LVM METHODS IN THE PRESENCE OF STRUCTURED NOISE

REGULARIZED LATENT VARIABLE METHODS IN THE PRESENCE OF STRUCTURED NOISE AND THEIR APPLICATION IN THE ANALYSIS OF ELECTROENCEPHALOGRAM DATA

BY

SIAMAK SALARI SHARIF, M.Sc., B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

DOCTOR OF PHILOSOPHY (2012) McMaster University
(Chemical Engineering) Hamilton, Ontario

TITLE: REGULARIZED LATENT VARIABLE METHODS IN THE PRESENCE OF STRUCTURED NOISE AND THEIR APPLICATION IN THE ANALYSIS OF ELECTROENCEPHALOGRAM DATA

AUTHOR: Siamak Salari Sharif
M.Sc., McMaster University,
Hamilton, Ontario, Canada
B.Sc., Sharif University of Technology,
Tehran, Iran

SUPERVISORS: Dr. John F. MacGregor
Dr. James P. Reilly

NUMBER OF PAGES: xxiv, 234

Abstract

This thesis provides new regression methods for the removal of structured noise in datasets. With multivariable data, the variables and the noise can be both temporally correlated (i.e. auto correlated in time) and contemporaneously correlated (i.e. cross-correlated at the same time). In many occasions it is possible to acquire measurements of the noise, or some function of it, during the data collection. Several new constrained latent variable methods (LVM) that are built upon previous LVM regression frameworks are introduced. These methods make use of the additional information available about the noise to decompose a dataset into basis for the noise and signal. The properties of these methods are investigated mathematically, and through both simulation and application to actual biomedical data.

In Chapter Two, linear, constrained LVM methods are introduced. The performance of these methods are compared to the other similar LVM methods as well as ordinary PLS throughout several simulation studies. In Chapter Three, a NIPALS type algorithm is developed for the soft constrained PLS method which is also able to account for missing data as well as datasets with large covariance matrices. Chapter Four introduces the nonlinear-kernelized constrained LVM methods. These methods are capable of handling severe nonlinearities in the datasets. The performance of these methods are compared to nonlinear kernel PLS method. In Chapter Five the constrained methods are used to remove ballistocardiographic and muscle artifacts from EEG datasets in combined EEG-

fMRI as well as single EEG experiments on patients. The results are shown and compared to the standard noise removal methods used in the field. Finally in Chapter Six, the overall conclusion and scope of the future work is laid out.

Index terms - LVM, PLS, EEG, Regression PCA, eigenvalues

Acknowledgements

I would like to express my deepest and most sincere gratitude to my supervisors, Dr. John F. MacGregor, and James Reilly whose excellent guidance, ceaseless encouragement, patience, and expertise supported me throughout the course of my research.

I would also like to thank my committee members, Dr. Michael Noseworthy, and Dr. Gary Hasey and Dr. Daniel Bosnyak for their invaluable questions and suggestions that enhanced the quality of this research. Also, many thanks go to administrative staff, Nanci Cole, Lynn Falkiner, Kathy Goodram, and Melissa Vasil.

My sincere thanks to my friends at McMaster University and MACC for making a friendly and warm working environment and for their invaluable discussions that helped me understand many concepts and overcome many problems.

Dedication

I would like to dedicate this thesis to my family who have always been there for me and been a link to the past and my bridge to the future.

Table of Contents

List of Figures

List of Tables

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Publication List

Chapter 2:

Salari Sharif, S.; Reilly, J. P.; MacGregor, J. Latent Variable Methods in the Presence of Structured Noise. *Journal of Chemometrics*.

Contributions: Siamak Salari developed a framework for constrained partial least square methods in the presence of structured noise. This method which is similar to ordinary partial least squares (PLS) regression takes advantage of the additional information available about the structured noise contaminating input and output data which improves the component selection and prediction of PLS methods

Chapter 4:

Salari Sharif, S.; Reilly, J. P.; MacGregor, J. Nonlinear Latent Variable Methods in the Presence of Structured Noise. *Journal of Chemometrics*.

Contributions: Siamak Salari extended the linear methods for the constrained latent variable methods into the nonlinear kernel framework which allow for utilization of nonlinear kernels and performing regression between the input-output when they are nonlinearly related.

Chapter 1
Introduction

In many regression situations structured noise or disturbances (**Z**) affects both the regressor variables (**X**) and the response variables (**Y**). If nothing is known about this noise then certain assumptions are made and default regression methods based on these assumptions are used (e.g. ordinary least squares, Partial Least Squares (PLS) regression, etc.). But if one has additional information by way of some independent measurements on parts of the noise (**Z**), then this information can be used to improve the estimation of the true relationships among **X** and **Y**. The goal of this thesis is to provide new regression methods for the removal of structured noise in datasets. Several new constrained latent variable methods are introduced that make use of the additional information, available about the noise, to decompose a dataset into subspaces belonging to noise or the signal. The properties of these new methods are investigated mathematically, and through both simulations and applications to actual data.

The success of a regression model depends on several issues such as the amount of information the input variable (**X**) holds about the response variable (**Y**), the degree of linearity between **X** and **Y** and also the magnitude, nature and distribution of the noise in **X** and **Y**. In general any variation in a dataset that is irrelevant of the desired response can be called noise. The presence of noise can severely degrade a regression model and the prediction quality for future observations. The model's success also depends on the knowledge available

concerning the noise and the signal. The regression method should be chosen based on the type of noise present in the dataset. The noise can be uncorrelated or correlated (i.e. structured).

With multivariate data, the variables and noise can be both temporally correlated (i.e. auto-correlated in time) and contemporaneously correlated (i.e. cross-correlated at the same time). An example of temporally correlated noise is the presence of the 60Hz oscillations in the power lines. This type of noise exhibits a structured power spectrum. An example of contemporaneously correlated noise is the ocular artifacts randomly contaminating many electroencephalogram (EEG) channels at the same time. The noise can have both temporal and contemporaneous structures. For example, the cardio artifacts in the EEG affect many channels at the same time and in addition they are periodic in nature and hence temporally correlated. In this thesis we refer to structured noise as contemporaneous noise in the signals and it is this noise that needs to be treated in order to uncover the true relationships among the signals.

When no information is available about the noise, perhaps the most convenient regression method is ordinary least squares (OLS) or its multivariable counterpart, multiple linear regression (MLR). In MLR it is assumed that the noise is only present in **Y** and it is independent and identically distributed (*iid*). The success of MLR regression, however, depends on the condition of the input matrix. Regular MLR only works when **X** is full rank and does not contain any missing elements. It also assumes that **X** is free of noise. When **X** is rank deficient, other regression

methods such as ridge regression or partial least squares (PLS) [1] should be used. The use of PLS has several advantages over the ridge regression; for example, it can handle the presence of missing values. It also makes use of the contemporaneous correlation among the **X** and **Y** variables to improve the model and can also provide a model for the variations in **X** that can be used for visual inspection or detection of outliers. If **X** is full rank and both **X** and **Y** contain no missing elements, then the final prediction for PLS, provided that enough principal components are extracted, will be the same as that for OLS.

The presence of noise in the input variable can lead to biased estimates of the model. One way to remove the noise is to use latent variable methods such as principal component regression (PCR) [2] or PLS which are less sensitive to random variations in **X**. However, in the presence of structured noise the situation changes. When the noise is structured, there is a chance that its variations can be modeled by the latent variables, leading to biased estimates and incorrect predictions. If the noise is only present in **X** or only in **Y**, the final prediction results will remain unchanged but more components may have to be extracted by PLS in order to achieve the same quality of fit. When there is common structured noise in both **X** and **Y**, building a model between **X** and **Y** will result in modeling all the structured variation, including the common structured noise

. In these situations, ordinary regression methods such as PLS and MLR cannot provide the best models for the true relationship between **X** and **Y** without being impacted by the presence of the common structured noise. In such situations, the

noise needs to be accounted for when building the model.

In many occasions it is possible to acquire measurements of the noise, or some function of it, during the data collection. For example in EEG recordings the ocular artifacts are recorded using the extra electrodes that are placed around the eye area [3]. These extra electrodes record information about the ocular noise (**Z**). This information can be used later to preprocess the data to identify the ocular artifacts such as eye movement or blinks and remove them from the EEG dataset. A simple yet very efficient way of removing the noise is to project the response data (**Y**) onto the orthogonal complement of the noise ($\hat{\mathbf{Y}} = \mathbf{Q_Z Y}$), which removes the noise components from **Y**, and then build a model between **X** and $\hat{\mathbf{Y}}$. There are several problems with this approach; one is that if the noise matrix **Z** is also noisy, then the projection into $\mathbf{Q_Z}$ will not be accurate. This situation will be discussed further throughout the thesis. Also, **Z** must be orthogonal to the signal's subspace or some of the useful information in **Y** or **X** will also be filtered out. For instance, in the EEG example, the ocular electrodes also record some of the useful signal coming from the brain. Therefore the use of straight projection in this example can reduce the signal to noise ratio (SNR) of the EEG electrodes [3]. Using PLS to build a model between **X** and $\hat{\mathbf{Y}}$ is also not a very good option, since the components of the noise still reside in **X**. The PLS method is likely to extract some of these noise components as latent variables, hence more principal components will be required to achieve the same quality of fit which will result in less reliable and biased estimates.

In this thesis one step latent variable-type solutions are proposed in which a constrained partial least squares model is built between **X** and **Y**, but simultaneously suppresses the effect of noise, using the auxiliary noise matrix provided (**Z**). Two variations of this method are proposed: Hard-Constrained PLS (HC-PLS) and the Soft-Constrained PLS (SC-PLS). In HC-PLS a set of principal components are extracted from **X** that are highly correlated with **Y**, but at the same time are orthogonal to **Z**. As in the PLS algorithm these scores can be used for visualization of the **X** space as well as for prediction of the future **Y** variables.

The components extracted by these constrained methods are less likely to contain noise components, which makes them more suitable for prediction of the true underlying response and interpretation of the data (such as in score plots).

Since these components are orthogonal to the noise space, they can be used for prediction of future noise free **Y** variables as well.

In the case where the noise matrix **Z** is not orthogonal to the signal's subspace or is not properly conditioned, a case which is discussed later, a less restrictive algorithm can be implemented to compromise between residual noise and better fit to the model. An example of this is the use latent variable methods for removal of the ballistocardiographic (BCG) artifacts [4] from EEG. The objective is to find a way to separate the EEG matrix into two subspaces: the noise subspace , and the brain signal subspace, each with their respective set of latent variables. This objective leads to the development of a soft constrained version of HC-PLS that is named Soft-Constrained PLS (SC-PLS). The advantage of the SC-PLS over HC-

PLS is that the rigidity and the sensitivity of the model to noise can be adjusted manually. This algorithm is also less sensitive to the condition of the noise matrix, such as the rank or its orthogonality to the signal subspace. These features as well as the other advantages of the SC-PLS algorithm will be discussed in detail in the upcoming chapters of the thesis. This algorithm was used to remove the ballistocardiographic (BCG) and muscle artifacts from the electroencephalogram (EEG) data in a study. The results are presented in chapter 5 of the thesis.

Wold et al. [5,6] developed the "non-linear iterative partial least squares" (NIPALS) algorithm for PLS which is an extension of the power iteration method [7] for extracting the eigenvalues in matrices. The advantage of using Wold's algorithm is an improved computational cost for large datasets as well as its inherent ability to handle missing points. Inspired by the NIPALS algorithm, we developed an iterative algorithm for SC –PLS that can also account for the missing values in **X**, **Y** and **Z** while building the model and while predicting the future responses. This algorithm is presented in chapter 4.

In many practical cases the relationship between **X** and **Y** and **Z** is linear or mildly nonlinear; however, there are cases in which the relationship between the **X** and **Y** or the **X** and **Z** can be defined using a nonlinear function $\Phi(.)$, as $\mathbf{Y} = \Phi(\mathbf{X})\mathbf{B}$. In other words **Y** is a linear mixture of nonlinear transformations of **X**. In such cases if the transformation is known, $\Phi(\mathbf{X})$ can be calculated and a simple LVM algorithm or regression can find the relationship between **Y** and $\Phi(\mathbf{X})$. However, if the nonlinearity is strong and the transformation function $\Phi(.)$ is

unknown, then a linear regression or LVM method may no longer provide satisfactory results. For such problems, nonlinear algorithms, known as kernel methods, have been developed for regression and have also been extended to the PLS algorithm known as the Kernel PLS (KPLS) [8–10]. These methods take advantage of a certain property known as the "kernel trick" [11,12] to perform regression in a nonlinear feature space without explicitly knowing the nonlinear transformation $\Phi(.)$. In chapter four the kernel idea is extended to SC-PLS and HC-PLS algorithms. The resulting nonlinear algorithms, named SC-KPLS and HC-KPLS for soft and hard constrained methods respectively, can also handle a nonlinear relationship between **X** and **Y** and **Z** matrices and simultaneously suppress the structured noise.

The organization of this thesis is as follows: In chapter two of the thesis a framework for regularized latent variable methods is provided. The hard constrained PLS and the soft constrained PLS methods are introduced. Later throughout the chapter, these methods are compared against a variation of Orthogonal Signal Correction methods (OSC) developed by Fearn et al. [13] and ordinary PLS regression using simulation studies. An industrial example in which HC-PLS is used to improve product optimization is also presented. In chapter three, the iterative NIPALS extension of the SC-PLS algorithm (NIP-SCPLS) is introduced and the quality of the model is validated using simulation studies. Later in that chapter it is shown and analyzed how the NIP-SCPLS is able to handle missing data. In chapter four the nonlinear constrained PLS methods, SC-

KPLS and HC-KPLS are introduced. Again the quality of these algorithms is compared to KPLS and linear LVM methods using simulation studies. Finally in chapter five, the SC-PLS algorithm is implemented on real EEG and simulated EEG data to remove ballistocardiographic and muscle artifacts from EEG datasets. Conclusion and suggestions for future work are presented in chapter 6.

.

REFERENCES

(1) BURNHAM, A.; MACGREGOR, J.; VIVEROS, R. A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of chemometrics* **1999**, *13*, 49-65.

(2) Burnham, A. J.; Viveros, R.; MacGregor, J. F. Frameworks for latent variable multivariate regression. *Journal of chemometrics* **1996**, *10*, 31-45.

(3) Croft, R. J.; Barry, R. J. Removal of ocular artifact from the EEG: a review. *Neurophysiologie Clinique/Clinical Neurophysiology* **2000**, *30*, 5-19.

(4) Niazy, R. K.; Beckmann, C. F.; Iannetti, G. D.; Brady, J. M.; Smith, S. M. Removal of FMRI environment artifacts from EEG data using optimal basis sets. *NeuroImage* **2005**, *28*, 720-737.

(5) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109-130.

(6) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37-52.

(7) Hua, Y.; Xiang, Y.; Chen, T.; Abed-Meraim, K.; Miao, Y. A New Look at the Power Method for Fast Subspace Tracking. *Digital Signal Processing* **1999**, *9*, 297-314.

(8) Kim, K.; Lee, J.-M.; Lee, I.-B. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems* **2005**, *79*, 22-30.

(9) Rosipal, R. Kernel partial least squares for nonlinear regression and discrimination. *Neural network world* **2003**, *13*, 291-300.

(10) Rosipal, R.; Trejo, L. Kernel partial least squares regression in reproducing kernel hilbert space. *The Journal of Machine Learning Research* **2002**, *2*, 97-123.

(11) Muller, K.; Mika, S.; Ratsch, G.; Tsuda, K.; Scholkopf, B. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks* **2001**, *12*, 181-201.

(12) Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **1998**, *10*, 1299-1319.

(13) Fearn, T. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems* **2000**, *50*, 47-52.

Chapter 2
Latent Variable Methods In the Presence of Structured Noise

Abstract

Latent variable methods are presented that are aimed at extracting models for true underlying relationships among a set of regressor and response variables when those measurements are contaminated by structured noise. It is assumed that one not only has measurements on the variables of interest (X, Y), but also information on some part of the structured noise, either as simultaneous measurements on variables (Z) that contain some of the structured noise. The latent variable methods are developed from objective function formulations that maximize the covariance explained among the X and Y variables, subject to hard or soft orthogonality constraints on the structured noise information. The various algorithms are evaluated on simulated and industrial data to illustrate and compare their performances against one another and against traditional PLS algorithms. The results illustrate the substantial improvements possible when one has such auxiliary information on some of the structured noise.

***Index terms*- Principal Component Analysis, Partial Least Squares, Structured noise, Modeling and Prediction, Constrained Optimization**

2.1 Introduction

The object of statistical inference is to extract information from measurements that are contaminated by noise (alternatively one might refer to noise as errors or disturbances). With no specific details on the noise, one usually makes certain assumptions about the nature of the noise and proceeds with the analysis under these assumptions. For example, with multiple linear regression one usually assume that all the errors reside in the response measurements ($\mathbf{Y}$) and are identically and independently (*iid*) distributed usually as a Normal distribution. With error-in-variables regression [1] errors with some assumed distribution are assumed to be present in both the regressors ($\mathbf{X}$) and response ($\mathbf{Y}$) variables.

However these methods require knowledge of the distribution of the errors. With large multivariate data sets Partial Least Squares (PLS) is often used to handle both the reduced rank nature of the data and to allow for errors of unspecified distribution in both the $\mathbf{X}$ and $\mathbf{Y}$ data. But if the noise or errors are structured, that is they are not independent of one another, then the PLS model will also extract latent variables that model the covariance structure of the noise. If the structured noise in the $\mathbf{X}$ and $\mathbf{Y}$ spaces is independent, then this should not influence the relationships extracted among the $\mathbf{X}$ and $\mathbf{Y}$, only the uncertainties in those inferences. However, if the structured noise has a strong presence in $\mathbf{X}$, even though the model may have the same prediction results, more components need to be extracted and results become harder to visually interpret.

Structured noise is considered to be multivariate noise where the elements of the noise vectors are correlated with one another (contemporaneous noise) as well as possibly temporally correlated, as opposed to white noise where the elements are independent and not auto-correlated. Such structure in the noise implies that PCA performed on it will have significant components, i.e. structure.

Structured noise can be exclusive to one subspace (uncommon structured noise) or can affect both **X** and **Y** with the same source (common structured noise). The reader should distinguish between these two terms. In the case of common structured noise any PLS regression model may also contain latent variables related to this noise. After all, PLS is simply providing models for the measured $\mathbf{X}$ and $\mathbf{Y}$ spaces, regardless of the source of the variations.

If the objective of the study is to uncover the true underlying relationships among the noise free $\mathbf{X}$ and $\mathbf{Y}$ spaces in the presence of common structured noise contaminating both $\mathbf{X}$ and $\mathbf{Y}$, then further information on the specific noise present in these measurements is necessary. This noise information may be available in different ways. In some cases, at the same time as $\mathbf{X}$ and $\mathbf{Y}$ are measured, one might have available measurements on additional variables ($\mathbf{Z}$) that contain information on the structured noise. Alternatively one might have data collected at other times that provide information on the covariance structure of the noise. To illustrate these situations, consider the following examples.

1. Bruwer et al [2] considered the problem of predicting the textural properties of snack foods (crispiness, hardness, surface texture, etc.) from measurements of vibrational spectra collected from multiple vibration sensors (accelerometers) affixed to a stainless steel chute onto which the snack food was falling in the process of being delivered to the next processing stage. The signals from these accelerometers ($\mathbf{X}$) are also affected by the structured noise from machines and other vibrating equipment located in the same vicinity as the chute, affecting the measurements on the vibrational spectra. This noise could be monitored either simultaneously with $\mathbf{X}$ and $\mathbf{Y}$ using microphones ($\mathbf{Z}$) or measured at other times and its covariance matrix (Λ) estimated.
2. Consider the situation of trying to relate process data ($\mathbf{X}$) to product characteristics ($\mathbf{Y}$) as measured from a NIR spectrometer. In many

such cases both the process data and the response data are affected by common disturbances such as temperatures and humidity and one might have a simultaneous measure of variables related to these environmental disturbances ($\mathbf{Z}$). An example, in the oil exploration industry was reported by [3] where soil samples collected from the ocean floor were subject to spectroscopy to detect oil and gas residues. These samples were contaminated with larger amounts of other organic residues that reduced the ability to detect the residues of interest. However, spectroscopic measurements on some of the major contaminating factors were available elsewhere and could be used to improve the final analysis.

3. Another example concerns the presence of ballistocardiographic noise (BCG) during the recording of electroencephalogram (EEG) signals [4]. Available simultaneous measurements from an electrocardiogram can be used to remove BCG from EEG signals giving a better measure of the true brain physiological activity. Another example from the same field is the presence of background brain activity in task- related electroencephalography. The background activity appears as structured noise. However, its covariance matrix can be calculated by performing resting estate EEG (Λ) and used to remove the structured noise.
4. As another example, while recording (EEG) signals from the brain an extra electrode is placed on the face to record eye movement (which in

this case can be considered the auxiliary noise matrix ($\mathbf{Z}$). This additional noise data can be later used to remove eye movement artifacts from the EEG data [5].

5. In combined EEG-fMRI analysis motion artifacts can affect both the fMRI and EEG data. Since the motion affects both EEG and fMRI data, inference between EEG and fMRI will result in a model that may explain the common motion artifact rather than the true underlying physiologic changes relating EEG and fMRI data. In such a case, extra measurements from motion sensors can be incorporated ($\mathbf{Z}$) and used to remove the common motion noise artifact from both EEG and fMRI

In this thesis we present and compare various latent variable methods that use this ancillary noise information collected simultaneously ($\mathbf{Z}$) with the regressor ($\mathbf{X}$) and response ($\mathbf{Y}$) measurements, to obtain improved estimates of the underlying noise free models for $\mathbf{X}$ and $\mathbf{Y}$.

In the realm of latent variable data analysis, several methods have been introduced by [6–9] and other authors to remove unrelated components from $\mathbf{X}$ prior to regressing against $\mathbf{Y}$. However, none of these methods exploit the additional knowledge that might be available on the noise. In the presence of common structured noise in both $\mathbf{X}$ and $\mathbf{Y}$, these methods fail to remove the impact of the structured noise. This issue will be discussed in further detail in the upcoming sections.

One way to eliminate the effect of structured noise, when an auxiliary noise matrix "$\mathbf{Z}$" exists, is to project the data onto the orthogonal subspace of $\mathbf{Z}$ prior to performing any latent variable data analysis. However, this method tends to complicate the process of predicting the future response values [7]. In addition, it can be shown that the condition and presence of random noise in the auxiliary matrix ($\mathbf{Z}$) can greatly reduce the efficacy of a simple projection method, leading to inaccurate results and predictions [10]. We shall discuss this problem in further details in the appendix.

This paper provides a framework for constrained latent variable methods that utilize information on some of the structured noise present in the data, either in the form of a simultaneous measurement matrix ($\mathbf{Z}$) or a covariance estimate (Λ) of the noise. It introduces the concept of soft and hard constrained (regularized) latent variable methods. What distinguishes this paper from previous articles is the concept of utilizing noise related matrices for directly extracting latent variables without any preprocessing (noise removal) steps. The methods provided here have the same properties as well known latent variable methods such as PCA and PLS, but with built-in noise constraints. Therefore, the principal components (or latent vectors) extracted can be used in the very same way they would have previously in regular methods such as normal partial least squares (PLS) or Reduced Rank Regression (RRR).

In Section II, hard-constrained latent variable methods are discussed and expanded into partial least squares and principal component type of regressions.

Later, we shall discuss some of the properties and shortcomings of hard constrained LVM methods. In section 2.3 we introduce the concept of soft constrained LVM. Finally in sections 2.4 and 2.5, we present some simulation results and well as an industrial optimization example to assess the properties of the methods proposed here and to compare them with standard methods. The Appendix discusses the extension of these constrained methods to other LVM models which are beyond the scope of this thesis but yet still fall within the concept of constraint LVM. As mentioned earlier we shall discuss the advantages of using the latent variable noise removal methods (i.e. those discussed in this thesis) as opposed to ordinary projection methods discussed earlier

Notation: Bold upper (lower) case Arabic symbols represent matrices or vectors respectively, and regular-faced symbols are scalars. Bold-faced, lower-case Greek symbols are matrices. The notation, e.g. $\mathbf{X}_i$ represents the i^{th} column of the matrix $\mathbf{X}$. The notations $\mathcal{N}(\cdot)$ and $\mathcal{R}(\cdot)$ denote the null-space and range of the argument respectively. In the following chapters we use the following naming scheme: methods that include hard constrints are accompanied by the prefix “HC” and those that include soft constrints will have the prefix “SC”.

2.2 Hard Constrained Latent variable regression in the Presence of Structured Noise

In ordinary LVM methods the assumption is that only two matrices, the matrix of regressor variables $\mathbf{X}\,(n\times k)$ and response variables $\mathbf{Y}\,(n\times m)$ are available. However, assuming the availability of an auxiliary noise matrix $\mathbf{Z}\,(n\times b)$ that contains information about some (or all) components of the structured noise, one can exploit $\mathbf{Z}$ to obtain a better model for the underlying noise-free behavior of $\mathbf{X}$ and $\mathbf{Y}$. In this section we look at several methods for finding latent structures that provide improved latent variable models for the underlying noise-free variations in $\mathbf{X}$ and $\mathbf{Y}$ by imposing various orthogonality constraints between the latent variable spaces of $\mathbf{Z}$ and $\mathbf{X}$ and $\mathbf{Y}$.

2.2.1 The OSC-PLS approach:

Orthogonal signal correction (OSC) and projections onto orthogonal latent structures (O-PLS and later extended to O2-PLS) were introduced [6,8,9,11] that prove effective under reasonable conditions. These methods remove variations in **X** (or **Y**) that are unrelated to **Y** (or **X**) before determining a model. As long as the structured noise resides in either **X** or **Y** these methods perform reasonably well. If both spaces are contaminated with structured noise but the noise structure is different in each dataset (different source), these methods still provide adequate results. However, in the presence of common structured noise, with a same (common) basis for structured variations in both **X** and **Y**, these methods fail. Take for example a variation of OSC PLS introduced by Fearn [7]: In this method the covariance matrix $\mathbf{X}'\mathbf{X}$ is projected onto the orthogonal complement of $\mathbf{X}'\mathbf{Y}$. In the case when common structured noise resides in both **X** and **Y**, $\mathbf{X}'\mathbf{Y}$ will also

be a basis for the noise as well and hence the components extracted from **X** will be also orthogonal to noise subspace. In such a case, when the common noise is not properly removed from the datasets, the resulting model will be biased towards the structured noise. In other words instead of building a relationship between the input and the true underlying responses, the resulting model may be better at predicting the variations in the noise rather than the desired **Y** values. Fearn's method does not have the essential ingredient that is the focus of this paper; namely, an independent measurement of a matrix **Z** that contains elements of structured noise in **X** and **Y**. However, Fearn's method has a solution structure that is parallel to those we consider here. It is also a methodology that many are familiar with and so provides a good starting point for this paper. Fearn's problem is formulated as:

$$\begin{aligned} & \max_{\mathbf{w}} \mathbf{w}_i^{'}\mathbf{X}^{'}\mathbf{X}\mathbf{w}_i, \\ \text{s.t.} \quad & \mathbf{w}_i^{'}\mathbf{w}_j = \delta_{ij} \\ & \mathbf{w}_i^{'}\mathbf{X}^{'}\mathbf{Y} = \mathbf{0}, i = 1,\ldots,q. \end{aligned} \tag{2-1}$$

The objective of the problem is to find a set of latent vectors ($\mathbf{w_i} \in \mathbb{R}^{k\times 1}$) that maximize the above objective function considering the constraints included. Here, δ_{ij} is the Kronecker delta. Rao [12] has shown that the solution $\mathbf{w_1},\cdots,\mathbf{w_q}$ to this problem is given by the q principal eigenvectors of the matrix $\mathbf{M_{XY}X'X}$. $\mathbf{M_{XY}}$ ($k\times k$) is a projection matrix (onto the orthogonal complement of $\mathbf{X'Y}$) defined as:

$$\mathbf{M_{XY}} = \mathbf{I} - \mathbf{X'Y}\left(\mathbf{Y'XX'Y}\right)^{\dagger}\mathbf{Y'X}. \tag{2-2}$$

Instruction on how to calculate the eigenvalues of $\mathbf{M_{XY}X'X}$ (a potentially non-

symmetric matrix) are given in Rao's publication mentioned above. The general idea common to most latent variable methods, including the Fearn method and those derived in this paper, is to sequentially extract a set of orthonormal weight vectors $\mathbf{w_i}, i = 1, \cdots, q$ according to some specified objective function. From the weight vectors, a set of orthogonal latent vectors $\mathbf{t}_i$ ($n \times 1$) are defined as a linear combinations of columns of $\mathbf{X}_i$ with $\mathbf{X}_1$ being the original $\mathbf{X}$

$$\mathbf{t}_i = \mathbf{X}_i \mathbf{w}_i. \tag{2-3}$$

Since the $\mathbf{t_i}$ are linear mixtures of $\mathbf{X}$, they provide a basis (latent vectors) for directions in $\mathbf{X}$ that carry the desired properties included in the objective function. To ensure the orthogonality between basis component vectors (eigenvectors), they are extracted iteratively and then $\mathbf{X}$ is deflated at each step by projecting it into the orthogonal complement of $\mathbf{t_i}$:

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i \mathbf{p}_i'. \tag{2-4}$$

where

$$\mathbf{p}'_i = (\mathbf{t}'_i \mathbf{t}_i)^{-1} \mathbf{t}'_i \mathbf{X}_i \tag{2-5}$$

This deflated $\mathbf{X}_{i+1}$ is then used in (2-1) - (2-5) until q eigenvectors are determined. The vector $\mathbf{p}_i$ ($k \times 1$) is the projection coefficient (loading) vector obtained by regressing $\mathbf{X}_i$ on $\mathbf{t}_i$ (2-5). In the OSC-PLS algorithm the principal components extracted (3) contain variations in $\mathbf{X}$ that are orthogonal to $\mathbf{Y}$. Once a sufficient number (q) of principal components are extracted, the residual $\mathbf{X}$

contains mainly variation that is relevant to $\mathbf{Y}$. The residual matrix ($\mathbf{X}_q$) can now be used to perform regression with $\mathbf{Y}$.

A problem with the OSC method is that the number of principal components (q) extracted from $\mathbf{X}$ must be determined separately. If the structured noise is only present in $\mathbf{X}$ this method will perform well (given that a sufficient number of components are extracted). However, in the case where an insufficient number of principal components are extracted or when the structured noise is present in both $\mathbf{X}$ and $\mathbf{Y}$ this method fails to provide the best model between the true (noise free) matrices since the structured noise will not be removed from $\mathbf{X}$. We illustrate this expected result later through some toy examples.

Other OSC-PLS methods that provide advantages over Fearn's method [6] have been proposed. However, at best the OSC-PLS method still provides the same prediction as ordinary PLS between $\mathbf{X}$ and $\mathbf{Y}$ and makes no attempt to remove any common structured noise in $\mathbf{X}$ and $\mathbf{Y}$.

In what follows, we will assume that we have a matrix of auxiliary measurements $\mathbf{Z}$ available that is related to at least some elements of the structured noise in $\mathbf{X}$ and $\mathbf{Y}$. In other words, the structured noise present in $\mathbf{X}$ and $\mathbf{Y}$ is partially within the range of $\mathbf{Z}$. We develop latent variable algorithms that are aimed at using this auxiliary noise measurement $\mathbf{Z}$ to account for some of the structured noise in $\mathbf{X}$ and $\mathbf{Y}$ during the latent variable regression. We refer to the first approach as Hard-Constrained Principal Component Regression (HC-PCR). This is an algorithm similar to Fearn's method but the orthogonalization is

forced between $\mathbf{X}$ and the auxiliary noise matrix $\mathbf{Z}$ rather than between $\mathbf{X}$ and $\mathbf{Y}$

.

2.2.2 HC-PCR

A modification of Fearn's method that accounts for the presence of structured noise is shown in (2-6) and (2-7). The orthogonal latent variable set $\mathbf{t}_i; i:1,\ldots,q$ extracted from this objective function explains maximum variance in the subspace of $\mathbf{X}$ that is orthogonal to the auxiliary measurement matrix $\mathbf{Z}$. Since the structured noise is (partially) within the range of $\mathbf{Z}$ the orthogonalization process will suppress the structured noise components in $\mathbf{X}$. The HC-PCR formulation can be cast as the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{w}} \ \mathbf{w}_i' \mathbf{X}' \mathbf{X} \mathbf{w}_i, \\ \text{s.t.} \quad &\mathbf{w}_i' \mathbf{w}_j = \delta_{ij} \end{aligned} \tag{2-6}$$

$$\mathbf{w}' \mathbf{X}' \mathbf{Z} = \mathbf{0}. \tag{2-7}$$

The $\mathbf{t}$'s form a basis for the subspace of $\mathbf{X}$ that is orthogonal to $\mathbf{Z}$ (and hence a partial subspace of the structured noise). Once a sufficient number of $\mathbf{t}$s are extracted they are then used for regression on $\mathbf{Y}$. Defining $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_q]$, a predicted value $\mathbf{Y}_p$ of $\mathbf{Y}$ can be obtained from:

$$\mathbf{Y_p} = \mathbf{T}\mathbf{Q}' \tag{2-8}$$

$$\mathbf{Q}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}, \tag{2-9}$$

where $\mathbf{Q}$ ($q \times m$) is the projection coefficient obtained by projecting $\mathbf{Y}$ onto $\mathbf{T}$. The quantity $\mathbf{X}_p$ is defined in a corresponding manner by projecting $\mathbf{X}$ onto $\mathbf{T}$ as:

$$\mathbf{X_P} = \mathbf{TP'} \tag{2-10}$$

where

$$\mathbf{P'} = (\mathbf{T'T})^{-1}\mathbf{T'X}. \tag{2-11}$$

This projection ($\mathbf{X_P}$) contains only variation in $\mathbf{X}$ that is orthogonal to the partial subspace of the structured noise that is within the range of $\mathbf{Z}$. We call this method Hard-Constrained Principal component regression (HC-PCR) due to its similarities to the conventional principal component regression. A problem with this method, as with any PCR-based method, is that it does not inherit the desirable properties of the PLS method; i.e., it only finds linear combinations of $\mathbf{X}$ that maximize the variance in the subspace of $\mathbf{X}$, but does not consider the relationship between $\mathbf{X}$ and the regressing variables $\mathbf{Y}$. If the variation within $\mathbf{X}$ is large compared to its correlation with $\mathbf{Y}$ there is no guarantee that $\mathbf{Y}$ will be explained properly by the number of principal components extracted. In the succeeding sections of this paper, we extend the proposed method to develop PLS-type algorithms that take into account the structured noise in both $\mathbf{X}$ and $\mathbf{Y}$.

2.2.3 The HC-PLS approach

To describe this approach, we consider the PLS objective function [13] but include a hard constraint to enforce orthogonality between $\mathbf{X}$ and $\mathbf{Z}$. Following the Burnham and Fearn formulations we can write:

$$\begin{aligned} & \max_{\mathbf{w}} \mathbf{w}_i^{'}\mathbf{X}^{'}\mathbf{Y}\mathbf{Y}^{'}\mathbf{X}\mathbf{w}_i, \\ \text{s.t.} \quad & \mathbf{w}_i^{'}\mathbf{w}_j = \delta_{ij} \\ & \mathbf{w}_i^{'}\mathbf{X}^{'}\mathbf{Z} = \mathbf{0}, i = 1, \ldots, q. \end{aligned} \tag{2-12}$$

This formulation is similar to that of ordinary PLS regression. However, the

addition of the last line ensures orthogonality of the latent variables with the auxiliary noise subspace ($\mathbf{Z}$). The solutions $\mathbf{w}_1,\cdots,\mathbf{w}_q$ to this constrained partial least squares problem are obtained by sequentially finding the principal eigenvalues of the matrix $\mathbf{M}_{XZ}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$ [12] where the $\mathbf{X}$ and $\mathbf{Y}$'s at each sequential stage are the deflated matrices obtained by projection into the orthogonal complement of the prior $\mathbf{t}_i$ as per equation (2-4). The matrix $\mathbf{M}_{XZ}$ is the projector onto the orthogonal complement of $\mathbf{X}'\mathbf{Z}$ defined similarly to that between $\mathbf{X}$ and $\mathbf{Y}$ in equation (2-2). This method, in addition to eliminating structured noise from the latent vectors extracted, also inherits the properties of conventional PLS, namely high covariance between the extracted $\mathbf{t}_i$ and the $\mathbf{Y}$ matrix.

2.3 Soft Constrained Latent variable regression in Presence of Structured Noise

In the previous section we showed how to generate latent variable structures that enforce orthogonality to the auxiliary noise matrix $\mathbf{Z}$, and therefore completely suppress any influence of $\mathbf{Z}$. However, as will be illustrated and discussed in more detail later in 2.4, in some cases it may be preferable only to penalize colinearity with the auxiliary noise rather than enforce exact orthogonality to it. In other words, we trade off the degree of influence of the structured noise into the estimated latent variable space against a model with potentially improved prediction capabilities. This will be further discussed in the

following subsections.

2.3.1 Soft Constrained PCR (SC-PCR[1])

In SC-PCR, we replace the hard constraint in (2-6) by a penalty on "squared" value of the covariance between $\mathbf{X}$ and $\mathbf{Z}$ (i.e. $\mathbf{X'ZZ'X}$). The term $\mathbf{w'X'Xw}$ is to be maximized while the term $\mathbf{w'X'ZZ'Xw}$, which explains the covariance with the auxiliary noise matrix, is to be simultaneously minimized. Hence a suitable form for the SC-PCR at each deflation step is defined by:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \left|\mathbf{w}_i'\mathbf{X'Xw}_i - \lambda \mathbf{w}_i'\mathbf{X'ZZ'Xw}_i\right| \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{w}_j = \delta_{ij} \end{aligned}. \tag{2-13}$$

The parameter λ is treated in this case as a manually-controlled meta-parameter which regulates the degree of trade-off between maximizing the variance of $\mathbf{X}$ (first term) and enforcing a small covariance with $\mathbf{Z}$ (the second term). The orthogonality between loading vectors ($\mathbf{w}_i$) is conserved through orthogonal projection at each deflation step. To solve the above problem a Lagrangian operator is constructed as follows

$$L(\mathbf{w}_i) = \left|\mathbf{w}_i'\mathbf{X'Xw}_i - \lambda \mathbf{w}_i'\mathbf{X'ZZ'Xw}_i\right| - \gamma_i(\mathbf{w}_i'\mathbf{w}_i - 1) \tag{2-14}$$

Applying the chain rule and recognizing that d/dα sgn(α) = 0 (except at α = 0), we have

$$\frac{\partial L(\mathbf{w}_i)}{\partial \mathbf{w}_i} = \text{sgn}(\alpha)\mathbf{X'}\left(\mathbf{I} - \lambda \mathbf{ZZ'}\right)\mathbf{Xw}_i - \gamma_i \mathbf{w}_i \tag{2-15}$$

where sgn(.) is the sign operator and α is the argument of the absolute value

[1] A prefix SC denotes the corresponding method is soft-constrained.

operator in (2-13) and γ_i is a Lagrange multiplier. Defining a new variable:

$$\hat{\gamma}_i = \text{sgn}(\alpha)\gamma_i, \tag{2-16}$$

And by setting the above to zero, we have

$$\mathbf{X}'\left(\mathbf{I} - \lambda\mathbf{Z}\mathbf{Z}'\right)\mathbf{X}\mathbf{w}_i = \hat{\gamma}_i\mathbf{w}_i. \tag{2-17}$$

Thus, we see the desired first latent variable for this case is the dominant eigenvector of the matrix

$$\mathbf{U}_1 = \mathbf{X}'\left(\mathbf{I} - \lambda\mathbf{Z}\mathbf{Z}'\right)\mathbf{X} \tag{2-18}$$

with corresponding eigenvalue $\hat{\gamma}_i$. Once again, the corresponding eigenvectors and principal components are extracted iteratively by deflating $\mathbf{X}$, the same way described in equation (2-4). The soft constraint approach no longer enforces orthogonality between $\mathbf{t_i}$ and $\mathbf{Z}$ but rather tries to achieve a softer version of it by penalizing the covariance between $\mathbf{t_i}$ and $\mathbf{Z}$. We can write the matrix argument of the first line of (2-13) as:

$$\mathbf{w}'\mathbf{X}'(\mathbf{I} - \lambda\mathbf{Z}\mathbf{Z}')\mathbf{X}\mathbf{w} = \mathbf{t}'(\mathbf{I} - \lambda\mathbf{Z}\mathbf{Z}')\mathbf{t} = \mathbf{t}'\mathbf{t} - \lambda\mathbf{t}'\mathbf{Z}\mathbf{Z}'\mathbf{t} \tag{2-19}$$

Now let's consider a particular iteration when the dominant eigenvalue $\hat{\gamma}$ of the matrix $\mathbf{U_1}$ in (2-18) is positive. Then $\text{sgn}(\alpha)$ in (2-15) is +1, and the solution $\mathbf{w}$ to (2-13) is the corresponding eigenvector, and from the right most equation of (2-19) we have $\mathbf{t}'\mathbf{t} > \lambda\mathbf{t}'\mathbf{Z}\mathbf{Z}'\mathbf{t}$, thus the corresponding $\mathbf{t}$ is dominated by the term $\mathbf{t}'\mathbf{t}$ and consequently $\mathbf{t}$ tends to be in a direction which explains maximum variation within $\mathbf{X}$ while suppressing the directions along $\mathbf{Z}$. We define $\mathbf{T}^s$ as

the matrix whose columns are the $\mathbf{t}$'s corresponding to positive values of $\hat{\gamma}$ obtained over all iterations of the process. The matrix $\mathbf{T}^{\mathbf{S}}$ is used in SC-PCR in place of $\mathbf{T}$ in (2-8)-(2-11) for predicting values for $\mathbf{X}$ and $\mathbf{Y}$.

We now consider the case in which the dominant eigenvalue γ is negative. In this case sgn(α) in (2-15) is -1 and the solution $\mathbf{w}$ maximizing (2-13) is again the eigenvector corresponding to the dominant eigenvalue. Here we have $\lambda\mathbf{t}'\mathbf{Z}\mathbf{Z}'\mathbf{t}' > \mathbf{t}'\mathbf{t}'$. So now the second term of the right most equation of (2-19) dominates. Using reasoning similar to that of the previous case, we see that $\mathbf{t}$ in this case corresponds to a direction in $\mathbf{X}$ which is most closely aligned with the noise matrix $\mathbf{Z}$. We define a matrix $\mathbf{T}^n$ in a manner similar to the way we have defined $\mathbf{T}^s$, whose columns consist of the $\mathbf{t}$ vectors obtained over the various iterations of the solution that are associated with negative eigenvalues. Since $\mathbf{T}^n$ is associated with the structured noise, components of this matrix are excluded from $\mathbf{T}$ in (2-8) – (2-11) for predicting values for $\mathbf{X}$ and $\mathbf{Y}$ using the SC-PCA method

The sign of the dominant eigenvalue depends on the value of the meta-parameter λ. The larger the value of λ, the more sensitive the equation will be to the colinearities between $\mathbf{X}$ and $\mathbf{Z}$. Further details on the properties of this phenomenon, as well as, on the choice of parameter values, is provided in the Appendix.

2.3.2 Soft Constrained PLS (SC-PLS)

We now consider the PLS approach in the soft-constrained, structured noise framework. In this case, we wish to find components $\mathbf{t}_i$ in $\mathbf{X}$ which are closely aligned with $\mathbf{Y}$ while simultaneously suppressing correlation with $\mathbf{Z}$. Following the arguments presented for the SC-PCA method, a suitable criterion in this respect is

$$\begin{aligned} \max_{\mathbf{w}} \quad & \left|\mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_i - \lambda\mathbf{w}_i'\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{w}_i\right| \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{w}_i = \delta_{ij}. \end{aligned} \tag{2-20}$$

The corresponding Lagrangian is given by

$$L(\mathbf{w}) = \left|\mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_i - \lambda\mathbf{w}_i'\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{w}_i\right| - \gamma_i(\mathbf{w}_i'\mathbf{w} - 1) \tag{2-21}$$

where the parameters λ and γ_i play the same role as in the SC-PCR case. After differentiating, rearranging and equating to zero, and assigning $\hat{\gamma_i} = \text{sgn}(\alpha)\gamma_i$ the solution in this case must satisfy

$$\mathbf{X}'\left(\mathbf{Y}\mathbf{Y}' - \lambda\mathbf{Z}\mathbf{Z}'\right)\mathbf{X}\mathbf{w}_i - \gamma_i\mathbf{w}_i = \mathbf{0}. \tag{2-22}$$

Thus the desired solution at each iteration is therefore the dominant eigenvector of the matrix

$$\mathbf{U}_2 = \mathbf{X}'\left(\mathbf{Y}\mathbf{Y}' - \lambda\mathbf{Z}\mathbf{Z}'\right)\mathbf{X} \tag{2-23}$$

In a manner similar to the previous section, we can rewrite (2-21) as:

$$\mathbf{w}'\mathbf{X}'(\mathbf{Y}\mathbf{Y}' - \lambda\mathbf{Z}\mathbf{Z}')\mathbf{X}\mathbf{w} = \mathbf{t}'(\mathbf{Y}\mathbf{Y}' - \lambda\mathbf{Z}\mathbf{Z}')\mathbf{t} = \mathbf{t}'\mathbf{Y}\mathbf{Y}'\mathbf{t} - \mathbf{t}'\lambda\mathbf{Z}\mathbf{Z}'\mathbf{t} \tag{2-24}$$

Thus in a particular iteration, when the dominant eigenvalue $\hat{\gamma}$ of the matrix $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X} - \lambda\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}$ is positive, from the right-most equation of (2-24), we have

$\mathbf{t}^{'}\mathbf{YY}^{'}\mathbf{t} > \mathbf{t}^{'}\mathbf{ZZ}^{'}\mathbf{t}$, Thus in the SC-PLS case the corresponding $\mathbf{t}$ is dominated by the term $\mathbf{t}^{'}\mathbf{YY}^{'}\mathbf{t}$, and consequently $\mathbf{t}$ tends to be in a direction which explains maximum variation along $\mathbf{Y}$. Since $\mathbf{t}$ itself is a linear combination of the columns of $\mathbf{X}$, $\mathbf{t}$ in this case corresponds to a direction in $\mathbf{X}$ which is most closely aligned with $\mathbf{Y}$. Similarly when γ is negative, $\mathbf{t}$ in this case is along a direction in $\mathbf{X}$ which is most closely aligned with the noise matrix $\mathbf{Z}$. The matrices $\mathbf{T}^{s}$ and $\mathbf{T}^{n}$ for the SC-PLS method are defined in an analogous manner to the SC-PCR case, and have corresponding roles in the prediction of $\mathbf{X}$ and $\mathbf{Y}$. This use of latent vectors associated with positive and negative eigenvalues in this SC-PLS method will be illustrated with examples in the application section.

2.4 Simulation Experiments:

In order to show the characteristics of the proposed latent variable methods we design toy problems to demonstrate the performance of our various formulations. We consider two separate cases; the first is where strong structured noise contaminates only $\mathbf{X}$, while the second is where the structured noise contaminates both $\mathbf{X}$ and $\mathbf{Y}$.

We consider a general structure for problems suffering from structured noise with system outputs $y_i \in \mathbb{R}^m$ corresponding to settings of variables

$\mathbf{x}_i \in \mathbb{R}^k, i = 1,\ldots,n$, where $\mathbf{x}_i$ is a row vector. We assume that our measurements of $\mathbf{X}$ and $\mathbf{Y}$ are contaminated with noise:

$$\mathbf{Y} = \mathbf{Y_0} + \mathbf{N_Y} \tag{2-25}$$

$$\mathbf{X} = \mathbf{X}_0 + \mathbf{N_X} \tag{2-26}$$

$\mathbf{Y} \in \mathbb{R}^{n\times m}$ and $\mathbf{X} \in \mathbb{R}^{n\times k}$ is a matrix of measured variables. $\mathbf{X_0}$ and $\mathbf{Y_0}$ are their true underlying values that are then contaminated with the noise $\mathbf{N_x}$ and $\mathbf{N_Y}$ to obtain the measured values of $\mathbf{X}$ and $\mathbf{Y}$. Without loss of generality it is assumed that $n \geq k \geq m$. The matrices $\mathbf{N_Y}$ and $\mathbf{N_X}$ are the noise terms, which contain both structured plus random noise components.

$$\mathbf{N}_Y = \mathbf{Z_Y} + \sigma_\mathbf{Y}\mathbf{E_Y}; \qquad \mathbf{Z_Y} \triangleq \mathbf{T_N}\mathbf{C_Y} \tag{2-27}$$

$$\mathbf{N_X} = \mathbf{Z_X} + \sigma_\mathbf{X}\mathbf{E_X}; \qquad \mathbf{Z_X} \triangleq \mathbf{T_N}\mathbf{C_X} \tag{2-28}$$

The $\mathbf{Z}$-terms are the "common" structured noise components; the columns of the matrix $\mathbf{T_N} \in \mathbb{R}^{n\times s}$ are assumed to be orthonormal latent vectors which describe the structured noise subspace, where $s < n$. The elements of the matrices $\mathbf{E_Y}$ and $\mathbf{E_X}$ are *iid* random variables with unit variance that represent the unstructured noise components. In addition to measurements of $\mathbf{X}$ and $\mathbf{Y}$ we assume that we also have available a matrix $\mathbf{Z}$, which contains measurements of a linear mixture of some components of the structured noise. In other words, it is a linear mixture of some components of $\mathbf{T_N}$ plus *iid* noise defined as:

$$\mathbf{Z} = \mathbf{T_N}\mathbf{C_Z} + \sigma_{\mathbf{Z}}\mathbf{E_Z} \tag{2-29}$$

In typical applications of interests, the matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are of low rank and a statistical relationship exists between $\mathbf{X_0}$ and $\mathbf{Y_0}$ due to the common basis $\mathbf{T_S}$:

$$\mathbf{X_0} = \mathbf{T_s}\mathbf{A_x} + \mathbf{U_X}\mathbf{B_X} \tag{2-30}$$

$$\mathbf{Y_0} = \mathbf{T_s}\mathbf{A_y} + \mathbf{U_Y}\mathbf{B_Y} \tag{2-31}$$

$\mathbf{T_S} \in \mathbb{R}^{n \times a}$ is the score matrix of latent vectors that defines the common basis (latent structure) of $\mathbf{X}$ and $\mathbf{Y}$ (common structured noise). The extra structured components $\mathbf{U_X}$ ($n \times v$) and $\mathbf{U_Y}$ ($n \times j$) are not correlated with each other nor with $\mathbf{T_S}$ and define the structured directions in $\mathbf{X}$ and $\mathbf{Y}$ that are unrelated to $\mathbf{Y}$ and $\mathbf{X}$ respectively (uncommon structured noise). The mixing matrices $\mathbf{A_X}$, $\mathbf{B_X}$, $\mathbf{A_Y}$ and $\mathbf{B_Y}$ are random mixing vectors with zero mean. The objective is to discover the structure of $\mathbf{X_0}$ and $\mathbf{Y_0}$ and the relationship between them by making use of not only the measurement matrices $\mathbf{X}$ and $\mathbf{Y}$ but also the measurement matrix $\mathbf{Z}$, which contains information on some of the structured noise present in $\mathbf{X}$ and $\mathbf{Y}$. We are interested in estimating the latent vectors from $\mathbf{X}$ that maximally explain $\mathbf{Y_0}$ and $\mathbf{X_0}$. To do so efficiently we must account for the influence of the structured noise components in $\mathbf{N_X}$ and $\mathbf{N_Y}$. In the case of large structured noise ($\mathbf{T_N}$) common to both $\mathbf{X}$ and $\mathbf{Y}$, the latent variables yielded by ordinary PLS may become more colinear with $\mathbf{T_N}$ than with $\mathbf{X_0}$ or $\mathbf{Y_0}$. We will demonstrate this case

in the toy example in section 2.4. To estimate the true latent variable structure of $\mathbf{X_0}$ and $\mathbf{Y_0}$ (contained in $\mathbf{T_S}$) it is therefore desirable to determine latent variables with maximum covariance with $\mathbf{X}$ and $\mathbf{Y}$, but also with minimum covariance with the structured noise effects contained in the auxiliary noise measurement matrix $\mathbf{Z}$.

Combining the above equations the matrices $\mathbf{X}$ and $\mathbf{Y}$ can be rewritten as:

$$\mathbf{Y} = \mathbf{T}_s\mathbf{A}_\mathbf{Y} + \mathbf{U}_\mathbf{Y}\mathbf{B}_\mathbf{Y} + \mathbf{T}_N\mathbf{C}_\mathbf{Y} + \sigma_\mathbf{Y}\mathbf{E}_\mathbf{Y} \tag{2-32}$$

$$\mathbf{X} = \mathbf{T}_s\mathbf{A}_\mathbf{X} + \mathbf{U}_\mathbf{X}\mathbf{B}_\mathbf{X} + \mathbf{T}_N\mathbf{C}_\mathbf{X} + \sigma_\mathbf{X}\mathbf{E}_\mathbf{X} \tag{2-33}$$

The overall latent variable and measurement structure of the data is illustrated in Figure 2-1.

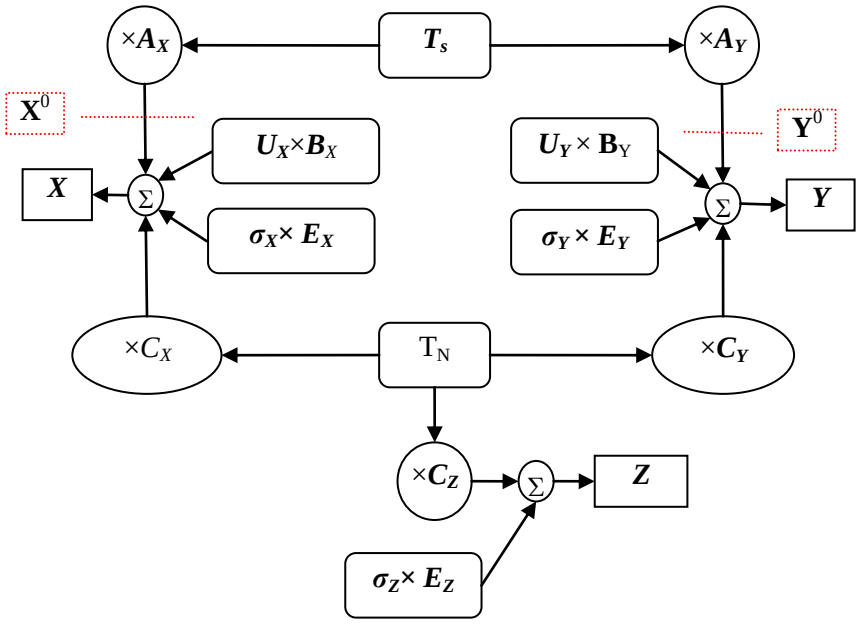


Figure 2-1: Latent structure and relationships between $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ and the noise structure

In order to generate the matrices for the experiments, four random orthonormal matrices are constructed for each simulation: $\mathbf{T}_S \in \mathbb{R}^{n\times 6}$ and $\mathbf{T}_N \in \mathbb{R}^{n\times 6}$, $\mathbf{U}_\mathbf{X} \in \mathbb{R}^{n\times 4}$ and $\mathbf{U}_\mathbf{Y} \in \mathbb{R}^{n\times 4}$, with n equal to 10000 elements (observations) from which the various latent structures in (2-25) - (2-33) are defined.

The objective of the modeling effort is to obtain the best latent variable model for the true underlying signals $\mathbf{X}_0$ and $\mathbf{Y}_0$ using measurements on **X**, **Y** and **Z** as defined in equations (2-25) through (2-33). Since the major interest in building a model between input and response variables is to explore the common subspace between them, two additional variables are defined as:

$$\mathbf{Y}^o = \mathbf{T}_s \mathbf{A}_\mathbf{Y} \tag{2-34}$$

$$\mathbf{X}^o = \mathbf{T}_s \mathbf{A}_\mathbf{X} \tag{2-35}$$

These two variables ($\mathbf{X}^0$ and $\mathbf{Y}^0$) are different from $\mathbf{X}_0$ and $\mathbf{Y}_0$ defined in equations (2-30) and (2-31) as they only define the common subspace of $\mathbf{X}$ and $\mathbf{Y}$ belonging to $\mathbf{T}_\mathbf{S}$. The rationale for choosing these variables for measuring the quality of fit is that $\mathbf{X}$, at its best, can relate to $\mathbf{Y}$ through the common basis identified by $\mathbf{T}_\mathbf{S}$ and therefore what we are interested in is to measure how much of this space can be captured using each method.

The individual quality of fit from a projection of some quantity $\mathbf{\Psi}$ onto the range of each principal component $\mathbf{t}_i$ can be calculated by:

$$R^2_{\mathbf{\Phi}} = 1 - \frac{\left\| \Psi - \hat{\Psi} \right\|^2_F}{\left\| \Psi \right\|^2_F} \tag{2-36}$$

where $\mathbf{\Psi}$ can be any matrix or vector corresponding to the model such as $\mathbf{X}$, $\mathbf{X}^0$, $\mathbf{Y}$, $\mathbf{Y}^0$, or $\mathbf{Z}$. The $\|.\|_F$ operator represents the Frobenius norm of a matrix. The quantity $\hat{\mathbf{\Psi}}$ is calculated by projecting $\mathbf{\Psi}$ onto the range of $\mathbf{t}_i$ as:

$$\hat{\mathbf{\Psi}} = \mathbf{t_i}(\mathbf{t_i}\mathbf{t_i}')^\dagger \mathbf{t_i}'\mathbf{\Psi}. \tag{2-37}$$

The cumulative quality of fit ($R^2_{\mathbf{\Psi}}(cum)$) into the first q latent variables is obtained by calculating $\hat{\mathbf{\Psi}}$ as:

$$\hat{\mathbf{\Psi}} = \mathbf{T}_q(\mathbf{T}_q'\mathbf{T}_q)^\dagger \mathbf{T}_q'\mathbf{\Psi} \tag{2-38}$$

where $\mathbf{T}_q = [\mathbf{t}_1, ..., \mathbf{t}_q]$.

We shall compare the quality of fit between standard PLS, Fearn's OSC-PLS (which performs PLS regression between $\mathbf{X}$ and $\mathbf{Y}$ after removing unrelated variance from $\mathbf{X}$), and the proposed methods: Hard Constrained PLS (HC-PLS), and Soft Constrained PLS (SC-PLS) with different levels of penalty coefficient (λ). More details on the choise of λ is provided in Appendix.

2.4.1 Toy Example: Case 1; Structured noise in X only

The latent variable structure for this example is explained with the aid of Table 2-2. We note that in this specific case, the coefficients $\mathbf{C}_Y$ multiplying $\mathbf{T}_\mathbf{N}$ in (2-32) are zero, meaning that only $\mathbf{X}$ is contaminated with the structured noise. It is assumed that simultaneous measurements of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are available. The statistical norms for the coefficients of the mixing matrices are given in Table 2-1:

TABLE 2-1 FROBENIUS NORMS OF THE MIXING MATRICES FOR THE FIRST EXAMPLE , CORRESPONDING TO $\sigma^2_{\mathbf{X}} \approx 50$, $\sigma^2_{\mathbf{X}^0} \approx 7$, $\sigma^2_{\mathbf{Y}} \approx 5.7$, $\sigma^2_{\mathbf{Y}^0} \approx 6.3$, $\sigma^2_{\mathbf{U_X B_X}} \approx 3.9$, $\sigma^2_{\mathbf{U_Y B_Y}} \approx 3.8$

$\|\mathbf{C}_\mathbf{X}\|_F = 50$	$\|\mathbf{C}_\mathbf{Y}\|_F = 37$
$\|\mathbf{A}_\mathbf{X}\|_F = 14$	$\|\mathbf{A}_\mathbf{Y}\|_F = 10$
$\|\mathbf{B}_\mathbf{X}\|_F = 12.5$	$\|\mathbf{B}_\mathbf{Y}\|_F = 8$
$\sigma_\mathbf{X} = 0.1$	$\sigma_\mathbf{Y} = 0.1$

The size of the mixing coefficient matrices $\mathbf{A}_\mathbf{Y}, \mathbf{B}_\mathbf{Y}, \mathbf{C}_\mathbf{Y}$ and $\mathbf{A}_\mathbf{X}, \mathbf{B}_\mathbf{X}, \mathbf{C}_\mathbf{X}$ are 6×18, 6×18, 4×18, 6×32,6×32 and 4×32 respectively. The mixing matrix for the noise matrix $\mathbf{Z}$ ($\mathbf{C}_\mathbf{Z}$) is 6×6. Hence the size of the produced datasets $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$

will be: 10000×32, 10000×18 and 10000×6 respectively. The number of components removed prior to performing PLS in the OSC-PLS method was 6 in all cases. We chose this number because the basis for the noise ($\mathbf{T}_N$) is constructed from 6 latent vectors.

TABLE 2-2: TOY EXAMPLE, CASE 1, STRUCTURED NOISE IS ONLY PRESENT IN $\mathbf{X}$

Columns of $\mathbf{T}$	$\mathbf{T}_s \in \mathbb{R}^{10000\times 6}$	$\mathbf{T}_N \in \mathbb{R}^{10000\times 6}$	$\mathbf{U}_\mathbf{X} \in \mathbb{R}^{10000\times 4}$	$\mathbf{U}_Y \in \mathbb{R}^{10000\times 4}$
$\mathbf{X} \in \mathbb{R}^{10000\times 32}$	←→	←→	←→	
$\mathbf{Y} \in \mathbb{R}^{10000\times 18}$	←→			←→
$\mathbf{Z} \in \mathbb{R}^{10000\times 6}$		←→		

Our hypothesis is that the presence of significant noise in $\mathbf{X}$ will impede the ability of ordinary PLS methods from extracting components that properly explain the common structure of the datasets, that is $\mathbf{X}^0$ and $\mathbf{Y}^0$ as defined in (2-34)-(2-35). In such cases due to the strong presence of the structured noise in $\mathbf{X}$ some of the extracted components will predominantly explain the structured noise within $\mathbf{X}$, even though it has weak correlation with $\mathbf{Y}$. This is not a major problem if enough PLS components are extracted. However the constraining methods provide a better basis for discovering the true underlying structure of $\mathbf{X}^0$ and $\mathbf{Y}^0$ as they better explain subspaces of $\mathbf{T}_\mathbf{S}$ in fewer extracted components, making the interpretation of the results easier and, in addition, the predicted results will have less variance due to the lack of noise related components. Our

results show that the use of the proposed constrained methods can more efficiently separate the components belonging to subspaces of $\mathbf{T_S}$ (common underlying structure) and $\mathbf{T_N}$ (of structured noise). In the soft constrained case, the association of the positive and negative eigenvalue components with either $\mathbf{T_N}$ or $\mathbf{T_S}$, as we have seen earlier, can greatly simplify the interpretation of the data. From Table 2-2 we note that $\mathbf{T_N}$ is orthogonal to X^0 and $\mathbf{Y}^0$. This is a favorable situation for our simulations. Later we discuss the case where this property does not hold.

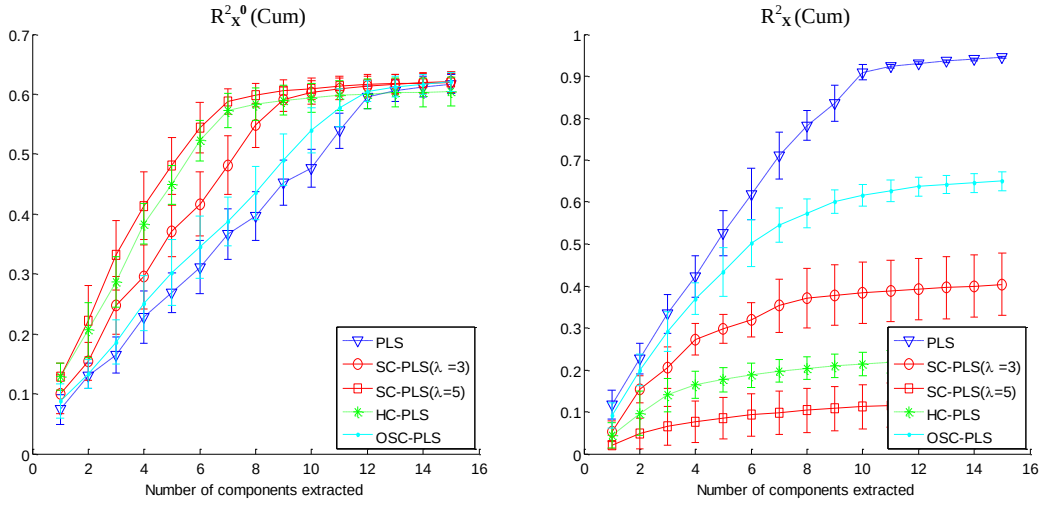


Figure 2-2: The cumulative quality of fit $R^2_{X^0}$ (left) and R^2_X (right) versus the first 15 (positive) LV components components extracted. 6 components have been removed prior to regression in OSC-PLS.

In Figure 2-2 we compare the quality of fit of the measured values of $\mathbf{X}$ and of $\mathbf{X}^o$ (subspace of $\mathbf{X}$ spanned by $\mathbf{T_S}$) for various partial least square methods described earlier, and for the first 15 positive components extracted from SC-PLS

methods (and PLS between **X** and **Y** after removing 6 components in OSC-PLS). These figures are the result of averaging over 100 independent Monte Carlo trials. The error bars indicate the one standard deviation on the results obtained. We can see that the methods which constrain **X** to be orthogonal to **Z** or penalize it, either as soft or hard constraints, have much higher values of $R^2_{\mathbf{X}^o}$ for the components extracted, even though these methods show lower quality of fit when capturing the overall variation of **X** itself. Of course this is an expected result since the constrained methods have removed some of the structured noise components in $\mathbf{T}_N$ through knowledge of the structured noise subspace made available by the measurement of **Z**.

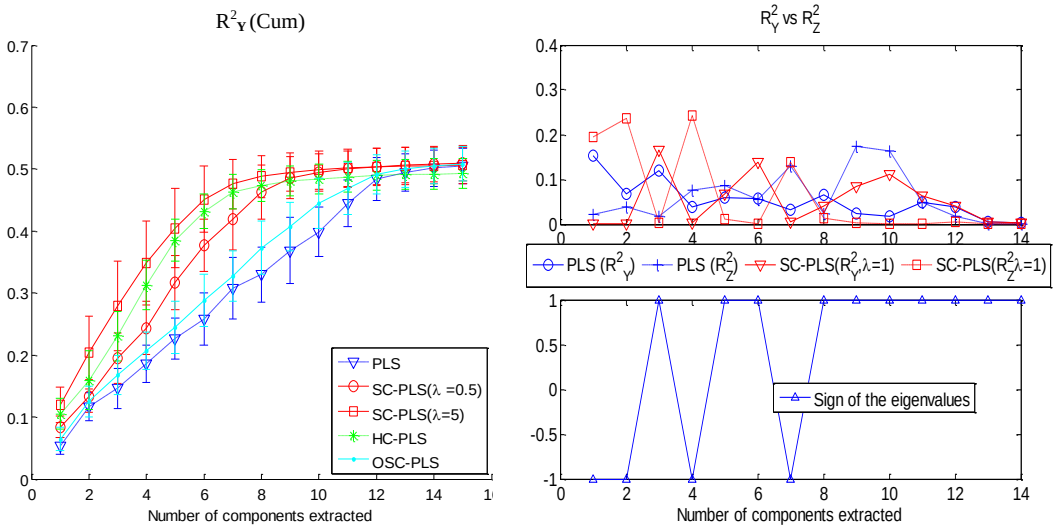


Figure 2-3:Left- The cumulative value of the quantity of fit ($R^2_{\mathbf{Y}}$) versus the number of positive components extracted for each PLS algorithm. Top-Right: quality of fit (non-cumulative) for individual components of PLS method versus SC-PLS ($\lambda = 1$). The lower-right figure shows the sign of each eigenvalue extracted in the SC-PLS method.

Figure 2-3 (left) shows the cumulative $R^2_{\mathbf{Y}}$ vs. the first 15 components extracted for each method. For the soft constrained method only the positive eigenvalue components are counted. We can see that the proposed methods that constrain the noise, i.e. HC-PLS and SC-PLS, for the same number of (positive) components, provide latent vectors that better explain $\mathbf{Y}$ compared to the methods which do not make use of the data on the structured noise, such as ordinary PLS. It should be mentioned again that if a sufficient number of components are selected all these methods (constrained or non-constrained), for this particular example, (when only $\mathbf{X}$ is contaminated with structured noise) will yield the same quality of fit eventually. However, constrained methods provide more information through less number of components.

In Figure 2-3 (right) we show the non-cumulative (individual) $R^2_{\mathbf{Y}}$ and $R^2_{\mathbf{Z}}$ against the number of components for the ordinary PLS method and the proposed SC-PLS approach ($\lambda = 1$). For a fixed value of λ in (2-21) (the penalty on auxiliary noise), the dominant eigenvalue of matrix $\mathbf{U}_2$ in (2-23) is either positive or negative after each sequence of extracting principal components. When this eigenvalue is positive, the corresponding eigenvalue component is weighted towards explaining variance in $\mathbf{Y}$. This is evident from the plot of R^2_Y in Figure 2-3 (right) which shows significant values when the eigenvalue sign (indicated in the bottom panel of the figure) is positive. From the corresponding plots of R^2_Z, we see these components explain relatively low variance in $\mathbf{Z}$. On the other hand,

when the eigenvalue is negative, the extracted component explains variance in $\mathbf{Z}$, (as is evident from the plot of R_Z^2), yet relatively small variance in $\mathbf{Y}$ (we should remind that the cumulative plots shown in Figure 2-2 and Figure 2-3 are obtained by only using latent variables that correspond to the positive eigenvalues explaining Y). The reason for such a behavior can be explain by the fact that the SC-PLS tends to find a linear combination of $\mathbf{X}$ that maximizes the covariance with Y while minimizing the covariance with $\mathbf{Z}$. Due to the nature of the algorithm, this problem can be solved by finding the largest eigenvalue (in magnitude) /eigenvector pair of the corresponding matrix in (2-23). If the corresponding matrix in (2-23) is (semi) negative-definite due to larger variation in $\mathbf{Z}$ (or higher covariance with $\mathbf{X}$, in each iteration), the largest eigenvalue (in magnitude) of $\mathbf{U}_2$ will be negative in sign which in turn means the latent vector and the corresponding component will maximize covariance with $\mathbf{Z}$ while minimizing its covariance with $\mathbf{Y}$.

A comparison between the quality of fit of each positive component extracted and the noise ($\mathbf{R}_\mathbf{Z}^2$) is shown in Figure 2-4. It is evident that the components extracted using the constrained PLS methods and in particular the Hard-Constrained PLS, explain little or no variation of noise (only positive components are shown for SC-PLS)

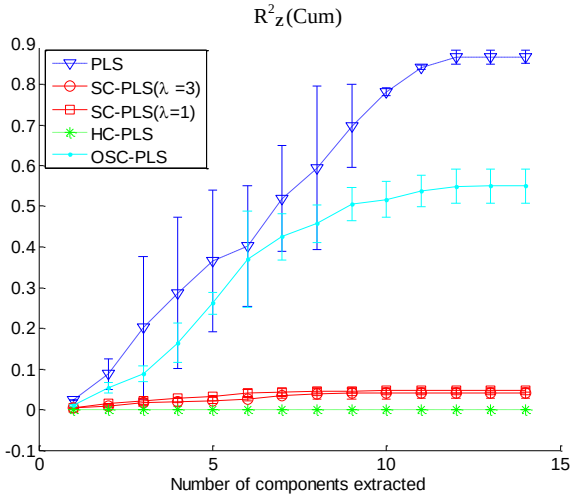


Figure 2-4: Cumulative quality of fit ($R^2_{\mathbf{Z}}$) for noise using the first 15 positive latent components extracted by each model

2.4.2 Toy Example Case 2: Common structured noise in both $\mathbf{X}$ and $\mathbf{Y}$

In this case we assume that both $\mathbf{X}$ and $\mathbf{Y}$ are contaminated with the structured noise ($\mathbf{T}_N$). This is a more important case for the use of the constrained methods outlined in this paper. We have mentioned earlier that the presence of structured noise in both $\mathbf{X}$ and $\mathbf{Y}$ can lead to misleading results. This example will illustrate the problem. The structure of the matrices generated is shown in Table 2-3:

TABLE 2-3: TOY EXAMPLE, CASE 1, STRUCTURED NOISE IS PRESENT IN BOTH $\mathbf{X}$ AND $\mathbf{Y}$

Columns of $\mathbf{T}$	$\mathbf{T}_s \in \mathbb{R}^{10000\times6}$	$\mathbf{T}_N \in \mathbb{R}^{10000\times6}$	$\mathbf{U}_X \in \mathbb{R}^{10000\times4}$	$\mathbf{U}_Y \in \mathbb{R}^{10000\times4}$
$\mathbf{X} \in \mathbb{R}^{10000\times32}$	←→	←→	←→	
$\mathbf{Y} \in \mathbb{R}^{10000\times18}$	←→	←→		←→
$\mathbf{Z} \in \mathbb{R}^{10000\times6}$		←→		

Our hypothesis is that the constrained LVM methods are more capable of capturing the true underlying structure of $\mathbf{X}^0$ and $\mathbf{Y}^0$ compared to the unconstrained methods. In addition, we also show that OSC-PLS (2-1) will not perform properly when common structured noise is contaminating both $\mathbf{X}$ and $\mathbf{Y}$ spaces. In such a case, OSC-PLS will only remove the uncommon structured noise components from $\mathbf{X}$ and will leave the common structured noise part in $\mathbf{X}$ intact. The settings used to construct the simulation data are given in Table 2-4

TABLE 2-4: SETTINGS USED FOR SECOND SIMULATION DATASET, $\sigma^2_{\mathbf{X}}\approx12.8$, $\sigma^2_{\mathbf{X}^0}\approx5.8$, $\sigma^2_{\mathbf{Y}}\approx11.3$, $\sigma^2_{\mathbf{Y}^0}\approx5.7$, $\sigma^2_{\mathbf{U_{XB_X}}}\approx3.8$, $\sigma^2_{\mathbf{U_{YB_Y}}}\approx3.8$

$\|\mathbf{C_X}\|_F = 13$	$\|\mathbf{C_Y}\|_F = 9$
$\|\mathbf{A_X}\|_F = 14$	$\|\mathbf{A_Y}\|_F = 10$
$\|\mathbf{B_X}\|_F = 12.5$	$\|\mathbf{B_Y}\|_F = 0.8$
$\sigma_{\mathbf{X}} = 0.1$	$\sigma_{\mathbf{Y}} = 0.1$

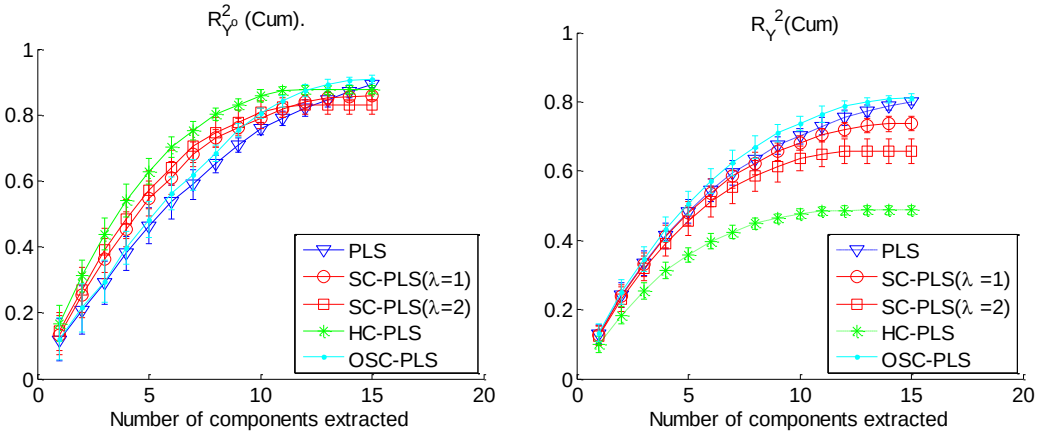


Figure 2-5: Right: cumulative ($R^2_{\mathbf{Y}}$) for the first 15 positive components extracted for various LVM methods explained earlier. Left: cumulative ($R^2_{\mathbf{Y}^0}$) for the first 15 positive components . Number of components removed before performing PLS in OSC-PLS was 6.

As in the previous case, the data for this experiment is simulated using a sequence of 100 Monte Carlo trials. Figure 2-5 shows the cumulative $\mathbf{R}^2$ for various methods explained in the second toy problem for the first 15 components extracted (only positive components in case of soft constrained methods and only the PLS components after deflating $\mathbf{X}$ in OSC-PLS). In Figure 2-5 (right) we can see that even though $\mathbf{R}^2_{\mathbf{Y}}$ for the constrained methods is lower, the extracted components in these methods explain $\mathbf{Y}^0$ (the part of $\mathbf{Y}$ in $\mathbf{T}_S$) better (within the confidence intervals) than the non constrained methods.. It should be noted that SC-PLS does not necessarily improve computational cost but rather improves the component selection criteria as proper components are identified by their eigenvalue signs. When both $\mathbf{Y}$ and $\mathbf{X}$ are contaminated, OSC-PLS will not remove the common structured noise components. This is apparent from the plots in Figure 2-5 (only slight improvement over PLS). Since OSC-PLS was unable to remove the common structured noise, the noise still remains in the subspace of $\mathbf{X}$ and will contaminate the extracted components. The reason for having very similar plots in both OSC-PLS and regular PLS is that the same noise components exist in both $\mathbf{X}$ and $\mathbf{Y}$ subspaces. Therefore projecting $\mathbf{X}$ into $\mathbf{Q}_{\mathbf{XY}}$ will not remove any components related to the structured noise at all. A major consequence of having common noise in both $\mathbf{X}$ and $\mathbf{Y}$ and not accounting for it in the model is

that the model does a presumably good job in modeling the relation between the measured $\mathbf{X}$ and $\mathbf{Y}$ matrices but in reality it is also modeling the relationship between the common structured noise components contaminating $\mathbf{X}$ and $\mathbf{Y}$ and not the true underlying structure of interest. This will result in lowered quality of prediction and higher number of components that need to be extracted. It should be noted that even when there is common structured noise in both **X** and **Y** subspaces the quality of fit ($\mathbf{R^2_{Y^0}}$ and $\mathbf{R^2_Y}$) will be high in all cases, however, the prediction rates will be very different. Such behavior is a clear case of over fitting data and should be avoided.

In order to shed more light on this subject a new set of data ("*test set*") was generated using the same parameters used for the current simulation study. The *test set* data ($\mathbf{X}^{ts}$, $\mathbf{Y}^{ts}$) was later used to measure the quality of each model to predict the corresponding values $\hat{\mathbf{Y}}^{ts}$ of $\mathbf{Y}^{ts}$'s for the test set. The cumulative quality of prediction for $\hat{\mathbf{Y}}^{ts}$ ($Q^2_{\mathbf{Y}}$) for the first *q* components was calculated from:

$$Q^2_{\mathbf{Y}} = 1 - \frac{\left\| \mathbf{Y^{ts}} - \hat{\mathbf{Y}}^{\mathbf{ts}} \right\|^2_F}{\left\| \mathbf{Y^{ts}} \right\|^2_F} \qquad (2\text{-}39)$$

In addition to $Q^2_{\mathbf{Y}}$ we define an additional quality parameter denoted by $Q^2_{\mathbf{Y}^0}$ as:

$$Q^2_{\mathbf{Y}^0} = 1 - \frac{\left\| \mathbf{Y^{0ts}} - \hat{\mathbf{Y}}^{\mathbf{ts}} \right\|^2_F}{\left\| \mathbf{Y^{0ts}} \right\|^2_F} \qquad (2\text{-}40)$$

This parameter measures how close the predicted $\hat{\mathbf{Y}}$ and $\mathbf{Y}^0$ (which is the only

part of **Y** that can be predicted by **X**) are to one another.

The quality of prediction (Q^2) is calculated for various methods (i.e. PLS, OSC-PLS, SC-PLS and HC-PLS). $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}^{ts}$ are calculated from $\hat{\mathbf{Y}}^{ts} = \mathbf{X}^{ts}\mathbf{W}^*\mathbf{Q}'$. The mixing matrix **W*** is obtained from the training dataset by regressing training set **T** on **X** [14]. $Q^2_{\mathbf{Y}}$ measures the overall quality of prediction for the test set values which also include noise components whereas the $Q^2_{\mathbf{Y^0}}$ measures the quality of fit to the true underlying relationship between **X** and **Y** which is defined by $\mathbf{T_S}$.

In Figure 2-6, We can see that despite higher $Q^2_{\mathbf{Y}}$ in the non constrained methods, the constrained methods provide much better prediction results for $\mathbf{Y^0}$ compared to regular PLS and OSC-PLS ($Q^2_{\mathbf{Y^0}}$).

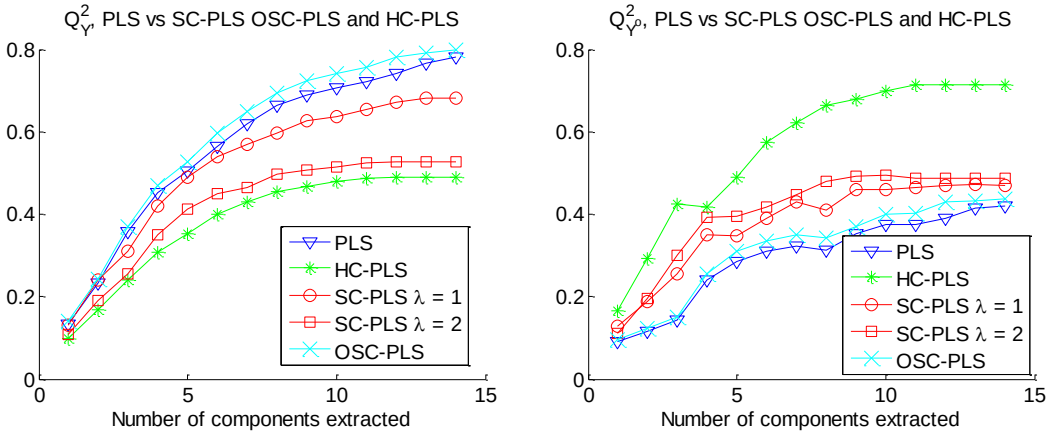


Figure 2-6: Left: cumulative quality of prediction for **Y** ($Q^2_{\mathbf{Y}}$). Normal PLS ostensibly provides the best results, but measuring the quality of prediction for $\mathbf{Y^0}$ ($Q^2_{\mathbf{Y}^0}$), on the right shows that non-constrained methods are actually modeling the common structured noise. The quality of fit to the noiseless data is much lower compared to the constrained methods.

2.4.3 Toy Example: comparison of performance between SC-PLS vs. HC-PLS

A few notes on Hard Constrained latent variable methods

Hard Constrained LVMs are powerful methods for removing unwanted variance from datasets. It is possible to assume the following general structure for hard constrained signal correction methods including OSC-PLS:

$$\begin{aligned} \max \quad & \mathbf{w}'\Sigma\mathbf{w} \quad (a) \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = \delta \quad (b) \\ & \mathbf{w}'\mathbf{X}'\mathbf{K} = \mathbf{w}'\Lambda = \mathbf{0} \, (c) \end{aligned} \tag{2-41}$$

where in HC-PLS $\mathbf{K} = \mathbf{Z}$ and in OSC-PLS method K = Y. (2-41)-(c) indicates that:

$$\Lambda'\mathbf{w} = 0 \rightarrow \mathbf{w} \in \mathcal{N}(\Lambda') \rightarrow \mathbf{w} = \mathrm{M}_{\Lambda}\mathbf{c} \tag{2-42}$$

This means that $\mathbf{w}$ lies in the null space of Λ' ($\mathbf{w}=\mathbf{M}_{\Lambda}\mathbf{c}$). A basis for the null space of Λ' is M_{Λ} defined by (2-2). The vector $\boldsymbol{c}$ is an arbitrary vector of appropriate length. The rank and dimensions of Λ can produce various outcomes. We consider the cases where Λ can be tall, square and rank deficient matrix, square and full rank or short. Each of these conditions will lead to a different outcome, as we now discuss:

Λ is tall or rank deficient:

This situation can arise when $\mathbf{Z}$ in (2-7) or (2-12) ($\mathbf{Y}$ in OSC-PLS case) has fewer columns than (rank of) $\mathbf{X}$. In this case Λ' has a non-empty null space, and hence Λ has an orthogonal complement which serves as a basis for the null space of Λ'.

Λ is full rank square or short:

This situation is the opposite of what was mentioned above when $\mathbf{Z}$ (in HC-PCR or HC-PLS) or $\mathbf{Y}$ (in OSC-PLS) have more columns than $\mathbf{X}$ has. In this case if the rank of $\mathbf{Y}$ or $\mathbf{Z}$ is larger than the rank of $\mathbf{X}$ then Λ' has an empty null space and hence an orthogonal complement for Λ does not exist. An example of such a situation is when multiple sensors are used to measure a low rank structured noise in a plant environment which are generated by fewer number of sources. Due to the contamination of the structured noise matrix ($\mathbf{Z}$) by random noise, it will become full rank and because there are more sensors than real sources of structured noise the Λ' in (2-41) will have fewer columns than rows and hence will be full rank. In this case a proper null space for Λ' cannot be found.

Why soft constraints?

If the auxiliary noise matrices $\mathbf{Z}$ or $\mathbf{Y}$ in OSC-PLS are conditioned the way we mentioned above ($\mathcal{N}(\mathbf{\Lambda}')$ is empty) then applying hard constrints may not provide a good solution. Again, let's assume that the noise subspace has a rank of "*s*" but $\mathbf{Z}$ has *r*>*s* columns and the measurements are noisy. Then $\mathbf{Z}$ will behave as a full rank matrix with rank *r* >*s*. Now if this matrix is used to impose hard constraints, it will actually remove some of the directions in $\mathbf{X}_0$ that are not relevant to the underlying structured noise. Another issue that can be raised while using hard constraint is the fact that in many applications the structured noise components cannot be measured purely (to contain only information about the noise subspace) and they might also contain some information on the latent structures of $\mathbf{X}$ and $\mathbf{Y}$

. In other words $\mathbf{Z}$ is not completely orthogonal to subspace of $\mathbf{T_S}$. In such a case, imposing a hard constraint on $\mathbf{Z}$ can actually remove some components of the underlying latent structure between $\mathbf{X}$ and $\mathbf{Y}$ ($\mathbf{T_S}$) which will lead to lowered efficacy of the model. In order to circumvent this difficulty we proposed incorporating soft constraints rather than the hard constraints.

In this toy example we compare the performance of soft constrained versus the hard constrained PLS (SC-PLS vs. HC-PLS) for two particular cases; 1- when the auxiliary noise matrix also contains some components of $\mathbf{T_S}$ (in other words $\mathbf{Z}$ is not orthogonal to $\mathbf{T_S}$) and 2- when the auxiliary matrix $\mathbf{Z}$ is full rank (mainly because of the addition of noise , not just because we have more columns in $\mathbf{T_N}$ than $\mathbf{X}$). The first case is a very common situation. An example of the first situation was mentioned earlier in the introduction (example 4). In this example the EOG electrode, in addition to the ocular artifacts from the eyes, also records some weakened (due to distance from the brain) brain signals ($\mathbf{T_S}$) arising from the stimulations. An example of the second case is when numerous sensors are located in various areas of a plant to record the instances of the systematic structured noise, with low rank. However, even though the actual number of noise source components may be limited, the rank of the noise matrix is full due the presence of unstructured noise picked up by the sensors.

For each case a new simulation dataset is generated. In the first case, some components of $\mathbf{T_s}$ are added to $\mathbf{Z}$ by adding $\mathbf{T_S A_Z}$ to equation (2-29), and in the

second case the rank of $\mathbf{Z}$ is changed by changing the number of columns of $\mathbf{C_Z}$ in (2-29) and adding white noise to it. As mentioned earlier, our hypothesis is that in these circumstances using soft constrained PLS can perform better than the hard-constrained method.

When $\mathbf{Z}$ contains components of $\mathbf{T_s}$ ($\mathbf{Z}$ not orthogonal to $\mathbf{T_S}$)

In this example we assume measurements of $\mathbf{Z}$ also contain components of $\mathbf{T_S}$. Hence $\mathbf{Z}$ is generated in the simulation as:

$$\mathbf{Z} = \mathbf{T_N C_Z} + \sigma_Z \mathbf{E_Z} + \mathbf{T_S A_Z} \tag{2-43}$$

The new structure and the corresponding combination tables are shown in Figure 2-7 and Table 2-5.

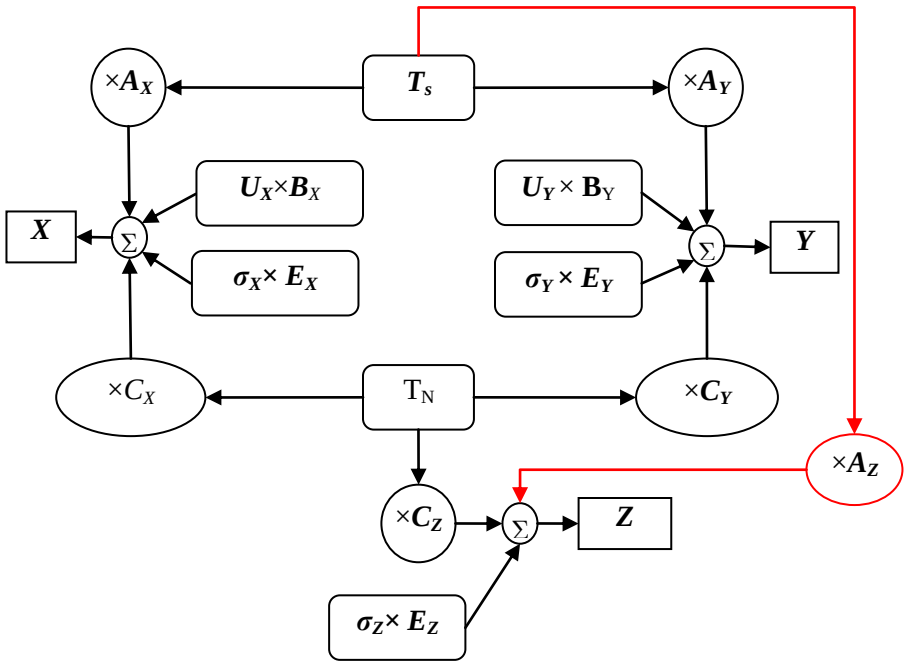


Figure 2-7: Relationship between system components when $\mathbf{Z}$ is contaminated with components of $\mathbf{T_S}$

The following table shows how these components are generated for this simulation:

TABLE 2-5: RELATIONSHIP BETWEEN $\mathbf{X}$, $\mathbf{Y}$ AND $\mathbf{Z}$ AFTER ADDITION OF COMPONENTS IN $\mathbf{T_S}$ TO $\mathbf{Z}$

Columns of $\mathbf{T}$	$[\mathbf{t}_{s1},...\mathbf{t}_{s3}] \in \mathbb{R}^{10000\times3}$	$[\mathbf{t}_{s4},...\mathbf{t}_{s6}] \in \mathbb{R}^{10000\times3}$	$\mathbf{T}_N \in \mathbb{R}^{10000\times6}$	$\mathbf{U}_X \in \mathbb{R}^{10000\times4}$	$\mathbf{U}_Y \in \mathbb{R}^{10000\times4}$
$\mathbf{X} \in \mathbb{R}^{10000\times32}$	↔	↔	↔	↔	
$\mathbf{Y} \in \mathbb{R}^{10000\times18}$	↔	↔			↔
$\mathbf{Z} \in \mathbb{R}^{10000\times6}$	↔		↔		

Figure 2-8-right shows the quality of prediction for the dataset we generated by adding components of $\mathbf{T_S}$ into $\mathbf{Z}$ (as in Table 2-5) and the plots on the left belong to a normal case where $\mathbf{Z}$ does not contain any components of $\mathbf{T_S}$ ($\mathbf{Z}$ being orthogonal to $\mathbf{X}$ as shown in Table 2-2). In this scenario, the following coefficient values are used:

$\|\mathbf{A_X}\|_F = 13, \|\mathbf{B_X}\|_F = 12, \|\mathbf{A_Y}\|_F = 9$, $\|\mathbf{B_Y}\|_F = 8$, $\sigma_x = \sigma_Y = 0.1$.

As hypothesized, imposing hard constrints on the model when $\mathbf{Z}$ is not fully orthogonal to $\mathbf{X_0}$ (when $\mathbf{Z}$ contains components of $\mathbf{T_S}$) does not produce very satisfactory results.

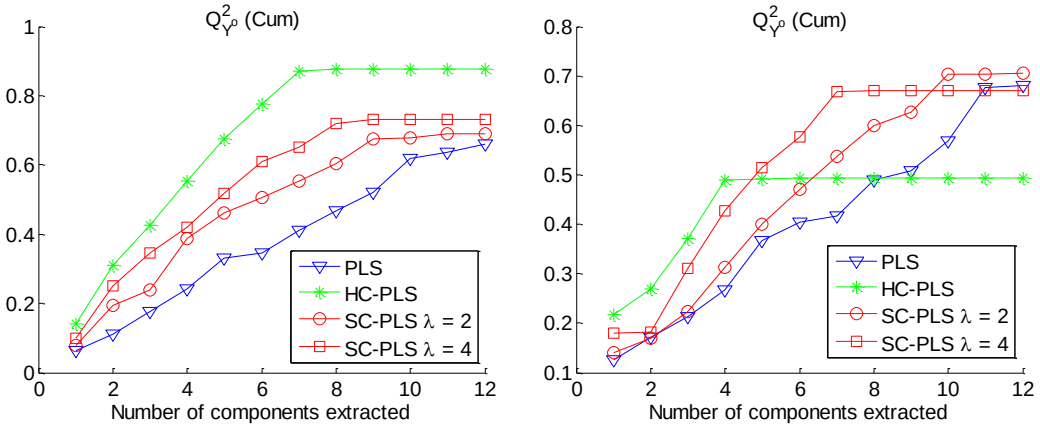


Figure 2-8: comparison of two cases; when **Z** is orthogonal to **X** (left plots) versus the case in which **Z** is not orthogonal to **X** (right plots). The quality of prediction ($Q^2_{Y}{}^0$ (cum)) for the hard constrained method degrades when there is not a true orthogonality between **Z** and **X**. The simulation data for the right plots were generated using the same setting as the left case but with three components of Ts randomly mixed and added to **Z**

We can see that compared to the hard constrained case, the soft constrained method provides better results because it does not fully orthogonalize $\mathbf{t}_i$'s to the subspace of $\mathbf{Z}$. Since $\mathbf{Z}$ contains components of $\mathbf{T}_S$, hard constraints will orthogonalize the $\mathbf{t}_i$ to those components of $\mathbf{T}_S$, thus removing some of the signal variation in $\mathbf{X}^0$ and lowering the quality of fit to subspace of $\mathbf{Y}$.

HC-PLS vs. SC-PLS when the rank of $\mathbf{Z}$ changes

The efficacy of hard constrained LVM methods relies on the proper collection of the auxiliary noise matrix $\mathbf{Z}$. As mentioned earlier, there are many cases where the number of columns in $\mathbf{Z}$ exceeds the number of columns in $\mathbf{X}$ and, even if $\mathbf{T}_N$ is very low rank, due to the presence of unstructured noise in $\mathbf{Z}$ (when recording the data) $\mathbf{Z}$ becomes full rank. In such cases imposing hard constraints can reduce the quality of fit as the orthogonal complement of $\mathbf{X}'\mathbf{Z}$ does not exist,

or spans very few directions. For these cases it is better to use soft constraints as they only penalize these extra dimensions and do not fully discard them. To show this, we generate a simulation dataset using the parameters shown in the caption of Figure 2-9. However in each simulation trial we change the rank of the auxiliary noise matrix $\mathbf{Z}$ by changing the number of columns in the mixing matrix ($\mathbf{C_Z}$) and adding *iid* noise to it. The simulation results show that as $\mathbf{Z}$ becomes larger (more columns), the quality of fit for HC-PLS degrades and eventually will have a lower quality of fit compared to SC-PLS. This example shows that in such cases it is advantageous to use soft constraints which are less sensitive to the conditions of the auxiliary noise matrix.

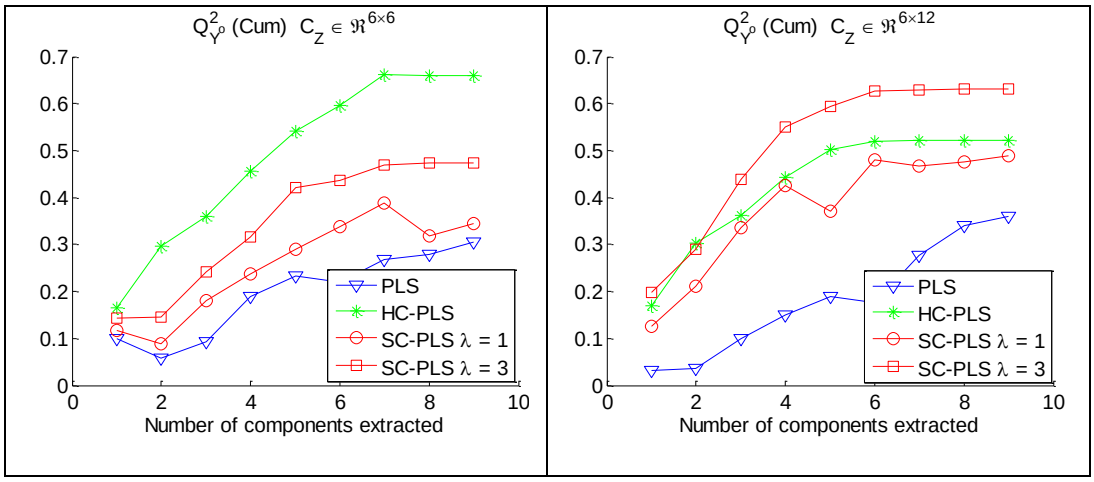


Figure 2-9: quality of prediction ($Q^2_{Y^0}$) for various mixing sizes of $\mathbf{Z}$. When $\mathbf{Z}$ gets larger and under determined SC-PLS outperforms HC-PLS. $\mathbf{C_Z}$ determines the size of $\mathbf{Z}$. (parameters used for this simulation: $\|\boldsymbol{C_X}\|_F = 15, \|\boldsymbol{C_Y}\|_F = 10, \|\boldsymbol{A_X}\|_F = 14, \|\boldsymbol{A_Y}\|_F = 14, \|\boldsymbol{B_X}\|_F = 12, \|\boldsymbol{B_Y}\|_F = 8,\ \sigma_{\boldsymbol{X}} = \sigma_{\boldsymbol{Y}} = 0.1$

2.5 Industrial example

In addition to latent variable methods being used for removing structured noise,

in latent variable methods, the constrained Latent Variable methods are also applicable to other cases such as constrained optimization problems. As an example, we will look at a problem where it is desired to optimize a response through adjusting the process variables, but at the same time constrain the values of other response variables. In such optimization problems the number of variables is often high and the models are reduced rank in nature (i.e. number of LV's is less that the number of $\mathbf{X}$ variables) and so the only space where one has a causal model within which one can optimize is the low dimensional LV space [15, 16, 17]. In this section we show how the constrained latent space methods developed in this paper can be used to create search spaces that conditionally optimize some of the final product quality properties ($\mathbf{Y}$-space), while retaining others at specified values. This approach does not allow for inclusion of other constraints on the solution such as hard constraints on some of the $\mathbf{X}$'s, etc. A more general optimization based approach is laid out in the references cited. However, if there are no operating constraints in the $\mathbf{X}$-space, the approach using the constrained latent variable methods proposed in this thesis provides a simple and less computationally intense solution. This example is provided here mainly to illustrate the extension of the proposed methods to additional types of problems, beyond noise removal.

The process considered is a high pressure process for the manufacture of low density polyethylene. A discussion of the process and PLS modeling are given in [18]. In this particular polymerization process example there are $n = 53$ samples of

the $\mathbf{X}$-variables (input measurements) and the corresponding observations in $\mathbf{Y}$, $k = 14$ process variables ($\mathbf{X}$) and $m = 5$ product quality variables ($\mathbf{Y}$). The response ($\mathbf{Y}$) variables are denoted "Conv, Mn, LCB, SCB and Mw" respectively. The objective of this study is to find conditions in $\mathbf{X}$ that will minimize the "Mw" variable while not affecting the other four $\mathbf{Y}$ variables. This problem can be reformulated into the context of the HC-PLS algorithm by dividing the $\mathbf{Y}$ columns into two groups: the first group consists only of the column corresponding to the observation Mw, ($\mathbf{Y}$= [Mw]) , while the second group (consisting of the columns corresponding to the remaining product quality variables) is assigned to the $\mathbf{Z}$ block ($\mathbf{Z}$ = [Conv, Mn, LCB, SCB]). The latent variables obtained using HC-PL are in $\mathcal{R}(\mathbf{X})$, and are correlated only with Mw, and are orthogonal to all the other quality variables. If $\mathbf{X}$ is decomposed by the HC-PLS algorithm into two subspaces, a latent variable space defining $\mathbf{Y_1}$ and $\mathbf{X}$ and a residual space:

$$\mathbf{X} = \mathbf{TP}' + ... + \mathbf{X_{res}} \tag{2-44}$$

then the components of $\mathbf{X}$ that cause variation in other quality variables ($\mathbf{Z}$) are captured in $\mathbf{X}_{\text{res}}$. Hence moving in the space defined by the components in $\mathbf{T}$ should only modify "Mw" while keeping the other product properties almost constant. An optimization problem (in our case minimization of "Mw" while keeping the other responses unchanged at their measured values for each observation) can be written for each observation *i* as:

$$\begin{aligned} &\min_{\underline{\mathbf{t}}_{new}} \; y_i \\ &\text{s.t.} \qquad \min(\mathbf{T}) < \underline{\mathbf{t}}_{new} < \max(\mathbf{T}) \end{aligned} \tag{2-45}$$

$$y_i = (\underline{\mathbf{t}}_{new}\mathbf{p}' + \underline{\mathbf{x}}_{i-res})\mathbf{B} \tag{2-46}$$

$\mathbf{B}$ is the regression coefficient obtained by projecting $\mathbf{Y}$ into $\mathbf{X}$ and can be obtained from any suitable regression method. The indiced, underlined variables represent i^{th} row of the matrix, corresponding to the i^{th} observation. If it is desirable to minimize quality variables for all observations, the optimization is looped through all observations (as in this example). Since the $\mathbf{X}$ components are highly correlated in most of such examples it is suggested to use PLS, PCR or regularized regression.

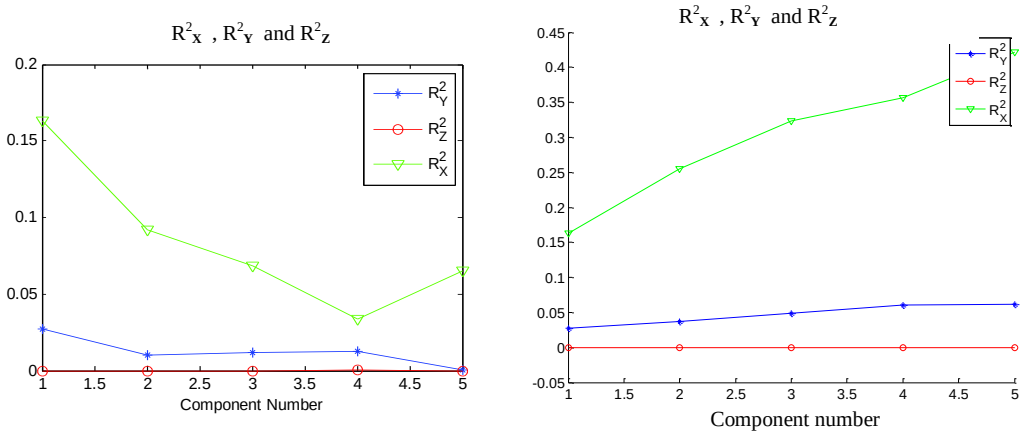


Figure 2-10: Left: individual component quality of fit. Right: cumulative quality of fit ($R^2_{\mathbf{X}}$, $R^2_{\mathbf{Y}}$ and $R^2_{\mathbf{Z}}$ versus q, the number of latent variables for the industrial simulation example). $\mathbf{Y}$ = [Mw], $\mathbf{Z}$ = [Conv, Mn, LCB, SCB]

To test the proposed method, the matrix $\mathbf{P}$ was determined using (2-12) and (2-11). Predicted values for $\mathbf{X}$ and $\mathbf{Y}$ were then calculated from (2-44) and (2-46). Figure 2-10 shows $R^2_{\mathbf{X}}$, $R^2_{\mathbf{Y}}$ and $R^2_{\mathbf{Z}}$ vs. q (number of components

extracted) resulting from this procedure. We can see that $R_{\mathbf{Z}}^2$ is near zero as desired, meaning that the extracted latent variables are orthogonal to $\mathbf{Z}$ (the matrix containing product qualities that are required to remain unchanged). We also see that the process yields latent variables which are predictive of that subspace of $\mathbf{X}$ that is both correlated to Mw ($\mathbf{Y}$) and orthogonal to the other responses ($\mathbf{Z}$). The lower $R^2_{\mathbf{Y}}$ is because the model is predicting only that subspace of $\mathbf{Y}$ that is orthogonal to $\mathbf{Z}$. In other words, it is the independent part of $\mathbf{X}$ affecting only "Mw" which was captured by $\mathbf{T}$. This means only this small change can be applied without affecting the quality of the other output variable from their current values.

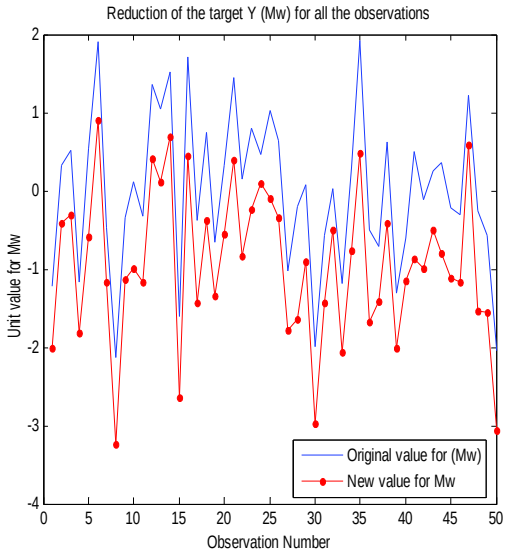


Figure 2-11: Normalized Mw vs. observation index, comparing pre-- vs. post--optimization values. The values are normalized by dividing the difference by the standard deviation of the original quality variables.

Figure 2-11 compares normalized pre- and post-optimization Mw values. It can be seen that the optimization process is indeed effective at reducing the level of this quantity. Further, Figure 2-12 shows normalized pre- and post-optimization changes (the plotted values are equal to $(y_{\text{pre}} - y_{\text{post}})/\sigma_{\text{pre}}$). We can see that this procedure preserves the levels of all the other quality variables except for Mw, which has been reduced in value by somewhat less than a standard deviation.

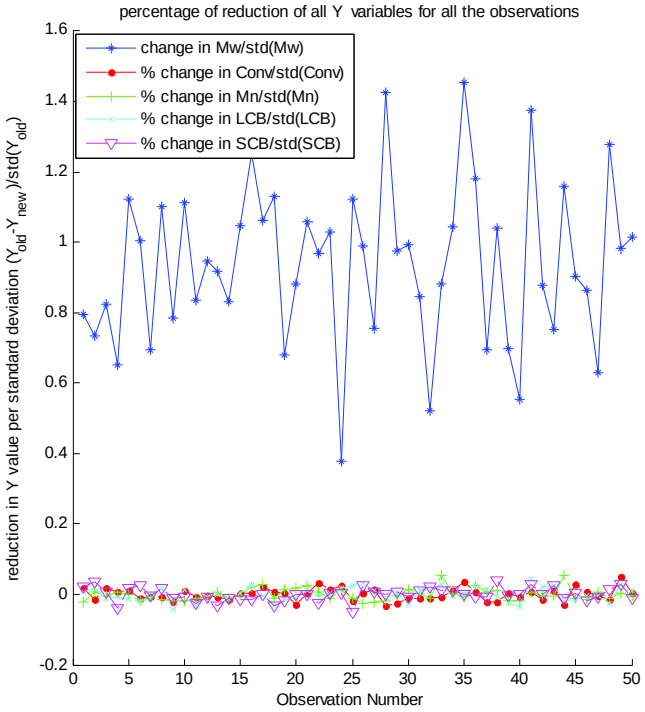


Figure 2-12: changes in output product quality values for all the observations in the dataset

In this example, HC-PLS method was used to optimize the variables in $\mathbf{X}$ to minimize the level of Mw, such that the values of all the other responses in each

experiment remained unchanged. This procedure results in a minimal change to all remaining quality variables. It may be possible to reduce the level of Mw further using one of the soft-constrained methods of Sect. 2.3.2 at the expense of some variation in the values of the remaining outputs.

2.6 Conclusion

When there is strong structured noise in either $\mathbf{X}$ or in both in $\mathbf{X}$ and $\mathbf{Y}$ and there is some information available on the noise, either as simultaneous measurements ($\mathbf{Z}$) or as a covariance matrix ($\mathbf{\Lambda}$), then constrained latent variable methods can provide improved models for the true (noise free) underlying $\mathbf{X_0}$ and $\mathbf{Y_0}$ spaces. In the case that there are low levels of noise or there is no information available about the noise, normal latent variable methods still provide the best solution. But, in the case where one has such auxiliary information on the noise, the constrained latent variable methods presented here, that make use of this information, are shown to provide better models. These latent variable algorithms are developed from an objective function framework in which the covariance among the $\mathbf{X}$ and $\mathbf{Y}$ variables are maximized subject to various hard or soft orthogonality constraints on the noise information ($\mathbf{Z}$ or $\mathbf{\Lambda}$).

Our results suggest that in the presence of common structured noise in both X and Y subspaces the use of PLS method can lead to misleading results. The method we have proposed here shows more stability and less over-fitting in such cases. However these methods prove to be effective when the signal to noise ratio

is very low and contamination from structured noise is so large that it can overcome the latent components of the signal (roughly speaking SNR<3 db). If the SNR is high, then normal PLS can provide adequate results. We also showed that if common structured noise is contaminating both X and Y then methods such as OSC-PLS will perform poorly. An specific example that we showed was the case where the same subspace of noise was contaminating both X and Y. In such case OSC-PLS will only remove structured noise (from X) that is uncommon to Y.

The effectiveness of SC-PLS depends on choice of λ. We have offered some insights into the choice of λ in the appendix; however, this matter requires further investigation and is a topic of future work. Another advantage of our proposed methods is better component selection. Even in the case that only $\mathbf{X}$ is contaminated by noise, incorporating the additional available information about the noise in our methods can reduce the number of components required to properly identify a system. Although regular PLS and OSC-PLS still provide good quality of fit and prediction they require more components to obtain the same quality of fit and prediction. Such performance is required in many cases such as visualizing the data through the score plots or in cases such as nonlinear kernel methods where the dimension of the datasets are potentially very large.

The SC-PLS method provides the ability to identify the noise and signal latent components through their corresponding eigenvalue sign. This property can be very useful in signal processing, for example, in removing ocular artifacts from

EEG data. In the case where only X is contaminated with structured noise (to which we have access through Z) using constrained methods and PLS eventually provides the same quality of prediction and fit as the proposed constrained methods. However in engineering we are not only interested in prediction of Y alone. This is often only a part of the problem. We are also interested in modeling the X space as this is used in monitoring, interpretation and optimization.

2.7 Appendix

In addition to the methods mentioned in the article, latent variable methods can be modified to suppress noise from the components extracted when auxiliary noise information is available. We shall introduce some further additional Hard and Soft Constrained methods in the following appendix. Additionally, more insights into the parameter selection in soft constrained methods will be provided.

2.7.1 Further insights into soft constrained methods

The choice of meta-parameter λ can change the outcome of the modeling. This topic requires further investigation and is a subject of future work. However, we can offer some insight into the choice of the meta parameter λ in (2-22) and (2-18) in section 2.3.1 as follows: Let $(\rho_i, \mathbf{v}_i)$ represent the eigen-decomposition of the matrix $\mathbf{ZZ}'$ where the eigenvectors are normalized to the unit norm. Then $\mathbf{U}_1$ can be written in the form

$$\begin{aligned} \mathbf{U}_1 &= \mathbf{X}'\left(\mathbf{I}-\lambda\mathbf{Z}\mathbf{Z}'\right)\mathbf{X} = \mathbf{X}'\left(\mathbf{I}-\lambda\sum_{i=1}^{q}\rho_i\mathbf{v}_i\mathbf{v}_i'\right)\mathbf{X} \\ &= \mathbf{X}'\left(\mathbf{I}-\lambda\rho_j\mathbf{v}_j\mathbf{v}_j'\right)\mathbf{X} - \lambda\mathbf{X}'\left[\sum_{i\neq j}\rho_i\mathbf{v}_i\mathbf{v}_i'\right]\mathbf{X} \end{aligned} \tag{2-47}$$

By choosing $\lambda = 1/\rho_j$, the term in the round brackets in (2-47) becomes a projector onto the orthogonal complement space of the j^{th} eigenvector $\mathbf{v}_j$ of $\mathbf{ZZ}'$. Thus (2-17) is equivalent to $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{w}_i = \gamma_i \mathbf{w}_i$, where $\tilde{\mathbf{X}} = \mathbf{M}_{\mathbf{v}_j}\mathbf{X}$ is the projection of $\mathbf{X}$ onto the orthogonal complement space of $\mathbf{v}_j$ and thus $\tilde{\mathbf{X}}$ has no components along the direction $\mathbf{v}_j$. Thus the latent variables $\mathbf{t}_i = \mathbf{X}\mathbf{w}_i$ are also orthogonal to $\mathbf{v}_j$. Therefore, by choosing λ in this manner, we can suppress one specific component of $\mathbf{Z}$.

Following subsection includes alternative formulation of regularized LVM method that are suggested for future work:

2.7.2 Hard-Soft-Constrained PLS (HSC-PLS)

The discussion in this paper has focused on the case where the auxiliary noise matrix $\mathbf{Z}$ is simultaneously measured along with the observations of $\mathbf{X}$ and $\mathbf{Y}$. There are many cases where such simultaneous, auxiliary noise measurements are not available, but additional information on the covariance structure of the background noise is available (possibly in addition to a measured $\mathbf{Z}$). For example, it is possible to collect the background noise when a system is running in idle mode (where the effects of $\mathbf{X}$ and/or $\mathbf{Y}$ are suppressed and the data recorded is mostly representative of the noise subspace and its covariance matrix can be calculated). We denote this covariance matrix of the noise by $\mathbf{\Lambda}$. It should be noted that $\mathbf{\Lambda}$ differs from $\mathbf{Z}$ even though it may contain components of Z in it,

it will have a different nature and is collected in a very different manner. A very relevant example is the presence of background brain activity in electroencephalography or functional MRI recording. This background noise can be measured in between the experiments while the brain is in resting state and its covariance matrix $\mathbf{\Lambda}$ estimated. Such a covariance matrix directly represents a major component of the background or the structured noise in $\mathbf{X}$ and $\mathbf{Y}$.

The reader should distinguish between the covariance matrix (Λ) computed from data recorded at a different time and the auxiliary structured noise ($\mathbf{Z}$) measured simultaneously with $\mathbf{X}$ and $\mathbf{Y}$. In addition the information content of $\mathbf{Z}$ and $\mathbf{\Lambda}$ is different in most cases. This additional background noise matrix can be added to equation (2-12) along with the hard constraint on $\mathbf{Z}$ to further suppress its components in addition to removing the effects of simultaneous noise collected in $\mathbf{Z}$ in the following manner:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_i - \lambda \mathbf{w}_i'\mathbf{\Lambda}\mathbf{w}_i \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{w}_j = \delta_{ij} \\ & \mathbf{w}'\mathbf{X}'\mathbf{Z} = \mathbf{0} \end{aligned} \tag{2-48}$$

λ controls the degree of influence by $\mathbf{\Lambda}$.

We name this method the HSC-PLS algorithm. In the MRI example the matrix $\mathbf{Z}$ contains the cardiac noise known as ballistocardiographic noise [5] whereas the matrix Λ contains information about the background brain activity such as waves known as alpha rhythms [19]. The information content of Λ is not the same as $\mathbf{Z}$. Therefore, each constraint suppresses a different type of effect and can be used

together when both are available.

2.7.3 Alternative formulation for HC-PLS method

An alternative form of PLS [13] modified, to include the constraints on both **X** and **Y**, is shown here. In this approach, we wish to find two sets of latent variables $\mathbf{w}_i$ and $\mathbf{c}_i, i = 1,\ldots,q$ so that $\mathbf{X}\mathbf{w}_i$ is maximally correlated with $\mathbf{Y}\mathbf{c}_i$, while rejecting the structured noise components. The objective function for this second formulation therefore becomes

$$\begin{aligned} & \max_{(\mathbf{w},\mathbf{c})} \mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{c}_i \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{w}_j = \mathbf{c}_i'\mathbf{c}_j = \delta_{ij} \\ & \mathbf{w}'\mathbf{X}'\mathbf{Z} = \mathbf{c}'\mathbf{Y}'\mathbf{Z} = \mathbf{0}. \end{aligned} \tag{2-49}$$

It is shown [20] that the solutions $\mathbf{w}, \mathbf{c}$ to the above objective function correspond to the principal right and left singular values respectively of $\mathbf{M}_{XZ}'\mathbf{X}'\mathbf{Y}\mathbf{M}_{YZ}$.

2.7.4 Hard Constrained Reduced Rank Regression (HC-RRR)

Reduced rank regression is defined as:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_i \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{X}'\mathbf{X}\mathbf{w}_j = \delta_{ij} \end{aligned} \tag{2-50}$$

The solution satisfying (2-50) is given as the q largest eigenvectors associated with the generalized eigenvalue problem.

$$\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} - \lambda\mathbf{X}'\mathbf{X}\mathbf{w} = \mathbf{0}. \tag{2-51}$$

In reduced rank regression the goal is to extract components that maximally correlate with $\mathbf{Y}$ and capture the maximum variance in $\mathbf{Y}$ and only $\mathbf{Y}$. This method is insensitive to the presence of noise in $\mathbf{X}$, however, in the presence of structured noise in both $\mathbf{X}$ and $\mathbf{Y}$ it can be extended to suppress the structured

noise in $\mathbf{Y}$ by adding the orthogonality constraint to (2-50) as follows:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{X}'\mathbf{X}\mathbf{w}_j = \delta_{ij} \\ & \mathbf{w}'\mathbf{X}'\mathbf{Z} = \mathbf{0}, \end{aligned} \tag{2-52}$$

for which the solution is given as the principal eigenvalues associated with the following generalized eigenvalue problem:

$$\mathbf{P}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} - \lambda\mathbf{X}'\mathbf{X}\mathbf{w} = \mathbf{0}, \tag{2-53}$$

Where

$$\mathbf{P} = \mathbf{I} - \mathbf{X}'\mathbf{Z}\left[\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\right]^{-1}\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{2-54}$$

which is the oblique projector onto the orthogonal complement subspace of $\mathcal{R}(\mathbf{X}'\mathbf{Z})$ in the metric $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

2.7.5 Alternate form of HSC-PLS

As mentioned in section 2.7.2, a hard constraint can be applied along with a soft constraint on the covariance of the background noise to account for both types of noise in a system. If we are not interested in obtaining orthonormal latent vectors, instead of applying the soft constraint, a hard constraint can be applied to the background noise by including the background noise in the following context:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_i \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{\Lambda}\mathbf{w}_j = \delta_{ij} \\ & \mathbf{w}'\mathbf{X}'\mathbf{Z} = \mathbf{0} \end{aligned} \tag{2-55}$$

which is similar to the hard-constrained reduced rank regression algorithm explained previously, with the latent vectors being chosen to be along the direction of the noise covariance matrix.

2.7.6 Hard-Constrained Canonical Correlation Regression (HC-CCR)

In the presence of structured noise and in the case of canonical correlation, we can extend the objective function for the ordinary canonical correlation to include an orthogonality constraint on the structured noise. That is, we wish to find a set $\mathbf{w}_i, \mathbf{c}_i$ of latent variables which satisfy the following conditions:

$$\begin{aligned} & \max_{(\mathbf{w}_i, \mathbf{c}_i)} \mathbf{w}_i' \mathbf{X}'\mathbf{Y}\mathbf{c}_i \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{X}'\mathbf{X}\mathbf{w}_j = \mathbf{c}_i'\mathbf{Y}'\mathbf{Y}\mathbf{c}_j = \delta_{ij} \\ & \mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{c}_j = \sigma_i \delta_{ij} \\ & \mathbf{w}'\mathbf{X}'\mathbf{Z} = \mathbf{c}'\mathbf{Y}'\mathbf{Z} = \mathbf{0} \end{aligned} \tag{2-56}$$

where σ_i are the canonical correlation coefficients. The constraint in the last line in the above objective function indicates that $\mathbf{w}, \mathbf{c}$ are in the nullspaces of $\mathbf{Z}'\mathbf{X}$ and $\mathbf{Z}'\mathbf{Y}$ respectively. Therefore if $\mathbf{H}, \mathbf{G}$ define a basis for $\mathcal{N}(\mathbf{Z}'\mathbf{X}), \mathcal{N}(\mathbf{Z}'\mathbf{Y})$ respectively, there exists a $\boldsymbol{\theta}, \boldsymbol{\gamma}$ for which $\mathbf{w} = \mathbf{H}\boldsymbol{\theta}$ and $\boldsymbol{\gamma} = \mathbf{G}\boldsymbol{\gamma}$. Therefore (2-56) can be replaced with [21]

$$\begin{aligned} & \max_{(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i)} \quad \boldsymbol{\theta}_i'\mathbf{H}'\mathbf{X}'\mathbf{Y}\mathbf{G}\boldsymbol{\gamma}_i \\ & \text{s.t.} \qquad \boldsymbol{\theta}_i'\mathbf{H}'\mathbf{X}'\mathbf{X}\mathbf{H}\boldsymbol{\theta}_j = \boldsymbol{\gamma}_i'\mathbf{G}'\mathbf{Y}'\mathbf{Y}\mathbf{G}\boldsymbol{\gamma}_j = \delta_{ij} \quad \boldsymbol{\theta}_i'\mathbf{H}'\mathbf{X}'\mathbf{Y}\mathbf{G}\boldsymbol{\gamma}_j = \sigma_i \delta_{ij} \end{aligned}$$

(2-57)

which is the ordinary canonical correlation specification on a new covariance matrix $\mathbf{H}'\mathbf{X}'\mathbf{Y}\mathbf{G}$. A suitable choice for $\mathbf{H}, \mathbf{G}$ is given by $\mathbf{M}_{\mathbf{XZ}}$ and $\mathbf{M}_{\mathbf{YZ}}$ respectively. By constructing the Lagrangian "$\mathbf{L}$" corresponding to this problem

and differentiating with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ respectively and setting the results to zero, we get

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial \boldsymbol{\theta}} &= \mathbf{M_{XZ}}'\mathbf{X}'\mathbf{Y}\mathbf{M_{YZ}}\boldsymbol{\gamma} - \rho_1 \mathbf{M_{XZ}}'\mathbf{X}'\mathbf{X}\mathbf{M_{XZ}}\boldsymbol{\theta} = \mathbf{0}, \\ \frac{\partial \mathbf{L}}{\partial \boldsymbol{\gamma}} &= \left[\mathbf{M_{XZ}}'\mathbf{X}'\mathbf{Y}\mathbf{M_{YZ}}\right]'\boldsymbol{\theta} - \rho_2 \mathbf{M_{YZ}}'\mathbf{Y}'\mathbf{X}\mathbf{M_{XZ}}\boldsymbol{\gamma} = \mathbf{0}, \end{aligned} \tag{2-58}$$

where ρ_1 and ρ_2 are Lagrange multipliers. Solving for $\boldsymbol{\theta}$ in the first equation and substituting into the second equation leads to the following generalized eigenvalue problem for the solution of $\boldsymbol{\gamma}$ [22]

$$\mathbf{A}'\mathbf{B}^{\dagger}\mathbf{A}\boldsymbol{\gamma} - \lambda_2 \mathbf{C}\boldsymbol{\gamma} = \mathbf{0} \tag{2-59}$$

where † represents pseudo inverse, and

$$\begin{aligned} \mathbf{A} &= \mathbf{M_{XZ}}'\mathbf{X}'\mathbf{Y}\mathbf{M_{YZ}} \\ \mathbf{B} &= \mathbf{M_{XZ}}'\mathbf{X}'\mathbf{X}\mathbf{M_{XZ}} \\ \mathbf{C} &= \mathbf{M_{YZ}}'\mathbf{Y}'\mathbf{Y}\mathbf{M_{YZ}} \end{aligned} \tag{2-60}$$

Using similar techniques, it is easily verified that the solution for $\boldsymbol{\theta}$ satisfies

$$\mathbf{A}\mathbf{C}^{\dagger}\mathbf{A}'\boldsymbol{\theta} - \lambda_1 \mathbf{B}\boldsymbol{\theta} = \mathbf{0}. \tag{2-61}$$

2.7.7 A few notes justifying the use of constrained LVM methods

We mentioned earlier that an alternative method for removing the structured noise is to project the input dataset (**X**) into the orthogonal complement of the noise matrix (**Z**) before performing a regression against **Y**. However, as we shall show in this section, this direct projection method is sensitive to the presence of

uncertainties in **Z** and can lead to biased estimates of the model whereas LVM methods are less sensitive to the presence of uncertainties.

This appendix offers some insights into the advantages of using OSC type signal correction as opposed to using direct projection of data into the orthogonal complement of the noise subspace. Here, we would like to show that methods such as orthogonal signal correction (OSC) or HC-PLS are less sensitive to the presence of noise in the auxiliary noise matrix (**Z**). This is a common situation as in many cases **Z** cannot be recorded without additive noise.

First we will show that in both methods the accuracy of the model depends on the quality of projectors into some orthogonal complement of a space. Later we will show that asymptotically in the first case, (direct projection method) the coefficients in the projector component are biased and will have less accuracy as the noise levels in **Z** increases. On the other hand we will show that asymptotically, in the constrained methods the projection coefficient matrix is unbiased and therefore, results in more accurate estimates of the projector and less sensitivity to the presence of noise in **Z.** Finally we will show some simulation results to back our hypothesis.

2.7.7.1 A typical direct projection vs. OSC

In direct projection method, prior to regressing $\mathbf{Y}$ against $\mathbf{X}$, either $\mathbf{X}$, $\mathbf{Y}$ or both $\mathbf{X}$ and $\mathbf{Y}$ are projected into the orthogonal complement of $\mathbf{Z}$. In other words we are interested in performing (PLS) regression between $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ where:

$$\tilde{\mathbf{X}} = \left(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\right)\mathbf{X} \tag{2-62}$$

$$\tilde{\mathbf{Y}} = \left(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\right)\mathbf{Y} \tag{2-63}$$

Once $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are calculated PLS regression can be performed between $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ (or $\tilde{\mathbf{Y}}$) as follows:

$$\begin{aligned} \max \quad & \mathbf{w}'\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = \mathbf{1}. \end{aligned} \tag{2-64}$$

Having:

$$\mathbf{Q}_\mathbf{z} = (\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \tag{2-65}$$

Or in other words $\mathbf{Q_z} = \mathbf{I} - \mathbf{P_z}$ where:

$$\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \tag{2-66}$$

And since $\mathbf{Q}_\mathbf{Z}\mathbf{Q}_\mathbf{Z} = \mathbf{Q}_\mathbf{Z}$ one can rewrite (2-64) as:

$$\begin{aligned} \max \quad & \mathbf{w}'\mathbf{X}'\mathbf{Q}_\mathbf{Z}'\mathbf{Y}\mathbf{Y}'\mathbf{Q}_\mathbf{Z}'\mathbf{X}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = \mathbf{1} \end{aligned} \tag{2-67}$$

Or:

$$\begin{aligned} \max_w \quad & \mathbf{w}'\tilde{\Sigma}\mathbf{w} \\ s.t. \quad & \mathbf{w}'\mathbf{w} = 1 \end{aligned} \tag{2-68}$$

Where $\tilde{\mathbf{\Sigma}}$ is defined by:

$$\Sigma = (\mathbf{Y}'(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{X})'(\mathbf{Y}'(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{X}) \tag{2-69}$$

In HC-PLS (and in general other OSC methods) we try to find principal

directions (components) that maximally explain a covariance structure $\mathbf{\Sigma}$ ($\mathbf{\Sigma}$ = $\mathbf{X'YY'X}$ in HC-PLS) constrained to be orthogonal to subspace of the structured noise ($\mathbf{Z}$).

In general the objective function can be written as:

$$\begin{array}{lll} \max_w & \mathbf{w}'\Sigma\mathbf{w} & (a) \\ s.t & \mathbf{w}'\mathbf{w} = \mathbf{1} & (b) \\ & \mathbf{w}'\mathbf{X}'\mathbf{Z} = 0 & (c) \end{array} \tag{2-70}$$

Equation (2-70)(c) can be written as:

$$\mathbf{Z}'\mathbf{X}\mathbf{w} = \mathbf{0} \tag{2-71}$$

which means $\mathbf{w}$ is in the nullspace of $\mathbf{Z'X}$:

$$\mathbf{w} \in \mathcal{N}(\mathbf{Z}'\mathbf{X}). \tag{2-72}$$

One subspace representing the $\mathcal{N}(\mathbf{Z}'\mathbf{X})$ is the orthogonal complement projector of $\mathbf{X}'\mathbf{Z}$ denoted by $\mathbf{Q}_{\mathbf{X}'\mathbf{Z}}$:

$$Q_{X'Z} = (\mathbf{I} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{X}\mathbf{X}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}). \tag{2-73}$$

Again, we can write $\mathbf{Q_{X'Z} = I - P_{X'Z}}$. where

$$\mathbf{P}_{\mathbf{X}'\mathbf{Z}} = \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{X}\mathbf{X}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \tag{2-74}$$

Assuming $\mathbf{w} = \mathbf{Q}_{\mathbf{X}'\mathbf{Z}}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a vector of appropriate length, we can rewrite (2-70) can reformatted as:

$$\begin{array}{lll} \max_\alpha & \boldsymbol{\alpha}'\mathbf{Q}'_{\mathbf{X}'\mathbf{Z}}\Sigma\mathbf{Q}_{\mathbf{X}'\mathbf{Z}}\boldsymbol{\alpha} & (a) \\ s.t. & \boldsymbol{\alpha}'\mathbf{Q}_{\mathbf{X}'\mathbf{Z}}\boldsymbol{\alpha} = 0 & (b) \end{array} \tag{2-75}$$

The solution to this maximization problem can be found by finding the dominant eigenvector of

$$\mathbf{Q}_{\mathbf{X'Z}}\Sigma \quad . \tag{2-76}$$

Where, with regards to the HC-PLS method:

$$\Sigma = \mathbf{X'YY'X} \quad . \tag{2-77}$$

Therefore in both equations (2-68) and (2-76), the accuracy of the model depends on proper estimation of the projectors $\mathbf{Q}_{\mathbf{Z}}$ and $\mathbf{Q}_{\mathbf{X'Z}}$ respectively. The quality of these projectors depend on how well $\mathbf{X}$ is projected into the range of $\mathbf{Z}$ (in normal projection method) or how well $\mathbf{X}$ is projected into the range of $\mathbf{X'Z}$ (in HC-PLS algorithm)

Hence, in ordinary projection method (direct projection onto $\mathbf{Q}_{\mathbf{Z}}$) the projected value of $\mathbf{X}$ into the range of $\mathbf{Z}$ ($\mathbf{P}_{\mathbf{Z}}$) can be obtained from:

$$\mathbf{Z}\hat{\beta}_o = \mathbf{X}, \tag{2-78}$$

where $\widehat{\boldsymbol{\beta}}_o$ are regression coefficients ("O" for ordinary). Therefore, it is essential to properly estimate the projection coefficient $\widehat{\boldsymbol{\beta}}_{\mathbf{o}}$. Similarly in the HC-PLS method the accuracy of the model depends on how well $\mathbf{X'Y}$ is projected into the orthogonal complement of $\mathbf{X'Z}$, which in turn, depends of the quality of the estimated projection coefficient $\widehat{\boldsymbol{\beta}}_{\mathbf{c}}$ (assuming **X, Y** and **Z** are mean centered):

$$(\mathbf{X'Z})\hat{\beta} = \mathbf{X'Y}. \tag{2-79}$$

Where again $\widehat{\boldsymbol{\beta}}_{\mathbf{c}}$ are regression coefficients ("C" refers to constrained).

2.7.7.2 ordinary projection method is biased:

In this section we show that the projection coefficients in $\mathbf{P}_{\mathbf{Z}}$ (ordinary

projection method) in the presence of error in **Z**, is biased, which can lead to less accurate results. To simplify the subject, let's assume a case where $\mathbf{x}$ ($m\times 1$) and $\mathbf{z}$ are uni-variate (they are $m\times 1$ vectors). In this example $\mathbf{x}$ denotes the input vector defined earlier. In a linear regression model we can write the regression of $\mathbf{x}$ into the range of $\mathbf{z}$ as:

$$\mathbf{x} = \mathbf{z}^{*}\beta + \epsilon \tag{2-80}$$

Where

$$\mathbb{E}(\mathbf{z}^{*}, \epsilon) = 0. \tag{2-81}$$

Assuming that **x** is error free but **z** is prone to errors we can write:

$$\mathbf{z} = \mathbf{z}^{*} + \boldsymbol{\eta} \tag{2-82}$$

where $\boldsymbol{\eta}$ ($m\times 1$) is a vector that contains the noise associated with $\mathbf{z}$ and $\mathbb{E}(.)$ is the expected value operator. Now, we can write (2-80) as:

$$\mathbf{x} = \mathbf{z}\beta + \mathbf{u} \tag{2-83}$$

where **u** ($m\times 1$) contains the error terms:

$$\mathbf{u} = \epsilon - \beta\boldsymbol{\eta}. \tag{2-84}$$

In the regression model (2-83) the coefficient $\hat{\beta}$ can be calculated from:

$$\hat{\beta} = \frac{\mathrm{Est}(\mathrm{cov}(\mathbf{z}, \mathbf{x}))}{\mathrm{Est}(\mathrm{var}(\mathbf{z}))} \tag{2-85}$$

Where Est(.) means the estimated value. Assuming that **z*** and **η** are independent, the covariance between **z** and **u** (the error term), is be given by:

$$\mathrm{cov}(\mathbf{z}, \mathbf{u}) = c\mathrm{ov}(\mathbf{z}^{*} + \boldsymbol{\eta}, \epsilon - \beta\boldsymbol{\eta}) = -\beta\,\mathrm{var}(\boldsymbol{\eta}). \tag{2-86}$$

The probability limit is defined as the expected value of an estimator as the number of observations reaches infinity. In other words, the probability limit for β as *n* reaches infinity can be calculated from:

$$\mathbb{E}(\hat{\beta}) = \beta + \frac{\text{cov}(\mathbf{z},\mathbf{u})}{\text{var}(\mathbf{z})} \tag{2-87}$$

Replacing (2-86) into (2-87) and multiplying by $\text{var}(\mathbf{z})$ we get

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \frac{\beta\text{var}(\mathbf{z}^* + \boldsymbol{\eta}) - \beta\text{var}(\boldsymbol{\eta})}{\text{var}(\mathbf{z}^* + \boldsymbol{\eta})} \\ &= \frac{\beta[\text{var}(\mathbf{z}^*) + \overbrace{2\,\text{cov}(\mathbf{z}^*, \boldsymbol{\eta})}^{0} + \text{var}(\boldsymbol{\eta}) - \text{var}(\boldsymbol{\eta})]}{\text{var}(\mathbf{z}^*) + 2\underbrace{\text{cov}(\mathbf{z}^*, \boldsymbol{\eta})}_{0} + \text{var}(\boldsymbol{\eta})}\end{aligned} \tag{2-88}$$

Since cov(**z*,η**)=0 above equation simplifies to:

$$\mathbb{E}(\hat{\beta}) = \beta\frac{\text{var}(\mathbf{z}^*)}{\text{var}(\mathbf{z}^*) + \text{var}(\boldsymbol{\eta})}. \tag{2-89}$$

We can see that as the noise level (**η**) increases in **z** the projection coefficient β becomes biased and hence the accuracy of the projector (into the range of **z** decreases).

The estimators in OSC algorithms are unbiased:

For the sake of simplicity, we now assume that **X** is a ($m \times n$) matrix and **Y** and **Z** are single mean-centered vectors.

In this case we assume:

$$\mathbf{X} = \mathbf{X}^* + \mathbf{E}. \tag{2-90}$$

Combining (2-90) and (2-82), now we shall have:

$$\mathbf{X}'\mathbf{z} = \mathbf{X}^{*\prime}\,\mathbf{z}^* + \mathbf{E}'\mathbf{z}^* + \mathbf{X}^{*\prime}\,\boldsymbol{\eta} + \mathbf{E}'\boldsymbol{\eta}, \tag{2-91}$$

and we can rewrite (2-79) as:

$$\mathbf{X}^{*}\prime\mathbf{y}+\mathbf{E}\prime\mathbf{y}=\mathbf{X}^{*}\prime\mathbf{z}*\beta+\mathbf{E}\prime\mathbf{z}*\beta+\mathbf{X}^{*}\prime\boldsymbol{\eta}\beta+\mathbf{E}\prime\boldsymbol{\eta}\beta. \tag{2-92}$$

In other words:

$$\mathbf{X}^{*}\prime\mathbf{y}=\mathbf{X}^{*}\prime\mathbf{z}^{*}\beta+\mathbf{u}, \tag{2-93}$$

where

$$\mathbf{u}=\mathbf{E}\prime\mathbf{z}^{*}\beta+\mathbf{X}^{*}\prime\boldsymbol{\eta}\beta+\mathbf{E}\prime\boldsymbol{\eta}\beta-\mathbf{E}\prime\mathbf{y}. \tag{2-94}$$

Since we have assumed that

$$\mathbb{E}(\mathbf{X},\mathbf{E})=\mathbb{E}(\mathbf{X},\boldsymbol{\eta})=\mathbb{E}(\mathbf{z},\mathbf{E})=\mathbb{E}(\mathbf{z},\boldsymbol{\eta})=0 \tag{2-95}$$

the bias term in (2-89) for the OSC or HC-PLS cases becomes:

$$\frac{\operatorname{cov}(\mathbf{X}\prime\mathbf{z},\mathbf{u})}{\operatorname{var}(\mathbf{X}\prime\mathbf{z})}=0 \tag{2-96}$$

Thus, the estimated covariance between **X'z** and **u** becomes zero, as the number of observations increases. Hence, the bias in the expected value of $\hat{\beta}$ will become zero, which means $\hat{\beta}$ in the HC-PLS method is an un-biased estimator of **β**. The reason that this fraction converges towards zeros is the fact that $\mathbb{E}(\mathbf{X},\boldsymbol{\eta})=0$ which will force the probability limit of the numerator to converge towards zero.

2.7.7.3 Simulation

The following figure shows simulation results for the quality of prediction when different levels of random noise are added to **Z.** The simulation is the same as that in section 2.4.2; with structured noise contaminating both **X** and **Y.** Three datasets **X, Y** and **Z** were generated using the same simulator program, and the

random noise levels in **Z** were changed in three steps; $\sigma_{\mathbf{Z}} = 0$, $\sigma_{\mathbf{Z}} = .5$ and $\sigma_{\mathbf{Z}} = 2$. In the first simulation $\sigma_{\mathbf{X}}$ and $\sigma_{\mathbf{Y}}$ were both equal to zero meaning no additive random noise was added to either **X** or **Y**

Following figure shows that when the random noise levels in **Z** are increased, the quality of prediction for $\mathbf{Y}^{\mathbf{0}}$ decreases drastically in the projection method, however they do not change as much in the HC-PLS method. These results support our hypothesis that the simple projector into orthogonal complement of **Z** is a biased estimated whereas the OSC method projection produces unbiased estimates, resulting in better models.

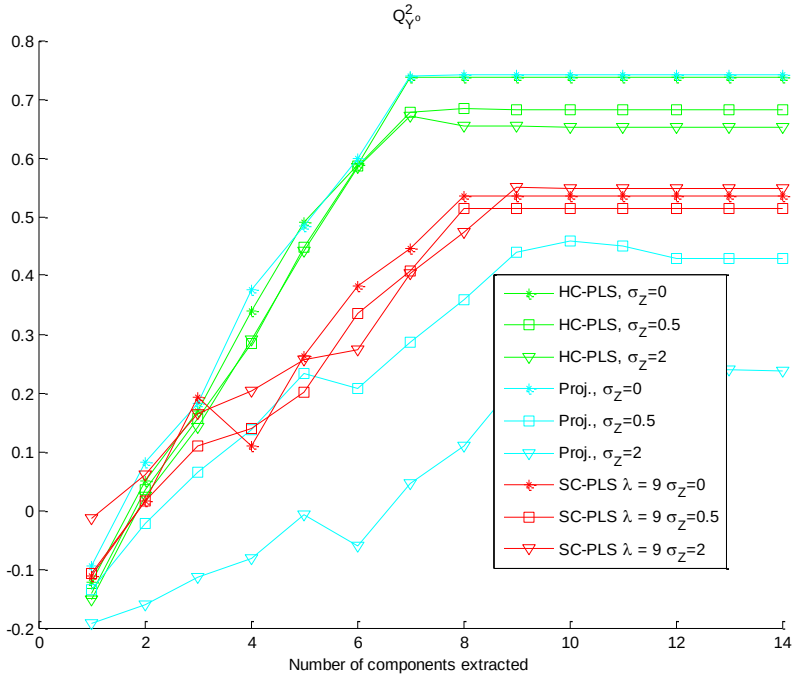


Figure 2-13: Quality of prediction ($Q^2_{\mathbf{Y}^{\mathbf{0}}}$) for future values of $\mathbf{Y}^0$. The results show that as the noise level increases in **Z**, the quality of prediction decreases in the projection model (Proj.) however the constrained methods (HC-PLS and SC-PLS) relatively keep the same level of prediction rate.

The next figure shows the simulation results when there is random noise in both

X and **Z**. We can see that even when there is substantial amount of noise in **X** σ_X = 1.5, the overall results for the HC-PLS method do not change very much.

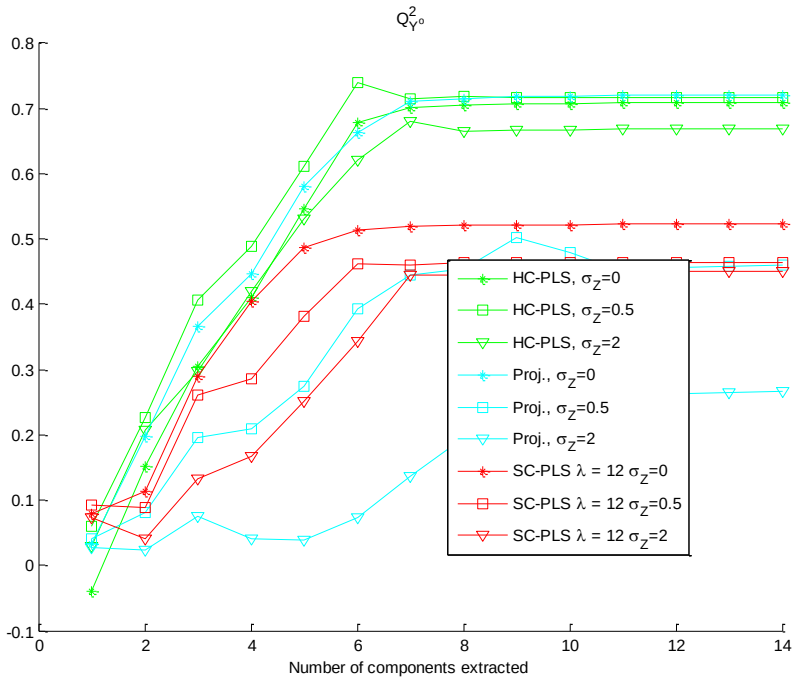


Figure 2-14: Quality of prediction ($Q^2_{\mathbf{Y}^0}$) for future values of $\mathbf{Y}^0$. In this second study both **X** and **Z** were contaminated with random noise. The results show that, same as in the first case, as the noise level in **Z** increases, the quality of prediction decreases in the projection model (Proj.) however the constrained methods (HC-PLS and SC-PLS) keep the same relatively level of prediction rate.

2.7.7.4 Conclusion:

It appears that using covariance projections, such as those used in orthogonal signal correction, will produce projection coefficients that are less sensitive to noise, as they are unbiased estimators. However in the case of normal projection, presence of noise in **Z** will result in attenuated and biased estimates in the data, resulting in lowered quality of fit and prediction rates.

REFERENCES

(1) Reilly, P. M.; Patino-Leal, H. A Bayesian Study of the Error-in-Variables Model. *Technometrics* **1981**, *23*, 221–231.

(2) Bruwer, M.-J.; MacGregor, J. F.; Bourg, W. M. Soft Sensor for Snack Food Textural Properties Using On-Line Vibrational Measurements. *Industrial & Engineering Chemistry Research* **2007**, *46*, 864–870.

(3) Christie, O. H. J. Data laundering by target rotation in chemistry-based oil exploration. *Journal of Chemometrics* **1996**, *10*, 453–461.

(4) Niazy, R. K.; Beckmann, C. F.; Iannetti, G. D.; Brady, J. M.; Smith, S. M. Removal of FMRI environment artifacts from EEG data using optimal basis sets. *NeuroImage* **2005**, *28*, 720–737.

(5) Croft, R. J.; Barry, R. J. Removal of ocular artifact from the EEG: a review. *Neurophysiologie Clinique/Clinical Neurophysiology* **2000**, *30*, 5–19.

(6) Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* **2002**, *16*, 119–128.

(7) Fearn, T. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems* **2000**, *50*, 47–52.

(8) Trygg, J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics* **2002**, *16*, 283–293.

(9) Trygg, J.; Wold, S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics* **2003**, *17*, 53–64.

(10) Salari Sharif, S. Regularized Latent Variable Methods in the Presence of Structured Noise and their Application in the Analysis Of ElectroEncephalogram Data, in Preparation. Ph.D. Thesis, McMaster University: Hamilton,Ontario,Canada.

(11) Wold, S.; Antti, H.; Lindgren, F.; Öhman, J. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems* **1998**, *44*, 175–185.

(12) Rao, C. R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics, Series A* **1964**, *26*, 329–358.

(13) Burnham, A. J.; Viveros, R.; MacGregor, J. F. Frameworks for latent variable multivariate regression. *Journal of chemometrics* **1996**, *10*, 31–45.

(14) de Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **1993**, *18*, 251–263.

(15) Yacoub, F.; MacGregor, J. F. Product optimization and control in the latent variable space of nonlinear PLS models. *Chemometrics and Intelligent Laboratory Systems* **2004**, *70*, 63–74.

(16) García-Muñoz, S.; Kourti, T.; MacGregor, J. F.; Apruzzese, F.; Champagne, M. Optimization of Batch Operating Policies. Part I. Handling Multiple Solutions#. *Industrial & Engineering Chemistry Research* **2006**, *45*, 7856–7866.

(17) García-Muñoz, S.; MacGregor, J. F.; Neogi, D.; Latshaw, B. E.; Mehta, S. Optimization of Batch Operating Policies. Part II. Incorporating Process Constraints and Industrial Applications. *Industrial & Engineering Chemistry Research* **2008**, *47*, 4202–4208.

(18) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal* **1994**, *40*, 826–838.

(19) Nunez, P. L.; Srinivasan, R. *Electric fields of the brain: the neurophysics of EEG*; Oxford University Press US, 2006.

(20) Golub, G. H. Some Modified Matrix Eigenvalue Problems. *SIAM Review* **1973**, *15*, 318–334.

(21) DeSarbo, W. S.; Hausman, R. E.; Lin, S.; Thompson, W. Constrained canonical correlation. *Psychometrika* **1982**, *47*, 489–516.

(22) Yanai, H.; Takane, Y. Canonical correlation analysis with linear constraints. *Linear Algebra and its Applications* **1992**, *176*, 75–89.

Chapter 3
An iterative NIPALS type algorithm for SC-PLS with ability to handle missing data

***Abstract*—An iterative algorithm, based on the NIPALS algorithm for the Soft Constrained PLS method is introduced. The advantage of the iterative algorithm is its ability to handle missing data during the model building process and also for prediction of the future observations.**

***Index Terms*—PLS, PCA, SC-PLS, constrained PLS, structured noise, NIPALS algorithm.**

3.1 INTRODUCTION

Latent variable methods (LVM) are used for regression between input variables (**X**) and response variables (**Y**) when the data is low rank and ill conditioned. The scores obtained from LVM methods can also be used for visual interpretation of the data. Another advantage of these methods is that they can handle missing points in a dataset. An example of the latent variable methods is the partial least squares PLS introduced by Wold et al. [1,2]. Wold introduced the NIPALS (Nonlinear Iterative Partial Least Squares) algorithm for iterative extraction of latent variable components in PLS. NIPALS is based on the power iteration used for extraction of eigenvectors in a matrix [3]. The advantage of NIPALS algorithm is its ability to handle datasets with large covariance matrices.

In a case where there are missing elements in the data matrix the ordinary PLS algorithm would work only if the rows of data containing the missing elements are removed prior to building the model. As well, they cannot handle any new

observations containing missing elements. During the model building process, if there are many observations, elimination of a few rows does not significantly degrade the model. However if the number of rows with missing elements, with respect to the size of **X** or **Y** is relatively high, the model will degrade. Even worse, if the missing points are scattered through many variables and observations, as opposed to be clustered in a few rows, many rows need to be deleted before building the model.

In this chapter an iterative algorithm for solving the soft constrained partial least squares (SC-PLS) [4] based on the NIPALS algorithm is introduced. This method is also capable of handling missing points during the model building process, and for the prediction of the future observations. In the following sections the PLS method and the NIPALS algorithm are reviewed. Then, one of the methods for recovering the missing points is discussed. Later it is shown how the NIPALS algorithm is modified to accommodate for SC-PLS case, enabling it to handle missing points. Finally some simulations are performed, showing how the algorithm behaves for various levels of missing elements in **X** , **Y** and **Z** matrices.

Notation: Bold upper (lower) case Arabic symbols represent matrices or vectors respectively, and regular-faced symbols are scalars. Bold-faced, upper-case Greek symbols are matrices. The notation, e.g. $\mathbf{x}_i$ represents the ith column of the matrix $\mathbf{X}$. The notation $\mathcal{N}(\cdot)$ and $\mathcal{R}(\cdot)$ denote the null-space and range respectively of the argument.

3.2 SC-PLS Iterative Algorithm

3.2.1 Partial Least Squares NIPALS algorithm

Partial Least Squares (PLS), is a latent variable regression method which is mostly suitable for building a latent model between **X** ($r \times k$) and **Y** ($r \times m$) when one or both are rank deficient. In PLS, a linear combination of **X** (**t = Xw**) and **Y** (**u = Yc**) are found, iteratively, that have maximum covariance with each other. This problem can be solved in the context of a maximization problem:

$$\begin{aligned} \max_{\mathbf{w},\mathbf{c}} \quad & \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{c} \\ s.t. \quad & \mathbf{w}'\mathbf{w} = \mathbf{1} \\ & \mathbf{c}'\mathbf{c} = \mathbf{1}. \end{aligned} \tag{3-1}$$

It can be shown that the solution to this problem ($\mathbf{w} \in \mathbb{R}^{k\times 1}, \mathbf{c} \in \mathbb{R}^{m\times 1}$) are the dominant eigenvalues of **X′YY′X** and **Y′XX′Y** respectively [5]. The eigenvectors can be found using the power iteration as:

$$\mathbf{w} \leftarrow \frac{\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}}{\left\|\mathbf{X}'\mathbf{Y}\mathbf{Y}\mathbf{X}\mathbf{w}\right\|} \tag{3-2}$$

and

$$\mathbf{c} \leftarrow \frac{\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{c}}{\left\|\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{c}\right\|}. \tag{3-3}$$

Where ||.|| is the two norm operator. The sequence is iterated until the difference between two subsequent values is less than a specified threshold. The associated eigenvalue, with **w**, can be obtained from:

$$\lambda = \frac{\mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}}{\mathbf{w}'\mathbf{w}}. \quad (3\text{-}4)$$

Wold et al. introduced the NIPALS algorithm, which is based on power iteration, for extraction of the principal components and the loadings in PLS algorithm. In the NIPALS algorithm the following steps are iterated until the iterations converge towards the largest eigenvectors of the respective matrices.

1. Initialize **u** as a column of **Y**
2. **w = X′ u**
3. **w ← w/||w||**
4. **t = Xw**
5. **c = Y′ t/(t′ t)**
6. **c ← c′ /||c||**
7. **u = Yc**
8. repeat steps 2 to 5 until convergence
9. **p = X′ t/(t′ t)**
10. **X ← X-tp′**
11. Go to step one and repeat for next principal component

The choice of normalization depends on the application and the algorithm. In many algorithms instead of **w** and **c**, **t** and **u** are normalized. However the end results at least in theory will be the same. The convergence criterion is usually calculated as the difference between two subsequent components and the iteration is stopped when it is less than a threshold value. Deflation of **X** at each step also ensures orthogonality between the subsequent principal components extracted.

Several methods have been developed for the handling of missing points through the PLS algorithm; such as *i*) Complete Object method [6] in which the rows that contain the missing elements are eliminated from the calculation and therefore only the data with no missing elements are included in the model building process, *ii*) the single component projection method used in the NIPALS

algorithm [7] which is a single pass method, Expectation maximization method [8] which is an iterative algorithm which can either start with a guessed value for the missing elements, or can start using the estimates of the NIPALS single component projection method and *iii*) Maximum Likelihood PCA method [9], are among such methods. The method implemented in our current algorithm is the expectation maximization (EM) method. In this algorithm the missing points are replaced with an initial guess value (for example mean value for each column) and the components are extracted using the NIPALS algorithm. Once all the principal components are extracted $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ are calculated as:

$$\hat{\mathbf{X}} = \mathbf{TP}' \tag{3-5}$$

and

$$\hat{\mathbf{Y}} = \mathbf{TC}'. \tag{3-6}$$

T, P and **C** are automatically provided by the NIPALS algorithm. The missing values in the original matrix are replaced with their estimated values from $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ and the algorithm is run again to obtain new values for the missing points that had been previously replaced by their estimates. The PLS algorithm is run again and a new estimate for the missing points is calculated. These new estimates are closer to the actual values of the missing points. This iteration is repeated until the change in the estimates of the missing values is less than a threshold level.

Once the algorithm has converged, the missing elements can be replaced using their final estimates from (3-5) and (3-6), and the remaining matrices, such s **W*** to be defined in (3-11), can be extracted at this point.

The NIPALS algorithm is specifically useful where the missing points are not clustered in a few rows or columns in the dataset [10]. A good rule of thumb is that the number of missing points should be less than 20% of the total number of elements in each dataset. However good results for larger number of missing points can be obtained when the dataset is large [10].

3.3 Soft Constrained PLS (SC-PLS).

In soft constrained PLS (SC-PLS), the PLS problem is reformulated to find a linear combination of **X** (**t = Xw**) that maximizes the difference between the covariance matrices **t′YY′t** and **t′ZZ′t** . As seen in chapter two, this problem can be formulated as a maximization problem:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \left|\mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} - \rho\mathbf{w}'\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{w}\right| \\ s.t. \quad & \mathbf{w}'\mathbf{w} = 1. \end{aligned} \tag{3-7}$$

The solution to this problem is found by finding the largest eigenvectors of the following matrix

$$\mathbf{H} = \mathbf{X}'(\mathbf{Y}\mathbf{Y}' - \rho\mathbf{Z}\mathbf{Z}')\mathbf{X}. \tag{3-8}$$

The latent vectors (**w**'s) are the eigenvectors associated with the eigenvalues λ_i of above equation (3-8) at each iteration. Subsequent eigenvectors are found by deflating the matrix **X** and performing the optimization using the new **X** value. The eigenvalues can become positive or negative depending on whether **H** is positive definite or negative definite. When the eigenvalue is positive the equation is equivalent to maximizing $\mathbf{t}'\mathbf{Y}\mathbf{Y}'\mathbf{t} - \rho_i\mathbf{t}'\mathbf{Z}\mathbf{Z}'\mathbf{t}$. However when $\lambda < 0$ then the problem is equivalent to maximizing $\rho_i\mathbf{t}'\mathbf{Z}\mathbf{Z}'\mathbf{t} - \mathbf{t}'\mathbf{Y}\mathbf{Y}'\mathbf{t}$. In the first case the

extracted components will be associated with **Y** while having low colinearity with **Z** and vice versa in the second case.

In order to derive the iterative algorithm for the soft constrained PLS, it can be rewritten as finding a linear combination of **X** (**t = Xw)** and **Y** (**u = Yc**) and **Z** (**f = Zk**) where:

$$\begin{aligned} \max_{\mathbf{w},\mathbf{c},\mathbf{k}} \quad & \left|\mathbf{w}'\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\mathbf{w} - \rho\mathbf{w}'\mathbf{X}'\mathbf{f}\mathbf{f}'\mathbf{X}\mathbf{w}\right| \\ & \mathbf{w}'\mathbf{w} = 1 \\ s.t. \quad & \mathbf{c}'\mathbf{c} = 1 \\ & \mathbf{k}'\mathbf{k} = 1. \end{aligned} \tag{3-9}$$

Compared to regular PLS algorithm, the objective function involved here is to find the difference of the two covariance matrices. Therefore the power method for finding **w** is used throughout the iteration steps. Hence the modified NIPALS algorithm will have the following form:

1. Initialize **u** as a column of **Y** and **f** as a column of **Z**, and **w** as a row of **X**
2. **w = (X′ uu′ X -** ρ **X′ ff′ X)w**
3. **w ← w/||w||**
4. **t = Xw**
5. **c = Y′ t/(t′ t)**
6. **c ← c /||c||**
7. **k= Z′ t/(t′ t)**
8. **k ← k /||k||**
9. **u = Yc**
10. **f = Zk**
11. λ = **w′ X′ (uu′** - ρ**ff′)Xw/(w′ w)**
12. repeat steps 2 to 10 until convergence
13. **p = X′ t/(t′ t)**
14. **X ← X- tp′ , Y←Y - tc′ , Z ← Z – tk′**
15. Go to step one and repeat for next principal component

The convergence is determined the same way as in the NIPALS algorithm for PLS. The second and third line of the iteration are adopted from the power

iteration in which "**w**" is updated sequentially until it converges towards the largest eigenvector of **H**. Once a convergence has been reached, **X** will be deflated as before (**X= X- tp′**) and a new component will be extracted by iterating through steps 1 to 11 using the new deflated **X,Y** and **Z**

As with the EM algorithm, when the missing points are present in the dataset, the missing points are initially replaced by an initial values (here, the mean value for each column) and then all the components are extracted sequentially. Once the components are extracted the estimates $\hat{\mathbf{X}}$,$\hat{\mathbf{Y}}$ and $\hat{\mathbf{Z}}$ are calculated using (3-5), (3-6) and (3-10)

$$\hat{\mathbf{Z}} = \mathbf{TK}' \tag{3-10}$$

and the missing points in the original matrix are replaced by these estimates. Again the SC-PLS is ran iteratively, for all the components, using the new **X,Y** and **Z** with the missing points replaced. Then the missing point values are updated. The iteration is repeated until convergence is achieved.

Following diagram shows the steps of the NIPALS SC-PLS algorithm incorporating the single component expectation maximizing idea.

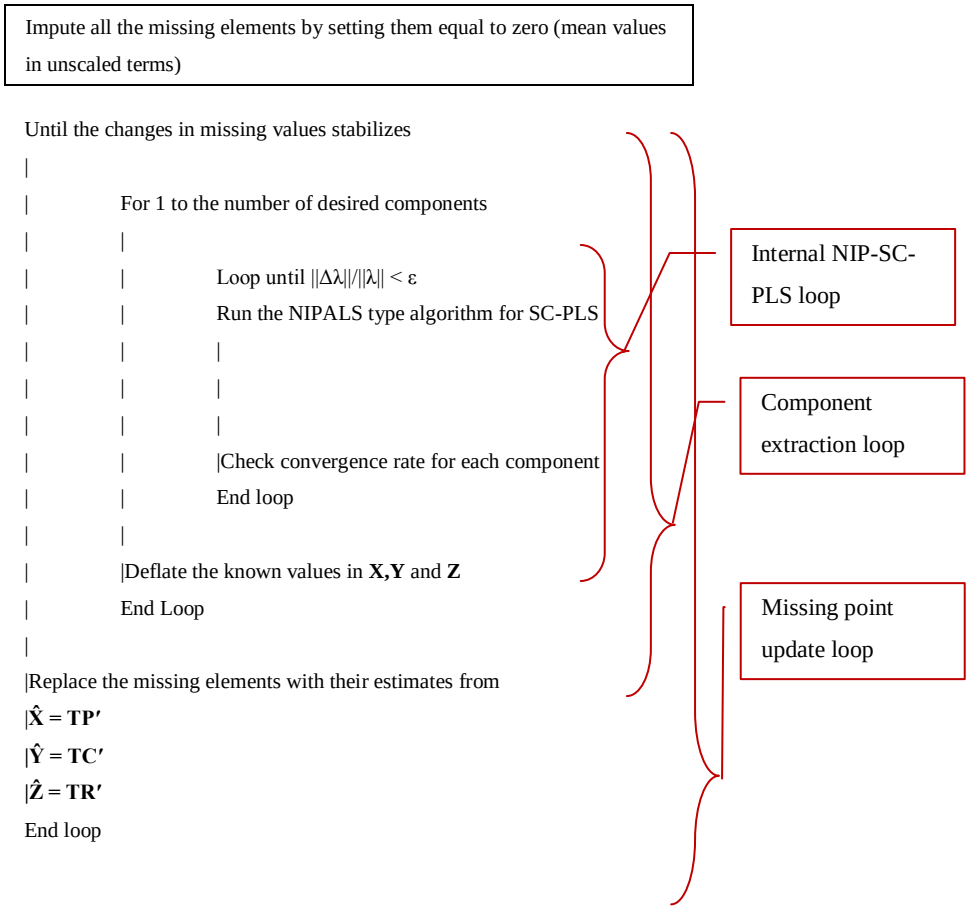


Figure 3-1: diagram of the different loops in the NIP-SC-PLS algorithm when there are missing elements in the datasets.

3.4 Future observations

Assuming that a model is already available (no missing data or the missing data problem has been resolved during the model building stage), the mixing components "**W**" are calculated in each step using deflated versions of **X**. Therefore, for conversion of the future observations a new mixing matrix that operates on the original **X** and obtains latent variables s the deflation method needs to be calculated. This new mixing matrix, called **W*** is calculated as [11]:

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}^{-1}\mathbf{W})^{-1}. \tag{3-11}$$

Once the **W*** is calculated, the score matrix $\mathbf{T}^{ts}$ for the new (future) observations is calculated from $\mathbf{X}^{ts}$ (test set):

$$\mathbf{T}^{ts} = \mathbf{X}^{ts}\mathbf{W}^*, \tag{3-12}$$

and using the projection coefficient "**C**" (the projection coefficient of **Y** onto the subspace of **T** during the model building process) it is possible to predict the previously unseen (future) values of $\mathbf{Y}^{ts}$.

$$\hat{\mathbf{Y}}^{ts} = \mathbf{T}^{ts}\mathbf{C}'. \tag{3-13}$$

When the model is obtained using the SC-PLS algorithms, some of its eigenvalues will be positive and some will have negative signs. It was mentioned earlier that the eigenvectors associated with positive components have high correlation with the **Y** variables whereas the components associated with negative eigenvalues have high correlation with the noise matrix (**Z**). Hence, when predicting the future values of **Y**, only the positive principal components should be used to predict the new **Y** variables:

$$\hat{\mathbf{Y}}_{new} = \mathbf{T}_+\mathbf{C}'_+ \tag{3-14}$$

where $\mathbf{T}_+$ is a matrix containing all the principal components associated with positive eigenvalues and $\mathbf{C}_+$ is the matrix containing their corresponding loading vectors.

In the event of missing points in the future observations (prediction set) a method known as the "future imputation" [12] is implemented. This method employs the following iterative steps to estimate the missing values for the new

observations.

1. Replace missing points in **X** with zeros (**W*** and **P** are already available from the model)
2. **T = XW***
3. $\hat{\mathbf{X}} = \mathbf{TP}'$
4. Replace missing points in **X** with the ones estimated in $\hat{\mathbf{X}}$
5. Repeat steps 2 and 3 until convergence is reached.

Convergence is reached when the changes in subsequent **T**'s are less than a threshold value. Once convergence is achieved, the new **T** values are used the same way as in (3-14) to predict the future response values (**Y**).

3.5 Simulation studies

The simulation dataset used for this study was exactly the same as the one used for the linear SC-PLS method (Chapter 2). For the examples in which datasets had missing points, the missing points for each of the matrices **X,Y** and **Z** were selected randomly using Matlab's random integer generator. The number of missing elements is chosen as a fraction of the total number of elements in each matrix. Once the random generator identified a set of random numbers, the corresponding element associated with that random element was be set to "NAN".

Throughout the study the goodness of fit for each dataset's results is measured as:

$$R^2_{\mathbf{\Phi}} = 1 - \frac{\sum\left(\sum\left(\phi_{ij} - \hat{\phi}_{ij}\right)^2\right)}{\sum\left(\sum\left(\phi_{ij}\right)^2\right)} \qquad (3\text{-}15)$$

where ϕ_{ij} is the element in the i^{th} column and the j^{th} row of the matrix $\mathbf{\Phi}$. For

our purposes $\mathbf{\Phi}$ can be either $\mathbf{X}$, $\mathbf{Y}$ or $\mathbf{Z}$. The quality of fit for NIP-SC-PLS, with various levels of missing points, for both cases: where there is noise only present in **X** and the case where structured noise is present in both **X** and **Y**, was compared.

For this simulation, once the datasets were generated, the observations in the datasets were divided in half. The first set was used for building the model, known as the "training set" (denoted by superscript "tr"), and the other set was used to measure the goodness of prediction, denoted by superscript "ts" and is called "test set". Each dataset contained 1000 observations. Quality of prediction is calculated as measure of the goodness of the model in predicting the future observation (i.e. *test set*) values. It can be calculated as:

$$Q^2_{\mathbf{Y}} = 1 - \frac{\sum\left(\sum\left(y^{ts}{}_{ij} - \hat{y}^{ts}_{ij}\right)^2\right)}{\sum\left(\sum\left(y^{ts}{}_{ij}\right)^2\right)} \tag{3-16}$$

where y^{ts}_{ij} is the test set's response variable and $\hat{y}^{ts}_{ij}$ is its corresponding predicted value obtained using (3-14). Overall the results of the SC-PLS and NIP-SC-PLS were identical whenever there were no missing points present. Therefore the remaining portion of this simulation section focuses on the properties of the NIP-SC-PLS algorithm in the presence of missing elements in **X**, **Y** and **Z** or multiple matrices

3.5.1 Convergence properties of the NIP-SC-PLS algorithm.

The NIPALS-SC-PLS algorithm is a stable algorithm; converging quickly for

most of the components. The simulations showed that in most cases the two-norm of the relative change of the obtained eigenvalue (from Equation (3-4)) reduced to less than 1E-7 after 30 to 40 iterations. In some cases minor instability was observed. However these instabilities usually are treated with a change in the initial guess and rerunning for that particular component.

As mentioned earlier, with the expectation maximization (EM) method, the SC-PLS (or PLS) algorithm is run with an initial guess replacing the missing elements in **X**, **Y** and **Z**. Once all the required components are extracted, the missing elements are replaced with their estimates from $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Z}}$, obtained from equations (3-5), (3-6) and (3-10) respectively. Once again the algorithm is run using the new updated **X**, **Y** and **Z** and a new estimate is obtained which is closer to the true estimate of the missing elements. This iteration is continued until a convergence is achieved.

Figure 3-2, (left) shows the changes in the first 5 eigenvalues of the extracted components during iteration steps mentioned above for the SC-PLS algorithm. Figure 3-2, (right) shows the quality of fit (R^2) for the missing elements in **X**, **Y** and **Z** as the iteration continues. It is observed that as the iteration advances, the quality of fit improves and eventually converges towards a certain value. Total percentage of missing elements (total number of elements) in **X**, **Y** and **Z** was 34, 22 and 27 percent respectively.

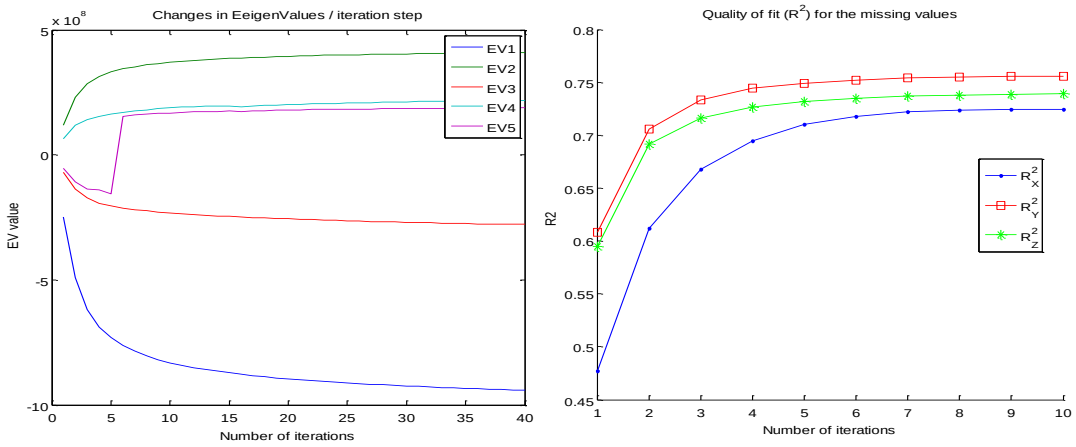


Figure 3-2: Left: changes in Eigenvalues (1 to 5) as the iteration proceeds. Right: changes in quality of fit (R^2) for the missing elements in **X,Y** and **Z** (5 principal components). Number of components extracted in total was 15. Total percentage of missing elements (to total number of elements) in **X**, **Y** and **Z** was: 34, 22 and 27 percent respectively.

This figure shows that as the iteration continues, some of the eigenvalues change sign, resulting in a jump in the eigenvalue and in turn a change in the value of the mixing matrix "**w**". The reason for such a behavior is that as eventually the missing elements get estimated, the matrix **H** in (3-8) changes from being positive definite to negative definite (or vice versa) and the components start converging using the new eigenvalues. However no radical behavior, such as instability, was noticed when the components changed sign within the iterations. This change in the eigenvalue sign usually appears in smaller eigenvalues as the model becomes more sensitive to noise. The results suggest that the convergence rates are almost exponential and usually stabilize after 10 to 15 iterations. The next figure shows the comparison between SVD based SC-PLS method (no missing points) and the SC-PLS-NIPALS algorithm with 25% missing points. It can be seen that the eigenvalues and the quality of prediction is in good

correspondence with the standard SC-PLS algorithm.

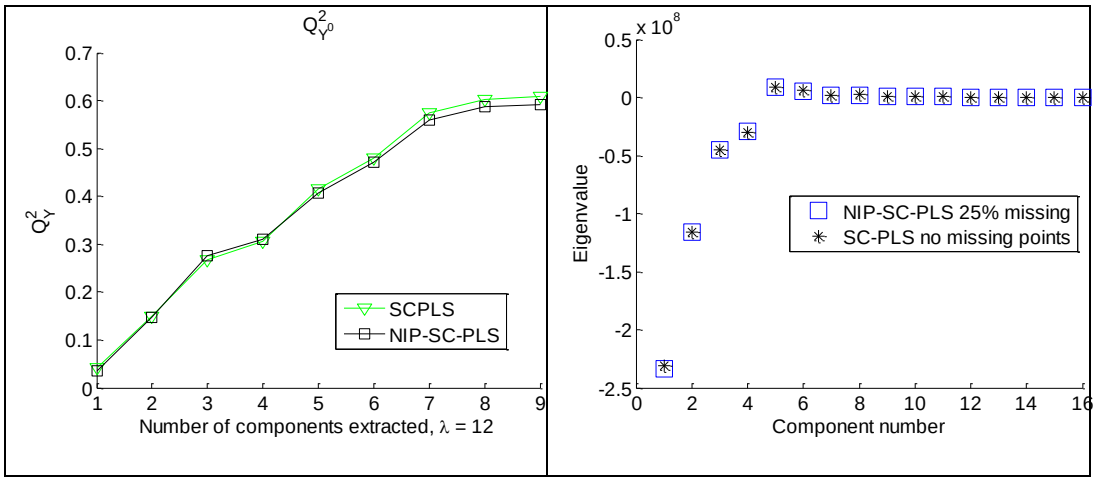


Figure 3-3: Left: Quality of prediction $Q^2_{\mathbf{Y}}$ for original SC-PLS (no missing values) and the NIP-SC-PLS with 25% missing data. Right: eivenvalues for the same simulation. Both plots show good agreement between the two methods even with 25% missing elements in the NIP-SC-PLS method. $\|\boldsymbol{C_X}\|_F = 40, \|\boldsymbol{C_Y}\|_F = 5, \|\boldsymbol{A_X}\|_F = 14, \|\boldsymbol{A_Y}\|_F = 10, \|\boldsymbol{B_X}\|_F = 6, \|\boldsymbol{B_Y}\|_F = 1,$

3.5.2 Effect of missing points in the quality of fit

The following plots in Figure 3-4 show how much of the missing elements are recovered during the model building process for various amounts of missing points. The algorithm was run with various levels of missing points in **X**, **Y**, **Z**. For all the runs the following parameters were chosen for the simulation datasets: $\|\mathbf{C_X}\|_F = 3, \|\mathbf{C_Y}\|_F = 0, \|\mathbf{B_X}\|_F = 0.1, \|\mathbf{B_Y}\|_F = 0.1$. The $\|.\|_F$ operator represents the Frobenius norm. $\|\mathbf{C_Y}\|_F = 0$ means that the noise is only present in **X**. In all simulations a total number of 15 components were extracted prior to measuring the quality of fit or prediction

The plots on the left show the quality of fit whereas the figure on the right shows the quality of prediction for new observations when the model was built

with a training set that contained missing points (same dataset as shown in plot of Figure 3-4-Left). It can be seen that the major determining component is the number of missing points in **X**. This of course is an expected result as the missing elements in **Y** and **Z** are also calculated using components (**T**) extracted from **X** from equations (3-5)(3-6) and (3-10).

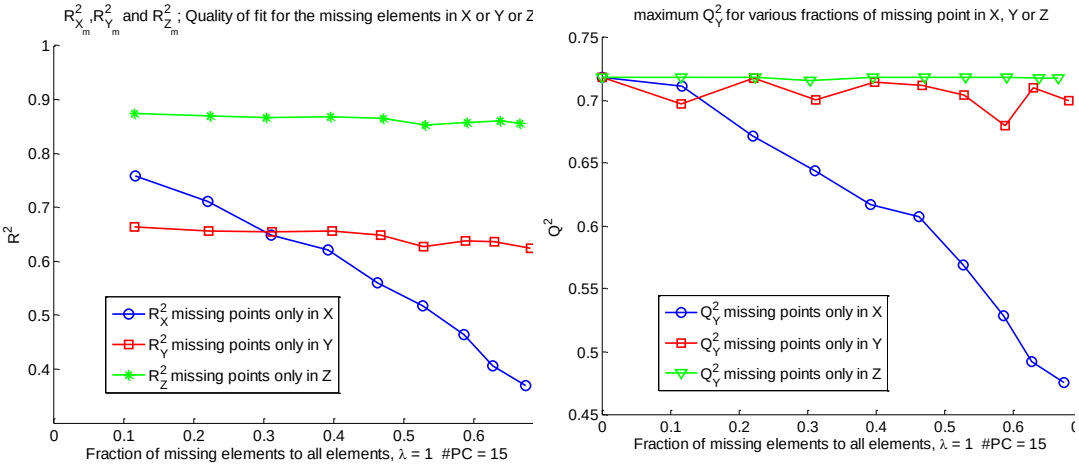


Figure 3-4: quality of fit using the SC-PLS algorithm for missing points in **X**,**Y** and **Z**. when the fraction of missing points (to all the points in the matrix) changes In each run a total of 15 components were extracted and λ was set to be equal to 1. Left figure shows the quality of fit to the missing elements (in **X**, **Y** or **Z**) while the right figure shows the quality of prediction for future observations (test set) when the model was built using a training set that contained missing elements (same dataset as left plots). There were no missing points in the future observations.

It should be noted that the quality of fit and prediction are also sensitive to the total number of components extracted. The following figure (Figure 3-5) shows the changes in the quality of fit for the missing values in **X**, for different simulation runs having a different number of components extracted. It is clear that if an insufficient number of components are extracted the prediction in the **X** variables can either under fit or over fit which will degrade the prediction results.

The strange behavior shown in Figure 3-5 is caused by the fact that only 10 latent variables were included in the original simulation model. Extracting more than 10 components from the model will result in over fitting and will ultimately add noise to the missing elements of **X**. However this added noise does not affect prediction of **Y** and **Z** as they are projections into the subspace of **X** and hence less sensitive to the error added into the missing elements in **X**. These results show that just like any other model, a proper number of components must be extracted. One of the methods that can be used in choosing the number of components is cross-validating the results on a test set to ensure proper number of components are extracted.

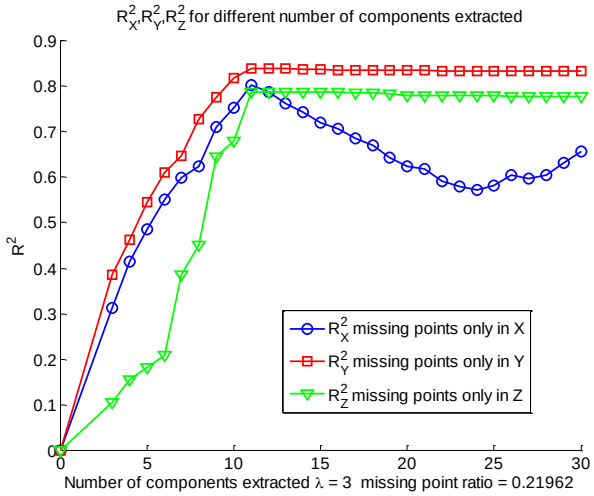


Figure 3-5: quality of prediction for future observations when the model was built using a training set that contained missing elements (corresponding to plots in Figure 3-4). If there are missing points in the training set, the prediction of the future observations can be affected accordingly.

3.6 Conclusion

In this chapter an iterative, NIPALS type algorithm (SC-PLS NIPALS) for

solving the SC-PLS problem was introduced. The advantage of the proposed algorithm compared to a direct component extraction approach is the same as the advantages of the ordinary NIPALS algorithm for selection of components compared to the methods based on direct SVD extraction.

The results of the NIPALS algorithm for SC-PLS are shown to be identical to the original SC-PLS algorithm when there are no missing data. The NIPALS algorithm (as well as the SVD based methods) can be altered to account for missing elements in the matrices. However, the advantage of NIPALS algorithm is its less computational cost when handling large datasets. The method used to recover the missing points in the algorithms presented here was the Expectation Maximization algorithm (EM). This method however is not the most advanced or statistically accurate method that is available. Other methods such as maximum likelihood method [9] have been shown to give better performance and therefore need to be investigated as alternatives to EM method. Nonetheless the EM method provides a simple and yet acceptable algorithm for recovering missing points during the model building process in NIP-SC-PLS algorithm. Overall the algorithm is stable, and in the case of large covariance matrices, can substantially reduce the computation cost compared to methods that use the whole matrix or extract all the components simultaneously (e.g. Singular Value Decomposition) .

REFERENCES

(1) Wold, S.; SjÃ strÃm, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109–130.

(2) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109–130.

(3) Saad, Y. *Numerical methods for large eigenvalue problems*; Manchester University Press ND, 1992.

(4) Salari Sharif, S.; Reilly, J. P.; MacGregor, J. Latent Variable Methods in the Presence of Structured Noise. *Journal of Chemometrics*.

(5) Burnham, A. J.; Viveros, R.; MacGregor, J. F. Frameworks for latent variable multivariate regression. *Journal of chemometrics* **1996**, *10*, 31–45.

(6) Nelson, P. The Treatment Of Missing Measurements In PCA And PLS Models. *Open Access Dissertations and Theses* **2002**.

(7) Wold, S. Estimation of Principal Components and Related Models by Iterative Least Squares. *Academic press* **1966**, 391–420.

(8) Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 2: Theory and algorithm. *Journal of Chemometrics* **1994**, *8*, 111–125.

(9) Andrews, D. T.; Wentzell, P. D. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer. *Analytica Chimica Acta* **1997**, *350*, 341–352.

(10) Nelson, P. R. C.; Taylor, P. A.; MacGregor, J. F. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems* **1996**, *35*, 45–65.

(11) Dayal, B. S.; MacGregor, J. F. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *Journal of Process Control* **1997**, *7*, 169–179.

(12) Arteaga, F.; Ferrer, A. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics* **2002**, *16*, 408–418.

Chapter 4
Constrained Nonlinear Latent Variable Methods

***Abstract*—Nonlinear Kernel methods have been widely used to deal with nonlinear problems in latent variable methods. However, in presence of structured noise these methods have reduced efficacy. We have previously introduced constrained latent variable methods that make use of any available additional knowledge about the structured noise. These methods improve performance by introducing additional constraints into the algorithm. In this paper we build upon our previous work and introduce hard and soft constrained nonlinear partial least squares methods using nonlinear kernels. The addition of nonlinear kernels reduces the effects of structured noise in nonlinear spaces and improves the regression performance between the input and the response variables.**

***Index Terms*— PCA, PLS, kernels, structured noise, latent variables**

4.1 INTRODUCTION

Latent variable methods such as principal component analysis (PCA) and partial least squares (PLS) [1] are very powerful techniques in de-noising datasets or performing regression in rank deficient environments. These methods tend to extract the major directions of variation within a dataset (PCA) or directions with most colinearity with respect to a response matrix (PLS). Despite their advantages in dealing with rank deficient datasets, these techniques also have certain shortcomings. One of the major problems with these techniques is their sensitivity to the presence of structured noise. Structured noise can be either temporally or contemporaneously correlated. Contemporaneous noise affects many variables of a dataset at the same time instant but does not necessarily have a structured

.

temporal power spectrum. Contemporaneous noise can also be referred to as unwanted systematic variations in a dataset. In this paper we take structured noise to be contemporaneous noise.

Large structured (systematic) variations affecting many variables of the input matrix(**X**) or both the input and response matrices (**X** and **Y**) can limit the extent to which latent variable methods (LVMs) can extract relevant latent components corresponding to the true underlying structure between **X** and **Y.** Regardless**,** orthogonal signal correction (OSC) and orthogonal projection to latent structures (O-PLS) methods [2,3,4,5] were introduced that prove effective under reasonable conditions. These methods remove variations in **X** (or **Y)** that are unrelated to **Y** (or **X**) before determining a model. As long as the structured noise resides in **X** or **Y**, these methods perform reasonably well. If there is structured noise in both **X** and **Y** but the systematic variation in each dataset is independent of the other one, these methods can still be applied with reasonable success [5,6] . However, in the presence of common structured noise , with the same (common) basis for structured variation in both **X** and **Y**, these methods fail. Take for example, OSC method by Fearn et al. [3] : In this method the covariance matrix (**X′X**) is projected into the orthogonal complement of **X′Y**. in a case when noise resides in both **X** and **Y, X′Y** will be a basis for the noise as well and hence the components extracted from **X** will be also orthogonal to noise subspace. In such a case, when the common noise is not properly removed from the datasets, the resulting model will be biased towards the structured noise. In other words, instead of building a

relationship between the input and the true underlying response, the resulting model may be better at predicting the variations in the noise rather than the desired **Y** values.

In the case where **X** and **Y** are related through some nonlinear relationship $\mathbf{Y} = \Phi(\mathbf{X})$, direct application of linear LVM methods or the orthogonal signal correction methods fail to provide optimal results. In order to overcome the nonlinearity problem, several methods such as principal curves [7], locally linear PCA methods [8], nonlinear NIPALS algorithm for PLS [9,10] and kernel PCA [11–13] have been introduced. The nonlinear kernel regression methods were further extended to kernel PLS methods [14–16;17]. Comprehensive reviews of linear and nonlinear PLS methods can be found in [18,19]. Kernel methods, which will be discussed further in this paper, use the so-called "kernel trick" to perform nonlinear expansion and regression in a potentially higher dimensional and nonlinear feature space without explicitly requiring direct access to **Φ**. In the presence of structured noise, especially when either **X** or both the **X** and **Y** spaces are contaminated, these methods suffer the same shortcomings as their linear counterparts.

We have previously addressed the problem of structured noise in the linear framework. In this methodology additional available information about the structured noise in the form of additional observations are imposed into the model to remove or reduce its effects [20] Several variations of constrained LVM, such as Hard-Constrained Principal Component Regression (HC-PCR), Hard-Constrained

Partial Least Squares (HC-PLS) and Soft-Constrained Partial Least Squares (SC-PLS) have been developed. These methods perform well when sufficient information about the noise subspace is available; however, they only perform well in the linear case. .

Nonlinear versions of OSC and OPLS [21,22,23] were introduced to overcome the undesired presence of nonlinearity in the underlying model. However, none of these methods exploit any additional information that might be available about the structured noise subspace and hence suffer the same shortcomings as their linear counterparts, especially in the presence of common structured noise residing in both **X** and **Y**.

In his article we expand our previously-developed linearly constrained LVM methodology to kernel-based nonlinear methods in the presence of structured noise. In this vein, we propose two LVM methods that can adapt to nonlinear models and are capable of exploiting any additional information available about structured noise that may be present in **X** and **Y.**

The remainder of this paper is organized as follows: Section II reviews the use of linear kernels in the context of the PLS algorithm, and then discusses the nonlinear kernel PLS algorithm. Section III reviews the previously-developed HC-PLS and SC-PLS algorithms, which extend the PLS concept to the case of structured noise. We then extend these PLS-based algorithms to the nonlinear case through the use of kernelization. We propose two algorithms for this purpose; the hard-constrained kernel PLS (HC-KPLS) and the soft-constrained

Kernel PLS (SC-KPLS) algorithms. Section IV explores the properties of these methods through simulations. Finally, in the Discussion and Conclusion section we address some of the advantages and issues associated with the proposed methods.

Notation: Bold upper (lower) case symbols represent matrices or vectors respectively, and regular-faced symbols are scalars. The notation, e.g. $\mathbf{x}_i$ represents the i^{th} column of the matrix $\mathbf{X}$ and $\underline{\mathbf{x}}_j$ represents the vector consisting of the j^{th} row of the matrix $\mathbf{X}$. The quantity $k(\underline{\mathbf{x}}_\mathbf{i}, \underline{\mathbf{y}}_\mathbf{j})$ represents a kernel operation between i^{th} and j^{th} rows of $\mathbf{X}$ and $\mathbf{Y}$ respectively. The notations $\mathcal{N}(\cdot)$ and $\mathcal{R}(\cdot)$ denote the nullspace and range the argument, respectively. In the following chapters the following naming scheme is used: methods that include hard constraints are accompanied by the prefix "HC" and those that include soft constrints include the prefix "SC". The prefix "K" denotes a nonlinear kernel version of the respective algorithm. Superscript $'$ denotes matrix transpose.

4.2 Kernel-latent variable methods

Linear Kernel methods were initially introduced in "linear" latent variable methods to reduce the calculation cost by performing feature extraction or regression directly from the covariance matrices rather than from the original variables [14,24,25]. To set the stage for a discussion on kernels, we first consider a simple partial least squares (PLS) algorithm:

4.2.1 The Ordinary PLS method

In partial least squares (PLS) the objective function is defined as finding a linear combination of the predictor matrix $\mathbf{X}$ ($n\times k$)

$$\mathbf{t} = \mathbf{Xw} \tag{4-1}$$

that maximizes the following objective function

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = 1 \end{aligned} \tag{4-2}$$

where $\mathbf{Y}$ ($n\times m$) contains the response values. The vector $\mathbf{w}$ ($k\times 1$) is called the loading vector and $\mathbf{t}$ ($n\times 1$) is called a *latent variable* (*latent component*) The solution to this problem can be found by differentiating the Lagrangian corresponding to (2) with respect to $\mathbf{w}$ and equating it to zero, to obtain

$$\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} - \rho\mathbf{w} = 0\,. \tag{4-3}$$

The $\mathbf{w}$ and ρ satisfying (4-3) are the dominant eigenvalue/eigenvector pair of $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$. The matrix $\mathbf{X}$ is then deflated according to

$$\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}\mathbf{p}' \tag{4-4}$$

where $\mathbf{p}$ ($k\times 1$), the loading vector, is found by projecting $\mathbf{X}$ into the range of $\mathbf{t}$:

$$\mathbf{p}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{X}\,. \tag{4-5}$$

The subsequent latent variables are found by reiterating through (4-1)-(5) . Once q principal components are calculated, $\mathbf{X}$ and $\mathbf{Y}$ can be written as

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}, \tag{4-6}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{C}' + \mathbf{F} \tag{4-7}$$

where $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_q]$ and $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_q]$ and the matrices $\mathbf{E}$ ($n\times k$) and $\mathbf{F}$ ($n\times m$)

are the remaining residuals after the decomposition. The projection coefficient $\mathbf{C}$ ($m \times q$) is calculated by projecting $\mathbf{Y}$ into the range of $\mathbf{T}$:

$$\mathbf{C}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}. \tag{4-8}$$

Typically in LVM analysis, a *training set* of data containing values of $\mathbf{X}$ and corresponding values of $\mathbf{Y}$ are assumed to be available. These are used in the training procedure to determine the matrices $\mathbf{P}$ and $\mathbf{C}$ in (4-6) and (4-7) respectively. Then in test or operational mode, a value $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ can be predicted from a new or previously unseen set (*the test set*) $\mathbf{X}^{ts}$ of $\mathbf{X}$ values according to

$$\hat{\mathbf{Y}} = \mathbf{T}^{ts}\mathbf{C}'. \tag{4-9}$$

In calculating the values $\hat{\mathbf{Y}}$, $\mathbf{T}^{ts}$ must be calculated directly from $\mathbf{X}^{ts}$. Since the $\mathbf{w}$'s were obtained from deflated $\mathbf{X}$ values, they cannot be used to calculate $\mathbf{T}^{ts}$ directly from $\mathbf{X}^{ts}$; however, it is possible to calculate a new matrix: $\mathbf{W^*} = \mathbf{WM}$ such that

$$\mathbf{T}^{ts} = \mathbf{X}^{ts}\mathbf{W}^*, \tag{4-10}$$

where $\mathbf{W^*}$ ($k \times q$) operates directly on $\mathbf{X}^{ts}$ (see Chapter Two) In the ordinary linear PLS method, the matrix $\mathbf{M}$ is calculated as:

$$\mathbf{M} = (\mathbf{T}'\mathbf{T}\mathbf{P}'\mathbf{W})^{-1}\mathbf{T}'\mathbf{T} \tag{4-11}$$

giving:

$$\mathbf{W}^* = \mathbf{W}(\mathbf{T}'\mathbf{T}\mathbf{P}'\mathbf{W})^{-1}\mathbf{T}'\mathbf{T}. \tag{4-12}$$

Predictions for the future observations can now be calculated directly from $\mathbf{X}^{ts}$

as:

$$\hat{\mathbf{Y}}^{ts} = \mathbf{X}^{ts}\mathbf{W}^{*}\mathbf{C}' . \tag{4-13}$$

4.2.2 Linear Kernels

Linear kernels were originally introduced to handle datasets with many variables and fewer observations in $\mathbf{X}$ [24]; i.e., $n<k$. In such datasets the generalized covariance matrix $\mathbf{X'\Sigma X}$ (where $\mathbf{\Sigma}$ is an arbitrary $n \times n$ positive semi-definite matrix) can become quite large and hence computationally expensive to calculate. These methods take advantage of certain properties of the loading vector $\mathbf{w}$ to reduce the size of the objective function's covariance matrix. It can be easily shown that when $\mathbf{T}=\mathbf{XW}$, $\mathbf{W}$ is within the range of $\mathbf{X'}$ [11],so that :

$$\mathbf{w} = \mathbf{X}'\boldsymbol{\alpha} . \tag{4-14}$$

Replacing $\mathbf{w}$ with $\mathbf{X}'\alpha$ in (4-2) results in:

$$\begin{aligned} \max_{\alpha} \quad & \boldsymbol{\alpha}'(\mathbf{XX}')\mathbf{YY}'(\mathbf{XX}')\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}'\mathbf{XX}'\boldsymbol{\alpha} = 1 \end{aligned} . \tag{4-15}$$

After differentiating the corresponding Lagrangian with respect to $\boldsymbol{\alpha}$ and equating it to zero, we have:

$$(\mathbf{XX}')\mathbf{YY}'(\mathbf{XX}')\boldsymbol{\alpha} - \lambda\mathbf{XX}'\boldsymbol{\alpha} = 0 \tag{4-16}$$

or

$$(\mathbf{K}')\mathbf{YY}'(\mathbf{K})\boldsymbol{\alpha} - \lambda\mathbf{K}\boldsymbol{\alpha} = 0 \tag{4-17}$$

where $\mathbf{K}$ ($n \times n$) equals

$$\mathbf{K} = \mathbf{XX}' . \tag{4-18}$$

The $\boldsymbol{\alpha}$ solving (4-15) is the eigenvector associated with the largest generalized eigenvalue of (4-17). Once α is calculated, **t** can be calculated directly from **K** as

$$\mathbf{t} = \mathbf{X}(\mathbf{X}'\boldsymbol{\alpha}) \tag{4-19}$$

or

$$\mathbf{t} = \mathbf{K}\boldsymbol{\alpha}. \tag{4-20}$$

The matrix **K** is called a kernel. The advantage of using kernels is two-fold. First, in this case where $n < k$, all matrix operations are performed in the smaller dimension (n) of **X**. The second is that once the kernel is constructed, there is no further need to access **X** directly and all calculations can be done through the kernel matrix. As in the ordinary PLS context, once each component is calculated, the kernel is deflated and the new kernel is used in (4-17) to extract the subsequent principal components. **K** is directly deflated as:

$$\mathbf{K} \leftarrow (\mathbf{X} - \mathbf{t}\mathbf{p}')(\mathbf{X} - \mathbf{t}\mathbf{p}')'. \tag{4-21}$$

Eq. (4-21) can be rewritten as:

$$\mathbf{K} \leftarrow \mathbf{K} - \mathbf{K}\frac{\mathbf{t}\mathbf{t}'}{\mathbf{t}'\mathbf{t}} - \frac{\mathbf{t}\mathbf{t}'}{\mathbf{t}'\mathbf{t}}\mathbf{K} + \frac{\mathbf{t}\mathbf{t}'\mathbf{K}\mathbf{t}\mathbf{t}'}{(\mathbf{t}'\mathbf{t})(\mathbf{t}'\mathbf{t})}. \tag{4-22}$$

This procedure of using kernels instead of **X** directly is known as the "kernel trick" . It plays a key role in the development of the nonlinear kernel methods, which are discussed next.

4.2.3 The Kernel Trick Applied to the Nonlinear PLS Problem

A linear LVM objective function which requires only inner product operations for its evaluation can be conveniently extended to the nonlinear case using nonlinear kernels. When the transformation $\Phi(\mathbf{X}): \mathrm{R}^{n\times k} \rightarrow \mathrm{R}^{n\times f}$ is known, where $\mathrm{R}^{n\times f}$ is referred to as the *feature space* (which is possibly of very high dimension ($f \gg n$) [11]), then a linear inference method such as regression or PLS can be used to build a model between $\mathbf{Y}$ and $\Phi(\mathbf{X})$, instead of between $\mathbf{Y}$ and $\mathbf{X}$ directly. The nonlinear version of the PLS problem may therefore be stated as:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\Phi(\mathbf{X})'\mathbf{Y}\mathbf{Y}'\Phi(\mathbf{X})\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = 1. \end{aligned} \tag{4-23}$$

The difficulty with this approach, even if $\Phi(\mathbf{X})$ is known, is the dimensionality of f. When $\Phi(\mathbf{X})$ involves larger polynomial orders, or when the dimension of $\mathbf{X}$ becomes large, the dimensionality of Φ grows very quickly and eventually the computation of $\Phi(\mathbf{X})$ becomes very expensive and impractical to calculate. This problem may be overcome using the kernel trick to implement the nonlinear transformation. Under certain conditions, the kernel trick permits implementation of the nonlinear transformation without the need to access $\Phi(\mathbf{X})$ directly. As indicated earlier, $\mathbf{w}$ can be written as:

$$\mathbf{w} = \Phi(\mathbf{X})'\boldsymbol{\alpha}. \tag{4-24}$$

By inserting (4-24) into (4-23):

$$\begin{aligned} \max_{\mathbf{w}} \quad & \boldsymbol{\alpha}'\Phi(\mathbf{X})\Phi(\mathbf{X})'\mathbf{Y}\mathbf{Y}'\Phi(\mathbf{X})\Phi(\mathbf{X})'\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}'\Phi(\mathbf{X})\Phi(\mathbf{X})'\boldsymbol{\alpha} = 1. \end{aligned} \tag{4-25}$$

And having $\mathbf{K} = \mathbf{\Phi(X)\Phi(X)'}$, the corresponding problem becomes:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \boldsymbol{\alpha}'\mathbf{K}'\mathbf{Y}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} = 1 \end{aligned} \quad . \tag{4-26}$$

The kernel trick allows us to implement the transformation $\Phi(\mathbf{X})$ indirectly in a very efficient manner, using only linear operations on **K**, in cases such as the current situation where the underlying algorithm is assumed to involves only inner product operations. A nonlinear transformation $\Phi(\mathbf{X})$ may be induced on the feature space by replacing an inner product in the feature space, such as $\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})'$, where $\boldsymbol{\phi}(\mathbf{X})$ is the i^{th} row of $\Phi(\mathbf{X})$. with a kernel function $k(\underline{\mathbf{x}}_i,\underline{\mathbf{x}}_j)$, which is defined later. In this case, there exists (see [11]) a nonlinear transformation $\Phi(\mathbf{x})$ such that $k(\underline{\mathbf{x}}_i, \underline{\mathbf{x}}_j) = \phi_i(\mathbf{X})\phi_j(\mathbf{X})'$ for any $\underline{\mathbf{x}}_i$, $\underline{\mathbf{x}}_j$, provided that the matrix **K,** whose elements are $k(\underline{\mathbf{x}}_i,\underline{\mathbf{x}}_j)$, is positive definite (Mercer's theorem) [26].

The ability to perform all computations directly from **K** allows us to impose nonlinearity in the model, using only linear operations on the kernel, in an effectively higher dimensional feature space, without having to access $\Phi(X)$ directly.

There are several well-known kernel functions that satisfy Mercer's theorem. Two such kernels are the polynomial kernel defined as:

$$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}'\mathbf{y} + \theta)^d \tag{4-27}$$

where θ and d are parameters to be determined, and the Gaussian kernel defined as:

$$k(\mathbf{x},\mathbf{y}) = exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{c^2}\right), \tag{4-28}$$

where c is also a parameter. Other kernels with different properties are also available [27]. See also [28].

Thus, a nonlinearity is induced on the feature space by assigning $\mathbf{K} = k_{ij}$, $(i,j) = 1,\ldots n$. The connection between the kernel function and the corresponding transformation $\Phi(X)$ is indirect and difficult to establish. Nevertheless, the choice of kernel function and the value of its parameters can be learned from the available training data (using, e.g., cross-validation methods) to best fit the underlying nonlinearity $\Phi(\mathbf{X})$. The use of kernel methods facilitates this process, due to the fact that the kernel is specified in terms of only a few parameters. In contrast, the determination of the nonlinear function $\Phi(\mathbf{X})$ can be very difficult due to the potential high dimensionality of Φ.

4.3 Regularized latent variable methods

Chapter 2 introduced the concept of regularized latent variable methods [20]. The advantage of these methods over previous LVM techniques is their ability to utilize auxiliary information that is sometimes available to improve component selection and regression in the presence of structured noise contaminating $\mathbf{X}$

and/or $\mathbf{Y}$. The Hard-Constrained Partial Least Squares (HC-PLS) method and its soft constrained counterpart (SC-PLS) are briefly reviewed here. Later, we extended these methods to the nonlinear case through the use of kernels.

In constrained LVM methods, in addition to the matrices $\mathbf{X}$ and $\mathbf{Y}$, it is assumed that an auxiliary noise matrix $\mathbf{Z}$ ($n \times b$) containing (partial) information about the noise is also available. The observations $\mathbf{Z}$ are assumed to be gathered simultaneously with the observations of $\mathbf{X}$ and $\mathbf{Y}$. It is further assumed that $\mathcal{R}(\mathbf{Z})$ intersects with the range of the structured noise components contaminating $\mathbf{X}$ and $\mathbf{Y}$. The information in $\mathbf{Z}$ can be used to suppress the effect of the structured noise when building a model between $\mathbf{X}$ and $\mathbf{Y}$, thereby improving the component selection and the prediction values. In the HC-PLS method, $\mathbf{Z}$ is used to formulate an additional constraint in the PLS objective function:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = 1 \end{aligned} \tag{4-29}$$

$$\mathbf{w}'\mathbf{X}'\mathbf{Z} = 0 . \tag{4-30}$$

Equation (4-30) restricts the extracted components to be orthogonal to the auxiliary noise matrix $\mathbf{Z}$, thereby suppressing the structured noise. The solution to the above optimization problem is obtained in an iterative fashion as before. First, $\mathbf{w}$ is determined as the dominant eigenvector of :

$$\mathbf{Q}_{\mathbf{XZ}}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X} \tag{4-31}$$

where:

$$\mathbf{Q}_{\mathbf{XZ}} = \mathbf{I} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{X}\mathbf{X}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}. \quad (4\text{-}32)$$

$\mathbf{Q}_{\mathbf{XZ}}$ ($k\times k$) is a projector into the orthogonal complement of $\mathbf{X}'\mathbf{Z}$. The details of calculating the eigenvalues, eigenvectors of the above non-symmetric matrix are given in [29] . Once the latent vector $\mathbf{w}$ is calculated, the corresponding principal component $\mathbf{t}$ is calculated as in (4-1). After each component is extracted, $\mathbf{X}$ is deflated and new components are extracted iteratively as shown in (4-4). The deflated $\mathbf{X}$ is used in equations (4-29)-(4-32) to extract the subsequent components. The projection vector $\mathbf{p}$ is calculated using (4-5).

Imposing hard constraints on the noise components can have various outcomes that depend on the condition and the size of the noise matrix. Rewriting equation (4-30) we get:

$$\mathbf{w}'\mathbf{X}'\mathbf{Z} = 0 \rightarrow \mathbf{\Lambda}'\mathbf{w} = 0, \quad (4\text{-}33)$$

where $\mathbf{\Lambda} = \mathbf{X}'\mathbf{Z}$. This means that $\mathbf{w}$ resides in the null space of $\mathbf{\Lambda}'$. Different outcomes are possible, depending on the structure of $\mathbf{\Lambda}$. If $\mathbf{\Lambda}$ is tall and full rank or rank deficient, $\mathbf{\Lambda}'$ has a non-empty nullspace. Therefore a vector $\mathbf{w}$ always exists that can satisfy (4-30). However, if $\mathbf{\Lambda}$ is full-rank square or short, then $\mathcal{N}(\mathbf{\Lambda}')$ will be empty. In this case, a $\mathbf{w}$ satisfying (4-30) does not exist.

In addition, if $\mathbf{Z}$ is not orthogonal to $\mathbf{X}$, imposing hard constrints may remove some of the variations in $\mathbf{X}$ that relate to $\mathbf{Y}$, reducing the efficacy of the results. Therefore when $\mathbf{\Lambda}$ is full-rank square, or short, or is not orthogonal to $\mathbf{X}$ it is

better to relax the orthogonality constrain and impose a soft constraint on the covariance of the auxiliary noise and the input data. We refer to this method as *soft constrained* PLS (SC-PLS). In SC-PLS the hard orthogonality constraint is replaced by a penalty on the square of the covariance between $\mathbf{X}$ and $\mathbf{Z}$ (i.e., $\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}$). The corresponding objective function is formulated as

$$\begin{aligned} \max_{\mathbf{w}} \quad & \left|\mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} - \lambda \mathbf{w}'\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{w}\right| \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = 1 \end{aligned}. \tag{4-34}$$

The $\mathbf{w}$ resulting from solving (4-34) is a vector that maximizes the absolute difference between the terms $\mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}$ and $\mathbf{w}'\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{w}$. The meta-parameter λ controls the relative weighting of the two terms.

To solve the above problem a Lagrangian operator is constructed, yielding

$$\max_{\mathbf{w}}(L) = \left| \mathbf{w}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X} - \lambda \mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X})\mathbf{w} \right| - \rho(\mathbf{w}'\mathbf{w} - 1). \tag{4-35}$$

The solution to this problem is found by differentiating the Lagrangian with respect to $\mathbf{w}$ and equating it to zero:

$$\operatorname{sgn}(\alpha)(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X} - \lambda \mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X})\mathbf{w} - \rho\mathbf{w} = 0 \tag{4-36}$$

where sgn(.) is the sign operator and α is the argument of the absolute value operator in (4-34). The solution to (4-34) is therefore the dominant eigenvector of the matrix $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X} - \lambda \mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}$.

The procedure to generate the latent variable model is as described previously. The components are extracted iteratively where $\mathbf{X}$ is deflated before each iteration. The matrix argument of the first line of (4-34) can be written as:

$$\mathbf{w}'\mathbf{X}'(\mathbf{Y}\mathbf{Y}' - \lambda \mathbf{Z}\mathbf{Z}')\mathbf{X}\mathbf{w} = \mathbf{t}'(\mathbf{Y}\mathbf{Y}' - \lambda \mathbf{Z}\mathbf{Z}')\mathbf{t} = \mathbf{t}'\mathbf{Y}\mathbf{Y}'\mathbf{t} - \lambda \mathbf{t}'\mathbf{Z}\mathbf{Z}'\mathbf{t}. \tag{4-37}$$

We first consider the case, in a particular iteration, when the dominant eigenvalue ρ of the matrix $\mathbf{X'YY'X} - \lambda\mathbf{X'ZZ'X}$ of (4-35) is positive. Then sgn(α) in (4-35) is +1, and the solution $\mathbf{w_0}$ to (4-34) is the corresponding eigenvector, and from the right-most equation of (4-37), we have $\mathbf{t_0'YY't_0} > \lambda\mathbf{t_0'ZZ't_0}$, where $\mathbf{t_0} = \mathbf{Xw_0}$. Thus the corresponding $\mathbf{t_0}$ is dominated by the term $\mathbf{t_0'YY't_0}$, and consequently $\mathbf{t_0}$ tends to be in a direction which explains maximum variation along **Y**. Since $\mathbf{t_0}$ itself is a linear combination of the columns of **X**, $\mathbf{t_0}$ in this case corresponds to a direction in **X** which is most closely aligned with **Y.** We can define the matrix $\mathbf{T^P}$ as that whose columns are the **t**'s corresponding to the positive values of ρ obtained over all iterations of the process.

Now, let us consider the case where the dominant eigenvalue ρ is negative. The solution $\mathbf{w_0}$ is again the eigenvector corresponding to the dominant eigenvalue. In this case;, sgn(α) in (4-35) is -1, leading to $\lambda\mathbf{t_0'ZZ't_0} > \mathbf{t_0'YY't_0}$. So now the second term of the right-most equation of (4-37) dominates. Using reasoning similar to that of the previous case, it can be seen that the $\mathbf{t_0}$ in this case corresponds to a direction in **X** which is most closely aligned with the noise matrix **Z**.

We define a matrix $\mathbf{T^N}$, in a manner similar to the way we have defined $\mathbf{T^P}$, whose columns consist of the $\mathbf{t_0}$ values associated with the negative eigenvalues obtained over the various iterations of the solution. Since $\mathbf{T^N}$ is associated with

the structured noise, no component of $\mathbf{T}^{\mathbf{N}}$ is used for the prediction of $\mathbf{Y}$. Thus, a predicted value $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ is obtained using only $\mathbf{T}^{\mathbf{P}}$ from (4-7) as

$$\hat{\mathbf{Y}} = \mathbf{T}^{\mathbf{P}}\mathbf{C}^{\mathbf{P}}, \tag{4-38}$$

where the rows of $\mathbf{C}^{\mathbf{P}}$ are a subset of those of $\mathbf{C}$ in (4-8)**,** corresponding to the columns in $\mathbf{T}^{\mathbf{P}}$. This use of latent vectors associated with the positive and negative eigenvalues in this SC-PLS method is illustrated further with examples in the application section.

4.3.1 Hard-Constrained Kernel PLS (HC-KPLS) method

Both the soft and hard constrained LVM methods can be modified to extract components using the "kernel trick". Assuming a nonlinear transformation of $\mathbf{X} \to \Phi(\mathbf{X})$, equations (4-29) and (4-30) can be rewritten as:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\Phi(\mathbf{X})'\mathbf{Y}\mathbf{Y}'\Phi(\mathbf{X})\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = 1 \end{aligned} \tag{4-39}$$

$$\mathbf{w}'\Phi(\mathbf{X})'\mathbf{Z} = 0. \tag{4-40}$$

We have shown previously that the loading vector $\mathbf{w}$ which solves above objective function is the eigenvector corresponding to the dominant eigenvalue of following equation:

$$\mathbf{Q}_{\Phi(\mathbf{X})\mathbf{Z}}\Phi(\mathbf{X})'\mathbf{Y}\mathbf{Y}'\Phi(\mathbf{X})\mathbf{w} - \lambda\mathbf{w} = 0. \tag{4-41}$$

Rewriting (4-41) using (4-24) we obtain:

$$\mathbf{K}'\mathbf{Y}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha} - \mathbf{K}'\mathbf{Z}(\mathbf{Z}'\mathbf{K}\mathbf{Z})^{\dagger}\mathbf{Z}'\mathbf{K}\mathbf{Y}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha} - \lambda\mathbf{K}\boldsymbol{\alpha} = 0. \tag{4-42}$$

Hence, α is calculated as the eigenvector associated with largest generalized eigenvalue of (4-42). The "†" symbol represents any suitable inverse. The solution for finding the eigenvectors of this problem is given in [29]

Once α is calculated, the corresponding latent vectors are calculated as described by equations (4-19) and (4-20). The latent vectors and the corresponding α vectors are extracted iteratively and $\mathbf{K}$ is deflated before each iteration as described in section 4.2. Once the latent vectors are extracted, the prediction coefficient $\mathbf{C}$ for $\mathbf{Y}$ is obtained by projecting $\mathbf{Y}$ into the range of $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_q]$ using (4-8).

We now discuss the prediction of values $\hat{\mathbf{Y}}$ with the kernel method. Nominally, the latent variables could be extracted directly from the original kernel ($\mathbf{K}_0$), in a manner similar to that shown in (11) – (13):

$$\mathbf{T} = \Phi(\mathbf{X})\mathbf{W}^* \rightarrow \mathbf{T} = \Phi(\mathbf{X})\mathbf{W}\mathbf{M}. \tag{4-43}$$

However, with the kernel method, there is no direct access to $\mathbf{W}$ and hence we must derive the latent variables using only the available matrix $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_q]$. To this end, we formulate a matrix $\mathbf{A}^*$ (analogous to $\mathbf{W}^*$) so that the latent vectors can be extracted using the original (undeflated) $\mathbf{K}_0$. We assign

$$\mathbf{T} = \mathbf{K}_0\mathbf{A}^*, \tag{4-44}$$

or in other words:

$$\mathbf{T} = \mathbf{K_0 A M}. \tag{4-45}$$

After multiplying both sides by $\mathbf{T}'$ and re-arranging, we have

$$\mathbf{M} = (\mathbf{T}'\mathbf{K_0}\mathbf{A})^{-1}\mathbf{T}'\mathbf{T} \rightarrow \mathbf{A}^* = \mathbf{A}(\mathbf{T}'\mathbf{K_0}\mathbf{A})^{-1}\mathbf{T}'\mathbf{T}. \tag{4-46}$$

Predicted values $\hat{\mathbf{Y}}$ can now predicted from previously unseen values $\mathbf{X^{ts}}$ of $\mathbf{X}$ as

$$\hat{\mathbf{Y}} = \mathbf{T}^{ts}\mathbf{C}'. \tag{4-47}$$

$\mathbf{T}^{ts}$ can be obtained directly from the original $\mathbf{X}$ ($\boldsymbol{X}^{tr}$) and $\mathbf{X}^{ts}$ as follows

$$\mathbf{T}^{ts} = \Phi(\mathbf{X}^{ts})\Phi(\mathbf{X})'\mathbf{A}^* \rightarrow \mathbf{T}^{ts} = \mathbf{K}^{ts}\mathbf{A}^* \tag{4-48}$$

where the $(i,j)th$ element of $\mathbf{K}^{ts}$ is defined as

$$\mathbf{K}^{ts}{}_{(ij)} = k(\underline{\mathbf{x}}_i^{ts}, \underline{\mathbf{x}}_j) \tag{4-49}$$

which is a kernel matrix constructed from the rows of the previously unseen observations and the rows of $\mathbf{X}^{tr}$,used in building the model prior to deflation. The matrix $\boldsymbol{C}$ in (4-47) is calculated as described previously from (4-8).

4.3.2 Soft Constrained KPLS (SC-KPLS)

As in the linear case, in the soft constrained kernel PLS (SC-KPLS) method, the orthogonality condition is replaced by a soft constraint, leading to the following equation:

$$\max_{\alpha} \quad \left|\boldsymbol{\alpha}'\mathbf{K}'\mathbf{Y}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha} - \lambda\boldsymbol{\alpha}'\mathbf{K}'\mathbf{Z}\mathbf{Z}'\mathbf{K}\boldsymbol{\alpha}\right| \quad \text{s.t.} \quad \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} = 1. \tag{4-50}$$

The rigidity of the penalty term is again controlled by changing the value of λ. We call this method "Soft Constrained Kernel PLS" (SC-KPLS). In (4-50), the quantities $\boldsymbol{\alpha}$ and ρ correspond to the largest generalized eigenvector, eigenvalue pair of the following equation:

$$(\mathbf{K}'\mathbf{Y}\mathbf{Y}'\mathbf{K} - \lambda\mathbf{K}'\mathbf{Z}\mathbf{Z}'\mathbf{K})\boldsymbol{\alpha} - \rho\mathbf{K}\boldsymbol{\alpha} = 0. \tag{4-51}$$

In a manner analogous to the linear case, it is assumed that $\mathbf{t} = \Phi(\mathbf{X})\mathbf{w}$ and the components are extracted iteratively by deflating the kernel matrix as described by (4-22). Again, the dominant eigenvalue of (4-51) can be either positive or negative, and using the same arguments as for the linear case, a predicted value $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ is obtained using only the latent vectors associated with the positive eigenvalues, using (4-38) as before.

$$\hat{\mathbf{Y}} = \mathbf{T}^{\mathbf{P}}\mathbf{C}^{\mathbf{P}} \tag{4-52}$$

where $\mathbf{T}^{\mathbf{P}}$ is the matrix consisting of only those $\mathbf{t}$'s that are associated with positive eigenvalues and $\mathbf{C}^{\mathbf{P}}$ consists of those rows of $\mathbf{C}$ that correspond to the $\mathbf{t}$'s in $\mathbf{T}^{\mathbf{P}}$. $\mathbf{T}$ for new prediction set is calculated from (4-44)

4.3.3 Kernels for noise

To provide extra flexibility in the underlying modeling process, it is also possible to kernelize the noise matrix $\mathbf{Z}$, when it appears in the form $\mathbf{ZZ}'$, e.g., as

in (4-34). In this case, the SC-KPS objective function becomes

$$(\mathbf{K}'\mathbf{Y}\mathbf{Y}'\mathbf{K} - \lambda\mathbf{K}'\mathbf{K}_{\mathbf{Z}}\mathbf{K})\boldsymbol{\alpha} - \rho\mathbf{K}\boldsymbol{\alpha} = 0 \tag{4-53}$$

where

$$\widehat{\mathbf{K}}_{\mathbf{Z}} = \widehat{\Phi}(\mathbf{Z})\widehat{\Phi}(\mathbf{Z})'. \tag{4-54}$$

In general, when modeling nonlinear relationships between **Y** and **X** or **Y** and **X** ,**Z**, it is not necessary to kernelize the **Y**-variables.

4.4 Experiments

In this section, we construct toy problems to compare the performances of the KPLS, HC-KPLS and SC-KPLS methods against their linear counterparts, which are PLS, SC-PLS and HC-PLS respectively.

In this toy problem it is assumed that 3 datasets; $\mathbf{X} \in \mathbb{R}^{n\times k}$, $\mathbf{Y} \in \mathbb{R}^{n\times m}$ and $\mathbf{Z} \in \mathbb{R}^{n\times b}$ are available where **X** and **Y** contain the measured values of the input and the response variables, respectively, and **Z** represents the dataset containing the additional auxiliary information about the structured noise contaminating **X** and/or **Y**. It is assumed that the relationship between **X**, **Y** and **Z** is nonlinear. The structure of the noise is discussed later. The purpose of the simulation is to compare performances of various LVM algorithms in extracting the principal components and in providing better relationship between the true and nonlinear underlying structure of **X** and **Y** while suppressing the noise effect.

Two separate cases are considered: the first is where strong structured noise contaminates only $\mathbf{X}$, while in the second case the structured noise contaminates

both $\mathbf{X}$ and $\mathbf{Y}$. The aim of the first study is to show how the inclusion of the auxiliary noise information can improve component selection by suppressing the covariance of $\mathbf{t}$ with the noise subspace. The second problem shows that the constrained methods perform better at revealing the true underlying structure between $\mathbf{X}$ and $\mathbf{Y}$, and also reduces error by suppressing the choice of latent variables which model the common noise between the input and the response variables.

In latent variable models the relationship between input and response variables is assumed to be non-causal, meaning that **X**, **Y** and, in this toy problem, **Z** are related to each other through a set of underlying, low rank, latent variables that are denoted by $\mathbf{T_S} \in \mathbb{R}^{n \times a}$ and $\mathbf{T_N} \in \mathbb{R}^{n \times s}$, for the signal and the noise subspaces respectively. The quantities **X**, **Y** and **Z** are nonlinear functions of $\mathbf{T_N}$ and $\mathbf{T_S}$.

As in chapter two, it is also assumed here that the measured values of **X** and **Y** consist of the true, uncontaminated, input and response values, $\mathbf{X_0} \in \mathbb{R}^{n \times k}$ and $\mathbf{Y_0} \in \mathbb{R}^{n \times m}$, plus the additive noise, $\mathbf{N_Y} \in \mathbb{R}^{n \times m}$ and $\mathbf{N_X} \in \mathbb{R}^{n \times k}$.

$$\mathbf{Y} = \mathbf{Y_0} + \mathbf{N_Y}, \tag{4-55}$$

$$\mathbf{X} = \mathbf{X}_0 + \mathbf{N_X}. \tag{4-56}$$

Without the loss of generality, it is assumed that $n \geq k \geq m$. The matrices $\mathbf{N_Y}$ and $\mathbf{N_X}$ contain the noise terms, which consists of both the structured plus random unstructured noise components. The structured noise components, $\mathbf{Z_Y} \in \mathbb{R}^{n \times m}$ and

$\mathbf{Z}_{\mathbf{X}} \in \mathbb{R}^{n \times k}$ in both $\mathbf{X}$ and $\mathbf{Y}$, are assumed to be linear mixtures of nonlinear functions of $\mathbf{T}_{\mathbf{N}}$.

Therefore:

$$\mathbf{N}_{Y} = \mathbf{Z}_{\mathbf{Y}} + \sigma_{\mathbf{Y}}\mathbf{E}_{\mathbf{Y}}; \qquad \mathbf{Z}_{\mathbf{Y}} \triangleq f_{\mathbf{Y}}(\mathbf{T}_{\mathbf{N}})\mathbf{J}_{\mathbf{Y}}, \tag{4-57}$$

$$\mathbf{N}_{\mathbf{X}} = \mathbf{Z}_{\mathbf{X}} + \sigma_{\mathbf{X}}\mathbf{E}_{\mathbf{X}}; \qquad \mathbf{Z}_{\mathbf{X}} \triangleq f_{\mathbf{X}}(\mathbf{T}_{\mathbf{N}})\mathbf{J}_{\mathbf{X}}. \tag{4-58}$$

In this toy problem it is assumed that the columns of the matrix $\mathbf{T}_{\mathbf{N}}$ are orthonormal vectors which describe the structured noise subspace with $s < n$. The elements of the matrices $\mathbf{E}_{\mathbf{Y}} \in \mathbb{R}^{n \times m}$ and $\mathbf{E}_{\mathbf{X}} \in \mathbb{R}^{n \times k}$ are *iid* random variables with unit variance that represent the unstructured noise components in $\mathbf{X}$ and $\mathbf{Y}$ respectively, scaled by values $\sigma_{\mathbf{X}}$ and $\sigma_{\mathbf{Y}}$ which set the variance levels for random noise components in $\mathbf{X}$ and $\mathbf{Y}$. The nonlinear functions $f_{\mathbf{X}}(.)$ and $f_{\mathbf{Y}}(.)$ map the latent variable $\mathbf{T}_{\mathbf{N}}$ into matrices of the same size. In addition to measurements of $\mathbf{X}$ and $\mathbf{Y}$, it is assumed a matrix $\mathbf{Z}$ containing measurements on the structured noise is also available. The matrix $\mathbf{Z}$ has an analogous structure to that of $\mathbf{X}$ and $\mathbf{Y}$, as follows:

$$\mathbf{Z} = f_{\mathbf{Z}}(\mathbf{T}_{\mathbf{N}})\mathbf{J}_{\mathbf{Z}} + \sigma_{\mathbf{Z}}\mathbf{E}_{\mathbf{Z}} \tag{4-59}$$

where $f_{\mathbf{Z}}(\mathbf{T}_{\mathbf{N}})$ scribes a nonlinear transformation of the structured components

of the noise. This means that some elements of the measured auxiliary noise matrix are nonlinearly related to the underlying structured noise affecting $\mathbf{X_0}$ and $\mathbf{Y_0}$. $\mathbf{J_Z} \in \mathbb{R}^{s \times b}$ is a mixing matrix constructed from *iid* (in dependant and identically distributed) random elements. $\mathbf{E_Z} \in \mathbb{R}^{n \times b}$ is an uncorrelated additive noise that is assumed to have contaminated the measurements of the auxiliary noise matrix. In typical applications of interest, the matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are low rank.

A statistical relationship exists between $\mathbf{X_0}$ and $\mathbf{Y_0}$ due to the common signal subspace latent variable $\mathbf{T_S}$ between them:

$$\mathbf{X_0} = g_{\mathbf{X}}(\mathbf{T_s})\mathbf{A_x} + \mathbf{U_X}\mathbf{B_X}, \tag{4-60}$$

$$\mathbf{Y_0} = g_{\mathbf{Y}}(\mathbf{T_s})\mathbf{A_y} + \mathbf{U_Y}\mathbf{B_Y}, \tag{4-61}$$

where it is seen these components are statistically dependant due to the common signal subspace latent variable $\mathbf{T_S}$ between them. The nonlinear terms $g_{\mathbf{X}}(\mathbf{T_S})$ and $g_{\mathbf{Y}}(\mathbf{T_S})$ again map the matrix argument $\mathbf{T_S}$ into one of the same size. The additional components $\mathbf{U_X} \in \mathbb{R}^{n \times v}$ and $\mathbf{U_Y} \in \mathbb{R}^{n \times j}$ are structured components in $\mathbf{X}$ and $\mathbf{Y}$ that are not correlated with each other nor with $\mathbf{T_S}$ and define the variations in $\mathbf{X}$ and $\mathbf{Y}$ that are uncorrelated to each other. The mixing matrices $\mathbf{A_X} \in \mathbb{R}^{a \times k}$, $\mathbf{B_X} \in \mathbb{R}^{v \times k}$, $\mathbf{A_Y} \in \mathbb{R}^{a \times m}$ and $\mathbf{B_Y} \in \mathbb{R}^{j \times m}$ are random mixing matrices with zero mean.

Our objective is to discover common structure between $\mathbf{X_0}$ and $\mathbf{Y_0}$ defined by $\mathbf{T_s}$, in the presence of unknown nonlinearity, by making use of not only the measurement matrices $\mathbf{X}$ and $\mathbf{Y}$ but also the available auxiliary noise matrix $\mathbf{Z}$, which is (partially) in the range of $\mathbf{T}_N$, which describes the structured noise present in $\mathbf{X}$ and $\mathbf{Y}$. It is desired to determine latent vectors in $\mathbf{X}$ that maximally explain $\mathbf{Y_0}$ and suppress the influence of $\mathbf{N_X}$ and $\mathbf{N_Y}$. Our hypothesis is that partial knowledge about $\mathbf{T}_N$ contained in the observations $\mathbf{Z}$ can improve the estimation of the latent variables $\mathbf{T_S}$. Suppressing these structured noise components improves the detection of the true underlying nonlinear relation between $\mathbf{X_0}$ and $\mathbf{Y_0}$ and leads to a better estimation of $\mathbf{T_S}$, and thus to an increase in the prediction accuracy of $\mathbf{Y_0}$. In the case of large structured noise in the linear or nonlinear case, the latent variables extracted by ordinary PLS or KPLS respectively may become more collinear with $\mathbf{T_N}$ than with $\mathbf{T_S}$, and thus the selected principal components will not optimally explain the true relationship between $\mathbf{X}_0$ and $\mathbf{Y}_0$.

Overall, the matrices $\mathbf{X}$ and $\mathbf{Y}$ can be written as:

$$\mathbf{Y} = g_{\mathbf{Y}}(\mathbf{T}_s)\mathbf{A}_{\mathbf{Y}} + \mathbf{U}_{\mathbf{Y}}\mathbf{B}_{\mathbf{Y}} + f_{\mathbf{Y}}(\mathbf{T}_N)\mathbf{J}_{\mathbf{Y}} + \sigma_{\mathbf{Y}}\mathbf{E}_{\mathbf{Y}} \qquad (4\text{-}62)$$

and

$$\mathbf{X} = g_{\mathbf{X}}(\mathbf{T}_s)\mathbf{A}_{\mathbf{X}} + \mathbf{U}_{\mathbf{X}}\mathbf{B}_{\mathbf{X}} + f_{\mathbf{X}}(\mathbf{T}_N)\mathbf{J}_{\mathbf{X}} + \sigma_{\mathbf{X}}\mathbf{E}_{\mathbf{X}}. \qquad (4\text{-}63)$$

This overall relationship between the components is explained in Figure 2-1.

The objective of the modeling effort is to obtain the best latent variable model for the true underlying signals $\mathbf{X_0}$ and $\mathbf{Y_0}$ as defined in equations (4-60) and (4-61) using the measurements of $\mathbf{X}$ and $\mathbf{Y}$ as defined in equations (4-62) and (4-63). Since the major interest in building a model between input and response variables is to explore the common subspace between them, two additional variables are defined as:

$$\mathbf{Y}^o = g_{\mathbf{Y}}(\mathbf{T}_s)\mathbf{A}_{\mathbf{Y}} \tag{4-64}$$

and

$$\mathbf{X}^o = g_{\mathbf{X}}(\mathbf{T}_s)\mathbf{A}_{\mathbf{X}}. \tag{4-65}$$

Hence , equations (2-30)-(2-31) can be rewritten using $\mathbf{X^0}$ and $\mathbf{Y^0}$ as:

$$\mathbf{X_0} = \mathbf{X^0} + \mathbf{U_X}\mathbf{B_X}, \tag{4-66}$$

$$\mathbf{Y_0} = \mathbf{Y^0} + \mathbf{U_Y}\mathbf{B_Y}. \tag{4-67}$$

These two variables only define the variations common to the subspace $\mathbf{T_S}$ and are different from $\mathbf{X_0}$ and $\mathbf{Y_0}$, in that $\mathbf{X_0}$ and $\mathbf{Y_0}$ contain $\mathbf{U_X}$ and $\mathbf{U_Y}$ which are unrelated to each other. The motive for defining such variables is that $\mathbf{X}$ is related to $\mathbf{Y}$ through $\mathbf{T_S}$, and therefore, at best, only that part of $\mathbf{Y}$ that is a function of $\mathbf{T_S}$ ($\mathbf{Y^0}$) can be explained by that part of $\mathbf{X}$ that is also a function of $\mathbf{T_S}$ ($\mathbf{X^0}$). Hence $\mathbf{Y^0}$ and $\mathbf{X^0}$ can be used as a measure of the true common variation

between **X** and **Y**.

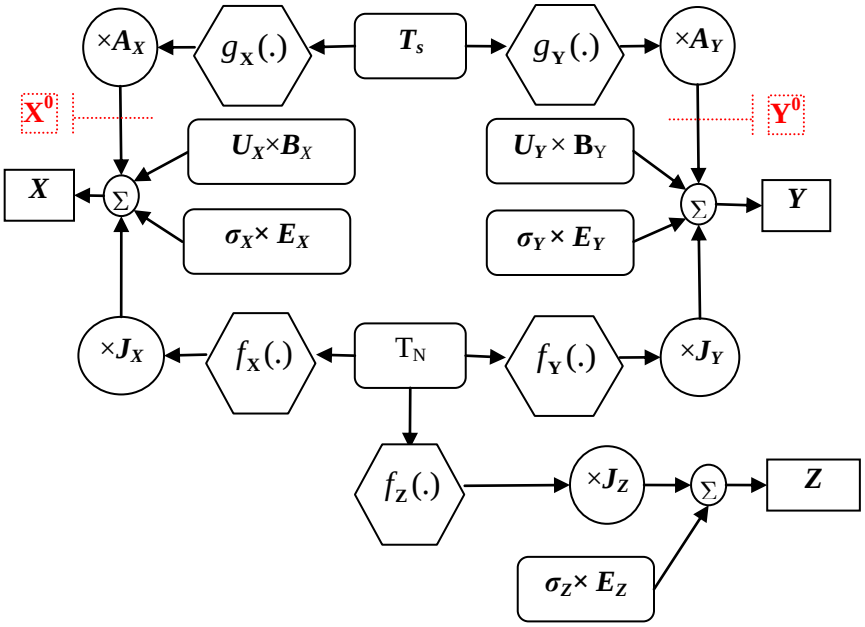


Figure 4-1 Relationship between $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ and the noise structure

In order to generate the matrices for the experiments, in each simulation four random orthonormal matrices are constructed: $\mathbf{T}_S \in \mathbb{R}^{n\times 6}$ and $\mathbf{T}_N \in \mathbb{R}^{n\times 6}$, $\mathbf{U_X} \in \mathbb{R}^{n\times 4}$ and $\mathbf{U_Y} \in \mathbb{R}^{n\times 4}$, with n equal to 2000 elements (unless otherwise stated), from which various latent structures in (2-32) - (2-33) are defined. The columns of $\mathbf{T_S}$, $\mathbf{T_N}$, $\mathbf{U_X}$ and $\mathbf{U_Y}$ are all mutually orthogonal.

The individual quality of fit from a projection of some quantity $\mathbf{\Psi}$ into the range of each principal component $\mathbf{t_i}$ can be calculated by:

$$R^2_{\mathbf{\Psi}} = 1 - \frac{\left\|\Psi - \hat{\Psi}\right\|^2_{\mathrm{F}}}{\left\|\Psi\right\|^2_{\mathrm{F}}} \qquad (4\text{-}68)$$

where $\mathbf{\Psi}$ can be any matrix or vector corresponding to the model such as $\mathbf{X}$, $\mathbf{X^0}$, $\mathbf{Y}$, $\mathbf{Y^0}$ or $\mathbf{Z}$. The quantity $\hat{\mathbf{\Psi}}$ is calculated by projecting $\mathbf{\Psi}$ into the range of $\mathbf{t_i}$ as:

$$\hat{\mathbf{\Psi}} = \mathbf{t}_i(\mathbf{t}_i'\mathbf{t}_i)^{\dagger}\mathbf{t}_i'\mathbf{\Psi}. \tag{4-69}$$

In the following simulations, we divide the simulation dataset into two equally-sized subsets. The first set of observations (the *training set*) is used to build the prediction model, whereas the second set (*the test set*) is used to test the quality of the model. In the experiments we denote the training dataset by the superscript "*tr*" and the test datasets are denoted by the "*ts*" superscript. All the predictions are performed on the test sets. The quality of prediction for previously unseen observations (i.e., those in the test set) is obtained from:

$$Q^2_{\mathbf{Y}} = 1 - \frac{\left\|\mathbf{Y^{ts}} - \hat{\mathbf{Y}}^{\mathbf{ts}}\right\|^2_{\mathrm{F}}}{\left\|\mathbf{Y^{ts}}\right\|^2_{\mathrm{F}}} \tag{4-70}$$

where $\hat{\mathbf{Y}}^{\mathbf{ts}}$ is the predicted value of $\mathbf{Y^{ts}}$. In addition to $Q^2_{\mathbf{Y}}$, we define an additional quality parameter denoted by $\mathbf{Q^2_{Y^0}}$ as:

$$Q^2_{\mathbf{Y}^0} = 1 - \frac{\left\|\mathbf{Y^{0ts}} - \hat{\mathbf{Y}}^{\mathbf{ts}}\right\|^2_{\mathrm{F}}}{\left\|\mathbf{Y^{0ts}}\right\|^2_{\mathrm{F}}}. \tag{4-71}$$

This parameter measures how close the predicted $\hat{\mathbf{Y}}$ and $\mathbf{Y^0}$, which is the only

part of **Y** that can be predicted by **X**, are to one another.

$\hat{\mathbf{Y}}^{ts}$ for the first (i) extracted latent variables is obtained from:

$$\hat{\mathbf{Y}}^{ts} = \mathbf{T}_i^{ts}\mathbf{C}'_i \tag{4-72}$$

where $\mathbf{T}_i^{ts} = [\mathbf{t}_1,\ldots, \mathbf{t}_i]$ and columns of C_i consists of the corresponding coefficient vectors. The cumulative quality of fit ($R^2(cum)$) is obtained by cumulating the quality of fit and prediction for the positive components respectively. The cumulative plots are generated using these cumulative quality variables.

We compare the quality of fit (and prediction) for ordinary PLS (i.e., PLS), soft constrained PLS (SC-PLS) with different levels of penalty coefficient (λ) and hard constrained PLS (HC-PLS), against their nonlinear correspondents: KPLS, SC-KPLS (different penalty values λ) and HC-KPLS. As previously mentioned the methods that utilize soft constraints such as SC-PLS and SC-KPLS may extract principal components with negative eigenvalues. Since these principal components are associated with variation in the noise subspace, they are excluded from the cumulative plots (i.e. $Q^2_{\mathbf{Y}}$, $Q^2_{\mathbf{Y}}{}^0$).

It is possible to include a nonlinear transformation for each function $g_{\mathbf{X}}(.)$, $f_{\mathbf{X}}(.)$ $g_{\mathbf{Y}}(.)$, $f_{\mathbf{Y}}(.)$ and $f_{\mathbf{Z}}(.)$. However, to simplify the presentation, only the following functions were (arbitrarily) chosen to undergo nonlinear transformations. The remaining transformations were kept linear. For the following transformations, the nonlinear transformations $g(\mathbf{T}_{\mathbf{s}})$ and $f(\mathbf{T}_{\mathbf{N}})$ are

chosen as:

$$g_{\mathbf{Y}}(\mathbf{T_S}) = [(\mathbf{t_{S1}})^1, (\mathbf{t_{S2}})^2, (\mathbf{t_{S3}})^3, (\mathbf{t_{S4}})^4, (\mathbf{t_{S1}} \odot \mathbf{t_{S5}}), (\mathbf{t_{S1}} \odot \mathbf{t_{S6}})], \quad (4\text{-}73)$$

$$f_{\mathbf{Z}}(\mathbf{T_N}) = [(\mathbf{t_{N1}})^4, (\mathbf{t_{N2}})^3, (\mathbf{t_{N3}})^2, (\mathbf{t_{N4}})^1, (\mathbf{t_{N1}} \odot \mathbf{t_{N5}}), (\mathbf{t_{N1}} \odot \mathbf{t_{N6}})], \quad (4\text{-}74)$$

$$f_{\mathbf{X}}(\mathbf{T_N}) = sin(\mathbf{T_N}). \quad (4\text{-}75)$$

The above transformations represent very common and relatively severe forms of nonlinearity in real situations. $(\mathbf{t}_{si})^r$ represents the i^{th} column of $\mathbf{T_S}$ to the r^{th} power (element-wise). The operator $(\odot)$ represents element-wise multiplication of its two vector arguments, resulting in a vector of the same length. In the Matlab context, this multiplication is denoted by the ".*" symbol. The power sign is also the same as the element-wise power operation (denoted by ".^" in Matlab software) for each element of the vector, and $sin(\mathbf{X})$ represents the sine transform of each element of the matrix. In practical situations the level of nonlinearity is usually much less severe. However for the sake of this study we chose a more complicated nonlinear relationship (especially between noise matrix **Z** and **X**) to study the ability of nonlinear kernels in such transformations

4.4.1.1 Case 1; structured noise only present in **X**

In nonlinear methods, due to the large size of the kernels, usually the number of components extracted will be much greater than linear methods, However, one advantage of latent variable methods is in compression of information using latent

variable methods. We will show that the use of constrained methods, in the presence of structured noise, can improve component selection (compression of more information into fewer latent variables).

When the structured noise is only present in $\mathbf{X}$, constraining the structured noise in a kernelized framework allows for selection of latent variables that are more likely to be in the subspace of $\mathbf{T_S}$ rather than $\mathbf{T_N}$. We shall see that latent variables extracted using the proposed method provide a better basis for the signal and explain both the input and response spaces through fewer components. In addition, in the soft constrained methods, the corresponding eigenvalue sign for each component determines whether the extracted component has a stronger covariance with the noise or the response variables. This allows for easier interpretation of the noise or the signal structure. It should be noted that if a large number of components are selected, the non-constrained methods eventually reach the same accuracy as the constrained methods; however, the presence of noise in the input will result in biased estimates of the response variables in the nonlinear case. In this simulation example the structured noise in $\mathbf{Y}$ is set to zero by setting $\mathbf{J_Y} = \mathbf{0}$. It is assumed that simultaneous measurements of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are available. The relation between the remaining components are determined through the following choice of parameters: $\sigma_{\mathbf{X}} = \sigma_{\mathbf{Y}} = 0.1, \|\mathbf{J_X}\|_F = 3, \mathbf{J_Y} = \mathbf{0}$. The remaining scalar coefficients are set to 1. The $\|.\|_F$ operator denotes the Frobenius norm of the matrix. The size of the mixing coefficient matrices

$\mathbf{A_Y}, \mathbf{B_Y}, \mathbf{J_Y}$ and $\mathbf{A_X}, \mathbf{B_X}, \mathbf{J_X}$ are 6×18, 6×18, 4×18, 6×32,6×32 and 4×32 respectively. The mixing matrix for the noise matrix $\mathbf{Z}$ ($\mathbf{J_Z}$) is 6×6. Hence the size of the produced datasets $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ will be: 2000×32, 2000×18 and 2000×6 respectively. The datasets were later divided into the equal size *training* and *test sets* (1000 observations each)

A Gaussian kernel (4-28) was used in the simulation. The kernel parameter *c* was chosen manually by trying several different values and choosing the one that provides the best prediction rate. In practice however, *c* must be selected using a formalized cross validation procedure. This is detailed further in the conclusion section.

The datasets $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ were constructed by first computing a 1000×20 orthonormal matrix. Then the latent variables $\mathbf{T_S}$, and $\mathbf{T_N}$ and the auxiliary variables $\mathbf{U_X}$ and $\mathbf{U_Y}$ were constructed by selecting columns from this matrix according to Table 2-2. Dataset variables were then synthesized from this orthonormal set in the manner described earlier in this section.

TABLE 6: TOY EXAMPLE, CASE 1, STRUCTURED NOISE IS ONLY PRESENT IN **X.** THE CELLS SHOW WHICH LATENT COMPONENTS WHERE COMBINED IN CONSTRUCTING **X, Y** AND **Z**

	$\mathbf{T}_S \in \mathbb{R}^{1000\times 6}$	$\mathbf{T}_N \in \mathbb{R}^{1000\times 6}$	$\mathbf{U}_X \in \mathbb{R}^{1000\times 4}$	$\mathbf{U}_Y \in \mathbb{R}^{1000\times 4}$

$\mathbf{X}^{tr}, \mathbf{X}^{ts} \in \mathbb{R}^{1000\times 32}$	←→	←→	←→	
$\mathbf{Y}^{tr}, \mathbf{Y}^{ts}, \in \mathbb{R}^{1000\times 18}$	←→			←→
$\mathbf{Z}^{tr}, \mathbf{Z}^{tr} \in \mathbb{R}^{1000\times 6}$		←→		

Figure 4-2 shows the quality of prediction between $\hat{\mathbf{Y}}^{ts}$ and $\mathbf{Y}$ for various linear and nonlinear methods described earlier. In this and subsequent examples, the model-building process (i.e., the determination of the latent variables and the kernel parameter values) is performed using only the training subset. The prediction (testing) procedure uses only the test subset. This figure shows that the methods based on a linear model (PLS and HC-PLS) fail to provide adequate predictions in this nonlinear scenario, since it may be observed that prediction quality does not improve with increasing number of latent variables. Thus, the linear methods fail to model the problem adequately. In contrast, it is apparent that the kernelized methods are capable of adapting to the nonlinear structure of the underlying model and consequently provide better predictions. A comparison between the KPLS and the constrained versions of KPLS (i.e., SC-KPLS and HC-KPLS) shows that imposing constraints further improves the quality of prediction, in the important case when the number of extracted components is moderate or low. On this basis, we may observe the proposed methods are capable of producing a latent variable basis that is effective in the presence of nonlinearity,

and that can adequately exploit additional information contained in structured noise observations.

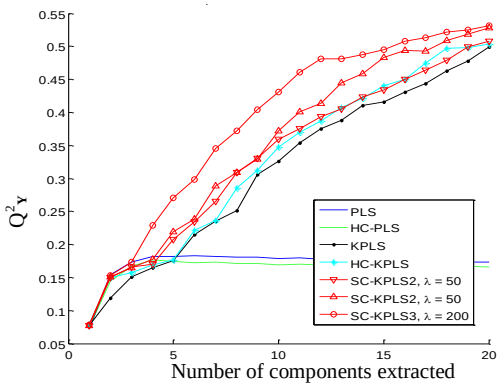


Figure 4-2: Quality of prediction as the number of components increase, for various linear and nonlinear models between $\hat{\mathbf{Y}}^{ts}$ and $\mathbf{Y}^{ts}$ $Q^2_{\mathbf{Y}}$

Figure 4-3 shows the individual quality of fit (by projecting into each individual latent variable $\mathbf{t}_i$) for the response values ($\hat{\mathbf{Y}}^{tr}$) and the auxiliary noise matrix ($\hat{\mathbf{Z}}^{tr}$) in the training set, for all components regardless of the sign of the corresponding eigenvalues, for the SC-KPLS method ($\lambda = 200$). The bottom chart shows the sign of the eigenvalue associated with each of the extracted components.

In the discussion on the SC-KPLS method in Sect. 4.3, we explained that when the dominant eigenvalue of the matrix $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X} - \lambda\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}$ in (4-33) is positive, the extracted latent variable explains variation in $\mathbf{Y}$, and in $\mathbf{Z}$ when negative. This behavior is demonstrated in Figure 4-3. It may be observed that in the positive

eigenvalue case, the fit of the corresponding component **t** to **Y** is significant, whereas the fit with **Z** can be seen to be negligible. The reverse behavior is noted when the eigenvalue is negative. The individual quality of fit for Figure 4-3 is obtained from (4-68), where $\hat{\mathbf{Y}}^{\mathbf{tr}} = \mathbf{t}_i(\mathbf{t}_{i'}\mathbf{t}_i)^{-1}\mathbf{t}_i\mathbf{Y}^{tr}$ and $\hat{\mathbf{Z}}^{\mathbf{tr}} = \mathbf{t}_i(\mathbf{t}_{i'}\mathbf{t}_i)^{-1}\mathbf{t}_i\mathbf{Z}^{tr}$ for individual components in $\mathbf{T}^{\mathbf{tr}}$.

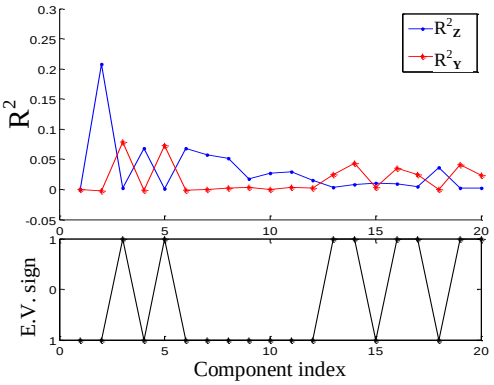


Figure 4-3: Quality of fit for $\hat{\mathbf{Y}}^{\mathrm{tr}}$ and $\hat{\mathbf{Z}}^{\mathrm{tr}}$ from projection into the individual components of $\mathbf{t}_i$,i = 1,… 20 respectively ($R^2_{\mathbf{Y}}$, $R^2_{\mathbf{Z}}$ from (4-68)), for the first 20 components extracted. The bottom plot shows the sign of the respective eigenvalue associated with each latent variable $\mathbf{t}_i$.

The plots in Figure 4-4 show the individual quality of fit into the subspace of $\mathbf{t}_i$, for $g_{\mathbf{Y}}(\mathbf{T}_{\mathbf{S}})$, $f_{\mathbf{Z}}(\mathbf{T}_{\mathbf{N}})$ and $f_{\mathbf{X}}(\mathbf{T}_{\mathbf{N}})$ which represent the nonlinear transformations of the signal subspace in **Y**, and the noise subspaces in **X** and **Z** respectively. The plots shows the contribution of each extracted latent variable ($\mathbf{t}_i$) to the nonlinear variations of noise and signal in **X** and **Y** and **Z**, in association with the respective eigenvalue sign. It is apparent that the latent variables do indeed capture substantial variation in each of these respective functions and that the components capturing variations in noise ($f_{\mathbf{Z}}(\mathbf{T}_{\mathbf{N}})$ and $f_{\mathbf{X}}(\mathbf{T}_{\mathbf{N}})$) capture very small variance in

the signal subspace ($g_{\mathbf{Y}}(\mathbf{T_S})$), and vice versa.

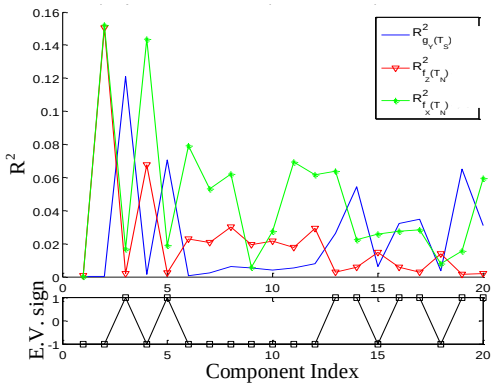


Figure 4-4: Training set's individual (non-cumulative) quality of fit from projecting the nonlinear transformations of $\mathbf{T_S}$ and $\mathbf{T_N}$, i.e. ($g_y(\mathbf{T}_s)$, fz($\mathbf{T}_n$) and $f_x(\mathbf{T}_n)$) for each of the datasets: **X**, **Y** and **Z** respectively into the range of each extracted principal components ($\mathbf{t}_i$) (*training set* data). It can be seen that latent variables corresponding to positive eigenvalues are associated with the signal subspace and the negative ones are associated with the noise subspace.

4.4.1.2 Case 2; structured noise present in both **X** and **Y**

When structured noise, with the same source, is affecting both **X** and **Y**, then regular non-constrained methods can not remove variations caused by noise (**Z**) without external information, using constrained methods proposed earlier allows for removal or reduction of the noise allowing for better selection of latent components and prediction of future true values of response. The latent variables extracted in the presence of structured noise in both **X** and **Y** (without constraining the noise subspace) are capable of modeling the noise components rather than the desired relationship between **X** and **Y**. The extent to which this is

possible is dependent on the structured noise level.

In such a case it is desirable to suppress the noise during the model construction procedure. The proposed constrained PLS methods take advantage of the additional knowledge that is available to reduce or eliminate the common structured noise. In this second toy problem new simulation datasets are generated by allowing both the input and response variables to be contaminated with structured noise, by assigning random (nonzero) values to the elements of both $\mathbf{J_X}$ and $\mathbf{J_Y}$ (the mixing matrices). Use of the proposed constrained methods leads to the extraction of components that better predict the true underlying relationship between $\mathbf{X}^0$ and $\mathbf{Y}^0$.

In this particular example the following parameters are used: $\sigma_{\mathbf{X}} = \sigma_{\mathbf{Y}} = 0.1$, $\|\mathbf{J_X}\|_F = \|\mathbf{J_Y}\|_F = 0.9$. Setting $\mathbf{J_Y}$ to nonzero values adds structured noise to $\mathbf{Y_0}$.

Figure 4-5 shows the cumulative quality of projection (Q^2_Y) between $\hat{\mathbf{Y}}^{ts}$ and $\mathbf{Y}$ and also the quality of prediction ($Q^2_Y{}^0$) between $\hat{\mathbf{Y}}^{ts}$ and $\mathbf{Y^0}$ (2-34) from various nonlinear methods mentioned earlier (Left and right respectively). Figure 4-5-right shows the quality of prediction between $\hat{\mathbf{Y}}^{ts}$ and $\mathbf{Y}^0$ which represent the only part of **Y** that can be explained by **X**. Figure 4-5-left shows the prediction rate between $\hat{\mathbf{Y}}^{ts}$ and the "measured" **Y** variables contaminated with the noise. Comparing the left and right plots shows that even though the quality of prediction to $\mathbf{Y}$ is comparable for the linear and nonlinear models, the true quality of fit, to $\mathbf{Y}^0$, is lower when the structured noise is not suppressed. The

reason is that non-constraining methods tend to model the common structured noise as well as the true underlying structure, leading to false predictions . It is also apparent from the left- and right-hand figures that the kernelized methods perform significantly better than their linear counterparts.

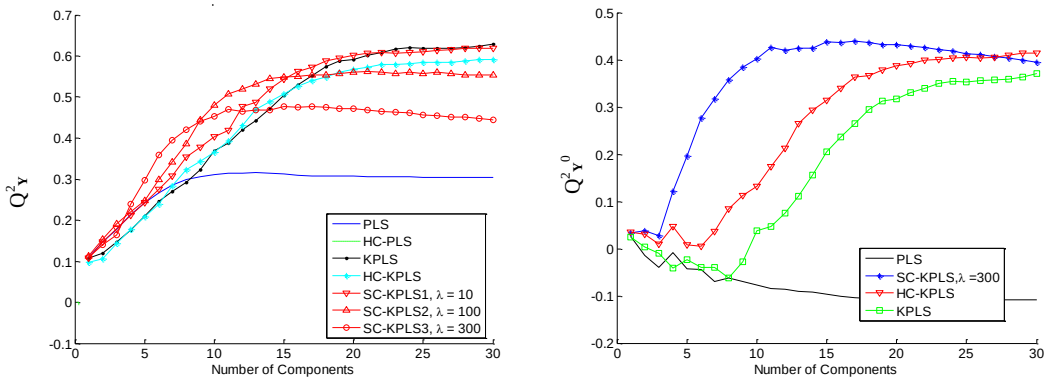


Figure 4-5. Left: Quality of prediction (Q^2_Y) between $\hat{\mathbf{Y}}^{ts}$ and the measured response value $\mathbf{Y}$. Right: quality of prediction ($Q^2_Y{}^0$) between $\hat{\mathbf{Y}}^{ts}$ and $\mathbf{Y^0}$ which is a function of $\mathbf{T_S}$. It is evident that the constrained methods provide better models for the true underlying structure common to both the input and response spaces, as they suppress the structured noise during the model construction.

4.4.1.3 Effect of noise on Quality of prediction

Compared to the linear methods, kernelized methods can be more sensitive to noise and therefore constrained models are critical for maintaining performance of the kernelized methods. In this section, we examine the sensitivity of the proposed methods to the level of structured noise and show that when the magnitude of the structured noise changes, the kernel PLS method's efficiency degrades rapidly whereas the constrained Kernel methods have less sensitivity to the noise contamination levels. In such cases, using constrained methods can

greatly improve the quality of component selection. In this study various datasets, with same values for **X** and **Y** and **Z**, but with different values of $\mathbf{J}_\mathbf{X}$ with different Frobenius norms were generated. $\mathbf{J}_\mathbf{Y}$ was kept constant for all the simulations.. Three magnitudes: $\|\mathbf{J}_\mathbf{X}\|_F = 0.3, 0.6, 0.9$ were chosen arbitrarily and $\|\mathbf{J}_\mathbf{Y}\|_F$ was chosen arbitrarily to be constant at 0.5. The following figure shows the cumulative quality of fit between $\hat{\mathbf{Y}}^{ts}$ and $\mathbf{Y}^{0ts}$ ($Q^2_{\mathbf{Y}^0}$) for KPLS and SC-KPLS. It can clearly be seen that as the level of noise increases, the quality of prediction degrades rapidly in the non constrained KPLS method. However, the quality of prediction for the SC-KPLS method remains relatively unchanged as the structured noise levels increase.

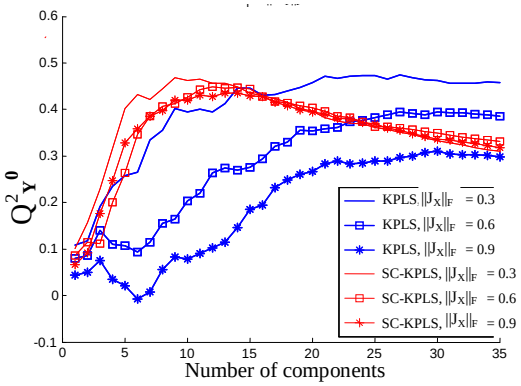


Figure 4-6: Cumulative quality of prediction between $\hat{\mathbf{Y}}$ and $\mathbf{Y}^0$ ($Q^2_{\mathbf{Y}^0}$). As the noise level increases, the prediction quality in the non-constrained methods (blue curves) reduces and more components are required to achieve same level of prediction. In the soft constrained KPLS (SC-KPLS $\lambda = 300$) methods (red curves) the quality of prediction and also the number of components required to achieve the same prediction level remains relatively unchanged.

4.4.1.4 Number of observations:

Nonlinear methods are also sensitive to the number of observations. As the

number of observations increases, the size of the kernel grows larger. Eventually, this requires the model to extract more components in order to properly model the underlying relationships. This behavior becomes more prominent in the presence of structured noise. The following simulations show that the constrained methods require fewer components to properly model the relationships between **X** and **Y** compared to their non-constrained counterparts. To illustrate this hypothesis, two datasets were generated, all using the settings: $\sigma_{\mathbf{X}} = \sigma_{\mathbf{Y}} = 0.1$, $\|\mathbf{J}_{\mathbf{X}}\|_F = 0.9, \|\mathbf{J}_{\mathbf{Y}}\|_F = 0.5$, with different number of observations. The plots in Figure 4-7 compare the cumulative predictions rates between $\hat{\mathbf{Y}}$ and $\mathbf{Y}^0$ ($Q^2_{\mathbf{Y}^0}$) when the number of observations is 600 versus the case in which the number of observations has increased to 1000. The figure shows that when the number of observations increases, the non-constrained methods compared to the constrained methods require more components to capture the same level of the prediction accuracy.

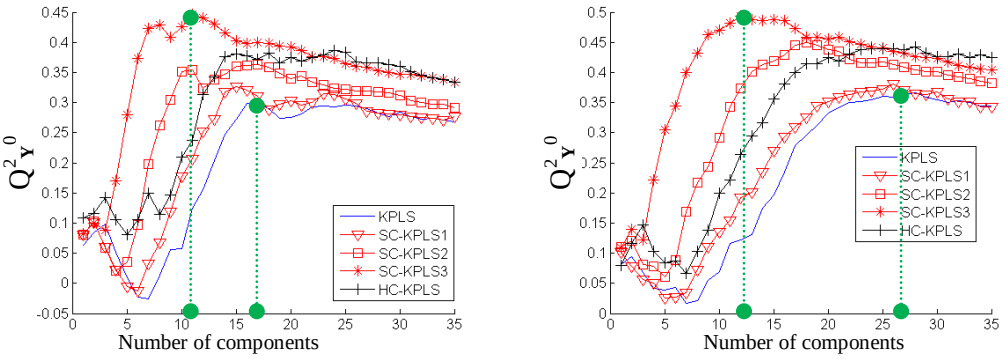


Figure 4-7: Cumulative quality of prediction between $\hat{\mathbf{Y}}$ and $\mathbf{Y}^0$ $Q^2_{\mathbf{Y}^0}$, comparing KPLS versus constrained methods when the number of observations increases. The model in the left figure was constructed using 600 observations, whereas the model in the right figure was constructed using

1000 observations. As the number of observations increases, the number of components required to capture the maximum prediction rate increases in the KPLS method. However this quantity remains relatively unchanged for the constrained methods. SC-KPLS values: SC-KPLS1: $\lambda = 10$, SC-KPLS2: $\lambda = 100$, SC-KPLS3: $\lambda = 300$

4.5 Discussion and Conclusion

We have demonstrated that when significant nonlinear relationships exist between **X** and **Y**, nonlinear kernel latent variable methods perform better than conventional linear methods in capturing the underlying model. However, as with their linear counterparts, they fail to provide the most reliable results when the model is contaminated with structured noise. In such cases, available information about the noise can be exploited to improve predictions. We have developed and demonstrated two such kernelized constrained methods; the HC-KPLS and the SC-KPLS methods. The latter has a parameter associated with it that controls the degree of constraint on the noise. We have shown that kernelized latent variable methods are effective at modeling non-linearity in the model, and that performance in the presence of structured noise is improved with the use of the proposed constrained methods. However, the improved performance of these methods comes at the cost of increased model complexity and the need to determine the optimal value of various parameters that affect performance.

We realized several issues that need to be addressed while implementing the nonlinear methods and, in particular, constrained ones;

4.5.1.1 Choice of kernels

Many kernels are available for model building. The type of kernel and the chosen values of the associated parameters can greatly change the effectiveness of the model. Thus another issue is that when there is structured noise in both **X** and **Y**, even cross validation results will not be reliable as the true underlying relationship between **X** and **Y** is hidden.

4.5.1.2 Large Kernels sub-sampling in the feature space

Kernel methods map the input data into a potentially much larger feature space, which is usually the size of the number of observations. Calculation of the kernel matrix in this condition, especially for a large number of observations, is not very cost effective and the obtained kernel can be ill-conditioned. Garcia et al., [17] proposed the implementation of a modified kernel constructed as:

$$\mathbf{K}_R = \Phi_R \Phi' \tag{4-76}$$

where Φ_R is constructed from only a subset of observations in $\mathbf{\Phi}$. Choosing the modified kernel can reduce the computation costs as well as benefiting from having better conditioned kernels. However, if too few observations are chosen, the effectiveness of the model is reduced, as this limits the level of nonlinearities permitted by the model. Using a very small subset of observations may not completely capture the nonlinear behavior. A very large observation set, in addition to being computationally expensive, may also decrease model accuracy. A simulation study comparing this method with regular constrained methods is presented in the Appendix.

4.5.1.3 Number of components to be extracted

The number of components to be extracted should be determined to avoid over-fitting (or under-fitting) of data. In our simulations, a good guess of the best number of components was determined using a cross validation procedure. We divided the observations into two equal sets. The first set was used to build the model, while the second portion was used to compute the statistical properties from which performance was evaluated. When structured noise is present in both the input and the response variables, the lack of access to the true response variables makes model selection more difficult. However, the use of constrained methods can improve the selection as these components are less likely to be correlated with the structured noise, leading to improved and more stable predictions.

4.5.1.4 Rank deficiency of the kernels

Since at each iteration step the kernels are deflated, eventually they become ill-conditioned. This problem is more prevalent in the kernel case as the optimization problem is a generalized eigenvalue problem. Our solution was to add a regularizing diagonal matrix to the second kernel in equations (4-16) and (4-17) in order to prevent rank deficiency. This additional diagonal matrix biases the estimates but its effect is very minimal and our studies showed the results are relatively insensitive to the value of the diagonal matrix.

4.6 Appendix

4.6.1 Subsampling in the feature space

When the number of observations is high, the dimensionality of the kernel can row as large as the sample size. This issue is addressed by running simulations comparing two cases. In the first case all observations are used to build the model and in the second case only a subset (25%) of observations are used. Figure 4-8 plots show the quality of fit to $\mathbf{Y}$ (left figure) and the prediction rate of $\mathbf{Y}^{\mathbf{0}}$ (right figure), which contains that part of $\mathbf{Y}$ that can be explained by $\mathbf{X}$. The figures show that despite using only 25% of the observations to train the model using the method proposed by Garcia et al. the prediction rates are roughly the same.

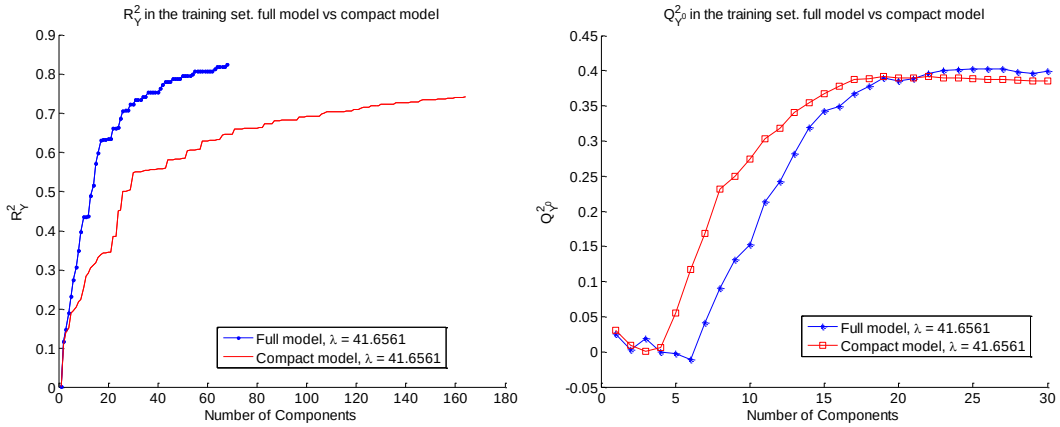


Figure 4-8: Left: Quality of fit for the training set when full range of $\mathbf{X}$ is used to build the kernel versus using only 25% of the observations (compact model). Right: Quality of prediction for the test set, comparing the full model and the compact model.

REFERENCES

1. Burnham, A. J., Viveros, R. & MacGregor, J. F. Frameworks for latent variable multivariate regression. *Journal of chemometrics* **10**, 31–45 (1996).
2. Wold, S., Antti, H., Lindgren, F. & Öhman, J. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems* **44**, 175–185 (1998).
3. Fearn, T. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems* **50**, 47–52 (2000).
4. Trygg, J. & Wold, S. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* **16**, 119–128 (2002).
5. Trygg, J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics* **16**, 283–293 (2002).
6. Trygg, J. & Wold, S. O2- PLS, a two- block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics* **17**, 53–64 (2003).
7. Hastie, T. & Stuetzle, W. Principal Curves. *Journal of the American Statistical Association* **84**, 502–516 (1989).
8. Bregler, C. & Omohundro, S. M. Surface learning with applications to lipreading. *Advances in neural information processing systems* 43 (1994).
9. Wold, S., Kettaneh-Wold, N. & Skagerberg, B. Nonlinear PLS modeling. *Chemometrics and Intelligent Laboratory Systems* **7**, 53–65 (1989).
10. Gnanadesikan, R. *Methods for statistical data analysis of multivariate observations*. **321**, (Wiley-Interscience: 1997).
11. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **10**, 1299–1319 (1998).
12. Rosipal, R., Girolami, M., Trejo, L. J. & Cichocki, A. Kernel PCA for Feature Extraction and De-Noising in Nonlinear Regression. *Neural Computing & Applications* **10**, 231–243 (2001).
13. Jade, A. M. *et al.* Feature extraction and denoising using kernel PCA. *Chemical Engineering Science* **58**, 4441–4448 (2003).
14. Rännar, S., Lindgren, F., Geladi, P. & Wold, S. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *Journal of Chemometrics* **8**, 111–125 (1994).
15. Rosipal, R. & Trejo, L. Kernel partial least squares regression in reproducing kernel hilbert space. *The Journal of Machine Learning Research* **2**, 97–123 (2002).
16. Rosipal, R. Kernel partial least squares for nonlinear regression and discrimination. *Neural network world* **13**, 291–300 (2003).
17. Arenas-García, J., Petersen, K. B. & Hansen, L. K. Sparse kernel orthonormalized PLS for feature extraction in large data sets. *Advances in Neural Information Processing Systems* **19**, (2006).
18. Fonville, J. M. *et al.* The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *Journal of Chemometrics* **24**, 636–649 (2010).
19. Rosipal, R. & Krämer, N. Overview and Recent Advances in Partial Least Squares. *Subspace, Latent Structure and Feature Selection* **3940**, 34–51 (2006).
20. Salari Sharif, S., Reilly, J. P. & MacGregor, J. Latent Variable Methods in the Presence of Structured Noise. *Journal of Chemometrics*
21. Rantalainen, M. *et al.* Kernel-based orthogonal projections to latent structures (K-OPLS). *Journal of Chemometrics* **21**, 376–385 (2007).
22. Kim, K., Lee, J.-M. & Lee, I.-B. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems* **79**, 22–30 (2005).
23. Zhang, Y. & Teng, Y. Process data modeling using modified kernel partial least squares. *Chemical Engineering Science* **65**, 6353–6361 (2010).
24. Rännar, S., Geladi, P., Lindgren, F. & Wold, S. A PLS kernel algorithm for data sets with many variables and few objects. Part II: Cross validation, missing data and examples. *Journal of chemometrics* **9**, 459–470 (1995).
25. Kettaneh, N., Berglund, A. & Wold, S. PCA and PLS with very large data sets. *Computational Statistics & Data Analysis* **48**, 69–85 (2005).
26. Mercer, J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **209**, 415 (1909).
27. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222 (2004).
28. Muller, K., Mika, S., Ratsch, G., Tsuda, K. & Scholkopf, B. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks* **12**, 181–201 (2001).
29. Rao, C. R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics, Series A* **26**, 329–358 (1964).

Chapter 5
Removing Structured Noise from Electroencephalogram Data

***Abstract*—An essential step prior to the analysis of electroencephalographic (EEG) data is the removal of noise artifact arising from muscle activity or mechanical disturbance of the electrode-skin junction. In combined EEG-fMRI (functional magnetic resonance imaging) analysis, another type of noise known as Ballistocardiographic noise (BCG) induced by the movement of blood in the static magnetic field also contaminates the EEG data. In both cases, the artifacts appear as structured noise affecting many channels. In most cases there is additional information about the noise that can be utilized to remove the artifacts from EEG data. In this article, we propose the use of a soft constrained PLS algorithm for removal of the noise-related components from the data. This method makes use of the additional knowledge about the event related potentials (ERPs) and latent variable methods (LVM) to extract EEG components that are noise free. We applied the SC-PLS algorithm to EEG data contaminated with muscle artifacts or BCG artifacts (both experimental data and simulated data) and compared the results to those obtained by conventional methods. In both studies the results were satisfying and comparable to the current noise removal methods.**

***Index Terms*— EEG, fMRI, Structured noise, PLS, ICA, PCA, SC-PLS, Ballistocardiographic noise, BCG, Muscle artifacts**

5.1 Introduction

This chapter looks into the problem of removing artifacts in the electroencephalogram (EEG). This includes muscle artifacts, as well as another type of artifact called, ballistocardiographic (BCG) noise, induced into EEG data in a medical procedure known as simultaneous EEG-fMRI (functional Magnetic resonance imaging). The EEG records the electric waves produced in the brain as a result of the neural activity in the brain. EEG data provides useful information about functional responses to stimuli including localization information.

In the first section of this chapter, the EEG and fMRI are each introduced

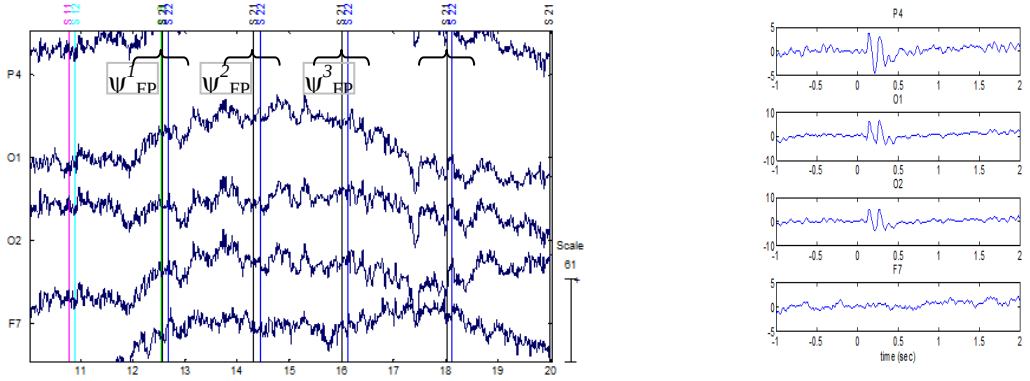
briefly. Next, the combined method and its advantages and difficulties are discussed. Later, the problem of additive BCG noise, which is a consequence of the combined method, its nature and the current methods for removing it are discussed. In the remainder of the chapter, two new methods for removing noise are discussed that take advantage of additional knowledge about the BCG noise. In addition, we apply the constrained methods as discussed in chapter two for removal of muscle movement induced artifacts during standard EEG recording outside of the MRI environment.

5.1.1 Electroencephalogram

Electroencephalography refers to the technique of measuring the brain's electric field from the surface of the scalp. Activity of the cells in the brain will cause intra- and extra-cellular current flows that can be measured using non-invasive methods. When large bundles of neurons fire synchronously, the changes in the local field potentials (LFP) can be sensed and recorded from the surface of the scalp. The measured voltage will be a weighted sum of the local field potentials throughout the brain. It is believed that the synchronous activations of the neurons that have laminar structure can be detected from the surface of the scalp [1]. The measured scalp voltage on each electrode will be a function of an activation site's strength, conductivity, distance from the active sources and their orientation. Therefore, each active site can be addressed as a dipole of specific orientation and strength. Most of the signal received by each electrode will be contributed by the local field potentials close to that electrode. However, activations from large

bundles of neurons at distant locations can also be detected [2]. Recorded background scalp levels of EEGs range around ±75 μV [1], peak to peak; however potentials induced by the introduction of a stimulus (known as event-related potentials or ERPs) are typically an order of magnitude smaller. Thus, in order to reliably detect the ERPs, repeated measurements are usually required. These ERPs are usually averaged by time locking them to an external stimulus (trigger). We refer to them as averaged ERPs or A-ERP. An example of an ERP signal are the Visually Evoked Potentials (VEP) produced as a result of response to visual stimuli. The duration of the information processing during an event-related task can take up to several hundred milliseconds. The measured response wave for each ERP will include several peaks and valleys. These peaks and valleys (ERP components) are known to be associated with different stages of cognitive data processing and usually are named according to their polarity and the latency at which they occur. For example, a P300 component stands for a positive peak occurring around 300 ms after the onset of the stimulus. The latency and amplitude of some of these components are known to be associated with functionality of the brain. For example, Loudness Dependence of auditory Evoked Potentials (LDAED) components have shown to be predictive of serotonergic neurotransmitter levels and it has been used in prediction of electrophysiological changes associated with neurological disorders such as changes in neurotransmitter levels in the brain [3,4]. Researchers have utilized these biomarkers to predict the outcome of drug treatment in brain disorders such as

depression. [5,6]. Electroencephalography has also been widely used in epilepsy research to detect the onset and the location of the seizures [7]. Another application of EEG is to localize the source of brain activity. Examples of these studies can be found in [8,9].



$$\overline{\psi(t)}_{ch} = \frac{1}{n}\sum_{i=1}^{nERP} \psi^{i}(t) \quad -1 < t < 2 \ (sec)$$

Figure 5-1: Extraction of the averaged ERPs. The EEG is time locked into the stimuli being presented to the patient (s11,s22 in left figure). In order to get Averaged-ERP response of the brain several repeated measurements (e.g. time locked to S11 component) are averaged over time to remove random noise and create a clear signal (right figure). The parameter nERP is the total number of ERP's present in the study and *ch* represents each EEG channel.

5.1.2 Functional Magnetic Resonance Imaging (fMRI)

fMRI consists of a series of MRI images that are recorded using an MRI protocol that is specifically sensitive to oxy/deoxyhemoglobin levels in the vessels and tissue. Simply put, the difference in magnetic properties of oxyhemoglobin and deoxyhemoglobin will cause contrast changes in the fMRI image. This contrast change is called Blood Oxygen Level Dependant contrast, or

the "BOLD" signal. There is a close relationship between local BOLD contrast and neural activity in the underlying brain region [10–12]. In general, fMRI BOLD contrast is a function of cerebral blood flow (CBF), cerebral blood volume (CBV) and metabolic oxygen consumption [13]. This relationship is not yet fully understood; however, the general consensus is that during local activation of large bundles of neurons, oxygen demand in that region increases, resulting in a rush of oxygen rich blood flow into that region. The regional oxygen concentration changes will cause a detectable BOLD signal. Activation of small or individual bundles of neurons, especially if not synchronized, is unlikely to cause any detectable BOLD signal changes [14]. It is also believed that in low MRI magnetic field strengths, the BOLD signal changes observed can be influenced by venous flow rather than by the oxygen concentration changes in the tissue [15].

Functional MRI has recently found increasing clinical usage. It is used to identify brain regions such as the motor and speech cortex for the pre surgery screening or for measuring the ability to stimulate the ear for cochlear implantations. Bartsch et al. [16] outlines a detailed survey on the clinical applications of fMRI. As well, fMRI data has been used in depression studies. Langenecker et al [17] performed a study using fMRI on two cohorts; controls and major depression (MD) patients. In that study, subjects were to complete a contextual inhibitory control test while they were scanned by MRI. The authors found that MD patients had a greater activation in the frontal area during correct rejection tasks. The fMRI hot spots have also been used as initial seeds in brain

tractography. [16].

5.1.3 Combined and Simultaneous EEG-fMRI

Currently, combined EEG-fMRI studies are gaining popularity amongst researchers. The reason is twofold: the complementary nature of the two modalities in terms of information content and the spatiotemporal resolution of each modality. Despite EEG's superior temporal resolution capabilities, it has very poor spatial resolution, and it relies on the solution of ill-posed boundary problems for localizing sources in the brain. Furthermore, it only reflects the subspace of the brain's characteristics that is directly related to normal electric currents. On the other hand, fMRI suffers from poor temporal resolution and is mainly reflective of the metabolites and perfusion activity inside the brain. These activities are in turn related to activation of the neurons inside the brain. Each of these modalities provides information that the other lacks.

Combined EEG-fMRI studies have been used to improve localization of the sources of brain activity in EEG studies [18]. In clinical applications, EEG-fMRI has been used to localize the sources of epileptic seizures [18–20] .

Combining EEG and fMRI data collected separately has some disadvantages due to the variabilities caused by the changes in patient's vigilance and also environmental factors [21]. Therefore, simultaneous recording is more favorable when such problems threaten the outcomes of the studies. In addition, in some applications such as seizure studies, it is necessary to perform EEG and fMRI simultaneously.

In simultaneous recording, besides the usual contaminations of the EEG data caused by blinking, eye movement, muscle activity or mechanical movement of the skin electrode junction caused by movement of the patient, the signals are also subject to additional noise from other origins. These additional noises must be removed before the signal becomes acceptable for further analysis. These additional artifacts are the Ballistocardiographic (BCG) and Gradient Artifact (GA) noise.

5.1.4 Gradient Artifacts

When an EEG is performed simultaneously with the MRI, the RF dissipation and the changes in the gradient fields of the MRI can create currents inside the EEG wires that will cause additional noise known as Gradient Artifact (GA). These artifacts can be avoided by recording the EEG during the silent phase of the MRI between each slice acquisition section. This method is called interleaved scanning. If the EEG is recorded continuously while the scanner is acquiring images, the gradient artifacts must be removed manually using signal processing methods prior to any further analysis. Using the proper steps during or prior to scanning can significantly reduce the GAs. For example, studies have shown that twisting the EEG wires and recording the EEG in bipolar format can greatly reduce GAs [22] by relying on the differential amplifier's common-mode rejection to remove artifact induced equally on both of the twisted wires. It is also a good practice to use short-length EEG leads to reduce the amount of RF dissipation into the wires. In addition to producing GAs, the RF dissipation in the EEG leads can

heat up the electrodes, which may even cause discomfort in the patients [23]. Non-ferromagnetic EEG leads such as pure silver or silver/silver-chloride are common options for preventing heating in electrode leads [24].

Since fMRI sequences are almost perfectly periodic, they produce periodically repeated artifacts. The most common signal processing method to remove these artifacts is to average the time-locked signals and subtract the estimated artifact template from the EEG data at each time point that the slices are acquired. This method, to the best of our knowledge, was first introduced by Allen et al. [25] and is known as average artifact subtraction (AAS). Gradient artifacts generated in the MRI are synchronized to much higher clock speeds compared to the sampling rate of the EEG, and therefore, a perfect time locking of the RF signals using conventional EEG hardware is not possible. Using conventional EEG systems will result in imperfect averaging of the time-locked RF pulses, which will lead to incomplete subtraction of the GAs. One way to improve the averaging procedure is to use EEG systems with a much higher sampling rate (above 5 kHz) or by synchronizing the EEG clock with the MRI's internal clock using additional hardware. After artifact subtraction, a low pass filter is usually required to be applied to remove any residual high-frequency artifacts remained in the data.

5.1.5 Ballistocardiographic noise

According the Maxwell's Law of electromagnetism, movement or vibration of a conductive loop inside a magnetic field induces current flow inside that medium (loop). In simultaneous EEG-fMRI recording, conductive loops exist that consist

of the EEG electrodes, wires, the EEG amplifier and the patient's body. The magnet's cooling pump, vibrations caused by the MRI's gradient coils and the cardiac pulsation of the patient or the body movements, all contribute to generation of BCG artifacts. The magnitude of these artifacts is proportional to the magnetic field of the MRI, and therefore, BCG artifacts can grow quite large in higher field magnets such as 3- or 7-Tesla machines. The exact sources of BCG artifact inside the patient bodies are not exactly known, but it is speculated to be generated by body motion caused by breathing and pulsation of the heart, as well as the flow of the blood (which is a conductive fluid) inside the vessels. Several procedures have been proposed to reduce these artifacts. For example, Gotman et al. [[26]] used sand bags to securely immobilize the EEG wires between the patient's head and the amplifier. Other methods include immobilizing the patient's head using cushions or bite bars [[27]]. Following these procedures can reduce the magnitude of the artifacts but will not eliminate them completely, and therefore, additional signal processing steps are required to eliminate these artifacts.

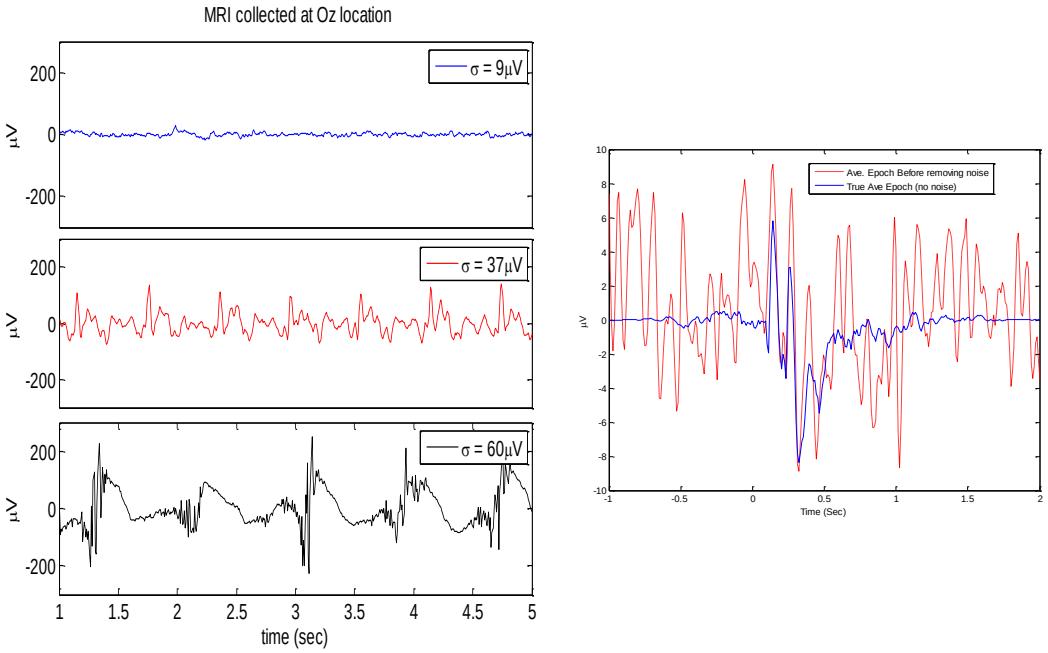


Figure 5-2: Left: BCG and GA noise affecting normal EEG. Top-Left: normal EEG recorded in a clean environment. Middle-Left: EEG recorded inside the MRI chamber. Bottom-Left: EEG recorded inside the MRI chamber while the MRI is running. Right: Averaged-ERP before (red) removal of BCG artifacts compared to a regular A-ERP obtained from a clean EEG dataset. Data shown has been obtained from an electrode located in occipital lobe (OZ).

Various signal processing methods have been proposed for removal of BCG artifacts. For example, one method is to estimate the BCG profile (the shape of the heart beat) for each channel and then to subtract this profile from each instance of a heart beat for each channel [26,28,29]. In this method, a peak finding algorithm, run on separate electrocardiogram (ECG) electrode data (placed on the patient's chest or back), finds the QRS peak in each cardiac cycle (the strong pulse at the start of the cycle), and then the time-locked data in each EEG channel is averaged to create a profile for the cardiac cycle. This cardiac profile is later subtracted from each cycle. This method is not, however, very effective, and its

accuracy depends on the performance of the peak detection algorithm. One problem with this method is the saturation of the ECG data by GAs during the MRI scanning periods. Because the ECG electrode lies far from the ground of the EEG loops and has much longer wire lengths it is prone to more RF interference. The blood flow and the breathing motion, in addition to the pulsation of the heart, all can create large enough voltages that might be mistaken by the algorithm to be the QRS peak. It is possible to detect the QRS peaks from the other electrodes placed on the head, but they are less likely to contain a significant QRS peak that can be clearly distinguished from other waveforms in the heart beat cycle. The following figure shows the ECG electrode waveform inside and outside of the scanner:

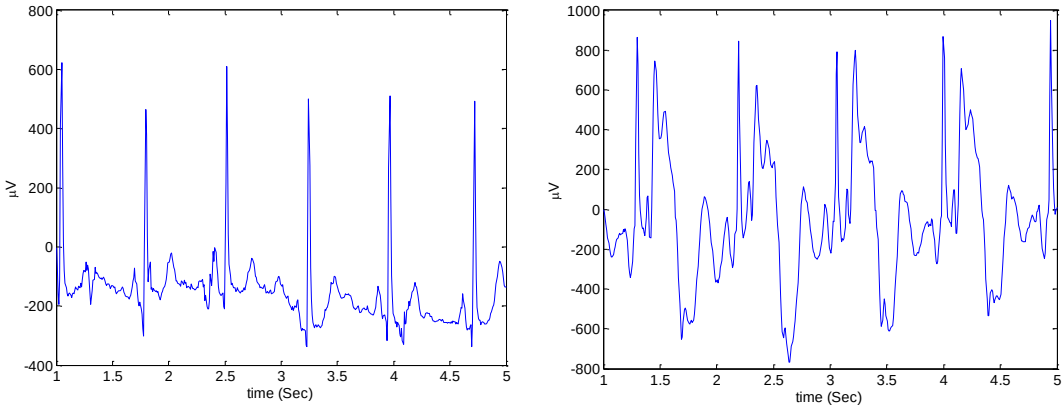


Figure 5-3: Left: ECG recorded outside the MRI. QRS peaks are clearly visible. Right: ECG recorded inside the MRI (while the scanner is inactive). Due to the presence of BCG artifacts, the ECG recorded inside MRI has a completely different shape, and detecting the QRS peaks is much more difficult.

Another problem with Average Artifact Subtraction (AAS)-related methods is

that they assume that the BCG artifacts have one unique profile. This is not entirely true, as the blood flow, heart rate and the breathing patterns change in patients over time, during the course of the scanning. Adaptive AAS methods have been proposed [29] to account for these temporal changes in the BCG artifact. Another method known as the Optimal Basis Sets [28] uses PCA to create several instances (principal components) of the BCG artifacts to account for the temporal changes. To our knowledge, this method is the most widely used artifact-removal algorithm. The success of this method again depends on successful detection of the QRS peaks and uniformity of the heart beats. Whenever the heart beats are not detected correctly, these methods fail to provide satisfactory results. Other researchers have proposed using independent component analysis (ICA) or PCA over the entire data matrix [30,31]. In these methods, the whole EEG matrix is decomposed into ICA or PCA components, and the components that correlate with the BCG artifacts are eliminated and a clean EEG signal is reconstructed using the remaining components. The problem with these methods is that they cannot create components that only isolate the noise-related or signal-related data, and in most cases, a loss of signal-to-noise ratio occurs as a result of removing some useful signals contained in noise components. In many of these methods, component selection is not automatic, and some sort of user intervention is required to identify components that correspond to noise or signal. Such correlation measures are not always available and there is always a possibility of choosing the wrong components for artifact rejection. Other researchers have also

proposed combining one of the peak detection methods with latent variable methods to provide better results [32].

5.1.6 Muscle Artifacts

Muscle artifacts in EEG recordings are mainly generated as a result of eye movement, blinking or jaw movement, as well as activation of other facial muscles. These artifacts usually have a frequency range that is similar to the components of the ERPs; therefore, simple frequency filtering will not remove them effectively. Several methods have been developed to reduce the artifacts, such as average artifact subtraction or the use of PCA, ICA, as well as the use of extra channels that record artifacts and subtract them from the rest of the EEG data [33,34]. Perhaps the most successful method in removing the artifacts is the ICA component-rejection method. However, as in the BCG case, in LVM methods it is not possible to entirely extract components that solely account for noise (muscle artifacts) and not the signal (ERP). Thus removing these components may result in deterioration of the signal-to-noise ratio. In addition, even if it were possible to use LVM methods to create components that are related to noise only, detecting such components without a proper reference would be difficult. Figure 5-4 shows the sample artifacts generated as a result of chewing gum on some arbitrarily selected EEG channels.

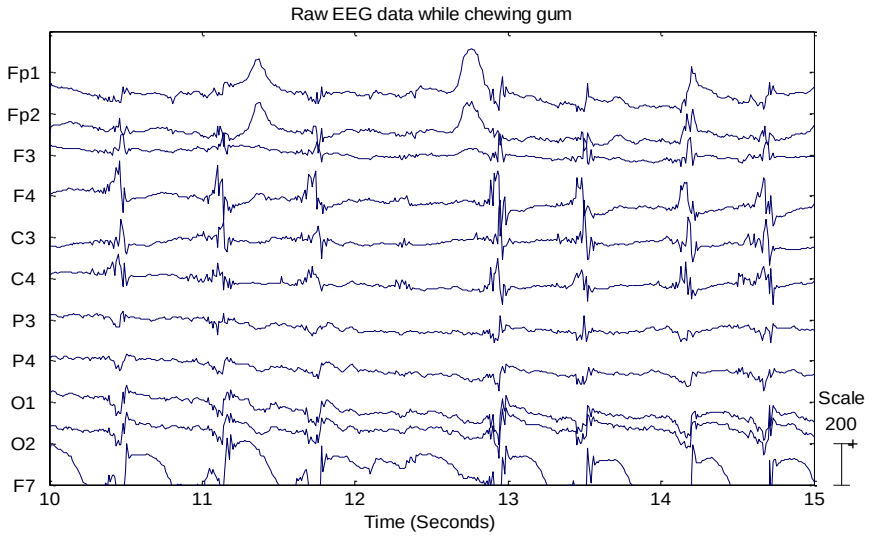


Figure 5-4: Example of muscle artifact when a patient is chewing gum during EEG recording. The EEG channels were selected arbitrarily. The FP1 and FP2 electrodes also exhibit several eye-blink induced artifacts.

5.1.7 Objectives

In this chapter, we propose the use of Soft Constrained PLS (SC-PLS) for removal of the BCG and muscle artifacts. This method exploits the additional knowledge about the noise (muscle or BCG artifacts) and the true signal (ERP) to improve the component selection and noise removal. Here we use the constrained PLS method to extract components of a basis matrix that accounts for all the variations in the EEG data other than those estimated by the ERP averages. The method first creates an initial estimate of the ERP using the noisy EEG dataset and iteratively improves the results by detecting and removing latent variable noise components from the EEG dataset. A method similar to ours was proposed by Bonmassar et al [27] which uses the estimates of the ERPs collected in a clean environment. They reduce the problem into a generalized eigenvalue problem and

extract components that maximize the covariance of the noise while minimizing the covariance of the stimuli-related data. In our proposed method, we use soft constrained PLS and its component-selection properties to iteratively build a basis for the noise that is later used to remove the noise from the EEG data and hence improve the ERP averages extracted from the updated EEG dataset. The improved ERP averages are reused in the constrained algorithm to provide even better estimates of the noise basis.

The remaining sections of this chapter are organized as follows: in Section II, we propose our algorithm and its mathematical explanation. In Section III, we define the details of the experiment used to produce datasets to test our algorithm. In Sections IV and V the results are shown, and finally, Section VI includes the conclusion and discussion of the results.

Two types of EEG noise are being investigated: BCG artifacts and muscle artifacts. In the BCG case, we compare our results to the results obtained from the OBS algorithm [28] and the reference EEG data recorded for each patient in a noise-free environment. In the OBS method the EEG data are reorganized according to the "qrs" peaks of the BCG artifact and after performing PCA on the rearranged EEG data a few principal components, which represent the BCG artifact, are projected out of the dataset and then the cleaned EEG is reorganized back to the previous arrangement. This method is very much the gold standard in BCG artifact removal methods.

In the muscle artifact case, the muscle artifacts induced by chewing gum during

data collection are removed using the SC-PLS algorithm and the results are compared to a manual ICA rejection method.

5.2 Algorithm

5.2.1 Signal Structure

In the following, we propose a method for estimation of an ERP signal in the presence of background EEG activity, muscle artifacts and BCG noise.

Briefly, the method works as follows: create an initial (rough) estimate of the ERPs (for each EEG channel) using the noisy data (or any external, clean data if available); use SC-PLS to find major variations in the data that are irrelevant of the ERP estimates; and improve the results by iteratively removing these components from the original EEG dataset. The advantage of using constrained PCA or PLS approaches for estimating the ERPs is that the information regarding the evoked responses will be retained while other variations irrelevant to the ERPs are removed.

First, the structure of an EEG dataset contaminated with noise is important to discuss. In an EEG experiment, the data recorded in the presence of noise can be decomposed into the following structure:

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z} + \mathbf{E} \tag{5-1}$$

$$\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{E} \in \mathbb{R}^{n \times ch} \tag{5-2}$$

where $\mathbf{X}$ is the raw EEG data recorded, n is the number of samples recorded (rows), and ch is the number of EEG channels (number of columns). $\mathbf{Y}$ is the actual brain-generated signals, which consist of the brain background noise as

well as the brain's response to the stimuli. $\mathbf{Z}$ is the structured external noise (such as BCG or muscle artifacts), and $\mathbf{E}$ is random noise with standard deviation σ. In many cases, it can be assumed that both $\mathbf{Y}$ and $\mathbf{Z}$ have latent structures with lower dimensionality than the number of channels recorded. Hence,

$$\mathbf{Y} = \mathbf{T_Y C_Y} + \tilde{\mathbf{E}}_\mathbf{Y} \tag{5-3}$$

$$\mathbf{Z} = \mathbf{T_Z C_Z} \tag{5-4}$$

where $\mathbf{T_Y}$ ($n \times r$) and $\mathbf{T_Z}$ ($n \times s$) are the latent vectors ($n,s < ch$), or sources of the brain waves and the external structured noise respectively, both with lower dimensionality than the actual number of channels recorded. $\tilde{\mathbf{E}}_\mathbf{Y}$ ($n \times ch$) in (5-3) is the background noise caused by background brain activity, and $\mathbf{T_Y}$ is the brain response to external, time-locked, stimulation such as the visual stimuli in the upcoming experiments. $\tilde{\mathbf{E}}_\mathbf{Y}$, which defines the additional brain activity signals, may or may not be structured. Since the event related (ER) brain response is almost within the same range of the background brain activity, event-related potentials, which we denote by (ψ), are usually averaged over several trials to provide reliable estimates known as the averaged ERPs ($\bar{\psi}$). The shape and latency of the averaged ERP (A-ERP) components provides valuable information about the brain response or the location of the stimulated sites in the brain. Hence, they are often required to be extracted for analysis and comparison reasons. In the presence of strong noise, a higher number of trials must be recorded and averaged, or the noise needs be removed prior to averaging the trials. Assuming

the presence of structured noise in the EEG data, it is possible to use methods such as Principal Component Analysis (PCA) [35] or Independent Component Analysis (ICA) to remove the components of noise ($\mathbf{T_Z}$) or to extract the low-rank components containing the ERPs, or in other words, to extract ERP components that belong to the column space of $\mathbf{T_Y}$

$$\Psi \in \mathcal{R}(\mathbf{T_Y}). \tag{5-5}$$

Because both the noise and ERPs are structured and have low rank, latent variable methods cannot distinguish between the noise and signal components, unless additional information and constraints are provided. When these methods are used, there is always a chance that some information overlap between the noise and ERP latent components will exist, resulting in reduced SNR in the averaged ERPs when the noise components are removed.

5.2.2 Formulating EEG problem as a constrained LVM method:

Since both $\mathbf{T_Z}$ and $\mathbf{T_Y}$ are low rank and reside in the subspace of $\mathbf{X}$, additional prior information about $\mathbf{Y}$ and/or $\mathbf{Z}$ can be implemented in the LVM method to decompose $\mathbf{X}$ into the noise and signal components, where the noise components are used to remove the structured noise without compromising much of the ERP's SNR. The idea is to find $\mathbf{t_i}$ ($n \times 1$) as a linear combination of $\mathbf{X}$:

$$\mathbf{t}_i = \mathbf{X}\mathbf{w}_i \tag{5-6}$$

that ideally, only belongs to the ERP subspace and contains minimal information about the noise subspace.

The soft constrained PLS method has been previously introduced, in which the data is decomposed into two latent variable sets that maximally separate noise and the signal subspaces. In these methods, a series of latent vectors are iteratively extracted by maximizing an objective function of the following form:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \left|\mathbf{w}_i^{'}\mathbf{X}^{'}\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_i - \lambda\mathbf{w}_i^{'}\mathbf{X}^{'}\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{w}_i\right| \\ \text{s.t.} \quad & \mathbf{w}_i^{'}\mathbf{w}_j = \delta_{ij} \end{aligned} \tag{5-7}$$

This optimization problem can be solved by constructing the Lagrangian as

$$L(\mathbf{w}_i) = \left|\mathbf{w}_i'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_i - \lambda\mathbf{w}_i'\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{w}_i\right| - \gamma_i(\mathbf{w}_i'\mathbf{w}_i - 1) \tag{5-8}$$

This objective function finds a linear combination of the $\mathbf{X}$ that maximizes the absolute covariance difference between $\mathbf{Y}$ and $\mathbf{Z}$. The parameter λ controls the tradeoff between covariance of $\mathbf{X}$ and $\mathbf{Y}$ and $\mathbf{Z}$. The solution to this problem ($\mathbf{w_i}$) can be found by differentiating with respect to $\mathbf{w}$ and equating it to zero, which leads to finding the dominant eigenvalues and eigenvectors associated with

$$\mathbf{U}_1 = \mathbf{X}'\left(\mathbf{Y}\mathbf{Y}' - \lambda\mathbf{Z}\mathbf{Z}'\right)\mathbf{X} \tag{5-9}$$

As discussed earlier in Chapter Two, an interesting feature of this method is that the sign of γ_i determines whether the components have a stronger correlation with $\mathbf{Y}$ or with $\mathbf{Z}$. This allows for choosing only those components that are either maximally correlated with the noise or the brain response's subspace. The

components can be extracted all at once using the eigen-decomposition algorithm, or they can be extracted one at the time using iterative procedures. Extracting all components iteratively results in principal components (**T**s) that are orthogonal to each other. Components extracted this way provide better visualization; however, the procedure will be much slower and more time consuming.

In the iterative procedure, before each step, **X** is deflated using

$$\mathbf{X} = \mathbf{X} - \mathbf{t}_i \mathbf{p}_i' \tag{5-10}$$

where **p** is a projection coefficient defined as:

$$\mathbf{p}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{X} \tag{5-11}$$

and subsequent components are extracted, using the deflated **X** in (5-7). This deflation process ensures orthogonality between principal components (**t**'s). In the eigen-decomposition method, all of the eigenvectors (**w**) are extracted at the same time without deflation. In such a case according to Rao [36] the objective function will translate into:

$$\begin{aligned} &\max_{\mathbf{w}} \sum_i \left| \mathbf{W}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{W}' - \lambda \mathbf{W}'\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}\mathbf{W}' \right| \\ &\text{s.t.} \qquad \mathbf{w}_i' \mathbf{w}_j = \delta_{ij} \end{aligned} \tag{5-12}$$

which is a more general case of the previously discussed soft constrained PLS. Again principal components can be calculated from the latent matrix **W** using:

$$\mathbf{T} = \mathbf{X}\mathbf{W}. \tag{5-13}$$

These principal components, depending on their eigenvalue sign, have strong covariance with **Y** and small covariance with **Z** (when their corresponding

eigenvalues or positive) and vice versa. However, the principal components **t**'s are no longer orthogonal to each other.

Let us assume that a matrix $\tilde{\mathbf{Y}} \in \mathbb{R}^{n \times ch}$ exists, with *n* samples and *ch* columns, that contains an initial (rough) estimate of the ERPs, obtained from averaging the noisy EEG data. The matrix is constructed by defining a matrix of zeros, the same size as **Y**. Once this matrix is constructed, at each instance and for each channel that a trigger was recorded, the zeros are replaced by an initial averaged ERP according to the following procedure:

$$\tilde{\mathbf{y}}_{ch}^{i}(t) \ = \bar{\boldsymbol{\psi}}_{ch}(t - \boldsymbol{\tau}(i)), \bar{\boldsymbol{\psi}}_{ch}(t) = 0 \,|\, \forall t < -1^{sec}, t > +2^{sec} \qquad (5\text{-}14)$$

where **τ**(i) is a vector containing the onset time for the stimuli recorded during the original experiment. Figure 5-5 shows how such a matrix can be constructed.

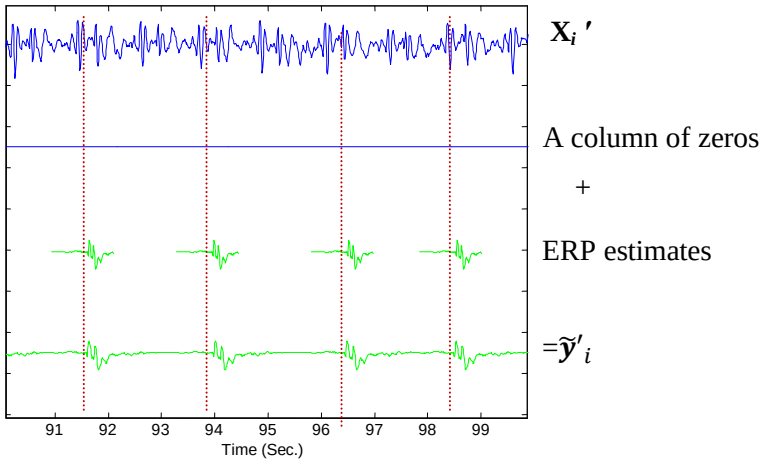


Figure 5-5: $\tilde{\boldsymbol{Y}}$ is constructed by replacing zeros with averaged ERP values for each channel from -1 seconds to +2 seconds after the trigger was recorded. This creates a vector resembling a rough estimate of the brain response to the stimuli.

Now the objective function defined in (5-12) is used to decompose the original EEG data (**X**), or any suitable basis of it, e.g. an ICA decomposition of the EEG data. denoted by $\tilde{\mathbf{X}}$, into two sets of components: those that have strong covariance with the epochs and those that have strong covariance with the noise. Having a rough estimate of the epochs in $\tilde{\mathbf{Y}}$, the noise in the EEG data can be roughly estimated as:

$$\tilde{\mathbf{Z}} = \mathbf{X} - \tilde{\mathbf{Y}} \tag{5-15}$$

Replacing X, Ỹ and Z̃ in (5-7) will result in:

$$\begin{aligned} \max_{\mathbf{W}} \quad & \sum_i \left| \mathbf{W}'\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\mathbf{W}' - \lambda \mathbf{W}'\tilde{\mathbf{X}}'\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}'\tilde{\mathbf{X}}\mathbf{W}' \right| \\ \text{s.t.} \quad & \mathbf{w}_i'\mathbf{w}_j = \delta_{ij}. \end{aligned} \tag{5-16}$$

By extracting the eigenvectors, a set of components are obtained from

$$\mathbf{T} = \tilde{\mathbf{X}}\mathbf{W} \tag{5-17}$$

that are either maximally correlated either with the noise or the ERPs.

Once a set of latent components is computed, the components correlated with noise, which are associated with negative eigenvalues, are used to deflate $\mathbf{X}$ and reconstruct $\tilde{\mathbf{Y}}$ using the new deflated EEG data ($\mathbf{X}_R$). Since $\mathbf{X}_R$ contains less noise, the estimated ERP values ($\tilde{\mathbf{Y}}$) will also contain fewer artifacts and will be closer to the true value of the ERPs. As the process iterates: the new $\tilde{\mathbf{Y}}$ can now be reused in the algorithm with the original EEG ($\mathbf{X}$) or $\tilde{\mathbf{X}}$ (e.g., the ICA

components of $\mathbf{X}$) to obtain an updated set of components that are more likely to estimate the noise in $\tilde{\mathbf{X}}$. The following flow chart shows the steps of the proposed algorithm:

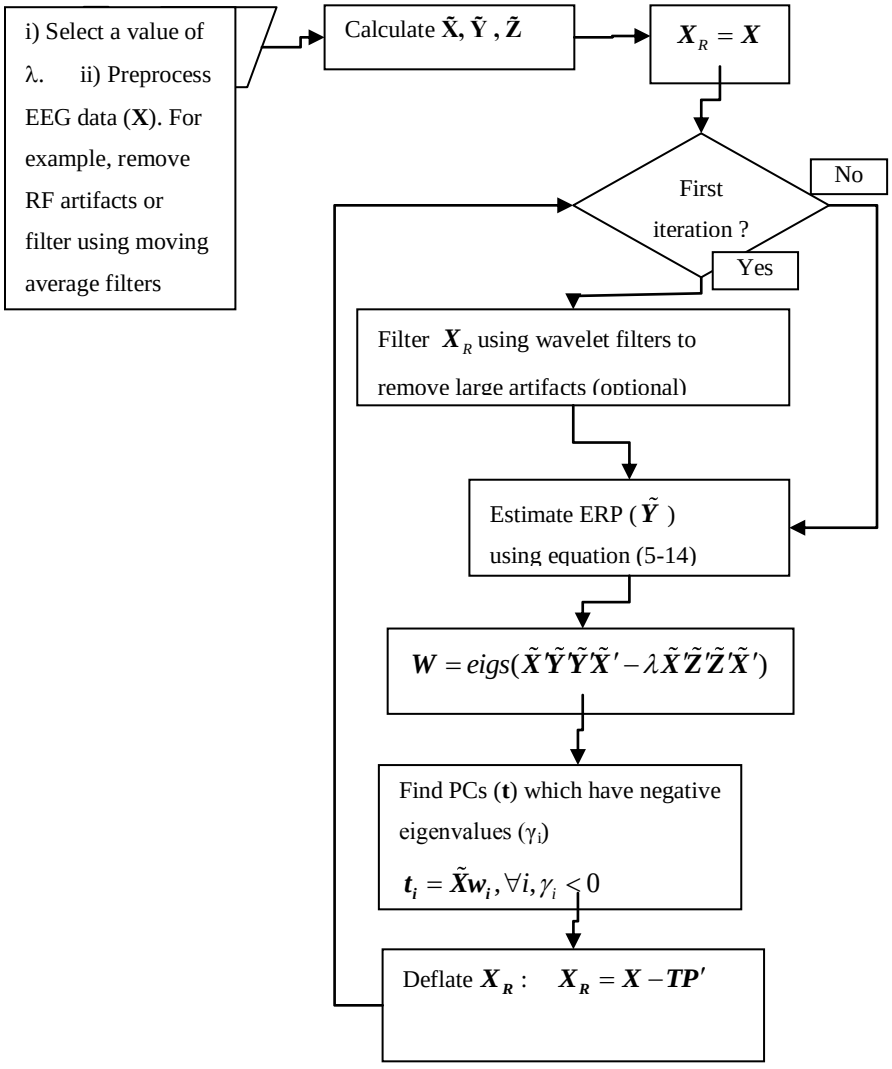


Figure 5-6: Flow diagram showing the steps of the noise removal algorithm.

The steps of the algorithm are hence as follows:

1. Preprocess the data and calculate $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Z}}$
2. Let $\mathbf{X_R} = \mathbf{X}$
3. Calculate a rough estimate of the ERP matrix ($\tilde{\mathbf{Y}}$) using the noisy EEG data ($\mathbf{X}$)
4. Calculate a rough estimate of the noise ($\tilde{\mathbf{Z}} = \mathbf{X}\text{-}\tilde{\mathbf{Y}}$ **)**
5. Use this estimate in equations (5-16) to (5-17) to calculate a basis for the noise
6. Deflate $\mathbf{X}_R$ by projecting onto the orthogonal subspace of noise as in (5-10)
7. Re-estimate the average ERP ($\tilde{\mathbf{Y}}$) from the new $\mathbf{X}_R$.
8. Repeat above step 3 to 7 to refine the estimates

If the noise components are extracted and identified correctly, and provided that a proper initial estimate of the ERPs ($\bar{\boldsymbol{\psi}}$) is available, this algorithm will eventually converge towards a set of components that capture the majority of the noise variance (e.g., BCG or muscle artifacts) and contain very little information about the ERPs. Compared to the regular PCA algorithm, the noise components extracted by the constrained method are much less likely to contain any information about the ERPs, as they have been penalized in the algorithm, and therefore, the estimated ERPs will have a much better signal-to-noise ratio.

5.2.3 Iterations and convergence:

At each step, $\mathbf{X_R}$ is refined by projecting original $\mathbf{X}$ (or $\tilde{\mathbf{X}}$) into the orthogonal complement of the noise basis. The iteration should be repeated until a stable result is obtained. However, we realized in most cases, best results are obtained after two or three iterations, and iterating too much may result in loss of SNR and lowered correlation. Throughout the upcoming experiments, at each step the SNR was estimated and the iterations were stopped once the maximum average SNR was reached.

5.2.4 Moving average (MA) filters:

For the MRI datasets, in the experiment section, the EEG datasets were preprocessed using a moving average filter. This filter, constructed from the ECG data "$\boldsymbol{\omega}$" ($n\times 1$) removes some of the BCG artifacts, which will improve the initial estimates of the ERP averages.

To construct this filter, equation (5-1) is re-written for individual EEG channel at time point t in $x_{ch}(t)$ as:

$$x_{ch}(t) = y_{ch}(t) + \sum_{r=0}^{m-1} \alpha_{ch}(r)\omega\,(t-r) + \gamma_{ch}(t) \qquad (5\text{-}18)$$

where $y_{ch}(t)$ is the noiseless signal at channel "ch" at time t and $\boldsymbol{\alpha}_{ch}$ $(m\times 1)$ is a weighting coefficient vector obtained by regressing $\mathbf{x}_{ch}$ against $\boldsymbol{\omega}$ and m is the number of lags in the MA model. $\gamma_{ch}(t)$ is the residual noise that cannot be removed using this method.. Once $\boldsymbol{\alpha}_{ch}$ is calculated, the projection of $\mathbf{X}$ into the ECG basis is subtracted from it. Experimental results indicate it is possible to

remove up to 50% of the BCG artifacts by applying this method alone. This preprocessing step can be used to remove some of the BCG artifacts before using the SC-PLS algorithm.

5.2.5 Wavelet filters

The initial estimates of the ERPs obtained by averaging the noisy data contain a significant amount of physiological artifacts such as muscle artifact or BCG noise. If the noise is strong, the initial estimates will contain large artifacts that will result in improper component selection, which may cause the algorithm to fail to converge. To provide better ERP estimates, the data in $\mathbf{X}_\mathbf{R}$ was pre-filtered at each iteration step using wavelet filters. This type of filter removes high-amplitude, impulse-shaped artifacts at different frequency bands. In wavelet filtering, first, each signal is decomposed into a set of wavelet coefficient vectors. After calculating the standard deviation of the coefficients at each level, any coefficient in the vector that has an absolute value larger than a threshold number of standard deviations is trimmed to the threshold value. Once all the coefficients have been trimmed, the filtered signal is reconstructed from the new coefficients. Figure 8 illustrates the process.

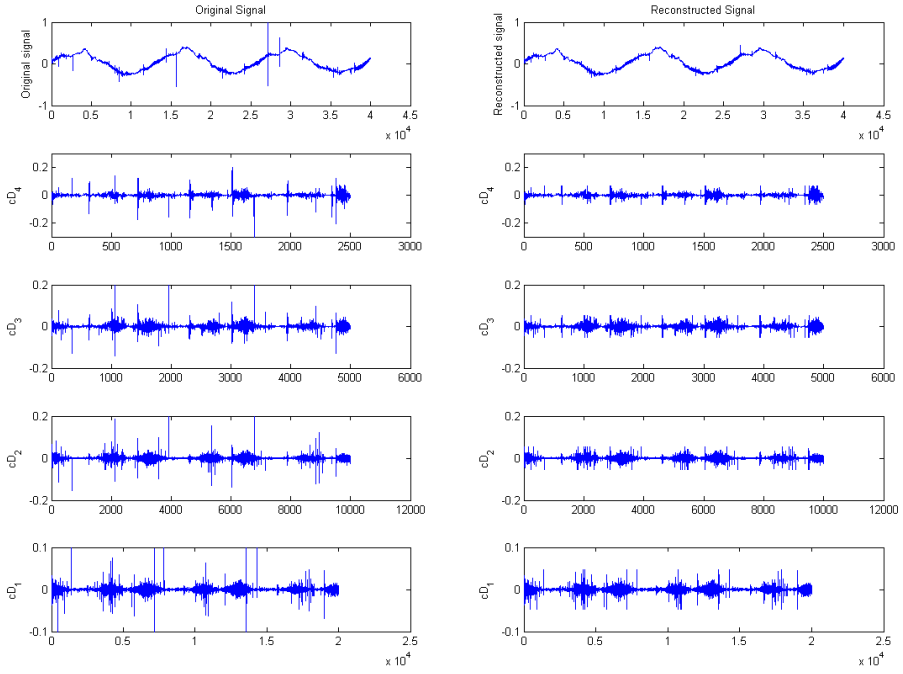


Figure 5-7: Wavelet thresholding of a signal (4 level, dB1). Original signal (top-left) is decomposed using wavelets at several stages (plots on the left below the original signal). Once the signal is decomposed at each stage (detail) the standard deviation of the signal is measured for each detail (σ) and any component having a value larger than a pre-determined threshold value (e.g. 2σ) is set to this threshold value (plots on the right below the reconstructed signal). The thresholded details are later used to reconstruct the signal. The reconstructed signal is shown in top-right

The advantage of using wavelet filtering compared to thresholding (setting large components of signal into a threshold value) is that the signal will have smooth transitions even in the thresholded segments. The thresholding only affects those elements of the wavelet coefficients that have very large deviations in that wavelet band. Therefore, normal variations in the data are much less likely to be affected by the wavelet thresholding. The following figure shows a portion of the EEG data before and after being filtered by wavelet filters:

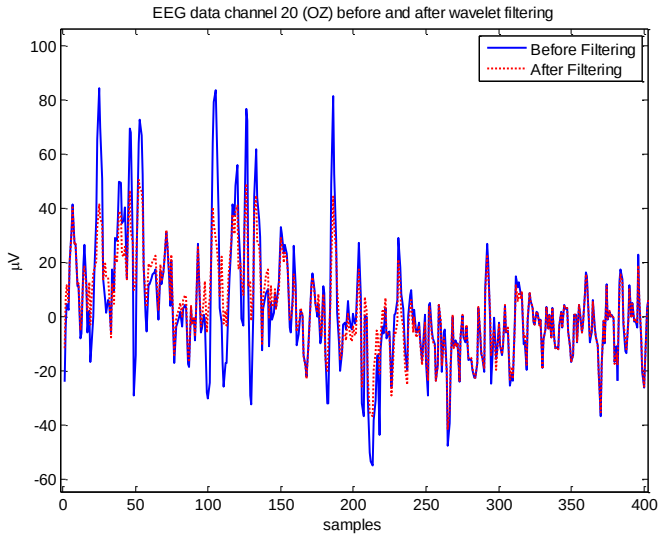


Figure 5-8: EEG data before and after filtering using wavelet filters (filter: 'bior4.4', levels: 6, threshold: 1.5 standard deviation).

5.3 **Experiments**

To test and compare the efficiency of the current noise removal algorithms against SC-PLS method proposed here, a visual evoked experiment (VEP) was designed. The EEG was recorded inside and outside of the MRI chamber, in a separate room, in various scenarios, introducing muscle or BCG artifacts. The experiments were conducted at St. Joseph's Hospital, (Hamilton, Ontario, Canada). The study was approved by the hospital's board of ethics. The EEG data was later used either directly or to create simulation datasets, which were used to test the performance of the proposed algorithms and to compare them against conventional noise removal methods. MRI studies were performed inside

a 3T scanner (General Electric), with a dedicated 8-channel head coil. The EEG system used in the study was manufactured by brain products GMBH (BrainAmp MR and BrainCap MR), consisting of an MR compatible amplifier, with 64-channel electrodes positioned on a cap according to the 10-20 standard [37]. The extra electrodes were located between the standard 10-20 channels as shown in Figure 5-9: 64 electrode locations used in the BrainCap MR EEG caps).

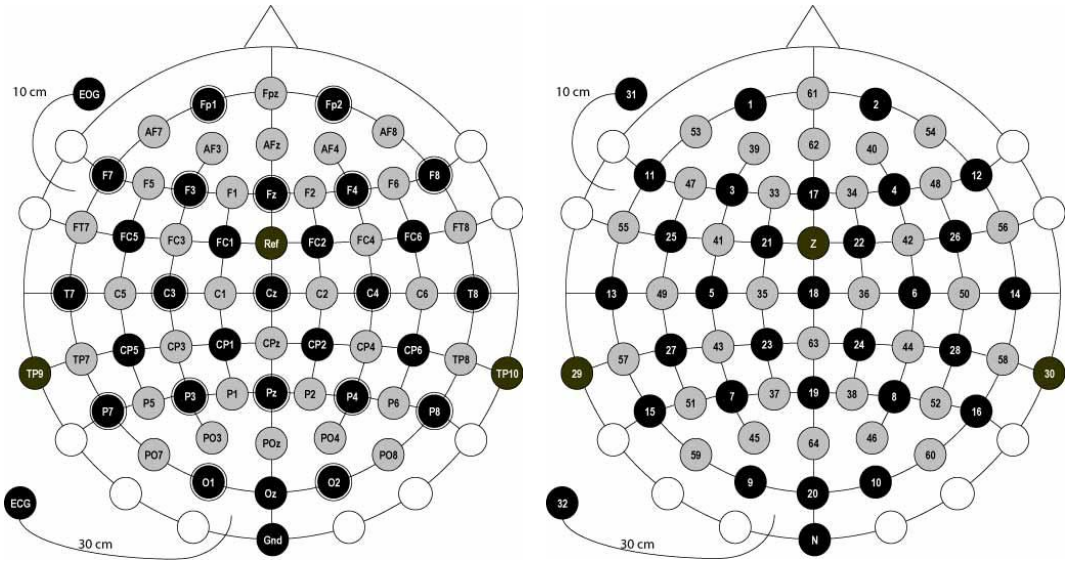


Figure 5-9: 64 electrode locations used in the BrainCap MR EEG caps. Left: 10-20 format, Right: corresponding electrode number

One of the 64 electrodes was attached to the patients' back (in mid section) to record the ECG signals, and one was attached to the patient's cheek to record the eye movements (EOG). All channels were referenced to an electrode located between channels 17, 18, 21 and 22. The EEG wires connecting the electrodes to the amplifier were sandwiched between a wooden board and pieces of memory foam strips (held by tape) to reduce and dampen the machine-induced vibrations.

The amplifier was placed in the back of the scanner 50 cm away from the scanner's bore. Twelve patients (7 male, 5 female, average age 29 yrs) participated in the study. Not all participants completed all the studies, and 3 of the MRI datasets were discarded due to bad equipment setup. All EEG data were low-pass filtered at 250 Hz and then recorded at a sampling rate of 5 kHz, later down-sampled to 100 Hz. During the VEP experiments, every time a visual trial was presented, a trigger was recorded in the EEG dataset indicating the exact temporal location at which the stimuli were presented. These triggers are denoted on the dataset with the label "S21" (shown in Figure 5-1). The experiments were carried out inside and outside of the scanner. In the experiments that were carried out inside the scanner, the patient was asked to lie down while his or her head was immobilized using memory foam cushions placed between the patient's head and the head coil.

5.3.1 Visual stimulation paradigm

A visual paradigm consisting of an alternating black and white target pattern, shown in Figure 5-10, was designed to induce visually evoked responses in the patients while keeping an iso-luminant visual field. Inside the MRI chamber, the visual stimulus was shown to the patient using an overhead projector projecting the stimuli into an oblique mirror located 6 inches away from patient's face, corresponding to a viewing angle of 90 degrees. The overhead projector was placed 5 meters away from the scanner. The head coil used in this experiment had a bar passing along the patient's nose that partially covered the patient's view. To

avoid partial viewing, the centre of the target board was shifted towards the right eye, and all patients had their left eye covered with a piece of foam, limiting the viewing to the right eye. The visual paradigm was constructed using "Presentation Software" developed by Neuro BS.

When the experiments were carried outside of the scanner, the patients were asked to sit in front of a 19-inch monitor located 1.3 meters away from them (which corresponds to a 40-degree viewing angle). The room was kept dark during the experiments.

The VEP paradigm consisted of 170 trials. During the experiments, the patients were asked to concentrate on the centre of the target board to reduce involuntary eye movements. Once the experiment was started, the target board would flicker (invert colors back and forth from black to white and vice versa) at random intervals of 1.8 to 2.2 seconds. The flicker time (time it took for the target board to invert and revert the colors) was fixed at 0.2 seconds. This type of VEP experiment produces visual responses in the brain in the occipital lobe, with a unique shape as a result of the flickering back and forth.

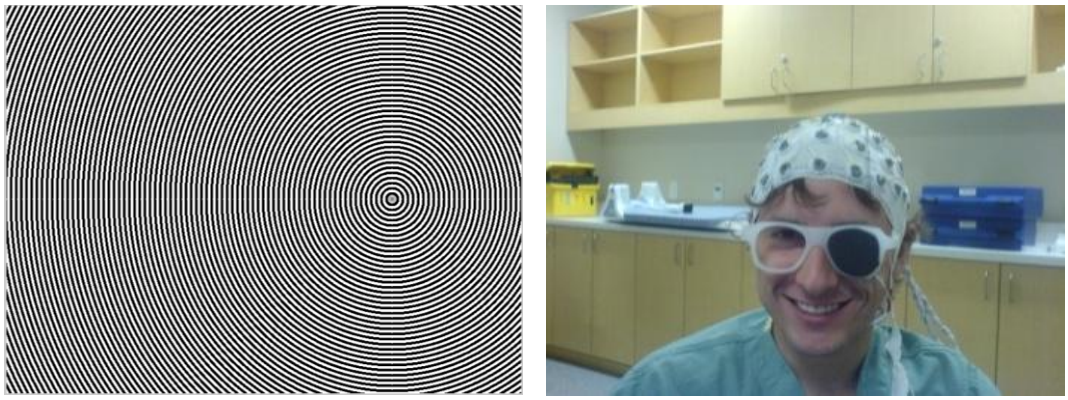


Figure 5-10: Visual paradigm, Left: the eccentric target board used for the VEP experiments. Right: one of the participants wearing the EEG cap with the left eye covered.

5.3.2 Procedure

The EEG data was recorded in several different conditions:

1. Noise-free EEG was recorded in a dark room while the patient was sitting 1.3 meters away from a computer screen watching the VEP presentation (40 degree viewing angle). This data, labeled as reference EEG, was used to obtain the reference average ERPs used for comparison between ERP obtained from the proposed algorithm (de-noised) data as well as to produce simulated EEG datasets that were used to further test each method's effectiveness.
2. EEG was recorded inside the MRI chamber while the patient watched the VEP experiment, without the MRI running, hence with no GA artifacts induced. SC-PLS and OBS algorithms were applied to this data set to recover the ERPs. These recovered ERPs were compared against the reference ERP collected in condition 1.
3. EEG was recorded inside the MRI chamber while the patient watched the VEP experiment with the MRI running, hence inducing GA artifacts. SC-

PLS and OBS algorithms were applied to this dataset to recover the ERPs from it. The quality of the recovered ERPs from each algorithm was compared to the reference ERPs collected in condition 1. The fMRI sequence consisted of a gradient echo sequence with a flip angle of 90 degree, 30 slices and TR equal to 62 msec. The gradient artifacts were removed using EEGLAB's [38] (version 10) built-in GA artifact removal toolbox. Details are provided later in the chapter.

4. EEG was recorded in a dark room while the patient was sitting 1.3 meters away from the computer screen and watching the VEP experiment. The patients were asked to chew a piece of gum during the experiment to induce muscle artifacts in the EEG dataset. SC-PLS and manual ICA rejection algorithms were applied to this dataset to clean and recover the ERPs. The quality of the recovered ERPs from each algorithm was compared to the reference ERPs collected in condition 1.
5. Background EEG data was recorded while the patients were inside the MRI. They were asked to relax and look at a bright computer screen without any stimuli being shown to the patient. These datasets were later used to create simulated EEG datasets contaminated with BCG artifacts.
6. Background EEG data was recorded while the patients were in a dark room watching a bright screen without any stimuli being shown to the patients. The patients were asked to chew a piece of gum during the EEG

data collection. These datasets were later used to create simulated EEG datasets contaminated with muscle artifacts.

Changing the environment in EEG experiments can influence the patient's vigilance, resulting in non-identical ERPs even for the same paradigm [39]. For this reason, the ERPs, extracted from EEG recorded inside the MRI are not identical to those acquired outside of MRI (Reference ERPs). Therefore, to have a better assessment of the properties of the proposed algorithms and to compare them to the other methods, simulated EEG datasets were created by adding reference ERPs to the noisy background EEG data collected in conditions 5 and 6 above. In addition to comparison between the extracted ERPs from experimental EEG datasets, the extracted ERPs from the simulated datasets, after de-noising, were also compared to the reference ERPs. Two sets of simulation datasets were created using the background EEG inside the MRI (BCG-EEG) and the background EEG recorded while the patients were chewing gum. Performing the experiments inside the scanner bore induces BCG artifacts onto the EEG data. The magnitude of the BCG artifacts depends on each patient's physiological status. Since the MRI study aims specifically at removal of the BCG artifacts, the data used for MRI simulation only consisted of the BCG artifacts and no fMRI gradient artifacts.

To create simulation datasets, the Reference ERPs were added to the background gum-EEG or the BCG-EEG background data at 100 random time

points. The minimum temporal distance between each trial was chosen to be at least 2.5 seconds to avoid any overlapping of the trials.

Overall, six dataset were obtained to study and compare the algorithms:

1. VEP experiment inside the scanner, no fMRI (VB-EEG)
2. VEP experiment inside the scanner, with fMRI (VBf-EEG)
3. Simulated VEP experiment with EEG data contaminated with BCG noise (SVB-EEG)
4. VEP experiment in a dark room while chewing gum (VG-EEG)
5. Simulated VEP experiment with the EEG data contaminated with muscle artifacts (SVG-EEG)
6. And finally, the Reference EEG (Ref-EEG) recorded in a dark room to extract the reference ERPs from

5.3.3 Quality measurement

Once the noise is removed from the EEG datasets, averaged ERPs ($\hat{\bar{\Psi}}$) are calculated from the de-noised datasets. To compare the quality of the ERPs before and after noise subtraction, several quality measures are acquired for each averaged ERP in each channel, for each experiment. In each experiment, the correlation (CRR_{ch}) between the Reference ERPs in each EEG channel and the de-noised ERPs was calculated. In addition, the standardized root mean-squared error between the Reference ERPs and the de-noised ERPs was calculated using the following formula:

$$RMSE_{ch} = \left(\frac{\sum_{n=1}^{v} \left(\left(\bar{\psi}_{ch}(n) - \frac{1}{v}\sum_{n=1}^{v} \bar{\psi}_{ch}(n) \right) - \left(\hat{\bar{\psi}}_{ch}(n) - \frac{1}{v}\sum_{n=1}^{v} \hat{\bar{\psi}}_{ch}(n) \right) \right)^2}{\sum_{n=1}^{v} \left(\bar{\psi}_{ch}(n) - \frac{1}{v}\sum_{n=1}^{v} \bar{\psi}_{ch}(n) \right)^2} \right)^{0.5}$$

(5-19)

where v=300 is the total number of points in each average ERP. $\bar{\psi}_{ch}(n)$ is the averaged Reference ERP vector from -1 (samples 1 to 100) to +2 seconds after the onset of the trigger (samples 101 to 300), and $\hat{\bar{\psi}}_{ch}(t)$ is the averaged estimated ERP for that particular channel, extracted from the noisy EEG after noise removal steps. The scalar values; $1/v \sum_{n=1}^{v} \bar{\psi}_{ch}(n)$ and $1/v \sum_{n=1}^{v} \hat{\bar{\psi}}_{ch}(n)$ are the mean values (baselines) for the original ERP and the extracted ERPs, respectively.

The averaged RMSE over all channels is defined as:

$$\overline{RMSE} = \frac{1}{62}\sum_{CH} RMSE(ch), ch \in [1,...30,32...64] \qquad (5\text{-}20)$$

Channels 31 and 32, corresponding to EOG and ECG channels, were excluded from averaging. In addition, the signal-to-noise ratio (SNR) for each channel was calculated as a logarithmic ratio of the averaged ERP variance until 2 seconds after the onset of the trigger (n = 101,…, 200) to averaged ERPs variance until one second before the onset of the trigger (n = 1, … , 100):

$$SNR_{ch} = 10\log_{10}\left(\frac{\sum_{n=101}^{200}\left(\hat{\bar{\psi}}_{ch}(n) - \frac{1}{100}\sum_{n=101}^{200}\hat{\bar{\psi}}_{ch}(n)\right)^2}{\sum_{n=1}^{100}\left(\hat{\bar{\psi}}_{ch}(n) - \frac{1}{100}\sum_{n=1}^{100}\hat{\bar{\psi}}_{ch}(n)\right)^2}\right) \qquad (5\text{-}21)$$

We assume that the averaged ERP prior to the unset of the stimulus should consist of random noise and hence have a small variance compared to after the onset of the stimulus. Therefore, a high SNR means good signal quality compared to the background noise.

The average SNR over channels near occipital lobe is calculated as:

$$\begin{aligned} &\overline{SNR_{OC}} = \frac{1}{16}\sum_{OCH} SNR(ch), \\ &OCH = [9, 20, 10, 59, 45, 64, 46, 60, 7, 37, 19, 38, 8, 23, 63, 24], \end{aligned} \qquad (5\text{-}22)$$

which are the electrodes positioned over the occipital region, at the back of the head. This is the region associated with the processing of visual stimulus.

The quantity Δvar_{ch} the variance of the difference between averaged odd and even ERP values for channel *ch*:

$$\Delta \mathrm{var}_{ch} = 1/N\sum_{N}(\hat{\bar{\psi}}_{ch,odd}(n) - \hat{\bar{\psi}}_{ch,even}(n))^2 \qquad (5\text{-}23)$$

Δvar measures the squared error in the ERP values. Ideally, the odd ERP average should be very similar to the even ERP average. Therefore, in the presence of noise, the value Δvar_{ch} from (5-23) gives an estimate of the noise variance on the respective channel. The scalar value *N* is the number of data points in the averaged ERP.

Average ΔVAR can be calculated over channels near occipital lobe as:

$$\overline{\Delta \text{var}_{OC}} = \frac{1}{16}\sum_{OCH} \Delta \text{var}(ch), \tag{5-24}$$

where "OCH" contains the 16 occipital channels defined in (5-22). Another quality indicator that was measured for each dataset after noise removal was the plus-minus ratio, originally proposed by Schimmel [40]. The positive-negative ratio (±R) represents an estimate of the SNR on the respective channel in db. It is calculated as a ratio of averaged ERP variance to the variance of the plus-minus ratio calculated as:

$$\pm R_{ch} = 20Log_{10}\frac{\left(\sum_n (\hat{\bar{\psi}}_{ch,odd}(n) + \hat{\bar{\psi}}_{ch,even}(n))^2\right)^{0.5}}{\left(\sum_n (\hat{\bar{\psi}}_{ch,odd}(n) - \hat{\bar{\psi}}_{ch,even}(n))^2\right)^{0.5}} \tag{5-25}$$

The quantity ±R is a measure of signal-to-noise power (variance) ratio, where the noise is estimated from the value Δvar_{ch}. Averaged ±R value over the channels near the occipital lobe is calculated as:

$$\overline{\pm R_{OC}} = \frac{1}{16}\sum_{OCH} \pm R(ch). \tag{5-26}$$

5.4 Results (muscle artifact)

5.4.1 Gum simulation data:

The gum simulation data for each patient was created by adding 100 occurrences of reference ERPs into the patient's background gum-EEG data (SVG-EEG) at randomly chosen intervals of 2 to 3 seconds. SC-PLS and manual

ICA rejection algorithms were used to remove the muscle artifacts from this dataset. The averaged ERPs calculated from the de-noised datasets were later compared against the original Reference ERPs to measure each method's success in removing the noise.

5.4.1.1 . Results for subject "S-S" (Simulation study)

The following figure shows the EEG signal before and after removal of the muscle artifacts in simulated EEG data for patient "S-S". The detailed results for the other participants are given later in the Appendix. It can be seen that SC-PLS has removed the majority of the variations in the dataset; however, because the algorithm is constrained by the ERP estimates, the residual signal preserves the information about the ERP values. The signal on the right-hand side of the traces are processed to extract the ERP components.

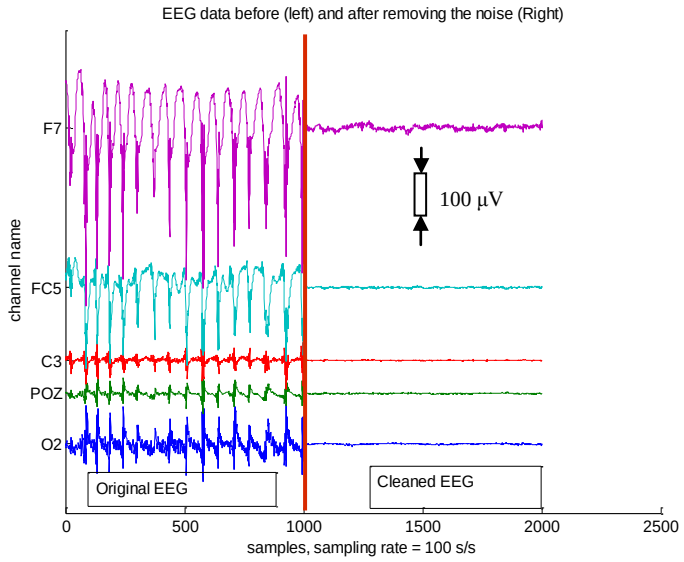


Figure 5-11: EEG signal before and after removal of noise for some of the channels. Signal on the left is the EEG data prior to removal of the noise; signal on the right of the screen represents the same portion of the EEG data after removal of the noise (subject S-S, simulation data)

Figure 5-12 shows the topographical maps of the epochs at a latency of 150 ms, corresponding to the P150 component of the ERPs. The figure on the left shows the topographical map of this component obtained from the Reference EEG data. The middle map shows the P150 component obtained from the noisy dataset prior to noise removal, and the topographical map on the right represents the ERP's P150 component obtained from the EEG dataset that was de-noised using SC-PLS algorithm. It is evident that removing the muscle artifacts allows visualization of the P150 component in the topographical maps.

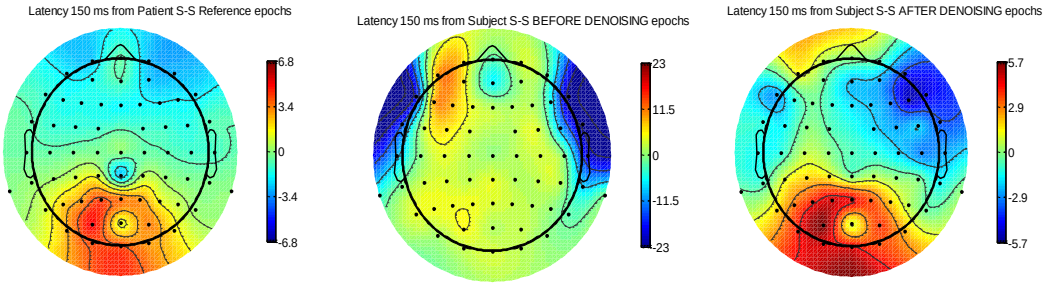


Figure 5-12: Topographical maps of the P150 component; Left: Reference EEG data. Middle: P150 component obtained from the noisy EEG dataset prior to noise removal. Right: P150 component map from the de-noised EEG dataset, using SC-PLS algorithm (subject S-S, simulation data). Note scale differences

Figure 5-13 shows the quality measurements, described earlier in equations (5-21) to (5-26), at each iteration of the SC-PLS algorithm. Figure 5-13-Left shows the average RMSE from (5-20), in dB, between the de-noised and the Reference ERPs. Figure 5-13-Right shows the averaged correlation between the Reference-averaged ERPs and the averaged ERPs obtained after noise removal using SC-PLS at each iteration. In all of the plots, the blue-square line shows the averaged statistical measures near the occipital lobe and the red-cross line shows the statistical parameters averaged over all EEG channels. These plots show that the de-noising process considerably enhances the ERP quality in the occipital channels. Since the source of the evoked responses is concentrated in this region of the brain, the electrodes standing farther from this region do not actually have any significant correlation with the visual stimuli, resulting in very small signal-to-noise ratio in the ERPs collected from these farther channels.

The plots in Figure 5-14 show the signal-to-noise ratio and ΔVAR calculated

using equations (5-23) to (5-26). These measures are independent of the Reference signal and can be used to estimate the convergence of the algorithm. For this patient for example, these figures suggest that the algorithm should be stopped after 4 iterations.

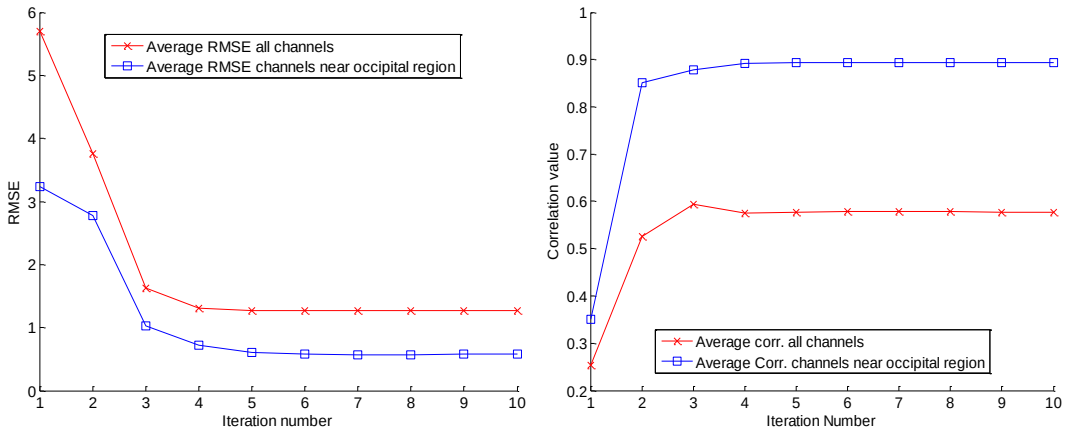


Figure 5-13: left: Averaged root mean-squared error between the Reference ERPs and the ERP estimates de-noised at each iteration step. Right: averaged correlation coefficient between the Reference ERPs and the ERP estimates de-noised at each iteration step. Crossed-Red plots show the average statistics over all EEG channels, and blue-square curves show the statistical values averaged over the channels near the occipital lobe (subject S-S, simulation data);

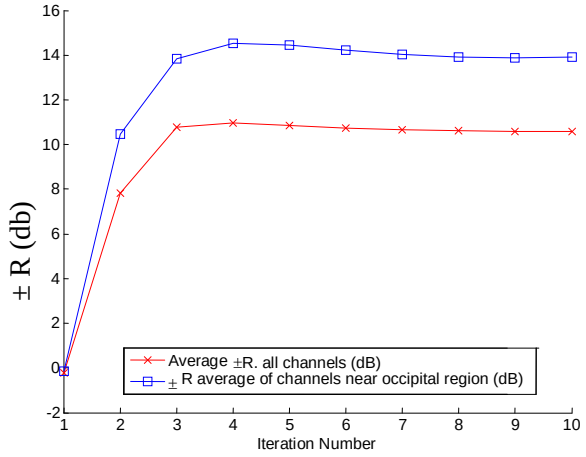


Figure 5-14: Left, Averaged ±R for the de-noised ERPs at each iteration step. Crossed-red plots show the average statistics over all EEG channels, and blue-square curves show the statistical values averaged over the channels near the occipital lobe (patient S-S, muscle simulation study)

The plots in Figure 5-15 show the averaged ERP in some channels before (dashed line) and after the noise removal process (dotted red line) compared to the averaged Reference ERPs shown in thick-solid blue line. Each plot shows the correlation, RMSE, ΔVAR and ±R values of the de-noised ERPs compared to the reference ERP in each channel shown.

Overall the ERPs from de-noised signal should have higher correlation and ±R, while having lower RMSE and Δvar. The results show that the SC-PLS algorithm is very effective in removing the noise while retaining the ERP information. However, in channels that have relatively small ERP amplitudes, the algorithm can induce artifacts not present in the original average. An example of such an induced artifact can be seen in channel 12 (F8). The same artifacts are also present when the ICA rejection method is used to de-noise the data, suggesting that the reason is somehow related to ICA decomposition and not because of the

SC-PLS algorithm. These similarities between ICA and SC-PLS results can also be seen in the plots shown in Figure 5-17.

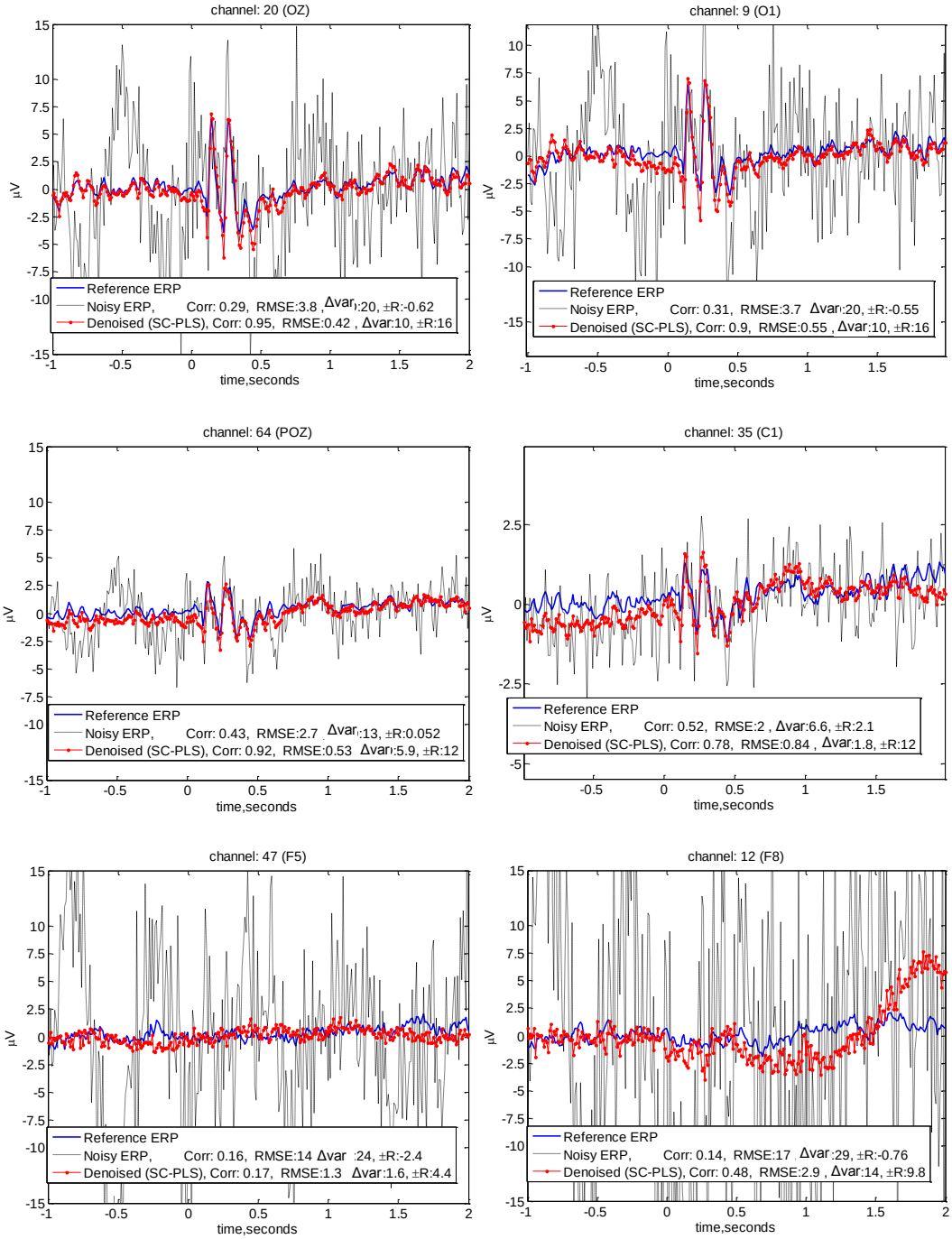


Figure 5-15: Averaged ERPs before and after removing noise from the EEG data, in channels 9, 10,64,35,47 and 11. (Subject S-S, simulation EEG)

The results obtained from the SC-PLS algorithm were also compared to those obtained from ICA component rejection. For ICA component rejection, 20 ICA components were rejected using visual inspection of the component maps (Topoplots; EEGLAB). The component maps and the rejected components, highlighted in red, are shown in Figure 5-16. Dominant ICA components with strong positive or negative amplitude in the temporal regions were removed. These components are most likely to be associated with jaw muscles.

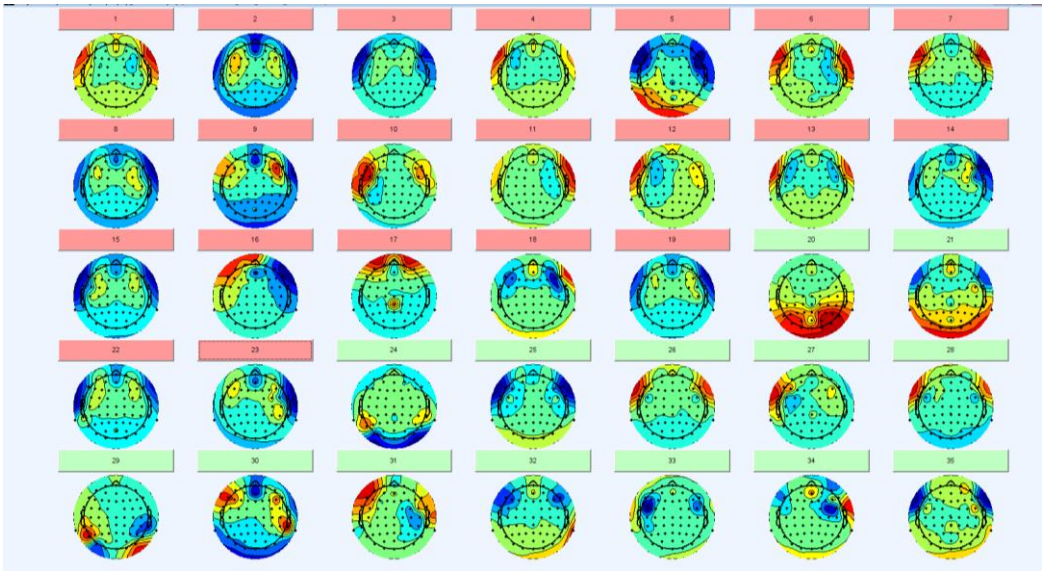
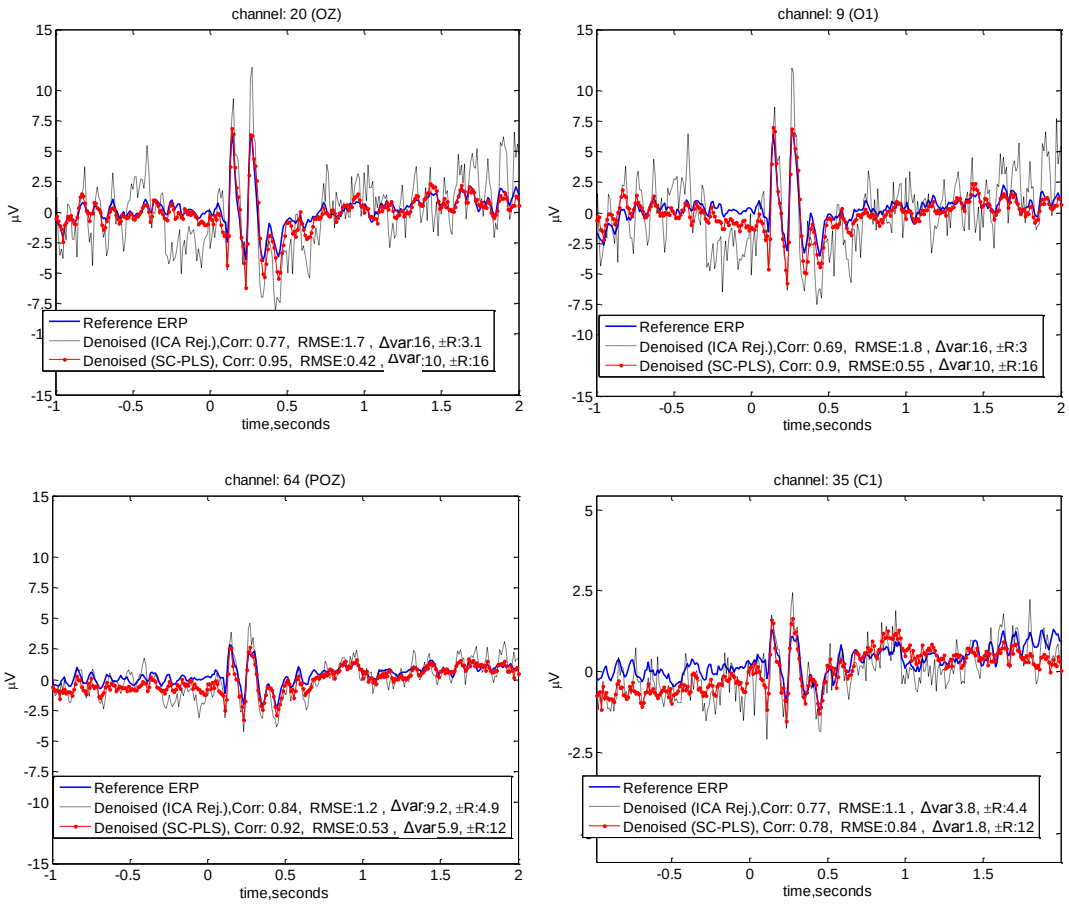


Figure 5-16: ICA component maps for the first 34 ICA components (sorted by RMS power). The ICA components highlighted in red boxes were rejected. (Subject S-S, simulation study)

The following plots show the averaged ERP obtained from the dataset after de-noising with ICA rejection superimposed with those obtained from de-noising with the SC-PLS method. Rejecting ICA components by inspection is a time-consuming task, and the results depend on the inspector's experience. In addition,

component rejection requires a reference for comparison; for example, in this case, the muscle artifacts are very likely to be produced by the temporal muscles, which will dominantly affect the temporal electrodes. There are many other cases for which such information is not available and component selection cannot be done efficiently by visual inspection. In these cases the advantage of SC-PLS is that the additional information about the ERPs is incorporated into the algorithm itself in the removal of irrelevant components. In the SC-PLS case, the components are extracted according to their statistical relevance to the ERPs.



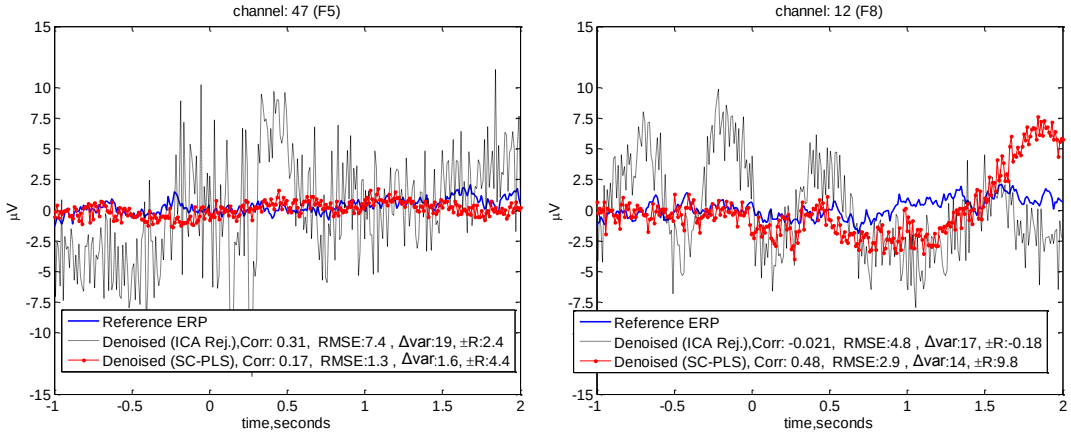


Figure 5-17: Averaged ERPs for some of the channels. Blue solid curve: Reference-averaged ERP. Dashed-black curve: ERPs extracted from the EEG data de-noised by manual ICA rejection. Dotted-red curve: Averaged ERP obtained from EEG dataset de-noised using SC-PLS algorithm.

5.4.1.2 Subject S-S (Experimental data)

SC-PLS and manual ICA rejection algorithms were applied to the experimental EEG data collected from the same participant. As mentioned earlier, experimental EEGs were collected by recording an EEG while the subjects were watching the VEP displays. The muscle artifacts were induced by asking the patient to chew a piece of gum during data collection. The ERPs obtained from experimental results were compared to the Reference ERPs. Because of the changes in the environmental factors, (i.e., chewing gum while watching the experiment), the Reference ERPs will not be quite the same as the experimental ERPs, but will be very similar. Therefore, they were used to compare the effectiveness of each de-noising algorithm. The plots in the following figure show the averaged correlation, RMSE and ±R of the algorithm at each iteration step:

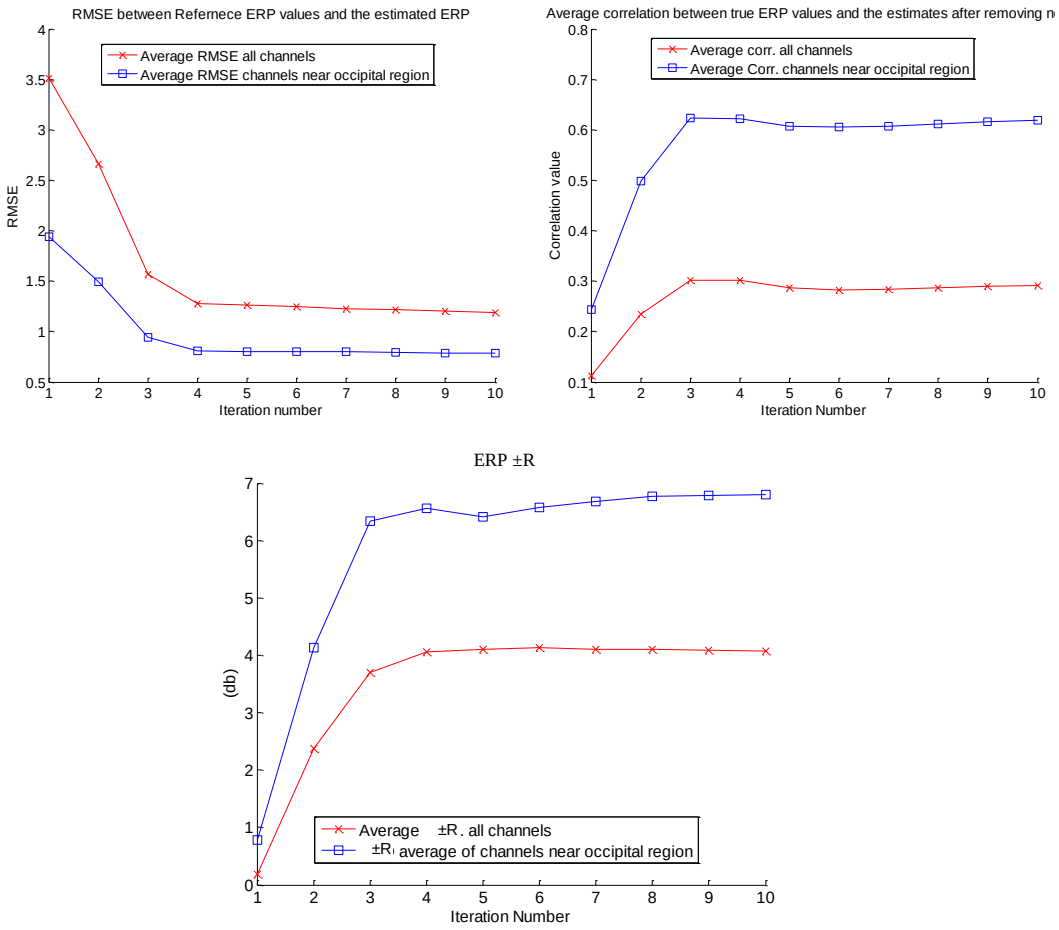


Figure 5-18 , Top-Left: RMSE between Reference ERPs and the ERP averages extracted from de-noised signal at each iteration step (SC-PLS). Top-Right: correlation coefficient between the averaged Reference ERPs and the ERPs extracted from de-noised signal at each step. Bottom Left and Right: average ±R of the de-noised signals at each iteration step (Subject S-S, actual EEG)

Similarly to the simulation study, the noisy experimental EEGs were cleaned using SC-PLS and manual ICA rejection. Figure 5-19 shows the topo-plots for the first 35 dominant components for the experimental data obtained from subject "S-S". For the ICA component rejection algorithm, 21 dominant components having strong temporal presence were visually identified and rejected from the EEG dataset. The rejected components are highlighted by red boxes.

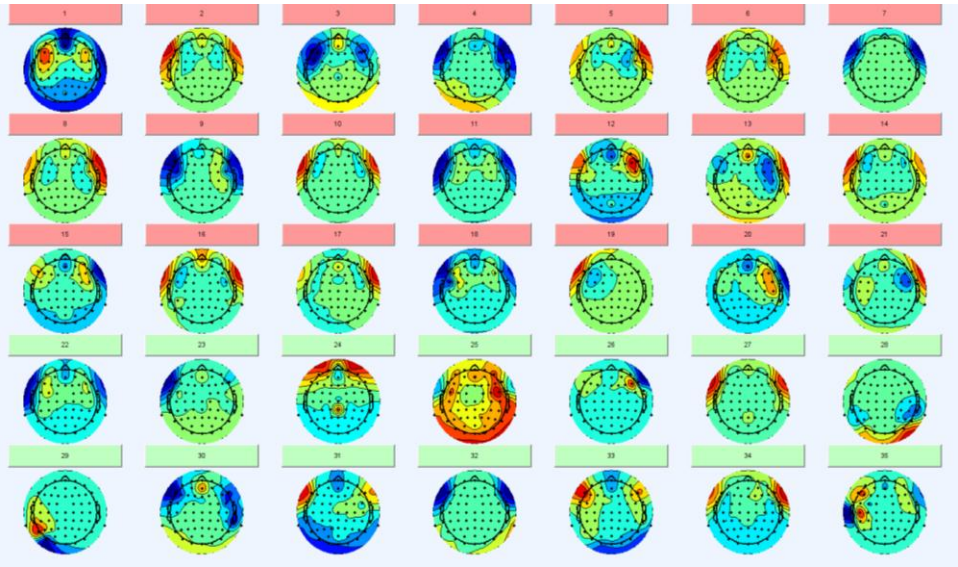


Figure 5-19: ICA component maps for the first 34 ICA components of the noisy EEG data, sorted by RMS power. The ICA components highlighted in red boxes were rejected (Subject S-S, experimental EEG).

The following figures show the ERP averages for some of the channels after removal of the noise using SC-PLS and manual ICA rejection algorithms. The ERP averages shown in solid blue represent the Reference-averaged ERPs, dashed black lines show the ERPs after cleaning the EEG using ICA rejection, and the dotted-red line shows the ERPs obtained by cleaning the EEG using SC-PLS. These results show that both methods obtain satisfactory results. However, the advantage of SC-PLS is that it is an automated procedure that requires minimal manual intervention. This method can perform as well as the manual ICA rejection and, in most cases, even better, as the components extracted are not just deleted ICA components but mixtures of various ICA components to achieve optimal results.

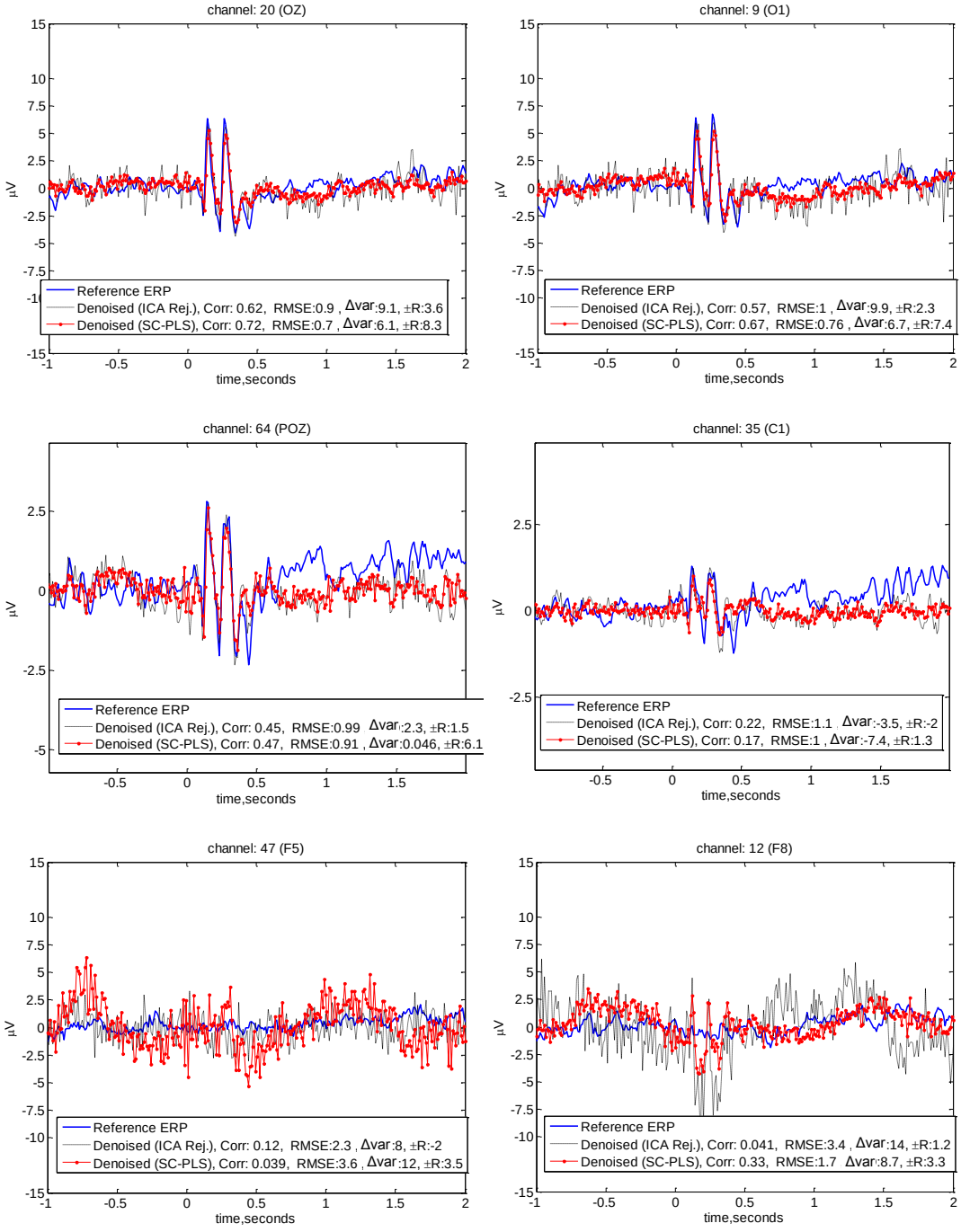


Figure 5-20: Averaged ERPs before and after removing noise from the experimental EEG data in channels 9,10,64,35,47 and 11. Dashed-black curve: manual ICA rejection. Solid blue curves: Reference ERP averages. Dotted-red curves: SC-PLS method. (Subject S-S, Experimental EEG)

5.4.1.3 Overall results for the muscle artifact study

The SC-PLS algorithm was applied to simulation and experimental data obtained from 6 participants. The number of iterations was determined by observing the ±R plots. The iterations were stopped at the first maximum peak in the ±R plot. The following charts show the changes in correlation, RMSE, ±R and SNR of the averaged ERPs after removing muscle artifacts from each simulation dataset. The quality measures were first averaged over either all EEG channels or the channels near the occipital lobe, and then the averaged statistics from the noisy ERPs were subtracted from the averaged statistics of the de-noised ERP statistics.

The quality statistics were measured for both SC-PLS as well as the ICA component rejection algorithms. The results shown for correlation, ±R and SNR are obtained by averaging the values over the channels near the occipital lobe. The RMSE values are obtained by averaging over all EEG channels. The simulation results show that, compared to ICA method, 4 of 6 subjects were found to have achieved much better correlation with the reference ERPs when SC-PLS was applied. Overall, the SC-PLS algorithm performed better than did manual ICA rejection in the simulation study.

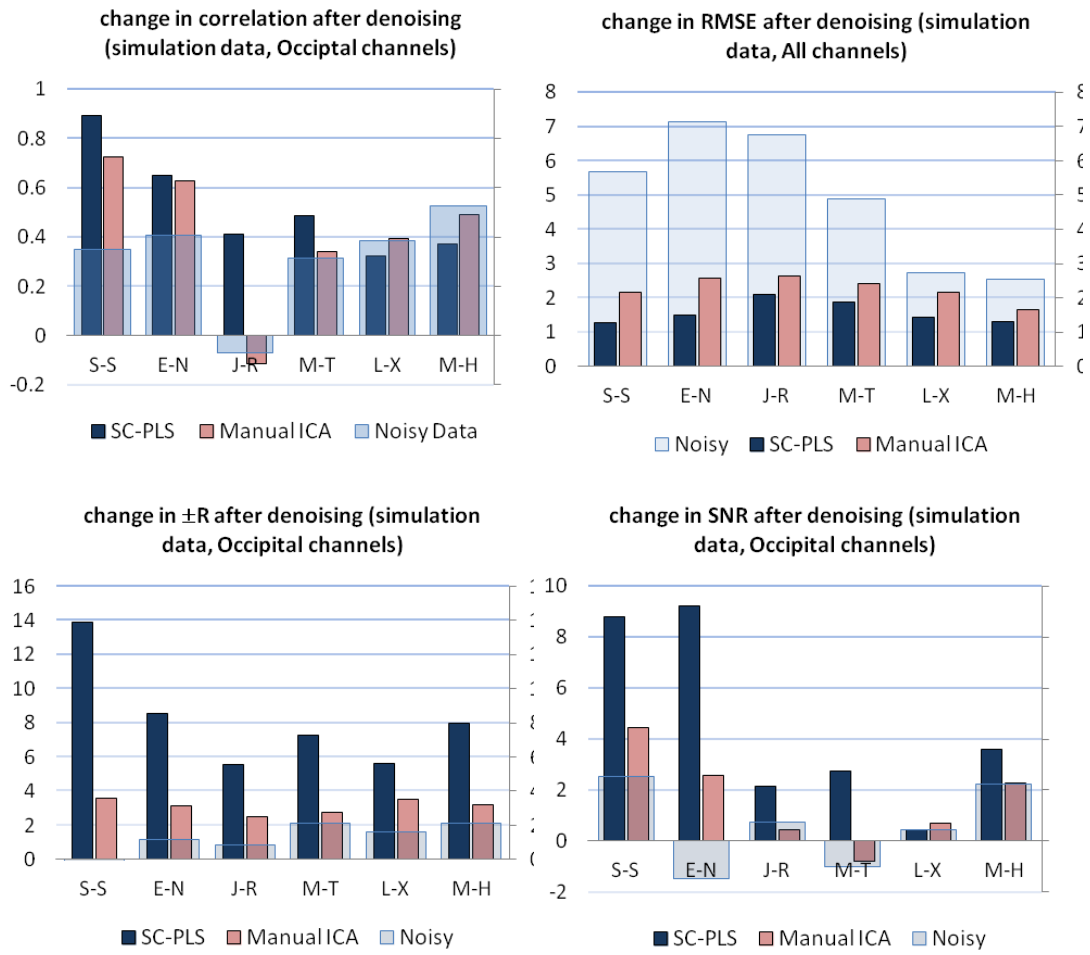
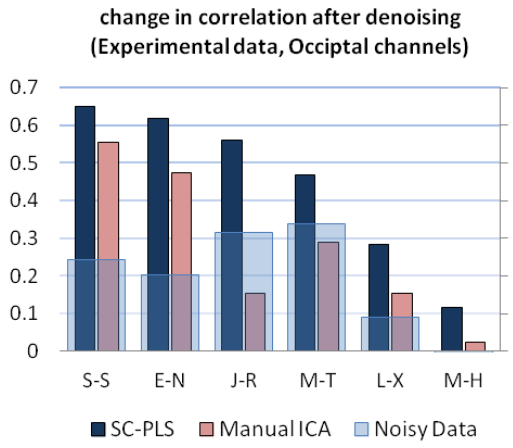
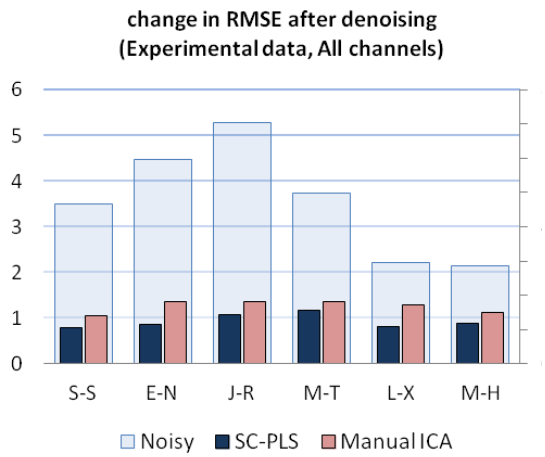


Figure 5-21: Changes in statistical values (De-noised – Noisy values) averaged over EEG channels near the occipital lobe (RMSE averaged over all channels). Top-Left; Average correlation of OCC channels, Top-Right; Average RMSE over all channels, Bottom-Left Average ±R (dB) -OCC Channels, Bottom-Right; SNR (dB), OCC-channels. (Results obtained from the Simulation Gum datasets)

Similarly to the simulation study, in the experimental EEGS, the average values for correlation, SNR and ±R over all electrodes near the occipital lobe and the average RMSE over all EEG electrodes, from the de-noised experimental EEG datasets for each participant was measured. The following charts show the changes in these parameters for each participant. The results show that for experimental data, like the simulation data, SC-PLS outperforms manual ICA

rejection across all subjects. However, in the experimental cases, the overall improvement for both ICA component rejection and SC-PLS are lower than the results achieved in the simulation data. Of course this is a expected outcome, as in the simulation study the dataset is obtained under a controlled situation, and the ERP waveform averages added to the dataset are all identical and do not change with time. In the experimental study, the ERP values are different at different time points, and hence, the experimental data shows poorer performance compared to the simulation study results.





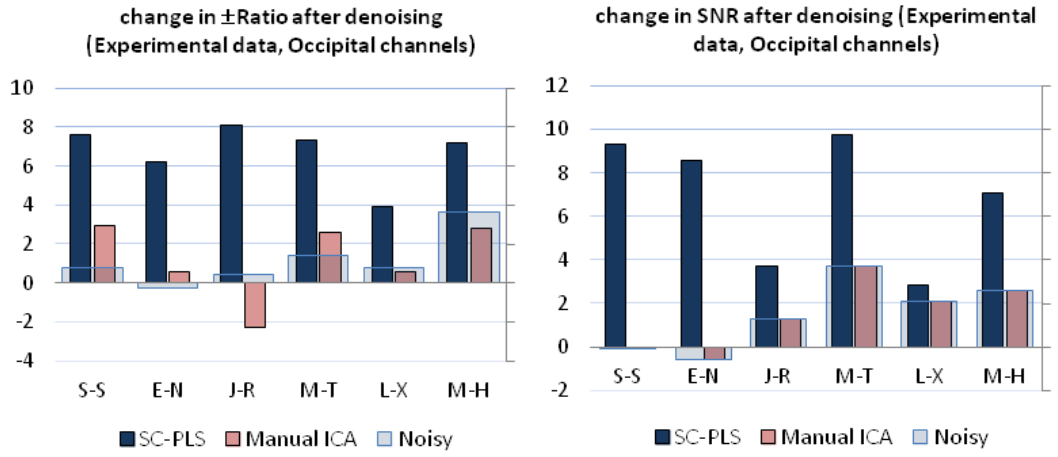


Figure 5-22: Changes in quality measures (De-noised – Noisy values) averaged over EEG channels near the occipital lobe (RMSE averaged over all channels). Top-Left; Average correlation of OCC channels, Top-Right; Average RMSE over all channels, Bottom-Left Average ±R (dB)-OCC channels, Bottom-Right; SNR (dB), OCC-channels. (Results obtained from the Experimental Gum datasets)

In both simulation and experimental studies, the results show superiority of the SC-PLS to manual component selection. Visual inspection required by the ICA method requires a great deal of experience and proper references for the artifact component maps. In addition, the outputs will vary based on the inspector's experience and the process is very time consuming. The SC-PLS algorithm does not reject components but rather finds linear combinations of them that explain the variations in the dataset, and at the same time minimizes their correlations with the ERPs. The new components are extracted based on their statistical relevance to the variation in the dataset as well as to their correlation with the ERP values. Another advantage of the SC-PLS is that it does not require any knowledge about the source of the artifacts, but rather uses the available information about the ERPs to constrain the component selection. However we

realized that it is more likely with SC-PLS to have additional artifacts added to the signal if the initial guess is far from the true values. This problem became more prevalent in the MRI study as the frequency content of the artifacts and the EEG signal overlap more closely. These problems will be discussed in further detail in the next section

5.5 **Results (MRI study)**

As in the previous study, a simulation dataset using the background EEG was recorded inside the MRI chamber. Reference ERPs collected outside were constructed for 6 of the participants. Averaged ERPs were randomly added 100 times to the background EEG at randomly chosen 2 to 3 seconds intervals. Both the OBS and SC-PLS algorithm were used to remove the BCG artifacts in the simulation datasets, and the results from each algorithm are compared against each other.

Since the objective of this study is the removal of BCG artifacts, the background EEG recorded for simulation was recorded inside the magnet but without the fMRI running. Therefore no Gradient artifacts were present and most of the artifact present was BCG noise. In the experimental studies, the EEG was recorded in two modes: without fMRI running and during fMRI image acquisition. Again in the first set, only BCG artifacts contaminate the experimental data, while in the latter case, both BCG and GA artifacts contaminate the EEG. In the datasets contaminated with GA artifacts, RF noise was removed using the built-in Gradient artifact removal tool provided by

EEGLAB package, for which the following parameters were used: up-sampling: 10 folds; slice triggers chosen and low pass filtering of 50 Hz after noise removal. In the OBS toolbox provided by the EEGLAB software, the OBS algorithm was run several times using 3 to 6 principal components, and the one producing the best results was chosen for comparison against SC-PLS method.

5.5.1.1 Simulation Study

This section presents detailed results for subject "E-N". the overall results and the detailed results for the other subjects are given in the latter sections. The overall results for the rest of the subjects for both the simulation and experimental studies are shown in the next section. Both the SC-PLS and OBS algorithms were applied to the simulation EEG data. The SC-PLS was applied iteratively until the SNR values reached their first maximum peak. Figure 5-23-Left shows the EEG data before and after removal of the BCG artifacts in subject "E-N". The amplitude of the BCG artifacts is almost an order of magnitude larger than the EEG data, thus can seriously affect the ERP shapes. Figure 5-23-Right shows the ERPs calculated from the contaminated EEG data for this participant prior to de-noising and are compared to the reference ERPs. It can be seen that unless the BCG artifacts are removed, the ERP values are hardly distinguishable from the noise artifacts.

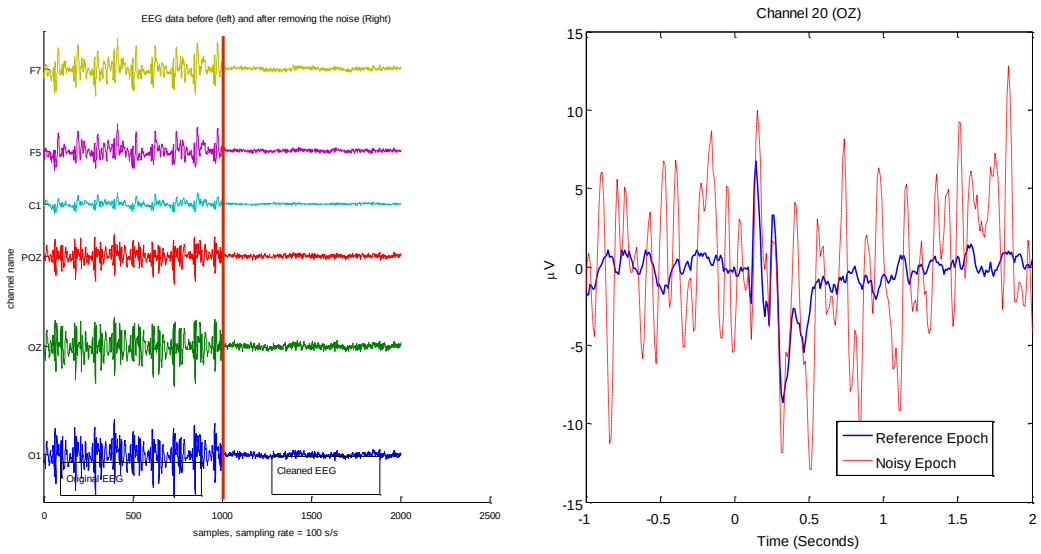


Figure 5-23: Left: EEG signal before and after removal of BCG artifacts for some of the channels. Signal on the left is the EEG data prior to removal of the noise, signal on the right of the screen represents the same portion of the EEG data after de-noising. Right: comparison of the averaged ERP in channel 20 before removal of the BCG artifacts against the reference ERP recorded outside the magnet in the same channel. (Subject E-N, simulation MRI study)

The next figure shows the averaged RMSE and overall correlation and the averaged correlation over the occipital lobe channels. Since the VEPs are concentrated near the occipital lobe, there is a higher signal-to-noise ratio and correlation between ERPs in that region with the Reference ERPs.

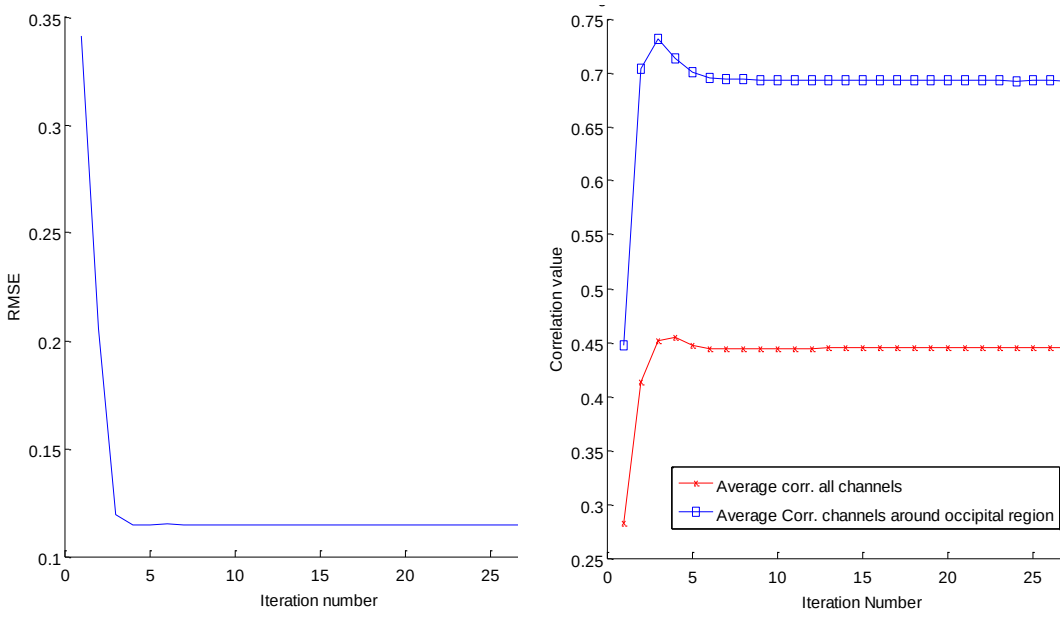
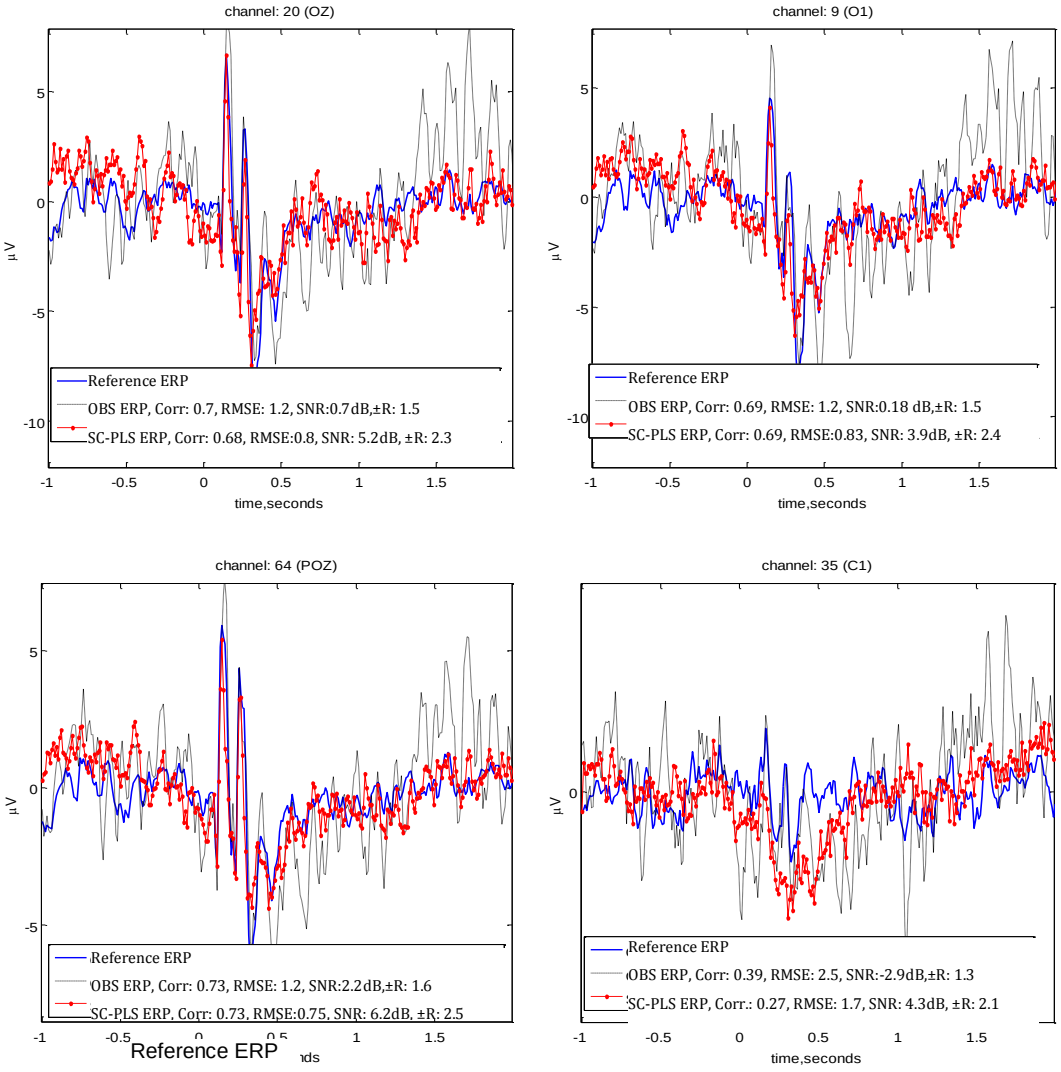


Figure 5-24: Average root mean-squared error between the Reference ERPs and the de-noised ERPs at each iteration step. Right: correlation value between the ERPs obtained after de-noising at each iteration step and the Reference ERPs. Correlation curves in crossed-red curve are averaged over all EEG channels, and the curves shown in square-blue are averaged over the channels near to occipital lob (Subject E-N, simulation MRI study)

The following figures show the averaged ERP values obtained from experimental data after removing the noise using SC-PLS and OBS algorithms. Five principal components were chosen to be removed in the OBS algorithm. Both methods show nearly the same overall correlation; however, the SC-PLS results have higher signal-to-noise ratio and lower RMSE values.


channel: 20 (OZ)
Reference ERP
OBS ERP, Corr: 0.7, RMSE: 1.2, SNR:0.7 dB,±R: 1.5
SC-PLS ERP, Corr: 0.68, RMSE:0.8, SNR: 5.2dB, ±R: 2.3
μ V
time,seconds
channel: 9 (O1)
Reference ERP
OBS ERP, Corr: 0.69, RMSE: 1.2, SNR:0.18 dB,±R: 1.5
SC-PLS ERP, Corr: 0.69, RMSE:0.83, SNR: 3.9dB, ±R: 2.4
μ V
time,seconds
channel: 64 (POZ)
Reference ERP
OBS ERP, Corr: 0.73, RMSE: 1.2, SNR:2.2dB,±R: 1.6
SC-PLS ERP, Corr: 0.73, RMSE:0.75, SNR: 6.2dB, ±R: 2.5
μ V
Reference ERP
channel: 35 (C1)
Reference ERP
OBS ERP, Corr: 0.39, RMSE: 2.5, SNR:-2.9dB,±R: 1.3
SC-PLS ERP, Corr.: 0.27, RMSE: 1.7, SNR: 4.3dB, ±R: 2.1
μ V
time,seconds

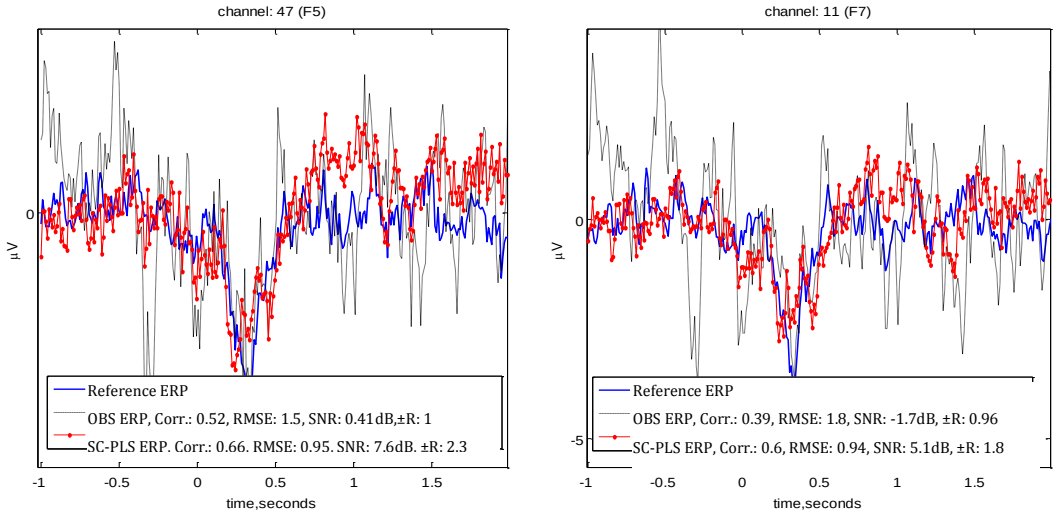


Figure 5-25: Averaged ERP values after removing noise using SC-PLS and OBS algorithm, compared to the reference ERPs shown in solid blue curve. (subject E-N, simulation MRI study)

5.5.1.2 Subject "E-N", Experimental results

Both the OBS and SC-PLS algorithms were applied to the experimental EEG data acquired inside the MRI while the patient "E-N' was watching the VEP presentation. The following results were obtained from the EEG datasets recorded without MRI running during the experiments. The extracted ERP averages after the noise removal process were compared to the Reference ERPs extracted from noise-free EEG for the same patient. Again, as mentioned earlier, since the Reference ERPs are collected in a different environment than the MRI chamber, they are not exactly the same as the noise-free epochs produced by the brain inside the MRI. However, since there is no access to the true brain response when the patients are inside the MRI, these Reference ERP averages are used for comparison of the results obtained from cleaning the EEG datasets. Figure 5-26 shows the average RMSE and correlation at each step of the SC-PLS algorithm. It

can be seen that in this particular dataset, very good results are obtained after three or four iteration steps. Figure 5-27 shows the averaged ERP values obtained from the experimental dataset after removing the noise using OBS and SC-PLS methods. The results are compared to the Reference ERP values for each channel. The graphs show that both the OBS and SC-PLS algorithms produce agreeable results. In both cases, the improvement in correlation value is the same; however, the signal-to-noise ratio and RMSE are slightly improved when the SC-PLS method is implemented.

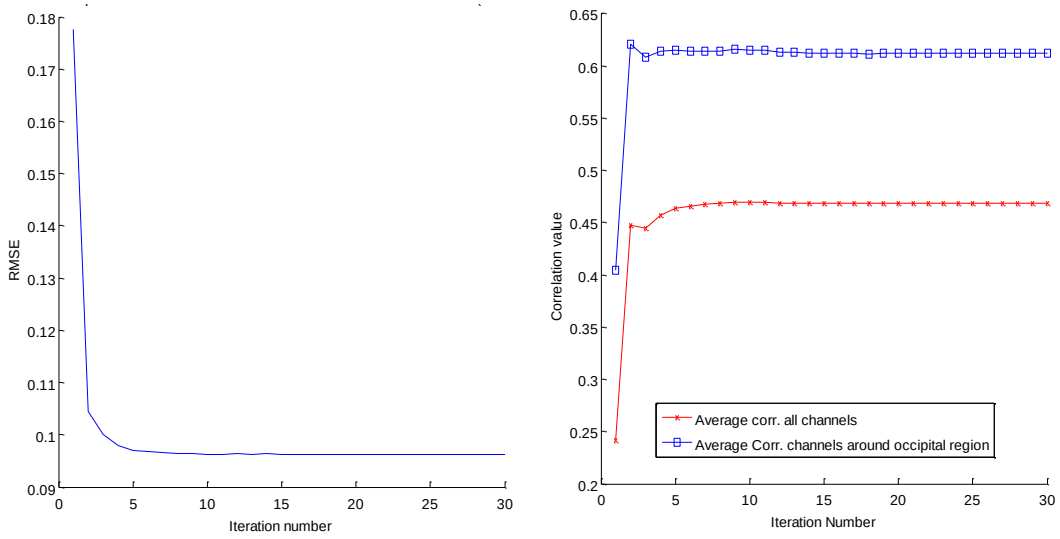
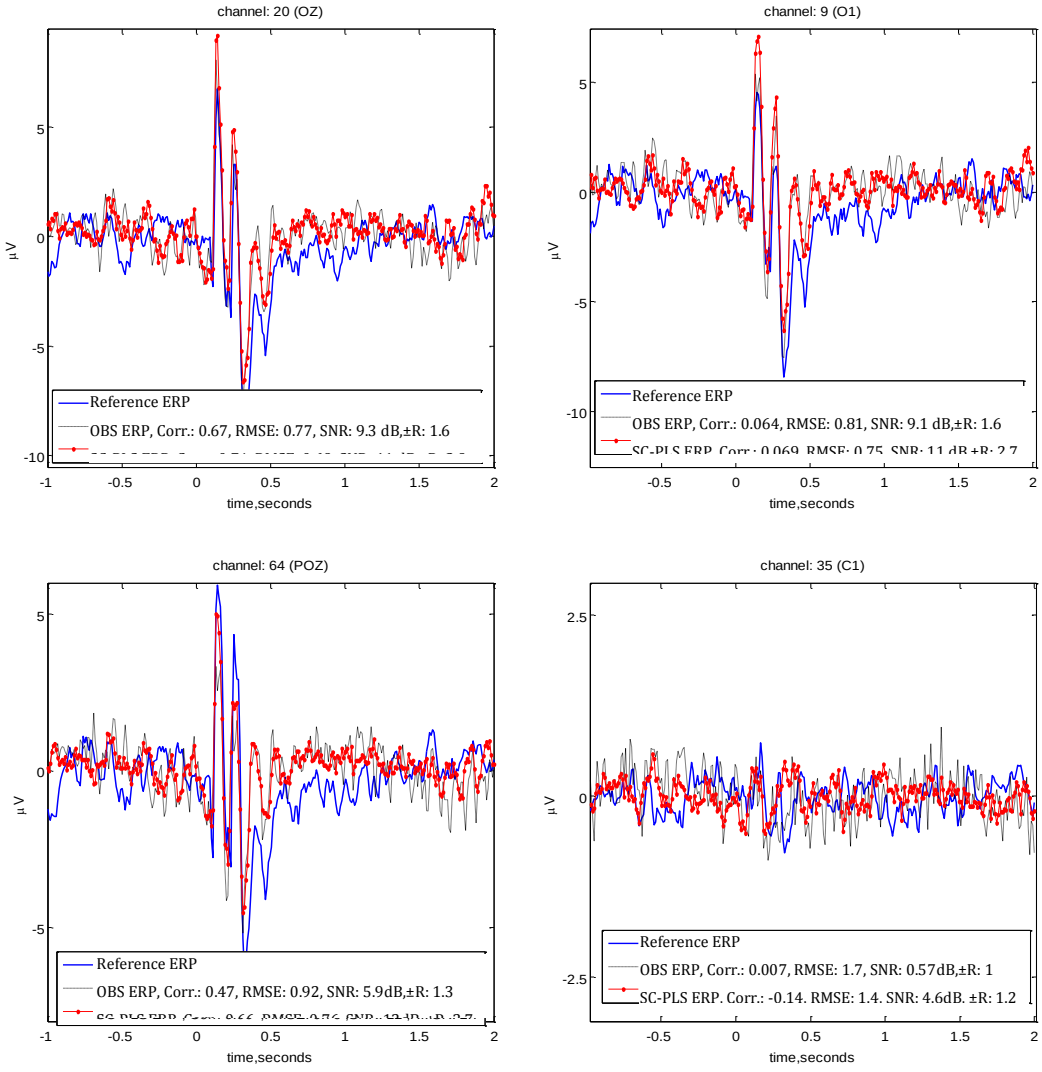


Figure 5-26: Root mean-squared error between the actual epochs (EEG without artifacts) and the de-noised epochs at each iteration step. Right: correlation coefficient between the de-noised epochs and the original epochs at each iteration step; averaged over all EEG channels and for EEG channels near to occipital area (subject E-N, experimental MRI study, no gradient artifacts)

The following figures show the epochs cleaned using the OBS algorithm and the SC-PLS algorithm for several channels for E-N:


channel: 20 (OZ)
Reference ERP
OBS ERP, Corr.: 0.67, RMSE: 0.77, SNR: 9.3 dB,±R: 1.6
μV
time,seconds
channel: 9 (O1)
Reference ERP
OBS ERP, Corr.: 0.064, RMSE: 0.81, SNR: 9.1 dB,±R: 1.6
μV
time,seconds
channel: 64 (POZ)
Reference ERP
OBS ERP, Corr.: 0.47, RMSE: 0.92, SNR: 5.9dB,±R: 1.3
μV
time,seconds
channel: 35 (C1)
Reference ERP
OBS ERP, Corr.: 0.007, RMSE: 1.7, SNR: 0.57dB,±R: 1
SC-PLS ERP. Corr.: -0.14. RMSE: 1.4. SNR: 4.6dB. ±R: 1.2
μV
time,seconds

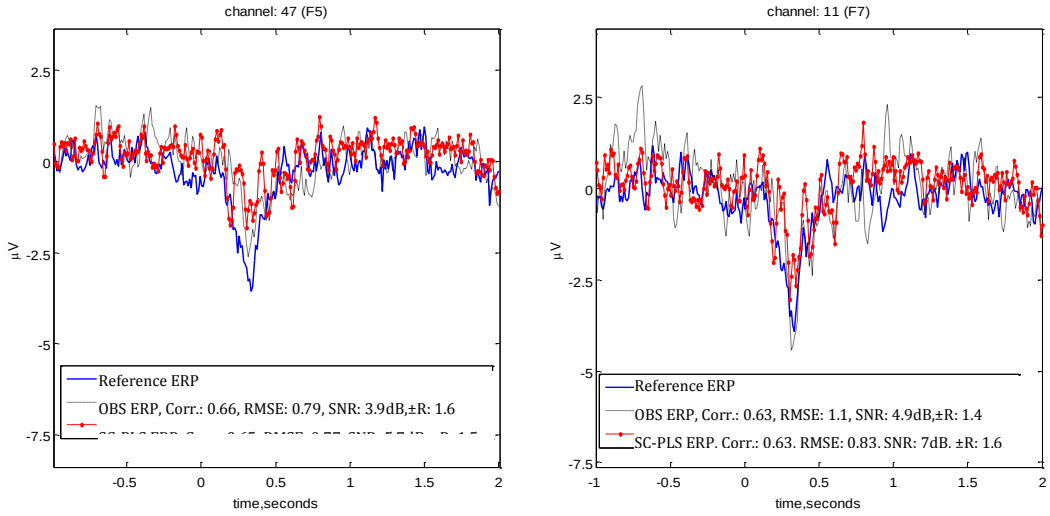


Figure 5-27: Epochs before and after removing noise from the EEG data in channels 9,10, 64,35,47 and 11 (subject E-N, experimental MRI study, no gradient artifact)

5.5.1.3 Overall Results

The following graphs show the quality measures obtained from the simulation study for six patients. The simulation results as well as the experimental results obtained without gradient artifacts (Figure 5-28 and Figure 5-29) show that SC-PLS provides slightly better results compared to the OBS algorithm. However, it appears that unless the datasets are pre-processed and a good initial estimate exists, the SC-PLS algorithm tends to retain some of the artifacts in the EEG signal. This issue mostly affects the high-frequency components of the ERP averages, such as the P150 components. However, the overall correlation values are better than the OBS method. Another issue that was encountered using SC-PLS was the problem of selecting proper penalty values (λ) for the algorithm. It appears that SC-PLS algorithm is more sensitive to the value of λ when it is used to remove BCG artifacts, and a poor choice of λ can affect the method's success.

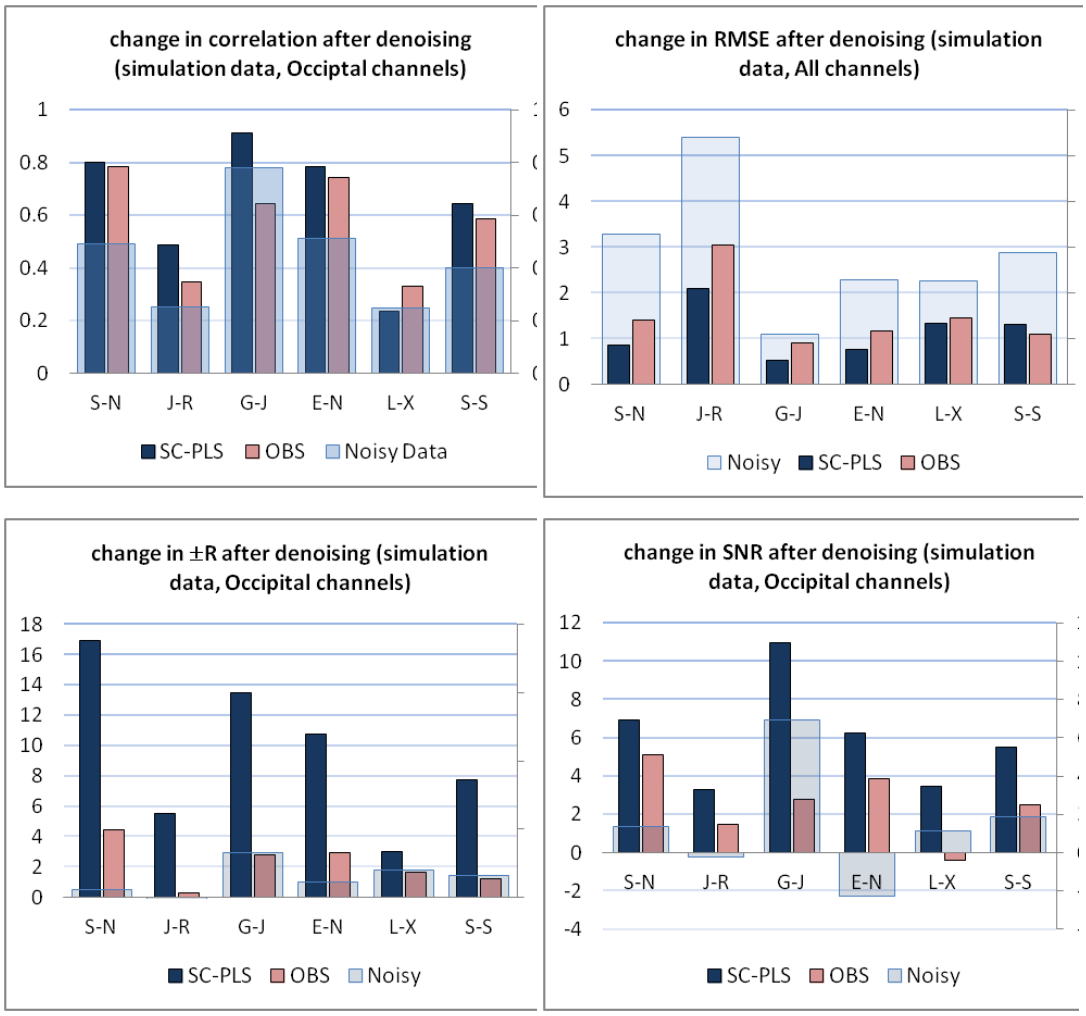


Figure 5-28: Simulatoin Data: Changes in quality measures (De-noised – Noisy values) averaged over EEG channels near the occipital lobe (RMSE averaged over all channels). Top-Left; Average correlation of OCC channels, Top-Right; Average RMSE over all channels, Bottom-Left Average ±Ratio (dB)-OCC Channels, Bottom-Right; SNR (dB), OCC-channels. (Results obtained from the Simulation MRI datasets)

The results after applying the SC-PLS and OBS in the experimental data without RF running are plotted below:

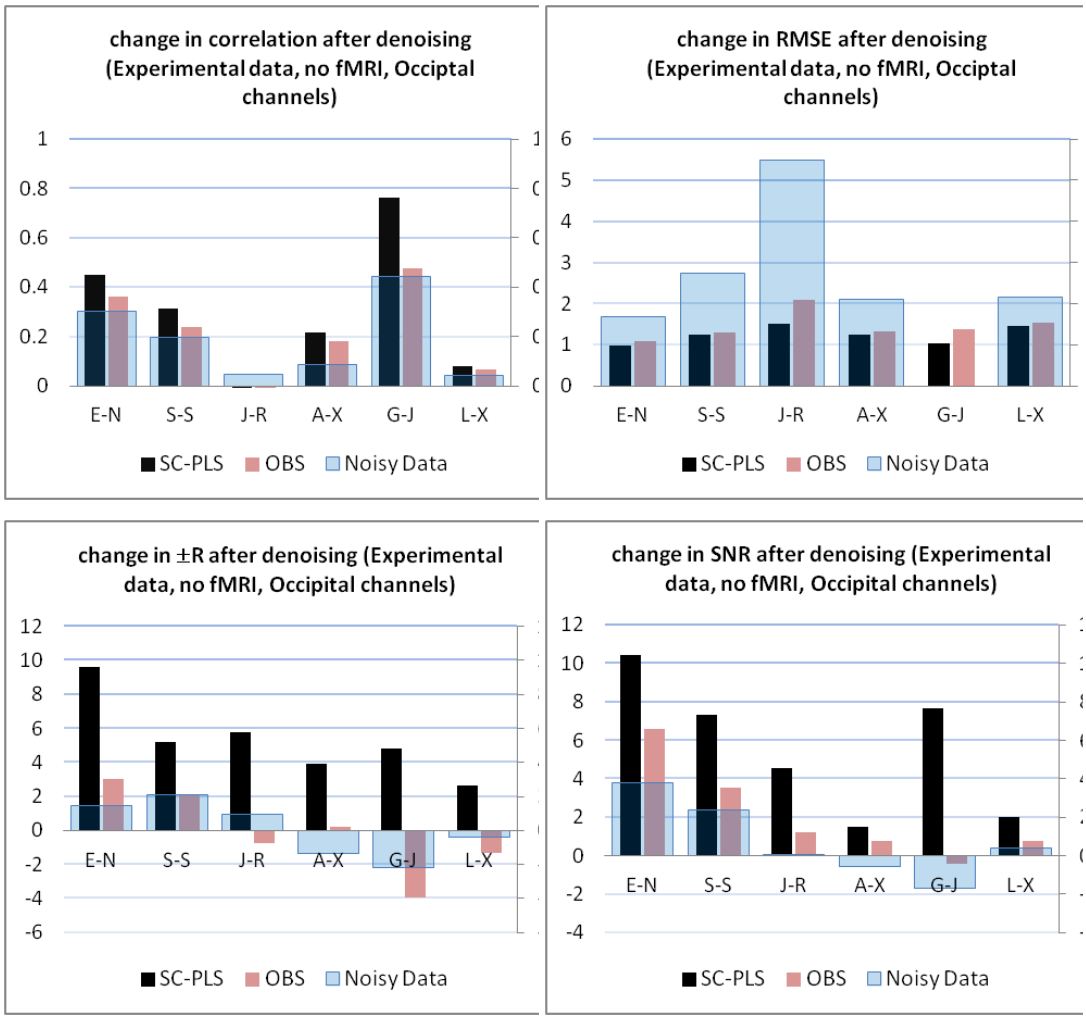


Figure 5-29: Experimental data: Changes in quality measures (De-noised – Noisy values) averaged over EEG channels near the occipital lobe (RMSE averaged over all channels). Top-Left; Average Correlation of OCC channels, Top-Right; Average RMSE over all channels, Bottom-Left Average ±Ratio (dB)-OCC Channels, Bottom-Right; SNR (dB), OCC-channels. (Results obtained from the experimental MRI datasets after removal of Gradient and BCG artifacts

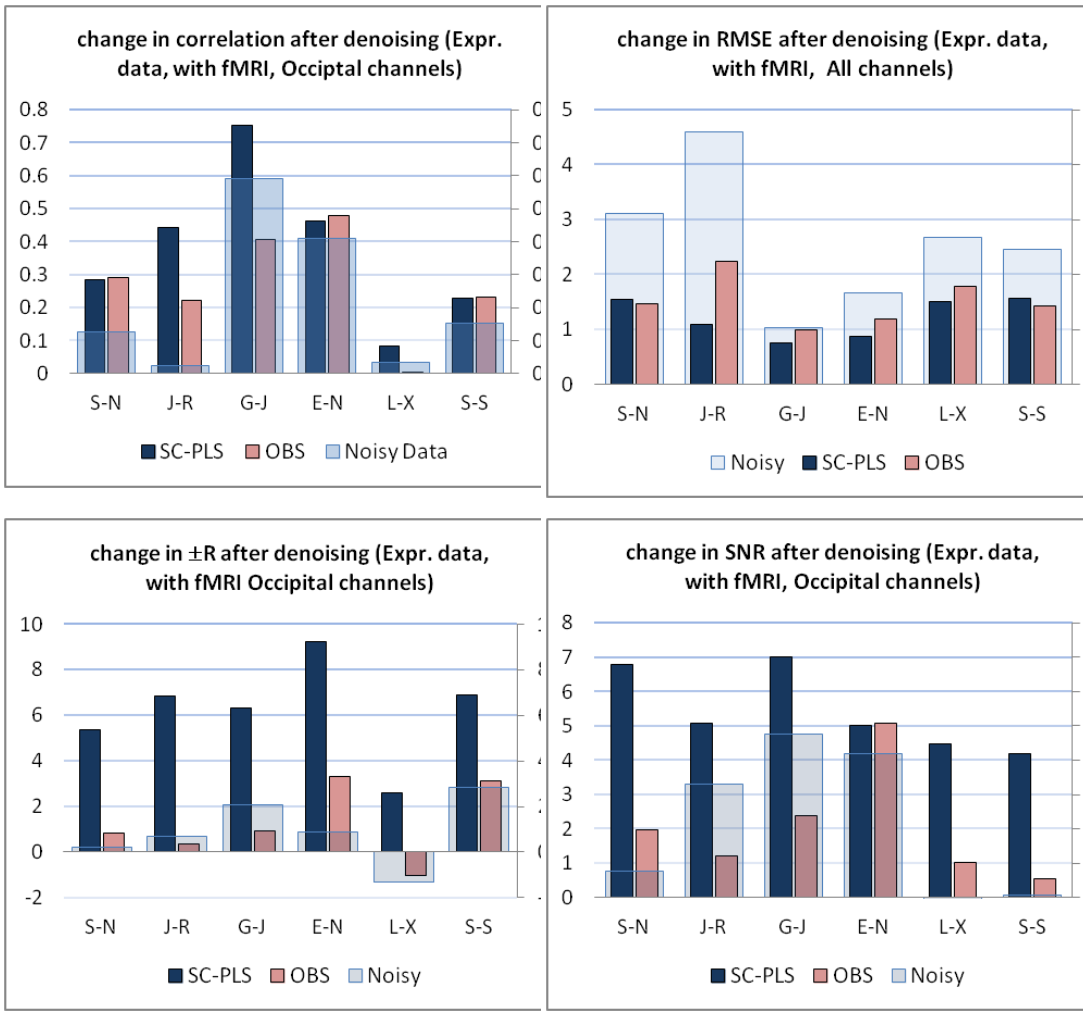


Figure 5-30: Changes in quality measures (De-noised – Noisy values) averaged over EEG channels near the occipital lobe (RMSE averaged over all channels). Top-Left; Average correlation of OCC channels, Top-Right; Average RMSE over all channels, Bottom-Left Average ±R (dB)-OCC Channels, Bottom-Right; SNR (dB), OCC-channels. (Results obtained from the experimental MRI datasets after removal of Gradient and BCG artifacts)

5.6 Conclusion

In this study, the SC-PLS algorithm was implemented to remove two different types of structured noise from EEG data: muscle artifacts and the BCG noise induced by recording the data inside the MRI chamber. In both cases, satisfactory

results were achieved; however, the accuracy of this method depends heavily on selection of the weighting coefficient (λ), which regulates the tradeoff between noise and the signal components. A very small λ value will result in loss of correlation, as the noise components selected will not be entirely free of stimuli-related signal. If the λ value is selected to be high, the components selected will not be strongly correlated with the noise. It was realized that in most cases choosing a value of

$$\lambda = \frac{\rho_{\mathbf{X'JJ'X}}}{\rho_{\mathbf{X'RR'X}}} \tag{5-27}$$

which is the ratio between $\rho_{\mathbf{X'\tilde{Y}\tilde{Y}'X}}$, the largest eigenvalue of $\mathbf{X'\tilde{Y}\ \tilde{Y}'X}$, and $\rho_{\mathbf{X'ZZ'X}}$, the largest eigenvalue of $\mathbf{X'ZZ'X}$, will provide satisfactory results. In most experiments (excluding those MRI experiments having to remove GA artifacts prior to applying SC-PLS), three to four iterations were enough to obtain the optimal results. It was also noted that a good initial guess is required for the SC-PLS algorithm to converge towards satisfactory results. In the MRI experiments, the data were preprocessed by applying a moving average filter to remove some of the BCG artifacts. This step is not necessary when the initial signal-to-noise ratio is high (roughly higher than 2 dB). However, it was realized that in most MRI experiment cases, the initial SNR was very low, and the algorithm did not achieve satisfactory results without the preprocessing step. Wavelet filters were also used to improve the value of $\mathbf{Z}$ at each iteration step.

A very good strategy for initializing the SC-PLS algorithm is to perform the

EEG outside the MRI and use the Reference ERP averages as the initial guess. This idea was used by [27] in an algorithm that is somewhat similar to SC-PLS to create spatial filters to remove BCG noise. Their method has been described earlier in Section I. However the advantage of our method is that it does not require an initial estimate acquired in a clean (noise free) environment (or condition) and also it improves the results by iteratively removing the noise and enhancing the signals.

Another issue to note here is the choice of the basis matrix ($\tilde{\boldsymbol{X}}$) in (5-15)-(5-17). The Basis matrix can either be the original noisy EEG data or a linear combination of the EEG data. Both datasets were tested in the above experiments. Using the ICA decomposition of the original EEG data as the basis for SC-PLS method seems to provide slightly better results. ICA is a very strong method for separating independent signals from each other. Our experiments showed that if the original EEG dataset is used as a basis, the convergence will less stable and the iterations should be stopped after two or three iterations or the SNR will reduce sharply.

Overall, the SC-PLS algorithm can be used to remove noise artifacts when there is some information available about the stimuli (for example the initial ERP values or estimates of the ERP obtained elsewhere). It does not require any knowledge about the noise, as it tries to retain the stimuli-related information while removing all the other variations in the dataset. In any case, where there is extra information available about the noise, it can be easily incorporated into the

algorithm by modifying matrix $\tilde{\mathbf{Y}}$.

5.6.1 GA artifact correction problems

It was realized that using gradient artifact correction can seriously obscure the shape of the ERP averages. A simulation study was conducted by adding Reference ERP averages randomly to the EEG data recorded with gradient artifacts. The gradient artifact removal tool from EEGLAB software was used to remove GA artifacts. We found that the algorithm greatly reduces the amplitude of the P100 components of the ERPs. This is probably due to the chosen number slices acquired per second (10 slices/sec) which was very close to the P100 component.

5.7 Appendix

Following section includes the results obtained from individual patients from the muscle artifact experiment.

5.7.1 subject M-T, Simulation results (muscle artifact)

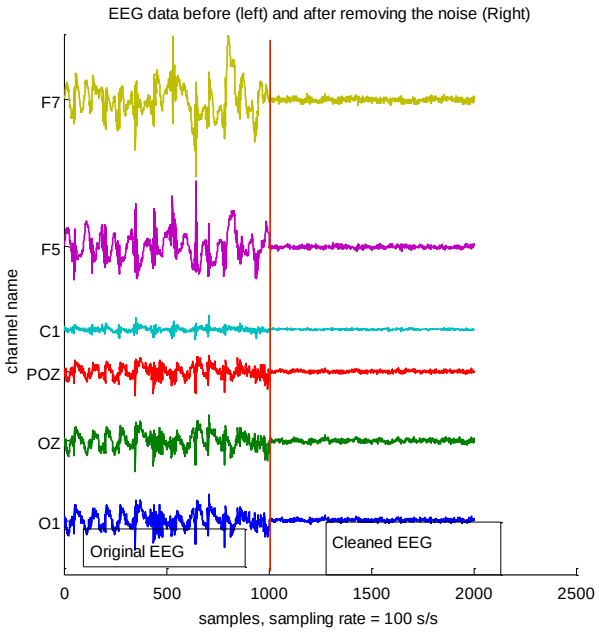


Figure 5-31: EEG signal before and after removal of noise for some of the channels. Signal on the left is the EEG data prior to removal of the noise, signal on the right of the screen represents the same portion of the EEG data after removal of the noise (subject M-T, simulation data, Muscle artifact removal)

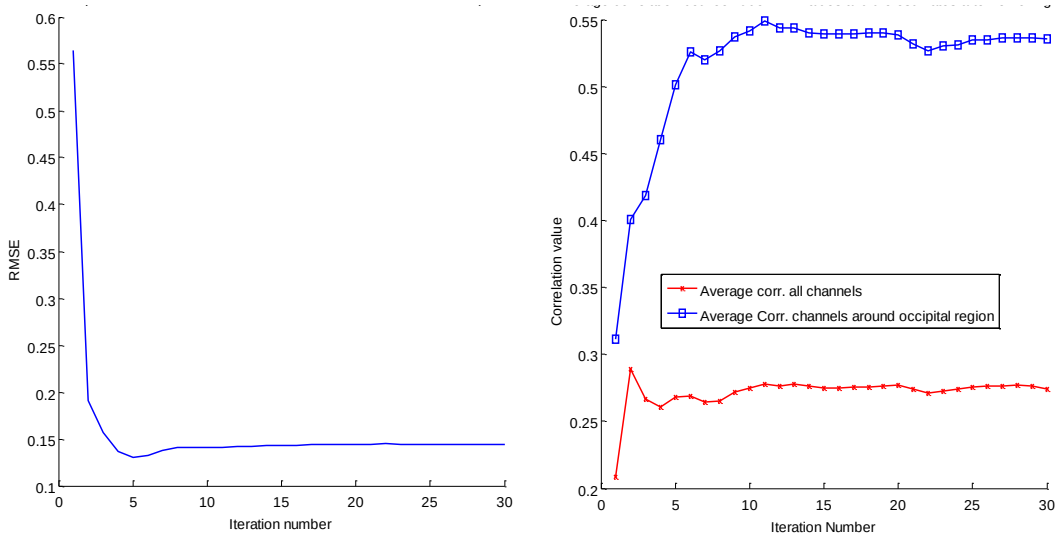
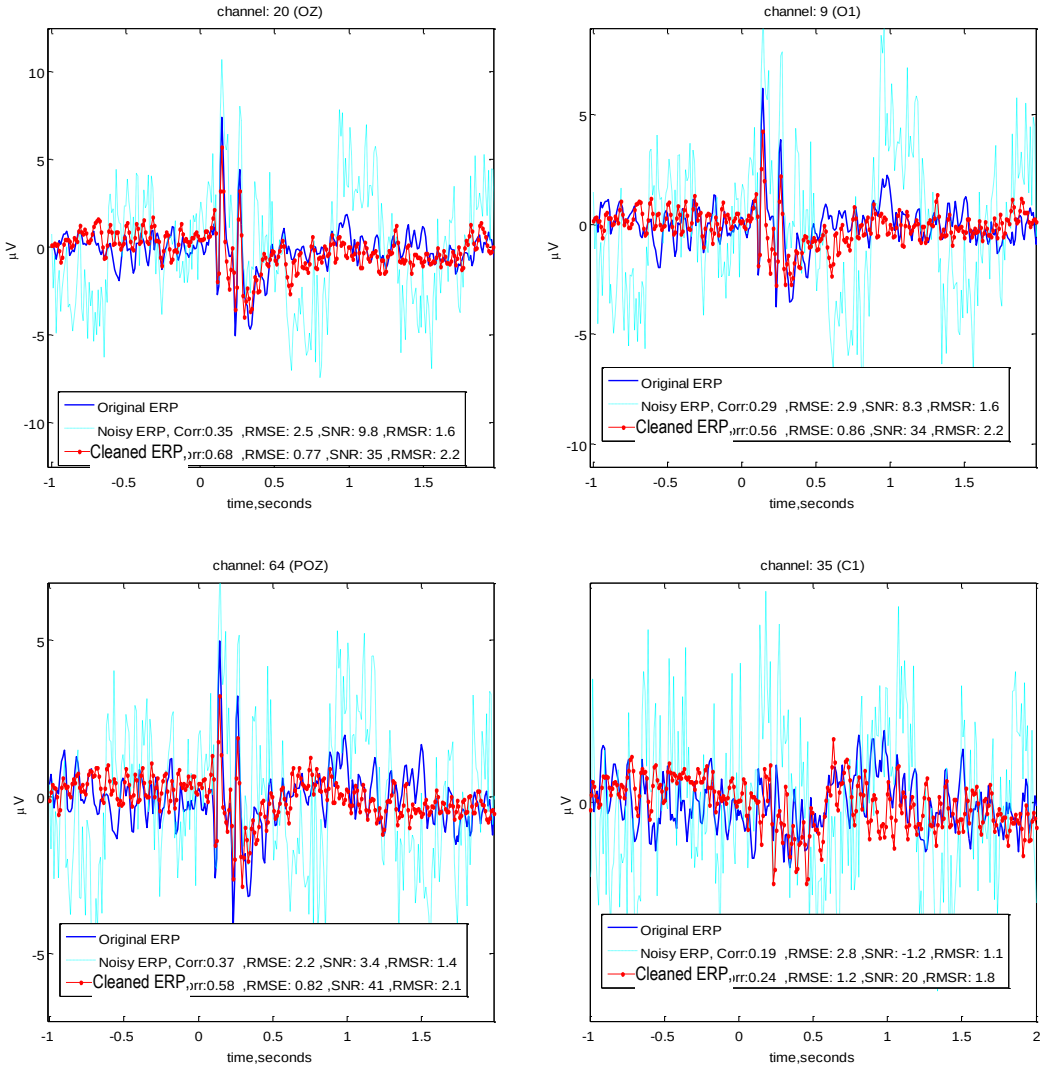


Figure 5-32: left: Root mean squared error between the averaged Reference ERP values and the averaged ERP values extracted after denoising the signal using SC-PLS at each iteration step. Right: averaged correlation (subject M-T, simulation study);



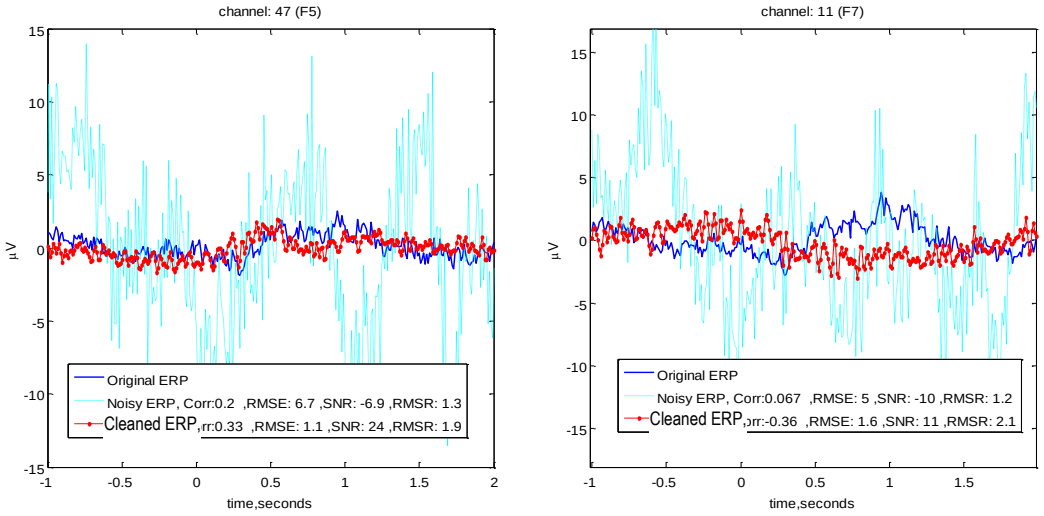


Figure 5-33: Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP (subject M-T, simulation EEG)

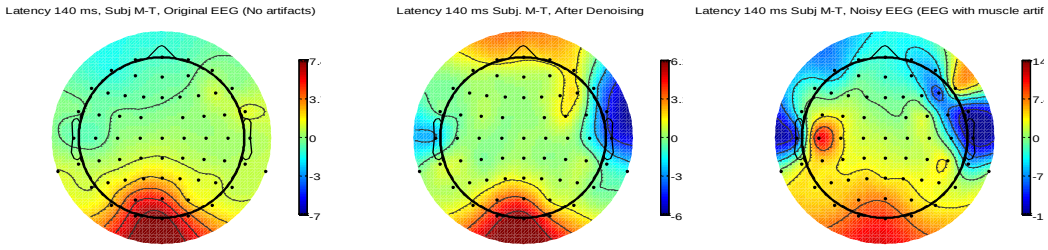


Figure 5-34: Topographical maps of the P150 component; Left: Reference EEG data. Middle: P150 component obtained from the noisy EEG dataset prior to noise removal. Right: P150 component map from the denoised EEG dataset, using SC-PLS algorithm (subject M-T, simulation EEG).

5.7.2 Subject "M-T", experimental results(muscle artifact removal) :

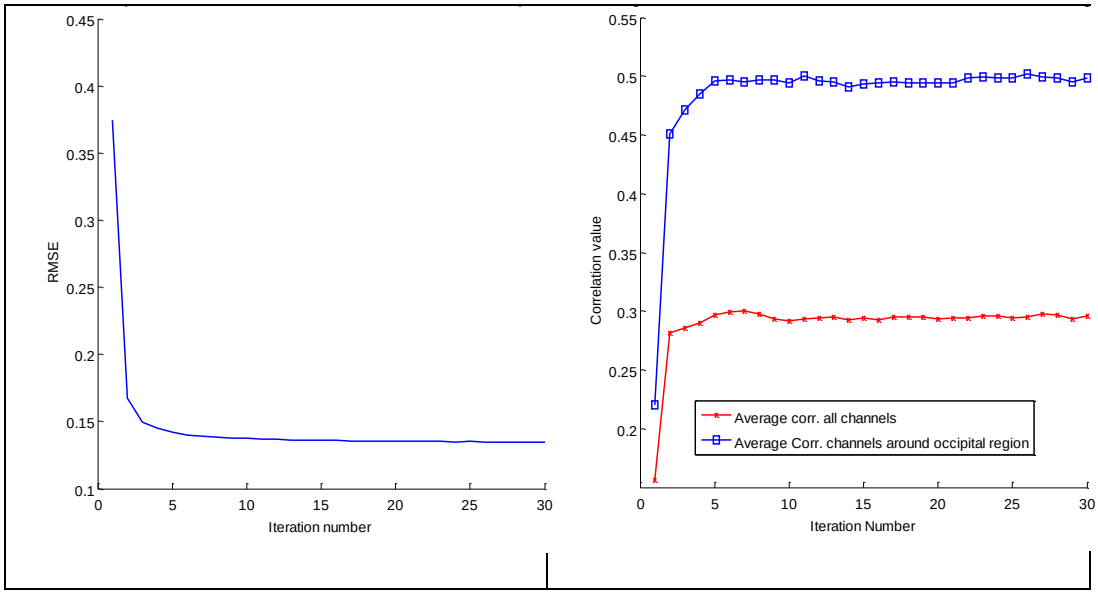
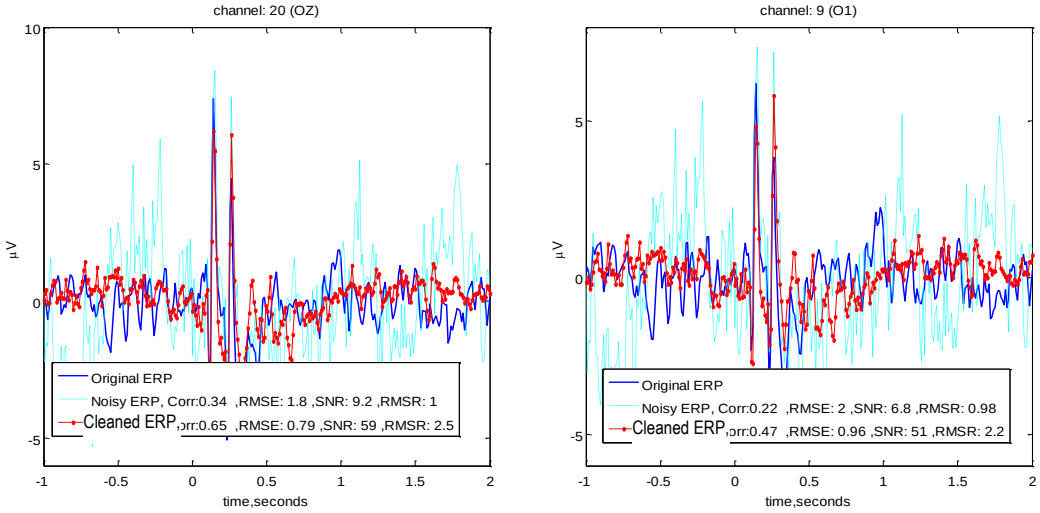


Figure 5-35. Left: averaged RMSE between Reference ERP values and the denoised ERP values using SC-PLS, at each iteration step. Right: averaged correlation values for all the channels and the channels near the occipital lobe at each iteration step (subject M-T, actual EEG);



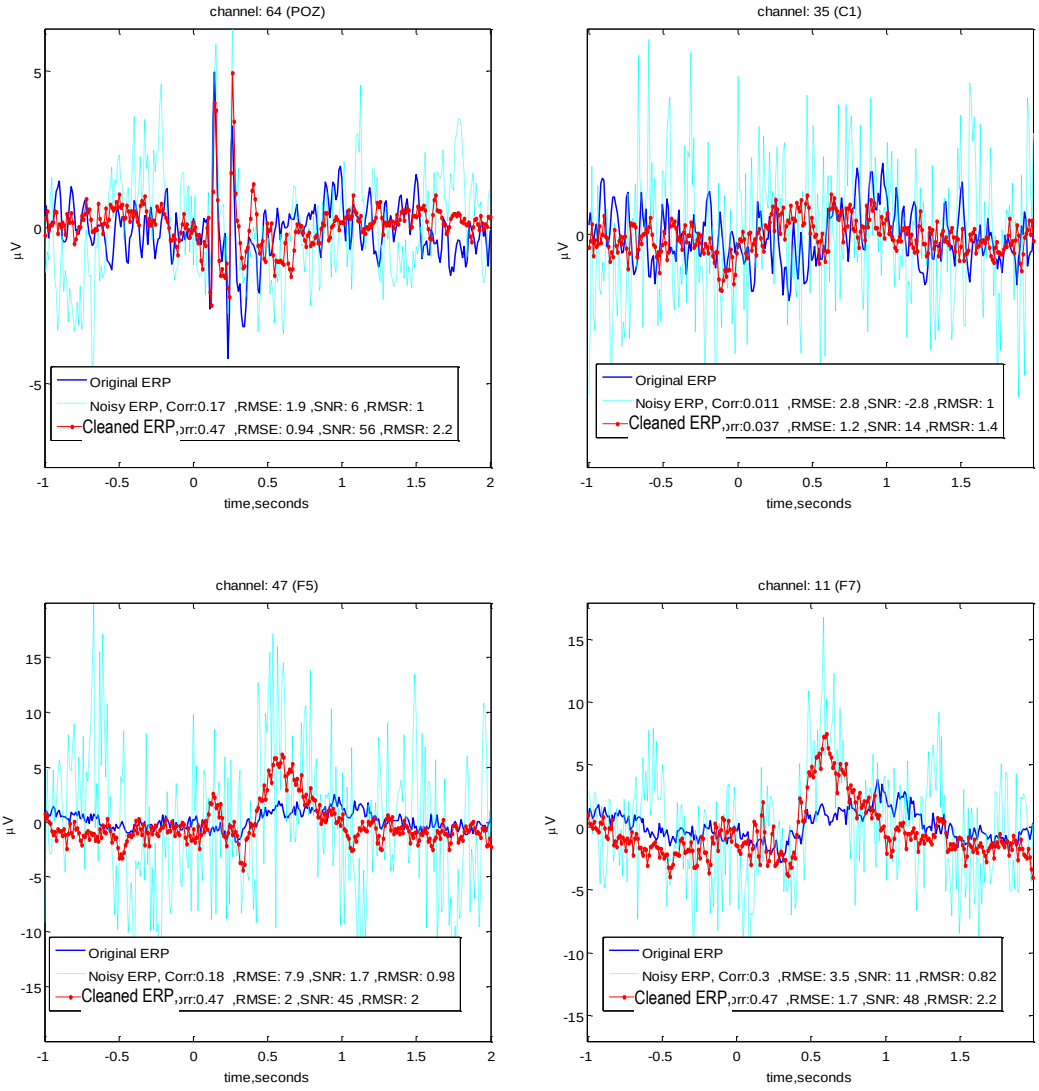


Figure 5-36 Averaged ERP values before and after removing noise using SC-PLS in channels 9,10,64,35,47 and 11. (subject M-T, Experimental EEG data)

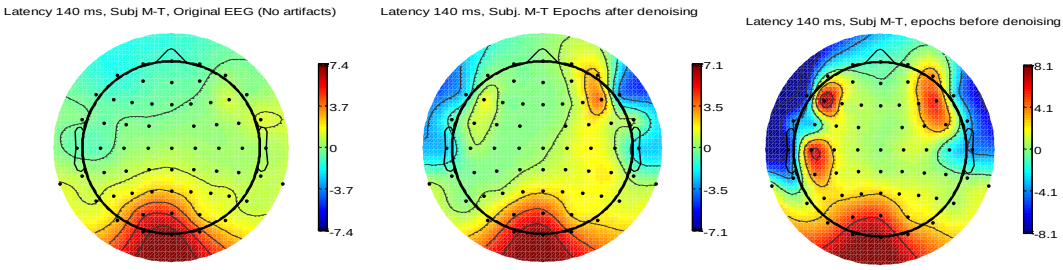


Figure 5-37: Topographical maps of the P150 component; Left: Reference EEG data. Middle: P150 component obtained from the noisy EEG dataset prior to noise removal. Right: P150 component map from the denoised EEG dataset, using SC-PLS algorithm (subject M-T, Experimental EEG data).

5.7.2.1 Subject "J-R", simulation results (gum experiment)

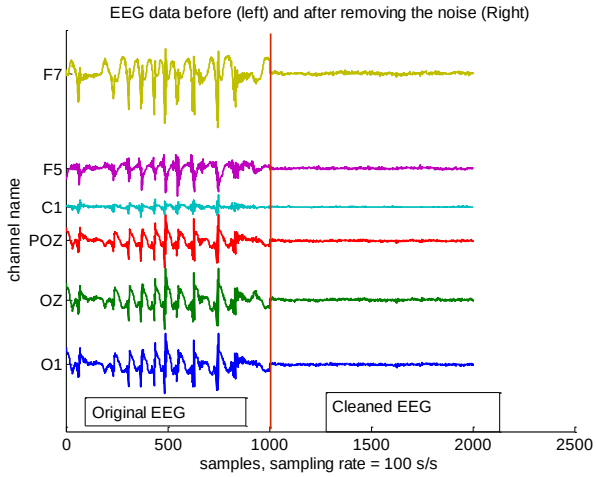


Figure 5-38: EEG signal before and after removal of noise for some of the channels. Signal on the left is the EEG data prior to removal of the noise, signal on the right of the screen represents the same portion of the EEG data after removal of the noise (Subject J-R, Simulation data).

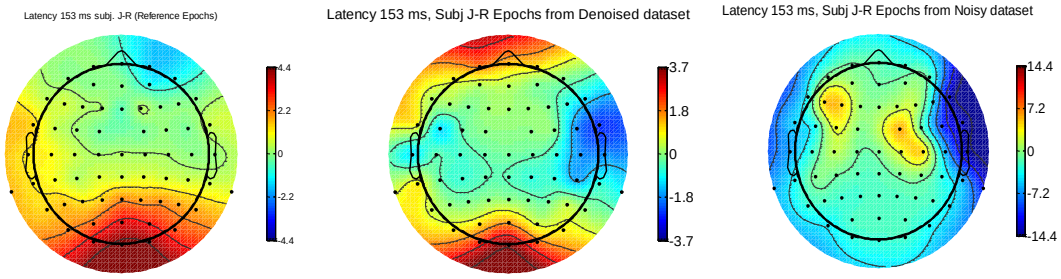


Figure 5-39: Topographical maps of the P150 component; Left: Reference EEG data. Middle: P150 component obtained from the noisy EEG dataset prior to noise removal. Right: P150 component map from the denoised EEG dataset, using SC-PLS algorithm (subject J-R, simulation data).

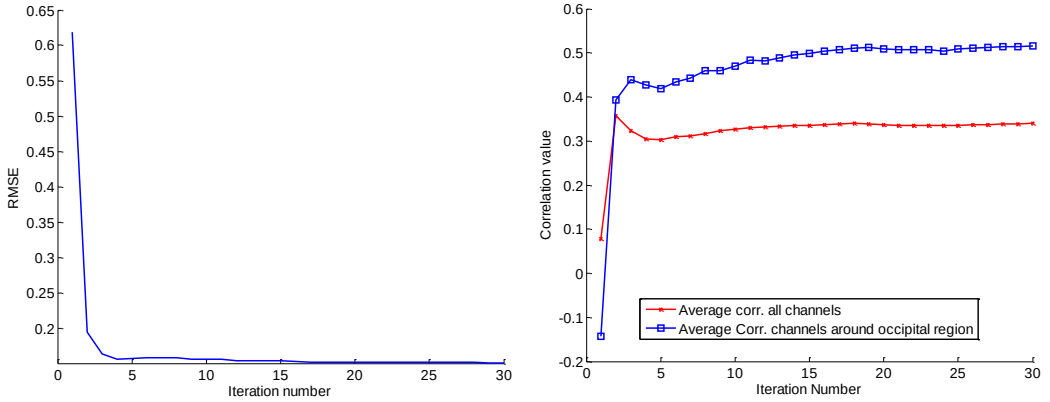
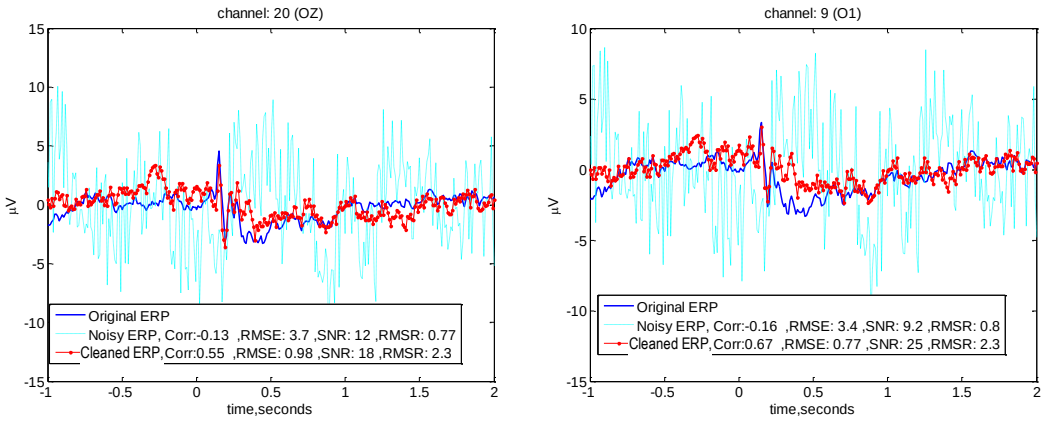


Figure 5-40: Left: averaged RMSE between Reference ERP values and the denoised ERP values using SC-PLS, at each iteration step. Right: averaged correlation values for all the channels and the channels near the occipital lobe at each iteration step (subject J-R, simulation data)



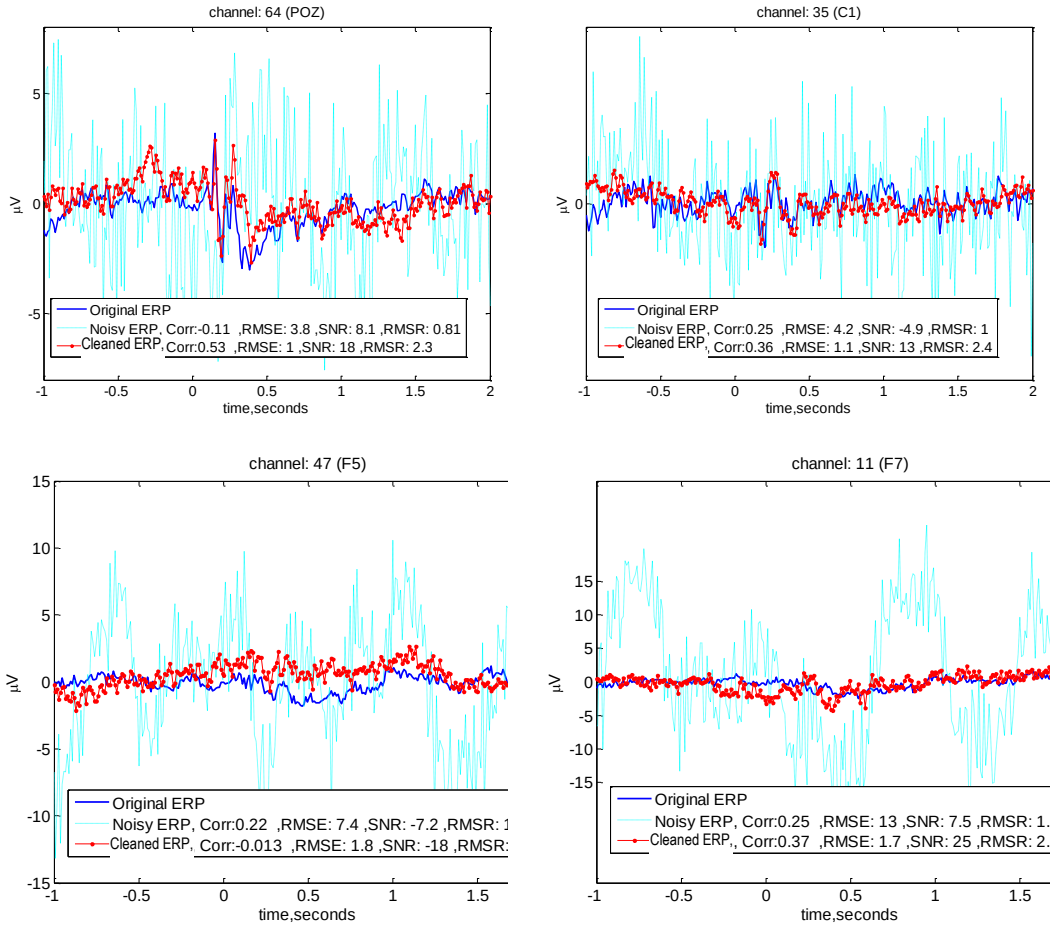


Figure 5-41: Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP . (subject J-R, simulation EEG).

5.7.3 Subject "J-R", experimental results (muscle artifact study):

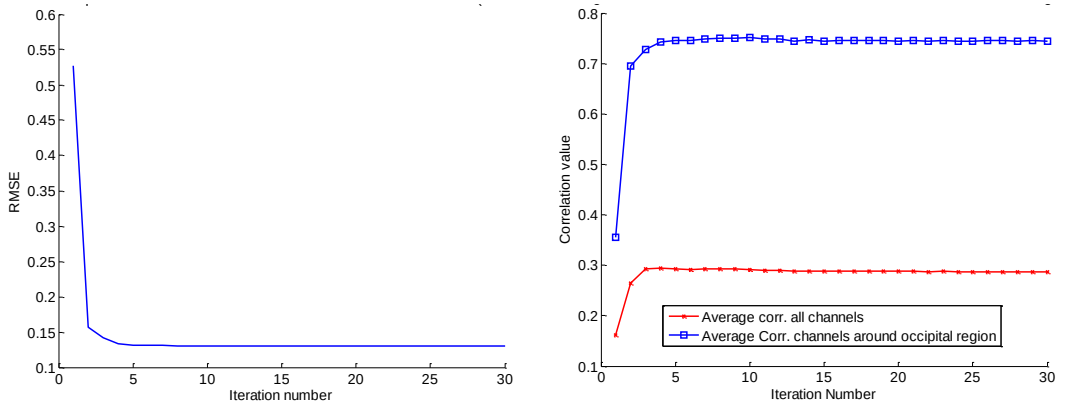


Figure 5-42: Left: averaged RMSE between Reference ERP values and the denoised ERP values using SC-PLS, at each iteration step. Right: averaged correlation values for all the channels and the channels near the occipital lobe at each iteration step (subject J-R, experimental EEG, muscle artifact study).

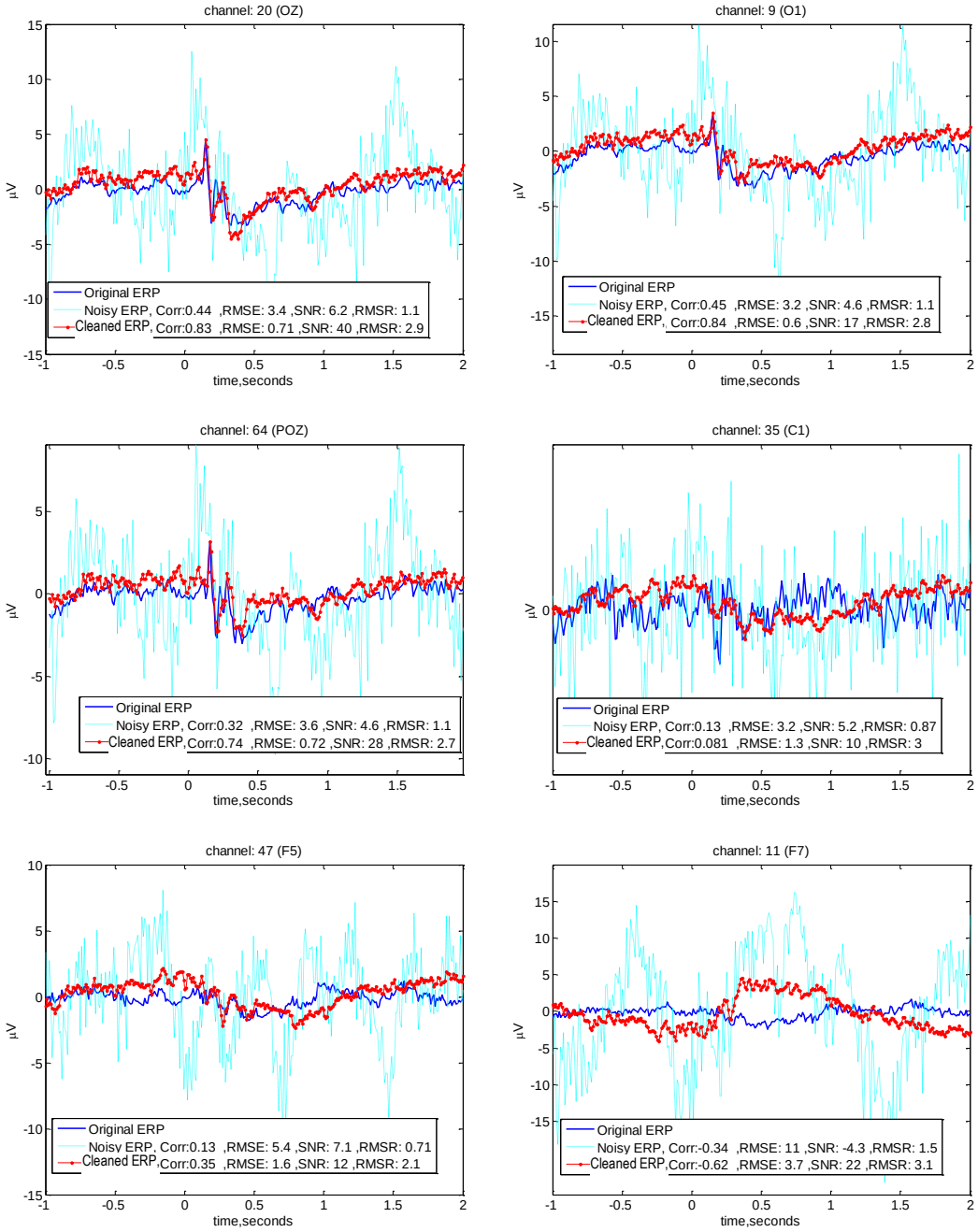


Figure 5-43: Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP .. (subject J-R, experimental muscle artifact EEG).

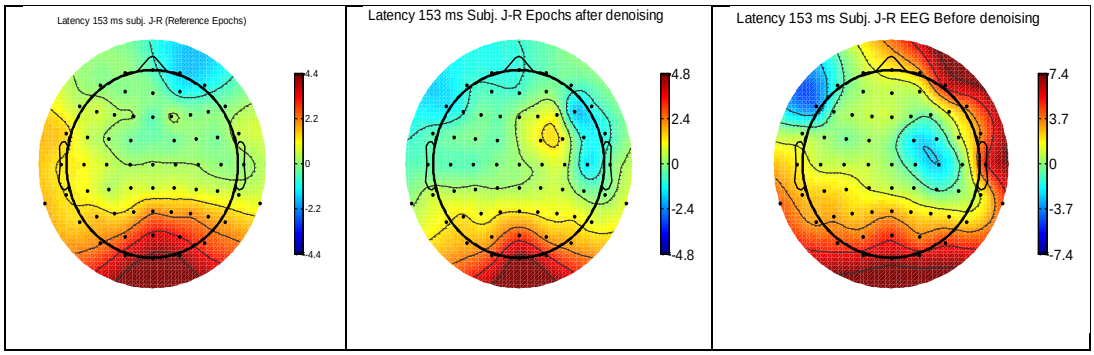


Figure 5-44: Topographical maps of the P150 component; Left: Reference EEG data. Middle: P150 component obtained from the noisy EEG dataset prior to noise removal. Right: P150 component map from the denoised EEG dataset, using SC-PLS algorithm (subject J-R, experimental muscle artifact EEG)

5.7.4 Subject "E-N" Simulation data (muscle artifact study)

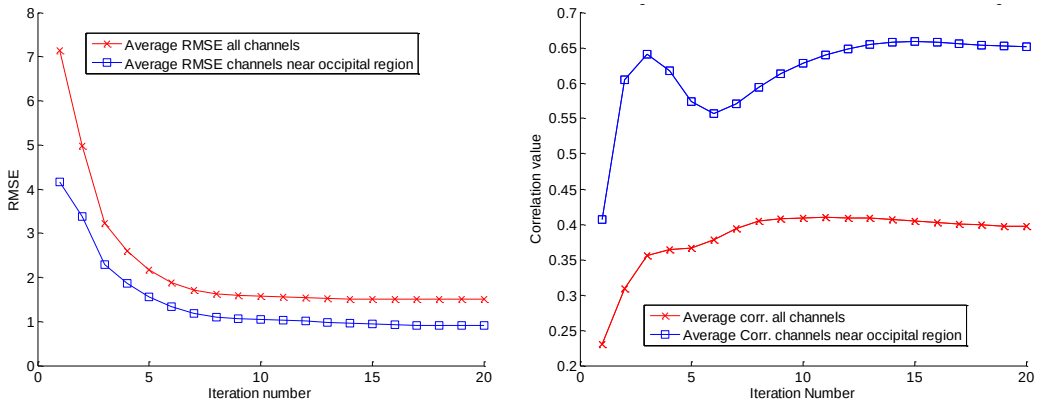


Figure 5-45: Left: averaged RMSE between Reference ERP values and the denoised ERP values using SC-PLS, at each iteration step. Right: averaged correlation values for all the channels and the channels near the occipital lobe at each iteration step. (subject "E-N" simulation muscle artifact study)

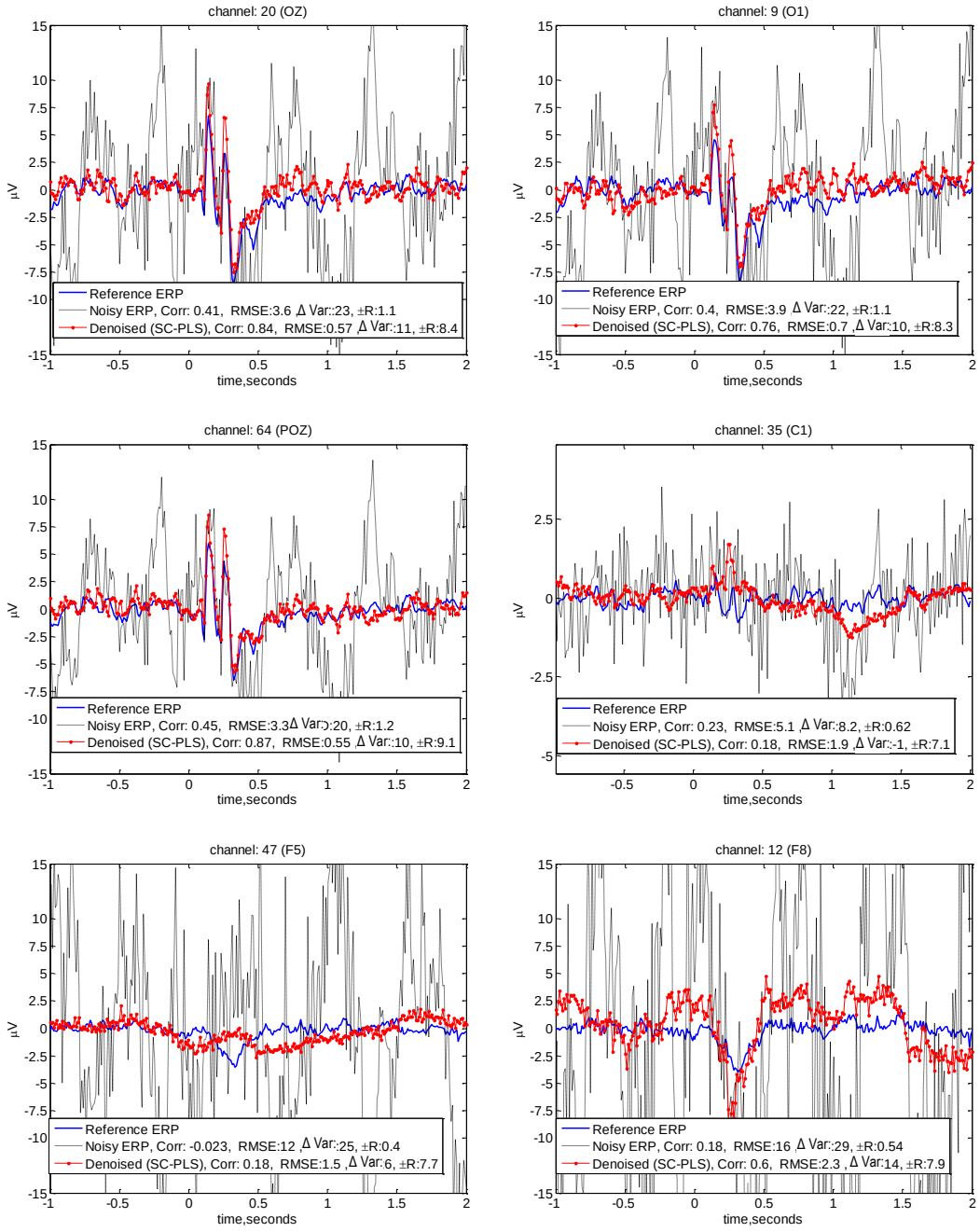


Figure 5-46: : Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP (Subject "E-N" simulation muscle artifact study)

5.7.5 Subject E-N , Experimental dataset (muscle artifact study)

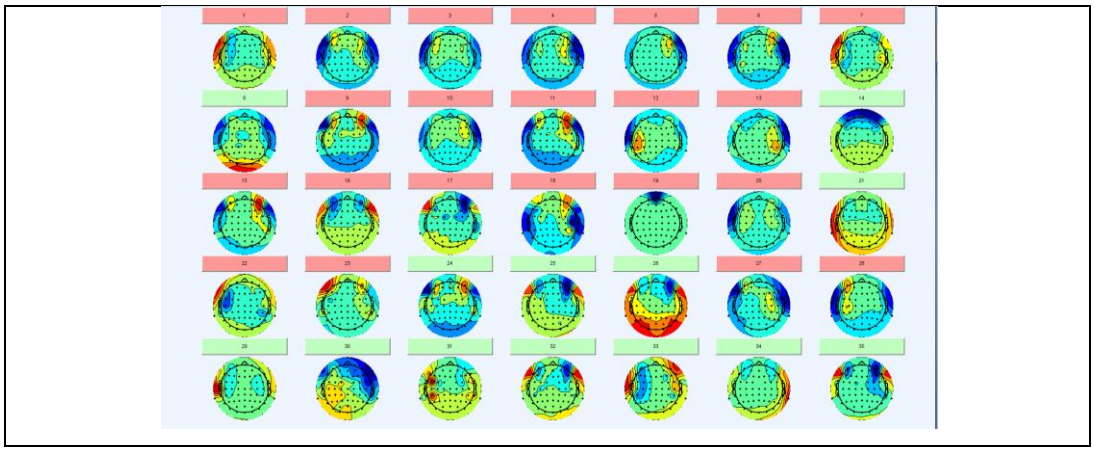


Figure 5-47: topographical maps of the ICA components for the noisy EEG data. The ICA components highlighted in red boxes discarded in the ICA component rejection algorithm (Subject "E-N", Experimental muscle artifact study)

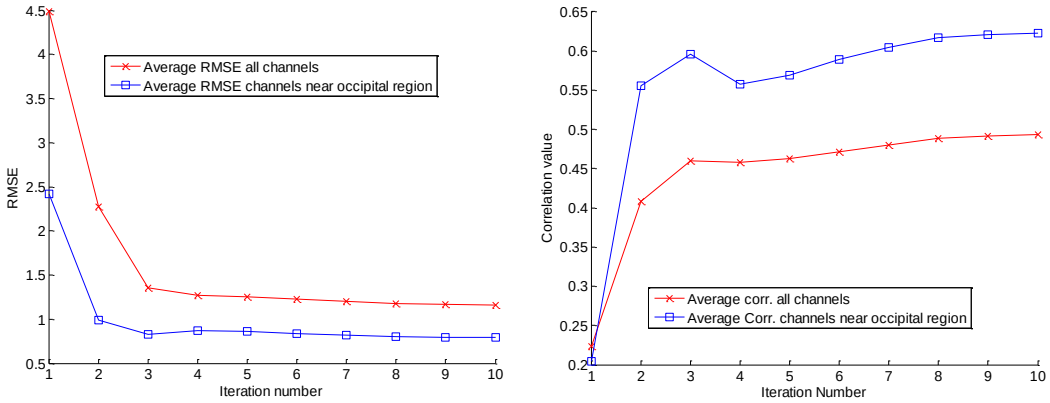


Figure 5-48: Left: averaged RMSE between Reference ERP values and the denoised ERP values using SC-PLS, at each iteration step. Right: averaged correlation values for all the channels and the channels near the occipital lobe at each iteration step. (Subject "E-N" experimental muscle artifact study)

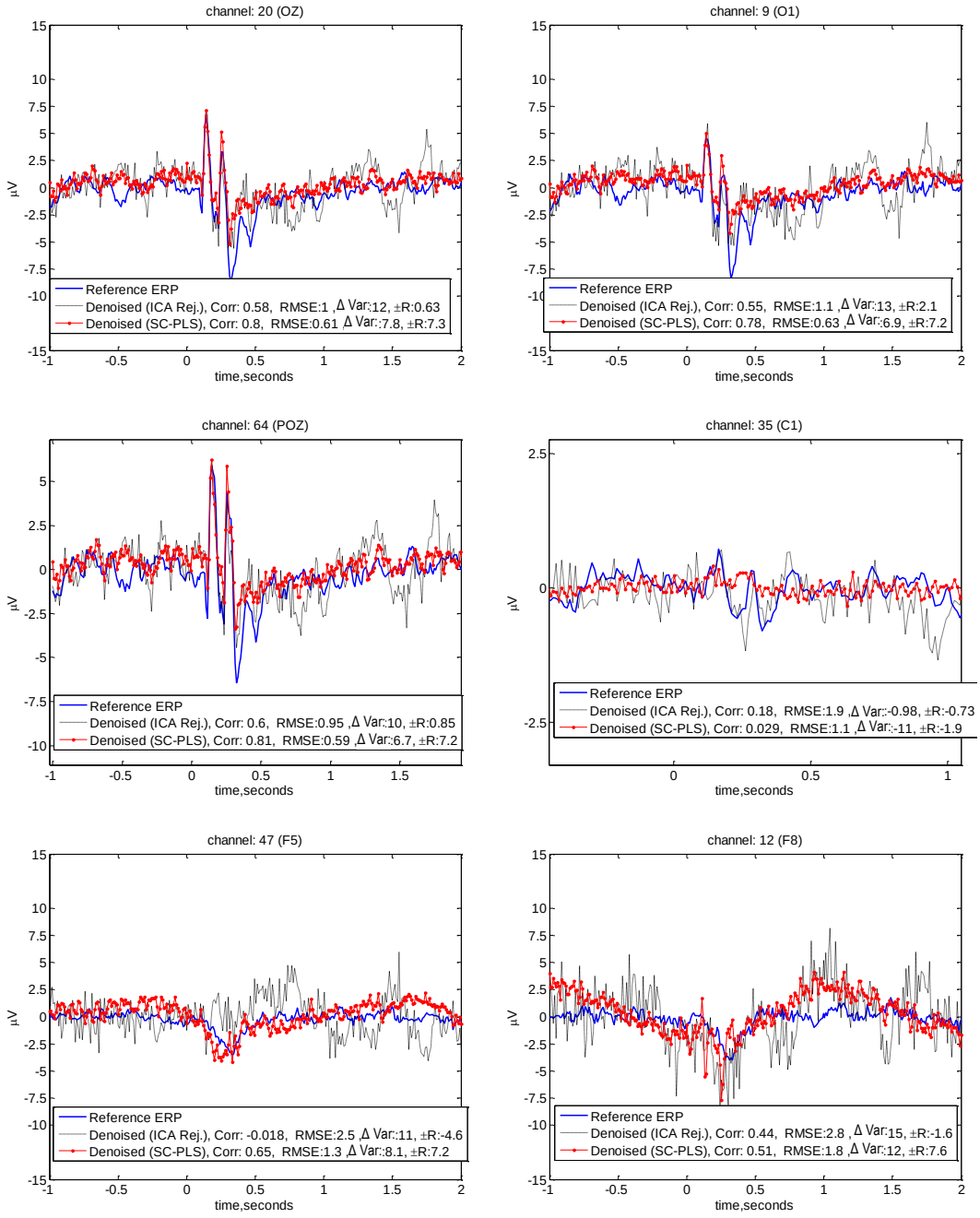


Figure 5-49: Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP (subject "E-N" experimental muscle artifact study)

5.7.6 Subject "L-X" simulation muscle artifact study

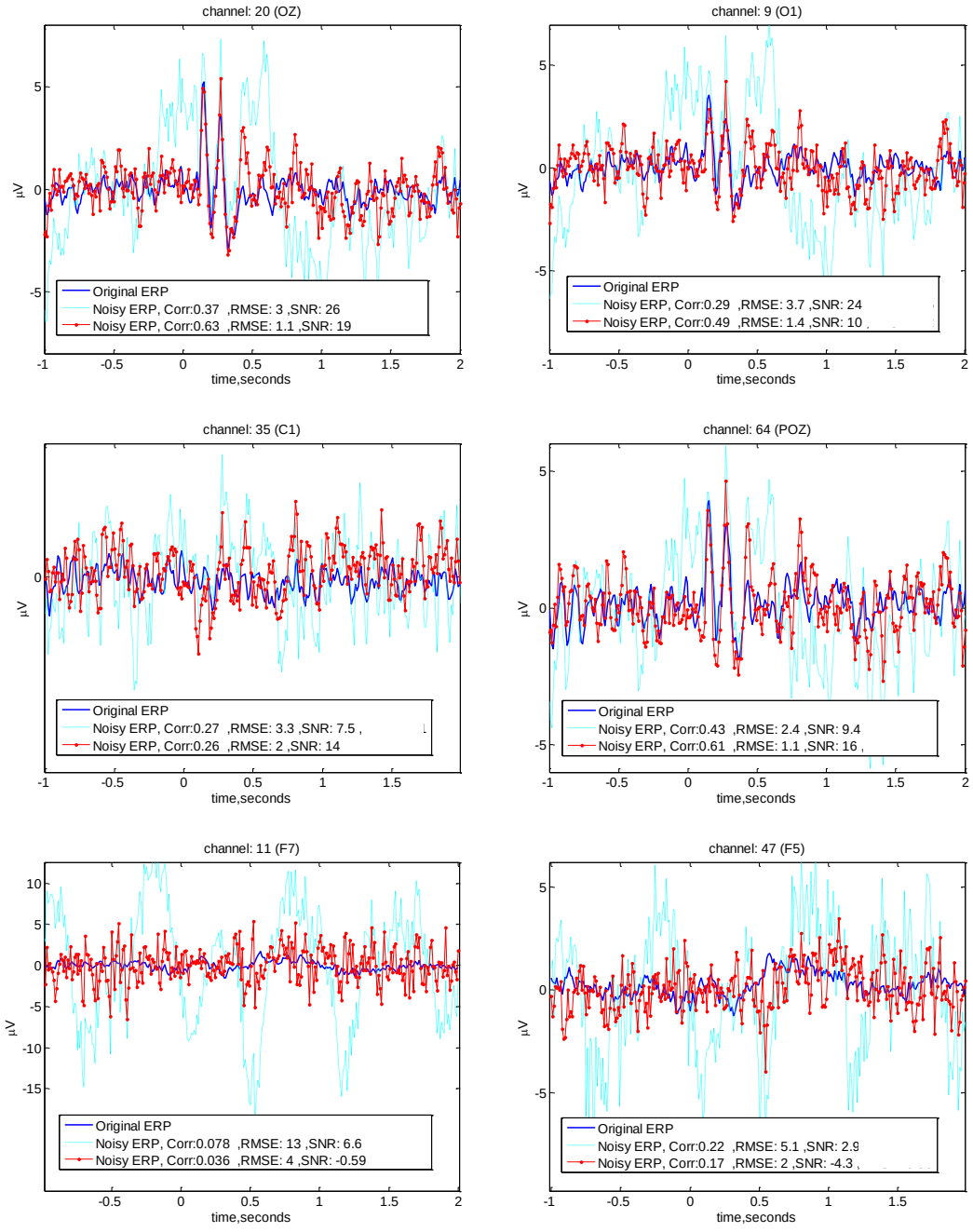


Figure 5-50: Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP (Subject "L-X" simulation muscle artifact study)

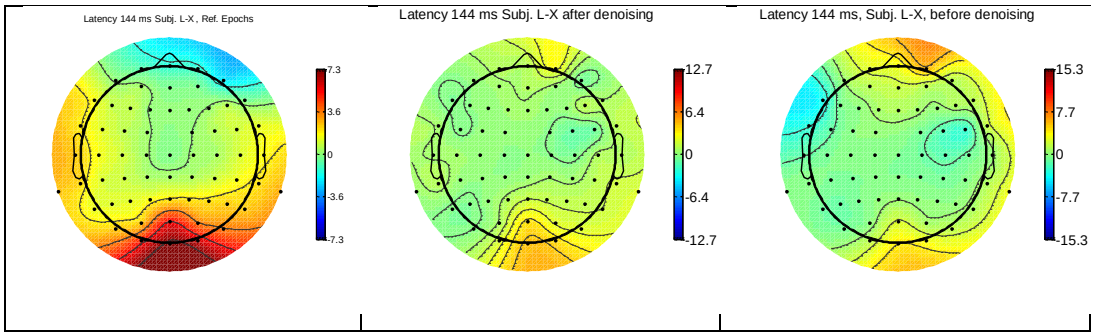


Figure 5-51: Topographical maps of the P150 component; Left: Reference EEG data. Middle: P150 component obtained from the noisy EEG dataset prior to noise removal. Right: P150 component map from the denoised EEG dataset, using SC-PLS algorithm (subject "L-X" simulation muscle artifact study)

5.7.7 Patient "L-X", BCG experimental study

5.7.7.1 Patient L-X

We applied the SC-PLS algorithm to remove the BCG noise from the EEG data for patient L-X.

5.7.7.1.1 Subject L-X, actual EEG data

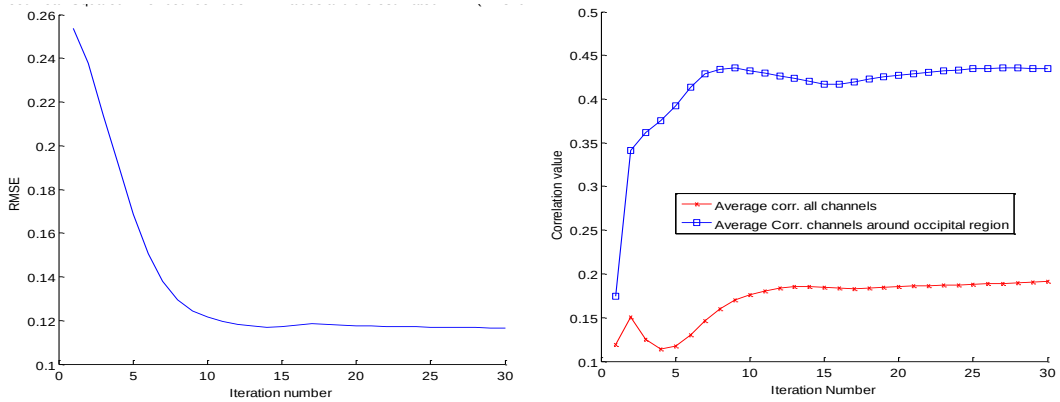
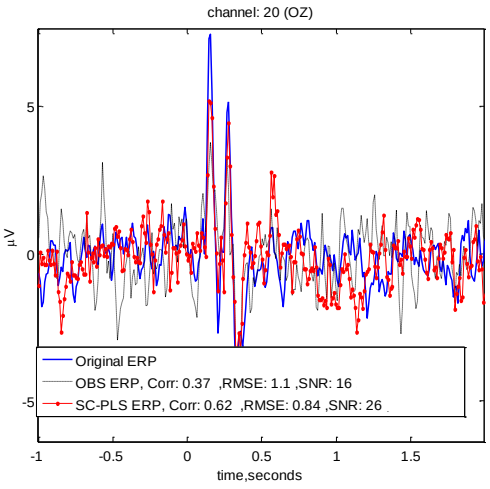
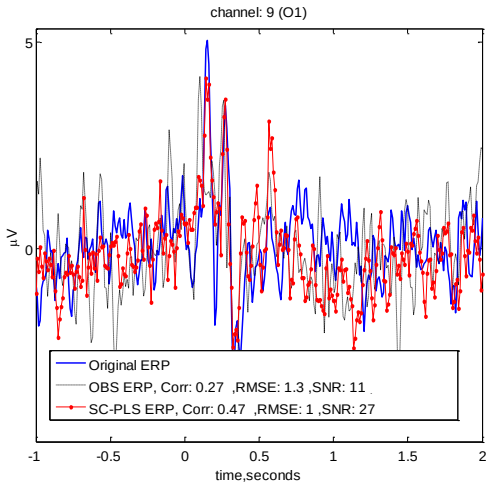
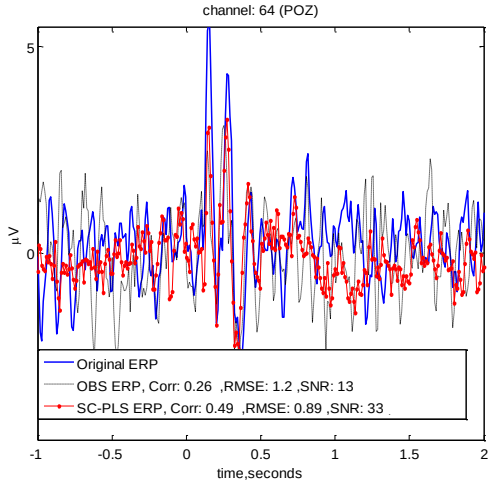
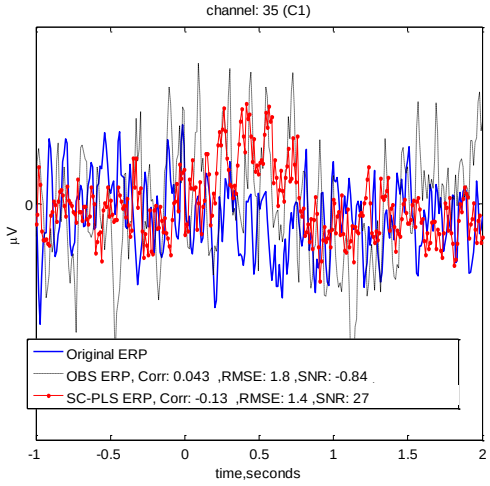


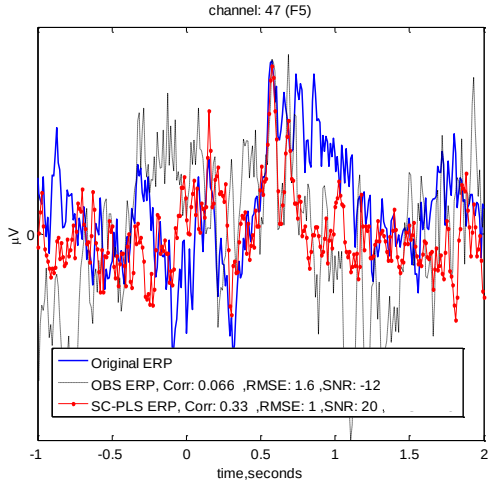
Figure 5-52: Left: averaged RMSE between Reference ERP values and the denoised ERP values using SC-PLS, at each iteration step. Right: averaged correlation values for all the channels and the channels near the occipital lobe at each iteration step (subject L-X, experimental EEG, MRI study)


channel: 20 (OZ)
Original ERP
OBS ERP, Corr: 0.37 ,RMSE: 1.1 ,SNR: 16
SC-PLS ERP, Corr: 0.62 ,RMSE: 0.84 ,SNR: 26
μV
time,seconds


channel: 9 (O1)
Original ERP
OBS ERP, Corr: 0.27 ,RMSE: 1.3 ,SNR: 11
SC-PLS ERP, Corr: 0.47 ,RMSE: 1 ,SNR: 27
μV
time,seconds


channel: 64 (POZ)
Original ERP
OBS ERP, Corr: 0.26 ,RMSE: 1.2 ,SNR: 13
SC-PLS ERP, Corr: 0.49 ,RMSE: 0.89 ,SNR: 33
μV
time,seconds


channel: 35 (C1)
Original ERP
OBS ERP, Corr: 0.043 ,RMSE: 1.8 ,SNR: -0.84
SC-PLS ERP, Corr: -0.13 ,RMSE: 1.4 ,SNR: 27
μV
time,seconds


channel: 47 (F5)
Original ERP
OBS ERP, Corr: 0.066 ,RMSE: 1.6 ,SNR: -12
SC-PLS ERP, Corr: 0.33 ,RMSE: 1 ,SNR: 20
μV
time,seconds

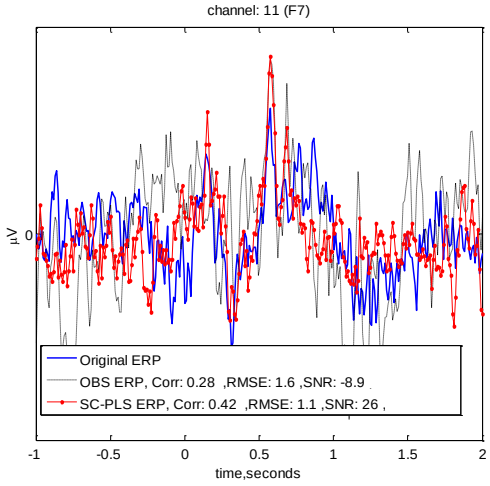

channel: 11 (F7)
Original ERP
OBS ERP, Corr: 0.28 ,RMSE: 1.6 ,SNR: -8.9
SC-PLS ERP, Corr: 0.42 ,RMSE: 1.1 ,SNR: 26
μV
time,seconds

Figure 5-53: Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP, (subject L-X, Experimental MRI study).

5.7.8 Subject "L-X" Simulation MRI study:

Following figures show the simulation study results for patient L-X , we used both SC-PLS and OBS algorithms to clean the EEG data contaminated with BCG artifacts, following figures show the results :

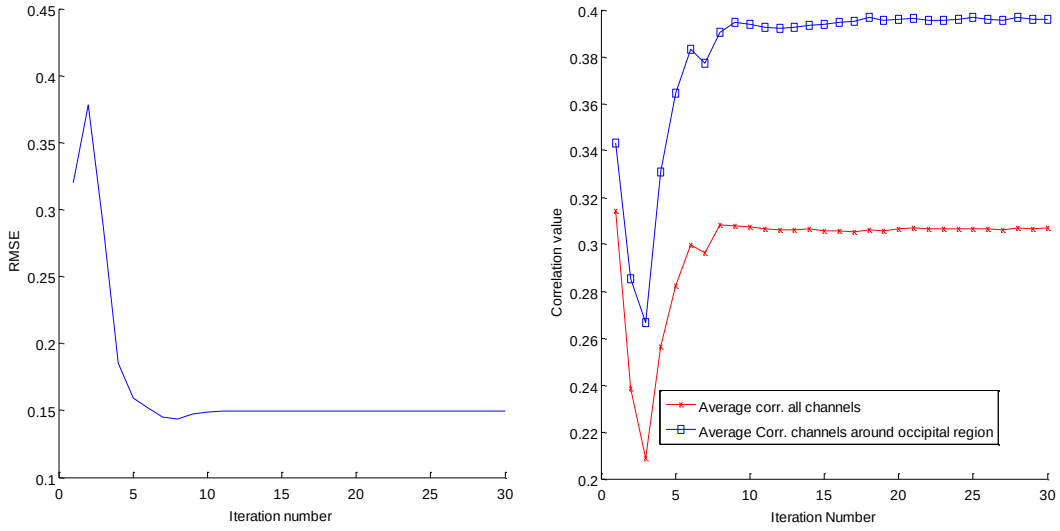


Figure 5-54: : Left: averaged RMSE between Reference ERP values and the denoised ERP values using SC-PLS, at each iteration step. Right: averaged correlation values for all the channels and the channels near the occipital lobe at each iteration step (subject L-X, simulation MRI study)

Following figures show the results from various channels cleaned using OBS and SC-PLS:

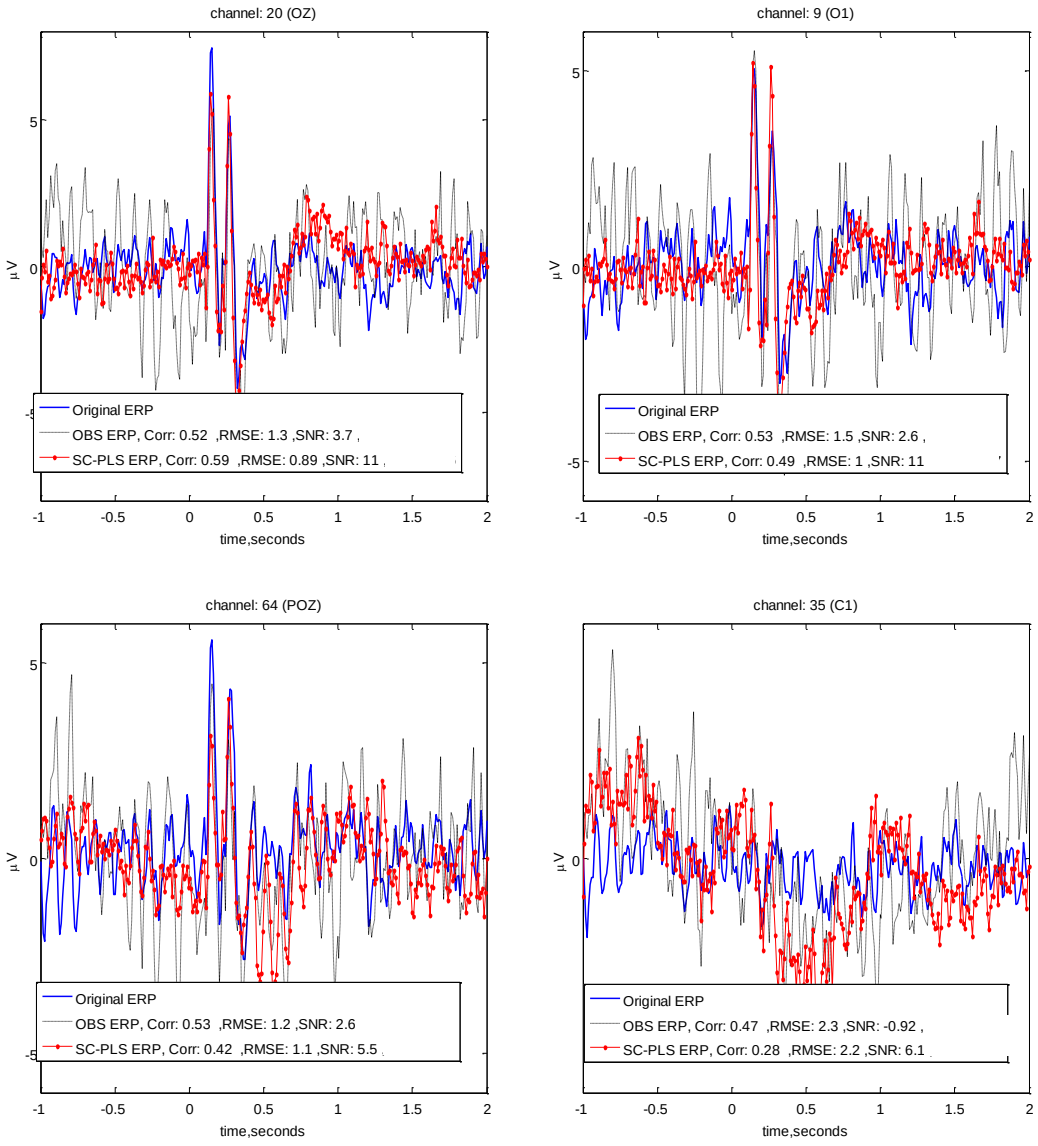


Figure 5-55: : Averaged ERP values before and after removing noise from the EEG data in channels 9,10,64,35,47 and 11. Solid blue curve shows the reference ERP (subject L-X, simulation MRI study)

REFERENCES

(1) Niedermeyer, E.; Silva, F. L. da *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*; Lippincott Williams & Wilkins, 2005.

(2) Liu, A. K.; Dale, A. M.; Belliveau, J. W. Monte Carlo simulation studies of EEG and MEG localization accuracy. *Human brain mapping* **2002**, *16*, 47–62.

(3) Hegerl, U.; Juckel, G. Identifying psychiatric patients with serotonergic dysfunctions by event-related potentials. *World Journal of Biological Psychiatry* **2000**, *1*, 112–118.

(4) Malaspina, D.; Devanand, D.; Krueger, R. B.; Prudic, J. The significance of clinical EEG abnormalities in depressed patients treated with ECT. *Convulsive therapy* **1994**.

(5) Luthringer, R.; Dago, K.; Patat, A.; Caille, P.; Curet, O.; Durieu, G.; Rinaudo, G.; Toussaint, M.; Granier, L.; Macher, J. Pharmacoelectroencephalographic profile of befloxatone, a new reversible MAO-A inhibitor, in healthy subjects. *Neuropsychobiology* **1996**, *34*, 98–105.

(6) Saletu, B.; Anderer, P.; Saletu-Zyhlarz, G.; Arnold, O.; Pascual-Marqui, R. Classification and evaluation of the pharmacodynamics of psychotropic drugs by single-lead pharmaco-EEG, EEG mapping and tomography (LORETA). *Methods and findings in experimental and clinical pharmacology* **2002**, *24*, 120.

(7) Salek-Haddadi, A.; Friston, K.; Lemieux, L.; Fish, D. Studying spontaneous EEG activity with fMRI. *Brain Res.Rev.* **2003**, *43*, 110–133.

(8) Halchenko12, Y. O.; Hanson, S. J.; Pearlmutter, B. A. Multimodal Integration: fMRI, MRI, EEG, MEG.

(9) Pascual-Marqui, R. D. Review of methods for solving the EEG inverse problem. *International Journal of Bioelectromagnetism* **1999**, *1*, 75–86.

(10) Niessing, J.; Ebisch, B.; Schmidt, K. E.; Niessing, M.; Singer, W.; Galuske, R. A. W. Hemodynamic Signals Correlate Tightly with Synchronized Gamma Oscillations. *Science* **2005**, *309*, 948–951.

(11) Logothetis, N. K.; Pauls, J.; Augath, M.; Trinath, T.; Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **2001**, *412*, 150–157.

(12) Leuchter, A. F.; Uijtdehaage, S. H. J.; Cook, I. A.; O'Hara, R.; Mandelkern, M. Relationship between brain electrical activity and cortical perfusion in normal subjects. *Psychiatry Research: Neuroimaging* **1999**, *90*, 125–140.

(13) Buxton, R. B.; Frank, L. R. A Model for the Coupling Between Cerebral Blood Flow and Oxygen Metabolism During Neural Stimulation. *Journal of Cerebral Blood Flow & Metabolism* **1997**, *17*, 64–72.

(14) Arthurs, O. J.; Boniface, S. How well do we understand the neural origins of the fMRI BOLD signal? *Trends in neurosciences* **2002**, *25*, 27–31.

(15) LAI, S.; HOPKINS, A.; HAACKE, E.; LI, D.; WASSERMAN, B.; BUCKLEY, P.; FRIEDMAN, L.; MELTZER, H.; HEDERA, P.; FRIEDLAND, R. Identification of vascular structures as a major source of signal contrast in high resolution 2 D and 3 D functional activation imaging of the motor cortex at 1. 5 T: preliminary results. *Magnetic resonance in medicine* **1993**, *30*, 387–392.

(16) Bartsch, A. J.; Homola, G.; Biller, A.; Solymosi, L.; Bendszus, M. Diagnostic functional MRI: Illustrated clinical applications and decision-making. *JOURNAL OF MAGNETIC RESONANCE IMAGING* **2006**, *23*, 921.

(17) Langenecker, S. A.; Kennedy, S. E.; Guidotti, L. M.; Briceno, E. M.; Own, L. S.; Hooven, T.; Young, E. A.; Akil, H.; Noll, D. C.; Zubieta, J. K. Frontal and limbic activation during inhibitory control predicts treatment response in major depressive disorder. *Biological psychiatry* **2007**, *62*, 1272–1280.

(18) Benar, C. G.; Grova, C.; Kobayashi, E.; Bagshaw, A. P.; Aghakhani, Y.; Dubeau, F.; Gotman, J. EEG-fMRI of epileptic spikes: concordance with EEG source localization and intracranial EEG. *Neuroimage* **2006**, *30*, 1161–1170.

(19) Gotman, J.; Kobayashi, E.; Bagshaw, A. P.; Benar, C. G.; Dubeau, F. Combining EEG and fMRI: A multimodal tool for epilepsy research. *JOURNAL OF MAGNETIC RESONANCE IMAGING* **2006**, *23*, 906.

(20) Gotman, J.; Bénar, C. G.; Dubeau, F. Combining EEG and fMRI in Epilepsy: Methodological Challenges and Clinical Results. *Journal of Clinical Neurophysiology* **2004**, *21*, 229.

(21) Menon, V.; Crottaz-Herbette, S. Combined EEG and fMRI Studies of Human Brain Function. *International review of neurobiology* **2005**, *66*, 292.

(22) Goldman, R. I.; Stern, J. M.; Engel, J.; Cohen, M. S. Acquiring simultaneous EEG and functional MRI. *Clinical Neurophysiology* **2000**, *111*, 1974–1980.

(23) CHOU, C.; BASSEN, H.; OSEPCHUK, J.; BALZANO, Q.; PETERSEN, R.; MELTZ, M.; CLEVELAND, R.; LIN, J.; HEYNICK, L. RADIO FREQUENCY ELECTROMAGNETIC EXPOSURE: TUTORIAL REVIEW ON EXPERIMENTALDOSIMETRY. *Bioelectromagnetics* **1996**, *17*, 195–208.

(24) Vasios, C. E.; Angelone, L. M.; Purdon, P. L.; Ahveninen, J.; Belliveau, J. W.; Bonmassar, G. EEG/(f) MRI measurements at 7 Tesla using a new EEG cap. *Neuroimage* **2006**, *33*, 1082–1092.

(25) Allen, P. J.; Josephs, O.; Turner, R. A Method for Removing Imaging Artifact from Continuous EEG Recorded during Functional MRI. *NeuroImage* **2000**, *12*, 230–239.

(26) Gotman, J.; Bénar, C. G.; Dubeau, F. Combining EEG and fMRI in Epilepsy: Methodological Challenges and Clinical Results. *Journal of Clinical Neurophysiology* **2004**, *21*, 229.

(27) Bonmassar, G.; Anami, K.; Ives, J.; Belliveau, J. W. Visual evoked potential (VEP) measured by simultaneous 64-channel EEG and 3T fMRI. *Neuroreport* **1999**, *10*, 1893.

(28) Niazy, R.; Beckmann, C.; Iannetti, G.; Brady, J.; Smith, S. Removal of FMRI environment artifacts from EEG data using optimal basis sets. *Neuroimage* **2005**, *28*, 720–737.

(29) Allen, P. J.; Polizzi, G.; Krakow, K.; Fish, D. R.; Lemieux, L. Identification of EEG Events in the MR Scanner: The Problem of Pulse Artifact and a Method for Its Subtraction. *NeuroImage* **1998**, *8*, 229–239.

(30) Mantini, D.; Perrucci, M. G.; Cugini, S.; Ferretti, A.; Romani, G. L.; Del Gratta, C. Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis. *NeuroImage* **2007**, *34*, 598–607.

(31) Srivastava, G.; Crottaz-Herbette, S.; Lau, K. M.; Glover, G. H.; Menon, V. ICA-based procedures for removing ballistocardiogram artifacts from EEG data acquired in the MRI scanner. *NeuroImage* **2005**, *24*, 50–60.

(32) Debener, S.; Ullsperger, M.; Siegel, M.; Engel, A. K. Single-trial EEG–fMRI reveals the dynamics of cognitive function. *Trends in cognitive sciences* **2006**, *10*, 558–563.

(33) Jung, T. P.; Makeig, S.; Humphries, C.; Lee, T. W.; Mckeown, M. J.; Iragui, V.; Sejnowski, T. J. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* **2000**, *37*, 163–178.

(34) Jung, T. P.; Makeig, S.; Westerfield, M.; Townsend, J.; Courchesne, E.; Sejnowski, T. J. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology* **2000**, *111*, 1745–1758.

(35) Burnham, A. J.; Viveros, R.; MacGregor, J. F. Frameworks for latent variable multivariate regression. *Journal of chemometrics* **1996**, *10*, 31–45.

(36) Rao, C. R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics, Series A* **1964**, *26*, 329–358.

(37) Herwig, U.; Satrapi, P.; Schönfeldt-Lecuona, C. Using the international 10-20 EEG system for positioning of transcranial magnetic stimulation. *Brain topography* **2003**, *16*, 95–99.

(38) Delorme, A.; Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods* **2004**, *134*, 9–21.

(39) Hoedlmoser, K.; Griessenberger, H.; Fellinger, R.; Freunberger, R.; Klimesch, W.; Gruber, W.; Schabus, M. Event- related activity and phase locking during a psychomotor vigilance task over the course of sleep deprivation. *Journal of Sleep Research* **2011**, *20*, 377–385.

(40) Schimmel, H. The (±) Reference: Accuracy of Estimated Mean Components in Average Response Studies. *Science* **1967**, *157*, 92 –94.

Chapter 6
Conclusion and Future work

This thesis introduced new methods for data regression and latent component selection in the presence of structured noise. These methods have applications in signal, image processing and optimization which need to be further explored.

In chapter two we introduced the constrained, linear latent variable methods which exploit additional knowledge about the noise to suppress its effect. These method are superior to regular projection-out method as they are more resistant to presence of random noise in collected auxiliary noise matrix and provide the same structure as regular LVM methods. The hard constrained method is useful when the collected noise matrix is well conditioned and the data collected is orthogonal to the signal subspace. The soft constrained method introduced provides a flexible way of noise removing when the noise matrix is not collected in the best way possible. We tested the efficiency and performance of these algorithms against regular PLS and OSC PLS algorithm using simulation data. in all cases the constrained methods outperformed the other methods when the structured noise was present in the system. Throughout the simulations we realized that the choice of the penalty term can change the outcome of the analysis. We mentioned earlier that the choice of penalty is related to the ratio of the eigenvalues of the datasets; however, we only investigated one case (SC-PCR) other cases such as SC-PLS need also be investigated to clarify the true relationship between component selection and choice of penalty term. In the appendix of Chapter Two several

other constrained, LVM methods have been introduced and the framework has been laid out mathematically. However, the performance of these algorithms need to be further investigated. and additional applications of these methods need to be discovered.

In chapter three we introduced the NIPALS SC-PLS which is a variation of the NIPALS algorithm. Like the original NIPALS algorithm, this method capable of handling large covariance structures as well as the cases when there are missing elements in the dataset. the advantage of the iterative algorithm is that it is much less sensitive to the size of the covariance matrix and hence, is more efficient in extracting principal components in such datasets with large number of variables. We tested the performance of the matrix for various levels of components missing in the dataset. Our simulation results show that the algorithm is much less sensitive to the missing points in either **Y** or **Z** matrices but is more sensitive to the presence of missing elements in **X**.

In chapter four we introduced the constrained, nonlinear, kernel LVM methods. The nonlinear methods introduced provide a simple way to account for complex, nonlinear interaction for model building. These methods use kernel methods to create a general nonlinear transformation of **X** which can be used to regress against **Y**. The advantage of using kernel methods is that the level of nonlinearity is generalized and does not require specific knowledge about the system and the level of nonlinearity can be adjusted using few parameters. We have introduced the nonlinear constrained KPLS algorithms that exploit these properties of the

kernel trick to build a nonlinear relationship between **X**, **Y** and at the same time, constrained the subspaces to be orthogonal to the auxiliary noise matrix **Z.** These methods are well capable of handling structured noise as well as the nonlinear relationship between the components. Again, the performance of these algorithsm were tested against non-constrained KPLS and linear PLS method and we showed that they outperform the latter methods. The choice of kernel parameter requires further investigation and additional methods such as cross validation need to be implemented for proper selection of kernel parameter. Our results show that these methods had very good performance in handling strong nonlinearities and outperformed regular KPLS method when the data was contaminated with structured noise.

In chapter five we implemented the constrained methods for removal of EEG artifacts. We introduced an iterative algorithm that uses SC-PLS to initially estimate a basis for the noise and signal subspaces and gradually, through iterations, improve the basis matrix for the noise and signal. The improved noise subspace matrix is later used to remove the noise from the EEG dataset by means of projection. We tested our algorithm against two types of structured noise contaminating EEG data; muscle artifacts and BCG artifacts. the results were later compared to the conventional noise removal methods such as ICA component rejection for muscle artifacts and optimal basis sets (OBS) for removal of BCG artifacts. We got very good results outperforming both ICA and OBS methods in the simulation data and in the experimental data our method outperformed manual

ICA rejection. In the experimental data for BCG artifacts we got comparable results to OBS method. the advantage of our method is that it is minimally operator dependant and unlike OBS method, it does not require the detection of QRS peaks in ECG signal. our studies showed that during the EEG data processing steps the choice of basis (**X**) can change the component selection outcome. For example, if instead of original **X** data another basis such as the ICA decomposition or PCA decomposition is used the results will change slightly, and then again the choice of penalty term can be influential in component selection.

In summary further research needs to be done on the role of the penalty coefficients in the SC-PLS algorithm to understand its behavior, the application of these methods for visualization, control and monitoring need to be further investigated, the stopping criteria in the iterative SC-PLS algorithm needs to be further improved and further applications for SC and HC algorithms need to be discovered.