum of squared penalty has dropped considerably (see Log$_{10} \lambda = -0.500$) the method still slightly tends to overfit the data by incorporating noise.
into the model (the fit is about 93% of the data that contains approximately 8.2% of white noise).

To find the optimal $\lambda$ for our simulated data set we have plotted the sum of squared penalty against the sum of squared residuals (Figure 5.3), according to the L-curve [144], as introduced by Lawson and Hanson [145]. The corner of the L-shape curve should provide the optimal smoothing parameter $\lambda$. The graph does, unfortunately, not display the characteristic L-shaped curve, therefore, making it not possible to determine an optimal $\lambda$ in this way for our WSPCA method.

The goal of WSPCA is by applying smoothness on the scores to obtain a better estimation of the true dynamic phenomena underlying the data. To get insight into what happens to the estimated smooth scores ($Z$) when increasing the smoothing parameter $\lambda$, we have calculated the angle between the subspace spanned by the true underlying smooth scores and the subspace spanned by the estimated smooth scores of the WSPCA models built on $X$ with three components for creating Figure 5.2. The angle between the subspaces has been calculated as described by Björck et al. [146] and describes the similarity (or resemblance) between two subspaces, with two subspaces being identical when the angle between them becomes zero. To simultaneously see what happens to the accompanying loadings ($P$) when increasing $\lambda$ we also calculated the angle between the subspace spanned by the true loadings and the subspaces spanned by the loadings of the same WSPCA models. The true underlying smooth scores and loadings have been obtained by performing WSPCA on $X_{true}$ with $\lambda = 0$ and three components ($R = 3$). Figure 5.4 displays the calculated angles as a function of $\lambda$ and clearly shows in Figure 5.4(a) that when increasing $\lambda$ the subspace spanned by the smooth scores of the calculated models increasingly resembles the subspace spanned by the true underlying scores. Maximal resemblance is reached around $\lambda = 0.7$, beyond this value of $\lambda$ the scores get over-smoothed. The over-smoothing of the scores causes the resemblance between the subspaces to drop, which is seen as an increase in the angle between the subspaces. Figure 5.4(b) shows that when penalizing the scores for their roughness the subspace spanned by the accompanying loadings also slowly adjusts towards the subspace spanned by the true loadings. However, maximal resemblance is reached at a much larger $\lambda$ value than found for the subspaces spanned by the scores. With increasing $\lambda$ not only the subspace spanned by the calculated smooth scores starts to resemble the subspace of the true underlying scores, but also the score profile is a better estimation of the true underlying score profile. This is clearly shown in Figure 5.5 in which the true smooth score profile and the smooth score ($Z$) profiles for two $\lambda$-values ($10^{-3}$ and 0.7) from Figure 5.2 are visualized. Comparison of the smooth score profiles in Figure 5.5(b) and 5.5(c) shows that the scores get nicely smoothed when changing from $\lambda = 10^{-3}$ to 0.7 and that the smooth scores at $\lambda = 0.7$ estimate the true underlying scores, shown in Figure 5.5(a), very well. We also looked at the accompanying loadings profiles, but as was to be expected from Figure 5.4(b) they hardly change when changing from $\lambda = 10^{-3}$ to 0.7.
5.4. Results and discussion

![L-curve graph](image)

Figure 5.3: L-curve: Sum of Squares penalty plotted against Sum of Squares residuals for $\lambda$ ranging from $10^{-4}$ to $10^{1.5}$.

![Angle graphs](image)

Figure 5.4: (a) Angle between the subspace spanned by the smooth scores obtained from WSPCA on $X_{true}$ ($\lambda = 0$ and three components $R = 3$) and the subspace spanned by the smooth scores of WSPCA models built with three components on $X$ for creating Figure 5.2 as a function of $\lambda$. (b) Angle between the subspace spanned by the loadings obtained from WSPCA on $X_{true}$ ($\lambda = 0$ and $R = 3$) and the subspace spanned by the loadings of WSPCA models built with three components on $X$ for creating Figure 5.2 as a function of $\lambda$. Although $\lambda = 0$ can not be shown on a common (base 10) logarithmic scale, it is shown here to indicate the non-regularized solution.
5. Weighted Smooth Principal Component Analysis (WSPCA)

Figure 5.5: Smooth score profiles for $R = 3$: (a) smooth score profiles ($Z$) calculated on $X_{true}$ with $\lambda = 0$; (b) smooth score profiles ($Z$) calculated on $X$ for $\lambda = 10^{-3}$; (c) smooth score profiles ($Z$) calculated on $X$ for $\lambda = 0.7$.

5.4.2 Leave elements out cross-validation

To determine the optimal model meta parameters for the WSPCA method the leave elements out cross-validation procedure, described in Section 5.2.2, has been applied in combination with the WSPCA method on the simulated data (see Section 5.3.1). For the cross-validation procedure we have chosen to leave out 5% of the elements per fold in a 20 fold cross validation, the number of components was varied from 2 to 4 and $\lambda$ was varied between 0 and 10.
Figure 5.6 shows two graphs in which the average PRESS values divided by the number of elements in $X$ are plotted against $\lambda$ for all possible combinations of $\lambda$ and the number of components used for modelling. In Figure 5.6(a) the average PRESS values are calculated with respect to $X$, whereas in Figure 5.6(b) the average PRESS values are calculated with respect to $X_{\text{true}}$. Both Figure 5.6(a) and Figure 5.6(b) show a minimum for three components between $\lambda = 0.4$ and 0.7 (the optimal meta parameters for the WSPCA method). This means that the used cross-validation procedure is also applicable in situations where the true underlying dynamic phenomena are unknown as normally is the case for measured data. Comparison of Figure 5.6(a) and Figure 5.6(b) also shows that the average PRESS with respect to $X_{\text{true}}$ is lower than the average PRESS with respect to $X$. The cross-validated predictions ($\hat{X}_{\text{leo}}$) closely represent the true data. This is exactly what the procedure is expected to do. It is able to capture the true underlying phenomena in the data. The optimal meta parameters to capture the real data as good as possible can be found by cross validation of measured data as the minimum PRESS obtained with the measured data in Figure 5.6(a) is obtained at the same meta parameters settings.

For four components (nc=$R$=4) the curves of the average PRESS against $\lambda$ is not so smooth as for two and three components, this is due to the existence of local minima in the calculation of the WSPCA model for several sets of left out elements in the cross-validation procedure. Another interesting feature can be seen in the curve for nc=4 of Figure 5.6(b) which is that the minimum in the curve is reached at a larger $\lambda$ value compared to the curves for two and three components. Modeling
Figure 5.7: Concentration profiles for the first three variables:
(a) estimated by using the leave elements out cross-validation procedure \( \hat{\mathbf{X}}_{\text{leo}} \) with \( \lambda = 10^{-3} \);
(b) estimated by using the leave elements out cross-validation procedure \( \hat{\mathbf{X}}_{\text{leo}} \) with \( \lambda = 0.7 \);
(c) estimated by using the leave elements out cross-validation procedure \( \hat{\mathbf{X}}_{\text{leo}} \) with \( \lambda = 10^{-1} \);
(d) \( \mathbf{X}_0 \) and;
(e) \( \mathbf{X}_{\text{true}} \).
5.4. Results and discussion

the data with 4 components leads to an over-fitted model as can be observed from the high PRESS value at $\lambda = 0$. When more noise get integrated into the model, more emphasis needs to be put on the roughness penalty to reach a minimum.

The observation, that $\hat{X}_{\text{leo}}$ is closer to $X_{\text{true}}$ than to $X$, is also clearly visualized by Figure 5.7 in which the concentration profiles of the first three variables are plotted as estimated for three different $\lambda$ values, the original data $X$ and the original noiseless data $X_{\text{true}}$. The concentration profiles for $\lambda = 0.7$ (close to the optimal $\lambda$) in Figure 5.7(b) resemble closely the concentration profiles of the original noiseless data ($X_{\text{true}}$), as shown in Figure 5.7(e). The concentration profiles for $\lambda = 10^{-3}$ (a very small value of $\lambda$), shown in Figure 5.7(a), look more like the concentration profiles of the original data, displayed in Figure 5.7(d). This can be explained by the fact that the solution is very similar to the non-regularized solution ($\lambda = 0$), or the PCA solution, which tries to model as much of the variation present in the data as possible. For $\lambda = 10$, visualized by Figure 5.7(c), the concentration profiles get smoothed too much resulting in an increase of the average PRESS values in Figure 5.6. Note that the maximum value of var 3 has dropped to 0.2 whereas at $\lambda = 0.7$ this maximum closely resembles the true value of 0.27.

5.4.3  *E. coli* metabolomics data from a batch fermentation

As an example of real life data we have applied our WSPCA method to the *E. coli* batch fermentation metabolomics data set described in Section 5.3.2 to estimate the missing data in this data set with application of smoothness over the time-resolved samples. To determine the model parameters for the WSPCA method we first applied the leave elements out cross-validation procedure of Section 5.2.2 on the data set, where the number of components have been varied from 3 to 6 and the smoothing parameter $\lambda$ has been varied between 0 and 1. Figure 5.8, showing the average PRESS divided by the number of elements in the data set against the smoothing parameter for all combinations, displays the results of the leave elements out cross-validation procedure. It shows that there is a trade-off between the number of components and the effect of the smoothness parameter. With 3 components the data matrix predicted by means of the leave elements out cross-validation (\(\hat{X}_{\text{leo}}\)) clearly underestimates the underlying noiseless data, as for 4 components the curve lies below the curve for 3 components. When the number of components is high, e.g. 6 components, the unrestricted solution at ($\lambda = 0$) leads to a high PRESS value. In these cases the introduction of more smoothness by increasing the smoothing parameter helps enormously in lowering the average PRESS. The minimum in the curve for six components, compared to four and five components, clearly shifts towards a higher smoothing parameter value. This is due to the fact that with six component more noise gets modelled and more smoothness needs to be applied to obtain a minimum in the PRESS curve, as was also described in the previous Section. The effect of the smoothing parameter $\lambda$ with four components is only small, since the average PRESS does not get reduced much when increasing the smoothing parameter. However, the optimal model parameters, leading to lowest average press per element in Figure 5.8, are found at four components and $\lambda = 0.06$. 


Figure 5.8: Average Predicted Error Sum of Squares (avg PRESS) divided by the number of elements in the data set as a function of the smoothing parameter $\lambda$ from applying the leave elements out cross-validation procedure on real life E. coli batch fermentation metabolomics data. Although $\lambda = 0$ can not be shown on a common (base 10) logarithmic scale, it is shown here to indicate the non-regularized solution.

With the optimal model parameters a WSPCA model was built on all available data to estimate the truly missing elements (described in Section 5.3.2). The missing elements were also estimated using a non-regularized WSPCA model ($\lambda = 0$) with four components. Figure 5.9 shows the time profiles of the range scaled intensities of the putative metabolites (measured variables) containing the missing elements substituted with estimates of both non-regularized and optimal smoothness for the missing elements. The missing elements predicted with smoothness are close to the non-regularized estimations. This result is not unexpected, since the cross-validation results of Figure 5.8 showed that incorporating smoothness with four components had only a very small effect on the average PRESS. However, based on the simulated data the estimations of the missing elements with incorporated smoothness should be closer to the true underlying noiseless data and, therefore, be better estimates.
5.5. Conclusion

In this paper we have shown that when a multivariate data set consists of samples, that are not independent of each other but linked via a clear relationship with expected smoothness in the underlying phenomena, incorporation of smoothness into data analysis can help to estimate the underlying phenomena. The proposed WSPCA method, which puts the smoothness on the calculated scores, is able to capture the true underlying scores and accompanying loadings very well as shown.

Figure 5.9: Range scaled measured intensities and predicted range scaled intensities for missing data in the time profiles of real life *E. coli* batch fermentation metabolomics data: (a) putative metabolite variable 18; (b) putative metabolite variable 19; (c) putative metabolite variable 220.
with our synthetic data set. Contrary to most other multivariate analysis methods we do not restrict the scores to the column space of the data as smooth solution most probably lie outside that range. Our leave elements out cross-validation procedure was able to find the number of components and the optimal smoothing parameter $\lambda$ and proved to give a good estimation of the underlying noiseless data. Application of the WSPCA method, with the model parameters determined with the leave elements out cross-validation procedure, can be applied for missing value estimation and give better estimations for the missing elements as they are calculated by incorporating the dynamic phenomena underlying the data.

Acknowledgement

The authors thank Peter J. Punt, TNO, Microbiology & Systems Biology, Zeist the Netherlands for permission to use the *E. coli* batch fermentation data.