Exploratory Studies For Gaussian Process Structural Equation Models

Yida Chiu

Department of Statistical Science
University College London
1-19 Torrington Place
London WC1E 7HB, United Kingdom

A dissertation submitted for
the degree of
Doctor of Philosophy

2014

I, Yida Chiu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Latent variable models (LVMs) are widely used in many scientific fields due to the ubiquitousness and feasibility of latent variables. Conventional LVMs, however, have limitations because they model relationships between covariates and latent variables or among latent variables with a parametric fashion. A more flexible model framework is therefore needed, especially without prior knowledge of sensible parametric forms.

This thesis proposes a new non-parametric LVM for the need. We define a model structure with particular features, including a multi-layered structure constituting of non-parametric Gaussian Processes regression and parametric factor analysis. The connections to existing popular LVMs approaches, such as structural equation models and latent curve models, are also discussed. The model structure is subsequently extended for observed binary responses and longitudinal application. It follows that model identifiability is examined through parameter constraints and algebraic manipulations.

The proposed model, despite convenient applicability, has a computational burden for analysing large data sets due to the computation of the inverse of a large covariance matrix. To address the issue, a sparse approximation method using a small number of M selected inputs (inducing inputs) is adopted. The associated computational cost can be reduced to $O(M^2NQ^2)$ (or $O(M^2NT^2)$) where N and Q are the numbers of data points and latent variables (or time points T), respectively.

Inference within this framework requires a series of algorithmic developments in a Bayesian paradigm. The algorithms, using Markov Chain Monte Carlo sampling-based methods and Expectation Maximisation optimisation methods with stochastic variant, are presented. A hybrid estimation procedure with two-step implementations is proposed as well, which can further reduce computational cost. Furthermore, a greedy selection scheme for inducing inputs is provided for better model predictive performance.

Empirical studies of the modelling framework are conducted for various experiments. Interest lies in inference, including parameter estimation and realization of distribution of latent variables; and assessments and comparisons of predictive performance with two baseline techniques. Discussion and suggestions for improvement are provided based on results.

Acknowledgments

This 4-year PhD life journey is full with adventure and challenge. It helps me to know myself and look the matters more deeply, especially when i confront difficulties. Without many people's aid, those obstacles can not be conquered and the study is unable to end. I am sincerely grateful to the people who kindly give me great support and good company during this journey. I will always remember their names and place the warm memory in a special place of my heart. I wish they live happily and enjoy every moment. All the best to their continuing life journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Development of Motivation

The estimation of latent factors explaining observed phenomena is one of the main goals in behavioural, educational, medical, and social sciences. The task usually can proceed by statistically examining whether or not reasonability can exist in a hypothesis – data from observed phenomena is generated by certain variables through a pre-assumed modelling mechanism. Latent variable models (LVMs) serve for the examination while the involving posited variables are unobserved or unmeasured directly in reality. Such unobserved variables are referred as latent (or hidden) variables (or constructs) in different disciplines. They also entitle LVMs as a conceptual framework to explore underlying factors along with possible information condensation, and their inter-relations with (multiple) observed variables.

Abundant literature about LVMs has been published and develops various model formulations and applications. For inter-relation of latent and observed variables, LVMs can be commonly linear or non-linear (Arminger & Muthén 1998); for model structure, LVMs can be non-directed (e.g., Markov random field with latent variable (Everitt 2012)) or directed (e.g., Hidden Markov Field (Ghahramani 2001), Autoregressive Latent Trajectory Model (Bollen 2006)). Conventionally, depending on characteristics of latent and observed variables, there are four well-studied development categories for LVMs with multiple observed variables (Bartholomew & Knott 1999). They are latent class analysis (categorical[1]

[1] Categorical variables contain nomial or ordinal variables.

latent variables and observed variables), latent trait analysis (metrical[2] latent variables and categorical observed variables), latent profile analysis (categorical latent variables and metrical observed variables), factor analysis (metrical latent variables and observed variables). Various types of observed variables and complex data structure motivate model development as well (Skrondal & Rabe-Hesketh 2004, Lee 2007). Recently, topic models have attracted a growing attention because of applications on a variety of response types in text document to search probability distribution of underlying latent variables (vocabularies describing themes) (Blei 2012).

Structure Equation Modelling (SEM) built on factor analysis (FA) is originally a linear, directed statistical modelling approach. Due to inheritance of FA (close to principal component analysis (PCA)), SEM can also be regarded as a dimension-reduction technique. It also allows inference on cause-effect relationships on two formulations – between many observed variables and few latent variables; and among latent variables. SEM enables incorporating additional variables into model structure, and this thereby increases its accountability for data. The connection between "covariates"[3] and latent variables also brings interest in their dependant latent structure, as well as distributions and patterns in latent-variable space. For example, multiple indicators multiple causes (MIMIC) model and latent curve model (LCM) (Bollen 2006), in the context of SEM, adopt a linear directed effect on latent variables from covariates. Moreover, in the former framework latent variables conventionally serves as random variables represented a posited concept, measured indirectly by observed variables (or indicators); in the latter, latent variables can work as modelling temporal change of observed variables, or of conventional latent variables.

Instead of using linear or non-linear direct sum on latent variables, utilization of non-parametric frameworks to model the functional relations can be appealing. This consideration may naturally emerge because practitioners need to deliberate what kind of parametric regression functional form is sensible. That especially rises in the case of modelling the change of latent variables and no prior knowledge about the growth pattern.

Various non-parametric methods for modelling regression relationship have been provided, such as kernel smoothing estimator, spline method, wavelet regression (Wasserman

[2] Metrical variables contain discrete or continuous variables.

[3] Covariates here can be response variables, which depend on practitioners research interests and hypotheses.

2006). But, there is only limited literature in the field of non-parametric SEM. The Bayesian P-spline, a Bayesian technique to penalized splines, is used to model the functional relationships between observed covariates and latent variables, and among latent variables (Fahrmeir & Raach 2007, Song & Lu 2010). These works, however, are based on the assumption that the effects of covariates and latent variables are additive. The functional relations there are expressed as univariate non-linear functions. By contrast, a Gaussian Processes (GPs) framework is utilised to model multivariate non-parametric function between latent variables (Silva & Gramacy 2010).

The GP non-parametric regression approach is emphasized in our works. It has been developed in spatial statistics over decades and become popular in machine learning field recently. Its convenience and natural incorporation of the Bayesian inferential work[4] allows one to automatically learn an appropriate posterior functional relation to fit data. It possibly has no need in undertaking trial-and-error experiments to tune the parameters controlling regression function.

Although GP provides great flexibility in modelling functional relation, computational expense is high in a large dataset. This increases cubically with dataset size, and the whole computing process can be extremely lengthy. Several GP researchers have worked on this issue (Rasmussen & Williams 2006, Seeger et al. 2003, Snelson & Ghahramani 2006*a*, Quinonero-Candela & Rasmussen 2005). Their ideas are based on a conditional distribution of GP function values given other variables under certain assumptions. Those given variables comprise of another set of GP function values evaluated at a finite number of inducing (or pseudo) inputs. The notions can be the sources to develop an efficient algorithm for the posited frameworks.

Model estimation can be fulfilled by several methods. One of the possible approaches is Markov Chain Monte Carlo (MCMC) methods, which has widely spread in statistics community and produced diverse applications in many disciplines, especially since Gibbs' sampling method was invented (Geman & Geman 1984). Due to the Bayesian feature of GP frameworks, the application of MCMC methods can be implemented. In fact, many authors have used this treatment on similar LVM frameworks (Lee 2007, Titsias &

[4] It means any modelling frameworks applying Bayesian rule – the posterior distribution of a hypothesis equals the quotient of the product of its prior distribution and data likelihood (formulated from posited model structure and distributional assumptions), and marginal likelihood (a constant factor for any hypotheses being considered).

Lawrence 2010). The Expectation Maximization (EM) method (Dempster et al. 1977) is another approach developed over decades. It addresses estimation problems with missing values and treats missing values as latent variables. The feasibility of this approach for many LVM frameworks has been identified. Essentially, it is a deterministic optimisation method, but several variants, including introducing sampling schemes, like MCMC techniques, can be applied in the case that distributions of latent variables are complicated. In addition, upon the multiple-output model structure, another estimation method, inspired by inference function of margin (IFM) (Joe & Xu 1996), can be developed to facilitate computing. It may bring more computational advantages than solely using the first two methods on the whole model.

Because of the multiple-output modelling structure, multiple-response prediction problem is of interest. This problem has been considered in geostatistics and machine learning research. For example, mining at several locations for various kinds of ore can be expensive, especially when some mines are difficult to detect. The issue may be alleviated via the correlations with other types of mine to achieve good prediction for locations. Instead of using independent prediction for each output, various methods (Teh et al. 2005, Álvarez et al. 2011, Bonilla et al. 2008) have been provided and attempt to capture the correlation between different output as conventional multiple-output regression does. It would be interesting to assess the predictive performance of the proposed modelling work, compared with those of baseline methods – using least-squared method, or GP regression framework independently.

Due to an analogy in model frameworks, potential model extension to the case of binary variables and applications on longitudinal studies may be taken into account as well.

1.2 Contributions, Goal and Scope

Our model frameworks make multi-fold contributions from different viewpoints. We expand the field of non-parametric SEM by using GP regression in the research direction of modelling functional relations between covariates and latent variables. Two features also distinguish between our model frameworks and the state-of-the-art multi-output GP regression. The first difference is that the outputs can be latent variables regressed on covariates. The second is that we allow the latent variables given covariates to be dependent according to arbitrary covariance structure. Due to non-parametric feature, our model can be considered as a generalisation of LCM for longitudinal analysis as well.

Our frameworks also follow a series of algorithmic developments and the exploratory works for model assessment. Note that the sampling estimation algorithm upon the MCMC approaches in Section 4.3 is particularly novel. The model assessments on empirical studies can be useful reference for future practitioners and the non-parametric LVM researchers.

The goal of this thesis is to accomplish exploratory works in computation, estimation, prediction, extension and application for the new modelling methodology - Gaussian Process Structure Equation Modelling (GP-SEM). To achieve that, we specify the associated tasks and the works having been done:

1. **Computation / Estimation**:

   - developed computational efficient algorithms via the MCMC and EM methods, and a hybrid scheme consisting of the EM and IFM approach.
   - modified a greedy selection scheme for inducing inputs.
   - examined model identification conditions and convergence diagnosis of parameters.
   - discovered the relation of parameters and latent variables between before and after data processing of standardization.
   - explored the estimation differences of the proposed algorithms on parameters and latent variables.

2. **Prediction**:

   - compared model predictive performance between the baseline frameworks (using linear regression and GP regression) and the proposed GP-SEM (with two algorithms).
   - contrasted the GP-SEM predictive performance under the different scenarios, including given two selection schemes for inducing inputs, as well as varying the number of pseudo inputs and latent variables
   - conducted posterior predictive checking to realize model appropriateness.

  - learnt the regression functional relationship between a covariate and a latent variable.

3. **Extension / Application**:

  - extended model frameworks to binary responses
  - carried out application on longitudinal studies.

We confine the scope of this thesis to continuous and binary response variables with continuous latent variables, continuous and categorical covariates. It is incidental that the latent variables mentioned in the thesis, depending on the context, only mean hypothetical constructs, or responses underlying binary variables, or missing values or unobserved heterogeneity.

1.3 Thesis Structure

The remainder of the thesis is at follows: Chapter 2 provides the background literature for our modelling framework. Chapter 3 illustrates the proposed model structure and examines identification problems. Chapter 4 presents our estimation methods including introduction for the MCMC methods, EM approaches and the associated limited-information algorithms, and their implementations. It also provides the computing procedures for prediction. Chapter 5 shows the experiment results for three studies to evaluate the performance in estimation, prediction and computation. Chapter 6 accentuates the case of longitudinal study with continuous and binary responses. It has tight connection with the preceding chapters, including model specification and estimation methods. Another three empirical studies are carried out and presented as well. Chapter 7 summarises our works, discusses possible improvement and future work. The relevant technical details and proofs are provided in the appendix.

Chapter 2

Background

This chapter provides the foundation of our unifying modelling frameworks which follows in the subsequent chapters. Here, we provide several related literature to characterise the main ideas of existing methods.

In Section 3.1, we first discuss the latent variable model - principal component analysis, factor analysis, structure equation modelling in order and point out their association. In Section 3.2, we introduce the definition of Gaussian Processes and its function space view for the regression problem. In Section 3.3 and 3.4, we respectively review several sparse approximation methods and multiple-outcome prediction methods for Gaussian Process models. Both proceed under existing general modelling frameworks. We conclude by summarising the whole chapter.

2.1 Latent Variable Models

2.1.1 Principal Component Analysis

Principal component analysis (PCA) is a classical dimension-reduction technique for a dataset in which moderate or high intercorrelation exists among variables. The prime objective is to substitute R continuous correlated variables for Q uncorrelated variables ($Q < R$) that explain more variance in the data through linear transformations (Jolliffe 2002). It therefore allows practitioners to learn a governing pattern from a high-dimensional data set.

The main idea is founded on the explainability of the new Q uncorrelated variables (called principle components) for the total variance (the sum of the variances of the original

R variables). More clearly, the accountability can be defined as the proportion sum of variances of the new variables over total variance. Alternatively, the central idea can also be based on the mean squared error between the original variables and the inverse image of the new variables (Jolliffe 2002).

Let $\mathbf{Y}^* = [\mathbf{y}_1^*, \ldots, \mathbf{y}_R^*]^\mathsf{T}$ be a $R \times N$ matrix which consists of N data points centred at $\mathbf{0}$. To find principle components, one can first obtain the singular value decomposition (SVD) of $\mathbf{Y}^*$. Let the SVD be $\mathbf{P}\boldsymbol{\Psi}\mathbf{R}^\mathsf{T}$, where $\boldsymbol{\Psi}$ is a $R \times N$ matrix with singular values (assigned in descending order) on the main diagonal; $\mathbf{P}$ and $\mathbf{R}$ are a $R \times R$ and a $N \times N$ matrix whose columns are arranged according to the descending order of singular values and are the orthonormal eigenvectors of $\mathbf{Y}^*\mathbf{Y}^{*\mathsf{T}}$ and of $\mathbf{Y}^{*\mathsf{T}}\mathbf{Y}^*$, respectively. The practitioners can select the Q largest singular values and the corresponding vectors to form the principle components and the linear transformation. Precisely,

$$\mathbf{Y}^* \approx \mathbf{P}_Q\mathbf{X}_Q, \tag{2.1}$$

where $\mathbf{P}_Q$ is a $R \times Q$ component loading matrix consisting of the first Q columns of $\mathbf{P}$; moreover, $\mathbf{P}_Q^\mathsf{T}$ transforms the original variables $y_1, y_2, \ldots, y_R$ to components $x_1, x_2, \ldots, x_Q$. $\mathbf{X}_Q$ is a $Q \times N$ matrix comprising the first Q rows of $\boldsymbol{\Psi}\mathbf{R}^\mathsf{T}$ which denotes the principal component scores of N data points. A principal component can be interpreted as a synthesis index of the original variables, or something else based on the component loadings. $\mathbf{P}_Q\mathbf{X}_Q$ additionally means the inverse image of $\mathbf{X}_Q$ through $\mathbf{P}_Q$ or approximate values of the original variables. In the above context, the PCA solution actually maximises the explainability of the components, $\frac{\sum_{i=1}^{Q}\phi_i}{\sum_{i=1}^{R}\phi_i}$ (ϕ_i is the variance of the i-th component); that equivalently minimises the error $||\mathbf{Y}^* - \mathbf{P}_Q\mathbf{X}_Q||^2$ (Jolliffe 2002). Note that there are infinite solutions for $\mathbf{P}_Q$ and $\mathbf{X}_Q$ because the SVD of $\mathbf{Y}^*$, $\mathbf{P}\boldsymbol{\Psi}\mathbf{R}^\mathsf{T}$ can be represented as $\mathbf{P}\mathbf{O}\mathbf{O}^{-1}\boldsymbol{\Psi}\mathbf{R}^\mathsf{T}$ using a non-singular matrix $\mathbf{O}$, which makes $\mathbf{X}_Q$ a unit covariance matrix.

There are several criteria to decide the number Q of principle components (Bartholomew et al. 2008). For example, one can select Q components whose sum of variance can account for a large proportion of the total variance (around 70-80 percentages,) or whose corresponding singular values starts decreasing abruptly. In addition to using K-fold cross-validation (CV), the number Q can be learnt based on the overall predictive performance over each test set by estimation of $\mathbf{P}_Q$ and $\mathbf{X}_Q$ upon the corresponding training set (Jolliffe 2002). The number Q also sometimes depends on whether the components have sensible and usual interpretation from the researcher's knowledge.

2.1.2 Factor Analysis

Factor analysis (FA), similar to PCA, is another dimension-reduction method. Its dimension-reduced characteristic, however, is induced from intercorrelation under the assumption that some observed variables (manifest variables or indicators) depend on the same unobserved variables (latent variables or factors) by a modelling mechanism. This model-based orientation can therefore distinguish FA from PCA (Bartholomew et al. 2008).

Closely related to regression analysis, latent variables in FA play the role of explanatory variables and account for the correlation among observed variables. Their regression relationship can be the basis for exploring the underlying patterns within data or testing causal hypotheses between observed and latent variables. Moreover, like component loadings in PCA, factor loadings represent the influence of latent variables on the observed ones in regression relationship. A noticeable difference is that FA also intends to discover inverse regression relationship given observed variables in order to learn distributions over latent variables (Bartholomew et al. 2008).

Classical FA is sometimes classified as exploratory or confirmatory (Lee 2007, Bollen 1989, Bartholomew et al. 2008). For exploring and understanding measured data, the number of latent variables is not fixed and all manifest variables are linked to all the latent variables (that means there are no constraints on factor loadings). If factor loadings are high, one can name the latent variables based on the common features of the corresponding observed variables, which may enable to be hypothesized as measured indicators of the factor. In this case, it is referred to as exploratory factor analysis (EFA). For hypothesis testing and theory development, the number of the latent variables is fixed and manifest variables are linked to a subset of latent variables only. That implies some of the factor loadings are zeros, which also reflects a research hypothesis. One can use goodness of fit test to justify whether the posited model structure is reasonable. In the context, it is called as confirmatory factor analysis (CFA). Bartholomew et al.(2008) point out that in practical applications the distinction between EFA and CFA is not absolute because researchers may adopt mixed strategies.

EFA and CFA also have some analogies in model assumptions. Measurement errors are typically assumed: (1) mutually uncorrelated with each other, (2) expected values of zero, (3) identical (or non-identical) variances and (4) uncorrelated with the latent variables. For EFA, latent variables can also be assumed to hold the assumptions (1)-(3)

although (1) could be relaxed depending on whether transformed factor loadings through oblique rotation can be interpreted sensibly or not. For CFA, latent variables are usually assumed correlated whereas constraints are sometimes imposed on correlation coefficients for necessity and hypothesis. FA typically adopts a normality assumption on measurement errors and latent variables.

The algebraic representation of FA model framework is given by:

$$\mathbf{Y} = \boldsymbol{\lambda}_{0,y} \otimes \mathbf{1}_N^{\mathsf{T}} + \boldsymbol{\Lambda}_y \mathbf{X} + \mathbf{E}_y, \tag{2.2}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_R]^{\mathsf{T}}$ is a $R \times N$ matrix which consists of N data points with R observed values each; $\boldsymbol{\lambda}_{0,y}$ and $\mathbf{1}_N$ are a $R \times 1$ intercept vector and a $N \times 1$ vector with all entries being 1, respectively. With Q latent variables, $\boldsymbol{\Lambda}_y$ is a $R \times Q$ factor loading matrix, $\mathbf{X}$ is a $Q \times N$ factor score matrix for all data points. $\mathbf{E}_y$ is a $R \times N$ matrix of measurement errors.

This equation relates to (2.1) through two procedures. At first, if $\mathbf{Y}$ is centered on its mean, then it would nullify the intercept term and lead $\mathbf{Y}$ to $\mathbf{Y}^*$ in (2.1). Then removing the error terms $\mathbf{E}_y$ causes the remaining term $\boldsymbol{\Lambda}_y\mathbf{X}$ to be the term in the right-hand side of (2.1), which means an approximate value of $\mathbf{Y}^*$ by a PCA solution. It is noted that measurement errors $\mathbf{E}_y$ can be interpreted as the part of SVD of $\mathbf{Y}^*$ corresponding to the insignificant components in the context of PCA. In particular, Tipping and Bishop (1998) found when those insignificant SVD values are roughly equal, a standard PCA solution would be the same as that estimated iteratively under a linear *isotropic* Gaussian noise FA model [1], referred to Probabilistic PCA in their work. All in all, this link reveals PCA can perhaps be a good guide before using FA.

The data analysis of FA models involves the two covariance matrices. One is the sample covariance matrix $\mathbf{S}$, which is used as a sufficient statistic. The other is the implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ (theoretical covariance matrix or population covariance matrix), which is formulised according to model structure and assumptions as a matrix-variate function of the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\Lambda}_y, \mathrm{Var}(\boldsymbol{\epsilon}_y), \mathrm{etc}\}$ (factor loadings and error variances). Parameter estimation proceeds by minimising (or maximising) some objective functions that measure the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, such as the weighed least square (WLS) or the generalized least square (GLS), or equivalently the maximum likelihood estimation

[1] Isotropic here means the variances of measurement errors are identical, and the covariance matrix is additionally assumed to have a diagonal structure.

(MLE). The derivation of asymptotic goodness-of-fit statistics for assessing whether $\mathbf{\Sigma}(\boldsymbol{\theta})$ fits $\mathbf{S}$ which depends heavily on the asymptotic multivariate normality of the discrepancy between $\mathbf{S}$ and $\mathbf{\Sigma}(\boldsymbol{\theta})$. These analysis procedures are referred to as covariance structure analysis (CSA). More detailed discussion can be found in (Bollen 1989, Lee 2007).

It is noted that an identification problem [2] should be considered before estimation.

2.1.3 Structural Equation Modelling

Structural Equation Modelling (SEM) is a statistical methodology for exploring, testing, and estimating causal relations by integrating data and qualitative structural and causal assumptions. Inheriting the characteristics of FA, it allows exploratory and confirmatory modelling to estimate functional or causal and probabilistic relationships, or to provide tools for dimension reduction (Bollen 1989, Pearl 2000).

Many families of SEMs are characterized by two sets of equations: a measurement model and a structural model (Bartholomew et al. 2008). A measurement model specifies the relationship between observed and latent variables. Like FA, it can be written as a set of regression equations where latent and measured variables serve as covariate and response variables respectively. The regression coefficients (factor loadings) imply how much effect the latent variables bring to the observed variables (manifest variables).

A structural model comprises relationships among latent variables and effects of design variables on latent variables – a distinction from classical FA. As exploratory variables and response variables in a regression equation, exogenous latent variables explain endogenous latent variables. This can allow testing relationships between factors under some hypothetical causal assumptions. Under explicit causal assumptions, the regression coefficients (called structural parameters) measure direct effects of exogenous latent variables on other latent variables (Bollen 1989). As the errors in measurement equations, the disturbances of endogenous latent variables may have the same assumptions of being uncorrelated with other variables.

[2] An identification problem is to investigate whether a statistical model is identifiable, in other words, whether the model parameters are uniquely determined by the model structure and the distributional information for the variables. Mathematically, a model is identifiable if then only if the function relation from parameters to probability distributions of the observed variables is a one-to-one map. We will discuss the problem further and conditions to achieve identifiability later.

The algebraic representation of a SEM is given by:

$$\mathbf{y}^{(n)} = \boldsymbol{\lambda}_{0,y} + \boldsymbol{\Lambda}_y \mathbf{x}^{(n)} + \boldsymbol{\epsilon}_y^{(n)}, \tag{2.3}$$

$$\mathbf{x_2}^{(n)} = \boldsymbol{\lambda}_{0,x} + \boldsymbol{\Lambda}_x \mathbf{x}^{(n)} + \boldsymbol{\epsilon}_x^{(n)} = \boldsymbol{\lambda}_{0,x} + \boldsymbol{\Lambda}_{x_1} \mathbf{x_1}^{(n)} + \boldsymbol{\Lambda}_{x_2} \mathbf{x_2}^{(n)} + \boldsymbol{\epsilon}_x^{(n)}, \tag{2.4}$$

where the superscript (n) means the n-th data point, $\mathbf{y}$ is a $R \times 1$ measured response vector; $\boldsymbol{\Lambda}$ is a $R \times Q$ factor loading matrix; $\mathbf{x} = (\mathbf{x_1}^\mathsf{T}, \mathbf{x_2}^\mathsf{T})^\mathsf{T}$ is a latent random vector ($\mathbf{x_1}$ and $\mathbf{x_2}$ are the exogenous and endogenous latent random vectors with size of $q_1 \times 1$ and $q_2 \times 1$ respectively); $\boldsymbol{\Lambda}_x = [\boldsymbol{\Lambda}_{x_1} \quad \boldsymbol{\Lambda}_{x_2}]$ is a $q_2 \times Q$ matrix of structure parameters that represent the causal effects among $\mathbf{x_1}$ and $\mathbf{x_2}$; $\boldsymbol{\epsilon}_y$ and $\boldsymbol{\epsilon}_x$ are $q_1 \times 1$ and $q_2 \times 1$ random vectors of measurement errors or residuals. $\boldsymbol{\lambda}_{0,y}$ and $\boldsymbol{\lambda}_{0,x}$ are intercept terms (if the measured and latent variables are taken as deviation from the mean, they are omitted). Equations (2.3) and (2.4) comprise a standard SEM, sometimes called a LISREL model (Linear Structural Relations Models) (Bartholomew et al. 2008).

Note that if SEM is used as a confirmatory tool, namely certain manifest variables are assumed to correlate a specific latent variable, the measurement equations (2.3) could be written as follows:

$$\begin{bmatrix} \mathbf{y}_1^{(n)} \\ \mathbf{y}_2^{(n)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}_{0,y_1} \\ \boldsymbol{\lambda}_{0,y_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Lambda}_{y_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{y_2} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^{(n)} \\ \mathbf{x}_2^{(n)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{y_1}^{(n)} \\ \boldsymbol{\epsilon}_{y_2}^{(n)} \end{bmatrix}, \tag{2.5}$$

where the block matrices $\mathbf{0}$ here reflect that disjoint subsets of manifest variables are assumed to measure only corresponding disjoint subsets of latent variables. The latent variables in the same subset may share certain manifest variables. One subset of latent variables may play a role of exogenous variables for the others.

The use of path diagrams equips SEM with a graph-based representation of the relationships between observed and latent variables, among latent variables, which can also be viewed as a directed graphical model. For example, Figure 2.1 shows a path diagram of political democracy and industrialization for developing countries, 1960 to 1965 (Bollen 1989). In the graph, boxes denote observed variables, circles denote latent variables and unenclosed characters denote errors. More specifically, democracy in 1960 (x_1) and that in 1965 (x_2) have four indicators respectively, freedom of the press (y_1, y_5), freedom of group opposition (y_2, y_6), fairness of elections (y_3, y_7) and the elective nature and effectiveness of the legislative body (y_4, y_8); industrialization in 1960 (x_3) has three indicators, the gross national product per capita (y_9), energy consumption per capita (y_{10}), and the percent of the labour force in industrial occupations (y_{11}). Curved lines with arrowheads at both ends denote correlations between exogenous variables and/or error terms (as showed in the graph). Relationships between observed variables and latent variables are represented by straight lines with an arrowhead pointing towards the dependent variable. In this example, the coefficients on straight lines additionally reveal how many effects a variable gives to its dependent variables. Such straight lines also represent assumed causal relationships where industrialization in 1960 influences democracy in 1960 and that in 1965.

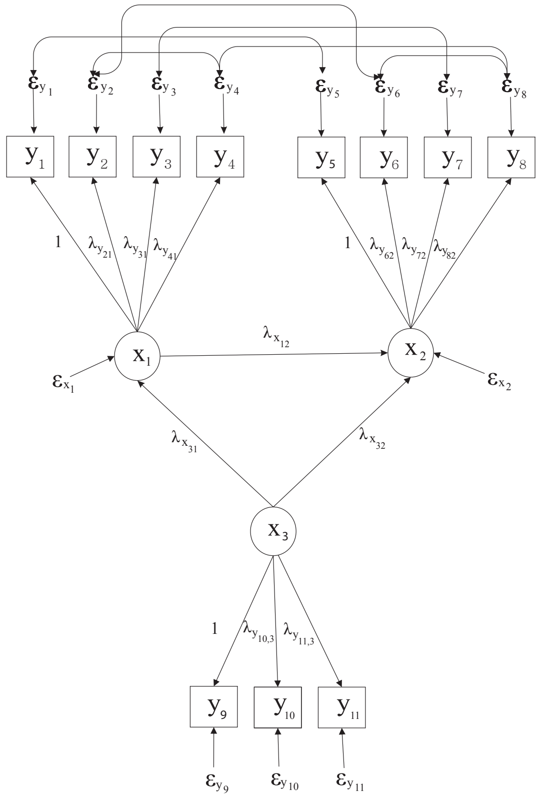


Figure 2.1: The path diagram of political democracy and industrialization for developing countries from 1960 to 1965

Covariance structure analysis (CSA) is a conventional method for modelling fitting in SEM, like its predecessor FA. With the development of Markov chain Monte Carlo (MCMC) samplings techniques, the Bayesian framework becomes another mainstream approach for parameter estimation. A Bayesian approach provides a convenient way to handle generalizations of such models and does not rely on asymptotic multivariate normality of the discrepancy (Lee 2007).

The general outline of a Bayesian procedure for SEM is: first, one has to specify the prior distribution model parameters. This can be done based on prior information from technical (or specific) expertise and analyses of analogous (or past) data. Often standard conjugate prior families are used: for example a normal distribution for factor loadings; an inverse gamma distribution for variances. For the situations without clear prior information, a non-informative prior could be adopted. Next, the posterior distributions of parameters and latent variables can be estimated by using a sufficiently large number of samples that are simulated from the posterior distribution of the unknown parameters through efficient statistical computing tools, such as MCMC methods. Some functional quantities of the posterior distribution, such as means or quantiles, can be estimated from the simulated samples. For more detailed discussion of Bayesian analysis procedure of various kinds of SEM; refer to (Lee 2007).

2.2 Gaussian Process For Regression

Since Gaussian Processes (GPs) was introduced in the machine learning community, it has been a popular approach for handling Bayesian non-parametric regression in that field (Rasmussen & Williams 2006). This probabilistic framework allows imposing a prior on a regression function by using a multivariate Gaussian distribution based on a specified covariance matrix. It gives more plausible functional forms to model regressions on which FA or SEM strongly counts. A GP model also provides another way to make predictions for test data points with less risk of poor predictions or underfitting compared to restricting the class of regression functions (for example, linear functions).

A GP can be interpreted as a distribution over functions. Following the definition in (Rasmussen & Williams 2006), a GP can be formally defined as a collection of normal variables with any finite number – it consists of a joint Gaussian distribution. More specifically, for a finite set of (multivariate) indices $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$, let $\mathbf{f} = (f(\mathbf{z}^{(1)}), \ldots, f(\mathbf{z}^{(N)}))^{\mathsf{T}}$, then $\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In a regression context, $\mathbf{z}^{(n)}$ corresponds to the covariates of the n-th data point[3]. $\boldsymbol{\mu}$ is a mean vector; $\boldsymbol{\Sigma}$ is a $N \times N$ covariance matrix specifying the dependency between any two function values, $f(\mathbf{z}^{(n)})$ and $f(\mathbf{z}^{(n^{'})})$. It should be noted that because N is any finite number, therefore one can informally represent a function as a lengthy vector, where each entry in the vector specifies the function value $f(\mathbf{z}^{(n)})$ at a particular covariate $\mathbf{z}^{(n)}$. The above definition also implies GP has the marginalization property. That means if GP specifies $(\mathbf{f}_1, \mathbf{f}_2)$, that is, $p(\mathbf{f}_1, \mathbf{f}_2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then it would be

$$p(\mathbf{f}_1) = \int p(\mathbf{f}_1, \mathbf{f}_2) d\mathbf{f}_2 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}),$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{11}$ are the submatrix of $\boldsymbol{\mu}$ and of $\boldsymbol{\Sigma}$, respectively.

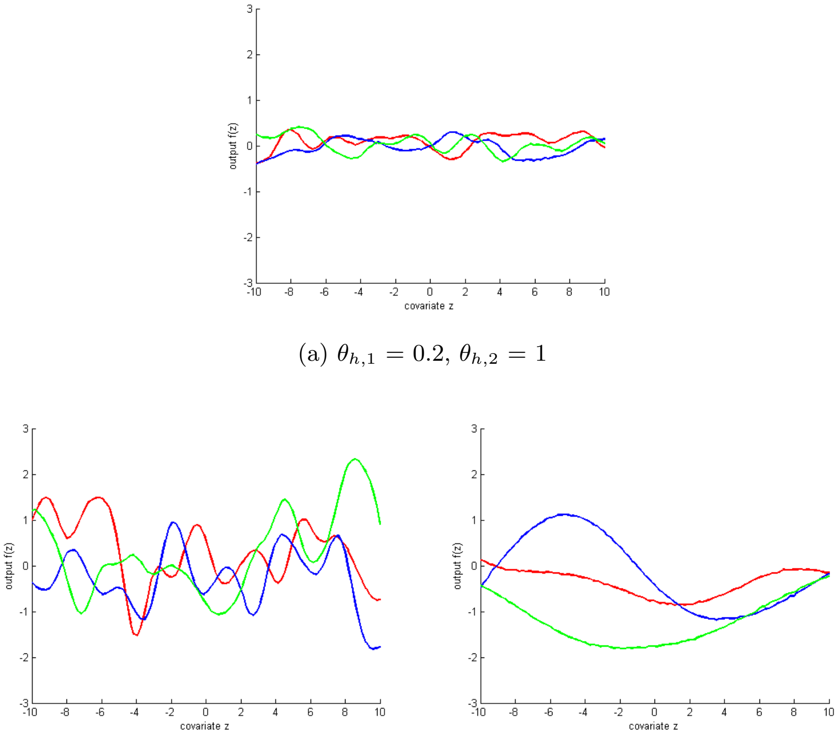
Generally, the GP of $f(\mathbf{z})$ can be written as $\mathbf{f}(\mathbf{z}) \sim \mathcal{GP}(m(\mathbf{z}), k(\mathbf{z}, \mathbf{z}^{'}))$ (Rasmussen & Williams 2006). $m(\mathbf{z})$ is the mean function of GP and describes the expected value of function f for input $\mathbf{z}$. It is usually set to be zero for notational simplicity; hereafter, we always consider this setup for a GP prior. $k(\mathbf{z}, \mathbf{z}^{'})$ is the covariance function controlling the variability of a function, defined as the covariance between two function values $f(\mathbf{z})$ and $f(\mathbf{z}^{'})$ based on the inputs $\mathbf{z}$ and $\mathbf{z}^{'}$. A covariance function can further be classified into two categories – stationary and non-stationary. The classification of the former depends on whether the function is parameterized by the inter-distance of inputs (implies invariance to translations of inputs).

A covariance function should be symmetric and positive semi-definite, that is, $k(\mathbf{z}, \mathbf{z}^{'}) = k(\mathbf{z}^{'}, \mathbf{z})$ and $\boldsymbol{\nu}^{\mathsf{T}} \mathbf{K} \boldsymbol{\nu} \geq 0$ for any $N \times 1$ vectors $\boldsymbol{\nu}$ (where $[\mathbf{K}]_{n,n'} = k(\mathbf{z}^{(n)}, \mathbf{z}^{(n^{'})})$). In addition, the resulting function of any covariance functions through algebraic operations, such as addition, multiplication and convolution, remains a valid covariance function (see Chapter 3 of (Rasmussen & Williams 2006)). Through this corollary, a researcher using GPs modelling formulation can attempt to create a new covariance function to obtain better data fitting and higher predictive precision.

The covariance function allows flexibility for setting high-level properties of a regression function (such as, smoothness, periodicity) and is characterised by its hyper-parameters (Rasmussen & Williams 2006). For example, one can use the squared-exponential (SE) covariance function

$$k(\mathbf{z}^{(n)}, \mathbf{z}^{(n^{'})}) = \theta_{h,1}^2 \exp\left(\frac{-1}{2\theta_{h,2}^2} |\mathbf{z}^{(n)} - \mathbf{z}^{(n^{'})}|^2 \right),$$

[3] From now on, we alternatively use the term “covariate” or “input”.



(a) $\theta_{h,1} = 0.2$, $\theta_{h,2} = 1$

(b) $\theta_{h,1} = 1$, $\theta_{h,2} = 1$

(c) $\theta_{h,1} = 1$, $\theta_{h,2} = 5$

Figure 2.2: Panel (a)-(c) show three functions randomly generated by a GP prior with zero mean function and squared-exponential covariance function with different values of signal variance $\theta_{h,1}^2$ and length-scale $\theta_{h,2}$.

where $\theta_{h,1}^2$ is the signal variance and $\theta_{h,2}$ is the length-scale, and both control the characteristic of the functions generated by a GP. The smaller $\theta_{h,1}$ is, the smaller amplitude a function varies with; and the larger $\theta_{h,2}$ is, the slower change a function varies with, as is showed in Figure 2.2; automatic relevance determination (ARD) is another common covariance function with a similar functional form as SE function. The difference is that it can model relevence of L inputs with different length-scales; in other words, $\theta_{h,l}$, $1 \leq l \leq L$, is the length-scale for the l-th dimension of input vectors $\mathbf{z}^{(n)}$. More covariance functions can be referred in Rasmussen and Williams (2006). For sake of computation, we use the SE covariance function in all the model exploratory experiments.

Given a dataset of N observations, one can consider a non-parametric regression model in a Bayesian formalism. The model can be written as follows:

$$y^{(n)} = f(\mathbf{z}^{(n)}) + \epsilon_y^{(n)},$$

where $\mathbf{z}$ is a covariate column vector of dimension L, $f(\cdot)$ is the function value (no functional form here indicates nonparametricity) and y is the observed noisy value, ϵ_y is assumed to follow an independent, identically distributed Gaussian distribution with zero mean and variance σ_y^2. The noise assumption implies the likelihood of the observations is factored over cases in the dataset. Furthermore, a Gaussian likelihood can be written as:

$$p(\mathbf{y}|\mathbf{f}, \mathcal{Z}) = \mathcal{N}(\mathbf{f}, \sigma_y^2 \mathbf{I}_N),$$

where $\mathbf{f} = (f(\mathbf{z}^{(1)}), \ldots, f(\mathbf{z}^{(N)}))^\mathsf{T}$ is a function vector where the n-th component is the function value at the input $\mathbf{z}^{(n)}$, $\mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^\mathsf{T}$ is a column response vector and $\mathcal{Z} = \{\mathbf{z}^{(1)} \ldots \mathbf{z}^{(N)}\}$ is a set of the covariate vectors. If integrating out $\mathbf{f}$ from joint distribution of $\mathbf{y}$ and $\mathbf{f}$ (or using (A.7)), one can obtain the marginal likelihood:

$$p(\mathbf{y}|\mathcal{Z}) = \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathcal{Z}, \mathcal{Z}) + \sigma_y^2 \mathbf{I}_N),$$

where $[\mathbf{K}(\mathcal{Z}, \mathcal{Z})]_{n,n'} = k(\mathbf{z}^{(n)}, \mathbf{z}^{(n^{'})})$, for $1 \leq n, n^{'} \leq N$, and $k(\cdot, \cdot)$ is the covariance function. The marginal likelihood can be used for learning hyper-parameters of a covariance function by gradient-based optimisation (Rasmussen & Williams 2006). For example, maximising the marginal likelihood over σ_y^2 gives a Bayesian estimate of the variance of the error terms.

Considering GP as a distribution over functions, one can specify a GP prior over the regression function to express our beliefs about the function before accessing the observations. The GP prior specifies the distribution of $\mathbf{f}$ as:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathcal{Z}, \mathcal{Z})). \tag{2.6}$$

By Bayes' rule, the posterior distribution of a regression function can be represented as the product of the likelihood and the GP prior divided by the normalizing constant (the marginal likelihood) for the Gaussian likelihood case. One can further derive its algebraic form as a Gaussian probability density:

$$p(\mathbf{f}|\mathbf{y}, \mathcal{Z}) = \mathcal{N}(\mathrm{mean}(\mathbf{f}_{post}), \mathrm{cov}(\mathbf{f}_{post})), \tag{2.7}$$

where

$$\mathrm{mean}(\mathbf{f}_{post}) = [\mathbf{K}(\mathcal{Z}, \mathcal{Z})^{-1} + \sigma_y^{-2} \mathbf{I}_N]^{-1} [\mathbf{K}(\mathcal{Z}, \mathcal{Z})^{-1} \mathbf{0} + \sigma_y^{-2} \mathbf{I}_N \cdot \mathbf{y}], \tag{2.8}$$

$$\mathrm{cov}(\mathbf{f}_{post}) = [\mathbf{K}(\mathcal{Z}, \mathcal{Z})^{-1} + \sigma_y^{-2} \mathbf{I}_N]^{-1}. \tag{2.9}$$

The mean and covariance matrix of the above posterior distribution can be derived by using the multiplication formula (A.5) of Gaussian distributions. It is noted that the posterior mean of $\mathbf{f}$ is a weighted average of the prior mean $\mathbf{0}$ and the data $\mathbf{y}$; the posterior inverse covariance matrix is the sum of the inverted covariance matrices in the prior and likelihood.

Let the subscript $*$ be the index of an unseen data point, such as data points in a test set whose function values we would like to predict. The predictive distribution can be derived by using Gaussian identities relating marginal and conditional distribution (A.6). More specifically, if a function vector $\mathbf{f}_*$ at the test-input set $\mathcal{Z}_*$ is a priori normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}(\mathcal{Z}_*, \mathcal{Z}_*)$, then the predictive distribution is

$$\mathbf{f}_*|\mathbf{y}, \mathcal{Z}, \mathcal{Z}_* \sim \mathcal{N}(\ \mathrm{mean}(\mathbf{f}_*), \mathrm{cov}(\mathbf{f}_*)), \tag{2.10}$$

where

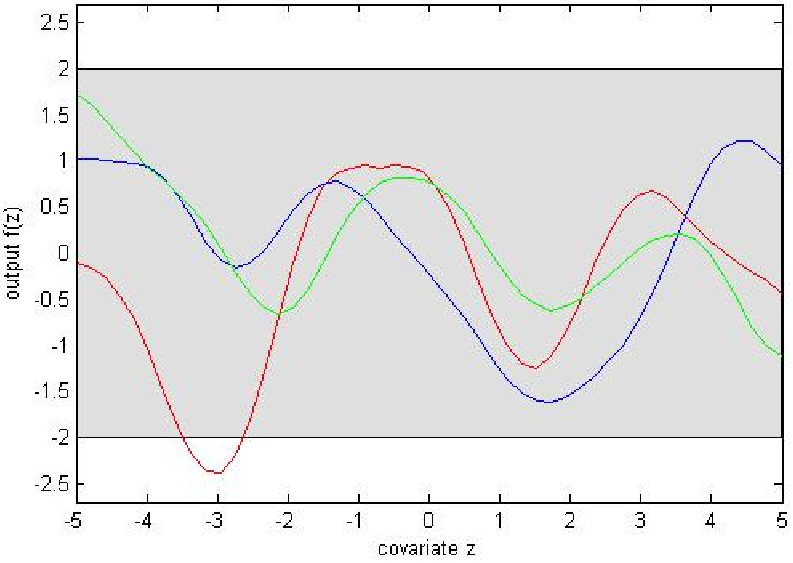
$$\begin{aligned} \mathrm{mean}(\mathbf{f}_*) &= \mathbf{K}(\mathcal{Z}_*, \mathcal{Z})[\mathbf{K}(\mathcal{Z}, \mathcal{Z}) + \sigma_y^2 \mathbf{I}_N]^{-1}\mathbf{y}, & (2.11) \\ \mathrm{cov}(\mathbf{f}_*) &= \mathbf{K}(\mathcal{Z}_*, \mathcal{Z}_*) - \mathbf{K}(\mathcal{Z}_*, \mathcal{Z})[\mathbf{K}(\mathcal{Z}, \mathcal{Z}) + \sigma_y^2 \mathbf{I}_N]^{-1}\mathbf{K}(\mathcal{Z}, \mathcal{Z}_*). & (2.12) \end{aligned}$$

Figure 2.3 shows an example of drawing samples from a prior distribution and samples from a posterior distribution. The grey area is the 95% confidence region for the prior and posterior. The region is based on the mean and the variance of a predictive distribution at test inputs equally located between -5 and 5.

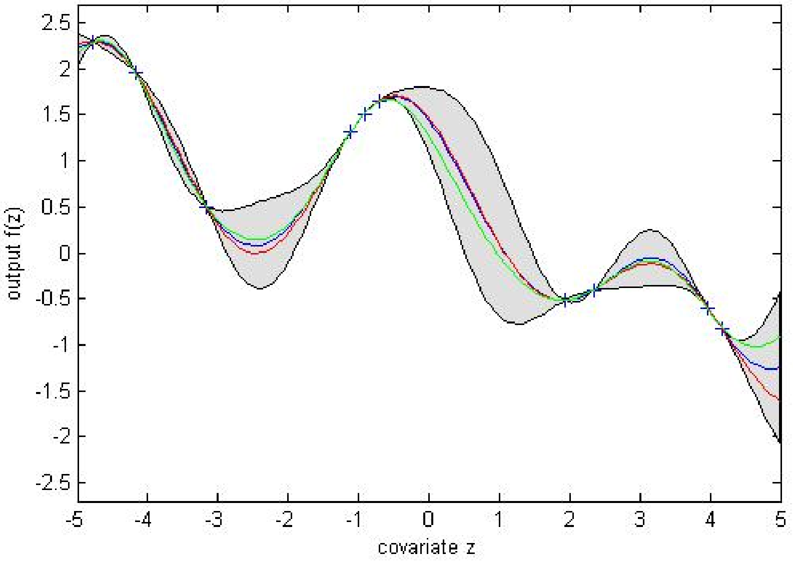
2.3 Sparse Approximations for GP regression

It is common that computational issues arise where using a GP model to calculate the mean and the covariance of the posterior distribution for a large dataset. Based on equations (2.8) and (2.9), the involved matrix inversion inevitably dominates computational cost because of a practically intractable scaling, $O(N^3)$. To address this issue, researchers have proposed several methods.

The Subset of Data (SD) method (Quinonero-Candela & Rasmussen 2005) is certainly the most naive approximation method one can consider. The main idea is described in the name. The associated inference, such as parameter estimation and prediction, is the same as that using full dataset. Total cost is therefore reduced to $O(M^3)$, where M is the size of a subset selected. Despite an easy implementation, it may be prone to poor



(a) Prior



(b) Posterior

Figure 2.3: Panel (a) shows three functions randomly drawn from a GP prior; Panel (b) shows three functions randomly drawn from the posterior distribution after adding 10 data points (denoted by +). The grey area is the 95% confidence region for the prior and posterior. The region is based on the mean and the variance of a predictive distribution at test inputs equally located between -5 and 5.

predictive performance because of lack of considering the links between the selected set and the corresponding function values. Some methods explained later could mitigate the issue.

There are common features on several approximate methods. For simplicity, we follow the unifying framework of (Quinonero-Candela & Rasmussen 2005) to describe the methods below.

Let $\mathbf{f}$ and $\mathbf{f}_*$ are latent function values on training and test inputs ($\mathscr{Z}$ and $\mathscr{Z}_*$, each with cardinal numbers of N and S respectively), and both are given a GP prior (2.6). Then their joint prior is

$$p(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{NN} & \mathbf{K}_{SN} \\ \mathbf{K}_{NS} & \mathbf{K}_{SS} \end{bmatrix}\right), \tag{2.13}$$

where $\mathbf{K}_{|\mathcal{A}||\mathcal{B}|}$ denotes the covariance matrix with entries evaluated at a pair of the sets $\mathcal{A}$ and $\mathcal{B}$.

The main idea of different algorithms is simply to introduce another set of latent variables $\bar{\mathbf{f}}$ with size M ($M < N$), and then modify the joint prior of $\mathbf{f}$ and $\mathbf{f}_*$ through $\bar{\mathbf{f}}$ by some assumptions.

By the marginalization property of GP, we know

$$p(\mathbf{f}, \mathbf{f}_*) = \int p(\mathbf{f}, \mathbf{f}_*, \bar{\mathbf{f}}) d\bar{\mathbf{f}} = \int p(\mathbf{f}, \mathbf{f}_* | \bar{\mathbf{f}}) p(\bar{\mathbf{f}}) d\bar{\mathbf{f}}, \tag{2.14}$$

where $\bar{\mathbf{f}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{MM})$, $\mathbf{K}_{MM} = \mathbf{K}(\bar{\mathscr{Z}}, \bar{\mathscr{Z}})$ and $\bar{\mathscr{Z}}$ is the set of inputs specifying $\bar{\mathbf{f}}$. Under the assumption that $\mathbf{f}$ and $\mathbf{f}_*$ are conditionally independent given $\bar{\mathbf{f}}$, then the approximate joint prior is

$$p(\mathbf{f}, \mathbf{f}_*) \simeq \tilde{p}(\mathbf{f}, \mathbf{f}_*) = \int \tilde{p}(\mathbf{f} | \bar{\mathbf{f}}) \tilde{p}(\mathbf{f}_* | \bar{\mathbf{f}}) p(\bar{\mathbf{f}}) d\bar{\mathbf{f}}. \tag{2.15}$$

The approximate conditional distributions $\tilde{p}(\mathbf{f}|\bar{\mathbf{f}})$ and $\tilde{p}(\mathbf{f}_*|\bar{\mathbf{f}})$ are the source of difference in various methods. One should notice that if there is no additional assumption on those conditionals, then the exact conditionals are simply predictive distributions of $\mathbf{f}$ and $\mathbf{f}_*$ given $\bar{\mathbf{f}}$ with means and covariance matrices derived by equations (2.11) and (2.12),

$$p(\mathbf{f}|\bar{\mathbf{f}}) = \mathcal{N}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\bar{\mathbf{f}}, \quad \mathbf{K}_{NN} - \mathbf{Q}_{NN}), \tag{2.16}$$

$$p(\mathbf{f}_*|\bar{\mathbf{f}}) = \mathcal{N}(\mathbf{K}_{SM}\mathbf{K}_{MM}^{-1}\bar{\mathbf{f}}, \quad \mathbf{K}_{SS} - \mathbf{Q}_{SS}), \tag{2.17}$$

where

$$\mathbf{Q}_{NN} = \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}, \tag{2.18}$$

$\mathbf{Q}_{SS}$ is defined in the same way.

The Subset of Regressors (SoR) approximate method (Smola & Bartlett 2001, Quinonero-Candela & Rasmussen 2005), or Deterministic Inducing Conditional (DIC), uses a deterministic way to approximate the conditionals. As is shown in the following equations, the approximate conditionals without any noise is simply determined by the mean of the exact predictive distributions given in (2.16) and (2.17).

$$\tilde{p}_{SoR}(\mathbf{f}|\bar{\mathbf{f}}) = \mathcal{N}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\bar{\mathbf{f}}, \mathbf{0}), \tag{2.19}$$

$$\tilde{p}_{SoR}(\mathbf{f}_*|\bar{\mathbf{f}}) = \mathcal{N}(\mathbf{K}_{SM}\mathbf{K}_{MM}^{-1}\bar{\mathbf{f}}, \mathbf{0}). \tag{2.20}$$

The resulting approximate joint prior of $\mathbf{f}$ and $\mathbf{f}_*$ can be derived by integrating out $\bar{\mathbf{f}}$ as did in (2.15):

$$\tilde{p}_{SoR}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q}_{NN} & \mathbf{Q}_{NS} \\ \mathbf{Q}_{SN} & \mathbf{Q}_{SS} \end{bmatrix}\right). \tag{2.21}$$

The Deterministic Training Conditional (DTC) method (Quinonero-Candela & Rasmussen 2005) has the approximate conditionals similar to those of SoR approximation, they are:

$$\tilde{p}_{DTC}(\mathbf{f}|\bar{\mathbf{f}}) = \mathcal{N}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\bar{\mathbf{f}}, \mathbf{0}), \tag{2.22}$$

$$\tilde{p}_{DTC}(\mathbf{f}_*|\bar{\mathbf{f}}) = p(\mathbf{f}_*|\bar{\mathbf{f}}). \tag{2.23}$$

Note that the approximation of $\mathbf{f}|\bar{\mathbf{f}}$ is exactly the same as that in SoR method and this inspires the descriptive name. The mean can be regarded as projection of $\mathbf{f}$ from $\mathbb{R}^N$ to $\bar{\mathbf{f}}$ on $\mathbb{R}^M$, this consideration leads to the alternative names of the DTC approximation, namely Projected Latent Variables (PLV) (Snelson & Ghahramani 2006*a*). The approximation of $\mathbf{f}_*|\bar{\mathbf{f}}$, on the other hand, has the exact form of $\mathbf{f}_*|\bar{\mathbf{f}}$ in equation (2.17). It improves flexibility (of function) when one undertakes posterior predictive task.

The associated joint prior of $\mathbf{f}$ and $\mathbf{f}_*$ is:

$$\tilde{p}_{DTC}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q}_{NN} & \mathbf{Q}_{NS} \\ \mathbf{Q}_{SN} & \mathbf{K}_{SS} \end{bmatrix}\right). \tag{2.24}$$

The Sparse Gaussian process (SPGP) is an outperforming approach on approximative GP models. Proposed by Snelson and Ghahramani (2006*a*), it is a more sophisticated likelihood approximation based on a conditional independence assumption given a set of latent variable $\bar{\mathbf{f}}$. Their work can be equivalently represented in the unifying framework of

Quinonero-Candela and Rasmussen (2005) with the name of Fully Independent Training Conditional (FITC):

$$\begin{aligned}
\tilde{p}_{FITC}(\mathbf{f}|\bar{\mathbf{f}}) &= \prod_{n=1}^{N} p(f^{(n)}|\bar{\mathbf{f}}) && (2.25)\\
&= \mathcal{N}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\bar{\mathbf{f}}, diag[\mathbf{K}_{NN} - \mathbf{Q}_{NN}]) && (2.26)\\
\tilde{p}_{FITC}(\mathbf{f}_*|\bar{\mathbf{f}}) &= p(\mathbf{f}_*|\bar{\mathbf{f}}). && (2.27)
\end{aligned}$$

The approximate joint prior of $\mathbf{f}$ and $\mathbf{f}_*$ is:

$$\tilde{p}_{FITC}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q}_{NN} - diag[\mathbf{Q}_{NN} - \mathbf{K}_{NN}] & \mathbf{Q}_{NS} \\ \mathbf{Q}_{SN} & \mathbf{K}_{SS} \end{bmatrix}\right), \tag{2.28}$$

where $diag[\cdot]$ is a linear operator to transform an input matrix into the corresponding diagonal matrix. From equations (2.24) and (2.28), one can recognize the difference between DTC and FITC is the top left block matrix. The replacement of the diagonal elements reveals the prior variance of latent function values $\mathbf{f}$ at training inputs is as the same as auto-covariances of $\mathbf{K}_{NN}$.

If a partial independence assumption is imposed on the training conditional $\mathbf{f}|\bar{\mathbf{f}}$ [4], then the Partial Independent Training Conditional (PITC) approximation can be achieved merely via replacing the diagonal operator by the block diagonal one. More details can be found in (Quinonero-Candela & Rasmussen 2005, Snelson & Ghahramani 2007).

The aforementioned methods can be categorized as likelihood approximations because they achieve computational merits by approximating conditional likelihood through a small number of inducing variables. In fact, one could consider the approximations in another viewpoint of matrix approximation. From the Mercer theorem (Rasmussen & Williams 2006), any non-degenerate covariance function (e.g. SE covariance function) can be represented in terms of infinite non-negative eigenvalues and the associated eigenfunctions. If one selects M pivot eigenvalues and the eigenfunctions to constitute a new covariance function, the resulting covariance matrix could achieve reduced low rank approximation to the exact matrix, that is

$$\mathbf{K}_{NN} \approx \mathbf{B}_1\mathbf{B}_1^{\mathsf{T}}, \tag{2.29}$$

where $\mathbf{B}_1$ is an $N \times M$ matrix. In practice it is very difficult to acquire an analytic closed-form of the eigen-decomposition. However, using a set of latent function values $\bar{\mathbf{f}}$ called

[4] Here"partial" means in full independence assumption only within a disjoint set of training inputs.

inducing variables could achieve the goal,

$$\mathbf{K}_{NN} \approx \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} = \mathbf{Q}_{NN}, \tag{2.30}$$

where $\mathbf{K}_{NM}$ is a matrix whose elements are evaluated at a pair of training set and inducing inputs. Note that the approximate joint prior involves inducing variables $\bar{\mathbf{f}}$ can also achieve reduction of time complexity from $O(N^3)$ to $O(NM^2)$ when one calculates the matrix inversion involving $\mathbf{K}_{NN}$ by the formula (A.1).

Selecting the inducing-input set $\bar{\mathcal{Z}}$ can be crucial for the approximation quality of sparse models. Traditionally, researchers can choose a subset from training set based on various methods. A subset could be made up of inputs within each cluster separated by using a classic K-mean method, or by using support vector machine (Cortes & Vapnik 1995) to choose inputs near the desired optimal separating hyperplane. Greedy posterior maximisation (Smola & Bartlett 2001) adopts a greedy forward selection scheme to find a subset that achieves the minimum of a quadratic objective function over transformed latent function values, where the function is related to the posterior distribution of those latent values. Seeger et al. (2003) also adopt a greedy forward selection method based on information gain that alternatively updates inducing inputs and hyper-parameters in two iterative steps. These sophisticated methods may be worthy of using in some applications. Here at starting stage for exploring our modelling methodology, we intend to adopt a scheme based on random selection and greedy selection of inputs.

In particular, if spatial features characterise inputs, one can select or design a set of knots in spatial space, as inducing inputs, in a regular or irregular way. For instance, one can set a uniform grid or different sizes of square grid on diverse areas. How to choose or design the knots in this case is not our focus and beyond our theme scope, one can refer more details in (Xia et al. 2006, Banerjee et al. 2008).

Another simple scheme one can use is the block Metropolis-Hastings method (Press 2003, Gilks et al. 1995). It simultaneously updates an inducing input (or pseudo input) and the associated inducing variable (or pseudo function variable) once a run. As this can be incorporated into a fully Bayesian framework, we will further describe and comment the sampling procedure in the Chapter 4.

In addition, Snelson and Ghahramani (2006*a*) consider inducing inputs as free estimated parameters of marginal likelihood. This thereby motivates them to find the location by simply using continuous optimisation. Instead of optimising marginal likelihood, Titsias et al. (2010) optimise total variance of the conditional distribution of $\mathbf{f}$ given inducing variables $\bar{\mathbf{f}}$ (they call control variables) for the selection of the associated input set $\bar{\mathcal{Z}}$. However, the computation loads could much increase on optimisation over extended parameter space due to more number of inducing inputs. For a computational reason, this optimisation scheme for the location of inputs would be not practised in our work.

2.4 Multiple Output Prediction Methods of GPR

Supervised learning on each output independently under GP framework is a naive approach for multi-output prediction tasks. Despite the simplicity of model formulation, this scheme likely underperforms in a scenario, for example, mentioned in Section 1.1. To boost predictability, one can construct covariance functions to capture information shared between multiple outputs. Several methods based on this idea have been proposed.

Linear models of coregionalization (LMC) (Goovaerts 1997), developed in the geostatistical community, models each of multiple output functions as a linear combination of independent random functions within a spatial domain. The observations of each output are simply noisy realizations of the output function at a specific location $\mathbf{z}$.

For convenience, we temporarily omit the noisy observed outputs and their error terms for the following explanations. Let $f_r(\mathbf{z})$, $1 \leq r \leq R$, be multiple output functions and $\nu_q(\mathbf{z})$, $1 \leq q \leq Q$, be independent random functions. The LMC framework can be expressed as:

$$f_r(\mathbf{z}) = \sum_{q=1}^{Q} \omega_{rq} \nu_q(\mathbf{z}), \tag{2.31}$$

where $\mathbf{z}$ lies in L-dimension input space, and ω_{rq} are the linear coefficients (or weights).

If $\nu_q(\mathbf{z})$ follows as a GP (that is, $\nu_q(\mathbf{z}) \sim \mathcal{GP}(0, k_q(\mathbf{z}, \mathbf{z}^{'}))$) and is independent of $\nu_{q^{'}}(\mathbf{z})$ for $q \neq q^{'}$, by the proposition about product of covariance functions (Rasmussen & Williams 2006), one can derive that $f_r(\mathbf{z})$ is distributed as another GP. More specifically, the covariance functions of $\nu_q(\mathbf{z})$ can be written as

$$cov(\nu_q(\mathbf{z}), \nu_{q^{'}}(\mathbf{z}^{'})) = \delta_{qq^{'}} k_q(\mathbf{z}, \mathbf{z}^{'}), \tag{2.32}$$

where $\delta_{qq^{'}}$ denotes the Kronector delta function. Then the covariance of $f_r(\mathbf{z})$ is

$$cov(f_r(\mathbf{z}), f_{r^{'}}(\mathbf{z}^{'})) = \sum_{q=1}^{Q} \omega_{rq} \omega_{r^{'}q} k_q(\mathbf{z}, \mathbf{z}^{'}), \tag{2.33}$$

Equation (2.31) can be written in a more general way so that each output enables to be expressed as a linear combination of groups of $\nu_q(\mathbf{z})$, and each group shares the same covariance function:

$$f_r(\mathbf{z}) = \sum_{q=1}^{Q} \sum_{j=1}^{S_q} \omega_{rq}^{j} \nu_q^{j}(\mathbf{z}), \tag{2.34}$$

then the covariance function of $f_r(\mathbf{z})$ is therefore

$$cov(f_r(\mathbf{z}), f_{r'}(\mathbf{z}^{'})) = \sum_{q=1}^{Q} \sum_{j=1}^{S_q} \omega_{rq}^{j} \omega_{r^{'}q}^{j} k_q(\mathbf{z}, \mathbf{z}^{'}). \tag{2.35}$$

Given a set of inputs $\mathcal{Z}$ with size N, $\mathbf{f}_r = (f_r(\mathbf{z}^{(1)}), \dots, f_r(\mathbf{z}^{(N)}))$ denotes the r-th output vector evaluated at $\mathcal{Z}$, and then the covariance of $\mathbf{f}_r$ and $\mathbf{f}_{r'}$ is

$$cov(\mathbf{f}_r, \mathbf{f}_{r'}) = \sum_{q=1}^{Q} \sum_{j=1}^{S_q} \omega_{rq}^{j} \omega_{r^{'}q}^{j} \mathbf{K}_{q;N}, \tag{2.36}$$

where $[\mathbf{K}_{q;N}]_{i,j}$ is the covariance evaluated at a pair of inputs $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$ through $k_q(\cdot, \cdot)$. The covariance matrix for a joint output vector $\mathbf{f} = (\mathbf{f}_1^{\mathsf{T}}, \dots, \mathbf{f}_R^{\mathsf{T}})^{\mathsf{T}}$ can be written

$$\mathbf{K}_{\mathbf{ff}} = \sum_{q=1}^{Q} \mathbf{\Omega}_q \mathbf{\Omega}_q^{\mathsf{T}} \otimes \mathbf{K}_{q;N} \tag{2.37}$$

$\mathbf{\Omega}_q$ is a $R \times S_q$ matrix consisting of all linear coefficients ω_{rq}^{j}. From equation (2.37), we can interpret the covariance matrix captures two sources of information: one from the dependence between output functions, represented by $\mathbf{\Omega}_q \mathbf{\Omega}_q^{\mathsf{T}}$; the other from the dependence between all inputs, given by $\mathbf{K}_{q;N}$.

The semiparametric latent factor model (SLFM) (Teh et al. 2005), can be regarded as a simplified LMC. Firstly, $S_q = 1$; this means that individual latent factor functions $\nu_q(\mathbf{z})$ ($\nu_q(\mathbf{z}) \sim \mathcal{GP}(0, k_q(\mathbf{z}, \mathbf{z}^{'}))$) linearly mix into each output function $f_r(\mathbf{z})$ as shown in Equation (2.31). Secondly, $R < Q$; it implies all R output functions linearly represented by a small number Q of latent functions. This scheme follows the spirit of FA and serves as a dimension-reduction technique.

The covariance matrix of SLFM for the joint output function vector $\mathbf{f}$ can be written as

$$\mathbf{K}_{\mathbf{ff}} = \sum_{q=1}^{Q} \boldsymbol{\omega}_q \boldsymbol{\omega}_q^{\mathsf{T}} \otimes \mathbf{K}_{q;N} = \sum_{q=1}^{Q} (\boldsymbol{\omega}_q \otimes \mathbf{I}_N) \mathbf{K}_{q;N} (\boldsymbol{\omega}_q^{\mathsf{T}} \otimes \mathbf{I}_N) \tag{2.38}$$

where $\boldsymbol{\omega}_q$ is a column vector with R entries. The second equality is obtained by the properties of Kronecker product.

Multi-task Gaussian Process (MGP) (Bonilla et al. 2008) for multi-task learning is another variant of the LMC framework. The main idea is that each output function is a linear combination of one group of latent functions (drawn from a GP). Moreover, their covariance function is written as a product of two scalars. One features a free estimated covariance of output functions for a pair of tasks, and the other the correlation of latent functions evaluated at a pair of inputs. In the context of the LMC, for $Q = 1$, the term $\sum_{j=1}^{S_1} \omega_{r1}^{j} \omega_{r'1}^{j}$ in equation (2.35) is factorized by two terms $c_{r,r'}$ and b_1. Then the covariance matrix for the joint output function vector $\mathbf{f}$ is

$$\mathbf{K}_{\mathbf{ff}} = \mathbf{C} \otimes (b_1 \mathbf{K}_{1;N}), \tag{2.39}$$

where $\mathbf{C}$ is $R \times R$ matrix with elements $c_{r,r'}$.

Besides constructing a covariance function for multiple output functions based on the mechanism of linearly mixing latent functions, utilization of convolution processes has been undertaken by several researchers (Boyle & Frean 2004, Higdon 2002, Álvarez et al. 2011, Álvarez & Lawrence 2009). The idea involves that each output function is expressed through a convolution integral between a smoothing kernel and a latent function. Like equation (2.34), one can generally express each output function as

$$f_r(\mathbf{z}) = \sum_{q=1}^{Q} \sum_{j=1}^{S_q} \int G_{rq}^{j}(\mathbf{z} - \mathbf{z}^{'}) \nu_q^{j}(\mathbf{z}) d\mathbf{z}^{'}. \tag{2.40}$$

In addition, if the convolving kernels $G_{rq}^{j}(\mathbf{z} - \mathbf{z}^{'})$ are the products of w_{rq}^{j} and Dirac delta functions $\delta_{\mathbf{z}}(\mathbf{z}^{'})$, the LMC framework is derived. The convolved modelling framework related to multiple outputs can therefore be regarded as a general version of the LMC and thereby proposes more general mixing fashion. Because our framework is related to FA, we thus do not use the above scheme to pursue solving multiple output prediction problems. More details, such as computationally efficient methods and approaches to address an issue caused by non-smooth latent functions, can be found in (Álvarez et al. 2011, Álvarez & Lawrence 2009).

Incidentally, the computational issue of the multiple-output prediction methods above could be generally mitigated by using the concept of sparse approximation approaches mentioned in section 2.3; that is, the utilization of a limited number of inputs and the associated latent functions [5].

[5] MGP in Bonllina et al. (2008) also adopts probabilistic PCA to approximate the linear combination matrix linking original GP functions to output functions.

2.5 Remarks

To sum up, PCA, FA and SEM not only can serve as a dimension-reduction technique with metric latent variables, but also can work for exploring a dominant data pattern in a low-dimension space constructing by latent variables. Despite the similarity, PCA is commonly regarded as extracting information from data summarisation through a linear transformation although in particular, probabilistic PCA can be viewed as a model-based method. FA, however, is considered as exploring inverse-regression relationship between latent variables and manifest variables through fitting a probabilistic model. FA and SEM could be viewed as a model inferring whether a hypothesis (casual relationship between latent variables and manifest variables and among latent variables) is reasonable.

A GP regression model gives one more flexibility to model functional relations between observed covariates and outputs for a supervised learning task. Given a Gaussian prior on function values, one can realize the most possible regression function fitting data points a posteriori. The characteristic of a regression function generated by a GP is governed by the covariance function and its hyper-parameters. This implies using a certain covariance function could perhaps achieve a better predictive performance.

A computational issue of a GP regression model for large dataset can be alleviated by several approximation methods. Their common idea is to introduce the conditional independence assumption, through inducing variables (a finite number set of latent function values), on the joint GP prior of latent function values at training and test inputs. The selection of inducing inputs, which generates inducing variables, can rely on some schemes, such as choosing randomly from training inputs, optimizing marginal likelihood over inducing inputs as free parameters.

Beside adopting an independent scheme, several prediction methods for a multiple-output GP model intend to capture relation between all outputs. To achieve that, one can construct a covariance function for each of multiple output functions by linearly mixing latent functions given GP priors.

Chapter 3

Framework of GP-SEM

This chapter aims to provide a new modelling framework, building on some works presented in the last chapter. The model structure is furthermore treated with different estimation methods in the subsequent chapters.

In Section 3.1, we describe the data structure suitable for our modelling framework and then specify the modelling formulation. We then specify its sparse version. In Section 3.2, we briefly examine the model identification in a simple example. Finally, we conclude with a general remark about the model structure.

3.1 Model Specification

3.1.1 Data Structure for Modelling Availability

The dataset (consisting of N data points) feasible to our modelling framework is of multiple-dimensional and each dimension represents a random variable of interest in an observational or experimental study. A subset of dimensions can be regarded as covariates that are assumed to indirectly or directly affect the rest of dimensions as responses or indicators. Furthermore, under our consideration those covariates can be metrical (continuous, discrete) or categorical (ordinal or nominal)[1] and responses can be metrical here. The responses or indicators are assumed to measure a certain latent characteristics regarded as metrical random variables. Take a simple example in an educational study that

[1] In this case, we adopt the treatment in Bonilla et al. (2008) to create dummy variables for the categorical features. For example, a covariate for four ethnic groups can be replaced by dummy variables"1000", "0100", "0010" and "0001". Note that it is unnecessary to create dummy variables for a binary covariate.

we may treat a student's age, family income, gender and grade as covariates and their mathematics, physics, history and literature examination scores as responses or indicators to measure their latent IQ scores.

From the above description, the n-th data point ($1 \leq n \leq N$) has a covariate-response pair denoted by $(\mathbf{z}^{(n)}, \mathbf{y}^{(n)})$. Each $L \times 1$ covariate vector $\mathbf{z}^{(n)}$ records the L realizations of the characteristics of the n-th case[2]. Each $R \times 1$ response vector $\mathbf{y}^{(n)}$ registers the measured values or the response outcomes of the n-th case to measure the n-th case's Q latent characteristics of interest, which denotes a $x_q^{(n)}$, for $1 \leq q \leq Q$.

3.1.2 Full Gaussian Process Structure Equation Modelling

The first part of the model framework can be regarded as a black box for the relationship between a covariate vector $\mathbf{z}^{(n)}$ and a latent variable $x_q^{(n)}$. In other words, $x_q^{(n)}$ can be viewed as a noisy outcome of an input $\mathbf{z}^{(n)}$ through a non-parametric function $f_q(\cdot)$. Let $f_q^{(n)}$ be a function value evaluated at $\mathbf{z}^{(n)}$, that is, $f_q(\mathbf{z}^{(n)})$, and $\mathbf{\Sigma}_x$ be a $Q \times Q$ noise covariance matrix of $\boldsymbol{\epsilon}_x^{(n)} = [\epsilon_{x_1}^{(n)}, \ldots, \epsilon_{x_Q}^{(n)}]^\intercal$, then the GP model formulation is

$$x_q^{(n)} = f_q^{(n)} + \epsilon_{x_q}^{(n)}, \tag{3.1}$$

$$\boldsymbol{\epsilon}_x^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_x), \tag{3.2}$$

where $\mathcal{N}(\mathbf{m}, \mathbf{C})$ is the multivariate Gaussian distribution with a mean vector $\mathbf{m}$ and covariance matrix $\mathbf{C}$. The function value $f_q^{(n)}$ and $\epsilon_{x_q}^{(n)}$, and the noise vectors $\boldsymbol{\epsilon}_x^{(1)}, \ldots, \boldsymbol{\epsilon}_x^{(N)}$ are assumed to be mutually independent, respectively. It is noted that $\mathbf{\Sigma}_x$ is also the conditional covariance matrix of the n-th latent variable vector consisting of all latent variables, denoted $\mathbf{x}^{(n)} = [x_1^{(n)}, \ldots, x_Q^{(n)}]^\intercal$, given all corresponding function values $f_1^{(n)}, \ldots, f_Q^{(n)}$. In addition, a GP prior for function $f_q(\cdot)$ is

$$\mathbf{f}_q | \mathbf{z}^{1:N} \sim \mathcal{N}(0, \mathbf{K}_{q;N}), \tag{3.3}$$

where $\mathbf{f}_q = [f_q^{(1)}, \ldots, f_q^{(N)}]^\intercal$ is a function value vector, a covariate set $\mathbf{z}^{1:N} \equiv \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}\}$, and $\mathbf{K}_{q;N}$ is a $N \times N$ covariance matrix, determined by $\mathbf{z}^{1:N}$ and the q-th covariance function $k_q(\cdot, \cdot)$.

The measurement model, the second part of the model structure, is to describe the distribution of the n-th case's noisy observation vector $\mathbf{y}^{(n)}$ given its latent variable vector

[2] We may alternatively use a word *case* or *subject* for data point.

$\mathbf{x}^{(n)}$. Let $\boldsymbol{\Sigma}_y$ be a $R \times R$ noise covariance matrix of $\boldsymbol{\epsilon}_y^{(n)} = [\epsilon_{y1}^{(n)}, \ldots, \epsilon_{yR}^{(n)}]^\top$, we have

$$\mathbf{y}^{(n)} = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\mathbf{x}^{(n)} + \boldsymbol{\epsilon}_y^{(n)}, \tag{3.4}$$

$$\boldsymbol{\epsilon}_y^{(n)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_y). \tag{3.5}$$

Here the error terms $\epsilon_{y1}^{(n)}, \ldots, \epsilon_{yR}^{(n)}$ can be assumed to be mutually independent (it implies $\boldsymbol{\Sigma}_y$ is a diagonal matrix) and to be independent of all latent variables $x_1^{(n)}, \ldots, x_Q^{(n)}$. Furthermore, a $R \times 1$ intercept vector $\boldsymbol{\lambda}_0$ and a $R \times Q$ factor loading matrix $\boldsymbol{\Lambda}$ can be interpreted in the same way as coefficients of a linear regression.

The model structure represented in Equations (3.1)-(3.5) is a GP latent variable model. Because it is of pure GP formulation (without others auxiliary variables), we add a *full* before the model name to indicate this characteristic. Furthermore, the GP formulation part can be viewed as a non-parametric structure model and we therefore refer to as full Gaussian Process Structure Equation Modelling (full GP-SEM).

The above GP-SEM has multilevel structure. The first level is for the multiple observed responses. The second level is for the general latent characteristics of interest of a case. These two levels follow a factor analysis framework to realize distributions of latent variables for exploring a possible pattern in a low-dimensional space. The second level and the covariates additionally follow a GP framework to enhance flexibility of direct-effect mechanism of covariates on latent variables for an implicit causal functional relation and a possible improvement of prediction on manifest responses.

The two-part formulation comprises a semi-parametric framework. In essence, this GP-SEM can be viewed as a semi-parametric multiple indicators multiple causes (MIMIC) model (a special SEM allowing covariates directly affect on latent or manifest variables) although not modelling direct effect on manifest variables. If the GP-SEM is for longitudinal analysis (that means the model with time scaling), it would be considered as a semi-parametric version of several popular methods, such as Latent Curve Model with Latent variable (Bollen 2006), multilevel SEM (Steele 2008). More details are presented in Chapter 6.

It is noted that the full GP-SEM might be non-identifiable and this can lead to some problems. For classical frequentist inference, failure of being an identifiable model leads a theoretic issue because of inconsistent parameter estimation – estimates starting at different initial points do not converge to the same value. For Bayesian inference, a non-identifiable model can have more computational issues because it might lead to poor mixing in MCMC simulations (Silva & Gramacy 2010). In practice, one can solve the issue by adding constraints on model coefficients and distributional assumptions on variables. For example, commonly a factor loading corresponding to an indicator can be set to one, or variances of noise terms can be set to be unity. Adding constraints can make a model globally or locally identifiable.[3]

However, sometimes to add what constraints could be another problem. To check model identification condition may give one an insight to solve. One could examine whether each free parameter is represented by a function of the moments of observed variables (such as means, covariances) and other fixed parameters, under the model structure and distributional information on variables. For full GP-SEM or its sparse version presented later, we demonstrate the algebraic check given the constraints in the Section 3-2.

Figure 3.1 shows a graphical representation of the model structure with two latent variables and four manifest variables as an illustrative example [4]. From the figure one can realize the information propagation between latent variables x_1 and of x_2 possibly is achieved through the linked noise terms ϵ_{x_1} and ϵ_{x_2}. If the cross-covariances of noise terms are free parameters, the correlations of indicators belonging to disjoint sets (each of which measure different latent variables) could perhaps be captured. It may enhance predictive performance on a certain indicator after being given the rest.

3.1.3 Sparse Gaussian Process Structural Equation Modelling

The above full GP-SEM is simple and model parameters can also be estimated using prevalent MCMC or EM methods. However, each estimation step, in practice, takes much computational costs. This is because calculating the covariance matrix of the posterior distribution of latent variables involves matrix inversion. It leads very low executive speed when N is simply in hundreds. To address this issue, we introduce an alternative model framework with a multilayered structure, adapted from the pseudo-input model – Sparse Gaussian Process (SPGP) Model (Snelson & Ghahramani 2006*a*) or Full Independent Training Condition (FITC) model (Quinonero-Candela & Rasmussen 2005). This model

[3] Being globally identifiable means that a model is identifiable for all the points in parameter space; being locally identifiable only for a certain neighbourhood of a point.

[4] The graphical representation mixes a path diagram commonly adopted in the context of SEM.

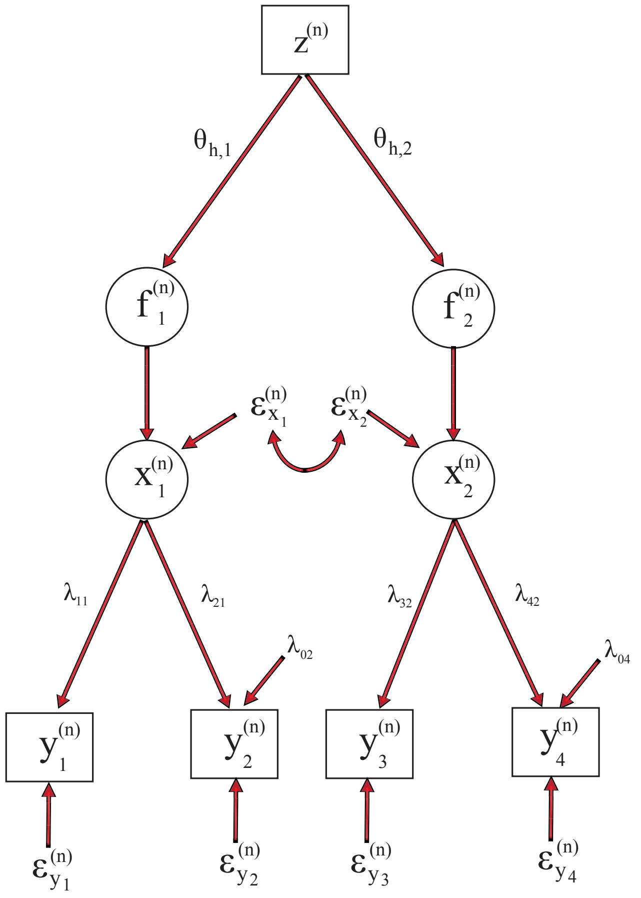


Figure 3.1: An illustrative full GP-SEM with 2 latent variable and 4 manifest variables

allows one to reduce the total computational cost from $O(N^3Q^3)$ down to $O(M^2NQ^2)$, where $M < N$ and M can be chosen according to the available computational resources. We explain the cost reduction in the next chapter.

For each q, define a pseudo-input set (or inducing-input set) $\bar{\mathbf{z}}_q^{1:M} \equiv \{\bar{\mathbf{z}}_q^{(1)}, \ldots, \bar{\mathbf{z}}_q^{(M)}\}$, and each $\bar{\mathbf{z}}_q^{(m)}$, $1 \leq m \leq M$, has the same dimension as $\mathbf{z}^{(n)}$. Moreover, we define a pseudo latent function value vector $\bar{\mathbf{f}}_q = [\bar{\mathbf{f}}_q^{(1)}, \ldots, \bar{\mathbf{f}}_q^{(M)}]^\mathsf{T}$, and then the pseudo-inputs model is

$$\mathbf{f}_q|\bar{\mathbf{f}}_q, \mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M} \sim \mathcal{N}(\mathbf{K}_{q;NM}\mathbf{K}_{q;M}^{-1}\bar{\mathbf{f}}_q, \mathbf{V}_q), \tag{3.6}$$

$$\bar{\mathbf{f}}_q|\bar{\mathbf{z}}_q^{1:M} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{q;M}), \tag{3.7}$$

where $\mathbf{K}_{q;NM}$ is a $N \times M$ matrix with $[\mathbf{K}_{q;NM}]_{n,m} = k_q(\mathbf{z}^{(n)}, \bar{\mathbf{z}}_q^{(m)})$ and $\mathbf{K}_{q;M}$ is a $M \times M$ matrix with $[\mathbf{K}_{q;M}]_{m,m'} = k_q(\bar{\mathbf{z}}_q^{(m)}, \bar{\mathbf{z}}_q^{(m')})$. Let $\mathbf{k}_{q;nM}$ be the n-th row of $\mathbf{K}_{q;NM}$, then $\mathbf{V}_q$ is a $N \times N$ diagonal matrix with entries $[\mathbf{V}_q]_{n,n} = k_q(\mathbf{z}^{(n)}, \mathbf{z}^{(n)}) - \mathbf{k}_{q;nM}\mathbf{K}_{q;M}^{-1}\mathbf{k}_{q;nM}^\mathsf{T}$. This implies all the components $f_q^{(1)}, \ldots, f_q^{(N)}$ of the q-th latent function value vector $\mathbf{f}_q$ are conditionally independent. One can find Equations (3.6) and (2.26) are almost the same but the multiple indices q and different covariance matrix notation[5].

Note that the pseudo data set, including the pseudo-inputs set $\bar{\mathbf{z}}_q^{1:M}$ and pseudo function values $\bar{\mathbf{z}}_q^{(m)}$, $1 \leq m \leq M$, works as another training set. If $M = N$ and $\bar{\mathbf{z}}_q^{1:M} = \mathbf{z}^{1:N}$, then Equation (3.6) would become the original predictive distribution as Equations (2.10)-(2.12).

Equations (3.1)-(3.7) comprise another kind of GP-SEM for dealing with a computational issue, we refer it to as Sparse Gaussian Process Structure Equation Modelling (Sparse GP-SEM). Figure 3.2 shows the model structure under the same scenario as Figure 3.1, one can find the differences are the additions of pseudo inputs and pseudo function values.

GP-SEM here is also close to the modelling framework proposed by (Silva & Gramacy 2010), named as Gaussian Process Structural Equation Modelling with Latent variables (GPSEM-LV). Their work focuses on the model structure between latent variables but ours aim to model relations between observed covariates and latent variables. The distinction can be alternatively viewed as whether to adopt latent covariates or observed covariates. Both additionally consider utilising pseudo inputs to speed up computation.

In addition, there are differences between sparse GP-SEM and SLFM, the multi-output

[5] For convenience, we change the notation of the covariance matrix and use that in the following chapters.

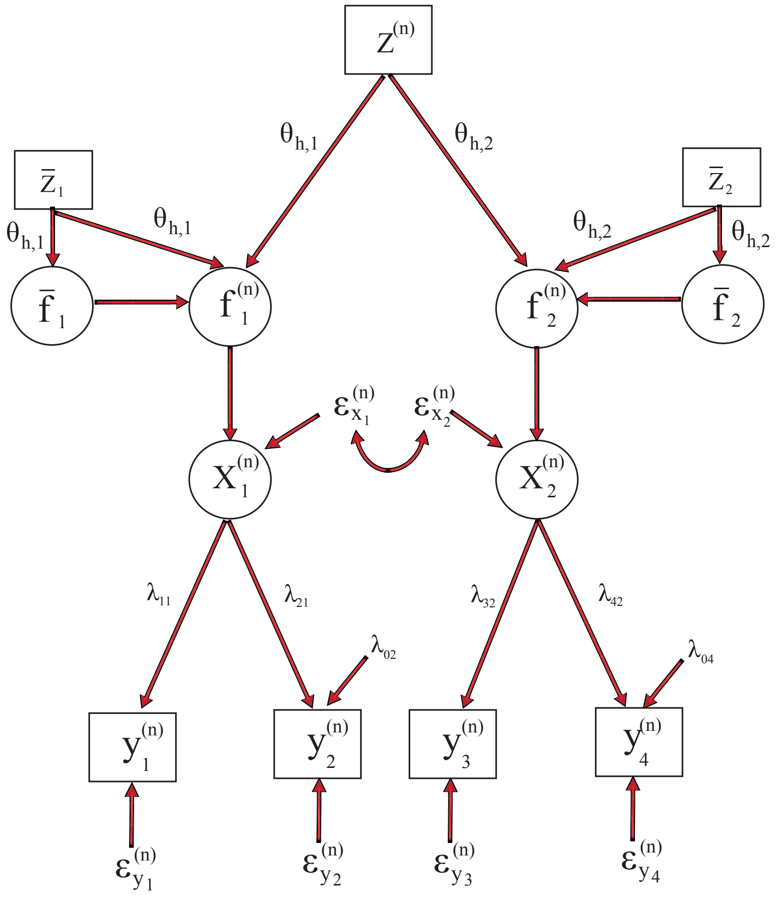


Figure 3.2: An illustrative sparse GP-SEM with 2 latent variable and 4 manifest variables

GP model proposed by Teh et al. (2005). First, although both exploit a factor-analysis linear formulation (served dimension reduction), that configuration is installed in different parts of the models. For (full or sparse) GP-SEM, the latent variables of the measurement model connect the observed variables in the linearly mixing way. For the SLFM, the latent variables are represented as a linear combination of the GP latent functions. Secondly, the SLFM does not involve latent errors $\epsilon_{x_1}, \cdots, \epsilon_{x_Q}$ in the GP regression formulation. It suggests that each of the latent variables is not an error-in-variable in the regression of the corresponding observed variable. However, our full or sparse GP-SEM incorporates inter-correlated latent errors. Thirdly, regarding computation, the SLFM uses a sophisticated technique involving greedy forward selection of inputs and outputs. For sparse GP-SEM, pseudo inputs are adopted but selected from the original input set through a MCMC sampling scheme, a random or greedy scheme, as discussed in Chapter 4.

An alternative expression of sparse GP-SEM can be formed by substituting $f_q^{(n)}$ in (3.1) with (3.6). It thus consists of Equations (3.2), (3.4),(3.5), (3.7) and the following new equation:

$$x_q^{(n)} = \mathbf{k}_{q;nM}\mathbf{K}_{q;M}^{-1}\bar{\mathbf{f}}_q + \epsilon_{x_q}^{(n)} + \epsilon_{\bar{f}_q}^{(n)}, \tag{3.8}$$

where $\epsilon_{\bar{f}_q}^{(n)}$ is a Gaussian distribution with mean 0 and variance $[\mathbf{V}_q]_{n,n}$.

Figure 3.3 shows the alternative sparse GP-SEM under the same setting as Figure 3.2. One can observe all the latent function values $f_1^{(n)}$ and $f_2^{(n)}$ are removed and all the involved straight lines are changed to point towards latent variables $x_1^{(n)}$ and $x_2^{(n)}$.

This alternative expression could be useful when one uses modelling estimation with MCMC methods. The reason is that high coupling correlation between latent variables $\mathbf{x}$ and latent function values $\mathbf{f}$ may lead poor sampling mixing. Therefore, integrating out latent function values $\mathbf{f}$ from the joint distribution of the original sparse GP-SEM can perhaps mitigate the issue for sampling latent variables. However, one still needs to use the original sparse GP-SEM for $\mathbf{f}$ and the error covariance matrix $\mathbf{\Sigma}_x$. More sampling details will be presented in the next chapter.

3.2 Examination of Identification Condition

Given model formulations and constraints on parameters, we attempt to examine whether the model identifiability is necessarily achieved by those known parameters and assump-

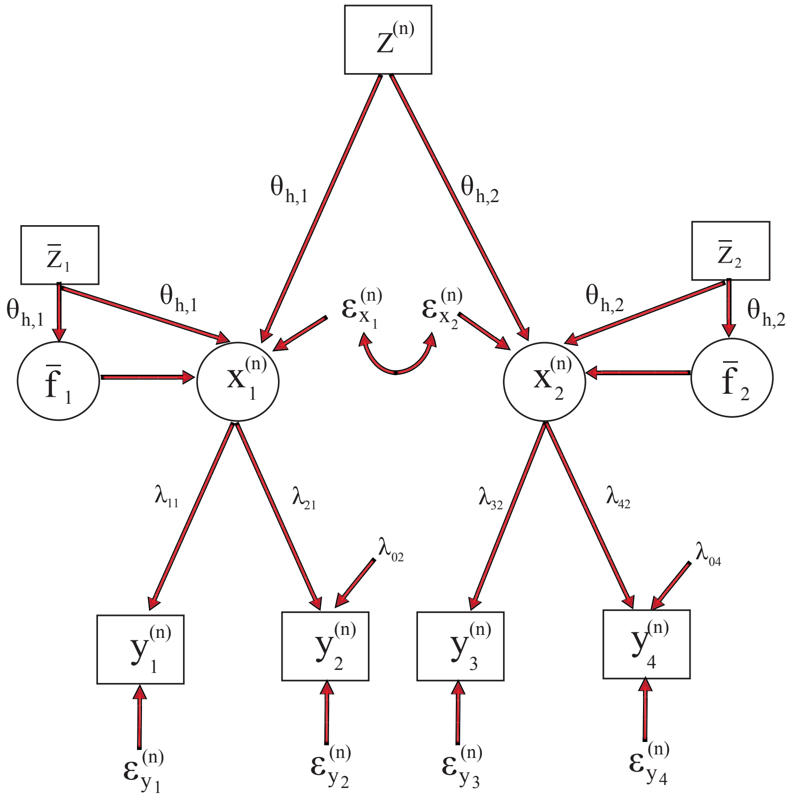


Figure 3.3: An illustrative alternative sparse GP-SEM with 2 latent variable and 4 manifest variables

tions. In other words, the goal is to search the existence of algebraic functional forms between known and unknown parameters. The necessity to examine model identifiability is because the lack may lead inefficient or inconsistent estimation for parameters.

Since the auxiliary pseudo inputs and pseudo latent functions work for computation and are irrelevant to the identification of GP latent function values, we only consider full GP-SEM here. Without specifying the data point, the model structure can be expressed as follows:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\epsilon}_x \tag{3.9}$$

$$\mathbf{y} = \boldsymbol{\Lambda}\mathbf{x} + \boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}_y, \tag{3.10}$$

where $\mathbf{f}(\mathbf{z}) = (f_1(\mathbf{z}), \dots, f_Q(\mathbf{z}))$ is a vector-valued functional between a covariate vector $\mathbf{z}$ and latent variables $\mathbf{x} = (x_1, \dots, x_Q)$. The remaining variables have the same denotations with Equations (3.1)-(3.5).

With the assumption of mutual independence between all measurement errors (given Equation (3.5)), we additionally make the constraints on the intercepts of the anchor variables [6], and variances of $\boldsymbol{\epsilon}_x$. Hence, these constrained parameters are known and the rest are classified as unknown parameters.

Here we take a simple example for demonstrating the model identifiability, where Q, the number of latent variables, is 2; and each latent variable has 3 indicators. With the 1st and 4th being anchor variables of the two latent variable, the unknown parameters are two functional values $f_1(\mathbf{z})$ and $f_2(\mathbf{z})$, all loadings $\lambda_{11}, \lambda_{21}, \lambda_{31}, \lambda_{42}, \lambda_{52}, \lambda_{62}$, intercept terms on non-anchor variables $\lambda_{02}, \lambda_{03}, \lambda_{05}, \lambda_{06}$, measurement error variances $\sigma^2_{y_1}, \dots, \sigma^2_{y_6}$ and conditional covariances of latent variables $\sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_{12}}$; the known parameters are intercept terms on anchor variables $\lambda_{01}, \lambda_{04}$, set to be 0, and conditional auto-covariances $\sigma^2_{x_1}, \sigma^2_{x_2}$, set to 1.

[6] An anchor variable is a variable whose path coefficient (or loading) with latent variables is 1. To select an anchor variable could depend on whether the corresponding latent variable shares the same measurement unit as an item of interest does, or on the biggest amplitude among all loadings through exploratory factor analysis. The anchor variables we used are different from convention to some degrees. In addition, we use a quick-and-dirty method to select an anchor variable for different latent variable models. For each latent variable, look at its observed children and pick the one with the highest coefficient of determination R^2 with respect to covariates. If this child is assigned to some other latent variables, try the next best child and so on.

Under the aforementioned model structure, assumptions and constraints, we can derive the following equations about the first and second moments of conditional distribution of observed variables given covariates (referred to as reduced-form distribution):

$$\begin{aligned}
E(y_1|\mathbf{z}) &= \lambda_{11}f_1(\mathbf{z}), \qquad\quad E(y_4|\mathbf{z}) = \lambda_{42}f_2(\mathbf{z}),\\
E(y_2|\mathbf{z}) &= \lambda_{21}f_1(\mathbf{z}) + \lambda_{02}, \ E(y_5|\mathbf{z}) = \lambda_{52}f_2(\mathbf{z}) + \lambda_{05},\\
E(y_3|\mathbf{z}) &= \lambda_{31}f_1(\mathbf{z}) + \lambda_{03}, \ E(y_6|\mathbf{z}) = \lambda_{62}f_2(\mathbf{z}) + \lambda_{06},
\end{aligned}$$

and

$$\begin{aligned}
Var(y_1|\mathbf{z}) &= \lambda_{11}^2 + \sigma_{y_1}^2, \ Var(y_4|\mathbf{z}) = \lambda_{42}^2 + \sigma_{y_4}^2,\\
Var(y_2|\mathbf{z}) &= \lambda_{21}^2 + \sigma_{y_2}^2, \ Var(y_5|\mathbf{z}) = \lambda_{52}^2 + \sigma_{y_5}^2,\\
Var(y_3|\mathbf{z}) &= \lambda_{31}^2 + \sigma_{y_3}^2, \ Var(y_6|\mathbf{z}) = \lambda_{62}^2 + \sigma_{y_6}^2.
\end{aligned}$$

The cross-covariances are

$$\begin{aligned}
Cov(y_1, y_2|\mathbf{z}) &= \lambda_{11}\lambda_{21}, \qquad\quad Cov(y_4, y_5|\mathbf{z}) = \lambda_{42}\lambda_{52},\\
Cov(y_1, y_3|\mathbf{z}) &= \lambda_{11}\lambda_{31}, \qquad\quad Cov(y_4, y_6|\mathbf{z}) = \lambda_{42}\lambda_{62},\\
Cov(y_2, y_3|\mathbf{z}) &= \lambda_{21}\lambda_{31}, \qquad\quad Cov(y_5, y_6|\mathbf{z}) = \lambda_{52}\lambda_{62},\\
Cov(y_1, y_4|\mathbf{z}) &= \lambda_{11}\lambda_{42}\sigma_{x_{12}}, \ Cov(y_2, y_4|\mathbf{z}) = \lambda_{21}\lambda_{42}\sigma_{x_{12}},\\
Cov(y_1, y_5|\mathbf{z}) &= \lambda_{11}\lambda_{52}\sigma_{x_{12}}, \ Cov(y_2, y_5|\mathbf{z}) = \lambda_{21}\lambda_{52}\sigma_{x_{12}},\\
Cov(y_1, y_6|\mathbf{z}) &= \lambda_{11}\lambda_{62}\sigma_{x_{12}}, \ Cov(y_2, y_6|\mathbf{z}) = \lambda_{21}\lambda_{62}\sigma_{x_{12}},\\
Cov(y_3, y_4|\mathbf{z}) &= \lambda_{31}\lambda_{42}\sigma_{x_{12}}, \ Cov(y_3, y_5|\mathbf{z}) = \lambda_{31}\lambda_{52}\sigma_{x_{12}},\\
Cov(y_3, y_6|\mathbf{z}) &= \lambda_{31}\lambda_{62}\sigma_{x_{12}}.
\end{aligned}$$

Through algebraic manipulations, the mathematical formulae of λ_{11} and λ_{42} are

$$\lambda_{11} = \pm\sqrt{\frac{Cov(y_1, y_2|\mathbf{z})Cov(y_1, y_3|\mathbf{z})}{Cov(y_2, y_3|\mathbf{z})}}, \ \lambda_{42} = \pm\sqrt{\frac{Cov(y_4, y_5|\mathbf{z})Cov(y_4, y_6|\mathbf{z})}{Cov(y_5, y_6|\mathbf{z})}}, \tag{3.11}$$

and thus λ_{11} and λ_{42} become known parameters. Based on the mathematical representation, it is sufficient to derive the functional relationship between the remaining unknown parameters and the known. For example, the conditional cross-covariance of latent variables $\sigma_{x_{12}}$, a loading λ_{31}, an intercept term λ_{03}, a measurement error variance $\sigma_{y_4}^2$ and the function value f_1 are

$$\begin{aligned} \sigma_{x_{12}} &= \frac{Cov(y_1, y_4|\mathbf{z})}{\lambda_{11}\lambda_{42}}, & \lambda_{31} &= \frac{Cov(y_1, y_3|\mathbf{z})}{\lambda_{11}}, \\ \lambda_{03} &= E(y_3|\mathbf{z}) - \lambda_{31}\frac{E(y_1|\mathbf{z})}{\lambda_{11}}, & \sigma_{y_4}^2 &= Var(y_4|\mathbf{z}) - \lambda_{42}^2, \\ f_1(\mathbf{z}) &= \frac{E(y_1|\mathbf{z})}{\lambda_{11}}. \end{aligned} \tag{3.12}$$

It is noted that because the sign of λ_{11} can be different, thereby it would affect the signs of the associated loadings. This scenario happens in λ_{42} and the rest loadings as well. Moreover, the sign differences of λ_{11} and λ_{42} can alter that of conditional cross-covariance of latent variables $\sigma_{x_{12}}$. The possible sign difference is not an issue because signs can always be changed the same as the associated parameters and variables. The parameters having two true values with opposite signs could efficiently converge to one of them if the initial values merely lie in the neighbourhood of the closest true value.

Instead of imposing constrains on conditional auto-covariance of latent variables $\sigma_{x_1}^2$ and $\sigma_{x_2}^2$, one can also use another option of constrains on the loadings (λ_{11} and λ_{42}) of the anchors variables. This does not lead to the situation of two possible estimated values due to sign difference as before. It implies the resulting model structure becomes global identifiable theoretically.

The assumption of independence among all measurement errors can be relaxed too. Here we only take an example for possible relaxations. It is to impose independence between the errors only on anchor variables, and among the errors corresponding to the same latent variable. This still leaves identification condition satisfied. The algebraic relations of the parameters (including factor loadings and intercepts, latent function, measurement error variances and latent error covariance) is formularised as those in Equations (3.11) and (3.12). The cross-covariances of the measurement errors corresponding different latent variables is mathematically represented in terms of known parameters. For example,

$$\begin{aligned} Cov(\boldsymbol{\epsilon}_{y_1}, \boldsymbol{\epsilon}_{y_5}) &= Cov(y_1, y_5|\mathbf{z}) - \lambda_{11}\lambda_{52}\sigma_{x_{12}}, \\ Cov(\boldsymbol{\epsilon}_{y_2}, \boldsymbol{\epsilon}_{y_6}) &= Cov(y_2, y_6|\mathbf{z}) - \lambda_{21}\lambda_{62}\sigma_{x_{12}}. \end{aligned} \tag{3.13}$$

Another analytic examination of model identification is to check the rank of a Jacobian matrix of all reduced-form parameters[7] over unknown parameters. If the rank is equal to

[7] The reduced-form parameters of a reduced-form distribution can characterize itself, such as the first and second moment.

that of unknown parameters, full rank is achieved and thus this can lead to local identifiability of model structure (Skrondal & Rabe-Hesketh 2004). In the previous example, the size of associated Jacobian matrix is 27×19 (here 27 and 19 are the numbers of reduced-form and unknown parameters, respectively) full rank can be derived through elementary matrix multiplication. The derivation process is tedious, so we would not present it here.

3.3 Remarks

Gaussian process (GP) and factor analysis (FA) establish the proposed model structure. The GP framework provides infinite possible classes of functional relationship between covariates and latent variables. The FA model endows the feature to explore distributions of latent variables given the multiple responses.

Through a finite number of pseudo (or inducing) inputs, the model framework reduces the computational cost from $O(N^3Q^3)$ to $O(NM^2Q^2)$. Under the assumption of conditional independence of GP functions, individual GP models (latent variables are regressed on covariates) share the same model structure as the SPGP model of Snelson and Ghahramani (2006*a*).

The examination of model identification can be implemented by using algebraic operations to check whether a functional relationship exists between unknown and known parameters. Another approach to conduct identification check is to calculate the rank of a Jacobian matrix of reduced-form parameters over unknown parameters. Full rank indicates the existence of a one-one map relationship between reduced-form and unknown parameters.

Chapter 4

Computation

This chapter gives computational treatments for sparse GP-SEM, which are applied in experiments afterwards. In Section 4.1, we briefly present a general idea of Monte Carlo Markov Chain methods, and focus on Metropolis-Hastings and Gibbs techniques. In Section 4.2, we provide the sampling scheme for Sparse GP-SEM. Section 4.3 presents improved samplers with efficient computational technique for several parameters and variables. Section 4.4 shortly discusses the expectation maximisation (EM) algorithm and its stochastic variants. Section 4.5 introduces inference function of margin (IFM) approach, which enables to reduce computation further. Section 4.6 provides the hybrid algorithm constituted by EM and IFM. Section 4.7 presents the associated predictive distribution for inference. Section 4.8 proposes a greedy selection scheme for the pseudo-input set, which aims to improve predictive performance. We remark the whole chapter in the final section.

4.1 MCMC sampling methods

Monte Carlo Markov Chain (MCMC) methods are literally Monte Carlo integration techniques (about numerical stochastic integration) using Markov Chains (a series of random variables with Markov property[1]). The methods aim to solve two practical computing problems - generating samples from a probability distribution of interest, and evaluating

[1] Denote time-indexed random variables $\nu^{[i]}, i = 1, 2, \ldots$; if they have the Markov property, then given the present state $\nu^{[i]}$, the conditional probability distribution of the future state $\nu^{[I+1]}$ only depends upon $\nu^{[I]}$ and is independent of the past states $\nu^{[I]}, i = 1, \ldots, I-1$.

the expectation of function under the distribution.

The idea to solve the above problems is to construct a Markov chain satisfying some conditions (aperiodic, positive recurrent and reversible) (Roberts 1995, Neal 1993). The constructed chain can guarantee the conditional distribution of the present state (given the initial state) eventually converges to the target distribution of interest after time transitions. Thereby successive correlated samples can be obtained from the target distribution beyond a certain time threshold, referred to as burn-in. The expectation of a function can thus be estimated by averaging all the function values evaluated on the after-burn-in samples. This estimator is so-called ergodic average and its approximation precision can be improved by increasing the size of the samples.

Various MCMC methods originate from the framework of Metropolis et al. (1953) and Hastings (1970), commonly referred to as the Metropolis-Hastings (MH) algorithm. Given a draw $\nu^{[1]}$ from the initial distribution, the next draw $\nu^{[2]}$ is obtained through a proposal distribution $g(\cdot|\nu^{[1]})$, and by this way one can produce a consecutive chain $\{\nu^{[1]}, \nu^{[2]}, \ldots\}$. The proposed draw ν given $\nu^{[i]}$ is a candidate point of $\nu^{[i+1]}$. It is possibly identical to $\nu^{[i]}$ if a uniform random number (between 0 and 1) is greater than the acceptance probability (or acceptance ratio)

$$r(\nu^{[i]}, \nu) = \min\left(1, \frac{h(\nu)g(\nu^{[i]}|\nu)}{h(\nu^{[i]})g(\nu|\nu^{[i]})}\right), \tag{4.1}$$

where $h(\cdot)$ is the target distribution or the distribution of interest. Note that in this case the candidate ν is rejected.

This ratio (4.1) can derive the detailed balance condition

$$h(\nu^{[i]})T(\nu^{[i+1]}; \nu^{[i]}) = h(\nu^{[i+1]})T(\nu^{[i]}; \nu^{[i+1]}), \tag{4.2}$$

here $T(\nu^{[i+1]}; \nu^{[i]})$ is the transition kernel proposing a probability from the state of $\nu^{[i]}$ to that of $\nu^{[i+1]}$. Equation (4.2) can derive that the target distribution $h(\cdot)$ is the stationary distribution (or invariant distribution) of the Markov chain by integrating both sides with respect to $\nu^{[i]}$

$$h(\nu^{[i+1]}) = \int h(\nu^{[i]})T(\nu^{[i+1]}; \nu^{[i]})d\nu^{[i]}. \tag{4.3}$$

It reveals to construct a transition kernel of a Markov chain holding detailed balance condition (4.2) is what MCMC researchers desire (Gilks et al. 1995, Neal 1993).

In multiple-dimension state space ($dim(\boldsymbol{\nu}) > 1$), the transition kernel is constructed by combining several base transition kernels where each holds detailed balance condition (Neal 1993).

A proposal distribution $g(\cdot|\cdot)$ in principle can be any probability distribution. However with close approximation to the target distribution, a proposal can enhance simulation mixing. In practice, a simple probability density may be experimentally used by tuning the involving parameters (Gilks et al. 1995). This may help one to realize the mixing for constructing a posteriori and appropriate proposal. Metropolis et al. (1953) initially considered a symmetric proposal having the form $g(\boldsymbol{\nu}|\boldsymbol{\omega}) = g(\boldsymbol{\omega}|\boldsymbol{\nu})$. For example, a multivariate normal distribution with a given state $\boldsymbol{\nu}$ as mean and a fixed covariance matrix $\boldsymbol{\Sigma}$ can propose a new state $\boldsymbol{\omega}$ to update the whole components of $\boldsymbol{\nu}$, or a simple normal proposal can be used to update each component of $\boldsymbol{\nu}$. Then the acceptance probability here is

$$r(\boldsymbol{\nu}, \boldsymbol{\omega}) = \min\left(1, \frac{h(\boldsymbol{\omega})}{h(\boldsymbol{\nu})}\right). \tag{4.4}$$

Hastings (1970) provides an insight of non-symmetric proposal distribution as generalisation of Metropolis algorithm. In this case, the probability evaluation of a candidate given current state is necessary and the acceptance ratio is exactly the one in Equation (4.1).

The multiple block Metropolis-Hastings method (Press 2003) or single component Metropolis-Hastings method (Gilks et al. 1995) is a generalized MH method. It is often used in high-dimension state space for tackling the difficulty of slow convergence to target distribution. It is actually cyclically to implement the MH method for each block through individual proposal distributions. And the updating block is drawn from the proposal conditioned on the updated and the other blocks. More specifically, to generate a new state vector from $\boldsymbol{\nu}^{[i]} = (\boldsymbol{\nu}_1^{[i]}, \boldsymbol{\nu}_2^{[i]}, \ldots, \boldsymbol{\nu}_J^{[i]})$ needs to update the J blocks in order. For $1 \leq j \leq J$, updating the j-th block MH method is implemented by the j-th proposal distribution $g_j(\boldsymbol{\omega}_j|\boldsymbol{\nu}_j^{[i]}, \boldsymbol{\nu}_{\setminus j}^{[i]})$, where $\boldsymbol{\omega}_j$ is a candidate of the j-th block $\boldsymbol{\nu}_j^{[i]}$ and $\boldsymbol{\nu}_{\setminus j}^{[i]} = (\boldsymbol{\nu}_1^{[i+1]}, \ldots, \boldsymbol{\nu}_{j-1}^{[i+1]}, \boldsymbol{\nu}_{j+1}^{[i]}, \ldots, \boldsymbol{\nu}_J^{[i]})$ denotes the vector comprising all the updated and the remaining blocks except the j-th one. Then the corresponding acceptance probability is

$$r(\boldsymbol{\nu}_j^{[i]}, \boldsymbol{\omega}_j; \boldsymbol{\nu}_{\setminus j}^{[i]}) = \min\left(1, \frac{h(\boldsymbol{\omega}_j|\boldsymbol{\nu}_{\setminus j}^{[i]}) g_j(\boldsymbol{\nu}_j^{[i]}|\boldsymbol{\omega}_j, \boldsymbol{\nu}_{\setminus j}^{[i]})}{h(\boldsymbol{\nu}_j^{[i]}|\boldsymbol{\nu}_{\setminus j}^{[i]}) g_j(\boldsymbol{\omega}_j|\boldsymbol{\nu}_j^{[i]}, \boldsymbol{\nu}_{\setminus j}^{[i]})}\right), \tag{4.5}$$

where $h(\boldsymbol{\nu}_j^{[i]}|\boldsymbol{\nu}_{\backslash j}^{[i]})$ is evaluated at $\boldsymbol{\nu}^{[i]}$ by the conditional density $h(\boldsymbol{\nu}_j|\boldsymbol{\nu}_{\backslash j})$ called the full conditional distribution for $\boldsymbol{\nu}_j$ under the target distribution $h(\boldsymbol{\nu})$. Because the full conditional distribution $h(\boldsymbol{\nu}_j|\boldsymbol{\nu}_{\backslash j})$ is proportional to the joint distribution $h(\boldsymbol{\nu}_j, \boldsymbol{\nu}_{\backslash j})$ by a normalising constant[2], the acceptance probability in (4.5) has the same expression as the target distribution. The use of full conditional densities not only avoids calculating normalising constant but also affords specifying and deriving the target distribution (Gilks 1995).

Gibbs sampling (GS) algorithm (Press 2003, Gelman et al. 2004) is a special class of multiple block MH methods. Because the full conditional distribution for $\boldsymbol{\nu}_j$ is the corresponding the j-th proposal, the acceptance probability in (4.5) becomes 1. This implies the proposed candidates are never rejected. Due to convenient random sampling from full conditional distributions, the GS approach is rather easily applied and has computing efficiency. However, if the full conditional density is not in standard exponential family, it may be inappropriate to use. In addition, it may be necessary to implement GS along with techniques, such as parameter expansion (Gelman et al. 2005, Liu & Wu 1999), reparametrisation (Gilks & Roberts 1995), to enhance mixing for sampling highly correlated variables.

Besides the above approaches, dynamical sampling algorithms, including the Hybrid Monte Carlo method (Neal 2010), propose a candidate point by a discretised dynamical system. These sophisticated schemes avoid random walk behaviour and can have fast convergence to the target distribution in some problems (see Neal (1993, 2000)).

4.2 Samplers

The Gibbs sampling (GS) and the Metropolis-Hastings (MH) schemes are used mixedly for sampling variables and parameters of sparse GP-SEM[3]. To be specific, the MH method is applied for sampling the hyper-parameters of the GP covariance functions, pseudo inputs and pseudo functions. And the GS scheme for the remainder of model parameters (including pseudo functions). For computational convenience, we adopt a typical conjugate prior distribution for most of sampled variables. More details of the prior we use can be

[2] This normalising constant is formed by integrating out $\boldsymbol{\nu}_j$ from the joint distribution. It is a function of $\boldsymbol{\nu}_{\backslash j}$ and ensures the integral value of the full conditional distribution over $\boldsymbol{\nu}_j$ is 1.

[3] Here we only treat Sparse GP-SEM (given by Equations (3.1)-(3.7)) because full GP-SEM (defined by Equations (3.1)-(3.5)) is its special case.

found in the text and Appendix A.5.

The derivations of full conditional distributions are based on the model structure and the distributional assumptions of sparse GP-SEM. Here we directly write the outcomes, some of derivations are provided in Appendix A.3.

4.2.1 Sampling Hyper-parameters of Covariance Function, Pseudo Inputs and Functions

Hyper-Parameters

For each q, we sample each component $\theta_{h,qj}$ of hyper-parameters $\boldsymbol{\theta}_{h,q}$ of the GP covariance function in turn from its non-canonical full conditional distribution. Once the new value for the j-th component is accepted by the MH method, then the hyper-parameters are updated for the next component.

Define a pseudo-function set $\bar{\mathbf{f}}_q^{1:M} \equiv \{\bar{f}_q^{(1)}, \ldots, \bar{f}_q^{(M)}\}$, a latent-variable set $\mathbf{x}_q^{1:N} \equiv \{x_q^{(1)}, \ldots, x_q^{(N)}\}$. We use a uniform proposal density over an interval $[a_w\theta_{h,qj}, (1/a_w)\theta_{h,qj}]$ with the pre-specified width parameter a_w $(0 < a_w < 1)$ to control the moving step of a candidate ν[4]. Then for the i-th sampling step the acceptance probability $r(\theta_{h,qj}^{[i]}, \nu)$ is given by

$$r(\theta_{h,qj}^{[i]}, \nu) = \min\left(1, \frac{h_{qj}(\nu)g(\theta_{h,qj}^{[i]}|\nu)}{h_{qj}(\theta_{h,qj}^{[i]})g(\nu|\theta_{h,qj}^{[i]})}\right), \tag{4.6}$$

where

$$h_{qj}(\theta_{h,qj}) = \pi_q(\theta_{h,qj}) \cdot p(\mathbf{x}_q^{1:N}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M}, \bar{\mathbf{f}}_q^{1:M}, \boldsymbol{\theta}_{h,q}) \cdot p(\bar{\mathbf{f}}_q^{1:M}|\bar{\mathbf{z}}_q^{1:M}, \boldsymbol{\theta}_{h,q}) \tag{4.7}$$

is the full conditional of $\theta_{h,qj}$ integrating out latent functions $\mathbf{f}_q^{1:N}$. $\pi_q(\theta_{h,qj})$ denotes the prior density[5] for θ_{iq}; $g(\theta_{h,qj}|\nu)$ is the proposal density of the current value $\theta_{h,qj}$ given a new state ν. Based on the aforementioned specification, the proposal density is an uniform distribution over $[a_w\nu, (1/a_w)\nu]$.

[4] In practice, the proposed values of the GP hyper-parameters can be all positive or negtive. This is because we use a logarithmic scale for computational convenience, and thereby the initial and successive values are so.

[5] Here we adopt a mixture of a gamma(1,20) and a gamma(10, 10) with equal probability for each density; the prior and the proposed distribution here are also used by Silva and Gramacy (2010) for modelling non-parametric regression between latent variables.

The calculation of the ratio $r(\theta_{h,qj}^{[i]}, \nu)$ can be facilitated through simple manipulations. First we use log-transformation, then sum all the terms (calculated from the individual densities) and finally recover by adopting exponential transformation. In addition, the factor $p(\mathbf{x}_q^{1:N}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M}, \bar{\mathbf{f}}_q^{1:M}, \boldsymbol{\theta}_{h,q})$ is obtained based on the alternative form of sparse GP-SEM[6]. And the determinant of the involved covariance matrix can be evaluated by using the matrix identity (A.2).

Pseudo inputs and pseudo latent functions

We sample pseudo inputs $\bar{\mathbf{z}}_q^{1:M}$ and the associated pseudo functions $\bar{\mathbf{f}}_q^{1:M}$ jointly by the multiple block MH method.

In brief, the sampling algorithm implements c sampling steps for updating. Before the implementations, the initial pseudo-input set $\bar{\mathbf{z}}_q^{1:M}$ is drawn uniformly at random without replacement, from the covariate set $\mathbf{z}^{1:N}$. Then, in each step we randomly select one member from $\bar{\mathbf{z}}_q^{1:M}$, that is, $\bar{\mathbf{z}}_q^{(m)}$ $(1 \leq m \leq M)$ as a updating pseudo input. We next choose another input $\boldsymbol{\nu}$ from the complement set of the pseudo inputs in the current step, as a candidate of $\bar{\mathbf{z}}_q^{(m)}$. Furthermore, we propose a new pseudo function value $\bar{f}_q^{\boldsymbol{\nu}}$ (this denotation merely notes the proposed input $\boldsymbol{\nu}$), through a proposal distribution given the rest of the pseudo inputs and pseudo functions. The proposed pair $(\boldsymbol{\nu}, \bar{f}_q^{\boldsymbol{\nu}})$ is accepted if an uniform random number is smaller than the associative acceptance probability.

More specifically, the proposal distribution $g(\cdot|\bar{\mathbf{z}}_q^{(m)}, \bar{\mathbf{z}}_q^{\backslash m}, \bar{\mathbf{f}}_q^{\backslash m})$ is an univariate Gaussian with a mean

$$[\mathbf{K}_{q;M}]_{m,\backslash m}([\mathbf{K}_{q;M}]_{\backslash m,\backslash m})^{-1}[\bar{\mathbf{f}}_q]_{\backslash m} \tag{4.8}$$

and a variance

$$[\mathbf{K}_{q;M}]_{m,m} - [\mathbf{K}_{q;M}]_{m,\backslash m}([\mathbf{K}_{q;M}]_{\backslash m,\backslash m})^{-1}[\mathbf{K}_{q;M}]_{\backslash m,m}, \tag{4.9}$$

where $\backslash m$ denotes the complement set consisting of all pseudo inputs except the m-th one and thus $\bar{\mathbf{z}}_q^{\backslash m} \equiv \{\bar{\mathbf{z}}_q^{(1)}, \ldots, \bar{\mathbf{z}}_q^{(m-1)}, \bar{\mathbf{z}}_q^{(m+1)}, \ldots, \bar{\mathbf{z}}_q^{(M)}\}$, $\bar{\mathbf{f}}_q^{\backslash m} \equiv \{\bar{f}_q^{(1)}, \ldots, \bar{f}_q^{(m-1)}, \bar{f}_q^{(m+1)}, \ldots, \bar{f}_q^{(M)}\}$. $\bar{\mathbf{f}}_q$ is a column vector consisting of the set $\bar{\mathbf{f}}_q^{1:M}$. Note that this proposal serves as the conditional distribution of $\bar{f}_q^{(m)}$ given $\bar{\mathbf{z}}_q^{(m)}, \bar{\mathbf{z}}_q^{\backslash m}, \bar{\mathbf{f}}_q^{\backslash m}$. It is also to calculate the density of a current value $\bar{f}_q^{(m)}$.

[6] That is represented by Equations (3.2), (3.4), (3.5), (3.7) and (3.8).

For the i-th sampling step, the sampling scheme has an acceptance probability

$$r((\bar{\mathbf{z}}_q^{(m),[i]}, f_q^{(m),[i]}), (\boldsymbol{\nu}, f_q^{\boldsymbol{\nu}})) = \min\left(1, \frac{h_q(\bar{\mathbf{f}}_q^{1:\boldsymbol{\nu}:M,[i]})g(\bar{f}_q^{(m),[i]}|\bar{\mathbf{z}}_q^{(m),[i]}, \bar{\mathbf{z}}_q^{\backslash m,[i]}, \bar{\mathbf{f}}_q^{\backslash m,[i]})}{h_q(\bar{\mathbf{f}}_q^{1:M,[i]})g(\bar{f}_q^{\boldsymbol{\nu}}|\boldsymbol{\nu}, \bar{\mathbf{z}}_q^{\backslash m,[i]}, \bar{\mathbf{f}}_q^{\backslash m,[i]})}\right), \tag{4.10}$$

where

$$h_q(\bar{\mathbf{f}}_q^{1:M}) = p(\mathbf{f}_q^{1:N}|\bar{\mathbf{f}}_q^{1:M}, \mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M}, \boldsymbol{\theta}_{h,q}) \cdot p(\bar{\mathbf{f}}_q^{1:M}|\bar{\mathbf{z}}_q^{1:M}, \boldsymbol{\theta}_{h,q}) \tag{4.11}$$

is the full conditional distribution of $\bar{\mathbf{f}}_q^{1:M}$. The value $h_q(\bar{\mathbf{f}}_q^{1:\boldsymbol{\nu}:M,[i]})$ in that ratio is evaluated at $\bar{\mathbf{f}}_q^{1:\boldsymbol{\nu}:M,[i]}$ (which denotes the m-th element of $\bar{\mathbf{f}}_q^{1:M,[i]}$ replaced by $\bar{f}_q^{\boldsymbol{\nu}}$) given inputs $\mathbf{z}^{1:N}$, the current GP hyper-parameters $\boldsymbol{\theta}_{h,q}$, current latent functions $\mathbf{f}_q^{1:N}$, the i-th-sampling-step pseudo inputs $\bar{\mathbf{z}}_q^{1:M,[i]}$, pseudo functions $\bar{\mathbf{f}}_q^{1:M,[i]}$ and the current GP hyper-parameters $\boldsymbol{\theta}_{h,q}$. Similarly, $h_q(\bar{\mathbf{f}}_q^{1:M,[i]})$ is evaluated at $\bar{\mathbf{f}}_q^{1:M,[i]}$ given $\mathbf{z}^{1:N}$, the current $\boldsymbol{\theta}_{h,q}$ and $\mathbf{f}_q^{1:N}$, the i-th-sampling-step proposed pseudo inputs $\bar{\mathbf{z}}_q^{1:\boldsymbol{\nu}:M,[i]}$ and proposed pseudo functions $\bar{\mathbf{f}}_q^{1:\boldsymbol{\nu}:M,[i]}$. In addition, $g(\cdot|\boldsymbol{\nu}, \bar{\mathbf{z}}_q^{\backslash m}, \bar{\mathbf{f}}_q^{\backslash m})$ has a mean and variance like those in Equations (4.8) and (4.9), where the elements of the covariance matrix associated with $\bar{\mathbf{z}}_q^{(m)}$ are evaluated at its candidate $\boldsymbol{\nu}$.

Note that here implementing the sampling scheme for multiple times is to acquire a "good" pseudo input set. We expect that the locations of the pseudo inputs can be diffuse enough to capture the prominent features of regression relationship between the original input sets and a latent variable.

Pseudo latent functions

Although updating pseudo inputs and pseudo functions jointly via the above sampling scheme, we consider to re-update pseudo functions by Gibbs sampling for enhancing the mixing. Basically, the new values $\bar{\mathbf{f}}_{\mathbf{q}}^{1:M}$ are drawn from their full conditional distribution $h_q(\cdot)$ in (4.11). It is noted that the full conditional is normally distributed with a covariance matrix

$$\boldsymbol{\Sigma}_{\bar{f}_q,post} \equiv (\mathbf{K}_{q;M}^{-1} + \mathbf{K}_{q;M}^{-1}\mathbf{K}_{q;MN}\mathbf{V}_q^{-1}\mathbf{K}_{q;MN}^{\mathsf{T}}\mathbf{K}_{q;M}^{-1})^{-1} \tag{4.12}$$

and a mean vector

$$\boldsymbol{\Sigma}_{\bar{f}_q,post}\mathbf{K}_{q;M}^{-1}\mathbf{K}_{q;MN}\mathbf{V}_q^{-1}\mathbf{f}_q. \tag{4.13}$$

The derivations of Equations (4.12) and (4.13) are placed in Appendix A.3.1. The computation of sampling $\bar{\mathbf{f}}_q^{1:M}$ is not an issue because the inversion of the $N \times N$ matrix $\mathbf{V}_q$

(defined in (3.6)) can be easily achieved due to its diagonal structure. $\mathbf{f}_q$ is a column vector consisting of the set $\mathbf{f}_q^{1:N}$.

4.2.2 Sampling Latent Variables and Latent Functions

Latent variables

In principle, we can sample all latent variables $\{\mathbf{x}_1^{1:N}, \ldots, \mathbf{x}_Q^{1:N}\}$ conditioning on all other variables and data. However, to improve mixing, we analytically marginalize out latent functions $\{\mathbf{f}_1^{1:N}, \ldots, \mathbf{f}_Q^{1:N}\}$ from the full conditional distribution. Hence, the resulting conditional density is the full conditional distribution of latent variables under the alternative sparse GP-SEM model structure. The conditional is a Gaussian distribution and Gaussianity is derived from the multiplication of two Gaussian densities (one from the measurement model represented by Equations (3.4) and (3.5), one from the sparse GP formulation given Equation (3.8)), by the identity (A.5). Then the sampling distribution of latent variable $\mathbf{x}^{(n)} = (x_1^{(n)}, \ldots, x_Q^{(n)})^\mathsf{T}$ has a covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{x}^{(n)}} = \left[\boldsymbol{\Lambda}^\mathsf{T}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Lambda} + (\mathbf{V}^{(n)} + \boldsymbol{\Sigma}_x)^{-1}\right]^{-1}, \tag{4.14}$$

where $\mathbf{V}^{(n)}$ is a $Q \times Q$ diagonal matrix consisting of the nn-th entry of $\mathbf{V}_q$ over all q; and a mean

$$\boldsymbol{\mu}_{\mathbf{x}^{(n)}} = \boldsymbol{\Sigma}_{\mathbf{x}^{(n)}}\left[\boldsymbol{\Lambda}^\mathsf{T}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}^{(n)} - \boldsymbol{\lambda}_0) + (\mathbf{V}^{(n)} + \boldsymbol{\Sigma}_x)^{-1}\widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\bar{\mathbf{f}}\right], \tag{4.15}$$

where $\widetilde{\mathbf{K}}_{nM}$ and $\widetilde{\mathbf{K}}_M$ are the block diagonal matrices consisting of all the n-th row of $\mathbf{K}_{q;NM}$ and with the matrix $\mathbf{K}_{q;M}$, across q, respectively. $\bar{\mathbf{f}}$ is a column vector whose elements are all pseudo latent functions $\{\bar{\mathbf{f}}_\mathbf{1}^{1:M}, \ldots, \bar{\mathbf{f}}_\mathbf{Q}^{1:M}\}$. See the derivations in Appendix A.3.2.

Latent functions

For all latent functions $\{\mathbf{f}_1^{1:N}, \ldots, \mathbf{f}_Q^{1:N}\}$, a new sample can be drawn explicitly from the full conditional distribution, which is normally distributed according to the original model structure of Spare GP-SEM represented by Equations (3.1)-(3.7). The conditional density of $\mathbf{f}^{(n)} = (f_1^{(n)}, \ldots, f_Q^{(n)})^\mathsf{T}$ has a covariance

$$\boldsymbol{\Sigma}_{\mathbf{f}^{(n)}} = \left[\boldsymbol{\Sigma}_x^{-1} + (\mathbf{V}^{(n)})^{-1}\right]^{-1}, \tag{4.16}$$

and a mean

$$\boldsymbol{\mu}_{\mathbf{f}^{(n)}} = \boldsymbol{\Sigma}_{\mathbf{f}^{(n)}} \big[\boldsymbol{\Sigma}_x^{-1}\mathbf{x}^{(n)} + (\mathbf{V}^{(n)})^{-1}\widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\overline{\mathbf{f}}\big]. \tag{4.17}$$

Again the normality is derived from two normality densities from Equations (3.1)-(3.2) and Equation (3.6) respectively by using the Gaussian identity (A.5). The derivations of (4.16) and (4.17) can be found in Appendix and A.3.3.

Here the sampling of latent functions $\{\mathbf{f}_1^{1:N}, \dots, \mathbf{f}_Q^{1:N}\}$ is used only as an intermediate step for more complicated sampling schemes of other variables.

4.2.3 Sampling Factor Loadings and Covariance Matrix of Measurement Errors

Factor loadings

The sampling scheme of factor loadings and intercept terms is identical to the case of classical Bayesian linear regression. For $1 \leq q \leq Q$ and $1 \leq r \leq R$, let λ_{qr} and λ_{0r} are the qr-th and r-th elements of factor loading matrix $\boldsymbol{\Lambda}$ and of intercepts $\boldsymbol{\lambda}_0$, respectively. Considering a Gaussian prior, the full conditional density of $(\lambda_{qr}, \lambda_{0r})$ is distributed normally with a covariance matrix

$$\boldsymbol{\Sigma}_{\lambda_r,post} \equiv \big(\frac{1}{\sigma_\lambda^2}\mathbf{I}_{|\mathcal{P}_r|} + \frac{1}{\sigma_{y_r}^2}[\widetilde{\mathbf{X}}^\mathsf{T}\widetilde{\mathbf{X}}]_{\mathcal{P}_r,\mathcal{P}_r}\big)^{-1}, \tag{4.18}$$

where σ_λ^2 denotes the prior variance of λ_{qr} and of λ_{0r}, $\mathbf{I}_{|\mathcal{P}_r|}$ is an identity matrix with the cardinality of the set $\mathcal{P}_r$; $\widetilde{\mathbf{X}} \equiv [\mathbf{x}_1, \dots, \mathbf{x}_Q, \mathbf{1}_N]$, here $\mathbf{x}_q$ is the column-wise rearrangement of $\mathbf{x}_q^{1:N}$ and $\mathbf{1}_N \equiv [1, \dots, 1]^\mathsf{T}$ with N entries; $\sigma_{y_r}^2$ is the rr-th element of noise covariance matrix $\boldsymbol{\Sigma}_y$. $\mathcal{P}_r$ denotes the parent set of the r-th indicator and the r-th component of the intercept vector, which indicates the associative indices of the latent variables and the constant 1. So the parent index of λ_{0r} is always $Q+1$, which corresponds to $\mathbf{1}_N$ in $\widetilde{\mathbf{X}}$. The expression of $[\cdot]_{\mathcal{P}_r,\mathcal{P}_r}$ denotes the sub-matrix whose entries are the ones of the original matrix corresponding to the elements of the mutual pair of the parent set. The sampler mean is

$$\frac{1}{\sigma_{y_r}^2}\boldsymbol{\Sigma}_{\lambda_r,post}[\widetilde{\mathbf{X}}^\mathsf{T}]_{\mathcal{P}_r,\cdot}\mathbf{y}_r \tag{4.19}$$

where $\mathbf{y}_r$ is a $N \times 1$ column vector with the r-th indicator of all the data points. $[\cdot]_{\mathcal{P}_r,\cdot}$ means the sub-matrix by extracting the rows from the original matrix, corresponding to the set $\mathcal{P}_r$. Readers can see Appendix A.3.4 for the derivations of (4.18)-(4.19).

Covariance matrix of errors

Sampling measurement errors $\mathbf{\Sigma}_y$ can be implemented from their individual full conditional distributions under the independence assumption imposed in Equation (3.5). Each conditional follows an inverse Gamma distribution when a conjugate prior is used. Then each diagonal element $\sigma^2_{y_r}$ of $\mathbf{\Sigma}_y$ has a distribution as

$$\sigma^2_{y_r}|e.e. \sim \mathcal{I}G(a_0 + \frac{N}{2}, b_0 + \sum_{n=1}^{N}(y_r^{(n)} - \lambda_{0,r} - \lambda_{j(r),r}x_{j(r)}^{(n)})^2) \tag{4.20}$$

where $\mathcal{IG}(a, b)$ represents an inverse gamma distribution with shape parameter a and scale parameter b. a_0 and b_0 are hyper-parameters of the prior of $\sigma^2_{y_r}$, $j(r)$ denotes the index of the latent variable corresponding to r-th indicator. Appendix A.3.5 provides the short derivation of (4.20).

4.2.4 Sampling the Correlation Matrix of GP Noise and Algorithm Summary

We sample the covariance matrix $\mathbf{\Sigma}_x$ by adopting the efficient Bayesian approach of (Talhouk et al. 2012), which involves sampling correlation matrix of multivariate Gaussian latent variables. $\mathbf{\Sigma}_x$ here is a correlation matrix as well because of constraining the variances of error terms ϵ_{x_q} being 1's. The utilisation of that sampling scheme satisfies identification condition and meanwhile can improve convergence of the correlation coefficients.

Instead of updating $\mathbf{\Sigma}_x$ directly, we sample a covariance matrix $\mathbf{\Sigma}_s$ under the factorization

$$\mathbf{\Sigma}_s = \mathbf{D}_s\mathbf{\Sigma}_x\mathbf{D}_s. \tag{4.21}$$

Then given the current correlation matrix, we sample $[\mathbf{D}_s]_{q,q}$ from the conditional density

$$[\mathbf{D}_s]_{q,q}\Big|\mathbf{\Sigma}_x \sim \mathcal{I}G(\frac{Q+1}{2}, \frac{\rho_{qq}}{2}) \tag{4.22}$$

where ρ_{qq} represents the qq-th entry of $(\mathbf{\Sigma}_x)^{-1}$. Employing this conditional density is because, with a marginally uniform prior for $\mathbf{\Sigma}_x$, one can derive an inverse-Wishart prior for the covariance matrix $\mathbf{\Sigma}_s$ (Barnard. et al. 2000), namely

$$\mathbf{\Sigma}_s \sim \mathcal{I}W(2, \mathbf{I}_Q). \tag{4.23}$$

All the sampled $[\mathbf{D}_s]_{q,q}$ serve as a expansion parameter to transform $\mathbf{E}_x$ to $\mathbf{W}_0 = \mathbf{D}_s\mathbf{E}_x$, where $\mathbf{E}_x$ is a $Q \times N$ residual matrix with each column being $\mathbf{x}^{(n)} - \mathbf{f}^{(n)}$. Here $\mathbf{x}^{(n)} = (x_1^{(n)}, \dots, x_Q^{(n)})^\mathsf{T}$ and $\mathbf{f}^{(n)} = (f_1^{(n)}, \dots, f_Q^{(n)})^\mathsf{T}$.

Then given $\mathbf{W}_0$, $\boldsymbol{\Sigma}_s$ can be sampled from the conditional distribution

$$\boldsymbol{\Sigma}_s \Big| \mathbf{W}_0 \sim \mathcal{I}W(2 + N, \mathbf{W}_0\mathbf{W}_0^\mathsf{T} + \mathbf{I}_Q), \tag{4.24}$$

where $\mathcal{IW}(\nu, \boldsymbol{\Psi})$ represents an inverse Wishart distribution with degree of freedom ν and the inverse scale matrix $\boldsymbol{\Psi}$ which is positive definite. The sampled covariance matrix can be projected back to the correlation matrix $\boldsymbol{\Sigma}_x$ by $\boldsymbol{\Sigma}_x = \mathbf{D}_s^{-1}\boldsymbol{\Sigma}_s\mathbf{D}_s^{-1}$. The brief derivation of (4.24) can be found in Appendix A.3.6.

Algorithm 1

For each MCMC iteration, the sampling scheme can be summarised as follows:

- Call a sampler of hyper-parameters of covariance functions, computing based on Equations (4.6) and (4.7).
- Call a sampler of latent variables using (4.14) and (4.15).
- Call a sampler of latent functions first using (4.16) and (4.17) and then sample pseudo inputs set and pseudo latent functions jointly using (4.8)-(4.11). Next re-sample pseudo latent functions using (4.12) and (4.13) after re-call a sampler of latent functions.
- Call a sampler of latent functions first and then call a sampler of correlation matrix of GP error terms using (4.21)-(4.24).
- Call a sampler of factor loadings using (4.18) and (4.19).
- Call a sampler of variances of measurement errors using (4.20).

Although we adopt a fixed updating order, the order per MCMC iteration can be a random permutation in principle. One can also update one of the above items with a fixed probability (Gilks et al. 1995).

4.3 Improved Samplers

From the features and practices of Algorithm 1, a few possible drawbacks had been noticed. The first problem is that we do not clear know the necessary number of implementing sampling scheme to acquire a pseudo input set with expected characteristics. Even using a medium number of steps, the computing process may be time-consuming as the dataset size increases. The second could be the intermediate step of sampling latent functions. It is implemented in the sampling procedures of several variables and therefore makes the algorithm complicated somewhat. The last one is that strong correlation between latent variables and factor loadings may cause mixing-slowly simulation chains.

Despite those potential issues, we can still modify the sampler for improvement. To specify it, the first problem can be mitigated by fixing the initial selected pseudo inputs set. The second and the third problems can be solved by modifying the associated samplers and introducing an extra parameter to improve mixing.

It can be reckoned that the original sampling procedure of pseudo latent functions basically does not lead a computational issue. However, we are still able to modify the procedure by merely selecting values from the latent functions. This fashion further enable to save some time.

In the following paragraphs, we only present the improved samplers for some parameters. For the rest, the samplers are the same as before and thus would not be mentioned. The associated derivations are placed in Appendix A.4.

4.3.1 Samplers of Latent variables and Latent Functions

Latent variables

Let $\mathbf{x}$, $\mathbf{f}$ and $\bar{\mathbf{f}}$ be $\{\mathbf{x}_1^{1:N}, \ldots, \mathbf{x}_Q^{1:N}\}$ be column vectors whose elements belong to the set of all latent variables $\{\mathbf{x}_1^{1:N}, \ldots, \mathbf{x}_Q^{1:N}\}$, latent functions $\{\mathbf{f}_1^{1:N}, \ldots, \mathbf{f}_Q^{1:N}\}$, and pseudo functions $\{\bar{\mathbf{f}}_1^{1:M}, \ldots, \bar{\mathbf{f}}_Q^{1:M}\}$. Then their sizes are $NQ \times 1$, $NQ \times 1$ and $MQ \times 1$. The collection of all pseudo input sets is denoted as $\bar{\mathbf{z}}_{1:Q}^{1:M} = \{\bar{\mathbf{z}}_1^{1:M}, \ldots, \bar{\mathbf{z}}_Q^{1:M}\}$.

In principle, we can sample $\mathbf{x}$ conditioning on all other variables and data. But, different from the sampler using (4.14) and (4.15), the modified sampler is the conditional distribution that $\mathbf{f}$ and $\bar{\mathbf{f}}$ are analytically marginalized out from the full conditional distribution. That conditional distribution has the probability density $p(\mathbf{x}|e.e. \setminus \mathbf{f}, \setminus\bar{\mathbf{f}})$, where everything else is abbreviated to *e.e.*.

This conditional density is proportional to the multiplication of two Gaussian densities. One is $p(\mathbf{x}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}^{1:M}_{1:Q}, \boldsymbol{\Theta}_h, \boldsymbol{\Sigma}_x)$, from the distribution represented by (3.8) and integrating out pseudo functions; the other is $p(\mathbf{Y}|\mathbf{x}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_y)$, from the measurement model represented by (3.4)-(3.5). They respectively have covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$,

$$\boldsymbol{\Sigma}_0 = \widetilde{\mathbf{K}}^{\mathsf{T}}_{MN}\widetilde{\mathbf{K}}^{-1}_{M}\widetilde{\mathbf{K}}_{MN} + \widetilde{\mathbf{V}} + \boldsymbol{\Sigma}_x \otimes \mathbf{I}_N, \tag{4.25}$$

$$\boldsymbol{\Sigma}_1 = (\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Lambda}) \otimes \mathbf{I}_N, \tag{4.26}$$

and means $\mathbf{0}$ (with size of $NQ \times 1$) and $\boldsymbol{\Sigma}_1\mathbf{m}_{x;post}$, where $\widetilde{\mathbf{K}}_{NM}$, $\widetilde{\mathbf{K}}_M$ and $\widetilde{\mathbf{V}}$ are respectively the block diagonal matrices with all matrices $\mathbf{K}_{q;NM}$, $\mathbf{K}_{q;M}$ and $\mathbf{V}_q$, for $1 \le q \le Q$. Hence, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are $NQ \times NQ$ matrices. The elements of the diagonal matrix $\mathbf{V}_q$ corresponding to the selection index (the pseudo inputs are chosen from the original covariates) are assigned as 0. The reason for this adjustment is to ensure that the values of $\mathbf{f}_q|\bar{\mathbf{f}}_q$ evaluated at the M selected pseudo inputs are deterministically decided by $\bar{\mathbf{f}}_q$ – see this from the mean in (3.6). The M latent function values are also the same as the values of $\mathbf{f}_q$ at the locations $\bar{\mathbf{z}}^{1:M}$.

In addition, the column vector $\mathbf{m}_{x;stack}$ is the column-wise rearrangement of a $N \times Q$ matrix $\mathbf{M}_x$,

$$\mathbf{M}_x = (\mathbf{Y} - \boldsymbol{\lambda}_0 \otimes \mathbf{1}^{\mathsf{T}}_N)^{\mathsf{T}}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Lambda}, \tag{4.27}$$

where $\mathbf{Y}$ is a $R \times N$ matrix consisting of all response vectors, $\mathbf{Y} = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}]$.

Through the Gaussian multiplication identity (A.5), the modified full conditional distribution of latent variables $\mathbf{x}$ can be derived with the Gaussianity (see Appendix A.4.1). It has the covariance matrix

$$\boldsymbol{\Sigma}_{x,post} = (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1})^{-1}, \tag{4.28}$$

and the mean

$$\boldsymbol{\mu}_{x,post} = \boldsymbol{\Sigma}_{x,post}\mathbf{m}_{x;stack}. \tag{4.29}$$

In practical estimation, computing $\boldsymbol{\Sigma}_{x,post}$ and sampling from the above conditional distribution (specified by Equation (4.28) and (4.29)) are challenges. A naive application of inversion or matrix division to is an impractical idea, since this would cost $O(N^3Q^3)$

operations, defeating the whole point of using pseudo inputs. To address the issue, we can use the matrix inversion identity (A.1) twice – one to obtain $\mathbf{\Sigma}_0^{-1}$, the other one to obtain $(\mathbf{\Sigma}_0^{-1} + \mathbf{\Sigma}_1^{-1})^{-1}$. After the applications of matrix inversion, $(\mathbf{\Sigma}_0^{-1} + \mathbf{\Sigma}_1^{-1})^{-1}$ can be decomposed into two covariance matrices of two new random variables. Readers can refer Appendix A.4.1 to understand the derivations and the positive definiteness of the two covariance matrices. Then the sampling procedure of latent variables turns into the generation of samples from the distributions of the new variables, which are

$$\begin{aligned} \mathbf{s}_1 &\sim \mathcal{N}(\boldsymbol{\mu}_{s_1}, \mathbf{A}_1^{-1}), & (4.30) \\ \mathbf{s}_2 &\sim \mathcal{N}(\boldsymbol{\mu}_{s_2}, \mathbf{A}_1^{-1}\mathbf{D}_1\mathbf{C}_1^{-1}\mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1}), & (4.31) \end{aligned}$$

where

$$\mathbf{A} = \widetilde{\mathbf{V}} + \mathbf{\Sigma}_x \otimes \mathbf{I}_N, \quad \mathbf{A}_1 = \mathbf{\Sigma}_1^{-1} + \mathbf{A}^{-1}, \tag{4.32}$$

$$\mathbf{D}_1 = \mathbf{A}^{-1}\widetilde{\mathbf{K}}_{MN}^{\mathsf{T}}, \quad \mathbf{C}_1 = \widetilde{\mathbf{K}}_M + \widetilde{\mathbf{K}}_{MN}\mathbf{A}^{-1}\widetilde{\mathbf{K}}_{MN}^{\mathsf{T}} - \mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1}\mathbf{D}_1, \tag{4.33}$$

$$\boldsymbol{\mu}_{s_1} = \mathbf{A}_1^{-1}\mathbf{m}_{x;stack}, \tag{4.34}$$

$$\boldsymbol{\mu}_{s_2} = (\mathbf{A}_1^{-1}\mathbf{D}_1\mathbf{C}_1^{-1}\mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1})\mathbf{m}_{x;stack}. \tag{4.35}$$

In addition, we can use Cholesky decomposition[7] to obtain a random sample from a multivariate normal distribution. As a result, a new draw of latent variables can be represented as

$$\mathbf{x} = \boldsymbol{\mu}_{s_1} + (chol(\mathbf{A}_1))^{-1}\mathbf{t}_1 + \boldsymbol{\mu}_{s_2} + \mathbf{A}_1^{-1}\mathbf{D}_1(chol(\mathbf{C}_1))^{-1}\mathbf{t}_2, \tag{4.36}$$

where $\mathbf{t}_1$ and $\mathbf{t}_2$ are the random vectors whose entries are sampled from a standard normal distribution.

Three points are noted for the above computation procedure. The first is the structure features of $\mathbf{A}$ and $\mathbf{A}_1$. $\mathbf{A}$ itself is a $NQ \times NQ$ matrix with a particular array structure, which consists of $Q \times Q$ block matrix structure and each block matrix with size $N \times N$ has only non-zero elements on the diagonal[8]. Two reasons for this structure are: 1. $\mathbf{A}$ is $\widetilde{\mathbf{V}} + \mathbf{\Sigma}_x \otimes \mathbf{I}_N$, where $\widetilde{\mathbf{V}}$ is a block diagonal matrix made of diagonal matrices $\mathbf{V}_q$; 2. $\mathbf{\Sigma}_x \otimes \mathbf{I}_N$ has the same structure as $\mathbf{A}$ due to the definition of Kronecker product. It follows

[7] A Cholesky decomposition of a positive definitive matrix $\mathbf{P}$ is to find a upper matrix $chol(\mathbf{P})$ with positive diagonal elements so that $\mathbf{P} = chol(\mathbf{P})^{\mathsf{T}} \cdot chol(\mathbf{P})$.

[8] We call this structure as *tiled diagonal structure* hereafter.

that $\mathbf{A}^{-1}$ has the same structure features, and so is $\mathbf{A}_1$ (owing to $\boldsymbol{\Sigma}_1^{-1}$ being a diagonal matrix).

The second is the matrix inversion of $\boldsymbol{\Sigma}_0^{-1}$ and $(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}_1^{-1})^{-1}$ by the identity (A.1). In the applications of (A.1), one also has to compute $\mathbf{A}^{-1}$ and $\mathbf{A}_1^{-1}$ by the identity of block matrix inversion (A.3). In practice, their tiled diagonal structures facilitate the computing.

The third is about calculating the Cholesky decompositions of $\mathbf{A}_1$ and $\mathbf{C}_1$. It could not be an computational issue. The reasons are that $\mathbf{A}_1$ has the tiled diagonal structure, and $\mathbf{C}_1$ is a $MQ \times MQ$ positive definite matrix, where the pseudo input size M can be rather smaller than N.

Latent functions

For all latent functions, a new sample of $\mathbf{f}$ can be drawn from its full conditional distribution integrating out pseudo latent function vector $\bar{\mathbf{f}}$. The resulting conditional density is $p(\mathbf{f}|e.e. \setminus \bar{\mathbf{f}})$.

This density is proportional to the multiplication of $p(\mathbf{f}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_{1:Q}^{1:M}, \boldsymbol{\Theta}_h)$ and $p(\mathbf{x}|\mathbf{f}, \boldsymbol{\Sigma}_x)$. The two factors are the densities of the Gaussian distributions with respective covariance matrices

$$\boldsymbol{\Sigma}_2 = \widetilde{\mathbf{K}}_{NM}^{\mathsf{T}}\widetilde{\mathbf{K}}_M^{-1}\widetilde{\mathbf{K}}_{NM} + \widetilde{\mathbf{V}}, \tag{4.37}$$

$$\boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_x \otimes \mathbf{I}_N, \tag{4.38}$$

and means $\mathbf{0}$ (with size of $NQ \times 1$) and $\mathbf{m}_{f;stack}$ (equal to $\boldsymbol{\Sigma}_3^{-1}\mathbf{x}$).

Then the modified full conditional distribution of $\mathbf{f}$ follows a Gaussian distribution based on the Gaussian multiplication identity (A.5). See the derivation in Appendix A.4.2, the covariance matrix and the mean can be respectively given

$$\boldsymbol{\Sigma}_{f,post} = (\boldsymbol{\Sigma}_2^{-1} + \boldsymbol{\Sigma}_3^{-1})^{-1}, \tag{4.39}$$

and

$$\boldsymbol{\mu}_{f,post} = \boldsymbol{\Sigma}_{f,post}\mathbf{m}_{f;stack}. \tag{4.40}$$

Similar to the computational scheme for $\boldsymbol{\Sigma}_{x,post}$, we again use an application of the matrix inversion identity (A.1) twice for $\boldsymbol{\Sigma}_{f,post}$. $\boldsymbol{\Sigma}_{f,post}$ can be decomposed into two covariance matrices of two new random variables for sampling. The derivations can be found in Appendix A.4.2 and the positive definite

$$\mathbf{s}_3 \sim \mathcal{N}(\boldsymbol{\mu}_{s_3}, \mathbf{A}_2), \tag{4.41}$$

$$\mathbf{s}_4 \sim \mathcal{N}(\boldsymbol{\mu}_{s_4}, \boldsymbol{\Sigma}_3\mathbf{D}_2\mathbf{C}_2^{-1}\mathbf{D}_2^{\mathsf{T}}\boldsymbol{\Sigma}_3), \tag{4.42}$$

where

$$\mathbf{A} = \widetilde{\mathbf{V}} + \boldsymbol{\Sigma}_3, \quad \mathbf{A}_2 = \boldsymbol{\Sigma}_3 - \boldsymbol{\Sigma}_3\mathbf{A}^{-1}\boldsymbol{\Sigma}_3, \tag{4.43}$$

$$\mathbf{D}_2 = \mathbf{A}^{-1}\widetilde{\mathbf{K}}_{MN}^{\mathsf{T}}, \quad \mathbf{C}_2 = \widetilde{\mathbf{K}}_M + \widetilde{\mathbf{K}}_{MN}\mathbf{A}^{-1}\widetilde{\mathbf{K}}_{MN}^{\mathsf{T}}, \tag{4.44}$$

$$\boldsymbol{\mu}_{s_3} = \mathbf{A}_2\mathbf{m}_{f;stack}, \tag{4.45}$$

$$\boldsymbol{\mu}_{s_4} = (\boldsymbol{\Sigma}_3\mathbf{D}_2\mathbf{C}_2^{-1}\mathbf{D}_2^{\mathsf{T}}\boldsymbol{\Sigma}_3)\mathbf{m}_{f;stack}. \tag{4.46}$$

A new draw of latent functions is

$$\mathbf{f} = \boldsymbol{\mu}_{s_3} + (chol(\mathbf{A}_2))^{\mathsf{T}}\mathbf{t}_1 + \boldsymbol{\mu}_{s_4} + \boldsymbol{\Sigma}_3\mathbf{D}_2(chol(\mathbf{C}_2))^{-1}\mathbf{t}_2. \tag{4.47}$$

Note that $\mathbf{A}_2$ has a tiled diagonal matrix structure as well because $\boldsymbol{\Sigma}_3$ and $\mathbf{A}$ shares the same array formation. Moreover, $\mathbf{A}_2$ is $(\widetilde{\mathbf{V}}^{-1} + \boldsymbol{\Sigma}_3^{-1})^{-1}$ (this fact can be found in the proof of positive definiteness of $\mathbf{A}_2$), and we can use block matrix inversion identity (A.3) to compute $\mathbf{A}_2$.

The above sampling schemes of latent variables and latent functions can achieve a better mixing in MCMC simulations. The reason is that the former does not depend on the associated strongly-correlated variables, and the latter only depends on latent variables. The resulting computational load may increase slightly due to matrix computation to jointly produce a new sample. By contrast, the original schemes in Algorithm 1 have dependence on latent variables and pseudo latent functions although having lighter computational cost.

4.3.2 Sampling Expanded Parameter and Factor loadings

We adopt the technique of parameter expansion (Gelman et al. 2005, Liu & Wu 1999) to improve sampling efficiency of loadings $\boldsymbol{\Lambda}$ by introducing a non-identical parameter α. The idea is to expand latent variables and to contract loadings by the parameter α. The acts thus produce the new variables w_q and new loadings $\boldsymbol{\Lambda}_\alpha$, where $w_q = \alpha x_q$ and $\boldsymbol{\Lambda}_\alpha = \alpha\boldsymbol{\Lambda}$. After the new α is sampled, by which latent variables and loadings can be transformed back from w_q and $\mathbf{\Lambda}_\alpha$. The latent variables and loadings recovered turn less correlated because of the expanded parameter α randomly generated in the immediate sampler. Note that w_q and new loadings $\mathbf{\Lambda}_\alpha$ are highly correlated because the latter's sampler related to the former.

For a computational reason, a conjugate prior (distributed as an inverse gamma density) is used for α. Then its resulting sampler can be obtained and shares the inverse gamma distribution $\mathcal{IG}^{-1}(a, b)$ with hyper-parameters

$$a = a_1 + NQ/2, \quad b = b_1 + \mathbf{w}\mathbf{\Sigma}_0^{-1}\mathbf{w}, \tag{4.48}$$

where a_1 and b_1 are hyper-parameters of the prior distribution; $\mathbf{w}$ is a column vector containing all the members of $\{\mathbf{w}_1, \ldots, \mathbf{w}_Q\}$ and $\mathbf{\Sigma}_0$ is used in the sampling procedure of latent variables. Here $\mathbf{\Sigma}_0$ adopts the derived result by matrix inversion identity during calculation. See the derivation in Appendix A.4.3.

The sampler of the shrunk factor loadings $\mathbf{\Lambda}_\alpha$ resembles that of the original $\mathbf{\Lambda}$, which simply replaces latent variables $\mathbf{x}$ by the transformed latent variables $\mathbf{w}$.

4.3.3 Sampling Hyper-parameters and Algorithm Summary

Hyper-parameters of GP Covariance Function

The sampling of hyper-parameters is similar to the original. The difference is that in Equation (4.6), the ratio has a different $h(\theta_{h,qj})$. Moreover, this $h(\theta_{h,qj})$ replaces $p(\mathbf{x}_q|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M}, \bar{\mathbf{f}}_q^{1:M}, \boldsymbol{\theta}_{h,q})$ by $p(\mathbf{f}_q|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M}, \boldsymbol{\theta}_{h,q})$. The latter is the probability density of the distribution represented by (3.6) and integrating out pseudo functions – the sampling efficiency could be improved further. It is also a Gaussian density with a mean $N \times 1$ zero vector and a covariance matrix $\mathbf{K}_{q;NM}\mathbf{K}_{q;M}^{-1}\mathbf{K}_{q;MN} + \mathbf{V}_q$.

The associated evaluation of that density can be facilitated by using the identities of block matrix for determinant (A.4) and inversion (A.3). Before the use, we can, without loss of generality, specify $\mathbf{K}_{q;NM}\mathbf{K}_{q;M}^{-1}\mathbf{K}_{q;MN} + \mathbf{V}_q$ as

$$\begin{bmatrix} \mathbf{V}_{q;N_0} + \mathbf{K}_{q;N_0M}\mathbf{K}_{q;M}^{-1}\mathbf{K}_{q;MN_0} & \mathbf{K}_{q;N_0M} \\ \mathbf{K}_{q;N_0M}^{\mathsf{T}} & \mathbf{K}_{q;M} \end{bmatrix},$$

where $\mathbf{V}_{q;N_0}$ and $\mathbf{K}_{q;MN_0}$ are the submatrices of $\mathbf{V}_q$ and $\mathbf{K}_{q;MN}$, excluding the rows or columns corresponding to the selection indices for the pseudo input set $\bar{\mathbf{z}}_q^{1:M}$. N_0 is the index set of the rest $N - M$ covariates not selected as pseudo inputs. Hence, $\mathbf{V}_{q;N_0}$ and $\mathbf{K}_{q;MN_0}$ respectively have sizes of $(N-M) \times (N-M)$ and $M \times (N-M)$. Note that the bottom-right block reflects the adjustment of the diagonal elements of $\mathbf{V}_q$ corresponding to the M selection indices, set to 0s. We have discussed the reason for the adjustment in Section 4.3.1.

Now, by using (A.4) and (A.2) the determinant of $\mathbf{K}_{q;NM}\mathbf{K}_{q;M}^{-1}\mathbf{K}_{q;MN} + \mathbf{V}_q$ is

$$|\mathbf{K}_{q;M}| \cdot |\mathbf{V}_{q;N_0}| = \prod_{m}^{M} eig_m(\mathbf{K}_{q;M}) \prod_{j \in N_0} [\mathbf{V}_{q;N_0}]_{j,j}, \tag{4.49}$$

and by using (A.3) the inverse matrix is

$$\begin{bmatrix} (\mathbf{V}_{q;N_0})^{-1} & -(\mathbf{V}_{q;N_0})^{-1}\mathbf{K}_{q;MN_0}\mathbf{K}_{q;M}^{-1} \\ -((\mathbf{V}_{q;N_0})^{-1}\mathbf{K}_{q;MN_0}\mathbf{K}_{q;M}^{-1})^{\mathsf{T}} & \mathbf{K}_{q;M}^{-1} + \mathbf{K}_{q;M}^{-1}\mathbf{K}_{q;MN_0}(\mathbf{V}_{N_0,q})^{-1}\mathbf{K}_{q;MN_0}^{\mathsf{T}}\mathbf{K}_{q;M}^{-1} \end{bmatrix}. \tag{4.50}$$

Algorithm 2

After randomly selecting Q sets of pseudo inputs, the modified algorithm for per MCMC iteration becomes as follows:

- Call a sampler of hyper-parameters of covariance functions, computing based on Equation (4.6) but using the modified conditional density for $h(\theta_{iq})$, where calculation involves utilization of Equations (4.49) and (4.50).
- Call a sampler of latent variables using (4.30), (4.31) and (4.36).
- Call a sampler of latent functions using (4.41), (4.42) and (4.47).
- Select pseudo latent functions from the sampled latent functions according to the selection of the pseudo inputs.
- Call a sampler of parameter expansion using (4.48) and produce transformed latent variables $\mathbf{w}$.
- Call a sampler of factor loadings using (4.18) and (4.19) with replacing the latent variables $\mathbf{x}$ by $\mathbf{w}$ to acquire transformed factor loadings $\mathbf{\Lambda}_\alpha$.
- Transform the factor loadings and latent variables back by the expansion parameter α.

- Call a sampler of measurement error variances using (4.20).

Sampling latent variables and latent functions need more computational steps than the others. It is reckoned that both have complexity of $O(M^2NQ^2)$. This is based on calculating the number of (multiplication) operations in matrix multiplication about $\mathbf{C}_1$ and $\mathbf{D}_1(chol(\mathbf{C}_1))^{-1}$ in Equation (4.36) for latent variables; $\mathbf{C}_2$ and $\mathbf{D}_2(chol(\mathbf{C}_2))^{-1}$ in Equation (4.46) for latent functions. Therefore, we could consider that the whole algorithm reduces the cost per iteration to $O(M^2NQ^2)$ (from $O(N^3Q^3)$).

4.4 Expectation Maximization methods and Its Stochastic Implementation

Besides fully Bayesian methods, the Expectation Maximization (EM) algorithm is another approach to estimate model parameters. Briefly speaking, via maximum likelihood on a complete data, a combination of realizations of observed variables $\mathbf{y}$ and unobserved variables $\mathbf{x}$ [9], EM methods allow one to run parameter estimation in an iterative way (Dempster et al. 1977, McLachlan & Krishnan 2008). They alternatively implement two steps: E-step, to calculate the “best” likelihood for model parameters; M-step, to optimise the model parameters from it.

More specifically, due to computational difficulty, parameter estimation is not based on calculating and optimising the log-marginal-likelihood of model parameters $\boldsymbol{\theta}$ given the observed data $\mathbf{y}^{1:N}$

$$l(\boldsymbol{\theta}) = \log \int p(\mathbf{y}^{1:N}, \mathbf{x}^{1:N}|\boldsymbol{\theta}) d\mathbf{x}^{1:N}.$$

By contrast, the joint log-likelihood under the complete data $\{\mathbf{y}^{1:N}, \mathbf{x}^{1:N}\}$,

$$l_c(\boldsymbol{\theta}) = \log(p(\mathbf{y}^{1:N}, \mathbf{x}^{1:N}|\boldsymbol{\theta}))$$

may have a closed-form expression for the maximum likelihood estimator (MLE) of model parameters and can facilitate the whole calculation and optimisation. To obtain the “best” likelihood for estimating $\boldsymbol{\theta}$, however, one needs to introduce an arbitrary distribution over latent variables $b(\mathbf{x})$ and then calculate the conditional expectation of $l_c(\boldsymbol{\theta})$ with respect to $b(\mathbf{x})$ and the current parameters $\boldsymbol{\theta}^{[i]}$. Through some derivations, the optimal option of $b(\mathbf{x})$ can be learned as the conditional density of unobserved variables $\mathbf{x}$ given the observed

[9] Here unobserved variables are latent variables or missing values.

data $\mathbf{y}^{1:N}$ and $\boldsymbol{\theta}^{[i]}$, that is, the posterior density of $\mathbf{x}$, $p(\mathbf{x}|\mathbf{y}^{1:N}, \boldsymbol{\theta}^{[i]})$. Thus, the E-step is to calculate the conditional expectation, which is defined as a function $Q(\cdot|\cdot)$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]}) = E_{\boldsymbol{\theta}^{[i]}}(l_c(\boldsymbol{\theta})|\boldsymbol{y}^{1:N}) = \int b(\mathbf{x}) \log p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}, \tag{4.51}$$

where $b(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}^{1:N}, \boldsymbol{\theta}^{[i]})$.

Following this notation, the M-step is to implement optimization as

$$\boldsymbol{\theta}^{[i+1]} = \arg\ \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]}). \tag{4.52}$$

It is noted that the Q-function is incremental over the estimation sequence of $\boldsymbol{\theta}$; in other words, it satisfies $Q(\boldsymbol{\theta}^{[i+1]}|\boldsymbol{\theta}^{[i]}) \geq Q(\boldsymbol{\theta}^{[i]}|\boldsymbol{\theta}^{[i]})$ (Dempster et al. 1977, McLachlan & Krishnan 2008).

Neal and Hinton (1999) provide an alternative perspective of EM methods that the E-step and M-step actually can be regarded as two maximization implementation on an objective function. Their insight is from the decomposition of the marginal log-likelihood $l(\boldsymbol{\theta})$,

$$\begin{aligned} l(\boldsymbol{\theta}) &= \int b(\mathbf{x}) \log \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})}{b(\mathbf{x})} d\mathbf{x} + \int b(\mathbf{x}) \log \frac{b(\mathbf{x})}{p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})} d\mathbf{x} && (4.53) \\ &= G(\boldsymbol{\theta}, b) + KL(b(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})), && (4.54) \end{aligned}$$

where $G(\boldsymbol{\theta}, b)$ is the aforementioned objective function for estimation; $KL(b\|p)$ denotes Kullback-Leibler divergence between two probabilistic distributions b and p, measuring the difference from b to p. Since $l(\boldsymbol{\theta})$ is fixed and $KL(\cdot\|\cdot)$ is non-negative, $G(\boldsymbol{\theta}, b)$ is a lower bound of the marginal likelihood $l(\boldsymbol{\theta})$. In fact, the lower bound can also be obtained via Jensen's inequality and concavity of a logarithm function.

The two optimization procedures are to maximize $G(\boldsymbol{\theta}, b)$ on one argument with fixing the other in turn. At the i-th iteration, one can firstly consider maximize $G(\boldsymbol{\theta}, b)$ on b given the current parameter $\boldsymbol{\theta}^{[i]}$. The distribution achieving the maximum is $b(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}^{1:N}, \boldsymbol{\theta}^{[i]})$, which can be learned because $l(\boldsymbol{\theta})$ is a constant and the conditional distribution $b(\mathbf{x})$ minimizes $KL(b(\mathbf{x})\|p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}))$. This procedure can actually be viewed as the E-step of the original EM method with a slight difference[10]. After the first step, one can consider maximize $G(\boldsymbol{\theta}, b)$ on $\boldsymbol{\theta}$ given the fixed distribution $b(\mathbf{x})$ - this is the M-step.

Note that no matter which viewpoint one adapts, the obtained sequence of parameters $\boldsymbol{\theta}$ will iteratively reach a local maximum of marginal likelihood $l(\boldsymbol{\theta})$. This is learnt by the

[10] The marginal difference is that $G(\boldsymbol{\theta}^{[i]}, b) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]}) - c$, where c is a constant, from $\int b(\mathbf{x}) \log b(\mathbf{x}) d\mathbf{x}$.

fact that with a fixed distribution $b(\cdot)$, $G(\boldsymbol{\theta}, b)$ or $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]})$ is a lower bound of $l(\boldsymbol{\theta})$ and incremental, that is,

$$l(\boldsymbol{\theta}^{[i+1]}) \geq G(\boldsymbol{\theta}^{[i+1]}, b) \geq G(\boldsymbol{\theta}^{[i]}, b) = l(\boldsymbol{\theta}^{[i]}), \tag{4.55}$$

or

$$l(\boldsymbol{\theta}^{[i+1]}) \geq Q(\boldsymbol{\theta}^{[i+1]}|\boldsymbol{\theta}^{[i]}) \geq Q(\boldsymbol{\theta}^{[i]}|\boldsymbol{\theta}^{[i]}) = l(\boldsymbol{\theta}^{[i]}). \tag{4.56}$$

When a statistical model is complex, that is, the variables involved are in high-dimensional space, calculating Q-function turns difficult. The reasons are lacking of an analytic closed form and evaluating an intractable high-dimensional integral. Conventional approaches including analytical approximation or quadrature have their limitations on this computational issue, especially the dimension is rather high (McLachlan & Krishnan 2008). Monte Carlo sampling-based methods, including MCMC algorithms, can address the issue from a high-dimensional integral and allow one to approximate the Q-function. Due to the characteristic of random sampling, such EM method becomes a stochastic approach. This implies that given the same starting values, repeating applications may not achieve the same value of the stationary point after a certain iterations. This stochastic version distinguishes itself from its deterministic origin, which mentioned in the preceding paragraphs.

The above stochastic EM algorithm where in the E-step the Monte Carlo methods are adopted is refer to as Monte Carlo EM (MCEM) algorithm (Wei & Tanner 1990). More precisely, at the i-th iteration MCEM approximates Q-function in the E-step by the Monte Carlo average, which is

$$\tilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}) = \frac{1}{m_i} \sum_{j=1}^{m_i} \log p(\mathbf{y}, \mathbf{x}^{[j]}|\boldsymbol{\theta}^{[i]}), \tag{4.57}$$

where $\{\mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[m_i]}\}$ are samples generated from the conditional distribution $b(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}^{1:N}, \boldsymbol{\theta}^{[i]})$ by using a Monte Carlo method given the sample size m_i, which is allowed to change over the EM iterations. If m_i is large enough, by the law of large numbers, the Q function can be approximated by $\tilde{Q}$ reasonably. The rest procedure of the MCEM method, the M-Step, is the same as that in the original EM method. If no analytic closed form for optimal solutions of $\boldsymbol{\theta}$ exists, typical iterative methods, such as Newton-Raphson or quasi-Newton methods, can be used to obtain an optimal solution (McLachlan & Krishnan 2008).

There are several stochastic versions of the EM methods, such as the stochastic EM method in (Celeux & Diebolt 1985, Celeux & Ip 1996); the stochastic approximation EM approach of (Delyon et al. 1999). The former could be regarded as a special case of MCEM where $m_i = 1$. The E-step is simply to replace latent variables $\mathbf{x}$ (or missing values) with a sample generated to form a complete data for maximization on model parameters in the next step. The latter provides another stochastic approximation for the Q-function by a recursive equation along with a step size and a small and constant of m_i (see details in (Delyon et al. 1999)).

Because MCEM is a basic stochastic version of the EM methods and it connects with MCMC samplers presented in the last section, we focus on that for practical implementation. If a simple closed form of posterior density of latent variables can be derived, one can evaluate the conditional expectation of the complete likelihood by calculating associated sufficient statistics rather than by using MCMC methods.

4.5 Inference Function for Margins

For a complicated multivariate model, the EM methods introduced before allow estimating model parameters simultaneously. However, the computational cost can still be expensive. For example, a practical implementation is conducted for a longitudinal data with multiple response variables. As for this issue, alternative approaches to make estimation easier are desired. Rather than estimating model parameters simultaneously, the method of inference function of margins (IFM) can serve as an alternative.

The IFM method (Joe & Xu 1996, Joe 1997) is to estimate model parameters through a system of estimating equations from the marginal and joint distributions of response variables. It is essentially a two-step optimisation approach. More specifically, one can first classify model parameters into two categories before implementation.

Given a response random vector $\mathbf{y} = (y_1, \ldots, y_R)$, one category is from univariate response models. For $1 \leq r \leq R$, the r-th model has the marginal cumulative distribution function (cdf), $B_r(y_r|\boldsymbol{\theta}_r)$, with the associated parameters $\boldsymbol{\theta}_r$. The other category is from the joint model of response variables. More precisely, those kinds of parameters are the dependence parameters characterising the dependence structure between response variables. The joint cdf of $\mathbf{y}$ is $B(\mathbf{y}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R, \boldsymbol{\phi})$ with all marginal parameters and the dependence parameters $\boldsymbol{\phi}$. And Sklar's theorem (see Section 1.1 of Ruschendorf (2013)) claims that the joint cdf can be expressed by a copula C[11] and written as follow:

$$B(\mathbf{y}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R, \boldsymbol{\phi}) = C\Big(B_1(y_1|\boldsymbol{\theta}_1), \ldots, B_R(y_R|\boldsymbol{\theta}_R)\Big|\boldsymbol{\phi}\Big).$$

Note that the copula model is parametrised by the dependence parameters $\boldsymbol{\phi}$. Thus in an example of multivariate normal distribution, the copula model can give structure information about parameters associated with each pairwise marginal distribution, like correlation coefficients.

Next, the estimation procedure of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$ proceeds separately based on the corresponding estimating equation. For each r, the r-th estimating equation is given by partial derivatives of a log-likelihood function from the marginal cdf B_r, that is

$$\frac{\partial}{\partial \boldsymbol{\theta}_r}\ell_r(\boldsymbol{\theta}_r) = \frac{\partial}{\partial \boldsymbol{\theta}_r}\sum_{n=1}^{N} \log b_r(y_r^{(n)}|\boldsymbol{\theta}_r), \tag{4.58}$$

where $\ell_r(\boldsymbol{\theta}_r)$ is the log-likelihood function of $\boldsymbol{\theta}_r$ and b_r is the probability density function (pdf) of the response variable y_r. An estimate $\tilde{\boldsymbol{\theta}}_r$ can be obtained by solving $\partial\ell_r(\boldsymbol{\theta}_r)/\partial\boldsymbol{\theta}_r = 0$, namely implementing MLE on $\boldsymbol{\theta}_r$.

At the final step, the way to estimate the dependence parameter $\boldsymbol{\phi}$ is similar to that of $\boldsymbol{\theta}_r$. The difference is that the associated estimating equation is the score function for the joint cdf B, which is

$$\frac{\partial}{\partial \boldsymbol{\phi}}\ell(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R, \boldsymbol{\phi}) = \frac{\partial}{\partial \boldsymbol{\phi}}\sum_{n=1}^{N} \log b(\mathbf{y}^{(n)}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R, \boldsymbol{\phi}), \tag{4.59}$$

where $\ell(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R, \boldsymbol{\phi})$ is the log-likelihood function of all model parameters and b is the joint pdf of the response random vector $\mathbf{y}$. An estimate $\tilde{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$ is obtained by maximising $\ell(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_R, \boldsymbol{\phi})$, where $\tilde{\boldsymbol{\theta}}_r$, $1 \leq r \leq R$, are given from the preceding estimation procedures.

Incidentally since the joint pdf b is the product of a copula density $c(\cdot)$ and all marginal pdf b_r, the RHD of Equation (4.58) can rewritten as

$$\frac{\partial}{\partial \boldsymbol{\phi}}\sum_{n=1}^{N} \log c\left(B_1(y_1^{(n)}|\boldsymbol{\theta}_1), \ldots, B_r(y_R^{(n)}|\boldsymbol{\theta}_R)\Big|\boldsymbol{\phi}\right).$$

As seen, the copula likelihood is the proxy to estimate the dependence parameters $\boldsymbol{\phi}$.

[11] A copula is a multivariate distribution function with the range a uniform interval $[0, 1]$, the domain on a multiple-dimensional unit cube defined by marginal distributions.

Multivariate margins[12] can be considered. According to (4.58), the multivariate-margin dependence parameters are able to be estimated by MLE on the corresponding log-likelihood. Thereby, the sum of the log-likelihood of all multivariate margins is the source for all dependent parameter estimates. This feature and the univariate-margin estimation procedure in (4.57) connects composite (marginal) likelihood methods (Varin et al. 2011).

The IFM method can also be applied to a multivariate model with covariates and latent variables. The margin parameters $\boldsymbol{\theta}_r$ there are not estimated by maximising the marginal likelihood (formed by the margin observations). Instead, the estimation is to maximise the conditional expectation of the complete likelihood (formed by the margin observations and latent variables). This follows the EM estimation steps. The same estimation fashion is implemented for dependence parameters $\boldsymbol{\phi}$. The conditional expectation of the joint complete likelihood (formed by all observations and latent variables) is calculated in the E-Step. This scheme is applied to parameter estimation of sparse GP-SEM in the next section.

4.6 Hybrid Algorithm for Sparse GP-SEM

In Section 4.4 and 4.5, two estimation methods (MCEM and IFM) are introduced. Combining the two features (convenient applicability and efficient computation on latent variable models with multiple outcomes), a hybrid algorithm is presented for sparse GP-SEM below.

Considering a structure of each outcome with only one parent latent variable, the hybrid algorithm is essentially to implement the EM method by the two-step estimation procedure. One step is for the marginal models – each of which consists of covariates and responses with the same latent variable. The other step is for the joint model – comprising covariates, all responses and their latent variables.

To specify the implementation, it is primary to formulise the objective function in E-step, which used in M-step later. The objective function is the conditional expectation of log complete likelihood with respect to the posterior density of latent variables and latent

[12] The term ”margin” means a marginal model or a sub-model associated with a subset of response variables.

functions. More specifically,

$$
\begin{aligned}
&E_{p(\mathbf{x},\mathbf{f}|\mathbf{y})}[\log p(\mathbf{x},\mathbf{f},\mathbf{y}|\mathbf{z}^{1:N},\bar{\mathbf{z}}^{1:M},\mathbf{\Theta}_h,\mathbf{\Lambda},\boldsymbol{\lambda}_0,\mathbf{\Sigma}_x,\mathbf{\Sigma}_y)]\\
&= E_{p(\mathbf{x},\mathbf{f}|\mathbf{y})}[\log p(\mathbf{y}|\mathbf{x},\mathbf{\Lambda},\boldsymbol{\lambda}_0,\mathbf{\Sigma}_y) + \log p(\mathbf{x}|\mathbf{z}^{1:N},\bar{\mathbf{z}}^{1:M},\mathbf{\Theta}_h,\mathbf{\Sigma}_x) + \log p(\mathbf{f}|\mathbf{x},\mathbf{\Sigma}_x)]\\
&= E_{p(\mathbf{x},\mathbf{f}|\mathbf{y})}[-\frac{N}{2}\log|\mathbf{\Sigma}_y| - \frac{1}{2}tr\big(\mathbf{\Sigma}_y^{-1}(\mathbf{Y}-\widetilde{\mathbf{\Lambda}}(\widetilde{\mathbf{X}})^{\mathsf{T}})(\mathbf{Y}-\widetilde{\mathbf{\Lambda}}(\widetilde{\mathbf{X}})^{\mathsf{T}})^{\mathsf{T}}\big)] \; +\\
&\quad E_{p(\mathbf{x},\mathbf{f}|\mathbf{y})}[-\frac{1}{2}\log|\mathbf{\Sigma}_0| - \frac{1}{2}tr(\mathbf{\Sigma}_0^{-1}\mathbf{x}\mathbf{x}^{\mathsf{T}})] \; +\\
&\quad E_{p(\mathbf{x},\mathbf{f}|\mathbf{y})}[-\frac{1}{2}\log|\mathbf{\Sigma}_x\otimes\mathbf{I}_N| - \frac{1}{2}tr((\mathbf{\Sigma}_x\otimes\mathbf{I}_N)^{-1}(\mathbf{f}-\mathbf{x})(\mathbf{f}-\mathbf{x})^{\mathsf{T}})]\\
&= (1) + (2) + (3),
\end{aligned}
$$

where the notations $\widetilde{\mathbf{X}}$ and $\mathbf{Y}$ happen in Equations (4.18) and (4.27); $\mathbf{\Sigma}_0$ is the same as that in (4.24). $\widetilde{\mathbf{\Lambda}}$ is a $R\times(Q+1)$ matrix that factor loadings $\mathbf{\Lambda}$ aggregates the intercept $\boldsymbol{\lambda}_0$ in columns; $\mathbf{x}$ and $\mathbf{f}$ are $(NQ)\times 1$ vectors, $\mathbf{x} = (\mathbf{x}_1^{\mathsf{T}},\ldots,\mathbf{x}_Q^{\mathsf{T}})^{\mathsf{T}}$ and $\mathbf{f} = (\mathbf{f}_1^{\mathsf{T}},\ldots,\mathbf{f}_Q^{\mathsf{T}})^{\mathsf{T}}$. $\mathbf{\Theta}_h = \{\boldsymbol{\theta}_{h,1},\ldots,\boldsymbol{\theta}_{h,Q}\}$ is the set of all GP hyper-parameters. Note that the conditional expectation depends on the current parameters (specifying the i-th optimisation step) and here we skip the notation for simplicity.

The sum of the three terms can be approximated by MCMC methods[13] as,

$$
(1) \approx \frac{1}{m_i}\sum_{j=1}^{m_i}\left[-\frac{N}{2}\log|\mathbf{\Sigma}_y| - \frac{1}{2}tr\big(\mathbf{\Sigma}_y^{-1}(\mathbf{Y}^{[j]}-\tilde{\mathbf{\Lambda}}(\tilde{\mathbf{X}}^{[j]})^{\mathsf{T}})(\mathbf{Y}^{[j]}-\tilde{\mathbf{\Lambda}}(\tilde{\mathbf{X}}^{[j]})^{\mathsf{T}})\big)\right], \tag{4.60}
$$

$$
(2) \approx \frac{1}{m_i}\sum_{j=1}^{m_i}\left[-\frac{1}{2}\log|\mathbf{\Sigma}_0| - \frac{1}{2}tr(\mathbf{\Sigma}_0^{-1}\mathbf{x}^{[j]}(\mathbf{x}^{[j]})^{\mathsf{T}}\right], \tag{4.61}
$$

$$
(3) \approx \frac{1}{m_i}\sum_{j=1}^{m_i}\left[-\frac{1}{2}\log|\mathbf{\Sigma}_x\otimes\mathbf{I}_N| - \frac{1}{2}tr((\mathbf{\Sigma}_x\otimes\mathbf{I}_N)^{-1}(\mathbf{f}^{[j]}-\mathbf{x}^{[j]})(\mathbf{f}^{[j]}-\mathbf{x}^{[j]})^{\mathsf{T}})\right], \tag{4.62}
$$

where $\mathbf{Y}^{[j]}$ is merely a replicate of $\mathbf{Y}$. And the joint sample of latent variables and latent functions can be produced by the individual samplers presented in Section 4.3.1.

Move to the M-step, maximum optimisation for parameter estimation is implemented upon the objective function presented in Equations (4.60)-(4.62). Moreover, (4.60) contributes the estimates for measurement error variances, factor loadings and intercepts. The analytic optimisation solutions can be derived through simple algebra, which are

[13] We discuss the necessity of using MCMC methods later.

$$\hat{\sigma}^2_{y_r} = \frac{1}{m_i}\sum_{j=1}^{m_i}\Big[\mathbf{y}_r^{\mathsf{T}}\mathbf{y}_r - 2([\tilde{\mathbf{\Lambda}}]_{r,\mathcal{P}_r})[(\tilde{\mathbf{X}}^{[j]})^{\mathsf{T}}(\mathbf{Y}^{[j]})^{\mathcal{T}}]_{\mathcal{P}_r,r}$$

$$+ ([\tilde{\mathbf{\Lambda}}]_{r,\mathcal{P}_r})[(\tilde{\mathbf{X}}^{[j]})^{\mathsf{T}}\tilde{\mathbf{X}}^{[j]}]_{\mathcal{P}_r,\mathcal{P}_r}[\tilde{\mathbf{\Lambda}}]_{r,\mathcal{P}_r})^{\mathsf{T}}\Big], \tag{4.63}$$

$$[\hat{\tilde{\mathbf{\Lambda}}}]_{r,\mathcal{P}_r} = \frac{1}{m_i}\sum_{j=1}^{m_i}\Big[(([(\widetilde{\mathbf{X}}^{[j]})^{\mathsf{T}}\widetilde{\mathbf{X}}^{[j]}]_{\mathcal{P}_r,\mathcal{P}_r})^{-1}[(\tilde{\mathbf{X}}^{[j]})^{\mathsf{T}}(\mathbf{Y}^{[j]})^{\mathsf{T}}]_{\mathcal{P}_r,r}\Big]^{\mathsf{T}}. \tag{4.64}$$

$\hat{\theta}_{h,ql}$, the estimator for the l-th hyper-parameter of the q-th GP covariance functions, would achieve a maximum of (4.61). It does not have a closed form, but by matrix calculus, the associated derivatives can facilitate the estimation. In practice we directly adopt an optimal routine (which uses quasi-Newton methods) for the solution to avoid the tedious derivation process.

The same manner is employed on the sum of (4.61) and (4.62) to obtain the optimiser for the correlation matrix of $\boldsymbol{\epsilon}_x$, which is $\mathbf{\Sigma}_x$ with constrains on the diagonal elements being 1's. It should be noted that the correlation coefficients are restricted between -1 and 1 to ensure positive definiteness when two latent variables are involved, that is, $Q = 2$. For $Q > 2$, more concerns need to be taken. One can first estimate $\mathbf{\Sigma}_x$ without constraints and then to solve a non-linear programming problem under a constraint that a solution matrix is positive definite. These procedures are to obtain the closest solution (under Frobenius norm) to the preceding estimate.

Define a response vector as $\mathbf{y} = (\mathbf{y}_{R_1}^{\mathsf{T}}, \ldots, \mathbf{y}_{R_Q}^{\mathsf{T}})^{\mathsf{T}}$ where the index R_q denotes the indicator number for the q-th latent variable $\mathbf{x}_q$, and therefore $\mathbf{y}_{R_q}$ is its indicator random vector. $\mathbf{x}$ and $\mathbf{f}$ are the same notation as usual, represented all latent variables and latent functions. Then the hybrid algorithm is listed as follows:

Algorithm 3

- Randomly select M inducing inputs $\bar{\mathbf{z}}_q$ from the N inputs $\mathbf{z}$ for the q-th marginal model.

For each iteration,

E-Step (margins)

- Generate m_i samples of latent variables $\mathbf{x}_q$ using Equations (4.30), (4.31) and (4.36), given the current estimates of the q-th model marginal parameters.

- (Option) Generate m_i samples of latent functions $\mathbf{f}_q$ using Equations (4.41), (4.42) and (4.47), given the current estimates of the q-th model marginal parameters.

- Calculate conditional expectation to achieve the objective functions using (4.60), (4.61), and (4.62)(Option). The objective function here is

$$E_{p(\mathbf{x}_q,\mathbf{f}_q|\mathbf{y}_{R_q})}[\log p(\mathbf{x}_q, \mathbf{f}_q, \mathbf{y}_{R_q}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M}, \hat{\boldsymbol{\theta}}_{h,q}, \hat{\lambda}_{rq}, \hat{\lambda}_{0q}, \hat{\sigma}^2_{y_r})].$$

M-Step (margins)

- Calculate optimal solutions of GP hyper-parameters, factor loadings (with intercepts) and measurement error variances associated with the q-th margin model; for the latter two parameters using (4.63) and (4.64).

- Implement the preceding procedures over all Q marginal models and then fix all estimated parameters for the next step.

E-Step (joint)

- Generate m_i samples of latent variables $\mathbf{x}$ of the joint model using (4.30), (4.31) and (4.36), given the current estimates of $\mathbf{\Sigma}_x$ and all the parameter estimates from fitting marginal models.

- Generate m_i samples of latent functions $\mathbf{f}$ of the joint model using (4.41), (4.42) and (4.47), given the current estimates of $\mathbf{\Sigma}_x$ and all the parameter estimates from fitting marginal models.

M-Step (joint)

- Calculate optimal solutions of the correlation matrix of latent errors associated with the joint model. The objective function here is

$$E_{p(\mathbf{x},\mathbf{f}|\mathbf{y})}[\log p(\mathbf{x}, \mathbf{f}, \mathbf{y}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_{1:Q}^{1:M}, \hat{\mathbf{\Theta}}_h, \hat{\mathbf{\Lambda}}, \hat{\boldsymbol{\lambda}}_0, \hat{\mathbf{\Sigma}}_x, \hat{\mathbf{\Sigma}}_y)].$$

The above algorithm divides into two estimation steps - one for margins, the other for the joint model. Each implements MCEM methods and the parameter estimates of marginal models are utilized for the estimation of the joint model. It is like the IFM approach, but the first step is to fit multivariate margins. In addition, all estimation procedures for the variances of latent errors are omitted due to the constrains of the auto-covariances being 1's. Therefore, it is optional to sample the latent function values $\mathbf{f}_q$.

In addition, the MCMC approximations in this algorithm (Equations (4.60)-(4.62)) can be not implemented. It is because the closed form of the objective function (the conditional complete likelihood) can be derived under distributional (Gaussian) assumptions of sparse GP-SEM. In this case, the conditional expectation of the sufficient statistics related to latent variables and latent functions are demanded for the derivation.

Note that the time complexities of the algorithm are roughly $O(M^2NQ)$ for the first step and $O(M^2NQ^2)$ for the second. This is because the cost of sampling latent variables and latent function values for per margin is $O(M^2N)$, and that for the joint model is $O(M^2NQ^2)$.

This hybrid algorithm also inspires us to implement the estimation methods presented in Section 4.3 (or 4.2) in the same fashion. The two-step procedure forms another computational scheme, and we raise its application in Chapter 5 and 6. The experiment results are further discussed there, compared with those of estimating all model parameters simultaneously.

4.7 Predictive Distribution

Given the estimates of model parameters, the predictive distribution of a new response vector $\mathbf{y_{new}}$ can be derived under the model structure and distributional assumptions of sparse GP-SEM[14]. More precisely, we are given a new covariate vector $\mathbf{z_{new}}$, the original dataset consisting of response vectors $\mathbf{y}^{1:N} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}\}$, covariate vectors $\mathbf{z}^{1:N}$, and all the estimated model parameters including GP hyper-parameters $\mathbf{\Theta}_h$, pseudo-input sets $\bar{\mathbf{z}}_{1:Q}^{1:M}$, factor loadings $\mathbf{\Lambda}$, intercept terms $\boldsymbol{\lambda}_0$, covariance matrix of latent errors $\mathbf{\Sigma}_x$ and of measurement errors $\mathbf{\Sigma}_y$. Then the predictive distribution of $\mathbf{y_{new}}$ can be written as:

[14] There is no difference between predictive distributions under the alternative version of GP-SEM defined by Eqns. (3.2), (3.4),(3.5), (3.7) and (3.8) and the original version defined by (3.1)-(3.5). Here we adapt the former.

$$
\begin{aligned}
&p(\mathbf{y}_{\mathbf{new}}|\mathbf{z}_{\mathbf{new}}, \mathbf{y}^{1:N}, \mathbf{z}^{1:N}, \boldsymbol{\Omega}) \\
&\qquad = \int p(\mathbf{y}_{\mathbf{new}}|\mathbf{x}_{\mathbf{new}}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}_0, \boldsymbol{\Sigma}_y) \cdot p(\mathbf{x}_{\mathbf{new}}|\mathbf{z}_{\mathbf{new}}, \bar{\mathbf{z}}^{1:M}, \bar{\mathbf{f}}^{1:M}, \boldsymbol{\Theta}_h, \boldsymbol{\Sigma}_x) \cdot \\
&\qquad\qquad p(\bar{\mathbf{f}}^{1:M}|\mathbf{y}^{1:N}, \bar{\mathbf{z}}^{1:M}, \mathbf{z}^{1:N}, \boldsymbol{\Theta}_h) d\mathbf{x}_{\mathbf{new}} d\mathbf{f}^{1\bar{:}M} \qquad (4.65) \\
&\qquad = \int p(\mathbf{y}_{\mathbf{new}}|\bar{\mathbf{f}}^{1:M}, \mathbf{z}_{\mathbf{new}}, \boldsymbol{\Omega}) \cdot p(\bar{\mathbf{f}}^{1:M}|\mathbf{y}^{1:N}, \bar{\mathbf{z}}^{1:M}, \mathbf{z}^{1:N}, \boldsymbol{\Theta}_h) d\bar{\mathbf{f}}^{1:M},
\end{aligned}
$$

where the set of pseudo latent function is denoted as $\bar{\mathbf{f}}^{1:M} = \{\bar{\mathbf{f}}_1^{1:M}, \ldots, \bar{\mathbf{f}}_Q^{1:M}\}$ and $\boldsymbol{\Omega} = \{\boldsymbol{\Theta}_h, \bar{\mathbf{z}}^{1:M}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}_0, \boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y\}$ is the set containing all the estimated model parameters.

In the first equality, the first two integrand factors consist a probabilistic likelihood under sparse GP-SEM; the last factor is a posterior density of the pseudo inputs set $\bar{\mathbf{f}}^{1:M}$. In the second equality, the new latent variables $\mathbf{x}_{\mathbf{new}}$ are integrated out, which is equivalent to merge Equation (3.8) to (3.4). As a result, the first integrand factor becomes a new probabilistic likelihood accordingly.

This predictive distribution has a closed form with a Gaussian density given training data and the estimates $\boldsymbol{\Omega}$. The reason is that the integrand factors are distributed normally (derived from the structure and assumptions of sparse GP-SEM). Rather than deriving the form analytically, the by-product of MCMC simulation can be utilised to achieve integral approximation, estimated by

$$
\frac{1}{N_{mcmc} - N_B} \sum_{i=N_B+1}^{N_{mcmc}} p(\mathbf{y}_{\mathbf{new}}|\bar{\mathbf{f}}^{1:M,[i]}, \mathbf{z}_{\mathbf{new}}, \boldsymbol{\Omega}^{[i]}), \qquad (4.66)
$$

where N_{mcmc} is the length of MCMC simulation, N_B is a burn-in threshold value. The superscript $[i]$ denotes the i-th MCMC iteration. Thus $\boldsymbol{\Omega}^{[i]}$ is the i-th sample of $\boldsymbol{\Omega}$, as the point estimates of the model parameters at the i-th iteration. Due to the Gaussianity of $p(\mathbf{y}_{\mathbf{new}}|\bar{\mathbf{f}}^{1:M}, \mathbf{z}_{\mathbf{new}}, \boldsymbol{\Omega})$, the approximate mean of the predictive distribution is

$$
\frac{1}{N_{mcmc} - N_B} \sum_{i=N_B+1}^{N_{mcmc}} \left[\bar{\mathbf{f}}_c^{[i]} (\boldsymbol{\Lambda}^{[i]})^\mathsf{T} + (\boldsymbol{\lambda}_0^{[i]})^\mathsf{T} \right], \qquad (4.67)
$$

where $\bar{\mathbf{f}}_c^{[i]}$ is a $1 \times Q$ row vector and $\bar{\mathbf{f}}_c^{[i]} = [\mathbf{K}_{1;1M}^{[i]} \mathbf{K}_{1;M}^{[i]} \bar{\mathbf{f}}_1^{[i]}, \ldots, \mathbf{K}_{Q;1M}^{[i]} \mathbf{K}_{Q;M}^{[i]} \bar{\mathbf{f}}_Q^{[i]}]$; for $1 \leq q \leq Q$, $\mathbf{K}_{q;1M}$ is a $1 \times M$ vector with $[\mathbf{K}_{q;1M}]_{1,m} = k_q(\mathbf{z}_{\mathbf{new}}, \bar{\mathbf{z}}_q^{(m)})$.

Furthermore, one can generate a new dataset through the predictive distribution, based on the mean and the covariance matrix of $\mathbf{y}_{\mathbf{new}}|\bar{\mathbf{f}}^{1:M,[i]}, \mathbf{z}_{\mathbf{new}}, \boldsymbol{\Omega}^{[i]}$, which are respectively

$$
\bar{\mathbf{f}}_c^{[i]} (\boldsymbol{\Lambda}^{[i]})^\mathsf{T} + (\boldsymbol{\lambda}_0^{[i]})^\mathsf{T}, \qquad (4.68)
$$

and

$$\mathbf{\Lambda}^{[i]}(\mathbf{\Sigma}_x^{[i]} + \mathbf{V}_{\mathbf{new}}^{[i]})(\mathbf{\Lambda}^{[i]})^{\mathsf{T}} + \mathbf{\Sigma}_y^{[i]}, \tag{4.69}$$

where $\mathbf{V}_{\mathbf{new}}^{[i]}$ is the i-th sample of $\mathbf{V}_{\mathbf{new}}$, where $\mathbf{V}_{\mathbf{new}}$ is a $Q \times Q$ diagonal matrix consisting of all $\mathbf{V}_q$, defined in Equation (3.6) but only evaluated at $\mathbf{z}_{\mathbf{new}}$.

Regarding the hybrid implementation, the approximate mean of the predictive distribution are almost identical to those represented by (4.67). The differences are that there is no burn-in threshold value and the MCMC sample size is the one used in the last optimisation step.

4.8 Greedy selection

The aforementioned procedure provides us a foundation to evaluate the model predictive performance of sparse GP-SEM. All the estimates are from fitting model on a randomly-selected pseudo-input set. Such pseudo-input set (with a certain size) likely capture the main characters of regression relationship between inputs and latent variables.

One of model improvement ideas may emerge is that based on a certain criterion, to choose a set of pseudo input can provide better input locations so that the prediction achieves decent or greater accuracy. Such greedy selection scheme is used frequently in sparse GP approximation methods (Seeger et al. 2003, Teh et al. 2005, Smola & Bartlett 2001, Titsias 2009). We consider a criterion related to entropy.

Entropy is a measure of information complexity of a random variable or process (MacKay 2003). Low entropy reveals less uncertainty. Intuitively, a random variable given more information from another variable can be more certainty for predicting behaviour of a stochastic phenomena. This implies the magnitude of entropy of a conditional distribution is smaller than that of the unconditional counterpart. The difference between with and without extra message can be referred to as information gain in entropy (IGE), which can be one of selection criterion for pseudo inputs. Choosing which input from the training set $\mathbf{z}^{1:N}$ is based on whether the selected input maximises the IGE sum for all the conditional latent function values.

More specifically, let $\mathcal{I}$ be a set consisting the first N integers which index the N training inputs; $\mathcal{S}$ be the current selection index set at the m-th selection step, where the size is $m-1$ and $\mathcal{S} \subseteq \mathcal{I}$. We choose a from $\mathcal{I}$ for maximising the set function $D(a)$

formulising the overall IGE of the conditional GP latent function value $f_q^{(i)}$, given the latent function values generating through the current selection set $\mathcal{S}$ and the additional input a. That is,

$$\max_{a\in\mathcal{I}} D(a) \quad = \quad \max_{a\in\mathcal{I}} \sum_{i\in\mathcal{I}} \Big[H(f_q^{(i)}|\mathbf{f}_q^{(\mathcal{S})}) - H(f_q^{(i)}|\mathbf{f}_q^{(\mathcal{S}\cup\{a\})})\Big]. \tag{4.70}$$

After the m-th input is selected, it is updated into the selection set $\mathcal{S}$ for the next selection step. This selection session stops until the size of $\mathcal{S}$ reaches the pre-specified number M.

Here the entropy H of the latent function value $f_q^{(i)}$ given the current selection set $\mathcal{S}$ and the additional input index a is

$$\begin{aligned} & H(f_q^{(i)}|\mathbf{f}_q^{(\mathcal{S}\cup\{a\})}) \\ = \; & \frac{1}{2}\ln\{(2\pi e)\Big|k_q(\{i\},\{i\}) - k_q(\{i\},\mathcal{S}\cup\{a\})\cdot[k_q(\mathcal{S}\cup\{a\},\mathcal{S}\cup\{a\})]^{-1}\cdot \\ & \qquad\qquad k_q(\mathcal{S}\cup\{a\},\{i\})\Big|\} - const. \end{aligned} \tag{4.71}$$

This equation is derived by the definition of entropy and the identity (A.6) about the marginal and conditional normal distribution. Furthermore, $k_q(\{i\},\{i\})$, $k_q(\{i\},\mathcal{S}\cup\{a\}))$ and $k_q(\mathcal{S}\cup\{a\},\mathcal{S}\cup\{a\})$ are respectively a scalar, a row vector and a matrix that the q-th covariance function $k_q(\cdot,\cdot)$ evaluates at the inputs indexed as $\{i\}$, the set $\mathcal{S}$ and $\{a\}$[15]. The involved hyper-parameters are the point estimates from fitting the q-th marginal model with a randomly-selected pseudo-input set; and fixed during selection.

Note that in practice, the IGEs in Equation (4.69) are calculated by matrix operations rather than entry by entry for saving time. All the matrices configured from $k_q(\{i\},\{i\})$, $k_q(\{i\},\mathcal{S}\cup\{a\})$, over all i and $[k_q(\mathcal{S}\cup\{a\},\mathcal{S}\cup\{a\})]^{-1}$ can be saved for calculation in the next selection step. The new matrices associated with $k_q(\{i\},\mathcal{S}\cup\{a\})$and $[k_q(\mathcal{S}\cup\{a\},\mathcal{S}\cup\{a\})]^{-1}$ are formed based on expanding the old matrices by adding one column or one row of the original covariance matrix $\mathbf{K}_{q;N}$, corresponding to the selected index.

The above greedy selection scheme proceeds with fixed GP hyper-parameters. This is different from the common EM-like manner that one step is to calculate criterion value and the other step is to select an input are implemented alternatively. The latter manner can be time-consuming under our model framework.

[15] Here the expression of $k_q(\cdot,\cdot)$ is different as before. For simplicity we do not use the notation for an input vector $\mathbf{z}^{(n)}$ but keep the superscript for covariates index.

4.9 Remarks

Based on the ergodic properties of the constructed Markov chains, MCMC methods provide great availability to estimate parameters for more complex modelling structure, though samples are dependant and a burn-in threshold value is unknown in prior. The first two estimation algorithms for sparse GP-SEM mainly depends on the feature of MCMC methods. More specifically, the MH random-walk sampling approach and its special version GS sampling methods are applied. As for potentially strongly-correlated parameters and variables, a strategy of integrating out the associative variables from probabilistic density of the target variable and then sampling it independently is adopted to improve sampling efficiency, which is similar to the collapsed Gibbs sampling framework of (Liu 1994). The second algorithm can achieve more efficiency because fixing pseudo inputs replaces the MH sampling scheme during iterations. In addition, for latent variables and latent function whose sampling distributions are Gaussian densities, we use matrix inversion identities to calculate the mean and covariance matrix boost mixing efficiency.

The third algorithm is mainly founded on optimisation frameworks of MCEM and IFM. Borrowing features of MCMC methods, the E-step evaluates the conditional expectation of the complete likelihood (consisting of observations and samples of latent variables) approximately, and the M-Step conducts parameter optimisation on the objective function (constructed by the preceding conditional expectation). This procedure can be implemented for each margin first and then for the joint model, which follows the IFM estimation scheme. The latter step is to obtain to optimal estimates of dependant parameters between margins, with fixed margin parameters.

Technically speaking, the predictive distribution of a new data point can be derived as a normal distribution. However, in practice the approximate evaluation of the mean can be done simply using the MCMC samples and estimated parameters produced in training process. In addition, the same instruments works for data generation as well.

Based on information gain in entropy, a greedy selection scheme for pseudo input set is designed with fixed GP hyper-parameters, estimated by fitting marginal models in prior. A reduction of predictive error can be possibly achieved.

Chapter 5

Experiments

This chapter provides practice of the proposed methodology and investigates the influence under different circumstances. The structure follows the empirical studies for three data sets. Two are the subjects of Section 5.1 and of Section 5.3 (synthetic data), and one is of Section 5.2 (real criminological data). Each section starts the introduction of dataset with a brief summary and possible preliminary data process. Each also focuses on two kinds of experiments - learning and prediction. In Section 5.1, we first conduct convergence diagnosis for the estimation results from implementing different parameter initialisation schemes. Then we investigate differences between posterior estimates and true values of latent variables, and differences in estimates between different model structures. The experiments for multiple-output prediction proceed with varied number of pseudo inputs, selection schemes and estimation methods. The assessment of whether the model fits the data is also implemented. In Section 5.2, we present the results from real data in a similar way as Section 5.1, and add the prediction study with varied latent variables. In Section 5.3, we investigate the experiment results by using Bayesian treatment with two computational strategies in the case of more latent variables involved. This chapter closes with remarks in Section 5.4.

Note that all experiments are implemented through our Matlab subroutines under a personal PC with a CPU Intel i5 core 3.2 GHz and 8 RAM.

5.1 Study I - Synthetic Data

5.1.1 Data description

The properties of the first multiple-output regression dataset are summarised in Table 5.1.

Table 5.1: Properties of Dataset I

Dataset	Input		Output	
Size (N)	Dimension (D)	character	Dimension (R)	character
2000	10	continuous	6	continuous

This data are basically generated by the procedures: 1. follow two functional forms $f_1(\mathbf{z}) = c_1 \sum_{l=1}^{10} z_l^2$ and $f_2(\mathbf{z}) = c_2 \sum_{l=1}^{10} cos(z_l)$ to produce latent function values, where c_1 and c_2 are constants; 2. add Gaussian noises to generate x_1 add x_2; multiply posited factor loadings; 3. add Gaussian measurement noises to generate all responses.

Figure 5.1 and Figure 5.2 provide the histograms of covariates and all response variables respectively. In Figure 5.1 each dimension of a covariate is distributed standard-normally. It maybe imply that the prominent pattern of regression relationship between covariates (ranged from -3 and 3) and latent variables, could be captured by random selection for some pseudo inputs. In Figure 5.2, the distributional shapes reveal that all responses seem to have a Gaussian density. In fact, the distributions of covariates and responses reflect the posited data generation mechanism.

The pattern of correlation coefficients shown in Table 5.2 reflects the model structure used in data generation. Two groups of response variables (y_1, y_2, y_3 and y_4, y_5, y_6) are designed to measure individual latent variables (that is, $Q = 2$). As seen, the variables within the groups have stronger inter-correlations than the variables between the groups.

In the following sections, we adopt a technique of standardisation on the dataset to improve computational efficiency further. We later explore the relationship of the estimated parameters and variables between before and after the data transformation.

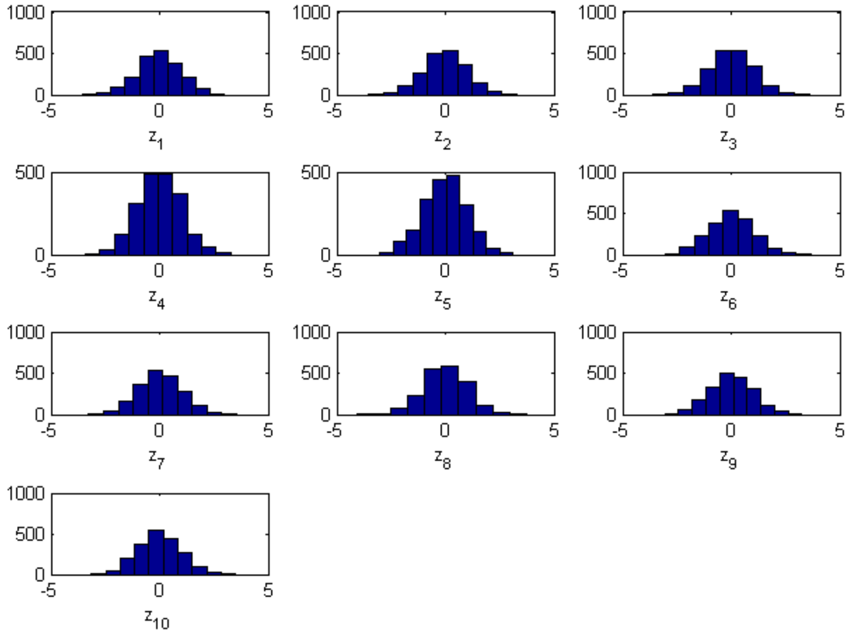


Figure 5.1: Distributions of the inputs for Dataset I.

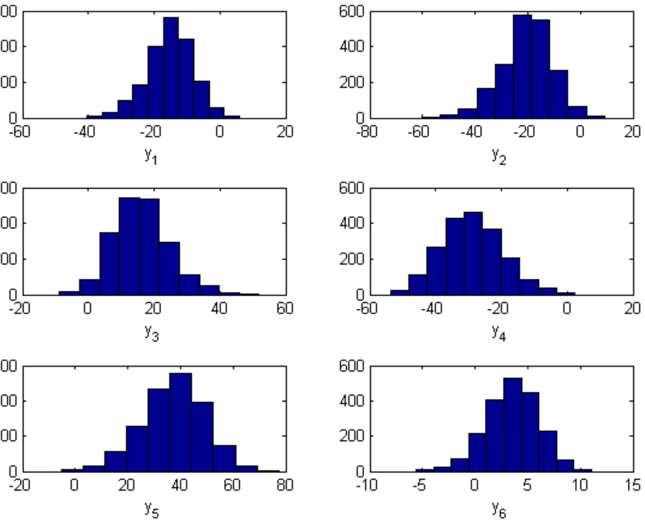


Figure 5.2: Distributions of the outputs for Dataset I.

Table 5.2: Correlation coefficients between response variables of Dataset I

	y_1	y_2	y_3	y_4	y_5	y_6
y_1	1	0.77	-0.77	-0.18	0.19	0.17
y_2		1	-0.76	-0.19	0.19	0.10
y_3			1	0.20	-0.21	-0.20
y_4				1	-0.77	-0.77
y_5					1	0.78
y_6						1

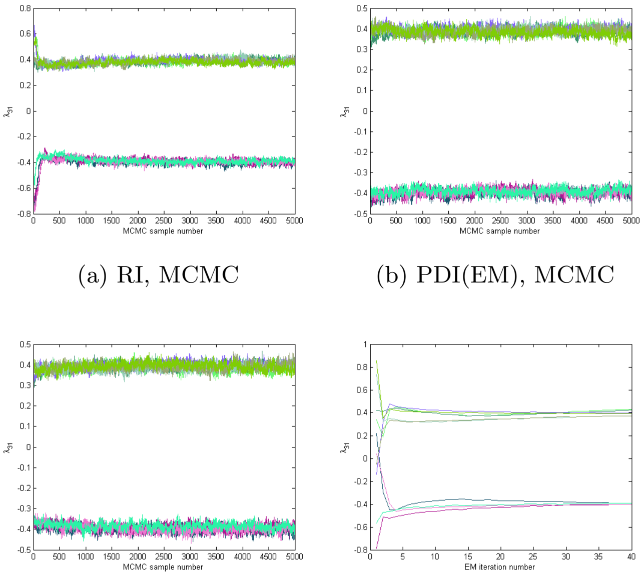
5.1.2 Learning

5.1.2.1 Examination of Parameter Initialization and Model Estimation

To set reasonable initial values for parameters could be necessary in the model estimation proceeding. We use two initialisation settings to investigate the difference of estimation results. They are random initialisation (RI) - randomly generate initial values for all parameters; partial deterministic initialisation (PDI) - deterministically assign initial values for the parameters of marginal models with the prior estimates (from the ergodic average or the point estimate by fitting marginal models using MCMC or MCEM methods, respectively). Although the initial values of correlations are set randomly here, they could be obtained from the estimates from the joint modelling fitting. The settings of random initialisation of correlations connect with the efficient computing strategy we would discuss later.

For other experiment settings, we set an initial number of pseudo inputs M as 200, the number of MCMC samples, N_{mcmc} as 5000, for applying the improved sampler in Section 4.3. And we set the number of samples for marginal model as 3000, the number of EM iterations as 40 with the varied number of inner MCMC samples (details mentioned later), for using the hybrid algorithm in Section 4.6.

In Figure 5.3-5.5, we provide the trace plots of 10-chain estimated parameters with different initialization settings and estimation algorithms - the improved sampling schemes with MCMC methods and the hybrid approach of IFM and MCEM. In the labels, the initialisation setting is specified before the comma. The words in the bracket point out the estimation method for fitting marginal models. The words after the comma indicate the estimation methods for fitting a joint model. For the bottom-right figures, the term



(a) RI, MCMC

(b) PDI(EM), MCMC

(c) PDI(MCMC), MCMC

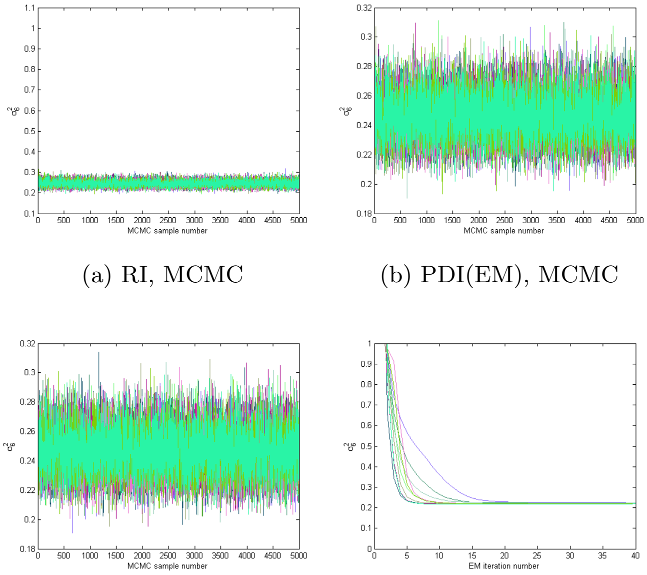
(d) RI,EM-EM

Figure 5.3: The trace plots of 10 simulations for factor loading λ_{31} with using different initialisation settings and estimation methods.

(EM-EM) after the comma represents the use of the hybrid method. Here because of space we only show the plots of one factor loading, a latent error correlation and one measurement error variance.

A simple visual inspection for estimation convergence can be implemented. All the top-left pictures show the 10 chains with different initial values (randomly generated from an appropriate probability density[1]) start fluctuating around a value after some iterations, based on which the value of burn-in can be roughly decided. This perhaps implies the setting of RI does not greatly affect the MCMC parameter estimation. The top-right and bottom-left figures indicate the parameter estimate seemingly remains within a range from the beginning of simulation. This could imply the initial values of parameters, which are achieved by fitting marginal models by both methods, are rather close to the true values. The conditional distributions seem unnecessary to transit to the target distribution after some iterations but remain in it already. The bottom-right plots show the estimates start increasing or decreasing slowly merely after some EM iterations.

[1] The random initial values for factor loadings, latent correlation (cross-covariance of GP latent errors) and measurement error variances, are generated respectively from a Gaussian, uniform, and inverse-gamma distributions.



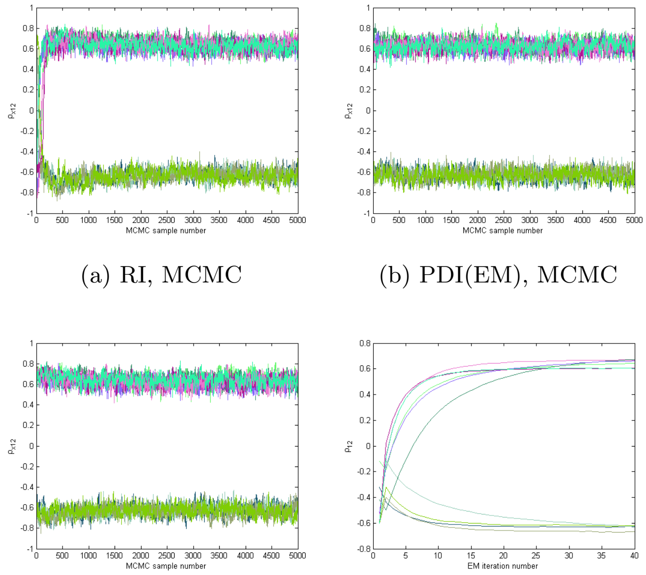
(a) RI, MCMC

(b) PDI(EM), MCMC

(c) PDI(MCMC), MCMC

(d) RI, EM-EM

Figure 5.4: The trace plots of 10 simulations for measurement error $\sigma^2_{y_6}$ with using different initialisation settings and estimation methods.



(a) RI, MCMC

(b) PDI(EM), MCMC

(c) PDI(MCMC), MCMC

(d) RI, EM-EM

Figure 5.5: The trace plots of 10 simulations for latent correlation $\sigma_{x_{12}}$ with using different initialisation settings and estimation methods.

Instead of simple visual convergence diagnosis, the estimated potential scale reduction (EPSR) values can be calculated (Lee 2007, Gelman & Rubin 1992, Gelman et al. 2004). The values assess convergence through variances of between and within multiple simulated sequences. Moreover, the EPSR value of a parameter is calculated in the quotient of the marginal posterior variance of the estimate, obtained by a weighted average consisting of the between-chain and the within-chain variances. If the EPSR values of all parameters are under 1.2, the simulation convergence can be assumed to be achieved. We report the results in the following table. Other convergence diagnostic methods for MCMC methods can be found in a review of (Cowles & Carlin 1996).

Table 5.3: The EPSR values and the relative change rates of estimated parameters under different circumstances for the standardized Dataset I.

parameter	RI, MCMC	PDI(MCMC), MCMC	PDI(EM), MCMC	RI, EM-EM
$\theta_{h,11}$	1.08	1.01	1.01	0.001
$\theta_{h,21}$	1.08	1.01	1.01	0.001
$\theta_{h,12}$	1.11	1.03	1.02	0.001
$\theta_{h,22}$	1.10	1.02	1.03	0.000
$\sigma_{x_{12}}$	1.03	1.01	1.01	0.002
λ_{11}	1.06	1.02	1.02	0.002
λ_{21}	1.06	1.02	1.01	0.002
λ_{31}	1.06	1.02	1.01	0.002
λ_{42}	1.09	1.03	1.02	0.002
λ_{52}	1.09	1.03	1.03	0.002
λ_{62}	1.09	1.03	1.03	0.002
λ_{02}	1.00	1.00	1.00	0.004
λ_{03}	1.00	1.00	1.00	0.004
λ_{04}	1.00	1.00	1.00	0.003
λ_{05}	1.00	1.00	1.00	0.003
$\sigma^2_{y_1}$	1.00	1.00	1.00	0.000
$\sigma^2_{y_2}$	1.00	1.00	1.00	0.000
$\sigma^2_{y_3}$	1.00	1.00	1.00	0.000
$\sigma^2_{y_4}$	1.00	1.00	1.00	0.000
$\sigma^2_{y_5}$	1.00	1.00	1.00	0.000
$\sigma^2_{y_6}$	1.00	1.00	1.00	0.000

As seen in Table 5.3 (except the first and the last column), all EPSR values[2] are under

[2] For latent correlation and factor loadings, due to possible sign differences under the parameter con-

1.2. Some EPSR scales under the setting RI are slightly higher than those under PDI. This could result from the transits of some chains of parameters starting initial values departing from the true values. The EPSR scales of the measurement error variances under RI and PDI are the same because of simulating efficiently.

For convergence diagnosis of the hybrid method, the EPSR values of parameters are unable to be produced due to the optimisation estimation. Instead, we can calculate the absolute change (the absolute value of difference of the two successive estimates), or relative change (the absolute value of the quotient of absolute change and current estimate) of the estimated values during iterations. Estimation convergence is able to be claimed if the errors are smaller than a pre-defined tolerance error. This stop rule should be examined for several iterations because a criterion value may be smaller than the tolerance by chance. Note that even when convergence is achieved, the estimates could be different due to the initial values, selected pseudo inputs and the nature of stochastic simulation.

With the same conditions used in the deterministic scheme[3], absolute or relative errors could still be greater than the pre-determined tolerance when the iterations finish. This may happen because the errors are dominated by the sampling errors, which are not diminished enough. One philosophy (Booth et al. 2001, Chan & Ledolter 1995, Wei & Tanner 1990) claims the errors can be reduced still by increasing the simulation (MC or MCMC) sample size steadily with iterations. Another philosophy claims that fixing the sample size for each iteration is sufficient (Delyon et al. 1999, Celeux & Diebolt 1992). Here, we adopt the former because of the conventional implementation of MCEM (Wei & Tanner 1990, McCulloch 1997). We set 200 MCMC samples for the first 30 iterations and 500 for the last 10 iterations.

We only report the average relative changes for the last 5 iterations in the last column in Table 5.3. As can be seen, all relative changes are smaller than 0.005. There are some other criteria; for example, we calculate the absolute change of values of the score function (or the Q-function) as an auxiliary tool, which was also considered in (Caffo et al. 2005).

We report the 10-chain average estimated parameters (acquired by calculating the

strains, we calculate the values after taking absolute values for the estimates.

[3] Those conditions contain the same initial values and the iteration number, which is sufficient to ensure convergence under deterministic scheme. And the deterministic scheme is the estimation procedure that the closed-form of conditional expectation of complete likelihood can be derived.

ergodic averages) and the average standard deviation of samples in Table 5.4.

Table 5.4: The true values and the averages of the 10-chain estimates under different circumstances.

	True	RI, MCMC		PDI(EM), MCMC		PDI(MCMC), MCMC		RI, EM-EM
parameter		mean	sd	mean	sd	mean	sd	mean
$\theta_{h,11}$	none	2.04	0.07	2.03	0.06	2.06	0.06	2.26
$\theta_{h,21}$	none	1.78	0.06	1.79	0.06	1.79	0.06	1.70
$\theta_{h,12}$	none	2.54	0.11	2.53	0.11	2.56	0.11	2.86
$\theta_{h,22}$	none	2.26	0.11	2.28	0.12	2.27	0.12	2.20
$\sigma_{x_{12}}$	0.60	0.64	0.05	0.64	0.05	0.64	0.05	0.68
λ_{11}	-0.39	-0.39	0.02	-0.39	0.02	-0.39	0.02	-0.38
λ_{21}	-0.38	-0.39	0.02	-0.39	0.02	-0.38	0.02	-0.38
λ_{31}	0.39	0.39	0.02	0.39	0.02	0.38	0.02	0.38
λ_{42}	-0.39	-0.36	0.02	-0.36	0.02	-0.36	0.02	-0.35
λ_{52}	0.38	0.37	0.02	0.37	0.02	0.37	0.02	0.35
λ_{62}	0.39	0.36	0.02	0.37	0.02	0.36	0.02	0.35
λ_{02}	0.00	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{03}	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{04}	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{05}	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.01
$\sigma^2_{y_1}$	0.23	0.24	0.01	0.24	0.01	0.24	0.01	0.23
$\sigma^2_{y_2}$	0.23	0.24	0.01	0.24	0.01	0.24	0.01	0.23
$\sigma^2_{y_3}$	0.24	0.25	0.01	0.25	0.01	0.25	0.01	0.23
$\sigma^2_{y_4}$	0.24	0.25	0.01	0.25	0.01	0.25	0.01	0.23
$\sigma^2_{y_5}$	0.23	0.23	0.01	0.23	0.01	0.23	0.01	0.22
$\sigma^2_{y_6}$	0.24	0.24	0.01	0.24	0.01	0.24	0.01	0.23

As we see from Table 5.4, under different initialisation of parameters, the estimates obtained by fitting a joint model with using the MCMC method are close to true values (listed in the first column[4]) and almost within the range of one standard derivation. In contrast, the hybrid-method estimates in few of factor loadings and $\sigma_{x_{12}}$ have a comparatively noticeable difference with the true values. The GP hyper-parameters in $\theta_{h,11}$ and $\theta_{h,12}$ differentiate by 0.2-0.3 from those estimated using MCMC methods under the three initialisation settings. The estimated loadings are obtained by averaging the absolute ergodic averages or final point estimates of the chains and then adding the same signs as

[4] In the next subsection, we would explain how to acquire the values.

those of the true ones. The same treatment is used for the latent correlation $\sigma_{x_{12}}$. Furthermore, the magnitude of the standard deviations can be reduced by generating more samples if one would like more precise estimates.

5.1.2.2 Comparison between Posterior Estimated and Exact Latent Variables

To learn the Bayesian estimates of latent variables is sometimes useful in applications, such as cluster analysis. Here we first intend to investigate the similarity between the estimated latent variables and the true ones. However, our posterior estimates are acquired by fitting model to a standardized data. It is necessary to discover the inter-relations of parameters and variables with and without data standardisation.

The posterior estimated latent variables acquired by using MCMC methods are given by

$$\hat{x}_q^{(n)} = E[x_q^{(n)})|\mathbf{Y}] = \frac{1}{N_{mcmc} - B} \sum_{i=B+1}^{N_{mcmc}} x_q^{(n),[i]},$$

where the notation $[i]$ indexes a MCMC sample at the i-th iteration. The calculation of the estimates obtain by using the hybrid algorithm is the same but the denominator is changed to N_{mcmc}, which indicates the sample size in the last optimisation iteration.

To realize the relationship between the estimates upon standardised and non-standardised data, the investigation could start from visual observations.

Figure 5.6 shows two evident features by comparing the three-color clusters, where the blue, green and red indicate the distributions of latent variables under three circumstances (true values, estimates before and after data standardisation). The first is that the prominent distribution shapes remain, and the second is that a translation seemingly exists among the green cluster and the others. The standard deviations of those estimates are rather similar, 2.25 in x_1 and 2.23 in x_2 with an error of ± 0.05, which verifies the analogue distribution feature. The distances between the individual distributional means in x_1 and x_2 reveal that the true latent points and the estimated ones without data standardisation are almost identical. The mean distances between the estimate with data standardisation and the true points are 4.39 in x_1 and 8.61 in x_2. Actually, these magnitudes could be obtained by some simple derivations below.

Assume the first response y_1 is an anchor variable, x_1 is the corresponding latent

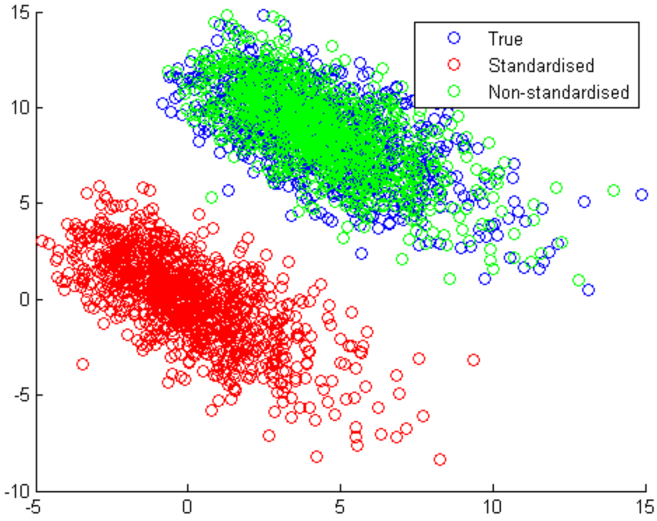


Figure 5.6: The distributions of the true latent variables and the posterior estimates before and after data standardisation.

variable, and let the mean and standard deviation of y_1 be μ_1 and σ_1. Then

$$\begin{aligned}
y_1 &= \lambda_{01} + \lambda_{11}x_1 + \epsilon_{y_1} \\
\Leftrightarrow \frac{y_1 - \mu_1}{\sigma_1} &= \frac{\lambda_{01} - \mu_1}{\sigma_1} + \frac{\lambda_{11}}{\sigma_1}x_1 + \frac{\epsilon_{y_1}}{\sigma_1} \\
\Leftrightarrow \frac{y_1 - \mu_1}{\sigma_1} &= \frac{\lambda_{11}}{\sigma_1}(x_1 + \frac{\lambda_{01} - \mu_1}{\lambda_{11}}) + \frac{\epsilon_{y_1}}{\sigma_1} \\
\Leftrightarrow y_1^* &= \lambda_{11}^* x_1^* + \epsilon_{y_1}^*,
\end{aligned}$$

where the superscript $*$ indicates the variables or parameters being transformed; y_1^* is a standardized response. Next, we assume the second response y_2 is a non-anchor variable which has the same parent latent variable, then

$$\begin{aligned}
y_2 &= \lambda_{02} + \lambda_{21}x_1 + \epsilon_{y_2} \\
\Leftrightarrow \frac{y_2 - \mu_2}{\sigma_2} &= \frac{\lambda_{02} - \mu_2}{\sigma_2} + \frac{\lambda_{21}}{\sigma_2}x_1 + \frac{\epsilon_{y_2}}{\sigma_2} \\
\Leftrightarrow \frac{y_2 - \mu_2}{\sigma_2} &= \frac{\lambda_{21}}{\sigma_2}(x_1 + \frac{\lambda_{01} - \mu_1}{\lambda_{11}}) + (-\frac{\lambda_{21}}{\sigma_2}\frac{\lambda_{01} - \mu_1}{\lambda_{11}} + \frac{\lambda_{02} - \mu_2}{\lambda_{21}}) + \frac{\epsilon_{y_2}}{\sigma_2} \\
\Leftrightarrow y_2^* &= \lambda_{21}^* x_1^* + \lambda_{02}^* + \epsilon_{y_2}^* \qquad .
\end{aligned}$$

The same derivations can be done for the other group of responses. For example, assume y_6 is an anchor variable, x_2 is the corresponding latent variable. Then the difference between the true and the transformed latent variables is $\frac{\lambda_{06}-\mu_6}{\lambda_{62}}$, which also accounts for the distance of 4.39 in x_1 and of 8.61 in x_2 shown in Figure 5.6.

The above derivations uncover the relationship between the original variables (or parameters) and the transformed ones. The transformed measurement error variance $\mathrm{Var}(\epsilon^*_{y_1})$ can also be derived by $\sigma^2_{y_1}/\sigma^2_1$. The differences of $\frac{\lambda_{01}-\mu_1}{\lambda_{11}}$ and $\frac{\lambda_{06}-\mu_6}{\lambda_{62}}$ mathematically explain a translation of the original latent variable and the transformed one. The correlation of latent errors $\sigma_{x_{12}}$ remain the same after data standardisation. The reason is that the standardised covariates and the translation of latent variables merely change the original functional relation, controlled by the new GP hyper-parameters.

Now we can quantitatively compare the transformed latent variables with the posterior estimated counterpart. The mean square error (MSE) is adopted, which is similar to residual sum of square in regression analysis to measure the dissimilarity; the small magnitude suggests learning the main feature of distributions of latent variables is capable. The similarity measure is

$$\frac{1}{N}\sum_{n=1}(\hat{x}_q^{(n)} - x_q^{(n)})^2.$$

It is also interesting to investigate what effects would be on latent variables if no association between them is assumed. In other words, we assume independence between latent errors $\boldsymbol{\epsilon}_x$. Thereby there is no links between latent variables in the model structure.

Table 5.5 summarises the 10 experiment results for that investigation. It contains the mean differences of MSEs under different scenarios and their p-values from using Wilcoxon signed rank test. The first result shows that without association, the MSE differences between two estimation methods (the MCMC and hybrid approaches[5]) is statistically non-significant. This suggests that the estimated latent variables from fitting marginal models using both methods are evidently similar. The second and third results indicate that whichever methods are adopted, statistically significant differences exist in the MSEs of latent variables from fitting model under the structures with and without the linkage. Although the mean differences in x_1 and x_2 are small, the effect of structure difference shows the dissimilarity in latent variables. The fourth comparison suggests that under the linked model structure, the two methods indeed have difference in the estimation of latent variables. This reflects the estimation characteristics distinction of the methods. We later

[5] In this case, one only needs to implement the MCEM method for the first step of the hybrid algorithm and also to carry out the MCMC method for fitting each marginal model.

discuss the distinction in the associated predictive task. It is noted that the last three comparison results perhaps explain the estimation difference in factor loadings and latent error correlation (refer to Table 5.4).

Table 5.5: Comparisons of the MSEs of the posterior estimates for latent variables using two estimation methods, based on 10 experiment results. The subscripts + and $-$ indicate the model structures with and without assuming associations between latent errors (or conditional latent variables given covariates).

	MCMC$^-$ vs. Hybrid$^-$		MCMC$^+$ vs. MCMC$^-$		Hybrid$^+$ vs. Hybrid$^-$		MCMC$^+$ vs. Hybrid$^+$	
	mean diff.	p-value	mean diff.	p-value	mean diff.	p-value	mean diff.	p-value
x_1	-0.001	0.275	-0.027	0.002	-0.017	0.002	-0.011	0.002
x_2	0.006	0.557	-0.076	0.002	-0.012	0.002	-0.069	0.002

5.1.3 Prediction

To evaluate model predictive performance, we use a common measure, root of mean square (RMSE), as our criteria for different predictive tasks. The definition of RMSE is given by

$$\sqrt{\frac{1}{N}\sum_{n=1}(\mathbf{y}_{\mathbf{new}}^{(n)} - \mathbf{y}^{(n)})^2}.$$

For this synthetic data, we create 100 independent instances that 2000 data points are evenly partitioned as training and test points.

5.1.3.1 Comparison of Models with Selection Quantity and Selection Scheme of Pseudo-inputs

Table 5.6 provides the comparison of experiment results. The first results merely suggest that there exist non-linear functional relationships between covariates and responses. This is because the RMSEs by using least square (LS) methods independently are greater than those by using GP regression (GPR). The second outcomes reveal that as the number of randomly selected pseudo inputs increases (with 10, 50, 100), the predictive performance of sparse GP-SEM using MCMC methods becomes parallel to that of GPR. This certainly makes sense because increasing the number more likely captures the non-linear functional relation between covariates and latent variables. The RMSEs of sparse GP-SEM with 100 pseudo inputs are statistically significantly smaller than those with fewer inputs by marginal differences. This may result from the precise capture of non-linear relation and good approximation of model parameters, such as factor loadings.

The third experiment shows that the RMSEs of using greedy selection (GS) scheme indeed outperform those of using random selection (RS) although the differences turn marginal with the number. The reason can be that the greedily-selected pseudo inputs somehow constitutes a set of inputs whose latent function values are overall diffuse and have large enough inter-distance. In other words, the latent function values may sketch the true functional relation to some extent. Increasing the randomly-selected pseudo-input number can make the latent function values compact enough to capture non-linear function relation. This can also happen for the latent function values through greedily-selected pseudo inputs too. In fact, the GS inputs are selected based on the estimates using the RS scheme. Hence, this leads almost identical RMSEs under both selection schemes when the pseudo-input number is large.

The last result shows that with any number of pseudo inputs, the RMSEs of using the MCMC method have difference with those of using hybrid approaches in terms of statistically significance. This reflects the small differences in estimated parameters (shown in Table 5.4) and in latent variables (shown in Table 5.5). It also implies the hybrid method may slightly lost accuracy in estimation and in prediction to some degree. —

Those RMSE difference could be due to the distinction of the methodological nature. For any algorithms, when the number is low, the GP latent functions may not capture the true regression relationship exactly. This may lead inappropriate estimates, such as the GP hyper-parameter estimates, factor loadings and intercepts. The first step of the hybrid method may produce more biased estimates because of limited information from data, consisting of subset of response variables. Fixing those averages for fitting the joint model (that means implementing the second step), the sampled latent functions $\mathbf{f}$ and latent variable $\mathbf{x}$ deviate from the locations where they are generated for fitting a marginal model (refer to the third comparison result in Table 5.5). The deviation force could be due to mutual influence among latent variables through the link of latent errors. The discrepancy of latent-variable location along with the fixed loadings and measurement error variances cause more predictive errors than those solely fitting marginal models for prediction.

In contrast, if predicting responses by fitting the joint model, one would not meet such kind of discrepancy problem. All parameters and variables seemingly reach relatively appropriate locations for prediction. This may be because we use full information of data.

In addition, the differences in RMSE between the two methods decrease with pseudo-input number. It perhaps could result from that the aforementioned discrepancy reduces. Also, the realised function values gradually capture the true regression functional and the estimated parameters become closed.

5.1.3.2 Model Checks with Posterior Predictive Checks

Instead of calculating RMSE for each response, we can also assess the discrepancy between the empirical training data and the replicated data for evaluating model predictive performance. The reason for the assessment is to know whether any features of the observed data are similar to those of the replicated counterparts generated by the posterior prediction distribution. If most of the quantitative feature values of the replicated data cover the observed feature value, it could be believed that model fits data properly. Thus, it implies the model predictive capability could allow one to capture subtle characteristics in outcome data given training inputs.

The assessment is based on posterior predictive check (PPC) (Gelman et al. 1996, Gelman et al. 2004, Gelman & Hill 2007). It follows three procedures: 1. generate a set of fake data by posterior predictive distribution given all inputs used in model fitting; 2. calculate "parameter-related" test statistics (or discrepancy) of the empirical data and of each replicated data, denoted by $T(\mathbf{Y})$ and $T(\mathbf{Y}^{rep})$ respectively; 3. compare the values and calculate the proportion of $T(\mathbf{Y}^{rep})$ greater than $T(\mathbf{Y})$ as posterior predictive (PP) p-value, which functions as similarly as does classical p-value. If the pp p-value is close to 0 or 1, say smaller than 0.05 or greater than 0.95, then it reveals that the replicated data cannot capture the empirical feature in the aspect of test statistics and signals model misfit with high likelihood. Gelman et al. (2004) also suggests practitioners to use multiple test statistics for assessment. This can help to realize model defeats and give an insight for possible model improvement or expansion.

We adopt χ^2-type discrepancy quantity which is commonly used to measure goodness of model fit. It is

$$T(\mathbf{Y}|\boldsymbol{\Omega}) = \sum_{n=1}^{N} \Big\{ [\mathbf{y}^{(n)} - E(\mathbf{y}^{(n)}|\boldsymbol{\Omega})]^{\mathsf{T}} Cov(\mathbf{y}^{(n)}|\boldsymbol{\Omega})^{-1} [\mathbf{y}^{(n)} - E(\mathbf{y}^{(n)}|\boldsymbol{\Omega})] \Big\}, \tag{5.1}$$

Table 5.6: Comparisons of the RMSEs under different circumstances, based on 100 experiment results. Here M means the number of psuedo inputs; the superscript IND, RS and GS indicate independent implementation, random and greedy selection schemes for pseudo inputs, respectively.

LR^{IND} vs. GPR^{IND}						
	mean diff.	p-value				
y_1	0.352	0.000				
y_2	0.339	0.000				
y_3	0.341	0.000				
y_4	0.339	0.000				
y_5	0.343	0.000				
y_6	0.345	0.000				
GPR^{IND} vs.	M = 10		M = 50		M = 100	
Sp. GP-SEM^{RS} (MCMC)	mean diff.	p-value	mean diff.	p-value	mean diff.	p-value
y_1	-0.280	0.000	-0.009	0.000	0.009	0.000
y_2	-0.266	0.000	-0.087	0.000	0.012	0.000
y_3	-0.276	0.000	-0.093	0.000	0.010	0.000
y_4	-0.264	0.000	-0.092	0.000	0.009	0.000
y_5	-0.259	0.000	-0.093	0.000	0.009	0.000
y_6	-0.262	0.000	-0.092	0.000	0.010	0.000
Sp. GP-SEM^{RS} (MCMC) vs.	M =10		M =50		M =100	
Sp. GP-SEM^{GS} (MCMC)	mean diff.	p-value	mean diff.	p-value	mean diff.	p-value
y_1	0.045	0.000	0.017	0.000	0.000	0.660
y_2	0.046	0.000	0.015	0.000	-0.000	0.315
y_3	0.047	0.000	0.016	0.000	0.000	0.332
y_4	0.052	0.000	0.022	0.000	0.001	0.000
y_5	0.059	0.000	0.026	0.000	0.001	0.003
y_6	0.058	0.000	0.022	0.000	0.001	0.000
Sp. GP-SEM^{RS} (MCMC) vs.	M = 10		M =50		M =100	
Sp. GP-SEM^{RS} (Hybrid)	mean diff.	p-value	mean diff.	p-value	mean diff.	p-value
y_1	-0.028	0.000	-0.016	0.000	-0.007	0.000
y_2	-0.026	0.000	-0.015	0.000	-0.008	0.000
y_3	-0.025	0.000	-0.017	0.000	-0.006	0.000
y_4	-0.043	0.000	-0.022	0.000	-0.011	0.000
y_5	-0.033	0.000	-0.021	0.000	-0.012	0.000
y_6	-0.041	0.000	-0.020	0.000	-0.012	0.000

where $\mathbf{\Omega}$ contains all estimated parameters; $E(\mathbf{y}^{(n)}|\mathbf{\Omega})$ and $Cov(\mathbf{y}^{(n)}|\mathbf{\Omega})$ are given by Equations (4.62) and (4.63). Based on this formulation, three test statistics (denoted by $T_1(\mathbf{Y}), T_2(\mathbf{Y})$, $T_3(\mathbf{Y})$[6]) can be used for model fitting assessment. They are respectively to calculate the statistic values for the indicators corresponding to the first, the second latent variable and both. The indicator vector $\mathbf{y}$ in Eqn.(5.1) is (y_1, y_2, y_3), $(y_4, y_5, y_6)$and $(y_1, y_2, y_3, y_4, y_5, y_6)$.

Figure 5.7 shows the replication distribution of the three test statistics for 2000 fake data sets. The PP p-values are 0.357, 0.405 and 0.793, which between 0.05 and 0.095. Also, the three red lines representing the test statistic values for observed data are in the region where more replicated values lie. All indicates the model fits data well and thus we could believe the observed data generated by the model and the factor structure.

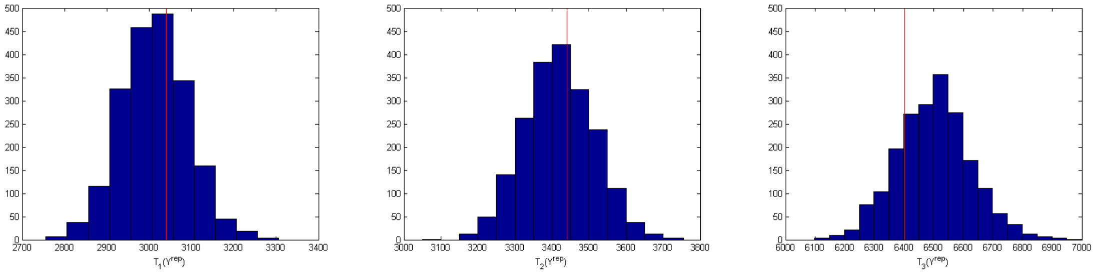


Figure 5.7: The replicated distributions and the observed value for the three test statistics $(T_1(\mathbf{Y}), T_2(\mathbf{Y}), T_3(\mathbf{Y}))$. The vertical red line represents the observed value of the test statistic. The replicated distributions are formed by 2000 replicated statistic values.

Another experiment can be conducted to investigate the effect on PP p-values given different factor structure. We deliberately set a factor structure that (y_1, y_5, y_6) and (y_2, y_3, y_4) measure two different latent variables respectively. After model fitting, the PPC procedure is conducted as the preceding experiment. Figure 5.8 exhibits a strong evidence of model misfitting. It is because for all test statistics the PP p-values are 0, and the observed values are far away from the replicated ones. This experiment suggests that the model checking (using the three test statistics) indeed has adequate power to detect inappropriateness of a “wrong” model[7].

[6] For notation simplicity, we skip the estimated parameters $\mathbf{\Omega}$.

[7] The idea that we use the word ”wrong” here is from George Box’s famous quote “essentially, all models are wrong, but some are useful.” In reality, we never know whether the factor structure used in model fitting is true or not, only through the model checking procedure one knows the inappropriateness.

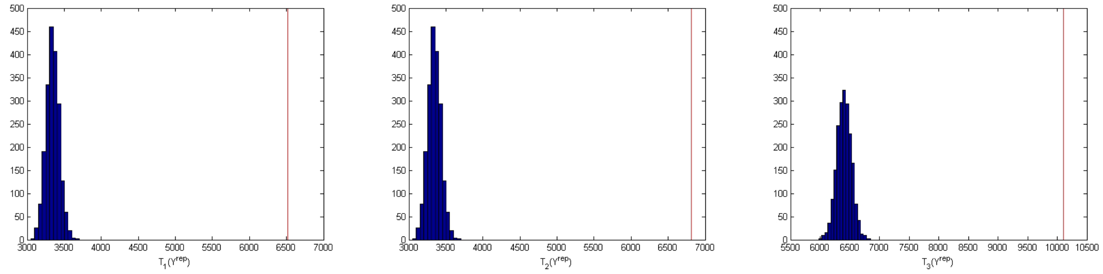


Figure 5.8: The replicated distributions and the observed value for the three test statistics ($T_1(\mathbf{Y}), T_2(\mathbf{Y}), T_3(\mathbf{Y})$) with another factor structure. The vertical red line represents the observed value of the test statistic. The replicated distributions are formed by 2000 replicated statistic values.

5.1.3.3 Discovering Functional Relationship

It is also interesting to realize the univariate functional relationship between a covariate and latent variables.

Since having multiple dimensions of covariates, we need a special procedure. Suppose the parameters estimated from training process are given first. Then one can use Equation (4.67) (without involving loadings and intercepts) to evaluate the predictive latent variable at a specified test input vector, where a constant c is set at a covariate of interest. Next, one has to average the resulting values consisting of all the evaluations over the specified test inputs to marginal out the effect of the other covariates. This is correct if covariates are mutually independent as in our example. Finally, the functional relationship could be realized by evaluating the mean at different values of c.

Figure 5.9 below shows the realized regression relationship between a covariate z_1 and x_1 through prediction. The ten curves are sketched upon the first ten folds of test sets. They all manifest the true quadratic relation. The results of mirror-reflected characteristics are due to the sign difference of loadings.

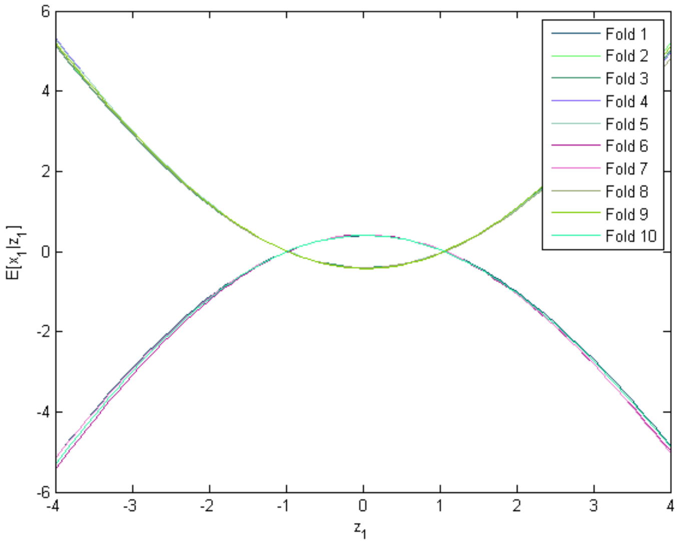


Figure 5.9: The predictive marginal expectation of the regression of latent variable x_1 against a covariate z_1. The expected value of $\mathrm{E}[x_1|z_1 = c]$ is evaluated by averaging all the predictive x_1 over test data points, given a test input vector $\mathbf{z}$, where $z_1 = c$.

5.2 Study II - Criminological Data

5.2.1 Data description

The second dataset is extracted from a US community-and-crime study which combines three associative data in 1990 and 1995.[8] The original study has 2215 data points in total, each represents a US city, and the goal is to investigate whether 129 covariates (including demographic and socioeconomic information) predict the occurrence of 18 target crime variables, such as the number and the rate of robbery incident. The data extraction is that of 129 covariates, we remove some variables which are nominal and of high proportion of missing value, and then only choose variables whose correlations are less than 0.95.[9]. We additionally use log-transformation to adjust the input scale. Next we select target variables whose missing rate is low and which can present the incident number per 100 thousand populations. Finally we delete data points having missing value in target vari-

[8] The data (communities and crime unnormalized data set) can be accessed in UCI machine learning repository website.

[9] Drop columns of the covariate dataset based on the amount of correlation among items. If two variables have an absolute correlation exceeding a particular value, drop the one with the highest index.

ables and extreme values in the predictors (for example a city has more than 100 thousand inhabitants) and then take log on target variables for scale adjustment. The properties of the resulting dataset are summarised in Table 5.7.

Table 5.7: Properties of the dataset II

Dataset	Input		Output	
Size (N)	Dimension (D)	character	Dimension (R)	character
1744	80	continuous	8	continuous

Note that 80 inputs might still be too high for nonparametric GP regression. Some inputs may be highly correlated and have weak influence on the outputs. Here we do not intend to do further variable selection or dimension reduction for covariates. The reason is that we would like to see how the whole framework tackles and what possible significant influence of those predictors on the targets are. Feasible treatments for high-dimensional covariates are left for discussion in conclusion.

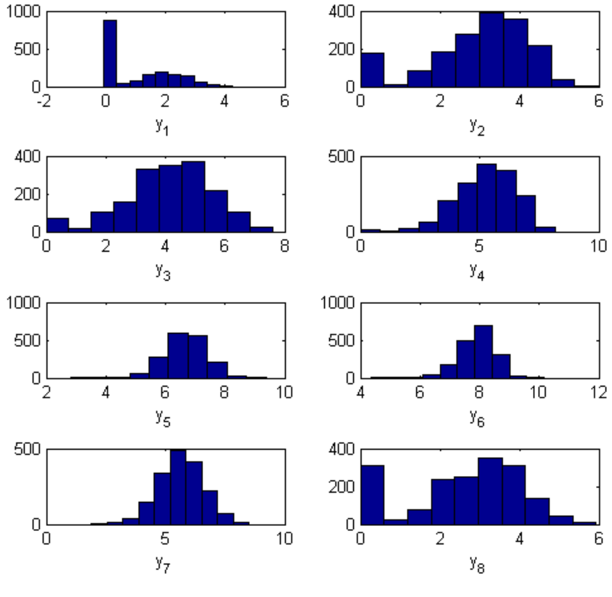


Figure 5.10: Distributions of the outputs for Dataset II.

Because of space, we do not provide visual distributional presentation for the 80 covariates here but brief information. Overall, the inputs are distributed variously. Most have (positive or negative) skewed distributions with different degrees of skewness, but some appear bell-shaped, could possibly assume as normal distributions.

Figure 5.10 shows the distributions of response variables. In the first histogram, high proportion of data points comprises a spike situated near 0, and the rest constitutes a hump located in the right side of the spike and distributed roughly normally. The similar scenario occurs in the second and eighth pictures but the spike is made up of less data points and the hump composes more. The rest of graphs exhibit data points seem to has an nearly normal distribution except the fourth can be regarded as a slightly negative-skewed distribution.

Table 5.8: Correlation coefficients between response variables of Dataset II

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
y_1	1	0.38	0.53	0.48	0.52	0.37	0.42	0.27
y_2		1	0.45	0.50	0.54	0.50	0.38	0.35
y_3			1	0.58	0.70	0.55	0.72	0.35
y_4				1	0.67	0.47	0.57	0.33
y_5					1	0.68	0.64	0.41
y_6						1	0.49	0.36
y_7							1	0.32
y_8								1

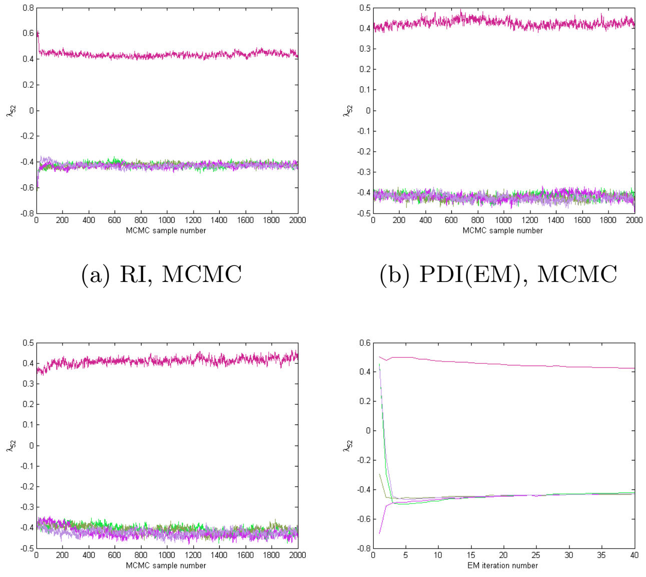
In Table 5.8, we can see most of correlation coefficients of the response variables are low and moderate. There is no clear pattern to classify the variables. However, the 8 target variables represent the numbers of incidents per 100000 people for murders, rapes, robberies, assaults, burglaries, larcenies, auto thefts and arsons. Hence, we could use their literal meanings for roughly grouping the response variables. Then murders, rapes, robberies and assaults are for violent crime; burglaries, larcenies, auto thefts and arsons are for non-violent crime. Our initial posited factor structure with two latent variables is used in most of the subsequent subsections.

5.2.2 Learning

For the experiment settings, the number of pseudo inputs M is set as 200, the number of MCMC samples, N_{mcmc} as 2000. The sample size for fitting a marginal model as 2000, the number of EM iterations as 40 and the inner sample size as 150 for the first 30 iterations and 400 for the last 10 iterations.

5.2.2.1 Examination of Parameter Initialization and Model Estimation

This section proceeds as Section 5.1.2.1 does. Simple visual inspection can be conducted for parameter estimation using different methods (the MCMC and hybrid approach) and two initialisation schemes (RI and PDI). Figure 5.11, 5.12 and 5.13 show the same story about estimation as before for the factor loading, latent correlation, and measurement error variance. The convergence recognition is based on that the absolute estimate values seem consistent. With the RI setting, the 5-chain estimated parameters seem to converge after some MCMC iterations, but with PDI the values seemingly remain. The estimation using the hybrid approach (labelled by (RI, EM-EM) appears convergent at different rates.



(a) RI, MCMC

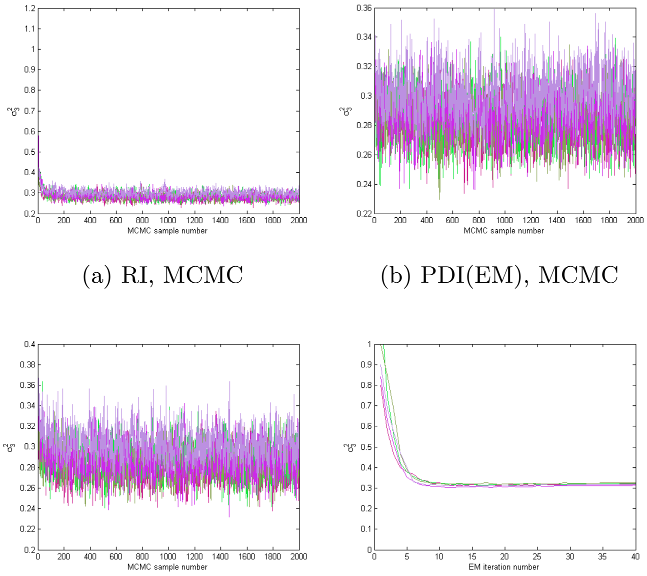
(b) PDI(EM), MCMC

(c) PDI(MCMC), MCMC

(d) RI, EM-EM

Figure 5.11: The trace plots of 5 simulations for factor loading λ_{52} with using different initialisation settings and estimation methods.

Table 5.9 provides quantitative diagnosis for 5 simulation chains. With RI initialisation, most of EPSR values are smaller than 1.2, which indicates setting 2000 MCMC samples seems reasonable. In contrast, all EPSR values with PDI scheme are lower than 1.2, which can result from the fair initial values. All the relative changes of parameter



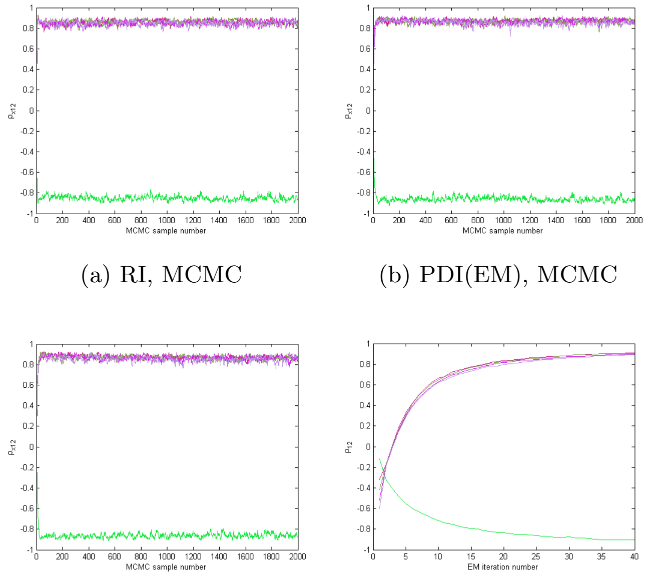
(a) RI, MCMC

(b) PDI(EM), MCMC

(c) PDI(MCMC), MCMC

(d) RI, EM-EM

Figure 5.12: The trace plots of 5 simulations for measurement error $\sigma_{y_3}^2$ with using different initialisation settings and estimation methods.



(a) RI, MCMC

(b) PDI(EM), MCMC

(c) PDI(MCMC), MCMC

(d) RI, EM-EM

Figure 5.13: The trace plots of 5 simulations for latent correlation $\sigma_{x_{12}}$ with using different initialisation settings and estimation methods.

estimates by using the hybrid method are smaller than 0.005, which could be considered convergence.

Table 5.9: The EPSR values and the relative change rates of estimated parameters under different circumstances for the standardized Dataset II.

parameter	RI, MCMC	PDI(EM), MCMC	PDI(MCMC), MCMC	RI, EM-EM
$\theta_{h,11}$	1.15	1.10	1.09	0.001
$\theta_{h,21}$	1.12	1.03	1.03	0.000
$\theta_{h,12}$	1.11	1.08	1.08	0.002
$\theta_{h,22}$	1.19	1.10	1.09	0.000
$\sigma_{x_{12}}$	1.01	1.01	1.01	0.005
λ_{11}	1.09	1.05	1.05	0.003
λ_{21}	1.07	1.04	1.04	0.003
λ_{31}	1.12	1.07	1.06	0.004
λ_{42}	1.06	1.02	1.02	0.005
λ_{52}	1.19	1.16	1.16	0.003
λ_{62}	1.09	1.05	1.05	0.001
λ_{72}	1.08	1.04	1.04	0.001
λ_{82}	1.14	1.05	1.05	0.001
λ_{01}	1.00	1.00	1.00	0.003
λ_{02}	1.00	1.00	1.00	0.004
λ_{04}	1.00	1.00	1.00	0.005
λ_{06}	1.00	1.00	1.00	0.004
λ_{07}	1.00	1.00	1.00	0.004
λ_{08}	1.00	1.00	1.00	0.004
$\sigma^2_{y_1}$	1.09	1.07	1.06	0.000
$\sigma^2_{y_2}$	1.08	1.05	1.05	0.000
$\sigma^2_{y_3}$	1.21	1.16	1.17	0.001
$\sigma^2_{y_4}$	1.10	1.04	1.04	0.001
$\sigma^2_{y_5}$	1.18	1.14	1.13	0.001
$\sigma^2_{y_6}$	1.20	1.12	1.12	0.000
$\sigma^2_{y_7}$	1.12	1.11	1.11	0.000
$\sigma^2_{y_8}$	1.04	1.04	1.04	0.000

Table 5.10 provides all average parameter estimates over 5 chains. With any initialisation schemes, all MCMC estimates excluding GP hyper-parameters are almost identical. By contrast, small differences exist in some of the parameter estimates for using the hybrid method. Note that all scenarios in estimation are rather similar to those for Dataset I. In

addition, there is one interesting point that the estimated latent correlation $\sigma_{x_{12}}$ is high, around 0.9. This may suggest the latent variables (violent crime factor x_1 and non-violent crime factor x_2) are highly positive associated, which can be observed in Figure 5.14. It perhaps concludes that a city with high score in violent crime factor also has high mark in non-violent crime factor. It may imply that if a city has large numbers of incidents in 100000 population for murders, rapes, robberies and assaults, then it would have high incident rates for burglaries, larcenies, auto thefts and arson; vice visa[10].

Figure 5.14 also reveals some information about the posterior estimated latent variables. Firstly, regarding the distributional shapes, a translation seems to occur between the standardised and non-standardised data. This re-verifies the derivation made in the last section for Dataset I. Secondly, the distribution also shows the existence of one cluster. It may indicate that over all data points, each latent variable follows a unimode Gaussian distribution.

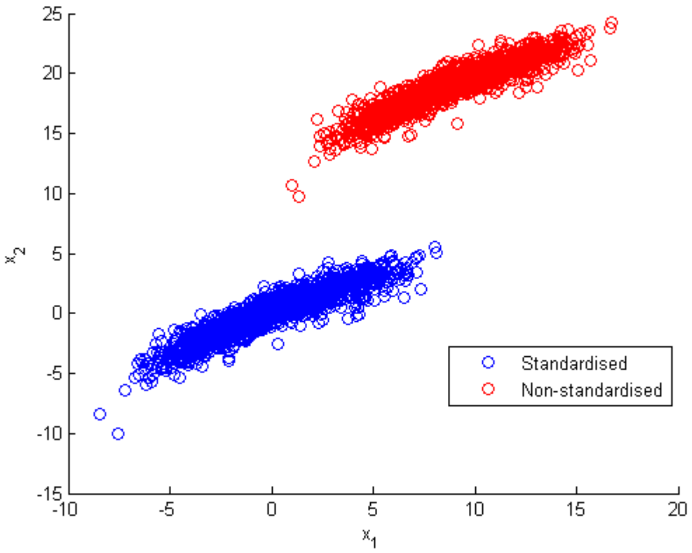


Figure 5.14: The distributions of the true latent variables and the posterior estimates before and after data standardisation for Dataset II.

[10] This interpretation is made under the positive factor loadings.

Table 5.10: The true values and the averages of the 5-chain estimates under different circumstances.

	RI, MCMC		PDI(EM), MCMC		PDI(MCMC), MCMC		RI, EM-EM
parameter	mean	sd	mean	sd	mean	sd	mean
$\theta_{h,11}$	2.78	0.08	2.96	0.04	2.75	0.04	2.89
$\theta_{h,21}$	2.64	0.08	2.73	0.05	2.62	0.06	2.84
$\theta_{h,12}$	1.16	0.10	1.21	0.07	1.18	0.09	1.39
$\theta_{h,22}$	0.70	0.11	0.82	0.09	0.76	0.08	0.92
$\sigma_{x_{12}}$	0.86	0.02	0.86	0.02	0.86	0.02	0.90
λ_{11}	0.21	0.01	0.21	0.01	0.21	0.02	0.19
λ_{21}	0.19	0.01	0.19	0.01	0.19	0.02	0.18
λ_{31}	0.27	0.01	0.27	0.02	0.27	0.02	0.25
λ_{41}	0.23	0.01	0.23	0.01	0.23	0.02	0.22
λ_{52}	0.42	0.02	0.42	0.02	0.42	0.02	0.43
λ_{62}	0.35	0.01	0.35	0.01	0.35	0.02	0.35
λ_{72}	0.33	0.01	0.33	0.01	0.33	0.02	0.32
λ_{82}	0.22	0.01	0.22	0.01	0.22	0.02	0.22
λ_{01}	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{02}	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{04}	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{06}	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{07}	0.02	0.01	0.02	0.01	0.02	0.01	0.01
λ_{08}	0.02	0.01	0.02	0.01	0.02	0.01	0.01
$\sigma^2_{y_1}$	0.56	0.02	0.56	0.02	0.56	0.02	0.55
$\sigma^2_{y_2}$	0.65	0.03	0.66	0.03	0.65	0.03	0.65
$\sigma^2_{y_3}$	0.28	0.02	0.28	0.02	0.28	0.02	0.29
$\sigma^2_{y_4}$	0.47	0.02	0.47	0.02	0.47	0.02	0.47
$\sigma^2_{y_5}$	0.19	0.01	0.20	0.01	0.19	0.01	0.18
$\sigma^2_{y_6}$	0.45	0.02	0.45	0.02	0.45	0.02	0.45
$\sigma^2_{y_7}$	0.51	0.02	0.51	0.02	0.51	0.02	0.52
$\sigma^2_{y_8}$	0.79	0.03	0.79	0.03	0.79	0.03	0.79

5.2.3 Prediction

We use 5-fold cross validation to evaluate the model predictive performance. Therefore, the training and testing sets have around 1395 and 349 data points, respectively.

5.2.3.1 Comparison of Model Predictive Performance with Quantity of Pseudo-inputs

Table 5.11 compares all model prediction results using least-square (LS), Gaussian process regression (GPR), and fitting sparse GP-SEM under the two methods with different numbers of pseudo inputs. The first comparison indicates the RMSE differences between using LS and GPR independently on all the responses have small magnitude. Despite the small differences, one could still regard that all the functional relationships between covariates and responses appear non-linear to some degree. The second comparison results reveal that the difference of RMSEs with GPR and sparse GP-SEM with 10 randomly selected pseudo inputs are statistically significant[11]. The differences turn smaller while one increases the number to 50. At 200 pseudo inputs the RMSE decrement seems not to occur evidently for all responses, but some do have statistically-significant reduction. Similar to the previous study, the RMSE decreases could result from few reasons. First is that when the inducing input number increases, the estimated functional relations between covariates and latent variables turn consistent over all the experiments. Second is that the relations are closed to the true underlying ones to some extent. It should be noted that the magnitudes of differences seem similar to those in the RMSE comparison of LS against GPR. This may imply certain kinds of inappropriateness for fitting sparse GP-SEM so that the RMSEs cannot be lower than those of GPR. The last comparison result re-testifies significant differences in the predictive performance between the MCMC and hybrid methods.

[11] Here we still use "statistically significant" because 0.062 is the lowest p-value one can obtain for 5 pairs of observations when using sign rank test – namely, the signs of the observation difference are only all positive or negative. We believe that if using more pairs of observations, the lowest p-values can decrease further.

Table 5.11: Comparisons of the RMSEs under different circumstances, based on 5 experiment results. Here M means the number of pseudo inputs; the superscript IND, RS and GS indicate independent implementation, random and greedy selection schemes for pseudo inputs, respectively.

LR^{IND} vs. GPR^{IND}						
	mean diff.	p-value				
y_1	0.024	0.062				
y_2	0.016	0.062				
y_3	0.008	0.313				
y_4	0.015	0.062				
y_5	0.012	0.062				
y_6	0.008	0.813				
y_7	0.021	0.062				
y_8	0.015	0.062				
GPR^{IND} vs.	M = 10		M = 50		M = 200	
Sp. GP-SEM^{RS} (MCMC)	mean diff.	p-value	mean diff.	p-value	mean diff.	p-value
y_1	-0.072	0.062	-0.031	0.062	-0.027	0.062
y_2	-0.052	0.062	-0.045	0.062	-0.020	0.062
y_3	-0.170	0.062	-0.059	0.062	-0.013	0.062
y_4	-0.064	0.062	-0.034	0.062	-0.063	0.125
y_5	-0.071	0.062	-0.015	0.062	-0.019	0.125
y_6	-0.080	0.062	-0.070	0.062	-0.044	0.062
y_7	-0.187	0.062	-0.141	0.062	-0.020	0.062
y_8	-0.045	0.062	-0.030	0.062	-0.025	0.062
Sp. GP-SEM^{RS} (MCMC) vs.	M = 10		M = 50		M = 200	
Sp. GP-SEM^{RS} (Hybrid)	mean diff.	p-value	mean diff.	p-value	mean diff.	p-value
y_1	-0.052	0.062	-0.023	0.062	-0.012	0.062
y_2	-0.046	0.062	-0.022	0.062	-0.011	0.062
y_3	-0.064	0.062	-0.028	0.062	-0.018	0.062
y_4	-0.042	0.062	-0.020	0.062	-0.010	0.062
y_5	-0.028	0.062	-0.014	0.062	-0.008	0.062
y_6	-0.023	0.062	-0.012	0.062	-0.004	0.062
y_7	-0.022	0.062	-0.014	0.062	-0.008	0.125
y_8	-0.024	0.062	-0.014	0.062	-0.007	0.062

5.2.3.2 Comparisons in Predictive Performance of Models with Varied Factors

Another experiment can be conducted to assess the effects on RMSE for various factor structures. The beginning model structure is to assign one latent variable to link with all responses. Next we divide responses into two groups $(y_1, \cdots, y_4)$ and $(y_5, \cdots, y_8)$, each of which measures one latent variable. Finally, we repeat even partition on each group and it turns out 4 pairs $(y_1,\ y_2)$, $(y_3,\ y_4), \cdots, (y_7,\ y_8)$ to measure 4 latent variables. All the the above model structures are nested. For the experiments, we use 200 pseudo inputs for fitting model.

Table 5.12 summarises the comparison results in RMSE between GPR and sparse GP-SEM. Overall, there is no clear pattern on RMSE differences for all responses with the number of latent variables although the differences for 4 responses appear to decrease. Furthermore, the RMSEs of GPR are smaller than those of sparse GP-SEM regardless of the latent factor structures. The reasons for the results could be multiple.

Table 5.12: Comparisons of 5 experiment results about the RMSEs under different circumstances. Here the superscript IND means independent implementation on each variable, and RS indicates random selection scheme for 200 pseudo inputs, respectively. Q is the number of latent variables. The model fitting is under the nested factor-loading structure.

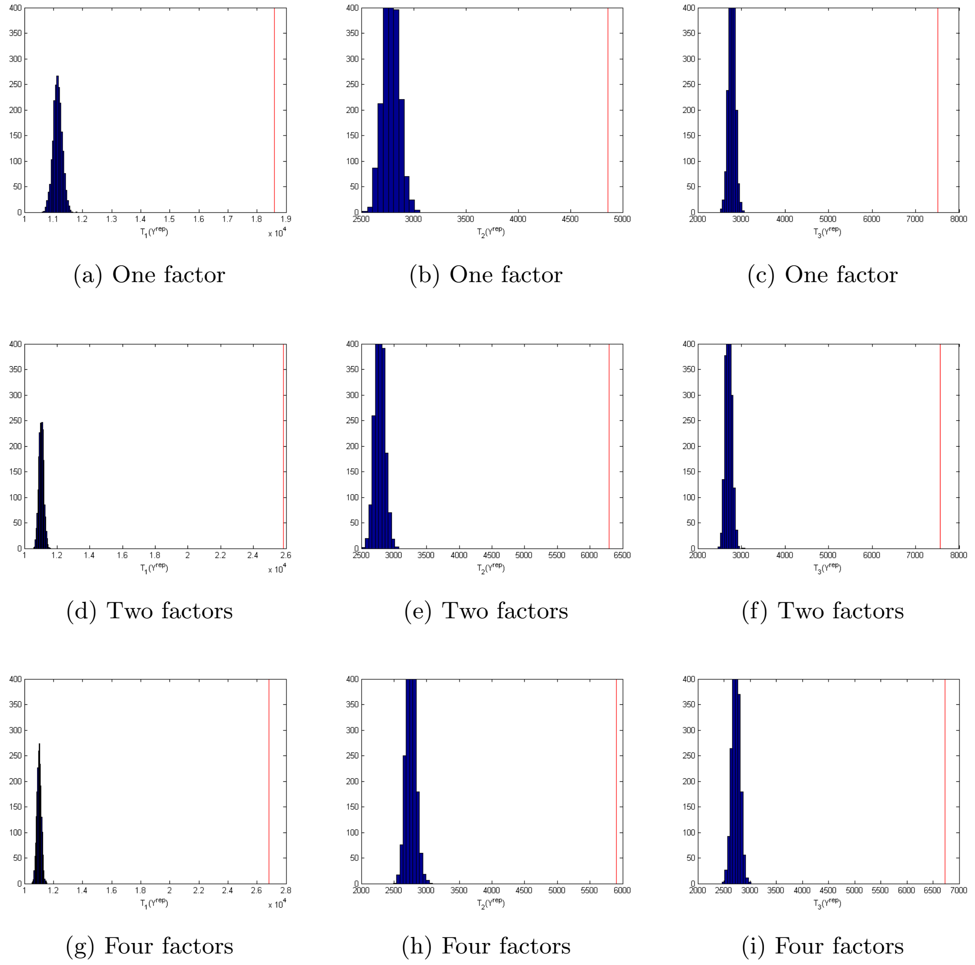
GPR^{IND} vs. Sp. GP-SEM^{RS} (MCMC)	$Q = 1$		$Q = 2$		$Q = 4$	
	diff. mean	p-value	diff. mean	p-value	diff. mean	p-value
y_1	-0.039	0.062	-0.026	0.062	-0.025	0.062
y_2	-0.046	0.062	-0.059	0.062	-0.020	0.062
y_3	-0.069	0.062	-0.040	0.062	-0.013	0.125
y_4	-0.026	0.062	-0.031	0.062	-0.029	0.062
y_5	-0.020	0.062	-0.013	0.062	-0.019	0.062
y_6	-0.089	0.062	-0.065	0.062	-0.044	0.062
y_7	-0.106	0.062	-0.061	0.062	-0.020	0.062
y_8	-0.031	0.062	-0.030	0.062	-0.055	0.062

One may be that the experimental model structures are not appropriate. Some of individual regression relations between response variables and covariates may be rather different and we inappropriately group them for fitting model. To verify the idea above, we perform PPC procedure with the three test discrepancy quantities. The three statistics

$(T_1(\mathbf{Y}), T_2(\mathbf{Y}), T_3(\mathbf{Y})$ are defined based on Equation (5.1) and calculate the values for different groups of responses - $(y_1, \cdots, y_8)$, $(y_1, \cdots, y_4\ )$ and $(y_5, y_6\ )$. The checking results are shown in Figure 5.15. Similar to Figure 5.8, all the 2000 replicated values of test statistics are much smaller than the observed value and extreme PP p-values occur. It recalls that we maybe adopt wrong factor structures and model fitting under all the experimental model structures could be inappropriate.

Another possible reason is the failure of model distributional assumptions – the normality of measurement errors. After obtaining the estimates of the residuals, we can utilise QQ-plots to detect the deviation of normality. Figure 5.16 shows the QQ-plot for the estimated errors from fitting model with the factor structure of 4 latent variables. As seen, overall the observations have severe deviation from the line, especially in those correponding to response variables. This strongly suggests the Gaussianity assumptions are violated[12]. In addition, the other two cases (the model structures for 1 and 2 latent variables) have the same scenarios and conclusions.

[12] We also use another normality testing for the check, such as Lilliefors test. The results consist with the the conclusions we draw for Figure 5.16, all suggest non-normal distributions.

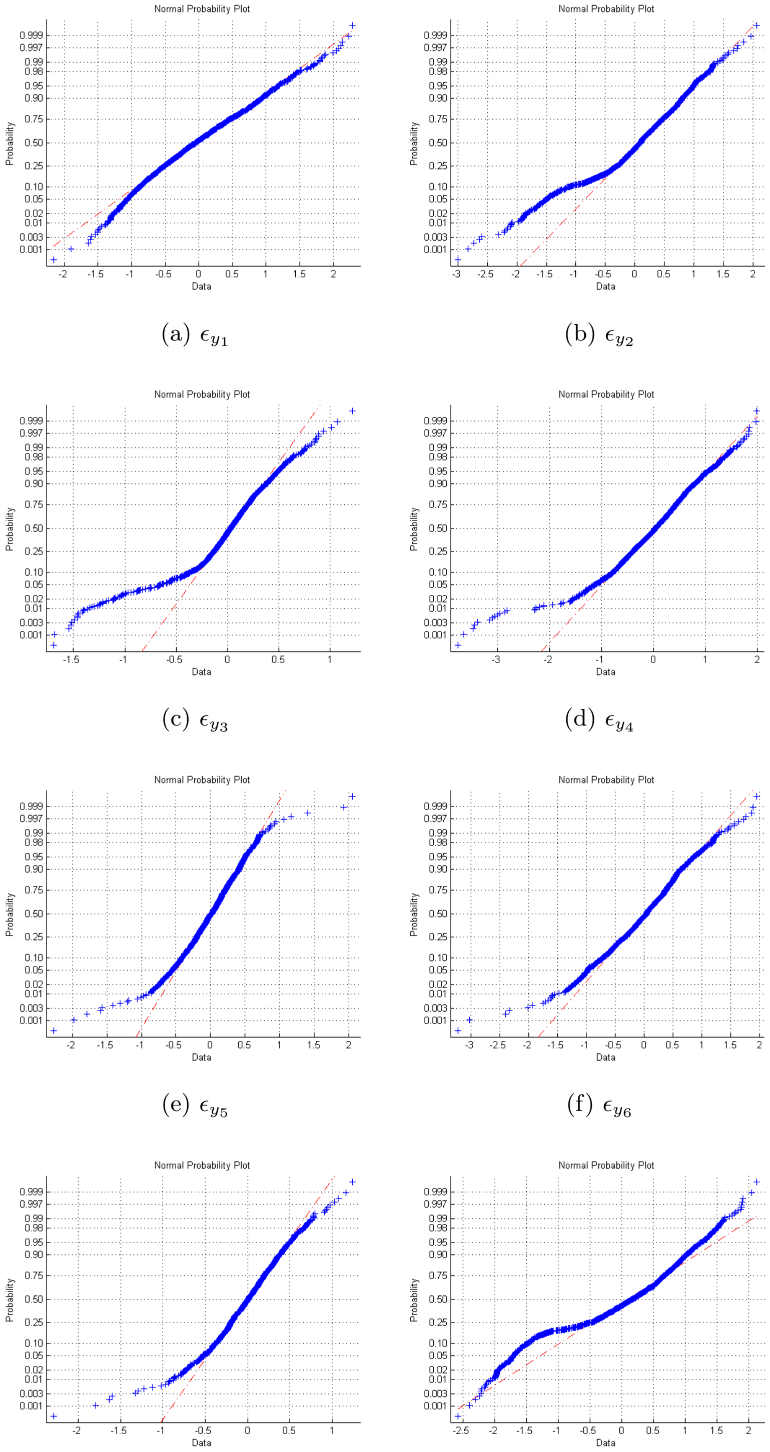


(a) One factor (b) One factor (c) One factor

(d) Two factors (e) Two factors (f) Two factors

(g) Four factors (h) Four factors (i) Four factors

Figure 5.15: The replicated distributions and the observed value for the three test statistics $(T_1(\mathbf{Y}), T_2(\mathbf{Y}), T_3(\mathbf{Y}))$, fitting model with three factor structures. The vertical red line represents the observed value of the test statistic. The replicated distributions are formed by 2000 replicated statistic values.



(a) ϵ_{y1}

(b) ϵ_{y2}

(c) ϵ_{y3}

(d) ϵ_{y4}

(e) ϵ_{y5}

(f) ϵ_{y6}

(g) ϵ_{y7}

(h) ϵ_{y8}

Figure 5.16: The QQ plots for all the estimated measurement errors from fitting model with the 4-factor model structure.

5.3 Study III - Synthetic Data

We move our interests to model fitting with more outcomes and latent variables. The primary reason is to realize the computational performance of our model framework. The secondary is to acquire a rough insight for parameter estimation and prediction under the circumstances.

5.3.1 Data description

The properties of the first multiple-output regression dataset are summarised in Table 5.13.

Table 5.13: Properties of Dataset III.

Dataset	Input		Output	
Size (N)	Dimension (D)	character	Dimension (R)	character
2500	10	continuous	20	continuous

Dataset II is generated based on the posited model structure and fixed parameters. Here we only specify the 10 functional for generating latent function values:

$$
\begin{aligned}
&f_1(\mathbf{z}) = c_1 \sum_{l=1}^{10} z_l^2, \quad f_2(\mathbf{z}) = c_2 \sum_{l=1}^{10} \cos(z_l),\\
&f_3(\mathbf{z}) = c_3 \sum_{l=1}^{10} \exp\{\frac{5}{2}|z_l|\}, \quad f_4(\mathbf{z}) = c_4[(1+\exp\{-\frac{1}{2}\sum_{l=1}^{10} z_l\})^{-1} + \cos(3\sum_{l=1}^{10} z_l)],\\
&f_5(\mathbf{z}) = c_5 \sum_{l=1}^{10} \frac{4}{1+z_l^2}, \quad f_6(\mathbf{z}) = c_6 \sum_{l=1}^{10} \sin(z_l),\\
&f_7(\mathbf{z}) = c_7 \sum_{l=1}^{10} \cos(0.8 z_l), \quad f_8(\mathbf{z}) = c_8 \cos(\frac{1}{10}\sum_{l=1}^{10} z_l),\\
&f_9(\mathbf{z}) = c_9 \cos(\frac{1}{30}\sum_{l=1}^{10} z_l^2), \quad f_{10}(\mathbf{z}) = c_{10} \sum_{l=1}^{10}(\sin(2z_l) + \cos(2z_l)),
\end{aligned}
$$

where $c_1, \ldots, c_{10}$ are constants.

Figure 5.17 and Figure 5.18 gives the histograms of covariates and all response variables respectively. Both reflect the data generation mechanisms. Each dimension of a covariate displays uniform distributions. This maybe imply that the behaviour of regression relationship between covariates and latent variables is possibly captured well near the

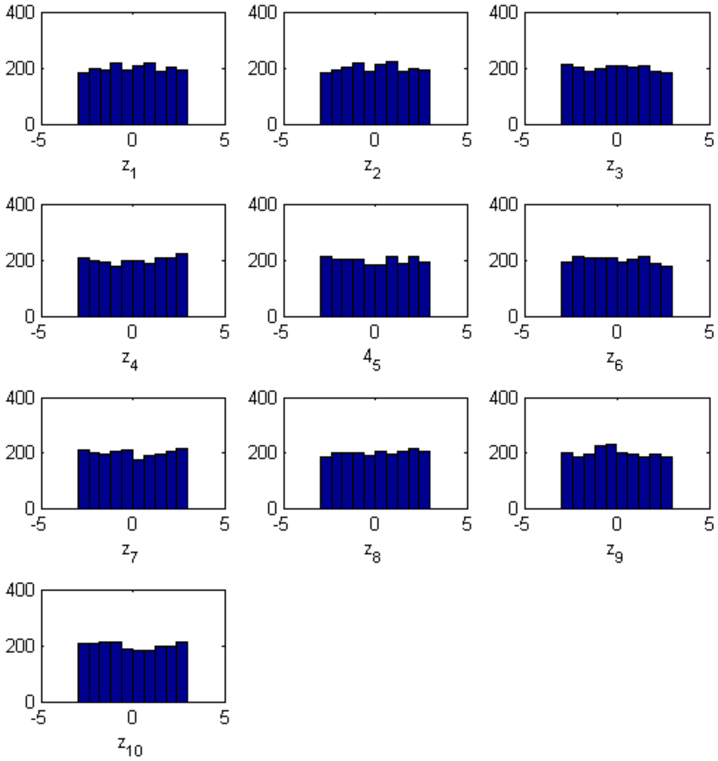


Figure 5.17: Distributions of the inputs for Dataset III.

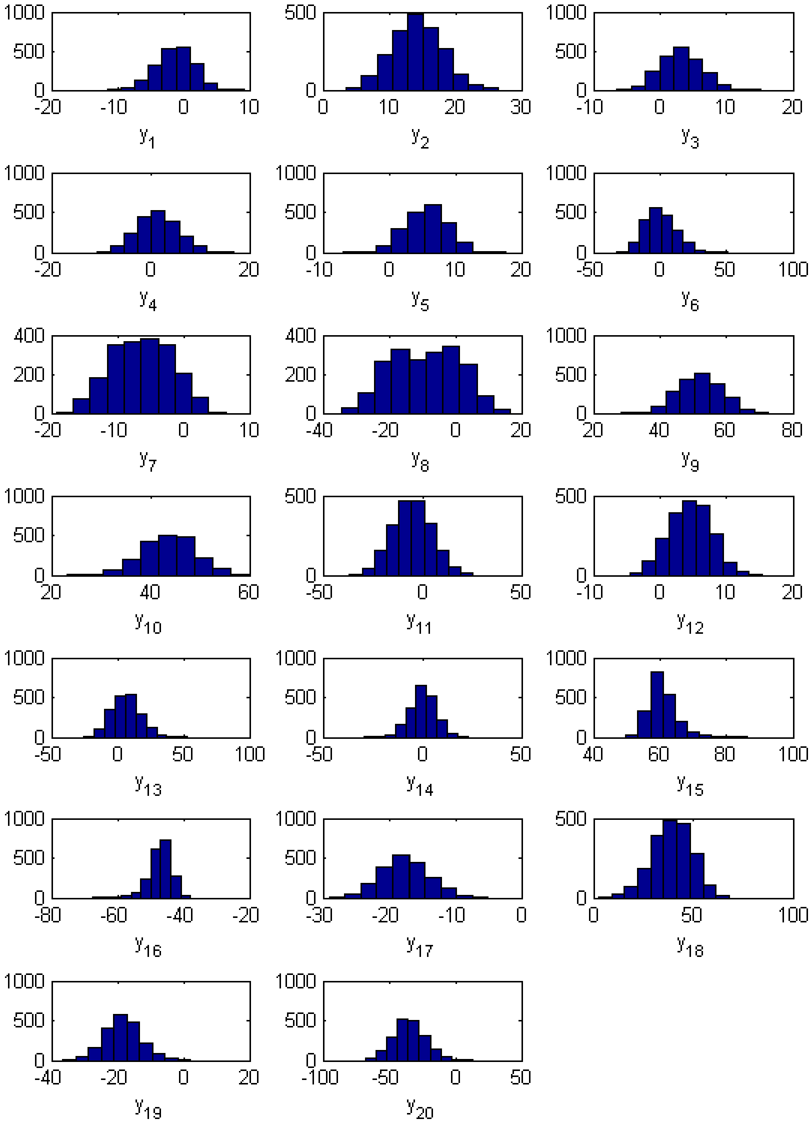


Figure 5.18: Distributions of the outputs for Dataset III.

boundaries if several random selected pseudo inputs are closed to the regions. In Figure 5.18, the distributional shapes show that all the responses have a normal distribution.

Table 5.14 exhibits the correlations of each pair of variables, $(y_1, y_2), (y_3, y_4), \ldots, (y_{19}, y_{20})$ are stronger than those of any other pairs, such as $(y_1, y_3), (y_1, y_5)$ and (y_{18}, y_{20}). The 10 groups of variables can be classified and each group is capable of being assumed to measure one latent variable, as we design. The 10 latent variables ($Q = 10$) are installed as a set-up of fitting model for experiments later.

In the following subsections we focus on the experiments of using Bayesian treatment for model fitting. Two computing strategies are adopted to investigate the difference. One is to implement the MCMC method (Algorithm 2 in Section 4.3) for fitting the joint model. The other is similar to adopt the hybrid approach (Algorithm 3 in Section 4.6), but the MCMC method is used for fitting marginal and joint models.

Table 5.14: Correlation coefficients between response variables of Dataset III.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{18}	y_{19}	y_{20}
y_1	1	-0.70	-0.26	-0.21	0.32	0.11	0.10	0.11	-0.36	-0.35	-0.10	-0.09	-0.28	0.27	-0.13	0.11	0.21	-0.26	-0.11	-0.12
y_2		1	0.20	0.29	0.29	-0.22	-0.16	-0.13	0.28	0.28	0.13	0.13	0.25	-0.24	0.16	-0.15	-0.30	0.38	0.13	0.14
y_3			1	0.68	0.22	-0.22	0.09	0.11	0.28	0.26	-0.11	-0.08	0.26	-0.25	-0.02	0.01	-0.10	0.12	0.35	0.36
y_4				1	0.25	-0.25	0.13	0.15	0.23	0.21	-0.15	-0.10	0.26	-0.24	-0.03	0.03	-0.14	0.16	0.39	0.41
y_5					1	-0.76	-0.11	-0.11	0.26	0.24	0.12	0.11	0.20	-0.28	0.11	-0.10	-0.20	0.22	0.06	0.06
y_6						1	0.17	0.17	-0.23	-0.20	-0.17	-0.17	-0.28	0.27	-0.17	0.16	0.20	-0.23	-0.07	-0.07
y_7							1	0.88	-0.17	-0.15	-0.27	-0.22	0.08	-0.08	-0.16	0.12	0.12	-0.15	0.36	0.33
y_8								1	-0.16	-0.15	-0.22	-0.28	0.11	-0.10	-0.15	0.11	0.11	-0.13	0.23	0.30
y_9									1	0.89	0.15	0.15	0.21	-0.29	0.21	-0.17	-0.20	0.25	0.12	0.11
y_{10}										1	0.13	0.13	0.20	-0.29	0.17	-0.15	-0.28	0.24	0.13	0.12
y_{11}											1	0.75	-0.12	0.12	0.19	-0.15	-0.11	0.11	-0.29	-0.25
y_{12}												1	-0.09	0.08	0.17	-0.10	-0.07	0.07	-0.25	-0.38
y_{13}													1	-0.93	-0.01	0.01	-0.26	0.20	0.20	0.22
y_{14}														1	0.04	-0.04	0.26	-0.29	-0.28	-0.21
y_{15}															1	-0.69	-0.16	0.17	-0.12	-0.12
y_{16}																1	0.14	-0.17	0.09	0.08
y_{17}																	1	-0.82	-0.03	-0.04
y_{18}																		1	0.03	-0.03
y_{19}																			1	0.91
y_{20}																				1

5.3.2 Learning

For the experiment settings, we set the number of pseudo inputs M as 250 for both scenario using two computational schemes. For Scheme 1, the sample size, N_{mcmc} for fitting the joint model as 5000 and adopting random initialisation for parameters. For Scheme 2, and the number of MCMC samples for fitting marginal models as 3000 but the sample size for fitting the joint model, N_{mcmc} as 3000.

5.3.2.1 Comparison in Computation, Parameter Estimation and Prediction

At first, the computation time of implementating the two schemes is reported in the Table 5.15.

Table 5.15: Execution time (second) for the implementation of the two strategies: Scheme 1 for a joint model with 5000 iterations; Scheme 2 for 10 margins with each of which 3000 iterations, and then for the joint model with 3000 iterations. The bold values in the brackets point out the average time per iteration.

	Scheme 1	Scheme 2
Margins	none	9496 (**0.32**)
Joint	231258 (**46.25**)	99188 (**33.06**)
Total Time	231258	108684

The execution time using Scheme 2 is obviously faster than Scheme 1. The reason is that the computational cost of fitting Q marginal models per iteration is $O(NM^2Q)$ for estimate the 70 marginal parameters; and the expense of fitting the joint model is $O(NM^2Q^2)$ for only 45 correlation coefficients. The cost of Scheme 1 is $O(NM^2Q^2)$ for all the 115 parameters. Moreover, these aforementioned costs are reckoned based on sampling latent variables and latent functions under fitting marginal or joint models, which occupies most of execution time.

We produce 5 chains to see differences in estimation. The simulation trace plots for some parameters are provided in Figures 5.19, 5.20 and 5.21. Each of the sub-figures has three coloured reference lines. The red, green and black are for the true value and the ergodic MCMC averages using Scheme 1 and Scheme 2, respectively. Here, the true values are calculated by the derivation similar to that in Section 5.1.1.2 after we inspect the translation of the posterior estimates of latent variables with standardizing and non-standardizing data.

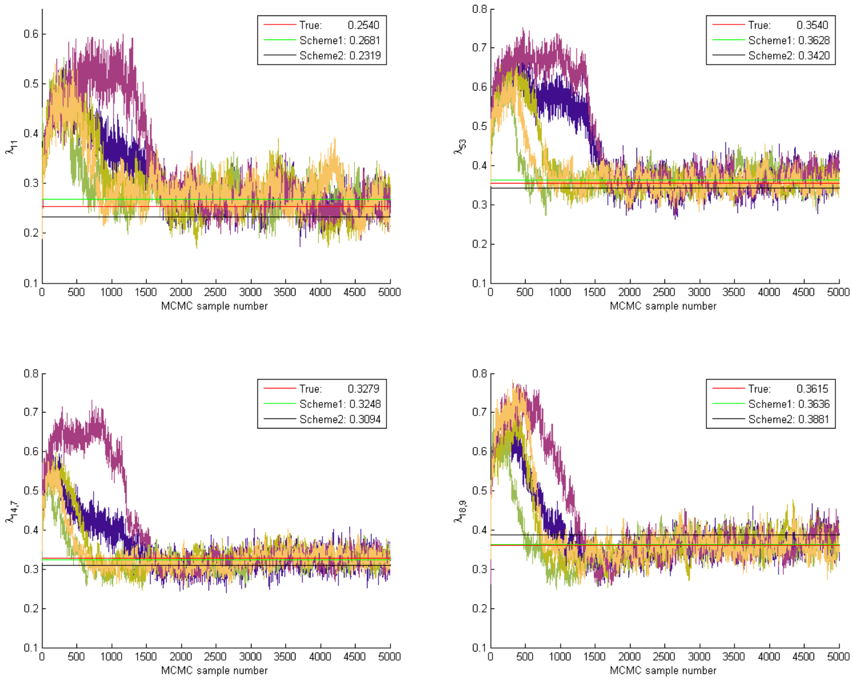


Figure 5.19: The trace plots of factor loading λ_{11}, λ_{53}, $\lambda_{14,7}$ and $\lambda_{18,9}$, from fitting model under Scheme 1.

Some facts can be noticed from the figures. In general, no matter what scheme is used, the ergodic averages are close to the true values. This basically consists with the observations in estimation results in Table 5.4 (for the first synthetic data study). Furthermore, the differences between the Scheme 1 estimates and the true values seem smaller than their counterparts although few parameters appear not, such as, latent correlation $\sigma_{x_{10,7}}$ in the bottom right panel of Figure 5.20. The reason could be that in higher-dimension parameter space, few components likely acquire extreme estimates while the sampling procedure is implemented through a multivariate distribution. Those incidents may affect the estimations; therefore, their estimates have some deviation from the true values. This scenario, for instance, could happen in the present case of 10 latent variables – 45 latent correlations are involved, sampled from a multivariate Wishart distribution.

Another observation is that given the same parameter initial values, the samplings

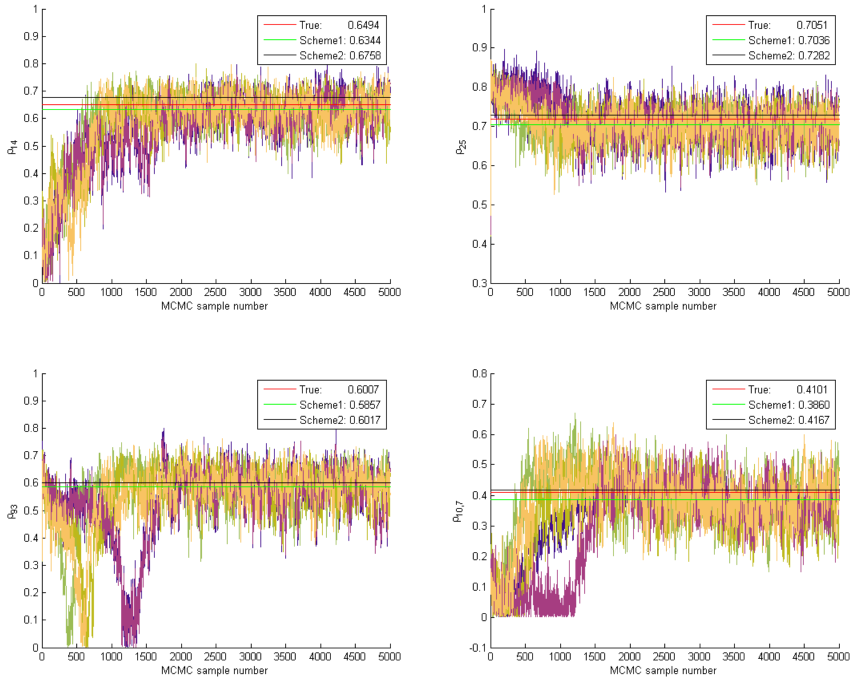


Figure 5.20: The trace plots of latent correlations σ_{x14}, σ_{x25}, σ_{x93} and $\sigma_{x10,7}$ from fitting model under Scheme 1.

of parameters using Scheme 1 may need to undergo more burn-in simulations to reach the target distribution. This frequently happens when the parameter space is of high dimension. By contrast, the two-step sampling procedure of Scheme 2 makes the dimension relatively lower in each step. Hence, the burn-in threshold values seem smaller, as seen in Figure 5.22 (here we only choose a few parameters presented in Figure 5.19 and Figure 5.20 for comparison).

Incidentally, the good mixing in sampling measurement error variances could be because the mutually independence assumptions imposed in the error components do match the data generation mechanism.

5.3.3 Prediction

5 prediction experiments are conducted to know the difference of the two schemes. The set-ups of each experiment are on 1250 data points as training set, the rest as test set and 125 pseudo inputs. The average total run-time for all prediction experiments is around 115578 seconds for Scheme 1, and 54342 seconds for Scheme 2. The times are roughly

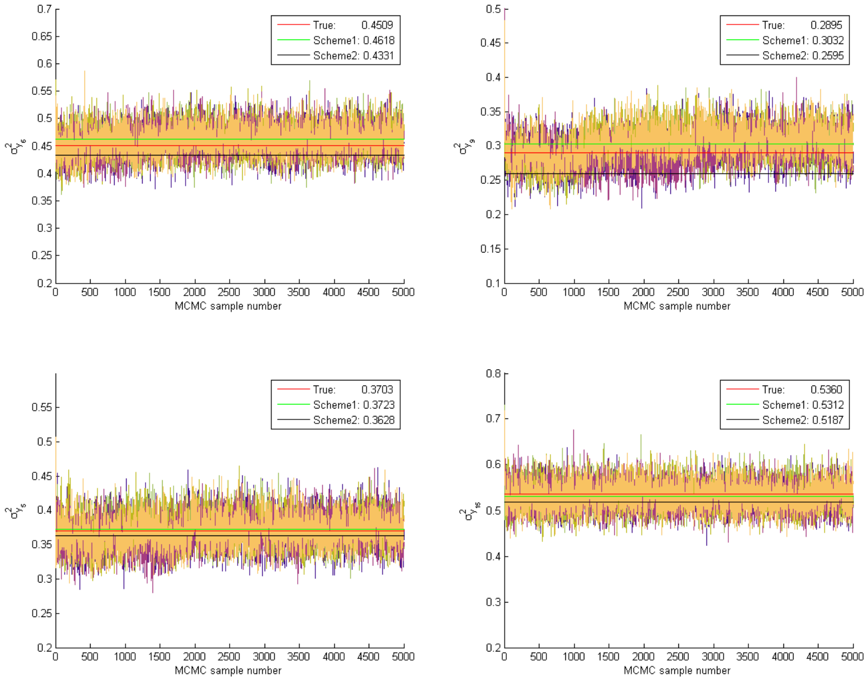


Figure 5.21: The trace plots of measurement errors σ^2_{y5}, σ^2_{y6}, σ^2_{y9} and σ^2_{y15}, , from fitting model under Scheme 1.

half of those shown in Table 5.15. The comparison result is summarised in Table 5.16. As seen, there are statistically significant small differences in most of response predictions. This observation is basically similar to the RMSE comparison between the MCMC and hybrid methods, shown in Table 5.6 for Dataset I and in Table 5.11 for Dataset II.

The PPC procedure can also be performed to see the differences of two computing schemes. Based on 2000 replicated data sets, we only use the χ^2 test statistics through all the response values (that is T_1 used before). The tail-area probabilities beyond the observed value of test statistics $T_1(\mathbf{Y})$ are 0.858 and 0.873 for Scheme 1 and 2 respectively. Here the slight difference may reflect the deviations in estimation and prediction, due to the methodological distinction. In addition the two PP p-valves are still within a reasonable range (between 0.05 and 0.95), which shows model appropriateness.

Considering trade-off on the estimation, computation and prediction, Scheme 2 can be an economical computational strategy in practice, especially more latent variables involved. Practitioners could increase more simulation iterations for fitting marginal models and decrease the iteration number for fitting the joint model. This helps one to obtain as

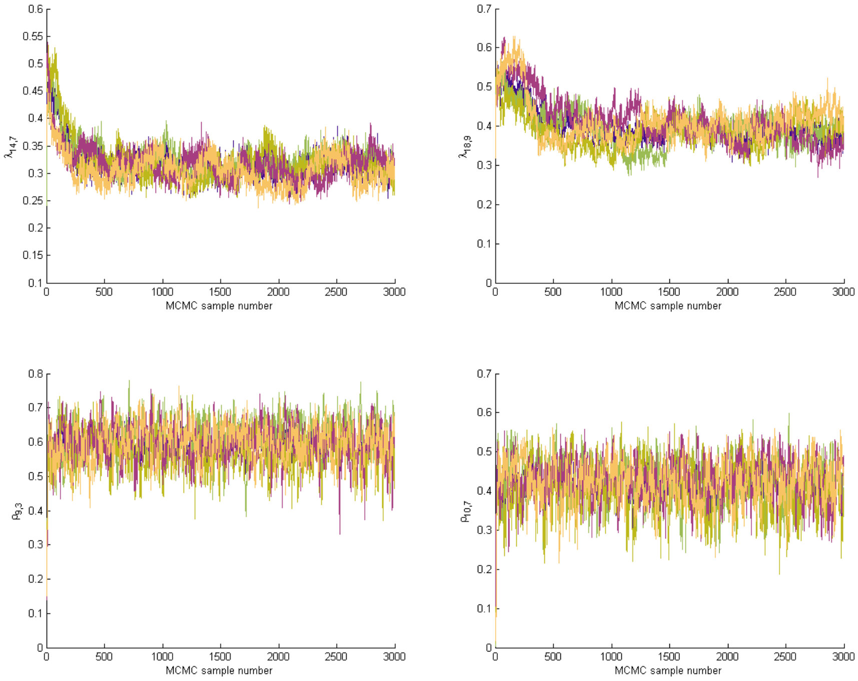


Figure 5.22: The trace plots of factor loading $\lambda_{14,7}$, $\lambda_{18,9}$, and latent correlations σ_{x93}, $\sigma_{x_{10,7}}$, from fitting model under Scheme 2.

Table 5.16: Comparisons of RMSEs for Scheme 1 against Scheme 2 based on 5 experiment results.

	mean diff.	p-value		mean diff.	p-value
y_1	-0.011	0.062	y_{11}	-0.004	0.125
y_2	-0.008	0.062	y_{12}	-0.002	0.250
y_3	-0.004	0.062	y_{13}	-0.006	0.062
y_4	-0.004	0.062	y_{14}	-0.005	0.062
y_5	-0.008	0.062	y_{15}	-0.005	0.062
y_6	-0.003	0.062	y_{16}	-0.005	0.062
y_7	-0.004	0.062	y_{17}	-0.011	0.062
y_8	-0.002	0.125	y_{18}	-0.028	0.062
y_9	-0.005	0.062	y_{19}	-0.004	0.062
y_{10}	-0.021	0.062	y_{20}	-0.003	0.062

similar estimation and prediction performance as fitting model using Scheme 1 does. If one aims to achieve more precise parameter estimation and lower predictive error, Scheme 2 provides reasonable initial values for model fitting using Scheme 1.

5.4 Remarks

The three empirical studies explore our modelling frameworks in aspects of estimation, prediction and computation. Following typical inspection and diagnosis, convergence can be assumed to be achieved. The transformation relations are able to be realized in the estimated latent variables and parameters between before and after data standardisation. The model structures with and without linking the latent errors (also refer to implement estimation for the joint model and for marginal models) reflect the effects on estimated latent variables and parameters.

Prediction results show increasing the number of pseudo inputs indeed reduces the predictive error. The predictive performance thereby could overtake those of using LS, or GPR on individual responses, especially when the regression relation between latent variables and covariates are highly non-linear. Using greedy selection for pseudo inputs empirically and significantly reduces predictive errors until the selected number increases to a certain value. Also, examining model appropriateness through the PPC procedure detects inappropriate factor structure.

The two estimation methods (the MCMC and hybrid approach) have differences in parameter estimation as well as in predictive error. Moreover, the magnitude of prediction difference turns smaller with pseudo input number. Despite small differences in estimation and prediction, using the two-step (hybrid) estimation procedure outperforms in computation, especially more responses and latent variables involved. Weighing the shortcomings and strengths, it can be a rather economical computational scheme by controlling iterations in each step if one is unintended to achieve superiority in estimation and prediction precision.

Chapter 6

Application On Longitudinal Studies

This chapter focuses on applications of sparse GP-SEM for longitudinal analysis. In Section 6.1 we focus on latent curve model (LCM) for development motivation and comparison reference of our framework. In Section 6.2 we briefly present longitudinal sparse GP-SEM with two kinds of response types – continuous and dichotomous. In Section 6.3, the relevant identification examination procedure is demonstrated. Section 6.4 briefly points out the schemes in estimation, prediction and computation. Section 6.5 examines the proposed longitudinal sparse GP-SEM on three data sets, two of which are synthetic and one real. The proceeding is the same as the last chapter: data summary, learning tasks (parameter estimation, growth curve of latent variables) and then prediction assessments. In the last section, we summarise the chapter.

6.1 Related Work

There are various methodologies for longitudinal empirical data, such as autoregressive models, repeated measures multivariate analysis of variance, generalized estimate equations and mixed effects model (Diggle et al. 2002, Skrondal & Rabe-Hesketh 2004).

Recently, Latent curve models (LCM) (or Latent growth models (LGM)) has received growing attention. It has developed for two decades but originates from a century's studying in individual and group difference (Bollen 2006). LCMs can be translated to a mixed effects model with a multilevel model structure, and its extensive applications exist in many fields of social and medical science, such as temporal relations between obsessive-compulsive cognition and disorder symptoms (Novara et al. 2011); the relationship between changes in socioeconomic status and changes in health (Hallerod & Gustafsson 2011); change in career satisfaction (Spurk et al. 2011) and in social and political attitude (Steele, 2008).

The simplest LCM can be represented as follows:

$$\mathbf{y}^{(n)} = \mathbf{\Lambda}_t \boldsymbol{\eta}^{(n)} + \boldsymbol{\epsilon}_y^{(n)}, \quad \boldsymbol{\eta}^{(n)} = \boldsymbol{\mu}_\eta + \boldsymbol{\epsilon}_\eta^{(n)}, \tag{6.1}$$

where $\mathbf{y}^{(n)} = (y^{(n1)}, \ldots, y^{(nT)})$ is the repeated measured metrical vector of the n-th individual, for $1 \leq n \leq N$. $\mathbf{\Lambda}_t$ is a time-dependent growth factor loading matrix and $\boldsymbol{\eta}^{(n)} = (\eta_1^{(n)}, \ldots, \eta_P^{(n)})$ is a latent growth factor vector for the n-th individual (P depends on which parametric functional form one uses). $\boldsymbol{\mu}_\eta$ is the expectation of the latent growth factor. For the case n, $\boldsymbol{\epsilon}_y^{(n)} = (\epsilon_y^{(n1)}, \ldots, \epsilon_y^{(nT)})$ and $\boldsymbol{\epsilon}_\zeta^{(n)} = (\epsilon_{\zeta_1}^{(n)}, \ldots, \epsilon_{\zeta_P}^{(n)})$ are the random disturbance deviations from $\mathbf{y}^{(n)}$ and $\boldsymbol{\mu}_\eta$ with means of zero and being uncorrelated with each other. Both are assumed uncorrelated for different cases and $\epsilon^{(nt)}$ further can be assumed optionally uncorrelated for different times. In terms of multilevel modelling (or hierarchical linear modelling), the first equation is a level-1 equation about the measures within the individual across time, and the second is a level-2 equation about the latent growth factors between the individuals. The two equations can further be combined into a linear mixed model. Although sharing the same specification framework with multilevel regression, LCM is more flexible on some features, such as the integration of the factorial structure of the repeated measured variable, extensions to larger structural models (Stoel et al. 2003).

Conceptually, the latent growth factor $\boldsymbol{\eta}^{(n)}$ specifies parametric relations between the observed variables at different time points. It enables the parameters to represent the functional form of the latent trajectory of the observed variable. For example, commonly one can use two parameters to represent a linear form - one for the intercept at the starting time, the other for the slope, showed respectively as η_1 and η_2 in Figure 6.1. The term $\boldsymbol{\mu}_\eta$ also summarises the starting intercepts and the rates of change across all cases in the group.

Through the corresponding factor loading matrix $\mathbf{\Lambda}_t$, latent growth factors $\boldsymbol{\eta}^{(n)}$ can exert the effects of time on the observed vector $\mathbf{y}^{(n)}$ at different time points, where time can be considered as a covariate. The loadings can be set by different metrics of time based on the associative factors. For a linear LCM in Figure 6.1, the loadings for the intercept factor can be set to 1 across all time points and those of the slope factor can be $t-1$ at the t-th wave.

Freeing factor loadings indicates that some loadings of the associated latent factor representing change (such as, the slope factor in linear LCMs) can be set to free parameters without parametric metrics of time. Noticeably, which factor loadings of the same latent factor are fixed leads to different interpretation of change of the trajectories.

LCM can also improve the accountability of the difference between patterns of the trajectories by including covariates. The predictors registered for individuals may be either constant over time (typically measured at the beginning of the study), or time-varying (taking on different values at each data collection time point). Figure 6.2 illustrates the time-invariant covariate $\mathbf{z}^{(n)}$ affects the latent growth factors η_1 (intercept) and η_2 (slope). For the time-varying covariates, their direct effects are straight on observed variables and they are introduced into the level-1 equations.

To investigate change across time in a latent variable (or construct) of interest, one of the state-of-art approach is latent variable LCM (LV-LVM). It is sometimes referred to as multiple-indicator growth curve models, curve-of-factors models, and second order latent growth curve model or latent variable longitudinal curve model. The simplest LV-LCM can be represented:

$$\mathbf{y}^{(nt)} = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda} x^{(nt)} + \boldsymbol{\epsilon}_y^{(nt)}, \quad \mathbf{x}^{(n)} = \boldsymbol{\Lambda}_t \boldsymbol{\eta}^{(n)} + \boldsymbol{\epsilon}_{\boldsymbol{\eta}}^{(n)}, \quad \boldsymbol{\eta}^{(n)} = \boldsymbol{\mu}_\eta + \boldsymbol{\epsilon}_\eta^{(n)}, \tag{6.2}$$

where the indicators of n-th respondent at Time t, $\mathbf{y}^{(nt)} = (y_1^{(nt)}, \ldots, y_R^{(nt)})^{\mathsf{T}}$, its measured error $\boldsymbol{\epsilon}_y^{(nt)} = (\epsilon_{y_1}^{(nt)}, \ldots, \epsilon_{y_R}^{(nt)})^{\mathsf{T}}$, the repeated measure latent variables $\mathbf{x}^{(n)} = (x^{(n1)}, \ldots, x^{(nT)})^{\mathsf{T}}$; and $\boldsymbol{\lambda}_0$ and $\boldsymbol{\Lambda}$ are a $R \times 1$ intercept vector and a $R \times 1$ factor loading matrix as in conventional factor analysis. The growth factor loading matrix $\boldsymbol{\Lambda}_t$, the growth latent factor $\boldsymbol{\eta}^{(n)}$, the random disturbances $\boldsymbol{\epsilon}_x^{(n)}$ and $\boldsymbol{\epsilon}_\eta^{(n)}$ have the same representation as in previous LCM equations. Non-correlation can additionally be assumed between latent variables and random disturbances, among random disturbances and between random errors of different cases (or optionally between different time points). LV-LCM is basically a SEM because one equation modelling growth of a latent variable and the other modelling measuring it, feature a structural model and measurement model.

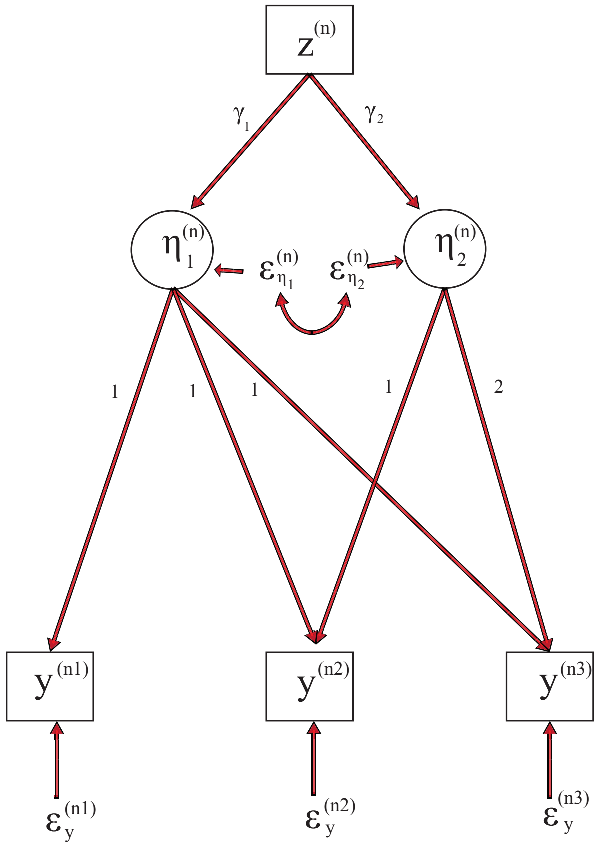


Figure 6.1: The path diagram of linear latent curve model with a time-invariant covariate and observed responses at three time points.

Figure 6.2 shows a simple case of LV-LVM with two indicators for a latent variable and a time-invariant covariate and responses across three time points.

The factor matrix $\mathbf{\Lambda}$ and the intercept $\boldsymbol{\lambda}_0$ can be set to be free estimated or time-dependent. This means the measurement (factorial) invariance assumption can be relaxed[1]. Although latent variables at different times can still be realised, it may cause an issue about interpretation for the growth (Ferrer et al. 2008). The reason is that varied factor loadings can confound temporal change of a latent construct.

Considering the growths of multiple latent variables ($Q > 1$), the structure of LV-LCM turns complicated. Rather than using the same growth latent factors, the framework is built on sets of indicators (each of which measures a latent variable) and modelling inter-correlation of growth latent factors (belonging to different LV-LCMs). This extension is referred to multivariate LCM (MacCallum et al. 1997).

In addition, to relax the form of a growth function can be one of research directions for LCMs. So far, very limited works have been done. To the best of our knowledge, there is only one paper close to LV-LCM or our modelling frameworks. Gaussian-Process factor analysis (GPFA) model (Yu et al. 2009) could be categorized as a variant of LV-LCM. Except utilizing a factor analysis formulation, it incorporates a GP framework to model regression relationship between the time covariate and latent variables (underlying neural states). It serves as an approach to reduce dimension of the recorded activity data (from large populations of neurons) to realize neural trajectories.

6.2 Model Specification

Sparse GP-SEM can model temporal tendency of multiple latent variables for a panel dataset with several indicators at each time point. The applications, for example, could include exploring the learning curves of school pupils on quantitative ability and language proficiency given various teaching methods and materials, based on an education cohort study; investigating into the temporal devoplement on morale and job satisfication for a large sample of employees given different leadership or management styles, upon internal questionnaire results.

[1] There are several types of measurement invariance applied on different components in the model (see (Ferrer et al. 2008, Cheung & Rensvold 2002, Meredith 1993)). Here we only consider the type about factor loadings and intercepts, which is called strong factorical invariance.

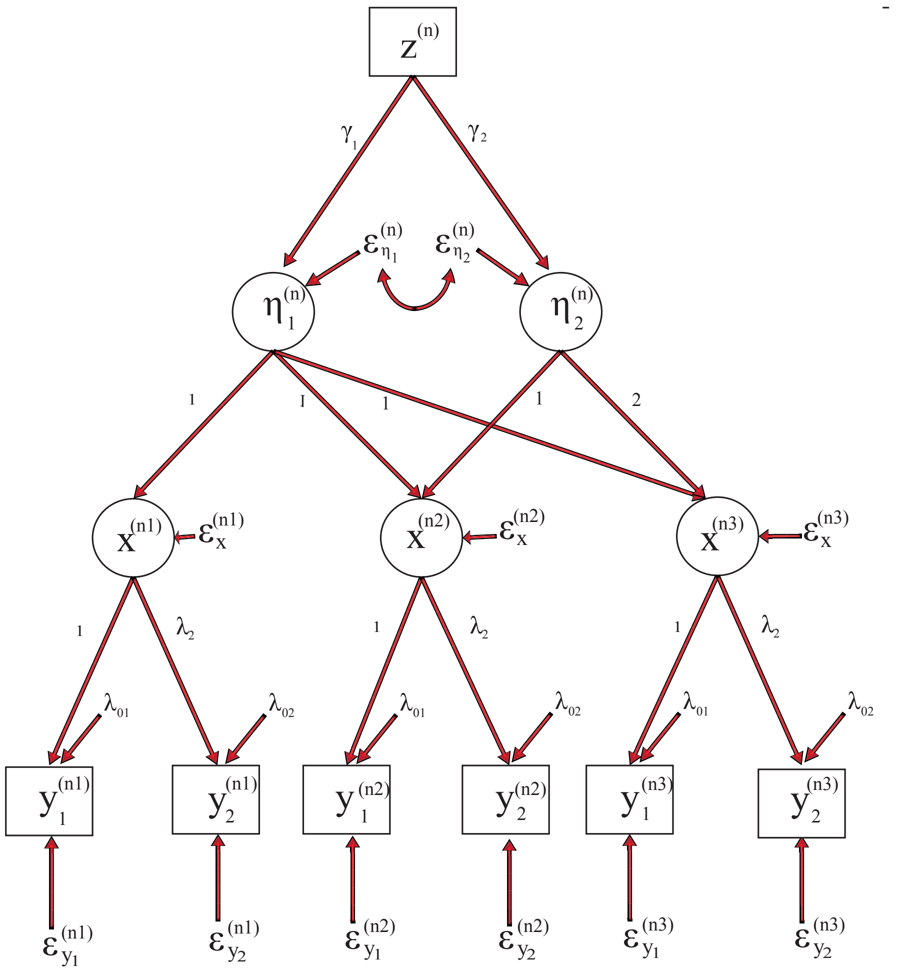


Figure 6.2: The path diagram of linear latent curve model with multiple indicators for two waves and a time-invariant covariate.

For clarity, we only consider the case of one latent variable, that is $Q = 1$. The latent variable of interest is still regarded as being continuous.

Continuous Responses

The model specification of sparse GP-SEM for longitudinal application is slightly different from the original (presented in Equations (3.1)-(3.7)). The number of the n-th subject's covariate vector is not single but multiple, identical to time point size. Each covariate vector is specified by adding a time scaling in superscript, that is, $\mathbf{z}^{(nt)}$, where t indexes time point. One of covariate dimensions is additionally registered for the time scaling as well. The rest of variables have the same notation difference. For instance, at the t-th time point, the n-th subject's the latent variable is $x^{(nt)}$, its GP latent function value $f^{(nt)}$ [2] and the r-th response continuous variable $y_r^{(nt)}$. There is another difference in modelling latent errors. We later point it out in the equations for description of model extension.

Two possible modelling notions can be adopted as well. One is from matrix-variate Gaussian models (Stegle et al. 2011) and is to build a composite GP covariance function for modelling regression relations between inputs and latent variables. That function is constructed by taking a Kronecker product for two covariance functions modelling dependence between latent function values at the inter-time and intra-time inputs, respectively. The other notion merely follows another GP regression, where only the time covariate is used (Rasmussen & Williams 2006). It is additionally to model association of latent GP errors at different time points. Due to the extra computational cost for estimating the GP hyper-parameters, the above two notions are not adopted here.

Sparse GP-SEM for longitudinal application has three differences from the GPFA model (Yu et al. 2009) mentioned before. First, GPFA only uses one input (representing time). Second, it models one subject's neural trajectories over T time points for several trials (the number of trials refer to the number of data points in our framework), but all trial results are treated as being generated independently. Third, it does not use sparse GP approximation methods to speed up computation.

[2] Compared with the notation in Equation (3.1), here because only considering the case of one latent variable, we remove the subscript of q. The other associated variables and matrices follows this fashion as well.

Binary responses

A simple model extention we can consider is to handle dichotomous outcome variables. The augmented model framework is adding underlying continuous responses $u_r^{(nt)}$ into the model structure – one of classical treatments (Albert & Chib 1993, Skrondal & Rabe-Hesketh 2004, Bartholomew et al. 2008, Chib & Greenberg 1998). Moreover, the addition is to link observed responses and to bear the direct effects of the higher-level latent variables (or latnet covariate). This postulates observed binary variables are generated by underlying normally-distributed latent variables with the connection that

$$y_r^{(nt)} = \begin{cases} 1, & \text{if } u_r^{(nt)} > 0 \ , \\ 0, & \text{otherwise.} \end{cases}$$

To clarify the model specification with binary responses, we present the equational description here. For $1 \leq n \leq N$, $1 \leq t \leq T$, the first three equations show mathematical formulation of regression relationship between covariates and a latent variable under GP framework per time point,

$$\begin{aligned} x^{(nt)} &= f^{(nt)} + \epsilon_x^{(nt)}, & (6.3) \\ \boldsymbol{\epsilon}_x^{(n)} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t) & (6.4) \\ \mathbf{f}^{(t)}|\mathbf{z}^{1:N,(t)} &\sim \mathcal{N}(0, \mathbf{K}_N^{(t)}), & (6.5) \end{aligned}$$

where $\boldsymbol{\Sigma}_t$ is the covariance matrix of the latent errors between time points, denoted as $\boldsymbol{\epsilon}_x^{(n)} = [\epsilon_x^{(n1)}, \ldots, \epsilon_x^{(nT)}]$. And there is an implicit and slight difference with the original sparse GP-SEM framework, which considers the covariance matrix $\boldsymbol{\Sigma}_x$ of the latent errors between latent-variable indexes. More specifically, the difference results from that only one latent variable is involved, and $\boldsymbol{\Sigma}_t$ indeed works as does $\boldsymbol{\Sigma}_x$. However, when multiple latent variables are considered, the difference turns evident because the resulting covariance matrix of latent errors is a Kronecker product of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_t$.

The measurement model has a modification on the original one (represented by (3.4) and (3.5)). The notation is changed from (n) to (nt), and the response variable vector $\mathbf{y}^{(n)}$ is replaced by the underlying latent variable vector $\mathbf{u}^{(nt)}$, denoted by $\mathbf{u}^{(nt)} = [u_1^{(nt)}, \ldots, u_R^{(nt)}]$. The model equation is

$$\begin{aligned} \mathbf{u}^{(nt)} &= \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda} x^{(nt)} + \boldsymbol{\epsilon}_u^{(nt)}, & (6.6) \\ \boldsymbol{\epsilon}_u^{(nt)} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u^{(t)}), & (6.7) \end{aligned}$$

where a measurement error vector is $\boldsymbol{\epsilon}_u^{(nt)} = [\epsilon_{u_1}^{(nt)}, \dots, \epsilon_{u_R}^{(nt)}]$ and $\boldsymbol{\Sigma}_u^{(t)}$ is its covariance matrix. All underlying responses follow a multivariate normal distribution because of Gaussianity of measurement errors $\boldsymbol{\epsilon}_u^{(nt)}$ and latent variables $x^{(nt)}$.

The conditional GP prior about sparse approximation

$$\begin{aligned} \mathbf{f}^{(t)} | \bar{\mathbf{f}}^{(t)}, \mathbf{z}^{1:N,(t)}, \bar{\mathbf{z}}^{1:M,(t)} &\sim \mathcal{N}(\mathbf{K}_{NM}^{(t)} (\mathbf{K}_M^{(t)})^{-1} \bar{\mathbf{f}}^{(t)}, \mathbf{V}^{(t)}), & (6.8) \\ \bar{\mathbf{f}}^{(t)} | \bar{\mathbf{z}}^{1:M,(t)} &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_M^{(t)}). & (6.9) \end{aligned}$$

The model assumptions are similar to the original ones. For simplicity, we introduce a notation II as being independent. Thus all the assumptions can be concisely written as $f^{(nt)} \perp \epsilon_x^{(nt)}$; $f^{(nt)} \perp f^{(nt^{'})}$; $\epsilon_x^{(nt)} \perp \epsilon_x^{(n^{'}t)}$; $\epsilon_{u_r}^{(nt)} \perp \epsilon_{u_{r^{'}}}^{(nt)}$; $\epsilon_{u_r}^{(nt)} \perp \epsilon_{u_r}^{(n^{'}t)}$; $\epsilon_{u_r}^{(nt)} \perp \epsilon_{u_r}^{(n^{'}t)}$; $x^{(nt)} \perp \epsilon_{u_r}^{(nt)}$, for any n, t, r and $n \neq n^{'}$, $t \neq t^{'}$, $r \neq r^{'}$.

We also implicitly assume the intercept terms and factor loadings are time-invariant. In other words, $\boldsymbol{\lambda}_0^{(t)} = \boldsymbol{\lambda}_0$ and $\boldsymbol{\Lambda}^{(t)} = \boldsymbol{\Lambda}$, where $\boldsymbol{\lambda}_0^{(t)}$ and $\boldsymbol{\Lambda}^{(t)}$ are the time-variant, intercept and factor loadings. Note that these two assumptions related to measurement invariance allow us to interpret the temporal-change of the latent variable of interest.

For satisfying identification condition, we further make some constraints on model paramters. We specify them in the next section.

The graphical representation of sparse GP-SEM with temporal dichotomous response variables is shown in Figure 6.3. Compared with Figure 3.2, the differences are evident. Beside slight changes in notations and covariates, the whole model structure merely has one more level, encoded between the higher-level latent variables and the lowest-level observed variables. Furthermore, more differences can be found between Figure 6.2 and figure 6.3. They are: 1. the GP latent errors $\boldsymbol{\epsilon}_x$ are correlated; 2. intercepts and loadings are imposed with different constraints; 3. the upper part of model structure (between covariates and a latent construct) are using different notions to model the regression relation.

6.3 Examination of identification

Given the modeling formulation, one still needs to consider whether the model is identifiable before estimation. The identification examination for longitudinal continuous responses is similar to its static counterpart, as shown in Section 3.3. Functional relations with unknown and known parameters can be obtained through algebraic derivation.

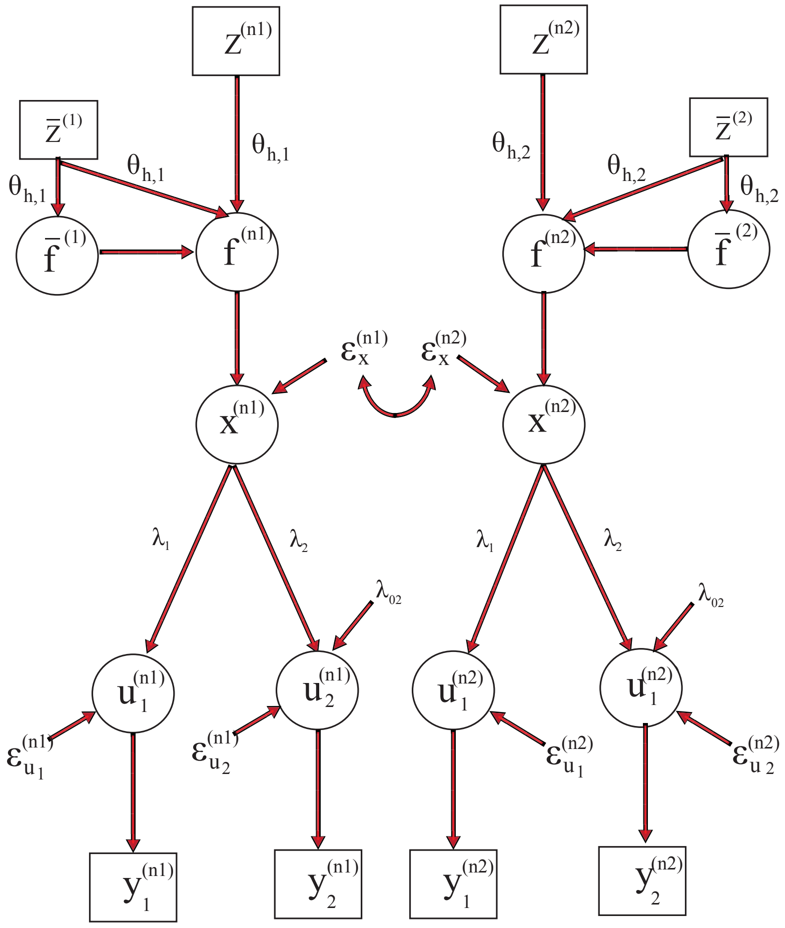


Figure 6.3: The path diagram of sparse GP-SEM with four dichotomous responses for two waves.

For the case of longitudinal (or static) binary responses, the model identification checking can be implemented by the second approach mentioned in Section 3.3. The method is based on the Jacobin matrix of reduced-form parameters over all unknown parameters. If the rank of the Jacobin matrix is the same as the number of unknown parameters, then local identification of model can be ensured. We exemplify the case of two time points and two binary variables per time for demonstrating the identification examination.

For $1 \leq t \leq 2$, without specifying data point the model structure can be expressed as follow:

$$\begin{aligned} x^{(t)} &= f(\mathbf{z}^{(t)}) + \boldsymbol{\epsilon}_x^{(t)} && (6.10)\\ \mathbf{u}^{(t)} &= \boldsymbol{\Lambda} x^{(t)} + \boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}_u^{(t)}, && (6.11)\end{aligned}$$

where the latent response vector $\mathbf{u}^{(t)} = [u_1^{(t)}, u_2^{(t)}]$ is related to the observed binary response variables by the indicator function: $y_r^{(t)} = 1$ if $u_r^{(t)} > 0$; and 0 otherwise, for $1 \leq r \leq 2$.

We set some constraints on parameters to ensure identification condition holds. Like the restrictions introduced in Section 3.3, the variances of $\boldsymbol{\epsilon}_x^{(t)}$ are ones, and the elements of intercept terms $\boldsymbol{\lambda}_0$ corresponding to anchors are zeros. In addition, the measurement-error variances are all set to ones. The necessity of the variance constraints is discussed later.

6 unknown parameters are therefore realised, which are $\{\lambda_{02}, \lambda_1, \lambda_2, f(\mathbf{z}^{(1)}), f(\mathbf{z}^{(2)}), \sigma_{t_{12}}\}$, where $\sigma_{t_{12}}$ is the cross-covariance of $\boldsymbol{\Sigma}_t$. Due to the measurement-invariance assumption, there is no need to estimate the intercept term and factor loadings at the 2-nd time points.

The marginal probability of the reduced-form distribution $p(y_r^{(t)}|\mathbf{z}^{(t)})$ becomes

$$\begin{aligned} p(y_r^{(t)} = 1|\mathbf{z}^{(t)}) &= \frac{1}{\sqrt{\lambda_r^2+1}} \int_0^\infty \phi\left(\frac{\lambda_{0r} + \lambda_r f(\mathbf{z}^{(t)}) + \xi_r}{\sqrt{\lambda_r^2+1}}\right) d\xi_r \\ &= \Phi\left(\frac{\lambda_{0r} + \lambda_r f(\mathbf{z}^{(t)})}{\sqrt{\lambda_r^2+1}}\right), \end{aligned}$$

where λ_{0r} and λ_r are, respectively, the r-th elements of intercept $\boldsymbol{\lambda}_0$ and factor loadings $\boldsymbol{\Lambda}$. Because of the constraints, λ_{01} is 0, where we assume the first response at each time point is an anchor. $\phi(\cdot)$ is the standard normal density, $\xi_r = \lambda_r \epsilon^{(t)}$, and Φ is the cdf of standard normal distribution. Hence, the mean $m_r^{(t)}$ of the underlying latent variable $u_r^{(t)}$ is

$$m_r^{(t)} = \frac{\lambda_{0r} + \lambda_r f(\mathbf{z}^{(t)})}{\sqrt{\lambda_r^2+1}}, \qquad (6.12)$$

and this is identified from the the marginal probability above.

Given the response vector $\mathbf{y} = [y_1^{(1)}, y_2^{(1)}, y_1^{(2)}, y_2^{(2)}]$ and a binary vector $\mathbf{b}$, the joint response probabilities of the reduced-form distribution is $p(\mathbf{y} = \mathbf{b}|\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$, determined by a multivariate four-dimensional Gaussian density with means from (6.12) and a correlation matrix $\mathbf{R} = diag(\mathbf{\Omega})^{-1/2} \cdot \mathbf{\Omega} \cdot diag(\mathbf{\Omega})^{-1/2}$, where $\mathbf{\Omega}$ is the covariance matrix of all the underlying latent variables. The resulting 8 reduced-form parameters contain

$$m_1^{(1)},\ m_2^{(1)},\ m_1^{(2)},\ m_2^{(2)},\ \frac{\lambda_1\lambda_2}{\sqrt{\lambda_1^2+1}\sqrt{\lambda_2^2+1}},\ \frac{\lambda_1\lambda_2\sigma_{t_{12}}}{\sqrt{\lambda_1^2+1}\sqrt{\lambda_2^2+1}},\ \frac{\lambda_1^2\sigma_{t_{12}}}{\lambda_1^2+1},\ \frac{\lambda_2^2\sigma_{t_{12}}}{\lambda_2^2+1},$$

where the last 4 parameters are from the unrepeated elements of $\mathbf{R}$.

Now the identification condition can be examined by checking full rank of the Jacobian matrix, built on all derivatives of reduced-form parameters over all unknown parameters. The matrix size is 8×6 and the rank is 6. Therefore, that ensures local identifiability (Skrondal & Rabe-Hesketh 2004). One should note that if no constraints on measurement-error variances are imposed, the number of the unknown parameters increases to 10 and thereby the local identifiability would not be guaranteed.

Incidentally, while measurement-invariance assumptions are relaxed, the identification condition of the example remains. The reason is that the associated Jacobian matrix with size of 10×9 has the rank of 9. Specifically, 2 more reduced-form parameters are the unrepeated entries of $\mathbf{R}$; 3 more unknown parameters are due to the unconstrained loadings and intercept.

In general cases, the identification condition is still ensured under the imposed constraints. It is because given $R \geq 2$ and $T \geq 2$, the number of the reduced-form parameters, $RT + R(R-1)/2 + RT(R+1)(T-1)/4$, is always greater than that of the unknown parameters, $(R-1)+R+T+T(T-1)/2$; and the rank of the associated Jacobian matrix can be checked as the latter number. Note that the former is comprised of the RT marginal means, and the rest from the unrepeated correlations of $\mathbf{R}$, where $R(R-1)/2$ are counted from correlations not involoving coefficients of $\mathbf{\Sigma}_t$ but $RT(R+1)(T-1)/4$; the latter is from the intercepts, factor loadings, latent function values and correlations of latent errors.

6.4 Computational Implementations

After model identifiability is ensured, we can consider how to implement computation for different tasks.

For Estimation

The estimation algorithm is rather similar to that in Section 4.3. We indicate the differences here.

For temporal continuous variables, if there are no measurement invariance assumptions imposed, the samplers remain identical except the minor notation differences and the change of latent-error covarince matrix, from $\mathbf{\Sigma}_x$ to $\mathbf{\Sigma}_t$. Imposing the assumptions indeed differentiates the samplers for intercepts and factor loadings despite the Gaussianities being reserved. All prior distributions are still adopted as before in Section 4.2 and 4.3.

The sampler of the r-th intercept and factor loading $(\lambda_{0r}, \lambda_r)$ is distributed normally with a covariance matrix

$$\mathbf{\Sigma}_{\lambda_r,post} \equiv (\frac{1}{\sigma^2_\lambda}\mathbf{I}_{|\mathcal{P}_r|} + \sum_{t=1}^{T} \frac{1}{\sigma^2_{y_r^{(t)}}}[(\widetilde{\mathbf{X}}^{(t)})^\intercal \widetilde{\mathbf{X}}^{(t)}]_{\mathcal{P}_r,\mathcal{P}_r})^{-1}, \tag{6.13}$$

where $\widetilde{\mathbf{X}}^{(t)} \equiv [(\mathbf{x}^{(t)})^\intercal, \mathbf{1}_N]^\intercal$, $\mathbf{x}^{(t)}$ is a row vector consisting of all scores of N subjects' latent variables at the t-th wave; the rest notations are the same as those in Equations (4.18) and (4.19), and a mean

$$\mathbf{\Sigma}_{\lambda_r,post} \cdot \sum_{t=1}^{T} [(\widetilde{\mathbf{X}}^{(t)})^\intercal]_{\mathcal{P}_r,\cdot} \mathbf{y}_r^{(t)}, \tag{6.14}$$

where $\mathbf{y}_r^{(t)}$ is a $N \times 1$ column vector with the r-th indicator of all the data points at the t-th time point.

One point should be reminded that (6.13) and (6.14) specify the full conditional distribution of $(\lambda_{0r}, \lambda_r)$, and the parameter expansion (PE) technique is not adopted there. If using PE to increase MCMC mixing efficiency, one needs to change the augmentated latent variable $\widetilde{\mathbf{X}}^{(t)}$ to the transformed one $\widetilde{\mathbf{W}}^{(t)}$ by a working parameter α, which is mentioned in Section 4.3. It follows that the resulting distribution can generate a sample of the transformed intercepts and factor loadings. Then Equations (6.13) and (6.14) have an alternative and equivalent expression which replaces the intercept, factor loading and latent variable by the transformed counterparts. Furthermore, the sample of the untransformed parameters can be obtained through α.

The model estimation for longitudinal binary responses incorporates the sampling scheme for the underlying response variables. The joint sampling scheme of $\mathbf{u}^{(nt)}$ can break into R samplers because its components are conditional independent given the latent variable $x^{(nt)}$. Following a classic Bayesian treatment for analysing binary and polychotomous response data (Albert & Chib 1993), the sampler of $u_r^{(nt)}$ is distributed truncated-normally as

$$\begin{cases} I_{[0,\infty]} \times \mathcal{N}(\lambda_{0r} + \lambda_r x^{(nt)}, 1), & \text{if } y_r^{(nt)} = 1 \\ I_{[-\infty,0]} \times \mathcal{N}(\lambda_{0r} + \lambda_r x^{(nt)}, 1), & \text{if } y_r^{(nt)} = 0 \ , \end{cases} \tag{6.15}$$

where $I_{[0,\infty]}$ is an indicator function over the interval $[0, \infty]$. This sampling scheme is also applied in multivarite probit models (Chib & Greenberg 1998).

For Prediction

Regarding temporal continuous responses, the calculation of predictive responses given a new covariate vector is still the same as that presented in Section 4.7. By contrast, the predictive longitudinal binary responses are decided by the predictive underlying responses, whose calculation is identical to continuous response. To be more specific,

$$\begin{cases} y_{r,new}^{(nt)} = 1, & \text{if } u_{r,new}^{(nt)} > 0 \\ y_{r,new}^{(nt)} = 0, & \text{if } u_{r,new}^{(nt)} < 0. \end{cases} \tag{6.16}$$

For Computational Schemes

Three computational schemes are adopted to investigate the influence of the measurement-invariance assumption on estimation and prediction. Scheme 1 relaxes restriction on loadings. Therefore, the related samplers are exactly the ones given by Equations (4.18) and (4.19) under the joint model. In contrast, Scheme 2 sets constraints on loadings. The samplers are used given by (6.13) and (6.14) under the joint model. Its sampling could be more efficient than Scheme 1. Scheme 3 is similar to the economical strategy examined in Section 5-3. Initially it is to implement estimation under marginal models and then to average the estimated loadings over time. It follows fitting the joint model by fixing all parameters obtained from the preceding estimations (including the averaged loading). This can boost more computational efficiency than the first two schemes if one manages the sample sizes on fitting the marginal and joint models.

6.5 Experiments

In this section, we pay more attentions to learning and comparing the temporal tendency of a posterior estimated latent variable, under three schemes. We also briefly report the prominent feature in parameter estimation and prediction. Incidentally, random selection for pseudo inputs is adopted in all the experiments.

6.5.1 Study IV - Synthetic Data

The properties of the first multiple-output regression longitudinal data set with metrical responses are summarised in Table 6.1.

Table 6.1: Properties of Dataset IV. Here p.t.p. is the abbreviation of "per time point".

Dataset		Input		Output	
Size (N)	Time (T)	Dim. p.t.p. (D)	character	Dim. p.t.p. (R)	character
2000	4	11	continuous	3	continuous

We use a function form $f^{(t)}(\mathbf{z}) = c\big[(z_1^{(t)} - 1)^2 + \sum_{l=2}^{10}(z_l^{(t)})^2\big]$ to generate the latent function values at the t-th time point, where $z_1^{(t)} = t$ and c is a positive contanst. Then through intercepts and loadings (fixed across time points), we produce all the responses.

The histograms for 44 covariates over time points would be not provided here for saving space. However, they (except the time covariates) indeed appear Gaussian densities due to normally random generation. As for all response variables, the histograms are presented in Figure 6.3. As seen, all distributional shapes seem roughly symmetric although some have a long tail. In general, the histograms reflect the normal-distribution data generation machenism.

In Table 6.2, there is an apparent classification in the correlation coefficients of all response variables. We can classify every three variables in order as one group, and thus obtain 4 groups in total, which corresponds to the time point number. The variables in a group have very strong correlation but low with a variable in other groups - further the decrement of correlation turn large as time proceeds.

Learning

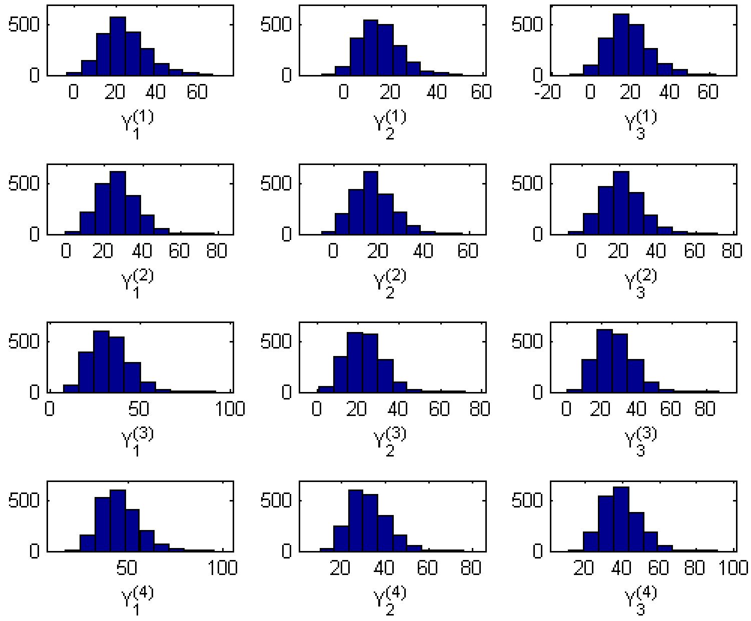


Figure 6.4: Distributions of the outputs for Dataset IV.

Table 6.2: Correlation coefficients between response variables of Dataset IV.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$	1	0.96	0.97	0.29	0.28	0.29	0.19	0.18	0.19	0.05	0.06	0.05
$y_2^{(1)}$		1	0.96	0.28	0.28	0.28	0.19	0.18	0.18	0.06	0.07	0.06
$y_3^{(1)}$			1	0.28	0.27	0.28	0.18	0.17	0.18	0.06	0.07	0.06
$y_1^{(2)}$				1	0.97	0.97	0.26	0.26	0.26	0.16	0.15	0.16
$y_2^{(2)}$					1	0.97	0.27	0.27	0.26	0.17	0.17	0.16
$y_3^{(2)}$						1	0.27	0.26	0.26	0.17	0.16	0.16
$y_1^{(3)}$							1	0.97	0.97	0.28	0.28	0.28
$y_2^{(3)}$								1	0.97	0.28	0.28	0.27
$y_3^{(3)}$									1	0.28	0.28	0.28
$y_1^{(4)}$										1	0.97	0.97
$y_2^{(4)}$											1	0.97
$y_3^{(4)}$												1

The set-ups of simulations, including the pseudo-input number and the MCMC sample size, are identical to those used in the learning task in Section 5.1.2. For Scheme 3, 5000 iterations are for fitting marginal models and 3000 for the joint model.

5-chain simulations are conducted. All EPSR values for estimated parameters are under 1.05. One could consider simulation convergence is achieved.

We only select the parameters related to the first time point and present the estimation results in Table 6.3. Note that the estimate of the loadings and intercepts for Scheme 3 are already averaged over time. The sample standard deviations after burn-in are only reported based on the estimates from fitting the first marginal model. In addition, the estimates for Scheme 1 are similar over time. There are only small differences between the estimates at the first time point and those at other times. For example, the maximal difference happens in the intercepts by the magnitude from 0.1 to 0.3.

Moreover, the estimates for all the schemes are overall similar to some extent. It is noted that the true values are in the 95% credible intervals, consisting of the values in the range centred at the estimated means with 2 standard derivations as radiuses. Comparing the results on loadings and intercepts for Scheme 2 with Scheme 3, the small mean differences can be due to the bias of the two-step estimation procedure. The estimates from the former appear closer to the true values.

Figure 6.5 shows the trend of the posterior means of the latent variable from model fitting using the three schemes. Each coloured trajectory depicts the temporal change of the ergodic-average estimated latent variable for a different data point. Most lie so densely that they consist of a bundle of lines. This suggests the estimates at each time point has a unimodal distribution. The green-square-red-dashed line represents the population trend pattern for all data points. The value at each time point is simply the mean of all estimated latent variable. As seen, all the population trend lines show a non-linear trend, which reveals quadratic growing.

Comparing the discrepancy with the true trend line, the mean absolute errors are reported in Table 6.4. The error magnitudes for all the schemes are rather small compared with the variance of the estimated latent variable, around 6.5. The mean absolute error for Scheme 3 is overall the largest and that for Scheme 1 seem the smallest.

Note that the general trend lines are obtained through calculating MCMC samples a posteriori. This is not like Multilevel-SEM or LV-LCM. The difference is that a set of

Table 6.3: Comparisons between true values and the 5-simulation-chain averages under three schemes.

	True	Scheme 1		Scheme 2		Scheme 3	
parameter		mean	sd	mean	sd	mean	sd
$\theta_{h,11}$	none	2.27	0.03	2.25	0.04	1.94	0.05
$\theta_{h,12}$	none	2.87	0.07	2.83	0.08	2.45	0.08
$\sigma_{t_{12}}$	0.85	0.83	0.02	0.83	0.02	0.82	0.02
$\sigma_{t_{13}}$	0.55	0.58	0.03	0.57	0.03	0.62	0.03
$\sigma_{t_{14}}$	0.30	0.32	0.03	0.32	0.04	0.37	0.04
λ_1	4.39	4.36	0.09	4.38	0.07	4.23	0.14
λ_2	3.58	3.62	0.08	3.60	0.06	3.48	0.12
λ_3	4.25	4.27	0.09	4.29	0.07	4.15	0.14
λ_{02}	-1.27	-1.34	0.16	-1.19	0.08	-1.13	0.16
λ_{03}	-1.04	-0.94	0.17	-1.01	0.09	-1.01	0.17
$\sigma^2_{y_1}$	3.00	3.05	0.22	3.04	0.23	3.20	0.24
$\sigma^2_{y_2}$	3.00	3.02	0.18	3.04	0.18	3.08	0.19
$\sigma^2_{y_3}$	3.00	2.95	0.20	2.96	0.21	3.07	0.23

specified growth factors is embedded into those modelling frameworks. The factors control the functional family of a growth line, and their means can feature an overall trend change over all data points.

Table 6.4: Mean absolute errors between the estimated trend and the true one of the latent variable under the three schemes.

	Time point t			
	1	2	3	4
Scheme 1	0.216	0.197	0.183	0.507
Scheme 2	0.204	0.229	0.315	0.413
Scheme 3	0.351	0.410	0.432	0.550

Prediction

The predictive performance of longitudinal sparse GP-SEM is assessed based on the 10 experiment results (upon 10 sets of even partitioned training-test dataset and 100 pseudo

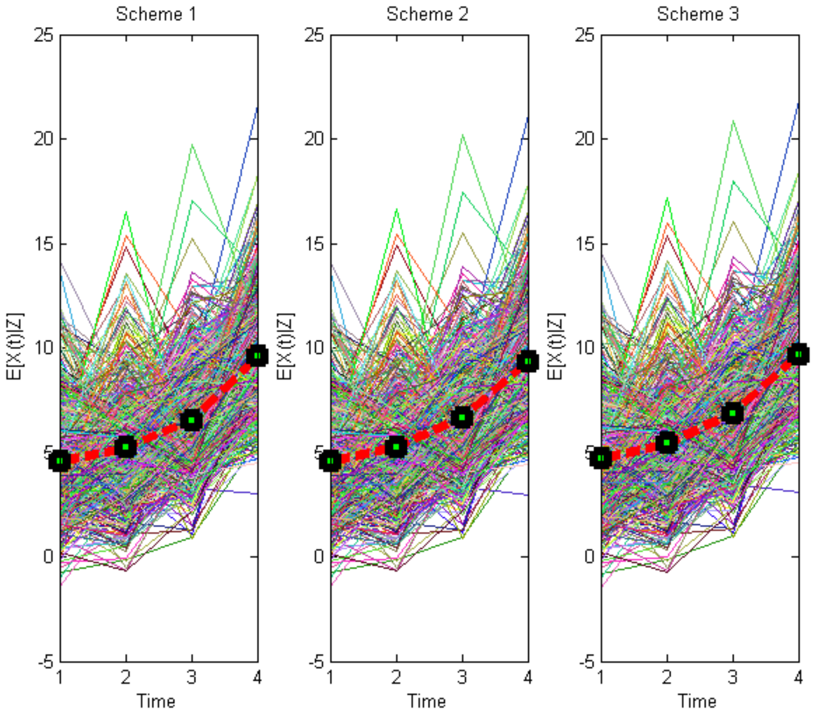


Figure 6.5: The trend plots of the posterior means of the latent variable under three schemes for Dataset IV. Each colour represents a different data point. The green-square-red-dashed line depicts the population trend pattern for all data points. Its value at each time point is simply the mean of all estimated latent variable.

inputs). Overall, the mean RMSEs for any schemes are much smaller (twice smaller) than those for using LS method (the values are around 11). Furthermore, the mean RMSEs for Scheme 1 are rather similar to those for Scheme 2. Both are also statistically significantly smaller than those for Scheme 3 by a difference magnitude from 0.12 to 0.2 over all responses.

Considering model checking, posterior predictive (PP) p-values can be calculated. Using the χ^2 test statistics in last chapter, the three PP p-values of the test statistics T_1 (involving all response variables) are 0.902, 0.891 and 0.925, corresponding to Scheme 1, 2 and 3, respectively. All are within a safe 0.05-0.95 range. The same scenario also happens in the p-values of another χ^2 test statistics (only involving the responses at the same time point). All indicate adequateness for the model structure and measurement invariance assumption.

6.5.2 Study V - Synthetic Data

The properties of the second longitudinal data set with dichotomous responses are summarised in Table 6.5.

Table 6.5: Properties of Dataset V. Here p.t.p. is the abbreviation of "per time point".

Dataset		Input		Output	
Size (N)	Time (T)	Dim. p.t.p. (D)	character	Dim. p.t.p. (R)	character
2000	4	11	continuous	3	binary

We generate the t-th latent function vector at through the function form $f^{(t)}(\mathbf{z}) = -c\big[z_1^{(t)} + \sum_{l=2}^{10}(z_l^{(t)})^2\big]$ to generate the latent function values at the t-th time point, where $z_1^{(t)} = t$ and c is a positive contanst. Then through intercepts and loadings (fixed across time points), we produce all the responses.

The 44 covariates over time points (except the time covariates) appear to have Gaussian densities. The proportions of all response variables are presented in Table 6.6. It shows the proportions for 0 decrease across time but vice versa for 1.

Table 6.6: Proportions (%) of response binary variables of Dataset V.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
0	84.2	84.9	90.1	78.9	73.3	80.0	58.7	58.0	67.6	41.2	42.7	49.4
1	15.8	15.1	9.9	26.1	26.7	20.0	41.3	42.0	32.4	58.8	58.3	50.6

Table 6.7 shows a similar scenario as Table 6.2 and 4 groups (consisting of 3 variables in order) can be identified. Group members have moderate or strong inter-correlations within the group, but the between-group correlations turn lower with time.

Learning

There is a difference in the set-ups of simulations -10000 MCMC sampling iterations for Scheme 1 and 2, but for Scheme 3, 10000 iterations for fitting marginal models, 3000 for the joint model.

Because all EPSR values for estimated parameters from 5-chain simulations are under 1.1, one could regard simulation convergence is achieved. The estimation results of

Table 6.7: Correlation coefficients between response variables of Dataset V.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$	1	0.61	0.66	0.20	0.20	0.19	0.13	0.14	0.14	0.02	0.04	0.04
$y_2^{(1)}$		1	0.59	0.21	0.20	0.21	0.12	0.12	0.14	0.01	0.03	0.03
$y_3^{(1)}$			1	0.20	0.19	0.16	0.15	0.13	0.14	0.07	0.05	0.07
$y_1^{(2)}$				1	0.64	0.60	0.21	0.19	0.19	0.11	0.09	0.10
$y_2^{(2)}$					1	0.62	0.20	0.19	0.18	0.08	0.12	0.09
$y_3^{(2)}$						1	0.18	0.19	0.18	0.09	0.08	0.11
$y_1^{(3)}$							1	0.69	0.61	0.20	0.19	0.21
$y_2^{(3)}$								1	0.63	0.22	0.20	0.21
$y_3^{(3)}$									1	0.20	0.21	0.20
$y_1^{(4)}$										1	0.73	0.66
$y_2^{(4)}$											1	0.70
$y_3^{(4)}$												1

some parameters related to the first time point are presented in Table 6.8. Although the estimates for Scheme 1 at the other time points are not reported in the table, all are similar over time to some degree. Compared with the estimates at the first time points, the maximal difference happens in the second intercept terms by the magnitude from 0.1 to 0.15 over time and the differences in the rest of parameters by 0.05 to 0.1. These differences could result from that the sampling error related to data generation, or the MCMC sampling errors due to underlying latent variables being involved.

The estimates for all the schemes are overall similar although there are the aforementioned differences in intercepts for Scheme 1 over time. All 95% credible intervals cover the true values. Furthermore, the estimates for Scheme 2 and Scheme 3, assuming measurement invariance, are similar despite the small differences.

Figure 6.5 exhibits the trajectories of the posterior estimated latent variable from model fitting using the three schemes. All the population trend lines (represented by a green-square-red-dashed object) show a linear decreasing trend. They cross through a bundle of trajectories from the middle, which may suggests all estimates are distributed with unimode Gaussianity across time.

To describe the discrepancy with the true trend line quantitatively, the mean absolute errors are reported in Table 6.9. For all time points, the differences for all the schemes are relatively small based on the variance of the estimated latent variable, around 6.7. It

Table 6.8: Comparisons between true values and the 5-simulation-chain averages over under three schemes.

	True	Scheme 1		Scheme 2		Scheme 3	
parameter		mean	sd	mean	sd	mean	sd
$\theta_{h,11}$	none	1.95	0.06	1.83	0.05	1.94	0.10
$\theta_{h,12}$	none	2.54	0.11	2.42	0.12	2.45	0.11
$\sigma_{t_{12}}$	0.70	0.63	0.07	0.63	0.06	0.62	0.07
$\sigma_{t_{13}}$	0.45	0.42	0.08	0.42	0.08	0.41	0.10
$\sigma_{t_{14}}$	0.20	0.23	0.09	0.22	0.08	0.26	0.10
λ_1	-1.21	-1.16	0.10	-1.15	0.07	-1.12	0.14
λ_2	-1.53	-1.45	0.10	-1.41	0.08	-1.39	0.12
λ_3	-1.10	-1.15	0.09	-1.08	0.07	-1.03	0.11
λ_{02}	1.75	1.83	0.11	1.77	0.08	1.72	0.10
λ_{03}	0.83	0.84	0.14	0.79	0.09	0.75	0.13

may suggest the green-square-red-dashed lines in Figure 6.6 indeed captures the true mean trend. Moreover, the mean absolute error for Scheme 3 is overall the largest. This seems to reflect the bias of using the two-step estimation procedure with limit information (which consists with the results in Table 5.5 before as well). In addition, the mean absolute errors for Scheme 2 seems the smallest.

Table 6.9: Mean absolute errors between the estimated trends and true one of the latent variable under the three schemes.

	Time point t			
	1	2	3	4
Scheme 1	0.493	0.541	0.369	0.392
Scheme 2	0.501	0.491	0.361	0.383
Scheme 3	0.512	0.587	0.385	0.325

Prediction

With the set-ups for previous prediction tasks, we conduct a 10 experiments for model prediction assessments. Instead of calculating RMSE, we use an accuracy rate, a percent-

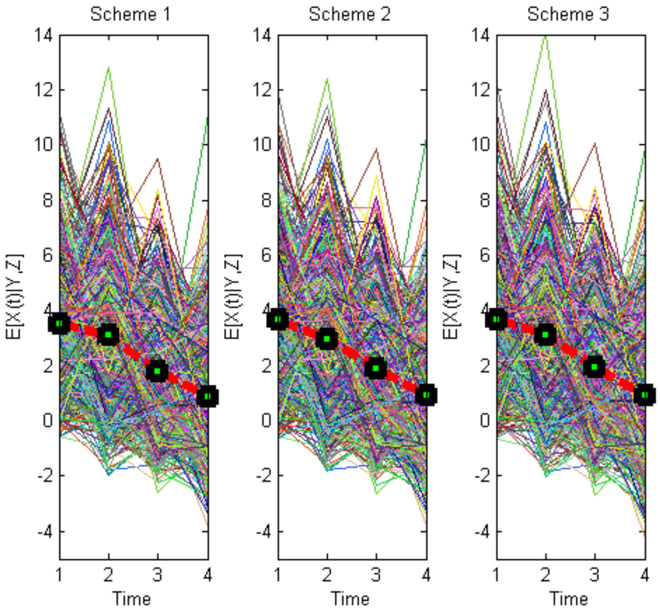


Figure 6.6: The trend plots of the posterior means of the latent variable under three schemes for Dataset V. Each colour represents a different data point. The population (green-square-red-dashed) line depicts the general trend pattern for all data points. Its value at each time point is simply the mean of all estimated latent variable.

age that the predictive response values match the true ones. The mean accuracy rates for all schemes are rather similar in terms of statistical level. All values are around 88% to 93%, higher than those for using logistic regression[3], where the rate is around 81%.

Except assessing predictive error on each variable, we also use a χ^2-type discrepancy quantity to examine the model suitability by PPC procedures. Similar to the χ^2 test statistic for goodness of fit in contingency table, the discrepancy quantity is

$$\begin{aligned} T(\mathbf{Y}|\mathbf{\Omega}) = \ & \frac{[P_{00}(y_r, y_{r'}) - E_{00}]^2}{E_{00}} + \frac{[P_{01}(y_r, y_{r'}) - E_{01}]^2}{E_{01}} + \\ & \frac{[P_{10}(y_r, y_{r'}) - E_{10}]^2}{E_{10}} + \frac{[P_{11}(y_r, y_{r'}) - E_{11}]^2}{E_{11}}, \end{aligned} \tag{6.17}$$

where $\mathbf{\Omega}$ contains all estimated parameters; $P_{00}(y_r, y_{r'})$ represents the proportion of the empirical data points with $y_r = 0$ and $y_{r'} = 0$. Likewise, E_{00} is the expectation proportion of $y_r = 0$ and $y_{r'} = 0$ calculated by the multivariate Gaussian distribution of underlying latent variables (u_r and $u_{r'}$) with the mean and covariance matrix given by Eqn. (4.28) and (4.29). The other observed proportions (P_{01}, P_{10}, P_{11}) and model expectation proportions (E_{01}, E_{10}, E_{11}) are defined in the respective way as well[4].

66 test statistics can therefore be invented based on the combinations for response pairs. All the discrepancy quantities are able to be calculated by Equation (6.17). Then the evaluation of the PP p-values follows by comparing the 2000 replicated test statistics values with their observed counterparts. Three tables of the p-values for all schemes are provided as a model fitting assessment. The intention is to see whether extreme values (less than 0.05 and more than 0.95) are overall observed or not, which reveals inappropriateness of model structure or assumptions[5].

Table 6.10, 6.11 and 6.12 show that all PP p-values are between 0.05 and 0.095, which indicates the model structure and the assumption of factor invariance are adequate. The conclusion of the checking is indeed not against the fact that the empirical data is actually generated by the model.

[3] we use 0.5 as a threshold value, if the predictive probability is larger than 0.5, then a response is valued as 1 and vice versa.

[4] The discrepancy quantity uses proportions, and the conventional the χ^2 test statistic use frequencies. The former is obtained from dividing the latter by the data size.

[5] We already used this procedure to realise the capacity of detecting model misfit. Due to space, we do not provide the detail experiment description and results here.

Table 6.10: The PP p-value results of model checking with Scheme 1 upon Dataset V. Each value is calculated based on comparing the values of chi-squared test statistic of empirical and replicated data for a pair of response variables.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$		0.170	0.273	0.110	0.101	0.140	0.162	0.158	0.173	0.124	0.125	0.133
$y_2^{(1)}$			0.213	0.121	0.128	0.127	0.179	0.137	0.185	0.106	0.192	0.142
$y_3^{(1)}$				0.161	0.157	0.148	0.246	0.154	0.214	0.197	0.269	0.217
$y_1^{(2)}$					0.279	0.320	0.350	0.269	0.240	0.316	0.176	0.288
$y_2^{(2)}$						0.321	0.322	0.289	0.215	0.160	0.212	0.282
$y_3^{(2)}$							0.306	0.271	0.234	0.322	0.239	0.371
$y_1^{(3)}$								0.532	0.698	0.751	0.530	0.710
$y_2^{(3)}$									0.560	0.906	0.828	0.634
$y_3^{(3)}$										0.812	0.834	0.504
$y_1^{(4)}$											0.524	0.558
$y_2^{(4)}$												0.707
$y_3^{(4)}$												

Table 6.11: The PP p-value results of model checking with Scheme 2 upon Dataset V.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$		0.350	0.785	0.194	0.102	0.466	0.700	0.634	0.528	0.826	0.752	0.871
$y_2^{(1)}$			0.279	0.144	0.127	0.185	0.482	0.699	0.172	0.710	0.332	0.680
$y_3^{(1)}$				0.277	0.315	0.839	0.261	0.769	0.338	0.569	0.182	0.168
$y_1^{(2)}$					0.738	0.365	0.137	0.327	0.499	0.291	0.635	0.286
$y_2^{(2)}$						0.530	0.188	0.195	0.597	0.577	0.459	0.295
$y_3^{(2)}$							0.359	0.429	0.625	0.355	0.538	0.173
$y_1^{(3)}$								0.139	0.803	0.125	0.432	0.174
$y_2^{(3)}$									0.793	0.143	0.090	0.288
$y_3^{(3)}$										0.085	0.092	0.450
$y_1^{(4)}$											0.522	0.849
$y_2^{(4)}$												0.748
$y_3^{(4)}$												

Table 6.12: The PP p-value results of model checking with Scheme 3 upon Dataset V.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$		0.671	0.862	0.070	0.081	0.265	0.504	0.502	0.386	0.883	0.724	0.812
$y_2^{(1)}$			0.534	0.062	0.106	0.104	0.333	0.627	0.109	0.771	0.308	0.628
$y_3^{(1)}$				0.130	0.195	0.698	0.174	0.641	0.209	0.419	0.149	0.898
$y_1^{(2)}$					0.901	0.772	0.116	0.408	0.446	0.368	0.674	0.389
$y_2^{(2)}$						0.861	0.207	0.415	0.607	0.783	0.613	0.415
$y_3^{(2)}$							0.329	0.438	0.617	0.435	0.701	0.237
$y_1^{(3)}$								0.442	0.901	0.069	0.308	0.104
$y_2^{(3)}$									0.871	0.106	0.069	0.235
$y_3^{(3)}$										0.057	0.088	0.429
$y_1^{(4)}$											0.535	0.881
$y_2^{(4)}$												0.694
$y_3^{(4)}$												

6.5.3 Study VI - Offense Crime Justice Data

The third longitudinal dataset is an extraction of Offense Crime Justice Study (OCJS) conducted from 2003 to 2006.[6] The original cohort study follows 2539 responders and documents 1044 items, which contain temporal demographic and socio-economic information, and responses of study questionnaires for offense history and risk factors. The extraction procedure is that for each year we first select 5 variables from a limited socio-economic variables, which are age, gender, household tenure, household income, employment status, and add one extra variable as time covariate. To simpify the empirical analysis, we make the house-income covariate a continuous variable by assigning a value through processing[7], and transform employment status as a binary variable by combining categories[8]. Next, we select 4 items and combine two variables (property offense and criminal damage offense) as a new variable. Therefore, the three responses represent other theft offense, vehicle theft offense, property damage offense. The final step is to delete the data points that missing value occurs on all covariates and responses over time.

Table 6.13 summaries the properties of the resulting dataset where p.t.p. means per

[6] The OCJS panel data can be accessed in the website of UK Data Archive.http://discover.ukdataservice.ac.uk/series/?sn=2000042

[7] A value is generated from a uniform density over the specific income ranges and then re-scaled.

[8] The status "student" is combined with employment, denoted by 1; "Economically inactive (others)" with non-employment, denoted by 0.

time point. Here the input character are mixed, we therefore separate the description into Figure 6.6 and Table 6.14.

In Figure 6.6, the distribution of responders' age (denoted by $z_1^{(t)}$, for $1 \leq t \leq 4$) shows that more teenagers aged around between 10 and 14 in the first year, the rest lightly spread ranged around from 14 to 25. The distributions for the successive years have similar scenario - more responders are teenagers but not merely transitions by years. For household income (denoted by $z_4^{(t)}$, for $1 \leq t \leq 4$), the temporal distributions seem different. In the first two years, more people's income are below scale 5 (which originally means 50000 pounds) and a spike is made of the most abundant family. In the remaining years more people's income are above scale 5 and the spikes at the top income family become more eminent. The lowest-income column also turns less.

Table 6.13: Properties of Dataset VI. Here p.t.p. is the abbreviation of per time point.

Dataset		Input		Output	
Size (N)	Time (T)	Dim. p.t.p. (D)	character	Dim. p.t.p. (R)	character
1274	4	6	mixed	3	binary

Table 6.14 shows that the proportions of female and male responders are almost 50-50, where the variables representing age ($z_2^{(t)}, 1 \leq t \leq 4$) are time-invariant. The household-tenure covariates ($z_3(t), 1 \leq t \leq 4$) have stable proportions over time - around two third of the responders are owners and one third are tenants. For employment status ($z_5(t), 1 \leq t \leq 4$), the steady time-varied proportions are about 90% being employed and 10% being unemployed.

Table 6.14: Proportions (%) of binary covariates of Dataset VI.

	$z_2^{(1)}$	$z_3^{(1)}$	$z_5^{(1)}$	$z_2^{(2)}$	$z_3^{(2)}$	$z_5^{(2)}$	$z_2^{(3)}$	$z_3^{(3)}$	$z_5^{(3)}$	$z_2^{(4)}$	$z_3^{(4)}$	$z_5^{(4)}$
0	47.9	68.5	91.1	47.9	69.9	90.4	47.9	70.0	89.9	47.9	69.9	89.5
1	52.1	31.5	8.9	52.1	30.1	9.6	52.1	30.0	10.1	52.1	30.1	10.5

Tables 6.15 reveals that most responders did not commit vehicle theft offense ($y_2^{(t)}, 1 \leq t \leq 4$) and property damage offense ($y_3^{(t)}, 1 \leq t \leq 4$). The proportions of the latter seem to increase with time. For other theft offense ($y_1^{(t)}, 1 \leq t \leq 4$), only almost one-tenth have

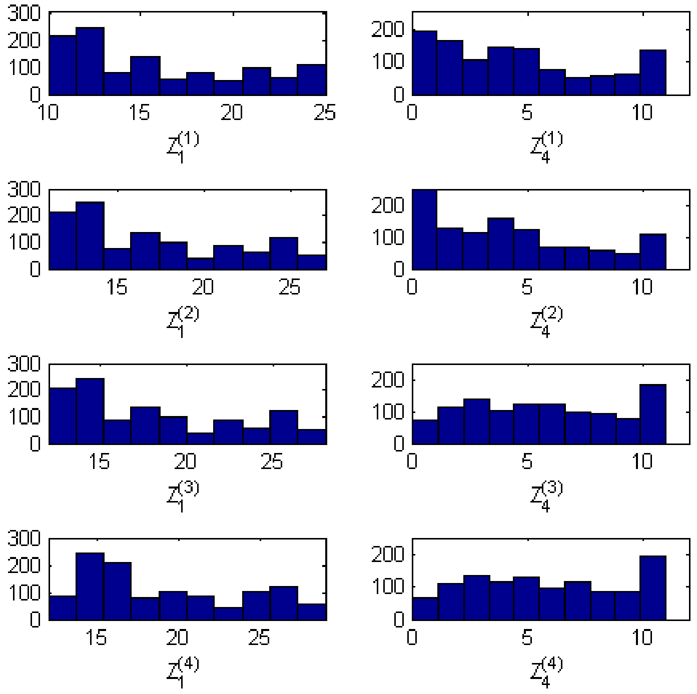


Figure 6.7: Distributions of the inputs for Dataset VI.

committed over time expect 13% at the second year.

Table 6.15: Proportions (%) of binary response variables of Dataset VI.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
0	90.7	98.7	97.2	86.9	98.7	98.6	90.3	98.8	98.8	90.3	97.2	98.9
1	9.3	1.3	2.8	13.1	1.3	1.4	9.7	1.2	1.2	9.7	2.8	1.1

Table 6.16 shows every three variables documented at the same time have lower-moderate inter-correlations. Overall, they have more even lower correlations with variables at other times, especially in further years. We assume them to measure one latent variable.

Table 6.16: Correlation coefficients between response variables of dataset VI.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$	1	0.27	0.26	0.33	0.12	0.14	0.20	0.10	0.02	0.16	0.04	0.02
$y_2^{(1)}$		1	0.29	0.12	0.05	-0.05	0.04	0.06	0.06	0.01	-0.01	-0.01
$y_3^{(1)}$			1	0.19	0.17	0.02	0.09	0.13	0.08	0.11	0.10	0.03
$y_1^{(2)}$				1	0.22	0.18	0.36	0.15	0.08	0.24	0.05	0.04
$y_2^{(2)}$					1	0.20	0.09	0.13	0.13	0.12	-0.01	-0.01
$y_3^{(2)}$						1	0.19	-0.01	0.05	0.09	0.07	0.13
$y_1^{(3)}$							1	0.21	0.21	0.31	0.10	0.17
$y_2^{(3)}$								1	0.28	0.15	0.26	0.14
$y_3^{(3)}$									1	0.13	0.26	0.43
$y_1^{(4)}$										1	0.21	0.23
$y_2^{(4)}$											1	0.19
$y_3^{(4)}$												1

Learning

200 pseudo inputs are used in these experiments. For Scheme 1 and 2, we set 30000 MCMC sampling iterations. But for Scheme 3, we fit a marginal model with 50000 samples and the joint model with 3000.

The EPSR values from 5-chain simulations are under 1.2 for all shemes. Achievement of simulation convergence can be considered.

Table 6.17 presents the estimation results of some parameters at the first time point. Overall the estimates of GP hyper-parameters and latent correlations are similar for all schemes. However, the rest manifest different results except the loading (corresponding to the third indicator). For those estimates that appear different, the overlap of the constructed 95% credible intervals are less.

The Scheme 1 estimates over the first three time points are similar. But, they are different from those at the fourth time point, except the parameters of the second loading and intercept. The maximal difference happens in one of loadings (corresponding to the first indicator) by the magnitude of around 1.6. This may suggest the measurement invariance assumption should not be imposed.

Table 6.17: Comparisons between true values and the 5-simulation-chain averages under three schemes.

	Scheme 1		Scheme 2		Scheme 3	
parameter	mean	sd	mean	sd	mean	sd
θ_{11}	2.37	0.15	2.30	0.19	2.35	0.14
θ_{12}	0.97	0.37	0.93	0.30	1.11	0.30
$\sigma_{t_{12}}$	0.65	0.06	0.66	0.06	0.64	0.07
$\sigma_{t_{13}}$	0.52	0.08	0.56	0.08	0.54	0.10
$\sigma_{t_{14}}$	0.52	0.09	0.51	0.08	0.52	0.10
λ_1	2.35	0.44	1.67	0.16	1.03	0.40
λ_2	1.44	0.30	0.91	0.19	1.97	0.36
λ_3	0.98	0.24	0.88	0.11	1.05	0.24
λ_{02}	-1.72	0.17	-1.80	0.09	-1.27	0.20
λ_{03}	-1.2	0.14	-1.60	0.07	-1.03	0.13

Figure 6.8 shows temporal changes of the latent-variable estimates under the three schemes. The trajectories, each of which forms by linking the ergodic average estimates at different time points, overall exhibit a similar pattern. A bundle of lines lie in the lower part of each subfigure, and some located over them have relatively huge fluctuations over time. This pattern can be more evidently found in the subfigures for Scheme 1 and 2. Also, it suggests the distribution of the estimates at each time point has multiple modes or a mixture distributional structure or a long tail. We later provide another figure to evidence the suggestion.

The three population trend lines are drawn into the upper edge of the bundle. They do not pass through it from the middle and remain level at the three time points. Moreover, the lines for Scheme 2 and 3 stay even at the end, but that for Scheme 1 has a drop.

It should be noted that the pattern of the population line for Scheme 1 may not be interpreted for the growth trend of the latent variables because measurement invariance assumption is not imposed. It only works as a reference line for the comparison with those for the other two schemes. The population trend lines for Scheme 2 and Scheme 3 overall remain negative and have no change under the measurement invariance assumption. One could name the latent variable as stealing potentials and interpret the general trend that subjects have had rather weak inclination in stealing or robbing acroos time[9].

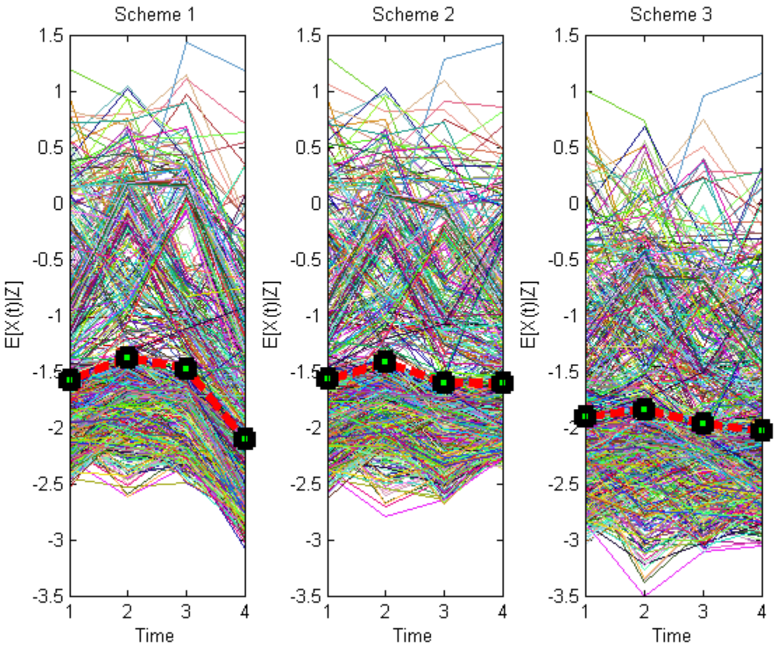


Figure 6.8: The trend plots of the posterior means of the latent variable under three schemes for Dataset VI. Each colour represents a different data point. The green-square-red-dashed line depicts the population trend pattern for all data points. Its value at each time point is simply the mean of all estimated latent variable.

Figure 6.9 provides another perspective to see posterior estimated latent variables at the first three waves under all schemes. Each subfigure is the projection of the four-dimensional latent variable into two-dimensional subspace. Two prominent features for all subfigures can be observed: a large cluster consisting of high proportions of data points is located in the bottom-left area corresponding to negative values; four small clusters

[9] The interpretation is made under the positive associated loadings are positive.

roughly lie in the centre or the top-right area. The second characteristic seems more evident over the points under Scheme 1 and 2; however, that is comparably not clear under Scheme 3.

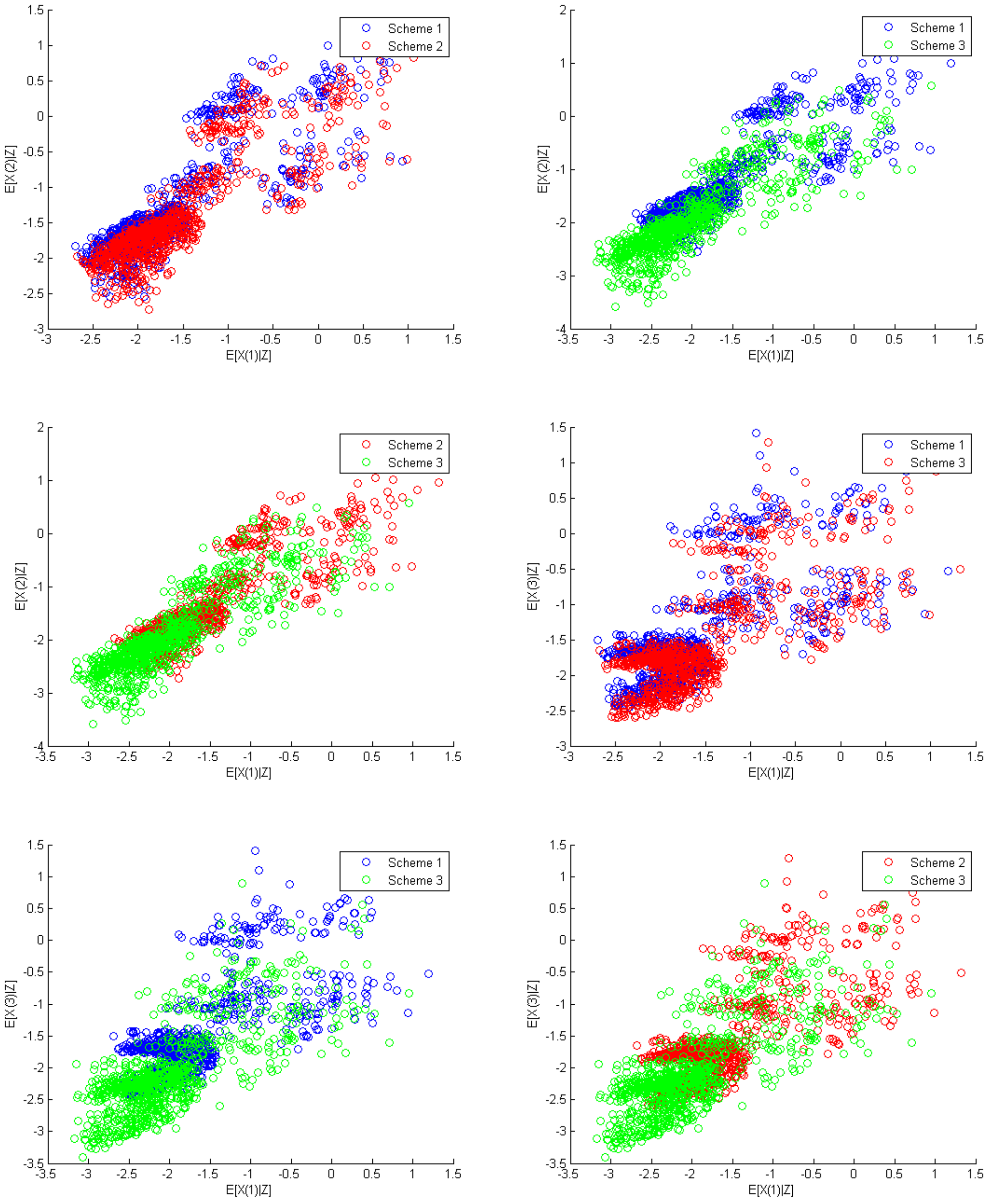


Figure 6.9: Scatter plots of the posterior estimates for latent variables $(x^{(1)}, x^{(2)})$ and for $(x^{(1)}, x^{(3)})$ under all schemes.

In fact, Hales et al. (2009) already uses latent class analysis to detect five distinct groups of offense behavioural patterns over all subjects [10]. This may reflect the number of the clusters we observe in Figure 6.9.

Prediction

We still conduct a 5-fold cross validation for model prediction assessment, where the sizes of each training set and test set are around 1019 and 255 and the pseudo-input size is 150.

As the previous study, we calculate the mean accuracy rates over 5 test sets under all schemes. All resulting values are identical across the schemes and are close to the proportion of 0 in Table 6.15. This is because all underlying predictive latent variables are valued left far from the axis origin. The accuracy rates of using logistic regression are lower to a degree, which may be due to the non-linear regression relation.

The results of PPC procedure for model fitting under three schemes are provided. Table 6.18 shows only one PP p-value is out of a reasonable range. Therefore we can still consider the appropriateness of the model is acceptable despite the defect. Table 6.19 and 6.20 reveals imposing the measurement invariance assumption may be not sensible. The reason is that more extreme p-values (coloured as red) occur – 14 are found in Table 6.19 and 17 in Table 6.20.

6.6 Remarks

Our model framework for longitudinal analysis shares some similarity on factor analysis model structure with LV-LCM. But, both frameworks adopt different concepts to model temporal change of a latent variable. LV-LCM with latent variables uses a set of pre-specified latent factors to restrict regression functional family between covariates and latent variables. The latent factors and the related random disturbances characterise an overall temporal pattern and the difference across all the units. Instead, longitudinal sparse GP-SEM can utilize non-parameteric probabilistic framework to increase the flexibility for modelling the regression function. Through post-processing calculations, the individual trajectories and general temporal pattern of a latent variable can be obtained.

[10] The responses they chose for data analysis are somewhat different from ours.

Table 6.18: The PP p-value results of model checking with Scheme 1 upon Dataset VI. Each value is calculated based on comparing the values of chi-squared test statistic of empirical and replicated data for a pair of response variables.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$		0.208	0.154	0.098	0.188	0.198	0.147	0.178	0.073	0.092	0.207	0.095
$y_2^{(1)}$			0.659	0.384	0.683	0.446	0.203	0.631	0.608	0.181	0.467	0.313
$y_3^{(1)}$				0.418	0.328	0.718	0.364	0.459	0.432	0.261	0.609	0.461
$y_1^{(2)}$					0.398	0.394	0.151	0.397	0.136	0.097	0.085	0.095
$y_2^{(2)}$						0.428	0.262	0.501	0.642	0.215	0.579	0.344
$y_3^{(2)}$							0.241	0.573	0.692	0.261	0.449	0.246
$y_1^{(3)}$								0.361	0.374	0.061	0.330	0.305
$y_2^{(3)}$									0.783	0.213	0.079	0.535
$y_3^{(3)}$										0.275	0.484	0.019
$y_1^{(4)}$											0.087	0.289
$y_2^{(4)}$												0.446
$y_3^{(4)}$												

Table 6.19: The PP p-values of model checking with Scheme 2 upon Dataset VI.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$		0.124	0.004	0.168	0.869	0.893	0.500	0.780	0.638	0.709	0.889	0.856
$y_2^{(1)}$			0.001	0.111	0.239	0.139	0.101	0.206	0.217	0.070	0.160	0.163
$y_3^{(1)}$				0.003	0.005	0.005	0.003	0.005	0.005	0.002	0.004	0.009
$y_1^{(2)}$					0.383	0.268	0.043	0.378	0.191	0.315	0.089	0.202
$y_2^{(2)}$						0.497	0.318	0.652	0.814	0.659	0.847	0.767
$y_3^{(2)}$							0.314	0.725	0.908	0.661	0.778	0.399
$y_1^{(3)}$								0.452	0.356	0.270	0.416	0.364
$y_2^{(3)}$									0.898	0.629	0.091	0.826
$y_3^{(3)}$										0.605	0.649	0.004
$y_1^{(4)}$											0.117	0.566
$y_2^{(4)}$												0.964
$y_3^{(4)}$												

Table 6.20: The PP p-values of model checking with Scheme 3 upon Dataset VI.

	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$y_1^{(3)}$	$y_2^{(3)}$	$y_3^{(3)}$	$y_1^{(4)}$	$y_2^{(4)}$	$y_3^{(4)}$
$y_1^{(1)}$		0.108	0.018	0.000	0.885	0.559	0.032	0.777	0.367	0.102	0.781	0.678
$y_2^{(1)}$			0.034	0.065	0.822	0.475	0.216	0.878	0.567	0.153	0.725	0.619
$y_3^{(1)}$				0.008	0.384	0.146	0.047	0.273	0.084	0.051	0.277	0.199
$y_1^{(2)}$					0.044	0.008	0.000	0.033	0.005	0.001	0.010	0.010
$y_2^{(2)}$						0.454	0.263	0.853	0.744	0.291	0.903	0.848
$y_3^{(2}$							0.081	0.667	0.478	0.119	0.732	0.468
$y_1^{(3)}$								0.197	0.053	0.001	0.206	0.116
$y_2^{(3)}$									0.562	0.211	0.063	0.916
$y_3^{(3)}$										0.081	0.569	0.032
$y_1^{(4)}$											0.040	0.082
$y_2^{(4)}$												0.991
$y_3^{(4)}$												

The proposed methodology addresses continuous or binary responses. The former model framework has analogy with its static version presented in Chapter 3. The latter is an extension of the former and adds another level for latent continuous responses. That level represents the underlying distribution of binary responses. In the aspects of estimation, prediction and computation, the modelling fitting with binary responses are similar to that with continuous ones. The differences are the augmented sampling scheme for the underlying latent variables, which adopts truncated Gaussian densities. The predictive binary values are based on the means of the predictive latent continuous responses.

Imposing the measurement invariance assumption or conducting the two-step estimation procedure differentiates computational schemes. The measures can give a genuine implication for temporal changes of a latent variable and endow a possible computational benefit. The appropriateness of the invariance assumption can be examined by initially inspecting the estimates and then conducting posterior predicitive checking.

Chapter 7

Discussion and Conclusion

In this thesis we have done exploratory works for GP-SEM in model estimation, efficient computation, multiple-output prediction and applications to longitudinal analysis. Here, we review our works and main contributions; also discuss possible improvement and future works.

We have presented our new modelling methodology (GP-SEM), which is built on GP probability framework and factor analysis model. Due to the constitution features, GP-SEM has capabilities including exploring distributions of latent variables (constructs), reducing response dimension, and realizing functional regressions between covariates and latent variables. It can serve as a causal model like SEM or MIMIC model (Bollen 1989, Pearl 2000) to examine a posited causal relationship between observed covariates and responses, where latent variables are medium. Although the structural model in GP-SEM is to describe assumed causal relations between observed covariates and latent variables rather than among latent variables, the framework can certainly extend to the latter case by simply incorporating another factor model.

We have adapted GP-SEM to addresses a computational issue on large dataset. The idea is motivated by the sparse GP (SPGP) approximation approach (Snelson & Ghahramani 2006*a*). We adopt a set of variables to modify the original GP prior of function values. Our sparse approximation treatment is different from the SPGP model. The pseudo (or inducing) inputs are selected from a training dataset rather than freely estimated by optimisation.

We also have demonstrated GP-SEM as a longitudinal analysis instrument to realise temporal change of latent variables under appropriate assumptions. Unlike LV-LCM to capture the mean trend simultaneously with estimation though, GP-SEM enables to achieve the task by post-processing calculations.

Given the model structure, we have demonstrated its identification examination under appropriate parameter constraints. We used two ways – algebraic derivations through moments and rank calculations of a Jacobian matrix – to ensure one-one relations between unknown and reduced-form parameters for local identifiability.

Regarding the computational algorithms we have provided three estimation methods. The first two mainly rely on Gibbs sampling and Metropolis-Hastlings simulation approaches. The second algorithm, parameter expansion, is additionally applied to enhance mixing for factor loadings and latent correlation matrix. For continuous responses, experiment results show simulations converge fast, but for binary responses, more MCMC samples are demanded, especially for factor loadings and intercepts. Instead of imposing more restrictions, it may be worth trying to apply the technique of parameter expansion data augmentation (PX-DA) (Liu & Wu 1999) to increase sampling efficiency on those parameters. The technique found successful for MCMC implementation of (multivariate) probit regression although in the applications the covariates are observed.

Furthermore, we have adapted an approach for the locations of inducing inputs. A random walk sampling scheme is initially created in Algorithm 1. And then considering possible computation cost reductions, a random selection scheme is utilized in Algorithm 2 before model fitting. We further used greedy selection based on information gain in entropy. The associated and fixed estimates of GP hyper-parameters for selection are from preliminary fitting by marginal models. We found that as the number of inducing inputs increase, the predictive performance of sparse GP-SEM (under either selection schemes) overtakes that of GP regression by independently fitting models for individual responses.

The third estimation method is a hybrid algorithm combining MCEM and IFM approaches. Though an estimation bias occurs and the predictive capability underperforms, parameter estimates, to some degree, are still close to the true values or the ones estimated merely under a joint model. The issue of slow convergence of parameters (for loadings and intercepts), usually happening in EM implementations, does not emerge in our experiments. The reason could be due to prior-centring responses and randomly generated starting points not departed from true values by chance. To avoid this, one can use a technique of parameter expansion expectation maximum (PX-EM) (Liu et al. 1998), which introduces a non-zero auxiliary parameter to make latent variables more uncertain.

The two-step estimation scheme of the hybrid algorithm is additionally adopted with MCMC methods in experiments as well. The estimates from marginal models are close to the true values. This suggests it is reasonable to set those estimates as starting values before fitting a joint model. We found that under that set-up, MCMC simulation converges indeed more efficiently than it does under random initialization. We also inspected that using the two-step computational scheme can be much less time-consuming than using the one-step scheme (fitting a joint model). This happens especially when more latent variables are involved.

We also did some other works in empirical studies. For learning tasks we explored distribution of latent variables before and after data standardisation and discovered the effect of processing. We gave mathematical explanation about the processing effect. The investigation into differences in latent variables estimated from marginal models and a joint model was conducted as well. For prediction tasks, instead of assessing model predictive performance for each response by RMSE, we further adopted predictive posterior checking to assess discrepancy between empirical and replicated datasets. This reflects whether or not model fitting is appropriate. Additionally, through a special procedure, we realized individual functional relationship between each covariate and a latent variable in high-dimensional input space.

The limit of the model frameworks results from huge computation cost in some circumstances. Although inventing efficient estimation methods to reduce computational expenses, we do not completely solve the computational problem. When the scale of a dataset is rather large, the size of inducing inputs may be increased relatively. In addition to more latent variables (constructs) or more time points involved, all make computation very slow. Practitioners could consider divide the dataset into relatively smaller-size subsets and directly implement estimation with the two-step computational strategy.

There are several points that our framework still demands concerns and the improvement can be carried out in the future. First is the GP covariance function and high-dimensional inputs. We only adopt square-exponential (SE) covariance function in all experiment implementations for computational convenience. It gives regression function equal smoothness on each dimension of covariates. This feature is unrealistic likely on some dataset, especially when some covariates may be inappropriate as a predictor. Using automatic relevance determination (ARD) (Rasmussen & Williams 2006) covariance function may lead a refined model and improve model predictive performance. It can diminish effect of covariates whose length scales are large in the corresponding input dimension. Furthermore, one could introduce a covariance function to project input space on a low-dimension subspace (Snelson & Ghahramani 2006*b*). If a SE covariance function is still considered, one can generate latent covariates by PCA for input dimension reduction.

Next point is about inducing inputs. When the number of pseudo input is low, the RMSEs of using greedy selection are smaller than those of using random selection. This may imply the locations by the former selection are more appropriate to capture the underlying regression functional relations. The greedy selection scheme depends on the fixed estimates of hyper-parameters. We could perhaps examine whether using an iteratively-varied[1] low-size pseudo inputs reduces predictive errors further.

On the other hand, if using a large number of inducing inputs is necessary, the model predictive performance may suffer from overfitting. That issue is reported in some of SPGP applications but does not occur in our few experiments. There are possible reasons why overfitting does not happen. It may be that the selected inducing inputs are good enough to sketch the underlying functional relationship between covariates and latent variables. It could be that the functional relationship is rather smooth, not too wiggly; or the inducing inputs are selected to be fixed rather than freely estimated by optimisation. To avoid that possible issue, our framework can be extended by adopting Titsias's variational approximation scheme (Titsias 2009). We could use greedy selection for inducing inputs to maximise the variational lower bound of the log marginal likelihood of latent variables.

Several modelling variants can be considered in circumstances for different types of response. Although only dealing with data with fully continuous or dichotomous outcomes, we can certainly apply GP-SEM to mixed type. One merely needs to modify the model structure compatible for the both types. The involved binary responses link latent continuous responses as before. For ordered categorical responses, we can extend GP-SEM by adding framework for sampling the categorical thresholds of latent responses. For re-

[1] "Iteratively-varied" means the low-size pseudo inputs are obtained by conducting alternative estimation and greedy selection with iterations. Note that each iteration produces new estimates of hyper-parameters and new pseudo inputs.

sponses of discrete event counts, the model structure can be changed to the one without measurement errors. The mean of outcome variables associated with the linear predictor of latent variables (constructs) by a link log. This is a typical formulation in generalized linear models (Skrondal & Rabe-Hesketh 2004).

Our framework can be applied to incomplete data. However, concerns and necessary data processing may need to be drawn. If missing values occur in covariates, practitioners demand to adopt some treatments, such as imputation or listwise deletion (Gelman & Hill 2007). If missing values happen on response variables, there could be no problem in our framework under the assumed missing-at-random (MAR) mechanism. For pure sampling estimation methods, one can simply regard missing values as latent variables and use data augmentation technique (Tanner & Wong 1987). Those missing responses are able to be sampled through the corresponding full conditional density. For the hybrid method about MCEM and IFM, one can still sample those missing responses and other latent variables to approximate conditional expectations in E-step. Incidentally, the MAR assumption maybe can be examined by assessing the MAR+ assumption (Potthoff et al. 2006).

Sensitivity analysis can be done for different choices of prior distributions and covariance functions. The instruments of analysis can be predictive posterior checks by various discrepancy statistics or cross-validation by different measure criteria.

The results of the final empirical study prompt us to develop an expanded framework for better modelling fit. One potential idea is to further expand GP-SEM into mixture model framework. We may need to introduce another set of latent variables for the proportion of different groups of data points. And also it would expand more model parameters, such as factor loadings and intercepts. The different components of latent constructs are possibly modelled by separate GP frameworks. The related estimation method can refer existing approaches, such as the sampling MCMC methods (Lee 2007).

Appendix A

A.1 Matrix identities

A.1.1 Matrix inversion lemma (Sherman-Morrison-Woodbury formula)

$$
\begin{aligned}
(\mathbf{A}+\mathbf{C}\mathbf{B}\mathbf{C}^{\mathsf{T}})^{-1} &= \mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1}+\mathbf{C}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}^{\mathsf{T}}\mathbf{A}^{-1} && \text{(A.1)}\\
|\mathbf{A}+\mathbf{C}\mathbf{B}\mathbf{C}^{\mathsf{T}}| &= |\mathbf{B}||\mathbf{A}||\mathbf{B}^{-1}+\mathbf{C}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{C}| && \text{(A.2)}
\end{aligned}
$$

A.1.2 Block Matrix inversion lemma

$$
\mathbf{A}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{12}^{\mathsf{T}} & \mathbf{A}_{22}\end{bmatrix},
$$

$$
\mathbf{A}^{-1}=\begin{bmatrix}\tilde{\mathbf{A}}_{11} & \tilde{\mathbf{A}}_{12}\\ \tilde{\mathbf{A}}_{12}^{\mathsf{T}} & \tilde{\mathbf{A}}_{22}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{11}^{-1}+\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\tilde{\mathbf{A}}_{22}\mathbf{A}_{21}^{\mathsf{T}}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\tilde{\mathbf{A}}_{22}\\ -(\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\tilde{\mathbf{A}}_{22})^{\mathsf{T}} & \tilde{\mathbf{A}}_{22}\end{bmatrix}, \tag{A.3}
$$

where $\tilde{\mathbf{A}}_{22}=(\mathbf{A}_{22}-\mathbf{A}_{12}^{\mathsf{T}}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}$

$$
|\mathbf{A}|=|\mathbf{A}_{11}||\tilde{\mathbf{A}}_{22}| \tag{A.4}
$$

A.2 Gaussian identity

A.2.1 Multiplication

$f_a(\boldsymbol{\nu})$ and $f_b(\boldsymbol{\nu})$ are the pdf of $\mathcal{N}(\mathbf{a},\mathbf{A})$ and $\mathcal{N}(\mathbf{b},\mathbf{B})$, then $f_a(\boldsymbol{\nu})\cdot f_b(\boldsymbol{\nu})$ is a Gaussian function proportional to the pdf of $\mathcal{N}(\mathbf{c},\mathbf{C})$, where

$$
\begin{aligned}
\mathbf{c} = &\ \mathbf{C}\mathbf{A}^{-1}\mathbf{a}+\mathbf{C}\mathbf{B}^{-1}\mathbf{b},\\
\mathbf{C} = &\ \quad(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}.
\end{aligned} \tag{A.5}
$$

A.2.2 Conditional distribution

If $\boldsymbol{\nu}_1 \sim \mathcal{N}(\mathbf{a}, \mathbf{A})$, $\boldsymbol{\nu}_2 \sim \mathcal{N}(\mathbf{b}, \mathbf{B})$ and

$$\begin{bmatrix} \boldsymbol{\nu}_1 \\ \boldsymbol{\nu}_2 \end{bmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{B} \end{bmatrix} \Big),$$

then

$$\boldsymbol{\nu}_1 | \boldsymbol{\nu}_2 \sim \mathcal{N}(\mathbf{a} + \mathbf{C}\mathbf{B}^{-1}(\boldsymbol{\nu}_2 - \mathbf{b}), \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\mathsf{T}) \qquad \text{(A.6)}$$

A.2.3 Integration

If $\boldsymbol{\nu}|\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{C}\boldsymbol{\omega}, \mathbf{A})$ and $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{B})$, then integrating out $\boldsymbol{\omega}$ from the joint likelihood of $\boldsymbol{\nu}$ and $\boldsymbol{\omega}$ leads to achieve the marginal likelihood of $\boldsymbol{\nu}$,

$$p(\boldsymbol{\nu}) \quad = \quad \int p(\boldsymbol{\nu}|\boldsymbol{\omega}) p(\boldsymbol{\omega}) d\boldsymbol{\omega},$$

and

$$\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{A} + \mathbf{C}\mathbf{B}\mathbf{C}^\mathsf{T}). \qquad \text{(A.7)}$$

A.3 Derivation of Samplers in Section 4.2

A.3.1 Pseudo latent functions

$\bar{\mathbf{f}}_q$ denotes the q-th pseudo latent function vector, the full conditionals is

$$\begin{aligned} p(\bar{\mathbf{f}}_q|e.e.) \quad &\propto \quad p(\mathbf{f}_q|\bar{\mathbf{f}}_q, \mathbf{z}^{1:N}, \bar{\mathbf{z}}_q^{1:M}, \boldsymbol{\theta}_{h,q}) p(\bar{\mathbf{f}}_q|\bar{\mathbf{z}}_q^{1:M}, \boldsymbol{\theta}_{h,q}) \\ &\propto \quad \exp\Big\{ -\frac{1}{2}(\mathbf{f}_q - \mathbf{K}_{q;NM}\mathbf{K}_{q;M}^{-1}\bar{\mathbf{f}}_q)^\mathsf{T} \mathbf{V}_q^{-1} (\mathbf{f}_q - \mathbf{K}_{q;NM}\mathbf{K}_{q;M}^{-1}\bar{\mathbf{f}}_q) \Big\} \cdot \\ &\qquad\qquad\qquad\qquad \exp\Big\{ -\frac{1}{2}\bar{\mathbf{f}}_q^\mathsf{T}\mathbf{K}_{q;M}^{-1}\bar{\mathbf{f}}_q \Big\} \\ &\propto \quad \exp\Big\{ -\frac{1}{2}(\bar{\mathbf{f}}_q - \boldsymbol{\mu}_{\bar{\mathbf{f}}_q,post})^\mathsf{T} \boldsymbol{\Sigma}_{\bar{\mathbf{f}}_q,post}^{-1} (\bar{\mathbf{f}}_q - \boldsymbol{\mu}_{\bar{\mathbf{f}}_q,post}) \Big\} \qquad \text{by (A.5)}, \end{aligned}$$

where $e.e.$ means eveything else.

A.3.2 Latent variables

Let all latent variables denoted $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}]$, $\mathbf{x}^{(n)} = [x_1^{(n)}, \ldots, x_Q^{(n)}]^\mathsf{T}$ then the full conditionals:

$$\begin{aligned}
p(\mathbf{X}|e.e.) &\propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_y) p(\mathbf{X}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_{1:Q}^{1:M}, \boldsymbol{\Theta}_h, \boldsymbol{\Sigma}_x, \bar{\mathbf{f}}) \\
&\propto \prod_{n=1}^{N} \Big[\exp\Big\{ -\frac{1}{2}(\mathbf{y}^{(n)} - \boldsymbol{\Lambda}\mathbf{x}^{(n)} - \boldsymbol{\lambda}_0)^\mathsf{T} \boldsymbol{\Sigma}_y^{-1} (\mathbf{y}^{(n)} - \boldsymbol{\Lambda}\mathbf{x}^{(n)} - \boldsymbol{\lambda}_0) \Big\} \cdot \\
&\qquad \exp\Big\{ -\frac{1}{2}(\mathbf{x}^{(n)} - \widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\bar{\mathbf{f}})^\mathsf{T} (\mathbf{V}^{(n)} + \boldsymbol{\Sigma}_x)^{-1} (\mathbf{x}^{(n)} - \widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\bar{\mathbf{f}}) \Big\} \Big] \\
&\propto \prod_{n=1}^{N} \Big[\exp\Big\{ -\frac{1}{2}[\boldsymbol{\Lambda}\mathbf{x}^{(n)} - (\mathbf{y}^{(n)} - \boldsymbol{\lambda}_0)]^\mathsf{T} \boldsymbol{\Sigma}_y^{-1} [\boldsymbol{\Lambda}\mathbf{x}^{(n)} - (\mathbf{y}^{(n)} - \boldsymbol{\lambda}_0)] \Big\} \cdot \\
&\qquad \exp\Big\{ -\frac{1}{2}(\mathbf{x}^{(n)} - \widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\bar{\mathbf{f}})^\mathsf{T} (\mathbf{V}^{(n)} + \boldsymbol{\Sigma}_x)^{-1} (\mathbf{x}^{(n)} - \widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\bar{\mathbf{f}}) \Big\} \Big] \\
&\propto \prod_{n=1}^{N} \Big[\exp\Big\{ -\frac{1}{2}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_{\mathbf{x}^{(n)}})^\mathsf{T} \boldsymbol{\Sigma}_{\mathbf{x}^{(n)}}^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_{\mathbf{x}^{(n)}}) \Big\} \Big] \qquad \text{by (A.5)}
\end{aligned}$$

A.3.3 Latent functions

Let latent functions be $\mathbf{F} = [\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(N)}]$, $\mathbf{f}^{(n)} = [f_1^{(n)}, \ldots, f_Q^{(n)}]^\mathsf{T}$, then the full conditionals:

$$\begin{aligned}
p(\mathbf{F}|e.e.) &\propto p(\mathbf{X}|\mathbf{F}, \boldsymbol{\Sigma}_x) p(\mathbf{F}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_{1:Q}^{1:M}, \boldsymbol{\Theta}_h, \bar{\mathbf{f}}) \\
&\propto \prod_{n=1}^{N} \Big[\exp\Big\{ -\frac{1}{2}(\mathbf{x}^{(n)} - \mathbf{f}^{(n)})^\mathsf{T} \boldsymbol{\Sigma}_x^{-1} (\mathbf{x}^{(n)} - \mathbf{f}^{(n)}) \Big\} \cdot \\
&\qquad \exp\Big\{ -\frac{1}{2}(\mathbf{f}^{(n)} - \widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\bar{\mathbf{f}})^\mathsf{T} (\mathbf{V}^{(n)})^{-1} (\mathbf{f}^{(n)} - \widetilde{\mathbf{K}}_{nM}\widetilde{\mathbf{K}}_M^{-1}\bar{\mathbf{f}}) \Big\} \Big] \\
&\propto \prod_{n=1}^{N} \Big[\exp\Big\{ -\frac{1}{2}(\mathbf{f}^{(n)} - \boldsymbol{\mu}_{\mathbf{f}^{(n)}})^\mathsf{T} \boldsymbol{\Sigma}_{\mathbf{f}^{(n)}}^{-1} (\mathbf{f}^{(n)} - \boldsymbol{\mu}_{\mathbf{f}^{(n)}}) \Big\} \Big] \qquad \text{by (A.5)}
\end{aligned}$$

A.3.4 Factor loadings and intercept

For any r, let $\boldsymbol{\lambda}_r = (\lambda_{qr}, \lambda_{0r})^\mathsf{T}$, and s indicates the corresponding index of latent variables to the indicator r. Let $\mathbf{y}_r = (y_r^{(1)}, \dots, y_r^{(N)})^\mathsf{T}$, $\mathbf{x}_s = (\mathbf{x}_s^{(1)}, \dots, \mathbf{x}_s^{(N)})^\mathsf{T}$ and $\tilde{\mathbf{X}}_s = [\mathbf{x}_s \mathbf{1}_N]$, then the full conditionals:

$$
\begin{aligned}
p(\boldsymbol{\lambda}_r|e.e.) &\propto p(\mathbf{y}_r|\mathbf{x}_s, \sigma^2_{y_r})p(\boldsymbol{\lambda}_r) \\
&\propto \exp\Big\{-\frac{1}{2}(\mathbf{y}_r - \tilde{\mathbf{X}}_s\boldsymbol{\lambda}_r)^\mathsf{T}(\sigma^2_{y_r})^{-1}(\mathbf{y}_r - \tilde{\mathbf{X}}_s\boldsymbol{\lambda}_r)\Big\}\exp\Big\{-\frac{1}{2}\boldsymbol{\lambda}_r^\mathsf{T}(\sigma^2_\lambda)^{-1}\boldsymbol{\lambda}_r\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\boldsymbol{\lambda}_r^\mathsf{T}[\tilde{\mathbf{X}}_s^\mathsf{T}\tilde{\mathbf{X}}_s(\sigma^2_{y_r})^{-1} + (\sigma^2_\lambda)^{-1}]^{-1}\boldsymbol{\lambda}_r - \mathbf{y}_r^\mathsf{T}(\sigma^2_{y_r})^{-1}\tilde{\mathbf{X}}_s\boldsymbol{\lambda}_r \\
&\quad -(\tilde{\mathbf{X}}_s\boldsymbol{\lambda}_r)^\mathsf{T}(\sigma^2_{y_r})^{-1}\mathbf{y}_r\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}(\boldsymbol{\lambda}_r - \boldsymbol{\mu}_{\lambda_r,post})^\mathsf{T}\boldsymbol{\Sigma}^{-1}_{\lambda_r,post}(\boldsymbol{\lambda}_r - \boldsymbol{\mu}_{\lambda_r,post})\Big\}
\end{aligned}
$$

where $\boldsymbol{\mu}_{\lambda_r,post} = (\sigma^2_{y_r})^{-1}\boldsymbol{\Sigma}_{\lambda_r,post}(\tilde{\mathbf{X}}_s)^\mathsf{T}\mathbf{y}_r$.

A.3.5 Measurement error variances

For any r, let measurement error variances be $\sigma^2_{y_r}$, denote $\mathbf{x}_s = (x_s^{(1)}, \dots, x_s^{(N)})^\mathsf{T}$ and $\tilde{\mathbf{x}}_s^{(n)} = (x_s^{(n)}, 1)$, then the full conditional:

$$
\begin{aligned}
p(\sigma^2_{y_r}|e.e.) &\propto p(\sigma^2_{y_r}) \cdot p(\mathbf{y}_r|\mathbf{x}_s, \boldsymbol{\lambda}_r, \sigma^2_{y_r},) \\
&\propto (\sigma^2_{y_r})^{-a_0-1}\exp\{-b_0/\sigma^2_{y_r}\} \cdot \\
&\qquad \prod_{n=1}^{N}(\sigma^2_{y_r})^{-1/2}\exp\Big\{-\frac{1}{2}(y_r^{(n)} - \tilde{\mathbf{x}}_s^{(n)}\boldsymbol{\lambda}_r)^\mathsf{T}(\sigma^2_{y_r})^{-1}(y_r^{(n)} - \tilde{\mathbf{x}}_s^{(n)}\boldsymbol{\lambda}_r)\Big\} \\
&\propto (\sigma^2_{y_r})^{-(a_0+N/2)-1}\exp\Bigg\{-\Big[b_0 + \frac{1}{2}\sum_{n=1}^{N}(y_r^{(n)} - \tilde{\mathbf{x}}_s^{(n)}\boldsymbol{\lambda}_r)^2\Big](\sigma^2_{y_r})^{-1}\Bigg\}
\end{aligned}
$$

A.3.6 Correlation matrix of GP noises

The full conditionals of $\boldsymbol{\Sigma}_s$ is

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_s|e.e.) &\propto p(\mathbf{W}_0|\boldsymbol{\Sigma}_s)p(\boldsymbol{\Sigma}_s) \\
&\propto |\Sigma_s|^{-N/2}\exp\Big\{-\frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{W}_0\mathbf{W}_0^\mathsf{T})\Big\} \cdot \\
&\qquad\qquad |\Sigma_s|^{-(2+Q+1)/2}\exp\Big\{-\frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{I}_Q)\Big\} \\
&\propto |\Sigma_s|^{-(N+2+Q+1)/2}\exp\Big\{-\frac{1}{2}tr\big(\boldsymbol{\Sigma}^{-1}(\mathbf{W}_0\mathbf{W}_0^\mathsf{T} + \mathbf{I}_Q)\big)\Big\}.
\end{aligned}
$$

A.4 Derivation of Samplers in Section 4.3

A.4.1 Latent variables

Let $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}]$, $\mathbf{x}$ is the columnwise rearrangement of $\mathbf{X}$. The full conditional of latent variable vector $\mathbf{x}$ is

$$\begin{aligned}
p(\mathbf{X}|e.e. \setminus \mathbf{f}, \setminus \bar{\mathbf{f}}) &\propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_y) p(\mathbf{X}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}^{1:M}_{1:Q}, \boldsymbol{\Theta}_h, \boldsymbol{\Sigma}_x) \\
&\propto \prod_{n=1}^{N} \left[\exp\left\{ -\frac{1}{2}(\mathbf{y}^{(n)} - \boldsymbol{\Lambda}\mathbf{x}^{(n)} - \boldsymbol{\lambda}_0)^{\mathsf{T}} \boldsymbol{\Sigma}_y^{-1} (\mathbf{y}^{(n)} - \boldsymbol{\Lambda}\mathbf{x}^{(n)} - \boldsymbol{\lambda}_0) \right\} \right] \cdot \\
&\quad \exp\left\{ -\frac{1}{2}\left[\mathbf{x}^{\mathsf{T}} (\widetilde{\mathbf{K}}_{MN}^{\mathsf{T}} \widetilde{\mathbf{K}}_{M}^{-1} \widetilde{\mathbf{K}}_{MN} + \widetilde{\mathbf{V}} + \boldsymbol{\Sigma}_x \otimes \mathbf{I}_N)^{-1} \mathbf{x} \right] \right\} \qquad \text{by (A.7)} \\
&\propto \prod_{n=1}^{N} \exp\left\{ -\frac{1}{2}\left[\mathbf{x}^{(n)} - (\boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{\Sigma}_y^{-1} (\mathbf{y}^{(n)} - \boldsymbol{\lambda}_0) \right]^{\mathsf{T}} \cdot \right. \\
&\quad \left. (\boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\Lambda}) \left[\mathbf{x}^{(n)} - (\boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{\Sigma}_y^{-1} (\mathbf{y}^{(n)} - \boldsymbol{\lambda}_0) \right] \right\} \cdot \\
&\quad \exp\left\{ -\frac{1}{2}\left[\mathbf{x}^{\mathsf{T}} (\widetilde{\mathbf{K}}_{MN}^{\mathsf{T}} \widetilde{\mathbf{K}}_{M}^{-1} \widetilde{\mathbf{K}}_{MN} + \widetilde{\mathbf{V}} + \boldsymbol{\Sigma}_x \otimes \mathbf{I}_N)^{-1} \mathbf{x} \right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2}\left[(\mathbf{x} - \boldsymbol{\Sigma}_1 \mathbf{m}_{x;stack})^{\mathsf{T}} \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\Sigma}_1 \mathbf{m}_{x;stack}) \right] \right\} \cdot \\
&\quad \exp\left\{ -\frac{1}{2}(\mathbf{x}^{\mathsf{T}} \boldsymbol{\Sigma}_0^{-1} \mathbf{x}) \right\} \\
&\propto \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{x,post})^{\mathsf{T}} \boldsymbol{\Sigma}_{x,post}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{x,post}) \right\} \qquad \text{by (A.5)}
\end{aligned}$$

$\boldsymbol{\Sigma}_{x,post}$ can be further written as the sum of two terms for computation.

$$\begin{aligned}
\boldsymbol{\Sigma}_{x,post}^{-1} &= \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1} \\
&= \left[\widetilde{\mathbf{K}}_{MN}^{\mathsf{T}} \widetilde{\mathbf{K}}_{M}^{-1} \widetilde{\mathbf{K}}_{MN} + (\widetilde{\mathbf{V}} + \boldsymbol{\Sigma}_x \otimes \mathbf{I}_N) \right]^{-1} + \boldsymbol{\Sigma}_1^{-1} \\
&= (\mathbf{D}\mathbf{B}\mathbf{D}^{\mathsf{T}} + \mathbf{A})^{-1} + \boldsymbol{\Sigma}_1^{-1} \\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{D}(\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1} + \boldsymbol{\Sigma}_1^{-1} \qquad \text{by (A.1)} \\
&= (\mathbf{A}^{-1} + \boldsymbol{\Sigma}_1^{-1}) - \mathbf{A}^{-1}\mathbf{D}(\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1} \\
&= \mathbf{A}_1 + \mathbf{D}_1\mathbf{B}_1\mathbf{D}_1^{\mathsf{T}} \\
\Longrightarrow \boldsymbol{\Sigma}_{x,post} &= \mathbf{A}_1^{-1} - \mathbf{A}_1^{-1}\mathbf{D}_1(\mathbf{B}_1^{-1} + \mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1}\mathbf{D}_1)^{-1}\mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1} \qquad \text{by (A.1)} \\
&= \mathbf{A}_1^{-1} + \mathbf{A}_1^{-1}\mathbf{D}_1(\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D} - \mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1}\mathbf{D}_1)^{-1}\mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1} \\
&= \mathbf{A}_1^{-1} + \mathbf{A}_1^{-1}\mathbf{D}_1\mathbf{C}_1^{-1}\mathbf{D}_1^{\mathsf{T}}\mathbf{A}_1^{-1}
\end{aligned}$$

In practical sampling scheme, it is necessary to ensure the positive-definiteness of $\mathbf{A}_1$ and $\mathbf{C}_1$ when one uses Cholesky decomposition. It is obvious for the first; but for the second term, it is the same to prove $\mathbf{B}^{-1} + \mathbf{D}^\mathsf{T}\mathbf{A}^{-1}\mathbf{D} - \mathbf{D}_1^\mathsf{T}\mathbf{A}_1^{-1}\mathbf{D}_1$ (namely, $\mathbf{C}_1$) is a positive-definite matrix as well, the proof is as follows:

for any non-zero $\mathbf{s}$, if $\mathbf{s}$ is a vector such that $\mathbf{s}^\mathsf{T}\mathbf{D} = \mathbf{0}$, then

$$\mathbf{s}[\mathbf{B}^{-1} + \mathbf{D}^\mathsf{T}\mathbf{A}^{-1}\mathbf{D} - \mathbf{D}_1^\mathsf{T}\mathbf{A}_1^{-1}\mathbf{D}_1]\mathbf{s} = \mathbf{s}^\mathsf{T}\mathbf{B}^{-1}\mathbf{s} > 0.$$

Here the positivity is from the positive-definiteness of $\mathbf{B}$ (namely, $\widetilde{\mathbf{K}}_M$).

For any non-zero $\mathbf{s}$, if $\mathbf{s}$ is a vector such that $\mathbf{s}^\mathsf{T}\mathbf{D} \neq \mathbf{0}$, then

$\mathbf{s}^\mathsf{T}[\mathbf{B}^{-1} + \mathbf{D}^\mathsf{T}\mathbf{A}^{-1}\mathbf{D} - \mathbf{D}_1^\mathsf{T}\mathbf{A}_1^{-1}\mathbf{D}_1]\mathbf{s}$

$$\begin{aligned}
&= \mathbf{s}^\mathsf{T}\mathbf{B}^{-1}\mathbf{s} + \mathbf{s}^\mathsf{T}\big[\mathbf{D}^\mathsf{T}(\mathbf{A}^{-1} - \mathbf{A}^{-\mathsf{T}}\mathbf{A}_1^{-1}\mathbf{A})\mathbf{D}\big]\mathbf{s} \\
&= \mathbf{s}^\mathsf{T}\mathbf{B}^{-1}\mathbf{s} + \mathbf{s}^\mathsf{T}\Big\{\mathbf{D}^\mathsf{T}\big[\mathbf{A}^{-1} - \mathbf{A}^{-\mathsf{T}}(\mathbf{A}^{-1} + \mathbf{\Sigma}_1^{-1})^{-1}\mathbf{A}\big]\mathbf{D}\Big\}\mathbf{s} \\
&= \mathbf{s}^\mathsf{T}\mathbf{B}^{-1}\mathbf{s} + \mathbf{s}_D^\mathsf{T}(\mathbf{A} + \mathbf{\Sigma}_1)^{-1}\mathbf{s}_D > 0, \qquad \text{by (A.1)}
\end{aligned}$$

where $\mathbf{s}_D = \mathbf{D}\mathbf{s}$ and $\mathbf{B}$ and $\mathbf{A} + \mathbf{\Sigma}_1$ allow $\mathbf{C}_1$ to achieve positive-definiteness.

A.4.2 Latent functions

Let $\mathbf{F} = [\mathbf{f}^{(1)}, \dots, \mathbf{f}^{(N)}]$, $\mathbf{f}$ is the column-wise rearrangement of $\mathbf{F}$. Then the full conditional is :

$$\begin{aligned}
p(\mathbf{F}|e.e. \setminus \bar{\mathbf{f}}) &\propto p(\mathbf{X}|\mathbf{F}, \mathbf{\Sigma}_x)p(\mathbf{F}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_{1:Q}^{1:M}, \mathbf{\Theta}_h) \\
&\propto \exp\Big\{ -\frac{1}{2}(\mathbf{x} - \mathbf{f})^\mathsf{T}(\mathbf{\Sigma}_x \otimes \mathbf{I}_N)^{-1}(\mathbf{x} - \mathbf{f})\Big\} \cdot \\
&\quad \exp\Big\{ -\frac{1}{2}\Big[\mathbf{f}^\mathsf{T}(\widetilde{\mathbf{K}}_{MN}^\mathsf{T}\widetilde{\mathbf{K}}_M^{-1}\widetilde{\mathbf{K}}_{MN} + \widetilde{\mathbf{V}})^{-1}\mathbf{f}\Big]\Big\} \qquad \text{by (A.7)} \\
&\propto \exp\Big\{ -\frac{1}{2}(\mathbf{f} - \mathbf{x})^\mathsf{T}\mathbf{\Sigma}_3^{-1}(\mathbf{f} - \mathbf{x})\Big\} \cdot \exp\Big\{ -\frac{1}{2}(\mathbf{f}^\mathsf{T}\mathbf{\Sigma}_2^{-1}\mathbf{f})\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}(\mathbf{f} - \boldsymbol{\mu}_{f,post})^\mathsf{T}\mathbf{\Sigma}_{f,post}^{-1}(\mathbf{f} - \boldsymbol{\mu}_{f,post})\Big\} \qquad \text{by (A.5)}
\end{aligned}$$

Analogous to the sampling procedure of latent variables, $\mathbf{\Sigma}_{f,post}$ can be decomposed into two terms.

$$\begin{aligned}
\mathbf{\Sigma}_{f,post}^{-1} &= \mathbf{\Sigma}_2^{-1} + \mathbf{\Sigma}_3^{-1} \\
\Longrightarrow \mathbf{\Sigma}_{f,post} &= (\mathbf{\Sigma}_3^{-1} + \mathbf{\Sigma}_2^{-1})^{-1} \\
&= \mathbf{\Sigma}_3 - \mathbf{\Sigma}_3(\mathbf{\Sigma}_2 + \mathbf{\Sigma}_3)^{-1}\mathbf{\Sigma}_3 && \text{by (A.1)} \\
&= \mathbf{\Sigma}_3 - \mathbf{\Sigma}_3(\widetilde{\mathbf{V}} + \widetilde{\mathbf{K}}_{MN}^{\mathsf{T}}\widetilde{\mathbf{K}}_M^{-1}\widetilde{\mathbf{K}}_{MN} + \mathbf{\Sigma}_3)^{-1}\mathbf{\Sigma}_3 \\
&= \mathbf{\Sigma}_3 - \mathbf{\Sigma}_3\Big[(\widetilde{\mathbf{V}} + \mathbf{\Sigma}_3) + \widetilde{\mathbf{K}}_{MN}^{\mathsf{T}}\widetilde{\mathbf{K}}_M^{-1}\widetilde{\mathbf{K}}_{MN}\Big]^{-1}\mathbf{\Sigma}_3 \\
&= \mathbf{\Sigma}_3 - \mathbf{\Sigma}_3(\mathbf{A} + \mathbf{DBD}^{\mathsf{T}})^{-1}\mathbf{\Sigma}_3 \\
&= \mathbf{\Sigma}_3 - \mathbf{\Sigma}_3\Big[\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{D}(\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\Big]\mathbf{\Sigma}_3 && \text{by (A.1)} \\
&= (\mathbf{\Sigma}_3 - \mathbf{\Sigma}_3\mathbf{A}^{-1}\mathbf{\Sigma}_3) + \mathbf{\Sigma}_3\mathbf{A}^{-1}\mathbf{D}(\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{\Sigma}_3 \\
&= \mathbf{A}_2 + \mathbf{\Sigma}_3\mathbf{D}_2\mathbf{C}_2^{-1}\mathbf{D}_2^{\mathsf{T}}\mathbf{\Sigma}_3
\end{aligned}$$

$\mathbf{A}_2$ and $\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D}$ (namely, $\mathbf{C}_2$) are positive-definite matrices. The proofs are straightforward. For the former, the proof is:

$$\begin{aligned}
\mathbf{A}_2 &= \mathbf{\Sigma}_3(\mathbf{\Sigma}_3^{-1} - \mathbf{A}^{-1})\mathbf{\Sigma}_3 \\
&= \mathbf{\Sigma}_3\Big\{\mathbf{\Sigma}_3^{-1} - \Big[\mathbf{\Sigma}_3^{-1} - \mathbf{\Sigma}_3^{-1}(\widetilde{\mathbf{V}}^{-1} + \mathbf{\Sigma}_3^{-1})^{-1}\mathbf{\Sigma}_3^{-1}\Big]\Big\}\mathbf{\Sigma}_3 && \text{by (A.1)} \\
&= (\widetilde{\mathbf{V}}^{-1} + \mathbf{\Sigma}_3^{-1})^{-1}.
\end{aligned}$$

As a result, positive-definiteness is achieved by the same properties $\widetilde{\mathbf{V}}$ and $\mathbf{\Sigma}_3$ have.

The proof of positive-definiteness of $\mathbf{C}_2$ is similar to that of $\mathbf{C}_1$. For any non-zero $\mathbf{s}$, if $\mathbf{s}$ is a vector such that $\mathbf{s}^{\mathsf{T}}\mathbf{D} = \mathbf{0}$, then

$$\mathbf{s}(\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D})\mathbf{s} = \mathbf{s}^{\mathsf{T}}\mathbf{B}^{-1}\mathbf{s} > 0.$$

Here the positivity is from the positive-definiteness of $\mathbf{B}$ (namely, $\widetilde{\mathbf{K}}_M$).

For any non-zero $\mathbf{s}$, if $\mathbf{s}$ is a vector such that $\mathbf{s}^{\mathsf{T}}\mathbf{D} \neq \mathbf{0}$, then

$$\begin{aligned}
\mathbf{s}^{\mathsf{T}}[\mathbf{B}^{-1} + \mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D}]\mathbf{s} &= \mathbf{s}^{\mathsf{T}}\mathbf{B}^{-1}\mathbf{s} + \mathbf{s}^{\mathsf{T}}\big(\mathbf{D}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{D}\big)\mathbf{s} \\
&= \mathbf{s}^{\mathsf{T}}\mathbf{B}^{-1}\mathbf{s} + \mathbf{s}_D^{\mathsf{T}}\mathbf{A}\mathbf{s}_D > 0,
\end{aligned}$$

where $\mathbf{s}_D = \mathbf{Ds}$, $\mathbf{B}$ and $\mathbf{A}$ (namely, $\widetilde{\mathbf{V}} + \mathbf{\Sigma}_3$) are positive-definite.

A.4.3 Parameter Expansion

The full conditional of parameter expansion α^2 is

$$\begin{aligned} p(\alpha^2|e.e.) &\propto p(\mathbf{W}|\mathbf{z}^{1:N}, \bar{\mathbf{z}}_{1:Q}^{1:M}, \mathbf{\Sigma}_x, \mathbf{\Theta}_h, \alpha) \cdot p(\alpha^2) \\ &\propto |\alpha^2 \cdot \mathbf{\Sigma}_0|^{-1/2} \exp\big[-\frac{1}{2}\mathbf{w}^\mathsf{T}(\mathbf{\Sigma}_0^{-1}/\alpha^2)\mathbf{w}\big] \cdot (\alpha^2)^{-a_1-1}\exp(-b_1/\alpha^2) \\ &\qquad\qquad\qquad\qquad \text{by } \mathbf{w} = \alpha\mathbf{x} \text{ and (A.7)} \\ &\propto (\alpha^2)^{-NQ/2} \exp\big[-\frac{1}{2}\mathbf{w}^\mathsf{T}(\mathbf{\Sigma}_0^{-1}/\alpha^2)\mathbf{w}\big] \cdot (\alpha^2)^{-a_1-1}\exp(-b_1/\alpha^2) \\ &= (\alpha^2)^{-(a_1+NQ/2)-1} \exp\big[-\frac{1}{\alpha^2}(b_1 + \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{\Sigma}_0^{-1}\mathbf{w})\big] \end{aligned}$$

A.5 Specification of the prior distributions

Pseudo inputs

$$\bar{\mathbf{z}}_q^{1:M} \sim \mathcal{U}(\mathscr{Z}),$$

where $\mathcal{U}$ denotes a uniform distribution, and $\mathscr{Z}$ is the collection of all pseudo input sets (with the size of M) selected from the original input set $\mathbf{z}^{1;N}$.

Hyper-parameters

$$\theta_{h,qj} \sim \frac{1}{2}\mathcal{G}(1, 20) + \frac{1}{2}\mathcal{G}(10, 10).$$

Pseudo functions

$$\bar{\mathbf{f}}_q|\bar{\mathbf{z}}^{1:M}, \boldsymbol{\theta}_{h;q} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{q;M}).$$

Latent functions

$$\mathbf{f}_q|\mathbf{z}^{1:N}, \boldsymbol{\theta}_{h;q} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{q;N}).$$

Factor loadings and intercepts

$$\boldsymbol{\lambda}_r| \sim \mathcal{N}(\mathbf{0}, \sigma_\lambda^2 \cdot \mathbf{I}_{|\mathcal{P}_r|}).$$

Measurement error variances

$$\sigma_{y_r}^2| \sim \mathcal{IG}(a_0, b_0).$$

Covariance matrix of GP latent errors

$$\mathbf{\Sigma}_s| \sim \mathcal{IW}(2, \mathbf{I}_Q).$$

Parameter expansion

$$\alpha^2| \sim \mathcal{IG}(a_1, b_1).$$

Bibliography

Albert, J. & Chib, S. (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association* pp. 669–679.

Álvarez, M. A. & Lawrence, N. (2009), 'Sparse convolved Gaussian processes for multi-output regression', *NIPS* **21**, 57–64.

Álvarez, M. A., Luengo, D., Titsias, M. K. & Lawrence, N. D. (2011), 'Efficient multi-output Gaussian Processes through Variational Inducing Kernels.', *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics, JMLR W and CP 9* **9**, 25–32.

Arminger, G. & Muthén, B. (1998), 'A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm', *Psychometrika* pp. 217–300.

Banerjee, S., Gelfand, A. E.and Finley, A. O. & Sang, H. (2008), 'Gaussian predictive process models for large spatial data sets', *Journal of the Royal Statistical Society. Series B(Methodological)* pp. 825–848.

Barnard., J., McCulloch, R. & Meng, X.-L. (2000), 'Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage', *Statistica Sinica* pp. 1281–1311.

Bartholomew, D. & Knott, M. (1999), *Latent Variable Models and Factor Analysis.*, Arnold, 2nd Edition.

Bartholomew, D., Steele, F., Moustaki, I. & Galbraith, J. (2008), *Analysis of Multivariate Social Science Data.*, Chapman& Hall, CRC.

Blei, D. M. (2012), 'Introduction to probabilistic topic models.', *Communications of the ACM* **55**, 77–84.

Bollen, K. (1989), *Structural Equtions with Latent Variables.*, John Willey & Sons.

Bollen, K. (2006), *Latent Curve Models: A Structural Equation Perspective.*, John Willey & Sons.

Bonilla, E. V., Chai, K. M. A. & Williams, C. K. I. (2008), 'Multi-task Gaussian Process Prediction', *Advances in Neural Information Processing Systems 20* .

Booth, J. G., Hobert, J. P., & Jank, W. S. (2001), 'A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model.', *Statistical Modelling* pp. 333–349.

Boyle, P. & Frean, M. (2004), 'Dependent Gaussian processes', *Advances in Neural Information Processing Systems 17* .

Caffo, B. S., Jank, W. S. & Jones, G. L. (2005), 'Ascent-based Monte Carlo expectation maximization', *Journal of the Royal Statistical Society, Series B* pp. 235–252.

Celeux, G. & Diebolt, J. (1985), 'The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem', *Computational Statistics Quaterly* pp. 73–82.

Celeux, G. & Diebolt, J. (1992), 'A stochastic approximation type EM algorithm for the mixture problem', *Stochastics and Stochastic Reparts* pp. 119–134.

Celeux, G. & Ip, E. H. S. (1996), 'Stochastic EM: method and appliation', *In Markov Chain Monte Carlo in Practice (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter.)* pp. 260–273.

Chan, K. S. & Ledolter, J. (1995), 'Monte Carlo EM Estimation for Time Series Models Involving Counts', *Journal of the American Statistical Association* pp. 242–252.

Cheung, G. W. & Rensvold, R. B. (2002), 'Evaluating goodness-of-fit indexes for testing measurement invariance', *Structural equation modeling* pp. 233–255.

Chib, S. & Greenberg, E. (1998), 'Analysis of multivariate probit models', *Biometrika* pp. 347–361.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Journal of Machine Learning Research* pp. 273–297.

Cowles, M. K. & Carlin, B. P. (1996), 'Markov chain Monte Carlo convergence diagnostics: a comparative review', *Journal of the American Statistical Association* pp. 883–904.

Delyon, B., Lavielle, M. & Moulines, E. (1999), 'Convergence of a Stochastic Approximation version of the EM algorithm', *The Annals of statistics* pp. 94–128.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society. Series B(Methodlological)* pp. 1–38.

Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002), *Analysis of Longitudinal Data (second edition)*, Oxford: OUP.

Everitt, R. E. (2012), 'Bayesian parameter estimation for latent Markov random fields and social networks', *Journal of Computational and Graphical Statistics* **21**, 940–960.

Fahrmeir, L. & Raach, A. (2007), 'A Bayesian semiparatric latent variable model for mixed responses', *Psychometrika* pp. 327–346.

Ferrer, E., Balluerka, N. & Widaman, K. F. (2008), 'Factorial invariance and the specification of second-order latent growth models', *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* pp. 22–36.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall.

Gelman, A. & Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

Gelman, A., Huang, Z., Van Dyk, D. & Boscardin, W. J. (2005), 'Transformed and parameter-expanded Gibbs samplers for multilevel linear and generalized linear models', *Technical report, Department of Statistics, Columbia University* .

Gelman, A., Meng, X.-L. & Stern, H. (1996), 'Posterior predictive assessment of model fitness via realized discrepancies', *Statistica Sinica* pp. 733–807.

Gelman, A. & Rubin, D. B. (1992), ‘Inference from iterative simulation using multiple sequences’, *Statistical Science* pp. 457–511.

Geman, S. & Geman, D. (1984), ‘Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 721–741.

Ghahramani, Z. (2001), ‘An introductino to hidden Markov models and Bayesian networks’, *International Journal of Pattern Recognition and Artificial Intelligence* **15**, 9–42.

Gilks, W. R. (1995), *Full conditional distributions*, Chapman & Hall/CRC.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1995), *Introducing Markov chain Monte Carlo in Practice*, Chapman & Hall/CRC.

Gilks, W. R. & Roberts, G. O. (1995), *Strategies for improving MCMC*, Chapman & Hall/CRC.

Goovaerts (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University press.

Hales, J., Nevill, C., S., P. & Tipping, S. (2009), ‘Longitudinal analysis of the offending, crime and justice survey 2003-2006’, *Home Office* .

Hallerod, B. & Gustafsson, E. (2011), ‘A longitudinal analysis of the relationship between changes in socio-economic status and changes in health’, *Social Science and Medicine* **72**, 116–123.

Hastings, W. (1970), ‘Monte Carlo sampling methods using Markov chains and their applications’, *Biometrika* pp. 97–109.

Higdon, D. M. (2002), ‘Space and space-time modelling using process convolutions’, *Quantitative Methods for Current Environmental Issue* pp. 37–56.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman& Hall, CRC.

Joe, H. & Xu, J. J. (1996), ‘The estimation method of inference funcitons for margins for multivariate models’, *Technical Report 166, Dpartment of Statistics, University of British Columbia* .

Jolliffe, I. (2002), *Principal Component Analysis, 2nd Edition.*, Springer.

Lee, S. (2007), *Structural Equation Modeling: A Bayesian Approach*, John Willey & Sons.

Liu, C. H., Rubin, D. B. & Wu, Y. N. (1998), 'Parameter expansion to accelerate EM: the PX-EM algorithm', *Biometrika* pp. 755–770.

Liu, C. H. & Wu, Y. N. (1999), 'Parameter expansion for data augmentation', *Journal of the American Statistical Association* pp. 1264–1274.

Liu, J. S. (1994), 'The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem', *Journal of the American Statistical Association* pp. 958–966.

MacCallum, R. C., Kim, C., Malarkey, B. & Kiecolt-Glaser, J. (1997), 'Studying multivariate change using multilevel models and latent curve models', *Multivariate Behavioral Research* **32**, 215–253.

MacKay, D. J. C. (2003), *Information Theory, Inference, and Learning Algorithms.*, Cambridge University Press.

McCulloch, C. E. (1997), 'Maximum likelihood algorithms for generalized linear mixed models', *Journal of the American Statistical Association* **92**, 162–170.

McLachlan, G. & Krishnan, T. (2008), *The EM Algorithm and Extensions*, John Wiley & Sons.

Meredith, W. M. (1993), 'Measurement invariance, factor analysis, and factorial invariance', *Psychometrika* pp. 525–543.

Metropolis, N., Resenbluth, A. W., Marshall, M. N. & Teller, A. H. (1953), 'Equation of state calculations by fast computing machines', *The Journal of Chemical Physics* pp. 1087–1093.

Neal, R. (2010), 'MCMC using Hamiltonian dynamics'.

Neal, R. & Hinton, G. (1999), 'A view of the EM algorithm that justifies incremental, sparse, and other variants', *In Learning in Graphical Model, edited by M. I. Jordan* .

Neal, R. M. (1993), 'Probabilistic Inference Using Markov Chain Monte Carlo Methods', *Technical Report CRG-TR-93-1, , Dept. of Computer Science, University of Toronto* .

Novara, C., Pastore, M., Ghisi, M., Sica, C., Sanavio, E. & McKay, D. (2011), 'Longitudinal aspects of obsessive compulsive cognitions in a non-clinical sample: a five-year follow-up study', *Journal of Behavior Therapy and Experimental Psychiatry* pp. 317–324.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

Potthoff, R. F., Tudor, G. E., Pleper, K. S. & Hasselblad, V. (2006), 'Can one assess whether missing data are missing at random in medical studies?', *Statistical Methods in Medical Research* pp. 213–234.

Press, S. (2003), *Subjective and Objective Bayesian Statistics: Principles, Models, and Appplications*, John Willey & Sons.

Quinonero-Candela, J. & Rasmussen, C. E. (2005), 'A unifying view of sparse approximate Gaussian process regression', *Journal of Machine Learning Research* pp. 1935–1959.

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

Roberts, G. O. (1995), *Markov chain concepts related to sampling algorithm*, Chapman & Hall/CRC.

Ruschendorf, L. (2013), *Mathematical Risk Analysis.*, Springer.

Seeger, M., Williams, K. I. C. & Lawrence, N. D. (2003), 'Fast forward selection to speed up sparse Gaussian process regression', *Processings of the Ninth International Workshop on Artificial Intelligence* .

Silva, R. & Gramacy, R. (2010), 'Gaussian process structural equation models with latent variables', *Proceedings of the 26th Conference on Uncertainty on Artificial Intelligence (UAI'10)* .

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structual Equation Models*, Chapman & Hall/CRC.

Smola, A. J. & Bartlett, P. L. (2001), 'Sparse greedy Gaussian process regression', *Advances in Neural Information Processing Systems 13* pp. 619–625.

Snelson, E. & Ghahramani, Z. (2006*a*), 'Sparse Gaussian process using pseudo inputs', *Advances in Neural Information Processing Systems, 13* .

Snelson, E. & Ghahramani, Z. (2006*b*), 'Variable noise and dimensionality reduction for sparse Gaussian processes', *Uncertainty in Artifical Intelligence 22 (UAI)* .

Snelson, E. & Ghahramani, Z. (2007), 'Local and global sparse Gaussian process approximations', *Artificial Intelligence and Statistics 11 (AISTATS)* .

Song, X. Y. & Lu, Z. H. (2010), 'Semiparametric latent variable models with Bayesian P-spline', *Journal of Computational and Graphical Statistics* pp. 590–608.

Spurk, D., Abele, A. E. & J., V. (2011), 'The career satisfaction scale: longitudinal measurement invariance and letent growth analysis', *Journal of Occupational and Organization Psychology* pp. 315–326.

Steele, F. (2008), 'Multilevel models for longitudinal data', *Journal of the Royal Statistical Society Series A-Statisics in Society* pp. 5–19.

Stegle, O., Lippert, C., Mooij, J., Lawrence, N. & Borgwardt, K. (2011), 'Efficient inference in matrix-variate gaussian models with iid observation noise', *Advances in Neural Information Processing Systems 24* pp. 630–638.

Stoel, R. D., Wittenboer, V. D. & Hox, J. J. (2003), 'Analyzing longitudinal data using multilevel regression and latent growth curve analysis', *Metodologia de las Ciencias del Comportamiento* .

Talhouk, A., Doucet, A. & Murphy, K. (2012), 'Efficient Bayesian Inference for Multivariate Probit Models with Sparse Inverse Correlation Matrices', *Journal of Computational and Graphical Statistics* pp. 739–757.

Tanner, M. A. & Wong, W. H. (1987), 'The calculation of posterior distributions by data augumentation', *Journal of the American Statistical Association* pp. 528–540.

Teh, Y. M., Seeger, M. & Jordan, M. I. (2005), 'Semiparametric latent factor models', *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTSTS'05)* pp. 333–340.

Tipping, M. E. & Bishop, C. M. (1998), 'Probablistic principal component analysis', *Journal of the Royal Statistical Society. Series B(Methodlological)* pp. 611–622.

Titsias, M. (2009), 'Variational learning of inducing variables in sparse Gaussian Processes', *The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP* pp. 567–574.

Titsias, M. K. & Lawrence, N. D. (2010), 'Bayesian Gaussian process latent variable model', *The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP* pp. 844–851.

Titsias, M. K., Rattray, M. & Lawrence, N. D. (2010), 'Markov chain Monte Carlo algorithms for Gaussian processes', *Inference and Learning in Dynamic Models* .

Varin, C., Reid, N. & Firth, D. (2011), 'An overview of composite likelihood methods', *Statistica Sinica* pp. 5–42.

Wasserman, L. (2006), *All of Nonparametric Statistics.*, Springer.

Wei, G. C. G. & Tanner, M. A. (1990), 'A Monte Carlo Implementation of the EM algorithm and the Poor Man's Data Augmentation Algorithms', *Journal of the American Statistical Association* pp. 699–704.

Xia, G., Miranda, M. L. & Gelfand, A. E. (2006), 'Approximately optimal spatial design approaches for environmental health data', *Environmentrics* pp. 363–385.

Yu, B. M., Cuningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V. & Sahani, M. (2009), 'Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity', *Journal of Neurophysiology* pp. 614–635.