Partial Sufficient Dimension Reduction in Regression

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Do Hyang Kim

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

R. Dennis Cook

June, 2011

Acknowledgements

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, my utmost gratitude goes to my adviser, R. Dennis Cook. With his encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of sufficient dimension reduction, I was able to complete this work. He was and remains my ultimate role model for a scientist, mentor, and teacher in my life. I also would like to thank my thesis committees, Prof. Meeden, Prof. Weisberg, and Prof. Reilly for their direction, dedication, and invaluable advice along this work.

I am deeply indebted to my many colleagues for providing a stimulating and fun environment in which to learn and grow. I am especially grateful to Leif, Sai, and Ka Young who always make me laugh. I have never been tired or bored in the office because of them. I also thank Dimension Reduction group members, Zhihua, Xin, Shanshan, Sen, Jie, Eric, Adam, and Liliana. They helped me to understand the ideas of research area and construct the effective and efficient group web site. I could complete the work on a plate as the group leader because of their continuous support and valuable advices.

Last but not the least, I would like to thank my family: my parents Manjin Kim and Miran Kim, for giving birth to me at the first place and supporting me spiritually throughout my life, and my younger brother Jihyuk, for sharing fresh thoughts and stimulating me to challenge for something new. As one of the family members, I want to give my special thanks to my roommate, Yonsil. She always helped me in every

possible way physically or emotionally to get over difficult situations and loneliness.

I offer my regards and blessings to all of those who supported me in any respect during the completion of this study.

Do Hyang Kim

Dedication

To

My family

For their constant support and unconditional love

I love you all dearly.

Abstract

In this thesis we propose a new model-based reduction method to reduce the dimension of one set of predictors while maintaining another set of predictors and a response if the response is present. Based on the probabilistic PCA model (Tipping and Bishop 1999) and the PFC model (Cook 2007), we develop new models in the partial dimension reduction context: partial probabilistic PCA models, partial PFC models, and combining models. We estimate the parameters of interest for the partial sufficient reduction using the maximum likelihood method. Methods are also proposed for prediction in partial PFC models.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Dimension reduction is a prevalent theme throughout the applied sciences, including statistics, engineering, astronomy, biology, remote sensing, economics, and consumer transactions. In particular, dimension reduction for regression has been extensively studied and is still an active research area since the introduction of sliced inverse regression (SIR, Li 1991) and sliced average variance estimation (SAVE, Cook and Weisberg 1991). Consider the regression of a univariate response Y on a $p \times 1$ random vector of predictors $\mathbf{X}$. A common goal of SIR, SAVE, and many other dimension reduction methods is to reduce the dimension of the predictor vector $\mathbf{X}$ without loss of information on the response Y. That is, the reduced vector $R(\mathbf{X}) : \mathbb{R}^p \to \mathbb{R}^q,\ q \leq p$, should capture all of the information that $\mathbf{X}$ contains about Y. The action of replacing $\mathbf{X}$ with a lower dimensional function $R(\mathbf{X})$ is *dimension reduction*. Furthermore, when $R(\mathbf{X})$ retains all the relevant information about Y, it is called *sufficient dimension reduction*. Cook (2007) described a general paradigm for sufficient reduction in regression, which emerges from the following definition:

Definition 1.1. *A reduction $R : \mathbb{R}^p \to \mathbb{R}^q,\ q \leq p$, is sufficient if it satisfies one of the*

following three statements:

$$\textit{(i) inverse reduction, } \mathbf{X}|(Y, R(\mathbf{X})) \sim \mathbf{X}|R(\mathbf{X}),$$
$$\textit{(ii) forward reduction, } Y|\mathbf{X} \sim Y|R(\mathbf{X}),$$
$$\textit{(iii) joint reduction, } Y \perp\!\!\!\perp \mathbf{X}|R(\mathbf{X}),$$

where $\perp\!\!\!\perp$ *indicates independence,* $\sim$ *means identically distributed and* $\mathbf{A}|\mathbf{B}$ *refers to the random vector* $\mathbf{A}$ *given the vector* $\mathbf{B}$*.*

If we consider a generic statistical problem and reinterpret $\mathbf{X}$ as the total data D and Y as the parameter θ, then condition *(i)* for inverse reduction becomes $D|(\theta, R) \sim D|R$ so that R is a sufficient statistic. In this way, the notion of a sufficient reduction encompasses Fisher's concept of sufficiency: If D represents the data, then a statistic $t(D)$ is sufficient if $D|(\theta, t) \sim D|t$ so that t contains all of the relevant information about θ. One difference is that sufficient statistics are observable, while a sufficient reduction may contain unknown parameters and thus needs to be estimated. The choice of a reductive paradigm depends on the stochastic nature of $\mathbf{X}$ and Y. If the values of $\mathbf{X}$ are fixed by design, then forward regression *(ii)* seems the natural choice. In discriminant analysis $\mathbf{X}|Y$ is a random vector of features observed in one of a number of subpopulations indicated by Y. If the values of Y are fixed by design then inverse regression *(i)* is perhaps the only reasonable reductive route. The three statements in Definition 1 are equivalent when $(Y, \mathbf{X})$ has a joint distribution. For example, we may determine a sufficient reduction from $\mathbf{X}|Y$ *(i)* and then pass that reduction to the forward regression *(ii)* or the joint distribution *(iii)* without specifying the marginal distribution of Y or the conditional distribution of $Y|\mathbf{X}$. We assume that Y and $\mathbf{X}$ are jointly distributed in this thesis, unless indicated otherwise.

1.1 Motivation

Conventional sufficient dimension reduction focuses on reducing the whole predictor vector $\mathbf{X}$ indiscriminately. However, we often encounter data which consist of two scientifically distinguishable sets of predictors in a large variety of applications. For

instance Massy (1965) studied the possibility of using 14 income predictors and 9 education predictors to estimate the influence on television and refrigerator ownership. Although he used regression on principal components for all predictors, we may want to reduce only the 14 income predictors in the presence of the response and education predictors or to reduce the income and education predictors simultaneously but separately. In studies of dietary patterns, lengthy questionnaires are often used to measure a subject's characteristics, like food frequency or attitude (see, for example, Butler *et al.* 2006). Here the goal is to reduce the questionnaire data for inclusion in a regression with response and additional predictors. More examples of this type can be found easily in applied science like the dataset distinguishing demographic information from numerical measurements. With this type of dataset, predictor vector $\mathbf{X}$ can be partitioned into two parts, $\mathbf{X}_1$ and $\mathbf{X}_2$ according to the characteristic of variables, and we can easily encounter the problem of reduction on $\mathbf{X}_1$ alone in the presence of the remaining part $\mathbf{X}_2$ and response—*partial dimension reduction*. This also would be of particular interest in applications in which some predictors play a particular role, and must therefore be shielded from the reduction process. This problem can be resolved through partial dimension reduction.

Although many dimension reduction methods have been proposed, less attention has been paid to partial dimension reduction. Chiaromonte, Cook, and Li (2002) and Li, Cook, and Chiaromonte (2003) explored sufficient dimension reduction in regression analyses involving both quantitative and categorical predictor variables. They proposed a moment-based dimension reduction method that is extended from SIR, while focusing on reducing dimension of $\mathbf{X}$ alone incorporating a categorical predictor variable $\mathbf{W}$. However, their method is limited to a categorical $\mathbf{W}$ and is difficult to generalize to continuous $\mathbf{W}$. Moreover, they did not consider the problem of reducing dimensions of both $\mathbf{X}$ and $\mathbf{W}$ simultaneously and separately. Yin and Zhu (2007) dealt with the issue of dimension reduction for both $\mathbf{X}$ and $\mathbf{W}$, where $\mathbf{W}$ can be categorical or continuous. Since their method requires nonparametric density estimation, it encounters practical difficulty when the number of predictors is not small. Li, Li, and Zhu (2010) investigated a groupwise dimension reduction method, moment-based approach, conducting

dimension reduction on the predictors that fall into several groups. They addressed partial dimension reduction as a special case of a groupwise dimension reduction. They proposed an iterative optimization algorithm to find the dimension reduction estimator, which involves the use of multidimensional kernel smoothers. So, this too can sometimes be a concern in high-dimensional setting.

One goal of this thesis is to propose new model-based reduction methods to reduce the dimension of one set of predictors given a specific response Y and another set of predictors, in which the reduced predictors preserve all the relevant information about Y and another set of predictors—*partial sufficient dimension reduction* (PSDR). To our knowledge, no model-based sufficient dimension reduction methods have been developed. We present the first model-based solution, leading to methods for partial sufficient dimension reduction in regressions. Our results build on Cook's (2007) formulation, which uses model-based inverse regression of $\mathbf{X}$ on Y to gain reductive information for the forward regression of Y on $\mathbf{X}$ using a likelihood-based objective function.

We concentrate on $\mathbf{X}_1|(\mathbf{X}_2, Y)$, although the goal is still to reduce the dimension of $\mathbf{X}_1$ with little or no loss of information on $Y|(\mathbf{X}_1, \mathbf{X}_2)$. Model-based forward regression analyses traditionally condition on the observed values of the predictors, even if $\mathbf{X}$ is random. Nevertheless, the conditional distribution of $\mathbf{X}_1|(\mathbf{X}_2, Y)$ may provide a better handle on reductive information since it can be linked usefully to $Y|(\mathbf{X}_1, \mathbf{X}_2)$ (Proposition 4.1 and Proposition 4.2).

1.2 Notation

To facilitate the exposition, let $\mathbf{X}_y$ denote a random variable distributed as $\mathbf{X}|(Y = y)$. $\mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{W}$ means that the random vectors $\mathbf{U}$ and $\mathbf{V}$ are conditionally independent given any value for the random vector $\mathbf{W}$. For positive integers p and q, $\mathbb{R}^{p\times q}$ stands for the class of real $p \times q$ matrices. For $\mathbf{A} \in \mathbb{R}^{p\times p}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, $\mathbf{A}\mathcal{S} \equiv \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{S}\}$. A basis matrix for a subspace $\mathcal{S}$ is any semi-orthogonal matrix whose columns are a basis for $\mathcal{S}$, where a semiorthogonal matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, $q < p$, has orthogonal columns, $\mathbf{A}^T\mathbf{A} = I_q$.

For $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathcal{S}_{\mathbf{B}} \equiv \text{span}(\mathbf{B})$ denotes the subspace of $\mathbb{R}^p$ spanned by the columns of $\mathbf{B}$. If $\mathbf{B} \in \mathbb{R}^{p\times q}$ and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is symmetric and positive definite, then the projection onto $\mathcal{S}_{\mathbf{B}}$ relative to $\mathbf{\Sigma}$ has the matrix representation $\mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})} \equiv \mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{\Sigma}$. $\mathbf{P}_{\mathcal{S}}$ indicates the projection onto the subspace $\mathcal{S}$ in the usual inner product, and $\mathbf{Q}_{\mathcal{S}} = \mathbf{I} - \mathbf{P}_{\mathcal{S}}$.

Let $\mathcal{S}_d(\mathbf{C}, \mathbf{D})$ denote the span of $\mathbf{C}^{-1/2}$ times the first d eigenvectors of $\mathbf{C}^{-1/2}\mathbf{D}\mathbf{C}^{-1/2}$, where $\mathbf{C}$ and $\mathbf{D}$ are symmetric matrices and, as used in this thesis, $\mathbf{C}$ will always be a nonsingular covariance matrix. Beginning with $\mathbf{D}$ we apply the transformation $\mathbf{C}^{-1/2}$ before computing the first d eigenvectors. Multiplying these eigenvector by $\mathbf{C}^{-1/2}$ then converts them to vectors that span the desired subspace in the original scale. The subspace $\mathcal{S}_d(\mathbf{C}, \mathbf{D})$ can also be described as the span of the first d eigenvectors of $\mathbf{D}$ relative to $\mathbf{C}$.

For notation convenience let $\mathbf{1}_p = (1, \ldots, 1)^T$ and $\mathbf{0}_p = (0, \ldots, 0)^T$ p-vectors with entries 1 and 0 respectively. Likewise $\mathbf{1}_{p\times q}$ and $\mathbf{0}_{p\times q}$ are $p \times q$ matrices with corresponding entries 1 and 0. Let $\lambda_i(\mathbf{Z})$ indicate the ith largest eigenvalue of matrix $\mathbf{Z}$. The symbol $\otimes$ denotes the Kronecker product, and $\oplus$ indicates the direct sum between two subspaces ($V_1 \oplus V_2 = \{v_1 + v_2;\ v_1 \in V_1,\ v_2 \in V_2\}$).

1.3 Outline

In this thesis we propose a new model-based reduction method to reduce the dimension of one set of predictors while maintaining another set of predictors and a response Y if the response is present. Starting with the probabilistic PCA model proposed by Tipping and Bishop (1999), we develop new models in the partial dimension reduction context. The methods of parameter estimation basically build on Cook's (2007) formulation, which uses maximum likelihood method under the inverse regression of $\mathbf{X}$ on Y. We will extend that to our specific situation, partial sufficient dimension reduction.

In Chapter 2, we review the probabilistic PCA model first and then develop that model to introduce a partial probabilistic PCA model. Partial sufficient dimension reduction will be defined based on the partial probabilistic PCA model.

In Chapter 3, using an orthogonality property of parameters, the partial probabilistic PCA model with isotropic error covariance matrix is specified and the estimators of the parameters on which it depends are obtained using the method of maximum likelihood.

Like the development of the PFC model, the adaptation of the reduction for a specific response Y occurs when introducing the partial PFC model in Chapter 4. In this chapter we consider four partial PFC models according to the type of error covariance matrices: isotropic, diagonal, and unstructured error covariance matrix and specific variance function. In the same manner as in Chapter 3 we estimate the parameters under each partial PFC model using the maximum likelihood method.

Sometimes selecting one particular model might be hard because of the nature of the dataset. For instance, only part of responses may be known. For this concern, we propose in Chapter 5 mixed models combining the isotropic partial probabilistic PCA model and the partial PFC model.

When dealing with a high dimensional predictor space there might be a relatively small set of relevant predictors and a large number of irrelevant ones. Chapter 6 illustrates several methods which can be considered to get rid of a substantial inactive subset of the predictors. In addition, the dimension of the sufficient reduction is inferred using various methods in this chapter.

Based on the models discussed in Chapters 3 through 5, the screening method is proposed in Chapter 6, and prediction methodology is given in Chapter 7.

Chapter 8 contains an illustration of how the proposed methodology might be used in practice. Two kinds of dataset with a number of predictors are used comparing the performance of various methods in the dimension reduction context. Proofs are given in the Appendix.

Chapter 2

Partial probabilistic PCA model

Principal component analysis (PCA) is one of the most widely used multivariate techniques for reducing the dimensionality of a large multivariate data set. The most common goal of PCA is to find uncorrelated linear combinations of the original variables, which account for the maximum variation. Suppose we have a multivariate random sample, represented here by the p-dimensional data vectors $\{\mathbf{X}^{(i)}\},\ i = 1, \ldots, n$. Then the q principal components of $\mathbf{X}$ are defined as linear combinations $\mathbf{v}_j^T\mathbf{X},\ j = 1, \ldots, q$, where the principal axes $\mathbf{v}_j$ are p-dimensional vectors which achieve the maximum of $\mathrm{Var}(\mathbf{v}_j^T\mathbf{X})$ among all linear combinations of $\mathbf{X}$ successively subject to $\mathbf{v}_j^T\mathbf{v}_j = 1$ and $\mathbf{v}_k^T\mathbf{v}_j = 0$ for $\forall k < j$. It can be shown that the vectors $\mathbf{v}_j,\ j = 1, \ldots, q$, are given by the q dominant eigenvectors of the covariance matrix of $\mathbf{X}$, which correspond to the q largest eigenvalues. Thus, in terms of dimension reduction, the basic idea of PCA method is to replace the vector $\mathbf{X}$ with a few of the principal components $\hat{\mathbf{v}}_j^T\mathbf{X},\ j = 1, \ldots, q\ (q < p)$, where $\hat{\mathbf{v}}_j^T$ is the eigenvector of the sample covariance matrix of $\mathbf{X}$ corresponding to its jth largest eigenvalue. The leading principal components, those corresponding to the larger eigenvalues, are typically chosen and have a number of properties that may be helpful for the subsequent analysis (Jolliffe, 2002).

PCA had been regarded as a method which is not associated with a probabilistic model for the observed data before probabilistic principal component analysis was proposed by Tipping and Bishop (1999). In that paper, they demonstrated how the

principal axes of a set of observed data vectors are determined through maximum likelihood estimation using a latent variable model. Subsequently, PCA was revisited by many authors with the probabilistic modeling point of view.

In this chapter we will review probabilistic principal component analysis briefly and then develop a new partial probabilistic PCA model using a text book conditional distribution theorem. As there is no response involved, this chapter is about unsupervised multivariate dimension reduction methods.

In order to distinguish the parameters in the non-partial model from those in the partial model, we put an asterisk to the parameters used in non-partial model throughout this thesis. The use of the error term $\boldsymbol{\epsilon}$ is different from $\boldsymbol{\varepsilon}$ in the thesis. The $p \times 1$ vector $\boldsymbol{\epsilon}$ is used for the error in non-partial models and $\boldsymbol{\varepsilon} \in \mathbb{R}^{p_1}$ is used in partial models.

2.1 Probabilistic principal component analysis revisited

A latent variable model that is closely related to factor analysis was proposed by Tipping and Bishop (1999) as follows:

$$\mathbf{X} = \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^* \boldsymbol{\nu}^* + \sigma \boldsymbol{\epsilon}, \tag{2.1}$$

where the parameter vector $\bar{\boldsymbol{\mu}}^*$ allows the model to have non-zero mean and the matrix $\boldsymbol{\Gamma}^* \in \mathbb{R}^{p \times d^*}$ relates the observable variable $\mathbf{X} \in \mathbb{R}^p$ and the latent variable $\boldsymbol{\nu}^* \in \mathbb{R}^{d^*}$, which is assumed to be normally distributed with mean 0 and identity covariance matrix. The error vector $\boldsymbol{\epsilon} \in \mathbb{R}^p$ is independent of $\boldsymbol{\nu}^*$ and normally distributed with mean 0 and covariance matrix $\mathbf{I}_p$. That is, with this isotropic Gaussian error term $\boldsymbol{\epsilon}$, the conditional distribution of $\mathbf{X}$ given the latent variables $\boldsymbol{\nu}^*$ is $\mathbf{X}|\boldsymbol{\nu}^* \sim \mathcal{N}(\boldsymbol{\mu}^*_{\boldsymbol{\nu}}, \sigma^2 \mathbf{I}_p)$, where $\boldsymbol{\mu}^*_{\boldsymbol{\nu}} = \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^* \boldsymbol{\nu}^*$. Model (2.1) can also be regarded as a formal statement about the regression of $\mathbf{X}$ on $\boldsymbol{\nu}^*$ when $\boldsymbol{\nu}^*$ is fixed. The latent variable $\boldsymbol{\nu}^*$ is intended to explain the correlation between observation variables (extrinsic variation) while $\boldsymbol{\epsilon}$ represents variability unique to a particular $\mathbf{X}^{(i)}$ (intrinsic variation).

We consider the dimension reduction problem based on model (2.1), but instead of

assuming an isotropic error structure, which is limiting, we assume

$$\mathbf{X} = \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^* \boldsymbol{\nu}^* + \boldsymbol{\Delta}^{1/2} \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{\Delta}$ is a positive definite matrix (Chen 2010). By positing various structures for $\boldsymbol{\Delta}$ we obtain a flexible class of models for dimension reduction. Intuitively, the dimension d^* of $\boldsymbol{\nu}^*$ will decrease as we allow more flexibility in $\boldsymbol{\Delta}$. For instance, if $\boldsymbol{\Delta}$ is a general positive definite matrix then the extrinsic variation is absorbed by the intrinsic variation and $d^* = 0$. Likewise, if we still require $\boldsymbol{\nu}^*$ to be normal with mean 0, but no longer require $\mathrm{Var}(\boldsymbol{\nu}^*) = \mathbf{I}$, then specific structures for $\boldsymbol{\Delta}$ can be considered. That is, $\boldsymbol{\nu}^*$ can be any unknown variable with a positive definite covariance matrix while letting $\boldsymbol{\Delta}$ have a special structure such as $\sigma^2\mathbf{I}$. Consequently, the investigator can tailor the partitioning of the variance of $\mathbf{X}$ between intrinsic and extrinsic variation, as appropriate for the application at hand.

Under model (2.2) $R(\mathbf{X}) = \boldsymbol{\Gamma}^{*T}\boldsymbol{\Delta}^{-1}\mathbf{X}$ is a sufficient reduction (Cook 2007) satisfying $\mathbf{X} \perp\!\!\!\perp \boldsymbol{\nu}^* | \boldsymbol{\Gamma}^{*T}\boldsymbol{\Delta}^{-1}\mathbf{X}$ based on the Definition 1.1. Here the parameter $\boldsymbol{\Gamma}^*$ is not identified under the model since multiplication by any nonsingular matrix $\mathbf{A} \in \mathbb{R}^{d^* \times d^*}$ leads to an equivalent model with different parameters, $\boldsymbol{\Gamma}^*\boldsymbol{\nu}^* = (\boldsymbol{\Gamma}^*\mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\nu}^*)$. However, the *reductive subspace* $\mathrm{span}(\boldsymbol{\Gamma}^*)$ is identified and estimable, and for this reason we assume without loss of generality that $\boldsymbol{\Gamma}^*$ is a semi-orthogonal matrix, $\boldsymbol{\Gamma}^{*T}\boldsymbol{\Gamma}^* = \mathbf{I}_{d^*}$. Therefore the goal is to estimate the *dimension reduction subspace* $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}^*} = \{\boldsymbol{\Delta}^{-1}\mathbf{z} : \mathbf{z} \in \mathcal{S}_{\boldsymbol{\Gamma}^*}\}$ on its parameter space in $\mathbb{R}^p$ with dimension $d^* \leq p$. This parameter space is a hyperplane through the origin and the set of such planes is called a Grassmann manifold $\mathcal{G}_{(d^*,\, p)}$ in $\mathbb{R}^p$, which is uniquely determined by choosing $d^*(p - d^*)$ real numbers. For background on Grassmann manifolds, see Edelman, Arias, and Smith (1998) and Chikuse (2003). Estimating $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}^*}$ on the d^*-dimensional Grassmann manifold $\mathcal{G}_{(d^*,\, p)}$ in $\mathbb{R}^p$ is our main goal.

2.2 Partial probabilistic PCA model

We assume that the predictor vector $\mathbf{X}$ is partitioned into $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$, $p_1 + p_2 = p$. The goal is to reduce the dimension of $\mathbf{X}_1$ in the presence of the latent variable $\boldsymbol{\nu}^*$ and the other predictors $\mathbf{X}_2$. Suppose that the partitioned $\mathbf{X}$ follows model (2.2) with $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$, $\boldsymbol{\Gamma}^* = (\boldsymbol{\Gamma}_1^{*T}, \boldsymbol{\Gamma}_2^{*T})^T$, and $\boldsymbol{\Delta} = ((\boldsymbol{\Delta}_{11}, \boldsymbol{\Delta}_{21})^T, (\boldsymbol{\Delta}_{12}, \boldsymbol{\Delta}_{22})^T)$. Then $\mathbf{X}_1|\boldsymbol{\nu}^*$ is normally distributed with mean $\bar{\boldsymbol{\mu}}_1^* + \boldsymbol{\Gamma}_1^*\boldsymbol{\nu}^*$ and constant variance $\boldsymbol{\Delta}_{11}$, and $\mathbf{X}_2|\boldsymbol{\nu}^*$ is normally distributed with mean $\bar{\boldsymbol{\mu}}_2^* + \boldsymbol{\Gamma}_2^*\boldsymbol{\nu}^*$ and constant variance $\boldsymbol{\Delta}_{22}$. Now we can consider the textbook conditional distribution theorem below (Johnson and Wichern 2007, Result 4.6.).

Theorem 2.1. *(Conditional Distribution) Assume an n-dimensional random vector* $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ *has a normal distribution* $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$ *where* $\mathbf{X}_1, \mathbf{X}_2$ *are two subvectors of respective dimensions* q *and* r *with* $q + r = n$. *Note that* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$, $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{21}^T$. *Then the conditional distribution of* $\mathbf{X}_1$ *given* $\mathbf{X}_2$ *is also normal with mean vector* $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ *and covariance matrix* $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

According to the Theorem 1, the conditional distribution $\mathbf{X}_1$ given $\mathbf{X}_2$ and $\boldsymbol{\nu}^*$ is normal with mean

$$\begin{aligned} & E(\mathbf{X}_1|\boldsymbol{\nu}^*) + \boldsymbol{\Delta}_{12}\boldsymbol{\Delta}_{22}^{-1}(\mathbf{X}_2 - E(\mathbf{X}_2|\boldsymbol{\nu}^*)) \\ &= \bar{\boldsymbol{\mu}}_1^* + \boldsymbol{\Gamma}_1^*\boldsymbol{\nu}^* + \underbrace{\boldsymbol{\Delta}_{12}\boldsymbol{\Delta}_{22}^{-1}}_{\boldsymbol{\beta}}(\mathbf{X}_2 - \bar{\boldsymbol{\mu}}_2^* - \boldsymbol{\Gamma}_2^*\boldsymbol{\nu}^*) \\ &= \underbrace{\bar{\boldsymbol{\mu}}_1^* - \boldsymbol{\beta}\bar{\boldsymbol{\mu}}_2^*}_{\bar{\boldsymbol{\mu}}_1} + \boldsymbol{\beta}\mathbf{X}_2 + (\underbrace{\boldsymbol{\Gamma}_1^* - \boldsymbol{\beta}\boldsymbol{\Gamma}_2^*}_{\boldsymbol{\Gamma}})\boldsymbol{\nu}^* \\ &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu}^* \end{aligned} \tag{2.3}$$

and covariance matrix $\boldsymbol{\Delta}_{11} - \boldsymbol{\Delta}_{12}\boldsymbol{\Delta}_{22}^{-1}\boldsymbol{\Delta}_{21} \stackrel{\text{def}}{=} \boldsymbol{\Psi}$. So, we obtain the new partial probabilistic model for $\mathbf{X}_1$ given $\mathbf{X}_2$ and $\boldsymbol{\nu}^*$,

$$\mathbf{X}_1|(\mathbf{X}_2, \boldsymbol{\nu}^*) = \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu}^* + \boldsymbol{\Psi}^{1/2}\boldsymbol{\varepsilon}, \tag{2.4}$$

where $\boldsymbol{\varepsilon}$ is normally distributed with mean 0 and covariance matrix $\mathbf{I}_{p_1}$. Here $\bar{\boldsymbol{\mu}}_1 \in \mathbb{R}^{p_1}$, $\boldsymbol{\beta} \in \mathbb{R}^{p_1 \times p_2}$, and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$. Although model (2.4) is derived from Theorem 1 in which $\mathbf{X}_2$ follows normal distribution, as a matter of fact $\mathbf{X}_2$ does not need to be normal for model (2.4) to hold. For instance, $\mathbf{X}_2$ could be binary variables indicating group characteristics such as gender or treatment status.

If $\mathbf{X}|\boldsymbol{\nu}^*$ has isotropic covariance matrix; that is, model (2.1) is considered as the original model in the derivation of the conditional distribution $\mathbf{X}_1$, then $\boldsymbol{\Delta}_{12} = 0$ and $\boldsymbol{\beta} = \boldsymbol{\Delta}_{12}\boldsymbol{\Delta}_{22}^{-1} = 0$ in the equation (2.3). Consequently model (2.4) becomes the conditional model $\mathbf{X}_1|\boldsymbol{\nu}^*$ not depending on $\mathbf{X}_2$ anymore. By positing various error covariance structure in the original model, however, our partial probabilistic model allows dependence between $\mathbf{X}_1$ and $\mathbf{X}_2$ given $\boldsymbol{\nu}^*$ and is thus less restrictive than ordinary probabilistic PCA model.

Define $\boldsymbol{\Gamma} \in \mathbb{R}^{p_1 \times d}$, $d \leq p_1$, to be a basis matrix for $\text{span}(\boldsymbol{\Gamma}_1^* - \boldsymbol{\beta}\boldsymbol{\Gamma}_2^*)$ and assume without loss of generality that $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$; that is, $\text{rank}(\boldsymbol{\Gamma}) = d \leq \min(p_1, d^*)$ and d becomes an unknown parameter. Using this $\boldsymbol{\Gamma}$ we can reformulate model (2.4) as

$$\mathbf{X}_1|(\mathbf{X}_2, \boldsymbol{\nu}') = \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu}' + \boldsymbol{\Psi}^{1/2}\boldsymbol{\varepsilon}, \tag{2.5}$$

where $\boldsymbol{\nu}' \in \mathbb{R}^d$ is the corresponding coordinate vector and has a positive definite covariance matrix. We will refer to model (2.5) as a *partial probabilistic model*. Our next task is to derive the sufficient reduction for model (2.5). The following proposition connects the partial probabilistic model (2.5) with the conditional distribution of the latent variable $\boldsymbol{\nu}'$ and $\mathbf{X}_2$ given $\mathbf{X}_1$. Its proof is given in Appendix A.1.

Proposition 2.1. *Under model* (2.5)*, the distribution of* $(\mathbf{X}_2,\ \boldsymbol{\nu}')|\mathbf{X}_1$ *is the same as the distribution of* $(\mathbf{X}_2,\ \boldsymbol{\nu}')|(\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1$ *for all values of* $\mathbf{X}_1$*. That is,* $R(\mathbf{X}_1) = (\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1$ *is a sufficient reduction. Let* $T(\mathbf{X}_1)$ *be any sufficient reduction. Then, under model* (2.5)*,* R *is a function of* T*.*

According to this proposition, $\mathbf{X}_1$ can be replaced by the sufficient reduction $R(\mathbf{X}_1) = (\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1$ without loss of information about $(\mathbf{X}_2,\ \boldsymbol{\nu}')$ given $\mathbf{X}_1$, and thus $R(\mathbf{X}_1)$ satisfies the conditional independence, $\mathbf{X}_1 \perp\!\!\!\perp (\mathbf{X}_2,\ \boldsymbol{\nu}')|R(\mathbf{X}_1)$. This relationship can be

extended to the following proposition restating the conditional independence proposition (Cook 1998, Proposition 4.6) in terms of $\mathbf{X}_1,\ \mathbf{X}_2,\ \boldsymbol{\nu}'$, and $R(\mathbf{X}_1)$.

Proposition 2.2. *The following pair of conditions a_1 and a_2 is equivalent to the pair of conditions b_1 and b_2 which is equivalent to condition c.*

$$(a_1)\ \mathbf{X}_1 \perp\!\!\!\perp \boldsymbol{\nu}'|(\mathbf{X}_2,\ R(\mathbf{X}_1)) \quad (a_2)\ \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|R(\mathbf{X}_1)$$
$$(b_1)\ \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|(\boldsymbol{\nu}',\ R(\mathbf{X}_1)) \quad (b_2)\ \mathbf{X}_1 \perp\!\!\!\perp \boldsymbol{\nu}'|R(\mathbf{X}_1)$$
$$(c)\ \mathbf{X}_1 \perp\!\!\!\perp (\mathbf{X}_2,\ \boldsymbol{\nu}')|R(\mathbf{X}_1)$$

Conditions (a_1, a_2) and (b_1, b_2) are really redundant, since each is implied by condition c. These conditions are stated separately to emphasize the symmetric roles of $\mathbf{X}_2$ and $\boldsymbol{\nu}'$. Based on the conditional independence condition (c), now we can extend Definition 1.1 motivated by the partial probabilistic model:

Definition 2.1.a. *Under model* (2.5)*, a partial reduction $R(\mathbf{X}_1)$, with $R : \mathbb{R}^{p_1} \to \mathbb{R}^{q_1},\ q_1 \leq p_1$, is sufficient for $(\mathbf{X}_2,\ \boldsymbol{\nu}')$ if it satisfies one of the following three statements:*

$$\textit{(i)}\ \mathbf{X}_1|(\mathbf{X}_2,\ \boldsymbol{\nu}',\ R(\mathbf{X}_1)) \sim \mathbf{X}_1|R(\mathbf{X}_1),$$
$$\textit{(ii)}\ (\mathbf{X}_2,\ \boldsymbol{\nu}')|\mathbf{X}_1 \sim (\mathbf{X}_2,\ \boldsymbol{\nu}')|R(\mathbf{X}_1),$$
$$\textit{(iii)}\ (\mathbf{X}_2,\ \boldsymbol{\nu}') \perp\!\!\!\perp \mathbf{X}_1|R(\mathbf{X}_1).$$

While we have considered the whole predictor set $\mathbf{X}$ in the Definition 1.1, the separated predictor sets, $\mathbf{X}_1$ and $\mathbf{X}_2$, are considered here. Definition 2.1.a can be also interpreted in the same manner as Definition 1.1 while $\mathbf{X}_1$ and $(\mathbf{X}_2, \boldsymbol{\nu}')$ play the role of $\mathbf{X}$ and Y in Definition 1.1 respectively. That is, statement (i) corresponds to partial probabilistic modeling and requires only the conditional distribution $\mathbf{X}_1|(\mathbf{X}_2, \boldsymbol{\nu}')$. Statement (ii) requires only the conditional distribution of $(\mathbf{X}_2, \boldsymbol{\nu}')|\mathbf{X}_1$, while statement (iii) requires the joint distribution of $(\mathbf{X}_1,\ \mathbf{X}_2,\ \boldsymbol{\nu}')$. These statements indicate that once we know $R(\mathbf{X}_1)$, the conditional distribution of $\mathbf{X}_1|(\mathbf{X}_2,\ \boldsymbol{\nu}',\ R(\mathbf{X}_1))$ is completely independent from $\mathbf{X}_2$ and $\boldsymbol{\nu}'$. In other words, $R(\mathbf{X}_1)$ contains all of the relevant information about $\mathbf{X}_2$ and $\boldsymbol{\nu}'$.

Based on the condition (a_1, a_2) in Proposition 2.2, Definition 2.1.a can be restated as follows:

Definition 2.1.b. *Under model* (2.5)*, a partial reduction* $(R(\mathbf{X}_1),\ \mathbf{X}_2)$*, with* $R : \mathbb{R}^{p_1} \to \mathbb{R}^{q_1},\ q_1 \leq p_1$*, is sufficient for* $\boldsymbol{\nu}'$*, and* $R(\mathbf{X}_1)$ *is sufficient for* $\mathbf{X}_2$ *if it satisfies one of the following three statements:*

(i) $\mathbf{X}_1|(\boldsymbol{\nu}',\ \mathbf{X}_2,\ R(\mathbf{X}_1)) \sim \mathbf{X}_1|(\mathbf{X}_2,\ R(\mathbf{X}_1)), \quad \mathbf{X}_1|(\mathbf{X}_2,\ R(\mathbf{X}_1)) \sim \mathbf{X}_1|R(\mathbf{X}_1)$

(ii) $\boldsymbol{\nu}'|(\mathbf{X}_2,\ \mathbf{X}_1) \sim \boldsymbol{\nu}'|(\mathbf{X}_2,\ R(\mathbf{X}_1)), \quad \mathbf{X}_2|\mathbf{X}_1 \sim \mathbf{X}_1|R(\mathbf{X}_1)$

(iii) $\mathbf{X}_1 \perp\!\!\!\perp \boldsymbol{\nu}'|(\mathbf{X}_2,\ R(\mathbf{X}_1)), \quad \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|R(\mathbf{X}_1)$

We can also think of the alternative way to state Definition 2.1.b, based on the conditions (b_1, b_2) in Proposition 2.2. While Definition 2.1.a focuses on the sufficiency for $(\mathbf{X}_2,\ \boldsymbol{\nu}')$, Definition 2.1.b considers the sufficiency for $\boldsymbol{\nu}'$ and $\mathbf{X}_2$ separately.

In the sense of reducing the dimension of $\mathbf{X}_1$ we assume $\text{rank}(\boldsymbol{\beta},\ \boldsymbol{\Gamma})$ is less than p_1, where the partial reduction is $R(\mathbf{X}_1) = (\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1$; that is, we assume $p_1 > p_2 + d^* \geq p_2 + d$. If $\boldsymbol{\beta}$ has full column rank, $\text{rank}(\boldsymbol{\beta}) = p_2$, we only need to consider inference on d, but otherwise we have to infer the rank of $(\boldsymbol{\beta},\ \boldsymbol{\Gamma})$. In this thesis we assume that $\boldsymbol{\beta}$ has full column rank and the other case of $\boldsymbol{\beta}$ will be studied in future work. We hold d fixed in Chapter 3 through Chapter 5 and Chapter 7, and consider inference for d in Chapter 6. Since the sufficient reduction contains unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$, reasonable estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ are needed for the sufficient reduction $R(\mathbf{X}_1)$ to be useful in practice. In the following chapters we will study how to estimate the parameters of interest.

Chapter 3

Isotropic Partial Probabilistic PCA Model and Estimation

In this chapter, beginning with the partial probabilistic PCA model (2.5), we will specify the structure of the model and reformulate it to estimate the parameters of interest. The maximum likelihood method will be used for estimation. The results of a small simulation will be given for better understanding of the isotropic partial probabilistic PCA model. In the last section we will study the use of categorical variables and propose an iterating algorithm to estimate the parameters for sufficient dimension reduction.

3.1 Isotropic partial probabilistic PCA Model

Consider model (2.1) as the regression of $\mathbf{X}$ on $\boldsymbol{\nu}^*$ where $\boldsymbol{\nu}^*$ is fixed and the errors are isotropic. Then model (2.1) is called a *PC model* since a sufficient reduction $R(\mathbf{X}) = \boldsymbol{\Gamma}^{*T}\mathbf{X}$ is estimated by the first d^* sample principal components of the sample covariance matrix of $\mathbf{X}$ (Cook 2007). Adopting the same idea of getting a sufficient reduction from the PC model, we consider an isotropic version of the partial probabilistic PCA model (2.5), which is the regression of $\mathbf{X}_1$ on $(\mathbf{X}_2,\ \boldsymbol{\nu}')$ with the isotropic error covariance matrix, and then estimate a sufficient reduction of $\mathbf{X}_1$ based on that model.

Assume that we have isotropic errors in model (2.5); that is, $\boldsymbol{\Psi} = \sigma^2\mathbf{I}_{p_1}$. Then the

sufficient reduction is $R(\mathbf{X}_1) = (\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\mathbf{X}_1$ from Proposition 1 and $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ are the focus of our inquiry. However, since $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}$ are confounded and we cannot estimate them simultaneously, we need to distinguish $\boldsymbol{\beta}$ from span($\boldsymbol{\Gamma}$). Substituting $\boldsymbol{\beta} = (\mathbf{Q}_{\boldsymbol{\Gamma}} + \mathbf{P}_{\boldsymbol{\Gamma}})\boldsymbol{\beta}$ and using the orthogonality of $\boldsymbol{\Gamma}$, the isotropic version of model (2.5) can be rewritten as

$$\begin{aligned}
\mathbf{X}_1|(\mathbf{X}_2, \boldsymbol{\nu}') &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu}' + \sigma\boldsymbol{\varepsilon} \\
&= \bar{\boldsymbol{\mu}}_1 + (\mathbf{Q}_{\boldsymbol{\Gamma}} + \mathbf{P}_{\boldsymbol{\Gamma}})\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu}' + \sigma\boldsymbol{\varepsilon} \\
&= \bar{\boldsymbol{\mu}}_1 + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\nu}') + \sigma\boldsymbol{\varepsilon} \\
&= \bar{\boldsymbol{\mu}}_1 + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}(\underbrace{\boldsymbol{\Gamma}^T\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\nu}'}_{\boldsymbol{\nu}}) + \sigma\boldsymbol{\varepsilon} \\
&= \bar{\boldsymbol{\mu}}_1 + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu} + \sigma\boldsymbol{\varepsilon}.
\end{aligned}$$

Define the semi-orthogonal basis matrix $\boldsymbol{\Gamma}_0$ such that span($\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}$) $\equiv$ span($\boldsymbol{\Gamma}_0$) and let $\boldsymbol{\beta}_0$ be the coordinates of $\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}$ in terms of $\boldsymbol{\Gamma}_0$. Then we obtain the new model

$$\mathbf{X}_1|(\mathbf{X}_2, \boldsymbol{\nu}) = \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu} + \sigma\boldsymbol{\varepsilon}, \tag{3.1}$$

where we assume that $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p_1\times(p_1-d)}$ has rank $(p_1 - d)$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^{(p_1-d)\times p_2}$. Based on the definition of $\boldsymbol{\Gamma}_0$, the matrices $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}$ are orthogonal. The latent variable $\boldsymbol{\nu}$ is assumed to be centered to have mean 0, $\sum_i \boldsymbol{\nu}_i = 0$, but is otherwise uncontrained. This centering of $\boldsymbol{\nu}$ in the sample is for later convenience and is not essential. We will refer to this model (3.1) as the *isotropic partial probabilistic PCA model*. The following corollary confirms that $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\mathbf{X}_1$ is a sufficient reduction under model (3.1). Its proof is given in Appendix A.2.

Corollary 3.1. span($\boldsymbol{\beta},\ \boldsymbol{\Gamma}$) = span($\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta},\ \boldsymbol{\Gamma}$) = span($\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}$), *where* $\boldsymbol{\Gamma}_0$ *is a completion of* $\boldsymbol{\Gamma}$.

From model (3.1) we can think of three trivial cases. First, when $d = 0$, $\mathbf{X}_1$ is no longer associated with the latent variable so that we can eliminate $\boldsymbol{\nu}$ from model (3.1). Second, when $\boldsymbol{\beta}_0 = 0$, $\mathbf{X}_2$ has no information about $\mathbf{X}_1$ and can be excluded from model (3.1). Third, when $d = 0$ and $\boldsymbol{\beta}_0 = 0$, $\mathbf{X}_1$ depends only on $\bar{\boldsymbol{\mu}}_1$.

3.2 Maximum likelihood estimators

We will use the method of maximum likelihood for estimation of the sufficient reduction. The full log likelihood for model (3.1) is

$$\begin{aligned} L_d(\bar{\boldsymbol{\mu}}_1, \boldsymbol{\beta}_0, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\nu}, \sigma^2) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{np_1}{2}\log(\sigma^2) \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \left\| \mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\nu}_i \right\|^2 . \end{aligned}$$

Let $\mathbb{X}_1$ denote the $n \times p_1$ matrix with rows $(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)^T$, and $\mathbb{X}_2$ denote the $n \times p_2$ matrix with rows $(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)^T$, and let $\widehat{\mathbb{X}}_1 = P_{\mathbb{X}_2}\mathbb{X}_1$ denote the $n \times p_1$ matrix of centered fitted values from the multivariate linear regression of $\mathbf{X}_1$ on $\mathbf{X}_2$, including an intercept. Then $\widehat{\boldsymbol{\Sigma}}_1 = \mathbb{X}_1^T\mathbb{X}_1/n$ is the sample covariance matrix and $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2} = \widehat{\mathbb{X}}_1^T\widehat{\mathbb{X}}_1/n$ is the sample covariance matrix of the fitted vectors. The maximum likelihood estimator of $\mathcal{S}_{\boldsymbol{\Gamma}}$ can be constructed by holding $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0$, and σ^2 fixed and then maximizing the log likelihood over $\bar{\boldsymbol{\mu}}_1$, $\boldsymbol{\nu}_i$, and $\boldsymbol{\beta}_0$. This yields $\widehat{\bar{\boldsymbol{\mu}}}_1 = \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_2$. To estimate $\boldsymbol{\nu}_i$ for a selected value of i with the remaining parameters held fixed we need to maximize

$$-\frac{1}{2\sigma^2}\left\| \mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\nu}_i \right\|^2 .$$

This is just ordinary least squares and consequently is maximized by setting $\widehat{\boldsymbol{\nu}}_i = \boldsymbol{\Gamma}^T(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)$ using $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}_0 = 0$. These automatically satisfy the constraint $\sum_{i=1}^n \boldsymbol{\nu}_i = 0$. Substituting $\widehat{\bar{\boldsymbol{\mu}}}_1$ and $\widehat{\boldsymbol{\nu}}_i$ back, the partially maximized log likelihood L_d is then

$$\begin{aligned} L_d(\mathcal{S}_{\boldsymbol{\Gamma}}, \sigma^2) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{np_1}{2}\log(\sigma^2) \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \left\| \mathbf{X}_{1i} - \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) - \mathbf{P}_{\boldsymbol{\Gamma}}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) \right\|^2 \\ =\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \left\| \mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) \right\|^2 . \end{aligned} \tag{3.2}$$

Based on this partially maximized log likelihood, the estimated $\boldsymbol{\beta}_0$ (see Appendix A.3 for details) is

$$\hat{\boldsymbol{\beta}}_0 = {\boldsymbol{\Gamma}_0}^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}. \tag{3.3}$$

Substituting $\hat{\boldsymbol{\beta}}_0$ into the log likelihood, we have the final partially maximized log likelihood (see Appendix A.4 for details):

$$L_d(\mathcal{S}_{\boldsymbol{\Gamma}_0}, \mathcal{S}_{\boldsymbol{\Gamma}}, \sigma^2) = -\frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\Big\{\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_1\right] - \mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Sigma}}_1\right] - \mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\right]\Big\}. \quad (3.4)$$

Since, by the definition, $\boldsymbol{\Gamma}_0$ is a completion of $\boldsymbol{\Gamma}$; that is $(\boldsymbol{\Gamma}_0,\ \boldsymbol{\Gamma}) \in \mathbb{R}^{p_1 \times p_1}$ is an orthogonal matrix, we can use $\mathbf{Q}_{\boldsymbol{\Gamma}} = \mathbf{I}_{p_1} - \mathbf{P}_{\boldsymbol{\Gamma}} = \mathbf{P}_{\boldsymbol{\Gamma}_0}$. This leads to the final partially maximized log likelihood

$$\begin{aligned} L_d(\mathcal{S}_{\boldsymbol{\Gamma}}, \sigma^2) &= -\frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\Big\{\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_1\right] - \mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Sigma}}_1\right] - \mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\right]\Big\} \\ &= -\frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\Big\{\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_1\right] - \mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Sigma}}_1\right] - \mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\right] + \mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\right]\Big\} \\ &= -\frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}\right] + \frac{n}{2\sigma^2}\mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}\right], \end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}} = \widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}} = \mathbb{X}_1^T\mathbb{X}_1/n - \widehat{\mathbb{X}}_1^T\widehat{\mathbb{X}}_1/n$, is the residual covariance matrix from the multivariate linear regression of $\mathbf{X}_1$ on $\mathbf{X}_2$. Holding σ^2 fixed, the likelihood depends only on $\mathcal{S}_{\boldsymbol{\Gamma}}$. The rank of $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}$ is at most p_1 and typically $\mathrm{rank}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}) = p_1$. In any event, we assume that $\mathrm{rank}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}) > d$. The likelihood is then maximized by setting $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ equal to the span of the eigenvectors $\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_d$ corresponding to the largest d eigenvalues $\lambda_i\left(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}\right),\ i = 1, \ldots, d$, where $\lambda_i(\mathbf{Z})$ indicates the ith eigenvalue of matrix $\mathbf{Z}$. Naturally $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_0}$ is estimated by the span of eigenvectors corresponding to the smallest $(p_1 - d)$ eigenvalues $\lambda_i\left(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}\right),\ i = d+1, \ldots, p_1$.

We call $\widehat{\boldsymbol{\xi}}_1^T\mathbf{X}_1, \ldots, \widehat{\boldsymbol{\xi}}_{p_1}^T\mathbf{X}_1$ *principal residual components of* $\mathbf{X}_1$ *on* $\mathbf{X}_2$ ($PRC^{1|2}$), which are the principal components of $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}$. Since $\boldsymbol{\Gamma}_0$ is the completion of $\boldsymbol{\Gamma}$, the estimation of $\boldsymbol{\Gamma}^T\mathbf{X}_1$ and $\boldsymbol{\Gamma}_0^T\mathbf{X}_1$ are obtained as the first d $\mathrm{PRC}^{1|2}$ and the remaining $(p_1 - d)$ $\mathrm{PRC}^{1|2}$ respectively. Thus, all p_1 $\mathrm{PRC}^{1|2}$ are used to estimate sufficient reduction $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\mathbf{X}_1$. The corresponding estimate of scale is $\widehat{\sigma}^2 = \sum_{i=d+1}^{p_1}\lambda_i\left(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}\right)/p_1$.

The following simulation example may provide some intuition and motivate further study. Observation on $\mathbf{X}_1$ were generated as $\mathbf{X}_{1i} = \widetilde{\boldsymbol{\Gamma}}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} + \widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\nu}}_i + \sigma\boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n$, where we set $\widetilde{\boldsymbol{\Gamma}}_0 = (\mathbf{I}_{p_1} - \widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Gamma}}^T)$ and $\widetilde{\boldsymbol{\nu}}_i$ was sampled uniformly from the boundary of the square $[-1, 1]^2$ given in Figure 3.1.

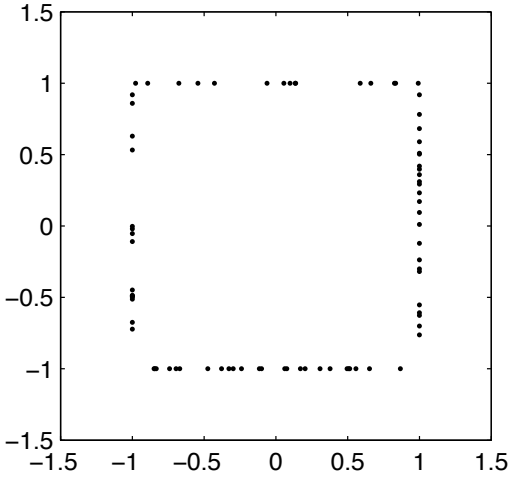


Figure 3.1: Plot of sampled $\widetilde{\boldsymbol{\nu}}$

The elements of the $p_1 \times p_2$ matrix $\boldsymbol{\beta}_0$, $p_2 \times 1$ matrix $\mathbf{X}_2$, and $p_1 \times 2$ matrix $\widetilde{\boldsymbol{\Gamma}}$ were sampled independently from a standard normal distribution, and the error vector $\boldsymbol{\varepsilon}$ was sampled from a normal distribution with mean 0 and variance matrix $\mathbf{I}_{p_1}$ with setting σ^2 equal to one. The sampling used to construct $\widetilde{\boldsymbol{\Gamma}}$ is for convenience only; the model is still conditional on $\widetilde{\boldsymbol{\Gamma}}$ regardless of how it was obtained. In terms of model (3.1), being attentive to the orthogonality of $\boldsymbol{\Gamma}$, we can see that $\bar{\boldsymbol{\mu}}_1 = 0$, $\boldsymbol{\Gamma} = \widetilde{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\Gamma}}^T\widetilde{\boldsymbol{\Gamma}})^{-1/2}$, and $\boldsymbol{\nu}_i = (\widetilde{\boldsymbol{\Gamma}}^T\widetilde{\boldsymbol{\Gamma}})^{1/2}\widetilde{\boldsymbol{\nu}}_i$. This sampling process was repeated $n = 80$ times for various values of p_1 and p_2 $(= p_1/5)$. Figure 3.2 shows plots of the first two principal residual components of $\mathbf{X}_1$ on $\mathbf{X}_2$ for four values of p_1. The plots show that for small p_1 the square is not recognizable, but for larger values of p_1 the square is quite clear. Especially the nice shape of the rotated square is shown in Figure 3.2d with $p_1 = 250$ predictors even though the number of observations is still $n = 80$. The sides of the estimated square in Figure 3.2d do not align with the coordinate axes because the method is designed to estimate only the subspace $\boldsymbol{\Psi}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$, which is equal to $\mathcal{S}_{\boldsymbol{\Gamma}}$ with isotropic errors.

From the construction of this simulation we can easily see why principal residual components of $\mathbf{X}_1$ on $\mathbf{X}_2$ reproduce the square for relatively large p_1. Starting with

model (3.1),

$$\begin{aligned}
\mathbf{X}_1|(\mathbf{X}_2,\boldsymbol{\nu}) &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\nu} + \sigma\boldsymbol{\varepsilon} \\
&= \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \widetilde{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\Gamma}}^T\widetilde{\boldsymbol{\Gamma}})^{-1/2}(\widetilde{\boldsymbol{\Gamma}}^T\widetilde{\boldsymbol{\Gamma}})^{1/2}\widetilde{\boldsymbol{\nu}} + \sigma\boldsymbol{\varepsilon} \\
&= \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \sqrt{p_1}\widetilde{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\Gamma}}^T\widetilde{\boldsymbol{\Gamma}})^{-1/2}\left(\frac{\widetilde{\boldsymbol{\Gamma}}^T\widetilde{\boldsymbol{\Gamma}}}{p_1}\right)^{1/2}\widetilde{\boldsymbol{\nu}} + \sigma\boldsymbol{\varepsilon},
\end{aligned}$$

where the mean term $\bar{\boldsymbol{\mu}}_1$ in the second equation is 0. Since $\widetilde{\boldsymbol{\Gamma}}$ was sampled independently from a standard normal distribution, $(\widetilde{\boldsymbol{\Gamma}}^T\widetilde{\boldsymbol{\Gamma}}/p_1)^{1/2} \to \mathbf{I}$ as $p_1 \to \infty$ in the last equation. Therefore,

$$\begin{aligned}
\mathbf{X}_1|(\mathbf{X}_2,\boldsymbol{\nu}) &\approx \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \sqrt{p_1}\boldsymbol{\Gamma}\widetilde{\boldsymbol{\nu}} \\
\boldsymbol{\Gamma}^T\mathbf{X}_1/\sqrt{p_1} &\approx \widetilde{\boldsymbol{\nu}}.
\end{aligned}$$

We can see that $\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}_1$ is able to recover all the properties of $\widetilde{\boldsymbol{\nu}}$ for large p_1.

3.3 Groupwise partial probabilistic PCA model and estimation

In a large variety of applications, the multivariate random sample may contain discrete variables taking on two or more possible values which indicate groups or levels as a natural partitioning of the sample into groups. Education level is an example of a discrete variable with several levels, high school, college, and graduate school. Besides, we may have the datasets with replication in the observed variable. This happens if the variable is supported on a finite number of points; for example, the average amount(hour) of sleep. The presence of replication is common in samples using experimental design. For instance, when an investigator wants to see the variability associated with a phenomenon, repetition of an experimental condition is usually conducted. In such samples, the variable observed by repeated values can be regarded as categorical. These types of dataset can be found easily in many application fields such as bioinformatics, geology, sociology, or engineering. With these applications it is more desirable to conduct dimension reduction incorporating the prior group or replication knowledge. For this

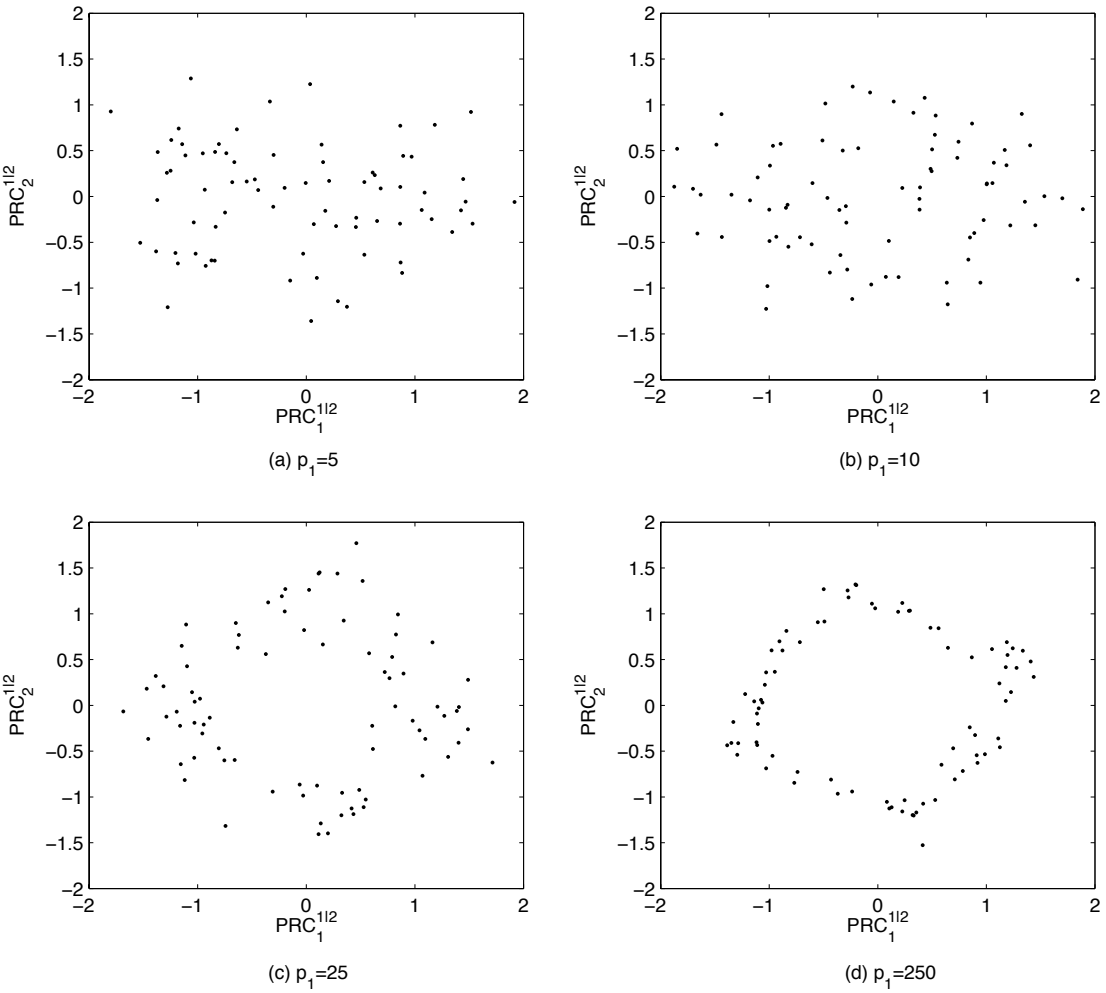


Figure 3.2: Plots of the estimated principal residual components of regression $\mathbf{X}_1$ on $\mathbf{X}_2$ ($\text{PRC}^{1|2}$) with n=80 observations and various number of predictors p_1

concern, Li, Li, and Zhu (2010) proposed groupwise dimension reduction method. They incorporate the group information by imposing a direct sum structure on the differential operator of the conditional mean $E(Y|\mathbf{X})$. Their method is the moment-based approach, and the proposed solution for the dimension reduction estimation involves multidimensional kernel smoothing. In this section we propose a model-based approach incorporating categorical variables into a partial probabilistic model.

As we mentioned in Chapter 2, model (2.5) can hold with a categorical variable $\mathbf{X}_2$. Suppose we have a categorical variable with g levels or groups. Then $\mathbf{X}_2$ becomes a $g-1$ dimensional vector of indicator variables representing each group. For example, a 2×1 vector is required to describe a categorical variable with 3 groups, $\mathbf{X}_2 = (1,0)^T$ for group one, $\mathbf{X}_2 = (0,1)^T$ for group two, and $\mathbf{X}_2 = (0,0)^T$ for group three. In order to emphasize that $\mathbf{X}_2$ is a group indicator, we will write $\mathbf{X}_2$ as $\mathbf{G}$. Therefore, our goal is to reduce the dimension of $\mathbf{X}$ in the presence of the latent variable $\boldsymbol{\nu}'$ and the group indicator variable $\mathbf{G}$. Specifying the group property, model (2.5) can be rewritten as

$$\mathbf{X}_{ij} = \bar{\boldsymbol{\mu}} + \boldsymbol{\beta}\mathbf{G}_i + \boldsymbol{\Gamma}_i\boldsymbol{\nu}'_{ij} + \sigma_i\boldsymbol{\varepsilon}_{ij}, \quad i = 1, \ldots, g,\ j = 1, \ldots, n_i, \tag{3.5}$$

where i indicates the group and j denotes the observations in each group i. Here $\mathbf{G}_i \in \mathbb{R}^{g-1}$ with g groups and $\boldsymbol{\Gamma}_i \in \mathbb{R}^{p \times d_i}$. The error vector $\boldsymbol{\varepsilon}_{ij}$ is normally distributed with mean 0 and covariance matrix $\mathbf{I}_p$ and the term $\sigma_i\boldsymbol{\varepsilon}_{ij}$ indicates that each group can have a different isotropic error covariance matrix. For later convenience we assume that the latent variable $\boldsymbol{\nu}'_{ij} \in \mathbb{R}^{d_i}$ satisfies $\sum_j \boldsymbol{\nu}'_{ij} = 0$ for all i, but is otherwise unconstrained. For fixed i, model (3.5) can also be regarded as a partial probabilistic PCA model taking the group information $\mathbf{G}$ instead of general continuous variable $\mathbf{X}_2$ into account. The sufficient reduction is easily obtained as $R(\mathbf{X}_{ij}) = \boldsymbol{\Gamma}_i^T\mathbf{X}_{ij}$ for each group i by Proposition 2.1. We can see that the parameters $\boldsymbol{\Gamma}_i,\ i = 1, \ldots, g$, determine the sufficient reduction for each group. That is, they contain all the groupwise sufficient reduction information.

Let $\boldsymbol{\Gamma}$ be a semi-orthogonal basis for $\text{span}(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_g)$. Then model (3.5) becomes

$$\begin{aligned}\mathbf{X}_{ij} &= \bar{\boldsymbol{\mu}} + \boldsymbol{\beta}\mathbf{G}_i + \boldsymbol{\Gamma}\boldsymbol{\gamma}_i\boldsymbol{\nu}'_{ij} + \sigma_i\boldsymbol{\varepsilon}_{ij}, \quad \text{where } \boldsymbol{\gamma}_i \in \mathbb{R}^{d \times d_i} \\ &= \bar{\boldsymbol{\mu}} + \boldsymbol{\beta}\mathbf{G}_i + \boldsymbol{\Gamma}\widetilde{\boldsymbol{\nu}}_{ij} + \sigma_i\boldsymbol{\varepsilon}_{ij}, \quad \text{where } \widetilde{\boldsymbol{\nu}}_{ij} \in \mathbb{R}^d. \end{aligned} \tag{3.6}$$

In the last equation, $\boldsymbol{\gamma}_i$ was absorbed into the latent variable $\widetilde{\boldsymbol{\nu}}_{ij}$ since it is not estimable, so $\widetilde{\boldsymbol{\nu}}_{ij}$ contains all the group information. Here the rank of $\boldsymbol{\Gamma}$ is $d \geq \max(d_1, \ldots, d_g)$ where each parameter $\boldsymbol{\Gamma}_i$ has rank d_i corresponding to group i. Consequently under model (3.6) we can derive the sufficient reduction for the whole $\mathbf{X}$ by using $\boldsymbol{\Gamma}$ which does not depend on i.

Applying the same rule as in Section 3.1 we have the model

$$\mathbf{X}_{ij} = \bar{\boldsymbol{\mu}} + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{G}_i + \boldsymbol{\Gamma}\boldsymbol{\nu}_{ij} + \sigma_i\boldsymbol{\varepsilon}_{ij}, \tag{3.7}$$

where we still assume $\sum_j \boldsymbol{\nu}_{ij} = 0$ for all i. Then the sufficient reduction for $\mathbf{X}$ is given by $R(\mathbf{X}) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\mathbf{X}$. We will refer to model (3.7) as a *groupwise partial probabilistic model*.

For the estimation of parameters we can think of using the maximum likelihood method again. The full log likelihood for model (3.7) is

$$L_d = -\frac{np}{2}\log(2\pi) - \frac{p}{2}\sum_{i=1}^{g} n_i \log\sigma_i^2 - \sum_{i=1}^{g}\frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i}\|\mathbf{X}_{ij} - \bar{\boldsymbol{\mu}} - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{G}_i - \boldsymbol{\Gamma}\boldsymbol{\nu}_{ij}\|^2, \tag{3.8}$$

where $n = n_1 + \cdots + n_g$. Here the σ_i's change according to group and it makes the estimation of parameters hard. Thus, we are not able to find a closed-form solution for the MLEs of all parameters which are related to the estimation of the sufficient reduction. It is necessary to use an iterative algorithm proposed by Adragi and Cook (2009). This algorithm is a straightforward alternating algorithm for estimating the parameters.

The alternating algorithm can be constructed with the following reasoning. Once the mean function is specified in model (3.7) then the variances $\sigma_i^2,\ i = 1, \ldots, g$ can be estimated by using the sample variance of the centered variables $\mathbf{X}_{ij} - \bar{\boldsymbol{\mu}} - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{G}_i - \boldsymbol{\Gamma}\boldsymbol{\nu}_{ij}$. If σ_i^2's are specified then we can standardize the predictor vector to obtain a standardized groupwise partial probabilistic model in $\mathbf{Z}_{ij} = (1/\sigma_i)\mathbf{X}_{ij}$:

$$\mathbf{Z}_{ij} = \sigma_i^{-1}\bar{\boldsymbol{\mu}} + \sigma_i^{-1}\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{G}_i + \sigma_i^{-1}\boldsymbol{\Gamma}\boldsymbol{\nu}_{ij} + \boldsymbol{\varepsilon}_{ij}. \tag{3.9}$$

Consequently, we can estimate span($\boldsymbol{\Gamma}$) from the estimate $\widetilde{\boldsymbol{\Gamma}}$ of $\sigma_i^{-1}\boldsymbol{\Gamma}$ from the above model and similarly span($\boldsymbol{\Gamma}_0$) from the estimate $\widetilde{\boldsymbol{\Gamma}}_0$ of $\sigma_i^{-1}\boldsymbol{\Gamma}_0$. Alternating between these two steps leads to the following algorithm:

1. Fit the standardized groupwise partial probabilistic model (3.9) to the original data, getting initial estimates $\widehat{\mathbf{\Gamma}}_{(1)}$, $\widehat{\mathbf{\Gamma}}_{0(1)}$, $\widehat{\boldsymbol{\beta}}_{0(1)}$, and $\widehat{\boldsymbol{\nu}}_{ij(1)}$. These estimates can be obtained by using the following equations.

   Let $\widehat{\mathbf{\Sigma}}_{\text{res}}^{\mathbf{X}|\mathbf{G}} = \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{\mathbf{X}|\mathbf{G}} = \mathbb{X}^T\mathbb{X}/n - \mathbb{X}^T P_{\mathbb{G}}\mathbb{X}/n$, where $\mathbb{G}$ is $n \times (g-1)$ matrix with $(\mathbf{1}_{n_1}^T \otimes (\mathbf{G}_1 - \bar{\mathbf{G}}), \cdots, \mathbf{1}_{n_g}^T \otimes (\mathbf{G}_g - \bar{\mathbf{G}}))^T$.

   Then estimators of each parameter are

   $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(1)}} = \text{span}\{\text{eigenvectors corresponding to the largest d eigenvalues of } \widehat{\mathbf{\Sigma}}_{\text{res}}^{\mathbf{X}|\mathbf{G}}\}$

   $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{0(1)}} = \text{span}\{\text{eigenvectors corresponding to the smallest p-d eigenvalues of } \widehat{\mathbf{\Sigma}}_{\text{res}}^{\mathbf{X}|\mathbf{G}}\}$

   $\widehat{\boldsymbol{\nu}}_{ij(1)} = \widehat{\mathbf{\Gamma}}^T(\mathbf{X}_{ij} - \bar{\mathbf{X}})$

   $\widehat{\boldsymbol{\beta}}_{0(1)} = \widehat{\mathbf{\Gamma}}_0^T \mathbb{X}^T\mathbb{G}(\mathbb{G}^T\mathbb{G})^{-1}$.

2. For some small $\delta > 0$, repeat for $k = 1, 2, \ldots$ until $\sum_{i=1}^{g}\left(\widehat{\sigma}_{i(k)}^2 - \widehat{\sigma}_{i(k+1)}^2\right)^2 < \delta$.

   (a) Calculate $\widehat{\sigma}_{i(k)}^2 = \dfrac{1}{n_i - 1}\displaystyle\sum_{j=1}^{n_i} \|\mathbf{X}_{ij} - \widehat{\mathbf{\Gamma}}_{0(k)}\widehat{\boldsymbol{\beta}}_{0(k)}\mathbf{G}_i - \widehat{\mathbf{\Gamma}}_{(k)}\widehat{\nu}_{ij(k)}\|^2$ for $i = 1, \ldots, g$.

   (b) Transform $\mathbf{Z}_{ij} = (1/\widehat{\sigma}_{i(k)})\mathbf{X}_{ij}$ for $i = 1, \ldots, g$, and $j = 1, \ldots, n_g$.

   (c) Fit the standardized groupwise partial probabilistic model to $\mathbf{Z}_{ij}$, yielding estimates $\widetilde{\mathbf{\Gamma}}$, $\widetilde{\mathbf{\Gamma}}_0$, $\widetilde{\boldsymbol{\beta}}_0$, and $\widetilde{\boldsymbol{\nu}}_{ij}$.

   (d) Backtransform the estimates to the original scale $\widehat{\mathbf{\Gamma}}_{(k+1)} = \widehat{\sigma}_{i(k)}\widetilde{\mathbf{\Gamma}}$, $\widehat{\mathbf{\Gamma}}_{0(k+1)} = \widehat{\sigma}_{i(k)}\widetilde{\mathbf{\Gamma}}_0$, $\widehat{\boldsymbol{\beta}}_{0(k+1)} = \widetilde{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\nu}}_{ij(k+1)} = \widetilde{\boldsymbol{\nu}}_{ij}$.

After going through all these algorithm steps, the sufficient reduction can be estimated as $\widehat{R}(\mathbf{X}) = (\widehat{\mathbf{\Gamma}}_{0(k+1)}\widehat{\boldsymbol{\beta}}_{0(k+1)}, \widehat{\mathbf{\Gamma}}_{(k+1)})^T\mathbf{X}$.

For illustration we use the same simulation setup as provided in Section 3.2. The only difference is that we have to include a group indicator variable in the sample here. Considering a groupwise partial probabilistic model (3.7), observations on $\mathbf{X}$ were generated as $\mathbf{X}_{ij} = \mathbf{\Gamma}_0\boldsymbol{\beta}_0\mathbf{G}_i + \mathbf{\Gamma}\boldsymbol{\nu}_{ij} + \sigma_i\boldsymbol{\varepsilon}_{ij}$. We took $g = 2$ and $n_1 = n_2 = 80$ with the total sample size 160. Therefore $i = 1, 2$ and $j = 1, \ldots, 80$ and the indicator variable $\mathbf{G}_1 = 1$ for group one and $\mathbf{G}_2 = 0$ for group two. We set $\sigma_1 = \sigma_2 = 1$, $\mathbf{\Gamma}_0 = (\mathbf{I}_p - \mathbf{\Gamma}\mathbf{\Gamma}^T)$, and $\boldsymbol{\nu}_{ij}$ was sampled uniformly from the boundary of the square $[-1, 1]^2$ for group one and the square $[-1, 1] \times [3, 5]$ for group two as shown in Figure 3.3.

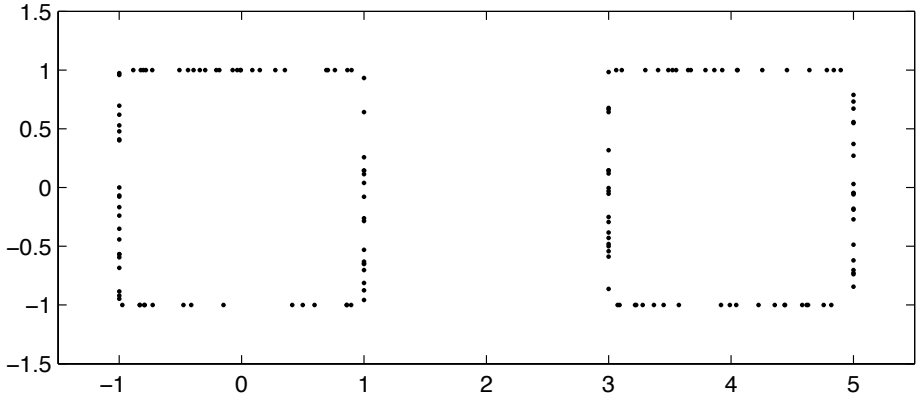


Figure 3.3: Plot of sampled $\boldsymbol{\nu}_{1j}$(left square) and $\boldsymbol{\nu}_{2j}$(right square)

The elements of $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times 2}$ were sampled independently from a standard normal distribution, and the error vector $\boldsymbol{\varepsilon}_{ij}$ was sampled from a normal distribution with mean 0 and variance with $\mathbf{I}_p$. This sampling process was conducted for various values of p. Figure 3.4 shows plots of the first two principal residual components of $\mathbf{X}$ on $\mathbf{G}$ for four values of p. The plots show the same results as in Section 3.2. When p is small, we cannot recognize the shape of a square or any structure indicating groups, as all points are mixed up without classification by group. When p is large, however, the two squares corresponding to each group are clear. Furthermore, with the same reasoning given in Section 3.2 these two principal residual components recover $\boldsymbol{\nu}_{ij}$ well, particularly when $p = 500$, which contains all the group information.

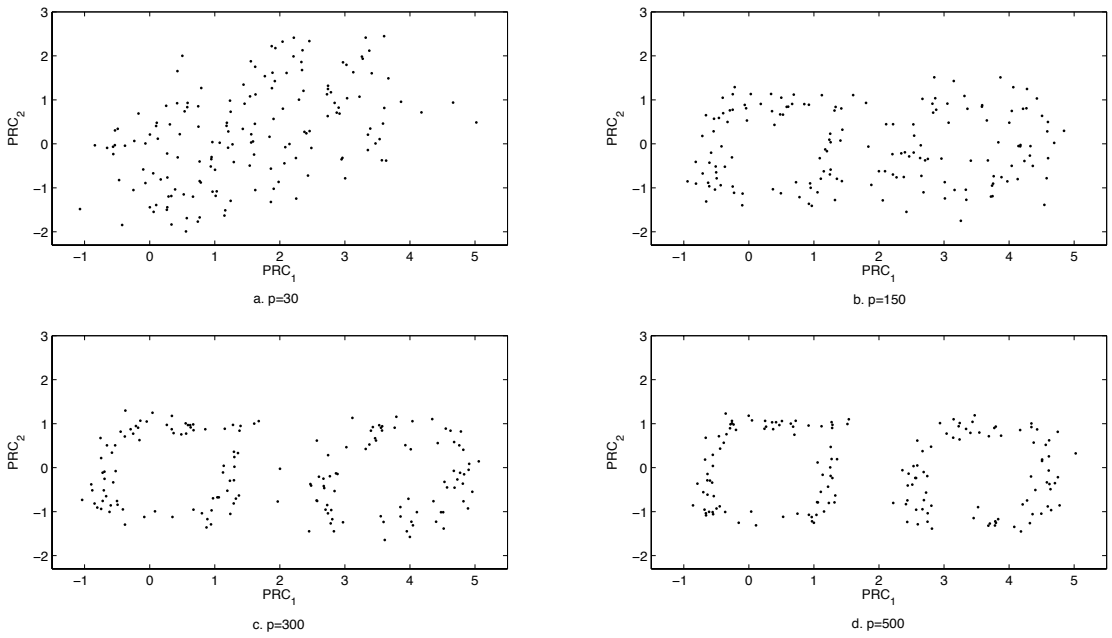


Figure 3.4: Plots of the estimated principal residual components of regression $\mathbf{X}$ on $\mathbf{G}$ ($\text{PRC}^{\mathbf{X}|\mathbf{G}}$) with $n_1 = n_2 = 80$ observations and various number of predictors p

Chapter 4

Partial PFC Models and Estimation

4.1 PFC models and partial PFC models

4.1.1 PFC models

Under the probabilistic PCA model (2.1), regardless of the randomness of the latent varialble $\boldsymbol{\nu}^*$, we have the same estimation results for the sufficient dimension reduction in the cases with fixed $\boldsymbol{\nu}^*$ and random $\boldsymbol{\nu}^*$ (Tipping and Bishop 1999, Cook 2007, and Chen 2010). This model does not include a response variable. In this chapter we assume that a response is known and study the impact of the randomness of the latent variable and the known response on the probabilistic PCA model (2.1).

An adaption of model (2.1) to accommodate a particular response is often possible and useful by modeling the latent variable $\boldsymbol{\nu}^*$. Assume that $\boldsymbol{\nu}^*$ is fixed and modeled by $\boldsymbol{\nu}^* = \boldsymbol{\alpha}^* \mathbf{f}_y^*$ without approximation error, where $\mathbf{f}_y^* \in \mathbb{R}^r$ is a known vector-valued function of y with linearly independent elements and $\boldsymbol{\alpha}^* \in \mathbb{R}^{d^* \times r}$, $d^* \leq \min(r, p)$, is an unrestricted rank d^* matrix. Then we have the model

$$\mathbf{X} = \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^* \boldsymbol{\alpha}^* \mathbf{f}_y^* + \sigma \boldsymbol{\epsilon}, \tag{4.1}$$

where $\bar{\boldsymbol{\mu}}^*$, $\boldsymbol{\Gamma}^*$, and $\boldsymbol{\epsilon}$ are as defined previously in (2.1). How to choose a suitable $\mathbf{f}_y^*$

was studied extensively with various simulations by Adragni and Cook (2009). Since we have a known response Y and predictors $\mathbf{X}$, model (4.1) can be regarded as an inverse regression model of $\mathbf{X}$ on Y. Under model (4.1) $\boldsymbol{\Gamma}^{*T}\mathbf{X}$ is a sufficient reduction. This model is called an *isotropic PFC model* because it has an isotropic error covariance matrix and the dimension reduction subspace $\mathcal{S}_{\boldsymbol{\Gamma}^*}$ is estimated by the span of the first d^* eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}$, where $\widehat{\boldsymbol{\Sigma}}_{\text{fit}} = \mathbb{X}^T\mathbf{P}_{\mathbf{F}^*}\mathbb{X}/n$ is the sample covariance matrix of the fitted vectors from the multivariate linear regression of $\mathbf{X}$ on $\mathbf{f}_y^*$, including an intercept (Cook 2007). Here $\mathbb{X}$ is the $n \times p$ matrix with rows $(\mathbf{X}_i - \bar{\mathbf{X}})^T$ and $\mathbf{P}_{\mathbf{F}^*}$ is the linear operator that projects onto the subspace spanned by the columns of $\mathbf{F}^*$, where $\mathbf{F}^*$ denotes the $n \times r$ matrix with rows $\mathbf{f}_{y_i}^{*T}$. If $\boldsymbol{\nu}^*$ is fixed and there is an approximation error when modeling $\boldsymbol{\nu}^*$ as $\boldsymbol{\alpha}^*\mathbf{f}_y^*$, then the bias term $\boldsymbol{\Gamma}^*(\boldsymbol{\nu}^* - \boldsymbol{\alpha}^*\mathbf{f}_y^*)$ is added in model (4.1) and should be considered.

When $\boldsymbol{\nu}^*$ is assumed to be random, a different formulation is applied to model (2.1) while taking the response into account:

$$\begin{aligned}\mathbf{X} &= \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^*\boldsymbol{\nu}^* + \sigma\boldsymbol{\epsilon} \\ &= \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^*E(\boldsymbol{\nu}^*|y) + \boldsymbol{\Gamma}^*(\boldsymbol{\nu}^* - E(\boldsymbol{\nu}^*|y)) + \sigma\boldsymbol{\epsilon} \\ &= \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_y^* + \boldsymbol{\Gamma}^*\boldsymbol{\omega}^* + \sigma\boldsymbol{\epsilon}, \end{aligned} \tag{4.2}$$

where $E(\boldsymbol{\nu}^*|y)$ is a function of y that can be modeled by $\boldsymbol{\alpha}^*\mathbf{f}_y^*$ using a known function of y, $\mathbf{f}_y^*$. The term $\boldsymbol{\omega}^* = \boldsymbol{\nu}^* - E(\boldsymbol{\nu}^*|y)$ can be regarded as an approximation error. That is, $E(\boldsymbol{\nu}^*|y)$ is only an approximation of $\boldsymbol{\nu}^*$ and there is some part which cannot be explained by a function of y. If $\boldsymbol{\nu}^*$ is exactly modeled by $E(\boldsymbol{\nu}^*|y)$ the term $\boldsymbol{\Gamma}^*\boldsymbol{\omega}^*$ in the last equation disappears and the model becomes equivalent to the isotropic PFC model (4.1). Suppose that $\boldsymbol{\omega}^*$ is a random vector assumed to be normally distributed with mean 0 and covariance matrix $\boldsymbol{\Phi}^*$ and independent of y. Then new normal error $\boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}$ with mean 0 and variance $\boldsymbol{\Gamma}^*\boldsymbol{\Phi}^*\boldsymbol{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$ is defined and used in model

$$\mathbf{X} = \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_y^* + \boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}. \tag{4.3}$$

The effect of approximation errors is reflected in the error term. Let $\boldsymbol{\Omega}^* = \boldsymbol{\Gamma}^*\boldsymbol{\Phi}^*\boldsymbol{\Gamma}^{*T} +$

$\sigma^2\mathbf{I}_p$. Then the sufficient reduction is $R(\mathbf{X}) = \mathbf{\Gamma}^{*T}\mathbf{\Omega}^{*-1}\mathbf{X}$. When $\mathbf{\Omega}^*$ has this special structure the reduction form can be simplified as $R(\mathbf{X}) = \mathbf{\Gamma}^{*T}\mathbf{X}$ by the following corollary. Its proof is given in Appendix A.5.

Corollary 4.1. *If* $\mathbf{\Omega}^* = \mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$, *then* $R(\mathbf{X}) = \mathbf{\Gamma}^{*T}\mathbf{\Omega}^{*-1}\mathbf{X}$ *is equivalent to* $R(\mathbf{X}) = \mathbf{\Gamma}^{*T}\mathbf{X}$.

Since the structure of $\boldsymbol{\omega}^*$ is not restrictive, we can consider various structures for the error covariance matrix $\mathbf{\Omega}^*$. First of all, consider model (4.3) as it is. That is, we have the variance $\mathbf{\Omega}^* > 0$ as one of the special structures. Then the sufficient reduction is obtained as $R(\mathbf{X}) = \mathbf{\Gamma}^{*T}\mathbf{X}$ by Corollary 4.1 and estimated by using a numerical optimization. Secondly, the isotropic version of model (4.3) with $\mathbf{\Omega}^* = \sigma^2\mathbf{I}_p$ can be considered as representing that the predictors are conditionally independent and have the same variance. Then model (4.3) is the same as the isotropic PFC model. The study for estimating the sufficient reduction under this model was nicely done by Cook (2007). Third, the scope of application can be expanded by permitting a diagonal covariance matrix $\mathbf{\Omega}^* = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. This allows for different measurement scales of the predictors, but still requires that they be conditionally independent. Adragni and Cook (2009) studied a PFC model with diagonal error covariance matrix using a straightforward alternating algorithm (the diagonal PFC model). Finally, we consider a more general version of model (4.3) by allowing for unstructured errors, $\mathbf{\Omega}^* > 0$. For the general error covariance matrix the predictors can be conditionally dependent with different variances. A PFC model with a general error structure was considered by Cook and Forzani (2008) (the general PFC model). While the isotonic and diagonal version of model (4.3) do not require $p < n$, the general version of model (4.3) works best in regressions with $p \ll n$ (Adragni and Cook 2009).

Consequently when the multivariate random sample contains a response variable four types of model can be considered according to the above development of model (2.1): (1) isotropic PFC model, (2) diagonal PFC model, (3) general PFC model, and (4) PFC model with the specific variance function $\mathbf{\Omega}^* = \mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$.

4.1.2 Partial PFC models

In Chapter 3, we studied the partial probabilistic model (2.5) with isotropic error covariance matrix $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}_{p_1}$. Starting with the same setting, now we assume that the dataset contains the response variable Y. Then model (2.5) can be adapted to accommodate Y by modeling $\boldsymbol{\nu}'$. Assume that $\boldsymbol{\nu}'$ is fixed and modeled by $\boldsymbol{\nu}' = \boldsymbol{\alpha}' \mathbf{f}_y'$ without approximation error where $\mathbf{f}_y' \in \mathbb{R}^r$ is a known function of y and $\boldsymbol{\alpha}' \in \mathbb{R}^{d \times r},\ d \leq \min(r, p_1 - p_2)$ is an unrestricted rank d matrix. Then we have the model

$$\mathbf{X}_1 | (\mathbf{X}_2, Y) = \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta} \mathbf{X}_2 + \boldsymbol{\Gamma} \boldsymbol{\alpha}' \mathbf{f}_y' + \sigma \boldsymbol{\varepsilon}, \tag{4.4}$$

where $\boldsymbol{\varepsilon}$ is normally distributed with mean 0 and covariance matrix $\mathbf{I}_{p_1}$. We assume that $\mathbf{f}_y'$ is centered throughout this thesis. If there is an approximation error when modeling $\boldsymbol{\nu}'$ as $\boldsymbol{\alpha}' \mathbf{f}_y'$, then the bias term $\boldsymbol{\Gamma}(\boldsymbol{\nu}' - \boldsymbol{\alpha}' \mathbf{f}_y')$ is added in model (4.4).

When $\boldsymbol{\nu}'$ is assumed to be random, following the same steps in equation (4.2), a different representation of model (2.5) is obtained as

$$\begin{aligned} \mathbf{X}_1 | (\mathbf{X}_2, Y) &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta} \mathbf{X}_2 + \boldsymbol{\Gamma} \boldsymbol{\nu}' + \sigma \boldsymbol{\varepsilon} \\ &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta} \mathbf{X}_2 + \boldsymbol{\Gamma} E(\boldsymbol{\nu}' | y) + \boldsymbol{\Gamma}(\boldsymbol{\nu}' - E(\boldsymbol{\nu}' | y)) + \sigma \boldsymbol{\varepsilon} \\ &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta} \mathbf{X}_2 + \boldsymbol{\Gamma} \boldsymbol{\alpha}' \mathbf{f}_y' + \boldsymbol{\Gamma} \boldsymbol{\omega} + \sigma \boldsymbol{\varepsilon}, \end{aligned} \tag{4.5}$$

where $\boldsymbol{\omega} = \boldsymbol{\nu}' - E(\boldsymbol{\nu}' | y)$ can be regarded as an approximation error and $E(\boldsymbol{\nu}' | y)$ is modeled by $\boldsymbol{\alpha}' \mathbf{f}_y'$. If there is no approximation error, then equation (4.5) is the same as model (4.4). Suppose that $\boldsymbol{\omega}$ is a random variable assumed to follow normal distribution with mean 0 and covariance matrix $\boldsymbol{\Phi}$ and independent of y. Equation (4.5) is rewritten as

$$\mathbf{X}_1 | (\mathbf{X}_2, Y) = \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta} \mathbf{X}_2 + \boldsymbol{\Gamma} \boldsymbol{\alpha}' \mathbf{f}_y' + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}}, \tag{4.6}$$

with new error term $\boldsymbol{\varepsilon}_{\boldsymbol{\omega}}$ that is normally distributed with mean 0 and covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Gamma} \boldsymbol{\Phi} \boldsymbol{\Gamma}^T + \sigma^2 \mathbf{I}_{p_1}$. Here the effect of approximation errors is absorbed in the error term.

Under model (4.6), we can now restate Proposition 2.1 connecting with the conditional distribution of $(Y,\ \mathbf{X}_2)$ given $\mathbf{X}_1$ and can define a sufficient reduction.

Proposition 4.1. *Under model* (4.6)*, the distribution of* $(Y,\ \mathbf{X}_2)|\mathbf{X}_1$ *is the same as the distribution of* $(Y,\ \mathbf{X}_2)|(\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1$ *for all values of* $\mathbf{X}_1$*. That is,* $R(\mathbf{X}_1) = (\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1$ *is a sufficient reduction. Let* $T(\mathbf{X}_1)$ *be any sufficient reduction. Then, under model* (4.6)*,* R *is a function of* T*.*

The proof is basically the same as in Proposition 2.1 and is omitted. According to this proposition, $\mathbf{X}_1$ can be replaced by the sufficient reduction $R(\mathbf{X}_1) = (\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1$ without loss of information about $(Y,\ \mathbf{X}_2)$ given $\mathbf{X}_1$, and thus the conditional independence $\mathbf{X}_1 \perp\!\!\!\perp (Y,\ \mathbf{X}_2)|R(\mathbf{X}_1)$ holds. This relationship can be extended to the following proposition restating Proposition 2.2 while considering the response Y instead of the latent variable $\boldsymbol{\nu}'$.

Proposition 4.2. *The following pair of conditions* a_1 *and* a_2 *is equivalent to the pair of conditions* b_1 *and* b_2 *which is equivalent to condition* c*.*

$$(a_1)\ \mathbf{X}_1 \perp\!\!\!\perp Y|(\mathbf{X}_2,\ R(\mathbf{X}_1)) \quad (a_2)\ \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|R(\mathbf{X}_1)$$
$$(b_1)\ \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|(Y,\ R(\mathbf{X}_1)) \quad (b_2)\ \mathbf{X}_1 \perp\!\!\!\perp Y|R(\mathbf{X}_1)$$
$$(c)\ \mathbf{X}_1 \perp\!\!\!\perp (Y,\ \mathbf{X}_2)|R(\mathbf{X}_1)$$

Again conditions (a_1, a_2) and (b_1, b_2) are redundant, but are given separately to emphasize the symmetric roles of $\mathbf{X}_2$ and Y. This description of the sufficient reduction emerges from the following definition which is a restatement of Definition 2.1.a:

Definition 4.1.a. *Under model* (2.5)*, a partial reduction* $R(\mathbf{X}_1)$*, with* $R : \mathbb{R}^{p_1} \to \mathbb{R}^{q_1},\ q_1 \leq p_1$*, is sufficient for* $(\mathbf{X}_2,\ Y)$ *if it satisfies one of the following three statements:*

$$\textit{(i)}\ \mathbf{X}_1|(\mathbf{X}_2,\ Y,\ R(\mathbf{X}_1)) \sim \mathbf{X}_1|R(\mathbf{X}_1),$$
$$\textit{(ii)}\ (\mathbf{X}_2,\ Y)|\mathbf{X}_1 \sim (\mathbf{X}_2,\ Y)|R(\mathbf{X}_1),$$
$$\textit{(iii)}\ (\mathbf{X}_2,\ Y) \perp\!\!\!\perp \mathbf{X}_1|R(\mathbf{X}_1).$$

The interpretation of Definition 4.1.a is the same as in Definition 2.1.a using Y instead of $\boldsymbol{\nu}'$ in the statements. Like Definition 2.1.b, based on the condition (a_1, a_2) in Proposition 4.2, Definition 4.1.a can be restated as follows:

Definition 4.1.b. *Under model* (4.6)*, a partial reduction* $(R(\mathbf{X}_1),\ \mathbf{X}_2)$*, with* $R : \mathbb{R}^{p_1} \to \mathbb{R}^{q_1},\ q_1 \leq p_1$*, is sufficient for* Y*, and* $R(\mathbf{X}_1)$ *is sufficient for* $\mathbf{X}_2$ *if it satisfies one of the following three statements:*

(i) inverse reduction, $\mathbf{X}_1|(Y,\ \mathbf{X}_2,\ R(\mathbf{X}_1)) \sim \mathbf{X}_1|(\mathbf{X}_2,\ R(\mathbf{X}_1))$
and $\mathbf{X}_1|(\mathbf{X}_2, R(\mathbf{X}_1)) \sim \mathbf{X}_1|R(\mathbf{X}_1)$

(ii) forward reduction, $Y|(\mathbf{X}_1,\ \mathbf{X}_2) \sim Y|(\mathbf{X}_2,\ R(\mathbf{X}_1))$
and $\mathbf{X}_2|\mathbf{X}_1 \sim \mathbf{X}_2|R(\mathbf{X}_1)$

(iii) joint reduction, $Y \perp\!\!\!\perp \mathbf{X}_1|(\mathbf{X}_2,\ R(\mathbf{X}_1))$ *and* $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|R(\mathbf{X}_1)$.

Again we can also think of the alternative way to state Definition 4.1.b, based on the conditions (b_1, b_2) in Proposition 4.2. While Definition 4.1.a focuses on the sufficiency for $(Y,\ \mathbf{X}_2)$, Definition 4.1.b considers the sufficiency for Y and $\mathbf{X}_2$ separately.

Since a sufficient reduction is defined as $R(\mathbf{X}_1) = (\boldsymbol{\beta},\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1$ by Proposition 4.1 under model (4.6), the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ are of interest. Following the same procedure as in Section 3.1 to distinguish the span($\boldsymbol{\Gamma}$) from $\boldsymbol{\beta}$, model (4.6) can be rewritten as

$$\begin{aligned}
\mathbf{X}_1|(\mathbf{X}_2, Y) &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\alpha}'\mathbf{f}_y' + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}} \\
&= \bar{\boldsymbol{\mu}}_1 + (\mathbf{Q}_{\boldsymbol{\Gamma}} + \mathbf{P}_{\boldsymbol{\Gamma}})\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\alpha}'\mathbf{f}_y' + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}} \\
&= \bar{\boldsymbol{\mu}}_1 + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}\mathbf{X}_2 + \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\alpha}'\mathbf{f}_y' + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}} \\
&= \bar{\boldsymbol{\mu}}_1 + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\alpha}'\mathbf{f}_y') + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}} \\
&= \bar{\boldsymbol{\mu}}_1 + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}\underbrace{(\boldsymbol{\Gamma}^T\boldsymbol{\beta} \quad \boldsymbol{\alpha}')}_{\boldsymbol{\alpha}}\underbrace{\begin{pmatrix}\mathbf{X}_2 \\ \mathbf{f}_y'\end{pmatrix}}_{\mathbf{f}_y} + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}}. \qquad (4.7)
\end{aligned}$$

Using the same definition of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\beta}_0$ as in Section 3.1, we can obtain the new model

$$\mathbf{X}_1|(\mathbf{X}_2, Y) = \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}}, \qquad (4.8)$$

where we are assuming without loss of generality that the $\mathbf{f}_{y_i}$'s are centered, $\sum_{i=1}^n \mathbf{f}_{y_i} = 0$. Model (4.8) is refered to as an *partial PFC model*. A sufficient reduction under model (4.8) is $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1$ by Corollary 3.1. When $\boldsymbol{\Omega}$ has the special structure

$\mathbf{\Omega} = \mathbf{\Gamma\Phi\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$, this reduction form can be simplified as $R(\mathbf{X}_1) = (\mathbf{\Gamma}_0\boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T\mathbf{X}_1$ by the following corollary. The proof is given in Appendix A.6.

Corollary 4.2. *If* $\mathbf{\Omega} = \mathbf{\Gamma\Phi\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$, *then* $R(\mathbf{X}_1) = (\mathbf{\Gamma}_0\boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T\mathbf{\Omega}^{-1}\mathbf{X}_1$ *is equivalent to* $R(\mathbf{X}_1) = (\mathbf{\Gamma}_0\boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T\mathbf{X}_1$.

Like the expansion of the PFC model, we can consider four types of partial PFC models as developing model (4.8) with different types of error structures: (1) isotropic partial PFC model, (2) diagonal partial PFC model, (3) general partial PFC model, and (4) partial PFC model with the specific variance $\mathbf{\Omega} = \mathbf{\Gamma\Phi\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$. In the following sections we will consider these four partial PFC models and estimate the parameters of interest for the sufficient reduction using the maximum likelihood method.

4.2 Isotropic partial PFC model and estimation

When $\mathbf{\Omega} = \sigma^2\mathbf{I}_{p_1}$, we have the isotropic version of model (4.8):

$$\mathbf{X}_1|(\mathbf{X}_2, Y) = \bar{\boldsymbol{\mu}}_1 + \mathbf{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \mathbf{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon}. \tag{4.9}$$

In fact, when model (4.4) and model (4.5) with $\boldsymbol{\omega} = 0$ go through all steps in equation (4.7), they result in the same equation as model (4.9). This means that model (4.9) can be considered when, regardless of the randomness of $\boldsymbol{\nu}'$, the latent variable $\boldsymbol{\nu}'$ is modeled by $\boldsymbol{\nu}' = \boldsymbol{\alpha}'\mathbf{f}_y'$ without approximation error. In addition, when in model (4.6) a random variable $\boldsymbol{\nu}'$ is modeled by $\boldsymbol{\alpha}'\mathbf{f}_y'$ and the error covariance $\mathbf{\Omega} = \sigma^2\mathbf{I}_{p_1}$ is constructed, model (4.9) is also considered

Under model (4.9), the maximum likelihood method is again used for estimation of a sufficient reduction. The full log likelihood is

$$\begin{aligned} L_d(\bar{\boldsymbol{\mu}}_1, \boldsymbol{\beta}_0, \mathcal{S}_{\mathbf{\Gamma}}, \boldsymbol{\alpha}, \sigma^2) \ &= \ -\frac{np_1}{2}\log(2\pi) - \frac{np_1}{2}\log(\sigma^2) \\ &\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\|\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \mathbf{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \mathbf{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\right\|^2. \end{aligned}$$

Let $\mathbf{F}$ denote the $n \times r$ matrix with rows $\mathbf{f}_{y_i}{}^T$. Then $n \times p_1$ matrix of fitted vectors from the regression of $\mathbf{X}_{1i}$ on $\mathbf{f}_{y_i}$ is $\mathbf{P}_{\mathbf{F}}\mathbb{X}_1$ and $\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} = \mathbb{X}_1^T\mathbf{P}_{\mathbf{F}}\mathbb{X}_1/n$, where $\mathbf{P}_{\mathbf{F}}$ denotes the

linear operator that projects onto the subspace spanned by the columns of $\mathbf{F}$. Holding $\mathbf{\Gamma}$, $\mathbf{\Gamma}_0$, and σ fixed and maximizing the likelihood over $\bar{\boldsymbol{\mu}}_1$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\alpha}$, we obtain $\widehat{\bar{\boldsymbol{\mu}}}_1 = \bar{\mathbf{X}}_1 - \mathbf{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_2$, $\widehat{\boldsymbol{\beta}}_0 = {\mathbf{\Gamma}_0}^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$, and $\widehat{\boldsymbol{\alpha}} = \mathbf{\Gamma}^T\mathbb{X}_1^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$.

Substituting back all these estimates and using $\mathbf{P}_{\mathbf{\Gamma}_0} = \mathbf{I}_{p_1} - \mathbf{P}_{\mathbf{\Gamma}}$, the final partially maximized log likelihood L_d is then, apart from constants (see Appendix A.7 for details),

$$\begin{aligned} & L_d(\mathcal{S}_{\mathbf{\Gamma}}, \sigma^2) \\ & = -\frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\Big\{\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_1\right] - \mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}\right] - \mathrm{tr}\left[\mathbf{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2})\right]\Big\}. \end{aligned} \tag{4.10}$$

Holding σ^2 fixed, the likelihood depends only on $\mathcal{S}_{\mathbf{\Gamma}}$. The rank of $\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}$ is at most $r + p_2$ and typically $\mathrm{rank}(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}) = r + p_2$. In any event, we assume that $\mathrm{rank}(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}) \geq d$. The likelihood is then maximized by setting $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ equal to the span of eigenvectors corresponding to the largest d eigenvalues $\lambda_i\left(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}\right)$, $i = 1, \ldots, d$. The corresponding estimate of scale is

$$\widehat{\sigma}^2 = \Big(\sum_{i=1}^{p_1}\lambda_i\left(\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2}\right) - \sum_{i=1}^{d}\lambda_i\Big(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}\Big)\Big)/p_1.$$

A sufficient reduction under model (4.9) is then estimated by $\widehat{R}(\mathbf{X}_1) = (\widehat{\mathbf{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0,\ \widehat{\mathbf{\Gamma}})^T\mathbf{X}_1$, where $\widehat{\mathbf{\Gamma}}$ is any orthonormal basis for the MLE of span($\mathbf{\Gamma}$) and $\widehat{\mathbf{\Gamma}}_0$ is a completion of $\widehat{\mathbf{\Gamma}}$. Since the sufficient reduction is estimated by using the sample partial principal fitted components (PFC) of $\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}$, we will refer to model (4.9) as an *isotropic partial PFC model*.

4.3 Diagonal partial PFC model and estimation

When $\boldsymbol{\nu}'$ is a random variable and model (4.8) with a diagonal error covariance matrix $\mathbf{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{p_1}^2)$ is considered, we have the model

$$\mathbf{X}_1|(\mathbf{X}_2, Y) = \bar{\boldsymbol{\mu}}_1 + \mathbf{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \mathbf{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \mathbf{\Omega}^{1/2}\boldsymbol{\varepsilon}, \tag{4.11}$$

where $\boldsymbol{\varepsilon}$ is normally distributed with mean 0 and covariance matrix $\mathbf{I}_{p_1}$. While the isotropic partial PFC model (4.9) requires that, given the response Y and $\mathbf{X}_2$, the

predictors $\mathbf{X}_1$ must be independent and have the same variance, this model allows for different measurement scales of the predictors, but the predictors are still conditionally independent given Y and $\mathbf{X}_2$. We call model (4.11) a *diagonal partial PFC model*. The log likelihood for this setting is

$$\begin{aligned} L_d(\bar{\boldsymbol{\mu}}_1, \boldsymbol{\beta}_0, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\alpha}, \boldsymbol{\Omega}) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Omega}| \\ & -\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\right)^T \boldsymbol{\Omega}^{-1} \\ & \qquad \left(\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\right), \end{aligned} \tag{4.12}$$

where we still assume that the $\mathbf{f}_{y_i}$'s are centered. For fixed $\boldsymbol{\Omega}$, $\boldsymbol{\Gamma}_0$, and $\boldsymbol{\Gamma}$, this log likelihood is maximized over $\bar{\boldsymbol{\mu}}_1$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\alpha}$ by the arguments $\widetilde{\bar{\boldsymbol{\mu}}}_1 = \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_2$, $\widetilde{\boldsymbol{\beta}}_0 = \boldsymbol{\Gamma}_0^T\mathbf{P}_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})}\widehat{\mathbf{E}}$, and $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\widehat{\mathbf{G}}$. Here $\widehat{\mathbf{E}} = \mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$ and $\widehat{\mathbf{G}} = \mathbb{X}_1^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$ are the coefficient matrices from the multivariate linear regression of $\mathbf{X}_1$ on $\mathbf{X}_2$ and that of $\mathbf{X}_1$ on $\mathbf{f}_y$ respectively. $\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1}$ is the projection onto $\mathcal{S}_{\boldsymbol{\Gamma}}$ in the $\boldsymbol{\Omega}^{-1}$ inner product. Similarly $\mathbf{P}_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})}$ is the projection onto $\mathcal{S}_{\boldsymbol{\Gamma}_0}$ with the same inner product. To find $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_0}$, $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{\Omega}}$ we substitute $\widetilde{\bar{\boldsymbol{\mu}}}_1$, $\widetilde{\boldsymbol{\beta}}_0$, and $\widetilde{\boldsymbol{\alpha}}$ into the log likelihood to obtain the partially maximized form (see Appendix A.8 for details)

$$\begin{aligned} L_d(\mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\Omega}) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Omega}| \\ & -\frac{n}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\boldsymbol{\Omega}^{-1/2} \\ & \qquad -\mathbf{P}_{\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}}\boldsymbol{\Omega}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|2}\right)\boldsymbol{\Omega}^{-1/2}\Big]. \end{aligned} \tag{4.13}$$

Holding $\boldsymbol{\Omega}$ fixed, this is maximized by choosing $\mathbf{P}_{\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}}$ to project onto the space spanned by the first d eigenvectors of $\boldsymbol{\Omega}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|2}\right)\boldsymbol{\Omega}^{-1/2}$. This leads to the final partially maximized log likelihood

$$\begin{aligned} L_d(\boldsymbol{\Omega}) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right] \\ & -\frac{n}{2}\sum_{i=1}^{d}\lambda_i\left(\boldsymbol{\Omega}^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|2}\right)\right). \end{aligned} \tag{4.14}$$

The MLE $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ is then obtained by maximizing (4.14). For a diagonal variance structure, (4.14) evidently does not have a closed form solution and iteration must be used.

An algorithm could be developed to maximize (4.14) directly, but a straightforward alternating algorithm has been developed for the diagonal PFC model (Adragni and Cook, 2009), and it is fast and easy to understand. We follow this algorithmic scheme to estimate $\boldsymbol{\Omega}$ in the diagonal partial PFC model (4.11).

A new alternating algorithm can be constructed by adapting the algorithm for the diagonal PFC model. Once the inverse mean function is specified, the variance $\sigma_i^2, i = 1, \ldots, p_1$, can be estimated by using the sample variances of the centered variables $\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}$. If $\boldsymbol{\Omega}$ is specified then we can standardize the predictor vector to obtain an isotropic partial PFC model in $\mathbf{Z}_1 = \boldsymbol{\Omega}^{-1/2}\mathbf{X}_1$:

$$\mathbf{Z}_1 = \boldsymbol{\Omega}^{-1/2}\bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \boldsymbol{\varepsilon}. \tag{4.15}$$

Consequently, we can estimate span($\boldsymbol{\Gamma}$) as $\boldsymbol{\Omega}^{1/2}$ times the estimate $\widetilde{\boldsymbol{\Gamma}}$ of $\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}$ from the isotropic model (4.15). Alternating between these two steps leads to the following algorithm:

1. Fit the isotropic partial PFC model to the original data, getting initial estimates $\widehat{\boldsymbol{\Gamma}}_{(1)}$, $\widehat{\boldsymbol{\Gamma}}_{0(1)}$, $\widehat{\boldsymbol{\beta}}_{0(1)}$, and $\widehat{\boldsymbol{\alpha}}_{(1)}$.
2. For some small $\delta > 0$, repeat for $j = 1, 2, \ldots$ until $\text{tr}\left[\left(\widehat{\boldsymbol{\Omega}}_{(j)} - \widehat{\boldsymbol{\Omega}}_{(j+1)}\right)^2\right] < \delta$
   (a) Calculate $\widehat{\boldsymbol{\Omega}}_{(j)} = (1/n)\text{diag}\{(\mathbb{X}_1 - \mathbb{X}_2\widehat{\boldsymbol{\beta}}_{0(j)}^T\widehat{\boldsymbol{\Gamma}}_{0(j)}^T - \mathbf{F}\widehat{\boldsymbol{\alpha}}_{(j)}^T\widehat{\boldsymbol{\Gamma}}_{(j)}^T)^T\,(\mathbb{X}_1 - \mathbb{X}_2\widehat{\boldsymbol{\beta}}_{0(j)}^T\widehat{\boldsymbol{\Gamma}}_{0(j)}^T - \mathbf{F}\widehat{\boldsymbol{\alpha}}_{(j)}^T\widehat{\boldsymbol{\Gamma}}_{(j)}^T)\}$.
   (b) Transform $\mathbf{Z}_1 = \widehat{\boldsymbol{\Omega}}_{(j)}^{-1/2}\mathbf{X}_1$,
   (c) Fit the isotropic Partial PFC model to $\mathbf{Z}_1$, yielding estimates $\widetilde{\boldsymbol{\Gamma}}$, $\widetilde{\boldsymbol{\Gamma}}_0$, $\widetilde{\boldsymbol{\beta}}_0$, and $\widetilde{\boldsymbol{\alpha}}$.
   (d) Backtransform the estimates to the original scale $\widehat{\boldsymbol{\Gamma}}_{(j+1)} = \widehat{\boldsymbol{\Omega}}_{(j)}^{1/2}\widetilde{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Gamma}}_{0(j+1)} = \widehat{\boldsymbol{\Omega}}_{(j)}^{1/2}\widetilde{\boldsymbol{\Gamma}}_0$, $\widehat{\boldsymbol{\beta}}_{0(j+1)} = \widetilde{\boldsymbol{\beta}}_0$, and $\widehat{\boldsymbol{\alpha}}_{(j+1)} = \widetilde{\boldsymbol{\alpha}}$.

Since the MLE $\widehat{\boldsymbol{\Omega}}$ is now obtained, we can develop the remaining estimated parameters. Let $\widehat{\boldsymbol{\Psi}} = (\widehat{\boldsymbol{\psi}}_1, \ldots, \widehat{\boldsymbol{\psi}}_{p_1})$ be the eigenvectors corresponding to the eigenvalues

$$\lambda_1\left(\widehat{\boldsymbol{\Omega}}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2})\widehat{\boldsymbol{\Omega}}^{-1/2}\right) > \cdots > \lambda_{p_1}\left(\widehat{\boldsymbol{\Omega}}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2})\widehat{\boldsymbol{\Omega}}^{-1/2}\right).$$

Let $\widehat{\boldsymbol{\Psi}}_d = (\widehat{\boldsymbol{\psi}}_1, \ldots, \widehat{\boldsymbol{\psi}}_d)$ and $\widehat{\boldsymbol{\Psi}}_{p_1-d} = (\widehat{\boldsymbol{\psi}}_{d+1}, \ldots, \widehat{\boldsymbol{\psi}}_{p_1})$. Then the MLEs of the parameters become $\text{span}(\widehat{\boldsymbol{\Omega}}^{-1/2}\widehat{\boldsymbol{\Gamma}}) = \text{span}(\widehat{\boldsymbol{\Psi}}_d)$, $\text{span}(\widehat{\boldsymbol{\Omega}}^{-1/2}\widehat{\boldsymbol{\Gamma}}_0) = \text{span}(\widehat{\boldsymbol{\Psi}}_{p_1-d})$, $\widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\Psi}}_{p_1-d}^T\widehat{\boldsymbol{\Omega}}^{-1/2}\widehat{\mathbf{E}}$, $\widehat{\boldsymbol{\mu}}_1 = \bar{\mathbf{X}}_1 - \widehat{\boldsymbol{\Omega}}^{1/2}\widehat{\boldsymbol{\Psi}}_{p_1-d}\widehat{\boldsymbol{\beta}}_0\bar{\mathbf{X}}_2$, and $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\Psi}}_d^T\widehat{\boldsymbol{\Omega}}^{-1/2}\widehat{\mathbf{G}}$. The sufficient reduction can be estimated as $\widehat{R}(\mathbf{X}_1) = (\widehat{\boldsymbol{\Psi}}_{p_1-d}\widehat{\boldsymbol{\beta}}_0,\ \widehat{\boldsymbol{\Psi}}_d)^T\widehat{\boldsymbol{\Omega}}^{-1/2}\mathbf{X}_1$.

4.4 General partial PFC model and estimation

We consider a more general version of model (4.8) by allowing for unstructured error covariance matrix, $\boldsymbol{\Omega} > 0$ while all other terms remain the same as defined previously in (4.11). Under the model with this general error covariance matrix, which will be called a *general partial PFC model*, the predictors can be conditionally dependent with different variances. Apart from the error structure, the model is exactly the same as model (4.11). Consequently, working with the same log likelihood, all MLEs for the parameters except $\boldsymbol{\Omega}$ are the same as those in the previous section: $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Omega}}\mathcal{S}_d\left(\widehat{\boldsymbol{\Omega}},\ \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2}\right)$, $\widehat{\boldsymbol{\beta}}_0 = (\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}}_0)^{-1}\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\mathbf{E}}$, $\widehat{\boldsymbol{\mu}}_1 = \bar{\mathbf{X}}_1 - \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0\bar{\mathbf{X}}_2$, and $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\Gamma}}^T\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\mathbf{G}}$, where $\widehat{\boldsymbol{\Gamma}}$ is any orthonormal basis for $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Gamma}}_0$ is the completion of $\widehat{\boldsymbol{\Gamma}}$. It follows that the sufficient reduction is of the form $\widehat{R}(\mathbf{X}_1) = (\widehat{\mathbf{V}}_{p_1-d}\widehat{\boldsymbol{\beta}}_0,\ \widehat{\mathbf{V}}_d)^T\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X}_1$, where $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_{p_1})$ are the eigenvectors of $\widehat{\boldsymbol{\Omega}}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2}\right)\widehat{\boldsymbol{\Omega}}^{-1/2}$ and $\widehat{\mathbf{V}}_d = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_d)$ and $\widehat{\mathbf{V}}_{p_1-d} = (\widehat{\mathbf{v}}_{d+1}, \ldots, \widehat{\mathbf{v}}_{p_1})$. Now we need to find the MLE of $\boldsymbol{\Omega}$. The following theorem shows how to construct $\widehat{\boldsymbol{\Omega}}$. Its proof is given in Appendix A.9.

Theorem 4.1. *Let $\widehat{\mathbf{V}}$ and $\widehat{\boldsymbol{\Lambda}} = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{p_1})$ be the matrices of the ordered eigenvectors and eigenvalues of $(\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2})^{-1/2}$, and assume that the nonzero $\widehat{\lambda}_i$'s are distinct. Then, the maximum of $L_d(\boldsymbol{\Omega})$ (4.14) over $\boldsymbol{\Omega} > 0$ is attained at $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2} + (\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2})^{1/2}\widehat{\mathbf{V}}\widehat{\mathbf{K}}\widehat{\mathbf{V}}^T(\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2})^{1/2}$, where $\widehat{\mathbf{K}} = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d, 0, \ldots, 0)$. The maximum value of the log likelihood is*

$$L_d = -\frac{np_1}{2} - \frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log\left|\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2}\right| - \frac{n}{2}\sum_{i=1}^{d}\log\left(1+\widehat{\lambda}_i\right). \tag{4.16}$$

The following corollary confirms the invariance of $\widehat{R}$ under full rank linear transformations of $\mathbf{X}_1$. Its proof is given in Appendix A.10.

Corollary 4.3. *If* $\mathbf{A} \in \mathbb{R}^{p_1 \times p_1}$ *has full rank, then* $\widehat{R}(\mathbf{X}_1) = \widehat{R}(\mathbf{A}\mathbf{X}_1)$.

The next corollary gives five equivalent forms for the MLE of $\boldsymbol{\Omega}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$. The proof is given in Appendix A.11.

Corollary 4.4. *The following are equivalent expressions for the MLE of* $\boldsymbol{\Omega}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ *under model* (4.11) *with* $\boldsymbol{\Omega} > 0$: $\mathcal{S}_d\left(\widehat{\boldsymbol{\Omega}},\ \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{res}}\right) = \mathcal{S}_d\left(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}},\ \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{res}}\right) = \mathcal{S}_d\left(\widehat{\boldsymbol{\Omega}},\ \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit}}\right) = \mathcal{S}_d\left(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}},\ \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit}}\right) = \mathcal{S}_d\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{res}},\ \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit}}\right)$, *where* $\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{res}} = \widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}$ *and* $\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{2}}_{\text{res}} = \widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{2}}_{\text{fit}}$.

All expressions for the MLE of $\boldsymbol{\Omega}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ use $\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} = \mathbb{X}_1^T\mathbf{P}_{\mathbf{F}}\mathbb{X}_1/n$ which contains the response information through the known function $\mathbf{f}_y$. The first and second forms indicate that the MLE of $\boldsymbol{\Omega}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ under the general partial PFC model can be computed by using the linearly transformed predictors $\widehat{\boldsymbol{\Omega}}^{-1/2}\mathbf{X}_1$ and $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}}\mathbf{X}_1$ respectively. The remaining forms indicate that the MLE of $\boldsymbol{\Omega}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ can be obtained also by using $\mathbf{A}^{-1/2}\mathbf{X}_1$, where $\mathbf{A}$ is $\widehat{\boldsymbol{\Omega}}$, $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}}$ or $\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{res}}$ for each corresponding term. Any of these five form can be used in practice since they each give the same estimated subspace, but we tend to use $\mathcal{S}_d\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{res}},\ \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit}}\right)$ for no compelling reason.

4.5 Partial PFC model with $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$

Consider model (4.8) with the variance function $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$. A sufficient reduction under this model is $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\mathbf{X}_1$ by Corollary 4.2. The full log likelihood is

$$
\begin{aligned}
& L_d(\bar{\boldsymbol{\mu}}_1, \boldsymbol{\beta}_0, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\alpha}, \boldsymbol{\Phi}, \sigma^2) \\
&= -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}| \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_1 - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y)^T(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1})^{-1} \\
&\qquad\qquad (\mathbf{X}_1 - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y).
\end{aligned}
$$

For fixed $\boldsymbol{\Gamma}_0$, $\boldsymbol{\Gamma}$, and σ, this log likelihood is maximized over $\bar{\boldsymbol{\mu}}_1$, $\boldsymbol{\beta}_0$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Phi}$ by the arguments $\widehat{\bar{\boldsymbol{\mu}}}_1 = \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_2$, $\widehat{\boldsymbol{\beta}}_0 = \boldsymbol{\Gamma}_0^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$, $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T\mathbb{X}_1^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$,

and $\widehat{\mathbf{\Phi}} = \mathbf{\Gamma}^T \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}} \mathbf{\Gamma} - \sigma^2 \mathbf{I}_d$ using the inverse equivalence $\mathbf{\Omega}^{-1} = (\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \sigma^2 \mathbf{I}_{p_1})^{-1} = \mathbf{\Gamma}(\mathbf{\Phi} + \sigma^2 \mathbf{I}_d)^{-1}\mathbf{\Gamma}^T + \sigma^{-2}\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T$. Substituting these estimated parameters into the log likelihood, the partially maximized form is (see Appendix A.12 for details),

$$\begin{aligned} L_d(\mathcal{S}_{\mathbf{\Gamma}}, \sigma^2) = & -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Gamma}^T \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}} \mathbf{\Gamma}| - \frac{n}{2}(p_1 - d)\log(\sigma^2) \\ & - \frac{nd}{2} - \frac{n}{2\sigma^2}\text{tr}\left[\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}\right] + \frac{n}{2\sigma^2}\text{tr}\left[\mathbf{\Gamma}\mathbf{\Gamma}^T \widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}\right]. \end{aligned} \tag{4.17}$$

Again this partially maximized log-likelihood function is maximized over σ^2 by $\widehat{\sigma}^2 = \left(\text{tr}\left[\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}\right] - \text{tr}\left[\mathbf{\Gamma}^T \widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} \mathbf{\Gamma}\right]\right)/(p_1 - d)$. Equation (4.17) can be rewritten as

$$\begin{aligned} L_d(\mathcal{S}_{\mathbf{\Gamma}}) = & -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Gamma}^T \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}} \mathbf{\Gamma}| - \frac{n}{2}(p_1 - d)\log\left(\text{tr}\left[\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}\right] - \text{tr}\left[\mathbf{\Gamma}^T \widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} \mathbf{\Gamma}\right]\right) \\ & - \frac{nd}{2} + \frac{n}{2}(p_1 - d)\log(p_1 - d) - \frac{n}{2}(p_1 - d). \end{aligned} \tag{4.18}$$

Now we are ready to estimate $\mathcal{S}_{\mathbf{\Gamma}}$. However, we were unable to find a closed-form solution to $\arg\max_{\mathcal{S}_{\mathbf{\Gamma}}} L_d(\mathcal{S}_{\mathbf{\Gamma}})$, and so it was necessary to use numerical optimization. Lippert's *sg_min 2.4.1* (*http://www-math.mit.edu/∼lippert/sgmin.html*) was adapted for this Grassmann optimization with analytic first derivatives and numerical second derivatives. We used a dog-leg step algorithm which interpolates a steepest descent and a Newton's method step on $\mathcal{G}_{(d,\, p_1)}$.

Here we show a small simulation of the numerical optimization using the suggested method to estimate $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$. With $p = p_1 + p_2$ and $d = 1$, where $p_1 = 10$ and $p_2 = 2$, data matrices were generated as $\mathbf{X}_2 = \mathbf{\Gamma}_2 y + \boldsymbol{\epsilon}_2$ and $\mathbf{X}_1 = \boldsymbol{\beta}\mathbf{X}_2 + \mathbf{\Gamma} e^y + \boldsymbol{\varepsilon}$, where $\mathbf{\Gamma}_2$, y, $\boldsymbol{\beta}$, and $\mathbf{\Gamma}$ were sampled independently from a standard normal distribution. The error term $\boldsymbol{\epsilon}_2 \in \mathbb{R}^{p_2}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^{p_1}$ are a vector of independent $N(\mathbf{0}, \mathbf{I}_{p_2})$ and $N(\mathbf{0}, \mathbf{I}_{p_1})$ variates respectively. We assumed that $\boldsymbol{\nu}'$ is modeled as $\mathbf{f}'_y = y^2$ in the model (2.5) with isotropic error covariance while expecting an approximation error. For a selected value of n we then computed the angle between the true $\mathbf{\Gamma}$ and the estimated basis for $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ obtained from the numerical optimization method. The entire process was repeated 100 times, yielding the results shown in Figure 4.1. These boxplots show that the variability and location of angles consistently decrease as n increases from 20 to 160, which supports that a numerical solution gives the reasonable estimation about $\mathcal{S}_{\mathbf{\Gamma}}$. When n is greater

than 20, we can see that there are relatively large angle values apart from whiskers. This indicates that the numerical optimization resulted in a local maximum rather than the global maximum.

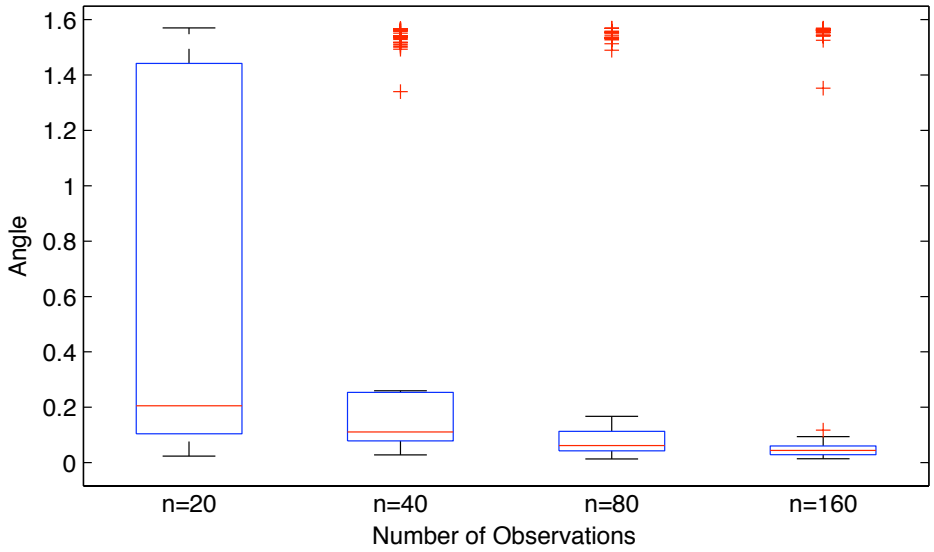


Figure 4.1: Simulation results showing angles in degrees for estimated value and true value of $\mathcal{S}_{\mathbf{\Gamma}}$ in model (4.8) with the variance function $\mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$.

Chapter 5

Combining PFC and probabilistic PCA models

In Chapter 4 we assumed that the responses for all cases are known. In this chapter we assume that only a part of responses is known; that is, the whole dataset can be divided into two parts:

(a) the dataset comprised of $\mathbf{X}$ and corresponding response.

(b) the dataset with $\mathbf{X}$ only.

When $\mathbf{X}$ comes from dataset (a), we can tailor our reduction to the known response with a PFC model. Otherwise, the probabilistic PCA model is the way to go. In Section 5.1 our focus is on the reduction of $\mathbf{X}$ when dealing with the problem of combining PFC and probabilistic PCA models. The focus moves on to the reduction on $\mathbf{X}_1$ in Section 5.2. All the procedures of the combining model derivation and parameter estimation in Section 5.1 are revisited in the context of partial reduction. That is, Section 5.2 covers how to combine partial PFC and partial probabilistic PCA models.

5.1 Combining models

Suppose that the data consist of two parts, one part with known responses and the other with unknown responses. Then under model (2.1), only for the cases with known responses can the latent variable $\boldsymbol{\nu}^*$ be modeled as a function of y, $\boldsymbol{\alpha}^*\mathbf{f}_y^*$; that is, model (4.1) is considered for such cases. On the other hand, for the cases without a response there is no change in model (2.1).

5.1.1 Latent variable $\boldsymbol{\nu}^*$ is fixed

Assume that $\boldsymbol{\nu}^*$ is fixed and, for the cases with known responses, modeled by $\boldsymbol{\alpha}^*\mathbf{f}_y^*$ without approximation error, where $\boldsymbol{\alpha}^*$ and $\mathbf{f}_y^*$ are as defined previously in (4.1). Simply by introducing an indicator function J, model (2.1) can be written as

$$\mathbf{X}|(\boldsymbol{\nu}^* \text{ or } Y) = \bar{\boldsymbol{\mu}}^* + \boldsymbol{\Gamma}^*(\boldsymbol{\alpha}^*\mathbf{f}_y^* J + \boldsymbol{\nu}^*(1-J)) + \sigma\boldsymbol{\epsilon}, \tag{5.1}$$

where $J = 1$ if Y is observed and $J = 0$ if Y is unobserved. If there is an approximation error when modeling this fixed $\boldsymbol{\nu}^*$ as $\boldsymbol{\alpha}^*\mathbf{f}_y^*$, the bias term $\boldsymbol{\Gamma}^*(\boldsymbol{\nu}^* - \boldsymbol{\alpha}^*\mathbf{f}_y^*)J$ is then added in model (5.1) and should be considered. The model with this bias term is not studied in the thesis and investigations into the estimation of a sufficient reduction is left for future work.

Model (5.1) can be seen as a combination of PC model (2.1) and isotropic PFC model (4.1) since it becomes the PC model when $J = 0$ and the isotropic PFC model when $J = 1$. In addition, a sufficient reduction is still $R(\mathbf{X}) = \boldsymbol{\Gamma}^{*T}\mathbf{X}$ because these two models have the same form. Suppose that l of n y's are known. Using the maximum likelihood method to estimate a sufficient reduction, the full log likelihood is

$$\begin{aligned} L_{d^*}(\bar{\boldsymbol{\mu}}^*, \mathcal{S}_{\boldsymbol{\Gamma}^*}, \boldsymbol{\alpha}^*, \boldsymbol{\nu}^*, \sigma^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{l}\left\|\mathbf{X}_i - \bar{\boldsymbol{\mu}}^* - \boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_{y_i}^*\right\|^2 \\ & - \frac{1}{2\sigma^2}\sum_{i=l+1}^{n}\left\|\mathbf{X}_i - \bar{\boldsymbol{\mu}}^* - \boldsymbol{\Gamma}^*\boldsymbol{\nu}_i^*\right\|^2. \end{aligned}$$

Let $\mathbb{X}_{(\mathrm{k})}$ denote the $l \times p$ matrix with rows $(\mathbf{X} - \bar{\mathbf{X}})^T$ where Y is observed, and let $\mathbb{X}_{(\mathrm{u})}$ denote the corresponding $(n-l) \times p$ matrix where Y is not observed.

The parentheses in the subscript indicate that the total mean of the variable with all n samples is subtracted from the variable. We will also use the same subscript without parenthesis to indicate that a partial mean of the variable is subtracted while using a part of samples. For example, $\mathbb{X}_{\mathrm{k}}$ denotes the $l \times p$ matrix with rows $(\mathbf{X} - \bar{\mathbf{X}}_{\mathrm{k}})^T$ for the cases with known responses, where $\bar{\mathbf{X}}_{\mathrm{k}} \in \mathbb{R}^p$ denotes the sample mean vector of the cases with known responses.

The maximum likelihood estimators of $\bar{\boldsymbol{\mu}}^*$, $\boldsymbol{\alpha}^*$, and $\boldsymbol{\nu}^*$ are obtained from the full log likelihood with the remaining parameters held fixed: $\widehat{\bar{\boldsymbol{\mu}}}^* = \bar{\mathbf{X}}$, $\widehat{\boldsymbol{\alpha}}^* = \boldsymbol{\Gamma}^{*T}\mathbb{X}_{(\mathrm{k})}^T\mathbf{F}_{\mathrm{k}}^*(\mathbf{F}_{\mathrm{k}}^{*T}\mathbf{F}_{\mathrm{k}}^*)^{-1}$, and $\widehat{\boldsymbol{\nu}}_i^* = \boldsymbol{\Gamma}^{*T}(\mathbf{X}_i - \bar{\mathbf{X}})$, where $\mathbf{F}_{\mathrm{k}}^*$ denotes the $l \times r$ matrix with rows $\mathbf{f}_y^{*T}$. Here $\mathbf{F}_{\mathrm{k}}^*$ has no parenthesis in the subscript, since it only belongs to the cases with known responses. Substituting back all these parameters, the final partially maximized log likelihood is then,

$$\begin{aligned} L_{d^*}(\mathcal{S}_{\boldsymbol{\Gamma}^*}, \sigma^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(\sigma^2) \\ & - \frac{n}{2\sigma^2}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}\right] + \frac{1}{2\sigma^2}\mathrm{tr}\left[\left\{l\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit(k)}} + (n-l)\widehat{\boldsymbol{\Sigma}}_{(\mathrm{u})}\right\}\mathbf{P}_{\boldsymbol{\Gamma}^*}\right], \end{aligned} \tag{5.2}$$

where $\widehat{\boldsymbol{\Sigma}}_{(\mathrm{u})} = \mathbb{X}_{(\mathrm{u})}^T\mathbb{X}_{(\mathrm{u})}/(n-l)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit(k)}} = \mathbb{X}_{(\mathrm{k})}^T\mathbf{P}_{\mathbf{F}_{\mathrm{k}}^*}\mathbb{X}_{(\mathrm{k})}/l$. Holding σ^2 fixed, the likelihood is then maximized by setting $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}^*}$ equal to the span of the eigenvectors corresponding to the largest d^* eigenvalues $\widehat{\tau}_i^*$, $i = 1, \ldots, d^*$, of $l\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit(k)}} + (n-l)\widehat{\boldsymbol{\Sigma}}_{(\mathrm{u})}$. This estimation makes sense in that the combining model (5.1) is just the combination of two models since $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}^*}$ was estimated by using the eigenvectors of $\widehat{\boldsymbol{\Sigma}}$ under the PC model and that of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$ under the isotropic PFC model respectively. The corresponding estimate of scale is $\widehat{\sigma}^2 = \sum_{i=1}^p \widehat{\lambda}_i\Big(\widehat{\boldsymbol{\Sigma}}\Big)/p - \sum_{i=1}^{d^*}\widehat{\tau}_i^*/(np)$, where the derivation requires that $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res(k)}} = \widehat{\boldsymbol{\Sigma}}_{(\mathrm{k})} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit(k)}} > 0$.

Under the PC model, the isotropic PFC model, and combining model (5.1) the sufficient reduction was obtained as $R(\mathbf{X}) = \boldsymbol{\Gamma}^{*T}\mathbf{X}$. Therefore, our interest has been mainly focused on estimating $\boldsymbol{\Gamma}^*$. As mentioned in Chapter 4 once the response is known we should be able to tailor our reduction to that response and by doing so we expect the better estimation for the sufficient reduction. We performed a small simulation to fix these ideas and gain insights into the behavior of the estimation of the reductive subspace $\mathcal{S}_{\boldsymbol{\Gamma}^*}$ in the combining model. A dataset with $n = 1000$, $d^* = 1$, $r = 1$,

$p = 200$ was generated as $\mathbf{X}_i = \mathbf{\Gamma}^* \boldsymbol{\alpha}^* \mathbf{f}^*_{y_i} + \boldsymbol{\epsilon}$, where $\mathbf{\Gamma}^* = \mathbf{1}_p/\sqrt{p}$, $\boldsymbol{\alpha}^* = 1$, $\mathbf{f}^*_y = y^2$ with $y \sim N(0,1)$, and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ is $N(\mathbf{0}, \mathbf{I}_p)$ vector. Using $\mathbf{f}^*_y = (y, |y|)^T$ we estimated $\mathcal{S}_{\mathbf{\Gamma}^*}$ by setting it equal to the span of the first d^* eigenvectors of $\widehat{\mathbf{\Sigma}}$ in the PC model, $\widehat{\mathbf{\Sigma}}_{\text{fit}}$ in the isotropic PFC model, and $l\widehat{\mathbf{\Sigma}}_{\text{fit(k)}} + (n-l)\widehat{\mathbf{\Sigma}}_{\text{(u)}}$ in the combining model. We assume that there is no response for the PC model, we know all the response information for the PFC model, and only some of responses are known in the combining model. While varying the number of the known responses from $l = 900$ to $l = 50$ in the combining model, we computed the angle between $\mathcal{S}_{\mathbf{\Gamma}^*}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*}$ with $d^* = 1$. The entire process was repeated 100 times, yielding the results shown in Figure 5.1.

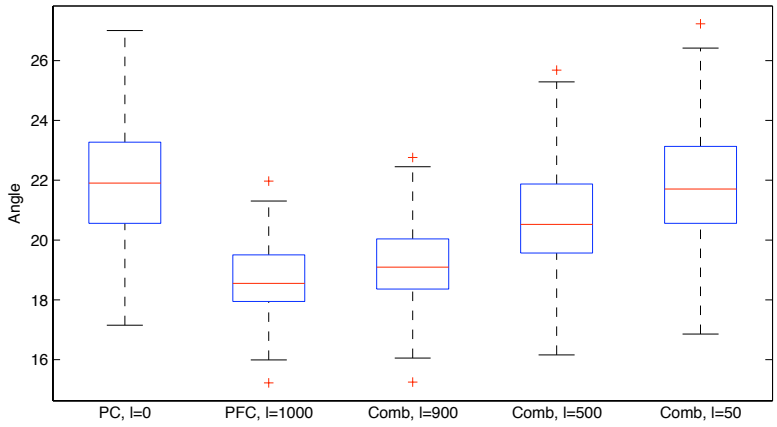


Figure 5.1: Simulation results showing the angles in degrees between $\mathcal{S}_{\mathbf{\Gamma}^*}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*}$ in each model with the various number of known responses l out of $n = 1000$, $l = \{900,\ 500,\ 50\}$ for the combining model.

This simulation results show how the amount of the known responses affects the estimation of $\mathcal{S}_{\mathbf{\Gamma}^*}$ in the combining model. First of all, the angles in the PFC model tend to be smaller than that in the PC model or the combining model as supporting our idea of having better estimation by using the response information. The boxplot of the combining model is very similar to that of the PFC model in the location and variance when the number of known responses is $l = 900$ which is closed to the number of whole response, $n = 1000$. As decreasing the number of known responses to $l = 50$

the combining model still performs a bit better than the PC model. Consequently the performance on the estimation of the sufficient reduction can be improved by using a combining model over PC model when a part of responses is observed.

When a dataset is divided into two parts, one part with known responses and the other without known responses, we might consider reducing the dimension of $\mathbf{X}$ separately using just each part. Let PC$(n-l)$ indicate that PC model is used for just $(n-l)$ samples where a response is not observed. Similarly PFC(l) indicates that we have only l samples with known responses and PFC model is used. In the same way the methods used in Figure 5.1 can be represented as PC(n) and PFC(n). We performed different simulation to compare the performance of PC$(n-l)$, PFC(l), and combining methods varying the number of known responses. The results of PC(n) and PFC(n) are also given to show the behavior where total samples are used. The methods used in the simulation are listed in Table 5.1 with how to estimate $\mathcal{S}_{\mathbf{\Gamma}^*}$ under the corresponding method.

Table 5.1: Dimension reduction methods used in the simulation

Method	Sample size	Number of known y	Number of unknown y	$\mathcal{S}_{\mathbf{\Gamma}^*}$ is estimated by the first d^* eigenvalues of
PCA(n)	n	0	n	$\widehat{\mathbf{\Sigma}}$
PFC(n)	n	n	0	$\widehat{\mathbf{\Sigma}}_{\text{fit}}$
Comb.	n	l	$n-l$	$l\widehat{\mathbf{\Sigma}}_{\text{fit(k)}} + (n-l)\widehat{\mathbf{\Sigma}}_{\text{(u)}}$
PCA$(n-l)$	$n-l$	0	$n-l$	$\widehat{\mathbf{\Sigma}}_{\text{(u)}}$
PFC(l)	l	l	0	$\widehat{\mathbf{\Sigma}}_{\text{fit(k)}}$

In this simulation we also considered different signal sizes to study the signal effect. Data were generated with $\mathbf{f}_y^* = \exp(y/10)$, where $y \sim N(0,1)$, for the weak signal case, $\mathbf{f}_y^* = \exp(y/2)$, where $y \sim N(0,1)$, for the moderate signal case, and $\mathbf{f}_y^* = \exp(y)$, where $y \sim N(0, 1.5^2)$, for the strong signal case. Except for this signal setting the data generation procedure was the same as for Figure 5.1 and $\mathbf{f}_y^* = (y, |y|, y^3, y^4, y^5)^T$ was still used for fitting.

Figure 5.2 shows how the amount of the known responses and the size of signal

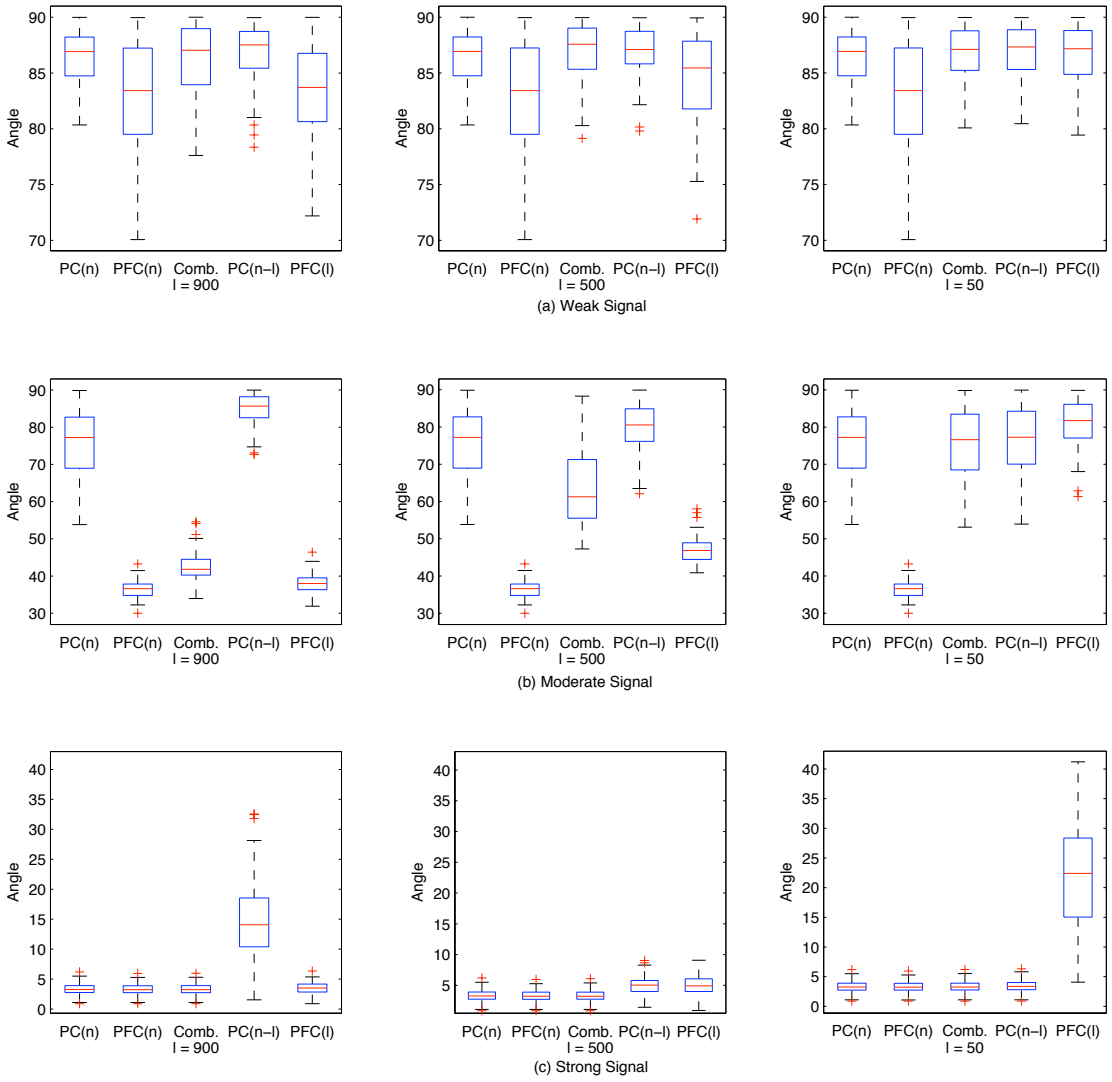


Figure 5.2: Simulation results showing the angles in degrees between $\mathcal{S}_{\mathbf{\Gamma}^*}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*}$ in each model with the various number of known responses l out of $n = 1000$, $l = \{900,\ 500,\ 50\}$ and three different signals, Weak, Moderate, Strong signals.

affect the estimation of $\mathcal{S}_{\mathbf{\Gamma}^*}$. Overall, boxplots of PFC(n) maintain the lowest location in every case supporting the importance of the response information in estimation.

When the number of known responses is 900, $l = 900$ and $n - l = 100$, PFC(900) is similar to PFC(1000), and PC(100) shows poor performance in every type of signals. If the small number of responses proportional to the total sample size are unknown and the signal is moderate or strong, combining model or PFC(900) are more desirable, showing better performance over PC(100). With the weak signal, PFC(900) might be a better choice than the combining or PCA(100) methods. Although the response information is unknown in PC(1000) and PC(100), better estimation for $\mathcal{S}_{\mathbf{\Gamma}^*}$ is obtained in PC(1000) by using more samples when the signal is moderate or strong. In the strong signal case, while degrees of PC(100) range from 0 to 30, other methods have degrees from 0 to 6 and the difference between boxplots is very slight.

When the number of known responses is 500, $l = 500$ and $n - l = 500$, PFC(1000) tends to have lower angles than PFC(500) as we expected. When the signal is weak PFC(500) is more desirable than the combining model or PC(500) method, and when the signal is strong there is no big difference between methods.

When the number of known response is 50, $l = 50$ and $n - l = 950$, the difference in angles between PFC(1000) and PFC(50) is quite large in every signal with poor performance in PFC(50). If we have a small number of known responses and the signal is strong, using PFC(50) is a bad idea, showing poor performance over the combining model and PC(950).

When the signal is weak, PC(n), PC($n-l$), and combining have very similar boxplots in variation and location with poor performance. It shows that the sample size used in PC model and the number of known responses in combining model does not affect the performance. When the signal is moderate, combining and PFC(l) gets close to PC($n-l$), showing poor performance as l is decreasing. When the signal is strong there is no big difference between PC(n), PFC(n), and combining methods having degrees as small as the range of 0 to 6.

To sum up, when the number of known response is large, PFC(l) is better than combining and PC($n - l$) methods. When the number of known response is small,

combining and PC($n-l$) are better choices over PFC(l). When the signal is weak we have to collect the response information as much as possible and then use PFC(l) rather than combining or PC($n-l$) methods. Although the difference is slight between methods when signal is strong, we suggest to use combining methods since they are not influenced by the number of known responses.

5.1.2 Latent variable $\boldsymbol{\nu}^*$ is random and modeled without approximation error

Assume that $\boldsymbol{\nu}^*$ is random and for the cases with known responses, modeled by $\boldsymbol{\alpha}^*\mathbf{f}^*$ without approximation error. Again introducing an indicator function J in model (2.1) we have the new model

$$\begin{aligned}\mathbf{X}|(\boldsymbol{\nu}^* \text{ or } Y) &= \bar{\boldsymbol{\mu}}^* + \big(\boldsymbol{\Gamma}^* E(\boldsymbol{\nu}^*|y) - \boldsymbol{\Gamma}^*(\boldsymbol{\nu}^* - E(\boldsymbol{\nu}^*|y))\big)J + \boldsymbol{\Gamma}^*\boldsymbol{\nu}^*(1-J) + \sigma\boldsymbol{\epsilon} \\ &= \bar{\boldsymbol{\mu}}^* + \big(\boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_y^* + \boldsymbol{\Gamma}^*\boldsymbol{\omega}^* + \sigma\boldsymbol{\epsilon}\big)J + \big(\boldsymbol{\Gamma}^*\boldsymbol{\nu}^* + \sigma\boldsymbol{\epsilon}\big)(1-J) \\ &= \bar{\boldsymbol{\mu}}^* + \big(\boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_y^* + \boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}\big)J + \boldsymbol{\epsilon}_{\mathrm{u}}(1-J), \end{aligned} \tag{5.3}$$

where $E(\boldsymbol{\nu}^*|y)$ can be written as $\boldsymbol{\alpha}^*\mathbf{f}_y^*$ because it is the function of y and $\boldsymbol{\omega}^*$ and $\boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}$ are as defined previously in Section 4.1.2. In the last equation, the new normal error $\boldsymbol{\epsilon}_{\mathrm{u}}$ has mean 0 and variance $\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$, since $\boldsymbol{\nu}^*$ is assumed to be normally distributed with mean 0 and identity covariance matrix. In fact $\boldsymbol{\omega}^*$, which represents the approximation error in the second equation, is equal to zero because we assume that there is no approximation error. Therefore, the combining model can be written as

$$\mathbf{X}|Y = \bar{\boldsymbol{\mu}}^* + \big(\boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_y^* + \sigma\boldsymbol{\epsilon}\big)J + \boldsymbol{\epsilon}_{\mathrm{u}}(1-J). \tag{5.4}$$

As we can see in the above model two different error covariance matrices are considered, $\sigma^2\mathbf{I}_p$ when Y is observed and $\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$ when Y is unobserved. This causes considerable difficulty when estimating the parameters, yielding a complicated estimator form for $\widehat{\bar{\boldsymbol{\mu}}}^*$. Therefore we allow for two different $\bar{\boldsymbol{\mu}}^*$'s according to the presence of the response; $\bar{\boldsymbol{\mu}}_{\mathrm{k}}^*$ where response is known and $\bar{\boldsymbol{\mu}}_{\mathrm{u}}^*$ where response is unknown. Then a combining model is obtained as

$$\mathbf{X}|Y = \big(\bar{\boldsymbol{\mu}}_{\mathrm{k}}^* + \boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_y^* + \sigma\boldsymbol{\epsilon}\big)J + \big(\bar{\boldsymbol{\mu}}_{\mathrm{u}}^* + \boldsymbol{\epsilon}_{\mathrm{u}}\big)(1-J). \tag{5.5}$$

Our goal is still to estimate $\mathcal{S}_{\mathbf{\Gamma}^*}$ for the sufficient reduction $R(\mathbf{X}) = \mathbf{\Gamma}^{*T}\mathbf{X}$. Under model (5.5) we have the full log likelihood

$$\begin{aligned} L_{d^*}(\bar{\boldsymbol{\mu}}^*_{\mathrm{k}}, \bar{\boldsymbol{\mu}}^*_{\mathrm{u}}, \mathcal{S}_{\mathbf{\Gamma}^*}, \boldsymbol{\alpha}^*, \sigma^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{lp}{2}\log(\sigma^2) - \frac{n-l}{2}\log|\mathbf{\Gamma}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p| \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{l}\left\|\mathbf{X}_i - \bar{\boldsymbol{\mu}}^*_{\mathrm{k}} - \mathbf{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}^*_{y_i}\right\|^2 \\ & - \frac{1}{2}\sum_{i=l+1}^{n}\left(\mathbf{X}_i - \bar{\boldsymbol{\mu}}^*_{\mathrm{u}}\right)^T\left(\mathbf{\Gamma}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p\right)^{-1}\left(\mathbf{X}_i - \bar{\boldsymbol{\mu}}^*_{\mathrm{u}}\right). \end{aligned}$$

Let $\bar{\mathbf{X}}_{\mathrm{k}} \in \mathbb{R}^p$ denote the sample mean vector of the cases with known responses and similarly let $\bar{\mathbf{X}}_{\mathrm{u}} \in \mathbb{R}^p$ denote that of the cases without a known response. Let $\mathbb{X}_{\mathrm{k}}$ denote the $l \times p$ matrix with rows $(\mathbf{X} - \bar{\mathbf{X}}_{\mathrm{k}})^T$ for the cases with known responses and likewise, $\mathbb{X}_{\mathrm{u}}$ denotes the $(n-l) \times p$ matrix with rows $(\mathbf{X} - \bar{\mathbf{X}}_{\mathrm{u}})^T$ for the cases without a known response. For fixed $\mathbf{\Gamma}^*$ and σ^2 the maximum likelihood estimators of $\bar{\boldsymbol{\mu}}^*_{\mathrm{k}}$, $\bar{\boldsymbol{\mu}}^*_{\mathrm{u}}$, and $\boldsymbol{\alpha}^*$ are $\widehat{\bar{\boldsymbol{\mu}}}^*_{\mathrm{k}} = \bar{\mathbf{X}}_{\mathrm{k}}$, $\widehat{\bar{\boldsymbol{\mu}}}^*_{\mathrm{u}} = \bar{\mathbf{X}}_{\mathrm{u}}$, and $\widehat{\boldsymbol{\alpha}}^* = \mathbf{\Gamma}^{*T}\mathbb{X}_{\mathrm{k}}^T\mathbf{F}^*_{\mathrm{k}}(\mathbf{F}^{*T}_{\mathrm{k}}\mathbf{F}^*_{\mathrm{k}})^{-1}$. To find $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*}$ and $\widehat{\sigma}^2$ we substitute these estimators into the log likelihood to obtain the partially maximized form

$$\begin{aligned} L_{d^*}(\mathcal{S}_{\mathbf{\Gamma}^*}, \sigma^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{np-(n-l)d^*}{2}\log(\sigma^2) - \frac{(n-l)d^*}{2}\log(\sigma^2+1) \\ & - \frac{l}{2\sigma^2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] + \frac{n-l}{2\sigma^2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right] \\ & + \frac{1}{2}\mathrm{tr}\left[\left\{\frac{1}{\sigma^2}\left(l\widehat{\mathbf{\Sigma}}_{\mathrm{fitk}} + (n-l)\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right) - \frac{n-l}{\sigma^2+1}\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right\}\mathbf{P}_{\mathbf{\Gamma}^*}\right], \end{aligned} \tag{5.6}$$

where $\widehat{\mathbf{\Sigma}}_{\mathrm{k}} = \mathbb{X}_{\mathrm{k}}^T\mathbb{X}_{\mathrm{k}}/l$, $\widehat{\mathbf{\Sigma}}_{\mathrm{u}} = \mathbb{X}_{\mathrm{u}}^T\mathbb{X}_{\mathrm{u}}/(n-l)$, and $\widehat{\mathbf{\Sigma}}_{\mathrm{fitk}} = \mathbb{X}_{\mathrm{k}}^T\mathbf{P}_{\mathbf{F}^*_{\mathrm{k}}}\mathbb{X}_{\mathrm{k}}/l$.

Now we need to estimate $\mathcal{S}_{\mathbf{\Gamma}^*}$ and σ^2. However, we are not able to find the closed-form solutions for them. Therefore it is necessary to use an alternating maximization algorithm. Holding σ^2 fixed, the likelihood is maximized by setting $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*}$ equal to the span of the first d^* eigenvectors of $\frac{1}{\sigma^2}\left(l\widehat{\mathbf{\Sigma}}_{\mathrm{fitk}} + (n-l)\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right) - \frac{n-l}{\sigma^2+1}\widehat{\mathbf{\Sigma}}_{\mathrm{u}}$. Once $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*}$ is determined, the estimator $\widehat{\sigma}^2$ can be obtained that maximizes likelihood (5.6). Then we set this value as the new designated σ^2 and find the new corresponding MLE of $\mathcal{S}_{\mathbf{\Gamma}^*}$. With this reasoning the alternating algorithm is defined as follows:

1. Assume there is no response and find the initial value $\widehat{\sigma}^2_{(1)}$ from the result of PC model, $\widehat{\sigma}^2_{(1)} = \sum_{j=d^*+1}^{p}\widehat{\lambda}_j\left(\widehat{\mathbf{\Sigma}}\right)/p$.

2. For some small $\delta > 0$, repeat for $j = 1, 2, \ldots$ until the maximum angle between $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j-1)}}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$ is less than δ,

   (a) For fixed $\widehat{\sigma}^2_{(j)}$, find $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$, which maximizes likelihood (5.6), as the span of the first d^* eigenvectors of $\dfrac{1}{\widehat{\sigma}^2_{(j)}}\left(l\widehat{\mathbf{\Sigma}}_{\text{fitk}} + (n-l)\widehat{\mathbf{\Sigma}}_{\text{u}}\right) - \dfrac{n-l}{\widehat{\sigma}^2_{(j)}+1}\widehat{\mathbf{\Sigma}}_{\text{u}}$.

   (b) With $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.6) (see Appendix A.13 for details).

Our ultimate goal with this algorithm is to find better estimation of $\mathcal{S}_{\mathbf{\Gamma}^*}$ rather than estimating σ^2. Hence the condition to stop the algorithm depends on the convergence of $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*}$ which achieves maximization of (5.6). Consequently, the sufficient reduction can be estimated as $\widehat{R}(\mathbf{X}) = \widehat{\mathbf{\Gamma}}^{*T}_{(j)}\mathbf{X}$ and the corresponding scale estimator is $\widehat{\sigma}^2_{(j+1)}$.

5.1.3 Latent variable $\boldsymbol{\nu}^*$ is random and modeled with approximation error

In this section we assume that $\boldsymbol{\nu}^*$ is random and, for the cases with known response, there is approximation error when modeling $\boldsymbol{\nu}^*$ as $\boldsymbol{\alpha}^*\mathbf{f}^*_y$. Then model (5.5) is reformulated as

$$\mathbf{X}|Y = \left(\bar{\boldsymbol{\mu}}^*_{\text{k}} + \mathbf{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}^*_y + \boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}\right)J + \left(\bar{\boldsymbol{\mu}}^*_{\text{u}} + \boldsymbol{\epsilon}_{\text{u}}\right)(1-J), \tag{5.7}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}$ is as defined in Chapter 4 and represented in (5.3), and is normally distributed with mean 0 and variance $\mathbf{\Omega}^*$. Like the four types of the error covariance structure of $\boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}$ considered in Chapter 4, four versions of model (5.7) can be discussed when $J = 1$: the model with (1) isotropic error, $\mathbf{\Omega}^* = \sigma^2_{\text{k}}\mathbf{I}_p$, (2) diagonal error, $\mathbf{\Omega}^* = \text{diag}\{\sigma_1, \ldots, \sigma_p\}$, (3) unstructured error, $\mathbf{\Omega}^* > 0$, and (4) the variance function $\mathbf{\Omega}^* = \mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$.

Isotropic error, $\mathbf{\Omega}^* = \sigma^2_{\text{k}}\mathbf{I}_p$. Assume that $\boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}$ has the isotropic covariance matrix $\sigma^2_{\text{k}}\mathbf{I}_p$. If σ^2_{k} is equal to σ^2 the result is the same as in Section 5.1.2. Here we allow σ^2_{k} to be different from σ^2. To emphasize this we will write σ^2 as σ^2_{u} because it is associated

with only the unknown responses. Then the full log likelihood is

$$\begin{aligned} L_{d^*}(\bar{\boldsymbol{\mu}}_{\mathrm{k}}^*, \bar{\boldsymbol{\mu}}_{\mathrm{u}}^*, \mathcal{S}_{\boldsymbol{\Gamma}^*}, \boldsymbol{\alpha}^*, \sigma_{\mathrm{k}}^2, \sigma_{\mathrm{u}}^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{lp}{2}\log(\sigma_{\mathrm{k}}^2) - \frac{n-l}{2}\log|\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{*T} + \sigma_{\mathrm{u}}^2\mathbf{I}_p| \\ & - \frac{1}{2\sigma_{\mathrm{k}}^2}\sum_{i=1}^{l}\left\|\mathbf{X}_i - \bar{\boldsymbol{\mu}}_{\mathrm{k}}^* - \boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_{y_i}^*\right\|^2 \\ & - \frac{1}{2}\sum_{i=l+1}^{n}\left(\mathbf{X}_i - \bar{\boldsymbol{\mu}}_{\mathrm{u}}^*\right)^T\left(\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{*T} + \sigma_{\mathrm{u}}^2\mathbf{I}_p\right)^{-1}\left(\mathbf{X}_i - \bar{\boldsymbol{\mu}}_{\mathrm{u}}^*\right). \end{aligned}$$

Holding $\boldsymbol{\Gamma}^*$, σ_{k}^2, and σ_{u}^2 fixed and substituting the estimators of $\bar{\boldsymbol{\mu}}_{\mathrm{k}}^*$, $\bar{\boldsymbol{\mu}}_{\mathrm{u}}^*$, and $\boldsymbol{\alpha}^*$, the partially maximized likelihood is

$$\begin{aligned} L_{d^*}(\mathcal{S}_{\boldsymbol{\Gamma}^*}, \sigma_{\mathrm{k}}^2, \sigma_{\mathrm{u}}^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{lp}{2}\log(\sigma_{\mathrm{k}}^2) - \frac{(n-l)d^*}{2}\log(\sigma_{\mathrm{u}}^2+1) \\ & - \frac{(n-l)(p-d^*)}{2}\log(\sigma_{\mathrm{u}}^2) - \frac{l}{2\sigma_{\mathrm{k}}^2}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{k}}\right] - \frac{n-l}{2\sigma_{\mathrm{u}}^2}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{u}}\right] \\ & + \frac{1}{2}\mathrm{tr}\left[\left\{\frac{l}{\sigma_{\mathrm{k}}^2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fitk}} + \frac{n-l}{\sigma_{\mathrm{u}}^2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{u}} - \frac{n-l}{\sigma_{\mathrm{u}}^2+1}\widehat{\boldsymbol{\Sigma}}_{\mathrm{u}}\right\}\mathbf{P}_{\boldsymbol{\Gamma}^*}\right]. \end{aligned} \tag{5.8}$$

Since $\mathcal{S}_{\boldsymbol{\Gamma}^*}$, σ_{k}^2, and σ_{u}^2 have no closed-form solutions, we propose the algorithm in the previous section to estimate them. The alternating algorithm should be modified with respect to likelihood (5.8).

1. Assume there is no response and $\sigma_{\mathrm{k}}^2 = \sigma_{\mathrm{u}}^2$. Set the initial values $\widehat{\sigma}_{\mathrm{k}(1)}^2 = \widehat{\sigma}_{\mathrm{u}(1)}^2$ equal to $\sum_{j=d^*+1}^{p}\widehat{\lambda}_j\left(\widehat{\boldsymbol{\Sigma}}\right)/p$.

2. For some small $\delta > 0$, repeat for $j = 1, 2, \ldots$ until the maximum angle between $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j-1)}^*}$ and $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j)}^*}$ is less than δ,

   (a) For fixed $\widehat{\sigma}_{\mathrm{k}(j)}^2$ and $\widehat{\sigma}_{\mathrm{u}(j)}^2$, find $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j)}^*}$, which maximizes likelihood (5.8), as the span of the first d^* eigenvectors of $\dfrac{l}{\widehat{\sigma}_{\mathrm{k}(j)}^2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fitk}} + \dfrac{n-l}{\widehat{\sigma}_{\mathrm{u}(j)}^2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{u}} - \dfrac{n-l}{\widehat{\sigma}_{\mathrm{u}(j)}^2+1}\widehat{\boldsymbol{\Sigma}}_{\mathrm{u}}$.

   (b) With $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j)}^*}$, find $\widehat{\sigma}_{\mathrm{k}(j+1)}^2$ and $\widehat{\sigma}_{\mathrm{u}(j+1)}^2$ which maximize likelihood (5.8) (see Appendix A.14 for details).

Consequently, the sufficient reduction can be estimated as $\widehat{R}(\mathbf{X}) = \widehat{\boldsymbol{\Gamma}}_{(j)}^{*T}\mathbf{X}$ and the corresponding scale estimators are $\widehat{\sigma}_{\mathrm{k}(j+1)}^2$ and $\widehat{\sigma}_{\mathrm{u}(j+1)}^2$.

Diagonal or general error, $\mathbf{\Omega}^* = \text{diag}\{\sigma_1, \ldots, \sigma_p\}$ or $\mathbf{\Omega}^* > 0$. Now we assume that $\boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}$ has the covariance matrix $\mathbf{\Omega}^*$. Then the full log likelihood can be written as

$$\begin{aligned} L_{d^*}(\bar{\boldsymbol{\mu}}^*_{\text{k}}, \bar{\boldsymbol{\mu}}^*_{\text{u}}, \mathcal{S}_{\mathbf{\Gamma}^*}, \boldsymbol{\alpha}^*, \mathbf{\Omega}^*, \sigma^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{l}{2}\log|\mathbf{\Omega}^*| - \frac{n-l}{2}\log|\mathbf{\Gamma}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p| \\ & - \frac{1}{2}\sum_{i=1}^{l}(\mathbf{X}_i - \bar{\boldsymbol{\mu}}^*_{\text{k}} - \mathbf{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}^*_{y_i})^T\mathbf{\Omega}^{*-1}(\mathbf{X}_i - \bar{\boldsymbol{\mu}}^*_{\text{k}} - \mathbf{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}^*_{y_i}) \\ & - \frac{1}{2}\sum_{i=l+1}^{n}\left(\mathbf{X}_i - \bar{\boldsymbol{\mu}}^*_{\text{u}}\right)^T\left(\mathbf{\Gamma}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p\right)^{-1}\left(\mathbf{X}_i - \bar{\boldsymbol{\mu}}^*_{\text{u}}\right). \end{aligned} \tag{5.9}$$

Substituting the MLEs of $\bar{\boldsymbol{\mu}}^*_{\text{k}}$, $\bar{\boldsymbol{\mu}}^*_{\text{u}}$, and $\boldsymbol{\alpha}^*$ into the log likelihood with fixed $\mathbf{\Gamma}^*$, $\mathbf{\Omega}^*$, and σ^2 the partially maximized likelihood is

$$\begin{aligned} L_{d^*}(\mathcal{S}_{\mathbf{\Gamma}^*}, \mathbf{\Omega}^*, \sigma^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{l}{2}\log|\mathbf{\Omega}^*| - \frac{(n-l)d^*}{2}\log(\sigma^2+1) \\ & - \frac{(n-l)(p-d^*)}{2}\log(\sigma^2) - \frac{l}{2}\text{tr}\left[\mathbf{\Omega}^{*-1/2}\widehat{\mathbf{\Sigma}}_{\text{k}}\mathbf{\Omega}^{*-1/2}\right] - \frac{n-l}{2\sigma^2}\text{tr}\left[\widehat{\mathbf{\Sigma}}_{\text{u}}\right] \\ & + \frac{1}{2}\text{tr}\left[l\mathbf{\Omega}^{*-1/2}\widehat{\mathbf{\Sigma}}_{\text{fitk}}\mathbf{\Omega}^{*-1/2}\mathbf{P}_{\mathbf{\Omega}^{*-1/2}\mathbf{\Gamma}^*} + \left\{\frac{n-l}{\sigma^2}\widehat{\mathbf{\Sigma}}_{\text{u}} - \frac{n-l}{\sigma^2+1}\widehat{\mathbf{\Sigma}}_{\text{u}}\right\}\mathbf{P}_{\mathbf{\Gamma}^*}\right]. \end{aligned} \tag{5.10}$$

Here we are not able to find explicit close-form solutions for $\mathcal{S}_{\mathbf{\Gamma}^*}$, $\mathbf{\Omega}^*$, and σ^2, while $\mathcal{S}_{\mathbf{\Gamma}^*}$ is associated with different projection matrices. With the diagonal structure of $\mathbf{\Omega}^*$ or unstructured $\mathbf{\Omega}^* > 0$, finding the MLEs along with the estimation of a sufficient reduction is work in the progress.

The variance function $\mathbf{\Omega}^* = \mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$. When we assume that $\boldsymbol{\epsilon}_{\boldsymbol{\omega}^*}$ has the variance $\mathbf{\Omega}^* = \mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$, the full log likelihood is obtained by using this variance in log likelihood (5.9). For fixed $\mathbf{\Gamma}^*$ and σ^2, using the estimated parameters of $\bar{\boldsymbol{\mu}}^*_{\text{k}}$, $\bar{\boldsymbol{\mu}}^*_{\text{u}}$, $\boldsymbol{\alpha}^*$, and $\mathbf{\Phi}^*$ we have the partially maximized likelihood

$$\begin{aligned} L_{d^*}(\mathcal{S}_{\mathbf{\Gamma}^*}, \sigma^2) \;=\; & -\frac{np}{2}\log(2\pi) - \frac{l}{2}\log|\mathbf{\Gamma}^{*T}\widehat{\mathbf{\Sigma}}_{\text{resk}}\mathbf{\Gamma}^*| - \frac{(n-l)d^*}{2}\log(\sigma^2+1) - \frac{ld^*}{2} \\ & - \frac{n(p-d^*)}{2}\log(\sigma^2) - \frac{l}{2\sigma^2}\text{tr}\left[\widehat{\mathbf{\Sigma}}_{\text{k}}\right] - \frac{n-l}{2\sigma^2}\text{tr}\left[\widehat{\mathbf{\Sigma}}_{\text{u}}\right] \\ & + \frac{1}{2}\text{tr}\left[\left\{\frac{l}{\sigma^2}\widehat{\mathbf{\Sigma}}_{\text{k}} + \frac{n-l}{\sigma^2}\widehat{\mathbf{\Sigma}}_{\text{u}} - \frac{n-l}{\sigma^2+1}\widehat{\mathbf{\Sigma}}_{\text{u}}\right\}\mathbf{P}_{\mathbf{\Gamma}^*}\right], \end{aligned} \tag{5.11}$$

where $\widehat{\mathbf{\Sigma}}_{\text{resk}} = \widehat{\mathbf{\Sigma}}_{\text{k}} - \widehat{\mathbf{\Sigma}}_{\text{fitk}}$. Again we have to use an alternating maximization algorithm for the estimation of $\mathcal{S}_{\mathbf{\Gamma}^*}$ and σ^2. The new alternating algorithm with likelihood (5.11) can be developed as follows:

1. Assume there is no response and find the initial value $\widehat{\sigma}^2_{(1)}$ from the result of PC model, $\widehat{\sigma}^2_{(1)} = \sum_{j=d^*+1}^{p} \widehat{\lambda}_j \left(\widehat{\mathbf{\Sigma}}\right)/p$.

2. For some small $\delta > 0$, repeat for $j = 1, 2, \ldots$ until the maximum angle between $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j-1)}}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$ is less than δ,

   (a) For fixed $\widehat{\sigma}^2_{(j)}$, find $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$ which maximizes likelihood (5.11) using the numerical optimization computer code, Lippert's *sg_min 2.4.1*, for Grassmann optimization.

   (b) With $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.11) (see Appendix A.15 for details).

Consequently, the sufficient reduction can be estimated as

$$\widehat{R}(\mathbf{X}) = \left(\widehat{\mathbf{\Gamma}}^{*T}_{(j)}(\widehat{\mathbf{\Gamma}}^{*}_{(j)}\widehat{\mathbf{\Phi}}^{*}\widehat{\mathbf{\Gamma}}^{*T}_{(j)} + \widehat{\sigma}^2_{(j+1)}\mathbf{I}_p)^{-1}J + \widehat{\mathbf{\Gamma}}^{*T}_{(j)}(1-J)\right)\mathbf{X}$$

with the corresponding scale estimators $\widehat{\sigma}^2_{(j+1)}$.

5.2 Combining partial models

We still assume that the dataset is divided into two parts, one part with observed Y and the other with unobserved Y. In this section, we focus on reducing the dimension of $\mathbf{X}_1$, while the goal was to reduce the dimension of $\mathbf{X}$ in Section 5.1. Starting with the partial probabilistic model (2.5) with isotropic error covariance matrix, we consider reparameterizing model (3.1). When the response is present the model can be adapted to accommodate Y by modeling $\boldsymbol{\nu}'$. That is, model (4.9) can be considered for the cases with known responses, but model (3.1) is otherwise considered.

In this section we will discuss six types of combining partial models,

1. $\boldsymbol{\nu}'$ is fixed and modeled by $\boldsymbol{\alpha}'\mathbf{f}_y'$ without approximation error only when the response is observed.

2. $\boldsymbol{\nu}'$ is random and modeled by $\boldsymbol{\alpha}'\mathbf{f}_y'$ without approximation error only when the response is observed.

3. $\boldsymbol{\nu}'$ is random and when the response is observed it is modeled by $\boldsymbol{\alpha}'\mathbf{f}_y'$ with the approximation error which has

   (a) an isotropic covariance matrix.

   (b) a diagonal covariance matrix.

   (c) a general covariance matrix.

   (d) the variance function $\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$.

5.2.1 Latent variable $\boldsymbol{\nu}'$ is fixed

In the same manner as in Section 5.1.1, assume that $\boldsymbol{\nu}'$ is fixed and for the cases with known responses, modeled by $\boldsymbol{\alpha}'\mathbf{f}_y'$ without approximation error, where $\boldsymbol{\alpha}'$ and $\mathbf{f}_y'$ are as defined previously in (4.4). With the same indicator function J, a combining model is obtained as

$$\begin{aligned}\mathbf{X}_1|(\mathbf{X}_2, Y \text{ or } \boldsymbol{\nu}') &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + \boldsymbol{\Gamma}(\boldsymbol{\alpha}'\mathbf{f}_y' J + \boldsymbol{\nu}'(1-J)) + \sigma\boldsymbol{\varepsilon} \\ &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + \boldsymbol{\Gamma}(\boldsymbol{\alpha}\mathbf{f}_y J + \boldsymbol{\nu}(1-J)) + \sigma\boldsymbol{\varepsilon}, \qquad (5.12)\end{aligned}$$

where the same process as in Section 3.1 and 4.2 is used in the second equation to distinguish span($\boldsymbol{\Gamma}$) from $\boldsymbol{\beta}$. Again the bias term $\boldsymbol{\Gamma}(\boldsymbol{\nu}' - \boldsymbol{\alpha}'\mathbf{f}_y')J$ is added in model (5.12) and should be considered if there is an approximation error when modeling $\boldsymbol{\nu}'$ as $\boldsymbol{\alpha}'\mathbf{f}_y'$. The model with the bias term is not discussed, but work is in progress.

This model can be also seen as a combination of the isotropic partial probabilistic PCA model (3.1) and the isotropic partial PFC model (4.9). Under model (5.12) a sufficient reduction is $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\mathbf{X}_1$ because the isotropic partial probabilistic PCA and the isotropic partial PFC model have the same reduction. We still assume

that l of n y's are observed. Then the full log likelihood is

$$\begin{aligned} L_d(\bar{\boldsymbol{\mu}}_1, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\beta}_0, \boldsymbol{\alpha}, \boldsymbol{\nu}, \sigma^2) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{np_1}{2}\log(\sigma^2) \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{l} \|\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\|^2 \\ & - \frac{1}{2\sigma^2}\sum_{i=l+1}^{n} \|\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\nu}_i\|^2 . \end{aligned}$$

Let $\mathbb{X}_{1(\mathrm{k})}$ denote the $l \times p_1$ matrix with rows $(\mathbf{X}_1 - \bar{\mathbf{X}}_1)^T$ where the corresponding response is observed and let $\mathbb{X}_{1(\mathrm{u})}$ denote the $(n-l) \times p_1$ matrix with that where the corresponding is unobserved. In the same way $\mathbb{X}_{2(\mathrm{k})} \in \mathbb{R}^{l \times p_2}$ and $\mathbb{X}_{2(\mathrm{u})} \in \mathbb{R}^{(n-l) \times p_2}$ can be defined.

In the same way as in Section 5.1.1, the parentheses in the subscript indicate that the total mean of the variable with all n samples is subtracted from the variable. In the later section, we will use the same subscript without parenthesis to indicate that a partial mean of the variable is subtracted while using a part of samples. For example, $\mathbb{X}_{1\mathrm{k}}$ denotes the $l \times p_1$ matrix with rows $(\mathbf{X}_1 - \bar{\mathbf{X}}_{1\mathrm{k}})^T$ for the cases with known responses, where $\bar{\mathbf{X}}_{1\mathrm{k}} \in \mathbb{R}^{p_1}$ denotes the sample mean vector of $\mathbf{X}_1$ where Y is observed.

The maximum likelihood estimators of $\bar{\boldsymbol{\mu}}_1$, $\boldsymbol{\beta}_0$, $\boldsymbol{\alpha}$, and $\boldsymbol{\nu}$ are obtained from the full log likelihood with fixed $\boldsymbol{\Gamma}$ and σ^2: $\widehat{\bar{\boldsymbol{\mu}}}_1 = \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_2$, $\widehat{\boldsymbol{\beta}}_0 = \boldsymbol{\Gamma}_0^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$, $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T\mathbb{X}_{1(\mathrm{k})}^T\mathbf{F}_{\mathrm{k}}({\mathbf{F}_{\mathrm{k}}}^T\mathbf{F}_{\mathrm{k}})^{-1}$ and $\widehat{\boldsymbol{\nu}}_i = \boldsymbol{\Gamma}^T(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)$, where $\mathbf{F}_{\mathrm{k}} \in \mathbb{R}^{l \times (p_2 + r)}$.

Substituting back all these estimates and using $\mathbf{P}_{\boldsymbol{\Gamma}_0} = \mathbf{I}_{p_1} - \mathbf{P}_{\boldsymbol{\Gamma}}$, the final partially maximized log likelihood L_d is then

$$\begin{aligned} L_d(\mathcal{S}_{\boldsymbol{\Gamma}}, \sigma^2) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right] \\ & + \frac{1}{2\sigma^2}\mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}}\left\{l\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit(k)}}^{1|\mathbf{f}} + (n-l)\widehat{\boldsymbol{\Sigma}}_{1(\mathrm{u})} - n\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|2}\right\}\right], \end{aligned} \tag{5.13}$$

where $\widehat{\boldsymbol{\Sigma}}_{1(\mathrm{u})} = \mathbb{X}_{1(\mathrm{u})}^T\mathbb{X}_{1(\mathrm{u})}/(n-l)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit(k)}}^{1|\mathbf{f}} = \mathbb{X}_{1(\mathrm{k})}^T\mathbf{P}_{\mathbf{F}_{\mathrm{k}}}\mathbb{X}_{1(\mathrm{k})}/l$. With fixed σ^2 the likelihood is then maximized by setting $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ equal to the span of eigenvectors corresponding to the largest d eigenvalues $\widehat{\tau}_i$, $i = 1, \ldots, d$, of $l\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit(k)}}^{1|\mathbf{f}} + (n-l)\widehat{\boldsymbol{\Sigma}}_{1(\mathrm{u})} - n\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|2}$. This estimation makes sense in that the combining model is just the combination of isotropic partial probabilistic PCA and isotropic partial PFC models since $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ was estimated by

using the eigenvectors of $\widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}$ in the isotropic partial probabilistic PCA model and that of $\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}$ in the isotropic partial PFC model respectively. The corresponding scale parameter σ^2 is estimated as $\widehat{\sigma}^2 = \sum_{i=1}^{p_1} \lambda_i \left(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2} \right) / p_1 - \sum_{i=1}^{d} \widehat{\tau}_i / (np_1)$.

Again we did the same simulation as in Section 5.1.1 while considering now isotropic partial probabilistic PCA(IPPCA), isotropic partial PFC(IPPFC), and combining models. With $n = 1000,\ d = 1,\ r = 1,\ p_1 = 200,\ p_2 = 10$ and $p = p_1 + p_2$, observations on $\mathbf{X}$ were generated as $\mathbf{X}_i = \mathbf{\Gamma}^* \boldsymbol{\alpha}^* \mathbf{f}_{y_i}^* + \boldsymbol{\epsilon}$, where $\mathbf{\Gamma}^* = \mathbf{1}_p / \sqrt{p}$, $\boldsymbol{\alpha}^* = 1$, $\mathbf{f}_y^* = y^2$ with $y \sim N(0,1)$, and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ is a $N(\mathbf{0}, \mathbf{I}_p)$ vector. $\mathbf{X}_1$ and $\mathbf{X}_2$ were obtained by dividing $\mathbf{X} \in \mathbb{R}^p$ into two vectors, $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ with the first p_1 elements and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$ with the remaining p_2 elements. Since the sufficient reduction is $R(\mathbf{X}_1) = (\mathbf{\Gamma}_0 \boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T \mathbf{X}_1$ our interest is still in estimating $\mathcal{S}_{\mathbf{\Gamma}}$. With this generated dataset and $\mathbf{f}_y^* = (y, |y|)^T$ used for fitting, we estimated $\mathcal{S}_{\mathbf{\Gamma}}$ by setting it equal to the span of the first d eigenvectors of $\widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}$ in the isotropic partial probabilistic PCA model, $\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}$ in the isotropic partial PFC model, and $l\widehat{\mathbf{\Sigma}}_{\text{fit(k)}}^{1|\mathbf{f}} + (n-l)\widehat{\mathbf{\Sigma}}_{1(\text{u})} - n\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}$ in the partial combining model (5.2.1).

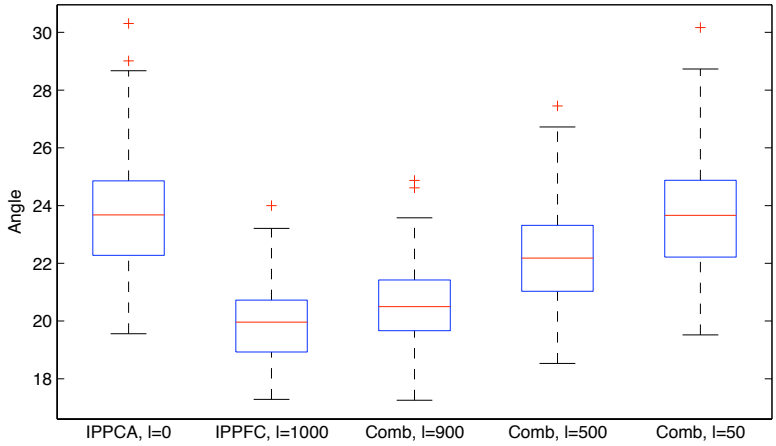


Figure 5.3: Simulation results showing the angle in degrees for $\mathcal{S}_{\mathbf{\Gamma}}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ in each model with the various number of known responses l out of $n = 1000$.

Figure 5.3 shows boxplots of the angle in degrees between $\mathcal{S}_{\mathbf{\Gamma}}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ varying the

number of the known responses, $l = 900,\ 500$, and 50. Here the true $\boldsymbol{\Gamma} \in \mathbb{R}^{p_1}$ is an orthogonalized vecor of the first p_1 elements of $\boldsymbol{\Gamma}^*$. We can see the differences in the performance according to the number of the known responses in Figure 5.3. While we do not have much improvement in the partial combining model with only few known responses, $l = 50$, showing the similar result to IPPCA, the performance in the partial combining model shows improvement with $l = 900$ taking after the boxplot of IPPFC. Consequently, we should be able to accommodate the response information as much as possible to have the better estimation.

Since we assumed that a dataset is divided into two parts, one part with observed Y and the other with unobserved Y, the same simulation as in Figure 5.2 can be performed. Let IPPCA$(n-l)$ indicate that the isotropic partial probabilistic model is used for just $(n-l)$ samples where a response is not observed. Similarly IPPFC(l) indicates that we have only l samples with known responses and the isotropic partial PFC model is used. In the same way we can define the methods used in Figure 5.3 as IPPCA(n) and IPPFC(n). The methods used in the simulation are listed in Table 5.2 with how to estimate $\mathcal{S}_{\boldsymbol{\Gamma}}$ under the corresponding method.

Table 5.2: Partial dimension reduction methods used in the simulation

Method	Sample size	Number of known y	Number of unknown y	$\mathcal{S}_{\boldsymbol{\Gamma}}$ is estimated by the first d eigenvalues of
IPPCA(n)	n	0	n	$\widehat{\boldsymbol{\Sigma}}^{1\|2}_{\text{res}} = \widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}}^{1\|2}_{\text{fit}}$
IPPFC(n)	n	n	0	$\widehat{\boldsymbol{\Sigma}}^{1\|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1\|2}_{\text{fit}}$
Comb.	n	l	$n-l$	$l\widehat{\boldsymbol{\Sigma}}^{1\|\mathbf{f}}_{\text{fit(k)}} + (n-l)\widehat{\boldsymbol{\Sigma}}_{1(\text{u})} - n\widehat{\boldsymbol{\Sigma}}^{1\|2}_{\text{fit}}$
IPPCA$(n-l)$	$n-l$	0	$n-l$	$\widehat{\boldsymbol{\Sigma}}^{1\|2}_{\text{res(u)}} = \widehat{\boldsymbol{\Sigma}}_{1(\text{u})} - \widehat{\boldsymbol{\Sigma}}^{1\|2}_{\text{fit(u)}}$
IPPFC(l)	l	l	0	$\widehat{\boldsymbol{\Sigma}}^{1\|\mathbf{f}}_{\text{fit(k)}} - \widehat{\boldsymbol{\Sigma}}^{1\|2}_{\text{fit(k)}}$

Data were generated by the same procedure as in Figure 5.2 with the signal sizes, weak, moderate, and strong. Figure 5.4 shows the influence of the number of known responses and size of signals on the estimation of $\mathcal{S}_{\boldsymbol{\Gamma}}$.

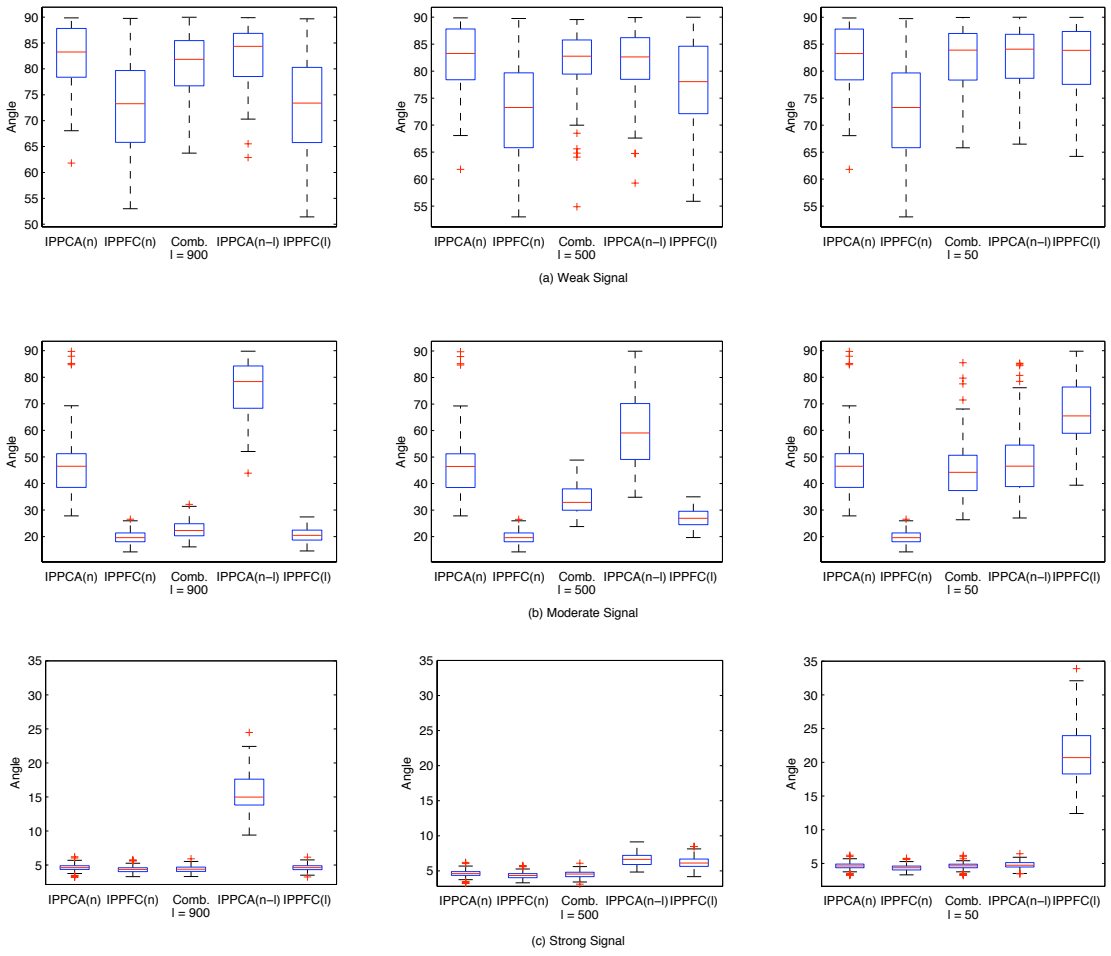


Figure 5.4: Simulation results showing the angles in degrees between $\mathcal{S}_{\mathbf{\Gamma}}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ in each model with the various number of known responses l out of $n = 1000$, $l = \{900,\ 500,\ 50\}$ and three different signals, Weak, Moderate, Strong signals.

Since Figure 5.4 shows results that are very similar to Figure 5.2 we can refer to the interpretation about the boxplots from Figure 5.2. In short, when the number of response is large, IPPFC(l) is better than combining or IPPCA($n-l$) particularly with a weak signal. When the number of responses is small, IPPFC(l) shows poor performance. With a weak signal IPPFC(l) might be a better choice than combining or IPPCA($n-l$), especially when the number of responses is large. With a strong signal the combining method guarantees good performance regardless of the number of known responses.

5.2.2 Latent variable $\boldsymbol{\nu}'$ is random and modeled without approximation error

In the same manner as in Section 5.1.2, assume that $\boldsymbol{\nu}'$ is random and for the cases with known responses, modeled by $\boldsymbol{\alpha}'\mathbf{f}_y'$ without approximation error. Then model (2.5) with isotropic error is rewritten as

$$\begin{aligned}
\mathbf{X}_1|(\mathbf{X}_2, Y \text{ or } \boldsymbol{\nu}') &= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + (\boldsymbol{\Gamma}E(\boldsymbol{\nu}'|y) + \boldsymbol{\Gamma}(\boldsymbol{\nu}' - E(\boldsymbol{\nu}'|y)) + \sigma\boldsymbol{\varepsilon})J + (\boldsymbol{\Gamma}\boldsymbol{\nu}' + \sigma\boldsymbol{\varepsilon})(1-J) \\
&= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\beta}\mathbf{X}_2 + (\boldsymbol{\Gamma}\boldsymbol{\alpha}'\mathbf{f}_y' + \boldsymbol{\Gamma}\boldsymbol{\omega} + \sigma\boldsymbol{\varepsilon})J + (\boldsymbol{\Gamma}\boldsymbol{\nu}' + \sigma\boldsymbol{\varepsilon})(1-J) \\
&= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + (\boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}})J + (\boldsymbol{\Gamma}\boldsymbol{\nu} + \sigma\boldsymbol{\varepsilon})(1-J) \\
&= \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + (\boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}})J + \boldsymbol{\varepsilon}_{\mathrm{u}}(1-J), \qquad (5.14)
\end{aligned}$$

where $E(\boldsymbol{\nu}'|y)$ is written as $\boldsymbol{\alpha}'\mathbf{f}_y'$ in the second equation and $\boldsymbol{\omega}$ and $\boldsymbol{\varepsilon}_{\boldsymbol{\omega}}$ are as defined previously in Section 4.1.2. The third equation is derived by following the same process as in Sections 3.1 and 4.2 to distinguish span($\boldsymbol{\Gamma}$) from $\boldsymbol{\beta}$. In the last equation, the new normal error $\boldsymbol{\varepsilon}_{\mathrm{u}}$ has mean 0 and variance $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$ because $\boldsymbol{\nu}$ is assumed to be normally distributed with mean 0 and identity covariance matrix. In fact $\boldsymbol{\omega}$ which represents the approximation error in the second equation is equal to zero because we assume that there is no approximation error. Then model (5.14) becomes

$$\mathbf{X}_1|Y = \bar{\boldsymbol{\mu}}_1 + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + (\boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon})J + \boldsymbol{\varepsilon}_{\mathrm{u}}(1-J). \qquad (5.15)$$

Here the two different covariance matrices according to the value of J cause considerable difficulty in the estimation of a sufficient reduction while yielding a complicated form

for $\bar{\boldsymbol{\mu}}_1$. So we again allow for two different $\bar{\boldsymbol{\mu}}_1$'s: $\bar{\boldsymbol{\mu}}_{1\mathrm{k}}$ for the cases with known responses and $\bar{\boldsymbol{\mu}}_{1\mathrm{u}}$ for the case without a known response. Then a combining model is obtained as

$$\mathbf{X}_1|Y = \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + (\bar{\boldsymbol{\mu}}_{1\mathrm{k}} + \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon})J + (\bar{\boldsymbol{\mu}}_{1\mathrm{u}} + \boldsymbol{\varepsilon}_{\mathrm{u}})(1-J). \tag{5.16}$$

Under model (5.16), $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\mathbf{X}_1$ is again a sufficient reduction, so we are still interested in estimating the subspace $\mathcal{S}_{\boldsymbol{\Gamma}}$. Let $\bar{\mathbf{X}}_{1\mathrm{k}} \in \mathbb{R}^{p_1}$ denote the sample mean vector of $\mathbf{X}_1$ where Y is observed and similarly let $\bar{\mathbf{X}}_{1\mathrm{u}} \in \mathbb{R}^{p_1}$ denote that where Y is unobserved. Let $\mathbb{X}_{1\mathrm{k}}$ denote the $l \times p_1$ matrix with rows $(\mathbf{X}_1 - \bar{\mathbf{X}}_{1\mathrm{k}})^T$ and $\mathbb{X}_{1\mathrm{u}}$ denote the $(n-l) \times p_1$ matrix with with rows $(\mathbf{X}_1 - \bar{\mathbf{X}}_{1\mathrm{u}})^T$. In the same way $\bar{\mathbf{X}}_{2\mathrm{k}} \in \mathbb{R}^{p_2}$, $\bar{\mathbf{X}}_{2\mathrm{u}} \in \mathbb{R}^{p_2}$, $\mathbb{X}_{2\mathrm{k}} \in \mathbb{R}^{l\times p_2}$, and $\mathbb{X}_{2\mathrm{u}} \in \mathbb{R}^{(n-l)\times p_2}$ can be defined. Maximizing the log likelihood over $\bar{\boldsymbol{\mu}}_{1\mathrm{k}}$, $\bar{\boldsymbol{\mu}}_{1\mathrm{u}}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}_0$, MLEs are obtained as $\widehat{\bar{\boldsymbol{\mu}}}_{1\mathrm{k}} = \bar{\mathbf{X}}_{1\mathrm{k}} - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_{2\mathrm{k}}$, $\widehat{\bar{\boldsymbol{\mu}}}_{1\mathrm{u}} = \bar{\mathbf{X}}_{1\mathrm{u}} - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_{2\mathrm{u}}$, $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T\mathbb{X}_{1\mathrm{k}}^T\mathbf{F}_{\mathrm{k}}(\mathbf{F}_{\mathrm{k}}^T\mathbf{F}_{\mathrm{k}})^{-1}$, and $\widehat{\boldsymbol{\beta}}_0 = \boldsymbol{\Gamma}_0^T\mathbf{B}$ where $\mathbf{B} = (\mathbb{X}_{1\mathrm{k}}^T\mathbb{X}_{2\mathrm{k}} + \mathbb{X}_{1\mathrm{u}}^T\mathbb{X}_{2\mathrm{u}})(\mathbb{X}_{2\mathrm{k}}^T\mathbb{X}_{2\mathrm{k}} + \mathbb{X}_{2\mathrm{u}}^T\mathbb{X}_{2\mathrm{u}})^{-1}$, and we find the partially maximized log likelihood,

$$\begin{aligned} L_d(\mathcal{S}_{\boldsymbol{\Gamma}}, \sigma^2) \ =\ & -\frac{np_1}{2}\log(2\pi) + \frac{(n-l)d - np_1}{2}\log(\sigma^2) - \frac{(n-l)d}{2}\log(\sigma^2+1) \\ & - \frac{1}{2\sigma^2}\Big\{l\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{1\mathrm{k}}\right] - \mathrm{tr}\left[\mathbf{A}_{\mathrm{k}}\right] + (n-l)\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}}\right] - \mathrm{tr}\left[\mathbf{A}_{\mathrm{u}}\right]\Big\} \\ & + \frac{1}{2}\mathrm{tr}\Big[\Big\{\frac{1}{\sigma^2}(l\widehat{\boldsymbol{\Sigma}}_{\mathrm{fitk}}^{1|\mathbf{f}} - \mathbf{A}_{\mathrm{k}} - \mathbf{A}_{\mathrm{u}} + (n-l)\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}}) - \frac{n-l}{(\sigma^2+1)}\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}}\Big\}\mathbf{P}_{\boldsymbol{\Gamma}}\Big], \end{aligned} \tag{5.17}$$

where $\mathbf{A}_{\mathrm{k}} = 2\mathbf{B}\mathbb{X}_{2\mathrm{k}}^T\mathbb{X}_{1\mathrm{k}} - \mathbf{B}\mathbb{X}_{2\mathrm{k}}^T\mathbb{X}_{2\mathrm{k}}\mathbf{B}^T$, $\mathbf{A}_{\mathrm{u}} = 2\mathbb{X}_{1\mathrm{u}}^T\mathbb{X}_{2\mathrm{u}}\mathbf{B}^T - \mathbf{B}\mathbb{X}_{2\mathrm{u}}^T\mathbb{X}_{2\mathrm{u}}\mathbf{B}^T$, $\widehat{\boldsymbol{\Sigma}}_{1\mathrm{k}} = \mathbb{X}_{1\mathrm{k}}^T\mathbb{X}_{1\mathrm{k}}/l$, $\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}} = \mathbb{X}_{1\mathrm{u}}^T\mathbb{X}_{1\mathrm{u}}/(n-l)$, and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fitk}}^{1|\mathbf{f}} = \mathbb{X}_{1\mathrm{k}}^T\mathbf{P}_{\mathbf{F}_{\mathrm{k}}}\mathbb{X}_{1\mathrm{k}}/l$. Like the likelihoods in Section 5.1.2, we are not able to find the closed-form solution for $\mathcal{S}_{\boldsymbol{\Gamma}}$ and σ^2 and the estimators of them should be obtained from the following alternating maximization algorithm:

1. Assume there is no response and find the initial value $\widehat{\sigma}_{(1)}^2$ from the result of isotropic partial probabilistic PCA model (3.1), $\widehat{\sigma}_{(1)}^2 = \sum_{i=d+1}^{p_1}\widehat{\lambda}_i\left(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right)/p_1$.

2. For some small $\delta > 0$, repeat for $j = 1, 2, \ldots$ until the maximum angle between $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j-1)}}$ and $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j)}}$ is less than δ,

(a) For fixed $\widehat{\sigma}^2_{(j)}$, find $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$, which maximizes likelihood (5.17), as the span of the first d eigenvectors of $\frac{1}{\widehat{\sigma}^2_{(j)}}(l\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fitk}} - \mathbf{A}_{\text{k}} - \mathbf{A}_{\text{u}} + (n-l)\widehat{\mathbf{\Sigma}}_{1\text{u}}) - \frac{n-l}{(\widehat{\sigma}^2_{(j)}+1)}\widehat{\mathbf{\Sigma}}_{1\text{u}}$.

(b) With $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.17) (see Appendix A.16 for details).

Like algorithms used in Section 5.1.2 and 5.1.3, our main goal of this algorithm is to estimate $\mathcal{S}_{\mathbf{\Gamma}}$ rather than σ^2. Hence the condition to stop the algorithm depends on the convergence of $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ which achieves maximization of (5.17). Consequently, the sufficient reduction can be estimated as $\widehat{R}(\mathbf{X}_1) = (\widehat{\mathbf{\Gamma}}_{0(j)}\widehat{\boldsymbol{\beta}}_0,\ \widehat{\mathbf{\Gamma}}_{(j)})^T\mathbf{X}_1$ and the corresponding scale estimator is $\widehat{\sigma}^2_{(j+1)}$.

5.2.3 Latent variable $\boldsymbol{\nu}'$ is random and modeled with approximation error

In the same manner as in Section 5.1.3 we assume that $\boldsymbol{\nu}'$ is random and for the cases with known responses, there is approximation error when modeling $\boldsymbol{\nu}'$ as $\boldsymbol{\alpha}'\mathbf{f}'_y$. Then model (5.16) is reformulated as

$$\mathbf{X}_1|Y = \mathbf{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 + (\bar{\boldsymbol{\mu}}_{1\text{k}} + \mathbf{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}})J + (\bar{\boldsymbol{\mu}}_{1\text{u}} + \boldsymbol{\varepsilon}_{\text{u}})(1-J), \tag{5.18}$$

where $\boldsymbol{\varepsilon}_{\boldsymbol{\omega}}$ is as defined in Section 4.2, and is normally distributed with mean 0 and variance $\mathbf{\Omega}$. Again we can consider four types of the error covariance structure when $J = 1$: model with (1) isotropic error, $\mathbf{\Omega} = \sigma^2_{\text{k}}\mathbf{I}_{p_1}$, (2) diagonal error, $\mathbf{\Omega} = \text{diag}\{\sigma_1, \ldots, \sigma_{p_1}\}$, (3) unstructured error, $\mathbf{\Omega} > 0$, and (4) the variance $\mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$.

Isotropic error, $\mathbf{\Omega} = \sigma^2_{\text{k}}\mathbf{I}_{p_1}$. Assume that $\boldsymbol{\varepsilon}_{\boldsymbol{\omega}}$ has the isotropic covariance matrix $\sigma^2_{\text{k}}\mathbf{I}_{p_1}$ and $\sigma^2_{\text{k}} \neq \sigma^2$, while the result for $\sigma^2_{\text{k}} = \sigma^2$ is the same as in Section 5.2.2. In order to emphasize that σ^2 is only related to the cases with unknown response, we will write σ^2 as σ^2_{u}. However, we are unable to find closed-form solutions for MLEs of parameters of interest with the same $\boldsymbol{\beta}_0$ for J and $(1-J)$. So we assume that there are two different $\boldsymbol{\beta}_0$'s: $\boldsymbol{\beta}_{0\text{k}}$ for the cases with known responses and $\boldsymbol{\beta}_{0\text{u}}$ for the cases without a known

response. Then the combining model can be written as

$$\mathbf{X}_1|Y = (\bar{\boldsymbol{\mu}}_{1\mathrm{k}} + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_{0\mathrm{k}}\mathbf{X}_2 + \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y + \boldsymbol{\varepsilon}_{\boldsymbol{\omega}})J + (\bar{\boldsymbol{\mu}}_{1\mathrm{u}} + \boldsymbol{\Gamma}_0\boldsymbol{\beta}_{0\mathrm{u}}\mathbf{X}_2 + \boldsymbol{\varepsilon}_{\mathrm{u}})(1-J). \quad (5.19)$$

Under this model (5.19) with $\boldsymbol{\varepsilon}_{\boldsymbol{\omega}} \sim N(0, \sigma^2_{\mathrm{k}}\mathbf{I}_{p_1})$, we have the full log likelihood

$$\begin{aligned} L_d \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{lp_1}{2}\log(\sigma^2_{\mathrm{k}}) - \frac{n-l}{2}\log|\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \sigma^2_{\mathrm{u}}\mathbf{I}_{p_1}| \\ & - \frac{1}{2\sigma^2_{\mathrm{k}}}\sum_{i=1}^{l}\|\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_{1\mathrm{k}} - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_{0\mathrm{k}}\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\|^2 \\ & - \frac{1}{2}\sum_{i=l+1}^{n}(\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_{1\mathrm{u}} - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_{0\mathrm{u}}\mathbf{X}_{2i})^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \sigma^2_{\mathrm{u}}\mathbf{I}_{p_1})^{-1}(\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_{1\mathrm{u}} - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_{0\mathrm{u}}\mathbf{X}_{2i}). \end{aligned}$$

Maximizing over $\bar{\boldsymbol{\mu}}_{1\mathrm{k}}$, $\boldsymbol{\beta}_{0\mathrm{k}}$, $\boldsymbol{\alpha}$, $\bar{\boldsymbol{\mu}}_{1\mathrm{u}}$, and $\boldsymbol{\beta}_{0\mathrm{u}}$, the partially maximized log likelihood is

$$\begin{aligned} L_d(\mathcal{S}_{\boldsymbol{\Gamma}}, \sigma^2_{\mathrm{k}}, \sigma^2_{\mathrm{u}}) \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{lp_1}{2}\log(\sigma^2_{\mathrm{k}}) - \frac{(n-l)d}{2}\log(\sigma^2_{\mathrm{u}}+1) \\ & - \frac{(n-l)(p_1-d)}{2}\log(\sigma^2_{\mathrm{u}}) - \frac{l}{2\sigma^2_{\mathrm{k}}}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resk}}\right] - \frac{n-l}{2\sigma^2_{\mathrm{u}}}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resu}}\right] \\ & - \frac{1}{2}\mathrm{tr}\Big[\Big\{\frac{l}{\sigma^2_{\mathrm{k}}}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitk}} - \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fitk}}) + \frac{n-l}{(\sigma^2_{\mathrm{u}}+1)}\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}} - \frac{n-l}{\sigma^2_{\mathrm{u}}}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resu}}\Big\}\mathbf{P}_{\boldsymbol{\Gamma}}\Big], \quad (5.20) \end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resk}} = \widehat{\boldsymbol{\Sigma}}_{1\mathrm{k}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitk}}$ and $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resu}} = \widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitu}}$ with $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitk}} = \mathbb{X}^T_{1\mathrm{k}}\mathbf{P}_{\mathbb{X}_{2\mathrm{k}}}\mathbb{X}_{1\mathrm{k}}/l$ and $\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitu}} = \mathbb{X}^T_{1\mathrm{u}}\mathbf{P}_{\mathbb{X}_{2\mathrm{u}}}\mathbb{X}_{1\mathrm{u}}/(n-l)$. With the same reasoning as in the previous sections, $\mathcal{S}_{\boldsymbol{\Gamma}}$, σ^2_{k}, and σ^2_{u} can be obtained from the following alternating maximization algorithm.

1. Assume there is no response and $\sigma^2_{\mathrm{k}} = \sigma^2_{\mathrm{u}}$. Set the initial values $\widehat{\sigma}^2_{\mathrm{k}(1)} = \widehat{\sigma}^2_{\mathrm{u}(1)}$ equal to $\sum_{i=d+1}^{p_1}\widehat{\lambda}_i\left(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}\right)/p_1$.

2. For some small $\delta > 0$, repeat for $j = 1, 2, \ldots$ until the maximum angle between $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j-1)}}$ and $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j)}}$ is less than δ,

   (a) For fixed $\widehat{\sigma}^2_{\mathrm{k}(j)}$ and $\widehat{\sigma}^2_{\mathrm{u}(j)}$, find $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j)}}$, which maximizes likelihood (5.20), as the span of the first d eigenvectors of

$$\frac{l}{\widehat{\sigma}^2_{\mathrm{k}(j)}}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitk}} - \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fitk}}) + \frac{n-l}{(\widehat{\sigma}^2_{\mathrm{u}(j)}+1)}\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}} - \frac{n-l}{\widehat{\sigma}^2_{\mathrm{u}(j)}}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resu}}.$$

(b) With $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$, find $\widehat{\sigma}^2_{\mathrm{k}(j+1)}$ and $\widehat{\sigma}^2_{\mathrm{u}(j+1)}$ which maximize likelihood (5.20) (see Appendix A.17 for details).

Consequently, the sufficient reduction can be estimated as

$$\widehat{R}(\mathbf{X}_1) = \left(\widehat{\mathbf{\Gamma}}_{0(j)}(\widehat{\boldsymbol{\beta}}_{0\mathrm{k}}J + \widehat{\boldsymbol{\beta}}_{0\mathrm{u}}(1-J)),\ \widehat{\mathbf{\Gamma}}_{(j)}\right)^T \mathbf{X}_1$$

and the corresponding scale estimators are $\widehat{\sigma}^2_{\mathrm{k}(j+1)}$ and $\widehat{\sigma}^2_{\mathrm{u}(j+1)}$.

Diagonal or general error, $\mathbf{\Omega} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_{p_1}\}$ or $\mathbf{\Omega} > 0$. Assuming that $\boldsymbol{\varepsilon}_{\boldsymbol{\omega}}$ has the covariance matrix $\mathbf{\Omega}$ in model (5.19), the partially maximized likelihood is

$$\begin{aligned} L_d(\mathcal{S}_{\mathbf{\Gamma}}, \mathbf{\Omega}, \sigma^2) \ = \ & -\frac{np_1}{2}\log(2\pi) - \frac{l}{2}\log|\mathbf{\Omega}| - \frac{(n-l)d}{2}\log(\sigma^2+1) - \frac{(n-l)(p_1-d)}{2}\log(\sigma^2) \\ & - \frac{l}{2}\mathrm{tr}\left[\mathbf{\Omega}^{-1/2}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{resk}}\mathbf{\Omega}^{-1/2}\right] - \frac{n-l}{2\sigma^2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{resu}}\right] \\ & + \frac{1}{2}\mathrm{tr}\left[l\mathbf{P}_{\mathbf{\Omega}^{-1/2}\mathbf{\Gamma}}\mathbf{\Omega}^{-1/2}(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fitk}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{fitk}})\mathbf{\Omega}^{-1/2} + \left\{\frac{n-l}{\sigma^2}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{resu}} - \frac{n-l}{\sigma^2+1}\widehat{\mathbf{\Sigma}}_{1\mathrm{u}}\right\}\mathbf{P}_{\mathbf{\Gamma}}\right]. \end{aligned}$$

Again we are not able to estimate $\mathcal{S}_{\mathbf{\Gamma}}$, $\mathbf{\Omega}$, and σ^2 separately with different projection matrices with respect to $\mathbf{\Gamma}$. Finding MLEs of parameters of interest for a sufficient reduction with the diagonal structure of $\mathbf{\Omega}$ and unstructured $\mathbf{\Omega} > 0$ is under study.

The variance function $\mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$. Assuming that $\boldsymbol{\varepsilon}_{\boldsymbol{\omega}}$ has variance $\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$ in model (5.19) and substituting the estimated parameters into the log likelihood with fixed $\mathbf{\Gamma}$ and σ^2, we have the partially maximized likelihood

$$\begin{aligned} L_d(\mathcal{S}_{\mathbf{\Gamma}}, \sigma^2) \ = \ & -\frac{np_1}{2}\log(2\pi) - \frac{l}{2}\log|\mathbf{\Gamma}^T\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\mathrm{resk}}\mathbf{\Gamma}| - \frac{(n-l)d}{2}\log(\sigma^2+1) - \frac{ld}{2} \\ & - \frac{n(p_1-d)}{2}\log(\sigma^2) - \frac{l}{2\sigma^2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{resk}}\right] - \frac{n-l}{2\sigma^2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{resu}}\right] \\ & + \frac{1}{2}\mathrm{tr}\left[\left\{\frac{1}{\sigma^2}(l\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{resk}} + (n-l)\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{resu}}) - \frac{n-l}{\sigma^2+1}\widehat{\mathbf{\Sigma}}_{1\mathrm{u}}\right\}\mathbf{P}_{\mathbf{\Gamma}}\right], \end{aligned} \tag{5.21}$$

where $\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\mathrm{resk}} = \widehat{\mathbf{\Sigma}}_{1\mathrm{k}} - \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fitk}}$. Again we have to use an alternating maximization algorithm for $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ and $\widehat{\sigma}^2$, which is

1. Assume there is no response and find the initial value $\widehat{\sigma}^2_{(1)}$ from the isotropic partial probabilistic PCA model (3.1), $\widehat{\sigma}^2_{(1)} = \sum_{i=d+1}^{p_1}\widehat{\lambda}_i\left(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res}}\right)/p_1$.

2. For some small $\delta > 0$, repeat for $j = 1, 2, \ldots$ until the maximum angle between $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j-1)}}$ and $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$ is less than δ,

   (a) For fixed $\widehat{\sigma}^2_{(j)}$, find $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$ which maximizes likelihood (5.21) using Lippert's *sg_min 2.4.1*, for Grassmann optimization.

   (b) With $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.21) (see Appendix A.18 for details).

Consequently, the sufficient reduction can be estimated as

$$\widehat{R}(\mathbf{X}_1) = \left(\widehat{\mathbf{\Gamma}}_{0(j)}(\widehat{\boldsymbol{\beta}}_{0\mathrm{k}}J + \widehat{\boldsymbol{\beta}}_{0\mathrm{u}}(1-J)),\ \widehat{\mathbf{\Gamma}}_{(j)}\right)^T \left((\widehat{\mathbf{\Gamma}}_{(j)}\widehat{\mathbf{\Phi}}_{(j)}\widehat{\mathbf{\Gamma}}^T_{(j)} + \widehat{\sigma}^2_{(j+1)}\mathbf{I}_{p_1})J + \mathbf{I}_{p_1}(1-J)\right)^{-1}\mathbf{X}_1$$

with the corresponding scale estimators $\widehat{\sigma}^2_{(j+1)}$.

Chapter 6

Considerations for implementation

The data we are dealing with consists of two distinguishable sets of predictors, $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$, $p = p_1 + p_2$. With the concept of partial dimension reduction on a high dimensional predictor space, we assume that the number of observations n is less than p_1, $n < p_1$ or much greater than p_1, $n \gg p_1$. We suppose that within this large pool of p_1 predictors, there is a relatively small set of relevant predictors and a large number of irrelevant ones. We hope to eventually find methods/formulations which can get rid of irrelevant predictors in $\mathbf{X}_1$ while retaining $\mathbf{X}_2$ as it is.

Screening by principal fitted components (SPFC) has been proposed (Adragni 2008), which is a versatile method that is more flexible and has better performance on screening than available leading methods for screening such as sure independence screening (SIS) of Fan and Lv (2008). In this chapter we introduce the method of screening by partial principal fitted Components (SPPFC) developing SPFC in the context of partial inverse regression.

The dimension d of the sufficient reduction was so far assumed to be known. There are many methods to choose d in practice. We will revisit those methods such as AIC, BIC, and likelihood ratio testing and compare results when applying to our data structure.

6.1 Screening by Partial Principal Fitted Components

6.1.1 Extraction scheme of active predictors

As illustrated by Adragni (2008), when dealing with regression problems with high-dimensional data, there may be a possibility that a substantial subset $\mathbf{X}_{12}$ of the predictor set $\mathbf{X}_1$ is inactive. Thus, we can develop tests for hypotheses of the form

$$Y \perp\!\!\!\perp \mathbf{X}_{12} | (\mathbf{X}_{11},\ \mathbf{X}_2), \tag{6.1}$$

where the predictor vector $\mathbf{X}_1$ is partitioned as $\mathbf{X}_1 = (\mathbf{X}_{11}^T,\ \mathbf{X}_{12}^T)^T$ with $\mathbf{X}_{11} \in \mathbb{R}^{p_{11}}$ and $\mathbf{X}_{12} \in \mathbb{R}^{p_{12}}$, $p_1 = p_{11} + p_{12}$. Under this hypothesis, $\mathbf{X}_{12}$ preserves no information once $\mathbf{X}_{11}$ and $\mathbf{X}_2$ are known. The following lemma helps us understand the structure in the partial PFC model (4.8). In order to conform to the partitioning of $\mathbf{X}_1$, partition $\boldsymbol{\Gamma}_0,\ \boldsymbol{\Gamma}$, and $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Gamma}_0 = \begin{pmatrix} \boldsymbol{\Gamma}_{01} \\ \boldsymbol{\Gamma}_{02} \end{pmatrix};\ \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_1 \\ \boldsymbol{\Gamma}_2 \end{pmatrix};\ \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix};\ \boldsymbol{\Omega}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}^{11} & \boldsymbol{\Omega}^{12} \\ \boldsymbol{\Omega}^{21} & \boldsymbol{\Omega}^{22} \end{pmatrix}. \tag{6.2}$$

Here $\boldsymbol{\Gamma}_{01} \in \mathbb{R}^{p_{11} \times (p_1 - d)}$, $\boldsymbol{\Gamma}_{02} \in \mathbb{R}^{p_{12} \times (p_1 - d)}$, $\boldsymbol{\Gamma}_1 \in \mathbb{R}^{p_{11} \times d}$, and $\boldsymbol{\Gamma}_2 \in \mathbb{R}^{p_{12} \times d}$. Let $\boldsymbol{\Omega}^{-ij} = (\boldsymbol{\Omega}^{ij})^{-1}$. Similarly, partition $\widehat{\boldsymbol{\Sigma}}_1 = (\widehat{\boldsymbol{\Sigma}}_{1,ij})$, $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2} = (\widehat{\boldsymbol{\Sigma}}_{\text{fit},ij}^{1|2})$, $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} = (\widehat{\boldsymbol{\Sigma}}_{\text{fit},ij}^{1|\mathbf{f}})$, and $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2} = (\widehat{\boldsymbol{\Sigma}}_{\text{res},ij}^{1|2})$. For a square partitioned matrix $\mathbf{A} = (\mathbf{A}_{ij})$, $i, j = 1, 2$, let $\mathbf{A}_{ii,j} = \mathbf{A}_{ii} - \mathbf{A}_{ij}\mathbf{A}_{jj}^{-1}\mathbf{A}_{ji}$.

The minimal sufficient reduction under the partial probabilistic PFC model is $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1$. Based on the partition (6.2), the reduction can be written as

$$\begin{aligned}(\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1 = &\left((\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_1)^T\boldsymbol{\Omega}^{11} + (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2)^T\boldsymbol{\Omega}^{21}\right)\mathbf{X}_{11} \\ &+ \left((\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_1)^T\boldsymbol{\Omega}^{12} + (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2)^T\boldsymbol{\Omega}^{22}\right)\mathbf{X}_{12}.\end{aligned} \tag{6.3}$$

Using this expression we can obtain the following lemma.

Lemma 6.1. *Assume model* (4.8). *Then* $Y \perp\!\!\!\perp \mathbf{X}_{12} | (\mathbf{X}_{11},\ \mathbf{X}_2)$ *if and only if* $\boldsymbol{\Gamma}_{02} = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_{01}$ *and* $\boldsymbol{\Gamma}_2 = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_1$.

The log likelihood for the alternative of dependence is as given in Theorem 4.1. The following theorem gives the maximum likelihood estimators under the hypothesis $Y \perp\!\!\!\perp \mathbf{X}_{12} | (\mathbf{X}_{11},\ \mathbf{X}_2)$ and unstructured $\boldsymbol{\Omega} > 0$. Its proof is given in Appendix A.20.

Theorem 6.1. *Assume that* $\boldsymbol{\Gamma}_{02} = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_{01}$ *and* $\boldsymbol{\Gamma}_2 = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_1$. *Then, the MLE of* $\boldsymbol{\Omega}$ *is given in blocks by* $\widehat{\boldsymbol{\Omega}}_{11} = (\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_d + \widehat{\mathbf{K}})\widehat{\mathbf{V}}^T(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{1/2}$, *with* $\widehat{\mathbf{K}} = \text{diag}(\widehat{\lambda}^1_1, \ldots, \widehat{\lambda}^1_d, 0, \ldots, 0)$ *and* $\widehat{\mathbf{V}}$ *and* $\widehat{\lambda}^1_1, \ldots, \widehat{\lambda}^1_{p_{11}}$ *the ordered eigenvectors and eigenvalues of* $(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1/2}\mathbf{B}_{\text{fit},11}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1/2}$, *where* $\mathbf{B}_{\text{fit},11} = \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit},11} - \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit},11}$; $\widehat{\boldsymbol{\Omega}}_{12} = \widehat{\boldsymbol{\Omega}}_{11}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},12}$ *and* $\widehat{\boldsymbol{\Omega}}_{22} = \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},22.1} + \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},21}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1}\widehat{\boldsymbol{\Omega}}_{11}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},12}$. *The MLE of the* $\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}$ *is the span of* $((\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1/2}\widehat{\boldsymbol{\Gamma}}_1,\ \mathbf{0})^T$, *with* $\widehat{\boldsymbol{\Gamma}}_1$ *the first* d *eigenvectors of* $(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1/2}\mathbf{B}_{\text{fit},11}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11})^{-1/2}$. *The maximum value of the log likelihood is*

$$L^1_d = -\frac{np_1}{2}\log(2\pi) - \frac{np_1}{2} - \frac{n}{2}\log\left|\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},11}\right| - \frac{n}{2}\log\left|\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res},22.1}\right| - \frac{n}{2}\sum_{i=1}^{d}\log(\widehat{\lambda}^1_i + 1). \tag{6.4}$$

Under the hypothesis $Y \perp\!\!\!\perp \mathbf{X}_{12}|(\mathbf{X}_{11},\ \mathbf{X}_2)$ the likelihood ratio statistic $\Theta_d = 2(L_d - L^1_d)$ has an asymptotic chi-squared distribution with $2dp_{12}$ degrees of freedom. Here L_d is given in equation (4.16) in Theorem 4.1. This test statistic can be expressed as

$$\Theta_d = n\sum_{i=1}^{d}\log\left(\frac{1+\widehat{\lambda}^1_i}{1+\widehat{\lambda}_i}\right),$$

where $\widehat{\lambda}_i$ are the eigenvalues of $(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}})^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit}}\right)(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}})^{-1/2}$.

Although we proposed the MLEs of all parameters and the likelihood ratio test statistic, the method requires a large sample with sufficiently small p_1. Moreover any asymptotic statistical test may not hold if we do not have a sufficiently large sample. Thus, we need to consider alternative approaches to test which predictors are inactive when n is less than p_1.

Assume that the relevant predictors are conditionally independent of the irrelevant ones, $\mathbf{X}_{11} \perp\!\!\!\perp \mathbf{X}_{12}|(Y,\ \mathbf{X}_2)$ equivalently $\boldsymbol{\Omega}_{12} = 0$. When $\boldsymbol{\Omega}_{12} = 0$, using the blockwise inversion formula, $\boldsymbol{\Omega}^{21} = -(\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^{-1}_{22}\boldsymbol{\Omega}_{21})^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}^{-1}_{22}$, Lemma 6.1 is represented as $Y \perp\!\!\!\perp \mathbf{X}_{12}|(\mathbf{X}_{11}, \mathbf{X}_2)$ if and only if $\boldsymbol{\Gamma}_{02} = 0$ and $\boldsymbol{\Gamma}_2 = 0$.

Consider another assumption for a different approach. Suppose that $\boldsymbol{\Omega}$ can be decomposed as $\boldsymbol{\Gamma}\mathbf{M}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\mathbf{M}_0\boldsymbol{\Gamma}^T_0$ using the fact that $(\boldsymbol{\Gamma},\ \boldsymbol{\Gamma}_0)$ is an orthogonal matrix. Here $\mathbf{M} > 0$ and $\mathbf{M}_0 > 0$. With this decomposition the sufficient reduction can be expressed as $(\boldsymbol{\Gamma}_0\mathbf{M}^{-1}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}\mathbf{M}^{-1})^T\mathbf{X}_1$ using the decomposition of $\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{M}^{-1}$ and

$\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}_0\mathbf{M}_0^{-1}$. Therefore, a row of $\boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})$ is zero if and only if the corresponding rows of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}$ are zero.

With partial PFC models, $\mathbf{X}_{12}$ is inactive if and only if the corresponding rows of $\boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})$ are all equal to zero. Theorem 6.1 addressed this possibility in terms of general partial PFC models with a data-rich regression. For the isotropic and diagonal partial PFC models and the partial PFC model with covariance $\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$, regardless of the number of observations, the predictors $\mathbf{X}_{12}$ are inactive if and only if the corresponding rows of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}$ are zero.

6.1.2 Testing procedure for predictors

The set of irrelevant predictors can be screened by testing the null hypothesis $\boldsymbol{\Gamma}_{02} = \boldsymbol{\Gamma}_2 = 0$ under the approaches in the previous section. This can be done by considering individual rows of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}$ and testing the hypothesis $\boldsymbol{\gamma}_{0j} = 0$ and $\boldsymbol{\gamma}_j = 0,\ j = 1, \ldots, p_1$, where $\boldsymbol{\gamma}_{0j}$ and $\boldsymbol{\gamma}_j$ are the jth row element of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}$ respectively.

Following the method proposed by Adragni (2008), consider using the univariate isotropic partial PFC model to determine whether individual predictors are relevant or not. Starting with model (4.9) we can derive the univariate isotropic partial PFC model.

$$\begin{aligned} X_{1j}|(Y, \mathbf{X}_2) &= \bar{\mu}_{1j} + \boldsymbol{\gamma}_{0j}^T\boldsymbol{\beta}_0\mathbf{X}_2 + \boldsymbol{\gamma}_j^T\boldsymbol{\alpha}\mathbf{f}_y + \sigma\varepsilon_j \\ &= \bar{\mu}_{1j} + \boldsymbol{\phi}_{1j}\mathbf{X}_2 + \boldsymbol{\phi}_{2j}\mathbf{f}_y + \sigma\varepsilon_j,\ \ j = 1, \ldots, p_1, \end{aligned} \tag{6.5}$$

where p_2-vector $\boldsymbol{\phi}_{1j}$ is equal to $\boldsymbol{\gamma}_{0j}^T\boldsymbol{\beta}_0$ and $(p_2 + r)$-vector $\boldsymbol{\phi}_{2j} = \boldsymbol{\gamma}_j^T\boldsymbol{\alpha}$.

This model is a linear regression model where X_{1j} is the conditional $X_{1j}|(Y = y, \mathbf{X}_2 = \mathbf{x}_2)$ and $\mathbf{f}_y$, defined as in model (4.8), is a known function of y and we assume $\varepsilon \sim N(0, 1)$. Since $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}$ have full column rank, $\boldsymbol{\phi}_{1j} = \boldsymbol{\phi}_{2j} = 0$ if and only if $\boldsymbol{\gamma}_{0j} = \boldsymbol{\gamma}_j = 0$. The relevance of a predictor X_{1j} is assessed by determining whether the mean function $E(X_{1j}|(Y,\ \mathbf{X}_2))$ depends on $\mathbf{X}_2$ and y. A nonconstant mean function can be evaluated by testing the hypotheses $\boldsymbol{\phi}_{1j} = \boldsymbol{\phi}_{2j} = 0$. A predictor is relevant when at least one of $\boldsymbol{\phi}_{1j}$ and $\boldsymbol{\phi}_{2j}$ is not equal to zero.

Model (6.5) is simply a forward linear model with predictors $\mathbf{X}_2$ and $\mathbf{f}_y$, and response X_{1j}. When this model is fitted, an F statistic can be used to test the null hypothesis

$$H_0 : \boldsymbol{\phi}_{1j} = \boldsymbol{\phi}_{2j} = 0.$$

The F test statistic can therefore be used as a criterion for selection. A predictor X_{1j} is relevant if the model yields an F statistic smaller than a user-specified cutoff value. The cutoff can correspond to a significance level α such as 0.1 or 0.05 for example. We will refer to this screening method as SPPFC.

Adragni (2008) pointed out that since the implementation of this screening method involves a wide variety of basis function for $\mathbf{f}_y$, screening by PFC model is more powerful, flexible and versatile as well as superior to many existing methods encountered in the literature.

6.2 Choice of d

The dimension d of the sufficient reduction was so far assumed to be known, but inference on d, which is in effect a model selection parameter, may be required in practice. Two general methods for choosing d in practice have been studied in literature. In this section we will study how to choose d under the isotropic partial PFC model. The process of choosing d under other partial probabilistic models follows the same procedure.

The first method for choosing d is based on using likelihood ratio statistics. Under the isotropic partial PFC model (4.9), let $\widehat{L}(d)$ denote the value of the maximized log likelihood given d,

$$\widehat{L}(d) = -\frac{np_1}{2}\log(2\pi) - \frac{np_1}{2}(\log(\widehat{\sigma}_d^2) + 1),$$

where $\widehat{\sigma}_d^2$ denotes the MLE of σ_d^2 given d. Then the likelihood test statistic to test the null hypothesis $d = d_0$ is

$$\Lambda(d_0, d_{\max}) = 2\{\widehat{L}(d_{\max}) - \widehat{L}(d_0)\} = np_1 \log\left(\frac{\widehat{\sigma}_{d_0}^2}{\widehat{\sigma}_{d_{\max}}^2}\right).$$

where $\widehat{L}(d_{\max})$ denotes the value of the maximized log likelihood for the full model with $d_0 = d_{\max} = \min(p_1 - p_2, r)$. Under the null hypothesis $\Lambda(d_0, d_{\max})$ has an asymptotic

chi-square distribution with $(p_1-d_0)(r-d_0)$ degrees of freedom. The likelihood ratio test statistic $\Lambda(d_0, d_{\max})$ can be used sequentially to estimate d, by starting with $d_0 = 0$ and estimating d as the first hypothesized value of d_0 that is not rejected. If the estimated dimension d is zero, then this in effect suggests that the conditional distribution of $Y|(\mathbf{X}_1, \mathbf{X}_2)$ does not depend on $\mathbf{X}_1$. In dimension reduction literature this method for dimension selection is commonly used (see Cook 1998, p.205, for background).

The second approach is based on an information criterion like AIC or BIC. BIC is consistent for d while AIC is minimax-rate optimal (Burnhan and Anderson 2002). For $d \in \{0, \ldots, \min(p_1 - p_2, r)\}$, the dimension is selected that minimizes the information criterion $-2\widehat{L}(d) + h(n)g(d)$, where $h(n)$ is equal to $\log(n)$ for BIC and 2 for AIC, and $g(d) = p_1 + (p_1 - d)(p_2 + d) + d(p_2 + r) + 1$ is the number of parameters to be estimated as a function of d.

Cook and Forzani (2008) demonstrated that reasonable inference on d is possible through selected results from a simulation study. We conclude this section with a similar simulation study to see if we can reasonably infer about d under the isotropic partial PFC model. We first generated $Y \sim N(0,\ \sigma_y^2)$, where σ_y^2 is given in each simulation. With $d = 2$, $\mathbf{X}$ was generated as $\mathbf{X} = \boldsymbol{\Gamma}^*\boldsymbol{\alpha}^*\mathbf{f}_y^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0,\ \sigma^2\mathbf{I}_{p_1})$, $\sigma^2 = 1$, $\boldsymbol{\alpha}^* = \mathbf{I}_2$, $\mathbf{f}_y^* = (y,\ |y|)^T$ and $\boldsymbol{\Gamma}^* = (\boldsymbol{\Gamma}_1^*, \boldsymbol{\Gamma}_2^*) \in \mathbb{R}^{p\times 2}$, with $\boldsymbol{\Gamma}_1^* = (1, 1, -1, -1, 0, \ldots, 0)^T/\sqrt{4}$ and $\boldsymbol{\Gamma}_2^* = (1, 0, 1, 0, 1, 0, \ldots, 0)^T/\sqrt{3}$. Then $\mathbf{X}_1$ and $\mathbf{X}_2$ was obtained by separating the generated $\mathbf{X}$ into two vectors, the p_1-vector $\mathbf{X}_1$ with the first p_1 elements and the p_2-vector $\mathbf{X}_2$ with the remaining elements. Let $F(2)$, $F(2,\ 3)$, and $F(2,\ 3,\ 4)$ denote the fraction of simulation runs in which d was estimated to be one of the integer arguments. Likelihood ratio test was done at the 5 percent level in all simulations.

Fitting model (4.9) with $\mathbf{f}_y = (y,\ |y|,\ y^3, y^4, y^5)^T$ and $p_1 = 10$, Figures 6.1(a)-(d) give the fraction $F(2)$ of runs in which the indicated procedure selected $d = 2$ as n varies from 30 to 800 for four values of $\sigma_y \in \{0.5,\ 1,\ 2,\ 5\}$ and the three mehods, AIC, BIC, and LRT, under consideration. The number of repetitions is 200. The variation in the results for adjacent sample size in figures reflects simulation error and systematic trends. The relative performance of the methods in Figure 6.1 depends on the sample size n and signal value σ_y, and all three methods improve as n and σ_y increase.

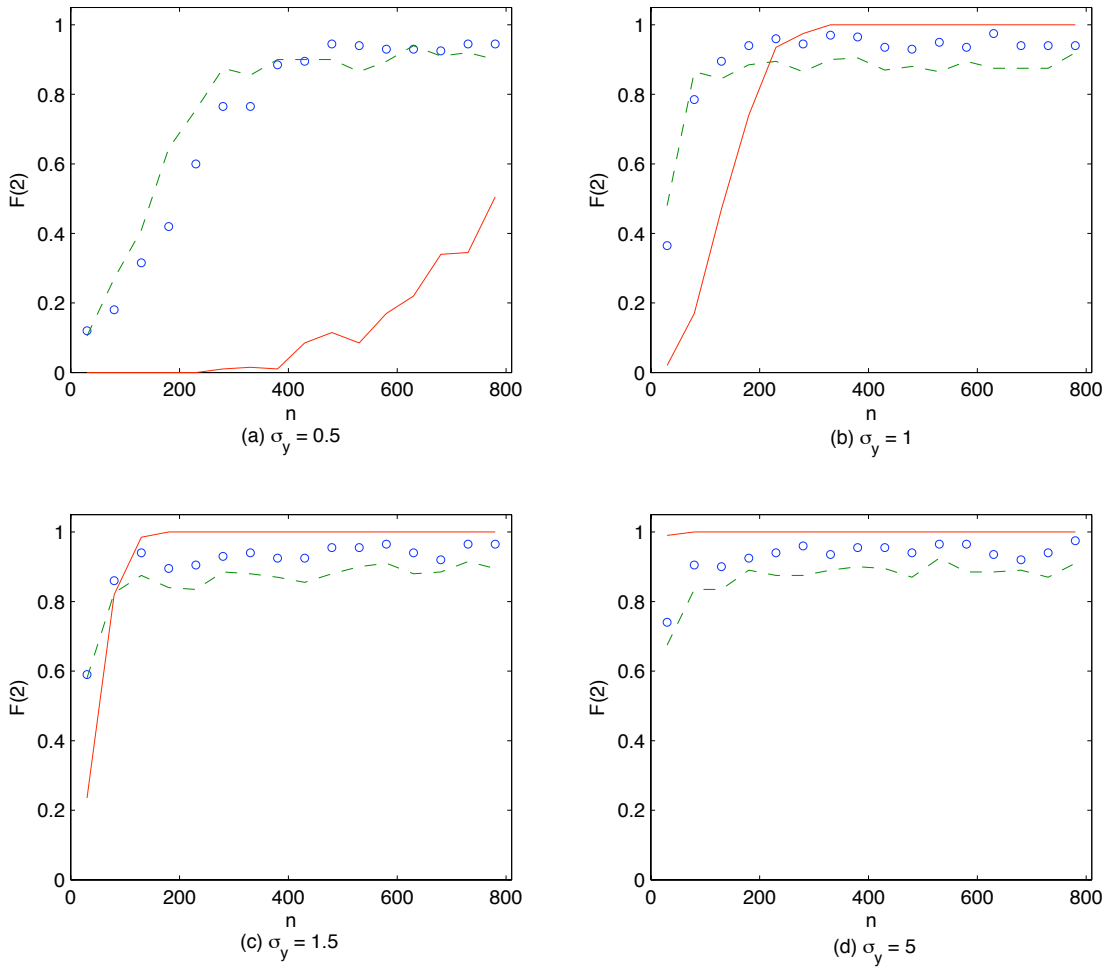


Figure 6.1: Inference about d: Fraction $F(2)$ of replications in which $d = 2$ is chosen by LRT, AIC and BIC versus the sample size n for four values of σ_y. BIC, $-$; AIC, $--$; LRT, $\circ$.

In Figure 6.2 $\sigma_y = 2$ and $n = 200$. Model (4.9) was fitted with $\mathbf{f}_y = (y,\ |y|,\ y^3, \ldots, y^5)^T$ for Figures 6.2(a) and (c) and $\mathbf{f}_y = (y,\ |y|,\ y^3, \ldots, y^{10})^T$ for the other two figures. Figures 6.2(a) and (b) show, as expected, that the chance of choosing the correct value of d decreases with p_1 for all procedures. Figures 6.2(c) and (d) show that, with increasing p_1, AIC and LRT tend to slightly overestimate d, while BIC underestimates d in all the cases estimating d as 0, 1, or 2 hardly taking more than 3. In the case of AIC, when p_1 is bigger than 20 it gave better results than BIC and LRT in Figures 6.2(a) and (b). Additionally it estimated nearly a 100 percent chance that the estimated d is 2, 3 or 4 in Figures 6.2(c) and (d). This simulation result is very similar to that of Cook and Forzani (2008). They mentioned that a little overestimation is not the serious problem in the context of reducing dimensions because $\widehat{R}$ will still estimate a sufficient reduction, but the reduction $\widehat{R}$ no longer estimates a sufficient reduction with underestimation. From this understanding they recommended to use AIC for selecting d.

As Adragni and Cook (2009) pointed out, these methods for choosing d have been shown to work well in data-rich regressions where $p \ll n$ with the reasonable $\mathbf{f}_y$ function, but can be unreliable otherwise like, as we have seen in our results. Therefore, they proposed to determine d using the k-fold cross-validation method (Adragni 2008, Section 4.1.). We will illustrate this method in Section 7.4.

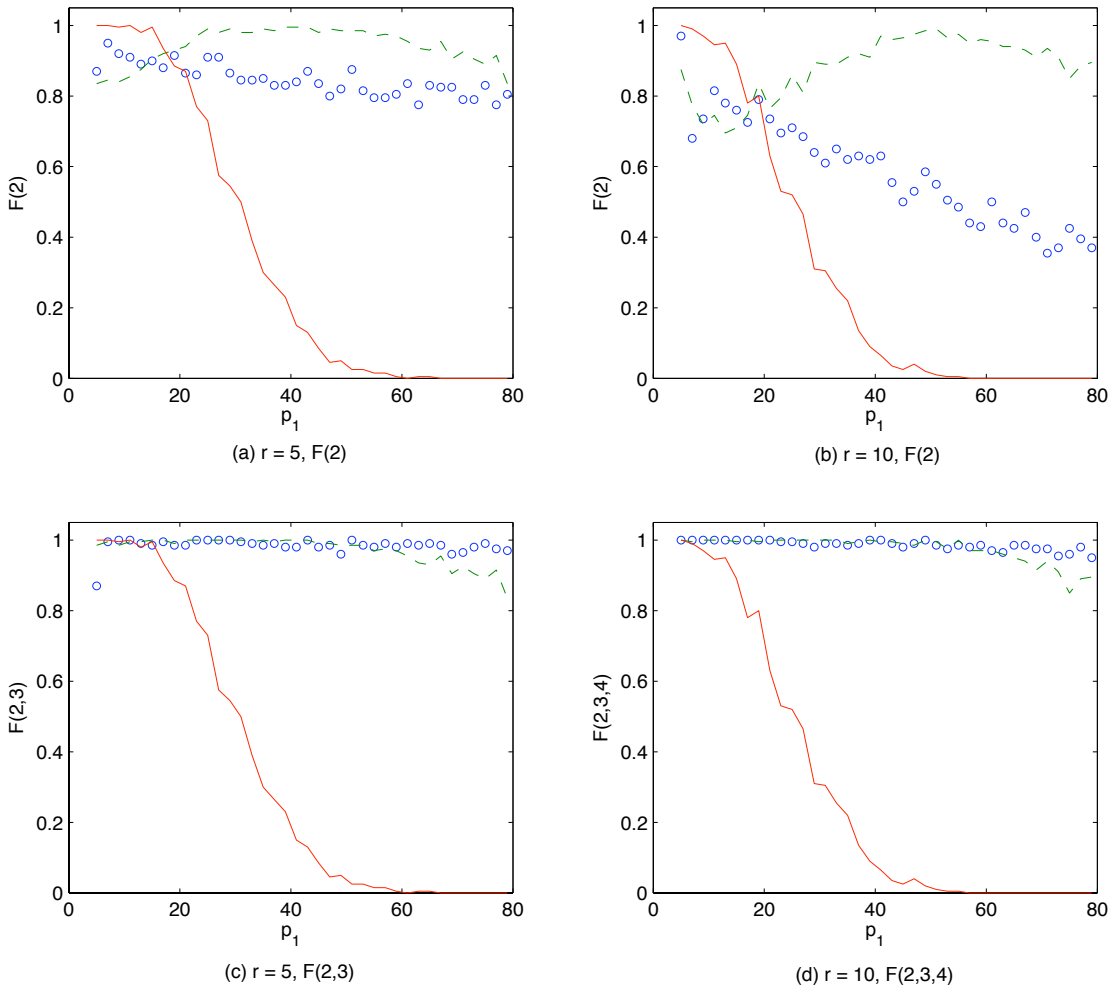


Figure 6.2: Inference about d with varying p_1 and two version of $\mathbf{f}$ used in fitting for LRT, AIC and BIC: BIC, $-$; AIC, $--$; LRT, $\circ$.

Chapter 7

Prediction under partial PFC models

We frequently encounter the problem of predicting a future observation of a continuos univariate response Y at the given continuous predictors $\mathbf{X}$. The prediction problem is to determine $m(\mathbf{x})$ at which the mean squared error $E(Y - m(\mathbf{x}))^2$ is minimized. When $m(\mathbf{x})$ is equal to the mean $E(Y|\mathbf{X} = \mathbf{x})$ of the conditional distribution of $Y|(\mathbf{X} = \mathbf{x})$ the minimization is accomplished. Consequently, the prediction goal is often specialized immediately to the task of estimating the conditional mean function $E(Y|\mathbf{X})$ from the regression of Y on $\mathbf{X}$. Since our reduction R is sufficient, following Definition 1.1 and provided that $R(\mathbf{X})$ captures all of the information that $\mathbf{X}$ contains about Y, the conditional mean function $E(Y|\mathbf{X})$ is the same as $E(Y|R(\mathbf{X}))$. In other words, the estimation of $E(Y|\mathbf{X})$ can be obtained directly through that of $E(Y|R(\mathbf{X}))$.

Generally, a model for $\mathbf{X}|Y$ can itself be inverted to provide a method for estimating the forward mean function $E(Y|\mathbf{X})$ without specifying a model for the full joint distribution of $(\mathbf{X}, Y)$. For convenience we denote the densities of $\mathbf{X}$ and $\mathbf{X}|Y$ by $g(\mathbf{X})$ and $g(\mathbf{X}|Y)$, and so on, where the symbol g indicated a different density in each case. Densities will always appear together with their arguments so this should cause no ambiguity. We assume that $R(\mathbf{X})$ has a density as well. With these understandings Adragni and Cook (2009) developed a method to estimate $E(Y|\mathbf{X})$ under various conditions.

Since we are dealing with the problem of predicting a future observation of a response, this issue only occurs when a response is present. That is, in this chapter the four partial PFC models from Chapter 4 will be discussed in the context of prediction. We will also discuss how to compute the mean functions under each model and then propose a method to calculate prediction errors based on the mean functions.

7.1 Mean function under partial PFC models

Under the partial PFC model (4.8) we are interested in predictions based on a function $R(\mathbf{X}_1)$ which has dimension less than p_1 and satisfies $E(Y|\mathbf{X}_1, \mathbf{X}_2) = E(Y|R(\mathbf{X}_1), \mathbf{X}_2)$. Then we can write a relationship between the forward mean function of $Y|(\mathbf{X}_1, \mathbf{X}_2)$ and the conditional density of $R(\mathbf{X}_1)|(\mathbf{X}_2, Y)$ as follows

$$E\{Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2\} = \frac{E\{Y g(\mathbf{x}_1, \mathbf{x}_2|Y)\}}{E\{g(\mathbf{x}_1, \mathbf{x}_2|Y)\}} \tag{7.1}$$

$$= E\{Y|R(\mathbf{x}_1), \mathbf{x}_2\} \tag{7.2}$$

$$= \frac{E\{Y g(R(\mathbf{x}_1), \mathbf{x}_2|Y)\}}{E\{g(R(\mathbf{x}_1), \mathbf{x}_2|Y)\}} \tag{7.3}$$

$$= \frac{E\{Y g(R(\mathbf{x}_1)|\mathbf{x}_2, Y) g(\mathbf{x}_2|Y)\}}{E\{g(R(\mathbf{x}_1)|\mathbf{x}_2, Y) g(\mathbf{x}_2|Y)\}}, \tag{7.4}$$

where all expectations on the right-hand side are with respect to the marginal distribution of Y. In equation (7.1) we find a relationship between the mean function $E(Y|\mathbf{X}_1, \mathbf{X}_2)$ and the conditional density of $\mathbf{X}_1, \mathbf{X}_2|Y$. The equality of $E(Y|\mathbf{X}_1, \mathbf{X}_2) = E(Y|R(\mathbf{X}_1), \mathbf{X}_2)$ is used in equation (7.2) and the same relationship as in equation (7.1) holds in equation (7.3) with respect to $R(\mathbf{X}_1)$ instead of $\mathbf{X}_1$. We used the conditional density equality, $g(R(\mathbf{x}_1), \mathbf{x}_2|Y) = g(R(\mathbf{x}_1)|\mathbf{x}_2, Y) g(\mathbf{x}_2|Y)$, in order to establish a relationship between the forward mean function and the conditional density of $R(\mathbf{X}_1)|(\mathbf{X}_2, Y)$ in equation (7.4). Since the partial PFC model for $\mathbf{X}_1$ given $\mathbf{X}_2$ and Y was dealt with in Chapter 4, we can derive the density of $R(\mathbf{X}_1)$ given $\mathbf{X}_2$ and Y easily.

The equation (7.4) can be rewritten when $\mathbf{X}_2$ is discrete.

$$\begin{aligned}
\frac{E\{Yg(R(\mathbf{x}_1)|\mathbf{x}_2,Y)g(\mathbf{x}_2|Y)\}}{E\{g(R(\mathbf{x}_1)|\mathbf{x}_2,Y)g(\mathbf{x}_2|Y)\}} &= \frac{\int yg(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(\mathbf{x}_2|y)g(y)dy}{\int g(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(\mathbf{x}_2|y)g(y)dy} \\
&= \frac{\int yg(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(\mathbf{x}_2,y)dy}{\int g(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(\mathbf{x}_2,y)dy} \\
&= \frac{\int yg(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(y|\mathbf{x}_2)g(\mathbf{x}_2)dy}{\int g(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(y|\mathbf{x}_2)g(\mathbf{x}_2)dy} \\
&= \frac{\int yg(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(y|\mathbf{x}_2)dy}{\int g(R(\mathbf{x}_1)|\mathbf{x}_2,y)g(y|\mathbf{x}_2)dy} \\
&= \frac{E_{Y|\mathbf{x}_2}Yg(R(\mathbf{x}_1)|\mathbf{x}_2,Y)}{E_{Y|\mathbf{x}_2}g(R(\mathbf{x}_1)|\mathbf{x}_2,Y)}.
\end{aligned}$$

Consequently,

$$E\{Y|\mathbf{X}_1=\mathbf{x}_1,\mathbf{X}_2=\mathbf{x}_2\} = \frac{E_{Y|\mathbf{x}_2}Yg(R(\mathbf{x}_1)|\mathbf{x}_2,Y)}{E_{Y|\mathbf{x}_2}g(R(\mathbf{x}_1)|\mathbf{x}_2,Y)}.$$

We can also use a forward model for $Y|(\mathbf{X}_2=\mathbf{x}_2)$ to have a relationship between $E(Y|\mathbf{X}_1,\mathbf{X}_2)$ and $E(R(\mathbf{X}_1)|(\mathbf{X}_2,Y))$. However, we assume that $\mathbf{X}_2$ is continuous variable and the conditional density of $\mathbf{X}_2|Y$ can be defined in this chapter. Thus, we will use the relationship in (7.4) through the chapter.

With the observed response $\mathbf{Y}=(Y_1,\ldots,Y_n)^T$, the predicted value $\hat{Y}$ for a given observation $\mathbf{X}_1=\mathbf{x}_1$ and $\mathbf{X}_2=\mathbf{x}_2$ is obtained as

$$\widehat{E}\{Y|\mathbf{X}_1=\mathbf{x}_1,\mathbf{X}_2=\mathbf{x}_2\} = \sum_{i=1}^{n} w_{y_i}(\mathbf{x}_1,\mathbf{x}_2)Y_i, \tag{7.5}$$

$$\text{where } w_{y_i}(\mathbf{x}_1,\mathbf{x}_2) = \frac{\widehat{g}(\widehat{R}(\mathbf{x}_1)|Y_i,\mathbf{x}_2)\widehat{g}(\mathbf{x}_2|Y_i)}{\sum_{i=1}^{n}\widehat{g}(\widehat{R}(\mathbf{x}_1)|Y_i,\mathbf{x}_2)\widehat{g}(\mathbf{x}_2|Y_i)}.$$

Here $\widehat{g}$ denotes an estimated density and $\widehat{R}$ is the estimated reduction. The estimated conditional expectation $\widehat{E}\{Y|\mathbf{X}_1=\mathbf{x}_1,\mathbf{X}_2=\mathbf{x}_2\}$ is a weighted mean function with weight $w_{y_i}(\mathbf{x}_1,\mathbf{x}_2)$. To calculate this weight we have to specify the estimated density function, $\widehat{g}(\mathbf{x}_2|Y_i)$. Since the conditional density of $\mathbf{X}_2|Y$ is not restrictive, any multivariate model for the inverse regression of $\mathbf{X}_2$ on Y can be considered. We assume that this inverse regression model follows an isotropic PFC model as follows

$$\mathbf{X}_2|Y = \bar{\boldsymbol{\mu}}_2^* + \boldsymbol{\Gamma}_2^*\boldsymbol{\alpha}_2^*\mathbf{h}_y^* + \sigma_2\boldsymbol{\epsilon}_2. \tag{7.6}$$

Here $\bar{\boldsymbol{\mu}}_2^* \in \mathbb{R}^{p_2}$, $\boldsymbol{\Gamma}_2^* \in \mathbb{R}^{p_2 \times d_2}$, $d_2 < p_2$, $\boldsymbol{\Gamma}_2^{*T}\boldsymbol{\Gamma}_2^* = \mathbf{I}_{d_2}$, $\sigma_2 > 0$, $\boldsymbol{\alpha}_2^* \in \mathbb{R}^{d_2 \times r_2}$, $d_2 \leq r_2$, has rank d_2 which is assumed to be known, and $\mathbf{h}_y^* \in \mathbb{R}^{r_2}$ is a known vector-valued function of the response with $\sum_y \mathbf{h}_y^* = 0$. The error vector $\boldsymbol{\epsilon}_2 \in \mathbb{R}^{p_2}$ is assumed to be independent of Y, and to be normally distributed with mean 0 and identity covariance matrix. Then the estimated density function is written as

$$\widehat{g}(\mathbf{x}_2|Y_i) \propto \exp\{ -(2\widehat{\sigma}_2)^{-1}\|\mathbf{x}_2 - \widehat{\bar{\boldsymbol{\mu}}}_2^* - \widehat{\boldsymbol{\Gamma}}_2^*\widehat{\boldsymbol{\alpha}}_2^*\mathbf{h}_{y_i}^*\|^2\}, \tag{7.7}$$

where $\widehat{\bar{\boldsymbol{\mu}}}_2^* = \bar{\mathbf{X}}_2$, $\widehat{\boldsymbol{\Gamma}}_2^*$ is constructed by the first d_2 eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{2|\mathbf{h}^*} = \mathbb{X}_2^T\mathbf{P}_{\mathbf{H}^*}\mathbb{X}_2/n$, $\widehat{\boldsymbol{\alpha}}^* = \widehat{\boldsymbol{\Gamma}}_2^{*T}\mathbb{X}_2^T\mathbf{H}^*(\mathbf{H}^{*T}\mathbf{H}^*)^{-1}$, where $\mathbb{X}_2$ denotes the $n \times p_2$ matrix with rows $(\mathbf{X}_2 - \bar{\mathbf{X}}_2)^T$ and $\mathbf{H}^*$ denotes the $n \times r_2$ matrix with rows $\mathbf{h}_y^{*T}$. The corresponding estimate of scale is $\widehat{\sigma}_2^2 = (\sum_{i=1}^{p_2}\lambda_i(\widehat{\boldsymbol{\Sigma}}_2) - \sum_{i=1}^{d_2}\lambda_i(\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{2|\mathbf{h}^*}))/p_2$, where $\widehat{\boldsymbol{\Sigma}}_2 = \mathbb{X}_2^T\mathbb{X}_2/n$.

Isotropic partial PFC model We studied four types of partial PFC model in Chapter 3 according to the different error covariance structures. The first was the isotropic partial PFC model (4.9). The sufficient reduction for this isotropic case is estimated as $\widehat{R}(\mathbf{x}_1) = (\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0,\ \widehat{\boldsymbol{\Gamma}})^T\mathbf{x}_1$. Let $\widetilde{\mathbf{X}}_{1i}^f = \bar{\mathbf{X}}_1 + \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0(\mathbf{x}_2 - \bar{\mathbf{X}}_2) + \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\alpha}}\mathbf{f}_{y_i}$. Then the weights on (7.5) are

$$w_{y_i}(\mathbf{x}_1, \mathbf{x}_2) \propto \exp\{ -(2\widehat{\sigma}^2)^{-1}(\widehat{R}(\mathbf{x}_1) - \widehat{R}(\widetilde{\mathbf{X}}_{1i}^f))^T\widehat{\mathbf{M}}^{-1}(\widehat{R}(\mathbf{x}_1) - \widehat{R}(\widetilde{\mathbf{X}}_{1i}^f))\}\ \widehat{g}(\mathbf{x}_2|Y_i), \tag{7.8}$$

where $\widehat{\mathbf{M}} = ((\widehat{\boldsymbol{\beta}}_0^T\widehat{\boldsymbol{\beta}}_0,\ \mathbf{0})^T,\ (\mathbf{0},\ \mathbf{I}_d)^T)$. The estimators, $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Gamma}}_0$, $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\sigma}^2$, can be obtained from Section 4.2.

Diagonal partial PFC model The weights for the mean function under the diagonal partial PFC model are similar in form to (7.8). The estimated sufficient reduction is given by $\widehat{R}_{\text{d}}(\mathbf{x}_1) = (\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0,\ \widehat{\boldsymbol{\Gamma}})^T\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{x}_1$ with $\widehat{\boldsymbol{\Omega}} = \text{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_{p_1}^2)$. Then the weights have the form

$$w_{y_i}(\mathbf{x}_1, \mathbf{x}_2) \propto \exp\{ -1/2(\widehat{R}_{\text{d}}(\mathbf{x}_1) - \widehat{R}_{\text{d}}(\widetilde{\mathbf{X}}_{1i}^f))^T\widehat{\mathbf{M}}_{\text{diag}}^{-1}(\widehat{R}_{\text{d}}(\mathbf{x}_1) - \widehat{R}_{\text{d}}(\widetilde{\mathbf{X}}_{1i}^f))\}\ \widehat{g}(\mathbf{x}_2|Y_i), \tag{7.9}$$

where $\widehat{\mathbf{M}}_{\text{diag}} = ((\widehat{\boldsymbol{\beta}}_0^T\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0,\ \mathbf{0})^T,\ (\mathbf{0},\ \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}})^T)$. Based on the results from Section 4.3, since $(\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Omega}}^{-1/2})(\widehat{\boldsymbol{\Omega}}^{-1/2}\widehat{\boldsymbol{\Gamma}}_0) = \mathbf{I}_{p_1-d}$ and $(\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Omega}}^{-1/2})(\widehat{\boldsymbol{\Omega}}^{-1/2}\widehat{\boldsymbol{\Gamma}}) = \mathbf{I}_d$, here $\widehat{\mathbf{M}}_{\text{diag}} = \widehat{\mathbf{M}}$ after all. The estimators, $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Gamma}}_0$, $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\Omega}}$, can be obtained from Section 4.3.

General partial PFC model The third case of partial PFC model is the model with general error structure, $\boldsymbol{\Omega} > 0$. We substituted all MLEs obtained as in Section 4.4 into the multivariate normal density for $\mathbf{X}_1|(\mathbf{X}_2, Y)$ and simplified the resulting expression by in part ignoring proportionally constant not depending on the observation i to obtain the estimated density function,

$$\widehat{g}(\widehat{R}(\mathbf{x}_1)|\mathbf{x}_2, Y) \propto \exp\{-\frac{1}{2}(\mathbf{x}_1 - \widetilde{\mathbf{X}}_{1i}^f)^T\widehat{\boldsymbol{\Omega}}^{-1}(\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0,\ \widehat{\boldsymbol{\Gamma}})$$
$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_0^T\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0 & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{\Gamma}}^{*T}\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_0^T\widehat{\boldsymbol{\Gamma}}_0^T \\ \widehat{\boldsymbol{\Gamma}}^T \end{pmatrix} \widehat{\boldsymbol{\Omega}}^{-1}(\mathbf{x}_1 - \widetilde{\mathbf{X}}_{1i}^f)\}.$$

We next discuss how this density function can be simplified to a more intuitive and computationally efficient form. Let $\widetilde{\mathbf{V}} = (\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}(\mathbf{I}_{p_1} + \mathbf{K})^{-1/2}$. Then it is easy to show that $\widehat{\boldsymbol{\Omega}}^{-1} = \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T$ since $\widehat{\boldsymbol{\Omega}} = (\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2})^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_{p_1} + \mathbf{K})\widehat{\mathbf{V}}^T(\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{1|2})^{1/2}$. The columns of $\widehat{\boldsymbol{\Omega}}^{1/2}\widetilde{\mathbf{V}}$ are the normalized eigenvectors of $\widehat{\boldsymbol{\Omega}}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2}\right)\widehat{\boldsymbol{\Omega}}^{-1/2}$ (see Lemma A.1. in Appendix A.11). Let $\widetilde{\mathbf{V}}_d$ and $\widehat{\mathbf{V}}_d$ denote the $p_1 \times d$ matrices consisting of the first d columns of $\widetilde{\mathbf{V}}$ and $\widehat{\mathbf{V}}$. Likewise, let $\widetilde{\mathbf{V}}_{p_1-d}$ and $\widehat{\mathbf{V}}_{p_1-d}$ denote the $p_1 \times (p_1 - d)$ matrices consisting of the last $p_1 - d$ columns of $\widetilde{\mathbf{V}}$ and $\widehat{\mathbf{V}}$. Then, since the MLE of $\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}$ is $\mathcal{S}_d\left(\widehat{\boldsymbol{\Omega}},\ \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2}\right)$, it follows that $\mathcal{S}_d\left(\widehat{\boldsymbol{\Omega}},\ \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2}\right) = \widehat{\boldsymbol{\Omega}}^{-1/2}\text{span}_d(\widehat{\boldsymbol{\Omega}}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2})\widehat{\boldsymbol{\Omega}}^{-1/2}) = \text{span}_d\left(\widehat{\boldsymbol{\Omega}}^{-1/2}\widehat{\boldsymbol{\Omega}}^{1/2}\widetilde{\mathbf{V}}\right) = \text{span}_d\left(\widetilde{\mathbf{V}}\right)$, where $\text{span}_d(\mathbf{A})$ indicates the span of the first d eigenvectors of $\mathbf{A}$. Consequently, we may take $\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}} = \widetilde{\mathbf{V}}_d$ and $\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}}_0 = \widetilde{\mathbf{V}}_{p_1-d}$, which implies that $\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Omega}}^{-1})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}}) = \widetilde{\mathbf{V}}_d^T\widehat{\boldsymbol{\Omega}}\widetilde{\mathbf{V}}_d = \mathbf{I}_d$ and $\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}}_0 = (\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Omega}}^{-1})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Gamma}}_0) = \widetilde{\mathbf{V}}_{p_1-d}^T\widehat{\boldsymbol{\Omega}}\widetilde{\mathbf{V}}_{p_1-d} = \mathbf{I}_{p_1-d}$. Using these results we can rewrite the

reduction as $\widehat{R}_{\rm g}(\mathbf{x}_1) = (\widetilde{\mathbf{V}}_{p_1-d}\widehat{\boldsymbol{\beta}}_0,\ \widetilde{\mathbf{V}}_d)^T\mathbf{x}_1$ and the estimated density function as

$$\begin{aligned}\widehat{g}(\widehat{R}(\mathbf{x}_1)|\mathbf{x}_2, Y) \propto \exp\Big\{ -\frac{1}{2}(\mathbf{x}_1 - \widetilde{\mathbf{X}}_{1i}^f)^T(\widetilde{\mathbf{V}}_{p_1-d}\widehat{\boldsymbol{\beta}}_0,\ \widetilde{\mathbf{V}}_d) \\ \begin{pmatrix} \widehat{\boldsymbol{\beta}}_0^T\widehat{\boldsymbol{\beta}}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_0^T\widetilde{\mathbf{V}}_{p_1-d}^T \\ \widetilde{\mathbf{V}}_d^T \end{pmatrix} (\mathbf{x}_1 - \widetilde{\mathbf{X}}_{1i}^f)\Big\} \\ = \exp\Big\{ -\frac{1}{2}(\widehat{R}_{\rm g}(\mathbf{x}_1) - \widehat{R}_{\rm g}(\widetilde{\mathbf{X}}_{1i}^f))^T \begin{pmatrix} \widehat{\boldsymbol{\beta}}_0^T\widehat{\boldsymbol{\beta}}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d \end{pmatrix}^{-1} (\widehat{R}_{\rm g}(\mathbf{x}_1) - \widehat{R}_{\rm g}(\widetilde{\mathbf{X}}_{1i}^f))\Big\}.\end{aligned}$$

Therefore, the weights become

$$w_{y_i}(\mathbf{x}_1, \mathbf{x}_2) \propto \exp\big\{ -1/2(\widehat{R}_{\rm g}(\mathbf{x}_1) - \widehat{R}_{\rm g}(\widetilde{\mathbf{X}}_{1i}^f))^T\widehat{\mathbf{M}}^{-1}(\widehat{R}_{\rm g}(\mathbf{x}_1) - \widehat{R}_{\rm g}(\widetilde{\mathbf{X}}_{1i}^f))\big\}\ \widehat{g}(\mathbf{x}_2|Y_i). \tag{7.10}$$

The estimators, $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Gamma}}_0$, $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\Omega}}$, are obtained from Section 4.4.

Partial PFC model with the variance function The fourth case of partial PFC model is the model with the variance function $\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$. The estimated sufficient reduction is given by $\widehat{R}(\mathbf{x}_1) = (\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\beta}}_0,\ \widehat{\boldsymbol{\Gamma}})^T\mathbf{x}_1$ by Corollary 4.2. Then the weights have the form

$$w_{y_i}(\mathbf{x}_1, \mathbf{x}_2) \propto \exp\big\{ -1/2(\widehat{R}(\mathbf{x}_1) - \widehat{R}(\widetilde{\mathbf{X}}_{1i}^f))^T\widehat{\mathbf{M}}_{\rm v}^{-1}(\widehat{R}(\mathbf{x}_1) - \widehat{R}(\widetilde{\mathbf{X}}_{1i}^f))\big\}\ \widehat{g}(\mathbf{x}_2|Y_i), \tag{7.11}$$

where $\widehat{\mathbf{M}}_{\rm v} = ((\widehat{\sigma}^2\widehat{\boldsymbol{\beta}}_0^T\widehat{\boldsymbol{\beta}}_0,\ \mathbf{0})^T,\ (\mathbf{0},\ (\widehat{\boldsymbol{\Phi}} + \widehat{\sigma}^2\mathbf{I}_d))^T)$. Here the estimators, $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Gamma}}_0$, $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\sigma}^2$ are obtained from Section 4.5.

7.2 Mean function under g-RMAVE

In the previous section we estimated the mean function $E(Y|\mathbf{X}_1, \mathbf{X}_2)$ by specifying a parametric model for the inverse regression of $\mathbf{X}_1$ on Y and $\mathbf{X}_2$. In this section we will consider a prediction problem under the groupwise refined minimum average variance estimation (g-RMAVE) method.

Li et al. (2010) proposed g-RMAVE, the refinement of a semiparametric estimator RMAVE proposed by Xia et al. (2002), which is based on local linear smoothing of the

forward regression with adaptive weights. g-MAVE focuses on conducting dimension reduction incorporating prior group knowledge in the predictors, while RMAVE cannot handle a groupwise dimension reduction problem with a priori group information.

Let $\{\mathcal{S}_l, \ldots, \mathcal{S}_g\}$ be subspaces of $\mathbb{R}^p$ that form an orthogonal decomposition of $\mathbb{R}^p$. That is, for $k \neq l$, $\mathbf{v}_k \in \mathcal{S}_k,\ \mathbf{v}_l \in \mathcal{S}_l$, we have $\mathbf{v}_k^T \mathbf{v}_l = 0$, and each vector $\mathbf{v} \in \mathbb{R}^p$ can be uniquely defined as $\mathbf{v}_1 + \cdots + \mathbf{v}_g$ for some $\mathbf{v}_1 \in \mathcal{S}_1, \ldots, \mathbf{v}_g \in \mathcal{S}_g$. This orthogonal decomposition is represented as

$$\mathbb{R}^p = \mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_g,$$

where $\oplus$ indicates the direct sum between two subspaces ($\mathcal{S}_1 \oplus \mathcal{S}_2 = \{\mathbf{v}_1 + \mathbf{v}_2;\ \mathbf{v}_1 \in \mathcal{S}_1,\ \mathbf{v}_2 \in \mathcal{S}_2\}$). For example, suppose $p = 6$ and we would like to reduce the dimension within the groups (X_1, X_3, X_4, X_6) and (X_2, X_5). Then $\mathcal{S}_1$ is the subspace of $\mathbb{R}^p$ spanned by $(\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_6)$, $\mathcal{S}_2$ is the subspace spanned by $(\mathbf{e}_2, \mathbf{e}_4)$, and $\mathbb{R}^6 = \mathcal{S}_1 \oplus \mathcal{S}_2$, where $\mathbf{e}_i \in \mathbb{R}^6$ denotes the vector whose ith element equal to one and other elements equal to zero. These subspaces are given as prior knowledge in general.

Li et al. define a groupwise mean dimension reduction subspace with respect to $\{\mathcal{S}_l, \ldots, \mathcal{S}_g\}$ as any subspace $\mathcal{T}_1 \oplus \cdots \oplus \mathcal{T}_g$ that satisfies $E(Y|\mathbf{X}) = E(Y|P_{\mathcal{T}_1}\mathbf{X}, \ldots, P_{\mathcal{T}_g}\mathbf{X})$. Here subspaces $\mathcal{T}_l \subseteq \mathcal{S}_l$ for $l = 1, \ldots, g$. They also define the intersection of all such subspaces as the groupwise central mean subspace with respect to the orthogonal decomposition $\{\mathcal{S}_l, \ldots, \mathcal{S}_g\}$, provided that the intersection itself is a groupwise mean dimension reduction subspace. It is denoted as $\mathcal{S}_{E(Y|\mathbf{X})}(\mathcal{S}_l, \ldots, \mathcal{S}_g)$. To find a groupwise central mean subspace is the main object of interest in that paper. By construction,

$$\mathcal{S}_{E(Y|\mathbf{X})}(\mathcal{S}_l, \ldots, \mathcal{S}_g) = \mathcal{T}_1^* \oplus \cdots \oplus \mathcal{T}_g^*$$

for some subspaces $\mathcal{T}_1^* \subseteq \mathcal{S}_1, \ldots, \mathcal{T}_g^* \subseteq \mathcal{S}_g$. In order to find a groupwise central mean subspace they proposed an iterative optimization algorithm in which each step has a closed form solution. R code (*http://www4.stat.ncsu.edu/ li/software/GroupDR.R*) for this algorithm is provided by Li.

Under our data structure, we have two groups of predictors and want to reduce the dimension of one group of predictors. With this setting, partial dimension reduction

can be implemented by shielding a subset of predictors from dimension reduction in g-RMAVE. That is, it is viewed as a special case of groupwise dimension reduction with some of the $\mathcal{T}_l^*$ equal to the subspace $\mathcal{S}_l$. In the context of partial dimension reduction we have the relationship $E(Y|\mathbf{X}) = E(Y|P_{\mathcal{T}_1^*}\mathbf{X}, P_{\mathcal{S}_2}\mathbf{X})$. Let $\boldsymbol{\gamma}_l \in \mathbb{R}^{p\times p_l}$ be a basis matrix of $\mathcal{S}_l$ for $l = 1, 2$ and $\boldsymbol{\beta}_1$ be a matrix in $\mathbb{R}^{p_1\times d}$ such that $\text{span}(\boldsymbol{\gamma}_1\boldsymbol{\beta}_1) = \mathcal{T}_1^*$. Then we have

$$E(Y|\mathbf{X}) = E(Y|\boldsymbol{\gamma}_1^T\mathbf{X}, \boldsymbol{\gamma}_2^T\mathbf{X}) = E(Y|\boldsymbol{\beta}_1^T\boldsymbol{\gamma}_1^T\mathbf{X}, \boldsymbol{\gamma}_2^T\mathbf{X}) = E(Y|\boldsymbol{\beta}_1^T\mathbf{X}_1, \mathbf{X}_2). \tag{7.12}$$

Consequently the estimator of $E(Y|\mathbf{X})$ is obtained as $\widehat{E}(Y|\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}_1, \mathbf{X}_2)$ where $\widehat{\boldsymbol{\beta}}_1$ is given by the g-RMAVE method.

7.3 Prediction error with mean functions

The estimation of $E(Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2)$ is the main step for predicting a future observation of a univariate response variable Y at the given value $(\mathbf{x}_1,\ \mathbf{x}_2)$ of $(\mathbf{X}_1,\ \mathbf{X}_2)$. The estimation of the mean function under the partial PFC models completely depends on the density function. Obtaining a good estimate of the density is the key to the success of our method. With the assumed models in the previous sections, the densities are known and well behaved. We can therefore use this method for prediction.

The performance of prediction method is evaluated by estimating the usual mean squared prediction error $E[Y - \widehat{E}(Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2)]^2$. Although there are numerous techniques for assessing the mean squared prediction error (PE), we estimate PE as

$$\widehat{\text{PE}} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i^* - \widehat{E}(Y|\mathbf{X}_1 = \mathbf{x}_{1i}^*, \mathbf{X}_2 = \mathbf{x}_{2i}^*)\right]^2,$$

where n new observations $(Y_i^*,\ \mathbf{x}_{1i}^*,\ \mathbf{x}_{2i}^*),\ i = 1, \ldots, n$ is generated under the various conditions based on the models. This setup can be used for example with datasets generated from known models. Here the estimated mean functions for the four partial PFC models are given in Section 7.1 according to the error structure.

For the estimated mean function under g-RMAVE, we assume that the forward

regression $Y|\mathbf{X}$ follows a textbook normal linear regression model,

$$Y = \alpha_0 + \boldsymbol{\eta}^T \mathbf{x}_1 + x_2 + \sigma_{Y|\mathbf{X}}\epsilon, \tag{7.13}$$

where $\mathbf{x}_1$ and x_2 denote observed values of $\mathbf{X}_1$ and X_2 respectively, $\sigma_{Y|\mathbf{X}}$ is constant, and ϵ is a standard normal random variable. For direct comparison with the inverse regression model we set $p_2 = 1$ so that there is no reduction on $\mathbf{X}_2$. Then span($\boldsymbol{\eta}$) is equivalent to span($\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}$), the reductive subspace under the isotropic partial PFC model. Thus, $E(Y|\mathbf{X})$ depends on $\mathbf{X}$ via $(\boldsymbol{\eta}^T\mathbf{X}_1,\ X_2)$ which is the conclusion seen by the relationship between mean functions in (7.12). The estimator of $E(Y|\mathbf{X})$ is obtained as

$$\widehat{E}(Y|\mathbf{X}) = \bar{Y} + \widehat{\boldsymbol{\eta}}^T_{\text{gRMV}}(\mathbf{X}_1 - \bar{\mathbf{X}}_1) + (x_2 - \bar{x}_2), \tag{7.14}$$

where $\boldsymbol{\eta}_{\text{gRMV}}$ can be estimated from g-RMAVE.

7.4 k-fold cross-validation

As we have mentioned in Section 6.2, estimating d using an information criteria like the AIC or BIC or a likelihood ratio statistics are suitable to cases where $p_1 \ll n$. We have to consider an alternative method to estimate d when the number of observations n is not large enough to use these asymptotic methods. Hence, Adragni (2008) proposed k-fold cross-validation.

With a dataset D, split the n observations randomly into K subsets of roughly equal size $D_1, \ldots, D_K$ and let $D_{(-k)}$ denote the set D with D_k being held out. $D_{(-k)}$ is used as a training set to estimate d, $d \in \{0, 1, \ldots, \min(p_1 - p_2, r)\}$. For each possible value d_m of d, the mean squared prediction error is calculated by cross-validation using training set which is considered as the whole data set D. We calculate

$$\widehat{\text{PE}}_{d_m,k} = \frac{1}{n_k}\sum_{k=1}^{K}\sum_{Y_j \in D_k}\left[(Y_j - \widehat{Y}_j)^2 | \mathbf{X}_1 = \mathbf{x}_{1j}^{(k)},\ \mathbf{X}_2 = \mathbf{x}_{2j}^{(k)}\right],$$

where $\mathbf{x}_{1j}^{(k)}$ and $\mathbf{x}_{2j}^{(k)}$ are from the testing set D_k, $j = 1, \ldots, n_k$, with n_k being the number of observations in D_k. The term $\widehat{Y}_j$ is obtained using the equation (7.5), which

is estimated as

$$\widehat{Y}_j = \sum_{Y_i \in D_{(-k)}} w_{y_i}(\mathbf{x}_{1j}^{(k)}, \mathbf{x}_{2j}^{(k)})\, Y_i,$$

where the weight w is obtained in Section 7.1 according to partial PFC models while using the training sets $D_{(-k)}$. The mean squared prediction error for $d = d_m$ is obtained as

$$\widehat{\mathrm{PE}}_{d_m} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathrm{PE}}_{d_m,k}.$$

The value $\hat{d}$ of d that yields the smallest mean squared prediction error is the value to be used. There is no specification of K for the K-fold cross-validation. When the number of observations is large enough, K can be 10. But with a small number of observations, a leave-one-out cross-validation can be used.

Chapter 8

Application on real datasets

We use two real examples in this chapter to illustrate the methodologies proposed in the previous chapters. The first has two points of view with different hypotheses, setting a different variable as a response. It is the most thorough. In the second example, we show results under the designated methodologies when the number of predictors is much larger than the sample size.

8.1 Body dimensions

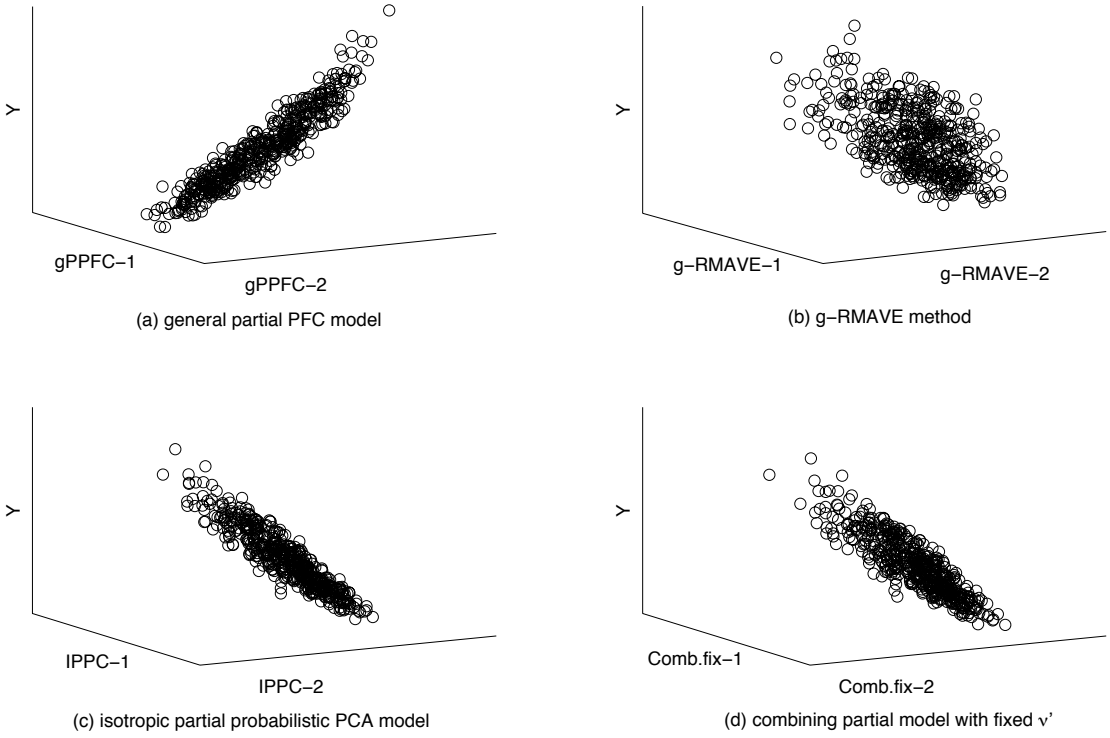
This illustration is from the research of Heinz et al. (2003), which investigated the correspondence between body build, weight, and girths in a group of physically active young men and women, most of whom were within normal weight range. Skeletal width and depth measurements of the trunk and limbs at nine well-defined body sites were used to characterize body build. They took measurements on the 247 men and 260 women ($n = 507$). The dataset including the description of variables can be found at *http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html*.

Weight is the response. Here the hypothesis that body build (skeletal) variables and height predict scale weight substantially better than height alone was affirmed for the group. The response is the weight and the predictors are 21 body girth measurements

and skeletal diameter measurements as well as height ($p = 22$). As scatterplots of weight against the predictors all look quite linear, it is no surprise that weight can be predicted well by the various girth measurement and height. They also found the linear combination of predictors obtaining an adjusted R^2 value of 97.3%.

We conducted dimension reduction on 21 body build measurements keeping the height without reduction. Since the height variable is significant in all reduced models there is no need for reduction on the height. To sum up, $\mathbf{X}_1 = \{21$ body build measurements$\}$ with $p_1 = 21$, $\mathbf{X}_2 =$ height with $p_2 = 1$ and $Y =$ weight. Fitting the general partial PFC model (gPPFC) is desirable because all predictors are somewhat correlated, with pairwise sample correlations ranging from 0.3 to 0.9. Plots of each predictor minus height versus the response suggested that we might use a linear function of Y to model $\boldsymbol{\nu}'$, but for this illustration we decided to allow more flexibility and so set $\mathbf{f}_y' = (y, y^2, y^3)$. The 5-fold cross validation method chose $d = 1$, suggesting that only two linear combinations of the predictors are sufficient with $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1 \in \mathbb{R}^{p_2+d} = \mathbb{R}^2$. The plots in Figure 8.1 show Y versus two linear combinations of $\widehat{R}(\mathbf{X}_1)$ based on the four methodologies, the general partial PFC model, g-RMAVE method, isotropic partial probabilistic PCA model, and combining partial model with fixed $\boldsymbol{\nu}'$. When fitting the combining partial model we assumed that randomly selected 253 cases out of 507 have the known responses. Figures 8.1 (a), (c), and (d) show a strong linear relationship as we expected even in Figure 8.1(c) having no information about the response. The reductions from the inverse models reflect the linearity property between response and predictors well. In contrast, there is weak linear relationship in the plot of Y versus two linear components from g-RMAVE shown in Figure 8.1(b). A plot of two SIR linear combinations (not shown) ignoring $\mathbf{X}_2$ and reducing $\mathbf{X}_1$ given Y is quite similar to that shown in Figure 8.1(b).

Gender is the response. Forensic scientists can fairly accurately determine the gender of adults given their skeletal remains; apparently an accuracy rate of 90% or more is possible if the skeletal remains are complete (see Joyce and Stover 1991 or Wingate 1922, for instance). As male and female skeletons show very little difference



(a) general partial PFC model

(b) g−RMAVE method

(c) isotropic partial probabilistic PCA model

(d) combining partial model with fixed ν'

Figure 8.1: Two components versus Y

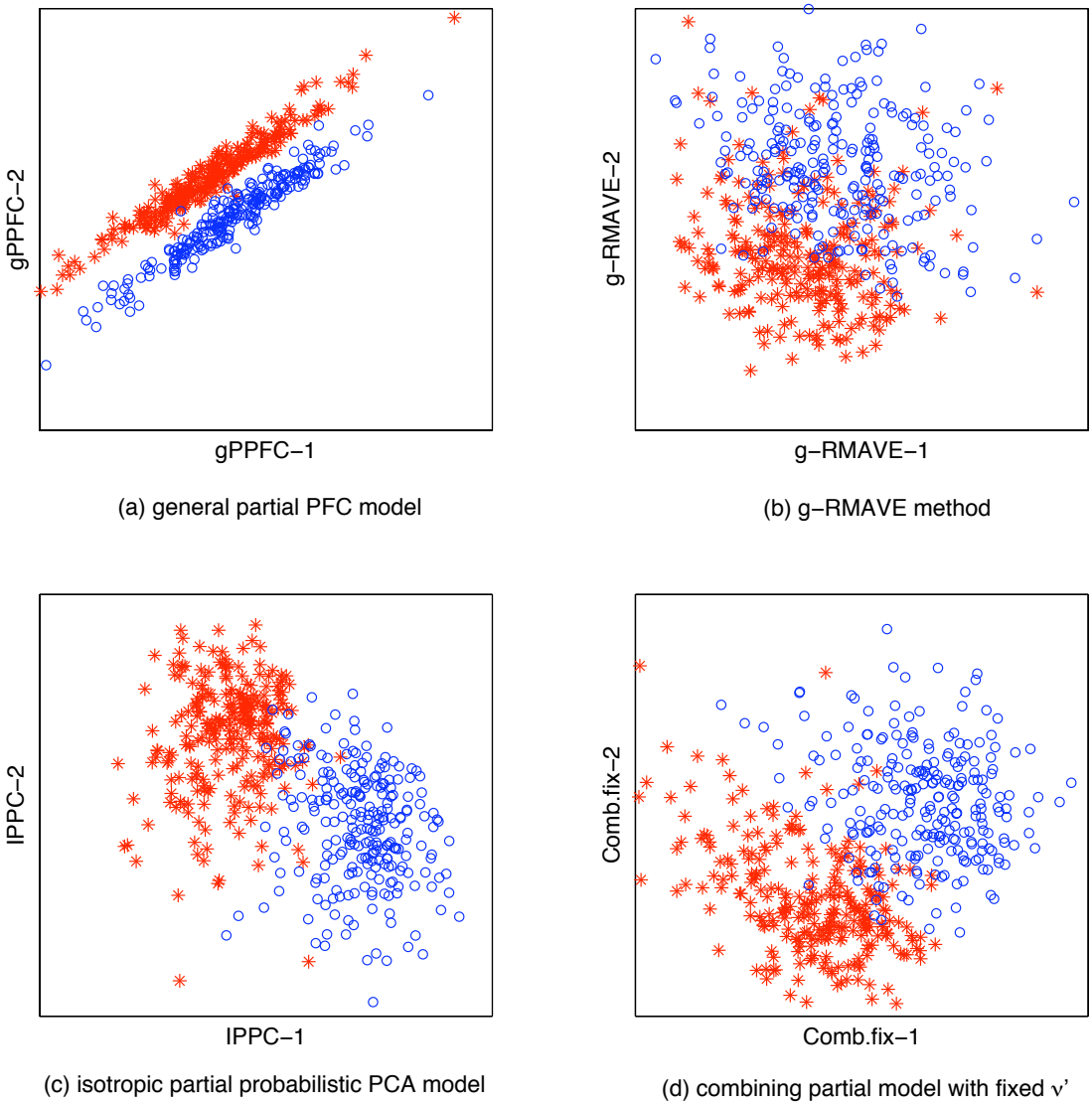
before puberty, gender determination from a child's skeleton is extremely difficult.

Lohman et al. (1988) indicates that biacromial diameter is useful in the evaluation of sex-associated differences in physique. This statement was confirmed in Heinz et al. (2003). Since biacromial diameter is useful to discriminate gender, we set it as $\mathbf{X}_2$ without reduction. Thus, we conducted reduction on 20 body build measurements, weight, and height in the presence of Y = gender and $\mathbf{X}_2$ = biacromial diameter. Dimension reduction on predictors might serve as a preparatory step for developing a classifier. Fitting the general partial PFC model is still desirable because of correlated predictors. Since the response variable is binary, $\mathbf{f}_y$ is naturally determined as $\mathbf{f}_y = (J(y \in C_1), J(y \in C_2))^T$, where J is the indicator function and C_1 denotes the category for female and C_2 for male. The 5-fold cross validation method chose $d = 1$ and then two linear combination of predictors are sufficient. The plots in Figure 8.2 show the first and second linear components marked by gender: female (red $*$), male (blue $\circ$).

A plot of first two gPPFC predictors is shown in Figure 8.2(a). It exhibits strong separation condensing into two discriminated lines according to gender. These first two gPPFC predictors perfectly separate gender, confirming that they are sufficient for discrimination. In contrast, as in Figure 8.2(b) female and male are overplotted while g-RMAVE shows low discrimination capacity. To expand the comparison we also added the results for isotropic partial PCA and for the combining method. Although female and male are still somewhat overplotted with large variance in Figure 8.2(c) and (d), the isotropic partial PCA model and combining partial model show better performance than g-RMAVE method in Figure 8.2(b), while separating female and male reasonably.

8.2 SBRCT gene expression

Gene expression arrays are an important new technology in biology. The dataset in this example is from Hastie et al. (2009). Data form a matrix of 2308 genes (variables) and 63 samples from a set of microarray experiments. Each expression value is a log-ratio $\log(R/G)$. R is the amount of gene-specific RNA in the sample that hybridizes to a particular (gene-specific) spot on the microarray, and G is the corresponding amount of



(a) general partial PFC model

(b) g−RMAVE method

(c) isotropic partial probabilistic PCA model

(d) combining partial model with fixed ν'

Figure 8.2: Two components categorized by Y: female, $*$; male, $\circ$.

RNA from a reference sample. The samples are collected from small, round blue-cell tumors (SRBCT) found in children, and are classified into four major types: BL (Burkitt lymphoma), EWS (Ewing's sarcoma), NB (neuroblastoma), and RMS (rhabdomysosarcoma).

Here we assume that the first predictor contributes highly to the class predictions for the purpose of illustration only, and set that predictor as $\mathbf{X}_2$ with $p_2 = 1$. Thus, our goal is to reduce the remaining $p_1 = 2307$ predictors in the presence of $Y=$ four tumor types and $\mathbf{X}_2$. In the book, they assumed that the genes are independent within each class, that is, the within-class covariance matrix is diagonal. They pointed out that despite the fact that genes will rarely be independent within a class, when the number of predictors is much greater than the sample size we don't have enough data to estimate their dependencies. Hence, here we assume that the predictors are conditionally independent with different scales. With $p_1 = 2298$, we may drop out irrelevant predictors by using the proposed screening method in Section 6.1.2. With significance level 0.01, 1542 genes were discarded, leaving 765 genes that might play a role in the classification process. Fitting the diagonal partial PFC model (dPPFC), the response variable is naturally determined as $\mathbf{f}_y = (J(y \in C_1), J(y \in C_2), J(y \in C_3), J(y \in C_4))^T$, where J is the indicator function and C_i denotes the category for each tumor type. The 5-fold cross validation method chose $d = 1$, suggesting that two linear combinations of the predictors are sufficient with $R(\mathbf{X}_1) = (\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1 \in \mathbb{R}^{p_2+d} = \mathbb{R}^2$. The plots in Figure 8.3 show the first two components marked by tumor types: EWS (green $\bullet$), RMS (blue $\times$), NB (black $\circ$), BL (red $*$).

A plot of first two dPPFC predictors is shown in Figure 8.3(a). It separates four tumor types clearly, confirming that just two dPPFC predictors are sufficient for discrimination. In contrast, as in Figure 8.3(b), the tumors EWS and RMS are overplotted, and NB and BL have no separation. g-RMAVE discriminates just two groups among four tumor types. To expand the comparison we also included the results for isotropic partial PFC model and for general partial PFC model. Figure 8.3(c) shows similar result to Figure 8.3(a) with clear separation, supporting that the assumption of the conditional independence in genes is adequate. However, fitting a general partial PFC

model is not a good idea, showing the poor discrimination performance in Figure 8.3(d). We might need more gPPFC predictors to see the apparent separation between tumor types.

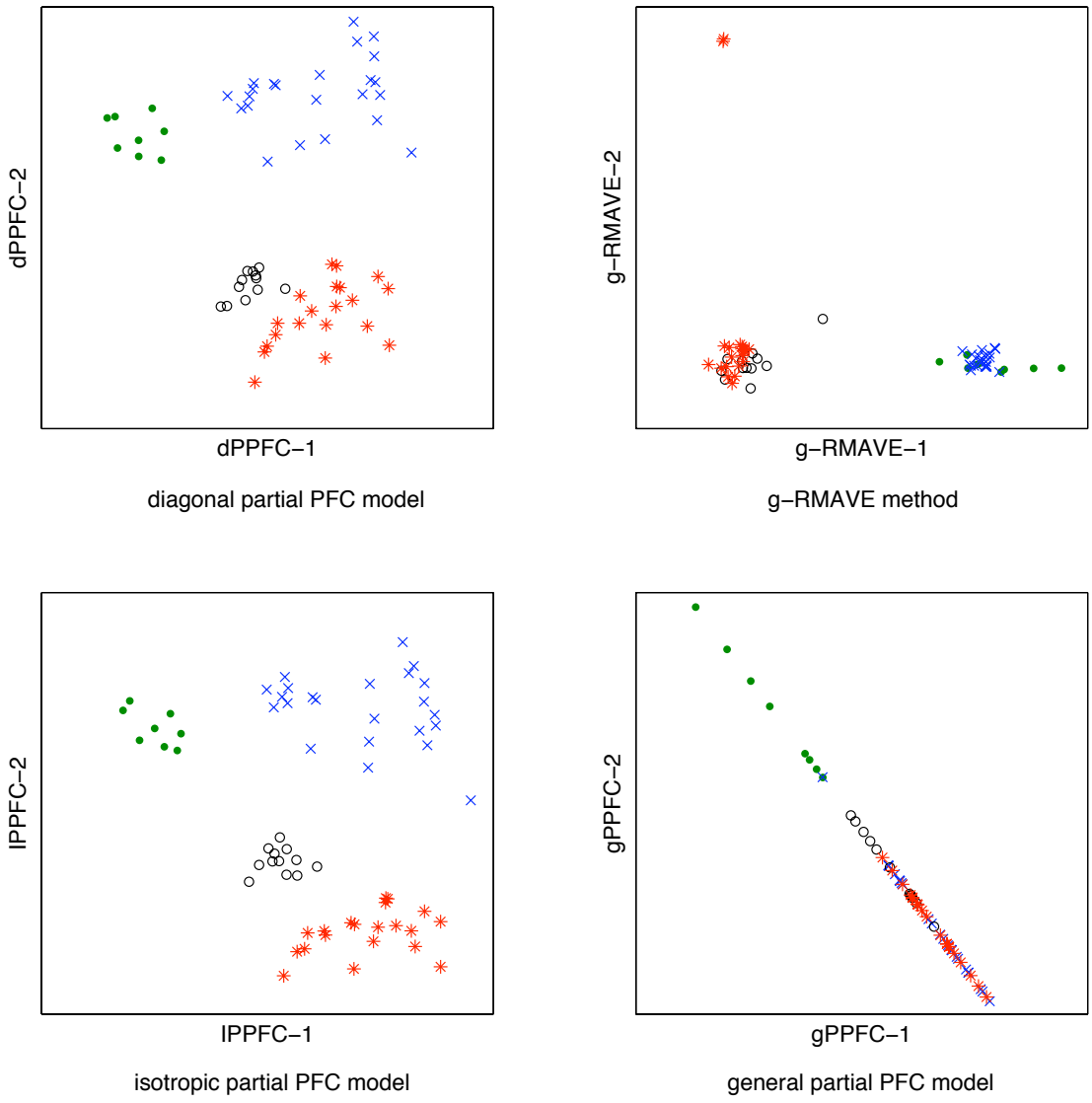


Figure 8.3: Two components categorized by Y: EWS, green •; RMS, blue ×; NB, black ○; BL, red ∗.

Chapter 9

Conclusion and Discussion

In this thesis, we described how to do partial dimension reduction, dimension reduction on one set of predictors for inclusion in a regression with response and another set of predictors. Several partial dimension reduction methods have been proposed since the work of Chiaromonte, Cook, and Li (2002). Mostly these are based on nonparametric or semiparametric method-of-moment arguments, and some of them restricted the type of predictors as categorical which is shielded from reduction process. Little attention was devoted to model-based approaches to the partial dimension reduction problem.

While proposing new model-based partial dimension reduction methods, we mainly considered three types of dataset: no responses are given, all responses are known, and a part of responses is known. While the partial probabilistic PCA model was proposed when there is no responses, the partial PFC model is the way to go when all responses are known. The combining method that combines the partial PFC and the partial probabilistic PCA model was suggested when a part of responses is known. Beside the discussion about the existence of response in the dataset, we developed four types of the partial inverse regression model with various error covariance structures, which widen the applicative scope of partial dimension reduction.

All methods proposed in this thesis provide likelihood-based solutions for partial sufficient dimension reduction estimators. When proposed models are accurate, the

methodology will inherit optimality properties from general likelihood theory. In addition, although methods are not designed for reducing discrete or categorical predictors, there are no restriction on the nature of the response, which may be continuous, categorical or even multivariate. We assumed that $\mathrm{Var}(\mathbf{X}_1|\mathbf{X}_2, Y)$ is constant throughout the thesis. Extensions to non-constant conditional variance are surely worth further research. The work on this issue by Cook and Forzani (2009) may be a good starting point while extending their proposed method to the partial dimension reduction context. In the thesis, we did not consider the problem of reducing dimensions of both $\mathbf{X}_1$ and $\mathbf{X}_2$ simultaneously, but separately. This issue will also be a direction for future study.

References

[1] Adragni, K. P. (2008). *Dimension Reduction and Prediction in Large p Regressions*, Ph.D. dissertation, School of Statistics, University of Minnesota.

[2] Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Statistical Society, A.*, **367**, 1–21.

[3] Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252–1258.

[4] Burges, C. (2010). Dimension Reduction: A Guided Tour. *Foundations and Trends in Machine Learning*, **2**, 275–365.

[5] Burnham, K. and Anderson, D. (2002). *Model section and Multimodel Inference*. Wiley, New York. MR1919620.

[6] Butler, L. M., Koh, W-P., Lee, H-P., Tseng, M., Mimi, C. Y. and London, S. J. (2006). Prospective study of dietary patterns and persistent cough with phlegm among Chinese Singaporeans. *American Journal of Respiratory Critical Care Medicine* **173**, 264–270.

[7] Chen, X. (2010). *Sufficient Dimension Reduction and Variable Selection*, Ph.D. dissertation, School of Statistics, University of Minnesota.

[8] Chen, X. and Cook, R.D. (2010). Some insights into continuum regression and its asymptotic properties. *Biometrika*, **97**, 985–989.

[9] Chen, X., Zou, C and Cook, R.D. (2010). Coordinate-Independent Sparse Sufficient Dimension Reduction and Variable Filtering. *The Annals of Statistics*, **38**, 3696–3723.

[10] Chiaromonte, F. and Cook, R. D. (2002). Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, **54**, 768–795.

[11] Chiaromonte, F., Cook, R.D. and Li, B. (2002). Sufficient dimension reduction in regression with categorical predictors. *The Annals of Statistics*, **30**, 475–497.

[12] Chikuse, Y. (2003). *Statistics on Special Manifolds.* Springer, New York. MR1960435.

[13] Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York, Wiley.

[14] Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, **22**, 1–26.

[15] Cook, R. D. and Li, B. (2002). Dimension Reduction for Conditional Mean in Regression. *The Annals of Statistics*, **30**, 455–474.

[16] Cook, R.D. and Forzani, L. (2008). Covariance reducing models: An alternative to spectral modeling of covariance matrices. *Biometrika*, **95**, 799–812.

[17] Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, **23**, 485–501.

[18] Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197–208.

[19] Cook, R. D., Forzani, L. and Tomassi, D. R. (2011). LDR: A Package for Likelihood-Based Sufficient Dimension Reduction. *Journal of Statistical Software*, **39**.

[20] Cook, R. D. and Li, L. (2009). Dimension reduction in regressions with exponential family predictors. *Journal of Computational and Graphical Statistics*, **18(3)**, 774–791.

[21] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression for Dimension Reduction," by Li, K. C. *Journal of the American Statistical Association*, **86**, 316–327.

[22] Cox, D. R. (1968). Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society: Series A*, **131**, 265–279.

[23] Donoho, D. L. (2000). High-dimensional data analysis: The curse and blessings of dimensionality. Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century.

[24] Edelman, A., Arias, T. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Mathematical Analysis*, **20**, 303–353.

[25] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society: Series B*, **70**, 849–911.

[26] Flury, B. N. (1984). Common Principal Components in K Groups. *Journal of the American Statistical Association*, **79**, 892–898.

[27] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition*, Springer.

[28] Heinz, G., Peterson, L. J., Johnson, R. W. and Kerk, C. J. (2003). Exploring Relationships in Body Dimensions. *Journal of Statistics Education*, **11**.

[29] Henderson, H. V. and Searle, S. R. (1979). Vec and Vech operators for matrices, with some uses in jacobians and multivariate statistics. *The Canadian Journal of Statistics*, **7**, 65–81.

[30] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis 6ed.* Pearson.

[31] Johnstone, I. M. and Titterington D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A*, **367**, 4237–4253.

[32] Jolliffe, I. T. (2002). *Principal component analysis*, 2nd ed. New York, Springer.

[33] Joyce, C. and Stover, E. (1991). *Witnesses from the Grave: The Stories Bones Tell*. Boston, MA: Little, Brown, and Company, 177–178.

[34] Li, B., Cook, R.D. and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *The Annals of Statistics*, **31**, 1636–1668.

[35] Li, K. C. (1991). Sliced inverse regression for dimension reduction with discussion. *Journal of the American Statistical Association*, **86**, 316–327.

[36] Li, L. (2009). Exploiting predictor domain information in sufficient dimension reduction. *Computational Statistics and Data Analysis*. **53**, 2665–2672.

[37] Li, L., Li, B., and Zhu, L.X. (2010). Groupwise dimension reduction. *Journal of the American Statistical Association*, **105**, 1188–1201.

[38] Lohman, T., Roche, A., and Martorell, R. (1988). *Anthropometric Standardization Reference Manual*, Champaign, IL: Human Kinetics Books.

[39] Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *The Annals of Statistics*, **7**, 381–394.

[40] Martinez, W. L. and Martinez, A. R.(2008). *Computational statistics handbook with MATLAB*. London, Chapman and Hall.

[41] Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60**, 234–256.

[42] Schönemann, P. H. (1985). On the formal differentiation of traces and determinants. *Multivariate behavioral Research*, **20**, 113–139.

[43] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B*, **61**, 611-622.

[44] Yin, X. and Zhu, L. X. (2007). Estimating Direction in Extending Generalized Partially Linear Single-Index Models. *Journal of Computational and Graphical Statistics*, **16**, 330–349.

[45] Wingate, A. (1992). *Scene of the Crime: A Writer's Guide to Crime-Scene Investigations*. Cincinnati, OH: Writer's Digest Books, p.148.

[46] Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Appendix A

Proofs of the results

A.1 Proposition 2.1

Proof. The model (2.5) can be rewritten as

$$\begin{aligned}\mathbf{X}_1|(\mathbf{X}_2,\boldsymbol{\nu}') &= \bar{\boldsymbol{\mu}}_1+\boldsymbol{\beta}\mathbf{X}_2+\boldsymbol{\Gamma}\boldsymbol{\nu}'+\boldsymbol{\Psi}^{1/2}\boldsymbol{\varepsilon}\\ &= \bar{\boldsymbol{\mu}}_1+\mathbf{K}\boldsymbol{\eta}+\boldsymbol{\Psi}^{1/2}\boldsymbol{\varepsilon},\end{aligned}$$

where $\mathbf{K}=(\boldsymbol{\beta}\ \ \boldsymbol{\Gamma})$, and $\boldsymbol{\eta}=\begin{pmatrix}\mathbf{X}_2\\ \boldsymbol{\nu}'\end{pmatrix}$.

We first show that the distribution of $\mathbf{X}_1|(\mathbf{K}^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1,\mathbf{X}_2,\boldsymbol{\nu}')$ is the same as the distribution of $\mathbf{X}_1|\mathbf{K}^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1$ for all $\mathbf{X}_2$ and $\boldsymbol{\nu}'$. According to the above model, $\mathbf{X}_1|(\mathbf{X}_2,\boldsymbol{\nu}')$ is normally distributed with mean $\bar{\boldsymbol{\mu}}_1+\boldsymbol{\beta}\mathbf{X}_2+\boldsymbol{\Gamma}\boldsymbol{\nu}'$ and constant variance. Thus $\mathbf{X}_1|(\mathbf{K}^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1,\mathbf{X}_2,\boldsymbol{\nu}')$ is normally distributed with constant variance and mean

$$\begin{aligned}&E(\mathbf{X}_1|\mathbf{K}^T\boldsymbol{\Psi}^{-1}\mathbf{X}_1,\mathbf{X}_2=\mathbf{x}_2,\boldsymbol{\nu}'=\boldsymbol{\nu}')\\ &= E(\mathbf{X}_1|\mathbf{x}_2,\boldsymbol{\nu}')+\mathbf{P}^T_{\boldsymbol{\Psi}^{-1}\mathbf{K}(\boldsymbol{\Psi})}(\mathbf{X}_1-E(\mathbf{X}_1|\mathbf{x}_2,\boldsymbol{\nu}'))\\ &= \bar{\boldsymbol{\mu}}_1+\boldsymbol{K}\boldsymbol{\eta}+\mathbf{P}^T_{\boldsymbol{\Psi}^{-1}\mathbf{K}(\boldsymbol{\Psi})}(\mathbf{X}_1-\bar{\boldsymbol{\mu}}_1-\boldsymbol{K}\boldsymbol{\eta})\\ &= (\mathbf{I}_{p_1}-\mathbf{P}^T_{\boldsymbol{\Psi}^{-1}\mathbf{K}(\boldsymbol{\Psi})})\bar{\boldsymbol{\mu}}_1+\mathbf{P}^T_{\boldsymbol{\Psi}^{-1}\mathbf{K}(\boldsymbol{\Psi})}\mathbf{X}_1+(\mathbf{I}_{p_1}-\mathbf{P}^T_{\boldsymbol{\Psi}^{-1}\mathbf{K}(\boldsymbol{\Psi})})\boldsymbol{K}\boldsymbol{\eta}\\ &= (\mathbf{I}_{p_1}-\mathbf{P}^T_{\boldsymbol{\Psi}^{-1}\mathbf{K}(\boldsymbol{\Psi})})\bar{\boldsymbol{\mu}}_1+\mathbf{P}^T_{\boldsymbol{\Psi}^{-1}\mathbf{K}(\boldsymbol{\Psi})}\mathbf{X}_1,\end{aligned}$$

where $\mathbf{P}_{\mathbf{\Psi}^{-1}\mathbf{K}(\mathbf{\Psi})}$ is the operator that projects onto span($\mathbf{\Psi}^{-1}\mathbf{K}$) in the $\mathbf{\Psi}$ inner product. From this the last term in the third equation is 0. Since $\mathbf{X}_1|(\mathbf{K}^T\mathbf{\Psi}^{-1}\mathbf{X}_1, \mathbf{X}_2, \boldsymbol{\nu}')$ is normally distributed with mean and variance that do not depend on $\mathbf{X}_2$ and $\boldsymbol{\nu}'$, it follows that the distribution of $\mathbf{X}_1|(\mathbf{K}^T\mathbf{\Psi}^{-1}\mathbf{X}_1, \mathbf{X}_2, \boldsymbol{\nu}')$ is the same as the distribution of $\mathbf{X}_1|\mathbf{K}^T\mathbf{\Psi}^{-1}\mathbf{X}_1$ for all $\mathbf{X}_2$ and $\boldsymbol{\nu}'$. Consequently, $(\mathbf{X}_2, \boldsymbol{\nu}') \perp\!\!\!\perp \mathbf{X}_1|\mathbf{K}^T\mathbf{\Psi}^{-1}\mathbf{X}_1$, which implies that $(\mathbf{X}_2, \boldsymbol{\nu}')|\mathbf{X}_1$ and $(\mathbf{X}_2, \boldsymbol{\nu}')|\mathbf{K}^T\mathbf{\Psi}^{-1}\mathbf{X}_1$ have identical distributions.

The second part of the conclusion will follow if we can show that R is a minimal sufficient statistic for $\mathbf{X}_1|(\mathbf{X}_2, \boldsymbol{\nu}')$. Note that in this treatment the actual unknown parameters $(\bar{\boldsymbol{\mu}}_1, \mathcal{S}_{\mathbf{\Gamma}_0}, \boldsymbol{\beta}_0, \mathcal{S}_{\mathbf{\Gamma}}, \boldsymbol{\nu}', \mathbf{\Psi})$ play no essential role. If $\boldsymbol{\eta}$ is fixed, we look for the function $T(\mathbf{X}_1)$ such that $\mathbf{X}_1|(T(\mathbf{X}_1), \boldsymbol{\eta}) \sim \mathbf{X}_1|T(\mathbf{X}_1)$. If $\boldsymbol{\eta}$ is random, this is equivalent to $\boldsymbol{\eta}|\mathbf{X}_1 \sim \boldsymbol{\eta}|T(\mathbf{X}_1)$. In either the fixed or random $\boldsymbol{\eta}$ cases, we can focus on the statement $\mathbf{X}_1|(T(\mathbf{X}_1), \boldsymbol{\eta}) \sim \mathbf{X}_1|T(\mathbf{X}_1)$. If we think of $\boldsymbol{\eta}$ as the parameter and $\mathbf{X}_1$ as the data, the construction of minimal $T(\mathbf{X}_1)$ is like the construction of a sufficient statistics. So the conclusion will follow if we can show that R is minimal sufficient statistic for $\mathbf{X}_1|\boldsymbol{\eta}$, thinking of $\boldsymbol{\eta}$ as the parameter.

Let $g(\mathbf{x}_1|\mathbf{x}_2, \boldsymbol{\nu}')$ denote the conditional density of $\mathbf{X}_1|(\mathbf{X}_2 = \mathbf{x}_2, \boldsymbol{\nu}' = \boldsymbol{\nu}')$. To show that R is a minimal sufficient statistic for $\mathbf{X}_1|(\mathbf{X}_2, \boldsymbol{\nu}')$ it is sufficient to consider the log likelihood ratio

$$\log\frac{g(\mathbf{z}|\mathbf{x}_2, \boldsymbol{\nu}')}{g(\mathbf{x}_1|\mathbf{x}_2, \boldsymbol{\nu}')} = -\frac{1}{2}\mathbf{z}^T\mathbf{\Psi}^{-1}\mathbf{z} + \frac{1}{2}\mathbf{x}_1^T\mathbf{\Psi}^{-1}\mathbf{x}_1 + (\mathbf{z} - \mathbf{x}_1)^T\mathbf{\Psi}^{-1}E(\mathbf{X}_1|\mathbf{x}_2, \boldsymbol{\nu}').$$

If $\log\frac{g(\mathbf{z}|\mathbf{x}_2, \boldsymbol{\nu}')}{g(\mathbf{x}_1|\mathbf{x}_2, \boldsymbol{\nu}')}$ is to be a constant in $\mathbf{x}_2$ and $\boldsymbol{\nu}'$ then we must have $\log\frac{g(\mathbf{z}|\mathbf{x}_2, \boldsymbol{\nu}')}{g(\mathbf{x}_1|\mathbf{x}_2, \boldsymbol{\nu}')} = E\left\{\log\frac{g(\mathbf{z}|\mathbf{x}_2, \boldsymbol{\nu}')}{g(\mathbf{x}_1|\mathbf{x}_2, \boldsymbol{\nu}')}\right\}$ for all $\mathbf{x}_2$ and $\boldsymbol{\nu}'$. Equivalently, we must have

$$\begin{aligned}&(\mathbf{z} - \mathbf{x}_1)^T\mathbf{\Psi}^{-1}\left(E(\mathbf{X}_1|\mathbf{x}_2, \boldsymbol{\nu}') - E(\mathbf{X}_1|\mathbf{X}_2, \boldsymbol{\nu}')\right)\\ &= (\mathbf{z} - \mathbf{x}_1)^T\mathbf{\Psi}^{-1}(\boldsymbol{\beta}\ \ \mathbf{\Gamma})\begin{pmatrix}\mathbf{x}_2 - \boldsymbol{\mu}_2\\ \boldsymbol{\nu}'\end{pmatrix} = 0.\end{aligned}$$

This result together with the assumption that $\text{var}(\mathbf{X}_2^T, \boldsymbol{\nu}'^T)^T > 0$ implies that $R(\mathbf{X}_1) = (\boldsymbol{\beta}\ \ \mathbf{\Gamma})^T\mathbf{\Psi}^{-1}\mathbf{X}_1$ is a minimal sufficient reduction. □

A.2 Corollary 3.1

Proof. Since $(\boldsymbol{\beta},\ \boldsymbol{\Gamma}) = (\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta} + \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\beta},\ \boldsymbol{\Gamma})$ is equivalent under full rank linear transformations, the span of this is the same as the span of $(\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta} + \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\beta},\ \boldsymbol{\Gamma})\begin{pmatrix} \mathbf{I}_{p_2} & \mathbf{0} \\ -\boldsymbol{\Gamma}^T\boldsymbol{\beta} & \mathbf{I}_d \end{pmatrix} = (\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta},\ \boldsymbol{\Gamma})$. By the setting of $\text{span}(\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}) \equiv \text{span}(\boldsymbol{\Gamma}_0)$, $\text{span}(\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\beta},\ \boldsymbol{\Gamma}) = \text{span}(\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})$ with the corresponding coordinate matrix $\boldsymbol{\beta}_0$.

Assume $\text{rank}(\boldsymbol{\Gamma}_0) = m < p_1 - d$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^{m\times p_2}$ has $\text{rank}(\boldsymbol{\beta}_0) \leq \min(m, p_2)$. Then $\boldsymbol{\Gamma}_0$ is not a completion of $\boldsymbol{\Gamma}$. However, we can think of $\text{span}(\boldsymbol{\Gamma}_0) \subseteq \text{span}(\boldsymbol{\Gamma}_0^*)$ with $\boldsymbol{\Gamma}_0^* = (\boldsymbol{\Gamma}_0,\ \boldsymbol{\Gamma}_{01})$ where $\boldsymbol{\Gamma}_{01}$ is an orthogonal matrix to make $\text{rank}(\boldsymbol{\Gamma}_0^*) = p_1 - d$. Then setting $\boldsymbol{\beta}_0^* = \binom{\boldsymbol{\beta}_0}{\mathbf{0}}$, $\boldsymbol{\Gamma}_0^*\boldsymbol{\beta}_0^* = (\boldsymbol{\Gamma}_0,\ \boldsymbol{\Gamma}_{01})\binom{\boldsymbol{\beta}_0}{\mathbf{0}} = \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0$. So we can always define the appropriate $\boldsymbol{\Gamma}_0^*\boldsymbol{\beta}_0^*$ which is the same as $\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0$ and satisfies $(\boldsymbol{\Gamma}_0^*,\ \boldsymbol{\Gamma}) \in \mathbb{R}^{p_1\times p_1}$. □

A.3 Equation (3.3)

Using vec operation, we can estimate $\hat{\boldsymbol{\beta}}_0$. The partially maximized log likelihood can be written as

$$\begin{aligned}
L_d &= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\|\mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)\right\|^2 \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\text{tr}\left[\sum_{i=1}^{n}\left\|\mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)\right\|^2\right] \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\text{tr}\left[\left\|\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbb{X}_2^T\right\|^2\right] \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\|\text{vec}\left(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbb{X}_2^T\right)\right\|^2 \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\|\text{vec}\left(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T\right) - (\mathbb{X}_2\otimes\boldsymbol{\Gamma}_0)\text{vec}(\boldsymbol{\beta}_0)\right\|^2.
\end{aligned}$$

Based on the least square method,

$$\begin{aligned}
\mathrm{vec}(\hat{\boldsymbol{\beta}}_0) &= \left((\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0)^T(\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0)\right)^{-1}(\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0)^T \mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T) \\
&= \left((\mathbb{X}_2^T \otimes \boldsymbol{\Gamma}_0^T)(\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0)\right)^{-1}(\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0)^T \mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T) \\
&= \left(\mathbb{X}_2^T\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma}_0\right)^{-1}(\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0)^T \mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T) \\
&= \left((\mathbb{X}_2^T\mathbb{X}_2)^{-1} \otimes I\right)(\mathbb{X}_2^T \otimes \boldsymbol{\Gamma}_0^T)\mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T) \\
&= \left((\mathbb{X}_2^T\mathbb{X}_2)^{-1}\mathbb{X}_2^T \otimes \boldsymbol{\Gamma}_0^T\right)\mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T) \\
&= \mathrm{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}).
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}_0 = \boldsymbol{\Gamma}_0^T\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1} = \boldsymbol{\Gamma}_0^T(I - \mathbf{P}_{\boldsymbol{\Gamma}})\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1} = \boldsymbol{\Gamma}_0^T(I - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$ and $\boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma} = 0$, we obtain $\hat{\boldsymbol{\beta}}_0 = {\boldsymbol{\Gamma}_0}^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$.

A.4 Equation (3.4)

Substituting $\hat{\boldsymbol{\beta}}_0 = {\boldsymbol{\Gamma}_0}^T \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1}$ into (3.2), we need to maximize

$$
\begin{aligned}
L_d \;=\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \left\| \mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) - \boldsymbol{\Gamma}_0 {\boldsymbol{\Gamma}_0}^T \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) \right\|^2 \\
= & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \left\| \mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) - \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) \right\|^2 \\
= & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Bigg\{ \operatorname{tr}\left[\sum_{i=1}^{n} (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)^T \mathbf{Q}_{\boldsymbol{\Gamma}} (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) \right] \\
& \qquad - \operatorname{tr}\left[\sum_{i=1}^{n} \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)^T \mathbf{Q}_{\boldsymbol{\Gamma}} \right] \\
& \qquad - \operatorname{tr}\left[\sum_{i=1}^{n} \mathbf{Q}_{\boldsymbol{\Gamma}} (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)^T (\mathbb{X}_2^T \mathbb{X}_2)^{-1} \mathbb{X}_2^T \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \right] \\
& \qquad + \operatorname{tr}\Bigg[\sum_{i=1}^{n} \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)^T \\
& \qquad\qquad (\mathbb{X}_2^T \mathbb{X}_2)^{-1} \mathbb{X}_2^T \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \Bigg] \Bigg\} \\
= & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Big\{ \operatorname{tr}\left[\mathbf{Q}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T \mathbb{X}_1 \right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{Q}_{\boldsymbol{\Gamma}} \right] \\
& \qquad - \operatorname{tr}\left[\mathbf{Q}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \right] + \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \right] \Big\} \\
= & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Big\{ \operatorname{tr}\left[\mathbf{Q}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T \mathbb{X}_1 \right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 (I - \mathbf{P}_{\boldsymbol{\Gamma}}) \right] \\
& \qquad - \operatorname{tr}\left[(I - \mathbf{P}_{\boldsymbol{\Gamma}}) \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \right] + \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \right] \Big\} \\
= & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Big\{ \operatorname{tr}\left[\mathbf{Q}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T \mathbb{X}_1 \right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \right] \\
& \qquad + \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}} \right] - \operatorname{tr}\left[\mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \right] \\
& \qquad + \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \right] + \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \right] \Big\} \\
\\
= & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Big\{ \operatorname{tr}\left[\mathbb{X}_1^T \mathbb{X}_1 \right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T \mathbb{X}_1 \right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1 \right] \Big\} \\
= & -\frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\Big\{ \operatorname{tr}\left[\widehat{\boldsymbol{\Sigma}}_1 \right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\Sigma}}_1 \right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2} \right] \Big\}.
\end{aligned}
$$

A.5 Corollary 4.1

Proof. If $\mathbf{\Omega}^*$ has the special structure $\mathbf{\Omega}^* = \mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$, then

$$\begin{aligned} R(\mathbf{X}) &= \mathbf{\Gamma}^{*T}\mathbf{\Omega}^{*-1}\mathbf{X} \\ &= \mathbf{\Gamma}^{*T}(\mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{X} \\ &= \mathbf{\Gamma}^{*T}(\mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2(\mathbf{\Gamma}^*\mathbf{\Gamma}^{*T} + \mathbf{\Gamma}_0^*\mathbf{\Gamma}_0^{*T}))^{-1}\mathbf{X} \\ &= \mathbf{\Gamma}^{*T}(\mathbf{\Gamma}^*(\mathbf{\Phi}^* + \sigma^2\mathbf{I}_p)\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{\Gamma}_0^*\mathbf{\Gamma}_0^{*T})^{-1}\mathbf{X} \\ &= \mathbf{\Gamma}^{*T}\mathbf{\Gamma}^*(\mathbf{\Phi}^* + \sigma^2\mathbf{I}_p)^{-1}\mathbf{\Gamma}^{*T}\mathbf{X} \\ &= (\mathbf{\Phi}^* + \sigma^2\mathbf{I}_p)^{-1}\mathbf{\Gamma}^{*T}\mathbf{X}, \end{aligned}$$

where $\mathbf{\Gamma}_0^* \in \mathbb{R}^{p\times(p-d)}$ denotes a completion of $\mathbf{\Gamma}^*$; that is, $(\mathbf{\Gamma}_0^*,\ \mathbf{\Gamma}^*) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix. Since $(\mathbf{\Phi}^*+\sigma^2\mathbf{I}_p)^{-1}$ is nonsingular here, $R(\mathbf{X}) = (\mathbf{\Phi}^*+\sigma^2\mathbf{I}_p)^{-1}\mathbf{\Gamma}^{*T}\mathbf{X}$ is equivalent to $R(\mathbf{X}) = \mathbf{\Gamma}^{*T}\mathbf{X}$. □

A.6 Corollary 4.2

Proof. When $\mathbf{\Omega}$ has the special structure $\mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$, the inverse of $\mathbf{\Omega}$ is written as $\mathbf{\Omega}^{-1} = \mathbf{\Gamma}(\mathbf{\Phi} + \sigma^2\mathbf{I}_{p_1})^{-1}\mathbf{\Gamma}^T + \sigma^{-2}\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T$. Then the reduction form can be simplified as

$$\begin{aligned} R(\mathbf{X}_1) &= (\mathbf{\Gamma}_0\boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T\mathbf{\Omega}^{-1}\mathbf{X}_1 \\ &= \begin{pmatrix} \boldsymbol{\beta}_0^T\mathbf{\Gamma}_0^T \\ \mathbf{\Gamma}^T \end{pmatrix} (\mathbf{\Gamma}(\mathbf{\Phi} + \sigma^2\mathbf{I}_{p_1})^{-1}\mathbf{\Gamma}^T + \sigma^{-2}\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T)\mathbf{X}_1 \\ &= \begin{pmatrix} \sigma^{-1}\boldsymbol{\beta}_0^T\mathbf{\Gamma}_0^T \\ (\mathbf{\Phi} + \sigma^2\mathbf{I}_{p_1})^{-1}\mathbf{\Gamma}^T \end{pmatrix} \mathbf{X}_1 \\ &= \underbrace{\begin{pmatrix} \sigma^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{\Phi} + \sigma^2\mathbf{I}_{p_1})^{-1} \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} \boldsymbol{\beta}_0^T\mathbf{\Gamma}_0^T \\ \mathbf{\Gamma}^T \end{pmatrix} \mathbf{X}_1 \\ &= \mathbf{A}(\mathbf{\Gamma}_0\boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T\mathbf{X}_1. \end{aligned}$$

Since $\mathbf{A}$ is nonsingular, $R(\mathbf{X}_1) = (\mathbf{\Gamma}_0\boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T\mathbf{\Omega}^{-1}\mathbf{X}_1$ is equivalent to $R(\mathbf{X}_1) = (\mathbf{\Gamma}_0\boldsymbol{\beta}_0,\ \mathbf{\Gamma})^T\mathbf{X}_1$.

□

A.7 Equation (4.10)

Substituting $\hat{\hat{\boldsymbol{\mu}}}_1 = \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_2$ back in the full log likelihood,

$$\begin{aligned}
L_d &= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\|\mathbf{X}_{1i} - \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\right\|^2 \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathrm{tr}\left[\sum_{i=1}^{n}\left\|\mathbf{X}_{1i} - \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\right\|^2\right] \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathrm{tr}\left[\left\|\mathbb{X}_1^T - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbb{X}_2^T - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{F}^T\right\|^2\right] \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\|\mathrm{vec}\left(\mathbb{X}_1^T - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{F}^T\right) - (\mathbb{X}_2 \otimes \boldsymbol{\Gamma}_0)\,\mathrm{vec}(\boldsymbol{\beta}_0)\right\|^2.
\end{aligned}$$

Following the same procedure as Appendix A.3 and using $\boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma} = 0$, we obtain $\hat{\boldsymbol{\beta}}_0 = {\boldsymbol{\Gamma}_0}^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$.

Substituting $\hat{\boldsymbol{\beta}}_0$ back,

$$\begin{aligned}
L_d &= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big\|\mathbf{X}_{1i} - \bar{\mathbf{X}}_1 \\
&\qquad\qquad - \boldsymbol{\Gamma}_0{\boldsymbol{\Gamma}_0}^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\Big\|^2 \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathrm{tr}\left[\left\|\mathbb{X}_1^T - \boldsymbol{\Gamma}_0{\boldsymbol{\Gamma}_0}^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}\mathbb{X}_2^T - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{F}^T\right\|^2\right] \\
&= -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\|\mathrm{vec}\left(\mathbb{X}_1^T - \boldsymbol{\Gamma}_0{\boldsymbol{\Gamma}_0}^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}\mathbb{X}_2^T\right) - (\mathbf{F} \otimes \boldsymbol{\Gamma})\,\mathrm{vec}(\boldsymbol{\alpha})\right\|^2.
\end{aligned}$$

Following the same procedure as Appendix A.3 and using $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}_0 = 0$, we obtain $\hat{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T\mathbb{X}_1^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$.

Substituting $\hat{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T \mathbb{X}_1^T \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$ into the log likelihood, we need to maximize

$$
\begin{aligned}
L_d \;=\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \big\| \mathbf{X}_{1i} - \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0 {\boldsymbol{\Gamma}_0}^T \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) \\
& \qquad - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T \mathbb{X}_1^T \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1} \mathbf{f}_{y_i} \big\|^2 \\
=\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \big\| \mathbf{X}_{1i} - \bar{\mathbf{X}}_1 - \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) \\
& \qquad - \mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1} \mathbf{f}_{y_i} \big\|^2 \\
=\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Bigg\{ \operatorname{tr}\left[\sum_{i=1}^{n} (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)^T (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) \right] \\
& \qquad - \operatorname{tr}\left[\sum_{i=1}^{n} \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)^T \right] \\
& \qquad - \operatorname{tr}\left[\sum_{i=1}^{n} \mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1} \mathbf{f}_{y_i} (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)^T \right] \\
& \qquad - \operatorname{tr}\left[\sum_{i=1}^{n} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)^T (\mathbb{X}_2^T \mathbb{X}_2)^{-1} \mathbb{X}_2^T \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) \right] \\
& \qquad + \operatorname{tr}\Bigg[\sum_{i=1}^{n} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)^T (\mathbb{X}_2^T \mathbb{X}_2)^{-1} \mathbb{X}_2^T \mathbb{X}_1 \\
& \qquad\qquad \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T \mathbb{X}_2 (\mathbb{X}_2^T \mathbb{X}_2)^{-1} (\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) \Bigg] \\
& \qquad - \operatorname{tr}\left[\sum_{i=1}^{n} {\mathbf{f}_{y_i}}^T (\mathbf{F}^T\mathbf{F})^{-1} \mathbf{F}^T \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}} (\mathbf{X}_{1i} - \bar{\mathbf{X}}_1) \right] \\
& \qquad + \operatorname{tr}\left[\sum_{i=1}^{n} {\mathbf{f}_{y_i}}^T (\mathbf{F}^T\mathbf{F})^{-1} \mathbf{F}^T \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1} \mathbf{f}_{y_i} \right] \Bigg\} \\
=\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Big\{ \operatorname{tr}\left[\mathbb{X}_1^T \mathbb{X}_1\right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1\right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T P_{\mathbf{F}} \mathbb{X}_1\right] \\
& \qquad - \operatorname{tr}\left[P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T\right] + \operatorname{tr}\left[P_{\mathbb{X}_2} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T\right] \\
& \qquad - \operatorname{tr}\left[P_{\mathbf{F}} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T\right] + \operatorname{tr}\left[P_{\mathbf{F}} \mathbb{X}_1 \mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T\right] \Big\} \\
=\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Big\{ \operatorname{tr}\left[\mathbb{X}_1^T \mathbb{X}_1\right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \mathbb{X}_1^T P_{\mathbb{X}_2} \mathbb{X}_1\right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}} \mathbb{X}_1^T P_{\mathbf{F}} \mathbb{X}_1\right] \Big\} \\
=\; & -\frac{np_1}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\Big\{ \operatorname{tr}\left[\widehat{\boldsymbol{\Sigma}}_1\right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0} \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|2}\right] - \operatorname{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\Sigma}}_{\text{fit}}^{1|\mathbf{f}}\right] \Big\}.
\end{aligned}
$$

A.8 Estimation of parameters for the diagonal partial probabilistic PFC model and Equation (4.13)

The MLE of $\bar{\boldsymbol{\mu}}_1$ is straightforward. To see the form of $\widetilde{\boldsymbol{\beta}}_0$ ignore the constant in the likelihood and write

$$\begin{aligned}
L_d &= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i})^T\boldsymbol{\Omega}^{-1} \\
&\qquad\qquad \cdot (\mathbf{X}_{1i} - \bar{\boldsymbol{\mu}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_{2i} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}) \\
&= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\right) \\
&\qquad\qquad \cdot \left(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2) - \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{y_i}\right)^T\Big] \\
&= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1/2}\left(\mathbb{X}_1 - \mathbb{X}_2\boldsymbol{\beta}_0^T\boldsymbol{\Gamma}_0^T - \mathbf{F}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T\right)^T \\
&\qquad\qquad \cdot \left(\mathbb{X}_1 - \mathbb{X}_2\boldsymbol{\beta}_0^T\boldsymbol{\Gamma}_0^T - \mathbf{F}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T\right)\boldsymbol{\Omega}^{-1/2}\Big] \\
&= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\left\|\mathrm{vec}\left(\mathbb{X}_1 - \mathbb{X}_2\boldsymbol{\beta}_0^T\boldsymbol{\Gamma}_0^T - \mathbf{F}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T\right)\boldsymbol{\Omega}^{-1/2}\right\|^2 \\
&= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\left\|\mathrm{vec}\left(\mathbb{X}_1\boldsymbol{\Omega}^{-1/2} - \mathbf{F}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1/2}\right) - \mathrm{vec}\left(\mathbb{X}_2\boldsymbol{\beta}_0^T\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1/2}\right)\right\|^2 \\
&= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\left\|\mathrm{vec}\left(\mathbb{X}_1\boldsymbol{\Omega}^{-1/2} - \mathbf{F}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1/2}\right) - \left(\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}_0 \otimes \mathbb{X}_2\right)\mathrm{vec}\left(\boldsymbol{\beta}_0^T\right)\right\|^2.
\end{aligned}$$

Because $\boldsymbol{\beta}_0$ is unconstrained, this is now just an ordinary least squares problem:

$$\begin{aligned}
\mathrm{vec}\left(\widetilde{\boldsymbol{\beta}}_0^T\right) &= \left(\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0 \otimes \mathbb{X}_2^T\mathbb{X}_2\right)^{-1}(\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1/2} \otimes \mathbb{X}_2^T)\mathrm{vec}\left(\mathbb{X}_1\boldsymbol{\Omega}^{-1/2} - \mathbf{F}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1/2}\right) \\
&= (\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0)^{-1} \otimes (\mathbb{X}_2^T\mathbb{X}_2)^{-1}\mathrm{vec}\left(\mathbb{X}_2^T\mathbb{X}_1\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0 - \mathbb{X}_2^T\mathbf{F}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0\right) \\
&= \mathrm{vec}\left((\mathbb{X}_2^T\mathbb{X}_2)^{-1}\mathbb{X}_2^T\mathbb{X}_1\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0)^{-1}\right).
\end{aligned}$$

Since $\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0 = 0$, the last term in the second equation is 0. Consequently,

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_0^T &= (\mathbb{X}_2^T\mathbb{X}_2)^{-1}\mathbb{X}_2^T\mathbb{X}_1\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0)^{-1} \\
\widetilde{\boldsymbol{\beta}}_0 &= (\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1} \\
&= \boldsymbol{\Gamma}_0^T\mathbf{P}_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})}\widehat{\mathbf{E}}.
\end{aligned}$$

Following the same procedure, we can easily get $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})} \mathbb{X}_1^T \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$. Substituting these estimates we need to maximize

$$\begin{aligned} L_d &= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1}\left(\mathbb{X}_1 - \mathbf{P}_{\mathbb{X}_2}\mathbb{X}_1\mathbf{P}^T_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})} - \mathbf{P}_{\mathbf{F}}\mathbb{X}_1\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\right)^T \\ &\qquad \cdot \left(\mathbb{X}_1 - \mathbf{P}_{\mathbb{X}_2}\mathbb{X}_1\mathbf{P}^T_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})} - \mathbf{P}_{\mathbf{F}}\mathbb{X}_1\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\right)\Big] \\ &= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1}\left(\mathbb{X}_1^T\mathbb{X}_1 - \mathbf{P}_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})}\mathbb{X}_1^T\mathbf{P}_{\mathbb{X}_2}\mathbb{X}_1 - \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\mathbb{X}_1^T\mathbf{P}_{\mathbf{F}}\mathbb{X}_1\right)\Big] \\ &= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1/2}\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{\Omega}^{-1/2} - \boldsymbol{\Omega}^{-1/2}\mathbf{P}_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\boldsymbol{\Omega}^{-1/2} \\ &\qquad - \boldsymbol{\Omega}^{-1/2}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fit}}\boldsymbol{\Omega}^{-1/2}\Big] \end{aligned}$$

Since we can rewrite the projection matrix as

$$\begin{aligned} \mathbf{P}_{\boldsymbol{\Gamma}_0(\boldsymbol{\Omega}^{-1})} &= \boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1} \\ &= \boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1} \\ &= \boldsymbol{\Omega}^{1/2}\mathbf{P}_{\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}_0}\boldsymbol{\Omega}^{-1/2} \\ &= \mathbf{I} - \boldsymbol{\Omega}^{1/2}\mathbf{P}_{\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}}\boldsymbol{\Omega}^{-1/2} \\ &= \mathbf{I} - \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}, \end{aligned}$$

the log likelihood is then

$$\begin{aligned} L_d &= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1/2}\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{\Omega}^{-1/2} - \boldsymbol{\Omega}^{-1/2}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\boldsymbol{\Omega}^{-1/2} \\ &\qquad - \boldsymbol{\Omega}^{-1/2}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\right)\boldsymbol{\Omega}^{-1/2}\Big] \\ &= -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}^{-1/2}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}}\boldsymbol{\Omega}^{-1/2} - \boldsymbol{\Omega}^{-1/2}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\right)\boldsymbol{\Omega}^{-1/2}\Big]. \end{aligned}$$

A.9 Theorem 4.1

Proof. We use f as a generic function whose definition changes and is given in context. We will make a series of changes of variables to rewrite the problem. Let $\mathbf{U} = (\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}})^{1/2}\boldsymbol{\Omega}^{-1}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}})^{1/2}$ so that maximizing (4.14) is equivalent to maximizing

$$f(\mathbf{U}) = \log|\mathbf{U}| - \mathrm{tr}(\mathbf{U}) - \sum_{i=1}^{d}\lambda_i\left(\mathbf{U}(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}})^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fit}}\right)(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{res}})^{-1/2}\right). \qquad (A.1)$$

Let $\tau = \text{rank}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)$ and use the singular value decomposition to write $(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}$. $\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2} = \widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}_\tau\widehat{\mathbf{V}}^T$ where $\widehat{\mathbf{V}} \in \mathbb{R}^{p_1 \times p_1}$ is an orthogonal matrix and $\widehat{\mathbf{\Lambda}}_\tau = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_\tau, 0, \ldots, 0)$, with $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \cdots > \widehat{\lambda}_\tau > 0$. Calling $\mathbf{H} = \widehat{\mathbf{V}}^T\mathbf{U}\widehat{\mathbf{V}} \in \mathbb{S}_{p_1}^+$, (A.1) becomes

$$f(\mathbf{H}) = \log|\mathbf{H}| - \text{tr}(\mathbf{H}) - \sum_{i=1}^{d} \lambda_i\left(\mathbf{H}\widehat{\mathbf{\Lambda}}_\tau\right). \tag{A.2}$$

We now partition $\mathbf{H}$ as $\mathbf{H} = (\mathbf{H}_{ij})$, $i, j = 1, 2$, with $\mathbf{H}_{11} \in \mathbb{S}_\tau^+$, $\mathbf{H}_{22} \in \mathbb{S}_{p_1-\tau}^+$ [for $p_1 = \tau$ we take $\mathbf{H} = \mathbf{H}_{11}$ and go directly to (A.3)]. Consider the transformation $\mathbb{S}_{p_1}^+$ to the space $\mathbb{S}_\tau^+ \times \mathbb{S}_{p_1-\tau}^+ \times \mathbb{R}^{\tau\times(p_1-\tau)}$ given by $\mathbf{V}_{11} = \mathbf{H}_{11}$, $\mathbf{V}_{22} = \mathbf{H}_{22} - \mathbf{H}_{12}^T\mathbf{H}_{11}^{-1}\mathbf{H}_{12}$ and $\mathbf{V}_{12} = \mathbf{H}_{11}^{-1}\mathbf{H}_{12}$. This transformation is one to one and onto (Eaton, 1983, Proposition 5.8). As a function of $\mathbf{V}_{11}$, $\mathbf{V}_{22}$, and $\mathbf{V}_{12}$, (A.2) can be written as

$$\log|\mathbf{V}_{11}||\mathbf{V}_{22}| - \text{tr}(\mathbf{V}_{11}) - \text{tr}(\mathbf{V}_{22}) - \text{tr}(\mathbf{V}_{12}^T\mathbf{V}_{11}\mathbf{V}_{12}) - \sum_{i=1}^{d} \lambda_i\left(\mathbf{V}_{11}\widetilde{\mathbf{\Lambda}}_\tau\right),$$

where $\widetilde{\mathbf{\Lambda}}_\tau = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_\tau)$, and we have used the fact that the nonzero eigenvalues of $\mathbf{H}\widehat{\mathbf{\Lambda}}_\tau$ are the same as those of $\mathbf{H}_{11}\widetilde{\mathbf{\Lambda}}_\tau$. The term $-\text{tr}(\mathbf{V}_{12}^T\mathbf{V}_{11}\mathbf{V}_{12})$ is the only one that depends on $\mathbf{V}_{12}$. Since $\mathbf{V}_{11}$ is positive definite, $\mathbf{V}_{12}^T\mathbf{V}_{11}\mathbf{V}_{12}$ is positive semidefinite. Thus, the maximum occurs when $\mathbf{V}_{12} = 0$. This implies that $\mathbf{H}_{12} = 0$, $\mathbf{H}_{11} = \mathbf{V}_{11}$, $\mathbf{H}_{22} = \mathbf{V}_{22}$, and we next need to maximize

$$f(\mathbf{H}_{11}, \mathbf{H}_{22}) = \log|\mathbf{H}_{11}| + \log|\mathbf{H}_{22}| - \text{tr}(\mathbf{H}_{11}) - \text{tr}(\mathbf{H}_{22}) - \sum_{i=1}^{d} \lambda_i\left(\mathbf{H}_{11}\widetilde{\mathbf{\Lambda}}_\tau\right).$$

This function is maximized over $\mathbf{H}_{22}$ at $\mathbf{H}_{22} = \mathbf{I}_{p_1-\tau}$, then we need to maximize

$$f(\mathbf{H}_{11}) = \log|\mathbf{H}_{11}| - \text{tr}(\mathbf{H}_{11}) - \sum_{i=1}^{d} \lambda_i\left(\mathbf{H}_{11}\widetilde{\mathbf{\Lambda}}_\tau\right). \tag{A.3}$$

Letting $\mathbf{Z} = \widetilde{\mathbf{\Lambda}}_\tau^{1/2}\mathbf{H}_{11}\widetilde{\mathbf{\Lambda}}_\tau^{1/2}$ leads us to maximise $f(\mathbf{Z}) = \log|\mathbf{Z}| - \text{tr}(\mathbf{Z}\widetilde{\mathbf{\Lambda}}_\tau^{-1}) - \sum_{i=1}^{d} \lambda_i(\mathbf{Z})$. Since $\mathbf{Z} \in \mathbb{S}_\tau^+$, there exists an $\mathbf{F} = \text{diag}(f_1, \ldots, f_\tau)$ with $f_i > 0$ in decreasing order and an orthogonal matrix $\mathbf{W}$ in $\mathbb{R}^{\tau\times\tau}$ such that $\mathbf{Z} = \mathbf{W}^T\mathbf{F}\mathbf{W}$. As a function of $\mathbf{W}$ and $\mathbf{F}$,

we can rewrite the function f as

$$\begin{aligned} f(\mathbf{F},\mathbf{W}) &= \log|\mathbf{F}| - \mathrm{tr}(\mathbf{W}^T\mathbf{F}\mathbf{W}\widetilde{\mathbf{\Lambda}}_\tau^{-1}) - \sum_{i=1}^{d} f_i \\ &= \log|\mathbf{F}| - \mathrm{tr}(\mathbf{F}\mathbf{W}\widetilde{\mathbf{\Lambda}}_\tau^{-1}\mathbf{W}^T) - \sum_{i=1}^{d} f_i. \end{aligned}$$

Now, using a lemma from Anderson (1971), Theorem A.4.7, $\min_{\mathbf{W}}\mathrm{tr}(\mathbf{F}\mathbf{W}\widetilde{\mathbf{\Lambda}}_\tau^{-1}\mathbf{W}^T) = \sum_{i=1}^{\tau} f_i\widehat{\lambda}_i^{-1}$, and if the diagonal element of $\mathbf{F}$ and $\widetilde{\mathbf{\Lambda}}_\tau$ are distinct, the minimum occur when $\widehat{\mathbf{W}} = \mathbf{I}_\tau$. Knowing this, we can rewrite the problem on last time, as that of maximizing in $(f_1,\ldots,f_\tau)$, all greater than zero, the function

$$f(f_1,\ldots,f_\tau) = \sum_{i=1}^{\tau}\log|f_i| - \sum_{i=1}^{\tau} f_i\widehat{\lambda}_i^{-1} - \sum_{i=1}^{d} f_i.$$

Clearly the maximum will occur at $f_i = \widehat{\lambda}_i/(\widehat{\lambda}_i+1)$ for $i = 1,\ldots,d$ and for $i = d+1,\ldots,\tau$, $f_i = \widehat{\lambda}_i$. Since $\widehat{\lambda}_i$ are positive and decreasing order, f_i are positive and decreasing in order. Since all the $\widehat{\lambda}_i$ are different, the f_i are different. Collecting all the results, the value of $\mathbf{\Omega}$ that maximizes (4.14) is

$$\begin{aligned} \widehat{\mathbf{\Omega}} &= (\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2})^{1/2}\widehat{\mathbf{U}}^{-1}(\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2})^{1/2} = (\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2})^{1/2}\widehat{\mathbf{V}}\widehat{\mathbf{H}}^{-1}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2})^{1/2} \\ &= (\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2})^{1/2}\widehat{\mathbf{V}}\begin{pmatrix} \widehat{\mathbf{\Lambda}}_\tau^{1/2}\widehat{\mathbf{Z}}^{-1}\widehat{\mathbf{\Lambda}}_\tau^{1/2} & \mathbf{0}_{\tau\times(p_1-\tau)} \\ \mathbf{0}_{(p_1-\tau)\times\tau} & \mathbf{I}_{(p_1-\tau)\times(p_1-\tau)} \end{pmatrix}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2})^{1/2}, \end{aligned}$$

where $\widehat{\mathbf{\Lambda}}_\tau^{1/2}\widehat{\mathbf{Z}}^{-1}\widehat{\mathbf{\Lambda}}_\tau^{1/2} = \mathrm{diag}(\widehat{\lambda}_1+1,\ldots,\widehat{\lambda}_d+1,\mathbf{I}_{\tau-d})$.

Now, to obtain the maximum value we replace $\mathbf{\Omega}$ by $\widehat{\mathbf{\Omega}}$ in (4.14),

$$\begin{aligned} L_d(\widehat{\mathbf{\Omega}}) &= -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log|\widehat{\mathbf{\Omega}}| - \frac{n}{2}\mathrm{tr}\left[\widehat{\mathbf{\Omega}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{1|2}\right] \\ &\quad - \frac{n}{2}\sum_{i=1}^{d}\lambda_i\left(\widehat{\mathbf{\Omega}}^{-1}\left(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}^{1|2}\right)\right). \end{aligned} \tag{A.4}$$

Since the trace and the eigenvalues are cyclic operations,

$$\begin{aligned}\operatorname{tr}(\widehat{\mathbf{\Omega}}^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}) &= \operatorname{tr}(\widehat{\mathbf{\Omega}}^{-1/2}\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}\widehat{\mathbf{\Omega}}^{-1/2}) \\ &= \operatorname{tr}(\widehat{\mathbf{V}}(\mathbf{I}_{p_1}+\widehat{\mathbf{K}})^{-1}\widehat{\mathbf{V}}^T) \\ &= \sum_{i=1}^{d}\frac{1}{\widehat{\lambda}_i+1}+(p_1-d), \end{aligned} \tag{A.5}$$

$$\begin{aligned}\sum_{i=1}^{d}\lambda_i\left\{\widehat{\mathbf{\Omega}}^{-1}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}-\widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right)\right\} &= \sum_{i=1}^{d}\lambda_i\Big\{\widehat{\mathbf{V}}(\mathbf{I}_{p_1}+\widehat{\mathbf{K}})^{-1}\widehat{\mathbf{V}}^T \\ &\qquad \cdot(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1/2}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}-\widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right)(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1/2}\Big\} \\ &= \sum_{i=1}^{d}\lambda_i\Big\{\widehat{\mathbf{V}}(\mathbf{I}_{p_1}+\widehat{\mathbf{K}})^{-1}\widehat{\mathbf{V}}^T\widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}_\tau\widehat{\mathbf{V}}^T\Big\} \\ &= \sum_{i=1}^{d}\lambda_i\Big\{(\mathbf{I}_{p_1}+\widehat{\mathbf{K}})^{-1}\widehat{\mathbf{\Lambda}}_\tau\Big\} \\ &= \sum_{i=1}^{d}\frac{\widehat{\lambda}_i}{\widehat{\lambda}_i+1} \end{aligned} \tag{A.6}$$

Since $\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} > 0$ we have

$$\begin{aligned}\log|\widehat{\mathbf{\Omega}}| &= \log|(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_{p_1}+\widehat{\mathbf{K}})\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{1/2}| \\ &= \log|\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}|+\sum_{i=1}^{d}\log(\widehat{\lambda}_i+1). \end{aligned} \tag{A.7}$$

Plugging (A.5), (A.6) and (A.7) into (A.4) we obtain (4.16). □

A.10 Corollary 4.3

Proof. Recall from Theorem 4.1 that $\widehat{\mathbf{\Omega}} = (\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_{p_1}+\widehat{\mathbf{K}})\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{1/2}$, where $\widehat{\mathbf{V}}$ contains the eigenvectors of $\mathbf{B} = (\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1/2}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}-\widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right)(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1/2}$. The transformation $\mathbf{X}_1 \to \mathbf{A}\mathbf{X}_1$ transforms $\mathbf{B} \to \mathbf{O}\mathbf{B}\mathbf{O}^T$, where $\mathbf{O} = (\mathbf{A}\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}\mathbf{A}^T)^{-1/2}\mathbf{A}\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}}$ is an orthogonal matrix. Consequently, under the transformation $\widehat{\mathbf{K}}$ is invariant, $\widehat{\mathbf{V}} \to \mathbf{O}\widehat{\mathbf{V}}$ and $\widehat{\mathbf{\Omega}} \to \mathbf{A}\widehat{\mathbf{\Omega}}\mathbf{A}^T$. The rest of the proof follows similarly. □

A.11 Corollary 4.4

To prove Corollary 4.4 we need a lemma.

Lemma A.1. *Let* $\widetilde{\mathbf{V}} = (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2}$*, where* $\mathbf{M} = (\mathbf{I}_{p_1} + \widehat{\mathbf{K}})^{-1}$*, with* $\widehat{\mathbf{V}}$ *and* $\widehat{\mathbf{K}}$ *as in Theorem 4.1. Then* $\widehat{\mathbf{\Omega}}^{1/2}\widetilde{\mathbf{V}}$ *are the normalized eigenvectors of* $\widehat{\mathbf{\Omega}}^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)\widehat{\mathbf{\Omega}}^{-1/2}$.

Proof. of Lemma A.1

From Theorem 4.1,

$$\begin{aligned}\widehat{\mathbf{\Omega}} &= \widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2} + (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{1/2}\widehat{\mathbf{V}}\widehat{\mathbf{K}}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{1/2} \\ &= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_{p_1} + \widehat{\mathbf{K}})\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{1/2}.\end{aligned}$$

Then, $\widehat{\mathbf{\Omega}}^{-1} = (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}(\mathbf{I}_{p_1} + \widehat{\mathbf{K}})^{-1}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2} = (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}$.
Using the fact that $\widehat{\mathbf{V}}$ are the eigenvectors of $(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}$ we get

$$\begin{aligned}\widehat{\mathbf{\Omega}}^{-1}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)\widetilde{\mathbf{V}} &= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} \\ &= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{\Lambda}}_\tau\mathbf{M}^{1/2} \\ &= \widetilde{\mathbf{V}}\mathbf{M}^{1/2}\widehat{\mathbf{\Lambda}}_\tau\mathbf{M}^{1/2},\end{aligned}$$

where $\mathbf{M}^{1/2}\widehat{\mathbf{\Lambda}}_\tau\mathbf{M}^{1/2} = \text{diag}\left(\frac{\widehat{\lambda}_1}{\widehat{\lambda}_1+1}, \ldots, \frac{\widehat{\lambda}_d}{\widehat{\lambda}_d+1}, \widehat{\lambda}_{d+1}, \ldots, \widehat{\lambda}_\tau, 0, \ldots, 0\right)$.
Therefore $\widehat{\mathbf{\Omega}}^{-1}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)$ has eigenvalues with eigenvectors $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{V}}^T\widehat{\mathbf{\Omega}}\widetilde{\mathbf{V}} = \mathbf{I}_{p_1}$.

□

Proof. of Corollary 4.4

From the development leading to Theorem 4.1, the MLE of span$(\mathbf{\Omega}^{-1}\mathbf{\Gamma})$ is $\mathcal{S}_d\left(\widehat{\mathbf{\Omega}},\ \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)$, which established the third form. Now, from Lemma A.1, span of the first d columns of $\widehat{\mathbf{\Omega}}^{-1/2}\widehat{\mathbf{\Omega}}^{1/2}\widetilde{\mathbf{V}} = \widetilde{\mathbf{V}}$ is the MLE for span$(\mathbf{\Omega}^{-1}\mathbf{\Gamma})$. Again we use that

$$\begin{aligned}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2} &= \widehat{\mathbf{\Lambda}}_\tau\widehat{\mathbf{V}}^T \\ (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2} &= \widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}_\tau\widehat{\mathbf{V}}^T \\ (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} &= \widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}_\tau\mathbf{M}^{1/2} \\ (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} &= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2}\widehat{\mathbf{\Lambda}}_\tau.\end{aligned}$$

Therefore, span $\left(\widehat{\mathbf{\Omega}}^{-1}(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}})\right) = \text{span}\left((\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2}\right) = \text{span}(\widetilde{\mathbf{V}})$
$= \text{span}\left((\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1}(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}})\right)$. This proves the fourth form. The proof of the fifth form can be found in the following relationship.

$$
\begin{aligned}
&(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}})^{-1}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda l \\
&\iff \left(\widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}\right)^{-1}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda l \\
&\iff \left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda\left(\widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}\right) l \\
&\iff \left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda\left(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} + \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}\right) l \\
&\iff (1+\lambda)\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} l \\
&\iff (\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \frac{\lambda}{1+\lambda} l.
\end{aligned}
$$

Therefore, the eigenvectors of $(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}})^{-1}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right)$ and $(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right)$ are identical, with corresponding eigenvalues $\widehat{\lambda}_i$ and $\widehat{\lambda}_i/(1+\widehat{\lambda}_i)$ because $\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} = \widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}} > 0$, $\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}} = \widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} > 0$, and $\lambda/(1+\lambda)$ is a strictly monotonic function of λ. The corollary follows now from the relation between the eigenvectors of the product of the symmetric matrices $\mathbf{AB}$ and the eigenvectors of $\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}$.

The second and first forms follow from the fourth and third forms with the facts, $\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} = \widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}$ and $\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}} = \widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}}$. For the second form we can use the following equivalence:

$$
\begin{aligned}
&(\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1}\left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda l \\
&\iff \left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} l \\
&\iff \left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}_1 + \widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}^{1|2}_{\text{fit}}\right) l = \lambda\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} l \\
&\iff \left(\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\mathbf{\Sigma}}_1\right) l = (\lambda-1)\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}} l \\
&\iff (\widehat{\mathbf{\Sigma}}^{1|2}_{\text{res}})^{-1}\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\text{res}} l = (1-\lambda) l.
\end{aligned}
$$

The first form can be obtained by the followings: Starting with the fact from Theorem 4.1,

$$
\begin{aligned}
\widehat{\mathbf{\Omega}}^{-1} &= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2} \\
\iff \widehat{\mathbf{\Omega}}^{-1}\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|\mathbf{f}} &= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|\mathbf{f}} \\
\iff \widehat{\mathbf{\Omega}}^{-1}\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|\mathbf{f}}\widetilde{\mathbf{V}} &= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|\mathbf{f}}\widetilde{\mathbf{V}} \\
&= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|\mathbf{f}}(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} \\
&= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} \\
&= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_1 - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2} + \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} \\
&= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2} + \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} \\
&= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2}(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} \\
&\quad - (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)(\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}^{1/2} \\
&= (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\mathbf{M}^{1/2} - (\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|2})^{-1/2}\widehat{\mathbf{V}}\mathbf{M}\widehat{\mathbf{\Lambda}}_\tau\mathbf{M}^{1/2} \\
&= \widetilde{\mathbf{V}}(\mathbf{M} - \mathbf{M}^{1/2}\widehat{\mathbf{\Lambda}}_\tau\mathbf{M}^{1/2}),
\end{aligned}
$$

since $\widehat{\mathbf{\Omega}}^{-1}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)\widetilde{\mathbf{V}} = \widetilde{\mathbf{V}}\mathbf{M}^{1/2}\widehat{\mathbf{\Lambda}}_\tau\mathbf{M}^{1/2}$, the eigenvectors of $\widehat{\mathbf{\Omega}}^{-1}\left(\widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|\mathbf{f}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}^{1|2}\right)$ and $\widehat{\mathbf{\Omega}}^{-1}\widehat{\mathbf{\Sigma}}_{\text{res}}^{1|\mathbf{f}}$ are identical. □

A.12 Equation (4.17)

The full log likelihood for the random combining model can be rewritten as

$$
\begin{aligned}
L_d \;=\; & -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log\left|\mathbf{\Gamma\Phi\Gamma}^T + \sigma^2\mathbf{I}_{p_1}\right| \\
& - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_1 - \bar{\boldsymbol{\mu}}_1 - \mathbf{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 - \mathbf{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y)^T(\mathbf{\Gamma\Phi\Gamma}^T + \sigma^2\mathbf{I}_{p_1})^{-1} \\
& \qquad \cdot (\mathbf{X}_1 - \bar{\boldsymbol{\mu}}_1 - \mathbf{\Gamma}_0\boldsymbol{\beta}_0\mathbf{X}_2 - \mathbf{\Gamma}\boldsymbol{\alpha}\mathbf{f}_y).
\end{aligned}
$$

From this equation, we can easily get $\widehat{\bar{\boldsymbol{\mu}}}_1 = \bar{\mathbf{X}}_1 - \boldsymbol{\Gamma}_0\boldsymbol{\beta}_0\bar{\mathbf{X}}_2$, $\widehat{\boldsymbol{\beta}}_0 = \boldsymbol{\Gamma}_0^T\mathbb{X}_1^T\mathbb{X}_2(\mathbb{X}_2^T\mathbb{X}_2)^{-1}$ and $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\Gamma}^T\mathbb{X}_1^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$. Substituting all these estimates back into the log likelihood,

$$\begin{aligned} L_d &= -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log\left|\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}\right| \\ &\quad - \frac{n}{2}\mathrm{tr}\left[\boldsymbol{\Gamma}(\boldsymbol{\Phi}+\sigma^2\mathbf{I}_{p_1})^{-1}\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right] - \frac{n}{2\sigma^2}\mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right] \\ &\quad + \frac{n}{2}\mathrm{tr}\left[\boldsymbol{\Gamma}(\boldsymbol{\Phi}+\sigma^2\mathbf{I}_{p_1})^{-1}\boldsymbol{\Gamma}^T(\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|\mathbf{f}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}^{1|2})\right] \\ &= -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log\left|\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}\right| \\ &\quad - \frac{n}{2}\mathrm{tr}\left[\boldsymbol{\Gamma}(\boldsymbol{\Phi}+\sigma^2\mathbf{I}_{p_1})^{-1}\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\right] - \frac{n}{2\sigma^2}\mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right]. \end{aligned}$$

In order to find $\widehat{\boldsymbol{\Phi}}$, we just consider two terms related to the $\boldsymbol{\Phi}$; $-\frac{n}{2}\log\left|\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}\right| - \frac{n}{2}\mathrm{tr}\left[\boldsymbol{\Gamma}(\boldsymbol{\Phi}+\sigma^2\mathbf{I})^{-1}\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\right]$. Then we can get $\widehat{\boldsymbol{\Phi}} = \boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\boldsymbol{\Gamma} - \sigma^2\mathbf{I}_d$.

Substituting $\widehat{\boldsymbol{\Phi}}$ back, the partially log likelihood is

$$\begin{aligned} L_d &= -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log\left|\mathbf{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}} - \sigma^2\mathbf{I}_{p1})\mathbf{P}_{\boldsymbol{\Gamma}} + \sigma^2\mathbf{I}_{p1}\right| \\ &\quad - \frac{nd}{2} - \frac{n}{2\sigma^2}\mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right]. \end{aligned}$$

Since

$$\begin{aligned} \log\left|\mathbf{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}} - \sigma^2\mathbf{I}_{p1})\mathbf{P}_{\boldsymbol{\Gamma}} + \sigma^2\mathbf{I}_{p1}\right| &= \log\left|\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\boldsymbol{\Gamma})\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\mathbf{I}_{(p_1-d)}\boldsymbol{\Gamma}_0^T\right| \\ &= \log\left|\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\boldsymbol{\Gamma}\right|\left|\sigma^2\mathbf{I}_{(p_1-d)}\right| \\ &= \log\left|\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\boldsymbol{\Gamma}\right| + \log\left|\sigma^2\mathbf{I}_{(p_1-d)}\right| \\ &= \log\left|\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\boldsymbol{\Gamma}\right| + (p_1-d)\log(\sigma^2), \end{aligned}$$

the partial log likelihood is

$$\begin{aligned} L_d &= -\frac{np_1}{2}\log(2\pi) - \frac{n}{2}\log\left|\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|\mathbf{f}}\boldsymbol{\Gamma}\right| - \frac{n}{2}(p_1-d)\log(\sigma^2) \\ &\quad - \frac{nd}{2} - \frac{n}{2\sigma^2}\mathrm{tr}\left[\mathbf{P}_{\boldsymbol{\Gamma}_0}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right]. \end{aligned}$$

and the corresponding estimate of scale is obtained as $\widehat{\sigma}^2 = \sum_{i=d+1}^{p_1}\lambda_i\left(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1|2}\right)/(p_1-d)$.

A.13 Maximization of σ^2 in the algorithm in Section 5.1.2

In step 2(b), with obtained $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.6). Taking the partial derivative of L_{d^*} with respect to σ^2,

$$\begin{aligned}\frac{\partial L_{d^*}}{\partial \sigma^2} =& - np\sigma^6 - \Big((n-l)d - 2np + l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] + l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{fitk}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right] + (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right]\Big)\sigma^4 \\ &+ 2\Big(\frac{(n-l)d - np}{2} + l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] - l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{fitk}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right] + (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right] - (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right]\Big)\sigma^2 \\ &+ l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] - l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{fitk}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right] + (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right] - (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right] = 0.\end{aligned}$$

We can easily find the one which maximizes log likelihood (5.6) among the obtained 3 solutions which can be minimizer or maximizer of the log likelihood.

A.14 Maximization of σ^2_{k} and σ^2_{u} in the algorithm with isotropic error $\sigma^2_{\mathrm{k}}\mathbf{I}_p$ in Section 5.1.3

In step 2(b), with obtained $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$, find $\widehat{\sigma}^2_{\mathrm{k}(j+1)}$ and $\widehat{\sigma}^2_{\mathrm{u}(j+1)}$ which maximize likelihood (5.8).

$\widehat{\sigma}^2_{\mathrm{k}(j+1)}$ is easily obtained as $\widehat{\sigma}^2_{\mathrm{k}(j+1)} = \left(\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] - \mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{fitk}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right]\right)/p$.

For $\widehat{\sigma}^2_{\mathrm{u}(j+1)}$, taking the partial derivative of L_{d^*} with respect to σ^2_{u},

$$\begin{aligned}\frac{\partial L_{d^*}}{\partial \sigma^2_{\mathrm{u}}} =& - (n-l)p\sigma^6_{\mathrm{u}} + \Big((n-l)(d-2p) + (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right]\Big)\sigma^4_{\mathrm{u}} \\ &- \Big((n-l)(p-d) - 2(n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right] + 2(n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right]\Big)\sigma^2_{\mathrm{u}} \\ &+ (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right] - (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\mathbf{P}_{\mathbf{\Gamma}^*_{(j)}}\right] = 0.\end{aligned}$$

We can find $\widehat{\sigma}^2_{u(j+1)}$ which maximizes log likelihood (5.8) among the obtained 3 solutions from the above cubic equation, which can be minimizer or maximizer of the log likelihood.

A.15 Maximization of σ^2 in the algorithm with specific variance function $\mathbf{\Gamma}^*\mathbf{\Phi}^*\mathbf{\Gamma}^{*T} + \sigma^2\mathbf{I}_p$ in Section 5.1.3

In step 2(b), with obtained $\widehat{\mathcal{S}}_{\mathbf{\Gamma}^*_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.11). Taking the partial derivative of L_{d^*} with respect to σ^2,

$$\begin{aligned}\frac{\partial L_{d^*}}{\partial \sigma^2} =& (ld^* - np)\sigma^6 + \left(\frac{l}{2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] - \frac{l}{2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\mathbf{P}_{\mathbf{\Gamma}^*}\right] - \frac{(n-l)d^*}{2} - n(p-d^*) + \frac{n-l}{2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right]\right)\sigma^4 \\ &+ \left(l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] - l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\mathbf{P}_{\mathbf{\Gamma}^*}\right] - \frac{n(p-d^*)}{2} + (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right] - (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\mathbf{P}_{\mathbf{\Gamma}^*}\right]\right)\sigma^2 \\ &+ \frac{l}{2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\right] - \frac{l}{2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{k}}\mathbf{P}_{\mathbf{\Gamma}^*}\right] + \frac{n-l}{2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\right] - \frac{n-l}{2}\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{\mathrm{u}}\mathbf{P}_{\mathbf{\Gamma}^*}\right] = 0.\end{aligned}$$

We can find the one which maximizes log likelihood (5.11) among the obtained 3 solutions which can be minimizer or maximizer of the log likelihood.

A.16 Maximization of σ^2 in the algorithm in Section 5.2.2

In step 2(b), with obtained $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.17). Taking the partial derivative of L_d with respect to σ^2,

$$\begin{aligned}\frac{\partial L_d}{\partial \sigma^2} =& -np_1\sigma^6 + \Big((n-l)d - 2np_1 + \mathbf{M} + (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{1\mathrm{u}}\mathbf{P}_{\mathbf{\Gamma}}\right]\Big)\sigma^4 \\ &+ \Big((n-l)d - np_1 + 2\mathbf{M}\Big)\sigma^2 + \mathbf{M},\end{aligned}$$

where $\mathbf{M} = l\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{1\mathrm{k}}\right] - \mathrm{tr}\left[\mathbf{A}_{\mathrm{k}}\right] + (n-l)\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{1\mathrm{u}}\right] - \mathrm{tr}\left[\mathbf{A}_{\mathrm{u}}\right] - \mathrm{tr}\left[\left\{l\widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fitk}} - \mathbf{A}_{\mathrm{k}} - \mathbf{A}_{\mathrm{u}} + (n-l)\widehat{\mathbf{\Sigma}}_{1\mathrm{u}}\right\}\mathbf{P}_{\mathbf{\Gamma}}\right]$

We can easily find the one which maximizes log likelihood (5.17) among the obtained 3 solutions which can be minimizer or maximizer of the log likelihood.

A.17 Maximization of σ^2_{k} and σ^2_{u} in the algorithm with isotropic error $\sigma^2_{\mathrm{k}}\mathbf{I}_p$ in Section 5.2.3

In step 2(b), with obtained $\widehat{\mathcal{S}}_{\mathbf{\Gamma}_{(j)}}$, find $\widehat{\sigma}^2_{\mathrm{k}(j+1)}$ and $\widehat{\sigma}^2_{\mathrm{u}(j+1)}$ which maximize likelihood (5.20).

$\widehat{\sigma}^2_{\mathrm{k}(j+1)}$ is easily obtained as $\widehat{\sigma}^2_{\mathrm{k}(j+1)} = \left(\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{1\mathrm{k}}\right] - \mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitk}}\right] + \mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitk}} - \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fitk}})\mathbf{P}_{\boldsymbol{\Gamma}_{(j)}}\right]\right)/p_1$.
For $\widehat{\sigma}^2_{\mathrm{u}(j+1)}$, taking the partial derivative of L_d with respect to σ^2_{u},

$$\begin{aligned}\frac{\partial L_d}{\partial \sigma^2_{\mathrm{u}}} =& -(n-l)p_1\sigma^6_{\mathrm{u}} + \left((n-l)(d-2p_1) + (n-l)\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{u}}\mathbf{P}_{\boldsymbol{\Gamma}_{(j)}}\right] + (n-l)\mathbf{M}\right)\sigma^4_{\mathrm{u}} \\ & -\left((n-l)(p_1-d) - 2(n-l)\mathbf{M}\right)\sigma^2_{\mathrm{u}} + (n-l)\mathbf{M},\end{aligned}$$

where $\mathbf{M} = \mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}}\right] - \mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitu}}\right] + \mathrm{tr}\left[(\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{fitu}} - \widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}})\mathbf{P}_{\boldsymbol{\Gamma}_{(j)}}\right]$. We can find the one which maximizes log likelihood (5.20) among the obtained 3 solutions which can be minimizer or maximizer of the log likelihood.

A.18 Maximization of σ^2 in the algorithm with specific variance function $\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_{p_1}$ in Section 5.2.3

In step 2(b), with obtained $\widehat{\mathcal{S}}_{\boldsymbol{\Gamma}_{(j)}}$, find $\widehat{\sigma}^2_{(j+1)}$ which maximizes likelihood (5.21). Taking the partial derivative of L_d with respect to σ^2,

$$\begin{aligned}\frac{\partial L_d}{\partial \sigma^2} =& -(n(p_1-d) - (n-l)d)\sigma^6 - \left(2n(p_1-d) - (n-l)d - \mathbf{M} - (n-l)\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{1\mathrm{u}}\mathbf{P}_{\boldsymbol{\Gamma}}\right]\right)\sigma^4 \\ & -\left(n(p_1-d) - 2\mathbf{M}\right)\sigma^2 + \mathbf{M} = 0,\end{aligned}$$

where $\mathbf{M} = l\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resk}}\right] + (n-l)\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resu}}\right] - l\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resk}}\mathbf{P}_{\boldsymbol{\Gamma}}\right] - (n-l)\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}^{1|2}_{\mathrm{resu}}\mathbf{P}_{\boldsymbol{\Gamma}}\right]$.

We can find the one which maximizes log likelihood (5.21) among the obtained 3 solutions which can be minimizer or maximizer of the log likelihood.

A.19 Lemma 6.1

Proof. $Y \perp\!\!\!\perp \mathbf{X}_{12}|(\mathbf{X}_{11}, \mathbf{X}_2)$ if and only if $Y \perp\!\!\!\perp \mathbf{X}_1|(\mathbf{X}_{11}, \mathbf{X}_2)$. Suppose that $Y \perp\!\!\!\perp \mathbf{X}_1|(\mathbf{X}_{11}, \mathbf{X}_2)$. We know that $(\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1$ is the minimal sufficient reduction and thus it should not depend on $\mathbf{X}_{12}$. Now, the equation (6.3) from section 6.1 ,

$$\begin{aligned}(\boldsymbol{\Gamma}_0\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma})^T\boldsymbol{\Omega}^{-1}\mathbf{X}_1 =& \left((\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_1)^T\boldsymbol{\Omega}^{11} + (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2)^T\boldsymbol{\Omega}^{21}\right)\mathbf{X}_{11} \\ & + \left((\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_1)^T\boldsymbol{\Omega}^{12} + (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2)^T\boldsymbol{\Omega}^{22}\right)\mathbf{X}_{12}\end{aligned}$$

will not depend on $\mathbf{X}_{12}$ if and only if $(\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_1)^T\boldsymbol{\Omega}^{12} + (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2)^T\boldsymbol{\Omega}^{22} = 0$. Since the following equivalence holds,

$$\begin{aligned}
&(\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_1)^T\boldsymbol{\Omega}^{12} + (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2)^T\boldsymbol{\Omega}^{22} = 0 \\
&\iff (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2) = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}(\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_1) \\
&\iff (\boldsymbol{\Gamma}_{02}\boldsymbol{\beta}_0,\ \boldsymbol{\Gamma}_2) = (-\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_{01}\boldsymbol{\beta}_0,\ -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_1) \\
&\iff \boldsymbol{\Gamma}_{02} = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_{01} \text{ and } \boldsymbol{\Gamma}_2 = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_1,
\end{aligned}$$

we can state that the predictors $\mathbf{X}_{12}$ are inactive if and only if $\boldsymbol{\Gamma}_{02} = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_{01}$ and $\boldsymbol{\Gamma}_2 = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_1$. The reciprocal follows directly if we replace $\boldsymbol{\Gamma}_{02}$ by $-\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_{01}$ and $\boldsymbol{\Gamma}_2$ by $-\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_1$ on equation (6.3). □

A.20 Theorem 6.1

Proof. After maximizing the log likelihood over $(\boldsymbol{\mu}_1,\ \boldsymbol{\beta}_0,\ \boldsymbol{\alpha})$, we need to maximize on $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}^{-1}$, apart from constants,

$$f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^{-1}) \ = \ \log\left|\boldsymbol{\Omega}^{-1}\right| - \operatorname{tr}\left[\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}}\right] + \operatorname{tr}\left[\boldsymbol{\Omega}^{-1}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Omega}^{-1})}\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit}}\right)\right].$$

From the hypotheses on $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_{02} = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_{01}$ and $\boldsymbol{\Gamma}_2 = -\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}\boldsymbol{\Gamma}_1$, we have $\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1} = (\boldsymbol{\Gamma}_1^T\boldsymbol{\Omega}^{11.2},\ \mathbf{0})$ where $\boldsymbol{\Omega}^{11.2} = \boldsymbol{\Omega}^{11} - \boldsymbol{\Omega}^{12}\boldsymbol{\Omega}^{-22}\boldsymbol{\Omega}^{21}$. Then $\boldsymbol{\Gamma}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1^T\boldsymbol{\Omega}^{11.2}\boldsymbol{\Gamma}_1$. The partially maximized log likelihood can be expressed as

$$\begin{aligned}
f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^{-1}) \ &= \log\left|\boldsymbol{\Omega}^{-1}\right| - \operatorname{tr}\left[\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}}\right] \\
&\quad + \operatorname{tr}\left[\mathbf{P}_{(\boldsymbol{\Omega}^{11.2})^{1/2}\boldsymbol{\Gamma}_1}(\boldsymbol{\Omega}^{11.2})^{1/2}\left(\widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit}} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit}}\right)(\boldsymbol{\Omega}^{11.2})^{1/2}\right].
\end{aligned}$$

For fixed $\boldsymbol{\Omega}$, the last term is maximized by choosing $(\boldsymbol{\Omega}^{11.2})^{1/2}\boldsymbol{\Gamma}_1$ to be a basis for the span the first d eigenvectors of $(\boldsymbol{\Omega}^{11.2})^{1/2}\mathbf{B}_{\text{fit},11}(\boldsymbol{\Omega}^{11.2})^{1/2}$ where $\mathbf{B}_{\text{fit},11} = \widehat{\boldsymbol{\Sigma}}^{1|\mathbf{f}}_{\text{fit},11} - \widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{fit},11}$, yielding another partially maximized log likelihood

$$f(\boldsymbol{\Omega}^{-1}) \ = \ \log\left|\boldsymbol{\Omega}^{-1}\right| - \operatorname{tr}\left[\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Sigma}}^{1|2}_{\text{res}}\right] + \sum_{i=1}^{d}\lambda_i\left((\boldsymbol{\Omega}^{11.2})^{1/2}\mathbf{B}_{\text{fit},11}(\boldsymbol{\Omega}^{11.2})^{1/2}\right).$$

Let us take the one-to-one and onto transformation defined by $\mathbf{L}_{11} = \mathbf{\Omega}^{11} - \mathbf{\Omega}^{12}\mathbf{\Omega}^{-22}\mathbf{\Omega}^{21}$, $\mathbf{L}_{22} = \mathbf{\Omega}^{22}$ and $\mathbf{L}_{12} = \mathbf{\Omega}^{12}\mathbf{\Omega}^{-22}$. As a function of $\mathbf{L}_{11},\ \mathbf{L}_{22},\ \mathbf{L}_{12}$ we get

$$\begin{aligned} f(\mathbf{L}_{11}, \mathbf{L}_{22}, \mathbf{L}_{12}) &= \log|\mathbf{L}_{11}| + \log|\mathbf{L}_{22}| \\ &\quad - \mathrm{tr}\Big[(\mathbf{L}_{11} + \mathbf{L}_{12}\mathbf{L}_{22}\mathbf{L}_{12}^T)\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11} + \mathbf{L}_{12}\mathbf{L}_{22}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},21}\Big] \\ &\quad - \mathrm{tr}\Big[\mathbf{L}_{22}\mathbf{L}_{12}^T\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12} + \mathbf{L}_{22}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},22}\Big] + \sum_{i=1}^{d}\lambda_i\left(\mathbf{L}_{11}^{1/2}\mathbf{B}_{\mathrm{fit},11}\mathbf{L}_{11}^{1/2}\right). \end{aligned}$$

Now, differentiating with respect to $\mathbf{L}_{12}$ in the last expression, we get that

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{L}_{12}} &= -2\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}\mathbf{L}_{22} - 2\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11}\mathbf{L}_{12}\mathbf{L}_{22} \quad \text{and} \\ \frac{\partial^2 f}{\partial \mathbf{L}_{12}^2} &= -2\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11} \otimes \mathbf{L}_{22}. \end{aligned}$$

Therefore the maximum occurs when $\mathbf{L}_{12} = -(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}$. Replacing this in the last log likelihood function we next need to maximize

$$\begin{aligned} f(\mathbf{L}_{11}, \mathbf{L}_{22}) &= \log|\mathbf{L}_{11}| + \log|\mathbf{L}_{22}| - \mathrm{tr}\Big[\mathbf{L}_{22}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},22}\Big] - \mathrm{tr}\Big[\mathbf{L}_{11}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11}\Big] \\ &\quad + \mathrm{tr}\Big[\mathbf{L}_{22}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}\Big] + \sum_{i=1}^{d}\lambda_i\left(\mathbf{L}_{11}^{1/2}\mathbf{B}_{\mathrm{fit},11}\mathbf{L}_{11}^{1/2}\right), \end{aligned}$$

since for $\mathbf{L}_{12} = -(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}$, $-2\mathrm{tr}\Big[\mathbf{L}_{22}\mathbf{L}_{12}^T\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}\Big] - \mathrm{tr}\Big[\mathbf{L}_{12}\mathbf{L}_{22}\mathbf{L}_{12}^T\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11}\Big] = \mathrm{tr}\Big[\mathbf{L}_{22}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}\Big]$. The maximum on $\mathbf{L}_{22}$ is at $\mathbf{L}_{22} = (\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},22.1})^{-1}$ so that we need to maximize on $\mathbf{L}_{11}$

$$f(\mathbf{L}_{11}) = \log|\mathbf{L}_{11}| - \mathrm{tr}\Big[\mathbf{L}_{11}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11}\Big] - \sum_{i=1}^{d}\lambda_i\left(\mathbf{L}_{11}^{1/2}\mathbf{B}_{\mathrm{fit},11}\mathbf{L}_{11}^{1/2}\right),$$

where $\mathbf{B}_{\mathrm{fit},11} = \widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{fit},11} - \widehat{\mathbf{\Sigma}}^{1|\mathbf{f}}_{\mathrm{fit},11}$. From Theorem 4.1 the MLE for $\mathbf{L}_{11}^{-1}$ is $(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_d + \widehat{\mathbf{K}})\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{1/2}$, where $\widehat{\mathbf{K}}$ and $\widehat{\mathbf{V}}^T$ are as defined in Theorem 6.1. Since $\mathbf{L}_{11} = \mathbf{\Omega}^{11} - \mathbf{\Omega}^{12}\mathbf{\Omega}^{-22}\mathbf{\Omega}^{21} = \mathbf{\Omega}_{11}^{-1}$ it follows that $\widehat{\mathbf{\Omega}}_{11} = (\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_d + \widehat{\mathbf{K}})\widehat{\mathbf{V}}^T(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{1/2}$. The MLE for $\mathbf{\Omega}^{22}$ is $(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},22.1})^{-1}$ and $\mathbf{\Omega}^{12}$ is $-(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},22.1})^{-1}$. Consequently, $\widehat{\mathbf{\Omega}}_{12} = \widehat{\mathbf{\Omega}}_{11}(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}$ and $\widehat{\mathbf{\Omega}}_{22} = \widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},22.1} + \widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},21}(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Omega}}_{11}(\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},11})^{-1}\widehat{\mathbf{\Sigma}}^{1|2}_{\mathrm{res},12}$.

The MLE for the span of $\mathbf{\Omega}^{-1}\mathbf{\Gamma} = (\mathbf{\Omega}^{11.2}\mathbf{\Gamma}_1,\ \mathbf{0})^T$ is the span of $(\widehat{\mathbf{\Omega}}_{11}^{-1/2}\widehat{\mathbf{\Gamma}}_1,\ \mathbf{0})^T$ with $\widehat{\mathbf{\Gamma}}_1$ the first d eigenvectors of $\widehat{\mathbf{\Omega}}_{11}^{-1/2}\mathbf{B}_{\text{fit},11}\widehat{\mathbf{\Omega}}_{11}^{-1/2}$. Using the logic of Corollary 4.4 it can be proved that the MLE of span$(\mathbf{\Omega}^{-1}\mathbf{\Gamma})$ is in this case the span of $((\widehat{\mathbf{\Sigma}}_{\text{res},11}^{1|2})^{-1/2}\widehat{\mathbf{\Gamma}}_1,\ \mathbf{0})^T$, with $\widehat{\mathbf{\Gamma}}_1$ the first d eigenvectors of $(\widehat{\mathbf{\Sigma}}_{\text{res},11}^{1|2})^{-1/2}\mathbf{B}_{\text{fit},11}(\widehat{\mathbf{\Sigma}}_{\text{res},11}^{1|2})^{-1/2}$.

The proof of (6.4) can be done essentially as the proof of (4.16). □