Probabilistic Models for Process Monitoring and Causality Analysis with Industrial Applications

by

Rahul Raveendran

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Process Control

Department of Chemical and Materials Engineering

University of Alberta

Abstract

Process monitoring involves ensuring that the process systems are run safely and operated in the most profitable manner. On the other hand, causal modelling involves studying the causal interactions among the variables in a process system. The knowledge of these interactions is useful in process monitoring, root cause analysis of process anomalies, and devising optimum operation strategies. Both these applications can greatly benefit from data-driven models when it is difficult to obtain models for the studied system based on the first principles.

In this thesis, we develop and present probabilistic models for process monitoring and causal modelling applications. The models developed in thesis enjoy an important benefit of probabilistic modelling that it allows one to define very general models that subsume several special cases. This in turn has two advantages, (i) a result derived for the general model can be reduced to special cases if required, alleviating the need to study special cases in isolation and (ii) if the special cases turn out to be different competing hypotheses about the data generating process, the users can then leverage Bayesian analysis to select between the competing hypotheses.

The probabilistic models developed for process monitoring address two extreme cases of monitoring problems, (i) monitoring unimodal systems and (ii) monitoring multi-modal systems. For monitoring unimodal systems, we define a general model that encompasses several linear Gaussian models as special cases. This allows us to develop a monitoring procedure based on the general model and reduce it to special cases if desired. In addition, we attempt to theoretically understand the connections

between the linear Gaussian models and classical multivariate techniques such as principal component analysis in the context of process monitoring. For monitoring multi-modal systems, we propose a two-layer model that consists of a convex combination of linear Gaussian models in the layers stacked one above the other. This model scales well when compared to the probabilistic models used for process monitoring in the literature to approximate non-Gaussian distributions. Furthermore, we illustrate the two-layer model for process monitoring using a lab-scale and an industrial case study.

In causal modelling, we address two important problems, (i) identification of time-lagged causal interactions in the presence of instantaneous/contemporaneous interactions among the variables and (ii) modelling long-term interactions for time-varying systems. Granger causality analysis is a most commonly used approach for studying time-lagged causal interactions. However, if the presence of contemporaneous interactions is not properly accounted for, the Granger causality analysis techniques tend to identify spurious time-lagged interactions. In this thesis, we propose a model for representing the time-lagged and contemporaneous interactions explicitly and perform Bayesian analysis to determine the presence and absence of both types of interactions. The approach is found to be more robust to the presence of contemporaneous interactions when compared to the traditional Granger causality analysis techniques. When studying the long-term effects of process variables on process performance indicators using the routine operation data from the process systems, time-varying nature of the process systems affects the correct identification of the effects. To address this problem, we propose a time-varying parameters model and a Bayesian analysis approach to recover the time-varying effects. We illustrate the causal modelling approaches developed in this thesis using industrial case studies.

Preface

This thesis is an original work conducted by Rahul Raveendran. The materials presented in this thesis resulted from the research projects conducted under the supervision of Dr. Biao Huang.

Chapters 1, 2 and 7 of this thesis was prepared by Rahul Raveendran.

Chapter 3 of this thesis has been published as "**Raveendran, R.**, Kodamana, H., & Huang, B. (2018). Process monitoring using a generalized probabilistic linear latent variable model" in Automatica, 96, 73-83.

Chapter 4 of this thesis has been published as "**Raveendran, R.**, & Huang, B. (2017). Two-layered mixture Bayesian probabilistic PCA for dynamic process monitoring" in Journal of Process Control, 57, 148-163.

Chapter 5 of this thesis has been published as "**Raveendran, R.**, & Huang, B. (2018). Variational Bayesian approach for causality and contemporaneous correlation features inference in industrial process data" in IEEE Transactions on Cybernetics, (99), 1-11.

Chapter 6 of this thesis has been submitted to IEEE Transactions on Control Systems Technology as "**Raveendran, R.**, Mitchell, W., & Huang, B. (2019). A variational Bayesian causal analysis approach for time-varying systems" and it is under review.

Rahul Raveendran was responsible for the idea development, deriving theoretical results, performing simulation studies and manuscript preparation for all the publications listed above. Dr. Biao Huang is a supervisory author of all the publications listed above. Dr. Hariprasad Kodamana helped by providing valuable and critical feedback on the work and correcting manuscript for publishing the materials in Chapter 3. Warren Mitchell played a similar role for submitting the materials in Chapter 6 for

review, as Dr. Hariprasad Kodamana did for Chapter 3.

Acknowledgements

First and foremost, I would like to take this opportunity to express my deep gratitude towards my thesis advisor Prof. Biao Huang for his guidance, motivation, giving me freedom to explore different research topics and yet making sure that I stayed focused on completing the tasks at hand in a timely manner and sparing his invaluable time reviewing my work promptly. I am also thankful to him for instilling confidence in me and entrusting me with challenging yet interesting industrial problems during my Ph. D. I am greatly indebted to Dr. Huang for his benevolent and unfailing financial support without which this work would not have been possible.

My special thanks to Dr. Hariprasad Kodamana and Warren Mitchell for their constant encouragement, critical feedback and spending their invaluable time helping me with reviewing my papers.

I am truly grateful to Spartan Controls for hosting me at their office and giving me exposure to several real-world problems. I learnt a lot from discussions and working with Warren Mitchell, Dr. Hailei Jiang, Dr. Anuj Narang, Shabnam Sedghi and others from the Advanced Process Control team at Spartan Controls, which I thoroughly enjoyed. It was great pleasure working with Agustin Vicente and Mengqi Fang from University of Alberta, who were also hosted by Spartan Controls at their office. I am also thankful to Eric Lau, Eliyya Shukeir, Dr. Fei Qi and Seraphina Kwak from Suncor Energy for again exposing me to challenging real-world industrial problems.

I would like to gratefully acknowledge the Department of Chemical Engineering, University of Alberta for giving me the opportunity to pursue my Doctoral degree and providing me with the best of the facilities and resources. Special thanks to Prof. Ken Cadian and Prof. Vinay Prasad, for giving me an opportunity to co-teach CH E 472 with Ajay Ganesh. I thoroughly enjoyed teaching and working with Ajay.

I would like to acknowledge the financial support from Natural Sciences and Engineering Research Council (NSERC) of Canada.

This work would not have been possible also without the help of several members of the Computer Process Control group. I would like to acknowledge the help given by Dr. Nabil Magbool Jan, Yanjun Ma, Dr. Fadi Ibrahim, Terry Runyon, Dr. Ruben Gonzalez and several other past and present members from the group.

Arnab, Richa, Anupam, Ashwin, Sanat, Gokul, Rishik, Nabil, Geetesh, Gail, Wesley had been great room-mates who took special care of me and made my stay in Edmonton enjoyable. My journey through graduate studies would have felt long and arduous without the great company of my friends, Shruti, Sushmitha, Vishal, Arun, Rishik, Shekar, Sahil and others.

Last but not the least, I would like to express my deepest gratitude towards my parents, my sister and her family for their unconditional love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistical Process Monitoring and Causality Analysis

Today, process industries as a major part of their digital transformation strategy strive to leverage cloud storage and analytics platforms to store and perform analytics on their operational data. Industries are on a constant lookout for use cases for their operational data to derive business value out of it. This thesis develops and presents new data-driven models for two important uses cases for the operational data namely, (i) statistical process monitoring (SPM) and (ii) causality analysis.

1.1.1 Statistical Process Monitoring

SPM involves defining statistical control or an operating limits for the measured process variables or the features derived from the process variables and ensuring that the variables and the features do not violate the defined limits. SPM techniques can benefit from stored or archived operational data for deriving these control limits. Consider that the stored operational data consists of data during the periods when the process operations were more cost-effective and energy efficient or when the process and process equipment were operated more reliably. Data-driven models can be used to identify these sweet spots in the data and the operating envelope for the process variables during these periods. Once the operating envelope is identified, the SPM techniques can be deployed online to monitor the process and alert the operations team when the process drifts away from these sweet spots. The operations team can

then take necessary actions to bring the process back to the desired operating conditions. Thus, SPM helps the operations team to (i) ensure safe operation of process and manufacturing facilities, (ii) reduce anomalies in product quality, (iii) reduce unprecedented production downtime, (iv) strictly meet emission standards and (v) ensure that the process and process equipment are operated reliably. SPM can be applied to process systems at many levels as discussed below,

1. Monitoring of quality variables or key performance indicators (KPIs): Performance of a process unit or a facility is often quantified by means of a manageable number of KPIs, which can be easily measure or estimated and monitored. The KPIs typically include production rates, cost of production and processing, energy efficiency, emission levels, production downtime, etc. SPM allows anomalies in the KPIs to be detected, addressed and eventually, the rate of anomalies to be decreased.

2. Monitoring of features generated from the intelligent control systems: Modern industrial control systems may possess several interconnected layers that make use of process measurements and interact with process systems to ensure that the systems perform well to meet targets set on the aforementioned KPIs. Data acquisition, inferential or soft sensors, regulatory control and supervisory control are some of the commonly found layers in the industrial process control systems and more sophisticated systems may also consist of fault detection, fault diagnosis, data reconciliation and equipment health monitoring layers, etc. These layers generate features that are representative of the performance of the process systems and may also generate features representative of their own performances. For example, controller error in the regulatory control layer, prediction error in the inferential sensors and fault signatures generated by the fault detection, diagnosis and equipment health monitoring layers. These layers can typically benefit from SPM techniques to monitor the generated features.

3. Monitoring of measured process variables: In addition to monitoring the KPIs and the features derived from the measured process variables, the process variables can themselves be monitored using the SPM techniques. In this case, SPM

techniques indirectly ensure that the KPIs and the features are within the desired limits by ensuring that the variables used to derive them are within the desired limits. This approach also helps the operations team and the control systems recognize when some of the process variables are approaching their safety limits, sensor failures when the measurements are out of the reasonable range, etc.

1.1.2 Causality Analysis

When it becomes challenging to understand and determine how the process variables interact with each other from the first principles knowledge of the process, data-driven causal modelling techniques prove to be vital. Causal models may allow us to reconstruct the process networks or identify the strengths of different interactions in the postulated process networks from the operational data. The knowledge of these causal interactions can help improve process operations in the following ways at the very least,

1. Assume that we can postulate or identify a reasonable network that encodes how the process variables affect the key performance indicators (KPIs) of a process. Quantifying the strengths of interactions in this network can give us a perspective on relative effect of each process variable on the KPIs. This information can further be utilized to optimize the process KPIs in multiple ways, for example, (i) the influential variables can be controlled such that they affect the KPIs in a desired manner, (ii) closed-loop control or optimization frameworks for the KPIs can be designed with the most influential variables as the manipulated variables, etc.

2. Consider an occurrence of an event where the process drifts away from the sweet spots mentioned earlier. The knowledge of where the problem originated and how it eventually led the process to drift away are imperative for the operations team to take the right sequence of actions to bring the process back to the normal operation. An approach that classifies the process variables as causes or effects or both or neither from the data during these upsets would provide

a starting point for the operations team to investigate, identify the root cause and speed up the recovery process.

3. A causal hierarchical network that assigns different hierarchies to the process variables based on their cause and effect relationships can give rise to a systematic statistical process monitoring strategy. Process variables can then be monitored from the top to the bottom of the hierarchical tree. This strategy allows the process abnormalities to be detected during its initial stages from the top layer variables in the tree before they propagate to the bottom levels of the tree.

Statistical process monitoring and causality analysis for applications in industrial process systems are among the active academic research areas. In the following sections, we review the techniques that are relevant to the models developed in this thesis.

The remainder of this chapter is organized as follows: In sections 1.2 and 1.3, we review the existing literature relevant to the models developed in this thesis and motivate the need for the developed models. In section 1.4, we provide the outline of this thesis. In section 1.5, we present the summary contributions made in this thesis.

1.2 Statistical Process Monitoring Techniques

There exist numerous techniques for statistical process monitoring in the literature, especially, the ones based on the data-driven models borrowed from chemometrics, statistics and different areas of machine learning. It would be incredibly challenging to review the existing techniques comprehensively and yet keep the discussion focused on the models developed in this thesis. For the readers who are interested in different available techniques, we refer them to more comprehensive texts and review articles [1, 2, 3, 4, 5, 6]. Instead, in this section, we stick to the techniques that are relevant to the development of the models presented in this thesis.

1.2.1 Univariate and Multivariate Control Charts

Univariate control charts (or Shewhart charts[7]) are the earliest of the SPM techniques. They are typically used to monitor the sample means or direct measurements of the quality variables. A simplest form of univariate charts relies on defining upper and lower control limits around the expected value of the monitored statistic and detecting the anomalies that lie outside the control limits as illustrated in Fig. 1.1. The control limits are derived either from process expertise or based on the cumulative distribution function (CDF) of the monitored statistic with a desired rejection rate. The rejection rate determines the theoretical value of the percentage of false detections when all the samples generated from the process follow the assumed null distribution. It is usual to assume the monitored variables are Gaussian distributed and determine the control limits based on their variance or standard deviation. When their variance is unknown, it is estimated from the data collected from the periods when the process is considered to be normal and the control limits are derived from the CDF of Student's t-distribution. Other common variants or relatively more advanced versions of univariate charts monitor exponentially weighted moving average (EWMA) [8, 9] or cumulative sum (CUMSUM) [10] of the quality variable. The EWMA and CUMSUM charts tend to be more sensitive and can be made to detect even small changes or gradual drifts in the sample mean.

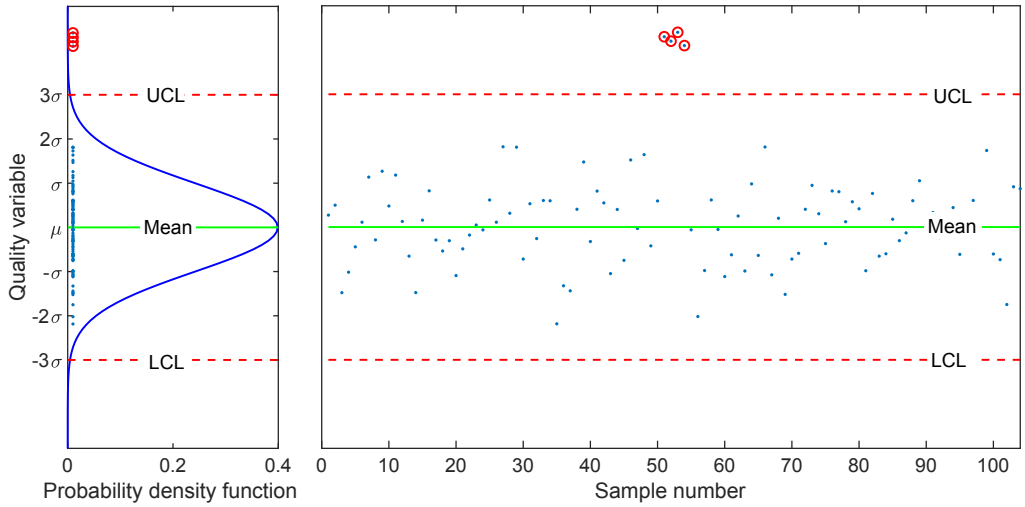


Figure 1.1: Example of a univariate control chart. Green solid line corresponds to the expected value of the process, red dashed lines correspond to the upper and lower control limits and the data points highlighted by the red circles are the anomalies detected by the chart.

When a large number of variables are monitored and the monitored variables are correlated, univariate control charts become ineffective as illustrated in Fig. 1.2.1. Instead, multivariate control charts can be utilized. Multivariate control charts convert multivariate variables into univariate or bivariate statistics that can then be managed and monitored with one or two control charts. A basic version [11] of multivariate charts involves monitoring the Hotelling's T^2 statistic [12], which is given by the covariance normalized quadratic distance between the actual observation and the expected values of the monitored variables. It defines an elliptical or a hyper-elliptical control limit around the expected values of the monitored variables depending upon the dimension of the monitored variable. When the covariance matrix of the monitored variables is available or estimated from a large number of samples, the control limits are derived from the CDF of χ^2 distribution. When it is estimated from a lesser number of samples, the control limits are derived from the CDF of Hotelling's T^2 distribution. The CUMSUM [13, 14, 15] or EMWA [16] version of multivariate charts are also available for process monitoring in the literature.

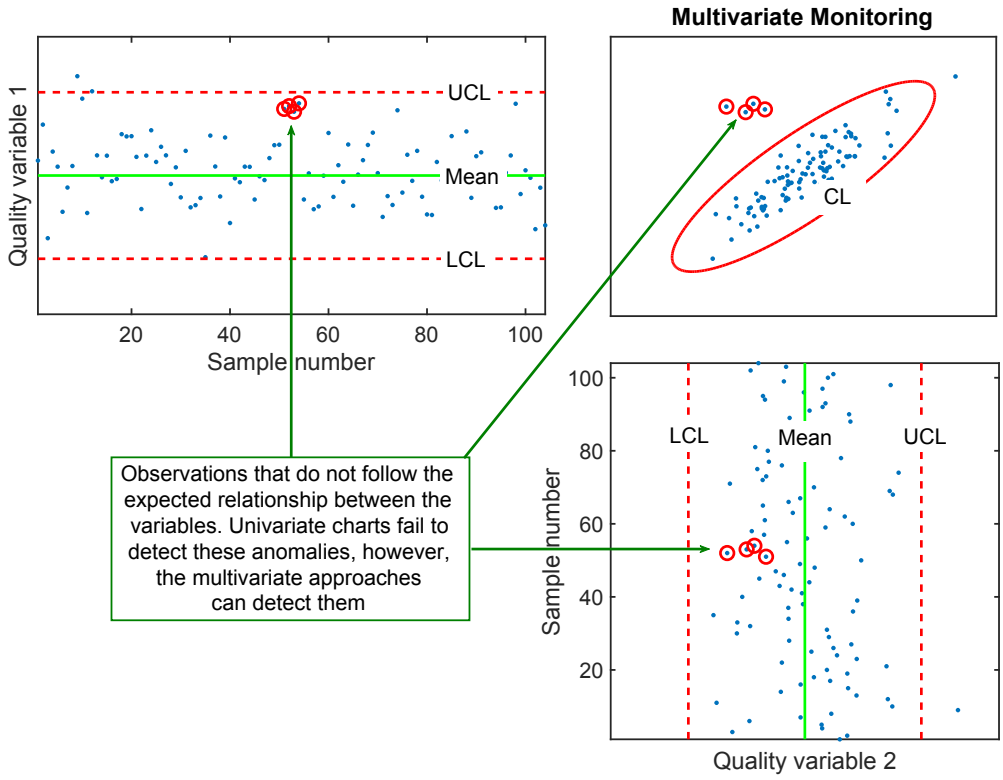


Figure 1.2: Multivariate monitoring approach.

1.2.2 Process Monitoring using Latent Variable Models

Latent variable models describe the observed variables as a function of the lower dimension latent or hidden variables. The popular multivariate latent variable techniques used for monitoring include principal component analysis (PCA), factor analysis (FA), partial least squares (PLS) method and canonical correlation analysis (CCA) [17, 18, 19, 5, 20, 21, 22, 23, 24, 3, 5, 25].

PCA is probably the most popular classical multivariate technique used for process monitoring applications in the literature. The PCA can be performed by subjecting the sample covariance matrix to eigendecomposition. From eigendecomposition, two sets of axes from the orthonormal basis that explains the spread of the observations can be obtained, namely, (i) principal components (PCs) and (ii) minor components (MCs) as illustrated in Fig. 1.3. Each axis in the basis explains a fraction of variance in the total variance of the observed data. The axes in the basis can be ordered from the

one that explains the maximum variance to the one that explains the minimum variance. The PCs explain the systematic variance in the observed variables. Projection of the actual observation on to the space spanned by the PCs gives the underlying latent variables. The MCs explain the measurement noise variance or the residual variance of the PCA model. For example, in the 2-dimension (2-D) observation case illustrated in Fig. 1.3, y^1 and y^2 represent the axes of the actual observed variables and PC and MC are axes in the orthogonal basis identified by the application of PCA. In this case, the axis PC explains the systematic variance in y^1 and y^2 and MC explains the variance of the residuals in the relationship between y^1 and y^2 identified by the PCA model.

PCA has certain advantages over the multivariate charts that we reviewed before including, (i) it models the covariance of the variables parsimoniously depending on the number of axes retained as PCs, (ii) control charts for Hotelling's T^2 statistic can be established when the sample covariance matrix is ill-conditioned, which is not possible with the previously described methods as they need the inverse of the sample covariance matrix, (iii) it can distinguish between the systematic variance and variance due to measurement noise when the measurement noise of the monitored variables have the same magnitude of variance and (iv) using PCA, high dimension variables can be mapped onto the lower dimension latent variables space, which in many cases aids data visualization.

Typically, two different control charts are used in the PCA based monitoring approach, (i) T^2 chart[12] and (iii) SPE or Q chart[26]. T^2 chart is used to monitor the latent variables. The latent variables extracted from the PCA are uncorrelated and have zero mean and a full rank diagonal covariance, which allows the implementation of T^2 chart. SPE chart monitors the space of model residuals. The T^2 chart can detect the instances when the latent variables drift far away from the origin and the SPE chart can detect the instances when the model residuals are out of the desired bounds as illustrated in Fig. 1.3.

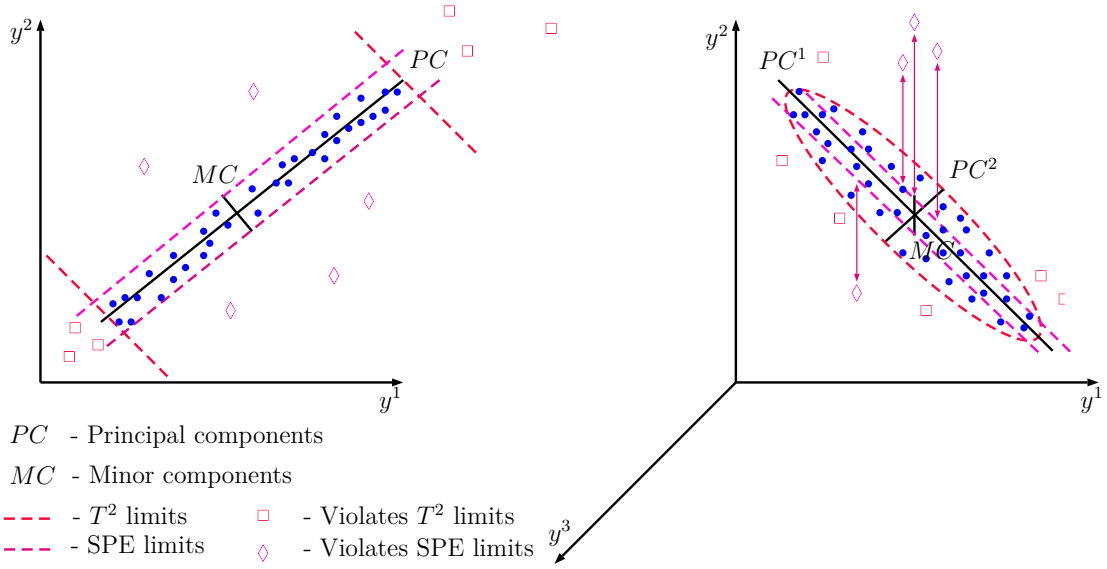


Figure 1.3: Illustrations of the PCA based monitoring approach. Left: Monitoring model from PCA with 1-D latent variable for monitoring 2-D observed variables. Right: Monitoring model from PCA with 2-D latent variables for monitoring 3-D observed variables.

The FA based approach is very similar to the PCA based approach, except that it allows monitored variables to have different magnitudes of noise variance. The PLS and CCA techniques are used when the observed variables can be split into two distinct subsets (Eg. inputs and outputs). They can provide dimension reduction on both the subsets. CCA extracts latent variables from both the subsets such that they are maximally correlated while PLS method maximizes the covariance between the latent variables obtained from both the subsets [27]. For process monitoring applications, in all these approaches, the extracted lower dimension variables and the model residuals or the reconstruction residuals of the original observations from the model are monitored. These techniques can also be used for process monitoring when the observations are serially/temporally correlated. In such cases, the usual trick is to treat a sequence of observation as a data point instead of treating every single observation as a data point. The resulting models are called the dynamic PCA model, dynamic CCA model and so on [18, 19].

1.2.3 Process Monitoring using Probabilistic Latent Variable Models

Classical latent variable techniques discussed above also have their probabilistic counterparts. For instance, probabilistic PCA (PPCA) [28] and probabilistic CCA (PCCA) [29] models are the probabilistic counterparts of PCA and CCA, respectively. Probabilistic models define a distribution over the observed variables. In addition to that, by the use of the rules of probability theory they allow one to assess all forms of uncertainties in the model including the uncertainty in the extracted latent variables, parameters and model structure.

Let us take an example of a probabilistic model, the PPCA model. It assumes that the observed variables ($y_n \in \mathbb{R}^D$) are generated by a linear combination of the lower dimension latent variables ($x_n \in \mathbb{R}^K$) and corrupted by the additive noise ($e_n \in \mathbb{R}^D$) as shown below,

$$\begin{aligned} y_n &= \mu_y + W z_n + e_n \\ z_n &\overset{i.i.d}{\sim} \mathcal{N}(0, I_K) \\ e_n &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma I_D) \end{aligned} \tag{1.1}$$

where W is a matrix of coefficients of the latent variables and μ_y is the mean of the observations. The PPCA model assigns a distribution (prior distribution) over the latent variable and the measurement noise, thereby, explicitly defining a distribution over the observed variables. The latent variable are considered to follow a multivariate Gaussian distribution with zero mean and identity covariance of size K and the measurement noise are considered to follow a multivariate Gaussian distribution with zero mean and diagonal covariance with all its diagonal elements equal to σ. The model allows the inference of the conditional distribution (posterior distribution) of the latent variables given the observations. Thus, one can assess the uncertainties in the extracted latent variables. Similarly, if one wishes to assess the uncertainties in the parameters of the model, W, μ_y and σ, one can also prior distributions over these parameters and infer their posteriors by the use of rules of probabilistic inference. We will return back to this topic in chapter 2 of this thesis.

Briefly, the advantages of the probabilistic modelling include: 1) They explicitly define the modelling assumptions to describe the data generation process, 2) they provide feasible frameworks to accommodate different distribution assumptions to

handle specific data characteristics, for instance, outliers [30], multi-modality [31], and missing data [32] and 3) they allow users to incorporate prior distributions for the parameters and perform Bayesian analysis [33], which can be used to select between multiple competing models that best describe the observed data. The probabilistic latent variable models have also been shown to be applicable for process monitoring in the literature by leveraging the above mentioned advantages [30, 34, 35, 36, 37, 38].

1.2.4 Process Monitoring Problems Addressed in this Thesis

This thesis addresses the following two problems in process monitoring using the probabilistic models,

1. Classical multivariate techniques and their probabilistic counterparts have been compared for process monitoring applications in the literature using simulation case studies. For example, comparison of PCA and PPCA for process monitoring [37]. However, there exists no standard or rigorous procedure that includes monitoring statistics, their null distributions and control limits for deploying probabilistic models for process monitoring unlike their classical counterparts. In addition, questions such as 'is there any advantage or difference in using probabilistic models when compared to using classical counterparts for process monitoring?' has not been answered previously. In this thesis, we address these problems for linear Gaussian models. We define a general model such that for most of the common probabilistic models including PPCA, PCCA and probabilistic factor analyser (PFA), these problems can be addressed under a unified framework. We derive the process monitoring procedure for the general model and show that this procedure can be reduced to special cases if required. Furthermore, through the derived results, we theoretically compare the classical and probabilistic models for process monitoring.

2. For monitoring multi-modal systems, mixture models that are expressed as a convex combination of linear Gaussian models have been proposed in the literature [30, 34, 35, 36]. These models retain a structural simplicity and let the users control the model complexity based on the number of local models

considered. Increasing the number of local models in the mixture allows one to approximate complex data distributions. However, such approaches do not scale well. They suffer from local optima convergence and increased computational complexity. Instead, we propose a modelling approach inspired from multi-layer neural networks, a two-layer mixture model formed by stacking the mixture models one above the other. We find the proposed approach to be more promising and outperform single layer models in multiple fronts.

1.3 Causal Modelling Techniques

Causal modelling is a broad area of research and there exists multiple definitions of causality and numerous causal modelling approaches. In this section, we review two causal modelling approaches that are relevant to the problems addressed in this thesis, one based on static Bayesian networks and the other is based on Granger causality networks, which can be seen as a special case of dynamic Bayesian networks.

1.3.1 Static Bayesian Networks

Bayesian networks are directed acyclic graphs (DAGs). In Bayesian networks, the variables are represented as nodes and the interactions among the variables are represented by means of directed arcs. Bayesian networks are a special class of probabilistic graphical models (PGMs) that are in general used to define a join distribution over a set of variables as we will see in chapter 2. However, in the context of causal modelling, the variables or nodes in a Bayesian network can be viewed as causes and effects. In a Bayesian network that consists of variables y^1 and y^2, variable y^1 is said to be a direct cause variable to y^2 if there exists a directed arc from y^1 pointing to y^2.

Bayesian networks can be used to construct a hierarchical tree of variables as the one shown in Fig. 1.4. In the hierarchical tree, the variables in level 1 are the direct cause variables to variables in level 2, the variables in level 2 are the the direct cause variables to variables in level 3 and so forth. In this network, any systematic change will first start with the variables in level 1, then propagates to level 2, then to level 3

and so on. The joint distribution over the system of these variables can be obtained as a product of marginal distribution of variables in level 1, conditional distribution of variables at level 2 given the ones at level 1, conditional of level 3 variables given the ones at level 2, and so on until the conditional of ones at level k given the ones at level $k-1$.

Modelling the joint distribution of the process variables from process data using the hierarchical Bayesian networks that assign cause and effect roles to the process variables has been found to be attractive in process monitoring applications in the literature [39, 40]. By establishing the statistical control limits from the marginal distribution of the level 1 variables and the conditionals of the lower level variables, any undesirable systematic change in the process system can be detected and tracked from the top of the network. However, these hierarchical networks can also be applied in industrial process systems for applications other than process monitoring. One other application could be studying the strengths of causal effects of one variable on the other variable in the network or inferring the effect of fixing or changing one variable on the other, known as the interventional analysis [41, 42].

The hierarchical network identification problem has two sub objectives, (i) determining the topology of the network and (ii) identification of the conditional distributions or relationships. Though there exist algorithms, determining the topology of the network from data is a relatively harder problem [43, 44, 45]. When there is sufficient domain knowledge available, one may potentially postulate different topologies and validate them against data. In this thesis, we are however interested in the problem of learning conditional distributions or parameters in a static causal network and study the strengths of causal interactions among the variables assuming the topology is available from the domain knowledge. Even though, this is a simpler problem compared to the topology identification problem, several data quality issues associated with the process routine operation data can confound the analysis significantly. In this thesis, we are interested in addressing such issues.

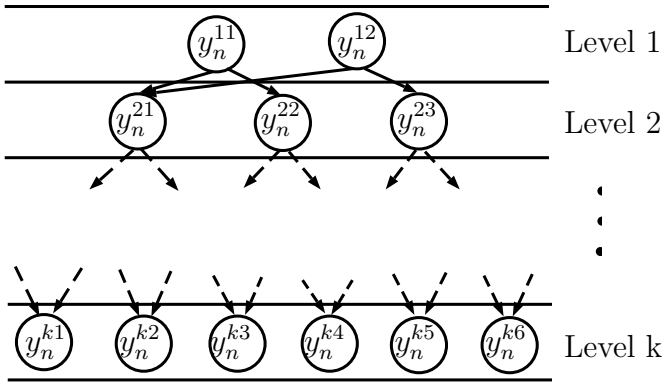


Figure 1.4: A hierarchical Bayesian network representing the interactions among the observed variables

1.3.2 Granger Causality Analysis

In dynamic systems, a notion of causality introduced by Granger [46] is widely used for defining the causal dependencies among the variables. The Granger causality is defined based on the idea that the cause improves the prediction accuracy of the effect. If a variable y^1 is said to Granger cause a variable y^2, y^2 should be predicted more accurately using the past states of both y^1 and y^2 than by using the past states of y^2 alone. The existence or absence of this causal relationship can be tested by constructing two different prediction models for y^2, one based on just the past observations of y^2 and the other based on the past observations of both y^1 and y^2. By subjecting the prediction accuracies of both the models to statistical tests to determine if the latter model really improves the prediction accuracy of y^2, one can comment on the presence or absence of the causal relationship. Typically, univariate or bivariate linear auto-regressive models are used for testing the presence of Granger causality.

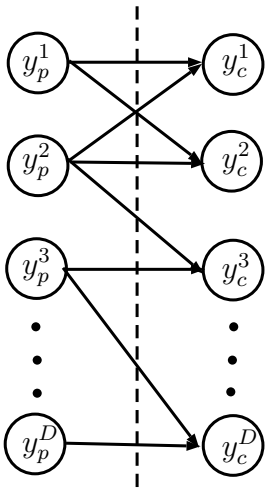
For multivariate systems, the idea of Granger causality analysis is extended to study the causal interactions with the use of linear vector auto-regressive (VAR) models [47, 48, 49, 50]. The linear VAR model for a multivariate process $y_t \in \mathbb{R}^D$ is defined as the following,

$$y_t = \sum_{l=1}^{L} W(l) y_{t-l} + e_t \tag{1.2}$$

where y_{t-l} are the lagged versions of y_t, L is the maximum lag, $W(l) \in \mathbb{R}^{D \times D}$ is a matrix of VAR model coefficients for lag l and $e_t \in \mathbb{R}^D$ is a multivariate Gaussian

white noise process. In the above model, a variable y^j (j^{th} dimension variable in y_t) is said to not directly Granger cause a variable y^i (i^{th} dimension variable in y_t) if the entry in i^{th} row and j^{th} column of $W(l)$ is effectively zero $\forall l \in [1, L]$ as in that case, the past observations of y^j do not help in predicting the future values of y^i [51]. Therefore, by testing the significance of the coefficients or by estimating the model through the penalized estimation approaches (to distinguish between zero and non-zero coefficients), one can comment on the presence or the absence of direct causal relationships between any pair of variables in the linear dynamic systems.

The resulting causal networks for this type of approach can be viewed as a special case of dynamic Bayesian networks as the one shown in Fig. 1.5. The model shown in Eqn. (1.2) defines a conditional distribution on the multivariate process given its past. This conditional distribution can be expressed in the form of Bayesian networks. In the resulting Bayesian networks, the direct causal relationships is expressed through direct arcs from the past states of the multivariate variables to the current states of those. In this representation shown in Fig. 1.5, the nodes with subscript p represent the past states of the variables and the nodes with subscript c represent the current states of the variables. In this network, a direct arc from the past states of a variable to the current state of another variable indicates the presence of a direct causal Granger causal influence from the former on the latter. For the example in Fig. 1.5, y^1 directly Granger causes y^2 but not y^3 and y^D.



y_p^D - Past states of y^D

y_c^D - Current state of y^D

Figure 1.5: Dynamic Bayesian network used to represent the interactions among the observed variables.

The Granger causality network reconstruction approaches have been found to be promising for process monitoring and data analysis applications such as root cause analysis of process abnormalities in the literature [52, 53, 54].

1.3.3 Causal Modelling Problems Addressed in this Thesis

In this thesis, we propose models to address two important issues in the causal modelling approaches discussed above,

1. The Granger causality analysis approaches only account for time-lagged causation. The existence of instantaneous/contemporaneous correlations among the variables can result in incorrect causal network reconstruction when using the traditional Granger causality analysis approaches. For instance, when identifying auto-regressive models from data, contemporaneous relationships tend to disguise themselves as time-lagged causations as shown in [55]. In process systems, it is hard to classify some of the interactions as time-lagged causations due to practical sampling and data historization rates. Slow sampling rates may not allow us to observe dead-time and dynamics in some of the interactions. Therefore, it is reasonable to expect both types of effective interactions to be present in the data. To address the problem of time-lagged causal network reconstruction in the presence of contemporaneous correlations among the variables in a linear system, we propose an approach in this thesis.

2. Both for causal network reconstruction and quantifying the interactions or learning the conditional distributions in a postulated causal network, experimental data are more suitable. However, it is expensive to conduct experiments in process systems as the experiments interfere with the routine operation. On the other hand, routine operation data are easy to obtain from the process historian. Nevertheless, one may have to be wary of data quality issues in routine operation data when they are used for causal modelling. One such issue is time-varying nature of the process systems and its effect on the observed data. For example, the relationship between a KPI of a plant and the process variables may vary with changes in the operating modes, physical plant modifications,

equipment fouling, etc. The effect of time-varying nature of the process can be severe especially when studying long-term data. In this thesis, we present a time-varying parameters model based approach that can be potentially applied to quantify long-term causal interactions in a postulated causal network.

1.4 Thesis Outline

The rest of this thesis is organized as follows,

In chapter 2, we present the preliminaries required to define and develop the estimation algorithms for a class of probabilistic graphical models known as the conjugate exponential family graphical models (CEFGMs), which is a special class of Bayesian networks. All the models developed and studied in this thesis can be shown to belong to CEFGMs. We introduce Bayesian networks and CEFGMs and illustrate how the data-driven models can be expressed graphically using the Bayesian networks. In this thesis, we utilize the expectation maximization (EM) algorithm for the maximum likelihood estimation and the variational Bayesian expectation maximization (VBEM) algorithm for approximate Bayesian analysis of the developed models. Therefore, we illustrate these algorithms using the PPCA model as a case study.

In chapter 3, we define a generalized probabilistic linear latent variable model (GPLLVM) that under specific restrictions reduces to various probabilistic linear models used for process monitoring. For the defined model, we rigorously derive the monitoring statistics and their respective null distributions. Monitoring statistics of the defined model also reduce to the monitoring statistics of various probabilistic models when restricted with the corresponding conditions. We show the equivalence between the classical multivariate techniques for process monitoring and their probabilistic counterparts, which is obtained by restricting the generalized model. We also provide an estimation approach based on the EM algorithm for the GPLLVM. The results presented in the chapter are verified using numerical simulation examples.

In chapter 4, a two-layer mixture Bayesian probabilistic principal component analyser model is developed and proposed for process monitoring. It is suitable for the data driven process monitoring applications where data with non-Gaussian distri-

bution and temporal correlated observations are encountered. Model development involves preprocessing the original observation matrix to make it suitable for building dynamic models, followed by two stages of estimation. In the first stage, the data is divided into a manageable number of clusters using a mixture model and in the second stage, a mixture model is built over each cluster. This strategy provides a scalable mixture model that is given by a convex combination of multiple local models. It has the potential to provide a parsimonious model and be less susceptible to local optima convergence compared to the existing approaches that build mixture models in a single stage. Dimension reduction during the estimation is automated using the Bayesian regularization approach. The proposed model essentially provides a probability density function for the monitored variables. It is deployed for process monitoring and the performance highlights are demonstrated in two real datasets, one is from a sulphur handling facility and the other is a publicly available experimental dataset.

In chapter 5, a hybrid model is proposed to simultaneously identify causal connections and features responsible for contemporaneous correlations in a multivariate process. The hybrid model is formed by combining the vector auto-regressive exogenous model (VARX) and the factor analysis (FA) model. The parameters of the resulting model are regularized using the hierarchical prior distributions. The model is estimated using the VBEM algorithm. The estimation is initiated with a complex model which is then systematically reduced to a simpler model that retains only the parameters corresponding to significant causal connections and contemporaneous correlations. Model reduction is carried out through a series of deterministic jumps from complex models to simpler models using a relevance criterion. The approach is illustrated with a number of simulated examples and an industrial case study using the data from the sulphur handling facility.

In chapter 6, we present a causal modelling approach for the time-varying systems. The approach relies on the time-varying parameter models (TVPMs) estimated using the VBEM algorithm. We incorporate a hypothesis switching procedure in combination with the VBEM algorithm that allows us to infer the time-varying strengths of causal effects of the inputs on the outputs of the system. We illustrate the proposed

approach using the production data from steam assisted gravity drainage (SAGD) wells. We find the time-varying model based approach to produce more consistent results across multiple case studies for the studied system as compared to the time-invariant model based approach.

In chapter 7, we conclude the thesis and present the recommendations for future research directions both in process monitoring and causal modelling.

1.5 Main Contributions

This thesis develops and presents the following four probabilistic models,

1. The generalized probabilistic linear latent variable model, which can be applied for unimodal process monitoring.

2. The two-layer mixture Bayesian probabilistic principal component analyser model, which scales well to describe the non-Gaussian data distributions and can be applied for multi-modal process monitoring.

3. A hybrid model that is formed by combining the vector auto-regressive and factor analyser models and the thesis also presents an approximate Bayesian analysis procedure for the developed model, which can be applied to study the casual and contemporaneous interactions among the variables in a linear system.

4. A time-varying parameters model and the thesis also presents an approximate Bayesian analysis procedure for the developed model, which can be applied to study the time-varying causal strengths in a postulated causal network.

Chapter 2

Preliminaries

The purpose of this chapter is to present the preliminaries required to easily follow the subsequent chapters in this thesis. All the models developed in this thesis for process monitoring and causal modelling applications can be defined using Bayesian networks (BNs), a class of probabilistic graphical models (PGMs). To be more specific, the developed models belong to the conjugate exponential family graphical models (CEFGMs), a special classs of BNs. We provide a brief introduction to BNs and CEFGMs. Then, we illustrate how the data-driven models can be defined using BNs with examples including time series models and multivariate statistical models. In addition, we take a popular example in process monitoring applications in the literature, principal component analysis (PCA) and show how a maximum likelihood (ML) and a Bayesian version of the PCA model can be defined using the BNs.

CEFGMs are amenable either to the exact maximum likelihood estimation through approaches such as the gradient descent algorithms and the expectation maximization (EM) algorithm [56, 57] or to approximate maximum likelihood estimation through algorithms such as the variational expectation maximization (VEM) approaches [58, 59]. They are also amenable either to the exact Bayesian analysis or to approximate Bayesian analysis through Markov chain Monte Carlo (MCMC) sampling approaches such Gibbs sampling and deterministic techniques such as the variational Bayesian expectation maximization (VBEM) algorithm [60]. This makes the whole exercise of defining and estimating new models that belong to CEFGMs more convenient. In this thesis, we predominantly utilize the EM and the VBEM algorithms for estimating the models developed. Therefore, we illustrate these algorithms in this

chapter for estimating the maximum likelihood (ML) and Bayesian versions of the PCA model, which were originally presented in [28, 33].

2.1 Bayesian Networks

Most data-driven models can be expressed graphically by means of the probabilistic graphical models (PGMs). Data-driven models involve many interacting variables that include observed data, latent variables, parameters and hyper-parameters. PGMs can be used to elegantly represent these interactions and visualize the dependence and independence relationships among these variables. The knowledge of these interactions in turn, aids in deriving a tractable estimation or inference or prediction algorithms for the data-driven models [41, 61, 62, 63, 64, 65]. There are two well-known classes of PGMs, namely, (i) Markov random field or Markov networks and (ii) Bayesian networks [41]. Markov networks are undirected graphical models and Bayesian networks are directed acyclic graphical models (DAGs).

A BN defines a joint distribution over a set of random variables as a product of set of conditional distributions. Fig. 2.1 shows an example of a BN. This particular network defines a joint distribution over the random variables, a, b, c, d and e. To understand how the joint distribution is defined by this network, we need to familiarize ourselves with some more terminologies associated with the BNs. In a BN, the random variables are denoted by encircled nodes and the interactions among the nodes are represented by directed arcs. The node at the tail end of a directed arc is commonly termed as a parent node of the node at the head end of the arc. Consequently, the node at the head end of an arc is termed as a child node of the node at the tail end of the arc. For example in Fig. 2.1, node d is a child node of a and a is a parent node of d. The joint distribution of all the variables in a BN can be expressed as a product of a set of conditional distributions of all the nodes given their respective parent nodes. In the case of Fig. 2.1, those conditional distributions are given by $p\left(e|c,d\right)$ (e given its parents c and d), $p\left(c|a\right)$ (c given its only parent a), $p\left(d|a,b\right)$ (d given its parents a and b), $p\left(a\right)$ and $p\left(b\right)$*. Conditional distributions of a and b are not conditioned on

*Note that the function $p(.)$ takes the probability mass function form in the case of discrete random variables and it take the probability density function form in the case of continues random

any of the other variables in the network as they have no parents.

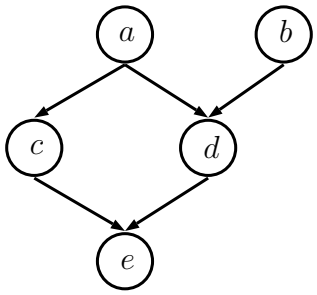


Figure 2.1: Example of a Bayesian network.

Mathematically, the joint distribution defined by the BN in Fig. 2.1 can be represented as the following,

$$p\left(a,b,c,d,e\right)=p\left(e|c,d\right)p\left(c|a\right)p\left(d|a,b\right)p\left(a\right)p\left(b\right) \tag{2.1}$$

where $p\left(a,b,c,d,e\right)$ is the joint distribution of a, b, c, d and e. BNs encode an important assumption about the interactions among the variables in the network. A node in a BN becomes independent of all of its non-descendants given its parent nodes. This can be understood by comparing the joint distribution defined by the chain rule of probability against the one defined by the BNs. For example, for the variables in the network shown in Fig. 2.1, one way of expressing the joint distribution by the chain rule of probability is as follows,

$$p\left(a,b,c,d,e\right)=p\left(e|c,d,a,b\right)p\left(c|d,a,b\right)p\left(d|a,b\right)p\left(a|b\right)p\left(b\right) \tag{2.2}$$

When comparing to the conditional distributions of e in Eqn. (2.2), the conditional distribution of e in Eqn. (2.1) does not include variables a and b. This is because the variables a and b are non-descendants of e and e becomes conditionally independent of them given it parents c and d. Similar simplification can be observed in the other conditional distributions in Eqn. (2.1). Therefore, we can ideally simplify the conditionals in Eqn. (2.2) with the knowledge of the BN to the ones in Eqn. (2.1).

The structure of a BN has to respect one important constraint. Only DAGs can be considered as BNs. In a DAG, by definition, if we were to start from an arbitrary

variables

node and travel along the direction of the directed arcs across the graph/network, we would never be able to reach the node that we initially started from. For example, in the BN shown in Fig. 2.1, if we start from the node a and by following the direction of the directed arcs, we can reach c, d and e, however, cannot reach back to a. Same will be the case when we start from any other node in the network. On the other hand, the graph shown in Fig. 2.2 is an example of a graph with cycles. By starting from node a, we can reach back to a. Such graphs cannot be considered as BNs.

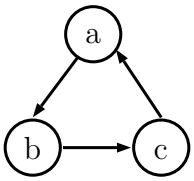


Figure 2.2: Example of a directed graph with cycles. Such graphs cannot be considered as Bayesian networks.

2.1.1 D-Separation

It is important to understand the nature of dependence and independence relationships among the variables to devise inference algorithms for the BNs. Inference refers to the problem identifying a probable state of a node or the probable states of a set of nodes in the network given a partial or complete information of the states of the rest of the nodes in the network. D-separation is a principle that defines the dependence and independence rules among the variables in PGMs. More precisely, D-separation allows one to infer the role of a subset of nodes that connects two distinct subsets of nodes in a network, whether the two distinct subsets remain independent or dependent under the following cases, (i) nodes in the connecting subsets are observed or given and (ii) nodes in the connecting subsets are unobserved or not given.

Let A, B and C be three distinct subsets of nodes in a BN, B be the subset that connects both A and C and there exist no direct connections between A and C. As illustrated in Fig. 2.3, both A and C can be connected through B in the following four possible configurations, (i) directed arcs from A to B and from B to C, (ii) directed arcs from C to B and from B to A, (iii) directed arcs from B to both A and C, and (iv) directed arcs from both A and C to B. In these cases, we are interested in

identifying if A and C are *a priori* independent (i.e., D-separated) or dependent on each other and if A and C are conditionally independent given B (i.e., D-separated by B) or conditionally dependent given B. For all four cases, the nature dependence and independence relationships or D-separation rules are described in proposition 1.

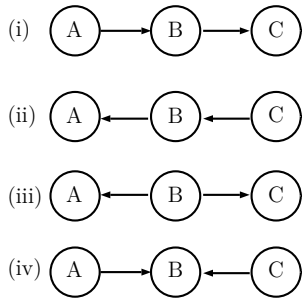


Figure 2.3: Four possible configurations that two directly unconnected distinct subsets of nodes A and C can be connected through a distinct subset of nodes B in a BN. The nodes within A, B and C can be connected among themselves through arbitrary BNs. Multiple nodes from A can be connected to multiple nodes in B, however, the direction of arcs has to remain the same across all the connections and the same applies to connections between B and C.

Proposition 1. *D-separation rules: For the four networks shown in Fig. 2.3, the following independence and dependence rules hold,*

- *For Fig. 2.3 (i), A and C may not be independent a priori, however, they become conditionally independent given B, i.e., $p(C|A,B) = p(C|B)$ and $p(A|B,C) = p(A|B)$ or in other words, B D-separates A and C.*
- *For Fig. 2.3 (ii), A and C may not be independent a priori, however, they become conditionally independent given B, i.e., $p(C|A,B) = p(C|B)$ and $p(A|B,C) = p(A|B)$ or in other words, B D-separates A and C.*
- *For Fig. 2.3 (iii), A and C may not be independent a priori, however, they become conditionally independent given B, i.e., $p(C|A,B) = p(C|B)$ and $p(A|B,C) = p(A|B)$ or in other words, B D-separates A and C.*
- *For Fig. 2.3 (iv), A and C are independent a priori, however, may become dependent on each other given B (i.e., B cannot D-separate A and C).*

A proof of proposition 1 is provided in section. A.1 of Appendix. A.

2.1.2 Markov Blanket

Markov blanket of an arbitrary subset of nodes in a PGM refers to a minimal distinct subset of nodes in the graph that D-separates the subset concerned from the rest of the graph. This D-separation implies that given the complete information or the states of its Markov blanket, any additional information about the nodes outside the Markov blanket will add no valuable information to infer the states of the subset concerned. Therefore, for the purpose of inference, it is useful to define the Markov blanket of a node or a subset of nodes in a graphical model. For a BN, the Markov blanket of an arbitrary subset of nodes can be defined as stated in proposition 2.

Proposition 2. *In a BN, Markov blanket of an arbitrary subset of nodes is given by a distinct subset consisting of all of its parent nodes, all of its children nodes and the set of all the other parent nodes of its children nodes as illustrated in Fig. 2.4.*

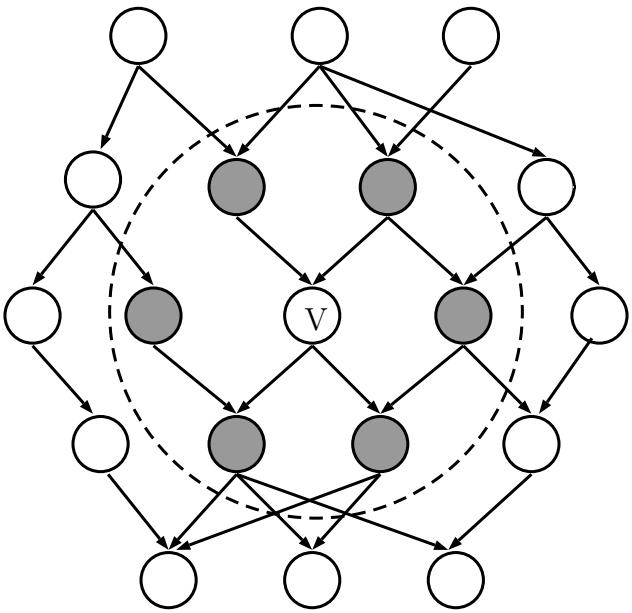


Figure 2.4: An arbitrary BN with a subset of nodes V. Markov blanket of V is given by the subset consisting of shaded nodes that are of all the parent nodes of V, all the children nodes of V and all the other parent nodes of its children nodes.

A proof of proposition 2 is provided in section A.2 of Appendix A. We will be using the concepts, D-separation and Markov blanket when deriving a tractable estimation

algorithms for the maximum likelihood and Bayesian version of the PCA model in this thesis chapter and for the models developed in the subsequent chapters.

2.1.3 Bayes Rule of Inference

The Bayes Rule of Inference is a fundamental building block in many algorithms devised for inference in BNs. The Bayes rule can be seen as a direct consequence of the chain of rule of probability. Consider a BN of two variables a and b shown in Fig. 2.5. The joint distribution of a and b is defined by the BN as the following,

$$p\left(a,b\right)=p\left(b|a\right)p\left(a\right) \tag{2.3}$$

where the conditional distribution, $p\left(b|a\right)$, can be used to predict the probable state of b given a. In addition, we are also often interested in the inverse inference problem, i.e., determining the probable state of a given b, which requires the conditional distribution $p\left(b|a\right)$.

Using the chain probability, the joint distribution of a and b can be expressed equivalently in two ways as the following

$$p\left(a,b\right)=p\left(b|a\right)p\left(a\right)=p\left(a|b\right)p\left(b\right) \tag{2.4}$$

We can rearrange the above equation to obtain the conditional distribution of a given b as the following,

$$p\left(a|b\right)=\frac{p\left(b|a\right)p\left(a\right)}{p\left(b\right)} \tag{2.5}$$

The above equation is known as the Bayes rule or the Bayes theorem. The terms $p\left(b|a\right)$, $p\left(a\right)$ and $p\left(a|b\right)$ are referred to as the likelihood, prior distribution and posterior distribution of a respectively. The term $p\left(b\right)$ is referred to as the marginal distribution of b, which can be obtained by marginalizing or integrating out a from the joint distribution of a and b as the following,

$$p\left(b\right)=\int_{b}p\left(b|a\right)p\left(a\right)db \tag{2.6}$$

Figure 2.5: A two variable Bayesian network.

2.1.4 Bayesian Network Representation of Data-Driven Models

There is an advantage in defining different models under a common framework such as the BNs. This would allow one to develop model estimation algorithms and state inference algorithms for the general framework and deduce the algorithms for any special case if needed. Furthermore, an algorithm developed for a particular model can be extended to be applied on a model that can be expressed under the same family. In this subsection, we will attempt to express and visualize some of the commonly used data-driven models in the process data analysis literature in the form of BNs. We pick the following four examples to show how they can be expressed as BNs, (i) multivariate linear regression model, (ii) vector autoregressive model, (iii) state space model, (iv) maximum likelihood or probabilistic principal component analysis (PPCA) model and (v) Bayesian PPCA model.

Example 1: Multivariate Linear Regression Model

Consider a set of N output observations $Y \triangleq \left\{y_1, ...y_n \in \mathbb{R}^D, ..., y_N\right\} \in \mathbb{R}^{D\times N}$, a set of N input observations $U \triangleq \left\{u_1, ...u_n \in \mathbb{R}^P, ..., u_N\right\} \in \mathbb{R}^{P\times N}$ obtained from a linear system, each output observation belongs to a D-dimension real space and each input observation belongs to a P-dimension real space. Any arbitrary input and output observation n can be related by the multivariate linear regression model as the following,

$$y_n = \theta u_n + e_n \tag{2.7}$$

where $\theta \in \mathbb{R}^{D\times P}$ is the parameter matrix, and $e_n \in \mathbb{R}^D$ is the D-dimension measurement noise in y_n. Assume e_n is independent and identically distributed and follows a

multivariate Gaussian distribution with zero mean and covariance Σ_e as the following,

$$e_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_e) \tag{2.8}$$

This model can be interpreted as a probabilistic model for the outputs of the system. Given the fixed parameters, θ and Σ_e and the inputs of the system, the probability distribution over the outputs can be expressed as shown below,

$$y_n \overset{i.i.d}{\sim} \mathcal{N}(\theta u_n, \Sigma_e) \tag{2.9}$$

where y_n is shown to follow multivariate Gaussian distribution with mean θu_n and covariance Σ_e. For the set of N observations from the system, the joint distribution can be expressed as the following,

$$p(Y|U, \theta, \Sigma_e) = \prod_{n=1}^{N} p(y_n|\theta, u_n, \Sigma_e) \tag{2.10}$$

where the joint distribution of all the output observations are expressed as the product of marginal distributions of the individual observations given the parameters θ and Σ_e, and the inputs. The individual marginal distributions take the distribution form shown in Eqn. (2.9). The joint distribution shown in Eqn. (2.10) can be represented using a BN as shown in Fig. 2.6. The only random variable in the network, y_n, is denoted by the encircled node and the deterministic variables and the parameters that take fixed values are denoted by the plain nodes. To denote the repetition of nodes for N different observations, those nodes that are to be repeated are contained within a rectangular plate in the graphical representation.

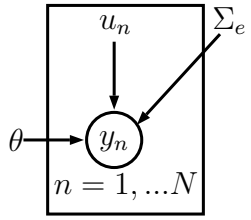


Figure 2.6: Bayesian network representation of the multivariate linear regression model with N observations.

Example 2: Vector Autoregressive Model

Consider a system with a D-dimension output and the generation of those outputs can be described by the first order auto regressive model of the following form,

$$y_t = \theta y_{t-1} + e_t \tag{2.11}$$

where $y_t \in \mathbb{R}^D$ is the output observed at time instant t, it depends on the output at the previous time instant $t-1$, y_{t-1}, $\theta \in \mathbb{R}^{D \times D}$ is the coefficient matrix, and e_t is the additive measurement noise at time instant t. The additive noise is independent and identically distributed and follows multivariate Gaussian distribution with zero mean and covariance Σ_e as the following,

$$e_t \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_e) \tag{2.12}$$

This would allow us to represent the probability distribution of the output at t as shown below,

$$y_t \overset{i.i.d}{\sim} \mathcal{N}(\theta y_{t-1}, \Sigma_e) \tag{2.13}$$

Further, the joint distribution of the time series of length T can be written as the product of conditional distributions of the outputs at each time instant given their previous observation as the following,

$$p(Y|\theta, \Sigma_e) = p(y_1|\Sigma_e) \prod_{t=2}^{T} p(y_t|\theta, y_{t-1}, \Sigma_e) \tag{2.14}$$

where $p(y_1|\Sigma_e)$ represents the distribution of the output at time instant one. It is simply assumed to follows multivariate Gaussian distribution with mean zero and covariance Σ_e. The joint distribution shown in Eqn. (2.14) can be represented using a BN as shown in Fig. 2.7.

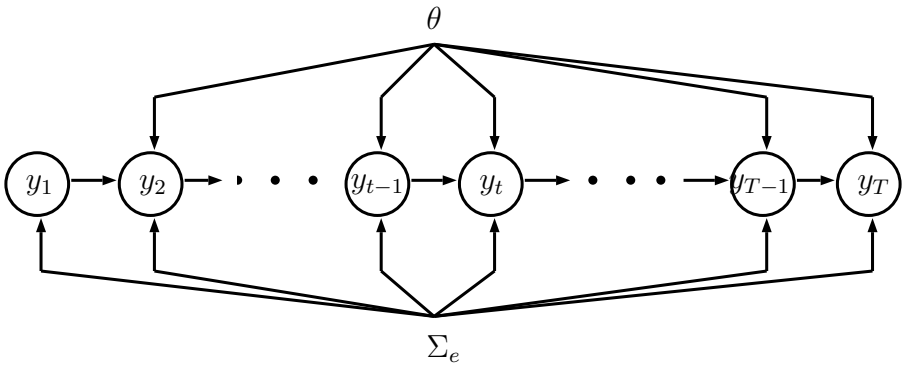


Figure 2.7: Bayesian network representation of a first order vector autoregressive model of a sequence of observations of length T.

Example 3: State Space Model

Consider a system that can be described using the state space model shown below,

$$z_t = \Omega_1 z_{t-1} + \Omega_2 u_t + \epsilon_t \tag{2.15}$$

$$y_t = \Omega_3 z_t + \Omega_4 u_t + e_t \tag{2.16}$$

where $z_t \in \mathbb{R}^K$ corresponds to the K-dimension states at time instant t and it can be expressed in terms of the states at previous time instant $t-1$, z_{t-1} and the P-dimension inputs at time instant t, $u_t \in \mathbb{R}^P$, $y_t \in \mathbb{R}^D$ are the outputs of the system at t and they can be expressed in terms of z_t and u_t, and ϵ_t and e_t are the additive noise terms in the state transition model and the output model at t respectively. The model parameters are given by $\Omega_1 \in \mathbb{R}^{K\times K}$, $\Omega_2 \in \mathbb{R}^{K\times P}$, $\Omega_3 \in \mathbb{R}^{D\times K}$, and $\Omega_4 \in \mathbb{R}^{D\times P}$.

The additive noise in both state transition and output models follow multivariate Gaussian distribution with zero mean, and Σ_ϵ and Σ_e covariances respectively as shown below,

$$\epsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_\epsilon) \tag{2.17}$$

$$e_t \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_e) \tag{2.18}$$

For this model, we can express the conditional distributions of the states and the outputs at t as the following,

$$z_t \sim \mathcal{N}(\Omega_1 z_{t-1} + \Omega_2 u_t, \Sigma_\epsilon) \tag{2.19}$$

$$y_t \sim \mathcal{N}(\Omega_3 z_t + \Omega_4 u_t, \Sigma_e) \tag{2.20}$$

The joint distribution of the states, $Z \triangleq \left\{z_0, z_1, ...z_t \in \mathbb{R}^K, ..., z_T\right\} \in \mathbb{R}^{K\times T+1}$ of length $T+1$ and the outputs, $Y \triangleq \left\{y_1, ...y_t \in \mathbb{R}^D, ..., y_T\right\} \in \mathbb{R}^{D\times T}$ of length T given the inputs, $U \triangleq \left\{u_1, ...u_t \in \mathbb{R}^P, ..., u_T\right\} \in \mathbb{R}^{P\times T}$ and the model parameters can be expressed as the following,

$$\begin{aligned} p\left(Y, Z|U, \Omega_1, \Omega_2, \Omega_3, \Omega_4, \Sigma_e, \Sigma_\epsilon\right) = \quad & p\left(z_0|\Sigma_\epsilon\right) \prod_{t=1}^{T} p\left(z_t|z_{t-1}, u_t, \Omega_1, \Omega_2, \Sigma_\epsilon\right) \times \\ & p\left(y_t|z_t, u_t, \Omega_3, \Omega_4, \Sigma_e\right) \end{aligned} \tag{2.21}$$

where z_0 is simply assumed to follow multivariate Gaussian distribution with mean zero and covariance Σ_e. This joint distribution can also be represented as a BN as shown in Fig. 2.8.

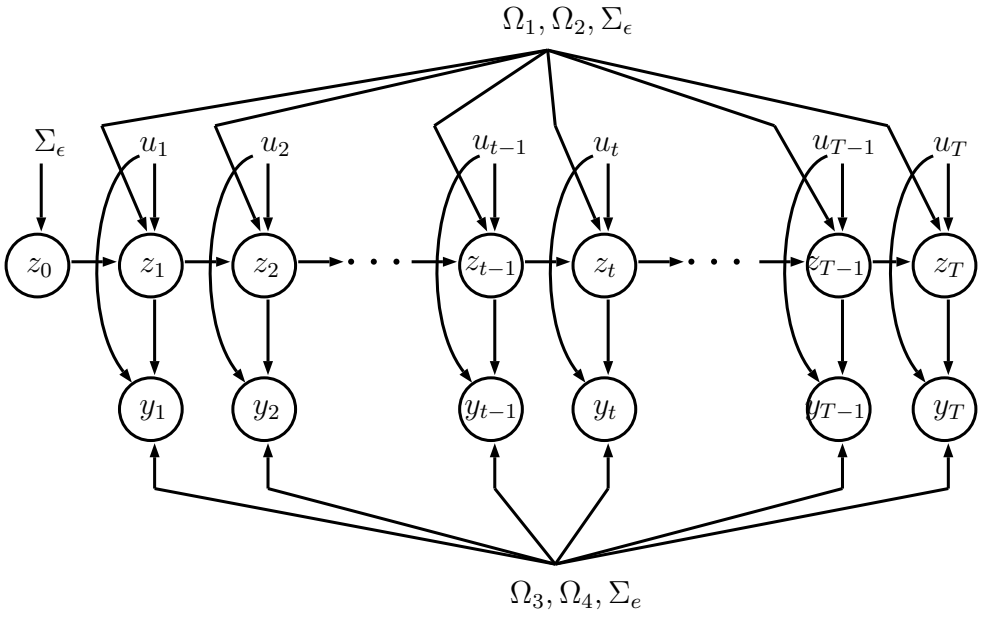


Figure 2.8: Bayesian network representation of a state-space model of a sequence of observations of length T.

Example 4: Probabilistic Principal Component Analyser Model

Consider a set of N output observations $Y = \left\{y_1, ..., y_n \in \mathbb{R}^D, ..., y_N\right\} \in \mathbb{R}^{D\times N}$ from a system. The PPCA model assumes that these observations are generated from the lower dimension latent variables. Let $Z = \left\{z_1, ..., z_n \in \mathbb{R}^K, ..., z_N\right\} \in \mathbb{R}^{K\times N}$ be the set of N lower dimension latent variables with $K < D$. In the PPCA model, each output observation is expressed as a linear combination of the corresponding lower

dimension latent variables as the following,

$$\begin{aligned} y_n &= W z_n + e_n \\ z_n &\overset{i.i.d}{\sim} \mathcal{N}(0, I_K) \\ e_n &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma I_D) \end{aligned} \tag{2.22}$$

where $W \in \mathbb{R}^{D \times K}$ is the projection or the coefficient or the loading matrix, e_n is the measurement noise in the observation and it follows multivariate Gaussian distribution with zero mean and covariance σI_D (diagonal matrix with all its diagonal elements given by σ), and z_n follows a multivariate Gaussian distribution with zero mean and identity covariance. The subscripts K and D represent the respective sizes of the identity matrices I_K and I_D, respectively. The measurement noise and latent variable are assumed to be mutually independent.

The joint distribution of the observations and the latent variables are given by the model shown in Equation (2.22) as the following,

$$p(Y, Z|W, \sigma) = p(Y|W, Z, \sigma)\, p(Z) = \prod_{n=1}^{N} p(y_n|W, z_n, \sigma)\, p(z_n) \tag{2.23}$$

where the conditional distribution of y_n, $p(y_n|W, z_n, \sigma)$ is a multivariate Gaussian distribution with mean $W z_n$ and covariance σI_D and the marginal distribution of z_n, $p(z_n)$ is a multivariate Gaussian distribution with mean zero and covariance σI_K. A BN representation of the PPCA model for N observations is shown in Fig. 2.9.

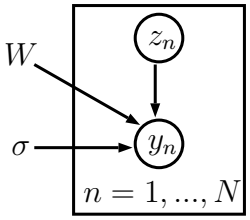


Figure 2.9: Bayesian network representation of the PPCA model. To obtain the actual network, the structure within the rectangular enclosure or within the rectangular plate has to be simply repeated $\forall n \in [1, N]$.

Example 5: Bayesian Probabilistic Principal Component Analysis

So far, we have treated the parameters in the models as fixed quantities. However, we can also impose our belief on the parameters by defining the probability distributions of the parameters and perform Bayesian analysis. For illustration, we extend the

PPCA model to a Bayesian version and call it the Bayesian PPCA model. We treat the parameter W as a random variable by defining the prior distribution for W. Consider that each column of W, w_k, is multivariate Gaussian distributed with zero mean and $\nu_k^{-1} I_D$ covariance as the following,

$$w_k \sim \mathcal{N}\left(0, \nu_k^{-1} I_D\right) \tag{2.24}$$

where ν_k is a scalar parameter representing the inverse of variance of each element in w_k and I_D is an identity matrix of dimension D. In addition, we can also treat ν_k as a random variable and assume it to follow a gamma distribution with shape parameter α^* and rate parameter β^* as shown below[†],

$$\nu_k \sim Ga\left(\alpha^*, \beta^*\right) \tag{2.25}$$

We can represent the resulting model as a BN as shown in Fig. 2.10. The parameter ν represents the collection of parameters ν_1 to ν_K, each corresponding to the respective column of W. We can expand the network to show each column of W and each element of ν. However, for simplicity, we will retain a simpler representation as shown in Fig. 2.10. If desired, we can go on and define a distribution for α^* and β^* and so forth. However, the analysis will then become more and more complex to infer or estimate the parameters given the data. Similarly, we can also define a distribution for σ. However, for the purpose of illustration, we will stick to the simpler case shown in Fig. 2.10.

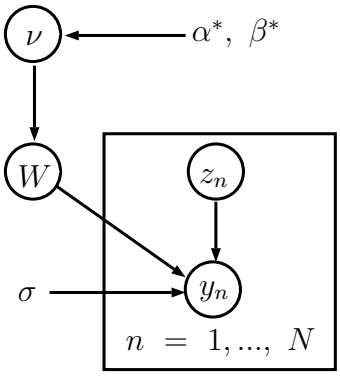


Figure 2.10: Bayesian network representation of the Bayesian probabilistic principal component analysis model.

[†]This particular choice of the prior distributions make the Bayesian PPCA fall under the class of CEFGMs as we will see in section 2.2 of this chapter.

2.2 Conjugate Exponential Family Graphical Models

Definition 1. *Exponential family distributions: They are a family of distributions that can be expressed as the following,*

$$p\left(x|\eta\right) = h\left(x\right)\exp\left\{\eta^T T\left(x\right) - A\left(\eta\right)\right\} \tag{2.26}$$

where $p\left(x|\eta\right)$ is the probability density or the mass function of the distribution of the random variable x with parameters η. $T(x)$ is a function of x that maps x to a real valued vector of the same size as η and $h(x)$ is a function of x that maps x to a real valued scalar. $A\left(\eta\right)$ is the cumulant function that normalizes the integration of $p\left(x|\eta\right)$ over its support to one. Given, η and $T(x)$, we can determine $A\left(\eta\right)$ as the following,

$$A\left(\eta\right) = \ln \int h\left(x\right)\exp\left\{\eta^T T\left(x\right)\right\} dx \tag{2.27}$$

In this thesis, all the considered prior distributions of the random variables belong to the exponential family. The exponential family includes most commonly used and many naturally occurring distributions such as normal, gamma, χ^2, exponential, beta, geometric, etc. As an example of an exponential family distribution, we discuss the univariate Gaussian distribution below,

Example: Univariate Gaussian Distribution

Let us consider a random variable x that follows a univariate Gaussian distribution with mean μ and variance $\frac{1}{\zeta}$. Its probability density function is given as the following,

$$p\left(x|\mu,\zeta\right) = \frac{\sqrt{\zeta}}{\sqrt{2\pi}}\exp\left\{-\frac{\zeta}{2}\left(x-\mu\right)^2\right\} \tag{2.28}$$

We can rewrite the above function in the form shown in Eqn. (2.26) as illustrated below,

$$p\left(x|\mu,\zeta\right) = \frac{\sqrt{\zeta}}{\sqrt{2\pi}}\exp\left\{-\frac{\zeta}{2}x^2 - \frac{\zeta\mu^2}{2} + \zeta\mu x\right\} \tag{2.29}$$

$$p\left(x|\mu,\zeta\right) = \frac{\sqrt{\zeta}}{\sqrt{2\pi}}\exp\left\{\begin{bmatrix} \zeta\mu \\ -\frac{\zeta}{2} \end{bmatrix}^T \begin{bmatrix} x \\ x^2 \end{bmatrix} - \frac{\zeta\mu^2}{2}\right\} \tag{2.30}$$

It can be seen that the above expression resembles the form shown in Eqn. (2.26) with the following parametrization,

$$h(x) = \frac{\sqrt{\zeta}}{\sqrt{2\pi}} \tag{2.31}$$

$$\eta = \begin{bmatrix} \zeta\mu \\ -\frac{\zeta}{2} \end{bmatrix} \tag{2.32}$$

$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \tag{2.33}$$

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} \tag{2.34}$$

Definition 2. *Conjugate prior distribution for the likelihood: Prior distribution of a random variable is called the conjugate prior distribution for its likelihood when both the posterior and the prior belong to the same distribution family.*

For illustration, let us recall the univariate Gaussian distribution shown in Eqn. (2.28). This forms the likelihood function for the parameters μ and ζ. Let μ be a fixed parameter and consider ζ to be a random variable that follows a gamma distribution with shape parameter κ^* and rate parameter ϕ^*. This prior distribution has the following density function,

$$p(\zeta|\kappa^*, \phi^*) = \frac{(\phi^*)^{\kappa^*}}{\Gamma(\kappa^*)} (\zeta)^{\kappa^*-1} \exp(-\phi^*\zeta) \tag{2.35}$$

where Γ represents the gamma function. Using the Bayes rule of probability, we express the posterior of ζ as the following,

$$p(\zeta|x, \mu, \kappa^*, \phi^*) = \frac{p(x|\mu, \zeta)\, p(\zeta|\kappa^*, \phi^*)}{p(x|\mu, \kappa^*, \phi^*)} \tag{2.36}$$

The term in the denominator is independent of ζ. Therefore, the posterior can be expressed as the following,

$$p(\zeta|x, \mu, \kappa^*, \phi^*) \propto p(x|\mu, \zeta)\, p(\zeta|\kappa^*, \phi^*) \tag{2.37}$$

Replacing the likelihood and the prior by their respective density functions and simplifying further leads to the following,

$$p(\zeta|x, \mu, \kappa^*, \phi^*) \propto \frac{\sqrt{\zeta}}{\sqrt{2\pi}} \exp\left\{-\frac{\zeta}{2}(x-\mu)^2\right\} \frac{(\phi^*)^{\kappa^*}}{\Gamma(\kappa^*)} (\zeta)^{\kappa^*-1} \exp\{-\phi^*\zeta\} \tag{2.38}$$

$$p\left(\zeta|x,\mu,\kappa^*,\phi^*\right) \propto \zeta^{\kappa^*+\frac{1}{2}-1} \exp\left\{-\left(\frac{(x-\mu)^2}{2}+\phi^*\right)\zeta\right\} \tag{2.39}$$

where the above expression very much resembles the density function of the gamma distribution family without the normalizing constant. Therefore, the posterior distribution of ζ can be expressed as the gamma distribution of the following form,

$$\zeta|x,\mu,\kappa^*,\phi^* \sim Ga\left(\kappa,\phi\right) \tag{2.40}$$

where

$$\kappa = \kappa^* + \frac{1}{2},\ \phi = \phi^* + \frac{1}{2}\left(x-\mu\right)^2 \tag{2.41}$$

In this example, both the prior distribution and the posterior distribution of ζ belong to the gamma distribution family. Therefore, the prior distribution of ζ can be called as the conjugate prior for its Gaussian likelihood.

Definition 3. *Conjugate exponential family graphical models (CEFGMs): It refers to the BNs with exponential family distributions as the prior distributions of all the random variable nodes and the prior distributions of the nodes are conjugate for their conditional likelihood, defined by the prior distributions of their children nodes.*

As an example for CEFGMs, the Bayesian PPCA model discussed in this chapter also belongs to CEFGMs among the others. All the prior distributions, $p\left(Y|Z,W,\sigma\right)$, $p\left(Z\right)$, $p\left(W|\nu\right)$, and $p\left(\nu|\alpha^*,\beta^*\right)$ fall under the exponential family. The prior $p\left(Z\right)$ is conjugate for its conditional likelihood $p\left(Y|Z,W,\sigma\right)$, $p\left(W|\nu\right)$ is conjugate for its conditional likelihood $p\left(Y|Z,W,\sigma\right)$, and $p\left(\nu|\alpha^*,\beta^*\right)$ is conjugate for $p\left(W|\nu\right)$.

2.3 Maximum Likelihood Estimation

When the parameters are treated as fixed quantities, the maximum likelihood (ML) approach is one of the commonly used approaches for parameter estimation in the case of probabilistic models. The ML approach seeks to obtain parameters that maximize the likelihood function, which is the conditional distribution of the observed data given the parameters. In this section, we illustrate the maximum likelihood estimation approach for the PPCA model defined in Fig. 2.9 using the EM algorithm.

For the PPCA model, the likelihood function of the parameters is given by the conditional distribution of Y given the parameters W and σ, $p\left(Y|W,\sigma\right)$. There the objective of the maximum likelihood estimation translates into the following,

$$W_{ML},\, \sigma_{ML} = \max_{W,\sigma} p\left(Y|W,\sigma\right) \tag{2.42}$$

where W_{ML} and σ_{ML} are the ML parameter estimates that maximize the likelihood function $p\left(Y|W,\sigma\right)$. Alternatively, we could also choose parameters that maximize any monotonic transformation of the likelihood function. For example, we could choose parameters that maximize the natural logarithm of the likelihood as the following,

$$W_{ML},\, \sigma_{ML} = \max_{W,\sigma} \ln p\left(Y|W,\sigma\right) \tag{2.43}$$

The estimates that maximize the monotonic transformation will also be the maximizers of the likelihood function. The transformation often makes the optimization more convenient. For example, the natural logarithm greatly simplifies the form of the likelihood functions such as the density function of the Gaussian distribution that includes an exponential term.

For our PPCA model, we can obtain the likelihood function or the conditional distribution of the observed data given the parameters by integrating out the latent variables from the joint distribution of the observations and the latent variables as the following,

$$p\left(Y|W,\sigma\right) = \int_{Z} p\left(Y|Z,W,\sigma\right) p\left(Z\right) dZ \tag{2.44}$$

Expanding the above expression for all the N observations leads to the following,

$$p\left(Y|W,\sigma\right) = \prod_{n=1}^{N} \int_{z_n} p\left(y_n|z_n,W,\sigma\right) p\left(z_n\right) dz_n \tag{2.45}$$

In the above expression, when substituting the conditional distributions dictated by the PPCA model, we can obtain the following expression,

$$p\left(Y|W,\sigma\right) = \prod_{n=1}^{N} \int_{z_n} \begin{array}{l} \frac{1}{(2\pi\sigma)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\sigma}\left(y_n - W z_n\right)^T \left(y_n - W z_n\right)\right\} \times \\ \frac{1}{(2\pi)^{\frac{K}{2}}} \exp\left\{-\frac{1}{2} z_n^T z_n\right\} dz_n \end{array} \tag{2.46}$$

Integrating out z_n from the above expression gives the form of the marginal distribution of the observations or the likelihood function of the model parameters as the

following,

$$p\left(Y|W,\sigma\right)=\prod_{n=1}^{N}\frac{1}{\left(2\pi\right)^{\frac{D}{2}}\left|WW^{T}+\sigma I_{D}\right|}\exp\left\{-\frac{1}{2}y_{n}^{T}\left(WW^{T}+\sigma I_{D}\right)^{-1}y_{n}\right\} \quad (2.47)$$

The resulting likelihood function is a product of individual conditional distributions of the observations. The individual conditional distributions are given by a multivariate Gaussian distribution with zero mean and $WW^T + \sigma I_D$ covariance. The functional form of the likelihood shown in Eqn. (2.47) is non-linear in terms of W and σ. We can utilize any optimization algorithm to obtain the parameters that maximize the likelihood. The natural log transform of the likelihood will simply it to an extent and yet would require implementation of an optimization algorithm to estimate the parameters of the model.

2.3.1 Expectation Maximization Algorithm

The EM algorithm [56] (for tutorial see [57]) is an optimization approach for the maximum likelihood estimation. For some models such as the PPCA model, it avoids implementation complexities associated with the gradient based optimization algorithms and it can provide a simple iterative procedure with explicit update expressions for the parameters in each iteration.

Implementation of the EM algorithm involves the following steps,

1. The first advancing step of deriving the EM algorithm involves identifying a lower bound function for the natural log of the likelihood function (log likelihood function) in terms of the joint distribution of the observed data and the latent variables in the model and a proxy posterior distribution of the latent variables.

2. The second advancing step involves deriving the posterior distribution of the latent variables in terms of the observed data and the model parameters using the Bayes rule of inference. This will not be achievable for CEFGMs where the latent variables cannot be marginalized from the joint distribution of the observed data and the latent variables. In such cases, approximate inference techniques can be applied (e.g. variational EM algorithm[58, 59]). However, for the PPCA model, we can marginalize the latent variables from the joint distribution as shown in Eqn. (2.47).

3. The third advancing step involves replacing the proxy posterior distribution with the posterior distribution derived in step 2 and deducing the lower bound expression. If the exact posterior distribution is available, the lower bound function becomes equal to the log likelihood function.

4. The fourth advancing step involves deriving the update expressions for the parameters of the models such that it maximizes the lower bound function. In the case of PPCA model, the parameter updates can be achieved through explicit update expressions. For some complex models, the parameter updates may require implementation of the optimization algorithms.

5. The fifth advancing step involves implementation of the EM algorithm through an iterative procedure where the parameters and the posterior distribution of the latent variables are updated recursively.

We illustrate these steps for the PPCA model in the rest of this section.

Lower Bound of the Log Likelihood Function

We derive the lower bound of the log likelihood function for the PPCA model using proposition 3 presented below,

Proposition 3. *A functional can be defined in terms of the joint distribution of the observed data and latent variables and $q(Z)$ that lower bounds the likelihood function of the parameters as the following,*

$$\ln p(Y|W,\sigma) \geq \int_Z q(Z) \ln \frac{p(Y|W,Z,\sigma)\, p(Z)}{q(Z)} dZ \tag{2.48}$$

where $q(Z)$ satisfies the properties of a probability density function defined over Z. When $q(Z)$ is equal to the actual posterior of Z, $p(Z|Y,W,\sigma)$, the lower bound defined above becomes exactly equal to the log likelihood function as shown below,

$$\ln p(Y|W,\sigma) = \int_Z q(Z) \ln \frac{p(Y|W,Z,\sigma)\, p(Z)}{q(Z)} dZ \tag{2.49}$$

Proof. Apart from expressing the joint distribution of the data and the latent variables using the BN defined in Fig. (2.9), we can also express it in terms of $p(Y|W,\sigma)$,

which is consistent with the chain rule of probability, as the following,

$$p\left(Y,Z|W,\sigma\right)=p\left(Z|Y,W,\sigma\right)p\left(Y|W,\sigma\right)=p\left(Y|W,Z,\sigma\right)p\left(Z\right) \tag{2.50}$$

The likelihood function can then be expressed by rearranging the terms in the above expression as the following,

$$p\left(Y|W,\sigma\right)=\frac{p\left(Y|W,Z,\sigma\right)p\left(Z\right)}{p\left(Z|Y,W,\sigma\right)} \tag{2.51}$$

By taking the natural logarithm on both sides results in the following,

$$\ln p\left(Y|W,\sigma\right)=\ln\frac{p\left(Y|W,Z,\sigma\right)p\left(Z\right)}{p\left(Z|Y,W,\sigma\right)} \tag{2.52}$$

Although the terms in the RHS involve Z, the log likelihood functions shown above in Eqn. (2.52) is independent of Z. This fact can also be verified from Eqn. (2.47). Therefore, taking the expectation of the log likelihood with respect to the function $q\left(Z\right)$ will results in the log likelihood itself. This leads to the following equality,

$$\ln p\left(Y|W,\sigma\right)=\int_{Z}q\left(Z\right)\ln\frac{p\left(Y|W,Z,\sigma\right)p\left(Z\right)}{p\left(Z|Y,W,\sigma\right)}dZ \tag{2.53}$$

Now, we can multiply and divide the terms inside the natural logarithm on the RHS by $q\left(Z\right)$ without altering the outcome to obtain the following,

$$\ln p\left(Y|W,\sigma\right)=\int_{Z}q\left(Z\right)\ln\frac{p\left(Y|W,Z,\sigma\right)p\left(Z\right)q\left(Z\right)}{p\left(Z|Y,W,\sigma\right)q\left(Z\right)}dZ \tag{2.54}$$

Further, the above expression can be split into the summation of the following two terms,

$$\ln p\left(Y|W,\sigma\right)=\underbrace{\int_{Z}q\left(Z\right)\ln\frac{p\left(Y|W,Z,\sigma\right)p\left(Z\right)}{q\left(Z\right)}dZ}_{\mathcal{L}_{LB}}+\underbrace{\int_{Z}q\left(Z\right)\ln\frac{q\left(Z\right)}{p\left(Z|Y,W,\sigma\right)}dZ}_{KL\ divergence} \tag{2.55}$$

where we name one of the terms in the summation as $\mathcal{L}_{LB}$ and the other term is the Kullback-Leibler (KL) divergence between the two distributions of Z, $q\left(Z\right)$ and $p\left(Z|Y,W,\sigma\right)$. The KL divergence is a measure of distance between two distributions and it is always positive. Therefore, the term $\mathcal{L}_{LB}$ lower bounds the log likelihood, $\ln p\left(Y|W,\sigma\right)$ as expressed below,

$$\ln p\left(Y|W,\sigma\right)\geq\mathcal{L}_{LB} \tag{2.56}$$

When $q(Z)$ becomes exactly equal to $p(Z|Y, W, \sigma)$, the KL divergence term in Eqn. (2.55) becomes zero as the term within the natural logarithm becomes one. A direct consequence of this is that, $\mathcal{L}_{LB}$ becomes equal to the log likelihood. This leads to the following results,

$$\ln p(Y|W, \sigma) = \mathcal{L}_{LB} \tag{2.57}$$

and

$$\ln p(Y|W, \sigma) = \int_Z q(Z) \ln p(Y|W, Z, \sigma)\, dZ + \int_Z q(Z) \ln \frac{p(Z)}{q(Z)} dZ \tag{2.58}$$

when

$$q(Z) = p(Z|Y, W, \sigma) \tag{2.59}$$

This completes the proof of proposition 3. ■

Posterior Distribution of the Latent Variables

We illustrate the derivation of the posterior distribution of the latent variables for the PPCA model through lemma 1.

Lemma 1. *The posterior distribution of the latent variable z_n corresponding to the observation y_n is given by a multivariate Gaussian distribution with mean $\hat{z}_n = \frac{1}{\sigma}\Sigma_z W^T y_n$ and covariance $\Sigma_z = \sigma \left[W^T W + \sigma I_K\right]^{-1}$*

Proof. From the knowledge of the BN of the model, we can express the posterior distribution of the latent variables, $p(Z|Y, W, \sigma)$ in a more simplified form. The posterior of the latent variable z_n is independent of the other latent variables as its Markov blanket in the network is given by its only child y_n and the other parents of y_n, W and σ. This holds true $\forall n$. Therefore, the posterior of each latent variable is independent of the posteriors of the other latent variables and the joint posterior can be expressed as the following,

$$p(Z|Y, W, \sigma) = \prod_{n=1}^{N} p(z_n|y_n, W, \sigma) \tag{2.60}$$

Using the Bayes rule, each individual posterior distribution can be expressed as,

$$p(z_n|y_n, W, \sigma) = \frac{p(y_n|z_n, W, \sigma)\, p(z_n)}{p(y_n|W, \sigma)} \tag{2.61}$$

The marginal distribution of y_n, $p\left(y_n|W,\sigma\right)$, in the denominator of the above expression is independent of z_n and constant with respect to z_n. Therefore, the posterior distribution is proportional to the terms in the numerator as the following,

$$p\left(z_n|y_n,W,\sigma\right) \propto p\left(y_n|z_n,W,\sigma\right) p\left(z_n\right) \tag{2.62}$$

Now, we can substitute the distribution forms of the likelihood and the prior to obtain the following expression,

$$p\left(z_n|y_n,W,\sigma\right) \propto$$
$$\frac{1}{\left(2\pi\sigma\right)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\sigma}\left(y_n - Wz_n\right)^T\left(y_n - Wz_n\right)\right\} \frac{1}{\left(2\pi\right)^{\frac{K}{2}}} \exp\left\{-\frac{1}{2}z_n^T z_n\right\} \tag{2.63}$$

The above expression can be further simplified by removing the constant terms and retaining only the terms that involve z_n.

$$p\left(z_n|y_n,W,\sigma\right) \propto \exp\left\{-\frac{1}{2\sigma}z_n^T W^T W z_n + \frac{1}{\sigma}y_n^T W z_n - \frac{1}{2}z_n^T z_n\right\} \tag{2.64}$$

$$p\left(z_n|y_n,W,\sigma\right) \propto \exp\left\{-\frac{1}{2}z_n^T\left(\frac{1}{\sigma}W^T W + I_K\right) z_n + \frac{1}{\sigma}y_n^T W z_n\right\} \tag{2.65}$$

The exponential term in the above expression is quadratic in z_n. We can simplify this in the form of multivariate Gaussian distribution as the following,

$$p\left(z_n|y_n,W,\sigma\right) \propto \exp\left\{-\frac{1}{2}\left(z_n - \hat{z}_n\right)^T\left(\Sigma_z\right)^{-1}\left(z_n - \hat{z}_n\right)\right\} \tag{2.66}$$

where

$$\hat{z}_n = \frac{1}{\sigma}\Sigma_z W^T y_n,\ \Sigma_z = \sigma\left[W^T W + \sigma I_K\right]^{-1} \tag{2.67}$$

This completes the proof of Lemma 1. ■

Deducing the lower bound expression

Now that we know the form of the posterior distribution, we can evaluate the lower bound using Eqn. (2.58). The step includes making use of the right distribution form for the proxy posterior of the latent variables in the $\mathcal{L}_{LB}$ expression and deducing $\mathcal{L}_{LB}$. For the PPCA case, this translates into $q(z_n)$ taking the form of the probability density function of a multivariate Gaussian distribution with mean $\hat{z}_n$ and covariance

Σ_z. Now, we can deduce the expression in Eqn. (2.58) that includes two terms. Substituting the right form of $q(Z)$ allows us to deduce the first term as the following,

$$\int_Z q\left(Z\right) \ln p\left(Y|W,Z,\sigma\right) = \sum_{n=1}^{N} \int_{z_n} q\left(z_n|\hat{z}_n,\Sigma_z\right) \ln p\left(y_n|W,z_n,\sigma\right)$$

$$= -\frac{DN}{2}\ln 2\pi - \frac{DN}{2}\ln\sigma - \frac{1}{2\sigma}\sum_{n=1}^{N} y_n^T y_n + \frac{1}{\sigma}\sum_{n=1}^{N} y_n^T W \hat{z}_n$$

$$-\frac{1}{2\sigma}\sum_{n=1}^{N} tr\left\{W^T W\left[\hat{z}_n \hat{z_n}^T + \Sigma_z\right]\right\} \tag{2.68}$$

The second term can be deduced as the following,

$$\int_Z q\left(Z\right) \ln \frac{p\left(Z\right)}{q\left(Z\right)} = -\sum_{n=1}^{N} KL\left(q\left(z_n|\hat{z}_n,\Sigma_z\right)||p\left(z_n|0,I_K\right)\right)$$

$$= -\frac{N}{2} tr\left(\Sigma_z\right) - \frac{1}{2}\sum_{n=1}^{N} \hat{z}_n^T \hat{z}_n + \frac{KN}{2} + \frac{N}{2}\ln|\Sigma_z| \tag{2.69}$$

The final expression for the log likelihood as a summation of the above two terms can be expressed as the following,

$$\ln p(Y|W,\sigma) = \mathcal{L}_{LB}$$

$$= -\frac{DN}{2}\ln 2\pi - \frac{DN}{2}\ln\sigma - \frac{1}{2\sigma}\sum_{n=1}^{N} y_n^T y_n + \frac{1}{\sigma}\sum_{n=1}^{N} y_n^T W \hat{z}_n$$

$$-\frac{1}{2\sigma}\sum_{n=1}^{N} tr\left\{W^T W\left[\hat{z}_n \hat{z_n}^T + \Sigma_z\right]\right\}$$

$$-\frac{N}{2} tr\left(\Sigma_z\right) - \frac{1}{2}\sum_{n=1}^{N} \hat{z}_n^T \hat{z}_n + \frac{KN}{2} + \frac{N}{2}\ln|\Sigma_z| \tag{2.70}$$

The above expression for the log likelihood is relatively simpler in terms of the parameters W and σ when compared to the one derived in Eqn. (2.47). However, this simplicity is a result of introducing additional parameters $\hat{z}$ and Σ_z in the likelihood expression. These additional parameters in turn depend on W and σ and therefore, the level of non-linearity with respect to W and σ remains the same as that of the original log likelihood expression. However, the advantage of the log likelihood expression shown in Eqn. (2.70) is that it allows deriving explicit update expressions for

the parameters in terms of the parameters of the posterior distribution of the latent variables and the posterior distribution of the latent variables in terms of the model parameters. These explicit update expressions removes the implementation complexities associated with the gradient based optimization approaches such as optimal step length selection.

Parameter Updates

The optimal values of the parameters that maximize the log likelihood can be obtained by equating the derivatives of the log likelihood with respect to the parameters to zero. For the PPCA case, When equating the derivative of the log likelihood with respect to W to zero yields the following update expression for W,

$$\frac{d\mathcal{L}_{LB}}{dW} = 0 \Rightarrow$$

$$W = \sum_{n=1}^{N} y_n \hat{z}_n^T \left\{ \sum_{n=1}^{N} \left[\hat{z}_n \hat{z}_n^T + \Sigma_z \right] \right\}^{-1} \tag{2.71}$$

When equating the derivative of the log likelihood with respect to σ to zero yields the following update expression for σ,

$$\frac{d\mathcal{L}_{LB}}{d\sigma} = 0 \Rightarrow$$

$$\sigma = \frac{\sum_{n=1}^{N} y_n^T y_n + \sum_{n=1}^{N} tr\left\{ W^T W \left[\hat{z}_n \hat{z}_n^T + \Sigma_z \right] \right\} - 2 \sum_{n=1}^{N} y_n^T W \hat{z}_n}{DN} \tag{2.72}$$

Implementation of the EM algorithm

The update expressions for W and σ can be seen to be dependent on the parameters of the posterior distribution of Z, $\hat{z}$ and Σ_z. Similarly, the posterior parameter updates for Z depends on the model parameters as shown in Eqn. (2.67). Each of these updates takes the log likelihood function to a stationary point that is a local maxima with respect to the updated quantity. This can be verified from the hessian of the log likelihood with respect to the updated quantity. The hessian will always remain negative definite in the case of the PPCA model. As with each update, the log likelihood function is maximized and it is bounded by the maximum value of the

log likelihood function, when the updates are implemented iteratively, the estimate of the log likelihood eventually converges.

The EM algorithm implements the expectation and the maximization steps alternatively through multiple iterations. The expectation step involves updating the posterior distribution of the latent variables and evaluating the log-likelihood function. The maximization step involves updating the parameters that maximizes the log-likelihood function. The EM algorithm for the PPCA model is shown in Table. 2.1. To start with, we have to initialize the parameters W and σ. In the expectation step, we update $q(Z)$ based on the recent estimates of W and σ as shown in Eqn. (2.67) and perform the expectation respect to $q(Z)$ to evaluate $\mathcal{L}_{LB}$ as shown in Eqn. (2.70). In the maximization step, we update the parameters W and σ.

Table 2.1: EM algorithm for the estimation of the PPCA model

Initialize W and σ
repeat until convergence (check for the convergence of $\mathcal{L}_{LB}$)
    Expectation step:
        Update $q(Z)$ based on the recent estimates of W and σ using Eqn. (2.67)
        Evaluate $\mathcal{L}_{LB}$ using the recent update of $q(Z)$ using Eqn. (2.70)
    Maximization step:
        Update W and σ using equations (2.71) and (2.72)
end repeat

Simulation Example

For the purpose of illustration we simulated a 4-dimension dataset from a PPCA model with the following parameters,

$$W = \begin{bmatrix} 0.36 & -0.07 & 0.86 & 0.36 \\ 0.67 & 0.71 & -0.16 & -0.13 \end{bmatrix}^T, \quad \sigma = 0.01 \tag{2.73}$$

where the parameter W contains two independent column vectors. From the dimensions of W, this model can be pictured to explain four dimension observations with two dimension latent variables. We simulated $N = 700$ observations. From the simulated observations, we estimated the model parameters. We assumed that we know the dimension of the latent variables *a priori* as two and proceeded to estimate the model shown in Eqn. (2.73). Fig. 2.11 shows the estimate of $\mathcal{L}_{LB}$ against the number

of iterations in the EM algorithm. It can be seen that the estimate of $\mathcal{L}_{LB}$ improves with each iteration. After around 400 iterations, it saturates. By having a threshold on the rate of increase, we can assess the convergence.

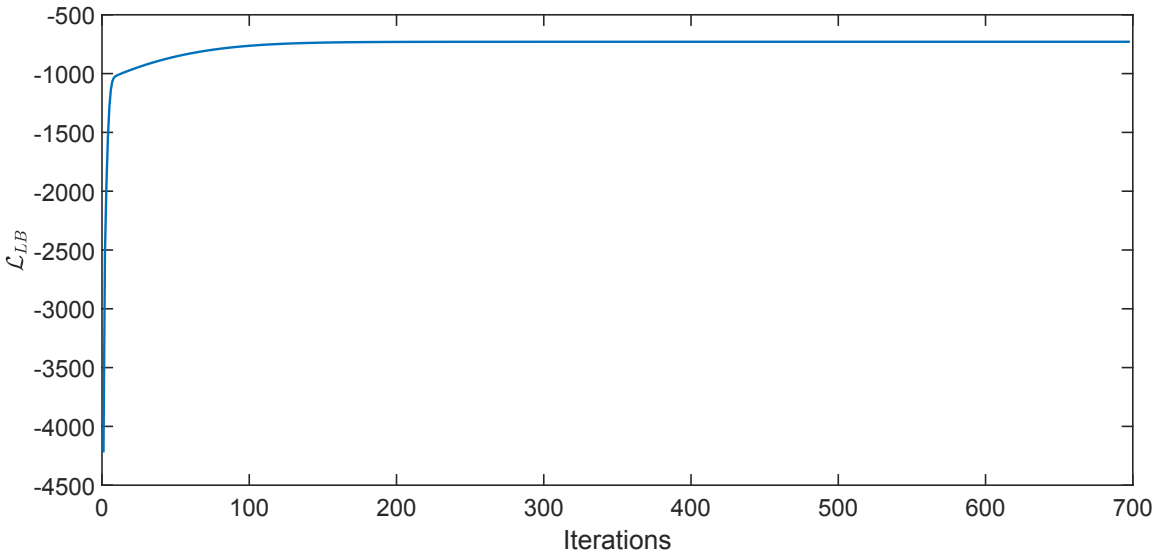


Figure 2.11: $\mathcal{L}_{LB}$ estimate vs. the number of iterations during the ML estimation of the PPCA model using the EM algorithm. It can be seen that $\mathcal{L}_{LB}$ estimate increases with each iteration.

The ML parameter estimates resulted from the estimation exercise is shown below,

$$W_{ML} \approx \begin{bmatrix} 0.1721 & -0.2559 & 0.8623 & 0.3763 \\ 0.7481 & 0.6740 & 0.0661 & -0.0397 \end{bmatrix}^T, \quad \sigma_{ML} \approx 0.0096 \tag{2.74}$$

where the subscript ML indicates that W_{ML} and σ_{ML} are the ML estimates. The estimate of the variance parameter σ_{ML} can be seen to be very close to the actual σ with which we simulated the data. The estimate of W_{ML} appears to be completely different from the actual W. However, this difference is due to the rotational ambiguity in the estimated coefficient matrix. We can estimate the rotational matrix $\mathcal{R}$ as shown below,

$$W \approx W_{ML}\mathcal{R} \tag{2.75}$$

$$\mathcal{R} \approx \begin{bmatrix} 0.9767 & -0.2567 \\ 0.2611 & 0.9550 \end{bmatrix} \tag{2.76}$$

The matrix $\mathcal{R}$, excluding minor numerical discrepancies, is an orthonormal matrix and this can be verified as shown below,

$$\mathcal{R}^T\mathcal{R} \approx I_2 \tag{2.77}$$

This suggests that the estimate W_{ML} spans the same subspace as W. This rotational ambiguity will be compensated by the rotation in Z without affecting the identification of the actual subspace.

2.4 Bayesian Analysis

In Bayesian analysis, the objective is to infer the posterior distributions of the unknowns in the model given the data. Let Y be the observed data, $\mathcal{M}$ be the model structure used to describe the data and Θ be the set of unknown parameters and latent variables in the model. In the case of Bayesian PPCA presented in Fig. 2.10, $\mathcal{M}$ corresponds to the space of PPCA models with different latent variable dimensions, K and Θ corresponds to the set $\{Z, W\}$.

The posterior distribution of the unknowns using the Bayes rule of probability can ideally be obtained as the following,

$$p\left(\Theta|Y,\mathcal{M}\right) = \frac{p\left(Y|\Theta,\mathcal{M}\right)p\left(\Theta|\mathcal{M}\right)}{p\left(Y|\mathcal{M}\right)} \tag{2.78}$$

where $p\left(\Theta|Y,\mathcal{M}\right)$, $p\left(Y|\Theta,\mathcal{M}\right)$ and $p\left(\Theta|\mathcal{M}\right)$ are the posterior, likelihood and prior of Θ respectively. The term in the denominator is the marginal distribution of the data given the model structure and it can also be interpreted as the likelihood of the considered model structure. It is also referred to as the model evidence. It can be obtained by integrating out Θ from the joint distribution term in the numerator as the following,

$$p\left(Y|\mathcal{M}\right) = \int_{\Theta} p\left(Y|\Theta,\mathcal{M}\right)p\left(\Theta|\mathcal{M}\right)d\Theta \tag{2.79}$$

When we have several different competing models, we can choose the model that has the maximum model evidence. The characteristics of the model evidence is shown Fig. 2.12. The illustration in Fig. 2.12 is adapted from [60, 66]. X-axis in the figure corresponds to space of observable data and Y-axis corresponds to the model evidence. Model evidences of three models with different complexities are illustrated. Let us say that we are interested in identifying a model that best describes the observed data Y_{given}. Let $\mathcal{M}_{simple}$ be a simpler model, $\mathcal{M}_{moderate}$ be a moderately complex model and $\mathcal{M}_{complex}$ be a complex model. For example, in our case, $\mathcal{M}_{simple}$, $\mathcal{M}_{moderate}$

and $\mathcal{M}_{complex}$ can be PPCA models with latent variable dimensions K_{simple}, $K_{moderate}$ and $K_{complex}$ where $K_{simple} < K_{moderate} < K_{complex}$. Since $p(Y|\mathcal{M})$ is a probability density or a mass function, integration/summation of $p(Y|\mathcal{M})$ over the space of Y should be equal to one. $\mathcal{M}_{complex}$ can describe a wider range of datasets as $K_{complex}$ will provide more degrees of freedom for the covariance matrix defined by the PPCA model. Therefore, $p(Y|\mathcal{M}_{complex})$ should be more spread out over the space of Y compared to the likelihoods of the other two models. The likelihood $p(Y|\mathcal{M}_{simple})$ would be more skewed as it can only explain specific datasets as K_{simple} will provide lesser number of degrees of freedom for the covariance matrix defined by the PPCA model. The likelihood $p(Y|\mathcal{M}_{moderate})$ will fall somewhere in-between, neither skewed nor more spread out. This property of $p(Y|\mathcal{M})$ becomes handy in model selection in Bayesian analysis. If our dataset can be described well by a simpler model, the likelihood of the simpler model will be much higher than the complex ones. If the dataset falls in the space where the likelihood of the simpler model is very low, then we would have no option other than describing the dataset with the complex ones. In this illustration, Y_{given} falls in the space where the simple model has a very low likelihood and the moderately complex model has a better likelihood the data than the complex model. If we were to pick one model among the three, we would pick $\mathcal{M}_{moderate}$ to be the appropriate model for describing Y_{given}. In this thesis, we make use of this particular advantage of Bayesian analysis for model selection.

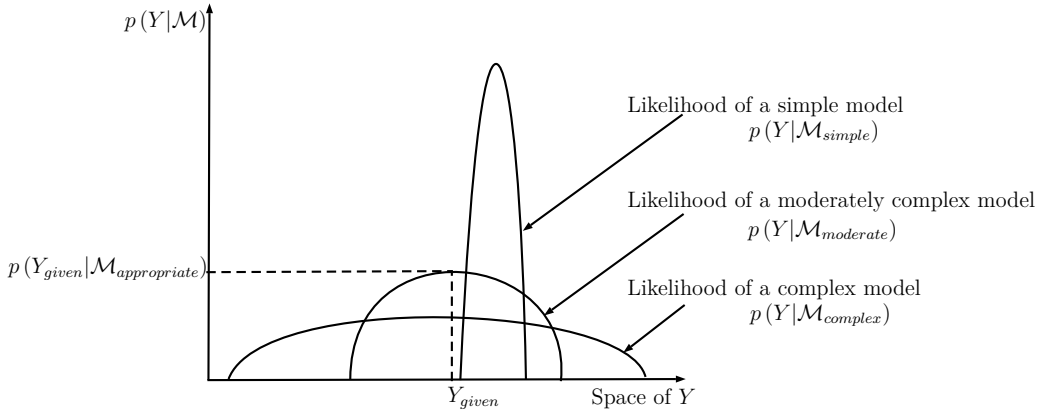


Figure 2.12: Model selection based on the model evidence/likelihood of the competing models. X-axis corresponds to the space of observable data. Y-axis corresponds to the model evidence.

One of the well recognized challenges in Bayesian estimation is that often the posterior distribution $p(\Theta|Y, \mathcal{M})$ will not have a recognizable form and the integral for estimating the model evidence shown in Eqn. (2.79) will be intractable. In these cases, we will have to move to approximate Bayesian estimation approaches. The approximate Bayesian estimation approaches can be classified into two categories, i) sampling based approaches and ii) deterministic approaches. In this thesis, we will predominantly make use of a deterministic algorithm called the VBEM algorithm. For the CEFGMs, the VBEM algorithm is one of the computationally efficient approaches and it provides a tractable estimation procedure [67, 68]. The VBEM approach has been illustrated for models that belong to the CEFGMs in the literature [67, 68, 69, 70, 71, 72] for its ability to identify the appropriate model structures and parameter estimates.

2.4.1 Variational Bayesian Expectation Maximization Algorithm

Recall the Bayesian PPCA model presented in Fig. 2.10. The joint distribution of all the variables in this model can be expressed as the following,

$$\begin{aligned} p(Y, Z, W, \nu|\alpha^*, \beta^*, \sigma) &= p(Y|Z, W, \sigma)\, p(Z)\, p(W|\nu)\, p(\nu|\alpha^*, \beta^*) \\ &= \prod_{k=1}^{K} p(w_k|\nu_k)\, p(\nu_k|\alpha^*, \beta^*) \prod_{n=1}^{N} p(y_n|W, z_n, \sigma)\, p(z_n) \end{aligned} \tag{2.80}$$

To be more accurate, we should include the model structure, $\mathcal{M}$, or the number of latent variables in the model, K, in the above joint distribution expression. However, let us take the case of a fixed model structure and ignore the specification of $\mathcal{M}$ for now. For this model, the Bayesian analysis translates into obtaining the posterior distribution $p(Z, W, \nu|Y, \alpha^*, \beta^*, \sigma)$. This posterior distribution can be expressed using the Bayes rule of probability as the following,

$$p(Z, W, \nu|Y, \alpha^*, \beta^*, \sigma) = \frac{p(Y|Z, W, \sigma)\, p(Z)\, p(W|\nu)\, p(\nu|\alpha^*, \beta^*)}{p(Y|\sigma, \alpha^*, \beta^*)} \tag{2.81}$$

We can express the posterior like in our previous examples in terms of the numerator in the above expression,

$$p(Z, W, \nu|Y, \alpha^*, \beta^*, \sigma) \propto p(Y|Z, W, \sigma)\, p(Z)\, p(W|\nu)\, p(\nu|\alpha^*, \beta^*) \tag{2.82}$$

However, unlike in our previous examples, recognizing the functional form of the posterior from the terms on the RHS of the above equation will be challenging. In fact, in the case of Bayesian PPCA model, it will not have a recognizable form. Also, the integration to obtain the model evidence in the denominator of Eqn. (2.81) is not tractable. Therefore, we will have to utilize approximate Bayesian analysis techniques such as the VBEM algorithm.

Similar to the EM algorithm, the VBEM algorithm can also be implemented for a model that belongs to CEFGMs using the following steps,

1. The first advancing step is deriving the VBEM algorithm requires approximating the posterior distributions of the unknowns and deciding the structure of the approximated posterior distributions.

2. The second advancing step involves deriving a functional that lower bounds the log model evidence in terms of the joint distribution of the variables in the model and the approximated posterior distributions.

3. The third advancing step involves determining the distribution families of the approximated posterior distributions.

4. The fourth advancing step involves deducing the lower bound expression with the determined posterior distribution families.

5. The fifth advancing step involves deriving the update expressions for the posterior distributions and the fixed parameters in the model.

6. The sixth advancing step involves implementation of an iterative procedure using the update expressions derived in the previous step.

We illustrate these steps for the Bayesian PPCA model in the rest of this section.

Approximate Posterior Distribution

The VBEM algorithm requires us to to approximate the joint posterior distribution as product of individual distributions of the unknowns. To estimate the Bayesian PPCA

model using the VBEM algorithm, we would have to make the following approximation,

$$p\left(Z, W, \nu | Y, \alpha^*, \beta^*, \sigma\right) \approx q\left(Z\right) q\left(W\right) q\left(\nu\right) \tag{2.83}$$

where $q\left(Z\right)$, $q\left(W\right)$ and $q\left(\nu\right)$ are the approximated posterior distributions of Z, W and ν respectively. By expressing the joint posterior as a product of individual posterior distributions, we made an approximation that Z, W and ν are mutually independent given Y. However for the network shown in Fig. 2.10, the D-separation rules suggest that W and Z are not independent given Y as they are co-parents of Y, and W and ν are not independent as they share a parent-child relationship. These approximations come as a trade off for the tractability.

Lower Bound on the Log Model Evidence

We present the derivation of the lower bound on the log model evidence of the Bayesian PPCA model through Proposition 4.

Proposition 4. *Lower bound on the model evidence: The natural logarithm of the model evidence can be expressed as a sum of two terms, i) a term that lower bounds the log model evidence and ii) the KL divergence between the approximated posterior distribution and the actual posterior distribution of the unknowns. For the Bayesian PPCA model, this translates into the following,*

$$\ln p\left(Y | \alpha^*, \beta^*, \sigma\right) = \mathcal{L}_{VB} + KL\left(q\left(\nu\right) q\left(W\right) q\left(Z\right) || p\left(\nu, W, Z | Y, \alpha^*, \beta^*, \sigma\right)\right) \tag{2.84}$$

and

$$\ln p\left(Y | \alpha^*, \beta^*, \sigma\right) \geq \mathcal{L}_{VB} \tag{2.85}$$

where $\mathcal{L}_{VB}$ is the functional of the posterior distributions $q\left(\nu\right)$, $q\left(W\right)$ and $q\left(Z\right)$ and the joint distribution of all the variables in the network, referred to as the variational lower bound.

Proof. From Eqn. (2.81), using the same trick in the proof of proposition 3, we can obtain the following expression,

$$\ln p\left(Y | \alpha^*, \beta^*, \sigma\right) = \int_{\nu} q\left(\nu\right) \int_{W} q\left(W\right) \int_{Z} q\left(Z\right) \ln \frac{p\left(Y, Z, W, \nu | \alpha^*, \beta^*, \sigma\right)}{q\left(\nu\right) q\left(W\right) q\left(Z\right)} d\nu dW dZ$$

$$+\int_{\nu} q(\nu) \int_{W} q(W) \int_{Z} q(Z) \ln \frac{q(\nu)\, q(W)\, q(Z)}{p(W, Z, \nu | Y, \alpha^*, \beta^*, \sigma)} d\nu dW dZ \tag{2.86}$$

where the first term is $\mathcal{L}_{VB}$ and the second term is the KL divergence between the approximated posterior and the actual posterior distributions of the unknowns as expressed below,

$$\ln p(Y|\alpha^*, \beta^*, \sigma) = \mathcal{L}_{VB} + KL\left(q(Z)\, q(W)\, q(\nu) \,||p(Z, W, \nu|Y, \alpha^*, \beta^*, \sigma)\right) \tag{2.87}$$

Since the KL divergence is always positive, the term $\mathcal{L}_{VB}$ lower bounds the log model evidence, $\ln p(Y|\alpha^*, \beta^*, \sigma)$. Therefore, we can state the following,

$$\ln p(Y|\alpha^*, \beta^*, \sigma) \geq \mathcal{L}_{VB} \tag{2.88}$$

This completes the proof of Proposition 4. ■

The term $\mathcal{L}_{VB}$ can further be expanded by splitting the integral as a summation of multiple integrals based on the assumption of factorized posterior distributions as the following,

$$\begin{aligned} \mathcal{L}_{VB} = &\int_{\nu} q(\nu) \ln \frac{p(\nu|\alpha^*, \beta^*)}{q(\nu)} d\nu + \int_{\nu} q(\nu) \int_{W} q(W) \ln \frac{p(W|\nu)}{q(W)} d\nu dW \\ &+ \int_{Z} q(Z) \ln \frac{p(Z)}{q(Z)} dZ + \int_{W} q(W) \int_{Z} q(Z) \ln p(Y|W, Z, \sigma)\, dW dZ \end{aligned} \tag{2.89}$$

Distribution Families of the Approximated Posteriors

D-separation rules dictate that each of $q(W)$, $q(\nu)$ and $q(Z)$ can be factored further. The rows of W become independent of each other given Y as each row is a parent of a particular of dimension of Y and does not share its child with any other rows of W. Additionally, W is already assumed to be independent of its parent ν and the co-parents Z with which it shares its children Y. Therefore, we can express $q(W)$ as a product of multiple factors as the following,

$$q(W) = \prod_{d=1}^{D} q\left(w^d | \hat{w}^d, \Sigma_{w^d}\right) \tag{2.90}$$

where w^d is the dth column of W. Similarly, $q(\nu)$ can be expressed as a product of multiple factors as the following,

$$q(\nu) = \prod_{k=1}^{K} q(\nu_k | \alpha, \beta_k) \tag{2.91}$$

where ν_k is the parent of column w_k and it does not share its child with any other $\nu_{i\neq k}$. The posterior of the latent variable Z can also be expressed as a product of multiple factors as the following,

$$q(Z) = \prod_{n=1}^{N} q(z_n|\hat{z}_n, \Sigma_z) \tag{2.92}$$

where $q(z_n)$ is the posterior of z_n and z_n does not share its only child y_n with any $z_{i\neq n}$. Further, these factors will also belong to the same distribution families as that of the respective priors due to conjugacy. Factor $q\left(w^d|\hat{w}^d, \Sigma_{w^d}\right)$ is a multivariate Gaussian distribution with mean $\hat{w}^d$ and covariance Σ_{w^d}, factor $q(\nu_k|\alpha, \beta_k)$ is a gamma distribution with parameters α and β_k, and factor $q(z_n|\hat{z}_n, \Sigma_z)$ is a multivariate Gaussian distribution with mean $\hat{z}_n$ and covariance Σ_z[‡],[§].

Deducing the Lower Bound Expression

For the Bayesian PPCA model, this amounts to deducing the $\mathcal{L}_{VB}$ expression from Eqn. (2.89). The integrals in Eqn. (2.89) are tractable. In fact, for all the CEFGMs, the variational Bayesian lower bound expressions are deducible. The details of deducing all the terms in Eqn. (2.89) are shown below,

Term I:

$$\int_{\nu} q(\nu) \ln \frac{p(\nu|\alpha^*, \beta^*)}{q(\nu)} d\nu = -\sum_{k=1}^{K} KL\left(q(\nu_k|\alpha, \beta_k)\,||p(\nu_k|\alpha^*, \beta^*)\right)$$

$$= -\sum_{k=1}^{K} \alpha \ln \beta_k + K\alpha^* \ln \beta^* + K \ln \frac{\Gamma(\alpha)}{\Gamma(\alpha^*)}$$

$$-\sum_{k=1}^{K} (\alpha - \alpha^*)(\Psi(\alpha) - \ln \beta_k) + \sum_{k=1}^{K} \alpha \left(1 - \frac{\beta^*}{\beta_k}\right) \tag{2.93}$$

Term II:

$$\int_{W} q(W) \ln \frac{p(W|\nu)}{q(W)} dW$$

[‡]Note that the shape parameter in the gamma distribution, α does not take any subscript. This is because $\alpha_1 = \alpha_2 = ... = \alpha_K = \alpha$ and it will become clear why this is the case in the subsequent derivations

[§]Note that the covariance of z_n, Σ_z is independent of the value that n takes. This is because $\Sigma_{z_1} = \Sigma_{z_2} = ... = \Sigma_{z_n} = \Sigma_z$ and it will become clear why this is the case in the subsequent derivations

$$= -\sum_{d=1}^{D} KL\left(q\left(w^d|\hat{w}^d,\Sigma_{w^d}\right)||p\left(w^d|0,\left(diag\left(\left[\nu_1,...,\nu_K\right]^T\right)\right)^{-1}\right)\right)$$

$$= -\frac{1}{2}\sum_{d=1}^{D} tr\left(diag\left(\left[\nu_1,...,\nu_K\right]\Sigma_{w^d}\right)\right) - \frac{1}{2}\sum_{d=1}^{D}\hat{w}^d diag\left(\left[\nu_1,...,\nu_K\right]\right)\left(\hat{w}^d\right)^T$$

$$+\frac{KD}{2} + \frac{D}{2}\sum_{k=1}^{K}\ln\nu_k + \frac{1}{2}\sum_{d=1}^{D}\ln|\Sigma_{W^d}| \tag{2.94}$$

$$\int_{\nu} q\left(\nu\right)\int_{W} q\left(W\right)\ln\frac{p\left(W|\nu\right)}{q\left(W\right)}d\nu dW$$

$$= -\frac{1}{2}\sum_{d=1}^{D} tr\left(\Lambda\Sigma_{w^d}\right) - \frac{1}{2}\sum_{d=1}^{D}\hat{w}^d\Lambda\left(\hat{w}^d\right)^T$$

$$+\frac{KD}{2} + \frac{D}{2}\sum_{k=1}^{K}\left(\Psi\left(\alpha\right) - \ln\beta_k\right) + \frac{1}{2}\sum_{d=1}^{D}\ln|\Sigma_{w^d}| \tag{2.95}$$

where

$$\Lambda = diag\left(\left[\frac{\alpha}{\beta_1},...,\frac{\alpha}{\beta_K}\right]\right) \tag{2.96}$$

Term III:

$$\int_{Z} q\left(Z\right)\ln\frac{p\left(Z\right)}{q\left(Z\right)}dZ = -\sum_{n=1}^{N} KL\left(q\left(z_n|\hat{z}_n,\Sigma_z\right)||p\left(z_n|0,I_K\right)\right)$$

$$= -\frac{N}{2}tr\left(\Sigma_z\right) - \frac{1}{2}\sum_{n=1}^{N}\hat{z}_n^T\hat{z}_n + \frac{KN}{2} + \frac{N}{2}\ln|\Sigma_z| \tag{2.97}$$

Term IV:

$$\int_{W} q\left(W\right)\int_{Z} q\left(Z\right)\ln p\left(Y|W,Z,\sigma\right)dZdW$$

$$= -\frac{DN}{2}\ln 2\pi - \frac{DN}{2}\ln\sigma - \frac{1}{2\sigma}\sum_{n=1}^{N} y_n^T y_n + \frac{1}{\sigma}\sum_{n=1}^{N} y_n^T\hat{W}\hat{z}_n$$

$$-\frac{1}{2\sigma}\sum_{n=1}^{N}\sum_{d=1}^{D} tr\left\{\left[\hat{z}_n\hat{z}_n^T + \Sigma_{z_n}\right]\left[\left(\hat{w}^d\right)^T\hat{w}^d + \Sigma_{w^d}\right]\right\} \tag{2.98}$$

The lower bound expression $\mathcal{L}_{VB}$ may never become equal to the model evidence. This is due to the gap between the approximated posterior distribution and the actual posterior distribution. The gap will be equal to the KL divergence between the approximated posterior distribution and the actual posterior distribution as shown in proposition 4. The summation of the lower bound term and the KL divergence term

is bounded by the log model evidence. If we minimize the KL divergence term, it will take the lower bound closer to the model evidence. Conversely, if we maximize the lower bound, it will minimize the KL divergence between the approximated posterior and the actual posterior. The VBEM algorithm maximizes the lower bound with respect to the approximated posterior distributions such that we get a reasonable approximation of the log model evidence in terms of the lower bound and also the posterior distributions of the unknowns that are practically closer to the actual posterior distribution.

Posterior and Parameter Updates

The lower bound expression is not necessarily concave for all CEFGMs. However, they are concave with respect to each individual posterior distributions and the model parameters when the rest of the posteriors and the model parameters are fixed. For instance, for the Bayesian PPCA model, this can be verified from the first order and second derivates of the lower bound expression with respect to the individual posteriors and the parameters. The VBEM algorithm makes use of this property and allows one to derive updates for the individual posterior distributions and the parameters. Implementing these updates iteratively maximizes the lower bound.

The updates for the posterior distributions and the parameters can be derived by taking the derivative of $\mathcal{L}_{VB}$ with respect to the posteriors distributions and the parameters and equating the derivatives to zero. For the Bayesian PPCA model, the update expressions can be obtained as shown below,

Update expression for $q(Z)$[¶]:

$$\frac{d\mathcal{L}_{VB}}{dq(Z)} = 0 \Rightarrow \ln q(Z) = \ln p(Z) + \int_W q(W) \ln p(Y|W, Z, \sigma)\, dW$$

$$\ln q(z_n|\hat{z}_n, \Sigma_z) \propto \ln p(z_n) + \int_W q(W) \ln p(y_n|W, z_n, \sigma)\, dW$$

$$\Sigma_z = \sigma \left[\sum_{d=1}^{D} \left\{ \left(\hat{w}^d\right)^T \hat{w}^d + \Sigma_{w^d} \right\} + \sigma I_K \right]^{-1}$$

$$\hat{z}_n = \frac{1}{\sigma} \Sigma_z \hat{W}^T y_n \tag{2.99}$$

[¶]Notice the expression for Σ_z, it is independent of n, which was previously pointed in foot note ‡

Update expression for $q(W)$:

$$\frac{d\mathcal{L}_{VB}}{dq(W)} = 0 \Rightarrow \ln q(W)$$

$$= \ln p(W|\nu) + \int_{\nu} q(\nu) \ln p(W|\nu)\, d\nu + \int_{Z} q(Z) \ln p(Y|W, Z, \sigma)\, dZ$$

$$\ln q\left(w^d|\hat{w}^d, \Sigma_{w^d}\right) \propto \ln p\left(w^d|\nu\right) + \int_{\nu} q(\nu) \ln p\left(w^d|\nu\right) d\nu$$

$$+ \int_{Z} q(Z) \ln p(Y|W, Z, \sigma)\, dZ$$

$$\Sigma_{w^d} = \sigma \left[\sum_{n=1}^{N} \left\{\hat{z}_n \hat{z}_n^T + \Sigma_z\right\} + \sigma\Lambda\right]^{-1}$$

$$\left(\hat{w}^d\right)^T = \frac{1}{\sigma} \Sigma_{w^d} \hat{z}_n y_n^d \tag{2.100}$$

Update expression for $q(\nu)$[‖]:

$$\frac{d\mathcal{L}_{VB}}{dq(\nu)} = 0 \Rightarrow \ln q(\nu) = \ln p(\nu|\alpha^*, \beta^*) + \int_{W} q(W) \ln \frac{p(W|\nu)}{q(W)} dW$$

$$\ln q(\nu_k|\alpha, \beta_k) \propto \ln p(\nu_k|\alpha^*, \beta^*) + \int_{W} q(W) \ln p(W|\nu)\, dW$$

$$\alpha = \alpha^* + \frac{D}{2}$$

$$\beta_k = \beta^* + \frac{1}{2} \sum_{d=1}^{D} \left[\left(\hat{w}_k^d\right)^2 + \Sigma_{w^d{}_k}^k\right] \tag{2.101}$$

Update expression for σ:

$$\frac{d\mathcal{L}_{VB}}{d\sigma} = 0 \Rightarrow$$

$$\sigma = \frac{\sum_{n=1}^{N} y_n^T y_n + \sum_{n=1}^{N} tr\left(\left[\sum_{d=1}^{D} \left\{\left(\hat{w}^d\right)^T \hat{w}^d + \Sigma_{w^d}\right\}\right] \left[\hat{z}_n \hat{z}_n^T + \Sigma_z\right]\right) - 2\sum_{n=1}^{N} y_n^T \hat{W} \hat{z}_n}{ND} \tag{2.102}$$

[‖]Notice the expression for α, it is independent of k , which was previously pointed in footnote §

Implementation of the VBEM Algorithm

Implementing the update expressions derived above iteratively amounts to the VBEM estimation algorithm for the Bayesian PPCA model. One form of implementation is shown in Table. 2.4.1. In the expectation step of the algorithm, all the posteriors are updated and in the maximization step, the parameters are updated. These updates are carried out until $\mathcal{L}_{VB}$ converges. Updates in each iteration maximizes $\mathcal{L}_{VB}$ and $\mathcal{L}_{VB}$ is bounded by the actual log model evidence. Therefore, the algorithm is guaranteed to converge. As stated before, the lower bound may not be concave for a given model and therefore, it may not converge to a global maxima. In this case, the estimation has to be repeated with several different initial guesses.

Table 2.2: VBEM algorithm for the estimation of the Bayesian PPCA model

Initialize $q(W)$, σ, α^* and β^*
repeat until convergence (check for the convergence of $\mathcal{L}_{VB}$)
    Expectation step:
        Update $q(\nu)$
        Update $q(Z)$
        Update $q(W)$
        Obtain the expression for $\mathcal{L}_{VB}$
    Maximization step
        Estimate σ that maximizes $\mathcal{L}_{VB}$
end repeat

Simulation Example

Recall our simulation example in Eqn. (2.73). We implemented the VBEM algorithm derived above to estimate this model from the simulated data. We fixed the number of latent variables in this case to 3 and we fixed the hyperparameters α^* and β^* to 1 and estimated the model parameters. Fig. 2.13 shows the $\mathcal{L}_{VB}$ estimates against the number of iterations. It can be seen that $\mathcal{L}_{VB}$ improves with each iteration and converges after a few hundreds of iterations. For this case, the parameter estimates obtained are shown in Eqn. (2.103). These parameters correspond to the posterior mean of the loading matrix ($\hat{W}$) and the optimal estimate of the variance parameter σ that maximizes $\mathcal{L}_{VB}$. We are not ready to interpret the parameters as we do not have the optimal choice for the hyperparameters yet. In the next subsection, we will

discuss hyperparameter selection.

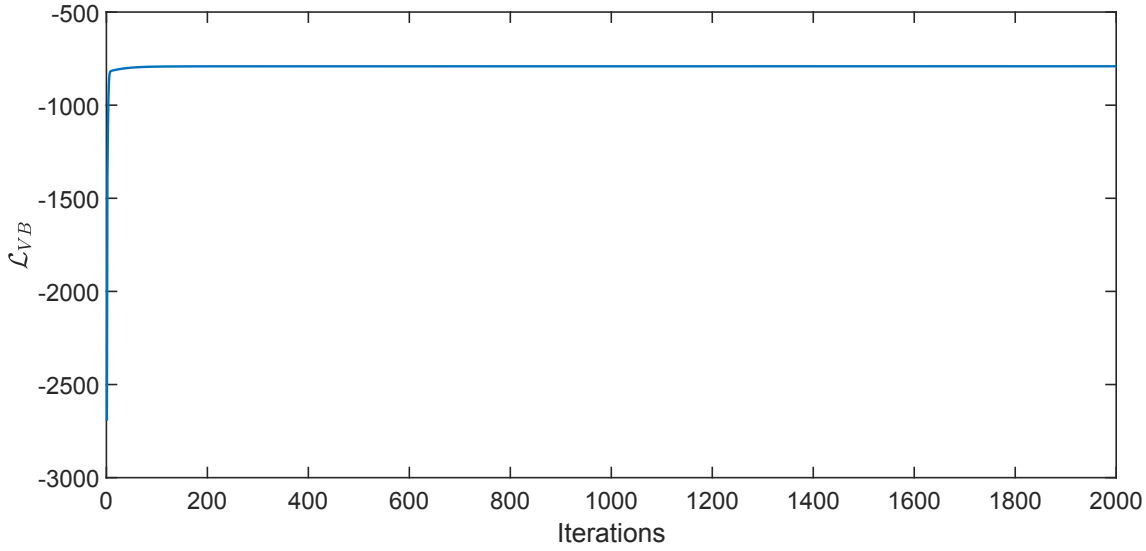


Figure 2.13: Lower bound estimate against the number of iterations during the VBEM estimation of the Bayesian PPCA model.

$$W_{VB} = \begin{bmatrix} -0.2306 & 0.2016 & -0.8642 & -0.3717 \\ -0.0154 & 0.0158 & 0.0095 & -0.0038 \\ 0.7316 & 0.6916 & -0.0024 & -0.0693 \end{bmatrix}^T, \quad \sigma_{VB} = 0.0094 \tag{2.103}$$

2.4.2 Hyperparameter Selection

Selection of hyperparameters plays an important role in the estimation and further analysis. If a researcher has a reasonable guess or belief about the range in which the parameters lie, the belief can be incorporated through the appropriate choice of hyperparameters. The effect of hyperparameters on the parameter estimates can be understood from the update expressions for the posterior of the parameters. When compared to the maximum likelihood update expression in Eqn. (2.71), the only extra term that appears in the update expression in the VBEM algorithm shown in Eqn. (2.100), is Λ. The term Λ adds a penalty or a regularization for each column of the coefficient matrix. For column one, it adds the penalty $\frac{\alpha}{\beta_1}$, for column two, it adds $\frac{\alpha}{\beta_2}$ and so forth. For an arbitrary column k, it adds the penalty $\frac{\alpha}{\beta_k}$, which is given by,

$$\frac{\alpha}{\beta_k} = \frac{\alpha^* + \frac{D}{2}}{\beta^* + \frac{1}{2}\sum_{d=1}^{D}\left[\left(\hat{w}_k^d\right)^2 + \Sigma_{w_k^d}^k\right]} \tag{2.104}$$

where α and β_k are expressed in terms of their update expressions shown in Eq. (2.101). This can also be equivalently represented as the following,

$$\frac{\alpha}{\beta_k} = \frac{\alpha^* + \frac{D}{2}}{\beta^* + \frac{1}{2}\sum_{d=1}^{D} E\left(\left(w_k^d\right)^2\right)} \tag{2.105}$$

where $E\left(\left(w_k^d\right)^2\right)$ is the posterior expectation of $\left(w_k^d\right)^2$, which is given by the sum of squared posterior mean and posterior variance.

The effect of choice of the hyperparameters on the penalty added to the parameter estimates is illustrated in Fig. 2.14. The plots in Fig. 2.14 illustrate this for two cases, one for decreasing β^* on the left panel and the other for increasing α^*. Y-axis in the plots corresponds to the penalty term added and the x-axis corresponds to the sum of expected value of the square of coefficients. For a fixed value of α^*, when β^* is decreased, as shown by the direction of the dashed arrow, the penalty on the smaller valued coefficients increases rapidly and the penalty on the larger valued coefficients does not increase appreciably. This effect regularizes the smaller valued coefficients significantly and forces them to converge to zero and leaves the larger valued coefficients relatively unaffected. The scale of parameters that one wants to regularize can be dictated by the choice of α^* and remains proportional to α^*. If we increase α^*, then the larger valued coefficients will also be penalized. This can be observed from the panel on the right. As we increase α^* for a fixed value of β^*, we can see that the penalty increases also for the larger valued coefficients. The penalty curve in this case inflates along the diagonal direction as indicated by the dashed arrow on the right panel. Therefore, the strategy for hyperparameter selection really depends on the purpose of the analysis. If one wants to infer the significance of the parameters in the model for a fixed value of alpha (depending on the scale of parameters that is considered to be significant), one can start with a higher value of β^* and continue to decrease β^* until all the insignificant parameters approach to zero. If one is interested in a variable selection problem as in the regression analysis, keeping β^* to a lower value and changing α^* incrementally would allow one to analyse the relative importance of the predictors in the model. At low values of α^*, lower valued coefficients converge close to zero, as we increase α^*, more and more parameters will start converging to zero.

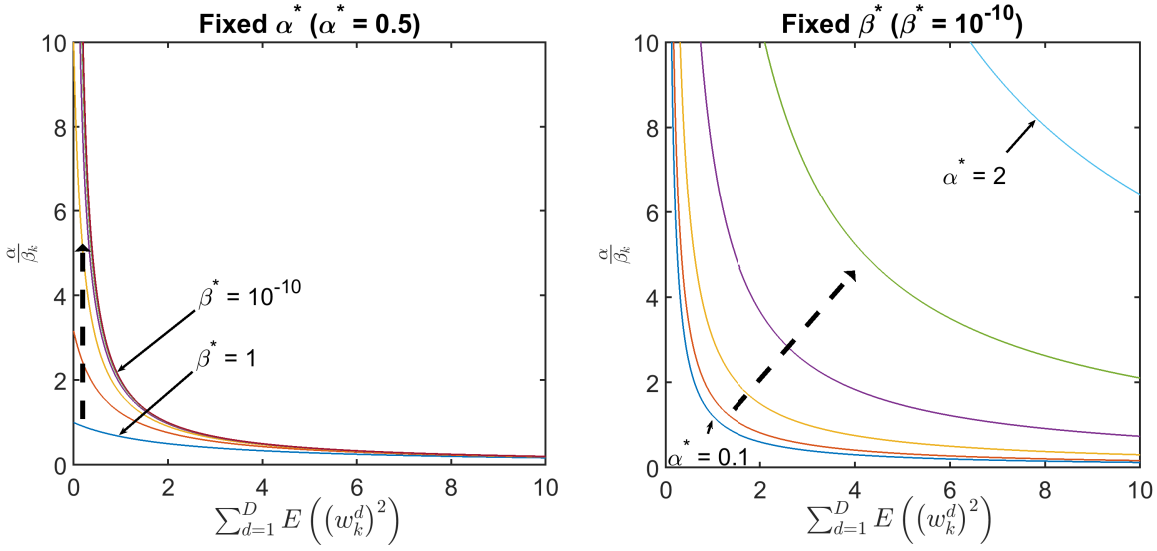


Figure 2.14: Effect of α^* and β^* on the penalty added to the parameter estimates. Left: Effect of decreasing β^* on the penalty and right: Effect of increasing α^* on the penalty. Dashed arrows indicate the direction of increase in the penalty term.

Hyperparameter Selection Through Cross-Validation

One of the strategies for hyperparameter selection is cross-validation. Simplest form of cross-validation involves splitting the available training data into two subsets, one for training the model, referred to as the training set and the other for validation, referred to as the validation set. The models are identified using the training set with different choices of hyperparameters and validated against the validation set. The choice of hyperparameters that provide the best validation performance is retained for further use. In this thesis, we use the log likelihood of the parameters in the validation set as the validation criteria. The optimal hyperparameter can be obtained using any stochastic optimization approach or a grid search technique that does not require the exact model between the log likelihood in the validation set and the hyperparameters. Here, we illustrate the use of Bayesian optimization using the 'Bayesopt' function in MATLAB for hyperparameter selection. Bayesian optimization is a surrogate model based optimization approach.

The 'Bayesopt' function in MATLAB performs minimization of the objective function with respect to the decision variable. Therefore, instead of maximizing the log likelihood, we pose the problem as a minimization problem for the negative log like-

lihood as shown below,

$$\alpha^*_{selected} = \min_{\alpha^*} -\ln \left\{ \prod_{n=1}^{N_{val}} p\left(y_n|0, W_{VB}W_{VB}^T + \sigma_{VB}I_D\right) \right\} \tag{2.106}$$

where $\prod_{n=1}^{N_{val}} p\left(y_n|0, W_{VB}W_{VB}^T + \sigma_{VB}I_D\right)$ refers to the Gaussian likelihood of the parameters with mean zero and covariance $W_{VB}W_{VB}^T + \sigma_{VB}I_D$ and N_{val} refers to the number of validation data points. The negative log likelihood is optimized with respect to α^* for a fixed value of β^*.

Simulation Example

We performed validation based hyperparameter selection for the Bayesian PPCA simulation example. We simulated additional 300 data points from the same model shown in Eqn. (2.73) and used them for validation. We fixed the value of β^* to 10^{-1} for this case and optimized the log likelihood with respect to α^*. Fig. 2.15 shows the optimization results for our simulation example. It shows the surrogate model prediction trend for the negative log likelihood with respect to α^* after 50 iterations (corresponding to the 50 sampled points). The blue dots correspond to the sampled points in those 50 iterations, the red trend shows the surrogate model predictions, the cyan trends correspond to the uncertainty bounds on the model prediction and the blue trends correspond to the uncertainty bounds on the modelling uncertainty. It can be seen that the negative log likelihood continues to decrease with increase in α^* till $\approx 10^{0.5}$ and increases with increase in α^* beyond $\approx 10^{0.5}$. For $\alpha^* \leq 10^{0.5}$, the model must be over-fitting the training samples. For $\alpha^* \geq 10^{0.5}$, the model must be under-fitting the training samples. Therefore, $\approx 10^{0.5}$ is a good choice for α^*.

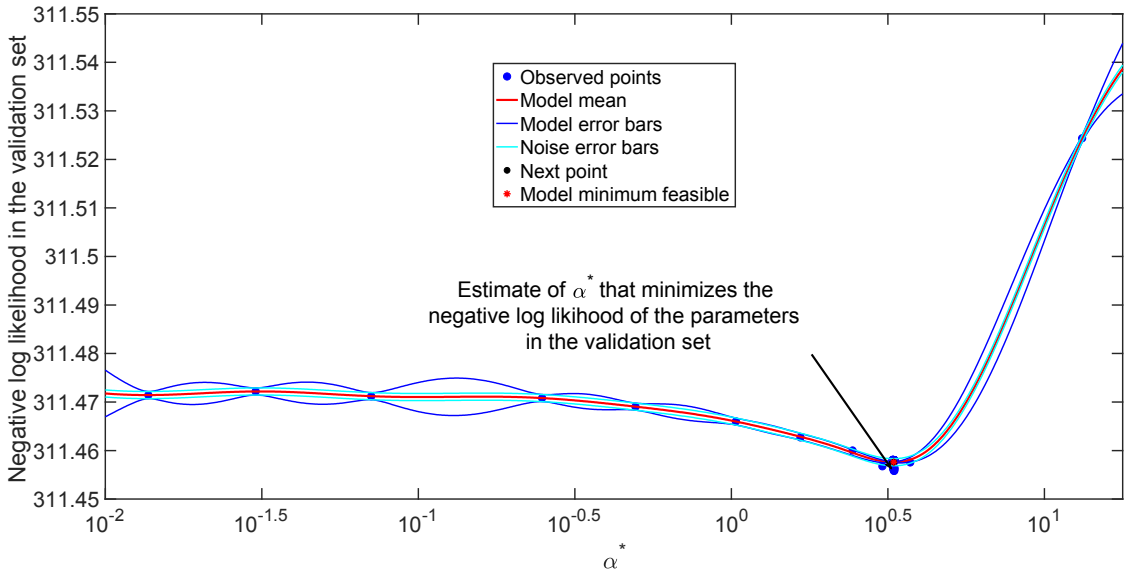


Figure 2.15: Selection of α^* through cross-validation. In this case, Bayesian optimization is employed to select α^* that minimizes the negative log likelihood in the validation dataset.

2.4.3 Model Selection or Dimension Reduction through Automatic Relevance Determination

For our simulation case study, we simulated data from a PPCA model with 2 dimension latent variables. In reality, we may not know the ideal number latent variables to choose when estimating the model from a given dataset. We started with a Bayesian PPCA model with 3-dimension latent variables. This model is the most complex PPCA model that we can choose for a 4 dimension dataset. However, this model may not be ideal for the given dataset. In this subsection, we address the problem of model selection.

We start with a reasonably complex model and estimate the model parameters for an optimal choice of the hyperparameters (through cross-validation). This complex model can be systematically simplified to a simpler version that is more suitable for the given dataset. For the Bayesian PPCA model, simplifying amounts to reducing the dimension of the latent variables or equivalently, to setting some of the columns of the coefficient matrix to zero. We can determine whether setting a particular column to zero fits the dataset better or leaving as it is fits the dataset better. This is a hypothesis selection problem for an arbitrary column k of the loading matrix i.e.,

which of the following hypothesis is more suitable for modelling the given dataset,

$$\begin{aligned} H_0 &: w_k \sim \mathcal{N}\left(0, \nu_k^{-1} I\right), \nu_k \sim Ga\left(\alpha^*, \beta^*\right) \\ H_1 &: w_k \sim \delta\left(0\right) \end{aligned} \tag{2.107}$$

where H_0 is the hypothesis defined by the Bayesian PPCA model for the column k of W and H_1 is a Dirac delta distribution which assigns zero density to any w_k other than zero column.

The VBEM algorithm provides us with a lower bound on the log model evidence for the complex model that was initially chosen. We switch the hypotheses of the columns of W one by one and assess if the variational lower bound of the reduced model improves with dimension reduction. If it improves, then we can retain the reduced dimension model. If it does not improve we can stop reducing the dimension of the model. The choice of which column of W should be removed or assigned to H_1 first can be determined by the relative significance of the columns of W. The sum of square of posterior means of the elements in a column ($\sum_{d=1}^{D} E\left(\left(w_k^d\right)^2\right)$) helps us determine the significance of that particular column. If it is high, then the column contains significantly non-zero coefficients. If it is close to zero, then the column contains coefficients close to zero. We prioritize and switch the hypothesis of those relatively insignificant columns first. If r number of columns to be removed from the loading matrix, then it will be the r relatively insignificant columns. We can formulate the selection of r as an optimization problem shown below,

$$K_{selected} = \min_{K=K_{initial}-r} -\mathcal{L}_{VB}\left(K\right) \tag{2.108}$$

where $K_{selected}$ is the number of retained columns or the dimension of the latent variables and $K_{initial}$ is the dimension of the latent variables of the complex model that we started with. This optimization picks the latent variable dimension for which the variational lower bound is maximum. We can use the same 'Bayesopt' function to solve the above optimization problem. Every time the dimension is reduced, the parameters have to the re-optimized. However, this re-optimization may not take many number of iterations as we would start from an already converged model.

Simulation Example

The optimization result for our simulation example is shown in Fig. 2.16. It can be seen from the trend that negative $\mathcal{L}_{VB}$ has the minimum value when r is equal to 1. We initially started with a model of 3 latent variables. Now, with $r = 1$, we exclude one of the latent variables from the model. Therefore, the latent variables dimension becomes 2. The parameter estimates of the reduced model are shown in Eqn. (2.109). The subscript VBR corresponds to the parameters estimates of the reduced model. The variance estimate σ_{VBR} lies close to the actual σ with which we generated the data. Similar to the ML estimates, W_{VBR} also has a rotational ambiguity, which can be verified from the estimated rotational matrix shown in Eqn. (2.110). The rotational matrix is approximately an orthonormal matrix. Therefore, the estimated loading matrix spans the same subspace as the original subspace from which the data was generated.

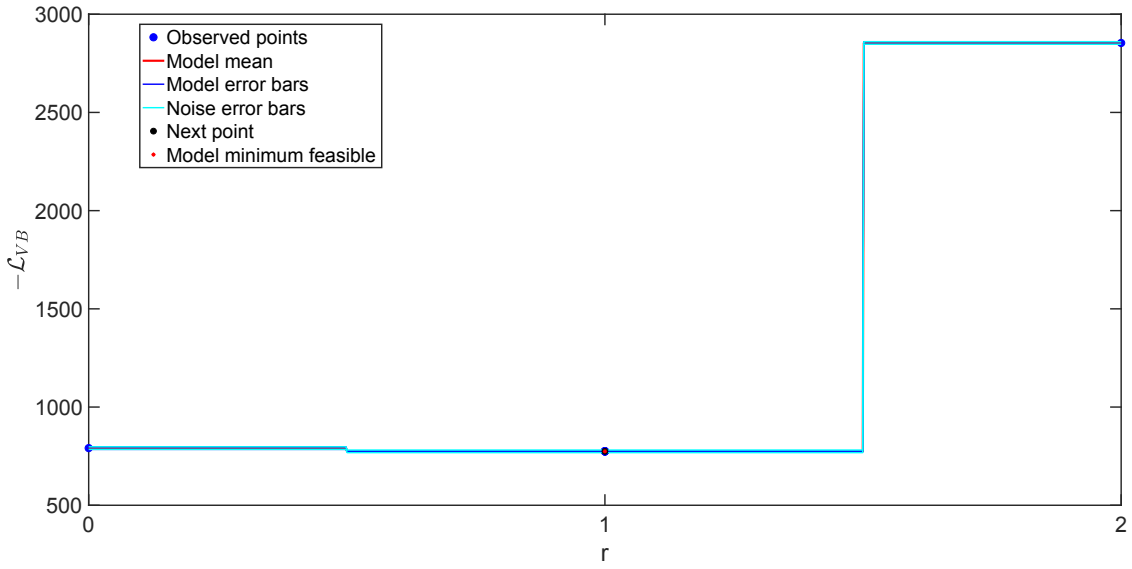


Figure 2.16: Model reduction by Bayesian optimization. The parameter r represents the number of latent variables excluded from the original model.

$$W_{VBR} \approx \begin{bmatrix} -0.2306 & 0.2016 & -0.8640 & -0.3716 \\ 0.7315 & 0.6915 & -0.0024 & -0.0693 \end{bmatrix}^T, \qquad \sigma_{VBR} \approx 0.0097 \tag{2.109}$$

$$\mathcal{R} \approx \begin{bmatrix} -0.9951 & 0.1805 \\ 0.1831 & 0.9732 \end{bmatrix} \tag{2.110}$$

2.5 Summary

This chapter reviewed the fundamentals associated with the BNs and CEFGMs. Then, we showed how data-driven models can be defined using the BNs. To illustrate the ML estimation and Bayesian analysis algorithms used in this thesis, we formulated ML and Bayesian version of the PCA model. The EM and VBEM algorithms are commonly used algorithms for the ML estimation and Bayesian analysis of data-driven models that belong to the CEFGMs. We illustrated these algorithms for the formulated PCA models. Followed by the estimation algorithms, we presented a strategy for hyperparameter selection in the case of Bayesian models and showed how Bayesian analysis can be utilized to obtain an appropriate model for the given dataset.

Chapter 3

Process monitoring using probabilistic models

In this chapter, we intend to investigate the use of probabilistic linear-in-parameters models for process monitoring and study their connection with the classical multivariate techniques in the context of process monitoring. We find that there lies an incentive for defining a general model that encompasses various probabilistic models. Instead of looking at monitoring based on individual models in isolation, it allows us to develop the monitoring approaches just for the general model, which can then be reduced effortlessly to the special cases if desired. This reduction is feasible due to linearity.

The following are the objectives of this chapter, 1) define a generalized probabilistic linear latent variable model (GPLLVM) that subsumes several probabilistic counterparts of the classical multivariate techniques, 2) develop monitoring statistics based on the GPLLVM, 3) restrict the model to the special cases and study the equivalence between the classical multivariate techniques and their probabilistic counterparts in the context of monitoring, and 4) present an approach based on the EM algorithm for estimating the maximum likelihood parameters of the GPLLVM. In addition, as a part of this exercise, we flag some common issues related to the monitoring statistics presented in the existing literature for monitoring approaches based on the probabilistic latent variable models.

This chapter is organized as follows: In section 3.1, we present the preliminaries where we briefly review classical multivariate techniques based monitoring ap-

proaches. In section 3.2, we define the GPLLVM. In section 3.3, we develop the monitoring charts for process monitoring based on the GPLLVM. In section 3.4, we show the equivalence between the monitoring methods based on the probabilistic latent variable models and the classical multivariate techniques. In section 3.5, we show the numerical simulations verifying the presented results and in section 3.6, we provide the concluding remarks. We also provide the EM algorithm for estimating the parameters of the GPLLVM in Appendix B.1 for completeness.

Recurring and the commonly used notations in this chapter: $\mathbb{R}$ is the space of real numbers, I_P is the identity matrix of size $P \times P$, $E(.)$ and $Cov(.)$ are the expectation and covariance operators respectively, $diag(.)$ is the operator that converts a vector into a diagonal matrix and vice versa, $\mathcal{N}(\mu, \Sigma)$ represents the multivariate normal distribution with mean μ and covariance Σ and superscript ^{T} corresponds to the transpose operator. Other notations used in this chapter are described when they are first introduced.

3.1 Preliminaries

In this section, we provide a brief review of the principal component analysis (PCA) and canonical correlation analysis (CCA) based monitoring approaches that is necessary to appreciate some of the key results presented in this chapter.

3.1.1 PCA based monitoring

Consider a system with P-dimension outputs. Let $Y \triangleq \{y_1, \ldots, y_n \in \mathbb{R}^P, \ldots, y_N\} \in \mathbb{R}^{P \times N}$ be a set of N mean-centred observations of the outputs with sample covariance matrix $\tilde{\Sigma}_{yy} \succeq 0$, measured during the normal operation of the system. In the PCA based monitoring approach, $\tilde{\Sigma}_{yy}$ is decomposed using SVD/eigendecomposition as the following,

$$\tilde{\Sigma}_{yy} = \eta_K \Lambda_K \eta_K^T + \eta_{\sim K} \Lambda_{\sim K} \eta_{\sim K}^T \tag{3.1}$$

where $\mathbb{R}^{K \times K} \ni \Lambda_K \succ 0$ is a diagonal matrix with K, $(K < P)$, principal eigenvalues of $\tilde{\Sigma}_{yy}$ as diagonal elements and $\mathbb{R}^{(P-K)\times(P-K)} \ni \Lambda_{\sim K} \succeq 0$ is a diagonal matrix with $(P - K)$ minor eigenvalues of $\tilde{\Sigma}_{yy}$ as diagonal elements. Matrices $\eta_K \in \mathbb{R}^{P \times K}$ and

$\eta_{\sim K} \in \mathbb{R}^{P \times \sim K}$ are composed of orthonormal eigenvectors as columns corresponding to the eigenvalues in Λ_K and $\Lambda_{\sim K}$ respectively. Then, the minor eigenvectors are discarded and using η_K, the data is projected onto the lower dimension latent space as the following,

$$S_K = \eta_K^T Y \tag{3.2}$$

where $S_K = \{s_1, \ldots, s_n \in \mathbb{R}^K, \ldots, s_N\} \in \mathbb{R}^{K \times N}$ is the set of N lower (K) dimension latent variables corresponding to the N observations. The latent variables are linearly uncorrelated and their covariance is given by $\Lambda_K = diag\,(\lambda_1,\, \lambda_2,\, ...,\, \lambda_K)$.

The model developed from the normal data is deployed to monitor the routine operation data. Typically, two different statistics, namely, 1) Hotelling's T^2 [12], and 2) Q or SPE [26] are monitored to check the conformity of the new data to the normal operation. T^2 is the normalized sum of square of latent variables. For an observation y_n, the T^2 statistic is defined as the following,

$$T_n^2 = ||\Lambda_K^{-\frac{1}{2}} s_n||_2 = s_n^T \Lambda_K^{-1} s_n = y_n^T \eta_K \Lambda_K^{-1} \eta_K^T y_n \tag{3.3}$$

The Q statistic is the sum of square of the residuals obtained with the optimal least squares reconstruction. When the observations are reconstructed from the lower dimensional latent variables, the optimal reconstruction for an observation in the least squares sense is given by,

$$\hat{y}_n = \eta_K s_{nK} = \eta_K \eta_K^T y_n \tag{3.4}$$

where $\hat{y}_n$ is the reconstruction of y_n given the model. The reconstruction residual, r_n is given by,

$$r_n = (I_P - \eta_K \eta_K^T) y_n \tag{3.5}$$

and the Q statistic based on the reconstruction residuals is defined as the following,

$$Q_n = ||r_n||_2 = r_n^T r_n = y_n^T \left(I_P - \eta_K \eta_K^T\right) y_n \tag{3.6}$$

Remark 1. *It can also be shown that PCA is an optimal solution to the following problem,*

$$\min_{\eta_K} \Sigma_{n=1}^N ||r_n||_2 \tag{3.7}$$

subject to

$$\eta_K^T \eta_K = I_K \tag{3.8}$$

The null distribution of the T^2 statistic is a χ^2 distribution with K degrees of freedom. The Q statistic is reducible to a nonnegative sum of χ^2 random variables, and for its cumulative distribution function, several approximations are available in the literature (for a recent review, see [73]). Generally, the approximation provided in [74] is used for obtaining the control limits by following [26].

3.1.2 CCA based monitoring

Consider a system with P-dimension outputs and L-dimension inputs. Let N mean-centred observations of output are given by Y and of inputs are given by $X \triangleq \{x_1, \ldots, x_n \in \mathbb{R}^L, \ldots, x_N\} \in \mathbb{R}^{L \times N}$ from the normal operation of the system, their respective sample covariance matrices are given by $\tilde{\Sigma}_{yy} \succ 0$ and $\tilde{\Sigma}_{xx} \succ 0$ and the sample cross-covariance matrix between the outputs and the inputs is given by $\tilde{\Sigma}_{yx}$. CCA extracts linearly independent latent variables from X and Y using the linear projection matrices $\zeta_x \in \mathbb{R}^{L \times J}$ and $\zeta_y \in \mathbb{R}^{P \times J}$, $(J = \min(L, P))$ respectively, such that the correlation between the latent variables from X and Y is maximized. When $\tilde{\Sigma}_{xx}$ and $\tilde{\Sigma}_{yy}$ are invertible, the projection matrices can be obtained as the following,

$$\zeta_y = \tilde{\Sigma}_{yy}^{-\frac{1}{2}} \beta_y, \ \zeta_x = \tilde{\Sigma}_{xx}^{-\frac{1}{2}} \beta_x \tag{3.9}$$

resulting from the decomposition shown below,

$$\tilde{\Sigma}_{yy}^{-\frac{1}{2}} \tilde{\Sigma}_{yx} \tilde{\Sigma}_{xx}^{-\frac{1}{2}} = \beta_y \Gamma \beta_x^T \tag{3.10}$$

where $\mathbb{R}^{J \times J} \ni \Gamma \succeq 0$ is a diagonal matrix with the singular values that are also the estimates of the correlations between the latent variables from Y and X, and the matrices $\beta_y \in \mathbb{R}^{P \times J}$ and $\beta_x \in \mathbb{R}^{L \times J}$ contain orthonormal eigenvectors spanning the basis of the row and the column spaces of the matrix on the left hand side of Eqn. (3.10) respectively. Due to orthonormal eigenvectors, the projection matrices obey the following relationships,

$$\zeta_y^T \tilde{\Sigma}_{yy} \zeta_y = I_J, \ \zeta_x^T \tilde{\Sigma}_{xx} \zeta_x = I_J \tag{3.11}$$

If $K, \ K < J$, correlations are found to be significant, the first K latent variables from X, S_{xK}, and Y, S_{yK}, are obtained as the following,

$$S_{yK} = \zeta_{yK}^T Y, \ S_{xK} = \zeta_{xK}^T X \tag{3.12}$$

where the matrices $\zeta_{yK} \in \mathbb{R}^{P\times K}$ and $\zeta_{xK} \in \mathbb{R}^{L\times K}$ are composed of first K retained columns of ζ_y and ζ_x, respectively.

The latent variables extracted from X and Y have identity covariance (which is evident from Eqn. (3.11)). For process monitoring, T^2 statistics are defined for the latent variables extracted from the inputs and the outputs separately as,

$$T^2_{yn} = y_n^T \zeta_{yK} \zeta_{yK}^T y_n,\ T^2_{xn} = x_n^T \zeta_{xK} \zeta_{xK}^T x_n \tag{3.13}$$

Remark 2. *It can also be shown that CCA is an optimal solution to the following problem,*

$$\min_{\zeta_{yK},\zeta_{yK}} \ \Sigma_{n=1}^N \ ||\zeta_{yK}^T y_n - \zeta_{xK}^T x_n||_2 \tag{3.14}$$

subject to

$$\zeta_{yK}^T \tilde{\Sigma}_{yy} \zeta_{yK} = \zeta_{xK}^T \tilde{\Sigma}_{xx} \zeta_{xK} = I_K \tag{3.15}$$

Hence, the monitoring statistic for the model residuals should be $||\zeta_{yK}^T y_n - \zeta_{xK}^T x_n||_2$*. However, as in Eqn. (3.6), the Q statistic based on the reconstruction residuals has also been used in the literature for CCA based monitoring[19].*

Similar to PCA, the null distributions of both T^2_{yn} and T^2_{xn} are a χ^2 distribution with K degrees of freedom. However, when PCA and CCA based monitoring models are developed from a finite number of observations, Hotelling's T^2 distribution is defined as the null distribution from which the control limits are obtained to monitor the T^2 statistic (for details, see [3]).

Remark 3. *Motivated by the use of PCA and CCA in subspace identification for dynamic modelling, both are also used in dynamic process monitoring. In fact, CCA is generally used for dynamic process monitoring. For modelling dynamic processes using PCA and CCA, the sample covariances and cross-covariances of the lag augmented observations are used.*

This completes our brief review of the two popular classical multivariate techniques used for monitoring. In the next section, we will define the GPLLVM and the special cases subsumed by the model, which will help us discuss probabilistic models based monitoring.

3.2 GPLLVM

Consider a system with the output observations Y that are affected by the input observations X and $U \triangleq \{u_1, \ldots, u_n \in \mathbb{R}^M, \ldots, u_N\} \in \mathbb{R}^{M \times N}$. We assume that Y, X, and U are measured and Y and X are mean centered. We define the GPLLVM that models Y as presented below:

Definition 4. ***Generalized probabilistic linear latent variable model:***
A GPLLVM that models output observations Y given the input observations X that are corrupted by noise, and the deterministic input observations U, is represented as:

$$\begin{aligned} y_n &= W z_n + F u_n + \epsilon_{yn},\ \epsilon_{yn} \overset{i.i.d}{\sim} \mathcal{N}(0, \psi_y) \\ x_n &= V z_n + \epsilon_{xn},\ \epsilon_{xn} \overset{i.i.d}{\sim} \mathcal{N}(0, \psi_x) \\ z_n &\overset{i.i.d}{\sim} \mathcal{N}(0, I_K) \end{aligned} \tag{3.16}$$

where $z_n \in \mathbb{R}^K$ is the latent variable such that $K < \min(P, L)$, $W \in \mathbb{R}^{P \times K}$ and $F \in \mathbb{R}^{P \times M}$ are the coefficient matrices of z_n and u_n in the relationship that generates the output y_n, $V \in \mathbb{R}^{L \times K}$ is the coefficient matrix of z_n in the relationship that generates the input x_n, and $\epsilon_{yn} \in \mathbb{R}^p$ and $\epsilon_{xn} \in \mathbb{R}^L$ are the noise terms that are multivariate Gaussian distributed with zero mean and covariances $\psi_y \succ 0$ and $\psi_x \succ 0$ respectively and W, F and V are full column rank matrices.

Fig. 3.1 shows the Bayesian network representation of the GPLLVM. GPLLVM also falls under the conjugate exponential family graphical models. The prior distribution of z_n is a multivariate Gaussian distribution and it is conjugate to its likelihood that defines the distributions of y_n and x_n.

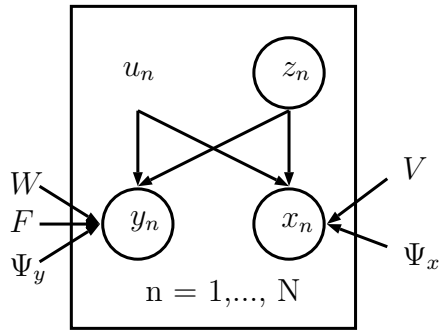


Figure 3.1: Bayesian network representation of the GPLLVM

Remark 4. *From Eqn. (3.16), the joint distribution of y_n and x_n given z_n takes the following form:*

$$\begin{bmatrix} y_n \\ x_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} Wz_n + Fu_n \\ Vz_n \end{bmatrix}, \begin{bmatrix} \psi_y & 0 \\ 0 & \psi_x \end{bmatrix} \right) \tag{3.17}$$

and when the latent variables are marginalized from Eqn. (3.17), the joint distribution of y_n and x_n takes the following form:

$$\begin{bmatrix} y_n \\ x_n \end{bmatrix} \overset{i.i.d}{\sim} \mathcal{N}\left(\begin{bmatrix} Fu_n \\ 0 \end{bmatrix}, \begin{bmatrix} WW^T + \psi_y & WV^T \\ VW^T & VV^T + \psi_x \end{bmatrix} \right) \tag{3.18}$$

The GPLLVM defined in Eqn. (3.16) subsumes probabilistic variants of several multivariate techniques used for process monitoring. We point out a few special cases of the model below. However, it should be noted that the special cases are not limited to the ones pointed out.

When there are no inputs and the output noise covariance, ψ_y is diagonal, the model reduces to the case of probabilistic factor analyzer model as shown below,

$$y_n = Wz_n + \epsilon_{yn},\ z_n \overset{i.i.d}{\sim} \mathcal{N}(0, I_K),\ \epsilon_{yn} \overset{i.i.d}{\sim} \mathcal{N}(0, \psi_y) \tag{3.19}$$

and when ψ_y is isotropic ($\psi_y = \sigma^2 I$), Eqn. (3.19) reduces to the case of PPCA model [28] as shown below,

$$y_n = Wz_n + \epsilon_{yn},\ z_n \overset{i.i.d}{\sim} \mathcal{N}(0, I_K),\ \epsilon_{yn} \overset{i.i.d}{\sim} \mathcal{N}\left(0, \sigma^2 I\right) \tag{3.20}$$

When there are no deterministic inputs, U, the model defined in Eqn. (3.16) reduces to the case of PCCA model [29] as shown below,

$$\begin{aligned} y_n &= Wz_n + \epsilon_{yn},\ \epsilon_{yn} \overset{i.i.d}{\sim} \mathcal{N}(0, \psi_y) \\ x_n &= Vz_n + \epsilon_{xn},\ \epsilon_{xn} \overset{i.i.d}{\sim} \mathcal{N}(0, \psi_x) \\ z_n &\overset{i.i.d}{\sim} \mathcal{N}(0, I_K) \end{aligned} \tag{3.21}$$

When there are only inputs u_n and the latent variables are dropped from the model, the model defined in Eqn. (3.16) reduces to the well known multiple linear regression (MLR) model,

$$y_n = Fu_n + \epsilon_{yn},\ \epsilon_{yn} \overset{i.i.d}{\sim} \mathcal{N}(0, \psi_y) \tag{3.22}$$

Also, the conditional distribution $p(y_n - Fu_n | x_n)$ obtained from the GPLLVM can be shown to be equivalent to an error in inputs and error in outputs MLR model with outputs $y_n - Fu_n$ and inputs x_n.

Remark 5. *There have been attempts in the literature [75, 76, 77] to develop probabilistic version of PLS. These models can also be seen as the special cases of the GPLLVM. However, we caution the readers from interpreting the models proposed in the mentioned references to be probabilistic counterparts of the traditional PLS models as their maximum likelihood estimates may not yield similar results as that of traditional PLS algorithms. Traditional PLS algorithms maximize the covariance between the latent variables extracted from the inputs and outputs [78], whereas the maximum likelihood estimation of the PCCA which is the special case of the GPLLVM maximizes the correlation between the latent variables extracted from the inputs and outputs [29].*

Given Y, X and U, the parameters (W, V, F, ψ_y and ψ_x) of the GPLLVM can be estimated using the EM algorithm. The E-step and the M-step recursive update expressions are provided in Appendix B.1.

Now that we have introduced the GPLLVM, we proceed to derive the control charts for process monitoring in the next section.

3.3 Control Charts based on the GPLLVM

To monitor the process based on the GPLLVM developed from the normal operation data, we propose various monitoring statistics. They include monitoring statistics based on, 1) the latent variables projected from the observed variables, and 2) the reconstruction residuals of the observed variables from the projected latent variables. We also enlist different possible monitoring statistics that leverage the general structure of the GPLLVM and can be used depending upon the user's need.

3.3.1 Monitoring the latent variables

In this subsection, we derive statistics for monitoring the variability in the latent variables extracted from the GPLLVM. Theorem 1 presented in this section helps us achieve that. Below, we present a lemma which will be useful in proving theorem 1.

Lemma 2. *Given y_n, x_n and u_n, the posterior distribution of z_n of a GPLLVM is a multivariate Gaussian distribution with mean, $\mu_{z_n|y_n,x_n,u_n}$ and covariance, $\Sigma_{z|y,x}$ as*

shown below,

$$z_n|y_n, x_n, u_n \overset{i.i.d}{\sim} \mathcal{N}\left(\mu_{z_n|y_n,x_n,u_n}, \Sigma_{z_n|y_n,x_n,u_n}\right) \tag{3.23}$$

where

$$\mu_{z_n|y_n,x_n,u_n} = \Phi\left[W^T\psi_y^{-1}\left(y_n - Fu_n\right) + V^T\psi_x^{-1}x_n\right] \tag{3.24}$$

$$\Sigma_{z_n|y_n,x_n,u_n} = \Sigma_{z|y,x} = \Phi \tag{3.25}$$

$$\Phi = \left[W^T\psi_y^{-1}W + V^T\psi_x^{-1}V + I_K\right]^{-1} \tag{3.26}$$

Proof. From the Bayes rule, we can infer the posterior distribution of z_n given x_n, u_n and y_n as the following,

$$p\left(z_n|y_n, x_n, u_n\right) \propto p\left(y_n, x_n|z_n, u_n\right) p\left(z_n\right) \tag{3.27}$$

Substituting the expression for the likelihood of z_n shown in Eqn. (3.17) and the prior distribution of z_n shown in Eqn. (3.16) in Eqn. (3.27) yields the following,

$$\begin{aligned} p\left(z_n|y_n, x_n, u_n\right) \propto \exp\left\{-\frac{1}{2}\left(y_n - Wz_n - Fu_n\right)^T\psi_y^{-1}\left(y_n - Wz_n - Fu_n\right)\right\} \\ \times \exp\left\{-\frac{1}{2}\left(x_n - Vz_n\right)^T\psi_x^{-1}\left(x_n - Vz_n\right)\right\}\exp\left\{-\frac{1}{2}z_n^Tz_n\right\} \end{aligned} \tag{3.28}$$

Rewriting the exponents on the RHS of the above equation as a quadratic function in z_n and dropping the other constant terms yield the following,

$$p\left(z_n|y_n, x_n, u_n\right) \propto \exp\left\{-\frac{1}{2}z_n^T\Phi^{-1}z_n\right\}\exp\left\{z_n^T\left[W^T\psi_y^{-1}\left(y_n - Fu_n\right) + V^T\psi_x^{-1}x_n\right]\right\} \tag{3.29}$$

Further, performing square completion yields,

$$p\left(z_n|y_n, x_n, u_n\right) \propto \exp\left\{-\frac{1}{2}\mu_{z_n|y_n,x_n,u_n}^T\Sigma_{z|y,x}^{-1}\mu_{z_n|y_n,x_n,u_n}\right\} \tag{3.30}$$

where the expressions for $\mu_{z_n|y_n,x_n,u_n}$, $\Sigma_{z|y,x}$ and Φ are the same as the ones shown in equations (3.24), (3.25) and (3.26) respectively. ■

Theorem 1. *When GPLLVM defines the true distribution of the observations with the exact parameters, to check if the latent variable z_n lies outside the normal operation region with $(1-\alpha)\times 100\%$ confidence level, $\alpha \in [0, 1]$, it is sufficient to verify if the following inequality is violated,*

$$\mu_{z_n|y_n,x_n,u_n}^T\left[I_K - \Phi\right]^{-1}\mu_{z_n|y_n,x_n,u_n} \leq \chi_{(1-\alpha;K)}^{-2} \tag{3.31}$$

where $\mu_{z_n|y_n,x_n,u_n}$ is the posterior mean and Φ is the posterior covariance of the latent variable given x_n, y_n and u_n, and $\chi^{-2}_{(1-\alpha;K)}$ refers to inverse of chi-squared distribution with K degrees of freedom.

Proof. The proof contains two parts: 1) We argue that it is sufficient to monitor $\mu_{z_n|y_n,x_n,u_n}$ by making use of the results presented in Lemma 2, and 2) we show that $\mu^T_{z_n|y_n,x_n,u_n}\left[I_K-\Phi\right]^{-1}\mu_{z_n|y_n,x_n,u_n}$ follows the chi-squared distribution with K-degrees of freedom.

Part 1: From equations (3.24), (3.25) and (3.26), it can be seen that $\mu_{z_n|y_n,x_n,u_n}$ changes with n and $\Sigma_{z_n|y_n,x_n,u_n}$ does not. Therefore, the only random component in the posterior distribution of z_n is the mean parameter and for monitoring the variability in z_n, it is sufficient to monitor only $\mu_{z_n|y_n,x_n,u_n}$.

Part 2: From Eqn. (3.24), $\mu_{z_n|y_n,x_n,u_n}$ is given by a linear combination of the observations. From Eqn. (3.18), the observations follow a multivariate normal distribution and therefore, $\mu_{z_n|y_n,x_n,u_n}$ will also follow a multivariate normal distribution. We proceed to show that the covariance of $\mu_{z_n|y_n,x_n,u_n}$ is $[I_K-\Phi]$ making the LHS of Eqn. (3.31) a χ^2 random variable with K degrees of freedom.

If the GPLLVM defines the true distribution of the observations, the expected value and the covariance of $\mu_{z_n|y_n,x_n,u_n}$ as the following,

$$E\left(\mu_{z_n|y_n,x_n,u_n}\right)=\Phi W^T\psi_y^{-1}E\left(y_n-Fu_n\right)+\Phi V^T\psi_x^{-1}E\left(x_n\right)=0 \tag{3.32}$$

$$\begin{aligned}Cov(\mu_{z_n|y_n,x_n,u_n})&=E\left(\mu_{z_n|y_n,x_n,u_n}\mu^T_{z_n|y_n,x_n,u_n}\right)\\&=\Phi W^T\psi_y^{-1}E\left[\left(y_n-Fu_n\right)\left(y_n-Fu_n\right)^T\right]\psi_y^{-1}W\Phi+\Phi V^T\psi_x^{-1}E\left[x_nx_n^T\right]\psi_x^{-1}V\Phi\\&+\Phi V^T\psi_x^{-1}E\left[x_n\left(y_n-Fu_n\right)^T\right]\psi_y^{-1}W\Phi+\Phi W^T\psi_y^{-1}E\left[\left(y_n-Fu_n\right)x_n^T\right]\psi_x^{-1}V\Phi\end{aligned} \tag{3.33}$$

Simplifying Eqn. (3.33) yields,

$$Cov(\mu_{z_n|y_n,x_n,u_n})=\left[W^T\psi_y^{-1}W+V^T\psi_x^{-1}V\right]\Phi$$

$$Cov(\mu_{z_n|y_n,x_n,u_n})=\left[\Phi^{-1}-I_K\right]\Phi$$

$$Cov(\mu_{z_n|y_n,x_n,u_n})=I_K-\Phi \tag{3.34}$$

As the dimension and the covariance of $\mu_{z_n|y_n,x_n,u_n}$ are K and $I_K - \Phi$ respectively, the term on the LHS of Eqn. (3.31) follows χ^2 distribution with K degrees of freedom. Therefore, verifying the inequality in Eqn. (3.31) that $\mu_{z_n|y_n,x_n,u_n}$ obtained from the normal operation data will be flagged as faulty with the rate exactly equal to α when the number of monitored samples tends to ∞. ■

3.3.2 Monitoring the model residuals

In this subsection, we derive the statistic for monitoring the variability in the model residuals of the GPLLVM. The GPLLVM can be used to predict/reconstruct y_n and x_n from z_n and u_n. As z_n is not observed, $\mu_{z_n|y_n,x_n,u_n}$ can be used to reconstruct y_n and x_n. However, the reconstruction from $\mu_{z_n|y_n,x_n,u_n}$ may not be deemed optimal as the distribution of $\mu_{z_n|y_n,x_n,u_n}$ is relatively skewed more toward the origin compared to the distribution of z_n. This can be seen from the difference between the covariances of z_n (Eqn. (3.16)) and $\mu_{z_n|y_n,x_n,u_n}$ (Eqn. (3.34)). Instead, we can obtain the optimal reconstruction for the observations in the weighted least squares sense from the following formulation,

$$\min_{z_n} \ (y_n - Fu_n - Wz_n)^T \psi_y^{-1} (y_n - Fu_n - Wz_n) + (x_n - Vz_n)^T \psi_x^{-1} (x_n - Vz_n) \tag{3.35}$$

Monitoring the value function of Eqn. (3.35) provides a means to check on how well the observation can be reconstructed by the GPLLVM. Theorem 2 presented in this section help us to achieve a framework for monitoring the model residuals. Before presenting Theorem 2, we derive the value function of Eqn. (3.35) in the lemma presented below, which is essentially the optimal weighted least squares estimate of z_n.

Lemma 3. *The optimal estimate of z_n that minimizes the weighted least squares residuals objective function shown in Eqn. (3.35) is given by,*

$$\hat{z}_{n|_{y_n,x_n}} = [I_K - \Phi]^{-1} \Phi \begin{bmatrix} \psi_y^{-1} W \\ \psi_x^{-1} V \end{bmatrix}^T \begin{bmatrix} y_n - Fu_n \\ x_n \end{bmatrix} \tag{3.36}$$

Proof. Taking the derivative of the objective function shown in Eqn. (3.35) and equating it to zero yields the following,

$$\hat{z_n}_{|x_n,y_n} = \left[W^T \psi_y^{-1} W + V^T \psi_x V\right]^{-1} \left[W^T \psi_y^{-1} (y_n - Fu_n) + V^T \psi_x^{-1} x_n\right] \tag{3.37}$$

From the expression for Φ shown in Eqn. (3.26), the following can be obtained,

$$W^T\psi_y^{-1}W + V^T\psi_x V = \Phi^{-1} - I_k = \Phi^{-1}\left[I_K - \Phi\right] \tag{3.38}$$

and

$$\left[W^T\psi_y^{-1}W + V^T\psi_x V\right]^{-1} = \left[I_K - \Phi\right]^{-1}\Phi \tag{3.39}$$

Substituting Eqn. (3.39) in Eqn. (3.37) and writing the second multiplier on the RHS of Eqn. (3.37) in a matrix form yield the estimate of z_n in the form shown in Eqn. (3.36). ■

Theorem 2. *The optimal value of the value function of Eqn. (3.35) reduces to a non-negative sum of χ^2 random variables. Further, when the GPLLVM defines the true distribution of the observations with the exact parameters, the value function of Eqn. (3.35) becomes a χ^2 random variable. If the degrees of freedom of the former is $\mathcal{K}_1$,and the latter is $\mathcal{K}_2$, then $\mathcal{K}_1, \mathcal{K}_2 < P + L$.*

Proof. The proof contains three parts, 1) we derive a compact representation of the value function of Eqn. (3.35), 2) we show that it indeed reduces to a non-negative sum of χ^2 random variables, and 3) when GPLLVM defines the true distribution of the observations, we show that it further simplifies to a χ^2 random variable.

Part 1: Substituting the optimal weighted least squares estimate of z_n shown in Eqn. (3.36) in Eqn. (3.35) leads the value function to a compact form give below,

$$\mathcal{Y}_n^T A \mathcal{Y}_n,\ \mathcal{Y}_n \overset{i.i.d}{\sim} \mathcal{N}\left(0, \Sigma\right) \tag{3.40}$$

where

$$\mathcal{Y}_n = \begin{bmatrix} y_n - Fu_n \\ x_n \end{bmatrix} \tag{3.41}$$

$$A = \begin{bmatrix} \psi_y & \\ & \psi_x \end{bmatrix}^{-1} - \begin{bmatrix} \psi_y^{-1}W \\ \psi_x^{-1}V \end{bmatrix}\left[I_K - \Phi\right]^{-1}\Phi\begin{bmatrix} \psi_y^{-1}W \\ \psi_x^{-1}V \end{bmatrix}^T \tag{3.42}$$

$A \succeq 0$ follows from the fact that the sum of square of residuals will always be greater than or equal to zero. Let $\Psi^{-1} = \begin{bmatrix} \psi_y^{-1} & \\ & \psi_x^{-1} \end{bmatrix}$ and $\mathcal{V} = \Psi^{-\frac{1}{2}}\begin{bmatrix} W \\ V \end{bmatrix}$. Then, A can be written as,

$$A = \Psi^{-\frac{1}{2}}\left\{I_{P+L} - \mathcal{V}(\mathcal{V}^T\mathcal{V})^{-1}\mathcal{V}^T\right\}\Psi^{-\frac{1}{2}} \tag{3.43}$$

where the term in the middle, $\left\{I_{P+L} - \mathcal{V}(\mathcal{V}^T\mathcal{V})^{-1}\mathcal{V}^T\right\}$ is an idempotent matrix.

Part 2: In this part, we show that the compact form in Eqn. (3.40) reduces to a non-negative sum of χ^2 random variables through a series of algebraic manipulations.

The compact form can be replaced by a quadratic function on the whitened observations as the following,

$$\mathcal{Y}_n^T A \mathcal{Y}_n = \mathcal{Y}_n^{'T} B \mathcal{Y}_n^{'} \tag{3.44}$$

where

$$B = \Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}, \tag{3.45}$$

and $\mathcal{Y}_n'$ is the whitened observation such that

$$\mathcal{Y}_n^{'} = \Sigma^{-\frac{1}{2}} \mathcal{Y}_n \tag{3.46}$$

Matrix $B \succeq 0$ is symmetric. Therefore, B can be decomposed and Eqn. (3.44) can be rewritten as the following,

$$\mathcal{Y}_n^T A \mathcal{Y}_n = \mathcal{Y}_n^{'T} \mathcal{U}_{\mathcal{K}_1} \mathcal{S}_{\mathcal{K}_1} \mathcal{U}_{\mathcal{K}_1}{}^T \mathcal{Y}_n^{'} = \mathcal{Z}_n^T \mathcal{S}_{\mathcal{K}_1} \mathcal{Z}_n \tag{3.47}$$

which is indeed a non-negative sum of $\mathcal{K}_1$ chi-square random variables with the diagonal elements of $\mathcal{S}_{\mathcal{K}_1}$ being the positive weights as the following results hold,

$$E\left(\mathcal{Z}_n \mathcal{Z}_n^T\right) = \mathcal{U}_{\mathcal{K}_1}{}^T I \mathcal{U}_{\mathcal{K}_1} = I \tag{3.48}$$

$$\mathcal{Z}_n \overset{i.i.d}{\sim} \mathcal{N}\left(0, I_{\mathcal{K}_1}\right) \tag{3.49}$$

$\mathcal{S}_{\mathcal{K}_1}$ is a diagonal matrix with $\mathcal{K}_1$ non-zero eigenvalues of B and $\mathcal{U}_{\mathcal{K}_1}$ is a matrix with corresponding eigenvectors. Now, recall Eqn. (3.43): As idempotent matrices are rank deficient unless they are identity matrices, the multiplication in Eqn. (3.43) always leads to rank deficient matrix since $\mathcal{V}(\mathcal{V}^T\mathcal{V})^{-1}\mathcal{V}$ is a non-zero matrix. Therefore, A and B are rank deficient and $\mathcal{K}_1 < P + L$.

A comment regarding the practical approach for monitoring the non-negative sum of χ^2 random variables is provided in Remark 6.

Part 3: When the GPLLVM defines the true distribution of the observations, the covariance of the observations in Eqn. (3.45) can be replaced by,

$$\Sigma = \begin{bmatrix} WW^T + \psi_y & WV^T \\ VW^T & VV^T + \psi_x \end{bmatrix} \tag{3.50}$$

The resulting matrix B becomes idempotent (For proof refer to Appendix B.3). As an idempotent matrix has all of its eigenvalues to be either zeros or ones, the following result emerges,

$$\mathcal{Y}_n^T A \mathcal{Y}_n = \mathcal{Z}_n^T \mathcal{S}_{\mathcal{K}_2} \mathcal{Z}_n = \mathcal{Z}_n^T I_{\mathcal{K}_2} \mathcal{Z}_n,\ \mathcal{Z}_n \overset{i.i.d}{\sim} \mathcal{N}(0, I_{\mathcal{K}_2}) \tag{3.51}$$

which is a χ^2 random variable with $\mathcal{K}_2$ degrees of freedom. With similar arguments presented in Part 2, $\mathcal{K}_2$ can be shown to be $< P + L$. ■

Remark 6. *Any reasonable approximation of cumulative distribution function of non-negative sum of χ^2 random variables can be used to define the control limit for the statistic in Eqn. (3.47) and subsequently for monitoring. For such cases in our numerical simulations, we use Imhof's method [79] to identify the control limit using the package 'CompQuadForm' [80].*

Remark 7. *From Theorem 2, an important implication of monitoring the model residuals of the GPLLVM (or any probabilistic linear latent variable model with multivariate Gaussian additive noise) is that the monitoring statistic simplifies to a χ^2 random variable. Hence, similar to the approach for monitoring the latent variables, the residuals can also be monitored through the following relationship,*

$$\mathcal{Y}_n^T A \mathcal{Y}_n \leq \chi^{-2}_{(1-\alpha;\mathcal{K}_2)} \tag{3.52}$$

3.3.3 Other possible monitoring statistics

In addition to the statistics presented in subsections 4.1 and 4.2, one can choose to monitor the other aspects of the system or monitor the system when only partial information is available by deriving the specific statistics. For instance, 1) monitor the variability in observed y_n and x_n directly as the model defines a distribution for the observations, 2) monitor the variability in z_n when it is inferred from a partial set of observations (either from x_n alone or from y_n and u_n alone), 3) monitor the discrepancy between z_n inferred just from x_n and z_n inferred just from y_n and u_n, among the other possibilities. Here, we illustrate the monitoring of variability in z_n when it is inferred only from y_n and u_n and enlist the above discussed possibilities in Table 3.1.

Eqn. (3.23) shows the posterior of z_n given y_n, x_n and u_n. We can also infer the posterior distribution of z_n when given only y_n and u_n by marginalizing x_n from Eqn. (3.23) as the following,

$$z_n|y_n, u_n \overset{i.i.d}{\sim} \mathcal{N}\left(\mu_{z_n|y_n,u_n}, \Sigma_{z_n|y_n,u_n}\right) \tag{3.53}$$

$$\mu_{z_n|y_n,u_n} = \Phi_y\left[W^T\psi_y^{-1}\left(y_n - Fu_n\right)\right] \tag{3.54}$$

$$\Sigma_{z_n|y_n,u_n} = \Sigma_{z|y} = \Phi_y \tag{3.55}$$

where

$$\Phi_y = \left[W^T\psi_y^{-1}W + I_K\right]^{-1} \tag{3.56}$$

Notice that when $V = 0$, Eqn. (3.23) reduces to Eqn. (3.53). Similarly, the posterior of z_n given x_n alone can be obtained by letting $W = 0$ in Eqn. (3.23). By following a similar procedure as illustrated in Theorem 1, we can derive the monitoring statistic for monitoring the variability in z_n given y_n and u_n to be the following,

$$\mu_{z_n|y_n,u_n}^T\left[I_K - \Phi_y\right]^{-1}\mu_{z_n|y_n,u_n} \tag{3.57}$$

as presented in row 1 of Table 3.1.

Now that we have discussed the monitoring options using the GPLLVM, in the following section, we proceed to show the connection between the proposed monitoring statistics and the classical monitoring statistics under specific conditions.

Table 3.1: A selected few other monitoring options that can be implemented from the GPLLVM of a system

S. NO.	Monitored aspect	Statistic to be monitored	Null distribution
1	Variability in $z_n$inferred from y_n and u_n alone	$\mu_{z_n\|y_n,u_n}^T\left[I_K - \Phi_y\right]^{-1}\mu_{z_n\|y_n,u_n}$	$\sim \chi_K^2$
2	Variability in z_n inferred from x_n alone	$\mu_{z_n\|x_n}^T\left[I_K - \Phi_x\right]^{-1}\mu_{z_n\|x_n}$	$\sim \chi_K^2$
3	Variability in observed y_n and x_n	$\mathcal{Y}_n^T\Sigma^{-1}\mathcal{Y}_n$	$\sim \chi_{(P+L)}^2$
4	Discrepancy between z_n inferred from x_n and from y_n and u_n	$\|\|[I_K - \Phi_y]^{-\frac{1}{2}}\mu_{z_n\|y_n,u_n} - [I_K - \Phi_x]^{-\frac{1}{2}}\mu_{z_n\|x_n}\|\|_2$	+ve sum of χ^2 RVs

Note: $\mu_{z_n|x_n} = \Phi_x\left[V^T\psi_x^{-1}x_n\right]$ and $\Phi_x = \left[V^T\psi_x^{-1}V + I_K\right]^{-1}$

3.4 Classical Multivariate Techniques vs. Their Probabilistic Counterparts

As introduced, the GPLLVM subsumes the probabilistic counterparts of many multivariate techniques used for monitoring. Specific restrictions imposed on the GPLLVM give rise to the probabilistic models such as PPCA, PCCA, etc., and the monitoring statistics derived in the previous section will also seamlessly reduce to their corresponding monitoring statistics. We compare the monitoring statistics and their null distributions corresponding to each special case against their classical counterparts. We restrict our presentation to two popular special cases of the GPLLVM namely, PPCA and PCCA through theorems 3 and 4. We start with claim 1, which will prove to be useful for the exposition of the proceeding results. Followed by that, we present lemmas 4 and 5 where we derive the important intermediate results to prove theorems 3 and 4. Then, we present theorems 3 and 4.

Claim 1. *In Eqn. (3.16) of the GPLLVM,*

1. *when $F = 0$, $V = 0$ and $\psi_y = \sigma^2 I$, the following holds true,*

$$W_{ML} = \eta_K \left(\Lambda_K - \sigma^2 I_K\right)^{\frac{1}{2}} R \tag{3.58}$$

$$\sigma^2_{ML} = \frac{1}{P-K} \sum_{i=K+1}^{P} \lambda_i \tag{3.59}$$

where W_{ML} and σ^2_{ML} are the maximum likelihood estimates of W and σ^2 of the model, respectively, R is an arbitrary rotational matrix and $\lambda_i, i \in [K+1, P]$ correspond to minor eigenvalues of $\tilde{\Sigma}_{yy}$ and

2. *when only $F = 0$, the following holds true,*

$$W_{ML} = \tilde{\Sigma}_{yy}\zeta_{yK}M_y,\ V_{ML} = \tilde{\Sigma}_{xx}\zeta_{xK}M_x \tag{3.60}$$

$$\psi_{yML} = \tilde{\Sigma}_{yy} - W_{ML}\left(W_{ML}\right)^T \tag{3.61}$$

$$\psi_{xML} = \tilde{\Sigma}_{xx} - V_{ML}\left(V_{ML}\right)^T \tag{3.62}$$

where V_{ML}, ψ_{xML} and ψ_{yML} are the maximum likelihood estimates of V, ψ_x and ψ_y, respectively, $M_y,\ M_x \in \mathcal{R}^{K\times K}$ are arbitrary matrices such that $M_y M_x^T = \Gamma_K$

with spectral norms lesser than one. Γ_K *is the diagonal matrix with first* K *canonical correlations as its diagonal elements.*

Proof.

1. Under the given restrictions in claim 1. 1, the GPLLVM reduces to the PPCA model presented in [28] and employing the results presented in section 3.2 of [28] yields the results presented above.

2. Under the given restrictions in claim 1. 2, the GPLLVM reduces to the PCCA model presented in [29] and employing the theorem 2 in [29] yields the results presented above.

■

Lemma 4. *Under the conditions stated and with the maximum likelihood estimates of the parameters shown in Claim 1.1, the following identities hold,*

$$[I_K - \Phi_y]^{-1} = I_K + \sigma^2 R^T \left(\Lambda_K - \sigma^2 I_K\right) R \tag{3.63}$$

$$\frac{1}{\sigma^2} W \Phi_y = \eta_K \left(\Lambda_K - \sigma^2 I_K\right)^{\frac{1}{2}} \left(\Lambda_K\right)^{-1} R \tag{3.64}$$

Proof. The expression for Φ_y shown in Eqn. (3.56) reduces to the following under the conditions stated in Claim 1.1,

$$\Phi_y = \left[\frac{1}{\sigma^2} W^T W + I_K\right]^{-1} \tag{3.65}$$

Substituting the maximum likelihood estimate of W shown in Eqn. (3.58) in the above equation leads to the following,

$$\Phi_y = \left[\frac{1}{\sigma^2} R^T \left(\Lambda_K - \sigma^2 I_K\right) R + I_K\right]^{-1} \tag{3.66}$$

Further simplification leads to

$$\Phi_y = \sigma^2 R^T \Lambda_K^{-1} R \tag{3.67}$$

Using the above expression for Φ_y, the LHS of Eqn. (3.63) can be written as,

$$[I_K - \Phi_y]^{-1} = \left[I_K - \sigma^2 R^T \Lambda_K^{-1} R\right]^{-1} \tag{3.68}$$

Applying the Woodbury matrix identity shown in Eqn. (B.1) of appendix B in the above equation with $\mathcal{M}_1 = I_K$, $\mathcal{M}_2 = -\sigma^2 R^T$, $\mathcal{M}_3 = \Lambda_K^{-1}$ and $\mathcal{M}_4 = R$ yields the result shown in Eqn. (3.63).

Similarly, using the expression for Φ_y shown in Eqn. (3.67) in the LHS of Eqn. (3.64) leads to the following,

$$\frac{1}{\sigma^2} W \Phi_y = W R^T \Lambda_K^{-1} R \tag{3.69}$$

Now, substituting the maximum likelihood estimate of W as shown in Eqn. (3.58) in the above equation yields the result shown in Eqn. (3.64). ■

Lemma 5. *Under the conditions stated and the maximum likelihood estimates of the parameters shown in Claim 1. 2, the following identities hold,*

$$[I_K - \Phi_y]^{-1} = \left[M_y^T M_y\right]^{-1} \tag{3.70}$$

$$\Phi_y W^T \psi_y^{-1} = M_y^T \zeta_{yK}^T \tag{3.71}$$

Proof. Under the conditions stated in Claim 1. 2 and from the result provided in Eqn. (3.61), the following holds,

$$W^T \psi_y^{-1} W = W^T \left[\tilde{\Sigma}_{yy} - W W^T\right]^{-1} W \tag{3.72}$$

Applying the identity shown in Eqn. (B.1) of appendix B in the above equation with $\mathcal{M}_1 = \tilde{\Sigma}_{yy}$, $\mathcal{M}_2 = -W$, $\mathcal{M}_3 = I_K$ and $\mathcal{M}_4 = W^T$, yields the following,

$$W^T \psi_y^{-1} W = W^T \tilde{\Sigma}_{yy}^{-1} W + W^T \tilde{\Sigma}_{yy}^{-1} W \left[I_K - W^T \tilde{\Sigma}_{yy}^{-1} W\right]^{-1} W^T \tilde{\Sigma}_{yy}^{-1} W \tag{3.73}$$

From the maximum likelihood estimate of W shown in Eqn. (3.61), the following holds,

$$W^T \tilde{\Sigma}_{yy}^{-1} W = M_y^T \zeta_{yk}^T \tilde{\Sigma}_{yy} \zeta_{yk} M_y = M_y^T M_y \tag{3.74}$$

Substituting the above equation in Eqn. (3.73) leads to the following,

$$W^T \psi_y^{-1} W = M_y^T \left\{I_K + M_y \left[I_K - M_y^T M_y\right]^{-1} M_y^T\right\} M_y \tag{3.75}$$

By applying the identity shown in Eqn. (B.1) of appendix B with $\mathcal{M}_1 = I_K$, $\mathcal{M}_2 = -M_y$, $\mathcal{M}_3 = I_K$ and $\mathcal{M}_4 = M_y^T$, the following can be obtained,

$$W^T \psi_y^{-1} W = M_y^T \left[I_K - M_y M_y^T\right]^{-1} M_y \tag{3.76}$$

Substituting the RHS of Eqn. (3.76) in the expression for Φ_y shown in Eqn. (3.56) yields the following,

$$\Phi_y = \left\{ I_K + M_y^T \left[I_K - M_y M_y^T \right]^{-1} M_y \right\}^{-1} \tag{3.77}$$

By applying the identity shown in Eqn. (B.1) of appendix B with $\mathcal{M}_1 = I_K$, $\mathcal{M}_2 = -M_y^T$, $\mathcal{M}_3 = I_K$ and $\mathcal{M}_4 = M_y$, the following can be obtained,

$$\Phi_y = \left\{ \left[I_K - M_y^T M_y \right]^{-1} \right\}^{-1} = \left[I_K - M_y^T M_y \right] \tag{3.78}$$

Using the above result for Φ_y, it is straightforward to see that the identity shown in Eqn. (3.70) holds.

Next we show that the identity shown in Eqn. (3.71) holds. From the maximum likelihood estimate of Φ_y, the following holds,

$$W^T \psi_y^{-1} = W^T \left[\tilde{\Sigma}_{yy} - W W^T \right]^{-1} \tag{3.79}$$

Applying the identity shown in Eqn. (B.1) of appendix B in the above equation with $\mathcal{M}_1 = \tilde{\Sigma}_{yy}$, $\mathcal{M}_2 = -W$, $\mathcal{M}_3 = I_K$ and $\mathcal{M}_4 = W^T$ and with the maximum likelihood estimate of W shown in Eqn. (3.60), the following can be obtained,

$$W^T \psi_y^{-1} = M_y^T \zeta_{yK}^T + M_y^T M_y \left[I_K - M_y^T M_y \right]^{-1} M_y^T \zeta_{yK}^T \tag{3.80}$$

Pre-multiplying the above expression by the expression for Φ_y obtained in Eqn. (3.78) yields the identity shown in Eqn. (3.71). ■

Theorem 3. *When the GPLLVM is restricted by the conditions: $F = 0$, $V = 0$ and $\psi_y = \sigma^2 I$, under the maximum likelihood parameter estimates, the statistics presented for monitoring the latent variables and the residuals in equations (3.31) and (3.40) are equivalent to the T^2 statistic in Eqn. (3.3) and the Q statistic in Eqn. (3.6) of the classical PCA based monitoring approach respectively.*

Proof. Under the restrictions given in theorem 3, the statistics presented in Eqn. (3.31) and (3.40) reduce to

$$\frac{1}{\sigma^2} y_n^T W \Phi_y \left[I_K - \Phi_y \right]^{-1} \Phi_y W^T y_n \frac{1}{\sigma^2}, \text{ and} \tag{3.81}$$

$$\frac{1}{\sigma^2} y_n^T \left[I_P - W \left[W^T W \right]^{-1} W^T \right] y_n, \tag{3.82}$$

respectively. Alternatively, these conditions can also be derived rigorously by starting from the restricted model instead of starting with the results obtained for the GPLLVM.

Making use of the results presented in lemma 4 simplifies the proof of equivalence. Substituting the results presented in equations (3.58), (3.63) and (3.64) to (3.81) and (3.82) reduces the statistics to the following,

$$y_n^T \eta_K \left(\Lambda_K \right)^{-1} \eta_K^T y_n, \text{ and} \tag{3.83}$$

$$\frac{1}{\sigma^2} y_n^T \left[I_P - \eta_K \eta_K^T \right] y_n, \tag{3.84}$$

respectively. We observe that Eqn. (3.83) is identical to the statistic presented in Eqn. (3.3) while Eqn. (3.84) is scaled by $1/\sigma^2$ with respect to the statistic presented in Eqn. (3.6). The scaling is due to the employment of the weighted least square solution as in Eqn. (3.35). However, it will not have any barring on the monitoring procedure as the control limit corresponding to the statistic will also be scaled with the same factor. ■

Remark 8. *In the classical PCA based monitoring, the Q statistic is shown to be non-negative sum of χ^2 random variables by making use of sample covariance of Y [26]. However, in the probabilistic model based monitoring, if the model covariance is used, it may lead to different control limits for monitoring the residuals unless the model covariance is same as the sample covariance.*

Remark 9. *In the literature, PPCA model based monitoring has been considered and the relevant statistics were presented as:*
For monitoring the latent variables [37]:

$$\frac{1}{\sigma^2} y_n^T W \Phi_y \Phi_y W^T y_n \frac{1}{\sigma^2}, \text{ and} \tag{3.85}$$

and for monitoring the model residuals [35, 37]:

$$\frac{1}{\sigma^2} y_n^T \left(I_P - \frac{1}{\sigma^2} W \Phi_y W^T \right) \left(I_P - \frac{1}{\sigma^2} W \Phi_y W^T \right) y_n \tag{3.86}$$

opposed to the statistics presented in equations (3.81) and (3.82) of this chapter. The differences arise from the following aspects:

1. *They assumed the null distribution of the estimate of the latent variable z_n to be $\mathcal{N}(0, I_K)$. However, as we have presented in Eqn. (3.81), the covariance of the estimate of z_n is $I_K - \Phi_y$, affecting the derived statistic in Eqn. (3.85)*

2. *They considered the estimate from the posterior, $p(\epsilon_{yn}|y_n)$ to be the residual, affecting the derived statistic in Eqn. (3.86). However, it should be noted that the likelihood $p(y_n|\epsilon_{yn})$ has a covariance WW^T which is non-invertible, rendering the posterior, $p(\epsilon_{yn}|y_n)$, to be intractable. Further, the computed residual ϵ_{yn} is a suboptimal reconstruction residual for y_n in the least squares sense.*

In the classical CCA based monitoring approach, two sets of latent variables are defined separately for the outputs and the inputs and can be monitored by monitoring the T_y^2 and T_x^2 statistics shown in Eqn. (3.13). As in Eqn. (3.16), the GPLLVM uses a same set of latent variables for both the inputs and the outputs. However, the latent variables from the GPLLVM can be inferred by conditioning either the outputs or the inputs separately and monitored as shown in Table 3.1.

Theorem 4. *When the GPLLVM is restricted by the conditions: $F = 0$, under the maximum likelihood parameter estimates, the statistics presented in rows 1 and 2 of Table 3.1 for monitoring the latent variables inferred separately from the outputs and the inputs are equivalent to T_y^2 and T_x^2 statistics of the CCA based monitoring approaches shown in Eqn. (3.13) respectively.*

Proof. For brevity, we only illustrate the equivalence between the statistic presented in row 1 of Table 3.1 under the conditions imposed by theorem 4 and the T_y^2 statistic of CCA presented in Eqn. (3.13).

The statistic presented in row 1 of Table 3.1 under the conditions imposed by theorem 4 reduces to,

$$y_n^T \psi_y^{-1} W \Phi_y \left[I_K - \Phi_y\right]^{-1} \Phi_y W^T \psi_y^{-1} y_n \tag{3.87}$$

Substituting the results shown in lemma 5, i.e., substituting equations (3.70) and (3.71) in Eqn. (3.87), the monitoring statistic reduces to the following form:

$$y_n^T \zeta_{yK} \zeta_{yK}^T y_n \tag{3.88}$$

which is identical to the T_y^2 statistic of CCA shown in Eqn. (3.13). Similar procedure would also yield equivalence of the statistic in row 2 of Table 3.1 and the T_x^2 statistic of the CCA based monitoring approach shown in Eqn. (3.13) are equivalent. ■

3.5 Simulation Example

In this section, we validate, 1) the results presented in section 3.3, and 2) the results presented in section 3.4, by means of numerical simulation examples.

1) We consider a GPLLVM with the following parameters:

$$W = \begin{bmatrix} 2.3 & -2.9 & 1.8 \\ 1.5 & 2.4 & -3.1 \end{bmatrix}^T,\ V = \begin{bmatrix} 1.2 & 3.2 & 1.3 \\ -2.3 & 1.7 & -2.4 \end{bmatrix}^T$$

$$F = \begin{bmatrix} 2.6 & 1.7 & -3.3 \end{bmatrix}^T,\ U \sim \mathcal{N}(0,1),\ Z \sim \mathcal{N}(0, I_2)$$

$$\psi_y = \begin{bmatrix} 0.8 & 0.2 & 0.3 \\ 0.2 & 0.5 & -0.4 \\ 0.3 & -0.4 & 0.9 \end{bmatrix},\ \psi_x = \begin{bmatrix} 0.8 & 0.4 & 0.3 \\ 0.4 & 0.9 & -0.2 \\ 0.3 & -0.2 & 0.8 \end{bmatrix}$$

Using the above model, we simulated Y and X. Followed by that, we performed monitoring on the simulated Y and X by means of various statistics presented in theorems 1 and 2 and Table 3.1. The statistics were monitored using the control limits derived from the corresponding null distributions with the desired α values. By performing 50 Monte-Carlo simulations each with 10^5 samples, we present the resulting fraction of false positives in Table 3.2. Results indicate that the fraction of false positives concurs with the considered α values. These numerical results verify the results presented in section 3.3.

Table 3.2: Fraction of type I error or false positives resulting from the control charts

S. NO	Monitored aspect	$\alpha = 5 \times 10^{-2}$	$\alpha = 5 \times 10^{-3}$
1	Variability in z_n inferred from y_n, u_n and x_n	$5 \times 10^{-2} \pm 1.9 \times 10^{-3}$	$5 \times 10^{-3} \pm 6.3 \times 10^{-4}$
2	Model residual statistic	$5 \times 10^{-2} \pm 1.9 \times 10^{-3}$	$5 \times 10^{-3} \pm 5.6 \times 10^{-4}$
3	Variability in z_n inferred from y_n and u_n alone	$4.9 \times 10^{-2} \pm 2.1 \times 10^{-3}$	$5 \times 10^{-3} \pm 6.4 \times 10^{-4}$
4	Variability in z_n inferred from x_n alone	$5 \times 10^{-2} \pm 1.8 \times 10^{-3}$	$5 \times 10^{-3} \pm 6.8 \times 10^{-4}$
5	Variability in observed y_n and x_n	$5 \times 10^{-2} \pm 1.9 \times 10^{-3}$	$5 \times 10^{-3} \pm 5.8 \times 10^{-4}$
6	Discrepancy between z_n inferred from x_n and from y_n and u_n	$5 \times 10^{-2} \pm 2.3 \times 10^{-3}$	$5 \times 10^{-3} \pm 5.4 \times 10^{-4}$

2) We simulated Y and X from a normal distribution with the following covariance matrices:

$$\Sigma_{yy} = \begin{bmatrix} 2.3 & 0.6 & -1.4 & 0.6 \\ 0.6 & 7 & -3.6 & 1.1 \\ -1.4 & -3.6 & 7.2 & -2.1 \\ 0.6 & 1.1 & -2.1 & 2.8 \end{bmatrix} \Sigma_{yx} = \begin{bmatrix} 1.2 & 0.3 & -0.7 & 0.3 \\ 0.3 & 3.5 & -1.8 & 0.6 \\ -0.7 & -1.8 & 3.6 & -1 \\ 0.3 & 0.6 & -1.1 & -1.4 \end{bmatrix}$$

$$\Sigma_{xx} = \begin{bmatrix} 3 & -0.8 & 1.8 & 0.6 \\ -0.8 & 2.8 & -2 & 0 \\ 1.8 & -2 & 4.6 & 0 \\ 0.6 & 0 & 0 & 1.3 \end{bmatrix}$$

We deployed the GPLLVM model based monitoring approach for monitoring Y with the PPCA restrictions and for monitoring both Y and X with the PCCA restrictions. In parallel, we also deployed the classical PCA and CCA based monitoring approaches for comparison. In both cases, we considered the number of latent variables to be 2 and computed the respective monitoring statistics. Figure 3.2 compares the T^2 and Q statistics obtained from PCA and PPCA respectively. Along the similar line, Figure 3.3 compares the T_x^2 and T_y^2 statistics obtained from CCA and PCCA. In both the figures the circle marker and asterisk marker indicate the statistics obtained from the classical technique and the probabilistic counterparts that are calculated from the GPLLVM under corresponding restrictions respectively. It can be seen from the figures that the statistics extracted from the classical techniques and the GPLLVM with corresponding restrictions are equivalent, verifying the results presented in section 3.4.

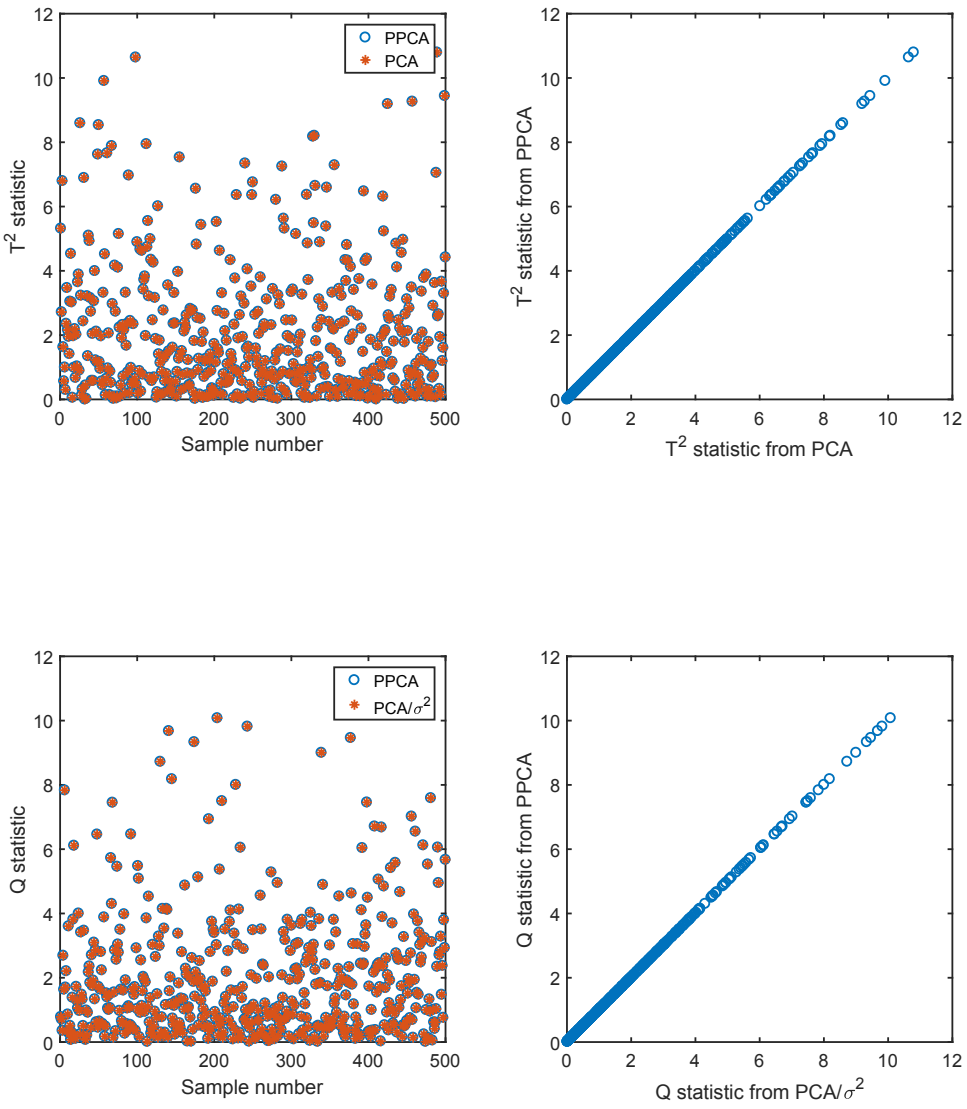


Figure 3.2: Statics obtained from PCA and PPCA: T^2 (top) and Q (bottom)

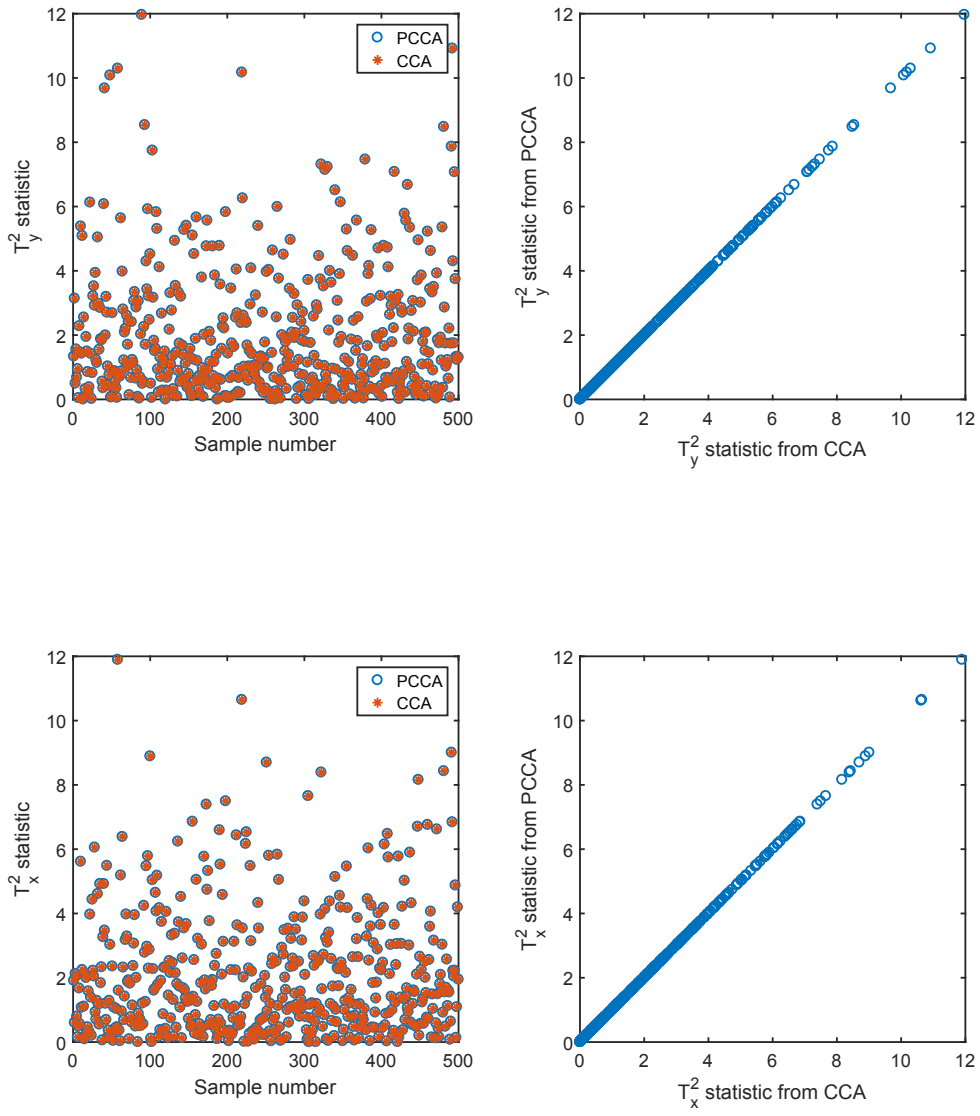


Figure 3.3: Statics obtained from CCA and PCCA: T_y^2 (top) and T_x^2 (bottom)

3.6 Summary

In this chapter, we proposed a unified probabilistic linear latent variable model that subsumes several commonly used linear models for process monitoring and derived monitoring statistics based on the same. This helps the researchers to view the various techniques for process monitoring under the same framework instead of looking

at them in isolation. The study provided insights into the equivalence of the classical multivariate techniques and their probabilistic counterparts by restricting the generalized model to the special cases. Ignoring the algorithmic differences in estimating the classical models and their probabilistic counterparts, there will not be any differences in the statistics derived for monitoring from both. Therefore, the real advantage of using the probabilistic models will only be with extending the models to deal with different data characteristics or distributions. The focus of the next chapter is to exploit this particular advantage of the probabilistic models.

Chapter 4

Multi-modal and dynamic process monitoring using probabilistic models

4.1 Introduction

In this chapter, we explore the possibility of extending the linear Gaussian models to obtain a model to approximate the non-Gaussian distributions. We develop a new probabilistic model that makes use of the probabilistic variant of the popular multivariate technique, principal component analysis (PCA), the probabilistic principal component analyser (PPCA) model, as a building block.

Due to its popularity, several extensions and modifications of the PCA based approach are available in the literature for process monitoring applications. One of the widely considered extensions is the application of PCA to monitor a finite sequence of observations [81, 19, 82]. The resulting monitoring model is called the dynamic PCA model and it is suitable for handling temporally correlated observations. The probabilistic version of PCA, the PPCA model allows the formulation of a mixture model, known as the mixture PPCA model [31]. The mixture PPCA model consists of a convex combination of several local PPCA models and it can be utilized for fault detection when the process variables tend to have multi-modal spreads or follow a non-Gaussian distribution as illustrated by several authors [83, 84, 35, 85]. One of the other approaches that applies PCA for monitoring multi-modal processes utilizes the Gaussian mixture model to section the data into several clusters and uses PCA to

model the data within each of the clusters [86]. A non-parametric PCA based model obtained through kernel trick, called the kernel PCA model can be used to describe the non-Gaussian distributions and has been previously considered for process monitoring [87]. The ideas of kernel PCA and dynamic PCA can be combined to address the process monitoring problems when the data with temporal correlated observations and non-Gaussian distribution are encountered [88]. However, for the industrial applications with a large historical database, implementation of the non-parametric approach is not computationally viable.

We find that combining the ideas of dynamic PCA and mixture PPCA could be a potential solution for monitoring multi-modal processes with temporally correlated observations. However, if we were to apply such models for process monitoring, we need to address the following challenges, 1) lack of scalability of the mixture PPCA model for large scale high dimension data without local optima convergence and over-fitting issues and 2) difficulty in selecting the appropriate dimension for the latent variables in the model. We address the scalability challenge by a two-stage estimation approach and the dimension selection challenge through the Bayesian regularization approach. The proposed solution strategy has the potential to provide a scalable mixture model. We call the resulting model from the proposed approach as the two-layer mixture Bayesian PPCA model. We illustrate the applicability of this model in a couple of case studies.

4.1.1 Organization of this chapter

In section 4.2, we provide a brief introduction to the PCA based models used in process monitoring application. In section 4.3, we introduce the proposed modelling strategy. In section 4.4, we present the formulation of the proposed model. In section 4.5, we show the process monitoring scheme using the proposed model. In sections 4.6 and 4.7, we present two case studies and highlight the performance of the proposed model by comparing it with the performances of the other PCA models. In section 4.8, we provide the concluding remarks.

4.2 Background

In this section, we provide a brief introduction to the PPCA, dynamic PCA and mixture PPCA models. These models are relevant for the development of the proposed model in this chapter.

4.2.1 PPCA

PPCA is a latent variable model. It uses lower dimensional latent variables to explain the generative process of the observations. Consider a set of observations $Y = \{y_1, y_2, .., y_N\} \in \mathbb{R}^{N\times D}$ and the corresponding latent variables $Z = \{z_1, z_2, .., z_N\} \in \mathbb{R}^{N\times M}$, the PPCA model relates the observations and the latent variables through a linear relationship of the following form,

$$y_n = W z_n + \mu + e_n, \qquad \forall n \tag{4.1}$$

where $y_n \in \mathbb{R}^D$, $z_n \in \mathbb{R}^M$, $M < D$, $W \in \mathbb{R}^{D\times M}$ is the loading matrix of the model and $\mu \in \mathbb{R}^D$ is the mean of the observations. Noise ($e_n \in \mathbb{R}^D$) in the observations are considered to be independent and identically distributed and follow a multivariate Gaussian distribution with zero mean and isotropic covariance $\sigma^2 I_D$. The latent variables are considered to follow a multivariate Gaussian with zero mean and identity covariance ($z_n \overset{i.i.d}{\sim} \mathcal{N}(0, I_M)$). From Eqn. (4.1), we can interpret the distribution of the observations given the latent variables as the following,

$$p\left(Y|Z, W, \mu, \sigma^2\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|W z_n + \mu, \sigma^2 I_D\right) \tag{4.2}$$

where each observation y_n is shown to follow a multivariate Gaussian distribution with mean $W z_n + \mu$ and covariance $\sigma^2 I_D$.

The marginal distribution of the observations can be obtained by marginalizing the latent variables from the joint distribution of the observations and latent variables as the following,

$$p\left(Y|W, \mu, \sigma^2\right) = \int_Z p\left(Y|Z, W, \mu, \sigma^2\right) p\left(Z\right) dZ$$

$$= \int_Z \prod_{n=1}^{N} \mathcal{N}\left(y_n|W z_n + \mu, \sigma^2 I_D\right) \mathcal{N}\left(z_n|0, I_M\right) dZ$$

$$= \mathcal{N}\left(y_n|\mu, WW^{'} + \sigma^2 I_D\right) \tag{4.3}$$

where the distribution of the set of observations is given by a product of marginal distributions of the individual observations, which is given by a multivariate Gaussian distribution with mean μ and covariance $WW^{'} + \sigma^2 I_D$.

4.2.2 Mixture PPCA

The PPCA model allows the construction of the mixture PPCA model. The mixture PPCA model consists of a convex combination of a finite number of local PPCA models. Let us consider a mixture PPCA model with S local models. When each local model $s \in S$ explains π^s proportion of the total observations, the prior probability of an observation y_n being explained by a local model s is given by,

$$p\left(s_n = s\right) = \pi^s \tag{4.4}$$

where

$$\sum_{s=1}^{S} p\left(s_n = s\right) = \sum_{s=1}^{S} \pi^s = 1 \tag{4.5}$$

where $s_n \in [1, S]$ is a categorical variable and it follows a categorical distribution with parameters, $\pi = \{\pi^1,\ \pi^2, ...,\ \pi^S\}$. Therefore, the distribution of the observations is given by a mixture PPCA model as the following,

$$\begin{aligned} p\left(Y|Z, W, \mu, \sigma^2\right) &= \prod_{n=1}^{N} \sum_{s=1}^{S} \mathcal{N}\left(y_n|W^s z_n^s + \mu^s, \sigma^2 I_D\right) p\left(s_n = s\right) \\ &= \prod_{n=1}^{N} \sum_{s=1}^{S} \mathcal{N}\left(y_n|W^s z_n + \mu^s, \sigma^2 I_D\right) \pi^s \end{aligned} \tag{4.6}$$

where each local model s has its own parameters, $\{W^s,\ \mu^s\}$. When fitted to a dataset, the mixture PPCA model divides the dataset into clusters (in this case, into S clusters). A probability measure $q(s_n = s) \forall s, n$ that indicates the posterior probability of an observation belonging to a particular cluster is also obtained along with the parameters of the model when fitted to a dataset. The mixture PPCA model is suitable for monitoring multi-modal processes where each local model can be used to describe a particular operating mode. In addition, when a sufficient number of local models are allowed, the mixture PPCA model can be used as an approximation for any non-Gaussian distributions.

4.2.3 Dynamic PCA

The dynamic PCA model is obtained through the following steps, 1) the observations over a time window are stacked together to form the lag augmented data matrix and 2) then, PCA is applied on the space of lag augmented observations. Consider an observation sequence, $\{y_1, y_2, ..., y_n, ..., y_N\}$, where y_n is a vector of observations at time instant n. When we augment the l past observations with the observations at each time instant, the resulting lag augmented observations form a matrix of the following form,

$$X = [x_1, x_2, ..., x_n, ..., x_{N-l-1}] = \begin{bmatrix} y_{l+1} & . & . & y_{N-1} & y_N \\ y_l & . & . & y_{N-2} & y_{N-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ y_1 & . & . & y_{N-l-1} & y_{N-l} \end{bmatrix} \tag{4.7}$$

where x_n corresponds to a vector of lag augmented observations. If large enough l is chosen, the columns of X will become mutually independent. Then, PCA is applied on the space of variable x_n to obtain the dynamic PCA model. This amounts to modelling the covariance of x_n. The model captures the temporal correlations among the variables up to lag l. In the probabilistic version, the dynamic PPCA model can be obtained by using Eqn. (4.1) to model x_n shown in Eqn. (4.7).

4.3 Proposed Model

4.3.1 A straightforward extension

Combining the ideas of dynamic PCA and mixture PPCA model to form a mixture model that can be named as the mixture dynamic PPCA model is a straightforward extension. It can be obtained by fitting the model shown in Eqn. (4.6) to X instead of Y.

X (instead of Y) in Eqn. (4.7) using the model shown in Eqn. (4.6). However, in doing so, one has to address the following challenges,

1. Scalability of the model to incorporate a large number of local models: In the past, the success of mixture PPCA and dynamic PCA models has been demonstrated only on the simulation studies that require a known and a smaller number of local

models. Data from industrial processes may require a mixture model with a large number of local models to approximate the true data distribution. Currently, the popular estimation approaches such as the expectation maximization (EM) algorithm and the variational Bayesian expectation maximization (VBEM) algorithm available for estimating the mixture models are prone to local optima convergence [31, 69]. They require a good initial guess for the model parameters and cluster identities. It is difficult to provide a reliable initial guess when we have a large number of local models and high dimension observations.

2. Dimension reduction: The dynamic model has to handle the augmented observations of higher dimension when compared to the original observations. Dimension reduction is essential to make sure that the model captures only the appropriate amount of variance and leaves the noise in the observations to the noise part of the model. Dimension reduction is an inbuilt property of these models as they model the observations as a function of lower dimension latent variables. However, the dimension of the latent variables has to be defined by the users *a priori*. In the mixture version, this means, the users have to choose the appropriate dimension for the latent variables of each local model. This would be a tedious task if carried out on a trial and error basis.

4.3.2 The proposed solution strategy

We address the above-mentioned challenges in building the mixture model in the following ways,

1. Divide and conquer strategy that addresses the scalability challenge: Instead of trying to fit a mixture model with a large number of local models at once, we divide this task between two stages. We split the data into several subsets known as the clusters at the first stage. In the second stage, we fit mixture models with a manageable number of local models to each of the identified clusters. By combining the mixture models fitted to each of the clusters, we obtain a model for the entire dataset with a large number of local models. Illustrative example of this approach is provided in Fig. 4.1. In the first stage, the data is divided into two clusters and in the second stage to each cluster, mixture models with three local models are fitted for

the case illustrated in Fig. 4.1. At the end, the number of local models for the data becomes six, however, at a given stage, only a model with two or three local models is identified. We are motivated to incorporate this idea mainly because, we will be in a better position to generate good initial guesses when we try to fit a mixture model with a smaller number of local models as opposed to a model with a large number of local models.

We achieve both clustering in the first stage and mixture model fitting in the second stage using the mixture PPCA models. In the first stage, the model divides the data into clusters and also provides dimension reduction. In the second stage, the models are fitted to the clusters with lower dimension latent variables obtained in the first stage. We show that the models fitted in the both stages can be collapsed to form a mixture Gaussian model that consists of a large number of local models. We point out the advantages of this model over a mixture dynamic PPCA model that is estimated in a single stage such as the role of divide and conquer strategy in reducing the risk of local optima convergence and obtaining a parsimonious model. The proposed model is similar to that of deep mixture of factor analysers proposed in [89], however, our model differs by the use of PPCA models as the building blocks.

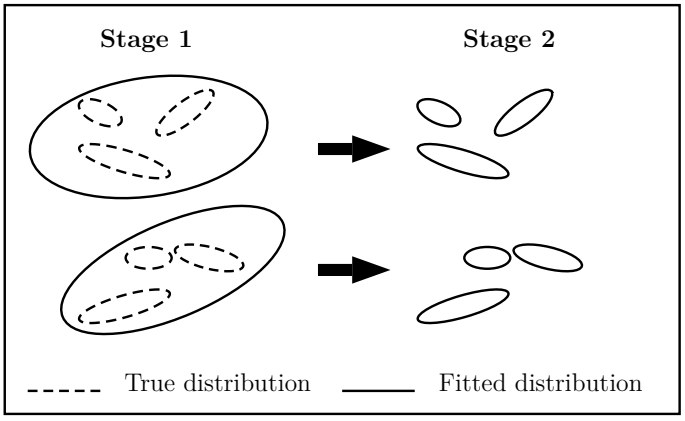


Figure 4.1: Illustrative representation of the two-layer mixture PPCA model.

2. The Bayesian regularization approach that addresses the dimension reduction challenge: To automate the process of dimension reduction in the both stages of model estimation, we incorporate a Bayesian regularization approach that penalizes the the loading parameters of the local models as previously illustrated in [69, 35,

33]. It incorporates a prior distribution for each column of the loading matrices of the local models. The prior distribution penalizes the insignificant columns of the loading matrices during the estimation stage. The columns corresponding to the latent variables that are insignificant in explaining the observations converge close to zero during the estimation. At the end of the estimation, the dimension of the latent variables in each local model can be inferred by identifying the number of non-zero columns in the loading matrices.

4.4 Formulation of the Proposed Model

In this section, we present the following, 1) incorporation of the Bayesian regularization approach to obtain the mixture Bayesian PPCA model, 2) formulation of the two-layer mixture Bayesian PPCA model, 3) a procedure for obtaining a mixture model by collapsing the two-layer model and 4) finally, the potential advantages and disadvantages of the proposed modelling strategy.

4.4.1 Mixture Bayesian PPCA

Loading matrix $W^s \in \mathbb{R}^{D\times M}$ is the important parameter in determining the latent variable dimension of the local model s in the mixture PPCA model shown in Eqn. (4.6). It can be seen that the contribution of the latent variables in explaining the observations solely depends on the loading parameters. When a particular column $m \in M$ of the loading matrix W^s contains only zero entries, the contribution from that latent dimension m in explaining the observations becomes negligible. Therefore, one way to achieve automatic dimension reduction is by regularizing the columns of W^s such that the insignificant entries are penalized. This in turn favours zero (or close to zero) entries on the columns of W^s corresponding to the insignificant latent variables. The regularization can be achieved by incorporating the hierarchical prior distributions for the columns of W^s.

In this work, each column of W^s is considered to follow a multivariate Gaussian distribution prior as the following,

$$W_m^s|\nu_m^s \sim \mathcal{N}\left(0, \nu_m^{s-1} I_D\right) \tag{4.8}$$

where W_m^s is the m^{th} column of the loading matrix, W^s and ν_m^s is the precision variable (inverse of variance) and I_D is an identity matrix of dimension D. The precision variable ν_m^s is assumed to follow gamma distribution as the following,

$$\nu_m^s | a^*, b^* \sim \Gamma\left(a^*, b^*\right) \tag{4.9}$$

where a^* and b^* are the scale and rate parameters of the Gamma distribution respectively. We call the resulting model, the mixture Bayesian PPCA model. For estimating this model, we follow a similar approach used to estimate Bayesian mixture latent variable models in [69, 35, 67] and the detailed derivation for this is given in C.1 of appendix C. The approach for inferring the zero columns in the loading matrix after the estimation is provided in C.1.5 of appendix C.

4.4.2 Two-layer mixture Bayesian PPCA

In this subsection, we discuss how the two-layer model is obtained. Originally, we want a model for our data X as a function of some latent variable T as the following,

$$X = f\left(T\right) \tag{4.10}$$

where f is the desired function, however, it may be complicated in nature. It is when the idea of two-layer model becomes useful. Instead of identifying f directly, we may identify two relatively simpler models and recover the original function f from the simpler models. The idea is to identify a model for an intermediate variable Z as a function of T and a model for X as a function of Z as the following,

$$X = g\left(Z\right), \; Z = h\left(T\right) \tag{4.11}$$

where g and h are simpler functions compared to f. Then, we recover the original function f as the following,

$$X = f\left(T\right) = g\left(h\left(T\right)\right) \tag{4.12}$$

When Z and T are observed variables, we can afford to estimate h first and then, g. However, in our case, both Z and T are latent variables and therefore, from X, we need to estimate g and infer Z at first and then, from Z we need to estimate h

and infer T. In our case, both g and f are mixture Bayesian PPCA models. X is the augmented observation matrix obtained using Eqn. (4.7). In the first layer model, X takes the position of the output Y and Z takes the position of the latent variables (Z) in Eqn. (4.6). In the second layer models, Z takes the position of the output Y and T takes the position of the latent variables in Eqn. (4.6).

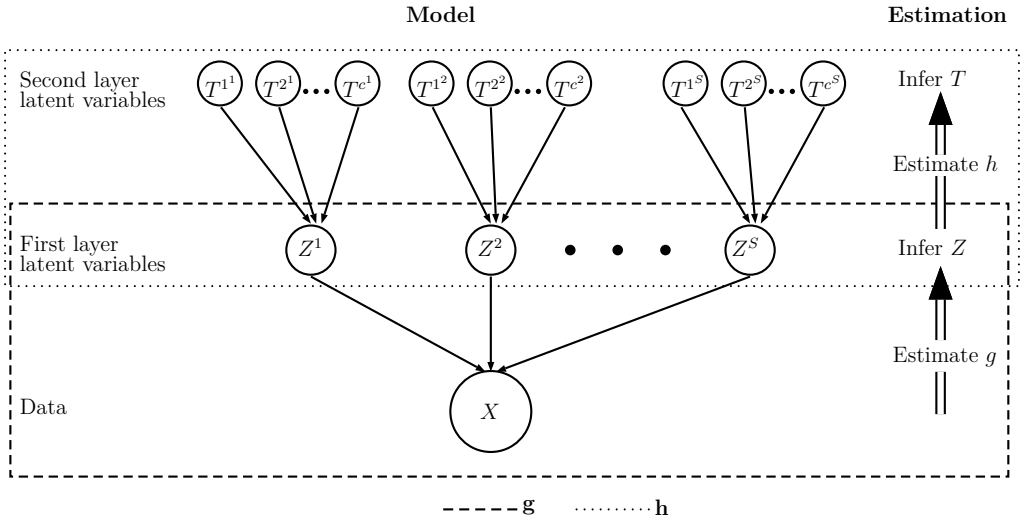


Figure 4.2: Schematic representation of the proposed model and the flow of estimation. Data and the latent variables in the model are represented by encircled nodes.

Schematic representation of the model and the estimation flow is shown in Fig. 4.2. We start with a mixture Bayesian PPCA model with an arbitrarily small number of local modes (S) to fit the data X. This model is equivalent to g in the above illustration. The resulting model is defined by the parameters $\{W^s,\ \mu^s,\ \pi^s,\ \sigma^2\}\ \forall s$ and also the posterior probability measure $q\left(s_n = s\right) \forall n, s$. We can split the data, X into S clusters as, $\{X^1,\ X^2, ..., X^s, ...,\ X^S\}$, where, X^s refers to a cluster of observations and $s \in [1, S]$. The clustering can be achieved using $q\left(s_n = s\right) \forall n, s$. For whichever s, an observation x_n has the highest value of $q(s_n = s)$, it is assigned to that particular cluster. From the first layer model, we can sample the latent variables $\{Z^1,\ Z^2, ...,\ Z^S\}$ corresponding to the clusters $\{X^1,\ X^2, ...,\ X^S\}$ from the posterior distribution of the latent variables.

In the second layer, to $Z^s, \forall s \in [1, S]$, we fit a mixture Bayesian PPCA model of

the following form,

$$p(z_n^s|t_n, W, \mu, \beta) = \sum_{c^s=1}^{C} \mathcal{N}\left(z_n^s|W^{c^s} t_n^{c^s} + \mu^{c^s}, \beta^s I\right) p\left(c_n^s = c^s\right), \qquad z_n^s \in Z^s \tag{4.13}$$

where $t_n^{c^s}$ is the latent variable for the second layer that describes $z_n^s \in Z^s$, the mixture model above contains C local models and the proportion of data explained by each model is given by $p\left(c_n^s = c^s\right) = \pi^{c^s}$. Sum of the proportions of the data explained by each local model should be equal to one as shown below,

$$\sum_{c^s=1}^{C} p\left(c_n^s = c^s\right) = \sum_{c^s=1}^{C} \pi^{c^s} = 1 \tag{4.14}$$

The approach discussed till here is summarized in Table. 4.1, which constitutes the procedure for estimating the two-layer model.

Table 4.1: The approach for estimating the two-layer mixture Bayesian PPCA model

Step 1	Define parameters: Lag l, cardinality of mixture models S and C, reasonable latent variable dimension M($<$ the dimension of augmented observations) for the first layer model and regularization parameters a^* and b^*
Step 2	Obtain the augmented observations X from the original observations as shown in Eqn. (4.7)
Step 3	Fit a mixture Bayesian PPCA model with S local models to X using the algorithm provided in section C.1 of appendix C
Step 4	Infer dimension of the latent variable in the local models using the procedure shown in section C.1.5 of appendix C
Step 5	Divide X into S clusters, $\left\{X^1, X^2, ..., X^S\right\}$ using $q\left(s_n = s\right) \forall n, s$
Step 6	Sample the latent variables $\left\{Z^1, Z^2, ..., Z^S\right\}$
Step 7	For $s = 1 : S$
Step 8	Define parameter: Reasonable latent variable dimension P ($<$ the dimension of Z^s)
Step 9	Fit a mixture Bayesian PPCA model with C local models to Z^s using the algorithm in section C.1 of appendix C
Step 10	Infer the dimension of the latent variables in the local models using the procedure in section C.1.5 of appendix C
Step 11	End For

4.4.3 Collapsing the two-layer model to form a mixture Gaussian model

In this subsection, we show that the models identified in two stages can be combined/collapsed to form a single layer mixture model. The models can be collapsed to obtain a model of the following form,

$$p(X|\mu, \Sigma, \pi) = \prod_{n=1}^{N} \sum_{k=1}^{K} \mathcal{N}(x_n|\mu^k, \Sigma^k) p(k_n = k) \tag{4.15}$$

where

$$\sum_{k=1}^{K} p(k_n = k) = \sum_{k=1}^{K} \pi^k = 1 \tag{4.16}$$

where the distribution of each observation x_n is given by a Gaussian mixture model with $K = S \times C$ local models, μ^k and Σ^k are the mean and covariance parameters of the local model $k \in K$ and π^k is the proportion of data explained by the local model $k \in K$.

From the first layer model, the observation x_n is Gaussian distributed when conditioned on z_n^s and s as the following,

$$p(x_n|z_n^s, W^s, \mu^s, \sigma^2) = \mathcal{N}(x_n|W^s z_n^s + \mu^s, \sigma^2 I) \tag{4.17}$$

The latent variable z_n^s, is also Gaussian distributed when conditioned on $t_n^{c^s}$ and c^s from Eqn. (4.13) of the second layer models as the following,

$$p(z_n^s|t_n^{c^s}, W^{c^s}, \mu^{c^s}, \beta^s) = \mathcal{N}\left(z_n^s|W^{c^s} t_n^{c^s} + \mu^{c^s}, \beta^s I\right) \tag{4.18}$$

The above two distributions when multiplied give the joint distribution of x_n and z_n^s. When the latent variable z_n^s is marginalized from the joint distribution, we can obtain,

$$\begin{aligned} & p\left(x_n|t_n^{c^s}, W^s, \mu^s, \sigma^2, W^{c^s}, \mu^{c^s}, \beta^s\right) \\ &= \int_{z_n^s} p\left(x_n, z_n^s|t_n^{c^s}, W^s, \mu^s, \sigma^2, W^{c^s}, \mu^{c^s}, \beta^s, \right) dz_n^s \\ &= \mathcal{N}\left(x_n|W^s\left(W^{c^s} t_n^{c^s} + \mu^{c^s}\right) + \mu^s, W^s \beta^s W^{s'} + \sigma^2 I\right) \end{aligned} \tag{4.19}$$

Further, we know that the prior of $t_n^{c^s}$ is a multivariate Gaussian with zero mean and identity covariance. Therefore, $t_n^{c^s}$ can also be marginalized to obtain the distribution of the observation x_n given a combination of s and c^s as the following,

$$p\left(x_n|W^s,\mu^s,\sigma^2,W^{c^s},\mu^{c^s},\beta^s\right) = \int_{t_n^{c^s}} p\left(x_n,t_n^{c^s}|W^s,\mu^s,\sigma^2,W^{c^s},\mu^{c^s},\beta^s\right) dt_n^{c^s}$$

$$= \mathcal{N}\left(x_n|W^s\mu^c+\mu^s, W^s\left(\beta^s + W^{c^s}W^{c^{s'}}\right)W^{s'} + \sigma^2 I\right) \tag{4.20}$$

Similarly, for each combination of s and c, we can obtain a local Gaussian model for the observations. In total, there will be $K = S \times C$ local models. When $k \in K$ corresponds to a particular combination of s and c^s, the proportion of the data explained by it (π^k) is given by,

$$\pi^k = \pi^s\pi^{c^s} \tag{4.21}$$

This is because the local model s in the first layer explains π^s portion of the data and of which, π^{c^s} portion of the data is explained by the local model c^s in the second layer. The mean and covariance parameters of each local model $k \in K$ from Eqn. (4.20) can be expressed as,

$$\mu^k = W^s\mu^{c^s} + \mu^s, \quad \Sigma^k = W^s\left(\beta^s + W^{c^s}W^{c^{s'}}\right)W^{s'} + \sigma^2 I \tag{4.22}$$

where k in the above equation corresponds to a particular combination of s and c^s.

4.4.4 Comments on the proposed model

By observing μ^k and Σ^k in Eqn. (4.22), it can be seen that the local models in the collapsed model resulting from a common first layer local model s share common parameters, whereas, in the conventional mixture PPCA model, each local model is defined by its own parameters. Therefore, in comparison, the proposed model may require a lesser number of parameters to define a mixture model with a single layer model of a similar complexity. This is true under certain conditions as shown in the following proposition,

Proposition 5. *A mixture PPCA model with K local models and P as the dimension of the latent variables in each local model fitted to observations with dimension D has more loading and mean parameters compared to a two-layer model with S local models*

in the first layer, C local models for each $s \in S$ in the second layer, M as the dimension of the latent variables in the first layer and P as the dimension of the latent variables in the second layer when,

$$C\left(1-\frac{1}{r_1}\right)\frac{1}{r_2} \geq 1 \tag{4.23}$$

$$K = SC \tag{4.24}$$

where $r_1 = \frac{M}{D}$ and $r_2 = \frac{P}{M}$ are the dimension reduction ratios in the first and second layers respectively.

The proof for the above proposition is provided in section C.2 of appendix C. From Eqn. (4.23), it can be seen that through appropriate dimension reduction ratios and the choice of number of local models in the second stage, we can always obtain a parsimonious model.

Estimating a mixture model requires a good initialization of cluster identities of the observations to avoid convergence to local optima. Generally, the K-means clustering algorithm is used as an initializer (more details on the initialization is provided in C.1.4). However, the K-means clustering algorithm is also susceptible to convergence to local optima as the number of clusters increases. Therefore, estimating a mixture model with a large number of local models in a single stage inevitably suffers from convergence to local optima. The proposed model development strategy involves only identifying a smaller number of local models in each of the two stages compared to the single stage model identification approaches. Therefore, we expect the proposed strategy to be more robust to convergence issues.

The proposed model also has some disadvantages. The above derivation of the collapsed mixture model is consistent only when the observations can be clustered into S perfect (hard) clusters in the first layer. If a set of observations are shared between the clusters, assigning observations strictly to individual clusters may not always be reasonable. Other drawback is that the numbers of local models in each layer have to be decided by the users. However, this problem can be addressed through cross validation. In our work, the number of local models in the first layer is arbitrarily chosen. The number of local models in the second layer is chosen based on the

log likelihood of the parameters of the collapsed mixture model in validation data ($\ln p(X^{val}|\mu, \Sigma, \pi)$, where, X^{val} is the validation set).

4.5 Fault Detection Using the Proposed Model

The model obtained in Eqn. (4.15) from the normal operating data of the process gives the probability density function (PDF) for the observations generated from the normal operating conditions (NOC). We define a likelihood based threshold for fault detection using the following definition of the probability density function,

$$\int_{x:\ p(x|\mu,\Sigma,\pi)\geq\delta} p\left(x|\mu, \Sigma, \pi\right) dx = \gamma, \qquad \gamma \in [0, 1] \tag{4.25}$$

where, $100\gamma\%$ represents the percentage of normal operating data that have the probability density value greater than or equal to δ. A δ value corresponding to a particular γ obtained using the definition in Eqn. (4.25) would contain $100\gamma\%$ of the normal operating data points and exclude $100(1-\gamma)\%$ of the normal operating data points. If the following rules in equation (4.26) and (4.27) are deployed for fault detection with an assumption that the new observations are always generated from the same distribution as of the data used for training the model, then we can have a fault detection system with a false positive rate of γ.

$$x^{new} \in Normal \quad if \quad p\left(x^{new}|\mu, \Sigma, \pi\right) > \delta \tag{4.26}$$

$$x^{new} \in Faulty \quad if \quad p\left(x^{new}|\mu, \Sigma, \pi\right) < \delta \tag{4.27}$$

This approach has been previously used for fault detection with the Gaussian mixture models and kernel density models in [90] and [39] respectively. To obtain a δ that corresponds to a specific γ, we need the integration in Eqn. (4.25) to be tractable. However, it cannot be achieved analytically for many PDFs. This problem is overcome by generating a large number of samples from the PDFs and identifying a δ that encompasses $100\gamma\%$ of the generated samples. In our implementation δ values are obtained correspond to $100\gamma\% = 99.97\%$. In our study, the proposed model and all the other compared models are deployed for fault detection using the above-mentioned procedure. In the implementation, we may encounter numerical underflow if we directly

use the likelihood values for monitoring. Instead, the log likelihood can be used. Here, we used the negative log likelihood values as a test static as shown below,

$$x^{new} \in Faulty \quad if \quad -\ln p\left(x^{new}|\mu, \Sigma, \pi\right) > -\ln \delta = \delta' \tag{4.28}$$

When the negative log likelihood of a new observation is greater than δ', the observation is declared faulty. Since log is a monotonic function, the rule in Eqn (4.27) remains intact.

4.5.1 Performance metrics

Three performance metrics were used to evaluate and compare the performance of the proposed model when deployed for fault detection.

Percentage of false positives:

The percentage of false positives is computed as,

$$False\,positives\,(\%) = \frac{Number\,of\,alarm\,instances\,in\,fault\,free\,test\,data}{Number\,of\,fault\,free\,observations\,in\,the\,test\,data} \times 100 \tag{4.29}$$

Lower values of the percentage of false positives correspond to better performance of the deployed technique as it gives less nuisance to the operators in the plant.

Fault detection rate:

The fault detection rate is computed as,

$$Detection\,rate(\%) = \frac{Number\,of\,alarm\,intants\,in\,the\,faulty\,observations}{Total\,number\,of\,faulty\,observations} \times 100 \tag{4.30}$$

Higher detection rate for a fault means that the technique is better at detecting that particular fault.

Time of fault detection:

For faults where the exact time of fault occurrence is unknown, we use the time instant at which the PPCA model detects the fault as a reference. This metric is calculated by taking the difference between the reference value and the time instant

at which the fault is detected by the proposed model. In these cases, higher values correspond earlier detection.

For faults where the exact time of fault occurrence is known, the time of fault detection is computed by taking the difference between the time of fault detection (first alarm) and the time of fault occurrence. In these cases, lower values correspond to earlier detection.

4.6 Case study 1: Sulphur Recovery Unit (SRU)

Our first case study is an SO_2 breakthrough detection problem in a sulphur handling plant. Previously, Gonzalez et al. [39] studied this problem using kernel density estimation and Bayesian networks by monitoring multiple sulphur recovery units (SRUs) and a tail gas treatment unit (TGTU). We monitor an SRU which is a known contributor for SO_2 breakthrough once in the past.

4.6.1 Process description

The sulphur handling plant is an integral part of the oil sands upgrading process. It provides control over the sulphur emission. The plant considered here recovers sulphur present in upstream amine acid gas (AAG) and sour water acid gas (SWAG). The units of the plant are shown in Fig. 4.3, which includes 1) multiple sulphur recovery units (SRUs), 2) a tail gas treatment unit (TGTU) and 3) a thermal oxidizer unit (TOU). The majority of the sulphur content present in the acid gases is recovered in the form of elemental sulphur in the SRUs. The gas leaving the SRUs is called the tail gas that contains sulphur in the form of SO_2, which is then reduced to H_2S by means of catalytic reactors in the TGTU and recovered through amine absorption. The remaining gas is thermally oxidized in the TOU and sent to stack.

SO_2 breakthrough, i.e, SO_2 leaving the TGTU unconverted is one of the undesirable events that happens in the TGTU. It was found that the faults in the SRUs have been the major contributors to the SO_2 breakthrough problems historically. When the faults in SRUs are detected early, the tail gas leaving the SRUs can be bypassed to TOU to avoid a potential SO_2 breakthrough problems in the TGTU. We deployed the

developed fault detection tool to detect one such fault that led to SO_2 breakthrough in the past.

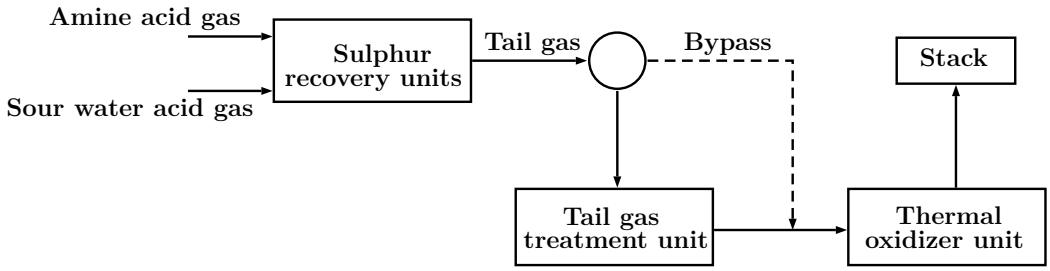


Figure 4.3: Units of sulphur handling plant

Sulphur recovery unit

A simplified schematic diagram of an SRU is shown in Fig. 4.4. The SRU draws combustion air in proportion to the acid gases entering the unit. Before the reaction stage, reactants entering the catalytic reactors are preheated in a preheating furnace. The reaction mixture leaving the furnace is sent through two catalytic reactors. Sulphur components present in the acid gases are converted to elemental sulphur form in the catalytic reactors. The elemental sulphur is then condensed in a sulphur condenser and recovered. The gas leaving the plant after the recovery of sulphur is called the tail gas. The tail gas leaving the plant contains traces of SO_2, which is further treated in the TGTU.

We are aware of one of the SO_2 breakthrough incidents that happened in the past because of a blockage in the sulphur condenser. This blockage led to increase in the concentration of sulphur components in the tail gas. Highly concentrated tail gas eventually led to SO_2 breakthrough problem in the TGTU. We set up the fault detection problem to detect this particular event from the data.

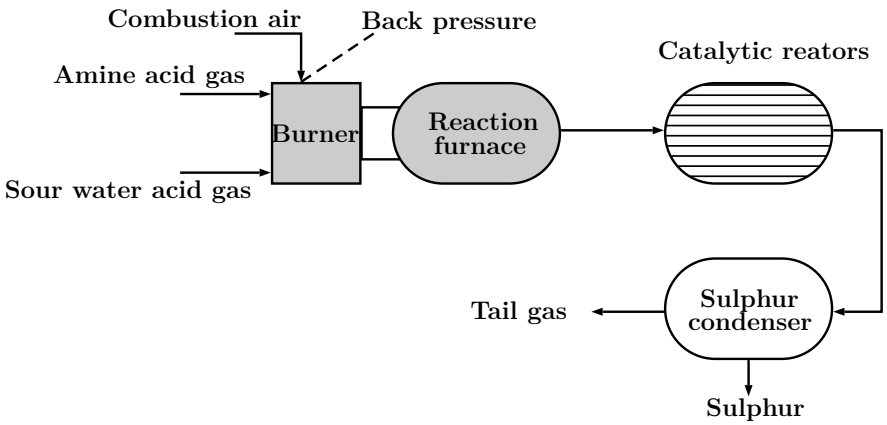


Figure 4.4: Schematic representation of a sulphur recovery unit

Data description

The data that belong to the year when the fault occurred was chosen for the study. We had the observations for the variables listed in Table. 4.2 available to us. Among the variables that are listed, pressure drop is a calculated variable from the pressure measurements upstream and downstream of the SRU and all the other variables are measured variables.

Table 4.2: Tags used for process monitoring

Tag number	Tag
1	AAG controller
2	SWAG controller
3	AAG flow rate
4	Air demand
5	Air controller
6	Pressure drop (Back pressure - down stream pressure)
7	SO_2 concentration
8	H_2S concentration

Table 4.3: Data summary

Plant	SRU
Number of tags	8
Period	1 year
Sampling interval	1 minute
Training data	4 months
Validation data	2 months
Testing data	6 months
Number of known faults	1

One year of data was split into two halves. The data from the first six months were used for training the model and from the second six months were used for testing. The first half of the data was further split randomly into two sets. Two thirds of it were used for the training and one third of it was for cross-validation. In the test set, the exact time of fault occurrence is unknown, however, the day of fault occurrence is known and the data belonging to that particular day were kept for evaluating the fault detection ability of the model and the rest of the testing data were used to evaluate the model in terms of false positives. These details are summarized in Table. 4.3.

4.6.2 Results and discussion

Base case

The PPCA and dynamic PPCA models were used as the base cases for comparison. The proposed model was evaluated based on the improvement in performance achieved from that of the PPCA and dynamic PPCA models. For the dynamic PPCA model, the lag was set as 14 and this amounts to monitoring a 15 minute multivariate sequence. The fault detection performances of the resulting models are shown in Table. 4.4.

The percentage false positives given by the dynamic PPCA model was slightly higher than that of the PPCA model. This could be due to the fact that the dynamic PPCA model extracts more information about the NOC and expects the test data to behave accordingly. This could also be the reason that the dynamic model performed better in terms of time of fault detection. It detected the fault 6 minutes before the static PPCA model could detect. The control charts obtained from the static and

the dynamic models are shown in figures 4.5 and 4.6 respectively. The control charts are shown for the observations closer to the period of fault occurrence. The black trends in the control charts correspond to the test statistics and the red solid lines correspond to the fault detection thresholds. It can be seen that the test statistics of the PPCA and dynamic PPCA models cross their respective thresholds at 565^{th} and 559^{th} minutes respectively.

Table 4.4: Performances achieved by the base case models

Model	Components	Lags	False positives (%)	Time of fault detection
PPCA	1	0	1.13	Base case
Dynamic PPCA	1	14	1.16	6 minutes before PPCA

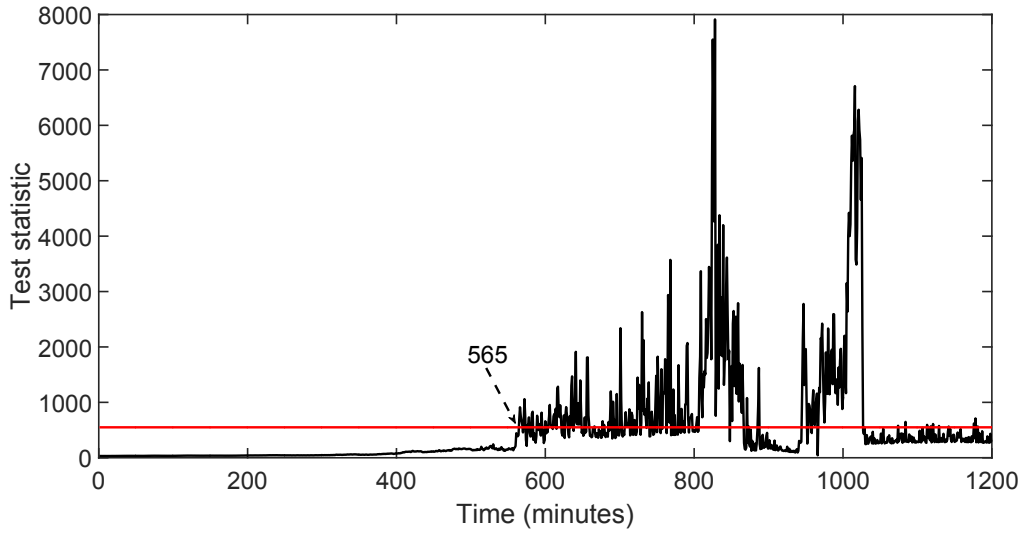


Figure 4.5: Control chart of the PPCA model

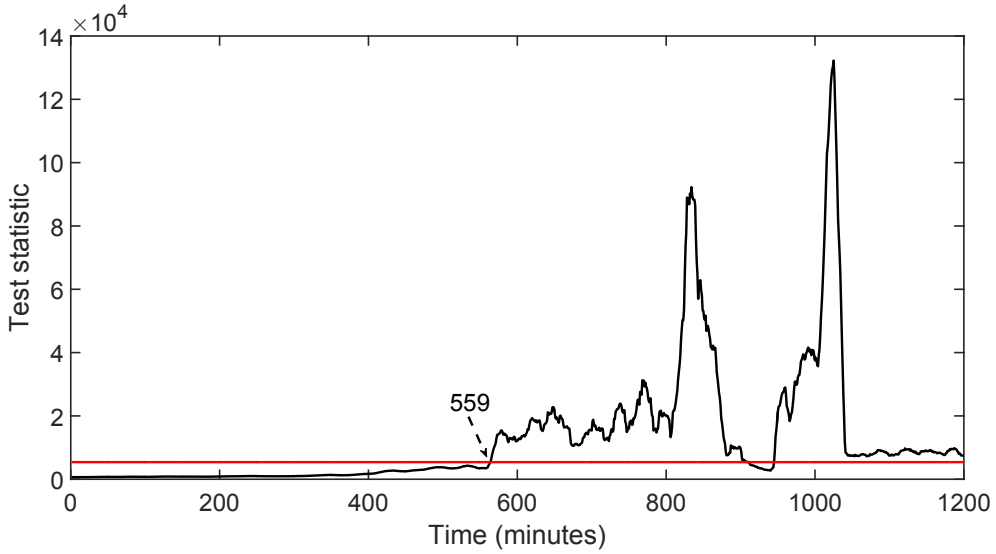


Figure 4.6: Control chart of the dynamic PPCA model

Model selection for the proposed model

For the proposed model, we need to select the number of local models. The first layer model was chosen to have four local models. The posterior probabilities of the local models given the observations are shown in Fig. 4.7. The plots show the posterior probabilities of the categorical latent variable s_n given the observations. Plots corresponding to each of the four local models are shown. If the posterior of a local model equals one for an observation, then the local model completely owns that particular observation. From the plots, it can be seen that all the four local models have either value of one or zero on most of the observations. This gives us an indication that fixing the number of local models to be four in this particular dataset almost results in four distinct clusters. Only very few points were shared between the local models two and three. However, increasing the local models from four to five resulted in more data points being shared between the clusters. Therefore, we decided to stick with four local models in the first layer. We started with an initial guess of 10 for the dimension of the latent variables of the local models. At the end, we obtained two local models with dimensions 7, one with 8 and the other with 6 through the Bayesian regularization.

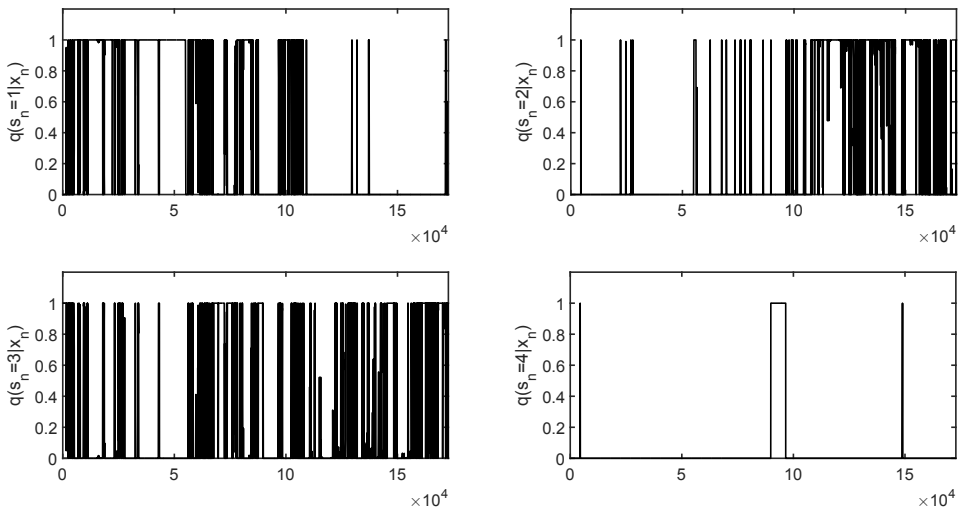


Figure 4.7: The posterior distribution of the local models given the observation. X - axis: training observations

In the second layer, as we increased the number of local models, we also computed the log likelihood of the parameters of the resulting collapsed model in the validation set. The log likelihood in the validation set is shown in Fig. 4.8. The log likelihood value kept increasing as we increased the number of local models. However, the value started to saturate as we increased the number of local models beyond three. Therefore, we chose second layer models with three local models. Dimension reduction did not occur in the second layer mixture Bayesian PPCA models. It remained the same as the initial values, which were chosen to be one less than the dimension of the latent variables inferred from the first layer model.

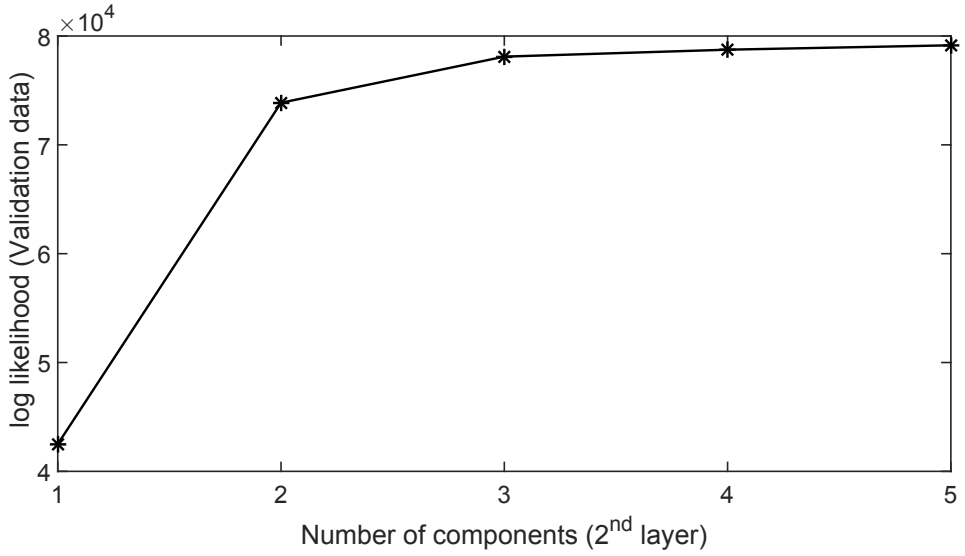


Figure 4.8: Log likelihood of the parameters in the validation set when the number of components in the second layer was increased

The above-discussion for model selection is for a particular initial guess. However, we fitted the model with 15 different initial guesses obtained from the k-means clustering algorithm. In most of the trials, these results were consistent. Therefore, a model structure with four local models in the first layer and three local models in the second layer was chosen. This led a mixture model with twelve local models for characterising the NOC.

4.6.3 Comparison

The proposed model was compared against the mixture PPCA and mixture dynamic PPCA models both with twelve local models. As the convergence of the models depends on the initial guesses, fifteen different initial guesses based on the k-means algorithm were provided for all the three models. The resulting models were tested for the percentage of false positives and the time taken for fault detection.

Table 4.5: Comparison of fault detection results

	Mixture PPCA	Mixture dynamic PPCA	Proposed Model
False positives (%)	3.15 ± 2.68	4.24 ± 3.42	**1.74 ± 1.17**
Time of detection (minutes)	5.4 ± 3.55	**93.1 ± 7.80**	91.4 ± 4.54

The percentage of false positives and the time taken for fault detection are shown in Table. 4.5. From Table. 4.5, it can be seen that the percentage of false positives given by the proposed model had less variability with respect to various initial guesses and also the percentage stayed lower compared to that of the mixture PPCA and the mixture dynamic PPCA models. One would expect both the proposed model and the mixture dynamic PPCA model to have similar performances. However, from the results, it is clearly not the case. The reason for more and the variability in the percentage of false positive was the difficulty in generating a good initial guesses and the convergence issue. Every time when the model parameters were estimated, they converged to different values and in turn, produced results that were very different from each other. However, as the proposed model fitted a maximum of four local models at a given stage, the variability in the converged parameters was low and in turn, produced more consistent results. In addition, when we checked the number of parameters required for both models, the proposed model was parsimonious. Reduced model complexity also adds to the lesser percentage of false positives. When it comes to the percentage of false positives, one would expect more complex model to have more false positives as the generalizing ability gets poorer with the complexity of the model. In this perspective, the false positives obtained by the mixture PPCA model should be lower compared to the proposed model. However, it remained higher except for few initial guesses where the percentage of false positives was close to 1.31%, whereas, the lowest that could be achieved by the proposed model was 1.44% and by that of the mixture dynamic PPCA model was 1.63% in the entire exercise.

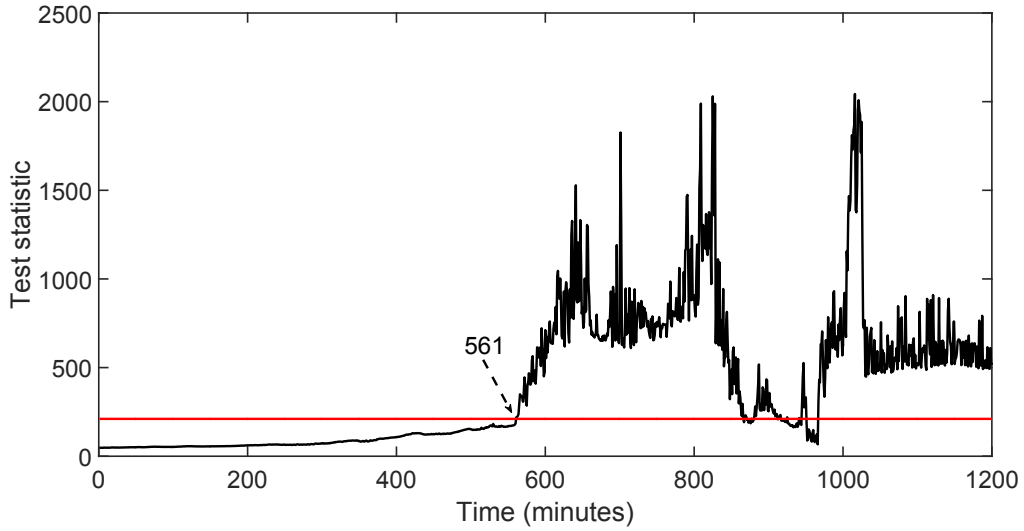


Figure 4.9: Typical control chart obtained using the mixture PPCA model

Table. 4.5 shows how early the fault could be detected by the mixture models compared to the base case PPCA model. As opposed to the false positives, the variability in the time of fault detection was low in all the three cases. Models fitted with different initial guesses reacted almost in a similar way to this particular fault. The time of fault detection in the case of mixture PPCA did not improve much compared to the base case model and the dynamic PPCA model. However, the mixture dynamic PPCA model and the proposed model were more sensitive to the fault and detected the fault almost 90 minutes earlier compared to the base case. In the whole exercise, the mixture dynamic PPCA model detected the fault 102 minutes ahead of the base case once, which was the best performance that could be obtained.

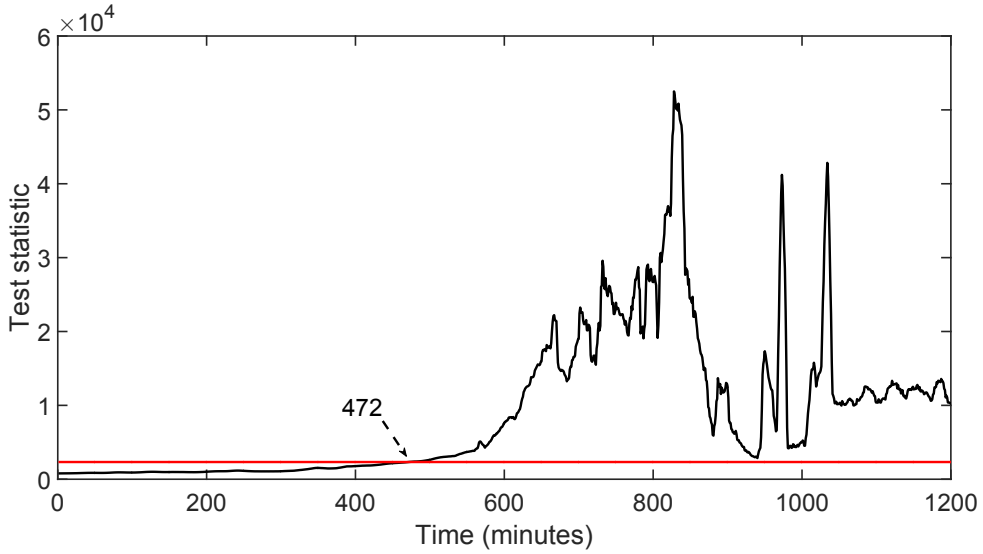


Figure 4.10: Typical control chart obtained using the mixture dynamic PPCA model

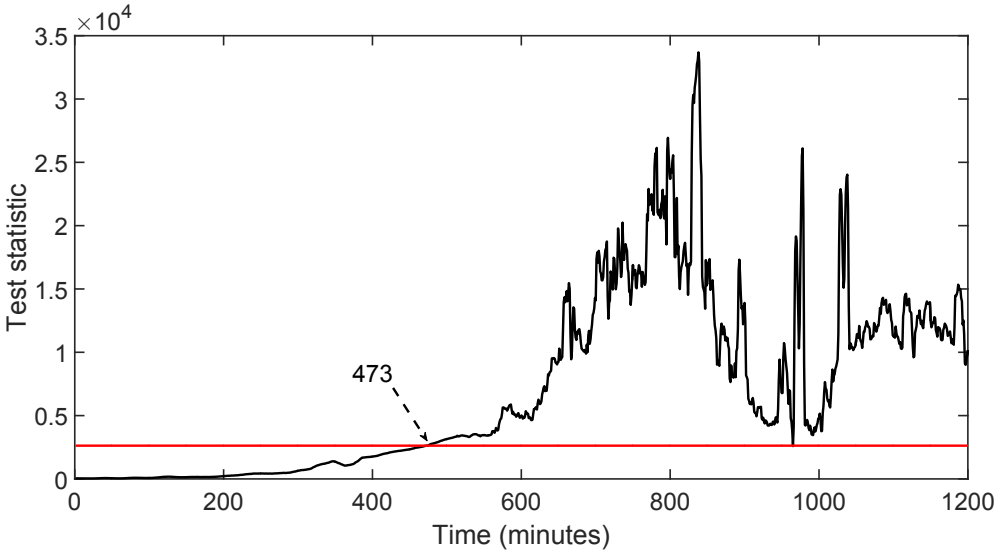


Figure 4.11: Typical control chart obtained using the two-layer mixture Bayesian PPCA model

The typical control charts obtained from all the three models are shown in figures 4.9, 4.10 and 4.11. It can be seen that the test statistics cross their respective thresholds much earlier in the case of dynamic models compared to the static mixture model. The other common behaviour of the test statistics of the dynamic models as opposed to the static models was that, once the test statistic crossed the threshold it remained above the threshold for the majority of the observations. It was the case

with the dynamic PPCA model, the mixture dynamic PPCA model and the proposed model, whereas, for the static models, the test statistic dropped below the threshold for some of the observations after the fault was detected. This corresponds to higher fault detection percentages given by the dynamic models. However, quantifying the detection rate was not possible as the exact time of fault occurrence was not known to us.

4.7 Case study 2: Three-phase flow system

Our second case study is a three-phase flow facility that is kept at Cranfield University. Ruiz-Cárcel et al. [91] had simulated datasets from the experimental setup to be used as a benchmark case study for fault detection algorithms and made it publicly available. A detailed description of the process can be found in [91]. Here, a brief description of the process and the datasets used for the fault detection study is provided.

4.7.1 Process description

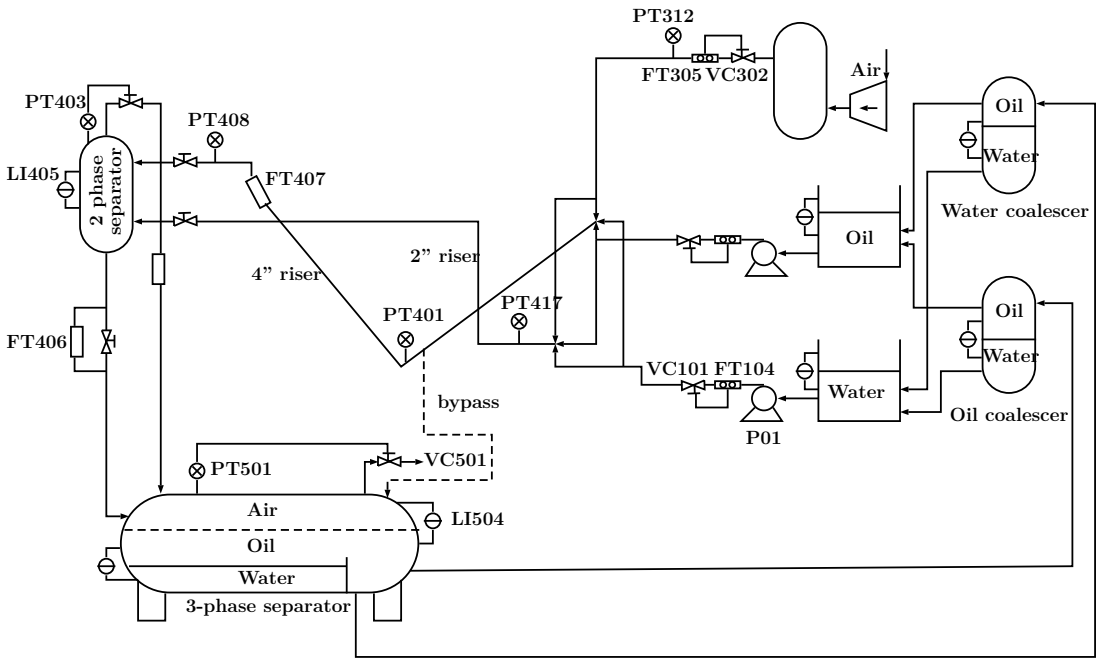


Figure 4.12: Schematic of the three-phase flow system.

A simplified schematic of the facility is shown in Fig. 4.12. The schematic shown in Fig. 4.12 is reproduced only with the minimal information that is required in the context of the description provided below. For a more detailed flow diagram, readers are referred to [91]. The set up in Fig. 4.12 is built to provide controlled and measured flow of air, water and oil to a pressurized system. Air is fed using two compressors to the system. Oil and water are fed by means of multistage pumps from their respective storage tanks. All these three components are fed to a 2-phase separator which is kept at a height of 10.5m from the ground. These three components get mixed and reach the separator through either of the two risers (2" and 4") shown in Fig. 4.12. From the 2-phase separator, partially separated mixture is sent to a 3-phase separator were almost 100 percentage separation is achieved. Air from the three-phase separator is released to the atmosphere. Water and oil emulsions are sent to coalescers from where almost pure water and oil are sent back to their respective storage tanks. The facility is equipped with sensors that allow various flow, pressure, level and density variables to be measured. The list of variables used for this study is provided in Table. 4.6. Ruiz-Cárcel et al. [91] used 23 or 24 (depending on the fault cases) variables for their fault detection study. However, we found that some of the temperature and level variables were in different ranges in the training and test data. Therefore, we removed those variables from the analysis. Also, the observations of tags indicated by * in Table. 4.6 were differenced from their previous observations and used for fault detection. They indicate changes in pressure and level measurements. This is again for the same reason that the observations were found to be in different ranges, however, the observations after differencing fell in a similar range both in the training and test data.

Table 4.6: Tags used for process monitoring

Tag number	Location	Tag
1	PT312	Air delivery pressure
2	PT401	Pressure at the bottom of the riser
3	PT408	Pressure at the top of the riser
4	PT403*	Pressure in top separator
5	PT501*	Pressure in 3 phase separator
6	PT408*	Pressure difference (PT401 - PT408)
7	PT403	Differential pressure over VC404
8	FT305	Flow rate of input air
9	FT104	Flow rate of input water
10	FT407	Flow rate at the 4" riser
11	LI405*	Level at the top separator
12	FT406	Flow rate of top separator output
13	FT407	Density of the fluid at the 4" riser
14	LI406	Density of top separator output
15	FT104	Density of input water
16	VC302	Position of valve VC302
17	VC101	Position of valve VC101

Data description

In total, seven data files were made available by [91]. Out of the seven, one file was of three datasets corresponding to the NOCs and the rest six were of files generated from the faulty operating conditions. Data that correspond to the NOCs were generated for twenty different combinations of air and water flow rates. Observations were recorded at one second frequency. Three datasets from the normal operating conditions had the data recorded for 10372s, 9825s and 13200s respectively. The set point values of air and water flow rates that were used to generate the data from the NOCs are listed in Table. 4.7.

Table 4.7: The set point values of air water flow rates used for generating the datasets from the NOCs

Air flow rate (m^3/s)	0.0208	0.0278	0.0347	0.0417	
Water flow rate (kg/s)	0.5	1	2	3.5	6

Out of the six fault conditions, the first four fault conditions were considered for the illustration here. The considered fault types are as follow,

1. Fault: Air line blockage. A blockage was introduced on the air line using a manual valve gradually just before the point where all the three components are mixed.

2. Fault: Water line blockage. A blockage was introduced on the water line using a manual valve gradually just before the point where all the three components are mixed.

3. Fault: 2-phase separator input blockage. A blockage was introduced on the input line to the two-phase separator.

4. Fault: Open direct bypass. The bypass line valve was opened gradually such that the 3-phase mixture bypasses the riser and the 2-phase separator and reaches the three-phase separator directly.

Three datasets were generated for each of the four fault cases, one with the changing operating conditions and the other two with the steady state operating conditions. Table. 4.8 shows the number of data points that were available, the start and end times of the faults introduced and the operating conditions under which the data were collected.

Table 4.8: Description of the datasets from the considered fault cases

Data set	Duration (s)	Fault start (s)	Fault end (s)	Operating conditions
1.1	5811	1566	5181	changing
1.2	4467	657	3777	steady state
1.3	4321	691	3691	steady state
2.1	9192	2244	6616	changing
2.2	3496	476	2656	steady state
2.3	3421	331	2467	steady state
3.1	9090	1136	8352	changing
3.2	6272	333	5871	steady state
3.3	10764	596	9566	steady state
4.1	7208	953	6294	changing
4.2	4451	851	3851	steady state
4.3	3661	241	3241	steady state

All the three normal operating datasets were concatenated. Two thirds of it were used for training and one third of it was used for validation. The trained model were

tested on all the twelve datasets shown in Table. 4.8.

4.7.2 Results and discussion

Base case

Similar to the previous case study, we used the PPCA and Dynamic PPCA models as the base cases for comparison. For the dynamic PPCA model, we set the lag to be 14. Except for the tags indicated by * Table 4.6, for all the other tags, 14 previous observations were augmented with the observations at all the time instants. The estimation of the static PPCA and dynamic PPCA models were initialized with latent variable dimensions 10 and 30 respectively and it converged to latent variable dimensions 7 and 18 for the respective cases. The overall performance in all the 12 datasets using the estimated models is shown in Table. 4.9. Between the static and the dynamic model, the dynamic model had a slight gain in detection rate. However, the obtained gain was accompanied by more false positives.

Table 4.9: The overall performance obtained from the base case models

	PPCA	DPPCA (lag: 14)
Detection rate (%)	56.75	57.65
False positives (%)	2.84	3.28

Model selection for the proposed model

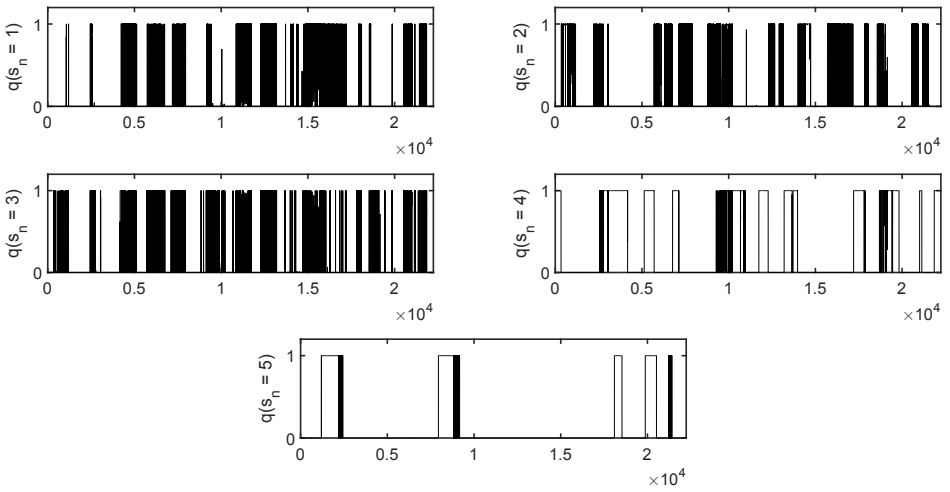


Figure 4.13: Posterior distribution of the local models given the observation. X - axis: training observations

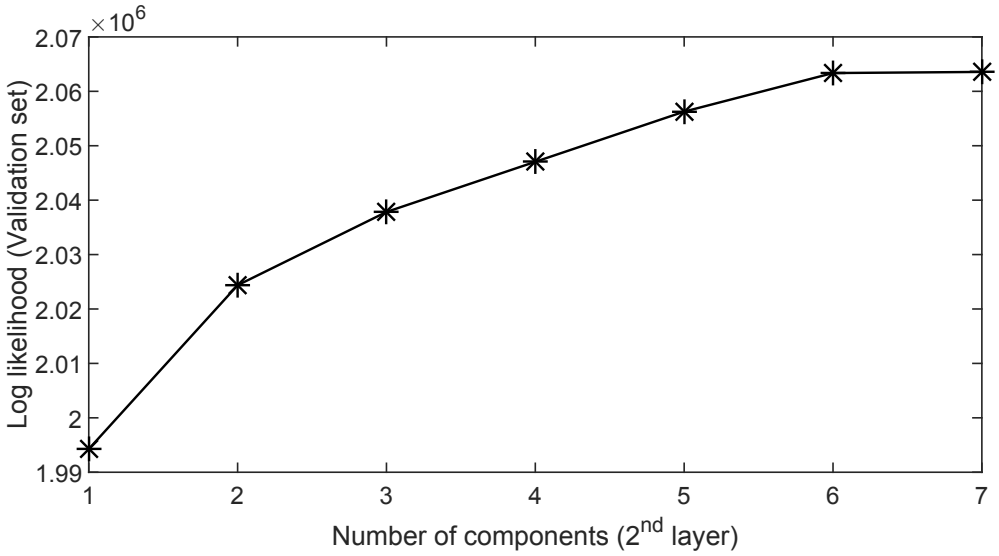


Figure 4.14: Log likelihood of the model parameters in the validation set when the number of components in the second layer was increased

We fitted a mixture model with five local models in the first layer. Each local was considered to have 30-dimension latent variables. Models converged to dimensions 11, 15, 19, 23, and 25 through the Bayesian regularization respectively. The posterior

probabilities of the local models given the observations are shown in Fig. 4.13. Again, for this case, it can be seen that the observations were clustered into almost five perfect clusters except for few observations that were shared by the local models one and three. We fixed the number of first layer local models to be five and started to fit mixture models to the latent variables obtained from the first layer. We kept increasing the number of second layer models from one. The resulting log likelihood values of the parameters in the validation set is shown in Fig. 4.14. It can be seen from the figure that after six local models, the log likelihood value starts to saturate. Taking this as an indication, the number of second layer local models was fixed to be six. The resulting collapsed model was of 30 local models.

Comparison

The proposed model was compared with the mixture PPCA and the mixture dynamic PPCA models consisting of thirty local models each. All the three models were initialized with 15 different initial guesses. The overall performances of all the three models are shown in Table. 4.10. It summarizes the percentage of false positives and the detection rates obtained by the models for all the 15 different initial guesses. It can be seen that the proposed model and the mixture dynamic PPCA model outperform the other models in terms of the fault detection rate. The proposed model and the mixture dynamic model detected more faulty observations accurately compared to the other models. Between the proposed model and the mixture dynamic PPCA model, in some cases, the mixture dynamic PPCA model outperforms the proposed model. However, when we look at the false positives resulted from the proposed model and the mixture dynamic PPCA model, it can be seen that the proposed model clearly had lesser number of false positives.

When compared with the dynamic PPCA model, the proposed model had a higher detection rate only at the cost of slight increase in number of false positives. However, at the same time, the mixture dynamic PPCA model had more number of false positives. Similar to the previous case study, the variability associated with the results obtained using the mixture PPCA and the mixture dynamic PPCA models were high when compared to the proposed model. Again, this indicates that the chances of

converging to different models with different initial guesses are very high when we try to fit a mixture model with a large number of local models.

The results for time of fault detection obtained for each individual fault case are presented in Table. 4.11. The key highlight of the results was, out of 12 cases, there were 9 cases for which the proposed model or the mixture dynamic PPCA model detected the faults earlier when compared to the other models.

Table 4.10: Comparison of the overall performances

	Mixture PPCA	Mixture dynamic PPCA	Proposed Model	PPCA	dynamic PPCA
False positives (%)	3.49 ± 1.43	5.54 ± 1.84	**3.38 ± 0.069**	2.84	3.28
Detection rate (%)	57.36 ± 3.86	**63.99 ± 4.25**	62.69 ± 0.10	56.75	57.65

Table 4.11: Comparison of the fault detection time by different models on different fault cases

Fault case	Mixture PPCA	Mixture dynamic PPCA	Proposed model	PPCA	Dynamic PPCA
1.1	1830.5±1.03	1573.3±187.78	**1487.3±0.87**	1832	1832
1.2	1803.6±1.01	1512.3±2.06	**1509±0.73**	1805	1805
1.3	1940.9±139.6	**1791.4±115.99**	1815.8±87.71	2106	2106
2.1	3.8452±106	**1707.3±316**	3641±12.21	3989	3770
2.2	1840±9.50	**1818.8±4.45**	**1818±2.43**	1850	1831
2.3	1541.7±14.56	693.5±7.34	**692**	1539	693
3.1	88.45±29.96	82±1.47	82	95	**62**
3.2	773±327	**663 ±1.07**	**663±2.67**	1090	665
3.3	57.50±30.97	42	59.61±16.66	67	**41**
4.1	457.38±9.81	471.61±15.99	472	483	**457**
4.2	428.28±5.85	**320.50±11.10**	343.46±4.52	436	428
4.3	314.5±1.04	**314**	**314**	315	315

4.8 Summary

The purpose of this chapter was to illustrate how the probabilistic latent variable models can be extended to model the multi-modal processes. In this chapter, the two-layered mixture Bayesian PPCA model for process monitoring was developed and evaluated. The model was developed mainly for fault detection applications where the process data with non-Gaussian distribution and temporally correlated observations are encountered. For the process data with the above-mentioned characteristics, the mixture dynamic PPCA model could have been a preferred choice. However, considering the shortcomings of the mixture dynamic PPCA model, the new model was

proposed. When the proposed model was applied to two different case studies, we found that the model manages to achieve the performance achieved by the mixture dynamic PPCA model in terms of the fault detection rates and the time of fault detection. The proposed model along with the mixture dynamic PPCA model was found to clearly outperform the PPCA, dynamic PPCA and mixture PPCA models in terms of the fault detection rates and the time of fault detection. However, it was also found that the proposed model had lower false positive percentages compared to the mixture dynamic PPCA model. The proposed model was also found to give more consistent fault detection performances with respect to different initial guesses for the parameters during the model estimation stage.

Chapter 5

An Approach for Causality Analysis and Contemporaneous Correlation Features Inference from Industrial Process Data

5.1 Introduction

In this chapter, we address the problem of causal network reconstruction from industrial process data in the presence of contemporaneous correlations among the variables in the data. Depending on the sampling rate and the presence of feedback loops, the measured process variables may have contemporaneous dependencies in addition to the casual interactions [92]. We propose a hybrid model to simultaneously mine causal connections and extract features responsible for contemporaneous correlations among the process variables from a finite window of observed data. The model consists of two components: A vector auto-regressive exogenous (VARX) model component as a predictor and a factor analysis (FA) model component used for modelling the prediction error. The causal connections are inferred through the VARX component and the contemporaneous correlation features are inferred from the FA component. The parameters of the resulting hybrid model are regularized using the hierarchical prior distributions for penalizing the insignificant parameters. It is then estimated under the variational Bayesian expectation maximization (VBEM) framework. The estimation is initiated with a complex model which is then systematically reduced to a simpler model that retains only the parameters corresponding to significant causal connec-

tions and contemporaneous correlations. Model reduction is carried out through a series of deterministic switches from complex models to simpler models using a relevance criterion. The approach is illustrated through a number of simulated examples and an industrial case study.

The vector auto-regressive (VAR) models are widely used to infer the causal connections in multivariate dynamic processes [47, 48, 49, 50]. The use of VAR models to infer causal connections without accounting for the contemporaneous correlations may lead to spurious findings as zero lag correlations tend to disguise themselves as the time-lagged correlations as shown in [55]. The process variables are said to be contemporaneously correlated when the prediction errors of those variables from the past observations tend to be significantly correlated. This in other words refers to the presence of non-zero off-diagonal elements in the prediction error covariance. There exist techniques to handle contemporaneous dependencies with the use of structural VAR modelling [92] or by explicitly modelling the contemporaneous dependencies [55] when inferring the causal connections. Our approach differs from the existing approaches by the use of FA model to represent the prediction error covariance. The use of FA model allows us to extract the latent variables responsible for the contemporaneous correlations and infer how they influence the observed variables. More importantly, in the FA model, the prediction error can be expressed as a linear function of the latent variables. This in turn lets us penalize the causal connections (defined by the linear VARX model) and the contemporaneous correlations with equal weights through the Bayesian regularization, without favouring one over the other.

Normally, the significance of the parameters after the convergence of the VBEM algorithm is determined through the posterior variance of the parameters, the procedure is known as the automatic relevance determination (ARD). For instance, the approach [93] that comes closer to our work in terms of the use of VBEM for causal inference, utilizes the ARD to determine the significant connections. However, the use of ARD requires a subjective threshold on the posterior variance. Instead we propose an automatic model reduction strategy that still makes use of the posterior variance, however, it does not involve subjective thresholds. At first, a reasonably complex model is estimated under the VBEM framework. It provides the posterior parameter

estimates and a surrogate estimate for the model evidence. In order to infer the causal connections and the contemporaneous correlation features, the complex model form is systematically reduced to a structure that retains only the relevant parameters in the model. Each switch to a simpler model from a relatively complex model is accepted only if the surrogate estimate of the model evidence is improved. In addition to the use of a new model reduction strategy and to the best of our knowledge, there has been no work done on combining the FA and VARX models to mine causal networks and features responsible for contemporaneous correlations simultaneously. The effectiveness of the VBEM framework and the inclusion of the FA component to the model is verified by comparing the results against the VAR based approaches under the maximum likelihood (ML) and the VBEM estimation frameworks.

The potential of the presented approach is not limited to industrial process data analysis. For instance, it could be utilized in connectivity analysis in brain networks where the contemporaneous dependency is termed as the functional connectivity and the causal dependency is termed as the effective connectivity [55] and other areas such as climatology [94], econometrics [46], human computer interaction, etc.

The remainder of this chapter is organized as follows: In section 5.2, the proposed model and the chosen Bayesian prior for the model parameters are discussed. In section 5.3, the Bayesian network that results after incorporating the prior distributions is discussed. In section 5.4, the VBEM framework for estimating the proposed model is discussed. In section 5.5, implementation details and the model reduction strategy are discussed. In section 5.6, the simulated and the industrial case studies are presented. In section 5.7, concluding remarks are presented.

5.2 Theory

5.2.1 Proposed Model

Assumptions

We assume that the studied multivariate process can be represented using a finite discrete linear VARX model as,

$$y(t) = \sum_{l=1}^{L} W(l) y(t-l) + W' u'(t) + \epsilon(t) \tag{5.1}$$

where $y(t) \in \mathbb{R}^D$ is an observation of the multivariate process at time, t, $u'(t) \in \mathbb{R}^P$ is the known exogenous input that affects the process at t and $W(l) \in \mathbb{R}^{D \times D} \forall l \in [1, L]$ and $W' \in \mathbb{R}^{D \times P}$ are the parameters of the model. Noise, $\epsilon(t)$ is independent and identically distributed and follows a multivariate Gaussian distribution with zero mean and Σ_ϵ covariance.

We represent the noise component of the model by the FA model as the following,

$$\epsilon(t) = V x(t) + \eta(t) \tag{5.2}$$

where

$$x(t) \overset{i.i.d}{\sim} \mathcal{N}(0, I_K), \ \eta(t) \overset{i.i.d}{\sim} \mathcal{N}\left(0, diag(\sigma)^{-1}\right) \tag{5.3}$$

where $x(t) \in \mathbb{R}^{K(<D)}$ is a vector of lower dimension latent variables that are multivariate Gaussian distributed with zero mean and identity covariance, $V \in \mathbb{R}^{D \times K}$ is the loading matrix of the FA model, $\eta(t) \in \mathbb{R}^D$ is the new noise term which follows a multivariate Gaussian distribution with zero mean and diagonal covariance, σ is a vector of precision (inverse of variance) parameters and the operator $diag(.)$ converts a vector into a diagonal matrix.

The addition of the FA model component does not change the form of the model shown in (5.1). It lets the noise to follow a conditional Gaussian distribution of the following form,

$$\epsilon(t)|x(t) \overset{i.i.d}{\sim} \mathcal{N}\left(Vx(t), diag(\sigma)^{-1}\right) \tag{5.4}$$

If the latent variables are marginalized from the joint distribution of $\epsilon(t)$ and $x(t)$, then the marginal distribution of $\epsilon(t)$ can be obtained as the following,

$$\epsilon(t) \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_\epsilon), \ \Sigma_\epsilon = VV^T + diag(\sigma)^{-1} \tag{5.5}$$

Therefore, it is just that the covariance of the noise is parametrized differently by the use of the FA model component.

Finally, the proposed model can be seen to take the following form,

$$y(t) = \sum_{l=1}^{L} W(l) y(t-l) + W' u'(t) + V x(t) + \eta(t) \tag{5.6}$$

The FA component in the model captures the correlations among the prediction errors in $Vx(t)$ term and the remaining variance in $\eta(t)$ term.

Causal connections

In the above model, variable y^j is does not directly Granger cause variable y^i if the entries in i^{th} row and j^{th} column of $W(l)$ $\forall l \in [1, L]$ are effectively zero as in that case, the past observations of y^j do not help in predicting the current values of y^i [51].

Contemporaneous Correlation

Variables y^i and y^j of the process are contemporaneously uncorrelated if,

$$E\left(\epsilon^i(t)\epsilon^j(t)\right) = 0 \tag{5.7}$$

where $E(.)$ refers to the expectation operation. These expectations give the off-diagonal elements of the prediction error covariance. It can be seen that the variables y^i and y^j of the process are contemporaneously uncorrelated if they do not have loading parameters that multiply a common latent variable in the above model. This is true because the covariance between ϵ^i and ϵ^j is quantified by the dot product between the i^{th} and the j^{th} rows of the loading matrix V. Thus, from the structure of the loading matrix (zeros and non-zero entries), one can infer the features responsible for contemporaneous correlations and how they influence the observed variables.

Therefore, the problem of mining causal connections and contemporaneous correlation features in the process boils down to inferring the effective non-zero parameters in the model.

5.2.2 Bayesian Regularization

In this work, we choose a reasonably complex model and incorporate the prior distributions for the parameters such that they penalize the model parameters of the VARX and the FA components with equal weights to let the model converge to a simpler form where only the set of significant parameters remain and the rest converge close to zero.

We assume that the model parameters follow Gaussian distribution and the precision parameters (inverse of variance) of the Gaussian distribution follow gamma distribution. For instance, let us say that θ is one of the model parameters, the prior distribution of θ is considered to be the following,

$$\theta \sim \mathcal{N}\left(0, \nu^{-1}\right), \quad \nu \sim Ga\left(\alpha^*, \beta^*\right) \tag{5.8}$$

where θ follows a Gaussian distribution with zero mean and ν^{-1} variance and ν follows a gamma distribution with shape and rate parameters α^* and β^* respectively.

5.3 Bayesian Network of the Proposed Model

In this section, we describe the prior distributions incorporated for the model parameters and the resulting Bayesian network of the model. For convenience, we rewrite the model shown in (5.6) in the following regression form,

$$Y = WU + VX + \eta' \tag{5.9}$$

where $Y = [y\left(N+L+1\right), ..., y(t), .., y(L+1)] \in \mathbb{R}^{D\times N}$,
$X = [x\left(N+L+1\right), ..., x(t), .., x(L+1)] \in \mathbb{R}^{K\times N}$,
$\eta' = [\eta\left(N+L+1\right), ..., \eta(t), .., \eta(L+1)] \in \mathbb{R}^{D\times N}$,
$W = [\mu, W(L), ..., W(1), W'] \in \mathbb{R}^{D\times Q}$ and $U \in \mathbb{R}^{Q\times N}$ is a matrix that contains all the known predictors for the output that includes the time-lagged outputs and exogenous inputs as shown below,

$$U = \begin{bmatrix} y\left(N+L\right) & . & . & . & y(t-1) & . & . & . & y(L) \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ y(N+1) & . & . & . & y(t-L) & . & . & . & y(1) \\ u'\left(N+L+1\right) & . & . & . & u'(t) & . & . & . & u'(L+1) \end{bmatrix} \tag{5.10}$$

Wherever necessary, we use the notations that combine all the parameters together and all the predictors together respectively as

$$F = [W, V] \in \mathbb{R}^{D \times M},\ Z = [U, X]^T \in \mathbb{R}^{M \times N} \tag{5.11}$$

From here onwards, the entries of the matrices Y, Z, X, U, F, W, and V will be represented by y, z, x, u, f, w, and v, respectively and the rows will be indexed in the superscript and the columns will be indexed in the subscript.

The joint distribution of the N output observations in terms of the parameters and the predictors can be represented as

$$p\left(Y|Z, F, \sigma\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|Fz_n, diag\left(\sigma\right)^{-1}\right) \tag{5.12}$$

where y_n and z_n are the n^{th} columns of Y and Z, respectively. As previously described, the latent variables follow a multivariate Gaussian distribution with zero mean and identity covariance. Consequently, the joint distribution of the latent variables is given by,

$$p\left(X\right) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|0, I\right) \tag{5.13}$$

where x_n is the n^{th} column of X.

Now, we arrive at the important part of the model where we define the prior distribution for the model parameters. Each regression parameter in the model is considered to follow a Gaussian distribution with zero mean and a certain precision. The resulting prior distribution of the parameters is given as follows,

$$p\left(F|\nu\right) = \prod_{m=1}^{M} \prod_{d=1}^{D} \mathcal{N}\left(f_m^d|0, \left(\nu_m^d\right)^{-1}\right) \tag{5.14}$$

where f_m^d is the entry in the d^{th} row and m^{th} column of F and its precision parameter, ν_m^d follows a gamma distribution with shape and rate parameters α^* and β^* respectively as the following,

$$p\left(\nu|\alpha^*, \beta^*\right) = \prod_{m=1}^{M} \prod_{d=1}^{D} Ga\left(\nu_m^d|\alpha^*, \beta^*\right) \tag{5.15}$$

where ν is the collection of precision parameters. As we impose prior on each parameter independently, at the end of the estimation of a fairly complex model, we would

be able to infer individual effective non zero parameters from where the model order and the causal connections can be inferred.

For completeness, we also consider the precision parameters of the noise to follow a gamma distribution with shape and rate parameters κ^* and ϕ^*, respectively, as the following,

$$p\left(\sigma|\kappa^*,\phi^*\right) = \prod_{d=1}^{D} Ga\left(\sigma^d|\kappa^*,\phi^*\right) \tag{5.16}$$

where σ^d is the precision of the noise of output in dimension d. With all the prior probabilities and the likelihood of the parameters defined, the resulting Bayesian network of the model is shown in Fig. 5.1. All the encircled nodes in the network are random variables and the other nodes are deterministic. Parameters that have * as superscript need to be either estimated or defined by the user. The joint likelihood of the network can be obtained by multiplying the conditional distributions shown in equations (5.12) - (5.16).

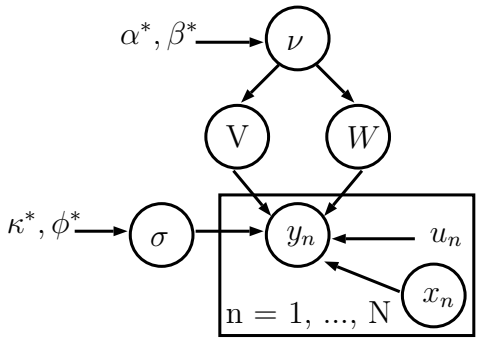


Figure 5.1: Bayesian network of the proposed model

5.4 Estimation

In this section, we discuss the VBEM framework for estimation. We describe how the posterior of the unknowns is factorized and provide the expressions for the variational lower bound and the posterior distributions.

5.4.1 Variational Posterior distribution

The posterior of the parameters and the latent variables in the model is assumed to be factorisable as the following,

$$p\left(F,\nu,X,\sigma|Y,U,\alpha^{*},\beta^{*},\kappa^{*},\phi^{*}\right)\approx q\left(F\right)q\left(\nu\right)q\left(X\right)q\left(\sigma\right) \tag{5.17}$$

Further, we consider the posterior of the precision parameters to be factorisable as the following,

$$q\left(\nu\right)=\prod_{m=1}^{M}\prod_{d=1}^{D}q\left(\nu_{m}^{d}\right) \tag{5.18}$$

The reason for the above factorization is that the posteriors of the precision parameters allow us to infer the relevance of each parameter in the model separately as it will be shown in section 5.5 of this chapter. We also assume that the noise precision along each dimension to be mutually independent. The resulting posterior of the noise preciosion parameters is given by,

$$q\left(\sigma\right)=\prod_{d=1}^{D}q\left(\sigma^{d}\right) \tag{5.19}$$

As a result of the aforementioned assumptions, some more factorizations will be induced in the network. The regression parameters will become independent along each output dimension as the following,

$$q\left(F\right)=\prod_{d=1}^{D}q\left(f^{d}\right) \tag{5.20}$$

where f^{d} is the d^{th} row of F. In addition, the posterior distribution of the latent variables will also become factorizable as the following,

$$q\left(X\right)=\prod_{n=1}^{N}q\left(x_{n}\right) \tag{5.21}$$

As result of the aforementioned factorizations, the posteriors of the precision parameters of the regression parameters and the precision parameters of the noise will follow gamma distributions as the following,

$$q\left(\nu_{m}^{d}\right)=Ga\left(\nu_{m}^{d}|\alpha,\beta_{m}^{d}\right),\ q\left(\sigma^{d}\right)=Ga\left(\sigma^{d}|\kappa,\phi^{d}\right) \tag{5.22}$$

where α and κ are the shape parameters and β_m^d and ϕ^d are the rate parameters of the posterior gamma distributions of ν_m^d and σ^d respectively. In addition, the posteriors of the regression parameters and the latent variables will follow multivariate Gaussian distributions as the following,

$$q\left(f^d\right) = \mathcal{N}\left(f^d|\hat{f}^d, \Sigma_{f^d}\right),\ q\left(x_n\right) = \mathcal{N}\left(x_n|\hat{x}_n, \Sigma_x\right) \tag{5.23}$$

where $\hat{f}^d$ and $\hat{x}_n$ are the mean vectors and Σ_{f^d} and Σ_x are the covariance matrices respectively.

5.4.2 Model Evidence and the Posterior Update Rules

Variational lower bound

The variational lower bound or the surrogate estimate for the model evidence is given by,

$$\ln p\left(Y|m\right) \geq \mathcal{L} = \int_F \int_\nu \int_X \int_\sigma q\left(F\right) q\left(\nu\right) q\left(X\right) q\left(\sigma\right) \times$$
$$\ln \frac{p\left(Y, F, \nu, X, \sigma|U, \kappa^*, \phi^*, \alpha^*, \beta^*\right)}{q\left(F\right) q\left(\nu\right) q\left(X\right) q\left(\sigma\right)} dF d\nu dX d\sigma \tag{5.24}$$

Further, this can be split into a summation of multiple terms as the following*,

$$\mathcal{L} = \sum_{m=1}^{M}\sum_{d=1}^{D} \int_{\nu_m^d} q\left(\nu_m^d\right) \ln \frac{p\left(\nu_m^d\right)}{q\left(\nu_m^d\right)} d\nu_m^d$$

$$+ \sum_{d=1}^{D} \int_{\nu_m^d}\int_{f^d} \prod_{m=1}^{M} q\left(\nu_m^d\right) q\left(f^d\right) \ln \frac{p\left(f^d\right)}{q\left(f^d\right)} d\nu_m^d df^d$$

$$+ \sum_{d=1}^{D} \int_{\sigma^d} q\left(\sigma^d\right) \ln \frac{p\left(\sigma^d\right)}{q\left(\sigma^d\right)} d\sigma^d + \sum_{n=1}^{N} \int_{x_n} q\left(x_n\right) \ln \frac{p\left(x_n\right)}{q\left(x_n\right)} dx_n$$

$$+ \sum_{n=1}^{N}\sum_{d=1}^{D} \int_{x_n}\int_{f^d}\int_{\sigma^d} q\left(x_n\right) q\left(f^d\right) q\left(\sigma^d\right) \ln p\left(y_n^d|f^d, z_n, \sigma^d\right) dx_n df^d d\sigma^d \tag{5.25}$$

where y_n^d is the entry in the d^{th} row and n^{th} column of Y. The explicit expression for the lower bound derived using the above equation is shown in Table. D.1 of Appendix D.

*Note that the prior and the posterior distributions should be conditioned on their respective parameters for an accurate representation; however, for the sake of representational simplicity we have not included them.

The posterior updates

The posterior update expressions for all the parameters are shown in Table. D.2 of appendix D. These update expressions can be obtained by taking the derivatives of the lower bound with respect to the posterior distributions and equating them to zero. For instance, to obtain the update expression for the posteriors of the precision parameters, the derivatives of the lower bound expression with respect to the posteriors of the precision parameters are equated to zero. It then leads to the update expressions of the following form,

$$\ln q\left(\nu_m^d\right) \propto \ln p\left(\nu_m^d\right) + \int_{f^d} q\left(f^d\right) \ln \frac{p\left(f^d\right)}{q\left(f^d\right)} df^d \quad \forall d, m \tag{5.26}$$

which can further be deduced to obtain the explicit update equations for the parameters α and β_m^d (of the posteriors) as shown in the first row of Table. D.2 of appendix D. Similarly, all the updates can be derived and shown to be the ones presented in Table. D.2 of appendix D.

Updates for the prior parameters

When a reasonable prior knowledge is not available, it is better to optimize the prior parameters with respect to the given data. For the proposed model, the lower bound is a concave function of the prior parameters and when setting its derivatives with respect to the prior parameters to zero, we can obtain the explicit updates that take the lower bound to its maxima with respect to the prior parameters. The update expressions for the prior parameters are shown in Table. D.2 of appendix D. Except for β^*, we optimize all the prior parameters in this fashion and β^* is chosen based on cross-validation. It is chosen such that the log likelihood of the model parameters in the validation set is maximized as illustrated below

$$\beta^*_{selected} = \max_{\beta^*} \sum_{n=1}^{N^{val}} log\left(\mathcal{N}\left(y_n|\hat{W}u_n, \hat{V}\hat{V}^T + diag\left(\frac{\phi}{\kappa}\right)\right)\right) \tag{5.27}$$

where N^{val} is the number of validation samples, $\hat{W}$, $\hat{V}$, ϕ and κ are the parameter estimates shown in Table. D.2 of Appendix D. The best value of β^* is chosen from the grid ranging from 10^{-15} to 1 with an interval of 10^{-1}.

5.5 Implementation Details and Model Reduction

In this section, we discuss how the VBEM framework is implemented to learn models from the data and how the learned model is reduced to a model that retains only the parameters relevant to significant causal connections and contemporaneous correlations.

For the VBEM implementation part, a model with a reasonably large L and K is chosen. Estimation proceeds with an objective of maximizing the lower bound expression through several iterations. Within each iteration, the parameters are updated recursively as the update expressions are dependent on the other. Convergence is assumed when the relative change in the lower bound estimate between the successive iterations becomes negligible. The whole procedure is summarized in steps from 1 to 12 shown in Table. 5.1. As the VBEM framework is prone to local maxima convergence, these steps are repeated several times with different initial guesses. Each time, the converged posterior estimates of the regression parameters are perturbed using Gaussian noise to form a new set of initial guess for the next round of estimation.

Table 5.1: Implementation details

Step	
1	Choose a model with a reasonably large L and K
2	Assign values for MaxIter and δ
3	Initial guess for all the posterior and the prior parameters
4	Set all the parameters active
5	$\mathcal{L}(0) = -\infty$
6	For Iter = 1:MaxIter
7	Update the active parameters recursively using the VBEM updates
8	Compute the lower bound with the active parameters and assign it to $\mathcal{L}(Iter)$
9	If $\lvert\mathcal{L}(Iter) - \mathcal{L}(Iter-1)\rvert / \lvert\mathcal{L}(Iter-1)\rvert \leq \delta$
10	Break For
11	End If
12	End For
13	$\mathcal{L}_{old} = \mathcal{L}(Iter)$
14	For p = 1:$(D \times M)$
15	switch to a simpler model by excluding the posterior parameters corresponding to an active regression parameter which has the lowest estimate of $\left(\hat{\nu}_m^d\right)^{-1}$ among all the active regression parameters
16	Compute the lower bound with the active parameters and assign it to $\mathcal{L}_{new}$
17	If $\mathcal{L}_{new} \geq \mathcal{L}_{old}$
18	Accept the switch
19	End If
20	Execute the steps from 5 to 12
21	$\mathcal{L}_{new} = \mathcal{L}(Iter)$
22	If $\mathcal{L}_{new} \geq \mathcal{L}_{old}$
23	Accept the switch
24	Else
25	Reject the switch and Break For
26	End If
27	$\mathcal{L}_{old} = \mathcal{L}_{new}$
28	End For

Once the estimation is complete, the next step is to determine the insignificant parameters in the model. Here, we follow the automatic relevance determination framework [66, 33, 69] where the estimates of the precision parameters are used to determine the relevance of the regression parameters. The estimate of the inverse of a precision

parameter, ν_m^d from the posterior gamma distribution is given by,

$$\left(\hat{\nu}_m^d\right)^{-1} = \frac{\beta_m^d}{\alpha} = \frac{\beta^* + \frac{1}{2}\left[\left(\hat{f}_m^d\right)^2 + \Sigma_{f_m^d}\right]}{\alpha^* + \frac{1}{2}} \tag{5.28}$$

where $\hat{f}_m^d$ and $\Sigma_{f_m^d}$ (m^{th} diagonal element of matrix Σ_{f^d}) are the posterior mean and the variance parameters of f_m^d respectively. The estimate shown above would be low for the regression parameters that have the posteriors concentrated around zero, and high for the rest. By setting a threshold on this estimate heuristically, one can split the regression parameters into two sets, (i) significant parameters and (ii) insignificant parameters.

Instead of setting a threshold heuristically, we automate the model reduction part. We use the estimate, $\left(\hat{\nu}_m^d\right)^{-1}$, to switch from a complex model to a simpler model. We proceed with irreversible deterministic switches from the converged complex models to simpler models. During each switch, the posterior parameters associated with a regression parameter that has the lowest estimate of $\left(\hat{\nu}_m^d\right)^{-1}$ is excluded from the model to form a relatively simpler model. The variational lower bound the simpler model is estimated after optimizing the parameters in the simpler form. The switch is accepted only if the new lower bound estimate is higher than the lower bound estimate for the previous model. When the lower bound estimate stops improving, the model reduction is also stopped. The entire procedure is summarized in the steps from 13 to 28 in Table. 5.1.

Although switching to a simpler model means excluding a set of parameters from the model, one should not have a problem with computing the lower bound estimates and carrying out the VBEM updates for the remaining parameters. All we have to do is to remove the parameters and the predictors explicitly from the update expressions or set them to zero accordingly. For example, to update Σ_x and $\hat{x}_n$, it is sufficient to set the parameters excluded in $\hat{v}^d$ (estimate of the d^{th} of column of V) and the columns and rows corresponding to parameters excluded in Σ_{v^d} (covariance of the estimate of the d^{th} of column of V) to zero. To update $\hat{f}^d$ and Σ_{f^d}, it is sufficient to remove the predictors corresponding to the dropped parameters. By following the steps outlined in Table. 5.1, we should be able to start with a complex model and switch to a simpler

model that retains only the relevant parameters. Ultimately, from the reduced model obtained using the steps in Table. 5.1, we can directly infer the presence of causal connections (from the non-zero parameters) and the contemporaneous correlation features.

5.6 Case Studies

We present two different case studies to demonstrate the proposed approach. One is a simulation case study and the other is a real industrial case study. The simulation case study is used to check whether the proposed approach can mine the true causal connections and determine the right number of contemporaneous correlation features. The industrial case study is used to check whether the model results can be interpreted physically. In both cases, we compare the results against the conditional Granger causal connections inferred from the VAR model estimated under the maximum likelihood (ML) framework. For this, we utilize the multivariate Granger causality toolbox [50]. To select the appropriate model order in the case of the ML approach, we use the Bayesian information criterion (BIC) available in the toolbox. In addition, we also compare the results against the causal connections inferred from the VAR model estimated under the VBEM framework for the simulation case study.

5.6.1 Simulation Case Study

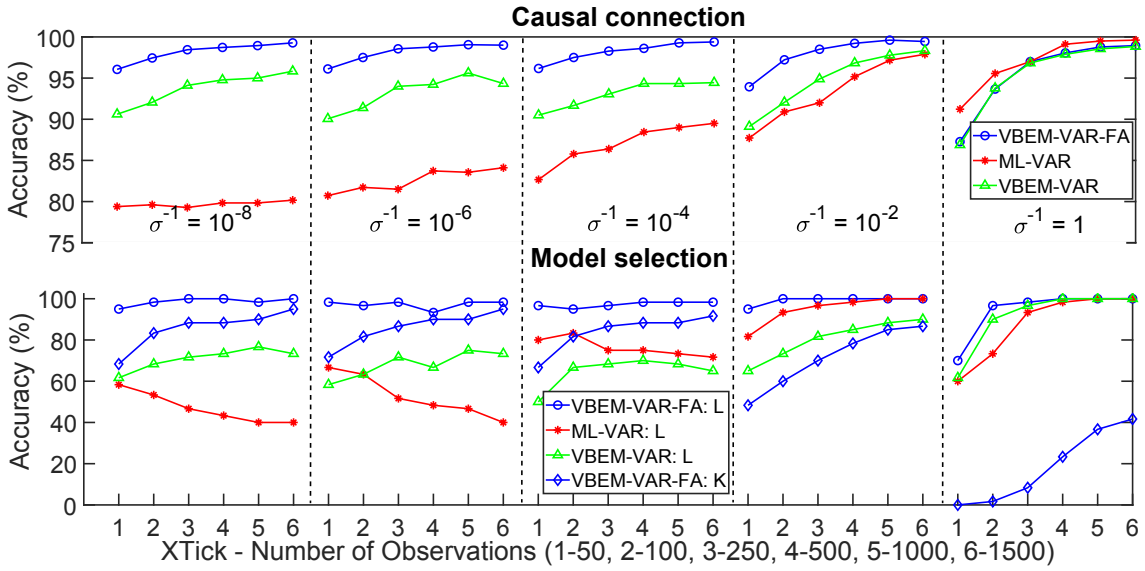


Figure 5.2: Summary of results for the simulation case study: the accuracy of causal connections inference (top) and accuracy of model selection (bottom). Panels separated by the dashed lines present result for different noise levels σ^{-1} and each panel presents the results for six different run lengths as indicated in x-label. Acronyms of the estimation approach followed by acronyms of model types are used as legends. For model selection, legends followed by L indicate the model order selection accuracy and the one followed by K indicates the correlation features selection accuracy

Model and data description

We constructed multiple stable VAR models and combined it with the FA models to simulate multiple datasets. The details of the models and the datasets are summarized in Table. 5.2. We constructed sixty models that have different combinations of L and K to generate six dimensional datasets. We chose a wide range of sparsity values for causal connections and the loading matrices of the FA models. Here, the sparsity is defined as the ratio between the number of active connections or the number of non-zero parameters and the maximum possible number of connections. These values were drawn from a uniform distribution with an interval from 0.4 to 0.9 for the VAR component, and from 0.33 to 0.66 for each column of the FA component. The sparsity values chosen for the FA model ensures that each column has at least two non-zero entries. The parameters of the models were drawn from two uniform distributions with equal probabilities that have intervals from -0.95 to -0.05 and from 0.05 to 0.95

respectively. Model draws were accepted only if the VAR part passes the stability test and the FA models have full column rank. We tested the approaches for different time series lengths and at different noise variances as listed in Table. 5.2. Since our models have the maximum values of L and K as 3, estimations for all the datasets were started with a fourth order model with four latent variables in the case of the proposed model. For the VAR models estimated under the ML and the VBEM framework, the maximum model order that considered was limited to 4.

Table 5.2: Simulated model and data characteristics

Attribute	Value
Number of models	60
Model order (K)	1 to 3
Number of latent variables (L)	1 to 3
Time series dimension (D)	6
Time series run length (N)	50, 100, 250, 500, 1000, 1500
Sparsity of the causal connections	$\sim \mathcal{U}(0.4, 0.9)$
Sparsity of the FA model parameters	$\sim \mathcal{U}(0.33, 0.66)$
Parameters (F)	$\sim \mathcal{U}(-0.95, 0.05)\,\&\,\mathcal{U}(0.05, 0.95)$
Noise variance (σ^{-1})	10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1

Results

Fig. 5.2 presents the summary of the results. It presents two different metrics, the accuracy of causal connections inference and the accuracy of model selection. The accuracy of causal connections inference is defined by the percentage of correct inferences (presence and absence of connections) made. In the case of the VAR and the proposed models estimated under the VBEM framework, there were 2160 inferences (with 36 inferences for each of the 60 models) under each combination of N and σ^{-1} to be made. For the ML approach, only the inferences of inter-causal connections among the variables are generated by the toolbox. Therefore, excluding the 6 self causal connections for each model, there were 1800 inferences to be made by the ML approach. Model selection accuracy is defined by the percentage of the correct model structures selected. There were 60 selections under each combination of N and σ^{-1} to be made.

In terms of the accuracy of causal connections inference, the proposed model significantly outperformed the other two approaches at $\sigma^{-1} \leq 10^{-2}$ irrespective of the

run lengths of the time-series data. At $\sigma^{-1} = 1$, accuracies of the VAR model and the proposed model under the VBEM framework were not distinguishable, however, the ML approach performed slightly better. In terms of the accuracy of model order selection, the proposed model under the VBEM framework performed better than the other two approaches. At higher noise levels, the accuracies of the other two started to approach the accuracy of the proposed approach.

The accuracy of contemporaneous correlation features selection by the proposed model improved with the increase in N, however, dropped significantly at higher noise levels.

There are two unobserved quantities, noise and the latent variables through which the system is excited. From the results, it appears that the relative strength of both plays a crucial role in the performance of each of these approaches. At higher noise levels, the diagonal elements of the prediction error covariance become more dominant and the excitation of the system is dominated more by the noise term. This also decreases the relative significance of the contemporaneous interactions which is the reason why the contemporaneous correlation features selection results deteriorate with the increased noise levels even though the causal connections were inferred with higher accuracies. At low noise levels, the latent variables start to dominate, leading to relatively significant contemporaneous interactions. This was responsible for the better performance of the proposed model at low noise levels.

The VBEM regularization itself tends to improve the performance of the VAR model by avoiding the spurious inference of causal connections when the influence of the latent variables is significant. However, the accuracy of causal connections inference by the VAR model estimated under the VBEM framework tends to be 4 to 6% lower when compared to the proposed model. This improvement in the performance of the proposed model could be attributed to the inclusion of the FA component.

Both false positives and miss detections contributed to the spurious inference by the ML approach when the contemporaneous interactions were relatively significant. At $\sigma^{-1} = 10^{-8}$, false positives and miss detections were 15 times and 6 times more respectively than at $\sigma^{-1} = 1$.

5.6.2 Industrial Case Study

Sulphur recovery unit (SRU) with Claus sulphur recovery process is our industrial case study. SRUs are very common in the sulphur handling plants that control the sulphur emission. In our case, the unit is a part of the sulphur handling plant that treats acid gases from the oil sands upgrading process.

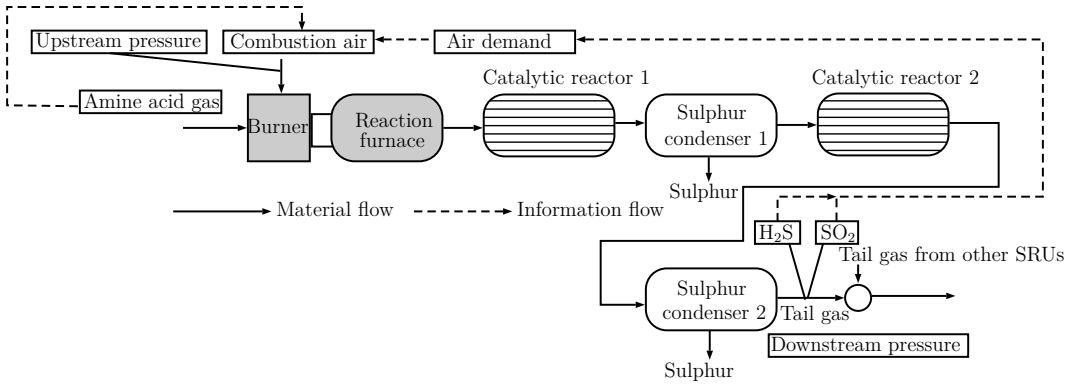


Figure 5.3: Simplified schematic diagram of the sulphur recovery unit

Process description

The schematic diagram of the unit is shown in Fig. 5.3. The unit recovers sulphur content from the upstream amine acid gas (AG). AG contains sulphur in the form of H_2S. H_2S is oxidized to elemental sulphur through a series of catalytic reactors in the SRU. The unit draws combustion air for oxidation. Following the oxidation step, the elemental sulphur is recovered in the sulphur condenser. The tail gas containing unconverted H_2S and SO_2 as major components leaves the unit as a by-product. It is critical to maintain a set ratio of H_2S and SO_2 concentrations in the tail gas to maintain a smooth operation of the downstream tail gas treatment unit in the plant. Solid arrows in Fig. 5.3 indicate the material flow through the equipment in the unit. Dashed arrows indicate the information flow in the unit which occurs mainly due to the presence of the control loops. The unit has a feed-forward ratio controller to draw combustion air in ratio with the AG flow rate. It also has a feedback controller that provides correction to the demanded combustion air flow rate by the feed forward loop, which is achieved using an air demand analyser that takes H_2S and SO_2 concentrations in the tail gas into account. Normally, the feedback controller

provides minor corrections which are limited by a threshold of 5% of the total combustion air inlet demanded by the feed-forward loop. From the process knowledge, one would expect both contemporaneous and causal connections to be present in the system, making the proposed approach more suitable as opposed to the traditional techniques. As the air demand is instantaneously calculated from the tail gas concentration measurements coming from a noisy instrument, we expect the presence of the contemporaneous correlation among these variables. Similarly, AG flow, combustion air flow and the pressure drop (DP) across the unit may have both contemporaneous and causal relationships depending on the choice of sampling rate. Downstream concentrations and air demand are expected to be influenced both causally and contemporaneously by either one or more of flow and DP variables due to the process dynamics.

Data description

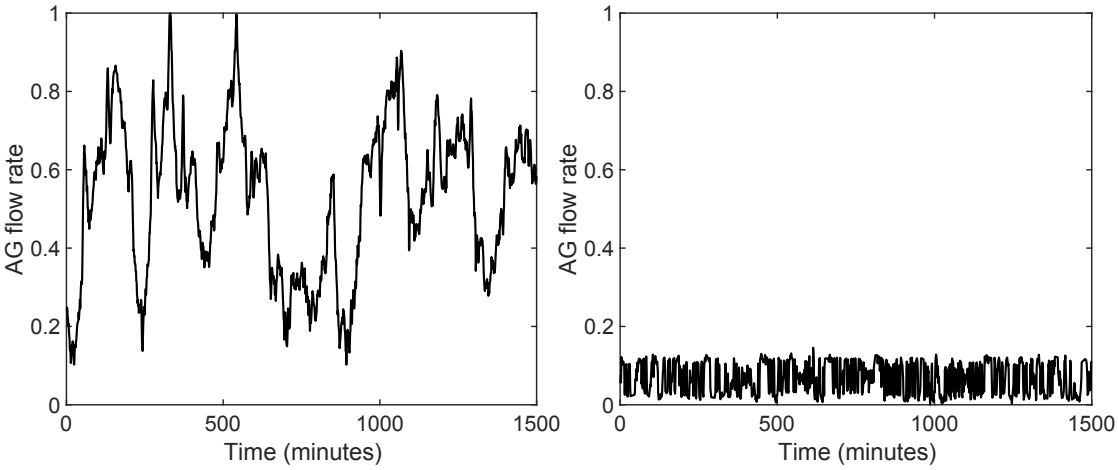


Figure 5.4: Normalized AG flow rate during two different periods of operation: Period I (left) and Period II (right).

We used the routine operation data for our analysis. The tags listed in Table. 5.3 were used for the analysis. Out of the used tags, AG depends mainly on the upstream operations and cannot be predicted using the other tags listed here. Therefore, it was used as an exogenous input in our analysis and the other tags were considered as outputs.

We considered data from two different operating periods for our analysis. During each period, we had data for 1500 minutes and the data samples were obtained from the plant historian at one minute frequency. The major and measured disturbance to the process comes in the form of AG. The normalized values of AG during these periods are shown in Fig. 5.4 for comparison. During Period I, we had the AG values varying significantly and during period II, we had a constant AG value with minor fluctuations. For our analysis, we randomly sampled multiple shorter windows of data from the operating periods and estimated models from those sampled windows. This was done for two reasons, 1) to check the consistency of the obtained causal connections and contemporaneous correlations across multiple windows and 2) to avoid the effect of non-linearity; assuming that within a shorter window, the process can be approximated by a linear model. Here, we present the results corresponding to 100 randomly sampled windows each with a run length of 500 minutes from both operating periods. The results presented here correspond to models obtained from the normalized data within each of the sampled windows.

Table 5.3: Tags used for the analysis and their descriptions

Tag	Description	Tag	Description
AG	Amine acid gas flow rate	H_2S	Tail gas H_2S concentration
Air	Combustion air flow rate	SO_2	Tail gas SO_2 concentration
DP	Pressure drop	ADA	Air demand analyzer

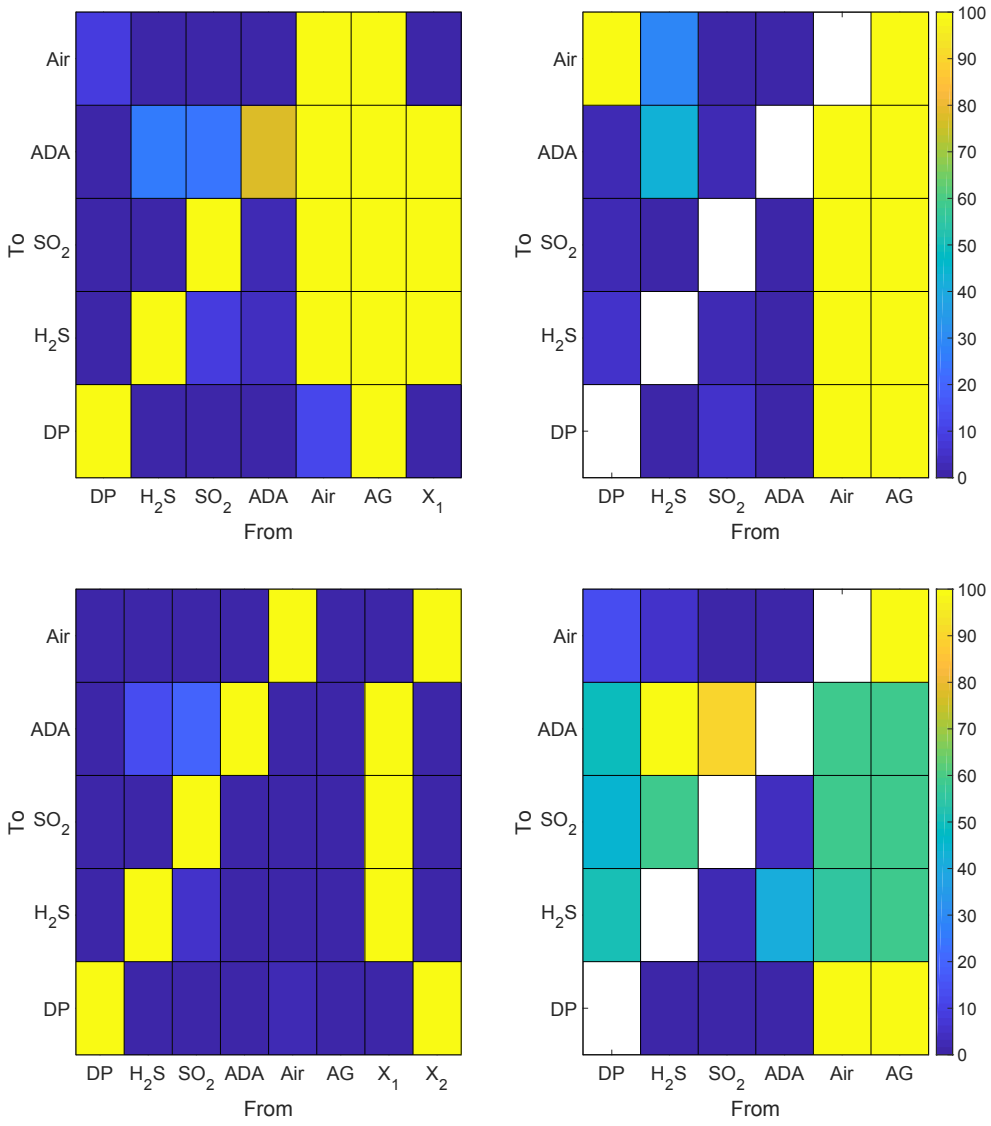


Figure 5.5: Summary of the results for the industrial case study: for period I using the proposed method (top left), for period I using the VAR model estimated under the ML framework (top right), for period II using the proposed method (bottom left) and for period II using the VAR model estimated under the ML framework (bottom right). Rows correspond tp outputs and columns correspond to inputs. Variables X_1 and X_1 in the inputs correspond to the latent variables. Bright yellow squares correspond to the presence of connections (non-zero coefficients) in all the 100 sampled windows, dark blue squares correspond to the absence of connections in all the windows and white squares correspond to the unavailability of the results.

Fig. 5.5 shows the summary of the analysis. In the case of ML approach, the toolbox provides only the results of inter causal relationships. Therefore, the results on the self causal connections are not reported for the ML approach. We compare and discuss the results from the both approaches below for both periods of data.

Period I: The proposed approach indicated that all the output variables were influenced by their own past observations. Both approaches indicated that Air and AG causally influence H_2S, SO_2 and ADA as expected. AG causally influences DP and Air which is due to the presence of the feed-forward loop and material flow causing changes in DP, which have been consistently identified by both approaches. H_2S, SO_2 and ADA are influenced by a common latent variable as found by the proposed approach, indicating the contemporaneous relationship as expected. This feature most likely represents the instrumentation noise.

The ML-VAR approach also indicated the presence of causal connections between Air and DP, which was not identified by the proposed approach. This, we suspect to be a contemporaneous relationship disguised as time-lagged causal dependency in the case of ML approach, however, appeared suppressed in the case of proposed approach. The potential reason for this could be due to the higher variability in AG causing the relative strength of the unexplained covariance of DP and Air to be less significant compared to their individual variances. The reason becomes clearer from the results for period II where AG has low variability and this contemporaneous relationship was explained by an additional feature by the proposed approach.

Period II: With low variability in AG, the proposed approach showed the absence of inter-causal relationships in the system. Two contemporaneous correlation features, one explaining the correlations among H_2S, SO_2 and Air and the other explaining the correlation between Air and DP were found. In the case of ML approach, these contemporaneous relationships can be seen to disguise themselves as the time lagged causal connections as both H_2S and SO_2 were shown to causally affect ADA, Air was shown to causally affect DP, and AG was shown to affect both Air and DP. It can also be observed that some of the causal connections were only present in the half of the sampled intervals (green boxes instead of yellow boxed) but not in all, indicating that the identified connections by the ML approach could be spurious.

In both periods, ADA was not found to affect Air by both approaches, indicating that the corrections provided by the feedback loop were not significant.

5.7 Summary

In this chapter, we presented a hybrid model formed by combining the VARX model and the FA model for causal network reconstruction. The hybrid model parameters were regularized by the hierarchical prior distributions and estimated under the VBEM approach. The approach allowed us to reconstruct causal networks in the presence of contemporaneous correlations among the variables in the data. The approach outperformed the VAR models estimated under the maximum likelihood approach for causal network reconstruction in the simulation case studies. In the industrial case study, the hybrid model reconstructed interpretable causal connections and contemporaneous correlation features.

Chapter 6

A Causal Analysis Approach for Time-Varying Systems

6.1 Introduction

In this chapter, we present a causal modelling approach for the time-varying systems. The approach relies on the time-varying parameter models (TVPMs) estimated under the variational Bayesian expectation maximization (VBEM) framework. We incorporate a hypothesis switching procedure followed by the VBEM estimation that allows us to infer the time-varying strengths of causal influence of the inputs on the outputs of the system. We illustrate the proposed approach using the production data from steam assisted gravity drainage (SAGD) wells. The proposed approach was found to extract consistent causal models of the SAGD system across multiple case studies and outperform the time-invariant models.

Study of long-term effects of process variables on the key performance indicators (KPIs) of process systems has the following advantages, it can help (i) optimize the process KPIs, (ii) design closed loop control or optimization framework for the KPIs, (iii) understand the anomalies in the KPIs, etc. Long-terms effects can be studied using causal models between the KPIs and the process variables. The causal models can be obtained from the first principles understanding of the process, however, when the process is poorly understood, data-driven models are often the only alternatives. One could set up experiments and collect reliable data for causal modelling, however, experiments affect routine operations and are expensive to conduct in process systems. On the other hand, operational data from routine operations are easy

and inexpensive to obtain. Nevertheless, one may have to be wary of multiple data quality issues when using the operational data for obtaining the causal models. One such issue is the time-varying nature of the process systems and its effect on the observed data. For example, the relationship between a KPI of a plant and the process variables may vary with changes in the operating modes, physical plant modifications, equipment fouling, etc.

In the exercise of causal modelling, the time-varying nature of the process system and its effect on the observed data are often ignored, which may lead to inaccurate findings. The time-varying nature of the process systems has been well acknowledged and addressed in some of the other process data-driven applications in the literature. Examples include: the use of multi-modal modelling and adaptive or recursive modelling strategies to account for changes in process operating modes or the time-varying nature in soft sensor development and process monitoring [95, 96, 97, 98, 99]. However, this is not the case with the causal modelling of process systems from data. To handle time-varying nature of the process systems and its effect on the observed data, we present a time-varying parameter model (TVPM) based causal modelling approach in this chapter. We briefly introduce the recurring symbols and notations below before we present the TVPM utilized in this chapter.

Recurring symbols and notations: The time instant at which a particular measurement is made is represented by a subscript, for instance, t in y_t corresponds to the instant at which the measurement y is made. A particular dimension of a vector is represented by a superscript, for instance, d in u_t^d corresponds to the dth element of the vector u_t. $\mathbb{R}$ and $\mathbb{R}_+$ represent the spaces of real and real positive numbers respectively, and the dimensions (rows $\times$ columns) of those spaces are indicated in superscripts. $\mathcal{N}$, Ga, and δ represent the multivariate normal, univariate gamma, and univariate delta distributions, respectively. $p(a|b)$ represents the conditional distribution of a given b. $q(a)$ represents the functional approximation of the posterior distribution of a. $E(.)$ represents the expectation operation, and $diag(.)$ represents the operation that converts a vector to a diagonal matrix and vice versa. $[a, b]$ represents a concatenated matrix or vector (along the rows) formed using the vectors or matrices a and b of appropriate dimensions. Subscript $1:t$ correspond to a collection

of a particular variable from time instant 1 to t.

Definition 5. *Time-varying parameters model (TVPM): Consider a system with a set of normalized output observations,* $Y \triangleq \{y_1, ..., y_t \in \mathbb{R}^1, ..., y_T\} \in \mathbb{R}^{1\times T}$ *and a set of normalized input observations,* $U \triangleq \{u_1, ..., u_t \in \mathbb{R}^D, ..., u_T\} \in \mathbb{R}^{D\times T}$. *The TVPM describing these observations is defined as follows,*

$$y_t = \theta_t^T u_t + e_t,\ e_t \sim \mathcal{N}\left(0, \sigma^2\right) \tag{6.1}$$

where $y_t \in \mathbb{R}^1$ *is the measurement of the output variable at time* t, $u_t \in \mathbb{R}^{D\times 1}$ *are the measurements of the input variables that are hypothesised to causally influence* y_t *at time* t, *and* $e_t \in \mathbb{R}^1$ *is the noise in the measurement,* y_t *and is Gaussian distributed with mean zero and variance* σ^2, *respectively.* θ_t *contains the model coefficients or the strengths of causal influences at* t. *The coefficients are allowed to vary with time and the set of coefficients at different time instants is given by* $\Theta \triangleq \{\theta_0, ..., \theta_t \in \mathbb{R}^D, ..., \theta_T\} \in \mathbb{R}^{D\times T+1}$.

Remark 10. *The parameters/coefficients in this model provide the direct causal effects of inputs on the output. This is because the parameter* θ_t^d *is essentially the partial derivative of the output with respect to input* u^d *at time instant* t *as given below,*

$$\frac{\partial y}{\partial u^d}\ |_t = \theta_t^d \forall\ t,\ d \tag{6.2}$$

We propose an approach to determine the hypotheses that best match the distributions from which the coefficients are drawn among the ones presented below. The coefficients for any given combination of $1 \leq t \leq T$ (time) and $1 \leq d \leq D$ (input dimension) are drawn from either of the following hypotheses that best fits the given observations,

$$\begin{aligned} H_0(t,d):\ & \theta_t^d = \theta_{t-1}^d + \epsilon_t^d \\ H_1(t,d):\ & \theta_t^d \sim \delta\left(\theta_t^d | \theta_{t-1}^d\right) \end{aligned} \tag{6.3}$$

and for any $1 \leq d \leq D$ at $t = 0$,

$$\begin{aligned} & H_0(0,d): \theta_0^d = \epsilon_0^d \\ & H_1(0,d):\ \theta_0^d \sim \delta\left(\theta_0^d | 0\right) \end{aligned} \tag{6.4}$$

where

$$\epsilon_t^d \sim \mathcal{N}\left(0, diag(\nu_t^d)^{-1}\right)\ \&\ \nu_t^d \sim Ga\left(\alpha^*, \beta^*\right) \forall t \tag{6.5}$$

$H_0(t,d)$ for any $1 \leq t \leq T$ considers that the coefficient of u_t^d, θ_t^d is time-varying and given by the summation of θ_{t-1}^d and the additive noise ϵ_t^d; ϵ_t^d is Gaussian distributed with zero mean and ν_t^d precision, and ν_t^d follows a gamma distribution with shape parameter α^* and rate parameter β^*. $H_1(t,d)$ for any $1 \leq t \leq T$ considers that, θ_t^d is drawn from the delta distribution, $\delta\left(\theta_t^d|\theta_{t-1}^d\right)$, i.e., $\theta_t^d = \theta_{t-1}^d$ (θ^d does not change at t). $H_0(0,d)$ considers that, θ_0^d is drawn from a Gaussian distribution with zero mean and ν_0^d precision, and $H_1(0,d)$ considers $\theta_0^d = 0$.

Remark 11. *Note that the TVPM shown in Eqn. (6.1) is a multivariate linear regression model with parameters that can vary with time. If u_t is of the time lagged inputs and outputs concatenated together, the model takes the form of the autoregressive exogenous (ARX) model with parameters that can vary with time.*

Remark 12. *Note that the hypotheses shown in equations (6.3) and (6.4) cover multiple possibilities. The causal strength of an input can change at all time instants if drawn from H_0 at all the instants or change at only few instants (if drawn from H_0 at those instants) and not change at the rest (if drawn from H_1 at the rest of the instants). If an input variable does not influence Y at all and its coefficients are drawn from H_1 at all the time instants, the causal strength of that particular variable may remain zero at all time instants from $t = 0$ to $t = T$.*

We present a methodology to estimate the TVPM and identify the hypotheses for the coefficients that best fit the data. In the first step, we estimate the TVPM under the VBEM framework assuming that the coefficients are drawn from $H_0(t,d)\ \forall\ 0 \leq t \leq T\ \&\ 1 \leq d \leq D$. The TVPM with the prior distributions defined by H_0 falls under the class of conjugate exponential family graphical models (CEFGMs). The VBEM estimation provides the posterior distribution of the model parameters and a lower bound of the log marginal distribution of the data. In the second step, we incorporate a hypothesis switching approach. The approach switches the hypothesis of a set of coefficients from H_0 to H_1 such that the lower bound on the log marginal distribution of the data is maximized.

6.1.1 Summary of the Main Contributions

Our approach to causal modelling using the TVPMs is similar to the concept of path analysis. Path analysis is a well established statistical method to test the consistency of the observed data with the graphical models containing the hypothesised causal connections. It has found applications in several fields including biology, psychology, sociology and linguistics. However, conventional path models seldom consider the time-varying nature of the process and its effect on the observed data. Therefore, our proposal can be viewed as a path modelling approach for the time-varying systems such as chemical processes.

We illustrate the proposed causal modelling approach using the production data from SAGD wells, commonly used in heavy oil extraction. SAGD is an in-situ thermal oil sands extraction technique, used to produce bitumen from the oil sands formation several hundred meters below the earth's surface. In this application, we define a causal model for the production rate from a SAGD well based on Darcy's law. We show using the case studies that the parameter invariant path models may lead to misleading conclusions that are inconsistent with the conventional understanding. However, the proposed approach recovers causal strengths that are physically interpretable and supports the theoretical understanding of the process causal relations and their strengths.

The main contributions in this chapter can be summarized as follows: 1) an approach for causal modelling of the time-varying systems using the TVPMs is proposed, and 2) an algorithm based on the VBEM framework and a hypothesis switching approach to infer the time-varying causal strengths is developed and presented, and 3) production data from SAGD wells has not been studied extensively in the literature previously and here, we study it using the proposed approach in this chapter.

6.1.2 Relevant Works

The linear TVPMs retain a simple interpretable structure and have the ability to approximate any form of non-linear process [100]. The TVPMs have been studied for decades (for example, see [101, 102]) and remains as an important subject of study, especially in the field of econometrics. Bayesian analysis of the TVPMs

with the prior distributions for the coefficients (similar to H_0) has garnered more attention recently. The prior distributions regularize the rate of change in coefficients between the successive time instants, thereby decreasing the prediction bias of the econometric time series models. The prior distributions for the TVPMs are designed or chosen to address the following objectives, (i) infer the evolution of coefficients over time, (ii) infer the sparsity of the models/variable selection at different time instants, (iii) constrain the evolution of the coefficients to be stationary, and (iv) infer the evolution of a specific subset of parameters in the model over time [103, 104, 105, 106, 107, 108, 109]. In most of these cited works, full blown Bayesian analysis has been employed through Markov chain Monte Carlo (MCMC) sampling approaches. However, the use of MCMC sampling approaches may prevent the application of the TVPMs in places where the computational load is of concern. The VBEM framework when compared to the MCMC sampling approaches, greatly simplifies the Bayesian analysis through functional approximations for the posterior distribution. Although it involves simplifying the actual posterior distribution through the mean-field approximation, the VBEM framework has been found to successfully recover the correct underlying model structures of several statistical models that belong to the CEFGMs [67]. Therefore, the implementation of the VBEM algorithm would make the application of the proposed causal modelling approach viable in many settings.

The TVPMs with the coefficients evolving through an autoregressive process are in essence state space or linear dynamic models with coefficients treated as states. Beal (2013) [67] proposed the treatment of state space models under the VBEM framework. However, they do not assign a prior distribution for the noise covariance of the states to determine whether the states evolve at a particular time instant or not. Koop and Korobilis (2018) [110] in their working paper proposed a variational Bayesian filter for the TVPMs with the prior distributions for the parameter covariance matrix. The filtering step may alone be suitable for the online prediction scenarios. However, if the accurate parameter inference is of importance, in addition to filtering, backward recursion or smoothing will also be necessary for the TVPMs. In this chapter, we propose a parameter estimation strategy based on the variational Bayesian filtering

and smoothing for the TVPMs. In addition to the parameter estimation, we provide a hypothesis switching strategy to infer if the coefficients of the system/causal strengths of the inputs change at a particular time instant. Our hypothesis testing strategy is based on the model reduction strategy presented in the previous chapter. In this chapter, we extend the strategy presented previously for the time-varying causal models.

The rest of the chapter is organized as follows: In section II, we present the VBEM approach to estimate the TVPM under H_0 as the prior distribution for all the coefficients at all the time instants. In section III, we present the hypothesis testing approach. In Section IV, we discuss the implementation details of the approach used. In section V, we present a number of SAGD case studies. Finally, in section VI, we present the concluding remarks.

6.2 Estimation

In this section, we present the VBEM approach for estimating the model shown in Eqn. (6.1) assuming that the coefficients are drawn from $H_0(t,d)\ \forall\ 0 \leq t \leq T,\ \&\ 1 \leq d \leq D$. The joint distribution of the variables given the inputs, noise variance and the hyper-parameters under this setting is given by,

$$p\left(Y,\Theta,\nu|U,\sigma,\alpha^*,\beta^*\right) = p\left(\theta_0|\nu_0\right)p\left(\nu_0|\alpha^*,\beta^*\right)\times$$
$$\prod_{t=1}^{T} p\left(\nu_t|\alpha^*,\beta^*\right)p\left(y_t|\theta_t,u_t,\sigma\right)p\left(\theta_t|\theta_{t-1},\nu_t\right) \tag{6.6}$$

where $\nu_t = \left[\nu_t^1,...,\nu_t^D\right]^T \in \mathbb{R}_+^{D\times 1}$ is the vector of precision parameters at time instant t, and $\nu = \{\nu_0,...,\nu_t,...,\nu_T\} \in \mathbb{R}_+^{D\times(T+1)}$ is the collection of precision parameters. Estimation of this model amounts to obtaining the posterior distribution of Θ and ν $\left(p\left(\Theta,\nu|Y,U,\sigma,\alpha^*,\beta^*\right)\right)$, and the optimal estimate of σ that maximizes the marginal distribution of the outputs given the inputs, noise variance, and the hyper-parameters $\left(p\left(Y|U,\sigma,\alpha^*,\beta^*\right)\right)$.

6.2.1 VBEM Algorithm

In the VBEM algorithm, the joint posterior distribution is approximated by a set of independent posterior distributions as shown below,

$$p\left(\Theta,\nu|Y,U,\sigma,\alpha^*,\beta^*\right) \approx q\left(\Theta\right)q\left(\nu\right) \tag{6.7}$$

where $q\left(\Theta\right)$ and $q\left(\nu\right)$ are the functional approximations of the posteriors of Θ and ν. The actual log marginal distribution of the data can be related to the approximated posterior distribution through a lower bound expression, $\mathcal{L}_{VB}$ as the following,

$$\ln p\left(Y|U,\sigma,\alpha^*,\beta^*\right) = \underbrace{\int_\nu\int_\Theta q\left(\nu\right)q\left(\Theta\right)\ln\frac{p\left(Y,\Theta,\nu|U,\sigma,\alpha^*,\beta^*\right)}{q\left(\nu\right)q\left(\Theta\right)}d\nu d\Theta}_{\mathcal{L}_{VB}} + \underbrace{\int_\nu\int_\Theta q\left(\nu\right)q\left(\Theta\right)\ln\frac{q\left(\nu\right)q\left(\Theta\right)}{p\left(\Theta,\nu|Y,U,\sigma,\alpha^*,\beta^*\right)}d\nu d\Theta}_{KL\ divergence} \tag{6.8}$$

The lower bound, $\mathcal{L}_{VB}$ will be lower than the actual log marginal distribution by the Kullback-Leibler (KL) divergence between the approximated posterior distribution and the actual posterior distribution as shown above. Therefore, optimizing the lower bound with respect to the approximated posterior distribution minimizes the gap between the lower bound and the actual log marginal distribution, and minimizes the KL divergence between the approximated posterior and the actual posterior distribution. $\mathcal{L}_{VB}$ can be further expanded as follows,

$$\begin{aligned}\mathcal{L}_{VB} = &\int_\nu q\left(\nu\right)\ln\frac{p\left(\nu|\alpha^*,\beta^*\right)}{q\left(\nu\right)}d\nu - \int_\Theta q\left(\Theta\right)\ln q\left(\Theta\right)d\Theta \\ &+ \int_\nu\int_\Theta q\left(\nu\right)q\left(\Theta\right)\ln p\left(Y,\Theta|\nu,\sigma\right)d\nu d\Theta\end{aligned} \tag{6.9}$$

The expression above becomes tractable when the posterior distribution belong to the same distribution families as that of the prior distributions i.e., when $q(\nu)$ is a gamma distribution as of $p(\nu)$ and $q(\Theta)$ is a multivariate normal distribution as of $p(\Theta)$.

$\mathcal{L}_{VB}$ can be optimized by sequentially updating $q\left(\nu\right)$, $q\left(\Theta\right)$, and σ. The sequential update expressions can be obtained by equating the derivatives of the lower bound with respect to $q\left(\nu\right)$, $q\left(\Theta\right)$, and σ to zero. These update expressions take $\mathcal{L}_{VB}$ to its

maxima with respect to the updated quantities. The update expression for $q(\nu)$ can be obtained as the following,

$$\frac{d\mathcal{L}_{VB}}{dq\left(\nu\right)} = 0 \Rightarrow \ln q\left(\nu\right) \propto \int_{\Theta} q\left(\Theta\right) \ln p\left(Y, \Theta, \nu | U, \sigma\right) d\Theta \tag{6.10}$$

The above expression can be simplified to obtain the posterior of $\nu_t^d \sim Ga(\alpha, \beta_t^d)\ \forall\, t,\ d$. The parameters of the posterior gamma distributions are given in Table E.2 of Appendix E. Similarly, the update expression for $q(\Theta)$ can be obtained as,

$$\frac{d\mathcal{L}_{VB}}{dq\left(\Theta\right)} = 0 \Rightarrow \ln q\left(\Theta\right) \propto \int_{\nu} q\left(\nu\right) \ln p\left(Y, \Theta, \nu | U, \sigma\right) d\nu \tag{6.11}$$

which leads to

$$q\left(\Theta\right) \propto p\left(\Theta_0 | \Lambda_0\right) \prod_{t=1}^{T} p\left(y_t | \theta_t, u_t, \sigma\right) p\left(\theta_t | \theta_{t-1}, \Lambda_t\right) \tag{6.12}$$

where $\Lambda_t = diag\left(\left[\frac{\alpha}{\beta_t^1}, ..., \frac{\alpha}{\beta_t^D}\right]^T\right)$. The above expression for $q\left(\Theta\right)$ is a linear Gaussian state space model with states Θ. Therefore, we can use the Kalman filter and smoother to estimate the posterior of Θ. We utilize the following forward recursion starting from $t = 0$ to $t = T$ to estimate the filtered mean, $\mu_t\ \forall\ t$ and covariance $\Sigma_t\ \forall\ t$ of the coefficients,

$$\mathcal{N}\left(\theta_t | \mu_t, \Sigma_t\right) \propto$$
$$\int_{\theta_{t-1}} p\left(\theta_{t-1} | \mu_{t-1}, \Sigma_{t-1}\right) p\left(y_t | \theta_t, u_t, \sigma\right) p\left(\theta_t | \theta_{t-1}, \Lambda_t\right) d\theta_{t-1} \tag{6.13}$$

The deduced expressions for μ_t and Σ_t are provided in Table E.2 of Appendix E. We utilize the backward recursion of the following form to estimate the joint posterior of the coefficients at two successive time instants,

$$\mathcal{N}\left(\left[\begin{array}{c} \theta_t \\ \theta_{t+1} \end{array}\right] | \left[\begin{array}{c} \hat{\theta}_t \\ \hat{\theta}_{t+1} \end{array}\right], \Sigma_{\theta_t \theta_{t+1}}\right) \propto$$
$$\frac{p\left(\theta_t | \mu_t, \Sigma_t\right) p\left(\theta_{t+1} | \theta_t, \Lambda_t\right) p\left(\theta_{t+1} | \hat{\theta}_{t+1}, \Sigma_{\theta_{t+1}}\right)}{\int_{\theta_t} p\left(\theta_t | \mu_t, \Sigma_t\right) p\left(\theta_{t+1} | \theta_t, \Lambda_{t+1}\right) d\theta_t} \tag{6.14}$$

where $\hat{\theta}_t$ and $\hat{\theta}_{t+1}$ are the posterior mean estimates of the coefficients at t and $t+1$, respectively, $\Sigma_{\theta_t \theta_{t+1}}$ is the variance-covariance matrix of the coefficients at the successive time instants t and $t+1$, and $\Sigma_{\theta_{t+1}}$ is the covariance matrix of the coefficients

at $t+1$. The expression for the joint posterior updates are provided in Table E.2 of Appendix E. Equating the derivative of $\mathcal{L}_{VB}$ with respect to σ to zero provides the update expression for σ which is provided in Table E.2 of Appendix E. Now, we have the update expressions for the posteriors of Θ and ν, and σ. Implementing these updates sequentially through multiple iterations takes the $\mathcal{L}_{VB}$ to a local optima. To assess the convergence of the algorithm, improvements in the lower bound can be tracked until it becomes negligible between the two successive iterations.

Deducing $\mathcal{L}_{VB}$ from Eqn. (6.9) is cumbersome due to the entropy term, $\int_{\Theta} q(\Theta) \ln q(\Theta) \, d\Theta$. The joint posterior distribution of Θ is a high dimension multivariate Gaussian distribution, making the evaluation of the entropy term challenging. To circumvent this, we use a modified lower bound where Θ is integrated out or marginalized from the joint distribution. This modified lower bound is called the KL corrected lower bound in the literature [111, 112] and it upper bounds $\mathcal{L}_{VB}$ for the CEFGMs. The KL corrected lower bound in this case is given by,

$$\mathcal{L}_{KL} = \ln \int_{\Theta} \exp \left\{ \int_{\nu} q(\nu) \ln p(Y, \Theta, \nu | U, \sigma) \, d\nu \right\} d\Theta \geq \mathcal{L}_{VB} \tag{6.15}$$

Both $\mathcal{L}_{VB}$ and $\mathcal{L}_{KL}$ become equal when $q(\Theta)$ is at its optimal value. It is a well established general result that $\mathcal{L}_{VB}$ and $\mathcal{L}_{KL}$ become equal when the posterior of the marginalized quantities is at its optimum for the CEFGMs [111, 112]. Therefore, at each iteration after the $q(\Theta)$ update, both lower bounds will be equal and $\mathcal{L}_{KL}$ after the $q(\Theta)$ update at each iteration can be assessed for the convergence. Expression for $\mathcal{L}_{KL}$ can be deduced from Eqn. (6.15) as the following,

$$\mathcal{L}_{KL} = R + \ln p(y_1 | u_1, \sigma) + \sum_{t=1}^{T} \ln p(y_t | y_{1:t-1}, u_t, \sigma) \tag{6.16}$$

where R is a constant term which is independent of the observations. The predictive distribution of y_t from the past measurements $y_{1:t-1}$ at each time instant can be recursively obtained by integrating out or marginalizing the coefficients as the following,

$$p(y_t | y_{1:t-1}, u_t, \sigma) \propto$$
$$\int_{\theta_t} \int_{\theta_{t-1}} p(\theta_{t-1} | \mu_{t-1}, \Sigma_{t-1}) \, p(y_t | \theta_t, u_t, \sigma) \, p(\theta_t | \theta_{t-1}, \Lambda_t) \, d\theta_{t-1} d\theta_t \tag{6.17}$$

The resulting expression for $\mathcal{L}_{KL}$ is provided in Table E.1 of Appendix E.

6.3 Hypothesis Switching

After the VBEM estimation, we have the posteriors $q(\nu)$ and $q(\Theta)$. The expected values of the square of change is parameter θ^d at time instant t can be inferred from the posterior estimate of the precision parameter ν_t^d. This procedure is known as the automatic relevance determination in the literature [66, 33]. In fact, the inverse of the posterior estimate of the precision parameter is given by,

$$\frac{1}{\hat{\nu}_t^d} = \frac{\beta_t^d}{\alpha} = \frac{\beta^* + \frac{1}{2}E\left(\left(\theta_t^d - \theta_{t-1}^d\right)^2\right)}{\alpha^* + \frac{1}{2}} \tag{6.18}$$

where $E\left(\left(\theta_t^d - \theta_{t-1}^d\right)^2\right)$ is the expected values of the square of change in parameter θ^d at time instant t. We utilize this estimate, which we call the relevance criterion to switch the prior distributions from H_0 to H_1. We set a threshold for this estimate and switch the hypotheses at all t and d with $\frac{1}{\hat{\nu}_t^d}$ less than the threshold. The threshold is selected such that $\mathcal{L}_{KL}$ is maximized. This optimization problem is solved using 'bayesopt', the Bayesian optimization routine in MATLAB [113, 114, 115].

In the process of switching, evaluation of $\mathcal{L}_{KL}$ is not very straightforward. We will be confronted with the following challenges, (i) some coefficients may not change at a time instant t and take the same values as at $t-1$, (ii) some coefficients which have remained as zero till $t-1$ may take non-zero values at t, and (iii) some coefficients may change at both time instants, $t-1$ and t. Without loss of generality, the coefficients can be rearranged and stacked together to form different sets at each time instant. In the $\mathcal{L}_{KL}$ estimation step, we define the following coefficient sets in reference to the coefficients at time instant t: (i) let θ_t^+ be the set of coefficients that are drawn from H_1 till $t-1$ and from H_0 at t; meaning, they remain as zero before and take non-zero values at t, (ii) let θ_t^- and θ_{t-1}^- be the sets of coefficients that are drawn from H_0 at least once till $t-1$ and from H_1 at t; meaning $\theta_t^- = \theta_{t-1}^-$, (iii) let $\theta_t^\sim$ be the set of coefficients that are drawn from H_0 at least once till $t-1$ and from H_0 at t, (iv) let $\theta_t^* = \left[\theta_t^\sim, \theta_t^-\right]^T$, (v) let $\theta_t^\# = \left[\theta_t^\sim, \theta_t^+\right]^T$, and (vi) let $\theta_t^\& = \left[\theta_t^\sim, \theta_t^-, \theta_t^+\right]^T$. This would allow us to estimate $p\left(y_t|y_{1:t-1}, u_t, \sigma\right)$ in the $\mathcal{L}_{KL}$ expression as shown below,

$$p\left(y_t|y_{1:t-1}, u_t, \sigma\right) \propto \int_{\theta_t}\int_{\theta_{t-1}} p\left(\theta_{t-1}^*|\mu_{t-1}^*, \Sigma_{t-1}^*\right) \times$$

$$p\left(y_t|\theta_t^{\&}, u_t^{\&}, \sigma\right) p\left(\theta^{\#}|\begin{bmatrix}\theta_{t-1}^{\sim}\\ 0\end{bmatrix}, \Lambda_t^{\#}\right) d\theta_{t-1} d\theta_t \tag{6.19}$$

where θ_{t-1}^{*} is the realisation of θ_t^{*} at $t-1$, μ_{t-1}^{*} and Σ_{t-1}^{*} are its filtered mean and covariance, $u_t^{\&}$ are the inputs corresponding to the coefficients $\theta^{\&}$, and Λ_t^{+} is a diagonal matrix with the posterior estimates of the precision parameters of θ_t^{+}, and $\theta_{t-1}^{\sim}$ is the realisation of $\theta_t^{\sim}$ at $t-1$. The resulting expression for $\mathcal{L}_{KL}$ is provided in Table E.1 of Appendix E. The constant term R in Eqn. (6.16) needs to be estimated only for the set $\theta_t^{\#}$ which are drawn from H_0 at t.

With hypothesis switching, the posterior estimates may have to be fine tuned. This fine tuning requires much lesser number of iterations compared to the VBEM estimation since it is done on an already converged model. To perform the VBEM iterations in the reduced model setup, we need the update expressions in the reduced model setup. The update expression for $q(\nu)$ is the same as before except that, the updates are performed only for the coefficient that are drawn from H_0. Using the same coefficient sets defined for the $\mathcal{L}_{KL}$ evaluation, the recursion for the filtering step can be modified as,

$$\mathcal{N}\left(\theta_t^{\&}|\mu_t^{\&}, \Sigma_t^{\&}\right) \propto \int_{\theta_{t-1}} p\left(\theta_{t-1}^{*}|\mu_{t-1}^{*}, \Sigma_{t-1}^{*}\right) \times$$
$$p\left(y_t|\theta_t^{\&}, u_t^{\&}, \sigma\right) p\left(\theta^{\#}|\begin{bmatrix}\theta_{t-1}^{\sim}\\ 0\end{bmatrix}, \Lambda_t^{\#}\right) d\theta_{t-1} \tag{6.20}$$

where $\mu_t^{\&}$ and $\Sigma_t^{\&}$ are the filtered mean and covariance of the set $\theta_t^{\&}$, and their expressions are provided in Table E.3 of Appendix E.

We define new coefficient sets for the recursion in the smoother step in reference to the coefficients at time instant $t+1$; (i) let $\theta_t^{\sim}$ be the set of coefficients that are drawn from H_0 at least once till t, from H_0 at $t+1$ and change to $\theta_{t+1}^{\sim}$, (ii) let θ_t^{-} and θ_{t+1}^{-} represent the set of coefficients that are drawn from H_0 at least once till t and from H_1 at $t+1$, and (iii) let $\theta_t^{*} = \left[\theta_t^{\sim}, \theta_t^{-}\right]^T$, the combined set of coefficients at t that are drawn from H_0 at least once till t. With these new notations, the smoother step can be expressed as,

$$\mathcal{N}\left(\begin{bmatrix}\theta_t^{*}\\ \theta_{t+1}^{\sim}\end{bmatrix}|\begin{bmatrix}\hat{\theta}_t^{*}\\ \hat{\theta}_{t+1}^{\sim}\end{bmatrix}, \Sigma_{\theta_t^{*}\theta_{t+1}^{\sim}}\right) \propto$$

$$\frac{p\left(\theta_t^*|\mu_t^*,\Sigma_t^*\right)p\left(\theta_{t+1}^{\sim}|\theta_t^{\sim},\Lambda_{t+1}^{\sim}\right)p\left(\left[\begin{array}{c}\theta_{t+1}^{\sim}\\ \theta_t^{-}\end{array}\right]|\left[\begin{array}{c}\hat{\theta}_{t+1}^{\sim}\\ \hat{\theta}_{t+1}^{-}\end{array}\right],\Sigma_{\theta_{t+1}^*}\right)}{\int_{\theta_t^*}p\left(\theta_t^*|\mu_t^*,\Sigma_t^*\right)p\left(\theta_{t+1}^{\sim}|\theta_t^{\sim},\Lambda_{t+1}^{\sim}\right)}d\theta_t^* \tag{6.21}$$

where $\hat{\theta}_t^*$ and $\hat{\theta}_{t+1}^{\sim}$ are the posterior means of θ_t^* and $\theta_{t+1}^{\sim}$, respectively, $\Sigma_{\theta_t^*\theta_{t+1}^{\sim}}$ is the variance-covariance matrix of θ_t^* and $\theta_{t+1}^{\sim}$, μ_t^* and Σ_t^* are the filtered mean and covariance of θ_t^*, $\Lambda_{t+1}^{\sim}$ is a diagonal matrix of the posterior precision parameter estimates for $\theta_{t+1}^{\sim}$, $\hat{\theta}_{t+1}^{-}$ is the posterior mean of θ_{t+1}^{-}, and $\Sigma_{\theta_{t+1}^*}$ is the posterior covariance matrix of θ_{t+1}^*. The expressions for these estimates are provided in Table E.3 of Appendix E.

6.4 Initialization and Hyper-Parameter Tuning

In our studies, we implement the VBEM updates in each iteration in the following order, (i) $q\left(\Theta\right)$ update; filtering step followed by the smoothing step, (ii) evaluation of $\mathcal{L}_{KL}$ for the convergence check; this may not be required for every single iteration and can be included after every few iterations, (iii) $q\left(\nu\right)$ update, and (iv) σ update. This sequence requires only the initialization of σ and $\beta_t^d\,\forall\, t,d$. We preform grid search for the initial values that maximize $\mathcal{L}_{KL}$ to avoid convergence to the local optima. To reduce the complexity, the initial values of $\beta_t^d\,\forall\, t,d$ are equated to a constant, κ. We utilize normalized data for our study, therefore, the choices of σ and κ within the interval of $[0,1]$ are reasonable as the actual variance of the output is equal to 1. We searched for the values of σ and κ with a grid interval of 0.1.

Depending on the choice of the hyper-parameters α^* and β^*, the prior defined by H_0 penalizes the changes in parameters. For the change in parameter from θ_{t-1}^d to θ_t^d, it adds the penalty $\frac{\alpha}{\beta_t^d}$, which is given by,

$$\frac{\alpha}{\beta_t^d}=\frac{\alpha^*+\frac{1}{2}}{\beta^*+\frac{1}{2}E\left(\left(\theta_t^d-\theta_{t-1}^d\right)^2\right)} \tag{6.22}$$

The penalty depends on the expected value of the square of change is parameter. The effect of choice of the hyper parameters on the penalty added to the parameter changes are illustrated in Fig. 6.1. The plots in Fig. 6.1 illustrate this for two cases, one for

decreasing β^* on the left panel and the other for increasing α^* on the right panel. Y-axis in the plots corresponds to the penalty term added and the x-axis corresponds to the expected value of the square of change in parameter. For a fixed value of α^*, when β^* is decreased, as shown by the direction of the dashed arrow, the penalty on the smaller valued changes increases rapidly and the penalty on the larger valued changes does not increase appreciably. This effect regularizes the smaller valued changes significantly and forces them to converge to zero and leaves the higher valued changes relatively unaffected. The scale of changes that one wants to regularize can be dictated by the choice of α^*. The scale remains proportional to α^*. If we increase α^*, then the larger valued changes will be penalized. This can be observed from the panel on the right. As we increase α^* with a fixed value of β^* we can see that the penalty increases even for the larger valued changes. Keeping β^* to a lower value and changing α^* incrementally would allow one to analyse the relative importance of the changes in parameter. At low values of α^*, lower valued changes converge close to zero, as we increase α^* more and more changes start to vanish.

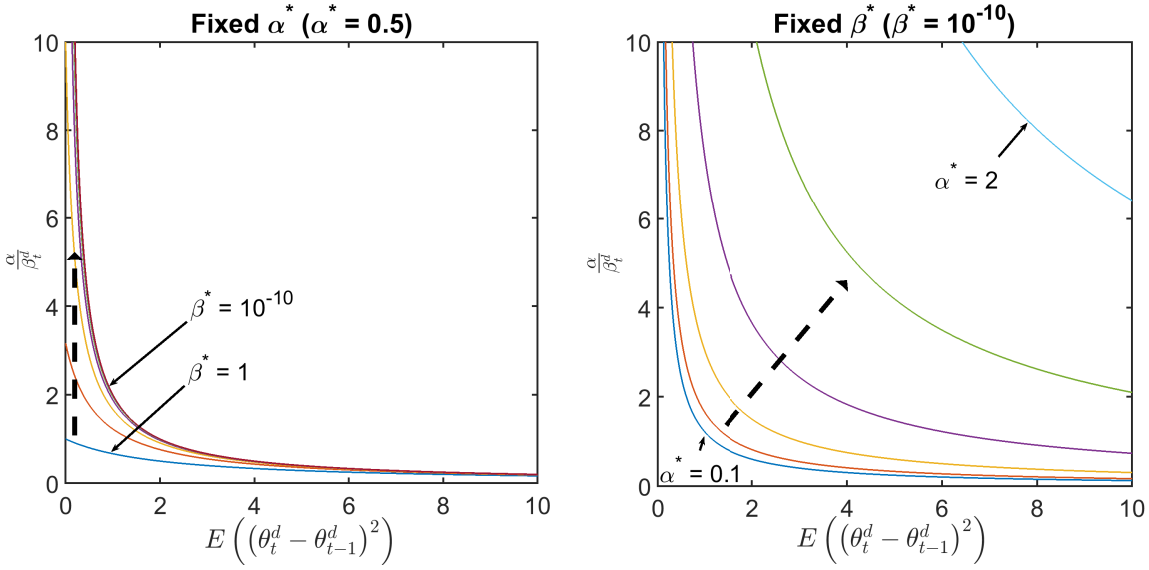


Figure 6.1: Penalty added to the changes in parameters: Left: for increasing values of β^* and Right: for increasing values of α^*

We fix β^* to low values such as 10^{-8} and we vary α^* and analyse the estimated values of the parameters. The hyper-parameter α^* can be selected through cross-validation, however, cross-validation for the time-varying systems is not feasible as

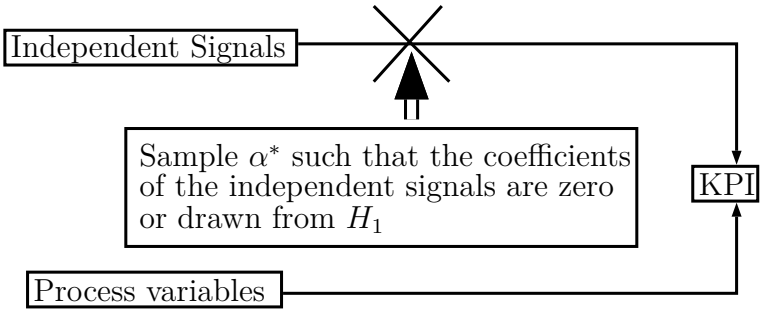


Figure 6.2: Hyper-parameter, α^* tuning strategy

the training and the test data sets will no longer be drawn from the same population. Therefore, we follow a different approach for selecting the range for α^*. We introduce new signals that are independent of the outputs and inputs of the system as additional inputs to the model. As we know *a priori* that the outputs are not dependent on the newly introduced signals, the coefficients of those signals should be zero, i.e, the coefficients of those signals should be drawn from H_1 at all t. We set the lowest value of sampled α^* such that the preferred choice of the prior distributions for those coefficients is H_1 at all t. This approach is illustrated in Fig. 6.2. In our simulations we include the following signals as the additional inputs, (i) white noise, (ii) random binary sequence and (iii) sum of sine waves. The maximum value for sampled α^* is constrained such that the number of effective parameters in the model is at least equal to the number of relevant inputs. The number of effective parameters are given by the number of parameters with H_0 as the preferred prior over H_1.

6.5 Application

In this section, we present an industrial case study. We provide the causal modelling results of multiple SAGD wells obtained using the proposed approach. The objective of this case study is to assess the causal influence of the reservoir parameters (inputs) on the production rates from the SAGD wells (output). We develop the hypotheses for the causal relations and the signs of causal strengths (positive or negative) based on Darcy's law. We compare the outcomes of the causal modelling exercise against the hypothesized causal relations. Further, to illustrate the importance of utilizing the TVPMs, we compare the outcomes against the results obtained using the time-

invariant multivariate linear regression based path models. The discussion in this section is organized in the following order, (A) at first, we present a description of the SAGD system and develop the hypotheses for the potential causal relations, (B) next, we provide the details of the data utilized in this study, and (C) in the end, we present the summary of the causal modelling results.

6.5.1 Steam Assisted Gravity Drainage Wells

A simplified schematic of a SAGD well pair is shown in Fig. 6.3. In SAGD, two parallel wells, one above the other, are drilled horizontally into the oil pay zone. Steam is injected continuously through the top well, which is called the injector well. Over time, a steam chamber develops around the injector well and it lowers the viscosity of the bitumen in the oil sands deposit. In addition to steam injection, residue gas is also injected through the injector well to help control the steam chamber pressure. Residue gas injection pressure is typically used as a proxy measure for the steam chamber pressure downhole by the operation. The produced oil and condensed steam form an available emulsion inventory that is mobilized and moved to the surface by the producer well. The accumulated emulsion is mobilized by means of an artificial lift system, quite often an electric submersible pump (ESP). The produced fluids feed a production separator where an initial gas-oil-water gravity separation occurs. To measure the production rate of each well in addition to the water cut, the emulsion stream from each well is diverted to a separator on regular periodic basis. Test separators provide volumetric measurements for each of the produced fluids including gas, water and oil for each well at the well pad.

Causal modelling of the SAGD system provides the following benefits, (i) it provides estimates of the relative causal strengths of the input reservoir parameters allowing production engineers to set production and operating strategies that will maximize the production from their field, (ii) helps in assessing the importance of closed loop production control strategies and (iii) in the cases of production anomalies (such as an anomalous decrease or increase in production), causal modelling can help determine the root cause, so that adjustments can be made accordingly.

There are two primary methods to maximize the production from the SAGD wells

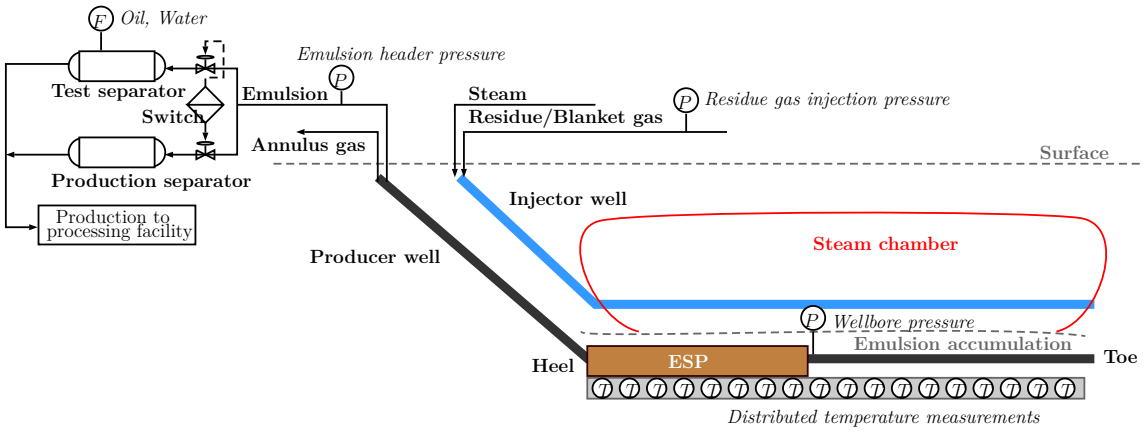


Figure 6.3: Schematic of a SAGD well pair

under normal operating conditions, (i) maximising the steam chamber pressure (P_R) while respecting cap-rock pressure constraints. This can be achieved by maximizing the steam and residue gas injection rates and (ii) optimizing the emulsion level downhole (sometimes referred to as subcool optimization). By effectively controlling the emulsion level slightly above the producer well, a larger reservoir area is exposed to the latent heat available in the steam, thereby mobilizing additional bitumen and maximizing production. Allowed to operate too high, and the energy in the steam is absorbed by emulsion layer rather than the reservoir itself causing suboptimal production rates. Operated too low and steam/gas break-through will occur pushing high temperature vapours and solids at high velocities into the producer well bore. These abnormal conditions are a known cause of pump and liner damage and is typically avoided by most producers.

The production from a SAGD well can be modelled using a simple form of Darcy's law shown below,

$$Q = \frac{kA\,(P_R - P_W)}{\mu L} \tag{6.23}$$

where Q is the oil inflow to the producer well bore, P_R is the reservoir or steam chamber pressure, P_W is the well bore pressure, μ is the viscosity of the emulsion, and k, A and L are the permeability of the reservoir, cross sectional area available to the flow and the length of the well bore, respectively. Based on Darcy's law, we postulate the effect of steam chamber pressure (P_R) and the well bore subcool (WS)

on the production rates (F) from SAGD well as presented below,

1. Effect of steam chamber pressure (P_R): Increasing the steam chamber pressure, increases the pressure gradient, $P_R - P_W$. As per Darcy's law, this increase in pressure gradient, increases the inflow (Q) from the chamber to the producer well bore. Therefore, increasing P_R leads to increased production, F. The upper limit for P_R is dictated by the pressure at which the cap-rock fractures. Increasing P_R, also increases the saturation temperature of steam at the steam chamber (T_S). This leads to the increase in the amount of heat transferred to bitumen in the reservoir, thus lowering its viscosity and increasing its mobility. Therefore, increasing P_R increases the condensate and bitumen inflow to the producer well bore and in turn, favours increased F. Increasing T_S also causes temperature at the producer well bore (T_W) to increase due heat transferred from the chamber to the emulsion at the well bore, lowering the viscosity and increasing the mobility of emulsion above the well bore.

2. Effect of well bore subcool (WS): Well bore subcool is defined as the difference between the saturation temperature of the steam ($T_{sat}(P_W)$) corresponding to the well bore or pump intake pressure (P_W) and the temperature of the emulsion at the producer well bore (T_W), $WS = T_{sat}(P_W) - T_W$. WS is used as a proxy for the emulsion level downhole as sensor technology to directly measure the emulsion level in SAGD wells does not yet exist. Practically speaking, a positive value for WS indicates that there is an available emulsion inventory downhole to be pumped. The greater the number the higher the emulsion level is above the producer well bore. WS can be lowered either by lowering P_W or by increasing T_W. Lowering P_W increases $P_R - P_W$ and in turn, favours higher inflow. Increasing T_W, lowers the viscosity of the emulsion accumulated in the well bore and increases its mobility through the well bore. Therefore, lowering WS leads to increased F. The lower limit for WS is dictated by the steam breakthrough limit. Steam breakthrough occurs when lowering WS below zero, allowing steam and produced vapour to flow into the producer well bore. Pushing WS to zero takes T_W closer to T_{sat} and exposes the liners and well bore to

the steam chamber.

Based on the above discussion, we postulated the causal model presented in Fig. 6.4. The graph in Fig. 6.4. (i) shows that WS has a negative effect (a_1) on F, corresponding to the hypothesis that lowering well bore subcool favours production. It also shows that P_R has a negative effect on WS (a_2) and positive effect on F (a_2). Increasing P_R increases TS, which in turn causes T_W to increase. Increasing T_W lowers WS. Therefore, P_R has a negative effect on WS. In addition to P_R, well bore pump speed, back pressure from the production header can also affect WS. P_R is affected by different factors including steam injection rates, steam quality and residue gas injection rates.

The total effects of WS and P_R on F can be quantified from the effects a_1, a_2 and a_3 as shown in Fig. 6.4. (ii). The total effect of WS is given by a_1 and it is hypothesised to have the negative sign and the total effect of P_R is given by $a_2 + a_1 \times a_3$ and it is hypothesised to have the positive sign. For the causal modelling exercise, we build a predictive model for WS from P_R and the other factors affecting WS and estimate a_1 and build a predictive model for F from P_R and WS and estimate a_2 and a_3. The predictive models are either based on TVPMs or based on time-invariant multivariate linear regression models. From the estimated effects a_1, a_2 and a_3, we study the total effects of WS and P_R on F.

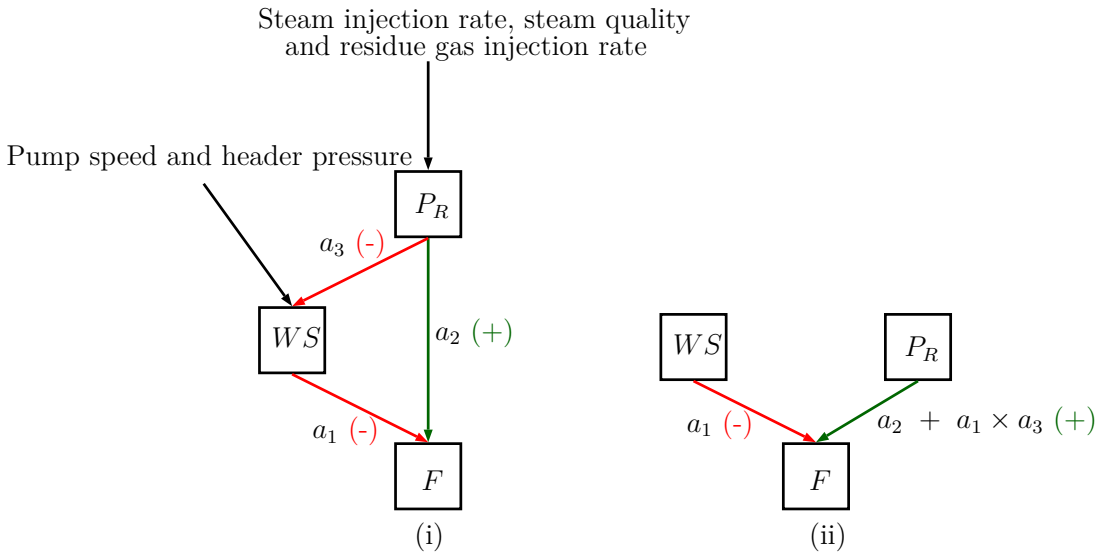


Figure 6.4: (i): Postulated graphical model among production rate, well bore subcool and steam chamber pressure and (ii): Total effect of well bore subcool and steam chamber pressure on production rate. Green arrows correspond to positive effect and red arrows correspond to negative effect.

6.5.2 Data Description

Fig. 6.5 shows the time series trends of the data for one of the case studies, well 1. The green, blue, cyan and magenta trends in the figure correspond to P_R, P_W, T_S and WS, respectively. The red trend correspond to F and the blue binary trend corresponds to the state of production. Periods when the binary trend goes to zero corresponds to the periods of well shutdown. Markers in the trends correspond to the instants when the test separator measurements are available. Each time the well is tested, the test separator's live production data are available for periods of 2 to 6 hours. The available data was averaged over those periods to calculate F. Data for the input variables were also obtained by averaging their measurements over those periods. Well bore temperatures are available at multiple locations along the well bore via fibre optice based distributed temperature sensors as shown in Fig. 6.3; however, as one would expect, they are highly correlated. Therefore, we utilized the median of these temperature measurements to calculate WS.

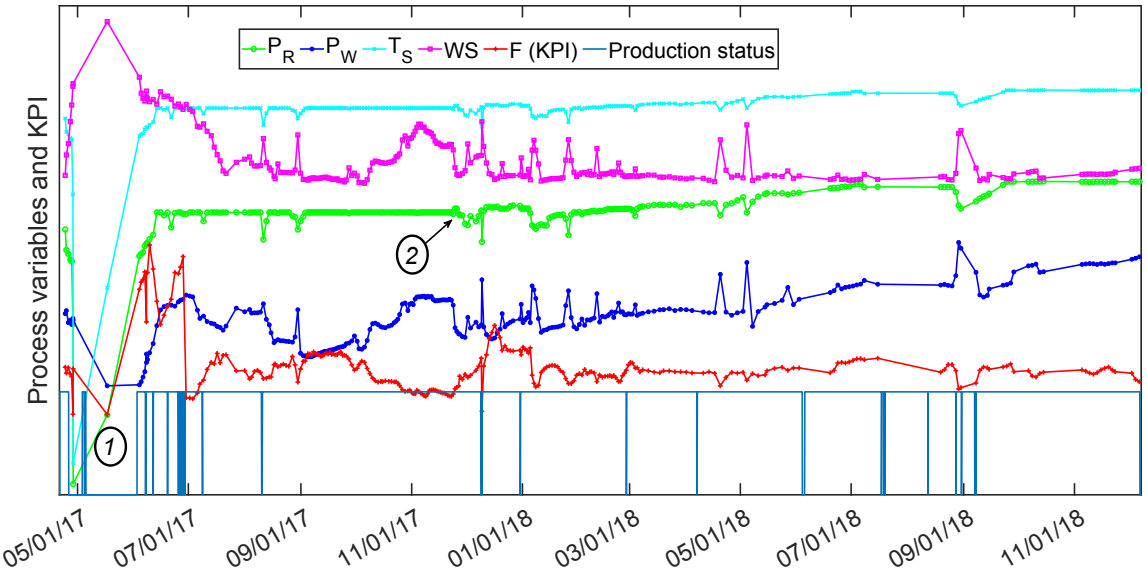


Figure 6.5: Well 1: Time trends of the process variables and KPI.

6.5.3 Results

As described in the procedure illustrated in section IV, we vary α^* and estimate the effects of P_R and WS with respect to their changing values. Additionally, with the use of TVPMs, the estimated effects may also vary with time. Therefore, we analyse the estimated effects under two different conditions in this section, (i) at different time instants, we assess the spread of the estimated effects that occurs due to changes in α^* and (ii) at different levels of α^*, we assess the spread of the estimated effects that occurs due to the time-varying nature of the causal coefficients.

Fig. 6.6 presents the spread of the total effect of well bore subcool on the production rates at different time instants obtained from the data presented in Fig. 6.5 using the TVPM based approach. At each time instant, the figure includes a box plot representing the spread of the estimated effect of the parameter of interest on production. As we vary the penalizing parameter α^*, with respect to each α^*, we obtain an estimate of the effect. The box plots essentially show the spread of these estimates due to changes in α^*. The box plots can be seen more clearly in the panel inside Fig. 6.6, which is a zoomed-in version of the plot from the 17th of August to 30th of August, 2017. The bottom boundary of the blue box corresponds to the lower quartile of the estimated effects and the top boundary of the blue box corresponds

to the upper quartile of the estimated effects. The red line in the middle of the box corresponds to the median of the estimated effects. The red dots that fall outside the boxes correspond to outliers in the estimates. A similar plot for the total effect of the steam chamber pressure on the production rates is shown in Fig. 6.7.

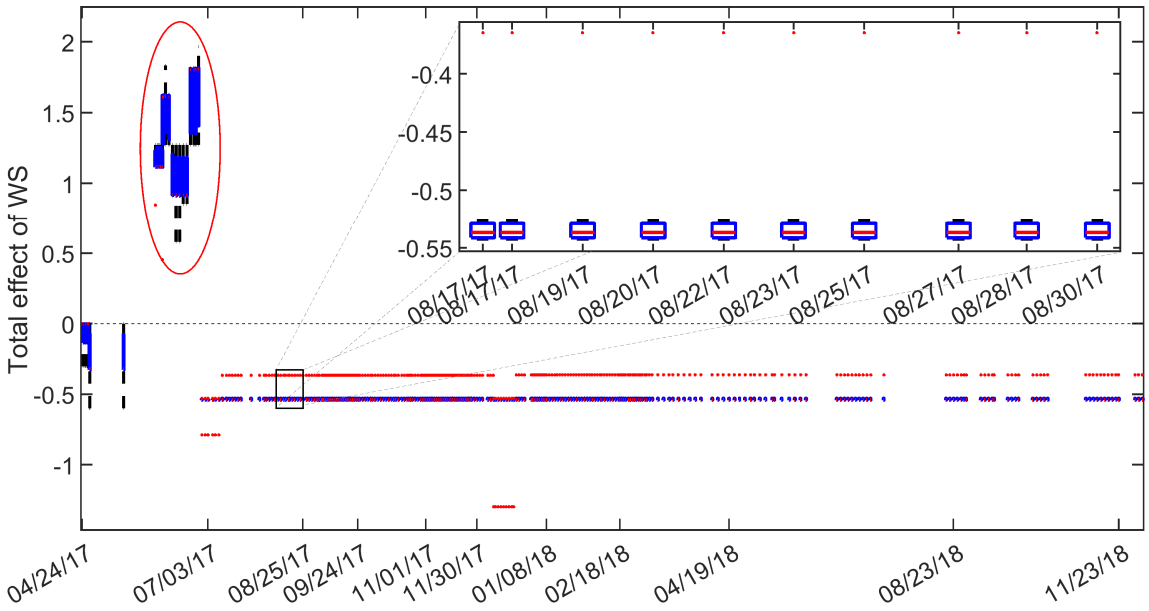


Figure 6.6: Well 1: Spread of the estimated total effect of well bore subcool on the production rates at different time instants. α^* is varied from 0.4 to 1.6.

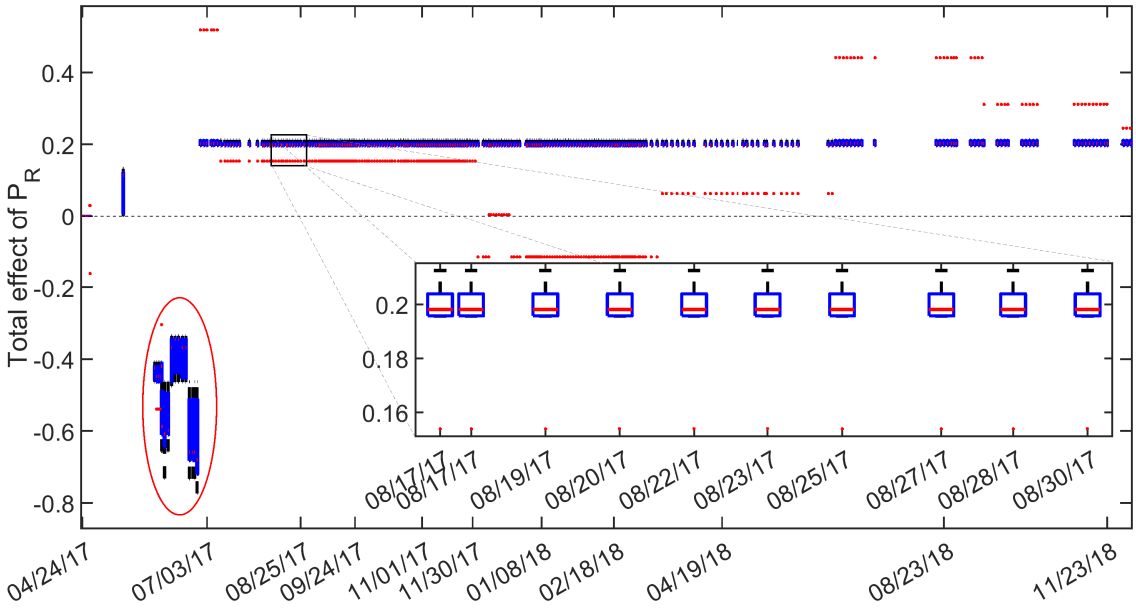


Figure 6.7: Well 1: Spread of the estimated total effect of steam chamber pressure on the production rates at different time instants. α^* is varied from 0.4 to 1.6.

From figures 6.6 and 6.7, it can be seen that during most time instants, the spreads of the total effect of WS falls on the negative side of the axis and the spreads of the total effect of P_R falls on the positive side of the axis. This is consistent with our initial belief shown in Fig. 6.4. However, during the period just before the 3rd of July, the spreads of the effect of WS falls on the positive side of the axis and the spreads of the effect of P_R falls on the negative side of the axis as indicated by the red ellipses. In Fig. 6.5, it can be seen that the well goes through a long production shutdown in the month of May and June as marked and indicated as 1. During this shutdown, an excessive amount of emulsion was accumulated down hole. As production was re-started, this inventory had to be pumped off prior to reaching stable operation. This can be seen from the production rate trend in Fig. 6.5, where the production rate during the end of June, 2017 is almost double in comparison to the production rates during remaining active production periods. Note that this abnormally high rate of production altered the correlation among the F, WS and P_R during that period. This was picked up by the TVPM based approach in terms of changes in signs of the estimated effects. Other than the described anomaly, during the normal operating conditions, the median and the spread of the effect remained consistent with our hypotheses. The patterns in the outliers can be seen to vary often after the 30th of November, 2017 in the effect of P_R on F even though the median and the spread remain the same during that period as can be seen in Fig. 6.7. From the 30th of November 2017, the production team had decided to ramp up the steam chamber pressure which was maintained almost constant until then as marked as 2 and indicated in Fig. 6.5. This ramp-up induced dynamics in the operation to which, the estimates under low values of α^* were sensitive. These outliers are the effects estimated at low values of α^*.

Fig. 6.8 presents the spreads of total effects of WS and P_R on F estimated at different levels of α^*. This plot is constructed by collecting the estimated effects at a fixed value of α^* (Ex. $\alpha^* = 0.4$) for the entire period during which the data are available and representing the spread of these estimates using a box plot. It is a better representation to evaluate the long-term effect of an input parameter on the KPI. Fig. 6.8 shows the spreads of the effects at seven different values of α^*. Except

for few outliers, the spread is tightly centred around the median that we hardly see the boxes. If the spread is tightly centred around the median, then the effect remains constant for long periods of time, only at some time instants it changes, which is captured by the outliers. If the spread is represented by a fat box as in the case of effect of P_R on F at $\alpha^* = 0.4$, then the estimated effect takes a value for some period and another value for some other period and so on.

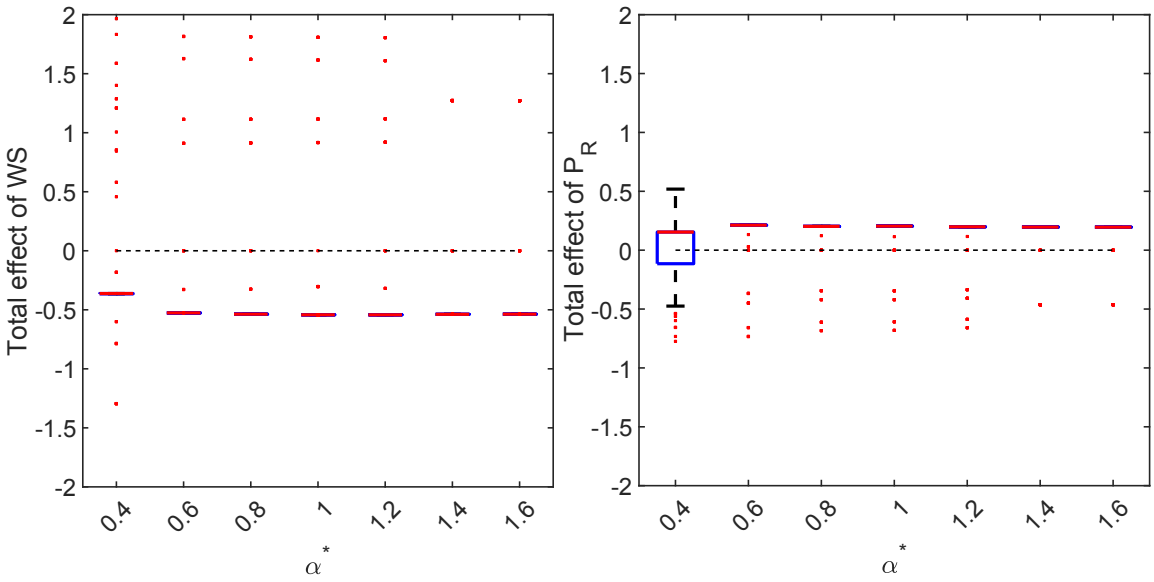


Figure 6.8: Well 1: Spreads of the estimated total effects of well bore subcool and steam chamber pressure on production at different values of α^*.

These box plots in Fig. 6.8 clearly suggests that irrespective of α^* values, the median and the spread of the effect of WS fall on the negative side of the axis and of the effect of P_R, they fall on the positive side of the axis. Figures E.1 to E.6 in the appendix, present this result for 6 additional SAGD wells. In all of the presented results, this observation is consistent, which goes on to validate our hypotheses that low subcool and high steam chamber pressure operations favour increased production. This finding is true on a long-term basis as the study included data from seven different wells for a period of nearly 18 months.

From the medians of the effects obtained from Fig. 6.8 and Figures E.1 to E.6 in

the appendix, we quantify the relative effect of WS and P_R on F as the following,

$$Relative\ effect\ of\ WS(\%) = \frac{|median\ of\ effect\ of\ WS| \times 100\%}{|median\ of\ effect\ of\ WS| - |median\ of\ effect\ of\ P_R|} \tag{6.24}$$

$$Relative\ effect\ of\ P_R(\%) = 100 - Relative\ effect\ of\ WS(\%) \tag{6.25}$$

The estimated relative effects are summarized in Table 6.1. We observe extreme values for the relative effects in well 3 and well 5. In well 3, the relative effect of WS reaches 100% and in well 5, the relative effect of P_R reaches 94%. Otherwise, the relative effects of WS and P_R range anywhere between 20% and 80%. Therefore, on an average, it is safe to conclude that both subcool and steam chamber pressure play important roles in influencing the production in SAGD operations. However as illustrated in Table 6.1, in each well, one parameter may influence production more than the other.

Table 6.1: The relative effects of well bore subcool and steam chamber pressure on production identified at different values of α^*

Well 1	α^*	0.4	0.6	0.8	1.2	1.4	1.6
	Relative effect of WS (%)	70.17	71.23	72.62	73.26	73.28	73.3
	Relative effect of P_R (%)	29.83	28.77	27.38	26.74	26.72	26.7
Well 2	α^*	0.4	0.5	0.7	0.9	1	1.1
	Relative effect of WS (%)	51.92	48.31	21.84	41.91	41.89	41.86
	Relative effect of P_R (%)	48.08	51.69	78.16	58.09	58.11	58.14
Well 3	α^*	0.3	0.5	0.7	0.9	1.1	1.3
	Relative effect of WS (%)	88.53	88.36	100	100	100	100
	Relative effect of P_R (%)	11.47	11.64	0	0	0	0
Well 4	α^*	0.2	0.25	0.3	0.35	0.4	0.45
	Relative effect of WS (%)	61.76	61.76	61.76	61.76	61.76	61.76
	Relative effect of P_R (%)	38.24	38.24	38.24	38.24	38.24	38.24
Well 5	α^*	0.32	0.33	0.34	0.35	0.36	0.37
	Relative effect of WS (%)	6.1	6.1	6.1	6.1	6.1	6.2
	Relative effect of P_R (%)	93.9	93.9	93.9	93.9	93.9	93.8
Well 6	α^*	0.7	0.8	0.9	1.1	1.2	1.3
	Relative effect of WS (%)	65.14	56.0	55.16	23	23.11	23.16
	Relative effect of P_R (%)	34.86	44.0	44.84	77	76.89	76.84
Well 7	α^*	0.48	0.54	0.6	0.66	0.72	0.78
	Relative effect of WS (%)	66.18	70.04	78.86	78.78	78.72	78.06
	Relative effect of P_R (%)	33.82	29.96	21.14	21.22	21.28	21.94

In the rest of this section, we argue why the TVPM based approach is better than a time-invariant model based approach for the case of analysing the production data from SAGD wells. Table 6.2 presents the standard deviation of the prediction error for TVPM based prediction model for F and time-invariant linear regression model for F. For lower values of α^*, we expect to see greater numbers of parameter changes in TVPM and therefore, the model should have more effective parameters. As we increase α^*, the frequency of parameter changes should decrease and as should the total number of effective parameters in the model. Standard deviation of the prediction error also increases with as α^* increases. At $\alpha^* = 1.6$, the model only has three effective parameters and the standard deviation of the prediction error is only

0.48. On the other hand, a time-invariant model of similar complexity with three parameters (coefficients of WS and F and the intercept parameter) has the standard deviation of the prediction error as 0.96, which is twice that generated by a TVPM of similar complexity. Therefore, TVPM clearly explains the data better than the time-invariant model.

Table 6.2: Well 1: Variability in production unexplained by the TVPM with different values of α^* and by the time-invariant linear regression model

	Time-varying model							Time-invariant model
α^*	0.4	0.6	0.8	1	1.2	1.4	1.6	-
Number of effective parameters	33	9	9	9	9	3	3	3
Standard deviation of prediction error	0.124	0.287	0.289	0.292	0.293	0.482	0.482	0.964

Fig. 6.9. (i) provides a graphical representation of the total effects based on the median values presented in Fig. 6.8 at an α^* value. Fig. 6.9. (ii) provides a graphical representation of the total effects estimated based on a time-invariant linear regression model. The graphs include the magnitudes of the total effects and their signs. Red arrow represents negative effect and green arrow represents positive effect. The thickness of the arrows in proportional to the magnitude of the effect. The graph obtained from the TVPM based approach in Fig. 6.9. (i) is consistent with the theoretical understanding of the process that low subcool and high steam chamber pressure favour production. The graph obtained from the time-invariant model in Fig. 6.9. (ii) is totally opposite to the theoretical understanding of the process. Fig. E.7 compares both approaches for six more wells. In all the wells, the sign of effects are consistent when inferred using the TVPM based approach, whereas the time-invariant model provides highly inconsistent results.

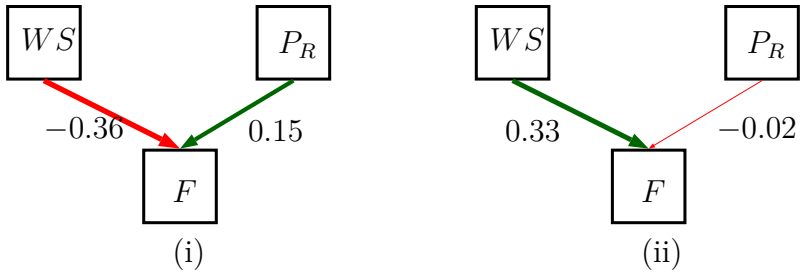


Figure 6.9: (i) Well 1: Median of total effects identified using the TVPM based approach with $\alpha^* = 0.4$ and (ii) total effect identified using the time-invariant linear regression models estimated under the ML approach.

6.6 Summary

In this chapter, we presented a causal modelling approach based on the VBEM framework for the time-varying systems. The data from the time-varying systems were modelled using the time-varying parameters models (TVPMs). The TVPMs were estimated under the VBEM framework. Followed by the VBEM estimation, a hypothesis switching procedure was utilized to infer the actual changes in the causal strengths. The whole approach was validated using the production data from seven SAGD wells. The approach provided theoretically consistent causal models in all the seven wells. The results obtained using the production data verified the following theoretical understandings of the SAGD wells, (i) increasing steam chamber pressure and (ii) lowering well bore subcool favour increased production. Comparisons against the time-invariant models revealed the importance of using the TVPMs for causal analysis of the time-varying systems. Time-invariant models provided inconsistent results with the understanding of the process as well as provided inconsistent results across seven SAGD wells.

Chapter 7

Conclusions and Recommendations

7.1 Conclusions

In this thesis, we developed and presented probabilistic models for data-driven process monitoring and causal modelling applications. The key findings of this thesis can be summarized as summarized below,

- In chapter 3, we showed that a generalized model can be defined such that it encompasses most of the linear Gaussian models used for process monitoring applications in the literature. This greatly simplified the effort required to derive the process monitoring procedure based on linear Gaussian models. The monitoring procedure was derived based on the generalized model and it was then shown to reduce to special cases when the model structure was constrained appropriately. Classical multivariate techniques such as principal component analysis and canonical correlation analysis can also be formulated as probabilistic models, which can be seen as special cases of the generalized model defined. By constraining the generalized model to these special cases, we showed that the resulting monitoring statistics of the probabilistic models will be exactly equivalent to the monitoring statistics derived from their classical counterparts. We verified the theoretical results using simulation examples. As a part of this exercise, we flagged some of the common misconceptions in the literature regarding the monitoring statistics and control charts derived based on the probabilistic models.

- In chapter 4, we discussed how a mixture model formed by convex combination

of linear Gaussian models is used for process monitoring. We showed stacking the mixture models one above the other can lead to a parameter efficient two-layer model. At a given moment, the two-layer modelling approach involved identification of lesser number of local models when compared to a single layer model of a similar complexity. This allowed us to provide better initial guesses for the model parameters and the identification converged to better results. During the model identification stage, we also leveraged a Bayesian regularization approach for model structure determination. We illustrated the applicability of the two-layer model in process monitoring applications using a lab-scale and an industrial case study. The industrial case study was on monitoring of a sulphur recovery units to predict downstream sulphur dioxide breakthrough problems. The two-layer model scaled well to approximate the data distributions in our case studies when compared to the single layer mixture models. It also had a better generalization ability when compared to the single layer model.

- In chapter 5, we presented a hybrid model that is formed by a combination of the vector auto-regressive model and the probabilistic factor analyser model. The model was used to represent the potential time-lagged and contemporaneous interactions among the variables in a linear system. The time-lagged causal interactions were represented by means of the vector auto-regressive model component and the contemporaneous correlations were represented by means of the probabilistic factor analyser model. We performed approximate Bayesian analysis by assigning normal-gamma prior distributions to the model parameters and using the variational Bayesian expectation maximization algorithm. In the linear case, the model parameters and causal and contemporaneous interactions have one to one correspondence. Therefore, determining the zero and non-zero parameters through the Bayesian analysis was useful in commenting the possible presence and absence of the causal and contemporaneous interactions. In the simulation case studies, the overall approach was found to be more robust to the presence of contemporaneous correlations when attempting to identify the time-lagged causal interactions as compared to the traditional techniques for identifying the time-lagged causal interactions. We also illustrated this ap-

proach by studying the data from the sulphur recovery unit.

- In chapter 6, we presented a time-varying parameters model. The model was used to study the interactions among the variables in a postulated casual network. We utilized approximate Bayesian analysis through the variational Bayesian expectation maximization approach to track the time varying parameters in the model. The approach can be utilized to study the causal interactions among the variables in time-varying systems. For linear systems, the parameters of the model has one to one correspondence to the strength of direct causal effects. We illustrated the whole approach in the production from multiple steam assisted gravity drainage wells. We postulated how the variables in the system affect the product rates and estimated the effects using the proposed approach. The time-invariant linear models were found to give inconsistent results across case studies. However, the proposed approach was found to produce consistent results for the signs (positive and negative) of causal coefficients across case studies.

7.2 Recommendations

7.2.1 Process Monitoring

The success of data-driven process monitoring applications rely on how much information about the desired operation characteristics can be learnt from the data using the data-driven models used. This thesis focused on one particular aspect, learning the distribution of the process variables from the operational data and determining the statistical bounds from the learnt distribution models. As illustrated in chapter 4 of this thesis, some applications may need approximating multi-modal or complex data distributions. To pursue in this direction of modelling the data distributions using the probabilistic models, we can extend the two-layer model presented in chapter 4 to a multi-layer model. Then, the model will become more powerful in approximating complex data distributions.

Learning the data distribution is one approach to characterizing the data generation process. The recent developments in the field of machine learning, particu-

larly in deep learning, has lead to more powerful methods of characterizing the data generation process. Two such promising methods that can be explored for process monitoring applications are variational auto encoders (VAEs) [116] and generative adversarial networks (GANs) [117]. However, VAEs, GANs and the multi-layer model discussed earlier are data hungry. These models may need large amounts of data to achieve better generalizing ability. Process industries do possess huge repositories of historical data. The questions, "how and how much the research communities can tap into those repositories?" will determine the level of success that we can achieve with modern machine learning approaches in process monitoring problems.

7.2.2 Causal Modelling

When applying causal modelling techniques to routine operation data as opposed to applying them to study experimental data, the users have to exercise caution. The routine operation data may contain multiple data quality issues such as slow sampling rates, outliers, data from biased or failed sensors, unobserved confounding variables, time-varying characteristics, data from multi-modal operations, etc. These data quality issues may be handled effectively by incorporating appropriate modelling assumptions. However, the assumptions may also lead to the scenarios where the users let their subjective belief or observational bias to significantly influence the identification results. Addressing this ambiguity will be an interesting pursuit.

Bibliography

[1] L. H. Chiang, E. L. Russell, and R. D. Braatz. *Fault detection and diagnosis in industrial systems*. Springer-Verlag London, 2000.

[2] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin. A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers & chemical engineering*, 27(3):327–346, 2003.

[3] J. S. Qin. Statistical process monitoring: Basics and beyond. *Journal of chemometrics*, 17(8-9):480–502, 2003.

[4] J. F. MacGregor and T. Kourti. Statistical process control of multivariate processes. *Control Engineering Practice*, 3(3):403–414, 1995.

[5] B. M. Wise and N. B. Gallagher. The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, 6(6):329–348, 1996.

[6] Z. Ge, Z. Song, and F. Gao. Review of recent research on data-based process monitoring. *Industrial & Engineering Chemistry Research*, 52(10):3543–3562, 2013.

[7] W. A. Shewhart. *Economic control of quality of manufactured product*. Van Nostrand, Princeton, NJ, 1931.

[8] S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959.

[9] J. S. Hunter. The exponentially weighted moving average. *Journal of quality technology*, 18(4):203–210, 1986.

[10] R. H. Woodward and P. L. Goldsmith. *Cumulative sum techniques*. Oliver and Boyd, London, 1964.

[11] F. B. Alt. *Economic design of control charts for correlated, multivariate observations*. PhD thesis, Georgia Institute of Technology, 1977.

[12] H. Hotelling. *Multivariate quality control, illustrated by the air testing of sample bombsights, in: C. Eisenhart, M. Hastay, W. Wallis (Eds.), Techniques of Statistical Analysis*. Mc Graw, New York, 1947.

[13] J. D. Healy. A note on multivariate CUSUM procedures. *Technometrics*, 29(4):409–412, 1987.

[14] R. B. Crosier. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30(3):291–303, 1988.

[15] J. J. Pignatiello Jr and G. C. Runger. Comparisons of multivariate CUSUM charts. *Journal of quality technology*, 22(3):173–186, 1990.

[16] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34(1):46–53, 1992.

[17] G. Li, S. J. Qin, and D. Zhou. Geometric properties of partial least squares for process monitoring. *Automatica*, 46(1):204–210, 2010.

[18] A. Negiz and A. Çlinar. Statistical monitoring of multivariable dynamic processes with state-space models. *AIChE Journal*, 43(8):2002–2020, 1997.

[19] E. L. Russell, L. H. Chiang, and R. D. Braatz. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 51(1):81–93, 2000.

[20] C. Xia, J. Howell, and N. F. Thornhill. Detecting and isolating multiple plant-wide oscillations via spectral independent component analysis. *Automatica*, 41(12):2067–2075, 2005.

[21] C. F. Alcala and S. J. Qin. Reconstruction-based contribution for process monitoring. *Automatica*, 45(7):1593–1600, 2009.

[22] L. H. Chiang, E. L. Russell, and R. D. Braatz. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics and intelligent laboratory systems*, 50(2):243–252, 2000.

[23] Q. Jiang, X. Yan, and B. Huang. Performance-driven distributed PCA process monitoring based on fault-relevant variable selection and Bayesian inference. *IEEE Transactions on Industrial Electronics*, 63(1):377–386, 2016.

[24] J. F. MacGregor and T. Kourti. Statistical process control of multivariate processes. *Control Engineering Practice*, 3(3):403–414, 1995.

[25] S. Yin, S. X. Ding, A. Haghani, H. Hao, and P. Zhang. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control*, 22(9):1567–1581, 2012.

[26] J. E. Jackson and G. S. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349, 1979.

[27] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley, 2003.

[28] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[29] F. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley, 2005.

[30] T. Chen, E. Martin, and G. Montague. Robust probabilistic PCA with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis*, 53(10):3706–3716, 2009.

[31] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[32] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 11(Jul):1957–2000, 2010.

[33] C. M. Bishop. Bayesian PCA. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 382–388. MIT; 1998, 1999.

[34] T. Chen and Y. Sun. Probabilistic contribution analysis for statistical process monitoring: A missing variable approach. *Control Engineering Practice*, 17(4):469–477, 2009.

[35] Z. Ge and Z. Song. Mixture Bayesian regularization method of PPCA for multimode process monitoring. *AIChE journal*, 56(11):2838–2849, 2010.

[36] Q. Jiang, B. Huang, and X. Yan. GMM and optimal principal components-based Bayesian method for multimode fault diagnosis. *Computers & Chemical Engineering*, 84:338–349, 2016.

[37] D. Kim and I. B. Lee. Process monitoring based on probabilistic PCA. *Chemometrics and intelligent laboratory systems*, 67(2):109–123, 2003.

[38] Z. Zhao, Q. Li, B. Huang, F. Liu, and Z. Ge. Process monitoring based on factor analysis: Probabilistic analysis of monitoring statistics in presence of both complete and incomplete measurements. *Chemometrics and Intelligent Laboratory Systems*, 142:18–27, 2015.

[39] R. Gonzalez, B. Huang, and E. Lau. Process monitoring using kernel density estimation and Bayesian networking with an industrial case study. *ISA transactions*, 58:330–347, 2015.

[40] J. Mori, V. Mahalec, and J. Yu. Identification of probabilistic graphical network model for root-cause diagnosis in industrial processes. *Computers & Chemical Engineering*, 71:171–209, 2014.

[41] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers, INC., San Francisco, CA, 1998.

[42] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

[43] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[44] D. M. Chickering. Learning Bayesian networks is NP-complete. In *Learning from data*, pages 121–130. Springer, 1996.

[45] R. E. Neapolitan. *Learning Bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[46] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[47] L. A. Baccalá and K. Sameshima. Partial directed coherence: A new concept in neural structure determination. *Biological Cybernetics*, 84(6):463–474, 2001.

[48] M. Kamiński. Determination of transmission patterns in multichannel data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1457):947–952, 2005.

[49] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, D. Klan, R. Bauer, J. Timmer, and H. Witte. Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Processing*, 85(11):2137–2160, 2005.

[50] L. Barnett and A. K. Seth. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *Journal of neuroscience methods*, 223:50–68, 2014.

[51] H. Lütkepohl. Stable Vector Autoregressive Processes. In *New introduction to multiple time series analysis*, pages 13–68. Springer-Verlag Berlin Heidelberg, 2005.

[52] F. Yang, P. Duan, S. L. Shah, and T. Chen. *Capturing connectivity and causality in complex industrial processes*. Springer Science & Business Media, 2014.

[53] T. Yuan and S. J. Qin. Root cause diagnosis of plant-wide oscillations using Granger causality. *Journal of Process Control*, 24(2):450–459, 2014.

[54] H. S. Chen, Z. Yan, Y. Yao, T. B. Huang, and Y. S. Wong. Systematic procedure for Granger-causality-based root cause diagnosis of chemical process faults. *Industrial & Engineering Chemistry Research*, 57(29):9500–9512, 2018.

[55] G. Deshpande, K. Sathian, and X. Hu. Assessing and compensating for zero-lag correlation effects in time-lagged Granger causality analysis of fMRI. *IEEE Transactions on Biomedical Engineering*, 57(6):1446–1456, 2010.

[56] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[57] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126, 1998.

[58] L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4:61–76, 1996.

[59] T. S. Jaakkola. *Variational methods for inference and estimation in graphical models*. PhD thesis, Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, 1997.

[60] M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, The Gatsby Computational Neuroscience Unit., University College London., London, 2003.

[61] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

[62] F. V. Jensen. *Introduction to Bayesian networks.* Springer-Verlag, New York, 1996.

[63] D. Heckerman. A tutorial on learning with Bayesian networks. In *Learning in graphical models*, pages 301–354. Springer, 1998.

[64] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer-Verlag, New York, 1999.

[65] M. I. (Ed.) Jordan. *Learning in graphical models*, volume 89. MIT Press, Cambridge, MA, 1999.

[66] D. J. C. MacKay. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469–505, 1995.

[67] M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, London, 2003.

[68] C. M. Bishop. *Chapter 10 in: Pattern recognition and machine Learning*. Springer-Verlag, New York, 2016.

[69] Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *NIPS'99 Proceedings of the 12th International Conference on Neural Information Processing Systems*, volume 12, pages 449–455, Denver, CO, 1999. MIT Press.

[70] Y. Lu, B. Huang, and S. Khatibisepehr. A variational Bayesian approach to robust identification of switched ARX models. *IEEE Transactions on Cybernetics*, 46(12):3195–3208, 2016.

[71] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor. Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals. *IEEE Transactions on Signal Processing*, 59(12):6257–6261, 2011.

[72] N. Nasios and A. G. Bors. Variational learning for Gaussian mixture models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4):849–862, 2006.

[73] D. A. Bodenham and N. M. Adams. A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, 26(4):917–928, 2016.

[74] D. R. Jensen and H. Solomon. A Gaussian approximation to the distribution of a definite quadratic form. *Journal of the American Statistical Association*, 67(340):898–902, 1972.

[75] S. Li, J. Gao, J. M. Nyagilo, and D. P. Dave. Probabilistic partial least square regression: A robust model for quantitative analysis of Raman spectroscopy data. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 526–531. IEEE, 2011.

[76] S. Li, J. Gao, J. O. Nyagilo, D. P. Dave, B. Zhang, and X. Wu. A unified probabilistic PLSR model for quantitative analysis of surface-enhanced Raman spectrum (SERS). In *The Proceedings of the Second International Conference on Communications, Signal Processing, and Systems*, pages 1095–1103. Springer, 2014.

[77] J. Zheng, Z. Song, and Z. Ge. Probabilistic learning of partial least squares regression model: Theory and industrial applications. *Chemometrics and Intelligent Laboratory Systems*, 158:80–90, 2016.

[78] T. D. Bie, N. Cristianini, and R. Rosipal. *Eigenproblems in pattern recognition*. Springer, Berlin, Heidelberg, 2005.

[79] J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4):419–426, 1961.

[80] P. Duchesne and P. L. De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

[81] W. Ku, R. H. Storer, and C. Georgakis. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and intelligent laboratory systems*, 30(1):179–196, 1995.

[82] J. Chen and K. Liu. On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chemical Engineering Science*, 57(1):63–75, 2002.

[83] J. Chen and J. Liu. Mixture principal component analysis models for process monitoring. *Industrial & engineering chemistry research*, 38(4):1478–1488, 1999.

[84] S. W. Choi, E. B. Martin, A. J. Morris, and I. B. Lee. Fault detection based on a maximum-likelihood principal component analysis (PCA) mixture. *Industrial & engineering chemistry research*, 44(7):2316–2327, 2005.

[85] R. Raveendran and B. Huang. Mixture probabilistic PCA for process monitoring collapsed variational Bayesian approach. In *11th IFAC Symposium on Dynamics and Control of Process SystemsIncluding Biosystems DYCOPS-CAB 2016*, volume 49, pages 1032–1037, Trondheim, Norway, 2016. Elsevier.

[86] S. W. Choi, J. H. Park, and I. B. Lee. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Computers & chemical engineering*, 28(8):1377–1387, 2004.

[87] J. Lee, C. Yoo, S. W. Choi, P. A. Vanrolleghem, and I. B. Lee. Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 59(1):223–234, 2004.

[88] S. W. Choi and I. B. Lee. Nonlinear dynamic process monitoring based on dynamic kernel PCA. *Chemical engineering science*, 59(24):5897–5908, 2004.

[89] Y. Tang, R. Salakhutdinov, and G. E. Hinton. Deep mixtures of factor analysers. In *Proceedings of the 29 th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012., 2012.

[90] T. Chen, J. Morris, and E. Martin. Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(5):699–715, 2006.

[91] C. Ruiz-Cárcel, Y. Cao, D. Mba, L. Lao, and R. T. Samuel. Statistical process monitoring of a multiphase flow facility. *Control Engineering Practice*, 42:74–88, 2015.

[92] E. Naghoosi and B. Huang. Interaction analysis of multivariate control systems under Bayesian framework. *IEEE Transactions on Control Systems Technology*, 25(5):1644–1655, 2017.

[93] S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fMRI. *Neuroimage*, 54(2):807–823, 2011.

[94] U. Triacca. Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature? *Theoretical and Applied Climatology*, 81(3-4):133–135, 2005.

[95] H. Kodamana, R. Raveendran, and B. Huang. Mixtures of probabilistic PCA with common structure latent bases for process monitoring. *IEEE Transactions on Control Systems Technology*, pages 1–9, 2018.

[96] R. Raveendran and B. Huang. Two layered mixture Bayesian probabilistic PCA for dynamic process monitoring. *Journal of Process Control*, 57:148–163, 2017.

[97] D. P. Filev, R. B. Chinnam, F. Tseng, and P. Baruah. An industrial strength novelty detection framework for autonomous equipment monitoring and diagnostics. *IEEE Transactions on Industrial Informatics*, 6(4):767–779, 2010.

[98] F. Souza and R. Araújo. Online mixture of univariate linear regression models for adaptive soft sensors. *IEEE Transactions on Industrial Informatics*, 10(2):937–945, 2014.

[99] W. Li, H. H. Yue, S. Valle-Cervantes, and S. J. Qin. Recursive PCA for adaptive process monitoring. *Journal of process control*, 10(5):471–486, 2000.

[100] C. W. J. Granger. Non-linear models: Where do we go next - Time varying parameter models? *Studies in Nonlinear Dynamics & Econometrics*, 12(3), 2008.

[101] R. Kalaba and L. Tesfatsion. Time-varying linear regression via flexible least squares. *Computers & Mathematics with Applications*, 17(8-9):1215–1245, 1989.

[102] N. Beck. Time-varying parameter regression models. *American Journal of Political Science*, pages 557–600, 1983.

[103] J. C. C. Chan, G. Koop, L. G. Roberto, and R. W. Strachan. Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3):358–367, 2012.

[104] J. J. J. Groen, R. Paap, and F. Ravazzolo. Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics*, 31(1):29–44, 2013.

[105] J. Nakajima and M. West. Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31(2):151–164, 2013.

[106] M. Kalli and J. E. Griffin. Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793, 2014.

[107] M. A. G. Belmonte, G. Koop, and D. Korobilis. Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33(1):80–94, 2014.

[108] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274, 2015.

[109] D. R. Kowal, D. S. Matteson, and D. Ruppert. Dynamic shrinkage processes. *arXiv preprint arXiv:1707.00763*, 2017.

[110] G. Koop and D. Korobilis. Variational Bayes inference in high-dimensional time-varying parameter models. *Working paper*, 2018.

[111] J. Hensman, M. Rattray, and N. D. Lawrence. Fast variational inference in the conjugate exponential family. In *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, pages 2888–2896, Lake Tahoe, Nevada, 2012.

[112] N. J. King and N. D. Lawrence. Fast variational inference for Gaussian process models through KL-correction. In *ECML'06 Proceedings of the 17th European conference on Machine Learning*, pages 270–281, Berlin, Germany, 2006. Springer-Verlag Berlin.

[113] A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(Oct):2879–2904, 2011.

[114] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[115] M. A. Gelbart, J. Snoek, and R. P. Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.

[116] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[117] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[118] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. *Neural computation*, 12(4):831–864, 2000.

Appendix A

Proofs of Propositions in Preliminaries

A.1 Proof of Proposition 1

Proof. The proof of the first three statements in proposition 1 can be achieved by the application of the chain rule of probability. We show the proof for the first statement and the proofs for the rest of the two can be achieved by following a similar procedure. The last statement in proposition 1 requires us to show that the statement cannot be disproved, which show followed by the proof of statement one.

Let us start with the proof of the first statement. The joint distribution of A, B and C from the structure of the BN shown in Fig. 2.3 (i) can be expressed as the following,

$$p\left(A,B,C\right)=p\left(C|B\right)p\left(B|A\right)p\left(A\right) \tag{A.1}$$

Using the chain rule probability, the conditional distribution of A given both B and C can be expressed as the following,

$$p\left(A|B,C\right)=\frac{p\left(A,B,C\right)}{p\left(B,C\right)} \tag{A.2}$$

where the numerator in the above expression corresponds to the joint distribution of A, B and C and the denominator corresponds to the joint distribution of B and C. The denominator term can be obtained by marginalizing A from the joint distribution of A, B and C as the following,

$$p\left(B,C\right)=\int_{A}p\left(A,B,C\right)dA=\int_{A}p\left(C|B\right)p\left(B|A\right)p\left(A\right)dA \tag{A.3}$$

where the integration with respect to A over the support of $p(A)$ corresponds to marginalization of A. Integration applies to the case when A is a continuous random variable and the integration has to replaced with the summation over the support of $p(A)$ when A is a discrete random variable. This marginalization leads to the joint distribution of B and C of the following form,

$$p\left(B,C\right)=p\left(C|B\right)p\left(B\right) \tag{A.4}$$

By substituting equations (A.2) and (A.4) in Eqn. (A.1), we can obtain the conditional distribution of A given both B and C as the following,

$$p\left(A|B,C\right)=\frac{p\left(C|B\right)p\left(B|A\right)p\left(A\right)}{p\left(C|B\right)p\left(B\right)} \tag{A.5}$$

From the above expression, we can cancel out the common terms in the denominator and the numerator. This cancels out the dependence on C and from the chain rule of probability we can obtain the following,

$$p\left(A|B,C\right)=\frac{p\left(B|A\right)p\left(A\right)}{p\left(B\right)}=p\left(A|B\right) \tag{A.6}$$

Therefore, given B, the dependence of A on C vanishes. By following a similar procedure to obtain the conditional distribution for C given both A and B, we can also show that given B, dependence of C on A vanishes. This completes the proof of the first statement. By following a similar procedure, the second and the third statements can also be proved.

For the fourth statement, we can start with the joint distribution defined by the BN and follow the same procedure as shown above to obtain the conditional distribution of A given both B and C as shown below,

$$p\left(A,B,C\right)=p\left(B|A,C\right)p\left(A\right)p\left(C\right) \tag{A.7}$$

$$p\left(A|B,C\right)=\frac{p\left(A,B,C\right)}{p\left(B,C\right)} \tag{A.8}$$

$$p\left(B,C\right)=\int_{A}p\left(A,B,C\right)dA=\int_{A}p\left(B|A,C\right)p\left(A\right)p\left(C\right)dA \tag{A.9}$$

$$p\left(B,C\right)=p\left(B|C\right)p\left(C\right) \tag{A.10}$$

$$p\left(A|B,C\right)=\frac{p\left(B|A,C\right)p\left(A\right)p\left(C\right)}{p\left(B|C\right)p\left(C\right)} \tag{A.11}$$

$$p\left(A|B,C\right)=\frac{p\left(B|A,C\right)p\left(A\right)}{p\left(B|C\right)} \tag{A.12}$$

From the exercise, we obtain the above expression for the conditional of A given both B and C. It can be seen from the above expression that the dependence of B on C both in the numerator and the denominator cannot be removed as C and B are directly connected. Therefore, in the conditional of A given both B and C, we cannot remove the dependence on C. We can show the same for the conditional of C given both A and B that the dependence of A cannot be removed by following the same procedure. This completes the proof of proposition 1. ■

A.2 Proof of Proposition 2

Proof. To prove this proposition, we consider an arbitrary BN with a subset of nodes V as the one shown in Fig. 2.4. Let us say that the set of all the nodes in the network except V is given by $\sim V$, the set of all the parent nodes of V is given by Pa, the set of all the children nodes of V is given by Ch, the set of all the other parent nodes of Ch is given by CPa (all the parents of Ch except V), and the set of all nodes that excludes just the subset V and the subset Ch is given by $\sim\{V,Ch\}$. In this setting, Proposition 2 translates into the following mathematical identity.

$$p\left(V|\sim V\right)=p\left(V|Ch,Pa,CPa\right) \tag{A.13}$$

where $p\left(V|\sim V\right)$ is the conditional distribution of V given the rest of the nodes in the network and it is equal to the conditional distribution of V given the set $\{Ch,Pa,CPa\}$. Given the set $\{Ch,Pa,CPa\}$, any information about the rest of the nodes in the network adds no valuable information for predicting or inferring the states of V.

Using the chain of probability, we can write the joint distribution of all the nodes in the network as the following,

$$p\left(V,\sim V\right)=p\left(V|\sim V\right)p\left(\sim V\right) \tag{A.14}$$

where $p\left(V,\sim V\right)$ is the joint distribution of all the nodes in the network and $p\left(\sim V\right)$ is the marginal distribution of the nodes $\sim V$. The marginal distribution $p\left(\sim V\right)$

can be obtained by integrating out V from the joint distribution $p(V, \sim V)$ as the following*,

$$p(\sim V) = \int_V p(V, \sim V)\, dV \tag{A.15}$$

By exploiting the structure of the BN, we can write the joint distribution as a product of multiple factors as the following,

$$p(V, \sim V) = p(V|Pa)\, p(Ch|V, CPa)\, p(\sim \{V, Ch\}) \tag{A.16}$$

where $p(V|Pa)$ is the conditional distribution of V given its parents, $p(Ch|V, CPa)$ is the conditional distribution of Ch given its parents, and $p(\sim \{V, Ch\})$ is the joint distribution of the set $\sim \{V, Ch\}$†. Marginalizing V from the joint distribution shown in Eqn. (A.16) as shown below,

$$p(\sim V) = \int_V p(V|Pa)\, p(Ch|V, CPa)\, p(\sim \{V, Ch\})\, dV \tag{A.17}$$

leads to the following,

$$p(\sim V) = p(Ch|Pa, CPa)\, p(\sim \{V, Ch\}) \tag{A.18}$$

Now, substituting the above expression in the expression shown for $p(V, \sim V)$ in Eqn. (A.14) and replacing the LHS of Eqn. (A.16) using the resulting expression leads to the following equality,

$$\begin{aligned} &p(V| \sim V)\, p(Ch|Pa, CPa)\, p(\sim \{V, Ch\}) \\ &= p(V|Pa)\, p(Ch|V, CPa)\, p(\sim \{V, Ch\}) \end{aligned} \tag{A.19}$$

Simplifying the above expression by cancelling out the common terms in both LHS and RHS results in the following,

$$p(V| \sim V)\, p(Ch|Pa, CPa) = p(V|Pa)\, p(Ch|V, CPa) \tag{A.20}$$

Further, rearranging the terms leads to the following,

$$p(V| \sim V) = \frac{p(V|Pa)\, p(Ch|V, CPa)}{p(Ch|Pa, CPa)} \tag{A.21}$$

*if V is of discrete random variables, integration has to be changed to summation over all the combination of states that V can take

†Of course, the joint distribution $p(\sim \{V, Ch\})$ could be factored into a product of multiple conditional distributions given the knowledge of the structure of the entire network. For our purposes, it could be any arbitrary structure that honours the DAG constraint.

Writing the numerator in the above expression as the joint conditional distribution of V and Ch using the chain rule results in the following,

$$p\left(V|\sim V\right)=\frac{p\left(V,Ch|Pa,CPa\right)}{p\left(Ch|Pa,CPa\right)} \tag{A.22}$$

Now, factoring the joint conditional distribution using the chain rule into the following product, $p\left(V|Ch,Pa,CPa\right)p\left(Ch|Pa,CPa\right)$ leads to

$$p\left(V|\sim V\right)=\frac{p\left(V|Ch,Pa,CPa\right)p\left(Ch|Pa,CPa\right)}{p\left(Ch|Pa,CPa\right)} \tag{A.23}$$

The common term in the numerator and the denominator can be cancelled to obtain,

$$p\left(V|\sim V\right)=p\left(V|Ch,Pa,CPa\right) \tag{A.24}$$

The above expression implies that V is D-separated from the rest of the network by Ch, Pa, and CPa. This completes the proof of proposition 2. ■

Appendix B

Estimation Approach for the GPLLVM

B.1 Maximum likelihood estimation of the GPLLVM using the EM algorithm

The update expressions for obtaining the parameters of GPLLVM using the EM algorithm are presented in Table B.1.

Table B.1: Recursive update expressions for estimating the parameters of GPLLVM

E-step updates
$\Sigma_{z\|y,x,u} = \Phi = \left[W^T\psi_y^{-1}W + V^T\psi_x^{-1}V + I_K\right]^{-1}$ $\mu_{z_n\|y_n,x_n,u_n} = \Phi\left[W^T\psi_y^{-1}\left(y_n - Fu_n\right) + V^T\psi_x^{-1}x_n\right]$
M-step updates
$\psi_x = \frac{1}{N}\sum_{n=1}^{N}\left[x_n x_n^T + V\left\{\mu_{z_n\|y_n,x_n,u_n}\mu_{z_n\|y_n,x_n,u_n}^T + \Phi\right\}V^T - 2V\mu_{z_n\|y_n,x_n,u_n}x_n^T\right]$ $\psi_y = \frac{1}{N}\sum_{n=1}^{N}\left[\left(y_n - Fu_n\right)\left(y_n - Fu_n\right)^T + W\left\{\mu_{z_n\|y_n,x_n,u_n}\mu_{z_n\|y_n,x_n,u_n}^T + \Phi\right\}W^T\right]$ $\quad - \frac{2}{N}\sum_{n=1}^{N}\left[W\mu_{z_n\|y_n,x_n,u_n}\left(y_n - Fu_n\right)^T\right]$ $W = \left[\frac{1}{N}\sum_{n=1}^{N}\left(y_n - Fu_n\right)\mu_{z_n\|y_n,x_n,u_n}^T\right]\left[\frac{1}{N}\sum_{n=1}^{N}\left\{\mu_{z_n\|y_n,x_n,u_n}\mu_{z_n\|y_n,x_n,u_n}^T + \Phi\right\}\right]^{-1}$ $V = \left[\frac{1}{N}\sum_{n=1}^{N}x_n\mu_{z_n\|y_n,x_n,u_n}^T\right]\left[\frac{1}{N}\sum_{n=1}^{N}\left\{\mu_{z_n\|y_n,x_n,u_n}\mu_{z_n\|y_n,x_n,u_n}^T + \Phi\right\}\right]^{-1}$ $F = \left[\frac{1}{N}\sum_{n=1}^{N}\left(y_n - W\mu_{z_n\|y_n,x_n,u_n}\right)u_n^T\right]\left[\frac{1}{N}\sum_{n=1}^{N}u_n u_n^T\right]^{-1}$

B.2 Woodbury Matrix Identity

Lemma 6. *Woodbury Matrix Identity: The following identity holds,*

$$\begin{aligned}&(\mathcal{M}_1 + \mathcal{M}_2\mathcal{M}_3\mathcal{M}_4)^{-1} = \\ &\mathcal{M}_1^{-1} - \mathcal{M}_1^{-1}\mathcal{M}_2\left(\mathcal{M}_3^{-1} + \mathcal{M}_4\mathcal{M}_1^{-1}\mathcal{M}_2\right)^{-1}\mathcal{M}_4\mathcal{M}_1^{-1}\end{aligned} \tag{B.1}$$

where matrices $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ are of appropriate sizes that allow matrix multiplications shown in the above equation and matrices $\mathcal{M}_1$ and $\mathcal{M}_3$ are invertible.

B.3 Matrix B is an Idempotent Matrix

We can show that the matrix B is an idempotent matrix by showing,

$$BB = B \tag{B.2}$$

From equation (3.45), BB can be expressed as,

$$BB = \Sigma^{\frac{1}{2}} A \Sigma A \Sigma^{\frac{1}{2}} \tag{B.3}$$

Therefore, the equality in equation (B.2) holds if the following is true,

$$A\Sigma A = A \tag{B.4}$$

Σ in equation (3.50) can be written in terms $\mathcal{V}$ and Ψ as the following,

$$\Sigma = \Psi^{\frac{1}{2}} \mathcal{V}\mathcal{V}^T \Psi^{\frac{1}{2}} + \Psi \tag{B.5}$$

Using the above expression for Σ and the expression for A shown in equation (3.43), the following can be shown,

$$A\Sigma A = \Psi^{-\frac{1}{2}} \left[I_{P+L} - \mathcal{V}\left(\mathcal{V}^T\mathcal{V}\right)^{-1}\mathcal{V}^T \right] \Psi^{-\frac{1}{2}} \tag{B.6}$$

which is nothing but A and this again can be verified from equation (3.43). Therefore, B is an idempotent matrix.

Appendix C

Supplementary Information for the Identification of the Two-Layer Mixture Bayesian PPCA model

C.1 Estimation of the mixture Bayesian PPCA model

The model estimation consists of estimating the following quantities, 1) the deterministic parameters that do not have prior distributions $\{\mu, \pi, \sigma^2\}$ and 2) the parameters and the latent variables that have prior distributions $\{W, \nu, \mathbf{s}, Z\}$. For the deterministic parameters, point estimates are obtained. For the others that have prior distributions, the posterior distribution ($p(W, \nu, Z, \mathbf{s}|X, \mu, \pi, \sigma^2, a^*, b^*)$) are obtained. We use the variational approach that is very popular in the case of mixture of latent variable models [69, 118]. For this model, the variation approach requires us to consider the following approximation for the posterior distribution,

$$p\left(Z, \mathbf{s}, W, \nu|X, \mu, \pi, \sigma^2, a^*, b^*\right) \approx q(W)\, q(\nu)\, q(Z|\mathbf{s})\, q(\mathbf{s}) \tag{C.1}$$

where the factors $q(W)$, $q(\nu)$, $q(Z|\mathbf{s})$ and $q(\mathbf{s})$ are the individual posteriors of the loading parameters, the precision parameters of the loading parameters, the latent variables given the model identities and the model identities respectively. Further, it requires the individual posteriors to have the same distribution forms of the priors. That is, $q(W)$ has to be a multivariate Gaussian, $q(\nu)$ has to be a Gamma distribution, $q(Z|\mathbf{s})$ has to be a multivariate Gaussian and $q(\mathbf{s})$ has to be a categor-

ical distribution. The detailed distribution forms of the posteriors are as follows,

$$q\left(Z|\mathbf{s}\right)=\prod_{n=1}^{N}\prod_{s=1}^{S}q\left(z_{n}^{s}|s_{n}=s\right)=\prod_{n=1}^{N}\prod_{s=1}^{S}\mathcal{N}\left(z_{n}^{s}|\hat{z_{n}^{s}},\Sigma_{z^{s}}\right) \tag{C.2}$$

$$q\left(\mathbf{s}\right)=\prod_{n=1}^{N}\prod_{s=1}^{S}q\left(s_{n}=s\right) \tag{C.3}$$

$$q\left(W\right)=\prod_{s=1}^{S}\prod_{d=1}^{D}q\left(W_{d}^{s}\right)=\prod_{s=1}^{S}\prod_{d=1}^{D}\mathcal{N}\left(W_{d}^{s}|\hat{W}_{d}^{s},\Sigma_{W_{d}^{s}}\right) \tag{C.4}$$

$$q\left(\nu\right)=\prod_{s=1}^{S}\prod_{m=1}^{M}q\left(\nu_{m}^{s}\right)=\prod_{s=1}^{S}\prod_{m=1}^{M}\Gamma\left(\nu_{m}^{s}|a,b_{m}^{s}\right) \tag{C.5}$$

where d and m indicate a particular dimension of the observation or a particular row of the loading matrix and a particular column of the loading matrix respectively. Further, the variational lower bound for the log marginal distribution of the observations is defined as a function of the posteriors and the parameters as the following,

$$\ln p\left(X|\mu,\pi,\sigma^{2},a*,b*\right)\geq\mathcal{F}\left(q\left(Z|\mathbf{s}\right),q\left(\mathbf{s}\right),q\left(W\right),q\left(\nu\right),\mu,\pi,\sigma^{2},a*,b*\right) \tag{C.6}$$

where

$$\mathcal{F}\geq\int q\left(W\right)q\left(\nu\right)q\left(Z|\mathbf{s}\right)q\left(\mathbf{s}\right)\ln\frac{p\left(X,Z,\mathbf{s},W,\nu|\mu,\pi,\sigma^{2},a*,b*\right)}{q\left(W\right)q\left(\nu\right)q\left(Z|\mathbf{s}\right)q\left(\mathbf{s}\right)}dZd\mathbf{s}dWd\nu \tag{C.7}$$

where the numerator term inside the logarithm is the joint distribution of the observations, the latent variables and the parameters which in our case is as follows,

$$\begin{aligned}p\left(X,Z,\mathbf{s},W,\nu|\mu,\pi,\sigma^{2},a^{*},b^{*}\right)&=\prod_{n=1}^{N}p\left(x_{n}|z_{n}^{s},s_{n},W^{s},\mu,\sigma^{2}\right)p\left(z_{n}^{s}\right)p\left(s_{n}=s\right)\\&\times\prod_{s=1}^{S}\prod_{m=1}^{M}p\left(W_{m}^{s}|\nu_{m}^{s}\right)p\left(\nu_{m}^{s}|a^{*},b^{*}\right)\end{aligned} \tag{C.8}$$

Estimation proceeds by maximizing the lower bound shown Eqn. (C.7). The E and M steps of the estimation algorithm are derived and shown below.

C.1.1 E-step

$$\frac{d\mathcal{F}}{dq\left(W_d^s\right)} = 0 \Rightarrow \tag{C.9}$$

$$q\left(W_d^s\right) \propto \int_{\nu^s} d\nu^s q\left(\nu^s\right) \ln p\left(W_d^s|\nu^s\right) + \sum_{\mathbf{s}} q\left(\mathbf{s}\right) \int dZ q\left(Z|\mathbf{s}\right) \ln p\left(X|Z,W,\mathbf{s},\mu\right) \tag{C.10}$$

$$\Sigma_{W_d^s} = \left[\sum_{n=1}^{N} \frac{q\left(s_n = s\right)}{\sigma^2}\left[\hat{z_n^s}\hat{z_n^s}' + \Sigma_{Z^s}\right] + \left[\begin{array}{ccccc} \frac{a}{b_1^s} & . & . & . & 0 \\ . & \frac{a}{b_2^s} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & \frac{a}{b_M^s} \end{array}\right]\right]^{-1} \tag{C.11}$$

$$\hat{W_d^s} = \Sigma_{W_d^s}\left[\sum_{n=1}^{N} \frac{q\left(s_n = s\right)}{\sigma^2}\left[x_{nd} - \mu_d^s\right]\hat{z_n^s}'\right] \tag{C.12}$$

$$\frac{d\mathcal{F}}{dq(\nu_m^s)} = 0 \Rightarrow \quad \ln q\left(\nu_m^s\right) \propto \ln p\left(\nu_m^s\right) + \int_W dW q\left(W\right) \ln \frac{p(W|\nu)}{q(W)} \tag{C.13}$$

$$a = a^* + \frac{D}{2} \tag{C.14}$$

$$b_m^s = b^* + \frac{1}{2} SS^s\left(m,m\right) \tag{C.15}$$

where

$$SS^s = \sum_{d=1}^{D}\left[\hat{W_d^s}'\hat{W_d^s} + \Sigma_{W_d^s}\right] \tag{C.16}$$

$$\frac{d\mathcal{F}}{dq(z_n^s)} = 0 \Rightarrow \quad \ln q\left(z_n^s\right) \propto \ln p\left(z_n^s\right) + \sum_{\mathbf{s}} q\left(\mathbf{s}\right) \int dW q\left(W\right) \ln p\left(X|Z,W,\mathbf{s},\mu\right) \tag{C.17}$$

$$\Sigma_{Z^s} = \left[\frac{1}{\sigma^2} SS^s + I\right]^{-1} \tag{C.18}$$

$$\hat{z_n^s} = \frac{1}{\sigma^2}\Sigma_{Z^s}\left[\left(x_n - \mu^s\right)\hat{W^s}\right] \tag{C.19}$$

$$\frac{d\mathcal{F}}{dq\left(s_n=s\right)}=0\Rightarrow$$

$$\ln q\left(s_n=s\right)\propto\ln p\left(s_n=s\right)+\int dz_n^s q\left(z_n^s|s_n=s\right)\ln\frac{p\left(z_n^s\right)}{q\left(z_n^s|s_n=s\right)}$$

$$+\int dz_n^s dW^s q\left(z_n^s|s_n=s\right)q(W^s)\ln p\left(x_n|z_n^s,W^s,s_n=s,\mu^s\right) \tag{C.20}$$

$$\ln q\left(s_n=s\right)\propto\ln|\Sigma_{Z^s}|+\frac{1}{2\sigma^2}\left[x_n-\mu^s\right]\hat{W^s}\hat{z_n^s}$$
$$-tr\left[\left(\frac{1}{2\sigma^2}SS^s+I\right)\left(\Sigma_{Z^s}+\hat{z_n^s}\hat{z_n^s}'\right)+\frac{1}{\sigma^2}\left(x_n-\mu^s\right)\left(x_n-\mu^s\right)\right] \tag{C.21}$$

where

$$\sum_{s=1}^{S}q\left(s_n=s\right)=1 \tag{C.22}$$

C.1.2 M-step

$$\frac{d\mathcal{F}}{d\pi^s}=0\Rightarrow\quad\pi^s=\frac{\sum_{n=1}^{N}q(s_n=s)}{N} \tag{C.23}$$

$$\frac{d\mathcal{F}}{d\mu^s}=0\Rightarrow\quad\mu^s=\frac{1}{\sum_{n=1}^{N}q(s_n=s)}\left[\sum_{n=1}^{N}q\left(s_n=s\right)\left(x_n-\hat{W^s}z_n^s\right)\right] \tag{C.24}$$

$$\frac{d\mathcal{F}}{d\sigma^2}=0\Rightarrow$$

$$\sigma^2=\frac{1}{ND}\sum_{s=1}^{S}\sum_{n=1}^{N}q\left(s_n=s\right)\left[\left(x_n-\mu^s\right)'\left(x_n-\mu^s-2\hat{W^s}z_n^s\right)\right]$$

$$+\frac{1}{ND}\sum_{s=1}^{S}\sum_{n=1}^{N}q\left(s_n=s\right)\left[SS^s\left[z_n^s z_n^{s'}+\Sigma_{Z^s}\right]\right] \tag{C.25}$$

C.1.3 Estimation

It can be seen from the update equations that the updates are dependent on each other. Therefore, the posteriors and the parameters need to be updated recursively in multiple iterations until convergence. The algorithm for estimation is presented in Table. C.1.

Table C.1: Estimation algorithm for mixture Bayesian PPCA

Step 1	Initial guess for the variational and deterministic parameters (refer to C.1.4)
Step 2	$\mathcal{F}(0) = -\infty$
Step 3	For Iter = 1:MaxIter
Step 4	E-step: update all the posteriors one after the other
Step 5	M-step: update all the deterministic parameter one after the other
Step 6	If $\lvert\mathcal{F}(Iter) - \mathcal{F}(Iter-1)\rvert / \lvert\mathcal{F}(Iter-1)\rvert \leq \epsilon$
Step 7	Break For
Step 8	End If
Step 9	End For

C.1.4 Initial guess

For estimating the mixture PPCA model using the variational EM or the EM algorithm, we need a good initial guess to avoid poor local maxima. Initial guesses for the parameters of all the local models in the mixture are needed. For any local model, this can be obtained from the observations belonging to that particular local model. However, for this, we need to split and assign the observations to the local models first. We used k-means clustering algorithm to split and assign the observations. A PPCA model was fit to the observations to obtain the parameters of all the local models separately. PPCA model can be obtained either using the EM algorithm or eigen decomposition as shown in [28]. We used the eigen decomposition. Still, this initialization strategy does not provide a unique initialization all the time. It is because the k-means clustering also requires randomized initialization. It may converge to different solutions with respect to different initializations of which observation belongs to which cluster. However, this is the maximum control that we can have over the initial guess. With the k-means clustering followed by estimation of PPCA for each of the clusters provided by the k-means algorithm, we can obtain

initial guesses for the mean parameter μ^s, the loading matrix W^s and the proportion of data explained π^s for each of the clusters and the noise variance σ^2. In addition to these, providing positive definite matrices as initial guesses for $\Sigma_{W_d^s} \forall s, d$ is sufficient to execute the algorithm presented in C.1.3.

C.1.5 Determining dimension of latent variables

Update equation for the posterior of the precision parameter is given by a Gamma distribution. By evaluating the parameters of the posterior Gamma distribution, decision about the dimension reduction can be made. The posterior expected value of the precision is given by,

$$< \nu_m^s >= \frac{a}{b_m^s} \tag{C.26}$$

where a and b_m^s are the posterior scale and rate parameter for the precision of m^{th} dimension in s^{th} local model. If the expected value is very high, then it means that the entries in that particular column of the loading matrix is still closer to zero. So by setting a threshold for the expected value of the precision variable, zero columns in the loading matrices can be removed. We utilize the posterior distributions only to infer the effective dimensions of the loading parameters and we do not retain the uncertainty information provided by the posteriors for obtaining the two layer model or for fault detection. Only the posterior means are retained.

C.2 Proof of Proposition 5

Proof. The number of loading parameters required for the mixture PPCA model is KDP and the number of loading parameters required for the two layer model is $SDM + SCMP$. The number of mean parameters required for the mixture PPCA model is KD and the number of mean parameters required for the two layer model is $SD + SCM$. Therefore, our objective is to show the following relationships is true,

$$KDP - SDM - SCMP \geq 0 \quad \& \quad KD - SD - SCM \geq 0 \tag{C.27}$$

where the relationships correspond to the difference between the number of loading parameters and the difference between the number of mean parameters respectively.

The first relationship can be reduced to the following by using Eqn. (4.24),

$$CDP - DM - CMP \geq 0 \tag{C.28}$$

which can then be reduced to,

$$\frac{C\,(D-M)\,P}{DM} \geq 1 \tag{C.29}$$

which can further be reduced to the following,

$$C\left(1 - \frac{1}{r_1}\right)\frac{1}{r_2} \geq 1 \tag{C.30}$$

which directly gives the condition shown in Eqn. (4.23) and therefore, it is true. Given that the first relationship is true, a lower bound for KD can be obtained as the following,

$$KD \geq \frac{SDM + SCMP}{P} \tag{C.31}$$

Replacing KD by its lower bound in the second relationship reduces it to the following,

$$D\,(M-P) \geq 0 \tag{C.32}$$

which is true because $M > P$ and therefore, the second relationship is also true. ■

Appendix D

Estimation Approach for the Hybrid Model

D.1 The VBEM algorithm for the estimation of the hybrid model

The expression obtained for the lower bound is provided in Table. D.1.

Table D.1: Lower Bound Expression

$$
\begin{aligned}
\mathcal{L} \geq \quad & -\alpha \sum_{d=1}^{D} \sum_{m=1}^{M} \ln \beta_m^d + DM\alpha^* \ln \beta^* + DM \ln \tfrac{\Gamma(\alpha)}{\Gamma(\alpha^*)} \\
& +\alpha \sum_{d=1}^{D} \sum_{m=1}^{M} \left(1 - \tfrac{\beta^*}{\beta_m^d}\right) - \tfrac{N}{2} tr\left(\Sigma_X\right) - \tfrac{1}{2} \sum_{n=1}^{N} \left(x_n\right)^T x_n + \tfrac{NK}{2} \\
& + \tfrac{N}{2} \ln |\Sigma_X| + \tfrac{DM}{2} + \tfrac{1}{2} \sum_{d=1}^{D} \ln |\Sigma_{f^d}| - \tfrac{1}{2} \sum_{d=1}^{D} tr\left(\lambda^d \left[\Sigma_{f^d} + \left(\hat{f}^d\right)^T \hat{f}^d\right]\right) \\
& - (\alpha - \alpha^*) \sum_{d=1}^{D} \sum_{m=1}^{M} \left(\psi\left(\alpha\right) - \ln \beta_m^d\right) \\
& + \tfrac{1}{2} \sum_{d=1}^{D} \sum_{m=1}^{M} \left(\psi\left(\alpha\right) - \ln \beta_m^d\right) - \kappa \sum_{d=1}^{D} \ln \phi^d + D\kappa^* \ln \phi^* + D \ln \tfrac{\Gamma(\kappa)}{\Gamma(\kappa^*)} \\
& +\kappa \sum_{d=1}^{D} \left(1 - \tfrac{\phi^*}{\phi^d}\right) - (\kappa - \kappa^*) \sum_{d=1}^{D} \left(\psi\left(\kappa\right) - \ln \phi^d\right) \\
& + \tfrac{N}{2} \sum_{d=1}^{D} \left(\psi\left(\kappa\right) - \ln \phi^d\right) - \tfrac{ND}{2} \ln\left(2\pi\right) - \sum_{n=1}^{N} \sum_{d=1}^{D} \tfrac{\left(y_n^d\right)^2 \kappa}{2\phi^d} \\
& + \sum_{n=1}^{N} \sum_{d=1}^{D} y_n^d \hat{f}^d \hat{z}_n \tfrac{\kappa}{\phi^d} - N \sum_{d=1}^{D} \tfrac{\kappa}{2\phi^d} tr\left\{\left[\left(\hat{f}^d\right)^T \hat{f}^d + \Sigma_{f^d}\right] \Sigma_z\right\} \\
& - \sum_{n=1}^{N} \sum_{d=1}^{D} \tfrac{\kappa}{2\phi^d} tr\left\{\left[\left(\hat{f}^d\right)^T \hat{f}^d + \Sigma_{f^d}\right] \hat{z}_n \left(\hat{z}_n\right)^T\right\}
\end{aligned}
$$

where $\hat{F} = \left[\hat{W}, \hat{V}\right]$, $\hat{Z} = \left[\ U, \hat{X}\ \right]^T$, $\Sigma_Z = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_X \end{bmatrix}$, $\epsilon^d = diag\left(\left[\nu_1^d, \nu_2^d, ..., \nu_M^d\right]\right)$,

$$\lambda^d = diag\left(\left[\tfrac{\alpha}{\beta_1^d}, \tfrac{\alpha}{\beta_2^d}, ..., \tfrac{\alpha}{\beta_M^d}\right]\right)$$

Update expressions for each of the posterior parameter and the prior parameters are listed in Table. D.2.

Table D.2: Update Expressions

Distribution	Parameters
$q\left(\nu_m^d\right)$	$\alpha = \alpha^* + \frac{1}{2},\ \beta_m^d = \beta^* + \frac{1}{2}\left[\left(\hat{f}_m^d\right)^2 + \Sigma_{f_m^d}\right]$
$q\left(\sigma^d\right)$	$\kappa = \kappa^* + \frac{N}{2},\ \phi^d = \phi^* + \frac{1}{2}\sum_{n=1}^N \left(y_n^d\right)^2 - \sum_{n=1}^N y_n^d \hat{f}^d \hat{z}_n$
	$+\frac{1}{2}\sum_{n=1}^N tr\left\{\left[\left(\hat{f}^d\right)^T \hat{f}^d + \Sigma_{f^d}\right]\left[\hat{z}_n\left(\hat{z}_n\right)^T\right]\right\}$
	$+\frac{1}{2}\sum_{n=1}^N tr\left\{\left[\left(\hat{f}^d\right)^T \hat{f}^d + \Sigma_{f^d}\right]\left[\Sigma_z\right]\right\}$
$q\left(f^d\right)$	$\Sigma_{f^d} = \left[\sum_{n=1}^N \frac{\kappa}{\phi^d}\hat{z}_n\left(\hat{z}_n\right)^T + \frac{N\kappa}{\phi^d}\Sigma_z + \lambda^d\right]^{-1}$
	$\hat{f}^{d^T} = \Sigma_{f^d}\left[\sum_{n=1}^N y_n^d \hat{z}_n \frac{\kappa}{\phi^d}\right]$
$q\left(x_n\right)$	$\Sigma_x = \left[\sum_{d=1}^D \left(\left(\hat{V}^d\right)^T \hat{V}^d + \Sigma_{V^d}\right)\frac{\kappa}{\phi^d} + I\right]^{-1}$
	$\hat{x}_n = \Sigma_x\left[\left(y_n - \hat{W}U\right)^T diag\left(\frac{\kappa}{\phi}\right)\hat{V}\right]$
Prior parameters	$\frac{1}{\beta^*} = \frac{\alpha}{DM\alpha^*}\sum_{m=1}^M \sum_{d=1}^D \frac{1}{\beta_m^d},\ \frac{1}{\phi^*} = \frac{\kappa}{D\kappa^*}\sum_{d=1}^D \frac{1}{\phi^d}$
	$\psi\left(\alpha^*\right) = \ln\beta^* + \frac{1}{DM}\sum_{m=1}^M \sum_{d=1}^D \left(\psi\left(\alpha\right) - \ln\left(\beta_m^d\right)\right)$
	$\psi\left(\kappa^*\right) = \ln\phi^* + \frac{1}{D}\sum_{d=1}^D \left(\psi\left(\kappa\right) - \ln\left(\phi^d\right)\right)$

where $\hat{F} = \left[\hat{W}, \hat{V}\right]$, $\hat{Z} = \left[\ U, \hat{X}\ \right]^T$, $\Sigma_Z = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_X \end{bmatrix}$, $\epsilon^d = diag\left(\left[\nu_1^d, \nu_2^d, ..., \nu_M^d\right]\right)$,

$$\lambda_d = diag\left(\left[\frac{\alpha}{\beta_1^d}, \frac{\alpha}{\beta_2^d}, ..., \frac{\alpha}{\beta_M^d}\right]\right)$$

Appendix E

Supplementary Materials for Causal Modelling Based on the TVPM

E.1 Additional Results

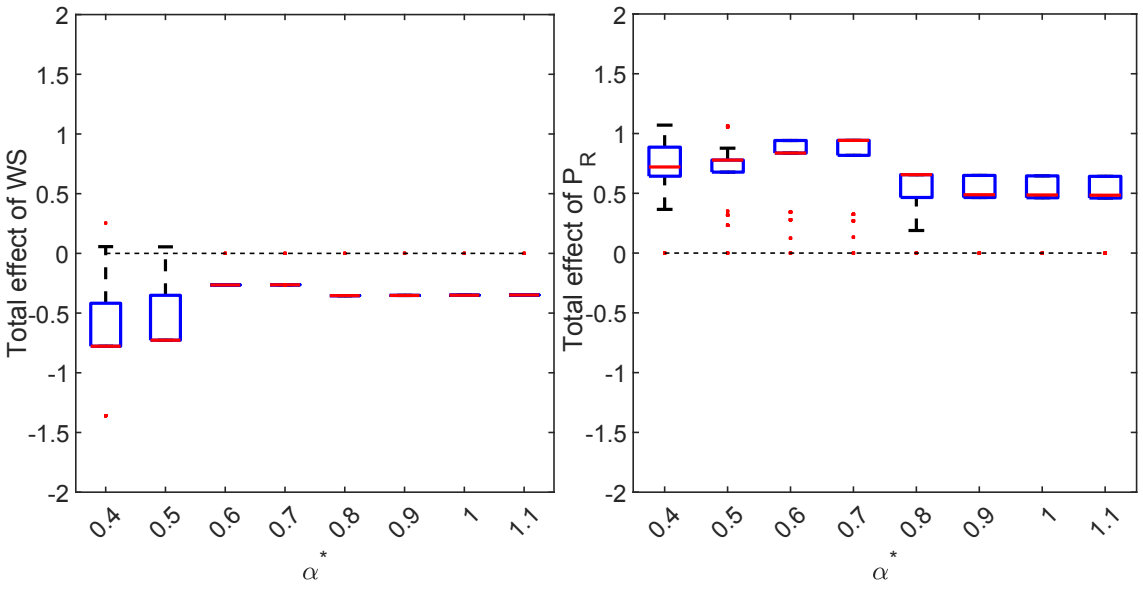


Figure E.1: Well 2: Spreads of the estimated total effects of well bore subcool and steam chamber pressure on production at different values of α^*.

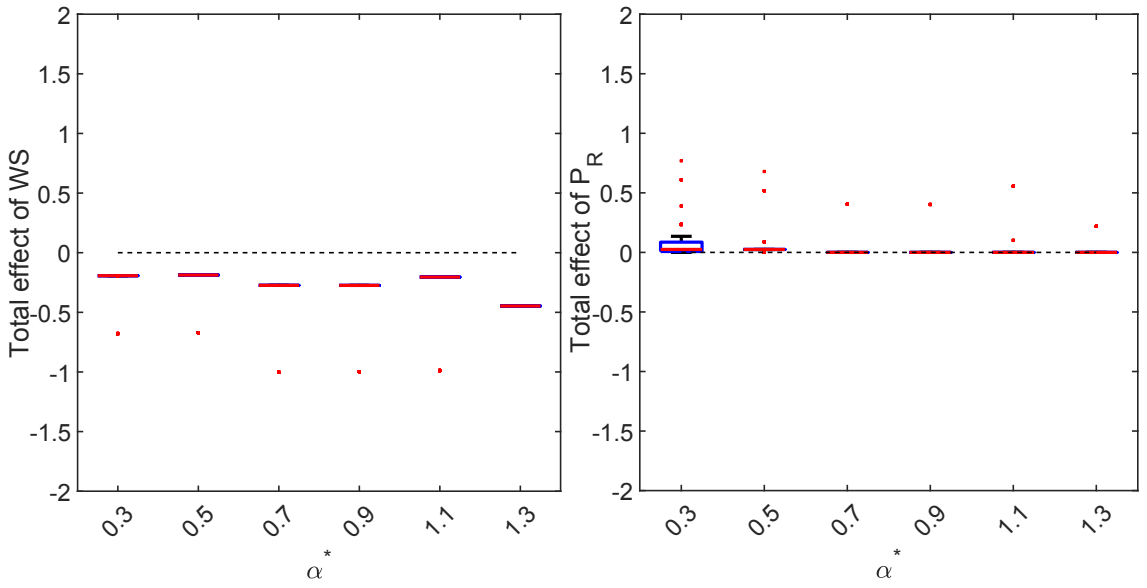


Figure E.2: Well 3: Spreads of the estimated total effects of well bore subcool and steam chamber pressure on production at different values of α^*.

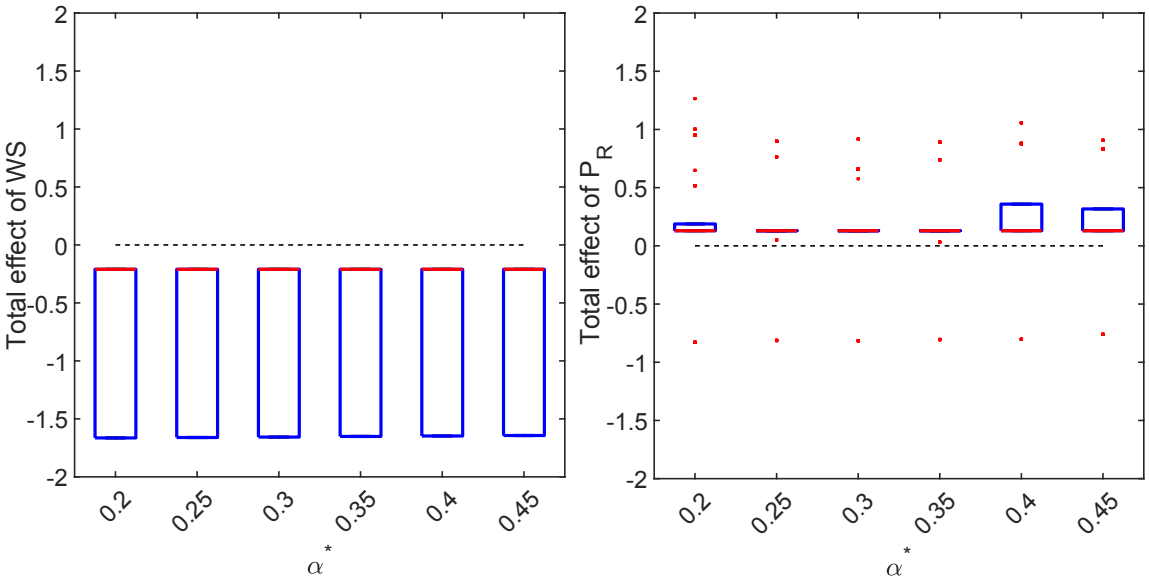


Figure E.3: Well 4: Spreads of the estimated total effects of well bore subcool and steam chamber pressure on production at different values of α^*.

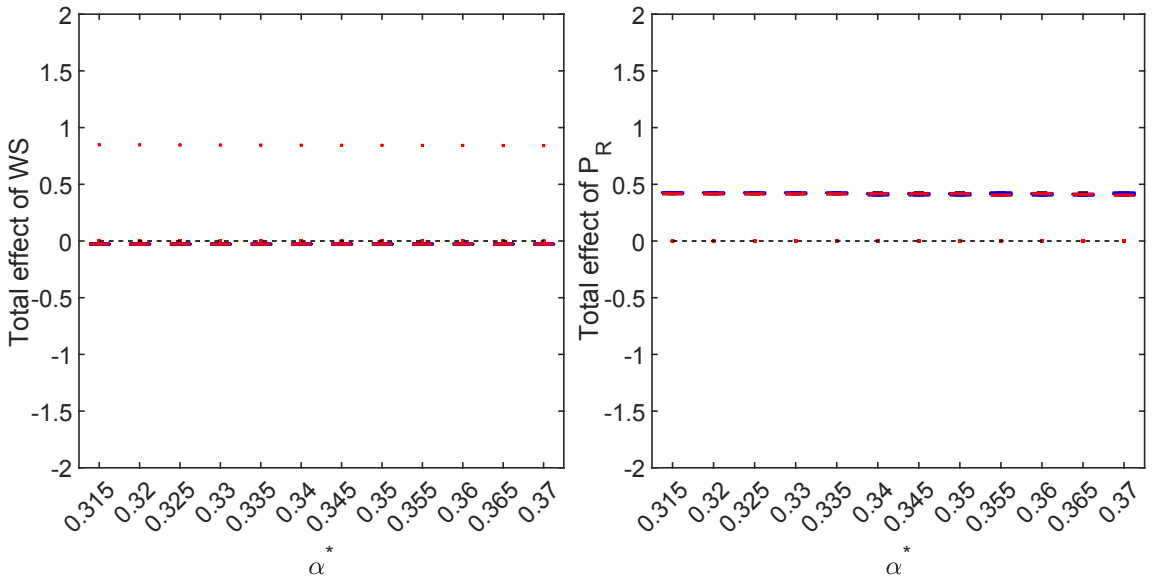


Figure E.4: Well 5: Spreads of the estimated total effects of well bore subcool and steam chamber pressure on production at different values of α^*.

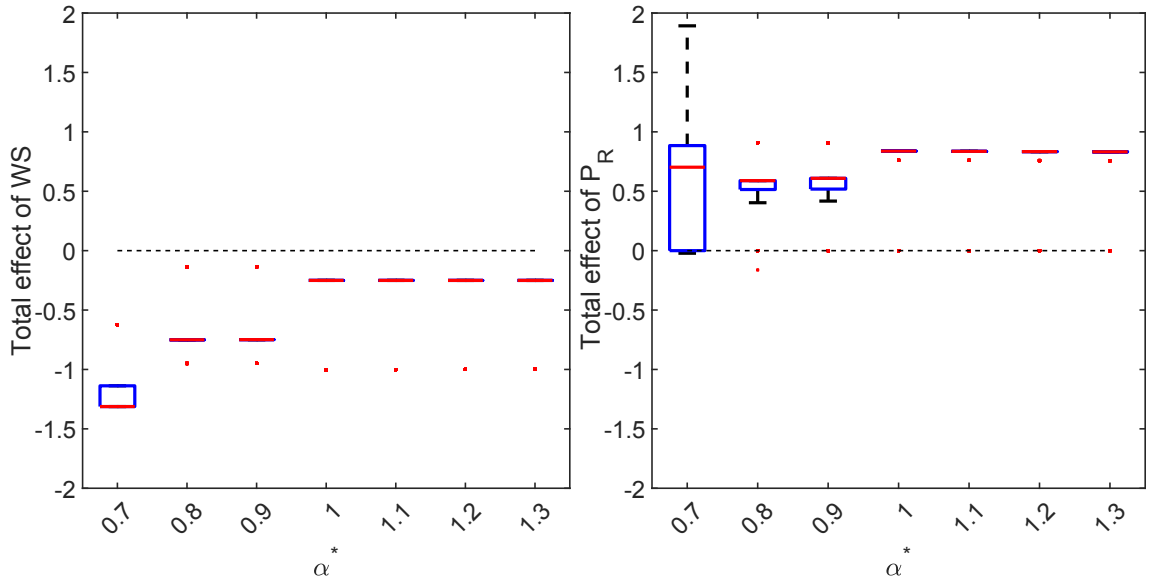


Figure E.5: Well 6: Spreads of the estimated total effects of well bore subcool and steam chamber pressure on production at different values of α^*.

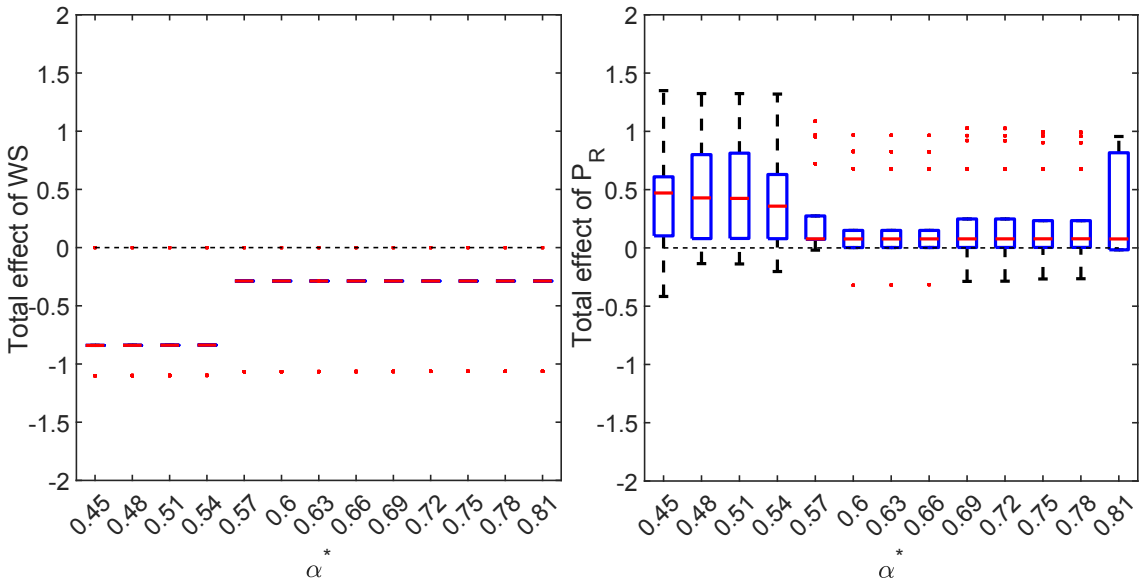


Figure E.6: Well 7: Spreads of the estimated total effects of well bore subcool and steam chamber pressure on production at different values of α^*.

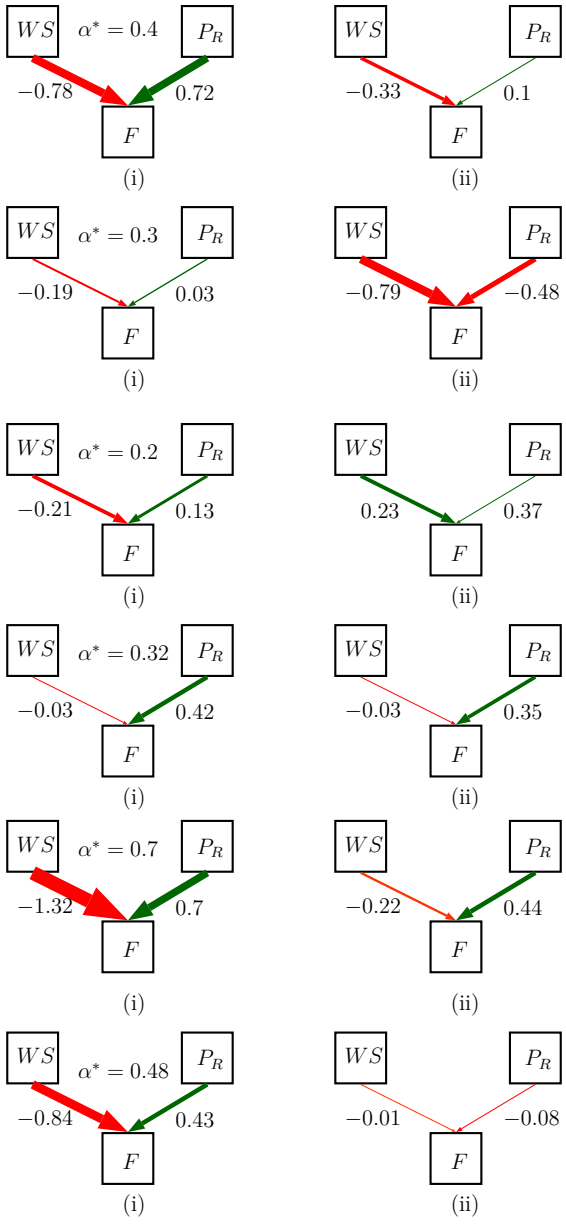


Figure E.7: For wells 2 to 7 (top to bottom): (i) Median of total effects identified using the TVPM based approach and (ii) total effect identified using the time-invariant linear regression models estimated under the ML approach.

E.2 VBEM Algorithm: Estimation of the TVPM

Table E.1: $\mathcal{L}_{KL}$: During the estimation stage (top) and the hypothesis testing stage (bottom)

$\mathcal{L}_{KL} =$	$-\alpha \sum_{t=0}^{T} \sum_{d=1}^{D} \ln \beta_t^d + (T+1) D \alpha^* \ln \beta^*$
	$+ (T+1) D \ln \frac{\Gamma(\alpha)}{\Gamma(\alpha^*)} + \alpha \sum_{t=0}^{T} \sum_{d=1}^{D} \left(1 - \frac{\beta^*}{\beta_t^d}\right)$
	$-\frac{1}{2} \sum_{t=0}^{T} \ln \lvert\Lambda_t\rvert + \ln p\left(y_1 \lvert u_1, \sigma\right) + \sum_{t=1}^{T} \ln p\left(y_t \lvert y_{1:t-1}, u_t, \sigma\right);$
$p\left(y_1 \lvert u_1, \sigma\right) =$	$\mathcal{N}\left(y_1 \lvert 0, \sigma^{-1} + u_1^T \left\{\Lambda_0^{-1} + \Lambda_1^{-1}\right\} u_1\right)$
$p\left(y_t \lvert y_{1:t-1}, u_t, \sigma\right) =$	$\mathcal{N}\left(y_t \lvert u_t^T \mu_{t-1}, \sigma^{-1} + u_t^T \left(\Lambda_t^{-1} + \Sigma_{t-1}\right) u_t\right)$
$\mathcal{L}_{KL} =$	$-\alpha \sum_{t=0}^{T} \sum_{d=1}^{D_t^{\#}} \ln \beta_t^d + \sum_{t=0}^{T} \sum_{d=1}^{D_t^{\#}} \alpha^* \ln \beta^*$
	$+ \sum_{t=0}^{T} \sum_{d=1}^{D_t^{\#}} \ln \frac{\Gamma(\alpha)}{\Gamma(\alpha^*)} + \alpha \sum_{t=0}^{T} \sum_{d=1}^{D_t^{\#}} \ln \left(1 - \frac{\beta^*}{\beta_t^d}\right)$
	$-\frac{1}{2} \sum_{t=0}^{T} \ln \lvert\Lambda_t^{\#}\rvert + \ln p\left(y_1 \lvert u_1^{\&}, \sigma\right) + \sum_{t=1}^{T} \ln p\left(y_t \lvert y_{1:t-1}, u_t^{\&}, \sigma\right);$
$p\left(y_1 \lvert u_1^{\&}, \sigma\right) =$	$\mathcal{N}\left(y_1 \lvert 0, \mathcal{M}_1\right)$
$\mathcal{M}_1 =$	$\sigma^{-1} + \left(u_1^{\&}\right)^T \begin{bmatrix} \left(\Lambda_0^{\sim}\right)^{-1} + \left(\Lambda_1^{\sim}\right)^{-1} & 0 & 0 \\ 0 & \left(\Lambda_0^{-}\right)^{-1} & 0 \\ 0 & 0 & \left(\Lambda_1^{+}\right)^{-1} \end{bmatrix} u_1^{\&}$
$p\left(y_t \lvert y_{1:t-1}, u_t^{\&}, \sigma\right) =$	$\mathcal{N}\left(y_t \lvert \left(u_t^{\&}\right)^T \begin{bmatrix} \mu_{t-1}^* \\ 0 \end{bmatrix}, \mathcal{M}_2\right)$
$\mathcal{M}_2 =$	$\sigma^{-1} + \left(u_t^{\&}\right)^T \left(\begin{bmatrix} \left(\Lambda_t^{\sim}\right)^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left(\Lambda_t^{+}\right)^{-1} \end{bmatrix} + \begin{bmatrix} \Sigma_{t-1}^* & 0 \\ 0 & 0 \end{bmatrix} \right) u_t^{\&}$

Table E.2: Update expressions: During the estimation stage

$$
\begin{array}{ll}
\alpha = & \alpha^* + \frac{1}{2} \\
\beta_0^d = & \beta^* + \frac{1}{2}\left[\Sigma_{\theta_0\theta_0}(d,d) + \left(\hat{\theta}_0^d\right)^2\right] \\
\beta_t^d = & \beta^* + \frac{1}{2}\left[\Sigma_{\theta_t\theta_t}(d,d) + \left(\hat{\theta}_t^d\right)^2\right] + \frac{1}{2}\left[\Sigma_{\theta_{t-1}\theta_{t-1}}(d,d) + \left(\hat{\theta}_{t-1}^d\right)^2\right] \\
 & - \left[\Sigma_{\theta_t\theta_{t-1}}(d,d) + \left(\hat{\theta}_t^d\hat{\theta}_{t-1}^d\right)\right] \ \forall\, t \in [1,T] \\
\text{Filter:} & \\
\Sigma_0 = & \Lambda_0^{-1} \\
\mu_0 = & 0 \\
\Sigma_t = & \left[\left\{\Lambda_t^{-1} + \Sigma_{t-1}\right\}^{-1} + \sigma u_t u_t^T\right]^{-1} \ \forall\, t \in [1,T] \\
\mu_t = & \Sigma_t\left[\left\{\Lambda_t^{-1} + \Sigma_{t-1}\right\}^{-1}\mu_{t-1} + \sigma u_t y_t\right] \ \forall\, t \in [1,T] \\
\text{Smoother:} & \\
\Sigma_{\theta_t\theta_{t+1}} = & \begin{bmatrix} C + C\Lambda_{t+1}\Sigma_{\theta_{t+1}}\Lambda_{t+1}C & C\Lambda_{t+1}\Sigma_{\theta_{t+1}} \\ \Sigma_{\theta_{t+1}}\Lambda_{t+1}C & \Sigma_{\theta_{t+1}} \end{bmatrix}; \\
 & C = \left[\Sigma_t^{-1} + \Lambda_{t+1}\right]^{-1} \ \forall\, t \in [T-1,0] \\
\begin{bmatrix} \hat{\theta}_t \\ \hat{\theta}_{t+1} \end{bmatrix} = & \Sigma_{\theta_t\theta_{t+1}}\begin{bmatrix} \Sigma_t^{-1}\mu_t \\ -\left(\Lambda_{t+1} - \Lambda_{t+1}C\Lambda_{t+1}\right)\mu_t + \Sigma_{\theta_{t+1}}^{-1}\hat{\theta}_{t+1} \end{bmatrix}; \\
 & C = \left[\Sigma_t^{-1} + \Lambda_{t+1}\right]^{-1} \ \forall\, t \in [T-1,0] \\
\sigma = & \frac{T}{\sum_{t=1}^{T}\left[y_t^T y_t + u_t^T\left\{\hat{\theta}_t\hat{\theta}_t^T + \Sigma_{\theta_t\theta_t}\right\}u_t - 2y_t^T u_t^T\hat{\theta}_t\right]}
\end{array}
$$

Table E.3: Update expressions: During the hypothesis testing stage

Filter:

$$\Sigma_0^{\&} = \left(\Lambda_0^{\&}\right)^{-1}$$

$$\mu_0^{\&} = 0$$

$$K_t^{\&} = \left[\begin{bmatrix} (\Lambda_t^{\sim})^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left(\Lambda_t^{+}\right)^{-1} \end{bmatrix} + \begin{bmatrix} \Sigma_{t-1}^{*} & 0 \\ 0 & 0 \end{bmatrix}\right]^{-1} \quad \forall\, t \in [1, T]$$

$$\Sigma_t^{\&} = \left[K_t^{\&} + \sigma u_t^{\&} \left(u_t^{\&}\right)^T\right]^{-1} \quad \forall\, t \in [1, T]$$

$$\mu_t^{\&} = \Sigma_t^{\&} \left[K_t^{\&} \mu_{t-1} + \sigma u_t^{\&} y_t\right] \quad \forall\, t \in [1, T]$$

Smoother:

$$\Sigma_{\theta_t^* \theta_{t+1}^{\sim}} = \begin{bmatrix} C_1^{\sim\sim} + C_2 & C_1^{\sim -} & -C_2 \\ C_1^{-\sim} & C_1^{--} + C_3^{--} - C_4^{--} & C_3^{-\sim} - C_4^{-\sim} \\ -C_2 & C_3^{\sim -} - C_4^{\sim -} & C_2 + C_3^{\sim\sim} - C_4^{\sim\sim} \end{bmatrix}^{-1}$$

$$\forall\, t \in [T-1, 0]$$

$$\begin{bmatrix} \hat{\theta}_t^* \\ \hat{\theta}_{t+1}^{\sim} \end{bmatrix} = \Sigma_{\theta_t^* \theta_{t+1}^{\sim}} \begin{bmatrix} C_1^{\sim\sim} \mu_t^{\sim} + C_1^{\sim -} \mu_t^{-} \\ C_1^{-\sim} \mu_t^{\sim} + C_1^{--} \mu_t^{-} + C_3^{-\sim} \hat{\theta}_{t+1}^{\sim} + C_3^{--} \hat{\theta}_{t+1}^{-} - C_4^{-\sim} \mu_t^{\sim} - C_4^{--} \mu_t^{-} \\ C_3^{\sim\sim} \hat{\theta}_{t+1}^{\sim} + C_3^{\sim -} \hat{\theta}_{t+1}^{-} - C_4^{\sim\sim} \mu_t^{\sim} - C_4^{\sim -} \mu_t^{-} \end{bmatrix}$$

$$\forall\, t \in [T-1, 0]$$

Constants

$$\begin{bmatrix} C_1^{\sim\sim} & C_1^{\sim -} \\ C_1^{-\sim} & C_1^{--} \end{bmatrix} = \left(\Sigma_t^*\right)^{-1},\ C_2 = \Lambda_{t+1}^{\sim},\ \begin{bmatrix} C_3^{\sim\sim} & C_3^{\sim -} \\ C_3^{-\sim} & C_3^{--} \end{bmatrix} = \left(\Sigma_{\theta_{t+1}^*}^*\right)^{-1}$$

$$\begin{bmatrix} C_4^{\sim\sim} & C_4^{\sim -} \\ C_4^{-\sim} & C_4^{--} \end{bmatrix} = \left[\Sigma_t^* + \begin{bmatrix} \left(\Lambda_{t+1}^{\sim}\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right]^{-1} \quad \forall\, t \in [T-1, 0]$$

$$\sigma = \frac{T}{\sum_{t=1}^{T} \left[y_t^T y_t + \left(u_t^{\&}\right)^T \left\{\hat{\theta}_t^{\&} \left(\hat{\theta}_t^{\&}\right)^T + \Sigma_{\theta_t^{\&} \theta_t^{\&}}\right\} u_t^{\&} - 2 y_t^T \left(u_t^{\&}\right)^T \hat{\theta}_t^{\&}\right]}$$