Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Variational Bayes Approximation for Inverse Regression Problems

A Thesis submitted to the University of Dublin, Trinity College

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

School of Computer Science and Statistics

Trinity College Dublin, Ireland

April 2011

Richa Vatsa

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Richa Vatsa.

Richa Vatsa

Richa Vatsa

Trinity College Dublin,

April 2011.

Summary

Inverse regression is a tool to predict an unknown explanatory variable for given observations of a response variable in a regression problem. The prediction problem is usually carried out in two stages: firstly, to fit the model relationship between the variables, and secondly, to predict the unknown explanatory variable. Both the problems, model fitting and prediction involve considerable computational burden. Previous work on the Bayesian approach to the problem have used MCMC, INLA and other numerical methods. This thesis aims to present an alternative fast variational Bayes (VB) approximation to Bayesian inference for inverse regression problems which claims to avoid the limitations of previous work. The VB method assumes independence between the parameters in the posterior distribution, thus provides fast approximations to Bayesian estimation problems. In contrast to INLA, it can be applied to models with many unknown parameters. In the thesis, the VB method is applied to a wider class of inverse regression problems classified into two classes: inverse latent regression and inverse non-latent regression which present challenges for the method's accuracy and tractability. The VB method itself is not without limitations. Quick VB solutions are obtained at the cost of some loss of accuracy. Also, tractable application of the method is limited to conjugate-exponential (CE) models. It is attempted to increase the accuracy and tractability of the method outside CE models with the use of further approximations, such as a Gaussian approximation.

Acknowledgements

First and foremost I would like to express my deep gratitude to my supervisor, Prof. Simon Wilson, for his invaluable guidance throughout the research. Without his support I would have never been able to come so far. He has always motivated me and has given every opportunity to explore and excel in the research field. His patient and organized way of working has taught me how to stay focused. Remembering those early days of my research, when sometimes I would become very impatient and start feeling that I had nothing to contribute, I can never forget his words, "Rome was not built in a day". I feel very lucky to work under Simon who has always time for his students, no matter how small a query is.

I am very grateful to Prof. John Haslett for helping me understand the palaeo-climate reconstruction used in the thesis as an important problem to explain the inverse regression. I would like to thank Dr. Brett Houlding and James Sweeney who have helped me with their useful discussions. A big thank to my colleagues for all the fun we had together.

I thank my very supportive parents for their enormous faith in me. Even miles away their love and care is always with me encouraging, motivating and making me feel very proud.

This work has been made possible by Science Foundation Ireland Research Frontiers Programme, grant number 05/RFP/MAT053 and the STATICA project, a Principal Investigator program of Science Foundation Ireland, Grant number 08/IN.1/I1879.

Richa Vatsa

Trinity College Dublin,

April 2011.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Inverse regression forms a class of statistical calibration problems that has important application in the fields of geology, biomedicine, archeology, palaeoclimatology, econometrics, astronomy etc. Interest lies in predicting the explanatory variable for a new observation(s) of the response variable. For example, consider a classical example of calibration for a dose of vitamins and weight gain of a chicken. It might be of great importance to infer the unknown amount of vitamins responsible for a desired weight gain. Another example of an inverse regression problem may be to estimate the unknown age of a specimen (in archeology), or to predict past climate in palaeoclimatology. Calibration problems can be very time consuming, expensive or very difficult to implement. Inverse regression performs the calibration in two stages. The first stage fits the model relationship between the variables, and the second stage uses the model fitting to predict the unknown explanatory variable for new values of the response variable. Both the stages of inference involve considerable computational burden. In this thesis the inference problem is performed in the Bayesian framework which is often analytically intractable or computationally expensive. The aim of the thesis to present fast approximations to Bayesian inference for multi-dimensional and complex inverse regression problems.

Hoadley (1970) proposed a Bayesian solution to an inverse linear regression problem, emphasizing the importance of a proper choice of the prior distribution in case of few data. The author did not follow the usual Bayesian approach for parameter estimation (model fitting) but stuck to least square estimates of the regression pa-

rameters to avoid the difficult Bayesian computation. Hunter & Lamboy (1981) also added their contribution to the Bayesian study of inverse linear regression problems. They used Bayes theorem and derived a joint posterior density of the regression parameters and the explanatory variable. The author had to depend on a Gaussian approximation to compare the result with previous Bayesian solutions to the inverse problem. The methods by Hoadley (1970) and Hunter & Lamboy (1981) are restricted to only simple lower-dimensional inverse linear regression problems. Racine-Poon (1988) discussed a Bayesian solution to a multi-dimensional inverse non-linear regression (non-linear calibration) problem that used a Gaussian approximation for the model fitting. This was based on the Gauss-quadrature approach and Laplace approximation to evaluate an integral which may not be appropriate if the parameters departs from Gaussianity. Also, it is suitable for only low-dimensional inverse regression problems as it depends on a numerical integration method to predict the explanatory variables.

The prediction problem of Haslett et al. (2006) can be termed as inverse latent regression. It used the inverse regression to predict past climate of a (palaeoclimate) latent model. Haslett et al. (2006) used Markov chain Monte Carlo (MCMC) but it did suffer from slow convergence. Salter-Townshend (2009) applied integrated nested Laplace approximation (INLA) of Rue. et al. (2009) for the palaeoclimate reconstruction problem of Haslett et al. (2006). The INLA method provides quite accurate results but is limited to Gaussian latent models with a few number of parameters. Vatsa & Wilson (2010) presented a variational Bayes approximation for the palaeoclimate reconstruction problem of Haslett et al. (2006). The method of Vatsa & Wilson (2010) can be applied to the models with several parameters but it may be slower if there are many predictions to be studied independently at a time.

This thesis aims to provide an alternative fast Bayesian approximation to a wider class of inverse regression problems (inverse latent and non-latent regression) using the variational Bayes (VB) method. Contrary to the method of Vatsa & Wilson (2010), the work in the thesis uses the concept of two stages of inference and presents the VB approximations for the parameter estimation and for the prediction problem separately. It is believed that the method avoids the limitations of previous work

in the field of inverse regression problems, though the method itself is not free from some drawbacks. The thesis explores the accuracy and the tractability of the method and attempts to make it amenable to complex models with a little compromise in the computational speed.

1.1 Overview of the chapters

A brief outline of the research presented by chapters in this thesis follows.

Chapter 2: Statistical Methodology

The inference work of the thesis is carried out in the Bayesian framework hence a brief introduction to the Bayesian statistics is given in the chapter. Some popular methods of Bayesian computation classified into two categories: simulation based method e.g. Monte Carlo methods, and functional approximation methods such as INLA, Laplace approximation and Gaussian approximation, are discussed, which provides a base to the comparative study of the variational Bayes method used in the thesis for Bayesian inference in the inverse regression problems .

Chapter 3: The variational Bayes Approximation

The variational Bayes (VB) method used throughout the thesis, is described in the chapter. The background and the past study of the method is briefly discussed. A comparative study of two approaches of the method by Beal (2003) and Šmídl & Quinn (2006) respectively is presented. Two other variational methods, Gaussian variational approach and variational tangent approach, are also discussed briefly which are used in the thesis to compare the results by the VB method. It is believed that the method provides fast approximations to Bayesian inference problems. Therefore, the VB method is compared with other Bayesian computation methods e.g. Laplace approximation, INLA, MCMC method, for its accuracy, speed and tractability.

Chapter 4: VB approximation for Inverse Non-latent Linear Regression

The aim of the chapter is to provide the VB approximations for non-latent regression problems. The inverse non-latent regression is introduced through two type of regression models: conjugate-exponential non-latent models and non-conjugate-exponential non-latent models. The conjugate-exponential non-latent models are explained via two regression models, simple linear and quadratic regression. The non-conjugate-exponential non-latent models are described through Poisson regression, mixture of Poisson regression and zero-inflated Poisson regression models. Bayesian analysis of inverse regression is described. The VB approximation to reduce the complexity of the Bayesian computational problems in inverse non-latent regression problem, is presented for inverse simple, quadratic, Poisson, mixture of Poisson and zero-inflated Poisson regression problems. The intractability issue of the VB method is explored for non-conjugate-exponential models. The VB approximations for the inverse non-latent regression problems are performed using simulated data. The VB result is further compared with the results from other methods, e.g. MCMC, variational tangent approach, Gaussian variational tangent approach and that of the previous work in the field of inverse linear regression problems.

Chapter 5: VB approximation for Inverse Latent Regression

Chapter 5 describes the VB approximation inverse latent regression models. Two types of latent regression models are considered: latent non-random effects models and latent random effect models. The Poisson latent regression model is considered to explain the latent non-random effects models. The latent random effects models are explained by two models: Poisson latent random effect model and zero-inflated Poisson random effect models. The same intractability issue of the VB method for non-conjugate-exponential model is further explored via these three complex models. An algorithm based on a Gaussian approximation of Rue & Held (2005) and the VB method is developed to deal with the intractability problem. The VB approxi-

mations for these inverse latent regression problems are performed using simulated data. The VB results are compared with the results by INLA for the accuracy and the computational time of the VB method.

Chapter 6: VB approximation for Palaeoclimate Reconstruction problem

The VB approximation for a complicated inverse latent regression is described via the palaeoclimate reconstruction problem of Haslett et al. (2006). The palaeoclimate reconstruction is an example of a complex latent regression problem that challenges the tractability and accuracy of the VB approximation. The VB solution for the inverse latent regression problem with the palaeoclimate data provides some insightful study of the limitations of the approximation which leaves some room for future work in this field.

1.2 Research Contributions

The research contributions of this thesis are listed as follows:

1. Inverse regression problems are explored and studied via two categories of models: latent models and non-latent models. In past literature, the inverse method is applied mostly in linear regression problems. The thesis extends the class of inverse regression problems to latent and non-linear models and presents a Bayesian solution.

2. The variational Bayes approximation is presented for the Bayesian inference in a wider class of the inverse regression problems. In Chapter 4, the VB approximations for inverse (non-latent) linear, quadratic and Poisson, mixture of Poisson, zero-inflated Poisson regression models are presented. In Chapter 5, The VB approximation for inverse latent regression problem is presented for three models: Poisson latent model, Poisson latent random effect model and zero-inflated Poisson latent random effect model. The VB approximation for a very complex multi-dimensional palaeoclimate reconstruction problem (an example of inverse latent regression models) is presented in Chapter 6.

3. The inverse non-latent (Poisson, mixture of Poisson and zero-inflated Poisson) and inverse latent regression models allow us to explore the tractability of the VB method for non-conjugate-exponential models. An attempt to reduce the intractability of the method for such complex models is made through further approximations such as a Gaussian approximation. The accuracy of this approximation is explored.

Chapter 2

Statistical Methodology

In this chapter, to have a better understanding of the inference procedures used throughout the thesis, relevant statistical methods are introduced briefly. In Section 2.1, the Bayesian statistical methodology is described, which is the inference approach of the thesis. Section 2.2 discusses Gaussian Markov random fields including the Markov property to help explaining the approximations to statistical inference methods described in the next sections. In Section 2.3, a directed acyclic graph is defined for a simple understanding of a complex model. Methods to approximate statistical inference are classified into two categories: simulation based and functional approximations, defined in Sections 2.4 and 2.5 respectively. Some other statistical tools are explained in Section 2.6.

2.1 Bayesian Inference

In Bayesian inference, probability quantifies uncertainty and is interpreted to be subjective. Therefore state of knowledge about any unknown is given by a probability distribution that is one's degree of belief. Bayesian inference is a way to modify one's belief in light of observed data by using Bayes' theorem. A modified belief is then called posterior belief and provides the probabilistic inference about an unknown parameter in light of data. Prior and posterior belief are quantified as prior and posterior distribution. A detailed introduction on Bayesian inference can be found in Lee (2004), Bernardo & Smith (1994), Box & Tiao (1992).

2.1.1 Prior Distribution

As mentioned above, the prior distribution is one's belief based on prior assumptions (or knowledge) about a parameter. Suppose interest lies in making inference on a set of unknown parameters denoted by $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\}$. If prior knowledge about the parameter θ from past study or expert's opinion is considered prior to observing data, it provides a probabilistic statement about θ and is denoted as $P(\theta)$.

As prior belief is subjective, it is an important question how to choose a prior density $P(\theta)$ in order to make an inference about θ. If the size of data is small, the nature of inference is much influenced by the choice of prior. On the other hand, the prior distribution has little impact on inference if the data size is large. Different choices of priors, termed as prior elicitation, are discussed in detail later in the chapter.

2.1.2 Likelihood

A model defining a relationship between a parameter θ and a set of observed data denoted as $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ can be expressed via a conditional probabilistic statement or a probability density function (p.d.f) denoted as $P(\mathbf{y}|\theta)$. The term $P(\mathbf{y}|\theta)$ is a function of data $\mathbf{y}$ given a fixed (unknown) value of θ. Likelihood is a function of θ given data $\mathbf{y}$, denoted by $L(\theta|\mathbf{y})$, to draw inference about θ such that it should extract all the possible information provided by data. It can be expressed as:

$$L(\theta|\mathbf{y}) = P(\mathbf{y}|\theta), \tag{2.1}$$

for i.i.d $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$,

$$L(\theta|\mathbf{y}) = \prod_{j=1}^{n} P(y_j|\theta). \tag{2.2}$$

The likelihood $L(\theta|\mathbf{y})$ is not a probabilistic statement over θ given $\mathbf{y}$. Unlike $P(\mathbf{y}|\theta)$, it is not a p.d.f of θ. It should not be expected to integrate to one.

2.1.3 Posterior Distribution

A posterior distribution can be interpreted as belief or knowledge about ar... parameter in light of observed data. It is a probability density function ... known parameter θ given data $\mathbf{y}$, denoted by $P(\theta|\mathbf{y})$. Given the likelihood $L(\theta|\mathbf{y})$ and a prior distribution $P(\theta)$, the posterior distribution $P(\theta|\mathbf{y})$ can be computed by Bayes' law as:

$$P(\theta|\mathbf{y}) = \frac{P(\theta)P(\mathbf{y}|\theta)}{P(\mathbf{y})}, \tag{2.3}$$

$$\propto P(\theta)P(\mathbf{y}|\theta). \tag{2.4}$$

The term $P(\mathbf{y})$ is called the marginal likelihood of data and is expressed as:

$$P(\mathbf{y}) = \int_{\theta} P(\theta)P(\mathbf{y}|\theta)\mathrm{d}\theta. \tag{2.5}$$

The marginal likelihood is c... ...ng, e.g. Bayes' factor.

The posterior distribution $P(\theta|\mathbf{y})$ can be expressed as proportional to the product of $P(\theta)$ and $P(\mathbf{y}|\theta)$. One should normalize $P(\theta|\mathbf{y})$ to express it as a p.d.f.

In most statistical problems, the parameter θ is multidimensional; $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\}$. Bayesian inference on the components of θ can be studied through their marginal posterior distributions. To compute the marginal distributions, integrals over the dimension of θ are required:

$$P(\theta_i|\mathbf{y}) = \int_{\theta_{-i}} P(\theta|\mathbf{y})\mathrm{d}\theta_{-i}, \tag{2.6}$$

where $\theta_{-i} = \{\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p\}$. Given the data and the prior belief, a summary of the knowledge about a parameter can be given by a point (posterior) estimate and an interval estimate. These estimates are some function of posterior distributions which again requires the computation of some multi-dimensional integrals. The expectation of a function of θ, denoted as $h(\theta)$ can be defined as:

$$\mathbb{E}(h(\theta)|\mathbf{y}) = \int_{\theta} h(\theta)P(\theta|\mathbf{y})\mathrm{d}\theta. \tag{2.7}$$

Solving integrals is therefore a necessary task in Bayesian computation. More often, the multi-dimensional integrals are either computationally intensive or intractable. In such cases, we rely on distributional or simulation type approximations discussed later in the chapter.

2.1.4 Prior Elicitation

Prior elicitation is a process of choosing a prior distribution that represents one's belief about an unknown. To make inference on an unknown, one's belief should be extracted and incorporated into a probabilistic statement. Selection of priors is an important question as it may have an impact on inference result. There are some important classes describing different properties of priors:

1. For computational ease, priors are often chosen such that the prior and the posterior belong to the same class of distributions. Such priors are called **conjugate** for the likelihood. Conjugate priors may not represent one's belief accurately, but are chosen for the tractability of the corresponding posterior distribution (Lee, 2004). Given a standard form of distributions a conjugate prior is proper, though, it may belong to a non-informative class of priors (see below) depending on the values of its hyper-parameters. The choice of hyper-parameters play a strong role in defining a prior. If substantial prior knowledge about an unknown is available, the choice of prior should reflect ones's prior belief. In case of 'no or little knowledge' about an unknown, the hyper-parameters should be defined so as to present a non-informative prior. For example, with a large variance, a normal or a Gaussian prior is proper but reflects little knowledge about an unknown parameter.

2. A proper prior with a strong prior belief about an unknown, not necessarily be conjugate, can be termed as an **informative** prior. It should match to one's belief, as with insufficient data a wrong selection of prior may lead to an inappropriate posterior distribution.

3. In the case of a little prior knowledge, a prior should be designed to express the ignorance about an unknown. Reference priors, locally uniform priors and

Jeffery's prior are some examples of **non-informative** priors. These priors do not belong to the family of standard distributions. They are **improper** densities, i.e. they do not sum or integrate to one. An example of such priors could be a uniform density over an infinite range, reflecting no specific prior knowledge about the parameter. Improper priors may yield proper posterior distributions with a sufficiently informative data likelihood. A detailed discussion on such non-informative and improper priors can be found in Box & Tiao (1992).

Different classes of priors are discussed in Lee (2004) in detail.

2.1.5 Predictive Posterior Distribution

In some cases, the interest also lies in making prediction on future observations. A predictive posterior distribution is a distribution of future observation given current data. The prior distribution of a future observation is the same as the p.d.f of current data. If $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ are current data and θ is the model parameter, the predictive posterior distribution of future observation $\mathbf{y}_{n+1}$ is:

$$\begin{aligned} P(\mathbf{y}_{n+1}|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) &= \int_\theta P(\mathbf{y}_{n+1}|\theta, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) P(\theta|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) \mathrm{d}\theta, \\ &= \int_\theta P(\mathbf{y}_{n+1}|\theta) P(\theta|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) \mathrm{d}\theta, \end{aligned} \tag{2.8}$$

under the assumption that the observations are independent conditional on θ.

The posterior distribution of θ given current data should be known in advance in order to compute the predictive posterior distribution.

2.1.6 Maximum a Posteriori estimate

The maximum a posteriori (MAP) estimate is the global mode of a posterior distribution of θ. It is defined as:

$$\theta_{MAP} = \arg\max_\theta P(\theta) P(\mathbf{y}|\theta). \tag{2.9}$$

A MAP estimate plays an important role in Bayesian analysis. Even if the posterior distribution is not in closed form a MAP estimate can be found by some optimization methods, e.g, Newton's optimization method. If the prior distribution is non-informative the MAP reduces to the ML estimation of the parameter. A MAP estimate may be used to approximate the intractable posterior distribution, such as a Laplace approximation or a Gaussian approximation.

2.1.7 Highest Posterior density region

The posterior knowledge of a parameter contained in its posterior distribution can be summarized through point and interval estimates. An interval estimate provides a summary of posterior uncertainty of the parameter. The interest may lie in specifying an interval which includes most of the posterior density. Such an interval should also be as short as possible. It should be defined in such a way that the density at any point inside the interval is greater than the density at any point outside it. A highest posterior density (HPD) region is an interval that is the shortest interval that contains a given probability mass.

2.2 Gaussian Markov Random Fields

Often in spatial statistics, a vector of latent variables is modelled as a multivariate Gaussian field. A latent field following a multivariate Gaussian distribution with a Markov property is termed as Gaussian Markov Random Field (GMRF) (Rue & Held, 2005).

More precisely, a latent field $\mathbf{Z} = (\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_N)^T$ is called a GMRF with respect to a labeled graph $\mathcal{G} = (\mathcal{V}, w)$ with mean μ and precision $Q > 0$ iff its density form is as follows:

$$P(\mathbf{Z}) = (2\pi)^{-\frac{N}{2}} (\det Q)^{\frac{1}{2}} \exp\left[(\mathbf{Z} - \mu)^T Q (\mathbf{Z} - \mu)\right], \tag{2.10}$$

where Q is the sparse precision matrix such that

$$Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in w; \forall i \neq j. \tag{2.11}$$

A labeled graph $\mathcal{G} = \{\mathcal{V}, w\}$ defines the Markov structure of $\mathbf{Z}$ in a discrete location space, where $\mathcal{V}$ indexes the locations and w is the set of edges reflecting dependency connection from one node to another in $\mathcal{G}$. If all the nodes are connected to each other, it makes the graph fully connected. If a $\eta(i)$ define a set of neighbours of the node i, a node j outside $\eta(i)$ is conditionally independent assuming a Markov property, i.e.

$$\eta(i) = \text{neighbours of } i, \tag{2.12}$$

$$\text{for all } j \neq \eta(i) \qquad \mathrm{Z}_i \perp \mathrm{Z}_j \mid \mathrm{Z}_{\eta(i)}. \tag{2.13}$$

This conditional independence makes the precision matrix sparse that eases the complexity associated with the matrix computations. In Bayesian computation, GMRFs are very useful. If a prior is a GMRF, the posterior is also a GMRF if the likelihood is Gaussian. The Markov property, that leads to sparseness in the precision matrix, greatly reduces the computational complexity in Bayesian estimation problems.

2.3 Directed Acyclic Graph

In multivariate statistics, a complex model often requires a lot of attention to understand its structure algebraically. It is also not always easy to grasp how the model factorizes over its variables. Graphical representation techniques are often used for the simple understanding of the multivariate models. A directed acyclic graph (DAG) is one of the graphical representations of the probabilistic models that visualizes the structure of the model and gives the useful insights into the properties of the model such as conditional independence. A DAG is shown by nodes and arrows (in a particular direction) which represent the random variables of the model and their causal relationships.

Consider a dense model with a set of random variables $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p\}$. The joint distribution of the variables is defined as:

$$P(\mathbf{X}) = P(\mathbf{X}_p|\mathbf{X}_{p-1}, \ldots, \mathbf{X}_2, \mathbf{X}_1)P(\mathbf{X}_2|\mathbf{X}_1)P(\mathbf{X}_1). \tag{2.14}$$

A graphical representation of the model presents a better understanding of the causal relationship between the random variables. With the help of a graphical representation it becomes easier to visualize the joint distribution decomposed into the conditional distributions. The joint distribution given in Eq. 2.14 can also be explained with a help of a DAG. Fig. 2.1 shows a DAG which represents the way the random variables are connected in the joint distribution. The variables in the DAG are fully connected since there is no link missing between the variables in acyclic order.

Consider a factored model as given below:

$$P(\mathbf{X}, \theta) = \prod_{i=1}^{p} [P(\mathbf{X}_i|\theta)]\, P(\theta) \tag{2.15}$$

Given θ, $\mathbf{X}_i'$s are independent of each other. The graphs shown in Fig. 2.1 is fully connected and Fig. 2.2 represents a factored model under the conditional independence property. The conditional property of a loosely (not fully) connected model will be described in the next section.

2.3.1 Conditional Independence

Consider a simple factored model (represented by the DAG shown in Figure 2.3) with variables a, b, c, d and e and missing links between some of the variables to explain the conditional property of the model. The joint distribution of the factored model can be expressed as:

$$P(a, b, c, d, e) = P(e|d, c, a)P(d|c, a)P(c|a)P(b|a)P(a) \tag{2.16}$$

The same model is represented by a DAG in Fig. 2.3. In the DAG, there is a missing link between the nodes b and c. They are connected only through the node

a. Observing a nodes b and c are conditionally independent which is also reflected in the algebraic expression of the joint distribution in Eq. 2.16.

For complex models (or models with several variables), there should be a defined rule (or a set of rules) according to which the conditional independence property of the graph could be explained. In the next section, the d-separation property of DAGs is explained in detail that describes the rules to explain the conditional property of the factored graphs as a result of missing links between variables and their observing status.

d-separation

The general framework to explain the conditional property of the multivariate models is called d-separation (direct separation). To understand the concept of d-separation, consider three types of graphs represented in Fig. 2.4, 2.5 and 2.6.

Example 1

The joint distribution corresponding to the graph in Fig. 2.4 can be defined as:

$$P(a, b, c) = P(a|c)P(b|c)P(c). \tag{2.17}$$

The joint distribution of a and b is obtained by marginalizing $P(a, b, c)$ over c.

$$P(a, b) = \sum_{c} P(a|c)P(b|c)P(c). \tag{2.18}$$

The joint distribution $P(a, b)$ does not factorize into marginal distributions $P(a)$ and $P(b)$. The variables a and b are independent only if the variable c is observed. Thus, the joint distribution of a and b given c can be explained via the product of the marginal distributions of a and b given c respectively,

$$P(a, b|c) = \frac{P(a, b, c)}{P(c)}, \tag{2.19}$$

$$= P(a|c)P(b|c). \tag{2.20}$$

Hence, the conditional property is achieved as $a \perp b \mid c$.

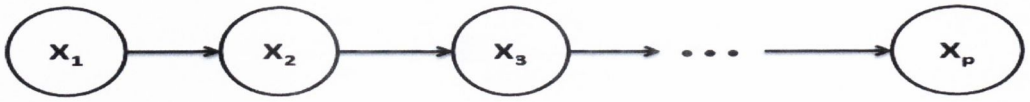

Fig. 2.1: A directed acyclic graph representing a fully connected model with variables $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p\}$.

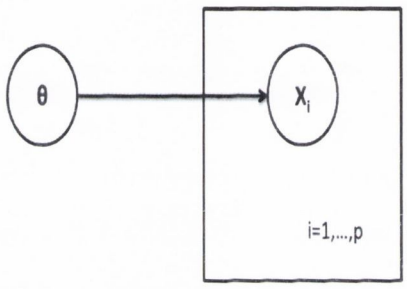

Fig. 2.2: A DAG representing a factorized plate model given in Eq. 2.15.

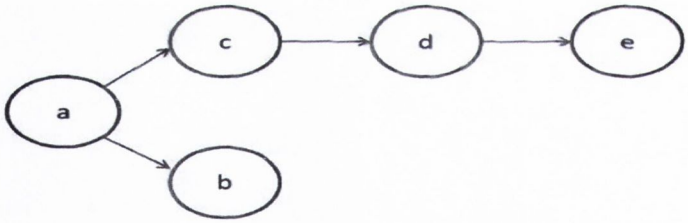

Fig. 2.3: A DAG representing a factored model defined in Eq. 2.16

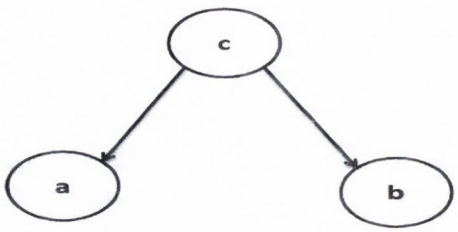

Fig. 2.4: Tail-to-tail DAG with variables a, b and c.

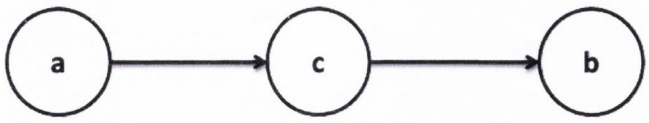

Fig. 2.5: Head-to-tail DAG with variables a, b and c.

In the graphical representation given in Fig. 2.4 the path from a to b is considered via c. The node c is said to be tail-to-tail with respect to the path from a to b as the nodes are connected to c through the tail of the two arrows. If c is not observed the path from a to b makes them dependent. However, if c is observed the path (conditioned on c) is blocked and it causes a and b independent.

Example 2

In Fig. 2.5, the nodes a and b are connected via c through the arrows from a to c and c to b respectively. The joint distribution of the variables in this graph is defined as:

$$P(a,b,c) = P(a)P(c|a)P(b|c). \tag{2.21}$$

Marginalizing over c, the joint distribution of a and b is given as:

$$\begin{aligned} P(a,b) &= P(a)\sum_{c} P(c|a)P(b|c), &(2.22)\\ &= P(a)P(b|a). &(2.23) \end{aligned}$$

Therefore if c is unobserved, the variables a and b are dependent. Now let c is observed. The joint distribution of a and b given c is as follows:

$$\begin{aligned} P(a,b|c) &= \frac{P(a)P(c|a)P(b|c)}{P(c)}, &(2.24)\\ &= P(a|c)P(b|c). &(2.25) \end{aligned}$$

Thus, we again obtain a and b independent given c with respect to the graph in Fig. 2.5.

The node c in the graph is called head-to-tail with respect to the path from a to b making them dependent. If c is observed, the path gets blocked and a and b

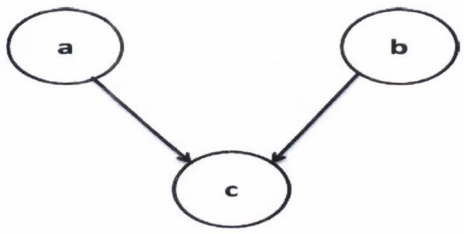

Fig. 2.6: Head-to-head DAG with variables a, b and c.

become independent of each other, i.e. $a \perp b \mid c$. Thus, this head-to-tail relationship defines the conditional property of the model.

Example 3

Now, consider the graph in Fig. 2.6 that shows that the nodes a and b are connected via c through the heads of the arrows from a to c and from b to c respectively. The joint distribution corresponding to this graph can be expressed as:

$$P(a, b, c) = P(a)P(b)P(c|a, b). \tag{2.26}$$

The joint distribution of a and b after marginalizing over c is given as:

$$\begin{aligned} P(a, b) &= P(a)P(b)\sum_{c} P(c|a, b), & (2.27) \\ &= P(a)P(b). & (2.28) \end{aligned}$$

This shows that if c is unobserved, a and b are independent in their joint distribution. Now suppose c is observed. Consider the joint distribution of a and b given c as:

$$P(a, b|c) = P(a)P(b)P(c|a, b). \tag{2.29}$$

Thus the joint distribution of a and b given c does not decomposed into the marginal distributions of a and b. Observing c unblocks the path from a to b making the variables dependent. The node c is said to be head-to-head with respect to the path from a to b via itself if the nodes are connected to c through the heads of the two arrows.

These three examples states the key concept of the d-separation for the directed acyclic graphs. Consider a DAG with non-intersecting sets of nodes A, B and C to define the d-separation property. To obtain the conditional independence between A and B given C, i.e. $A \perp B \mid C$, all the paths from A to B should be blocked. A path is blocked if it includes a node (or nodes) satisfying either of the following two conditions:

1. if the node is in C and the path via the node is defined through tail-to-tail or head-to-tail property,

2. if the node and its decedent nodes are not in C and the path via the node is defined through head-to-head property.

Thus A is said to be d-separated from B by C if all the paths are blocked which follows to the independence property between A and B given C, $A \perp B \mid C$.

Conditional independence and Markov property

The conditional independence property of a DAG can also be understood via the Markov property. Consider a model with a set of random variables $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p\}$. The joint distribution of $\mathbf{X}$ is given as:

$$P(\mathbf{X}) = P(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p), \tag{2.30}$$

$$= \prod_{i=1}^{p} P(\mathbf{X}_i | pa_i). \tag{2.31}$$

The term pa_i in the conditional distribution $P(\mathbf{X}_i|pa_i)$ includes the parents, the children and the co-parent nodes of $\mathbf{X}_i$. In other words, given pa_i $\mathbf{X}_i$ is independent of other nodes which are not included in pa_i. This simplifies the complex structure of a conditional distribution can be simplified. The set of nodes in pa_i, including parent, children and co-parent, are called Markov blanket of node $\mathbf{X}_i$. For example, in Fig. 2.7 the set of parent nodes of $\mathbf{X}_3$ is $pa_i = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_4, \mathbf{X}_5\}$.

Thus the concept of the Markov blanket stating the conditional independence property of a DAG makes it very useful for the methods/models based on the Markov property such as the MCMC method (described in the next section).

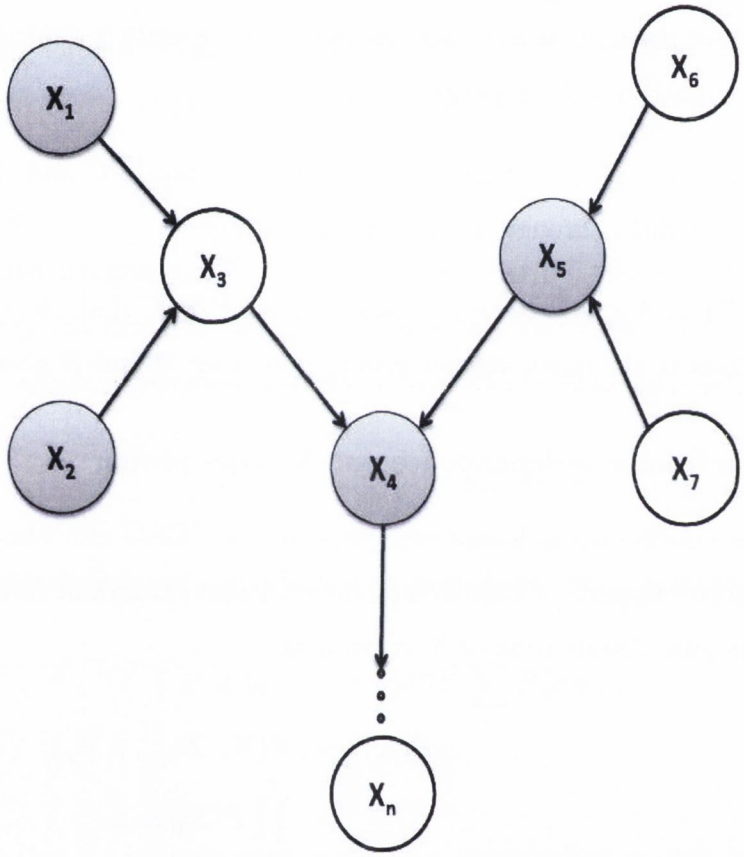

Fig. 2.7: A DAG showing conditional independence by the Markov property for a multivariate model. The DAG represents a Markov blanket around the node $\mathbf{X}_3$ with the shaded nodes as its parents, children and co-parents. Node $\mathbf{X}_4$ is a child, nodes $\mathbf{X}_1$ and $\mathbf{X}_2$ are parents and $\mathbf{X}_5$ is the co-parent. $\mathbf{X}_3$ is conditionally independent of nodes $\mathbf{X}_6$ and $\mathbf{X}_7$ given its parents.

2.4 Monte Carlo methods

Monte Carlo methods are numerical methods to simulate random realizations from distributions. These simulations are further used in the evaluation of integrals that are the cause of intractability of Bayesian inference problems, e.g. in the computation of the normalizing constant in the Bayes' law, posterior expectations, marginal posterior distributions etc. There are two commonly used classes of Monte Carlo methods: direct Monte Carlo and Markov chain Monte Carlo (MCMC) methods, depending on the nature of the posterior distribution to simulate from. The MCMC method can be described by the algorithm Metropolis Hastings and its special case, Gibbs sampler, discussed in detail later in this section.

The accuracy of the Monte Carlo approximations is based on the property of the strong law of large numbers (SLLN).

Strong Law of Large Numbers: If $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is an infinite sequence of i.i.d random variables with finite expected value, i.e. $\mathbb{E}(\mathbf{X}_1) = \mathbb{E}(\mathbf{X}_2) = \ldots = \mathbb{E}(\mathbf{X}) < \infty$. Then

$$\frac{\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_n}{n} \xrightarrow{a.s} \mathbb{E}(\mathbf{X}) \text{ as } n \to \infty. \tag{2.32}$$

This means that the Monte Carlo methods converge almost surely for a large number of simulations by the strong law of large numbers (SLLN).

The rate of convergence of the Monte Carlo methods can be defined through the central limit theorem (CLT).

Central Limit Theorem (CLT): If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is a sequence of i.i.d finite random variables with finite mean and variance, i.e. $\mathbb{E}(\mathbf{X}_1) = \mathbb{E}(\mathbf{X}_2) = \ldots = \mathbb{E}(\mathbf{X}_n) = \mu < \infty$ and $V(\mathbf{X}_1) = V(\mathbf{X}_2) = \ldots = \mathbb{E}(\mathbf{X}_n) = \sigma^2 < \infty$, then

$$\frac{\sqrt{n}}{\sigma}\left(\frac{\sum_i \mathbf{X}_i}{n} - \mu\right) \xrightarrow{D} N(0,1). \tag{2.33}$$

This follows that the average of the sequence of i.i.d random variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ asymptotically follows a Normal distribution with some finite mean μ and variance

$\frac{\sigma^2}{n}$, as n tends to infinity.

The condition of the identical distribution can be relaxed under Lindeberg's condition (Billingsley, 1986) which is a sufficient condition for CLT to hold for the sequence of the (finite) independent random variables.

Lindeberg's condition: If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is a sequence of finite independent variables with finite mean $\mathbb{E}(\mathbf{X}_k) = \mu_k$ and variance $V(\mathbf{X}_k) = \sigma_k^2$, suppose $s_n^2 = \sum_{k=1}^n \sigma_k^2$, then

$$\frac{1}{s_n^2} \sum_{k=1}^{n} (\mathbf{X}_k - \mu_k)^2 \xrightarrow{D} N(0, 1). \tag{2.34}$$

That is, the random variable $\frac{\sum_{k=1}^n (\mathbf{X}_k - \mu_k)}{s_n} = \mathbf{Z}_n$ (say) converges in distribution to a standard normal distribution as n tends to infinity.

By CLT, the rate of the convergence of the Monte Carlo methods is of order $n^{-\frac{1}{2}}$ and the error of the approximation can be given by $\frac{\sigma}{\sqrt{n}}$ (or s_n more generally). This means that the error in the approximation approaches to a small value with a large number of simulations.

There are other simulation based methods such as importance sampling, rejection method, inverse transformation methods that simulate directly from the posterior distribution. These may only work for low-dimensional problems but are used within MCMC algorithms.

2.4.1 Monte Carlo Integration

Monte Carlo integration approximates integrals by sample averages. The accuracy of the Monte Carlo approximation depends on the number of independent samples used (Robert & Casella, 1999). For example, the integrals in Eq. 2.7 for the posterior expectation can be approximated as:

$$\mathbb{E}(h(\theta)|\mathbf{y}) \approx \frac{1}{M} \sum_{m=1}^{M} h(\theta^m), \tag{2.35}$$

where $\theta = (\theta^m;\ m = 1 : M)$ are independent samples from the standard posterior distribution of θ.

The marginal posterior distributions can also be approximated in the same way,

as:

$$P(\theta_i|\mathbf{y}) \approx \frac{1}{M}\sum_{m=1}^{M} P(\theta_i|\theta_{-i}^m, \mathbf{y}). \tag{2.36}$$

The term θ_{-i}^m in the above equation, represents m^{th} independent samples from the joint posterior distribution $\pi(\cdot)$ of θ_{-i}.

By the strong law of large numbers, the accuracy can be increased by making M very large. The Monte Carlo integration will work even if the samples are not independent, as in the case with MCMC methods (Gilks et al., 1996). Monte Carlo integration is widely used in Bayesian inference since many integrals of interest can be approximated by expectations e.g. in evaluating a predictive posterior distribution for a new observation, calculating posterior moments, computing marginal posterior distributions.

2.4.2 Markov chain Monte Carlo methods

It may be difficult to efficiently draw independent samples from a distribution $\pi(\cdot)$. However, the samples need not necessarily be independent under the *ergodic theorem* defined later in the section . In such cases, Markov chain Monte Carlo methods (MCMC) can be used to draw dependent samples from an approximation of $\pi(\cdot)$.

MCMC uses the concept of the Markov property to generate from intractable or non-standard distributions. The samples of θ are drawn sequentially from a conditional distribution, $P(\cdot|\cdot)$, such that given current sample θ^t, the conditional distribution of future sample θ^{t+1} is independent of all past samples. Assuming θ a discrete variable (belonging to a discrete state space), the Markov property is defined as:

$$P(\theta^{t+1} = i|\theta^t, \theta^{t-1}, \ldots, \theta^1) = P(\theta^{t+1} = i|\theta^t).$$

Such a sequence of samples is called a Markov chain and $P(\cdot|\cdot)$ with the Markov property is called a transition distribution. The property can be similarly defined for a continuous state space.

It is assumed that the chain is time-homogeneous, i.e. $P(\cdot|\cdot)$ is independent of t. As t grows, after a period of transitions often called the burn-in period, the Markov chain of dependent samples forgets its initial state and converges asymptotically to a

unique stationary distribution which is the target distribution $\pi(\cdot)$. To converge to a unique stationary distribution the chain needs to satisfy three important properties: irreducibility, aperiodicity and positive recurrence. The definition of these properties are given below:

1. **Positive recurrence:** A state i (of a Markov chain) is called positive recurrent state if the Markov chain starting in state i will return back to i in a finite time with probability one. If all the states of a Markov chain are positive recurrent, it is called positive recurrent.

2. **Irreducibility:** A Markov chain is irreducible if for any starting state i of a state space S, there exits a t such that $P_{ij}^t = P(\theta^t = j|\theta^0 = i) > 0 \; \forall \; j \in S$.

   That is, from any starting point the irreducible Markov chain is eventually able to reach to any region of the state space with a positive probability.

3. **Aperiodicity:** A period d_i of a state i is a greatest common divisor of the set $\{t : P_{ij}^t > 0\}$. Hence it is only possible to return to the state i in a d_i multiple number of steps. If $d_i = 1 \; \forall i \in S$, the Markov chain is called aperiodic. That is, all the states of the chain return to themselves only at irregular intervals. The property of aperiodicity ensures that the chain does not oscillate between disjoint subsets of the state space in a periodic (regular) manner.

A Markov chain is said to be ergodic if it is irreducible, positive recurrent and aperiodic.

Definition: A **stationary distribution** $\pi(\cdot)$ of a positive recurrent and aperiodic Markov chain defined by a time-homogeneous transition probability matrix $P(\cdot|\cdot)$ has the following property:

$$\sum_i \pi(i) P_{ij}^t = \pi(j) \; \forall i, j \text{ and } t \geq 0. \tag{2.37}$$

where $P_{ij}^t = P(\theta_t = j|\theta_{t-1} = i)$. Then the Markov chain with a unique stationary distribution $\pi(\cdot)$ is said to be ergodic and following results hold:

1. $P_{ij}^t \to \pi(j)$ as $t \to \infty$ for all i, j. This means an ergodic Markov chain asymptotically converge to a unique stationary distribution irrespective to the initial state.

2. If $\mathbb{E}_\pi(h(\theta)) < \infty$, then

$$P(S_n \to \mathbb{E}_\pi(h(\theta))) = 1,$$

where $\mathbb{E}_\pi(h(\theta)) = \int_\theta h(\theta)\pi(\theta)\mathrm{d}\theta$ and S_n is the average of n sample of an ergodic Markov chain satisfying above properties. This is also called the ***ergodic theorem*** which defines the strong law of large number for an ergodic Markov chain.

These two above properties of an ergodic Markov chain are the key concepts of the convergence of the MCMC method. The rate of convergence of the MCMC sample can be given by the CLT theorem (defined before) with the Lindeberg's condition for non-identical random variables. More details can be found in Gilks et al. (1996).

The dependent MCMC samples after a burn-in period, drawn from the support of the target distribution may then be used in Monte Carlo integration to approximate intractable posterior expectations, such as:

$$\mathbb{E}(h(\theta)|\mathbf{y}) \approx \frac{1}{M-R}\sum_{t=R+1}^{M} h(\theta^t), \tag{2.38}$$

where the burn-in period is before time R.

There are two main algorithms of the MCMC method to simulate Markov chains from multi-dimensional, intractable posterior distributions: the Metropolis-Hastings algorithm and the Gibbs sampler.

2.4.3 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm developed by Hastings (1970), is a generalization of the Markov chain method first proposed by Metropolis et al. (1953). Given a state θ^t of a Markov chain, the next state θ^{t+1} is chosen by first sampling a proposal $\acute{\theta}$

from a proposal density $q(\cdot|\theta^t)$. This is accepted with a probability

$$\alpha(\theta^t, \acute{\theta}) = \min\left(1, \frac{\pi(\acute{\theta})q(\theta^t|\acute{\theta})}{\pi(\theta^t)q(\acute{\theta}|\theta^t)}\right). \tag{2.39}$$

The term $\alpha(\cdot,\cdot)$ in Eq. 2.39 is called the acceptance probability. If the proposal is accepted, the next state θ^{t+1} equals the proposal $\acute{\theta}$ else the chain remains at the current state $\theta^{t+1} = \theta^t$.

In terms of using this method for Bayesian inference, the important property of Eq. 2.39 is that if $\pi(\theta)$ is a posterior distribution then, because it only appears as a ratio, the normalizing constant cancels out and the ratio $\frac{\pi(\acute{\theta})}{\pi(\theta)}$ is simply the ratio of prior$\times$likelihood. This is key to sampling from posterior distributions.

With any form of $q(\cdot|\theta^t)$, the stationary distribution will be $\pi(\cdot)$. This can be proved with the concept of detailed balance (Tierney, 1994). The transition density for the algorithm is

$$P(\theta^{t+1}|\theta^t) = q(\theta^{t+1}|\theta^t)\alpha(\theta^t, \theta^{t+1}) + I(\theta^{t+1} = \theta^t)\left[1 - \int q(\acute{\theta}|\theta^t)\alpha(\theta^t, \acute{\theta})\mathrm{d}\acute{\theta}\right], \tag{2.40}$$

where $I(\cdot)$ is the indicator function. The first term in the above equation is for the acceptance of $\acute{\theta}$, whereas the second term comes from the rejection. From Eq. 2.39:

$$\pi(\theta^t)q(\theta^{t+1}|\theta^t)\alpha(\theta^t, \theta^{t+1}) = \pi(\theta^{t+1})q(\theta^t|\theta^{t+1})\alpha(\theta^{t+1}, \theta^t) \tag{2.41}$$

which gives the **detailed balance** equation:

$$\pi(\theta^t)P(\theta^{t+1}|\theta^t) = \pi(\theta^{t+1})P(\theta^t|\theta^{t+1}). \tag{2.42}$$

Integrating both the sides of the equation with respect to θ^t:

$$\int \pi(\theta^t)P(\theta^{t+1}|\theta^t)\mathrm{d}\theta^t = \pi(\theta^{t+1}). \tag{2.43}$$

Under the assumption that θ^t is from $\pi(\cdot)$, the marginal posterior of θ^{t+1}, $\pi(\theta^{t+1})$, is equal to $\pi(\cdot)$. That is, once the stationarity is obtained at a state, all the samples at future states will also come from the same stationary distribution. The stationary

distribution is then equal to the target distribution (Gilks et al., 1996).

A good proposal distribution has the following properties:

1. It should be easy to sample from.

2. It is easy to compute the probability of acceptance.

3. Each move of the Markov chain should be at a reasonable distance in the support of the distribution, otherwise, the chain will move very slowly. Suppose the proposal distribution $q(\mathbf{X}_t|\mathbf{X}_{t-1}) = g(\mathbf{X}_t - \mathbf{X}_{t-1})$ then,

$$\mathbf{X}_t = \mathbf{X}_{t-1} + \epsilon;\ \epsilon \sim g.$$

   The absolute value of ϵ should not be very small resulting in slow convergence.

4. The proposals should not be rejected too frequently. For example, if ϵ is very far from $\mathbf{X}_{t-1}$, it might reject the proposal leading to slow mixing.

If the proposal distribution is symmetric, e.g. a multivariate normal distribution or a random walk, the algorithm reduces to the Metropolis algorithm, a special case with the acceptance probability as:

$$\alpha(\theta^t, \acute{\theta}) = \min\left(1, \frac{\pi(\acute{\theta})}{\pi(\theta^t)}\right). \tag{2.44}$$

For more details on the forms of the proposal density, see Chib & Greenberg (1995), Gilks et al. (1996) and Lee (2004). Chib & Greenberg (1995) describes the Metropolis-Hastings algorithm in detail.

2.4.4 Gibbs sampling

Gibbs sampling is a widely used MCMC technique in Bayesian statistics to draw samples from posterior distributions. It was first developed for use in Bayesian inference by Geman & Geman (1984) to analyze Gibbs distributions on lattices. However, it was a well known method in statistical physics with the name *heat bath algorithm*. The method was introduced to mainstream statistics by Gelfand & Smith (1990) and Gelfand et al. (1990).

The method is a special case of Metropolis-Hastings algorithm with the proposal distribution equal to the product of all the full conditional posterior distributions. In the algorithm, the proposal distribution for updating the i^{th} component of θ,

$$q(\acute{\theta}_i|\theta_i, \theta_{-i}) = P(\acute{\theta}_i|\theta_{-i}, \mathbf{y}), \tag{2.45}$$

which means that the Gibbs sampling proposals are always accepted.

The steps of the Gibbs sampling are given as follows:

1. Choose arbitrary initial values θ^0.

2. For t=1:T,

   (a) generate θ_1^t from $P(\theta_1|\theta_{-1}^{t-1}, \mathbf{y})$,

   (b) generate θ_i^t from $P(\theta_i|\theta_1^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1} \ldots, \theta_p^{t-1}, \mathbf{y})$; $i = 2 : P$.

   The number of iteration T should be set to a value to reach to the convergence.

Gibbs sampler and the Markov property:

It is worth mentioning the Markov property (conditional independence) for the Gibbs sampler algorithm under which the conditional distribution $P(\theta_i|\theta_1^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1} \ldots, \theta_P^{t-}$ is simplified and it makes it easier to draw samples from it. As described in the previous section, let us consider a Markov blanket to describe the Markov property of a factored model through a DAG represented in Figure 2.8. Suppose at any iteration of Gibbs sample algorithm, we are interested in generating samples of θ_3 from its conditional distribution $P(\theta_3|\theta_1, \theta_2, \theta_4, \theta_5, \theta_6, \mathbf{y})$. By the conditional property shown via the Markov blanket,

$$P(\theta_3|\theta_1, \theta_2, \theta_4, \theta_5, \theta_6, \mathbf{y}) = P(\theta_3|\theta_1, \theta_2, \theta_4, \mathbf{y}).$$

That is, we do not require the samples of θ_5 and θ_6 to generate θ_3 from its conditional distribution, as θ_3 is independent of θ_5 and θ_6 given the parents $\theta_1, \theta_2, \theta_4, \mathbf{y}$, .

Thus in general by the Markov property, any θ_i is independent of all those θ_j $j \neq i$

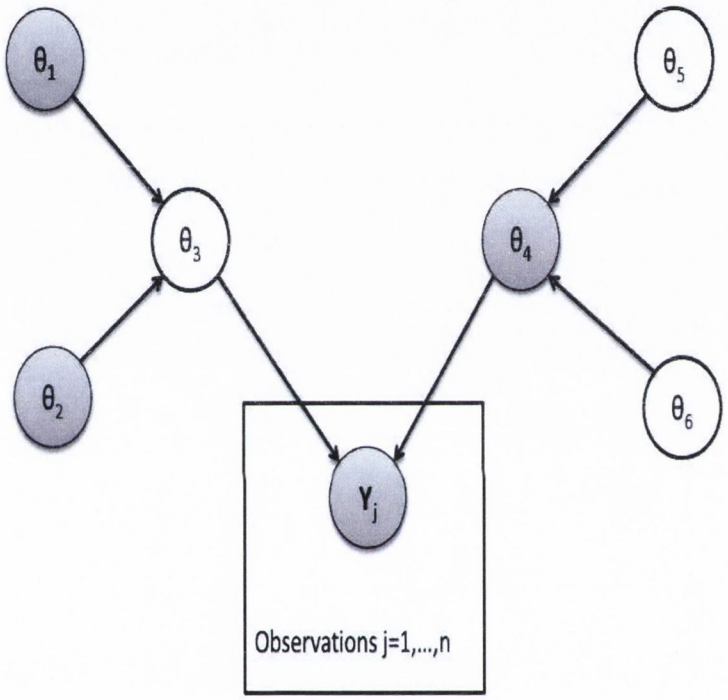

Fig. 2.8: A DAG showing conditional independence by the Markov property for a multivariate model. The DAG represents a Markov blanket around the node θ_3 with the shaded nodes as its parents, children and co-parents. Node $\mathbf{Y}_j$'s are children, nodes θ_1 and θ_2 are parents and θ_4 is the co-parent. θ_3 is conditionally independent of nodes θ_5 and θ_6 given its parents.

which are not its parents. Or,

$$pa_{\theta_i} = \text{parents of } \theta_i, \tag{2.46}$$

$$\theta_i \perp \theta_j \mid pa_{\theta_i} \; \forall \; j \neq i \text{ and } \theta_j \neq pa_{\theta_i}. \tag{2.47}$$

Thus, the Markov property simplifies the structure of the conditional distributions to simulate samples from the conditional distribution for a Gibbs sampler algorithm. A detailed approached to the Gibbs sampler can be found in Casella & George (1992).

2.5 Methods of Functional approximations

Functional approximation attempts to find a tractable function of θ that approximates the intractable posterior distribution. Some popular methods of such type are discussed below.

2.5.1 Gaussian approximation

The concept of the Gaussian approximation has been widely used for distributions that are uni-modal and roughly symmetric. A Gaussian approximation of a log-posterior density, $\log P(\theta|\mathbf{y})$, is a quadratic function that is usually centered at its mode, $\hat{\theta}$. A quadratic function of $\log P(\theta|\mathbf{y})$ can be found by its second order Taylor's expansion:

$$\log P(\theta|\mathbf{y}) \approx \log P(\hat{\theta}|\mathbf{y}) + (\theta-\hat{\theta}) \left[\frac{d}{d\theta} \log P(\theta|\mathbf{y})\right]_{\theta=\hat{\theta}} + \frac{1}{2}(\theta-\hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log P(\theta|\mathbf{y})\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta}) \tag{2.48}$$

where the first derivative term in the R.H.S of the expression should be zero since the log-posterior density has zero first derivative at its mode. The first term in the expansion is constant and the term related to the second derivative of $\log P(\theta|\mathbf{y})$ is proportion to a log-normal density, providing a normal approximation to $P(\theta|\mathbf{y})$ with

mean equal to $\hat{\theta}$ and variance as a function of the second derivative of $\log P(\theta|\mathbf{y})$.

$$\log P(\theta|\mathbf{y}) \approx -\frac{1}{2}(\theta-\hat{\theta})^T \left[-\frac{d^2}{d\theta^2}\log P(\theta|\mathbf{y})\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta}), \tag{2.49}$$

$$P(\theta|\mathbf{y}) = \log N(\hat{\theta}, V(\hat{\theta}|\mathbf{y})), \tag{2.50}$$

where, the variance of the approximation $V(\hat{\theta}|\mathbf{y}) = -\left[\frac{d^2}{d\theta^2}\log P(\theta|\mathbf{y})\right]^{-1}_{\theta=\hat{\theta}}$.

The Gaussian approximation can be very useful in Bayesian computation provided the second derivative of the density is defined. There are disadvantages to Gaussian approximation. According to Gelman et al. (2003), a Gaussian approximation to the marginal or conditional distribution of a variable is more accurate than that of the joint distribution as some variables might deviate from Gaussianity than others. Whereas, Murphy et al. (1999) argue that the Gaussian approximation of the joint posteriors can sometimes be well defined even if some of its conditional distributions are very far from the symmetry.

2.5.2 Laplace Approximation

The Laplace approximation, based on the Gaussian approximation, is a method of approximating integrals (Tierney & Kadane, 1986). It evaluates the posterior distribution $P(\theta|\mathbf{y})$ by approximating its normalizing constant. Suppose,

$$P(\theta|\mathbf{y}) = \frac{P(\theta,\mathbf{y})}{P(\mathbf{y})}, \tag{2.51}$$

To approximate the normalizing constant $P(\mathbf{y})$, a second-order Taylor's expansion of $\log P(\theta,\mathbf{y})$ around its mode $\hat{\theta}$ is considered:

$$\log P(\theta,\mathbf{y}) \approx \log P(\hat{\theta},\mathbf{y}) + \frac{1}{2}(\theta-\hat{\theta})^T \left[\frac{d^2}{d\theta^2}\log P(\theta,\mathbf{y})\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta}), \tag{2.52}$$

$$P(\theta,\mathbf{y}) \approx P(\hat{\theta},\mathbf{y}) \exp\left[-\frac{1}{2}(\theta-\hat{\theta})^T A(\theta-\hat{\theta})\right], \tag{2.53}$$

$$\text{where } A \approx -\left[\frac{d^2}{d\theta^2}\log P(\theta,\mathbf{y})\right]_{\theta=\hat{\theta}} \tag{2.54}$$

Thus, $P(\theta, \mathbf{y})$ is approximated with a normal density whose normalizing constant is:

$$\begin{aligned} P(\mathbf{y}) &= \int P(\theta, \mathbf{y}) \mathrm{d}\theta, & (2.55) \\ &\approx P(\hat{\theta}, \mathbf{y}) \int_{\theta} \exp\left[-\frac{1}{2}(\theta - \hat{\theta})^T A (\theta - \hat{\theta})\right] \mathrm{d}\theta, & (2.56) \\ &= P(\hat{\theta}, \mathbf{y})(\det A)^{-\frac{1}{2}}(2\pi)^{\frac{P}{2}}. & (2.57) \end{aligned}$$

Then, the Laplace approximation to $P(\theta|\mathbf{y})$ is given as:

$$P(\theta|\mathbf{y}) \approx \frac{(\det A)^{\frac{1}{2}}(2\pi)^{-\frac{P}{2}}}{P(\hat{\theta}, \mathbf{y})} P(\theta, \mathbf{y}). \tag{2.58}$$

The Laplace approximation of $P(\theta|\mathbf{y})$ can be compared with its Gaussian approximation. If the normal approximation of $P(\theta, \mathbf{y})$ is also considered in the numerator of Eq. 2.58, the Laplace approximation of $P(\theta|\mathbf{y})$ becomes a Gaussian approximation to the distribution.

The Laplace approximation to the marginal posterior distribution $P(\theta_i|\mathbf{y})$ requires Gaussian approximations for terms in both the numerator and the denominator of the R.H.S of the expression given below:

$$P(\theta_i|\mathbf{y}) = \frac{\int_{\theta_{-i}} P(\theta_i, \theta_{-i}, \mathbf{y}) \mathrm{d}\theta_{-i}}{\int_{\theta} P(\theta, \mathbf{y}) \mathrm{d}\theta}. \tag{2.59}$$

The term in the denominator of the R.H.S. of the expression in Eq. 2.59 is the normalizing constant of $P(\theta|\mathbf{y})$, which can be approximated as described before. To approximate the integral in the numerator, find a second-order Taylor's expansion of $P(\theta_i, \theta_{-i}, \mathbf{y})$ as a function of θ_{-i} around its mode $\hat{\theta}_{-i}$. The marginal posterior $P(\theta_i|\mathbf{y})$ is then approximated as:

$$\begin{aligned} P(\theta_i|\mathbf{y}) &\approx \frac{P(\theta_i, \hat{\theta}_{-i}, \mathbf{y})(\det A')^{-\frac{1}{2}}(2\pi)^{\frac{(P-1)}{2}}}{P(\hat{\theta}, \mathbf{y})(\det A)^{-\frac{1}{2}}(2\pi)^{\frac{P}{2}}}, & (2.60) \\ &= (2\pi)^{-\frac{1}{2}} \frac{P(\theta_i, \hat{\theta}_{-i}, \mathbf{y})(\det A)^{\frac{1}{2}}}{P(\hat{\theta}, \mathbf{y})(\det A')^{\frac{1}{2}}}, & (2.61) \end{aligned}$$

where $A' = -\left[\frac{d^2}{d\theta_{-i}^2} \log P(\theta_i, \theta_{-i}|\mathbf{y})\right]_{\theta=\hat{\theta}}$. For more details on approximating marginal posterior distribution see Tierney & Kadane (1986).

Rue. et al. (2009) presents the Laplace approximation of marginal posterior distributions in a different way. The marginal posterior distribution can be defined in terms of joint and conditional posterior distribution as:

$$P(\theta_i|\mathbf{y}) = \frac{P(\theta_i, \theta_{-i}, \mathbf{y})}{P(\theta_{-i}|\theta_i, \mathbf{y})P(\mathbf{y})}, \tag{2.62}$$

$$\propto \frac{P(\theta_i, \theta_{-i}, \mathbf{y})}{P(\theta_{-i}|\theta_i, \mathbf{y})}, \tag{2.63}$$

It should be noted that in above equation $P(\theta_{-i}|\theta_i, \mathbf{y})$ depends on θ_i as well. Therefore this definition of the Laplace approximation is defined for given values of θ_i, i.e for discrete θ_i (or defined over grid values).

To present the R.H.S of the above equation as a function of θ_i only, consider a Gaussian approximation of $P(\theta_{-i}|\theta_i, \mathbf{y})$ and plug-in its posterior mode $\hat{\theta}_{-i}$ in the expression. Then, the Laplace approximation of $P(\theta_i|\mathbf{y})$ can be defined as:

$$P(\theta_i|\mathbf{y}) \propto \left.\frac{P(\theta_i, \theta_{-i}, \mathbf{y})}{P_G(\theta_{-i}|\theta_i, \mathbf{y})}\right|_{\theta_{-i}=\hat{\theta}_{-i}}, \tag{2.64}$$

where $P_G(\theta_{-i}|\theta_i, \mathbf{y})$ is the Gaussian approximation of $P(\theta_{-i}|\theta_i, \mathbf{y})$. This approach of the Laplace approximation by Rue. et al. (2009) approximates $P(\theta_i|\mathbf{y})$ only up to a proportionality constant. If θ_i is a low-dimensional parameter, the normalizing constant of the approximation can be found by numerical integration.

The Laplace approximation can be quite accurate if the integrand is uni-modal or at least dominated by a single mode and the sample size is large.

2.5.3 Integrated Nested Laplace Method

The Integrated Nested Laplace approximation(INLA) developed by Rue. et al. (2009) is a method of distributional approximation that yields quick and very accurate Bayesian approximations for latent Gaussian models. To describe the structure of the method, consider a latent model with latent variables $\mathbf{Z} = (\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_N)^T$

as:

$$P(\mathbf{y}|\mathbf{Z},\theta) = \prod_{i}^{N} P(\mathrm{y}_i|\mathrm{Z}_i,\theta), \tag{2.65}$$

$$P(\mathbf{Z}|\theta) = GMRF(\theta), \tag{2.66}$$

$$\theta \sim P(\theta). \tag{2.67}$$

The latent field $\mathbf{Z}$ follows a GMRF (defined in Section 2.2) parameterized by θ. The likelihood of data $\mathbf{y}$ is assumed to be an i.i.d non-Gaussian distribution given θ. The unknown θ is a set of unknown model parameters and hyper-parameters of $\mathbf{Z}$. The aim is to compute the marginal posterior distributions of Z_i; $i = 1 : N$ and θ:

$$P(\mathbf{Z},\theta|\mathbf{y}) \propto P(\theta)P(\mathbf{Z}|\theta)\prod_{i} P(\mathrm{y}_i|\mathrm{Z}_i,\theta), \tag{2.68}$$

$$P(\mathbf{Z}|\mathbf{y}) = \int_{\theta} P(\mathbf{Z},\theta|\mathbf{y})\mathrm{d}\theta, \tag{2.69}$$

$$P(\theta|\mathbf{y}) = \int_{\mathbf{Z}} P(\mathbf{Z},\theta|\mathbf{y})\mathrm{d}\mathbf{Z}. \tag{2.70}$$

The posterior marginals are often not available in closed form due to the non-Gaussian likelihood. The INLA method, developed as an alternative to the MCMC method, provides a quick and tractable solution for such models by assuming two basic properties. Firstly, the latent field $\mathbf{Z}$ is assumed to be a GMRF. Secondly, the number of unknown parameters should be very small. The computational steps of the method are described as follows:

1. To approximate the marginal posterior distribution of θ, the method uses the Laplace approximation as:

$$P(\theta|\mathbf{y}) \propto \left.\frac{P(\mathbf{Z},\theta,\mathbf{y})}{P_G(\mathbf{Z}|\theta,\mathbf{y})}\right|_{\mathbf{Z}=\mathbf{Z}^\star(\theta)}, \tag{2.71}$$

   where $P_G(\mathbf{Z}|\theta,\mathbf{y})$ is a Gaussian approximation of the conditional distribution $P(\mathbf{Z}|\theta,\mathbf{y})$ and $\mathbf{Z}^\star(\theta)$ is the mode of the distribution.

   If θ is a low-dimensional parameter, the proportionality constant of $P(\theta|\mathbf{y})$ can be computed numerically.

2. The Gaussian approximation to $P(\mathbf{Z}|\theta,\mathbf{y})$ found by an algorithm developed for latent fields by Rue & Held (2005).

$$P(\mathbf{Z}|\theta,\mathbf{y}) \approx N(\mathbf{Z};\ \mu(\theta),\Sigma(\theta)). \tag{2.72}$$

Then, the conditional marginal $P(\mathrm{Z}_i|\theta,\mathbf{y})$ will be a univariate Gaussian distribution:

$$P(\mathrm{Z}_i|\theta,\mathbf{y}) \approx N(\mathrm{Z}_i;\ \mu_i(\theta),\sigma_i(\theta)). \tag{2.73}$$

3. The marginal posterior distribution of Z_i can be computed numerically with respect to θ as:

$$P(\mathrm{Z}_i|\mathbf{y}) \approx \sum_k P(\mathrm{Z}_i|\theta^k,\mathbf{y})P(\theta^k|\mathbf{y})\delta^k, \tag{2.74}$$

where $P(\mathrm{Z}_i|\theta^k,\mathbf{y})$ is approximated at the previous step. The term θ^k denotes a discrete point in the space of θ and δ^k is the area weight.

The posterior marginal of $\theta_j;\ j = 1:p$, can also be computed numerically:

$$P(\theta_j|\mathbf{y}) \approx \sum_k P(\theta_j^k,\theta_{-j}^k|\mathbf{y})\delta_j^k. \tag{2.75}$$

Eq. 2.75 is evaluated on a discrete grid that should cover the support of θ. For this reason the dimension of θ cannot be large.

4. The conditional marginal posterior distribution of Z_i, $P(\mathrm{Z}_i|\theta,\mathbf{y})$ can also be approximated by the Laplace approximation, as:

$$P(\mathrm{Z}_i|\theta,\mathbf{y}) \propto \left.\frac{P(\mathbf{Z},\theta,\mathbf{y})}{P_{GG}(\mathbf{Z}_{-i}|\mathrm{Z}_i,\theta,\mathbf{y})}\right|_{\mathbf{Z}_{-i}=\mathrm{Z}^{\star}_{-i}(\theta,\mathrm{Z}_i)}, \tag{2.76}$$

where $P_{GG}(\mathbf{Z}_{-i}|\mathrm{Z}_i,\theta,\mathbf{y})$ is a Gaussian approximation to $P(\mathbf{Z}_{-i}|\mathrm{Z}_i,\theta,\mathbf{y})$ (and is different from the conditional density corresponding to $P_G(\mathbf{Z}|\theta,\mathbf{y})$) and needs to be computed for each Z_i. This may increase the computational burden, though it produces more accurate result than a Gaussian approximation in the case of departure from Gaussianity.

The assumption of conditional independence through the Gaussian Markov random field makes INLA very quick for highly dimensional latent models. Based on the Laplace approximation, INLA provides very accurate solutions. The error associated with the Gaussian approximation of the conditional distribution is corrected by replacing it by Laplace approximation. Rue. et al. (2009) have shown that the method outperforms MCMC in term of accuracy and computational time.

The main disadvantage of the method is its non-applicability to models with large number of parameters (≥ 6). The method uses numerical integration to approximate the marginal posterior distribution of the parameters which limits the use of the method to latent models with only few parameters.

2.6 Other Methods

In this section, some techniques or terms used later in the thesis are described very briefly.

2.6.1 Cross-Validation for model checking

Cross-validation is a technique to assess how accurately a model assumption fits a data set. For this purpose, it partitions the data set into two samples, a training data set to fit the model and a test data set to check the accuracy of the model fit. The accuracy of the model fitting is generally performed using a predictive posterior distribution. If the predictive probability of test data sets is very small, the model does not fit the data accurately. This accuracy measure can be defined in a variety of ways. One example is the highest (predictive) posterior density (HPD) region. The percentage of samples falls inside the HPD region decides the accuracy of the model fitting.

There are mainly two types of cross-validation techniques used in statistical analysis: the K-fold cross validation and the leave-one-out cross validation. The *K-fold cross-validation* divides the data set into K non-overlapping samples of which $(K-1)$ are used as training data set and the remaining K^{th} sample is used as the test data set to perform the accuracy check. This is repeated for every K^{th} sample

such that each of the K samples is used once as the test data set. Then, the total K results of the accuracy check from K different test data set are averaged to give a single estimate.

The *Leave-one-out cross-validation* technique uses all but one datum for the training purpose so that each sample point can be used exactly once for the validation. This is actually a special case of *K-fold cross-validation* with K equal to the total number of sample points in the data. With this technique of cross-validation, repeating the accuracy check for each and every sample points can be quite expensive in the case of large sample sizes.

Cross-validation can present inaccurate results if the test and training data do not belong to same population or if the size of the test data set is very small.

2.6.2 Conjugate-exponential (CE) models

A Bayesian parametric model is called conjugate-exponential if it satisfies two conditions:

1. The likelihood of the i.i.d data $\mathbf{y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_N\}$ given the unknown parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\}$ should be in the **exponential** family:

$$\begin{aligned} P(\mathbf{y}|\theta) &= \prod_{j=1}^{N} \left[f(\mathrm{y}_j) g(\theta) \exp(\phi(\theta)^T u(\mathrm{y}_j)) \right], \\ &= \prod_{j=1}^{N} \left[f(\mathrm{y}_j) \right] g(\theta)^N \exp(\phi(\theta)^T \sum_{i=1}^{N} u(\mathrm{y}_j)) \end{aligned} \tag{2.77}$$

   where $\mathbf{t}(\mathbf{y}) = \sum_{j=1}^{N} u(\mathrm{y}_j))$ (say) is a vector of sufficient statistics of θ and the terms $f(\cdot)$, $g(\cdot)$, $\phi(\cdot)$ and $u(\cdot)$ are known functions. The term $\phi(\theta)$ is a vector of the natural parameters that linearly relates to $t(\mathbf{y})$.

2. The prior distribution of θ should be **conjugate** to the data-likelihood:

$$P(\theta|\eta, \nu) = h(\eta, \nu) g(\theta)^{\eta} \exp(\phi(\theta)^T \nu), \tag{2.78}$$

   where η and ν are hyper-parameters of θ. The prior is conjugate to the likeli-

hood if the posterior has the same form as the prior:

$$P(\theta|\mathbf{y}) = g(\theta)^{\eta+n} \exp(\phi(\theta)^T(\nu + \mathbf{t}(\mathbf{y}))). \tag{2.79}$$

CE models play an important role in Bayesian statistics. For most of the models, the analytical solutions to Bayesian computation are not available. The problem of Bayesian computation can be avoided for the CE models with the known standard forms of the posterior distributions. Most of the distributions of the exponential family, for example Normal, Bernoulli, Gamma, exponential, Poisson, belong to the CE models. Whereas, the logistic distribution (in the exponential family) do not have natural conjugate priors.

The CE models have important significance for the VB method as the models provide natural factorized form for the method to provide tractable solutions (see Beal (2003)). The tractability issue of the VB method due to the CE models will be discussed in the next chapter.

Chapter 3

The Variational Bayes Approximation

In the previous chapter, some of the Bayesian computational methods have been discussed briefly. This chapter describes variational Bayes (VB), a functional approximation method, in detail. The VB method is used throughout the thesis as a tool for finding approximations to (Bayesian) inverse regression problems. Section 3.1 discusses the background and is a literature review of the method. In Section 3.2.1, a detailed introduction to the method is presented. Two possible approaches to the method, named as the classical approach and the Šmídl and Quinn approach (explained in the same section), describe the procedures to find VB approximations to Bayesian inference problems. The Gaussian variational approach and the variational tangent approach are discussed briefly in Sections 3.2.2 and 3.2.3 respectively which are later used in the thesis to compare the VB results. Further in the chapter, the VB method is compared with other methods e.g. the Laplace approximation, INLA, MCMC.

3.1 Background and Literature Review

The VB method has been widely used in machine learning, neural networks, artificial intelligence, signal processing, Bayesian statistics etc. for the past two decades. The method was first proposed by Hinton & van Camp (1993) in neural networks.

The authors approximated the posterior distribution with a multivariate Gaussian distribution, assuming a diagonal covariance matrix. They showed that the approximation can be made more accurate by minimizing a measure of discrepancy between the true and the approximation iteratively. The method was applied to latent models by Waterhouse & Robinson (1996), MacKay (1998) under the name *ensemble learning*, optimizing variational free energy with no restriction on the distributional form of the approximation. Attias (1999) presented a variational Bayesian framework for graphical models. The author showed that the VB method could be interpreted as a generalization of the famous *EM* algorithm. Beal (2003) explored the algorithm by Attias (1999) and presented a general framework for variational Bayes learning for latent models. He developed a variational Bayes *EM* algorithm which combines the concept of the *EM* algorithm and the mean field approximation (Saul et al., 1996; Jaakkola, 2000). At the E-step of the algorithm, it approximates the posterior of the latent variables given the VB posterior-estimates of other unknown parameters and the M-step finds a variational approximation for the posterior distribution of the parameters by optimizing a lower bound on the marginal likelihood. Šmídl & Quinn (2006) presented a different perspective of the method in signal processing. Assuming both the parameters and latent variables as unknowns in the Bayesian framework, the authors developed a variational iterative algorithm which is also applicable to non-latent models. The VB algorithm by Šmídl & Quinn (2006) uses the concept of minimizing a discrepancy between the approximation and the true posterior distribution.

The key assumption of the method is similar to the mean field approximation (Saul et al., 1996; Jaakkola, 2000). It also ignores the dependence between the components of unknowns in the posterior distribution. The mean field approximation assumes complete independence between the unknown parameters. Whereas, the independence assumption of Šmídl & Quinn (2006) may still assume some dependence between the parameters, for example if the dependence between some of the parameters does not lead to intractability in the approximation or if the independence between them leads to a very poor approximation.

3.2 Introduction

In Bayesian inference, the evaluation of a multivariate posterior distribution is often intractable. Variational approximation is a functional approximation method that tries to minimize a distance measure between two functions, true and approximation. Based on the assumptions that helps finding a tractable approximation, the variational approximations can be described by many different variational methods, for example, the **Gaussian variational approach**, the **variational tangent approach** or the **variational Bayes approach**. The variational Bayes (VB) is considered to be the main tool for providing tractable solutions to inverse estimation throughout the thesis. The other two variational methods are used for the comparison of the VB results for inverse estimations problems considered in the thesis.

In the next section, we discuss the VB method in detail. After that the other variational methods, Gaussian variational approach and tangent variational approach, are described briefly.

3.2.1 VB method

The VB method is a popular method of variational approximation in which an intractable posterior distribution is approximated by a factored distribution assuming the mean field approximation (Saul et al., 1996; Jaakkola, 2000). The key assumption of the method is to ignore the posterior dependencies between the components. It facilitates the method to provide a tractable and quick solution.

We present two approaches of the VB method based on the definition and the assumption of the method by Beal (2003) and Šmídl & Quinn (2006) respectively. These two different approaches of the method provide the same VB result.

Classical approach:

Beal (2003) describes what we call the classical approach to the VB method as a generalization of the *EM* algorithm. To derive this approach, we consider a latent model with an i.i.d data set, $\mathbf{y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_N\}$, a vector of latent variables,

$\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_N\}$ and a set of model parameters and hyper-parameters, $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\}$. The joint posterior distribution of $\mathbf{Z}$ and θ given data $\mathbf{y}$ can be defined by Bayes' law as:

$$P(\mathbf{Z}, \theta|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{Z}, \theta)P(\mathbf{Z}, \theta)}{P(\mathbf{y})}, \tag{3.1}$$

$$\text{where } P(\mathbf{y}) = \int P(\mathbf{y}|\mathbf{Z}, \theta)P(\mathbf{Z}, \theta)\mathrm{d}\mathbf{Z}\mathrm{d}\theta. \tag{3.2}$$

In the above equation, $P(\mathbf{y}|\mathbf{Z}, \theta)$ represents the likelihood of θ and $\mathbf{Z}$ and $P(\mathbf{Z}, \theta)$ stands for their joint prior distribution. If $\mathbf{Z}$ and θ are a priori independent, the joint prior can be expressed as the product of marginal priors: $P(\mathbf{Z}, \theta) = P(\mathbf{Z})P(\theta)$. The term $P(\mathbf{y})$ denotes the marginal likelihood of $\mathbf{y}$ which is the normalizing constant of the posterior distribution. It is often computationally intractable to evaluate.

To avoid the computational burden, Beal (2003) finds a lower bound on $\log P(\mathbf{y})$ using Jensen's inequality. For any density function of $\mathbf{Z}$ and θ given $\mathbf{y}$, $q(\mathbf{Z}, \theta|\mathbf{y})$, we have:

$$\log P(\mathbf{y}) = \log \int P(\mathbf{y}, \mathbf{Z}, \theta)\mathrm{d}\mathbf{Z}\mathrm{d}\theta, \tag{3.3}$$

$$= \log \int q(\mathbf{Z}, \theta|\mathbf{y}) \frac{P(\mathbf{y}, \mathbf{Z}, \theta)}{q(\mathbf{Z}, \theta|\mathbf{y})}\mathrm{d}\mathbf{Z}\mathrm{d}\theta, \tag{3.4}$$

$$\geq \int q(\mathbf{Z}, \theta|\mathbf{y}) \log \frac{P(\mathbf{y}, \mathbf{Z}, \theta)}{q(\mathbf{Z}, \theta|\mathbf{y})}\mathrm{d}\mathbf{Z}\mathrm{d}\theta, \tag{3.5}$$

$$= \int q(\mathbf{Z}, \theta|\mathbf{y}) \log \frac{P(\mathbf{Z}, \theta|\mathbf{y})P(\mathbf{y})}{q(\mathbf{Z}, \theta|\mathbf{y})}\mathrm{d}\mathbf{Z}\mathrm{d}\theta. \tag{3.6}$$

Maximizing the lower bound with respect to $q(\mathbf{Z}, \theta|\mathbf{y})$ gives $q(\mathbf{Z}, \theta|\mathbf{y}) = P(\mathbf{Z}, \theta|\mathbf{y})$ which does not solve our problem, as $P(\mathbf{Z}, \theta|\mathbf{y})$ is known only up to a proportionality constant. Instead, a factorized approximation $q(\mathbf{Z}, \theta|\mathbf{y}) = q(\mathbf{Z}|\mathbf{y})q(\theta|\mathbf{y})$ is considered that ignores the dependence between $\mathbf{Z}$ and θ, which may provide a tractable lower bound on $\log P(\mathbf{y})$:

$$\log P(\mathbf{y}) \geq \int q(\mathbf{Z}|\mathbf{y})q(\theta|\mathbf{y}) \log \frac{P(\mathbf{y}, \mathbf{Z}, \theta)}{q(\mathbf{Z}|\mathbf{y})q(\theta|\mathbf{y})}\mathrm{d}\mathbf{Z}\mathrm{d}\theta, \tag{3.7}$$

$$= F(q(\mathbf{Z}|\mathbf{y}), q(\theta|\mathbf{y})), \tag{3.8}$$

where F is a lower bound on $\log P(\mathbf{y})$ depending on the factored approximations $q(\mathbf{Z}|\mathbf{y})$ and $q(\theta|\mathbf{y})$. Beal (2003) developed an algorithm, variational Bayes *EM*, which iteratively maximizes F with respect to $q(\mathbf{Z}|\mathbf{y})$ and $q(\theta|\mathbf{y})$ respectively. As in the *EM* algorithm, there are two steps of the iterative variational Bayesian *EM* algorithm: VBE-step and VBM-step. At the VBE-step, maximizing $F(q(\mathbf{Z}|\mathbf{y}), q(\theta|\mathbf{y}))$ w.r.t $q(\mathbf{Z}|\mathbf{y})$ gives the expression for $q(\mathbf{Z}|\mathbf{y})$. At the VBM-step, it maximizes $F(q(\mathbf{Z}|\mathbf{y}), q(\theta|\mathbf{y}))$ w.r.t $q(\theta|\mathbf{y})$ and gives a solution for $q(\theta|\mathbf{y})$. At any VB-iteration t, the two steps of the algorithm are given as follows:

$$\textbf{VBE-step}: \; q^t(\mathbf{Z}_i|\mathbf{y}) \propto \int q^{t-1}(\theta|\mathbf{y}) \log P(\mathbf{y}_i, \mathbf{Z}_i, \theta) \; \forall i, \tag{3.9}$$

$$\text{where } q^t(\mathbf{Z}|\mathbf{y}) = \prod_i q^t(\mathbf{Z}_i|\mathbf{y}), \text{ and} \tag{3.10}$$

$$\textbf{VBM-step}: \; q^t(\theta|\mathbf{y}) \propto \int q^t(\mathbf{Z}|\mathbf{y}) \log P(\mathbf{Z}, \theta, \mathbf{y}) \mathrm{d}\mathbf{Z}. \tag{3.11}$$

These two steps of computation are equivalent to finding a possibly tight lower bound on the marginal likelihood. At each iteration, the lower bound on the marginal likelihood gets increased or remains unchanged with the update in the VB approximations. A tight bound is obtained when the approximation is very close to the true posterior distribution. Thus, the posterior distribution can be approximated by the VB method as:

$$P(\mathbf{Z}, \theta|\mathbf{y}) \approx q(\mathbf{Z}, \theta|\mathbf{y}), \tag{3.12}$$

$$= \left[\prod_i q(\mathrm{Z}_i|\mathbf{y})\right] q(\theta|\mathbf{y}). \tag{3.13}$$

The lower bound F can also be expressed in the terms of a Kullback-Leibler divergence:

$$F(q(\mathbf{Z}|\mathbf{y}), q(\theta|\mathbf{y})) = \int q(\mathbf{Z}|\mathbf{y}) q(\theta|\mathbf{y}) \log \frac{P(\mathbf{Z}, \theta|\mathbf{y}) P(\mathbf{y})}{q(\mathbf{Z}|\mathbf{y}) q(\theta|\mathbf{y})} \mathrm{d}\mathbf{Z} \mathrm{d}\theta, \tag{3.14}$$

$$= \int q(\mathbf{Z}|\mathbf{y}) q(\theta|\mathbf{y}) \log \frac{P(\mathbf{Z}, \theta|\mathbf{y})}{q(\mathbf{Z}|\mathbf{y}) q(\theta|\mathbf{y})} \mathrm{d}\mathbf{Z} \mathrm{d}\theta + \log P(\mathbf{y}),$$

$$= -KL(q(\mathbf{Z}|\mathbf{y}) q(\theta|\mathbf{y}) \parallel P(\mathbf{Z}, \theta|\mathbf{y})) + \log P(\mathbf{y}),$$

$$\Rightarrow \log P(\mathbf{y}) - F(q(\mathbf{Z}|\mathbf{y}), q(\theta|\mathbf{y})) = KL(q(\mathbf{Z}|\mathbf{y}) q(\theta|\mathbf{y}) \parallel P(\mathbf{Z}, \theta|\mathbf{y})). \tag{3.15}$$

It follows that maximizing F with respect to $q(\mathbf{Z}|\mathbf{y})q(\theta|\mathbf{y})$ is equivalent to minimizing the Kullback-Leibler (KL) divergence $KL(q(\mathbf{Z}|\mathbf{y})q(\theta|\mathbf{y})||P(\theta, \mathbf{Z}|\mathbf{y}))$. KL divergence is a measure of discrepancy from one distribution to another. It is always a non-negative quantity. Šmídl & Quinn (2006) have discussed the KL divergence and its properties. Detailed discussion on KL divergence can be found in Kullback & Leibler (1951) and Kullback (1997).

Šmídl and Quinn approach:

Šmídl & Quinn (2006) develop an iterative-VB algorithm which is applicable to both latent and non-latent models. Consider an unknown ψ representing the latent variables $\mathbf{Z}$ and parameters θ: $\psi = \{\mathbf{Z}, \theta\}$. In the Šmídl and Quinn approach, the VB method attempts to find a distribution $q(\psi|\mathbf{y})$ as an approximation to the intractable posterior distribution $P(\psi|\mathbf{y})$ by minimizing a KL-divergence $KL(q(\psi|\mathbf{y}) \parallel P(\psi|\mathbf{y}))$:

$$KL(q(\psi|\mathbf{y}) \parallel P(\psi|\mathbf{y})) = \int q(\psi|\mathbf{y}) \log \frac{q(\psi|\mathbf{y})}{P(\psi|\mathbf{y})} \mathrm{d}\psi, \tag{3.16}$$

In practice, it is not possible to find a minimum of $KL(q(\psi|\mathbf{y}) \parallel P(\psi|\mathbf{y}))$ with respect to $q(\psi|\mathbf{y})$, i.e.

$$q(\psi|\mathbf{y}) = \arg \min_{q(\psi|\mathbf{y})} KL(q(\psi|\mathbf{y}) \parallel P(\psi|\mathbf{y})), \tag{3.17}$$

as the posterior distribution $P(\psi|\mathbf{y})$ is known only up to a proportionality constant. So, exactly as in the classical approach by Beal (2003) it assumes the posterior independence between the components of the unknowns:

$$q(\psi|\mathbf{y}) = \prod_{i=1}^{d} q(\psi_i|\mathbf{y}), \tag{3.18}$$

which reduces the minimization of $KL(q(\psi|\mathbf{y}) \parallel P(\psi|\mathbf{y}))$ with respect to $q(\psi|\mathbf{y})$ to the minimization of $KL(q(\psi|\mathbf{y}) \parallel P(\psi|\mathbf{y}))$ with respect to $q(\psi_i); i = 1, \ldots, d$. If

$q(\psi_{-i}|\mathbf{y}) = \prod_{\substack{j=1 \\ j\neq i}}^{d} q(\psi_j|\mathbf{y})$ is fixed, $q(\psi_i|\mathbf{y})$ can be found to be:

$$q(\psi_i|\mathbf{y}) \propto \exp\left[\int q(\psi_{-i}|\mathbf{y})[\log P(\mathrm{y},\psi)]\mathrm{d}\psi_{-i}\right]; \qquad i = 1, \cdots, d, \tag{3.19}$$

(see Theorem 3.3.1 of Šmídl & Quinn (2006)). The term $q(\psi_i|\mathbf{y})$ is called the VB-marginal of ψ_i.

The expression for $q(\psi_i|\mathbf{y})$ involves a $(d-1)$ dimensional integral. This integral may not be computable. However, since it is assumed that $q(\psi_{-i}|\mathbf{y}) = \prod_{\substack{j=1 \\ j\neq i}}^{d} q(\psi_j|\mathbf{y})$, this integral becomes a tractable product of $(d-1)$ 1-dimensional integrals if the log-joint density $\log P(\psi, \mathbf{y})$ factorizes over the components of ψ. For example, it is common to have factorized log-joint density over the components of the parameter for the conjugate-exponential family of distributions. If the likelihood in the components of the parameter belongs to the exponential family, we may have the log-joint density factorized over the components of the parameter. There are some examples where this factorization does not occur, such as logistic or Poisson regression models with non-conjugate priors.

It can be seen from Eq. 3.19 that the solution for the VB-marginal $q(\psi_i|\mathbf{y}); \forall i$ is available given other VB-marginals $q(\psi_j|\mathbf{y})$; $j \neq i$. Šmídl & Quinn (2006) give a way to avoid the requirement of the prior knowledge of the VB-marginals and the computation of integrals involved in their definition. Eq. 3.19 can be written as:

$$q(\psi_i|\mathbf{y}) \propto \exp[\mathbb{E}_{q(\psi_{-i}|\mathbf{y})}[\log P(\mathbf{y},\psi)]], \qquad i = 1, \ldots, d, \tag{3.20}$$

and so, we see that $q(\psi_i|\mathbf{y})$ is a function of the moments of ψ_j, $j \neq i$, with respect to $q(\psi_{-i}|\mathbf{y})$. It can be understood from Eq. 3.20 that the VB-marginals interact with each other via their moments, or in other words, the moments of one VB-marginal are some function of the moments of other VB-marginals.

There are two ways in which the iterative algorithm can be implemented:

1. **Šmídl and Quinn Iterative VB algorithm:** The first way is to follow the VB-iterative algorithm given by Šmídl & Quinn (2006). This requires the

recognition of the VB-marginals as standard distributions. If this is the case then, from Eq. 3.20, the parameters of each VB-marginal will be the functions of the moments of other VB-marginals. Critically, this means that all we need to compute and store at each VB-iteration are the parameters and/or the moments of the VB-marginals. The iterative VB algorithm is reduced to iteratively updating the parameters of each VB-marginal which are a function of the moments of other marginals, until these values converge.

At any VB-iteration t, let $H_i^t(\psi_i)$ and $M_i^t(\psi_i)$ represent the parameters and the moments of $q^t(\psi_i|\mathbf{y})$ (eg. $\mathbb{E}(\psi_i)$, $\mathbb{E}(\psi_i^2)$, $\mathbb{E}(\log\psi_i)$ etc.) required for the evaluation of $q^t(\psi_j|\mathbf{y})$; $j \neq i$ respectively, $f_i(\cdot)$'s are the functions through which the parameters of one VB-marginal depend on the moments of other VB-marginals and $g_i(\cdot)$'s are the functions defining the relation between the moments and the parameters of the VB-marginals. The form of the function $f_i(\cdot)$'s and $g_i(\cdot)$'s can only be known if the VB-marginals can be recognized as standard distributions. Thus, $f_i(\cdot)$ will be a function of all current values of the moments of $q(\psi_j|\mathbf{y})$, $j \neq i$ and $g_i(\cdot)$ will be a function of the moments of $q(\psi_i|\mathbf{y})$.

For $i = 1, \ldots, d$, we can write,

$$H_i^t(\psi_i) = f_i(M_1^t(\psi_1), M_2^t(\psi_2), \ldots, M_{i-1}^t(\psi_{i-1}), M_{i+1}^{t-1}(\psi_{i+1}), \ldots, M_d^{t-1}(\psi_d), \mathbf{y}), \tag{3.21}$$

$$M_i^t(\psi_i) = g_i(H_i^t(\psi_i)) \tag{3.22}$$

Recognition of the standard form of the VB-marginal also solves the problem of computation of its normalizing constant (as VB-marginals are defined up to proportionality, see Eq. 3.20).

2. **Iterative VB algorithm using numerical integration:** Rather than identifying the VB-marginals as standard distributions, the second way uses a numerical approximation to compute the normalizing constants of the VB-

marginals defined in Eq. 3.20. The required moments are also computed numerically. In other words, in this implementation, the basic iterative VB algorithm is used and where any integration is needed, it is done by numerical approximation. We use a finite sum (Riemann) approximation on a finite support that is $\mathbb{E}(\psi_i) \pm 4\ \mathrm{sd}(\psi_i)$ with respect to the previous VB-marginal $q^{t-1}(\psi_i|\mathbf{y})$. At each iteration, the VB-marginals are modified and then their moments and the support (the range of the unknowns) are also updated accordingly.

Suppose at any VB-iteration t, $q^{\star t}(\psi_i)$ is the un-normalized VB-marginal obtained, Z_i^t is the normalizing constant for the VB-marginal $q^t(\psi_i|\mathbf{y})$, $m_i^t(\psi_i)$ stands for the moments of $q^t(\psi_i|\mathbf{y})$, $h_i(\psi_i)$ is some function of ψ_i required to evaluate its moments, $E_{q^t}(\psi_i)$ is the VB-mean and $Var_{q^t}(\psi_i)$ is the VB-variance obtained from the definition of the moments, $R_i^t(\psi_i)$ shows the approximated range of ψ_i and $\delta^t(\psi_i)$ is the step-size of the regular grid on $R_i^t(\psi_i)$. Then, for $i = 1, \ldots, d$,

$$q^{\star t}(\psi_i|\mathbf{y}) = \exp[\mathbb{E}_{q^{(t-1)}(\psi_{-i}|\mathbf{y})}[\log P(\mathbf{y}, \psi)]], \tag{3.23}$$

$$Z_i^t = \sum_{R^{t-1}(\psi_i)} q^{\star t}(\psi_i|\mathbf{y})\delta^t(\psi_i), \tag{3.24}$$

$$q^t(\psi_i|\mathbf{y}) = \frac{1}{Z_i^t} q^{\star t}(\psi_i|\mathbf{y}), \tag{3.25}$$

$$m_i^t(\psi_i) = \sum_{R^{t-1}(\psi_i)} h_i(\psi_i) q^t(\psi_i|\mathbf{y})\delta^t(\psi_i), \tag{3.26}$$

$$R_i^t(\psi_i) = \mathbb{E}_{q^t}(\psi_i) \pm 4\sqrt{Var_{q^t}(\psi_i)}, \tag{3.27}$$

and for $j \neq i$,

$$q^{\star t}(\psi_j) = \exp[\mathbb{E}_{q^{(t-1)}(\psi_{-i,-j}|\mathbf{y}),q^t(\psi_i|\mathbf{y})}[\log P(\mathbf{y},\psi)]], \tag{3.28}$$

$$Z_j^t = \sum_{R^{t-1}(\psi_j)} q^{\star t}(\psi_j|\mathbf{y})\delta^t(\psi_j), \tag{3.29}$$

$$q^t(\psi_j|\mathbf{y}) = \frac{1}{Z_j^t} q^{\star t}(\psi_j|\mathbf{y}), \tag{3.30}$$

$$m_j^t(\psi_j) = \sum_{R^{t-1}(\psi_j)} h_j(\psi_j) q^t(\psi_j|\mathbf{y})\delta^t(\psi_j), \tag{3.31}$$

$$R^t(\psi_j) = \mathbb{E}_{q^t}(\psi_j) \pm 4\sqrt{Var_{q^t}(\psi_j)}. \tag{3.32}$$

This way of carrying out the VB-iteration is useful for the cases where the standard form of the VB-marginals cannot be identified.

Šmídl & Quinn (2006) presented a general approach to the VB method which can be applied to non-latent models. The VB algorithm by Beal (2003) is suitable for latent models only. For non-latent models, the algorithm can still be used by considering only the VBM-step for approximating the posterior distribution of the unknown parameters of the model. Both Beal (2003) and Šmídl & Quinn (2006) commented on the intractability issue of the VB method for complex models. Beal (2003) showed that the conjugate-exponential (CE) models result in tractable VB approximation but the non-CE models may not be amenable to the method. Šmídl & Quinn (2006) mention that the VB approximation is tractable only if the log-joint likelihood can be factorized over the unknowns. For example, the log-joint likelihood for CE-models may be separable over the components of the unknowns. The approach of the method by Šmídl & Quinn (2006) presented the VB marginals interacting with each other via their moments. Beal (2003) showed that the VB approximation for CE models are some function of the variational posterior expectations or VB-moments and sufficient statistics.

The approach by Beal (2003) finds the VB approximation by maximizing a lower bound on the marginal likelihood. The lower bound on the marginal likelihood as a function of VB approximation which may be used for model selection or check the accuracy of the approximation. The Šmídl and Quinn approach of the method

lacks such a measure of accuracy of the approximation. Šmídl & Quinn (2006) commented on the under-estimation of the variance of VB approximations (Wang & Titterington, 2005), but Beal (2003) overlooked it.

3.2.2 The restricted VB method

The VB method may be time consuming or intractable in some cases, e.g., the iterative VB algorithm may take longer to converge for complex models or highly multidimensional problems. The VB marginal may result in an intractable distribution or a non-standard distribution with complicated moments for non-CE models. To tackle such situations, Šmídl & Quinn (2006) proposed the restricted VB method.

Let the posterior distribution of unknowns $\psi = \{\psi_1, \psi'\}$ approximated with the VB approximation:

$$
\begin{aligned}
P(\psi_1, \psi'|\mathbf{y}) &\approx q(\psi_1, \psi'|\mathbf{y}), && (3.33)\\
&= q(\psi_1|\mathbf{y})q(\psi'|\mathbf{y}), && (3.34)\\
\text{where } q(\psi_1|\mathbf{y}) &\propto \exp\left[\mathbb{E}_{q(\psi'|\mathbf{y})} \log P(\psi_1, \psi', \mathbf{y})\right] \text{ and} && (3.35)\\
q(\psi'|\mathbf{y}) &\propto \exp\left[\mathbb{E}_{q(\psi_1|\mathbf{y})} \log P(\psi_1, \psi', \mathbf{y})\right]. && (3.36)
\end{aligned}
$$

Suppose, the VB marginal $q(\psi_1|\mathbf{y})$ is either intractable or leads to a time consuming iterative VB algorithm. If $q(\psi_1|\mathbf{y})$ is restricted or kept fixed to a known standard distribution, we need to compute only $q(\psi'|\mathbf{y})$. This greatly reduces the computational burden of VB approximation as we no longer need the iterative VB algorithm for a converged VB approximation as a minimum of a KL divergence. It also does not require to compute complicated VB marginals needed to compute other VB marginals.

The accuracy of the restricted VB approximation depends on the choice of the distribution for the intractable VB marginal. The accuracy of the VB approximation may reduce if the restricted VB-marginal increases the divergence between the approximation and the true posterior distribution. However, it may be useful in such cases, where computational difficulty mean that further compromise on approximation accuracy is needed.

It is an important question that how the error in the VB-approximation can be monitored during the VB-runs. As the method is based on the KL divergence which can provide an accuracy measure of the approximation. Since the true posterior distribution is known up to a proportionally constant so is this measure of divergence. It is mentioned before that minimizing a KL divergence is equivalent to maximizing a lower bound on the log-likelihood which can be computed easily as a function of the variational Bayes approximation and the complete log-likelihood. As the lower bound improves after every VB-iteration, the difference of the lower bounds at two successive VB-iterations can also provide a measure of error in approximation during the VB runs.

3.2.3 Gaussian variational approach

The Gaussian variational approach is a special type of density transform approach (Ormerod & Wand, 2010) to variational approximations in which a posterior density is approximated by a Gaussian density that minimizes a KL-divergence between the densities.

Consider the problem of determining the marginal distribution $P(\mathbf{y})$ (as given in Eq. 3.2) which is a key factor in the computation of the posterior distribution $P(\psi|\mathbf{y})$:

$$P(\psi|\mathbf{y}) = \frac{P(\psi,\mathbf{y})}{P(\mathbf{y})}, \tag{3.37}$$

$$P(\mathbf{y}) = \int_{\psi} P(\psi,\mathbf{y})\mathrm{d}\psi. \tag{3.38}$$

A lower bound on $\log P(\mathbf{y})$ which is often intractable as below:

$$\begin{aligned}
\log P(\mathbf{y}) &= \log P(\mathbf{y}) \int_{\psi} q(\psi|\mathbf{y})\mathrm{d}\psi, \\
&= \int_{\psi} q(\psi|\mathbf{y}) \log P(\mathbf{y})\mathrm{d}\psi, \\
&= \int_{\psi} q(\psi|\mathbf{y}) \log \frac{P(\psi,\mathbf{y})q(\psi|\mathbf{y})}{P(\psi|\mathbf{y})q(\psi|\mathbf{y})}\mathrm{d}\psi, \\
&= \int_{\psi} q(\psi|\mathbf{y}) \log \frac{P(\psi,\mathbf{y})}{q(\psi|\mathbf{y})}\mathrm{d}\psi + \int_{\psi} q(\psi|\mathbf{y}) \log \frac{q(\psi|\mathbf{y})}{P(\psi|\mathbf{y})}\mathrm{d}\psi,
\end{aligned} \tag{3.39}$$

where the second term in the R.H.S of above expression is a KL-divergence between the true posterior distribution $P(\psi|\mathbf{y})$ and its variational approximation $q(\psi|\mathbf{y})$. Since it is always a non-negative quantity, we get a lower bound on $\log P(\mathbf{y})$ as:

$$\log P(\mathbf{y}) \geq \log P(\mathbf{y}; q), \tag{3.40}$$

$$P(\mathbf{y}; q) = \exp\left[\int_{\psi} q(\psi|\mathbf{y}) \log \frac{P(\psi, \mathbf{y})}{q(\psi|\mathbf{y})} \mathrm{d}\psi\right]. \tag{3.41}$$

If $q(\psi|\mathbf{y}) = P(\psi|\mathbf{y})$, the lower bound $P(\mathbf{y}; q)$ becomes exactly equal to $P(\mathbf{y})$.

The Gaussian variational approach restricts $q(\psi|\mathbf{y})$ to a Gaussian density as:

$$q(\psi|\mathbf{y}) \equiv N(\psi; \mu_{\psi}, \Sigma_{\psi}), \tag{3.42}$$

where μ_{ψ} and Σ_{ψ} are the variational parameters to found such that the lower bound $P(\mathbf{y}; q)$ is maximized to get a tight bound on $P(\mathbf{y})$:

$$\mu_{\psi} = \arg\max_{\mu_{\psi}} \log P(\mathbf{y}; q), \tag{3.43}$$

$$\Sigma_{\psi} = \arg\max_{\Sigma_{\psi}} \log P(\mathbf{y}; q). \tag{3.44}$$

The Gaussian variational approximation may not be tractable if the above two expressions do not provide tractable solutions to μ_{ψ} and Σ_{ψ}.

The main advantage of the Gaussian variational approach over other variational methods is that it provides a standard (Gaussian) density approximation to an intractable distribution. It can further be helpful computing other posterior estimates, point or interval estimates. Unlike a simple Gaussian approximation, it is not based on a posterior mode which often requires a (computationally expensive) mode-finding iterative algorithm.

A mean field approximation (or the assumption of independence between the unknown components of ψ) may make the Gaussian approximation more tractable. More details on the approach can be found in Ormerod & Wand (2010) and Hall et al. (2011).

3.2.4 Variational tangent approach

As mentioned before, the VB method might fail to provide a tractable solution outside the conjugate-exponential family of models. In such a situation, the variational tangent approach may be employed in which the non-conjugate-exponential model is transformed into a conjugate-exponential form.

Suppose the complete-likelihood $P(\mathbf{y}, \psi)$ has a complex form such that a tractable variational approximation may not be obtained even when a suitable or preferred prior distribution is assumed (if the prior distribution is complex, it may be changed to a simple form). The variational tangent approach considers a tangent to the complex complete log-likelihood function $\log P(\mathbf{y}, \psi)$ and transforms it into a simple quadratic or linear form which facilitates the variational method for a tractable approximation.

The key to the tangent approach is to find a tangent to obtain a lower or upper bound of the function. There can be many ways to do this, e.g. convex duality, Taylor's expansion. We will discuss only the Taylor's expansion for the variational tangent approach. For details on convex duality see Jordan (1999) and Ormerod & Wand (2010).

Let us denote $\log P(\mathbf{y}, \psi)$ by $l(\psi, \mathbf{y})$ as a function of ψ given $\mathbf{y}$. To obtain a tangent (say $l(\psi_q, \mathbf{y})$) of $l(\psi, \mathbf{y})$ (a quadratic function in ψ), consider a Taylor's expansion of $l(\psi, \mathbf{y})$ around $\psi = \psi_q$:

$$\begin{aligned} l(\psi, \mathbf{y}) &\geq l(\psi, \mathbf{y}; \psi_q), \\ l(\psi, \mathbf{y}; \psi_q) &= l(\psi_q, \mathbf{y}) + (\psi - \psi_q)\frac{\partial}{\partial \psi} l(\psi, \mathbf{y})|_{\psi=\psi_q} \\ &\quad +0.5(\psi - \psi_q)^T \frac{\partial^2}{\partial \psi^2} l(\psi, \mathbf{y})|_{\psi=\psi_q}(\psi - \psi_q). \end{aligned} \tag{3.45}$$

This bound (tangent) may now be used instead of $\log P(\mathbf{y}, \psi)$ for a tractable varia-

tional approximation:

$$\begin{aligned} P(\psi|\mathbf{y}) &= \frac{P(\mathbf{y}|\psi)P(\psi)}{\int_{\psi} P(\mathbf{y}|\psi)P(\psi)\mathrm{d}\psi}, \\ &\approx \frac{\exp\{l(\psi,\mathbf{y};\psi_q)\}}{\int_{\psi}\exp\{l(\psi,\mathbf{y};\psi_q)\}\mathrm{d}\psi}, \\ &\equiv q(\psi|\mathbf{y}). \end{aligned} \tag{3.46}$$

Often $l(\psi,\mathbf{y};\psi_q)$ is quadratic which makes the variational tangent approximation $q(\psi|\mathbf{y})$ as a Gaussian density with a conjugate prior $P(\psi)$. The posterior parameters of this approximation are some functions of ψ_q, an unknown variational parameter to be found. To obtain ψ_q Jaakkola & Jordon (1999) derived an iterative algorithm based on the *EM* algorithm includes the E-step and M-step as described below:

1. E-step: at this step the expectation over $l(\psi,\mathbf{y};\psi_q)$ is considered with respect to $q(\psi|\mathbf{y})$:

$$Q(\psi_q;\mathbf{y}) = \int_{\psi} q(\psi|\mathbf{y})l(\psi,\mathbf{y};\psi_q)\mathrm{d}\psi, \tag{3.47}$$

2. M-step: this step computes ψ_q by maximizing $Q(\psi_q;\mathbf{y})$ with respect to ψ_q:

$$\psi_q = \arg\max_{\psi_q} Q(\psi_q;\mathbf{y}). \tag{3.48}$$

The function $Q(\psi_q;\mathbf{y})$ depends on the posterior-parameters of $q(\psi|\mathbf{y})$ and they are some function of ψ_q which requires to compute the functions iteratively.

This is the classical approach of variational tangent approximation which avoids KL divergence. Jaakkola & Jordon (1999) have described the variational tangent approach (with convex duality) for a Bayesian logistic regression model.

For multivariate ψ, the approximation of $P(\psi|\mathbf{y})$ might be computationally intensive or intractable even with this approach of variational tangent method.

3.3 VB approximation for CE models

Earlier in the chapter, it is mentioned many times that the VB method provides tractable solutions for the CE models and the models for which the log- joint like-

lihood can be factorized over the components of the unknown parameters. The natural form of the CE models let the log-joint likelihood explained in a factorized form. We proceed with the Šmídl and Quinn approach and derive a standard form of VB approximations for CE models. See Beal (2003) for the standard form of the approximation for latent CE models.

The standard forms of the likelihood $P(\mathbf{y}|\psi)$ and the prior distribution $P(\psi|\eta,\nu)$ of the CE models, as defined already in Chapter 2, are given below:

$$\begin{aligned} P(\mathbf{y}|\psi) &= \prod_{i=1}^{N} \left[g(\psi_1,\psi') \exp\left(\phi(\psi_1,\psi')^T u(\mathrm{y}_i)\right) f(\mathrm{y}_i) \right], \\ &= g(\psi_1,\psi')^N \exp\left(\phi(\psi_1,\psi')^T \sum_{i=1}^{N} u(\mathrm{y}_i) \right) \prod_{i=1}^{N} \left[f(\mathrm{y}_i) \right], \qquad (3.49) \\ P(\psi|\eta,\nu) &= h(\eta,\nu) g(\psi_1,\psi')^{\eta} \exp\left(\phi(\psi_1,\psi')^T \nu\right), \qquad (3.50) \end{aligned}$$

where $\sum_{i=1}^{N} u(\mathrm{y}_i) = \mathbf{t}(\mathbf{y})$ (say) is the sufficient statistics of $\psi = \{\psi_1,\psi'\}$, $\phi(\psi_1,\psi')$ is the natural parameter that linearly relates to $\mathbf{t}(\mathbf{y})$, $g(\psi_1,\psi')$ is the normalizing constant and η and ν are the hyper-parameters of ψ.

Consider the VB marginals of ψ_1 and ψ' as given in Eq. 3.35 and 3.36:

$$\begin{aligned} q(\psi_1|\mathbf{y}) &\propto \exp\Big[(\eta+N)\mathbb{E}_{q(\psi'|\mathbf{y})} \log g(\psi_1,\psi') + (\mathbf{t}(\mathbf{y})+\nu) \\ &\qquad \times \mathbb{E}_{q(\psi'|\mathbf{y})}\{\phi(\psi_1,\psi')^T\} \Big], \qquad (3.51) \\ q(\psi'|\mathbf{y}) &\propto \exp\Big[(\eta+N)\mathbb{E}_{q(\psi_1|\mathbf{y})} \log g(\psi_1,\psi') + (\mathbf{t}(\mathbf{y})+\nu) \\ &\qquad \times \mathbb{E}_{q(\psi_1|\mathbf{y})}\{\phi(\psi_1,\psi')^T\} \Big]. \qquad (3.52) \end{aligned}$$

The standard form (or the tractability) of the VB marginals $q(\psi_1|\mathbf{y})$ and $q(\psi'|\mathbf{y})$ depends on the form of the functions $\log g(\psi_1,\psi')$ and $\phi(\psi_1,\psi')$. If both functions factorized over ψ_1 and ψ' in such a way that could define the VB expectations (given in the above two equations) and which is possible for the CE models.

To demonstrate the standard form of the VB marginals for CE models, consider an example of an i.i.d univariate Gaussian model with the mean μ and the variance

σ^2. Suppose $\psi_1 = \mu$ and $\psi' = \sigma^2$. The likelihood of the model is defined as:

$$P(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{\mathrm{y}_i^2}{2\sigma^2} + \frac{\mu \mathrm{y}_i}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}. \tag{3.53}$$

Comparing the standard form of the exponential models (likelihood),

$$\begin{aligned} g(\mu, \sigma^2) &= \exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}, & (3.54)\\ \phi(\mu, \sigma^2) &= \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right), & (3.55)\\ u(\mathrm{y}_i) &= \left(-\frac{\mathrm{y}_i^2}{2}, \mathrm{y}_i\right). & (3.56) \end{aligned}$$

Defined below is the conjugate joint prior over μ and σ^2,

$$\begin{aligned} P(\mu, \sigma^2) &= P(\mu|\sigma^2)P(\sigma^2), & (3.57)\\ P(\mu|\sigma^2) &= N\left(\mu; m, \sigma^2\right), \\ &= \frac{1}{2\pi\sigma^2} \exp\{-\frac{\mu^2}{2\sigma^2} + \frac{m\mu}{\sigma^2} - \frac{m^2}{2\sigma^2}\}, & (3.58)\\ P(\sigma^2) &= 1. \end{aligned}$$

Comparing the standard form of the conjugate prior,

$$\begin{aligned} h(\eta, \nu) &= \frac{b^a}{\Gamma a}, & (3.59)\\ \eta &= 1, & (3.60)\\ \nu &= (-\frac{m^2}{2}, m). & (3.61) \end{aligned}$$

The functions $\log g(\mu, \sigma^2)$ and $\log \phi(\mu, \sigma^2)$ are factorized over μ and σ^2 such that the required VB expectations can be defined. The VB marginals over μ and σ^2 are

obtained as:

$$
\begin{aligned}
q(\mu|\mathbf{y}) &= N(\mu; m^\star, S^{\star\,2}), && (3.62)\\
S^{\star\,2} &= \frac{1}{N+1}\left[\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\right]^{-1}, && (3.63)\\
m^\star &= \frac{\sum_{i=1}^{N} \mathrm{y}_i + m}{N+1}, && (3.64)\\
q(\sigma^2|\mathbf{y}) &= \mathit{Inverse\ Gamma}(\sigma^2; a^\star, b^\star), && (3.65)\\
a^\star &= \frac{N+3}{2}, && (3.66)\\
b^\star &= \left[\frac{N+1}{2}\mathbb{E}_q(\mu^2) + \frac{1}{2}\left(\sum_{i=1}^{N}\mathrm{y}^2 + m^2\right) - \mathbb{E}_q(\mu)\left(\sum_{i=1}^{N}\mathrm{y}_i + m\right)\right] && (3.67)
\end{aligned}
$$

3.4 VB approximation and the Markov Property

The Markov property, along with the independence assumption and the factorized form of the log-joint likelihood, is an important factor to present the VB marginals in simple and tractable forms. Defined earlier in Chapter 2, the Markov property explains the conditional independence between the unknowns of a multivariate model. A VB marginal of an unknown, say ψ_i, depends on the VB moments of those other unknowns among ψ_{-i} (all other unknowns but ψ_i) which directly influence ψ_i under the conditional independence property.

Consider the DAG presented in Figure 3.1. Suppose we are interested in the computation of the VB marginal $q(\psi_1|\mathbf{y})$:

$$
\begin{aligned}
q(\psi_1|\mathbf{y}) &= q(\psi_1|\mathbf{y}, \bar{\psi}_{-1}), && (3.68)\\
&= q(\psi_1|\mathbf{y}, \bar{\psi}', \bar{\alpha}, \bar{\beta}). && (3.69)
\end{aligned}
$$

The VB marginal $q(\psi_1|\mathbf{y})$ depends on the VB-moments of ψ', α, β denoted by $\bar{\psi}'$, $\bar{\alpha}$ and $\bar{\beta}$ and the data $\mathbf{y}$ under the Markov property which says that an unknown ψ_1 is conditionally independent of all other unknowns that are not its parents. That is,

$$
q(\psi_1|\mathbf{y}) = q(\psi_1|\mathbf{y}, \bar{pa}_{\psi_1}), \tag{3.70}
$$

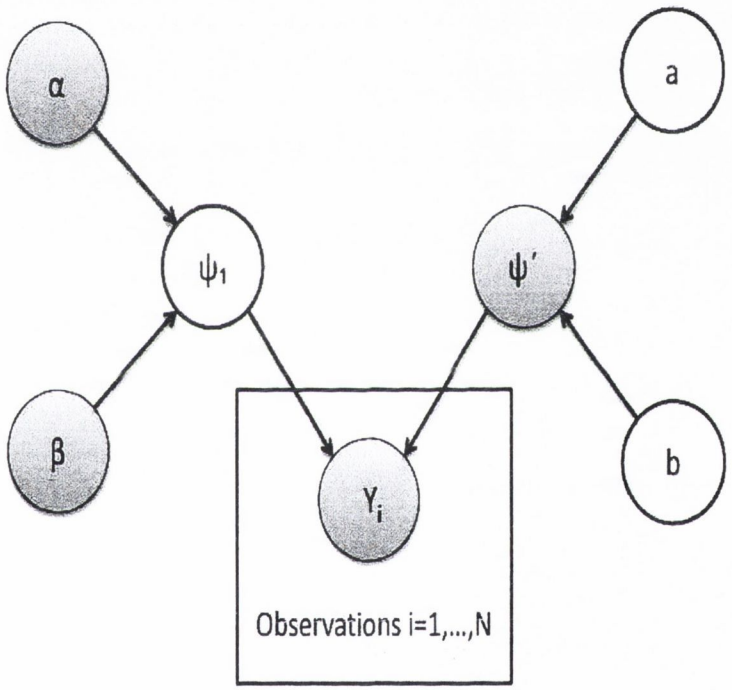

Fig. 3.1: A DAG showing conditional independence by the Markov property for a multivariate model. The DAG represents a Markov blanket around the node ψ_1 with the shaded nodes as its parents, children and co-parents. Nodes $\mathbf{y}_i$'s are children, α and β are parents and ψ' is the co-parent. ψ_1 is conditionally independent of nodes a and b given its parents.

where $\bar{pa}_{\psi_1} = \{\bar{\psi}', \bar{\alpha}, \bar{\beta}\}$ denotes the set of VB moments of the parents of ψ_1.

Thus without writing down the VB equations, due to the Markov property we are still able to find the unknowns which interact with a particular unknown in its VB marginal via their VB moments. The VB moments have important role in the definition of the VB marginals as they carry important information about a particular parameters contained with its parents. This property of information passing through VB moments is called variational message passing which describes the simple forms of the VB approximations for latent models through the Markov property, for details see Winn & Bishop (2005).

3.5 The VB method vs other methods of approximation of Bayesian computation

The VB method is a quick and straightforward method to apply. In this section, a comparison between the VB method and the other popular methods of approximation of Bayesian computation is presented for its accuracy, convergence and the computational time.

***EM* algorithm:** The *EM* algorithm is an optimization method which provides a maximum a posteriori (MAP) estimate (a modal value of a posterior density) of parameters for latent models. The algorithm converges to a local maximum of the posterior only. It may converge slowly for multi-modal and highly multi-dimensional models. Just like the *EM* algorithm, the VB method is also more effective for exponential models. The VB method approximates the uncertainty of the parameters, whereas the *EM* algorithm provides only a point estimate. Likewise the *EM* algorithm, the method reaches a local minimum of a KL divergence.

Laplace approximation: The Laplace approximation uses a quadratic approximation around the posterior mode of the unknowns. Hence, it may be inappropriate for multi-modal and non-symmetric posterior distributions. It needs mode-finding or optimization algorithms to find a posterior mode, e.g. Newton's method (Nocedal

& Wright, 2006). It requires to compute the Hessian matrix needed for finding the posterior mode (usually found with the Newton's method) which may be very difficult in large dimensions. Unlike the Laplace approximation, the VB method may provide non-Gaussian approximations to the posterior distributions. It also does not require to compute the Hessian matrix.

Integrated Nested Laplace Approximation (INLA): The INLA method is applicable for latent Gaussian models with only a few parameters. It uses a numerical integration to approximate the posterior marginals of parameters, which requires a limited number of unknown parameters in the model. As it is based on the Laplace approximation, it requires to compute the Hessian matrix of the posterior distribution. Unlike the INLA, the VB method with the assumption of posterior independence can be applied to a model with several parameters. Rue. et al. (2009) comments that the VB method may under-estimate the posterior variance, which makes it less accurate than the INLA method for models with a few parameters. Both the methods provide quick solutions to Bayesian computational problems.

MCMC method: MCMC is a simulation based method that can yield very accurate results in the long run. For multi-modal and multi-dimensional distributions, it suffers from slow convergence. The label switching problem for mixture models is a major drawback of the method. The VB method, in contrast, converges fast even for multi-dimensional models, but it lack some accuracy in the approximation due to posterior independence assumption. It may provide uni-modal approximations to multi-modal posterior distributions (for mixture models), therefore, it does not suffer from the label switching or symmetry problem, see the variational mean field approximation by Jaakkola (2000).

A pictorial representation of the comparison of the VB method with INLA and MCMC is given in Fig. 3.2. In the figure, VB, INLA and MCMC are qualitatively compared for their computational speed, accuracy and applicability to different variety of models. The MCMC method can be applied to a large variety of models generating very accurate results, but is very slow. INLA is fast and very accurate.

But, it is applicable to only a restricted class of models (latent Gaussian models with few parameters). The VB method is also very fast and can be applied to a wider class of models as compared to INLA, though it is less accurate. The next two chapters attempt to increase the accuracy and applicability (tractability) of the VB method for complex models while maintaining the computational speed.

3.6 Discussion

In short, the key idea of the VB method is to approximate the posterior density in terms of approximate marginals. The VB method enforces posterior independence between the subsets of the components of the unknown parameters. The method does not reach to a unique solution, however it can provide a local minimum of a KL-divergence. Approximate marginal posteriors of subsets of parameters, called VB-marginals, interact with each other via their moments. The assumption of posterior independence is necessary for the method to be computationally tractable, though it may lead to the under-estimation of the posterior variance. However, the independence assumption is not solely responsible for this under-estimation problem. A KL-divergence (used in the definition of the VB approximation) requires the coverage of the VB approximation smaller than that of the true posterior distribution to maintain its property of non-negativity. Thus, the definition of the VB approximation without the independence assumption itself leads to a smaller posterior variance. To solve this problem one could think of using the alternative definition of KL-divergence but that will result in the VB marginals equal to the true (unknown) marginal posterior distributions (Šmídl & Quinn, 2006). Some other methods that avoid this problem are, for example, the expectation propagation (Minka, 2001) or use of Fisher's information matrices (Wang & Titterington, 2005). Other alternative approaches of the VB method to tackle this problem need to be explored.

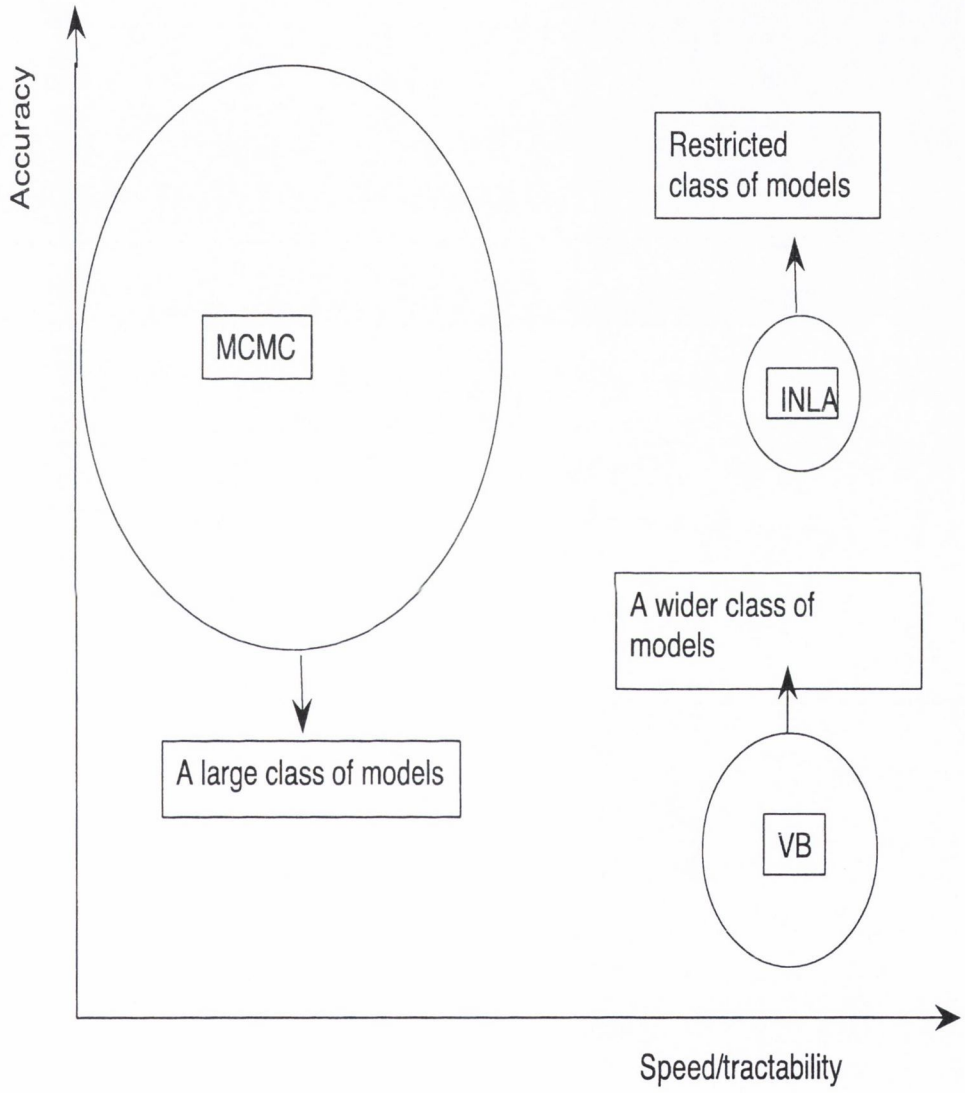

Fig. 3.2: A comparison between VB, INLA and MCMC for their accuracy, computational speed and applicability to models is shown. The area of circles represents the variety of models.

Chapter 4

VB approximation for Inverse Non-latent Regression

This chapter presents the VB approximations for inverse non-latent regression problems. Two types of non-latent regression problems are considered: conjugate-exponential regression model and non-conjugate-exponential model. To describe conjugate-exponential models, two regression models are considered: simple linear regression and quadratic regression. Non-conjugate-exponential models are explained by Poisson (log-linear) regression, mixture of Poisson and zero-inflated Poisson regression model. Inverse regression is described using Bayesian approach of inference. The aim of the chapter is to explain the VB approximation for the Bayesian inference in inverse non-latent regression problems. The chapter also builds up a simple understanding of the application of the VB method to a more complex inverse latent regression problem described in the next chapter.

4.1 Introduction

Inverse non-latent regression can be described as a method of finding predictions of an explanatory variable given observations on a response variable of a regression model. A regression model reflecting a relationship between a response variable (say

Y) and an explanatory variable (say X) can be defined as follows:

$$\mathrm{Y} = f(\mathrm{X}; \theta) \tag{4.1}$$

In the above equation, f is a parametric model that defines a relationship between a response variable, Y and an explanatory variable, X. The term θ represents unknown parameters in the model. To predict X for a given value(s) of Y, the model f should be learnt beforehand. In a statistical parametric model, model (f) learning relates to fitting a parametric model to data by estimating the model parameters (θ). The model fitting can further be used to predict X. It is often of interest to predict the explanatory variable(s) used to derive observations on a response variable. For example, consider a classical example of regression for dose of vitamins and weight gain. It might be of great importance to infer the amount of vitamins responsible for a desired weight gain.

Consider a Bayesian analysis of the inverse non-latent regression problem. Suppose, n i.i.d observations $\mathbf{y} = \{\mathrm{y}_i;\ i = 1 : n\}$ of Y corresponding to n values of X, $\mathbf{x} = \{\mathrm{x}_i;\ i = 1 : n\}$, are obtained from a distribution of Y given X and θ:

$$\begin{aligned} \mathbf{y} &\sim P(\mathbf{y}|\mathbf{x}, \theta), \\ P(\mathbf{y}|\mathbf{x}, \theta) &= \prod_{i=1}^{n} P(\mathrm{y}_i|\mathrm{x}_i, \theta), \end{aligned} \tag{4.2}$$

where θ is unknown. The inverse regression of an unknown X, denoted as X_{new}, for a new value of Y, y_{new}, can be found through its posterior distribution as given all the observations on X and Y:

$$P(\mathrm{X}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) = \int P(\mathrm{X}_{\text{new}}, \theta|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x})\mathrm{d}\theta, \tag{4.3}$$

$$\propto \int P(\mathrm{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \theta)P(\mathrm{X}_{\text{new}}|\theta)P(\theta|\mathbf{y}, \mathbf{x})\mathrm{d}\theta, \tag{4.4}$$

where $P(\mathrm{X}_{\text{new}}|\theta)$ is the prior distribution of X_{new} and $P(\mathrm{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \theta)$ is the likelihood of X_{new} given θ and y_{new}. The term $P(\theta|\mathbf{y}, \mathbf{x})$ is the posterior distribution of θ given $(\mathbf{y}, \mathbf{x})$ which should be computed in advance, since it does not depend on y_{new}.

Thus, the inverse regression of X_{new} can be carried out in two stages:

1. **Forward stage**: At this stage, the aim is to compute the posterior distribution of the unknown parameters θ of the regression model given data set $(\mathbf{x}, \mathbf{y})$ by Bayes' law as:

$$P(\theta|\mathbf{x}, \mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x}, \theta)P(\theta)}{\int_\theta P(\mathbf{y}|\mathbf{x}, \theta)P(\theta)\mathrm{d}\theta}. \tag{4.5}$$

   In the R.H.S of the above equation, $P(\mathbf{y}|\mathbf{x}, \theta)$ stands for the likelihood of θ given data on $\mathbf{x}$ and $\mathbf{y}$ and $P(\theta)$ denotes a prior distribution of θ.

2. **Inverse stage:** At the inverse stage, the posterior distribution of X_{new} given the data $(y_{new}, \mathbf{y}, \mathbf{x})$, $P(X_{new}|y_{new}, \mathbf{y}, \mathbf{x})$, is obtained by integrating out θ from the joint posterior distribution of X_{new} and θ, $P(X_{new}|y_{new}, \mathbf{y}, \mathbf{x})$, as expressed in Eq. 4.3–4.4.

To compute the posterior distribution $P(X_{new}|y_{new}, y, x)$, the R.H.S of Eq. 4.4 should be normalized and this requires the evaluation of an integral over the dimension of the parameters, which often remains intractable. However, if the integrands (posterior distribution of θ, prior distribution of X_{new} and its likelihood) belong to conjugate family of distributions, the posterior distribution $P(X_{new}|y_{new}, y, x)$ may be obtained as a standard distribution and hence the computation of the normalizing constant can be avoided.

It is not always necessary to describe Bayesian inference on the inverse regression in two stages. The previous work in the field of Bayesian inverse regression (Hunter & Lamboy, 1981; Racine-Poon, 1988) study θ and X_{new} jointly through their posterior distribution. The joint posterior distribution combines both the stages. As defined in Eq. 4.4, the marginal posterior distribution θ, $P(\theta|y, x)$, does not depend on y_{new}, hence can be computed in advance. The VB marginal of θ may depend on y_{new} and the VB moments of X_{new}. The combined VB approximation of θ and y_{new} is explained in Vatsa & Wilson (2010). The method may be straightforward as it avoids splitting the inference problem, though can be time consuming in case of several predictions to be studied independently. The inference method of this thesis is compared with that of Vatsa & Wilson (2010) in detail in Chapter 6.

Inverse probability and inverse regression:

The Bayesian inverse regression can be compared with the concept of 'inverse probability' (Isserlis, 1936; Stigler, 1986; Zabell, 1989; Dale, 1999) as both describe the inverse inference on an unknown. The inverse probability is a method of inverse inference that uses Bayes' theorem to find posterior distribution of an unobserved quantity given data. The use of inverse probability in inverse problems can be found in Dale (1999). Whereas, (Bayesian) inverse regression refers to learning only an unknown explanatory variable for a given observed value(s) of a response variable.

4.1.1 Models to explain the Inverse Non-Latent regression problem

Five non-latent regression models categorized in two classes of models:

- conjugate-exponential and
- non-conjugate-exponential (non-latent) models,

are described below to illustrate how the VB method performs for the inverse estimation for these sets of models.

Under the conjugate-exponential non-latent models, two regression models are considered:

1. **Simple linear regression problem:**

   Simple linear regression defines a linear relationship between an explanatory variable X and a response variable Y. A simple linear relation between X and Y can be defined as:

   $$\mathbb{E}(\mathrm{Y}|\mathrm{X} = \mathrm{x}) = \beta_0 + \beta_1 \mathrm{x} \tag{4.6}$$

   Eq. 4.6 can be rephrased as

   $$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{x}_i + \epsilon_i;\ i = 1 : n. \tag{4.7}$$

   In the above equation, the terms β_0 and β_1 are regression parameters. Corre-

sponding to given a value of X, x_i, Y_i is observed with some discrepancy. The ϵ' s are assumed to be independently and identically normally distributed:

$$\epsilon_i \sim N(0, \sigma^2);\ i = 1 : n, \tag{4.8}$$

The aim of the simple linear regression analysis is to estimate the unknown parameters β_0, β_1 and σ^2 to fit a linear relationship between explanatory and response variables. The estimation of unknown β_0, β_1 and σ^2 of simple linear regression is then used for inverse prediction of the explanatory variable.

2. **Quadratic regression problem:**

   A quadratic regression models a quadratic relationship between a response variable, Y, and an explanatory variable X:

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x + \beta_0 x^2. \tag{4.9}$$

   Given a sample of data points on Y and X, Eq. 4.9 can be represented as

$$Y_i = \beta_0 + \beta_1 x_i + \beta_0 x_i^2 + \epsilon_i;\ i = 1 : n. \tag{4.10}$$

   The term ϵ's in Eq. 4.10 are identically and independently normally distributed error terms with mean zero and variance σ^2 (same as defined in Eq. 4.8).

   At the forward stage of the inverse inference, the unknown parameters β_0, β_1, β_2 and σ^2 are estimated and then used to predict X for a new observation of Y at the inverse stage.

The inverse estimation by the VB method is discussed for non-conjugate-exponential models via the following models:

3. **Poisson Regression problem**

   Inverse Poisson regression can be described as a method of inverse regression to study an explanatory variable X for given observations (counts) on response

variable Y of a Poisson regression model. A Poisson regression fits a Poisson model to a set of count data on Y given data on X of the model. The log mean of Y is linearly related to X:

$$\log(\mathbb{E}(\mathrm{Y}|\mathrm{X} = \mathrm{x})) = \beta_0 + \beta_1 \mathrm{x}. \tag{4.11}$$

In the above equation, β_0 and β_1 are the unknown regression parameters. The regression parameters β_0 and β_1 are estimated at the forward stage given data on Y and X. The knowledge of β_0 and β_1 is then used for the inverse prediction of X for new counts Y.

4. **Mixture of Poisson Regression problem**

   In a mixture of Poisson regression problem the response variable are assumed be following a mixture of Poisson distributions. In this model, the counts are a mixture of regression component and a component that is independent of the explanatory variable. The mean of the mixture Poisson model is given as follows:

$$\mathbb{E}(\mathrm{Y}|\mathrm{X} = \mathrm{x}) = \pi \exp(\beta_0 + \beta_1 \mathrm{x}) + (1 - \pi)\mu. \tag{4.12}$$

   In the above equation, β_0 and β_1 are the unknown regression parameters and μ is the rate parameter of the second distribution of the mixture model. The term π is the probability that an observation of Y is obtained from the Poisson distribution with the rate parameter as a function of regression parameters. Therefore, there are four parameters to estimated at the forward stage of the problem given data on Y and X. Then the inverse prediction of X for new counts Y can be studied using the knowledge of these parameters already obtained at the forward stage.

5. **Zero-Inflated (ZI) Poisson Regression problem**

   Some real life count data with mixed behaviour are zero excessive. Such type of data can be well modeled with the ZI-Poisson distribution. In the ZI-Poisson model is a mixture of two generative process, one fits the excess of zero counts and another models the non-zero counts. The ZI-Poisson regression model is

defined as below:

$$P(\mathbf{Y}|\mathbf{x}, \beta_0, \beta_1, \pi) = \prod_{n=1}^{N} ZIP(\mathrm{Y}_i;\ \lambda_i, \pi_i), \tag{4.13}$$

$$\mathrm{ZIP}(\mathrm{Y}_i;\ \lambda_i, \pi_i) = \begin{cases} (1-\pi_i) + \pi_i e^{-\lambda_i}, & \text{if } \mathrm{Y}_i = 0; \\ \pi_i Poiss(\mathrm{Y}_i;\ \lambda_i), & \text{if } \mathrm{Y}_i > 0 \end{cases}$$

where the term ZIP stands for the ZI-Poisson distribution, λ is the rate parameter of the Poisson density in the ZI-Poisson model which relates the response variable $\mathbf{Y}$ to the explanatory variable $\mathbf{X}$ and $(1-\pi)$ is the probability of observing essential zero counts.

$$\lambda_i = \exp(\beta_0 + \beta_1 \mathrm{x}_i)\ \forall i, \tag{4.14}$$

$$\pi_i = \frac{\lambda_i}{1+\lambda_i}. \tag{4.15}$$

The inverse ZI-Poisson regression problem studies the inverse estimation of the explanatory variable $\mathrm{X}_{\mathrm{new}}$ given a new count y_{y} on the response variable of a ZI-Poisson regression model.

The Poisson regression problem describes the non-conjugate and exponential family of models whereas the mixture of Poisson and the zero-inflated Poisson regression problems are the examples of the non-conjugate and non-exponential family.

In the next section, the VB approximation to the Bayesian inference for the inverse non-latent regression problems is discussed. It should be noted that the analytical solutions to the Bayesian inverse problem can also be found for low dimensional problems numerically, that allows us to evaluate the performance of the VB approximation.

4.2 VB approximation to Inverse non-latent Regression Problem

The VB approximation for the inverse regression is described separately for the two stages:

VB approximation at the forward stage:

The aim of the forward stage is to compute the posterior distribution of θ given data. The unknown parameter θ of the non-latent regression models (discussed before) is multivariate, $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\}$. A VB approximation to the joint posterior distribution of θ is given as:

$$P(\theta_1, \theta_2, \ldots, \theta_p|\mathbf{x}, \mathbf{y}) \approx q_\theta(\theta_1, \theta_2, \ldots, \theta_p|\mathbf{x}, \mathbf{y}), \tag{4.16}$$

$$q_\theta(\theta_1, \theta_2, \ldots, \theta_P|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{p} q_{\theta_i}(\theta_i|\mathbf{x}, \mathbf{y}), \tag{4.17}$$

$$q_{\theta_i}(\theta_i|\mathbf{x}, \mathbf{y}) \propto \exp\left[\mathbb{E}_{q_{\theta_{-i}}(\theta_{-i}|\mathbf{x},\mathbf{y})}[\log(P(\mathbf{y}|\mathbf{x}, \theta)P(\theta))]\right], \tag{4.18}$$

$$\text{where } q_{\theta_{-i}}(\theta_{-i}|\mathbf{x}, \mathbf{y}) = \prod_{j\neq i}^{p} q_{\theta_j}(\theta_j|\mathbf{x}, \mathbf{y}). \tag{4.19}$$

The functional form of the VB approximation for the different models are described in the section for VB solution.

VB approximation at the inverse stage:

At the inverse stage, the aim is to compute the marginal posterior distribution $P(\mathrm{X}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x})$ by integrating out θ from the joint posterior distribution $P(\mathrm{X}_{\text{new}}, \theta|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x})$. The regular VB method to approximate $P(\mathrm{X}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x})$ suggests to find an approximation to $P(\mathrm{X}_{\text{new}}, \theta|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x})$ and present it as a product of approximations to the marginals:

$$P(\mathrm{X}_{\text{new}}, \theta|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) \approx q(\mathrm{X}_{\text{new}}, \theta|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x}), \tag{4.20}$$

$$P(\mathrm{X}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) \approx q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x}), \tag{4.21}$$

$$q(\mathrm{X}_{\text{new}}, \theta|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) = q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x})q_\theta(\theta|\mathrm{y}_{\text{new}}, \mathbf{y}, \mathbf{x}). \tag{4.22}$$

An iterative VB method makes the problem of parameter estimation computationally expensive. It requires an approximation of $P(\mathrm{X}_{\mathrm{new}}, \theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})$ jointly every time a prediction of $\mathrm{X}_{\mathrm{new}}$ to be studied. Also, it does not fulfill the definition of inverse regression method to estimate θ first then invert the model to predict $\mathrm{X}_{\mathrm{new}}$.

The restricted VB method can be used to approximate the marginal posterior distribution of $\mathrm{X}_{\mathrm{new}}$. The idea of the restricted VB method is to compute a non-iterative VB approximation of $P(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})$. By the definition of the method, $P(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})$ can be approximated as:

$$
\begin{aligned}
P(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) &\approx \bar{q}_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}), && (4.23)\\
\bar{q}_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) &\propto \exp\left[\mathbb{E}_{P(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})} \log P(\theta, \mathrm{X}_{\mathrm{new}}, \mathrm{y}_{\mathrm{new}}|\mathbf{y}, \mathbf{x})\right], && (4.24)\\
&= \exp[\mathbb{E}_{P(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})} \log(P(\mathrm{y}_{\mathrm{new}}|\theta, \mathrm{X}_{\mathrm{new}}) \\
&\qquad \times P(\mathrm{X}_{\mathrm{new}}) P(\theta|\mathbf{x}, \mathbf{y}))], && (4.25)\\
&\propto \exp\left[\mathbb{E}_{P(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})} \log[P(\mathrm{y}_{\mathrm{new}}|\theta, \mathrm{X}_{\mathrm{new}}) P(\mathrm{X}_{\mathrm{new}})]\right] && (4.26)\\
\bar{q}_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) &\approx \exp\left[\mathbb{E}_{\tilde{P}(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})} \log[P(\mathrm{y}_{\mathrm{new}}|\theta, \mathrm{X}_{\mathrm{new}}) P(\mathrm{X}_{\mathrm{new}})]\right] && (4.27)
\end{aligned}
$$

A possible choice for $\tilde{P}(\theta|\mathrm{y}_{\mathrm{new}}, \mathrm{y}, \mathrm{x})$ is

$$\tilde{P}(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) = q_\theta(\theta|\mathbf{y}, \mathbf{x}). \tag{4.28}$$

The approximation $q_\theta(\theta|\mathbf{y}, \mathbf{x})$ is found at the forward stage. It should be noted that it is not conditioned on the new observation $\mathrm{y}_{\mathrm{new}}$. If y is a large set of observations compared to $\mathrm{y}_{\mathrm{new}}$, $\bar{q}_\theta(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})$ is very close to $q_\theta(\theta|\mathbf{y}, \mathbf{x})$.

4.2.1 Comparison of the VB approximation for $\mathrm{X}_{\mathrm{new}}$ with the results from other methods.

For the comparison of the restricted VB approximation, the marginal posterior distribution of $\mathrm{X}_{\mathrm{new}}$ can be computed in following ways:

1. Compute the true posterior distributions with a numerical integration method at both the stages.

2. Approximate the joint posterior distribution of parameters with the VB method at the forward stage and then compute the true marginal posterior distribution of $\mathrm{X_{new}}$ by a numerical integration method at the inverse stage.

3. Use the VB method at both the stages (restricted VB at the inverse stage) to approximate the posterior distributions.

4. Compute the true posterior distribution of the unknown parameters by the MCMC method at the forward stage. At the inverse stage, use the MCMC samples of the parameters to compute the posterior distribution of $\mathrm{X_{new}}$ by the Monte Carlo integration.

5. Use other variational methods, such as the variational tangent approach and Gaussian variational approach to compare with the VB approximations in case of the non-conjugate-exponential models.

The posterior distribution of $\mathrm{X_{new}}$ by the Monte Carlo methods:
As described in Chapter 2, the Monte Carlo methods may provide true results in the long run. To obtain the true posterior distribution of $\mathrm{X_{new}}$ the Monte Carlo integration may be considered which uses the MCMC posterior samples of the parameter θ. The posterior distribution of $\mathrm{X_{new}}$ by the Monte Carlo method is given as follows:

$$P_{MC}(\mathrm{X_{new}}|\mathrm{y_{new}}, \mathbf{x}, \mathbf{y}) \propto \sum_{m=1}^{MS} \left[P(\mathrm{y_{new}}|\mathrm{X_{new}}, \theta^{mcmc}) P(\mathrm{X_{new}}) \right] \tag{4.29}$$

where θ^{mcmc} is the MCMC posterior samples of θ. The term MS denotes the number of the MCMC samples. If MS is large enough, the posterior distribution of $\mathrm{X_{new}}$, $P_{MC}(\mathrm{X_{new}}|\mathrm{y_{new}}, \mathbf{x}, \mathbf{y})$, is by the Central Limit theorem close to the true posterior distribution of $\mathrm{X_{new}}$. It should be noted that the Monte Carlo methods may provide very accurate result but they are very time consuming.

The approximation by other variational methods for non-conjugate-exponential models are discussed later in the chapter.

4.2.2 Evaluation of VB approximation

A leave-one-out cross validation technique is used to test the accuracy of the VB approximation. A data set on Y and X generated from the true distribution is partitioned into two sets: training and test data set. The training data set consists all but one data point. The left out data point is considered for the validation of the approximation at the inverse stage. The training data set is used to estimate the regression parameters. Then, each of the data of Y is used once independently to predict the corresponding values of X. A measure of accuracy is described in terms of percentage of (all leave-one-out) test data of X lie inside its 95% Highest Posterior Density (HPD) region.

4.2.3 VB solution to Inverse Simple Linear Regression

In a simple linear regression model, the parameter θ is a set of unknown regression parameters β_0 and β_1 and a variance parameter σ^2;

$$\theta = \{\beta_0, \beta_1, \sigma^2\}.$$

As the underlying model is linear with normally distributed error terms, the likelihood of the parameters is Gaussian:

$$P(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} N(\mathrm{y}_i; \mu_{\mathrm{y}_i}, \sigma^2), \tag{4.30}$$

$$\text{where } \mu_{\mathrm{y}_i} = \beta_0 + \beta_1 \mathrm{x}_i. \tag{4.31}$$

The prior distributions on β_0 and β_1 are assumed to be Gaussian with zero means and variances large enough to show ignorance about the parameters. The variance parameter σ^2 is assumed to have an non-informative inverse Gamma prior.

$$P(\beta_0) = N(0, S^2_{\beta_0}), \tag{4.32}$$

$$P(\beta_1) = N(0, S^2_{\beta_1}), \tag{4.33}$$

$$P(\sigma^2) = \textit{Inverse Gamma}(a, b). \tag{4.34}$$

In the regression models of this chapter, the explanatory variables are assumed to be controlled and so possess no random behavior. With the Bayesian analysis, any unknown is assumed to be random. To show lack of prior knowledge on randomness in X_{new}, an improper prior on X_{new} is assumed:

$$P(\mathrm{X}_{\text{new}}) \propto 1. \tag{4.35}$$

VB solution at Forward stage:

A VB approximation to the joint posterior distribution $P(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x})$ is presented as follows:

$$P(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x}) \approx q(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x}), \tag{4.36}$$

where $q(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x})$ is found by the VB method as:

$$\begin{aligned}
q(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x}) &= q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x}) q_{\sigma^2}(\sigma^2|\mathbf{y}, \mathbf{x}), && (4.37)\\
\text{where, } q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) &= N(\mu^{\star}_{\beta_0}, S^{2\,\star}_{\beta_0}), && (4.38)\\
q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x}) &= N(\mu^{\star}_{\beta_1}, S^{2\,\star}_{\beta_1}), && (4.39)\\
q_{\sigma^2}(\sigma^2|\mathbf{y}, \mathbf{x}) &= \textit{Inverse Gamma}(a^{\star}, b^{\star}). && (4.40)
\end{aligned}$$

The VB-parameters (defining the VB marginals) are presented below:

$$\begin{aligned}
S^{2\,\star}_{\beta_0} &= \left[\frac{1}{S^2_{\beta_0}} + n\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\right]^{-1},\\
\mu^{\star}_{\beta_0} &= \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\left[\sum_{i=1}^{n} \mathrm{y}_i - \mathbb{E}_q(\beta_1)\sum_{i=1}^{n} \mathrm{x}_i\right] S^{2\,\star}_{\beta_0},\\
S^{2\,\star}_{\beta_1} &= \left[\frac{1}{S^2_{\beta_1}} + \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\sum_{i=1}^{n} \mathrm{x}_i^2\right]^{-1},\\
\mu^{\star}_{\beta_1} &= \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\left[\sum_{i=1}^{n} \mathrm{y}_i\mathrm{x}_i - \mathbb{E}_q(\beta_0)\sum_{i=1}^{n} \mathrm{x}_i\right] S^{2\,\star}_{\beta_1},\\
a^{\star} &= a + \frac{n}{2},
\end{aligned}$$

$$b^{\star} = b + \frac{1}{2}[\sum_{i=1}^{n} \mathrm{y}_i^2 + n\mathbb{E}_q(\beta_0^2) + \mathbb{E}_q(\beta_1^2)\sum_{i=1}^{n} \mathrm{x}_i^2 - 2\mathbb{E}_q(\beta_0)\sum_{i=1}^{n} \mathrm{y}_i - 2\mathbb{E}_q(\beta_1)\sum_{i=1}^{n} \mathrm{y}_i\mathrm{x}_i + 2\mathbb{E}_q(\beta_0)\mathbb{E}_q(\beta_1)\sum_{i=1}^{n} \mathrm{x}_i]. \quad (4.41)$$

VB marginals on the regression parameters β_0 and β_1 are recognized as Gaussian densities and VB marginal of σ^2 is an inverse Gamma density. The (posterior) parameters of the VB marginals are presented as functions of moments (expectation of particular functions) of other unknown parameters.

VB approximation at the inverse stage:

The posterior distribution of $\mathrm{X}_{\mathrm{new}}$ given data $\{\mathrm{y}, \mathrm{x}, \mathrm{y}_{\mathrm{new}}\}$ is computed as

$$\begin{aligned} P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y}_{\mathrm{new}}) &\propto \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}}, \beta_0, \beta_1, \sigma^2)P(\mathrm{X}_{\mathrm{new}}) \\ &\quad \times P(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\sigma^2, \quad (4.42) \\ &\approx \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}}, \beta_0, \beta_1, \sigma^2)P(\mathrm{X}_{\mathrm{new}}) \\ &\quad \times q(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\sigma^2. \quad (4.43) \end{aligned}$$

Using $q(\beta_0, \beta_1, \sigma^2|\mathbf{y}, \mathbf{x}) = q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x})q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x})q_{\sigma^2}(\sigma^2|\mathbf{y}, \mathbf{x})$ in the above integral:

$$P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y}_{\mathrm{new}}) \approx \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}}, \beta_0, \beta_1, \sigma^2)P(\mathrm{X}_{\mathrm{new}})q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) \times q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x})q_{\sigma^2}(\sigma^2|\mathbf{y}, \mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\sigma^2. \quad (4.44)$$

The integral in Eq. 4.44 is not in a closed form. As it is a low dimensional integration problem, an analytical solution can be found by a numerical integration. For a quick and tractable VB approximation, the restricted VB method is applied to

approximate $P(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$:

$$P(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}) \approx \tilde{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}), \tag{4.45}$$

$$\tilde{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}) \propto \exp[\mathbb{E}_{q(\beta_0,\beta_1,\sigma^2|\mathbf{y},\mathbf{x})} \log[P(\mathrm{y_{new}}|\beta_0, \beta_1, \sigma^2, \mathrm{X_{new}}) \times P(\mathrm{X_{new}})]], \tag{4.46}$$

$$\tilde{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}) = N(\mu_{\mathrm{X}^\star}, S^2_{\mathrm{X}^\star}), \tag{4.47}$$

$$S^2_{\mathrm{X}^\star} = \left[\mathbb{E}_q(\beta_1^2)\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\right]^{-1}, \tag{4.48}$$

$$\mu_{\mathrm{X}^\star} = \mathbb{E}_q(\beta_1)\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)[\mathrm{y_{new}} - \mathbb{E}_q(\beta_0)]\, S^2_{\mathrm{X}^\star}. \tag{4.49}$$

The restricted VB approximation to $P(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$ obtained is a Gaussian density with a Gaussian likelihood and an improper prior.

Evaluation of VB approximation:

For the accuracy check of the VB approximation, the 95% HPD region of $\mathrm{X_{new}}$ is obtained as $\mu_{\mathrm{X}^\star} \pm 2\sqrt{S^2_{\mathrm{X}^\star}}$.

Result:

A sample of fifty equally-spaced values of an explanatory variable X is assumed, fifty corresponding values on a response variable Y are generated from a Gaussian distribution with mean $\beta_0 + \beta_1 \mathrm{X}$ and variance σ^2. The data are simulated with parameters $\beta_0 = 0.1$, $\beta_1 = 0.9$ and $\sigma^2 = 0.02$. The VB method is applied to carry out inverse regression analysis of X for a new value of Y.

At the forward stage, the VB method is used to approximate the joint posterior distribution of the regression and variance parameters. The true posterior distribution by the MCMC and a numerical integration method is computed for the comparison of the VB results. The comparison between the true and the VB marginals are shown in Fig. 4.1. It is clear that the VB-variance of β_0 and β_1 are underestimated. This under-estimation of posterior variance may be a result of the independence assumption of the method. By this assumption, the method assumes

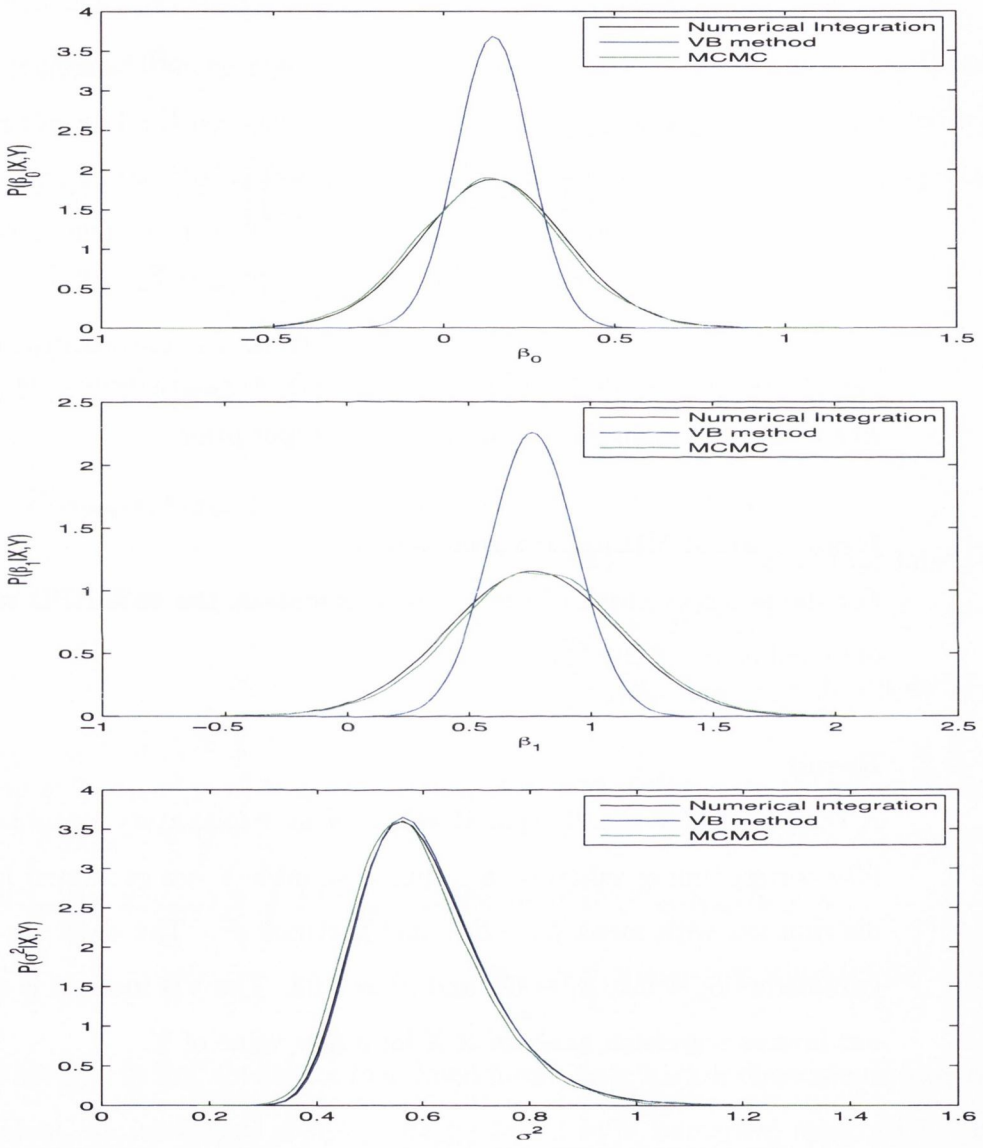

Fig. 4.1: The comparison of true posterior distribution of the regression parameters β_0 (top), β_1 (middle) and variance σ^2 (bottom) of a simple linear regression problem by a numerical integration (black) and by the MCMC (green) and VB approximation (blue) is shown. The VB marginal of β_0 and β_1 show under-estimated posterior variance, whereas the VB marginal of σ^2 is quite close to its true marginal posterior distribution.

the posterior covariance between the parameters zero which might make the posterior variances small if the parameters are not independent. The VB marginal over σ^2 matches with the true result. It shows that σ^2 is less correlated with β_0 and β_1 given sufficient data on X and Y and so is less effected with the independence assumption of the VB method. The approximation at the forward stage can be improved by allowing dependence between β_0, β_1.

The true marginal posterior distribution of X by the Monte Carlo integration and by a numerical integration and its VB approximation by the restricted VB approximation are shown in Fig. 4.2. The underestimation of posterior (VB) variance of the regression parameters is clearly reflected in the inverse estimation. As shown in Fig. 4.3, the true posterior distribution of $\mathrm{X_{new}}$ has a very large variance. But the posterior variance of $\mathrm{X_{new}}$ by the VB method is under-estimated. However, the under-estimation of the posterior variance is unavoidable with the VB method even if the independence assumption is ignored.

A validity check on the inverse regression is performed with a leave-one-out-cross validation technique (as explained in Section 4.2.2). There are no data points of $\mathrm{X_{new}}$ outside their 95% HPD region, i.e. a 100% coverage is achieved for the true posterior distribution and the VB approximation. It is an indication of large posterior variance of $\mathrm{X_{new}}$. Fig. 4.3 also suggests that the 95% HPD of $\mathrm{X_{new}}$ (given a new data point, $\mathrm{y_{new}}$) is too wide.

If estimated by classical approach of inference (least square method) as shown in Eq. 4.50–4.51, the variance of $\mathrm{X_{new}}$ is a function of σ^2 and β_1. As β_1 tends to zero $V(\mathrm{X_{new}})$ approaches to ∞. A detailed description on the larger posterior variance can be found on (Hoadley, 1970; Hunter & Lamboy, 1981).

$$\hat{\mathrm{X}} = \frac{\mathrm{y}}{\beta_1} - \frac{\beta_0}{\beta_1}, \tag{4.50}$$

$$V(\mathrm{X}) = \frac{\sigma^2}{\beta_1^2}, \tag{4.51}$$

An improper prior on $\mathrm{X_{new}}$ is assumed in the Bayesian prediction. Therefore, all the posterior information on $\mathrm{X_{new}}$ comes from its likelihood given data. Other parameters have been integrated out. Hence, its posterior variance is a function of

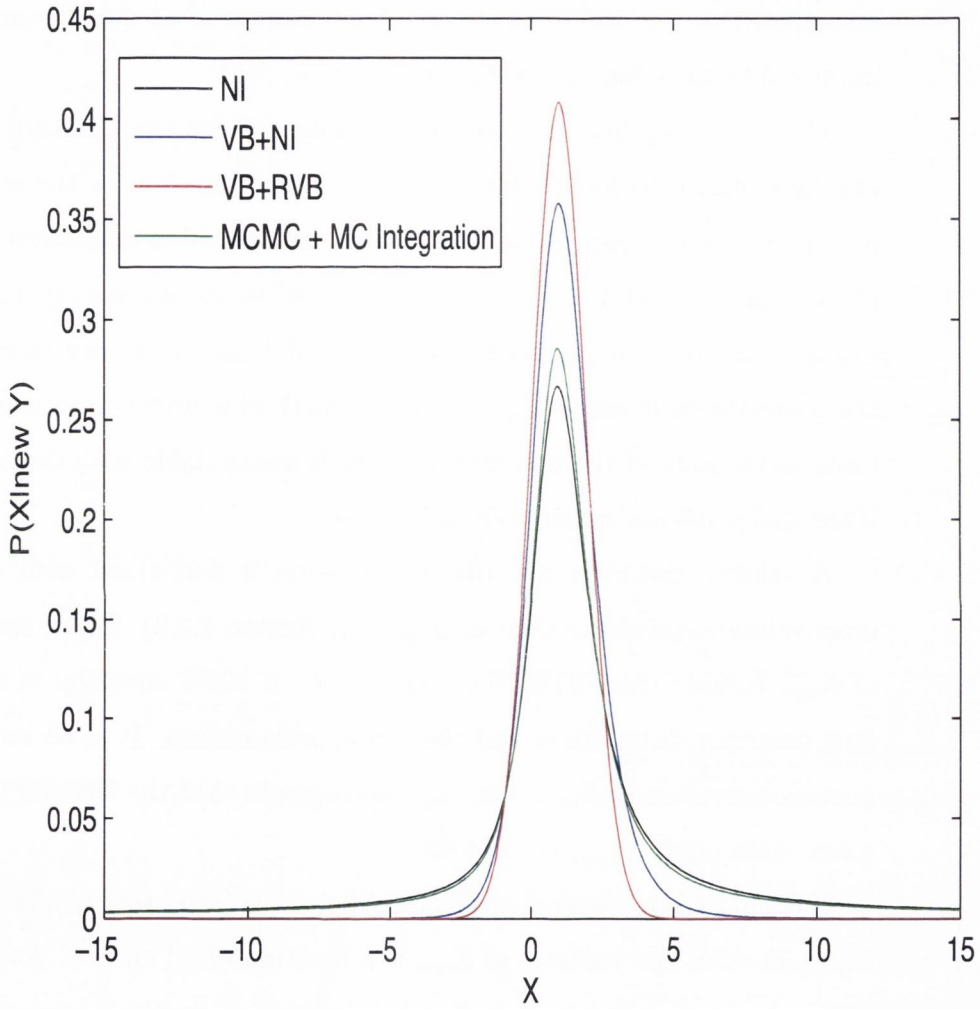

Fig. 4.2: The comparison of the true posterior distributions of an explanatory variable X_{new} (given $Y_{new} = 0.93$) of a simple linear regression problem (with an improper prior) by a numerical integration (black) at both the stages of inference, by the MCMC at the forward and the Monte Carlo integration (green) at the inverse stage of inference, the approximations by the VB method at the forward stage and a numerical integration (Red) and by the restricted VB method (Blue) at the inverse stage is shown. The true posterior uncertainty is large, whereas the VB variance is under-estimating the true variance.

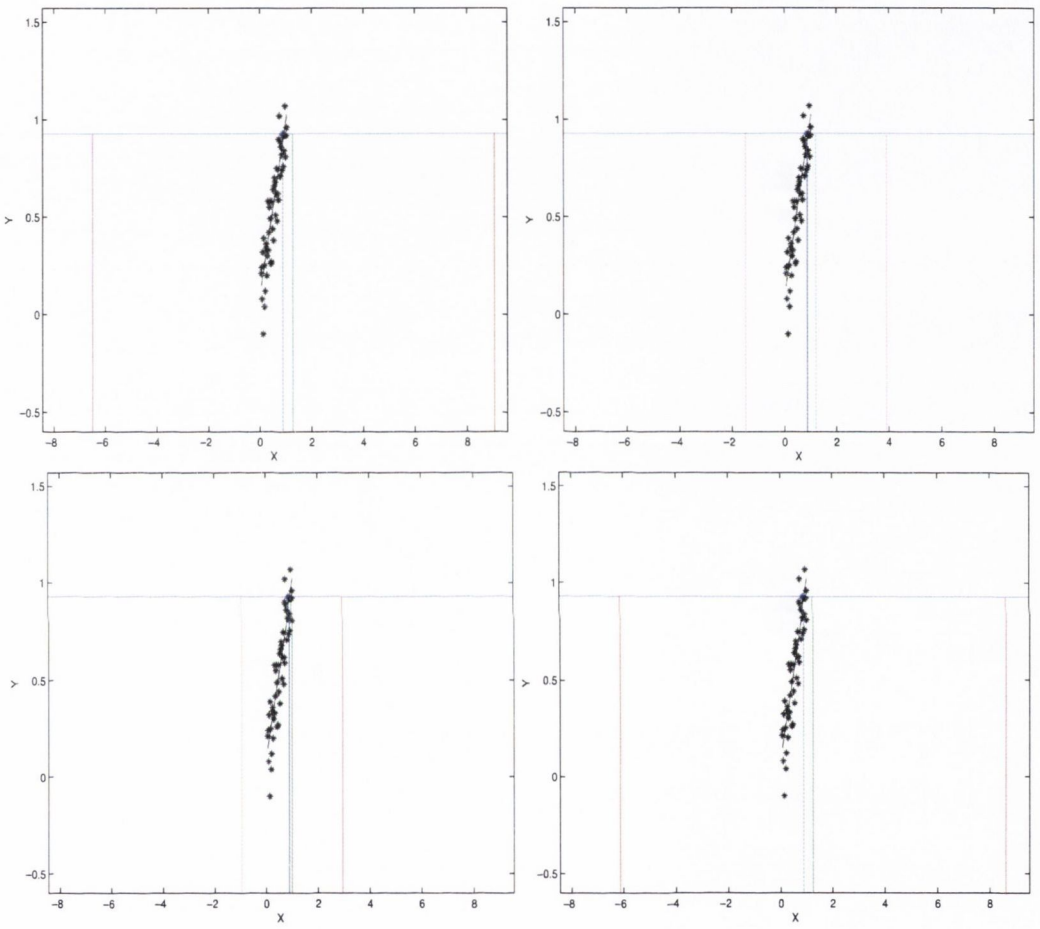

Fig. 4.3: Posterior estimates of an explanatory variable X_{new} of a simple linear regression problem with an improper prior by numerical integration (NI) (the top left), by the VB method and NI at forward and inverse stage respectively (the top right), by the VB method and RVB at the forward and inverse stage respectively (the bottom left) and by the MCMC and the Monte Carlo integration at forward and inverse stage respectively (the bottom right). The comparison of the true values (blue) and the estimation (green) of X_{new}, with the lower and upper bounds (in red) of 95% HPD region, is shown. The black asterisks are training data and the blue is for a new data Y_{new}. The 95% HPD region are very wide due to the large posterior variance which shows a 100% coverage.

β_1 (true)	$\mathbb{E}_q(\beta_1)$	$\mathbb{E}_q(\beta_1^2)$	$V(\mathrm{X_{new}})$ (VB + NI)	$V(\mathrm{X_{new}})$ (VB + RVB)
0.5	0.5478	0.3321	8.4096	1.846
0.9	0.8887	0.8213	1.0686	0.7339
1.5	1.2659	1.6343	0.4559	0.3730
2.5	2.161	4.7024	0.1476	0.1327

Table 4.1: For different values of the regression parameter β_1 of a simple linear regression problem, different results by the VB method and numerical integration and by the regular and restricted VB method are presented. It shows that a change in the value of β_1 inversely affects the posterior variance of an explanatory $\mathrm{X_{new}}$. As the value of β_1 increases, the difference between the posterior variance by the two methods decreases. However, the values of VB variance by the restricted VB method is moderately small for large or small values of β_1.

posterior mean of β_1 and σ^2. As $\mathbb{E}(\beta_1) \to 0$, $V(\mathrm{X_{new}}) \to \infty$, reflected Fig. 4.3 and Table 4.1. From Table 4.1, it can be also concluded that for small values of β_1 the posterior variances of $\mathrm{X_{new}}$, $V(\mathrm{X_{new}})$, by the VB method and a numerical integration method differ a lot. As β_1 increases, the difference between the values of the posterior variances decreases. Clearly, if $\beta_1 = 0$, data on $\mathbf{Y}_{\mathrm{new}}$ provides no information about $\mathrm{X_{new}}$ and it makes the likelihood flat. If the prior is also non-informative or improper, the posterior distribution is also flat with heavy tails as is experienced in the result. A strong prior over $\mathrm{X_{new}}$ is needed in order to obtain a well-defined posterior.

It is experienced from running many such VB-experiments that a big discrepancy in $\mathrm{y_{new}}$ (departure from the mean value) may increase the bias in the inverse estimation if the data size is not big.

Both the problems, the bigger variance and an increase in bias can be solved (to some extent) with the assumption of a suitable and informative prior on $\mathrm{X_{new}}$ (Hoadley, 1970). It may be suggested to use more than one new data point to predict $\mathrm{X_{new}}$ accurately.

Fig 4.4 and 4.5 shows the VB approximation and its comparison with the results by other methods for the same regression model with a proper prior over $\mathrm{X_{new}}$. It can understood from the figure that the assumption of a proper prior can solve the problem of large posterior variance in case of the non-informative data. Even with a proper prior and a smaller posterior variance, the coverage under the 95%

HPD region of X_{new} is still 100% for the true posterior distribution and for the VB approximation (checked by the leave-one-out cross-validation technique) which shows that there is no error in the inverse estimation of the explanatory variable for given observations on the response variable of this particular problem.

Hence, it can be suggested that the under-estimation of the variance by the VB method (in the case of the assumption of an improper prior) for the inverse simple linear regression problem is a blessing not a curse.

4.2.4 VB solution to Inverse Quadratic Regression

In a quadratic regression model, there are four unknown parameters to be estimated: regression parameters β_0,β_1 and β_2 and a variance parameter σ^2;

$$\theta = \{\beta_0, \beta_1, \beta_2, \sigma^2\}.$$

The likelihood of the parameters is Gaussian:

$$P(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1, \beta_2, \sigma^2) = \prod_{i=1}^{n} N(\mathrm{y}_i; \mu_{\mathrm{y}_i}, \sigma^2), \tag{4.52}$$

$$\text{where } \mu_{\mathrm{y}_i} = \beta_0 + \beta_1 \mathrm{x}_i + \beta_2 \mathrm{x}_i^2. \tag{4.53}$$

As in the simple linear regression problem, the prior distributions of β_0, β_1 and β_2 are assumed to be (non-informative) Gaussian with zero means and large variances. A non-informative inverse Gamma prior is assumed over σ^2:

$$P(\beta_0) = N(0, S_{\beta_0}^2), \tag{4.54}$$

$$P(\beta_1) = N(0, S_{\beta_1}^2), \tag{4.55}$$

$$P(\beta_2) = N(0, S_{\beta_2}^2), \tag{4.56}$$

$$P(\sigma^2) = \textit{Inverse Gamma}(a, b). \tag{4.57}$$

An improper prior on X_{new} is assumed:

$$P(\mathrm{X}_{\mathrm{new}}) \propto 1. \tag{4.58}$$

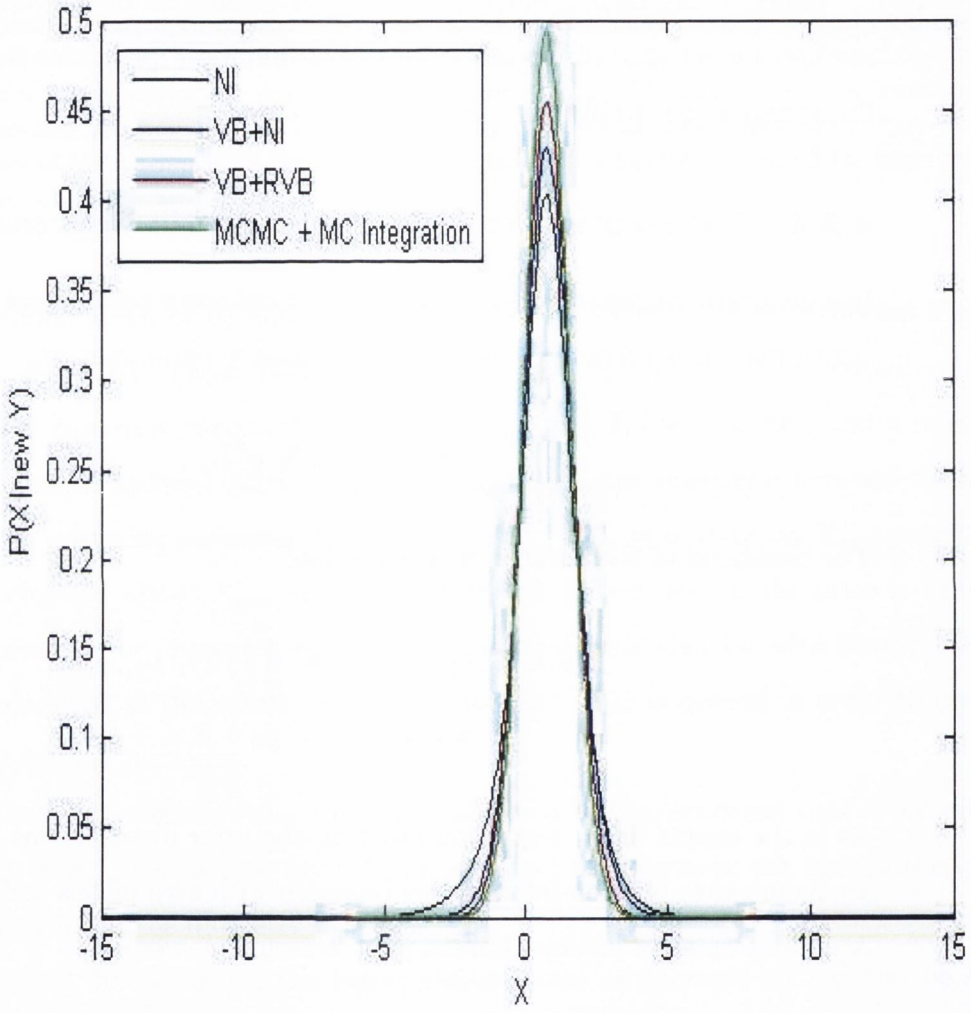

Fig. 4.4: The comparison of the true posterior distributions of an explanatory variable X_{new} (given $Y_{new} = 0.93$) of a simple linear regression problem (with a normal prior) by a numerical integration (black) at both the stages of inference, by the MCMC at the forward and the Monte Carlo integration (green) at the inverse stage of inference, the approximations by the VB method at the forward stage and a numerical integration (Red) and by the restricted VB method (Blue) at the inverse stage is shown.

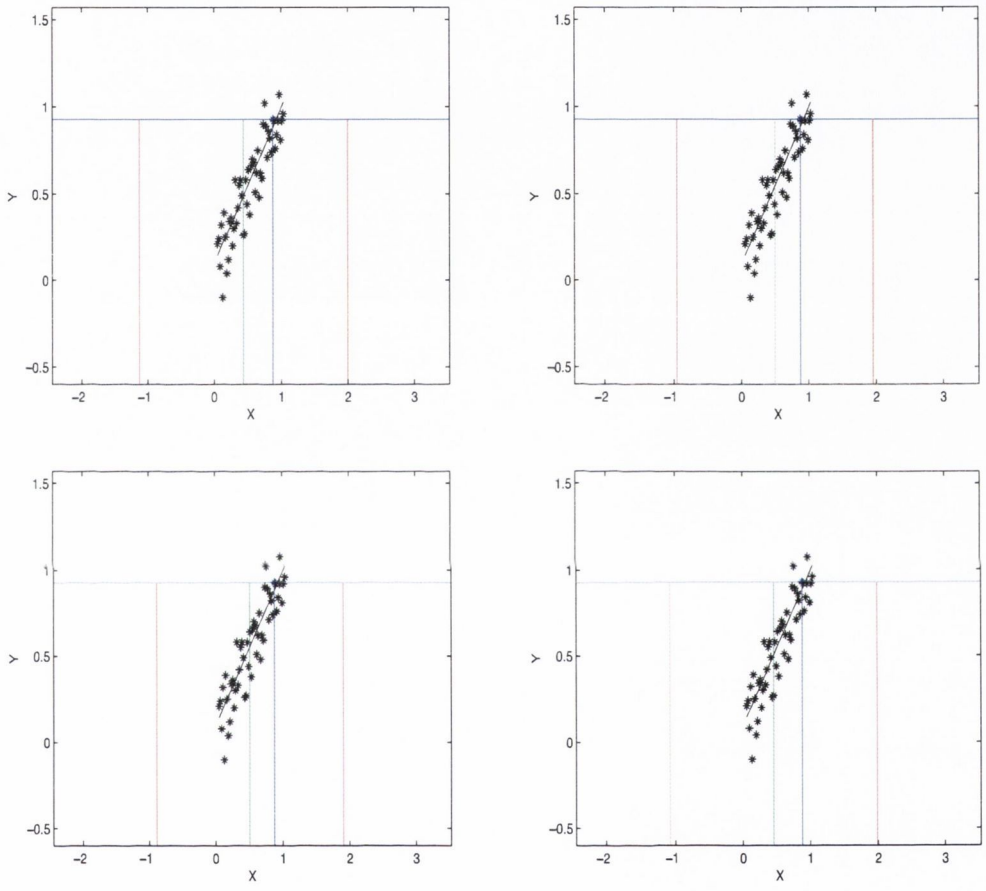

Fig. 4.5: Posterior estimates of an explanatory variable X_{new} of a simple linear regression problem with a normal prior by numerical integration (NI) (the top left), by the VB method and NI at forward and inverse stage respectively (the top right), by the VB method and RVB at the forward and inverse stage respectively (the bottom right) and by the MCMC and the Monte Carlo integration at forward and inverse stage respectively (the bottom left). The comparison of the true values (blue) and the VB-estimation (green) of X_{new} , with the lower and upper bounds (in red) of 95% HPD region, is shown. The black asterisks are training data and the blue is for a new data Y_{new}.

VB solution at Forward stage:

The VB method is applied to find a VB approximation of the joint posterior distribution $P(\beta_0, \beta_1, \beta_2, \sigma^2|\mathrm{y}, \mathrm{x})$ presented as follows:

$$
\begin{aligned}
P(\beta_0, \beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) &\approx q(\beta_0, \beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) && (4.59)\\
&= q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x}) q_{\beta_2}(\beta_2|\mathbf{y}, \mathbf{x}) q_{\sigma^2}(\sigma^2|\mathbf{y}, \mathbf{x}), && (4.60)\\
\text{where } q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) &= N(\mu^{\star}_{\beta_0}, S^{2\,\star}_{\beta_0}), && (4.61)\\
q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x}) &= N(\mu^{\star}_{\beta_1}, S^{2\,\star}_{\beta_1}), && (4.62)\\
q_{\beta_2}(\beta_2|\mathbf{y}, \mathbf{x}) &= N(\mu^{\star}_{\beta_2}, S^{2\,\star}_{\beta_2}), && (4.63)\\
q_{\sigma^2}(\sigma^2|\mathbf{y}, \mathbf{x}) &= \textit{Inverse Gamma}(a^{\star}, b^{\star}). && (4.64)
\end{aligned}
$$

The VB-parameters (defining the VB marginals) are obtained as:

$$
\begin{aligned}
S^{2\,\star}_{\beta_0} &= \left[\frac{1}{S^2_{\beta_0}} + n\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\right]^{-1},\\
\mu^{\star}_{\beta_0} &= \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\left[\sum_{i=1}^{n} \mathrm{y}_i - \mathbb{E}_q(\beta_1)\sum_{i=1}^{n} \mathrm{x}_i - \mathbb{E}_q(\beta_2)\sum_{i=1}^{n} \mathrm{x}_i^2\right] S^{2\,\star}_{\beta_0},\\
S^{2\,\star}_{\beta_1} &= \left[\frac{1}{S^2_{\beta_1}} + n\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\sum_{i=1}^{n} \mathrm{x}_i^2\right]^{-1},\\
\mu^{\star}_{\beta_1} &= \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\left[\sum_{i=1}^{n} \mathrm{y}_i\mathrm{x}_i - \mathbb{E}_q(\beta_0)\sum_{i=1}^{n} \mathrm{x}_i - \mathbb{E}_q(\beta_2)\sum_{i=1}^{n} \mathrm{x}_i^3\right] S^{2\,\star}_{\beta_1},\\
S^{2\,\star}_{\beta_2} &= \left[\frac{1}{S^2_{\beta_2}} + n\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\sum_{i=1}^{n} \mathrm{x}_i^4\right]^{-1},\\
\mu^{\star}_{\beta_2} &= \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\left[\sum_{i=1}^{n} \mathrm{y}_i\mathrm{x}_i^2 - \mathbb{E}_q(\beta_0)\sum_{i=1}^{n} \mathrm{x}_i^2 - \mathbb{E}_q(\beta_1)\sum_{i=1}^{n} \mathrm{x}_i^3\right] S^{2\,\star}_{\beta_1},\\
a^{\star} &= a + \frac{n}{2},\\
b^{\star} &= b + \frac{1}{2}g(\mathrm{y}, \mathrm{x}, M_{\beta_0}M_{\beta_1}, M_{\beta_2}),
\end{aligned}
$$

where,

$$g(\mathrm{y},\mathrm{x},M_{\beta_0}M_{\beta_1},M_{\beta_2}) = \sum_{i=1}^{n}\mathrm{y}_i^2 + n\mathbb{E}_q(\beta_0^2) + \mathbb{E}_q(\beta_1^2)\sum_{i=1}^{n}\mathrm{x}_i^2 + \mathbb{E}_q(\beta_2^2)\sum_{i=1}^{n}\mathrm{x}_i^4$$
$$-2\mathbb{E}_q(\beta_0)\left(\sum_{i=1}^{n}\mathrm{y}_i - \mathbb{E}_q(\beta_1)\sum_{i=1}^{n}\mathrm{x}_i - \mathbb{E}_q(\beta_2)\sum_{i=1}^{n}\mathrm{x}_i^2\right)$$
$$-2\mathbb{E}_q(\beta_1)\left(\sum_{i=1}^{n}\mathrm{y}_i\mathrm{x}_i - \mathbb{E}_q(\beta_2)\mathrm{x}_i^3\right) - 2\mathbb{E}_q(\beta_2)\sum_{i=1}^{n}\mathrm{y}_i\mathrm{x}_i^2. \quad (4.65)$$

Just like in the simple linear regression example, the VB marginals of regression parameters (of a quadratic regression problem) are Gaussian with the Gaussian likelihood and conjugate priors. The VB marginal of σ^2 is a conjugate inverse Gamma density. The functional form of the posterior parameters of the VB marginals are also presented in above equations.

VB approximation at the inverse stage:

The posterior distribution of $\mathrm{X}_{\mathrm{new}}$ given data $\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}$ is

$$P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) \propto \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1,\beta_2,\sigma^2)P(\mathrm{X}_{\mathrm{new}})$$
$$P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) \propto \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1,\beta_2,\sigma^2)P(\mathrm{X}_{\mathrm{new}})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\beta_2\mathrm{d}\sigma^2,$$
$$\approx \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1,\beta_2,\sigma^2)P(\mathrm{X}_{\mathrm{new}})$$
$$\times q(\beta_0,\beta_1,\beta_2,\sigma^2|\mathbf{y},\mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\beta_2\mathrm{d}\sigma^2,$$

$$P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) \propto \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1,\beta_2,\sigma^2)P(\mathrm{X}_{\mathrm{new}})q_{\beta_0}(\beta_0|\mathrm{y},\mathrm{x})$$
$$\times\, q_{\beta_1}(\beta_1|\mathbf{y},\mathbf{x})q_{\beta_2}(\beta_2|\mathbf{y},\mathbf{x})q_{\sigma^2}(\sigma^2|\mathbf{y},\mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\beta_2\mathrm{d}\sigma^2. \quad (4.66)$$

The expression in the R.H.S of Eq. 4.66 is not in a closed form. Therefore, the restricted VB method is applied to approximate the posterior distribution $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}})$ as described in Section 4.2. The (restricted) VB approximation of $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}})$ is given as:

$$P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) \approx \bar{q}(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}), \quad (4.67)$$

$$\log \bar{q}(\mathrm{X_{new}}|\mathbf{y},\mathbf{x},\mathrm{y_{new}}) \approx -\frac{1}{2}\Big[-2\mathrm{X_{new}}\left(\mathrm{y_{new}}-\mathbb{E}_q(\beta_0)\right)+2\mathrm{X_{new}^2}\left(0.5\mathbb{E}_q(\beta_1^2)-\mathrm{y_{new}}\mathbb{E}_q(\beta_2)\right.$$
$$\left.+\mathbb{E}(\beta_2)\mathbb{E}(\beta_0)\right)$$
$$+2\mathbb{E}_q(\beta_1)\mathbb{E}_q(\beta_2)\mathrm{X_{new}^3}+\mathbb{E}_q(\beta_1)\mathbb{E}_q(\beta_2^2)\mathrm{X_{new}^4}\Big]\mathbb{E}_q\left(\frac{1}{\sigma^2}\right). \quad (4.68)$$

The VB marginal $\bar{q}(\mathrm{X_{new}}|\mathbf{y},\mathbf{x},\mathrm{y_{new}})$ is an un-normalized density and cannot be recognized as a standard distribution. Its normalizing constant of can be computed numerically.

Evaluation of VB approximation :

For the accuracy check of the VB approximation at the inverse stage, the 95% HPD region of $\mathrm{X_{new}}$ is approximated as $\mathbb{E}_q(\mathrm{X_{new}}) \pm 2\sqrt{V_q(\mathrm{X_{new}})}$. The terms $\mathbb{E}_q(\mathrm{X_{new}})$ and $V_q(\mathrm{X_{new}})$ are the mean and variance of the VB approximation $\tilde{q}(\mathrm{X_{new}}|\mathbf{y},\mathbf{x},\mathrm{y_{new}})$.

Result:

A set of fifty equally-spaced values is assumed on X. Corresponding to each value of X, a value of Y is generated from a Gaussian density with variance $\sigma^2 = 0.01$ and mean $\beta_0 + \beta_1\mathrm{x} + \beta_2\mathrm{x}^2$. The true values the regression parameters are set as: $\beta_0 = 0.1$, $\beta_1 = 0.1$, $\beta_2 = 1.5$.

In Fig. 4.6, the VB approximations are compared with the true posterior distributions obtained by a numerical integration and by the MCMC. The posterior distribution of X by a numerical integration is defined on a coarse grid. A finer grid may result in a numerical approximation closer to the true posterior distribution (it is expensive to use a finer grid with many unknown parameters to be estimated). The VB-marginals of β_1 and σ^2 are quite close to the true marginal posterior densities, though the posterior VB-variances are under-estimated. The posterior distribution of $\mathrm{X_{new}}$ is a bimodal density as shown in Fig. 4.7. As the quadratic regression equation $\beta_2\mathrm{X}^2 + \beta_1\mathrm{X} + \beta_0 - \mathrm{y} = 0$, X has two roots:

$$\mathrm{X} = -\frac{\beta_1}{2\beta_2} \pm \frac{1}{2\beta_2}\sqrt{(\beta_1^2 - 4\beta_2(\beta_0 - \mathrm{y}))}, \quad (4.69)$$

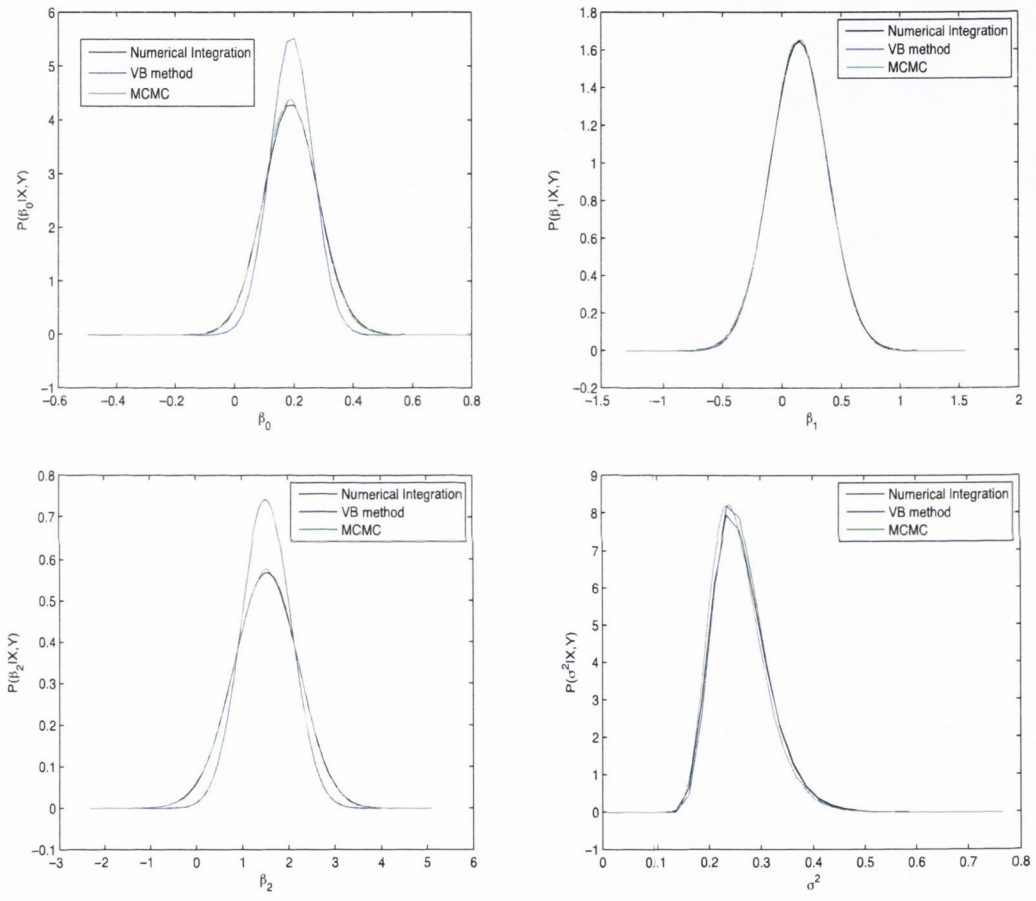

Fig. 4.6: The comparison of the true marginal posterior distributions (by a numerical integration (black) and by the MCMC (green)) and the VB approximations (blue) of the regression parameters β_0 (the top left), β_1 (the top right), β_2 (the bottom left) and variance σ^2 (the bottom right) of a quadratic regression problem is shown. The VB marginals of β_1 and σ^2 are close to the approximation by a numerical integration method. A finer grid may lead to a more accurate result by a numerical integration method.

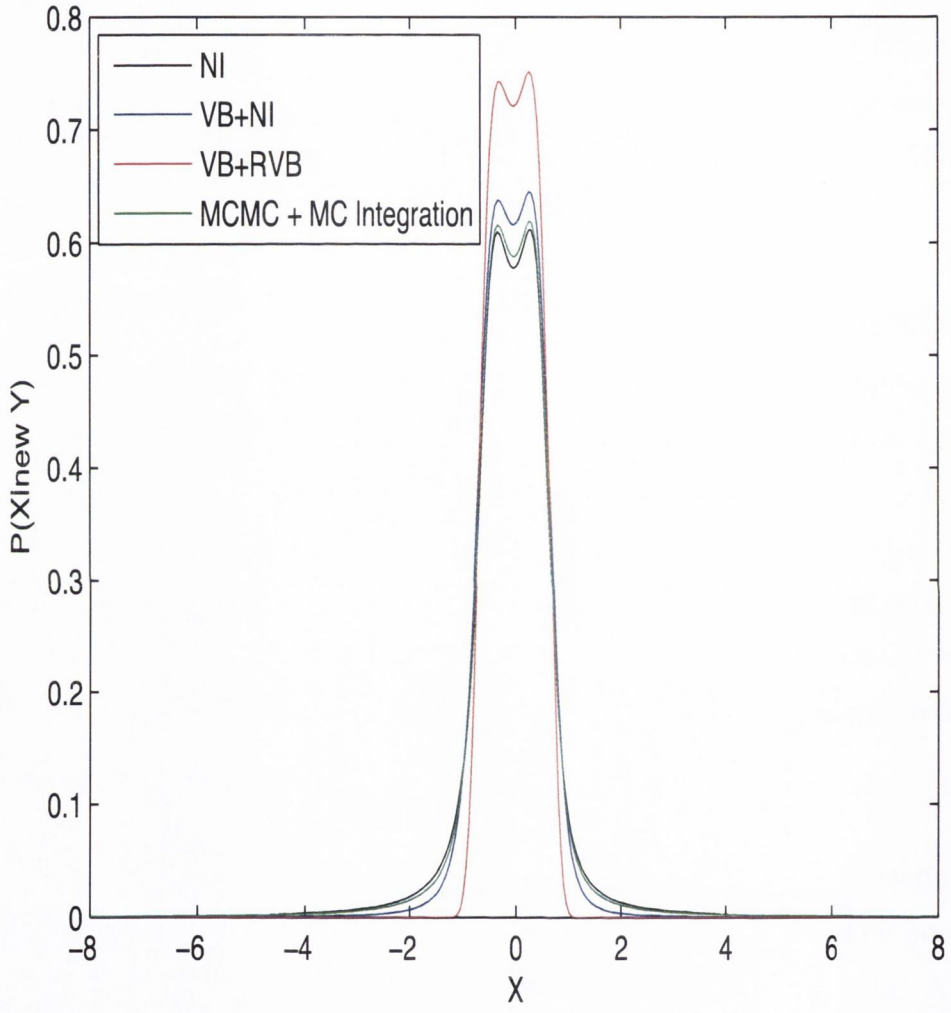

Fig. 4.7: The comparison of the true posterior distributions of X_{new} (of a quadratic regression problem with an improper prior over X_{new}) by a numerical integration (black) and by the MCMC and the Monte Carlo integration (green), the approximations by the VB method and a numerical integration (red), the regular and restricted VB method (blue) and by the MCMC and the Monte Carlo integration (green), is shown. The VB marginal by the restricted VB method is peaked due to the under-estimation of the posterior variance.

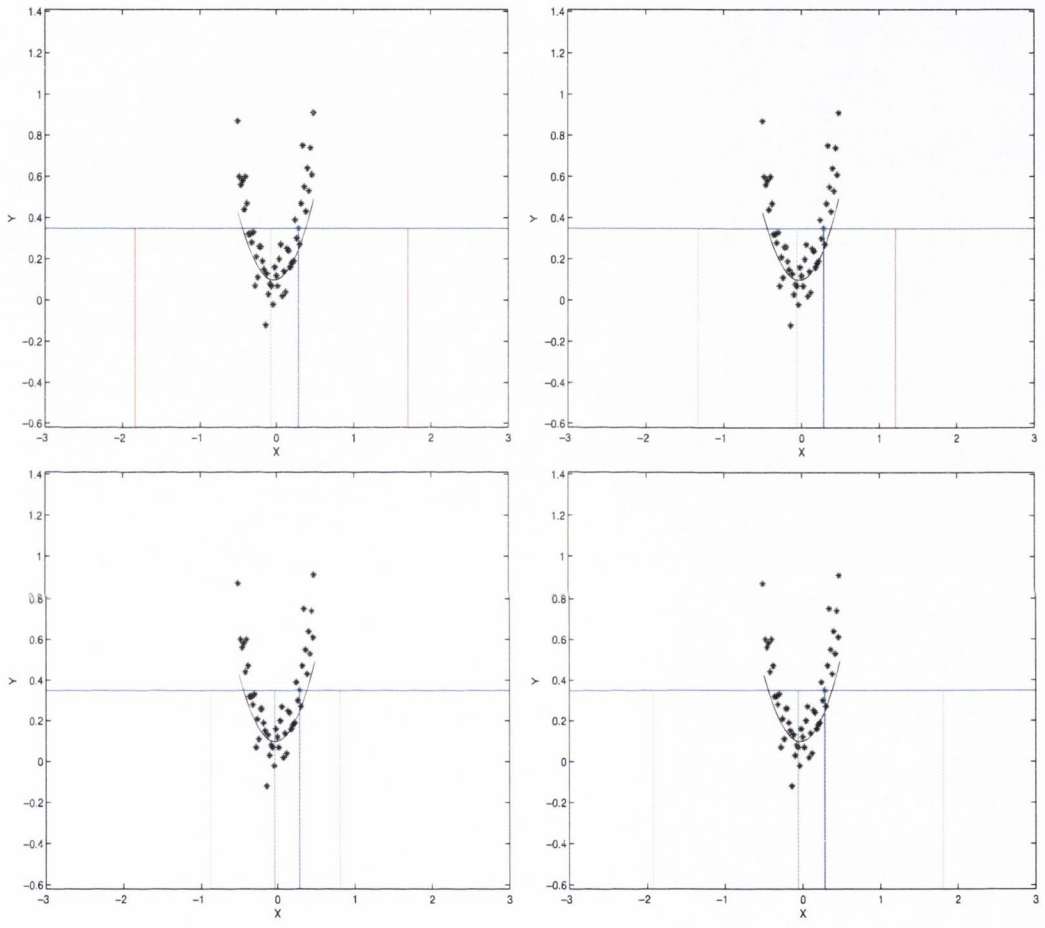

Fig. 4.8: Posterior estimates of X_{new} of a quadratic regression problem with an improper prior by numerical integration (NI) (the top left), by the VB method and NI at forward and inverse stage respectively (the top right), by the VB method and RVB at the forward and inverse stage respectively (the bottom left) and by the MCMC and the Monte Carlo integration at forward and inverse stage respectively (the bottom right). The comparison of the true values (blue) and the estimation (green) of X_{new}, with the lower and upper bounds (in red) of 95% HPD region, is shown. The black asterisks are training data and the blue is for new data Y_{new}.

β_2 (true)	$\mathbb{E}_q(\beta_2)$	$\mathbb{E}_q(\beta_2^2)$	$V(\mathrm{X_{new}})$ (VB + NI)	$V(\mathrm{X_{new}})$ (VB + RVB)
1.5	0.558	0.6193	2.8491	0.2508
2	1.1166	1.5508	0.9101	0.2048
2.5	1.3451	2.1171	0.5921	0.1853
3.0	1.5034	2.5804	0.4210	0.1540

Table 4.2: Table shows that a change in the value of the regression parameter β_2 inversely affects the posterior variance of an explanatory variable $\mathrm{X_{new}}$ of a quadratic linear regression problem. For different values of β_2, the VB variance remains very small when compared to the posterior variance by a numerical integration method.

the two possible roots of X may give rise to two modes in the posterior.

The nature of Bayesian inverse predictions for the quadratic model is quite similar to that of the simple regression problem. The effect of the underestimation of the variance at the forward stage is much reflected in the restricted VB approximation to the posterior distribution of $\mathrm{X_{new}}$. The variance of the approximation by the restricted VB method is much smaller than the true. Fig. 4.7 and 4.8 show large (true) posterior variance of $\mathrm{X_{new}}$. Though an explicit functional form of the VB-variance of $\mathrm{X_{new}}$ is not known, it can be understood from Table 4.2 that the variance of $\mathrm{X_{new}}$ is a decreasing function of VB-estimates of β_2.

A leave-one-out cross validation technique described in Section 4.2.2 is used to check the accuracy of the approximation at the inverse stage. Similar to the VB approximation for inverse simple linear problem, 100% coverage is achieved when the results are compared with the true (test data) values of $\mathrm{X_{new}}$. All the test data of $\mathrm{X_{new}}$ fall inside their 95% HPD regions, shown in Fig. 4.8. That is a 100% coverage is achieved for the true posterior distribution and the VB approximation. It is a result of the large posterior variance of $\mathrm{X_{new}}$.

It is of a great interest if the assumption of a proper prior over $\mathrm{X_{new}}$ reduces the effect of the estimate of the regression parameter on the posterior variance of $\mathrm{X_{new}}$ with an improper prior. Fig. 4.9 and 4.10 represent the results of the inverse estimation for the inverse quadratic model with the assumption of a proper (normal distribution) prior over $\mathrm{X_{new}}$. It can be understood from the figures that the assumption of the proper prior over $\mathrm{X_{new}}$ improves the true result, as the problem of the large variance is solved. The result from the MCMC and from the VB method

are very close. Hence, it can be concluded that the VB method performs well in the context of the nature of the posterior variance no matter if the data is informative or the prior is strong, in case of the weak data.

4.2.5 VB solution to Inverse Poisson Regression

In a Poisson regression model, there are only two unknown parameters to estimate; β_0 and β_1.

$$\theta = \{\beta_0, \beta_1\}.$$

The likelihood of the parameters is defined as follows:

$$P(\mathrm{y}|\mathbf{x}, \theta) = \prod_{i=1}^{n} Poisson(\mathrm{y}_i; \lambda_{\mathrm{y}_i}), \tag{4.70}$$

$$\text{where } \log \lambda_{\mathrm{y}_i} = \beta_0 + \beta_1 \mathrm{x}_i. \tag{4.71}$$

The prior distributions on the regression parameters β_0 and β_1 are assumed to be Gaussian as in the simple regression problem.

$$P(\beta_0) = N(0, S^2_{\beta_0}), \tag{4.72}$$

$$P(\beta_1) = N(0, S^2_{\beta_1}). \tag{4.73}$$

Assuming a proper prior provides a better result in case of a weak data. Therefore to avoid such a situation, as experienced in the case of the inverse simple linear and the inverse quadratic regression problems, a normal prior on $\mathrm{X}_{\mathrm{new}}$ is assumed as:

$$P(\mathrm{X}_{\mathrm{new}}) = N(\mathrm{X}_{\mathrm{new}}; \mu_{\mathrm{X}}, S^2_{\mathrm{X}}). \tag{4.74}$$

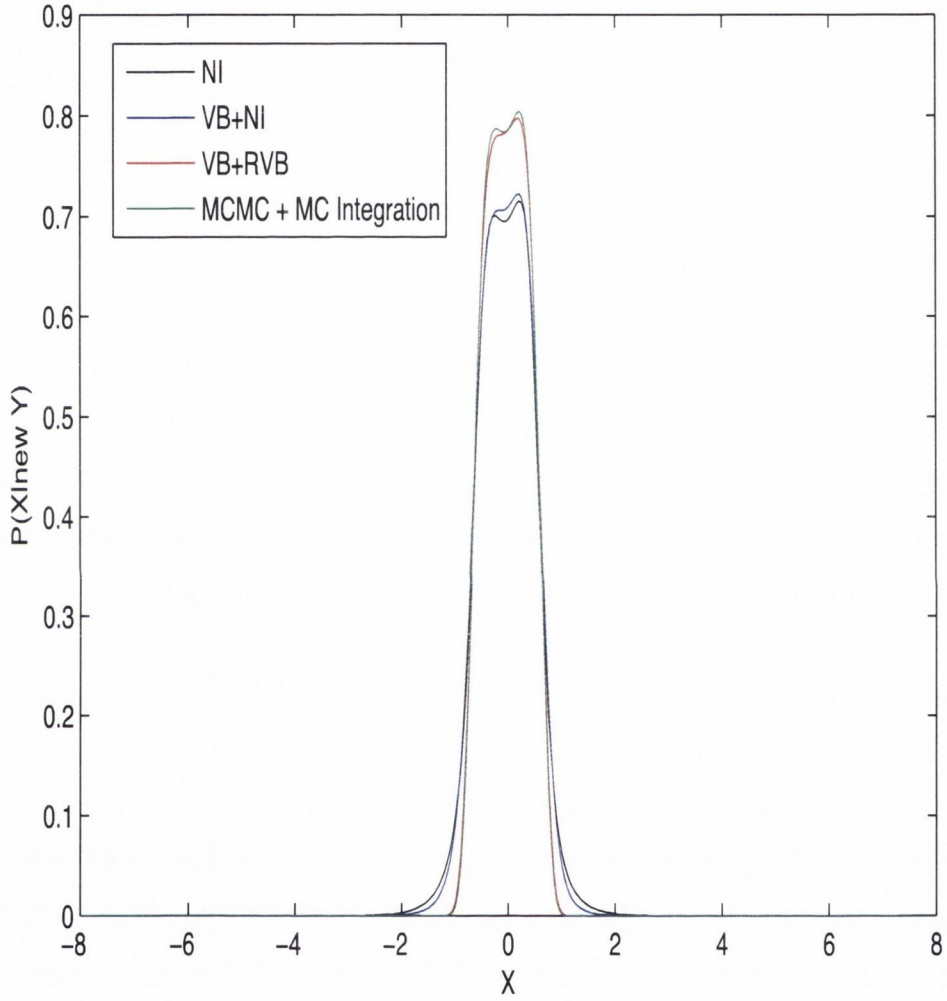

Fig. 4.9: The comparison of the true posterior distributions of X_{new} (of a quadratic regression problem with a normal prior over X_{new}) by a numerical integration (black) and by the MCMC and the Monte Carlo integration (green), the VB approximations by the VB method and a numerical integration (red), the regular and restricted VB method (blue), is shown. The VB marginal by the restricted VB method is peaked due to the under-estimation of the posterior variance.

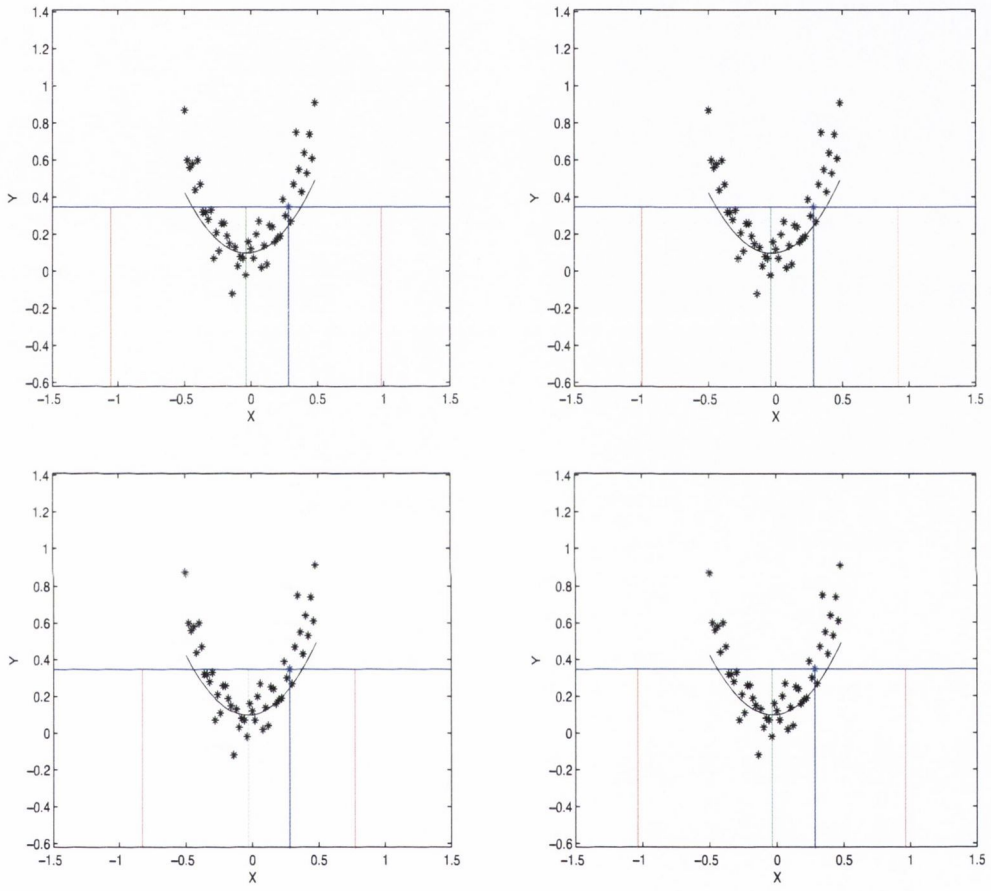

Fig. 4.10: Posterior estimates of X_{new} of a quadratic regression problem with a normal prior by numerical integration (NI) (the top left), by the VB method and NI at forward and inverse stage respectively (the top right), by the VB method and RVB at the forward and inverse stage respectively (the bottom left) and by the MCMC and the Monte Carlo integration at forward and inverse stage respectively (the bottom right). The comparison of the true values (blue) and the estimation (green) of X_{new}, with the lower and upper bounds (in red) of 95% HPD region, is shown. The black asterisks are training data and the blue is for new data Y_{new}.

VB solution at Forward stage:

A VB approximation to the joint posterior distribution $P(\beta_0, \beta_1|\mathrm{y}, \mathrm{x})$ is:

$$
\begin{aligned}
P(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) &\approx q(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}), && (4.75)\\
q(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) &= q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x}), && (4.76)\\
\text{where } \log q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) &\approx \left[-\frac{\beta_0^2}{2S_{\beta_0}^2} - e^{\beta_0} \sum_{i=1}^{n} \mathbb{E}_q \left(e^{\beta_1 \mathrm{x}_i} \right) + \beta_0 \sum_{i=1}^{n} \mathrm{y}_i \right], && (4.77)\\
\log q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x}) &\approx \left[-\frac{\beta_1^2}{2S_{\beta_1}^2} - \mathbb{E}_q \left(e^{\beta_0} \right) \sum_{i=1}^{n} e^{\beta_1 \mathrm{x}_i} + \beta_1 \sum_{i=1}^{n} \mathrm{y}_i \mathrm{x}_i \right]. && (4.78)
\end{aligned}
$$

The VB marginals $q_{\beta_1}(\beta_1|\mathrm{y}, \mathrm{x})$ and $q_{\beta_0}(\beta_0|\mathrm{y}, \mathrm{x})$ are not standard distributions. The proportionality constants of the approximations and the required VB-moments can be computed numerically.

VB approximation at the inverse stage:

The posterior distribution of X_{new} is

$$
\begin{aligned}
P(\mathrm{X}_{\text{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y}_{\text{new}}) &\propto \int P(\mathrm{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \beta_0, \beta_1) P(\mathrm{X}_{\text{new}}) \\
&\qquad \times P(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) \mathrm{d}\beta_0 \mathrm{d}\beta_1, && (4.79)\\
&\approx \int P(\mathrm{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \beta_0, \beta_1) P(\mathrm{X}_{\text{new}}) \\
&\qquad \times q(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) \mathrm{d}\beta_0 \mathrm{d}\beta_1, && (4.80)\\
&= \int P(\mathrm{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \beta_0, \beta_1) P(\mathrm{X}_{\text{new}}) q_{\beta_0}(\beta_0|\mathbf{y}, \mathbf{x}) \\
&\qquad \times q_{\beta_1}(\beta_1|\mathbf{y}, \mathbf{x}) \mathrm{d}\beta_0 \mathrm{d}\beta_1. && (4.81)
\end{aligned}
$$

A closed form solution to the integrals is not available. However it is only a 2-dimensional integral, therefore, an analytical solution to the posterior distribution $P(\mathrm{X}_{\text{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y}_{\text{new}})$ can be found by a 2-dimensional numerical integration. The

restricted VB method is used to find a tractable approximation to $P(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$:

$$P(\mathrm{X_{new}}|\mathrm{y}, \mathbf{x}, \mathbf{y}_{\mathrm{new}}) \approx \bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}), \tag{4.82}$$

$$\log \bar{q}(\mathrm{X_{new}}|\mathrm{y}, \mathbf{x}, \mathbf{y}_{\mathrm{new}}) = \mathbb{E}_{q(\beta_0|\mathbf{y},\mathbf{x})q(\beta_1|\mathbf{y},\mathbf{x})} \log[P(\mathrm{y_{new}}|\beta_0, \beta_1, \mathrm{X_{new}}) \times P(\mathrm{X_{new}})], \tag{4.83}$$

$$\log \bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}) \approx \left[-\mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(e^{\beta_1 \mathrm{X_{new}}}) + \mathbb{E}_q(\beta_1)\mathrm{y_{new}}\mathrm{X_{new}}\right]. \tag{4.84}$$

The restricted VB approximation $\bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$ as expressed in the above equation is not tractable because the term $e^{\beta_1 \mathrm{X_{new}}}$ in the expression for $\bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$ cannot be factorized over β_1 and $\mathrm{X_{new}}$ separately. A Gaussian approximation is considered to find a tractable solution to $\bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$. Considering the Taylor's expansion of $\bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$ as a function of $\mathrm{X_{new}}$ around its posterior mode (denoted as X_a), a Gaussian approximation is found as:

$$\bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}) = N(\mu_X^\star, S_X^{2\star}), \tag{4.85}$$

$$S_X^{2\star} = [\mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(\beta_1^2 e^{\beta_1 \mathrm{X}_a})]^{-1}, \tag{4.86}$$

$$\mu_X^\star = [\mathbb{E}_q(\beta_1)\mathrm{y_{new}} + \mathbb{E}_q(e^{\beta_0})[\mathrm{X}_a\mathbb{E}_q(\beta_1^2 e^{\beta_1 \mathrm{X}_a}) - \mathbb{E}_q(\beta_1 e^{\beta_1 \mathrm{X}_a})]]S_X^{2\star} \tag{4.87}$$

The VB-moments of β_1 in Eq. 4.86 and 4.87 can be computed numerically given the posterior mode X_a. To compute a Gaussian approximation to $\log \bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$, the posterior mode X_a should be found by an optimization method. A quick and easy way to find a Gaussian approximation with an algorithm by Rue & Held (2005). The algorithm avoids use of any optimization method to find a posterior mode. It uses a simple property of a Gaussian distribution i.e. the mean and the mode of the distribution are equal. It assumes an initial value of posterior mode, finds a Gaussian approximation around that initial value and computes the mean of the approximation as given by some expression (e.g. given in Eq. 4.87). It sets the mode of the posterior distribution to the mean of the approximation and proceeds in the same manner until convergence.

Comparison of the VB approximation with the result from other methods:

The Gaussian approximation with the VB method for a non-conjugate family of distributions may provide a tractable VB solution. But the quality of the approximation should be examined to see whether it degrades the VB approximation or improves it. Two methods are considered to check the accuracy of the VB approximation: the Monte Carlo method and the variational tangent approach. The Monte Carlo approach for the inverse estimation is already discussed in the chapter.

The variational tangent approach to the problem:

For the comparison of the Gaussian approximation with VB method, a variational tangent approach may be employed. As defined earlier in Chapter 3, the variational tangent approach finds an approximation of a posterior distribution (or just of a function) by considering a tangent (a lower bound) of a non-linear or non-quadratic log-likelihood (or joint log-likelihood). The aim of considering a tangent transformation to the problem is to make the problem conjugate-exponential so that a tractable variational approximation could be found.

The variational tangent approximation of the posterior distribution of $\mathrm{X}_{\mathrm{new}}$ is defined as below:

$$\begin{aligned}
\log P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) &\approx \log P(\mathrm{X}_{\mathrm{new}},\mathrm{y}_{\mathrm{new}}|\mathbf{x},\mathbf{y}),\\
&\approx \mathbb{E}_{q(\beta_0|\mathbf{y},\mathbf{x})q(\beta_1|\mathbf{y},\mathbf{x})}\Big[\log P(\mathrm{X}_{\mathrm{new}}) + \log P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1)\\
&\qquad + \log q(\beta_0,\beta_1|\mathbf{x},\mathbf{y})\Big],\\
&\approx \log P(\mathrm{X}_{\mathrm{new}}) + \mathbb{E}_{q(\beta_0|\mathbf{y},\mathbf{x})q(\beta_1|\mathbf{y},\mathbf{x})}\Big[\log P(\mathrm{y}_{\mathrm{new}}|\beta_0,\beta_1,\mathrm{X}_{\mathrm{new}})\Big],\\
&\approx -\frac{1}{2S_{\mathrm{X}}^2}\left(\mathrm{X}_{\mathrm{new}}-\mu_{\mathrm{X}}\right)^2 - \mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(e^{\beta_1\mathrm{X}_{\mathrm{new}}}) + \mathbb{E}_q(\beta_1)\mathrm{y}_{\mathrm{new}}\mathrm{X}_{\mathrm{new}}\\
&\qquad + \mathbb{E}_q(\beta_0)\mathrm{y}_{\mathrm{new}}. \qquad (4.88)
\end{aligned}$$

The only term $e^{\beta_1\mathrm{X}_{\mathrm{new}}}$ in the R.H.S of the above expression is not quadratic or linear in $\mathrm{X}_{\mathrm{new}}$. Therefore, a tangent transformation of the joint log-likelihood $\log P(\mathrm{X}_{\mathrm{new}},\mathrm{y}_{\mathrm{new}}|\mathbf{x},\mathbf{y})$ is equivalent to considering a lower bound (linear or quadratic)

on $e^{\beta_1 \mathrm{X}_{\mathrm{new}}}$.

The second order Taylor's expansion of $e^{\beta_1 \mathrm{X}_{\mathrm{new}}}$ around $\beta_1 \mathrm{X}_{\mathrm{new}} = \xi$ is considered to find a quadratic lower bound (tangent) of the function:

1. if $\hat{\beta}_1 \mathrm{X}_{\mathrm{new}} > 0$ ($\hat{\beta}_1$ is the posterior modal value of β_1),

$$e^{\beta_1 \mathrm{X}_{\mathrm{new}}} \approx e^{\xi} + e^{\xi}(\beta_1 \mathrm{X}_{\mathrm{new}} - \xi) + 0.5e^{\xi}(\beta_1 \mathrm{X}_{\mathrm{new}} - \xi)^2 \tag{4.89}$$

2. otherwise, if $\hat{\beta}_1 \mathrm{X}_{\mathrm{new}} < 0$,

$$e^{\beta_1 \mathrm{X}_{\mathrm{new}}} \approx e^{\xi} - e^{\xi}(\beta_1 \mathrm{X}_{\mathrm{new}} - \xi) + 0.5e^{\xi}(\beta_1 \mathrm{X}_{\mathrm{new}} - \xi)^2 \tag{4.90}$$

Therefore,

$$\log P(\mathrm{X}_{\mathrm{new}}, \mathrm{y}_{\mathrm{new}} | \mathbf{x}, \mathbf{y}) \geq \log P(\mathrm{X}_{\mathrm{new}}, \mathrm{y}_{\mathrm{new}} | \mathbf{x}, \mathbf{y}, \xi). \tag{4.91}$$

Considering the lower bound in the VB approximation the variational tangent approximation we get is as follows:

$$P(\mathrm{X}_{\mathrm{new}} | \mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) \approx q_T(\mathrm{X}_{\mathrm{new}} | \mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}), \tag{4.92}$$

$$q_T(\mathrm{X}_{\mathrm{new}} | \mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) = N(\mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\,\star}), \tag{4.93}$$

where $q_T(\mathrm{X}_{\mathrm{new}} | \mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y})$ is the variational tangent approximation to the posterior distribution of $\mathrm{X}_{\mathrm{new}}$.

The mean $\mu_{\mathrm{X}}^{\star}$ and the variance $S_{\mathrm{X}}^{2\,\star}$ parameter of the approximation is given as follows:

1. if $\hat{\beta}_1 \mathrm{X}_{\mathrm{new}} > 0$ then,

$$S_{\mathrm{X}}^{2\,\star} = \left[\mathbb{E}_q(e^{\beta_0}) \mathbb{E}_q(\beta_1^2) e^{\xi} + \frac{1}{S_{\mathrm{X}}^2} \right]^{-1}, \tag{4.94}$$

$$\mu_{\mathrm{X}}^{\star} = \left[\mathbb{E}_q(e^{\beta_0}) \left(e^{\xi}(\xi - 1) \mathbb{E}_q(\beta_1) \right) + \mathrm{y}_{\mathrm{new}} \mathbb{E}_q(\beta_1) + \frac{\mu_{\mathrm{X}}}{S_{\mathrm{X}}^2} \right] S_{\mathrm{X}}^{2\,\star}. \tag{4.95}$$

2. if $\hat{\beta}_1 \mathrm{X}_{\mathrm{new}} < 0$ then,

$$S^{2\star}_{\mathrm{X}} = \left[\mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(\beta_1^2)e^{\xi} + \frac{1}{S^2_{\mathrm{X}}}\right]^{-1}, \tag{4.96}$$

$$\mu^{\star}_{\mathrm{X}} = \left[\mathbb{E}_q(e^{\beta_0})\left(e^{\xi}(\xi+1)\mathbb{E}_q(\beta_1)\right) + \mathrm{y}_{\mathrm{new}}\mathbb{E}_q(\beta_1) + \frac{\mu_{\mathrm{X}}}{S^2_{\mathrm{X}}}\right] S^{2\star}_{\mathrm{X}}. \tag{4.97}$$

The variable ξ is still to be determined. Consider the expectation of the lower bound $\log P(\mathrm{X}_{\mathrm{new}}, \mathrm{y}_{\mathrm{new}}|\mathbf{x}, \mathbf{y}, \xi)$ with respect to $\mathrm{X}_{\mathrm{new}}$ and maximize with respect to ξ.

$$Q(\xi) = \mathbb{E}_{q(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}},\mathbf{x},\mathbf{y})}[P(\mathrm{X}_{\mathrm{new}}, \mathrm{y}_{\mathrm{new}}|\mathbf{x}, \mathbf{y}, \xi)], \tag{4.98}$$

$$\xi = \arg\max_{\xi} Q(\xi). \tag{4.99}$$

The value of ξ as a maximum of $Q(\xi)$ is obtained as follows:

1. if $\hat{\beta}_1 \mathrm{X}_{\mathrm{new}} > 0$, $\xi = \mathbb{E}_q(\beta_1)\mathrm{X}_{\mathrm{new}}$,

2. otherwise, if $\hat{\beta}_1 \mathrm{X}_{\mathrm{new}} < 0$, $\xi = \mathbb{E}_q(\beta_1)\mathrm{X}_{\mathrm{new}}$-2.

As the hyper-parameters of the variational tangent approach $\mu^{\star}_{\mathrm{X}}$, $S^{2\star}_{\mathrm{X}_{new}}$ and the variational parameter ξ are mutually dependent, they can be computed iteratively.

Evaluation of VB approximation:
Another accuracy check of the VB approximation is found in terms of its 95% HPD region. The 95% HPD region of $\mathrm{X}_{\mathrm{new}}$ is obtained as $\mu^{\star}_X \pm 2\sqrt{S^{2\star}_X}$.

Result:

Corresponding to fifty equally spaced values of X, fifty observations on Y are generated from a Poisson density given true values of regression parameters as $\beta_0 = 0.15$ and $\beta_1 = 1.5$. Results from the VB approximation for an inverse Poisson regression problem are shown in Fig. 4.11, 4.12, 4.13. The VB approximation at the forward stage of the problem is compared with the (true) results from the MCMC method and a numerical integration method in Fig. 4.11. It can be understood from the figure that the VB-variance of the regression parameters β_0 and β_1 are underestimated. The parameters β_0 and β_1 are strongly correlated in the Poisson

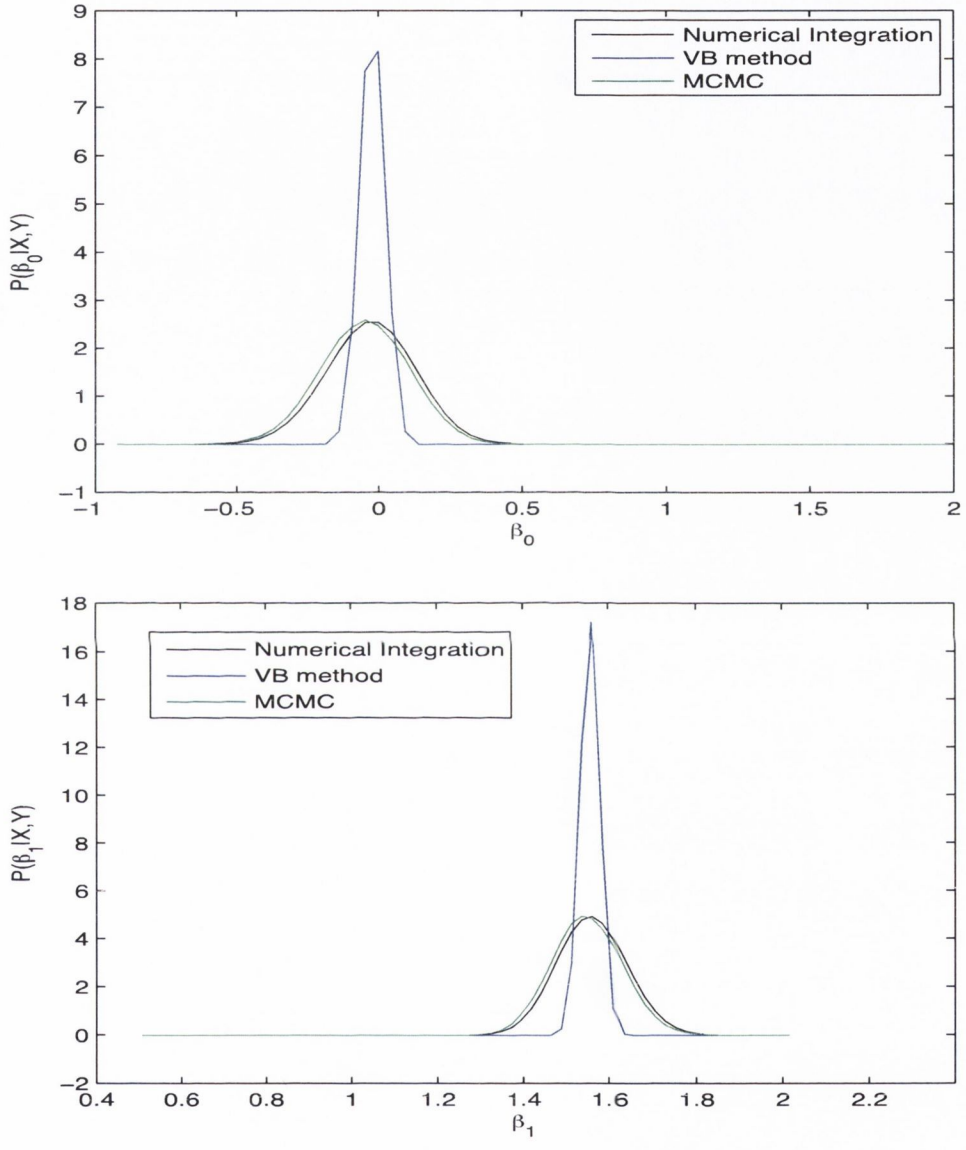

Fig. 4.11: The comparison of the true posterior distribution of the regression parameters β_0 (top), β_1 (bottom) of a Poisson regression problem by a numerical integration (black) and MCMC (green), and their VB approximation (blue) of the marginal posterior distributions, is shown. The VB marginal of β_0 and β_1 are peaked. The regression parameters β_0, β_1 are too correlated to allow for the posterior independence.

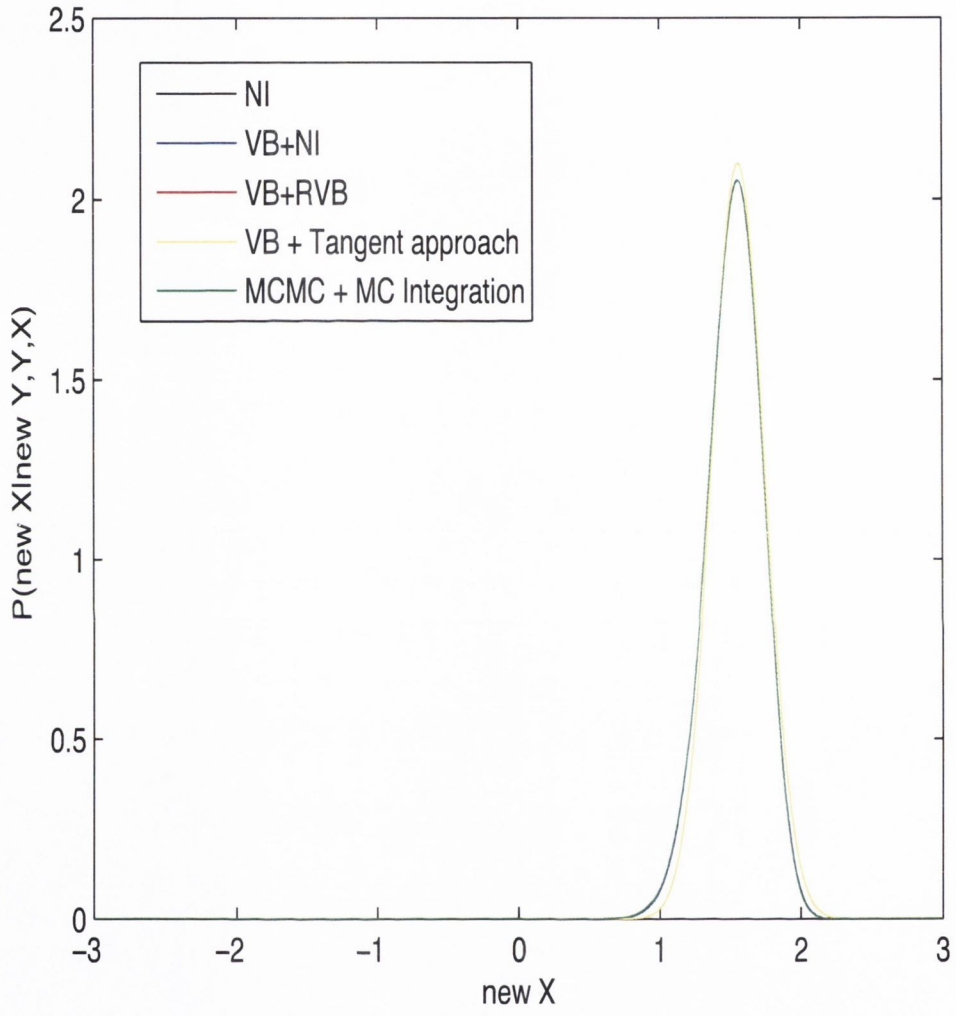

Fig. 4.12: The comparison of the true posterior distributions of X_{new} (with a normal prior) of a Poisson regression problem given a big count on Y_{new} (=28) by a numerical integration (black) and by the MCMC method at the forward stage and Monte Carlo integration at the inverse stage (green), the VB approximations at the forward stage and a numerical integration at the inverse stage (blue), the restricted VB approximation at the inverse stage (red), the variational tangent approximation at the inverse stage (yellow) , is shown. The VB marginal by the restricted VB method matches the result by a numerical integration method.

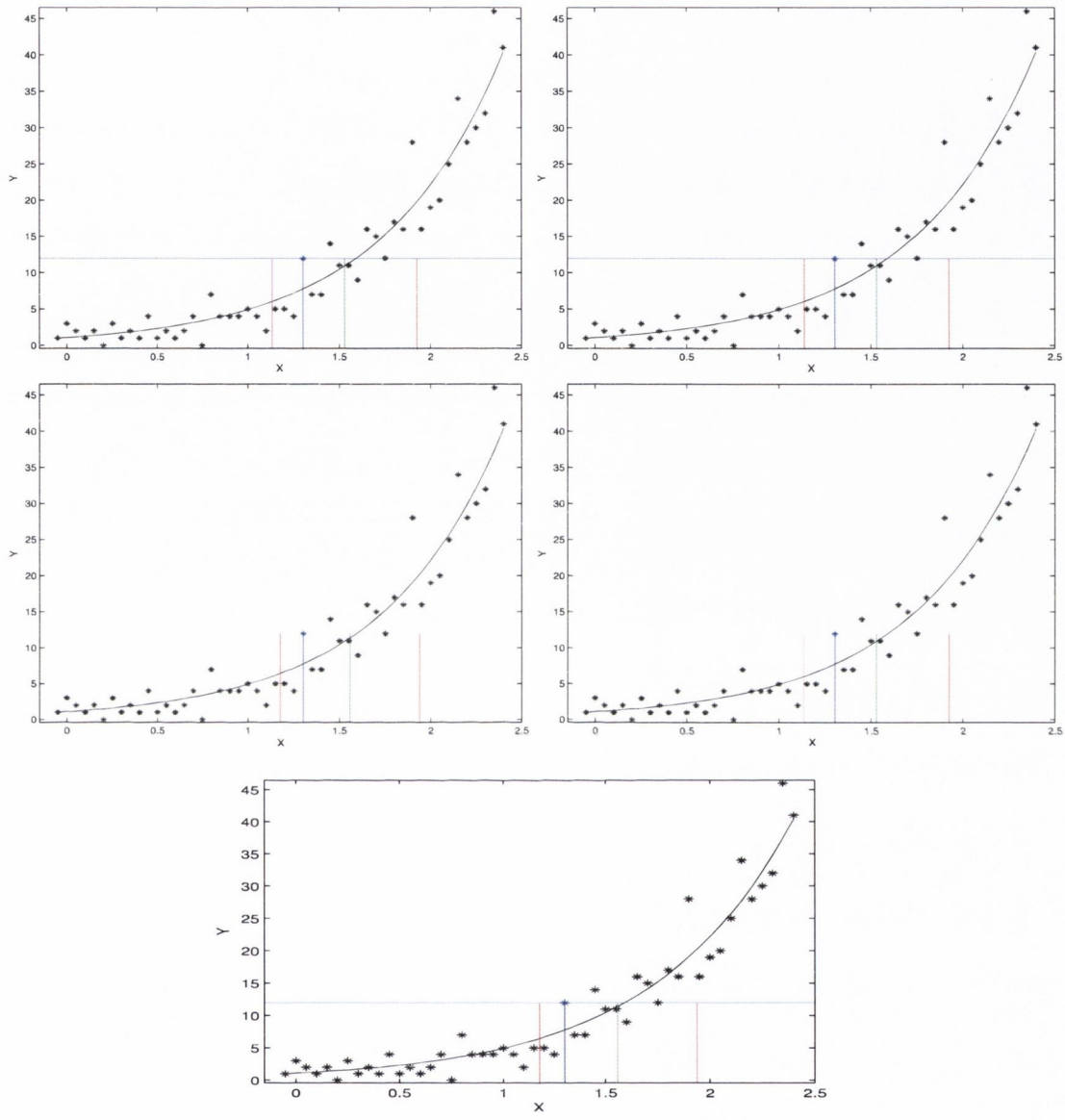

Fig. 4.13: Posterior estimates of an explanatory variable X_{new} of a Poisson regression problem with a normal prior by numerical integration (NI) (the top left), by the VB method and NI at forward and inverse stage respectively (the top right), by the VB method and RVB at the forward and inverse stage respectively (the middle left), by the MCMC and the Monte Carlo integration at forward and inverse stage respectively (the middle right) and by the VB and tangent approach at forward and inverse stage respectively (the bottom). The comparison of the true value (blue) and the estimates (green) of X_{new}, with the lower and upper bounds (in red) of 95% HPD region, is shown. The black asterisks are training data and the blue is for a new data on the response variable Y. The posterior variance of X_{new} are very small when compared to that of the simple and quadratic regression.

likelihood and so to assume them independent in the VB approximation leads to this under-estimation.

In Fig. 4.12 and 4.13, the VB approximation at the inverse stage is compared with the results obtained from the Monte Carlo integration, a numerical integration and the variational tangent approach. Assuming the result from the MCMC true, the VB methods provides better result than the variational tangent approach as the VB approximation is matching the true result quite well . At the inverse stage, the effect of the posterior independence assumption of the VB method is not reflected. It is experienced with several such experiments with different sample sizes that the VB approximation of X_{new} is close to its true posterior distribution for sufficiently large (training) data used at the forward stage to estimate the regression parameters. In Fig. 4.12, the posterior distribution of X_{new} is conditioned on a fairly big count on Y_{new} (=28). Fig. 4.13 displays the 95% HPD regions of X_{new}. Clearly, the problem of infinite variance is not noticed in the case of the inverse Poisson regression problem.

As a result of accuracy check via leave-one-out cross validation technique, 94% counts are inside their 95% HPD regions for VB and restricted VB (RVB) approximation. Whereas, the accuracy is 96% for both the true predictions (by a numerical integration) and the approximated (by VB at the forward and a numerical integration at the inverse stage).

Fig. 4.14 and 4.15 represent the posterior distribution of X_{new} given a small count on Y_{new} (=1) with a normal prior and an improper prior assumed over X_{new} respectively. It can be concluded that a small number of training data set leads to a less accurate VB approximation of X_{new}, but it provides a better VB approximation than the variational Tangent approximation for a small count.

Comparing the Fig 4.14 and 4.15, it can be concluded that the assumption of a proper prior distribution over X_{new} for the weak (response) data leads to the small posterior variance hence improves the accuracy of the result. This means, even if some accuracy is lost in VB approximation (in terms of the variance) at the forward stage, an accurate approximation of inverse estimation can be found for a sufficiently large training data set or for weak data with an informative prior distribution.

It should be noted that the considered Poisson regression problem is only a two

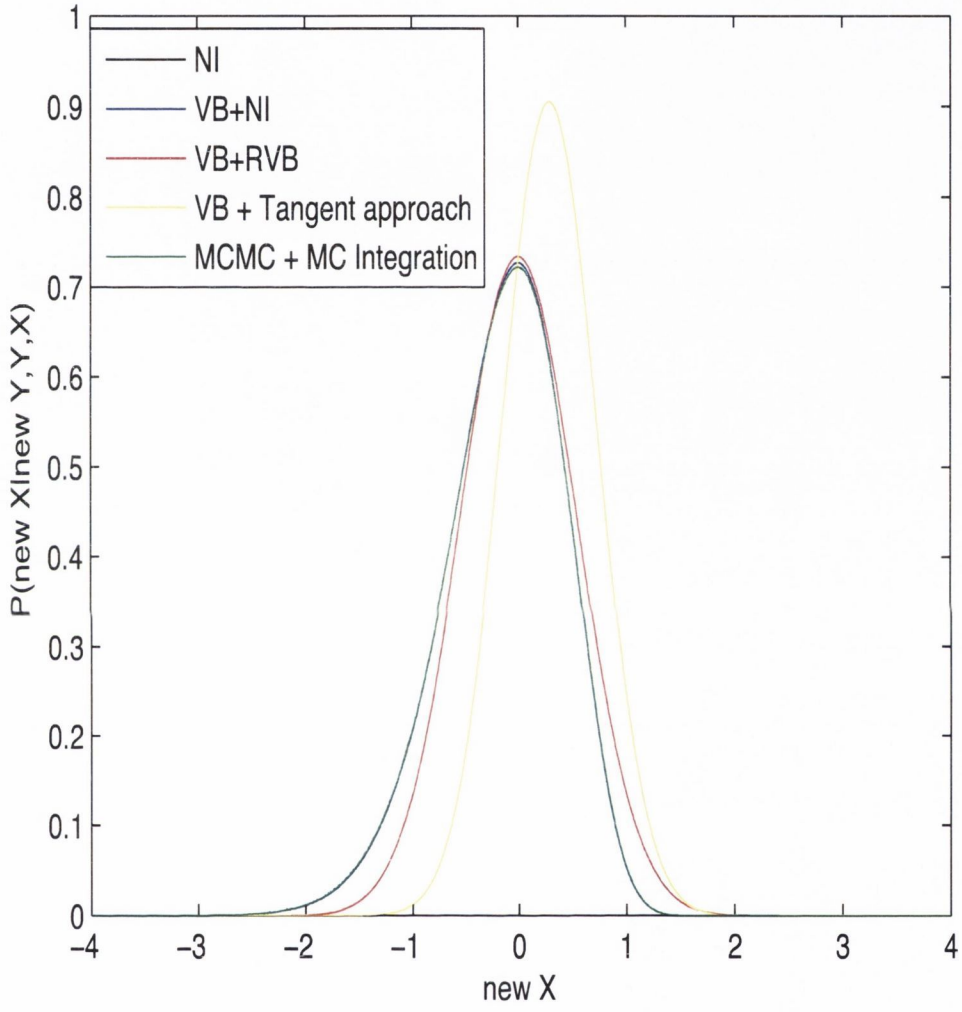

Fig. 4.14: The comparison of the true posterior distributions of X_{new} (with a normal prior) of a Poisson regression problem given a small count on Y_{new}(=1) by a numerical integration (black) and by the MCMC method at the forward stage and Monte Carlo integration at the inverse stage (green), the VB approximations at the forward stage and a numerical integration at the inverse stage (blue), the restricted VB approximation at the inverse stage (red), the variational tangent approximation at the inverse stage (yellow) , is shown. The VB marginal by the restricted VB method matches the result by a numerical integration method.

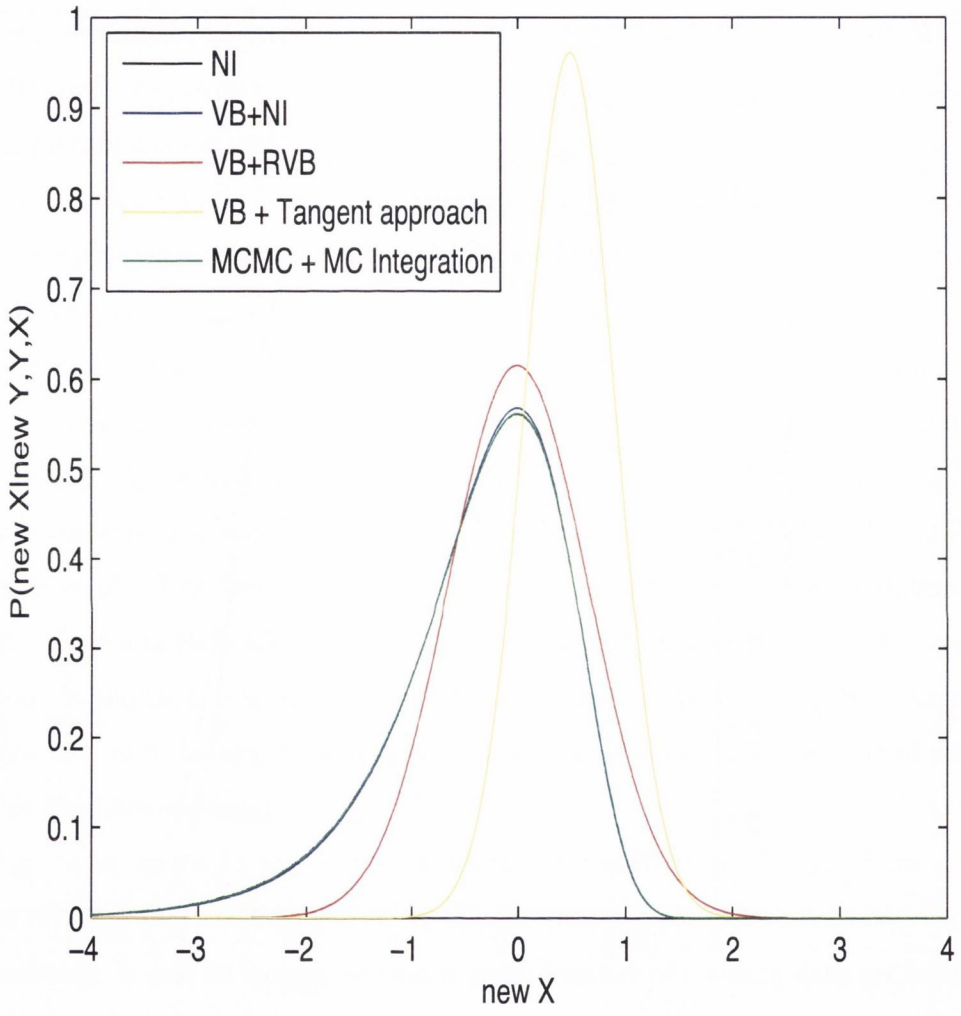

Fig. 4.15: The comparison of the true posterior distributions of X_{new} (with an improper prior) of a Poisson regression problem given a small count on $y_{new}(=1)$ by a numerical integration (black) and by the MCMC method at the forward stage and Monte Carlo integration at the inverse stage (green), the VB approximations at the forward stage and a numerical integration at the inverse stage (blue), the restricted VB approximation at the inverse stage (red), the variational tangent approximation at the inverse stage (yellow) , is shown. The VB marginal by the restricted VB method matches the result by a numerical integration method.

dimensional problem. Hence, the independence assumption of the VB method (used at the forward stage) does not affect the inverse inference poorly. It is also an interesting fact that for a big value of new (test) data $\mathrm{Y}_{\mathrm{new}}$ the posterior distribution of $\mathrm{X}_{\mathrm{new}}$ is nearly symmetric. Therefore, a Gaussian approximation of the VB approximation of $\mathrm{X}_{\mathrm{new}}$ is close to the true posterior distribution for large values of the test data. Thus a Gaussian approximation to the restrictive VB approximation helps to improve the accuracy of the approximation.

It is noticed that for a zero count on $\mathrm{Y}_{\mathrm{new}}$, the Bayesian prediction of $\mathrm{X}_{\mathrm{new}}$ by the restricted VB method suggests a very small negative estimate of $\mathrm{X}_{\mathrm{new}}$ with a very large variance for the Poisson with an improper prior. If compared with the result from the classical method of inference, a very small negative value of $\mathrm{X}_{\mathrm{new}}$ should give rise to a zero count of $\mathrm{Y}_{\mathrm{new}}$.

An iterative optimization method is applied to compute a Gaussian approximation (Rue & Held, 2005) of the intractable restricted VB approximation around the posterior mode of $\mathrm{X}_{\mathrm{new}}$. The iterative algorithm does not converge to a unique result for zero count on $\mathrm{Y}_{\mathrm{new}}$ and gives rise to many very small negative posterior estimates (mean) of $\mathrm{X}_{\mathrm{new}}$ with a very large posterior variances. To avoid this situation, it can be suggested to use multiple counts on Y to predict a single X.

4.2.6 Inverse Mixture of Poisson regression problem

The mixture of Poisson model considered in the chapter is of the following form:

$$P(\mathbf{Y}|\beta_0, \beta_1, \mu, , \mathbf{Z}, \mathbf{x}) = \prod_{i=1}^{n} \left[\frac{e^{-\lambda_i}\lambda_i^{\mathrm{Y}_i}}{\mathrm{Y}_i!}\right]^{\mathrm{Z}_i} \left[\frac{e^{-\mu}\mu^{\mathrm{Y}_i}}{\mathrm{Y}_i!}\right]^{(1-\mathrm{Z}_i)}, \tag{4.100}$$

$$\lambda_i = \exp(\beta_0 + \beta_1 \mathrm{x}_i), \tag{4.101}$$

$$P(\mathbf{Z}) = \prod_{i=1}^{N} \pi^{\mathrm{Z}_i}\pi^{(1-\mathrm{Z}_i)}. \tag{4.102}$$

where the terms β_0 and β_1 are the unknown regression parameters, μ is the unknown rate parameter of the one of the Poisson distribution of the mixture, π is the unknown parameter that denotes the probability that a count on $\mathbf{Y}$ follows the Poisson distribution with the rate parameter λ. The parameter λ relates the response

variable $\mathbf{Y}$ to the independent variable $\mathbf{X}$. The term $\mathbf{Z}$ is the set of the unknown indicator variables Z_i's which take the value 1 if Y_i is regressed on X_i or value 0 if it follows the distribution with the rate parameter μ which is independent of the explanatory variable X_i.

At the forward stage of the inference, the mixture of the Poisson model has four parameters β_0, β_1, μ and π, and a set of auxiliary variables $\mathbf{Z}$ to be estimated.

The prior distributions over the regression parameters β_0 and β_1 assumed as in the case of the inverse Poisson regression:

$$P(\beta_0) = N(0, S^2_{\beta_0}), \tag{4.103}$$

$$P(\beta_1) = N(0, S^2_{\beta_1}). \tag{4.104}$$

The Prior distribution over the parameters μ and π are

$$P(\mu) = log\ Normal(m, S^2_{\mu});\ \mu > 0, \tag{4.105}$$

$$P(\pi) = Beta(a, b). \tag{4.106}$$

The Prior distribution over X_{new} corresponding to a new count on Y_{new} is assumed to be a normal prior as given below:

$$P(X_{new}) = N(X_{new};\ \mu_X, S^2_X). \tag{4.107}$$

It is shown for the inverse simple, quadratic and Poisson regression modes that a proper prior may lead to a better result for inverse prediction if the data is weak. Therefore, the inverse inference over X_{new} will be shown using only a proper prior.

VB solution at Forward stage:

A VB approximation to the joint posterior distribution $P(\beta_0, \beta_1, \mu, \pi, \mathbf{Z}|\mathbf{y}, \mathbf{x})$ is as

follows:

$$P(\beta_0, \beta_1, \mu, \pi, \mathbf{Z}|\mathbf{y}, \mathbf{x}) = q(\beta_0, \beta_1, \mu, \pi, \mathbf{Z}|\mathbf{y}, \mathbf{x}), \tag{4.108}$$

$$q(\beta_0, \beta_1, \mu, \pi, \mathbf{Z}|\mathbf{y}, \mathbf{x}) = q(\beta_0|\mathbf{y}, \mathbf{x})q(\beta_1|\mathbf{y}, \mathbf{x})q(\mu|\mathbf{y}, \mathbf{x})q(\pi|\mathbf{y}, \mathbf{x}) \times \left[\prod_{i=1}^{n} q(\mathrm{Z}_i|\mathbf{y}, \mathbf{x})\right], \tag{4.109}$$

Suppose N_{β_0}, N_{β_1} and N_μ are the normalizing constants of the VB marginals $q(\beta_0|\mathbf{y}, \mathbf{x})$, $q(\beta_1|\mathbf{y}, \mathbf{x})$ and $q(\mu|\mathbf{y}, \mathbf{x})$ respectively, then

$$\log q(\beta_0|\mathbf{y}, \mathbf{x}) = -\log(N_{\beta_0}) + \left[\sum_{i=1}^{n} -e^{\beta_0}\mathbb{E}_q(e^{\beta_1 \mathrm{x}_i})\mathbb{E}_q(\mathrm{Z}_i) + \beta_0 \mathrm{y}_i \mathbb{E}_q(\mathrm{Z}_i)\right] - \frac{\beta_0^2}{2S_{\beta_0}^2}, \tag{4.110}$$

$$\log q(\beta_1|\mathbf{y}, \mathbf{x}) = -\log(N_{\beta_1}) + \left[\sum_{i=1}^{n} -\mathbb{E}_q(e^{\beta_0})e^{\beta_1 \mathrm{x}_i}\mathbb{E}_q(\mathrm{Z}_i) + \beta_1 \mathrm{y}_i \mathrm{x}_i \mathbb{E}_q(\mathrm{Z}_i)\right] - \frac{\beta_1^2}{2S_{\beta_1}^2}, \tag{4.111}$$

$$\log q(\mu|\mathbf{y}, \mathbf{x}) = -\log(N_\mu) - \frac{1}{2S_\mu^2}(\log\mu - m)^2 - \log\mu - \mu\sum_{i=1}^{n}(1 - \mathbb{E}_q(\mathrm{Z}_i)) + \log\mu \sum_{i=1}^{n} \mathrm{y}_i(1 - \mathbb{E}_q(\mathrm{Z}_i)), \tag{4.112}$$

$$q(\pi|\mathbf{y}, \mathbf{x}) = Beta(a', b'), \tag{4.113}$$

$$a' = a + \sum_{i=1}^{n} \mathbb{E}_q(\mathrm{Z}_i), \tag{4.114}$$

$$b' = b + n - \sum_{i=1}^{n} \mathbb{E}_q(\mathrm{Z}_i), \tag{4.115}$$

$$\log q(\mathrm{Z}_i|\mathbf{y}, \mathbf{x}) = -\log N_{\mathrm{Z}_i} + \mathrm{Z}_i\left[\mathbb{E}_q(\log\pi) - \mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(e^{\beta_1 \mathrm{x}_i}) + \mathrm{y}_i(\mathbb{E}_q(\beta_0) + \mathbb{E}_q(\beta_1)\mathrm{x}_i)\right] + (1 - \mathrm{Z}_i)\left[\mathbb{E}_q(\log(1 - q)) - \mathbb{E}_q(e^{\mu}) + \mathrm{y}_i\mathbb{E}_q(\mu)\right]. \tag{4.116}$$

The VB marginal $q(\pi|\mathbf{y}, \mathbf{x})$ is a standard *Beta* distribution, hence requires no further approximation. The VB marginals $q(\beta_0|\mathbf{y}, \mathbf{x})$, $q(\beta_1|\mathbf{y}, \mathbf{x})$, $q(\mu|\mathbf{y}, \mathbf{x})$ and $q(\mathrm{Z}_i|\mathbf{y}, \mathbf{x})$ do not belong to the standard family of the distributions but can be computed numerically. The VB expectation of π that appears in the VB marginal $q(\mathrm{Z}_i|\mathbf{y}, \mathbf{x})$

can be obtained as:

$$\begin{aligned}
\mathbb{E}_q(\log\pi) &= \psi(a') - \psi(a'+b'), \\
&= \frac{\partial}{\partial a'}\log\gamma(a') - \frac{\partial}{\partial(a'+b')}\log\gamma(a'+b'), \qquad (4.117)\\
\mathbb{E}_q(1-\log\pi) &= \psi(b') - \psi(a'+b'), \\
&= \frac{\partial}{\partial b'}\log\gamma(b') - \frac{\partial}{\partial(a'+b')}\log\gamma(a'+b'). \qquad (4.118)
\end{aligned}$$

VB solution at the Inverse stage:

The posterior distribution of $\mathrm{X}_{\mathrm{new}}$ is

$$\begin{aligned}
P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) &\propto \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1,\mu,\pi,\mathbf{Z})P(\mathrm{X}_{\mathrm{new}}) \\
&\quad \times P(\beta_0,\beta_1,\mu,\pi,\mathbf{Z}|\mathbf{y},\mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\mu\mathrm{d}\pi\mathrm{d}\mathbf{Z}, \qquad (4.119)\\
&\approx \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1,\mu,\pi,\mathbf{Z})P(\mathrm{X}_{\mathrm{new}}) \\
&\quad \times q(\beta_0,\beta_1,\mu,\pi,\mathbf{Z}|\mathbf{y},\mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\mu\mathrm{d}\pi\mathrm{d}\mathbf{Z}. \qquad (4.120)
\end{aligned}$$

The integral in the R.H.S of the expression is not available in closed form. For a tractable VB approximation of $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}})$ the restricted VB method is applied:

$$\begin{aligned}
P(\mathrm{X}_{\mathrm{new}}|\mathrm{y},\mathbf{x},\mathbf{y}_{\mathrm{new}}) &\approx \bar{q}(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}), \qquad (4.121)\\
\log\bar{q}(\mathrm{X}_{\mathrm{new}}|\mathrm{y},\mathbf{x},\mathbf{y}_{\mathrm{new}}) &= \mathbb{E}_{q(\beta_0,\beta_1,\mu,\pi,\mathbf{Z}|\mathbf{y},\mathbf{x})}\log[P(\mathrm{y}_{\mathrm{new}}|\beta_0,\beta_1,\mu,\pi,\mathbf{Z},\mathrm{X}_{\mathrm{new}}) \\
&\quad \times P(\mathrm{X}_{\mathrm{new}})], \qquad (4.122)\\
\log\bar{q}(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) &\approx \mathbb{E}_q(\mathrm{Z}_{\mathrm{new}})\left[-\mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(e^{\beta_1\mathrm{X}_{\mathrm{new}}}) + \mathbb{E}_q(\beta_1)\mathrm{y}_{\mathrm{new}}\mathrm{X}_{\mathrm{new}}\right] \\
&\quad + \log P(\mathrm{X}_{\mathrm{new}}). \qquad (4.123)
\end{aligned}$$

The term $e^{\beta_1\mathrm{X}_{\mathrm{new}}}$ in the R.H.S of the expression for the the restricted VB approximation $\bar{q}(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}})$ is not factorized over β_1 and $\mathrm{X}_{\mathrm{new}}$, therefore the VB-expectation of the term is not defined in a closed form which makes $\bar{q}(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}})$ intractable.

As for the inverse Poisson regression, a Gaussian approximation may be applied to find a tractable solution to $\bar{q}(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}})$. To compute a Gaussian approx-

imation consider the Taylor's expansion of $\bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}})$ as a function of $\mathrm{X_{new}}$ around its posterior mode (denoted as X_a). The Gaussian approximation to the restricted VB approximation of $\mathrm{X_{new}}$ is given as:

$$\begin{aligned}
\bar{q}(\mathrm{X_{new}}|\mathbf{y}, \mathbf{x}, \mathrm{y_{new}}) &= N(\mu_X^\star, S_X^{2\star}), && (4.124)\\
S_X^{2\star} &= \left[\frac{1}{S_{\mathrm{X}}^2} + \mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(\beta_1^2 e^{\beta_1 \mathrm{X}_a})\right]^{-1}, && (4.125)\\
\mu_X^\star &= S_X^{2\star}\left[\frac{\mu_{\mathrm{X}}}{S_{\mathrm{X}}^2} + \mathbb{E}_q(\mathrm{Z_{new}})(-\mathbb{E}_q(e^{\beta_0})\mathbb{E}_q(\beta_1 e^{\beta_1 \mathrm{X}_a})\right.\\
&\qquad \left. + \mathrm{y_{new}}\mathbb{E}_q(\beta_1))\right]. && (4.126)
\end{aligned}$$

The VB-expectations with respect to β_1 and β_0 are not explicitly but can be computed numerically. The posterior mode X_a may be found by an optimization method. But as mentioned in the previous section for the inverse Poisson regression problem, an algorithm by Rue & Held (2005) provides a quick and easy Gaussian approximation without using any other optimization technique to find the posterior mode. The same algorithm can also be applied to compute a Gaussian approximation to the intractable restricted VB approximation over $\mathrm{X_{new}}$ of the mixture of Poisson regression problem.

Comparison of the VB approximation with the result from other methods:

Two types of methods are considered to check the accuracy of the VB approximation for the mixture of Poisson regression problem. First method is simulation based, the Monte Carlo method, and the second method is the Gaussian variational approximation.

The mixture of Poisson regression model is an example of a complex model for which the true result is feasible to find. A numerical integration method is even very expensive for the model with many parameters to be estimated. Though the Monte Carlo integration may take a long time to provide an accurate result, it is tempting to compare results with the VB approximation for this non-conjugate-exponential model to see if the VB method provides a good trade off between the accuracy and

the time consumption.

Gaussian variational approximation to the inverse problem:

The Gaussian variational approach for the non-conjugate-exponential models is to restrict the approximation to a Gaussian density. The variational method with a Gaussian approximation (described in the chapter) provides a Gaussian approximation of the intractable variational approximation which uses the Taylor's expansion of the non-linear or non-quadratic function around the posterior mode of the variable. Whereas with the Gaussian variational method, one has to compute the mean and the variance parameters of the Gaussian variational approximation (also called the variational parameters) by optimizing the lower bound (as a function of the variational parameters) on the$\log P(\mathrm{y}_{\mathrm{new}}) \geq F(\mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star})$,

The Gaussian variational a $q(\mathrm{X}_{\mathrm{new}}) = N(\mathrm{X}_{\mathrm{new}}; \mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star})$.ibution of $\mathrm{X}_{\mathrm{new}}$ is described by considering a lower bound on the log-marginal likelihood:

$$\log P(\mathrm{y}_{\mathrm{new}}) \geq \int q(\beta_0)q(\beta_1)q(\mu)q(\pi)q(\mathrm{Z}_{\mathrm{new}})q(\mathrm{X}_{\mathrm{new}})$$

$$\log P(\mathrm{y}_{\mathrm{new}}) \geq \int q(\beta_0)q(\beta_1)q(\mu)q(\pi)q(\mathrm{Z}_{\mathrm{new}})q(\mathrm{X}_{\mathrm{new}}) \qquad \frac{\mu, \pi)}{(\mathrm{X}_{\mathrm{new}})}\Bigg] \mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\mu\mathrm{d}\pi. \quad (4.127)$$

$$\left[\quad P(\mathrm{y}_{\mathrm{new}}, \mathrm{X}_{\mathrm{new}}, \mathrm{Z}_{\mathrm{new}}, \beta_1, \beta_0, \mu, \pi)\right.$$

The terms $\mathbf{y}$, $\mathbf{x}$ and $\mathrm{y}_{\mathrm{new}}$ has been dropped from $q(\cdot)$ for the simplicity of the notation. If the variational marginal $q(\mathrm{X}_{\mathrm{new}})$ is restricted to a Gaussian distribution with the unknown mean $\mu_{\mathrm{X}}^{\star}$ and the variance $S_{\mathrm{X}}^{2\star}$. Then the lower bound is a function of the these unknown parameters, also called as the variational parameters:

$$\begin{aligned} F(\mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star}) &= \int q(\beta_0)q(\beta_1)q(\mu)q(\pi)q(\mathrm{Z}_{\mathrm{new}})q(\mathrm{X}_{\mathrm{new}}) \\ &\quad \times \log\left[\frac{P(\mathrm{y}_{\mathrm{new}}, \mathrm{X}_{\mathrm{new}}, \mathrm{Z}_{\mathrm{new}}, \beta_1, \beta_0, \mu, \pi)}{q(\beta_0)q(\beta_1)q(\mu)q(\pi)q(\mathrm{Z}_{\mathrm{new}})q(\mathrm{X}_{\mathrm{new}})}\right] \\ &\quad \times \mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\mu\mathrm{d}\pi, \end{aligned} \quad (4.128)$$

$$\log P(\mathrm{y}_{\mathrm{new}}) \geq F(\mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star}), \quad (4.129)$$

$$q(\mathrm{X}_{\mathrm{new}}) = N(\mathrm{X}_{\mathrm{new}}; \mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star}). \quad (4.130)$$

It should be noted that the auxiliary variable $\mathrm{Z}_{\mathrm{new}}$ is related to the new count $\mathrm{y}_{\mathrm{new}}$

and so the VB marginal $q(\mathrm{Z_{new}})$ is not found at the forward stage of the inference problem. The VB marginal can be computed at the inverse stage by the restricted VB method:

$$q(\mathrm{Z_{new}}) \propto \mathbb{E}_{q(\beta_0)q(\beta_1)q(\mu)q(\pi)q(\mathrm{X_{new}})}\Big[P(\mathrm{y_{new}}, |\mathrm{X_{new}}, \beta_1, \beta_0, \mu)P(\mathrm{Z_{new}}|\pi)\Big], \tag{4.131}$$

where the VB marginal $q(\mathrm{X_{new}})$ as to be found by the Gaussian variational approach.

The variational parameters $\mu_{\mathrm{X}}^{\star}$ and $S_{\mathrm{X}}^{2\star}$ can be found by maximizing the lower bound so that it reaches to the true value of the log-marginal likelihood.

$$S_{\mathrm{X}}^{2\star} = \arg\max_{S_{\mathrm{X}}^{2\star}} F(\mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star}), \tag{4.132}$$

$$\mu_{\mathrm{X}}^{\star} = \arg\max_{\mu_{\mathrm{X}}^{\star}} F(\mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star}). \tag{4.133}$$

The Gaussian variational approximation for the inverse estimation problem for the mixture of the Poisson model is given as follows:

$$q_{GV}(\mathrm{X_{new}}|\mathrm{y_{new}}, \mathbf{x}, \mathbf{y}) = N(\mathrm{X_{new}}; \mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\star}), \tag{4.134}$$

Where $\mu_{\mathrm{X}}^{\star}$ and $S_{\mathrm{X}}^{2\star}$ can be found from the expressions below

$$S_{\mathrm{X}}^{2\star}\left[\mathbb{E}_q(\mathrm{Z_{new}})\mathbb{E}_q(e^{\beta_0})\hat{\beta}_1{}^2 \exp(\hat{\beta}_1\mu_{\mathrm{X}}^{\star} + 0.5\hat{\beta}_1{}^2 S_{\mathrm{X}}^{2\star}) + \frac{1}{S_{\mathrm{X}}^2}\right] = 1, \tag{4.135}$$

$$\mu_{\mathrm{X}}^{\star} - \mu_{\mathrm{X}} + \frac{1}{\hat{\beta}_1}\left(\frac{S_{\mathrm{X}}^2}{S_{\mathrm{X}}^{2\star}} - 1\right) - \mathrm{y_{new}}\mathbb{E}_q(\beta_1)S_{\mathrm{X}}^2 = 0, \tag{4.136}$$

where $\hat{\beta}_1$ is the posterior mode of β_1 used to avoid its complicated VB-expectations.

Evaluation of VB approximation:

Another accuracy check of the VB approximation is found in terms of its 95% HPD region. The 95% HPD region of $\mathrm{X_{new}}$ is obtained as $\mu_X^{\star} \pm 2\sqrt{S_X^{2\star}}$.

Result

Fifty observations of Y are generated from a mixture of Poisson distribution given the true values of the regression parameters as $\beta_0 = 0.1$ and $\beta_1 = 1.5$ and

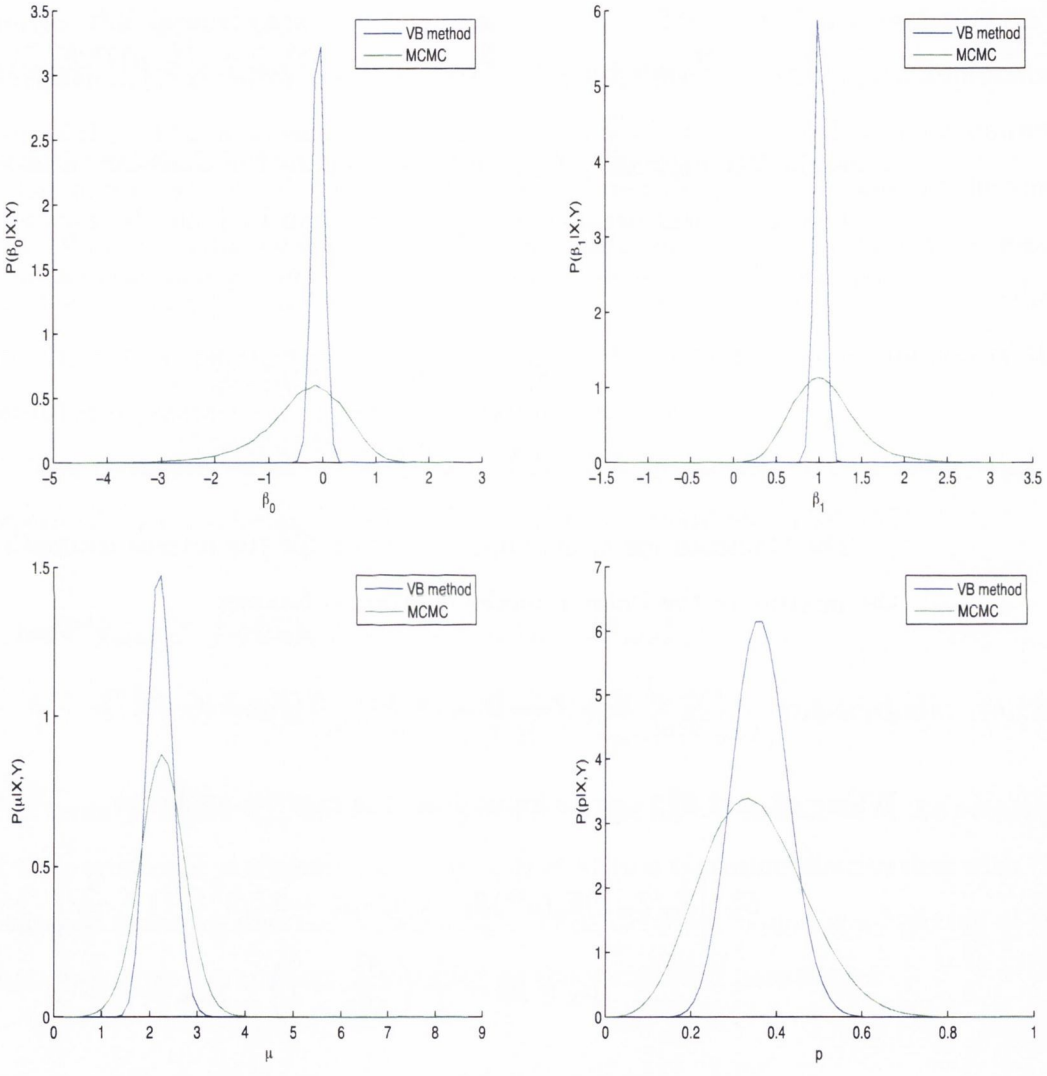

Fig. 4.16: The comparison of the true by the MCMC (green) and the VB approximations (blue) of the marginal posterior distributions of the parameters β_0 (the top left), β_1 (the top right), μ (the bottom left) and π (the bottom right) of a mixture of Poisson regression problem, is shown. The VB marginal of β_0 and β_1 are peaked. The regression parameters β_0, β_1 are too correlated to allow for the posterior independence.

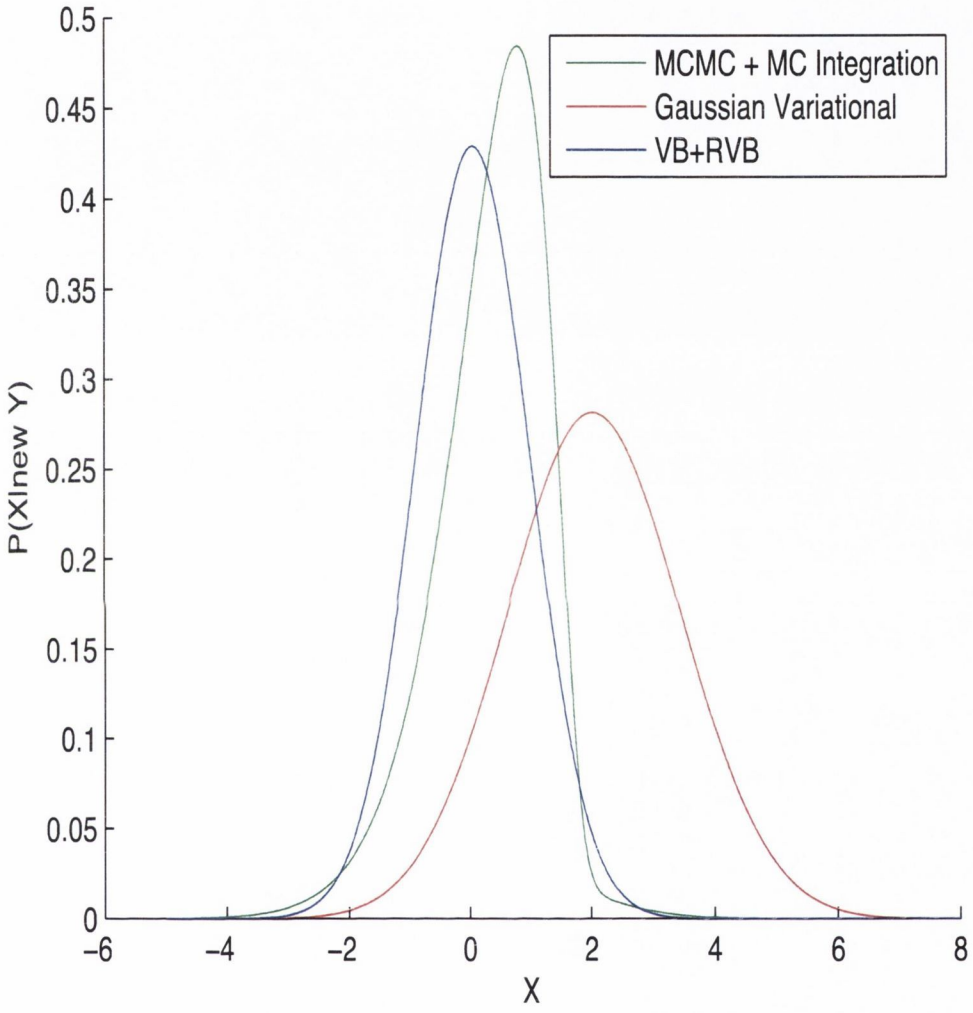

Fig. 4.17: The comparison of the true posterior distributions of an explanatory variable X_{new} of a mixture of Poisson regression problem given a small count on $Y_{new}(=1)$ by the MCMC method at the forward stage and Monte Carlo integration at the inverse stage (green), the VB approximations at the forward stage and the restricted VB approximation at the inverse stage (blue), the Gaussian variational approximation at the inverse stage (red), is shown. The VB marginal by the restricted VB method matches with the Gaussian variational approximation.

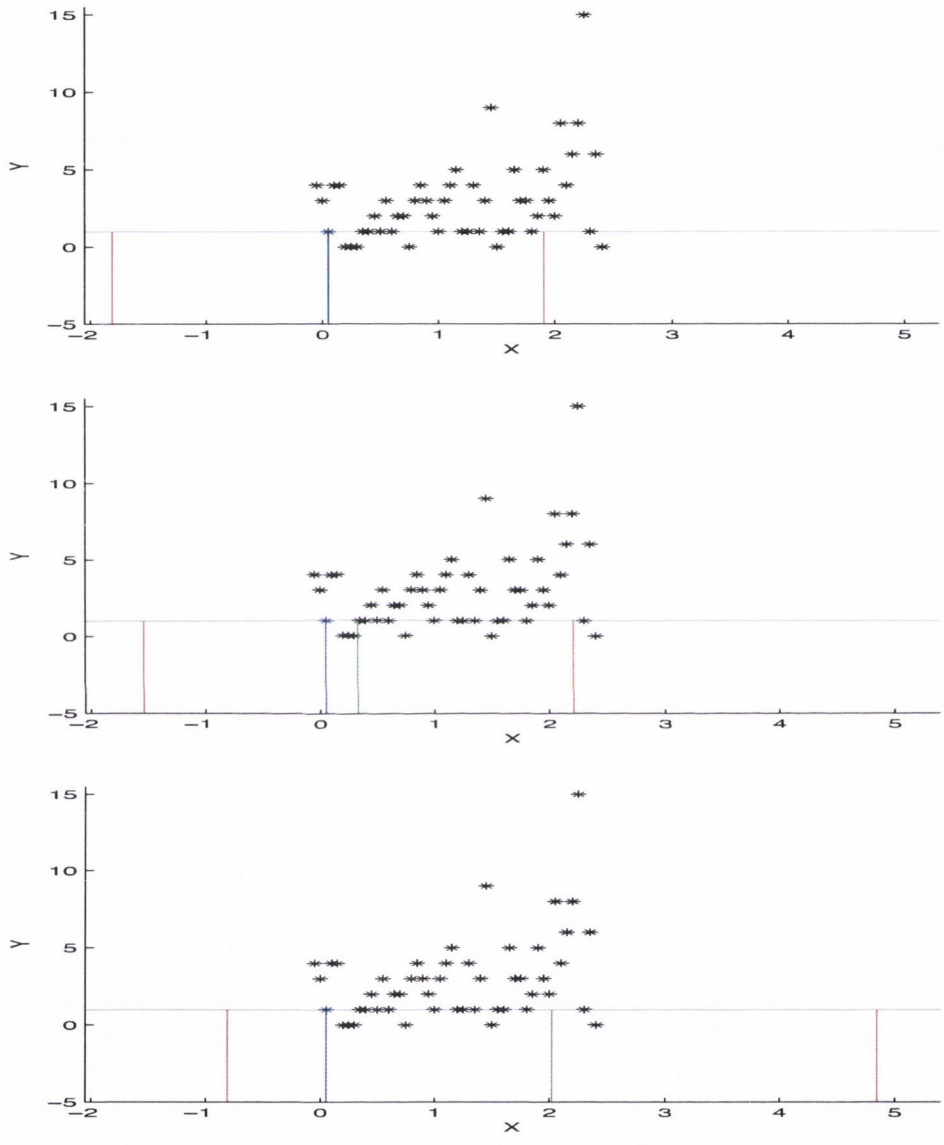

Fig. 4.18: Posterior estimates of an explanatory variable X_{new} of a mixture of Poisson regression problem with a normal prior by by the VB method and RVB at forward and inverse stage respectively (the uppermost), by the MCMC and the Monte Carlo integration at forward and inverse stage respectively (the middle) and by the VB and the Gaussian variational approximation at the forward and inverse stage respectively (the last). The comparison of the true value (blue) and the estimates (green) of X_{new}, with the lower and upper bounds (in red) of 95% HPD region, is shown. The black asterisks are training data and the blue is for a new data on the response variable Y.

the fifty equi-spaced values of X. Results from the VB approximation for an inverse mixture of Poisson regression problem are shown in Fig. 4.16, 4.17 and 4.18. Fig. 4.16 shows the VB approximation at the forward stage of the problem and the (true) results from the MCMC. Just like in the Poisson regression problem, the VB method under-estimates the true posterior variance of the regression parameters β_0 and β_1. The independence assumption of the VB method might be a reason to this under-estimation of the posterior variance of the regression parameters which are strongly correlated in the model. The VB marginals of the parameters μ and π do not show the under-estimation of the variance, though they do not coincide with the result from the MCMC.

Fig. 4.17 and 4.18 represent the VB approximation of X_{new} given a small count ($Y_{new} = 1$) at the inverse stage. In Fig 4.17, the VB marginal of X_{new} is compared with the results from the Monte Carlo integration and the Gaussian variational approach. Fig. 4.18 shows the VB approximation and the true result (by Monte Carlo integration) for a small count on X_{new}. Comparing with the results (true) from the Monte Carlo integration, the VB method provides a better approximation to the posterior distribution than the Gaussian variational approximation. At the inverse stage, the effect of the posterior independence assumption of the VB method is not very much reflected in the result. As a result of accuracy check via leave-one-out cross validation technique, 98% counts are inside their 95% HPD regions for VB and restricted VB (RVB) approximation.

Hence, the VB method provides a good approximation to inverse prediction for the mixture of Poisson regression model.

4.2.7 VB solution to Inverse Zero-Inflated Poisson regression

In the zero-inflated Poisson regression model (defined earlier in the chapter), there are only two parameters to be estimated:

$$\theta = \{\beta_0, \beta_1\}.$$

The prior distributions over β_0 and β_1 assumed to be a Gaussian as follows:

$$P(\beta_0) = N(0, S^2_{\beta_0}), \tag{4.137}$$

$$P(\beta_1) = N(0, S^2_{\beta_1}). \tag{4.138}$$

The prior distribution over the explanatory variable $\mathrm{X}_{\mathrm{new}}$ given a new count on $\mathrm{Y}_{\mathrm{new}}$ is given as:

$$P(\mathrm{X}_{\mathrm{new}}) = N(\mu_{\mathrm{X}}, S^2_{\mathrm{X}}). \tag{4.139}$$

VB solution at Forward stage:

A VB $q(\beta_0|\mathbf{y}, \mathbf{x})$ approximation to the joint posterior distribution $P(\beta_0, \beta_1|\mathbf{y}, \mathbf{x})$ is explained as follows:

$$P(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) = q(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}), \tag{4.140}$$

$$q(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) = q(\beta_0|\mathbf{y}, \mathbf{x})q(\beta_1|\mathbf{y}, \mathbf{x}), \tag{4.141}$$

$$\log q(\beta_0|\mathbf{y}, \mathbf{x}) \approx \log P(\beta_0) + \mathbb{E}_{q(\beta_1|\mathbf{y},\mathbf{x})}[\log P(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1)], \tag{4.142}$$

$$\log q(\beta_1|\mathbf{y}, \mathbf{x}) \approx \log P(\beta_1) + \mathbb{E}_{q(\beta_0|\mathbf{y},\mathbf{x})}[\log P(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1)], \tag{4.143}$$

where

$$\begin{aligned} \log P(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1) = & \sum_{i=1}^{n} \mathrm{z}_i \log\left(1 - \pi_i + \pi_i e^{-\exp(\beta_0 + \beta_1 \mathrm{x}_i)}\right) \\ & + (1 - \mathrm{z}_i)\Big[\log \pi_i - \exp(\beta_0 + \beta_1 \mathrm{x}_i) \\ & + \mathrm{y}_i(\beta_0 + \beta_1 \mathrm{x}_i) - \log \gamma(1 + \mathrm{y}_i)\Big], \end{aligned} \tag{4.144}$$

where z_i's are an indicator function such that $\mathrm{z}_i = 1$; if $\mathrm{y}_i = 0$, otherwise, $\mathrm{z}_i = 0$; if $\mathrm{y}_i > 0$ The expectations of the log-likelihood $\log P(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1)$ with respect to $q(\beta_1|\mathbf{y}, \mathbf{x})$ and $q(\beta_0|\mathbf{y}, \mathbf{x})$ respectively, are not available in closed form as the log-likelihood is not factorized over β_0 and β_1. For a tractable VB approximation, the algorithm for a Gaussian approximation by Rue & Held (2005) can be applied in the same way as discussed for the Poisson and the mixture of Poisson regression problems.

For a Gaussian approximation, consider a Taylor's expansion of the (full) log-likelihood as a function of the regression parameter (for which the VB marginal is to be computed) around its posterior mode. The Gaussian approximation to the VB marginals of β_0 ,$q(\beta_0|\mathbf{y},\mathbf{x})$ is obtained as:

$$
\begin{aligned}
q_g(\beta_0|\mathbf{y},\mathbf{x}) &= N(\mu^{\star}_{\beta_0}, S^{2\,\star}_{\beta_0}), && (4.145)\\
S^{2\,\star}_{\beta_0} &= \left[\frac{1}{S^2_{\beta_0}} - f_2\right], && (4.146)\\
\mu^{\star}_{\beta_0} &= \left(\frac{\mu_{\beta_0}}{S^2_{\beta_0}} + f_1 - \beta_0^m f_2\right), && (4.147)\\
f_1 &= \frac{\partial}{\partial\beta_0}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, && (4.148)\\
f_2 &= \frac{\partial^2}{\partial\beta_0^2}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, && (4.149)
\end{aligned}
$$

where the term $q_g(\beta_0|\mathbf{y},\mathbf{x})$ denotes the Gaussian approximation, $\mu^{\star}_{\beta_0}$ and $S^{2\,\star}_{\beta_0}$ are the mean and variance of the approximation respectively, f_1 and f_2 are the first and second derivative of the log-likelihood with respect to β_0 around its posterior mode β_0^m.

Similarly, the Gaussian approximation to $q(\beta_1|\mathbf{y},\mathbf{x})$ is found as:

$$
\begin{aligned}
q_g(\beta_1|\mathbf{y},\mathbf{x}) &= N(\mu^{\star}_{\beta_1}, S^{2\,\star}_{\beta_1}), && (4.150)\\
S^{2\,\star}_{\beta_1} &= \left[\frac{1}{S^2_{\beta_1}} - g_2\right], && (4.151)\\
\mu^{\star}_{\beta_1} &= \left(\frac{\mu_{\beta_1}}{S^2_{\beta_1}} + g_1 - \beta_1^m g_2\right), && (4.152)\\
g_1 &= \frac{\partial}{\partial\beta_1}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, && (4.153)\\
g_2 &= \frac{\partial^2}{\partial\beta_1^2}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, && (4.154)
\end{aligned}
$$

where the term $q_g(\beta_1|\mathbf{y},\mathbf{x})$ denotes the Gaussian approximation, $\mu^{\star}_{\beta_1}$ and $S^{2\,\star}_{\beta_0}$ are the mean and variance of the approximation respectively, g_1 and g_2 are the first and second derivative of the log-likelihood with respect to β_0 around its posterior mode

β_1^m.

The functions f_1, f_2, g_1 and g_2 are defined in Appendix.

It should be noted that even after considering the Taylor's expansion of the log-likelihood, it still does not factorize over the parameters. To simplify this problem, the posterior mode of the parameters are plugged-in wherever the VB-expectation is needed. Considering a Gaussian approximation may not simplify the VB expectations, however it provides the VB approximation in a standard form of distribution.

VB solution at the Inverse stage:

The posterior distribution of $\mathrm{X}_{\mathrm{new}}$ is

$$\begin{aligned} P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}) &\propto \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1)P(\mathrm{X}_{\mathrm{new}}) \\ &\qquad \times P(\beta_0,\beta_1|\mathbf{y},\mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1, \qquad (4.155) \\ &\approx \int P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1)P(\mathrm{X}_{\mathrm{new}}) \\ &\qquad \times q(\beta_0,\beta_1|\mathbf{y},\mathbf{x})\mathrm{d}\beta_0\mathrm{d}\beta_1. \qquad (4.156) \end{aligned}$$

The posterior distribution is not available in closed form. The restricted VB method is applied to find a tractable VB approximation of $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}})$:

$$\begin{aligned} P(\mathrm{X}_{\mathrm{new}}|\mathrm{y},\mathbf{x},\mathbf{y}_{\mathrm{new}}) &\approx \bar{q}(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x},\mathrm{y}_{\mathrm{new}}), \qquad (4.157) \\ \log\bar{q}(\mathrm{X}_{\mathrm{new}}|\mathrm{y},\mathbf{x},\mathbf{y}_{\mathrm{new}}) &= \mathbb{E}_{q(\beta_0,\beta_1|\mathbf{y},\mathbf{x})}\log[P(\mathrm{y}_{\mathrm{new}}|\beta_0,\beta_1,\mathrm{X}_{\mathrm{new}}) \\ &\qquad \times P(\mathrm{X}_{\mathrm{new}})]. \qquad (4.158) \end{aligned}$$

As discussed at the forward stage of the inference that the $\log P(\mathrm{y}_{\mathrm{new}}|\beta_0,\beta_1,\mathrm{X}_{\mathrm{new}})$ is not factorized over β_0 and β_1, the posterior modes of β_0 and β_1 are plugged-in to avoid the intractable VB-expectations. Even after this the VB approximation of $\mathrm{X}_{\mathrm{new}}$ remains intractable. To solve the issue of the intractability, the Gaussian approximation approach as discussed at the forward stage is applied. The Gaussian approximation to the intractable VB approximation of $\mathrm{X}_{\mathrm{new}}$ by following the same

algorithm by Rue & Held (2005) is found as:

$$q_g(\mathrm{X_{new}}|\mathbf{y},\mathbf{x}) = N(\mu_{\mathrm{X}}^{\star}, S_{\mathrm{X}}^{2\,\star}), \tag{4.159}$$

$$S_{\mathrm{X}}^{2\,\star} = \left[\frac{1}{S_{\mathrm{X}}^2} - h_2\right], \tag{4.160}$$

$$\mu_{\mathrm{X}}^{\star} = \left(\frac{\mu_{\mathrm{X}}}{S_{\mathrm{X}}^2} + h_1 - \mathrm{X}_{\mathrm{new}}^{m} h_2\right), \tag{4.161}$$

$$h_1 = \frac{\partial}{\partial \mathrm{X_{new}}} \log P(\mathrm{y_{new}}|\mathrm{X_{new}}, \beta_0, \beta_1)\Big|_{\mathrm{X}_{\mathrm{new}}^{m}, \beta_0^{m}, \beta_1^{m}}, \tag{4.162}$$

$$h_2 = \frac{\partial^2}{\partial \mathrm{X_{new}^2}} \log P(\mathrm{y_{new}}|\mathrm{X_{new}}, \beta_0, \beta_1)\Big|_{\mathrm{X}_{\mathrm{new}}^{m}, \beta_0^{m}, \beta_1^{m}}, \tag{4.163}$$

where the term $q_g(\mathrm{X_{new}}|\mathbf{y},\mathbf{x})$ denotes the Gaussian approximation, $\mu_{\mathrm{X}}^{\star}$ and $S_{\mathrm{X}}^{2\,\star}$ are the mean and variance of the approximation respectively, h_1 and h_2 are the first and second derivative of $\log P(\mathrm{y_{new}}|\mathrm{X_{new}}, \beta_0, \beta_1)$ with respect to $\mathrm{X_{new}}$ around its posterior mode $\mathrm{X}_{\mathrm{new}}^{m}$.

Comparison of the VB approximation with other methods:

The Gaussian variational approximation to the inference problem for the ZI-Poisson model is not tractable. Therefore, the VB results for the ZI-Poisson regression model are compared only with the results from MCMC and from a numerical integration method.

Evaluation of VB approximation:

Another accuracy check of the VB approximation is found in terms of its 95% HPD region. The 95% HPD region of $\mathrm{X_{new}}$ is obtained as $\mu_X^{\star} \pm 2\sqrt{S_X^{2\star}}$.

Result

Two hundred observations of Y are generated from a ZI-Poisson distribution corresponding to two hundred values of X with the true values of the regression parameters as $\beta_0 = 0.5$ and $\beta_1 = 1.5$. A comparison of the VB approximations of β_0 and β_1 with their true posterior distributions by a numerical integration and the approximation by the MCMC for an inverse ZI-Poisson regression problem is shown

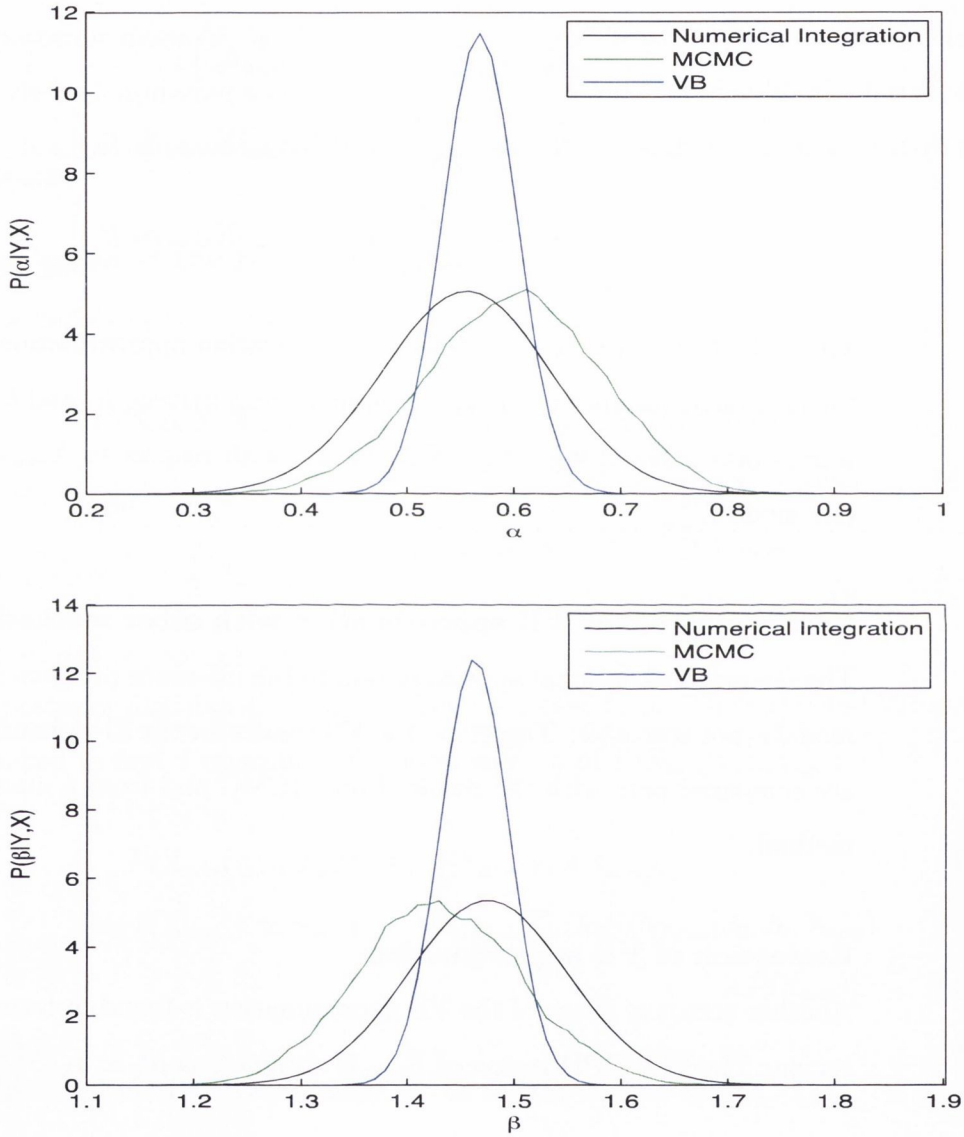

Fig. 4.19: The comparison of the true by the MCMC (green) and the VB approximations (blue) of the marginal posterior distributions of the parameters β_0 (top) and β_1 (bottom) of a ZI-Poisson regression problem, is shown. The VB marginal of β_0 and β_1 are peaked. The regression parameters β_0, β_1 are too correlated to allow for the posterior independence.

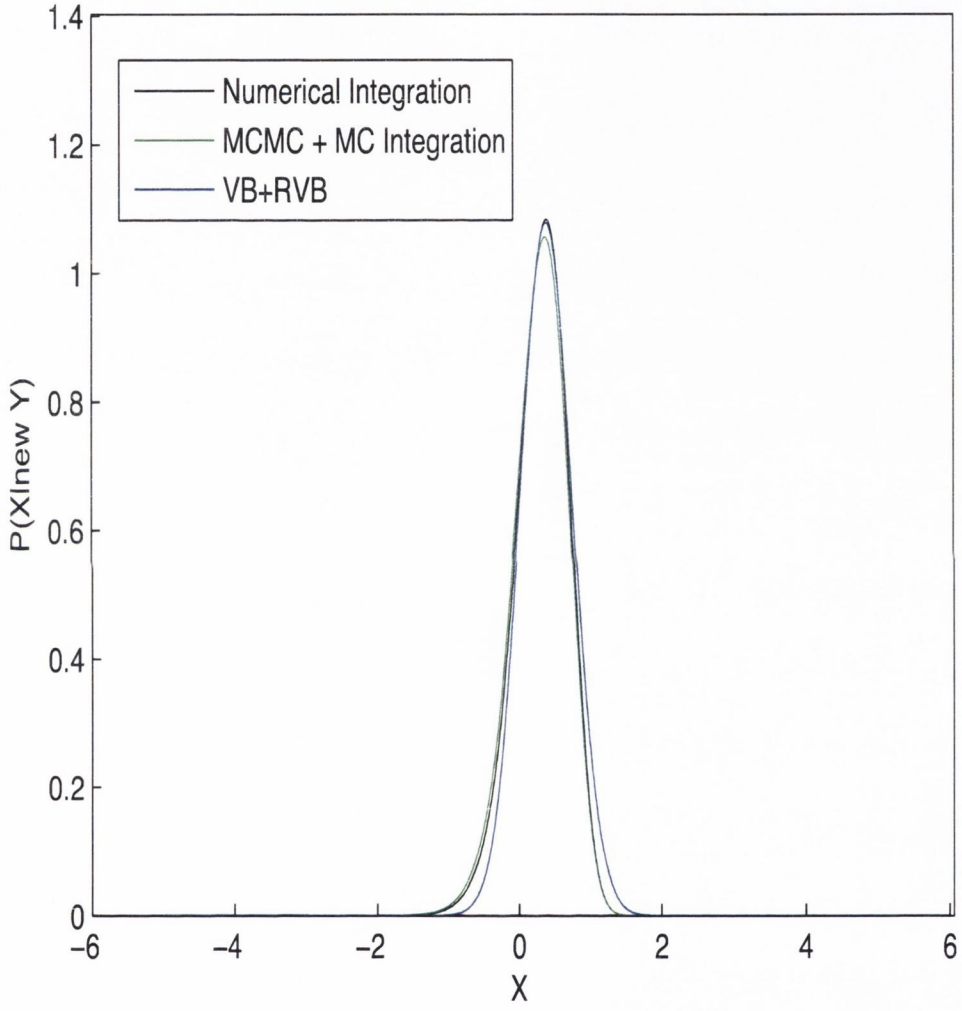

Fig. 4.20: The comparison of the true posterior distributions of an explanatory variable X_{new} of a ZI-Poisson regression problem by the MCMC method at the forward stage and Monte Carlo integration at the inverse stage (green), and by a numerical integration (black) and the VB approximations at the forward stage and the restricted VB approximation at the inverse stage (blue), is shown.

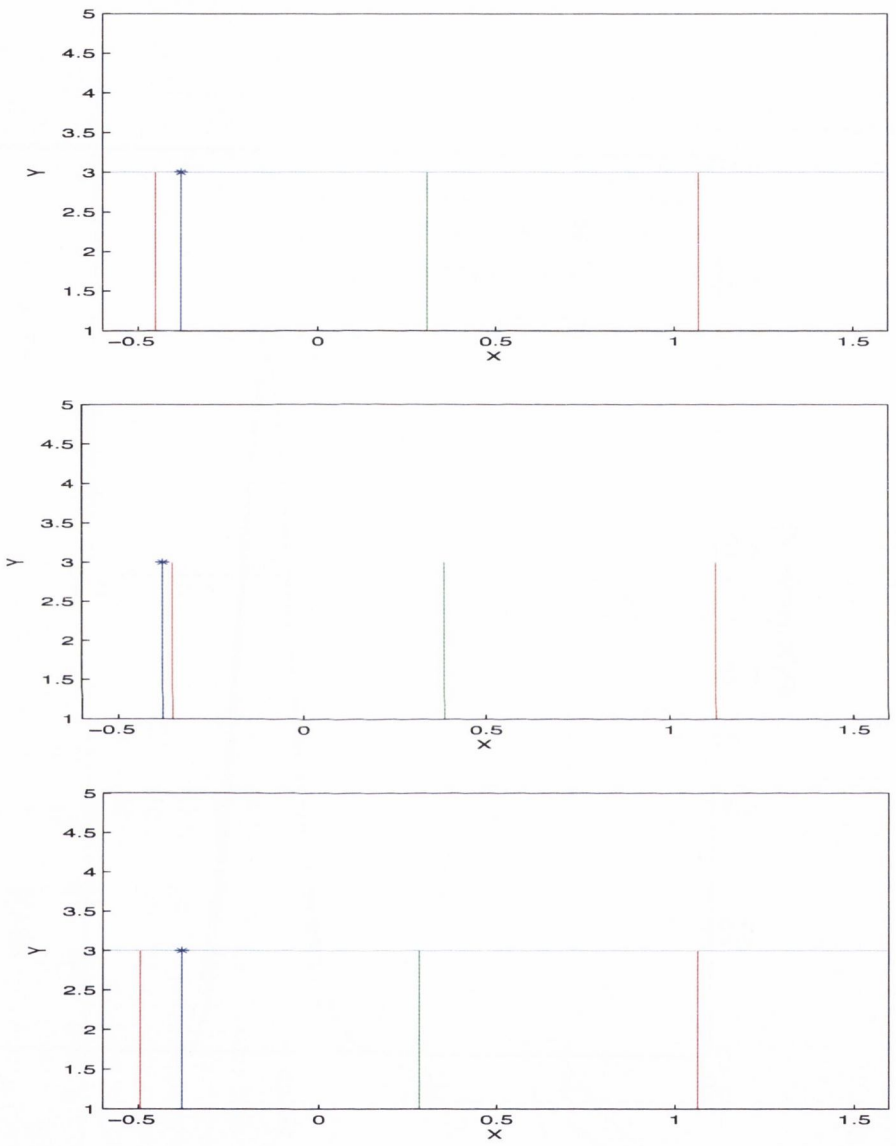

Fig. 4.21: Posterior estimates of an explanatory variable X_{new} (given $Y_{new} = 3$) of a ZI-Poisson regression problem with a normal prior by by the VB method and NI at forward and inverse stage respectively (the uppermost), by the VB and the RVB at the forward and inverse stage respectively (the middle) and by the MCMC and the Monte Carlo integration at forward and inverse stage respectively (the last). The comparison of the true value (blue) and the estimates (green) of X_{new} of a ZI-Poisson regression problem, with the lower and upper bounds (in red) of 95% HPD region, is shown. The black asterisks are training data and the blue is for a new data on the response variable Y.

in Fig. 4.19. It is clear from the figure that the posterior variance is under-estimated by the VB method.

In Fig. 4.20 and 4.21, the comparison of the VB approximation of X_{new} with the results from the MCMC and a numerical integration, is shown. In Fig 4.20, the VB marginal of X_{new} is compared with the results from the Monte Carlo integration and a numerical integration. Since, a Gaussian approximation of the VB approximation is considered, the VB marginal is symmetric. It is very close to the results from the MCMC and from a numerical integration. Therefore, the VB method provides a good approximation to the true posterior distribution for this particular ZI-Poisson model. The Fig. 4.21 shows the comparison of the true value and the estimates of X_{new} and its 95% HPD region. The estimates by the MCMC and by a numerical integration lie within the 95% HPD region whereas the VB-estimate appears outside the region. This requires to perform a leave-one-out cross validation, to check the accuracy of the result through the percentage of the points lie inside the 95% HPD region. There are only 5% of points outside the 95% HPD (VB) region of X_{new} that favours the VB method with a Gaussian approximation (explained in the chapter) for the inverse inference for the ZI-Poisson model.

4.3 Discussion

The VB method for inverse prediction performs well when compared to the results on prediction from an accurate numerical integration method and from other methods such as MCMC, Gaussian variational approximation and variational tangent approach. VB approximations for the quadratic and simple linear regression problem show that the method is straightforward to apply and provides a quick and tractable solution to exponential family and conjugate prior distributions. Šmídl & Quinn (2006) comment on the limitation of the VB method for non-CE (non conjugate-exponential) models. VB approximations for non-CE models might not be tractable due to an un-factorized log-joint likelihood over the unknowns for the model. It may require further other approximations to intractable VB approximations. This is experienced in the approximation for the inverse Poisson regression

model, mixture of Poisson regression model and ZI-Poisson regression model. In this case, it is shown that a further Gaussian approximation with the VB method improves the approximation.

For inverse quadratic and simple linear regression problems, it is found that the VB method under-estimates the posterior variance of the regression parameters due to the posterior independence assumption (Šmídl & Quinn, 2006). However, the under-estimation of the VB variance is unavoidable with the VB method even if the independence assumption is not considered (it is already discussed in chapter 3). As a result of under-estimation of the variance of the parameters, the variance (VB) of the inverse prediction is also under-estimated. It is observed that for small values of the slope parameter, the variance of the true prediction is large if an improper prior is assumed. Krutchkoff (1967) and Hoadley (1970) discussed the issue of infinite variance of prediction as a decreasing function of the slope parameter. The VB-variance of prediction of X is an inverse function of $\mathbb{E}_q(\beta^2)$ and $\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)$.

$$\mathbb{V}_q(\mathrm{X}) = \left[\mathbb{E}_q(\beta^2)\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\right]^{-1}.$$

Posterior variance of prediction with non-informative prior can be represented by variance from the classical approach of inference.

$$\mathbb{V}(\mathrm{X}) = \frac{\hat{\beta}^2}{\hat{\sigma^2}}.$$

Comparing VB with the classical result; $\mathbb{E}_q(\beta^2) > \hat{\beta}^2$, even if the VB-variance of β is underestimated, also, $\hat{\sigma^2} > \frac{1}{\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)}$. Thus, $\mathbb{V}(\mathrm{X}) > \mathbb{V}_q(\mathrm{X})$. In the examples of non-latent regression problems, σ^2 is kept small to discuss the role of only β, the slope parameter in the uncertainty of estimation. It should be noted that a small $\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)$ (a VB estimate of $\frac{1}{\sigma^2}$) can also make the variance of prediction very large. The problem of large variance with an improper prior is not very well studied for inverse quadratic regression models. However, it is further seen that an assumption of a proper prior may solve the issue of the large variance of inverse inference for simple linear and quadratic regression problems.

It is shown that the VB approximation for prediction of X of the quadratic

regression problem is bimodal. It does not lack any modes of the true posterior distribution. It should be noted that the bi-modality behaviour of the posterior distribution of X is due to quadratic nature of the model and not from the identifiability or label switching problem as described in Jaakkola (2000) and Šmídl & Quinn (2006).

For the inverse Poisson regression model, it is observed that the VB solution (with Gaussian approximation) is very accurate for large (training) data. Thus for inverse Poisson regression, an accurate VB approximation of inverse estimation can be achieved if large data is used for the estimation of the regression parameters. Also, a Gaussian approximation for a large new data of the response variable Y improves the accuracy of the VB approximation. For small or weak data, an assumption of a proper prior avoids the problem of large (true) variance. It is shown that the Gaussian approximation of the VB approximation also improves the accuracy even if an informative prior is not available. Though, it was also experienced that the inverse estimation is not unique or finite for zero counts of a Y with an improper prior. In case of zero counts, the approximation tries to find a very small value of X responsible for the zero counts and there is no unique and finite solution. In such a situation, an assumption of a proper informative prior with finite variance over X should be a reasonable solution. However in general, to avoid such situations, it can also be suggested to use more than one data point to predict X. It also results in a reduction in bias of the prediction.

It is experienced that the Gaussian approximation of the VB approximations also works very well as compared to the result from the MCMC for the mixture of Poisson regression problem and for the ZI-Poisson regression problem. The method outperforms the variational Tangent approach for the mixture of Poisson regression problem. Therefore, it can be concluded that the method, the Gaussian approximation with the VB method discussed in the chapter provides a good approximation for the low-dimensional inference problem for which the direct VB approximation is not tractable. This encourages us to use the method for high dimensional and complex problems for its accuracy and tractability.

Hoadley (1970) derived the marginal posterior distribution of X of a simple linear

regression model, assuming a t-distribution, a proper prior on X and a Jeffrey's prior on regression parameters β_1, β_2 and variance parameter σ^2. He avoided the complex analytical Bayesian solution and used the posterior estimates of the parameters to compute the marginal posterior distribution of X. To discuss the issue of infinite variance of X, he suggested a test statistic F as an inverse function of data. F is also an increasing function of mean estimate of β_2. He showed that the bias in the estimation of X is a decreasing function of F. Thus, for greater accuracy of estimation the data should be very informative. He strongly suggested to use a proper-informative prior in case of weak data. Hunter & Lamboy (1981) came up with a complicated function form for the posterior distribution for the mean of X of simple linear regression model, ξ; $\mathrm{X} = \xi + \delta$, with δ a small error, assuming a bivariate normal prior on β_1 and β_2. They advocated that with the Bayesian approach of inference, one should not worry much about infinite variance. Infinite variance is irrelevant for an appropriate model if all the information about the unknown is inherent in the posterior. But it is also true that infinite variance increases uncertainty in the estimation to infinity. They further commented that the variance of X will become large with over-dispersion in σ^2. They did not comment on the role of β estimate in infinite variance of X. Besides these works, in best of my knowledge there is no previous work on inverse regression in the field of non-conjugate-exponential models to compare with the VB results on the inverse inference found in the chapter.

Contrary to these previous attempts mentioned, the VB method provides tractable approximations to intractable and computationally intensive posterior distributions of the unknowns. It presents a simple understanding of Bayesian inference for both, inverse non-latent conjugate-exponential and non-latent non-conjugate-exponential regression models.

Chapter 5

VB approximation for Inverse Latent Regression

This chapter describes the VB approximations for inverse latent regression problems. In the previous chapter, the intractability issue of the method is discussed for non-conjugate-exponential inverse non-latent regression models. This chapter proceeds with the same intractability issue of the method and describes how to deal with the problem for a successful VB approximation to multi-dimensional, complex inverse latent regression problems. Three models are considered to explain the problem: Poisson latent regression, Poisson latent regression with random effects, zero-inflated Poisson regression with random effects regression.

5.1 Introduction

In the previous chapter, non-latent regression models are discussed briefly and inverse prediction on unknown non-latent explanatory variables is explained. In most of the real life examples, the interest lies in some unobservable objects or latent variables for which we have indirect observations. A latent regression model describes a functional relationship between non-latent variable(s) and its indirect output variable(s) through a set of latent variables:

$$\mathbf{Y} = f(\mathbf{Z}(\mathbf{X}); \theta). \tag{5.1}$$

Eq. 5.1 is similar to Eq. 4.1. The only difference is that the model deals with latent variables denoted by $\mathbf{Z}$. The model f defines a statistical relation between $\mathbf{Y}$ and $\mathbf{X}$ through $\mathbf{Z}$:

$$\mathbf{Y} \sim P(\mathbf{Y}|\mathbf{Z}(\mathbf{X}), \theta) \tag{5.2}$$

If $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ are N i.i.d observations of $\mathbf{Y}$ corresponding to N values $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of $\mathbf{X}$:

$$P(\mathbf{Y}|\mathbf{Z}(\mathbf{X}), \theta) = \prod_{i=1}^{N} P(\mathbf{y}_i|\mathbf{Z}(\mathbf{x}_i), \theta) \tag{5.3}$$

Inverse latent regression can be described as a method of prediction of an unknown explanatory variable $\mathrm{X}_{\mathrm{new}}$ for some new observations of $\mathbf{Y}$, $\mathbf{y}_{\mathrm{new}}$, through the knowledge of the latent variables $\mathbf{Z}$ (and the parameters θ). A Bayesian analysis therefore computes:

$$\begin{aligned} P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{new}, \mathbf{y}, \mathbf{x}) &= \int P(\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \theta|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) \mathrm{d}\mathbf{Z}\mathrm{d}\theta, & (5.4) \\ &\propto \int P(\mathbf{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \theta) P(\mathrm{X}_{\mathrm{new}}|\mathbf{Z}, \theta) \\ &\qquad \times P(\mathbf{Z}, \theta|\mathbf{y}, \mathbf{x}) \mathrm{d}\mathbf{Z}\mathrm{d}\theta. & (5.5) \end{aligned}$$

Assuming that $\mathbf{X}_{new}$ is independent of $\mathbf{Z}$ a priori, this becomes:

$$P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) \propto \int P(\mathbf{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \theta) P(\mathbf{Z}, \theta|\mathbf{y}, \mathbf{x}) P(\mathrm{X}_{\mathrm{new}}|\theta) \mathrm{d}\mathbf{Z}\mathrm{d}\theta. \tag{5.6}$$

We emphasize that the explanatory variable $\mathbf{x}$ indexes the latent variable $\mathbf{Z}$. We consider $\mathbf{Z}$ defined on a discrete grid, hence $\mathbf{x}$ is also discrete. One might consider a continuous version $\mathbf{Z}(\mathbf{x})$ with a real valued $\mathbf{x}$ but that is not pursued here.

There seems to be no literature on the use of the VB method for inverse latent regression problems. Salter-Townshend (2009) has used INLA for the inverse estimation for Poisson latent and zero-inflated Poisson latent model. The author has shown that a good approximation from INLA depends on the structure of the model and the availability of data. In the next section, we explore the VB approximation for inverse latent regression and compared te result from INLA.

5.1.1 Models to explain the Inverse Latent regression problem

Three models categorized in two classes: **latent non-random effect models** and **latent random effect models**. A latent random effect model represents the complex latent models for which the VB approximation to the inverse estimation problem could be challenged. The idea of considering the random effect model is to explain the complex model in a simpler form with random effects capturing extra variation or over dispersion in the data. A latent non-random effect model is considered to build up simple understanding of the VB approximation.

Inverse Poisson latent regression model:

A one dimensional Poisson latent regression problem is considered. To explain the intractability issue of VB approximation for inverse estimation for the latent non-conjugate exponential regression models, the Poisson latent regression model presents the simplest example to start with. The Poisson latent model (likelihood) is defined as below:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{Z}, \beta_0, \beta_1) = \prod_{j=1}^{n} \frac{\lambda_j^{\mathrm{y}_j} e^{-\lambda_j}}{\mathrm{y}_j!}, \tag{5.7}$$

$$\lambda_j = \exp(\beta_0 + \beta_1 \mathbf{Z}(\mathrm{x}_j)), \tag{5.8}$$

where $\{\mathrm{y}_j;\ j = 1 : n\}$ are n observations on the one-dimensional response variable $\mathbf{Y}$, β_0 and β_1 are the regression parameters.

The latent variable $\mathbf{Z}$ is the one-dimensional latent variable defined on a grid of size p, $\mathrm{x}_j;\ j = 1 : n$ are n (discrete) values on the explanatory variable $\mathbf{X}$ representing the grid locations of the latent variable $\mathbf{Z}$. Therefore,

$$\mathrm{x}_j \in \{1, \ldots, p\};\ \forall j.$$

The model is shown in Fig. 5.1 via a DAG. In the DAG, shown are the following variables:

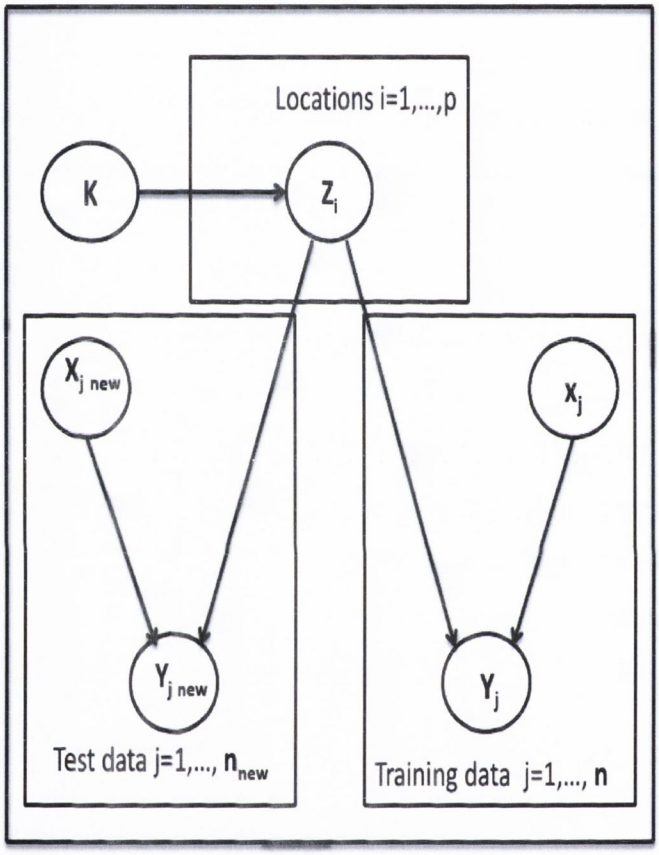

Fig. 5.1: The Poisson latent regression model represented as a DAG.

p : size of the discrete grid,

$\mathbf{Z}_i$: latent variable at grid location $i = 1, \ldots, p$,

κ : hyper-parameter in the prior over $\mathbf{Z}$,

x_j : j^{th} observation/training data point on the explanatory variable X $\forall j = 1 : n$,

Y_j : j^{th} observation/training data point on the response variable Y $\forall j = 1 : n$,

$\mathrm{Y}_{j\,\mathrm{new}}$: j^{th} observation/test data point on the response variable Y $\forall j = 1 : n_{\mathrm{new}}$,

$\mathrm{X}_{j\,\mathrm{new}}$: Unknown explanatory variable corresponding to the new observation $\mathrm{Y}_{j\,\mathrm{new}}$.

It is assumed that given $\mathbf{Z}(\mathrm{x}_j)$, the observations y_j's are independent of each other. This assumption enables us to estimate $\mathbf{Z}$ at the forward stage considering training data Y_j's only and then infer $\mathrm{X}_{j\,\mathrm{new}}$'s at the inverse stage given the test data $\mathrm{Y}_{j\,\mathrm{new}}$.

The regression parameters are assumed to be known ($\beta_0 = 0$, $\beta_1 = 1$) to show the effect of the latent variable only on the inverse estimation of $\mathrm{X}_{\mathrm{new}}$. For the next two models described below, the regression parameters will be assumed to known as given before.

The latent random effect model is described through two models: Poisson latent random effects model and zero-inflated random effects model. Given random effects, the model factorizes over multi-dimensional response variable and hence the model can be described in a simple form which also makes the problem feasible for a tractable VB approximation.

1. **Inverse Poisson latent random effect regression model:**

   A Poisson latent random effect model provides a simple understanding of more complex latent models such as a zero-inflated Poisson latent random effect model. A three dimensional Poisson latent random effect model is considered. The likelihood of the model is given below:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{Z}, \mathbf{U}) = \prod_{k=1}^{K} \prod_{j=1}^{n} \frac{\lambda_{kj}^{\mathrm{y}_{kj}} e^{-\lambda_{kj}}}{\mathrm{y}_{kj}!}, \tag{5.9}$$

$$\lambda_{kj} = \exp[\mathbf{Z}_k(\mathrm{x}_j) + \mathrm{U}_{kj}], \tag{5.10}$$

   where K is the dimension of the response variable $\mathbf{Y}$, y_{kj} is the j^{th} observation

on the k^{th} response variable $\mathbf{Y}_k$, $\mathbf{U}_j = [\mathrm{U}_{kj};\ k = 1 : K]^T$ is the K-dimensional j^{th} random effect corresponding to the j^{th} observation on the response variable $\mathbf{Y}_j$ for all $j = 1 : n$. The term $\mathbf{Z}_k$ is the k^{th} latent variable corresponding to the k^{th} response variable $\mathbf{Y}_k$ for all $k = 1 : K$. The latent variables $\mathbf{Z}_k;\ k = 1 : K$ are defined on a grid of size p represented by the discrete one dimensional explanatory variable $\mathbf{X}$.

The model is explained through a DAG in Fig. 5.2. The variables shown in the DAG are defined as below:

p : size of the discrete grid,

x_j : j^{th} observation/training data on the explanatory variable X $\forall j = 1 :$

$\mathbf{Z}_{ik}$: latent variable corresponding to k^{th} response variable $\mathbf{Y}_k$ at grid location $i = 1 : p$,

κ_k : hyper-parameter in the prior over $\mathbf{Z}_k$,

$\mathbf{Q}_\mathrm{U}$: precision parameter in the prior over $\mathbf{U}_j$ $\forall j = 1 : n$ and $\mathbf{U}_{j\ \mathrm{new}}$ $\forall j = 1 : n_{\mathrm{new}}$,

$\mathrm{Y}_{kj\ \mathrm{new}}$: j^{th} observation/test data on the response variable Y_k $\forall j = 1, \ldots, n_{\mathrm{ne}}$

$\mathrm{X}_{j\ \mathrm{new}}$: Unknown explanatory variable corresponding to the new observation dimensional response variable, $\mathrm{Y}_{j\ \mathrm{new}}$.

It is assumed that given $\mathbf{Z}$ and $\mathbf{U}$, response variables $\mathbf{Y}$ are independent of each other (within dimension and across locations). Hence, we can consider test data and training data set separately to estimate $\mathbf{Z}$ and $\mathbf{U}$ at the forward stage and then infer $\mathbf{X}_{\mathrm{new}}$ at the inverse stage.

2. **Inverse zero-inflated Poisson latent with random effects model:**
A zero-inflated Poisson latent random effect model represent the complex latent models to explain the challenges of the VB method for the inverse estimation for the complex latent models. A zero-inflated Poisson models count data with excess of zeros (Ridout et al., 1998). The likelihood of the model is

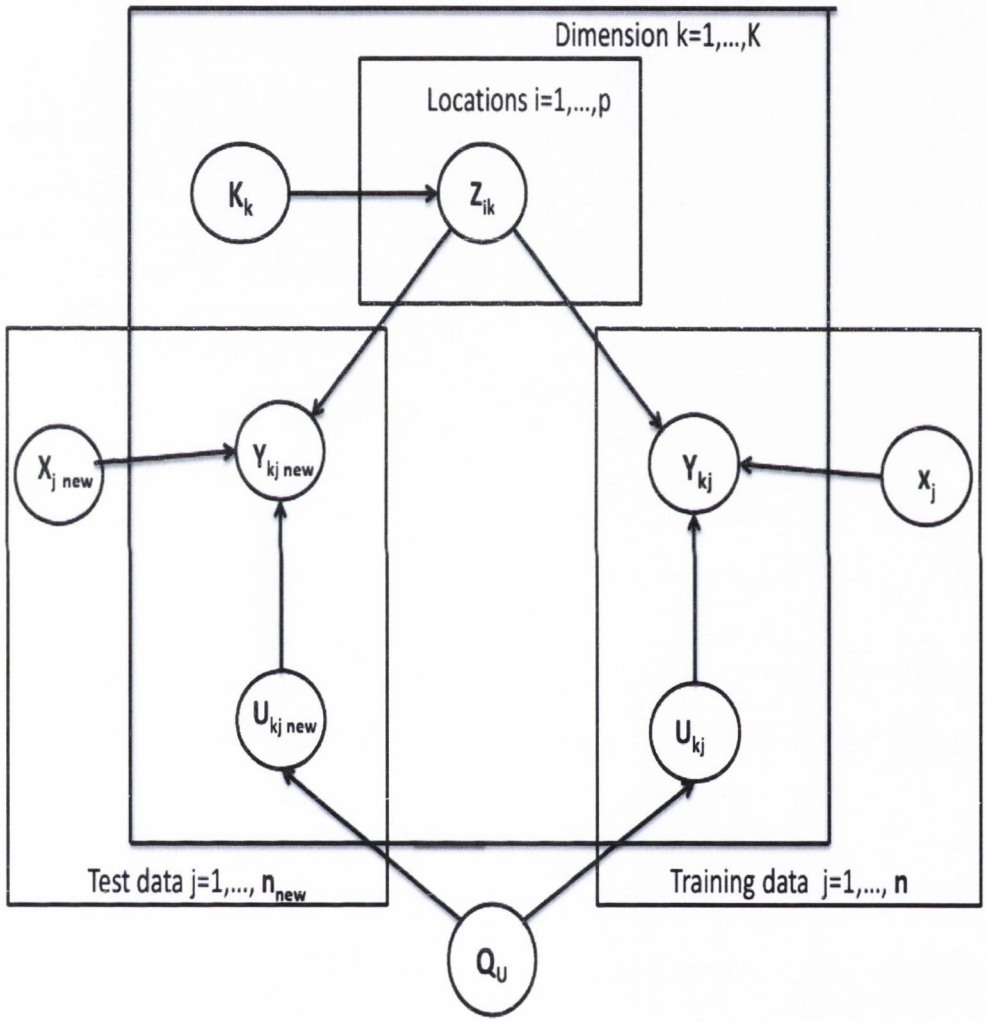

Fig. 5.2: The Poisson latent random regression model represented as a DAG.

given below:

$$P(\mathbf{y}|\mathbf{Z},\mathbf{U},\mathbf{x},\alpha) = \prod_{k=1}^{K}\prod_{j=1}^{n} ZIP(\mathrm{y}_{kj};\ \lambda_{jk}, q_{kj}), \tag{5.11}$$

$$\mathrm{ZIP}(\mathrm{y}_{kj};\ \lambda_{jk}, q_{kj}) = \begin{cases} 1 - q_{kj} + q_{kj}\, e^{-\lambda_{kj}}, & \text{if } \mathrm{y}_{kj} = 0; \\ q_{kj}\mathrm{Poiss}(\mathrm{y}_{kj};\ \lambda_{kj}), & \text{if } \mathrm{y}_{kj} > 0 \end{cases}$$

The term $(1 - q_{kj})$ is the probability of observing essential zero counts. Salter-Townshend (2009) used a power law functional relationship to define the probability q_{kj} as in terms of λ_{jk}:

$$q_{kj} = \left(\frac{\lambda_{kj}}{1+\lambda_{kj}}\right)^{\alpha_k}, \ \forall k,\ j, \tag{5.12}$$

$$\lambda_{kj} = \exp(\mathrm{Z}_k(\mathrm{x}_j) + \mathrm{U}_{kj}),\ \forall k,\ j \tag{5.13}$$

where λ_{kj} is mean of the Poisson term in the likelihood. The term $\mathrm{Z}_k(\mathrm{x}_j)$ represents the k^{th} latent variable of indexed by the j^{the} discrete value of the explanatory variable x_j. The term U_{kj} is the random effect corresponding to the j^{th} observation on the k^{th} response variable y_{kj} that induces dependence between the multi-dimensional response variable and captures the extra variation in the data.

The power index $\alpha_k\ \forall k$ in the ZI-Poisson likelihood takes values from 0 to ∞ such that a big value should induce many zero counts. It should be noted that both the essential zero probability and Poisson mean are defined in terms of λ_{kj} rather than separately. This allows Salter-Townshend (2009) to use INLA to implement Bayesian inference for a similar model.

The pictorial representation of the model is given through a DAG in Fig. 5.3. The variables shown in the DAG (shown in Fig. 5.3) have similar interpretation as given for the Poisson latent random effect model. In the ZI-Poisson latent random effect model, α_k is the extra parameter.

VB approximation for the inverse estimation for the these models are discussed in

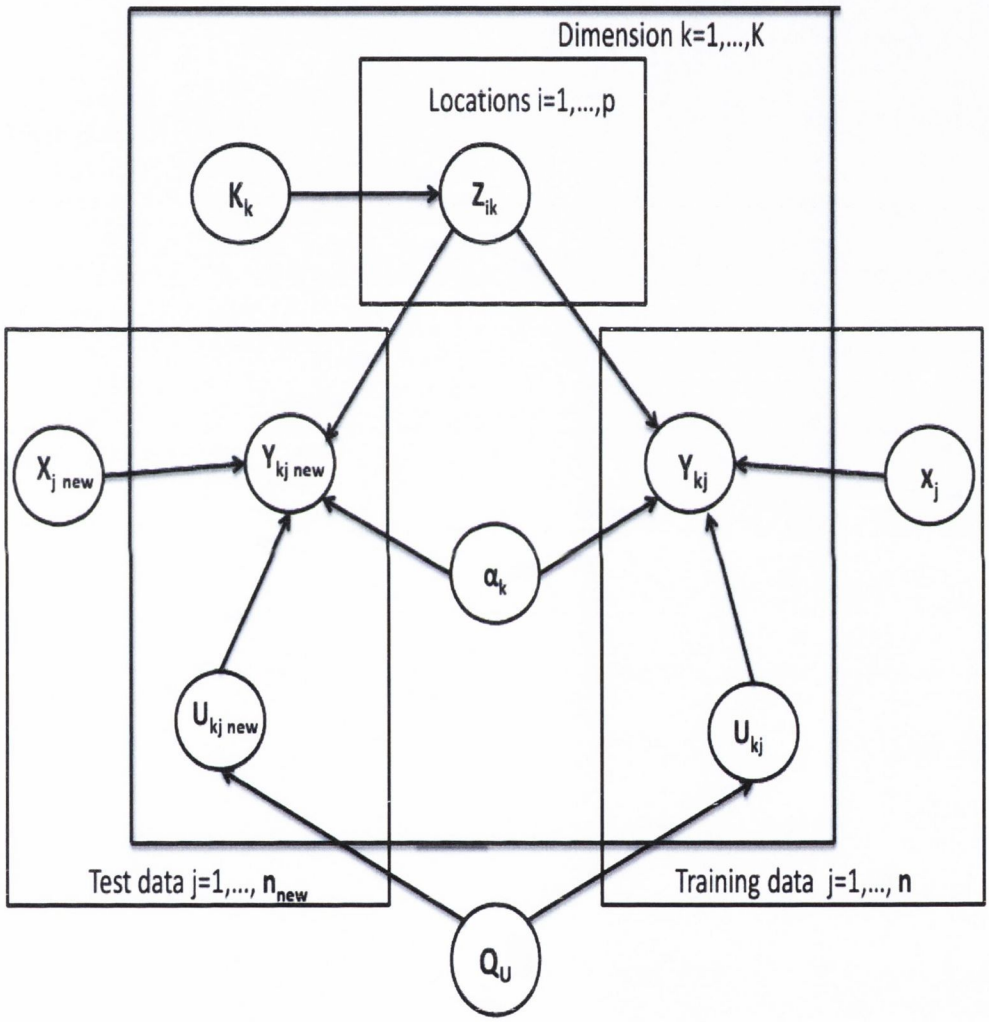

Fig. 5.3: The Poisson latent random regression model represented as a DAG.

the next sections.

5.2 Inference procedure for the Inverse Latent Regression Problem (Poisson Latent regression model)

The prior distribution on $\mathbf{Z}$ is assumed to be a GMRF with mean vector zero and precision $Q_{\mathbf{Z}}$:

$$P(\mathbf{Z}) = GMRF_p(\mathbf{Z}; \underline{0}, Q_{\mathbf{Z}}^{-1}), \tag{5.14}$$

$$Q_{\mathbf{Z}} = \kappa R_{p\times p}. \tag{5.15}$$

The term κ is an unknown smoothing parameter in the precision and the matrix R s defined as

$$R = \begin{pmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix},$$

The precision $Q_{\mathbf{Z}}$ is structured such that the GMRF density of $\mathbf{Z}$ is proper.

The prior distribution over κ is assumed as a Gamma distribution:

$$P(\kappa) = Gamma(\kappa; a, b). \tag{5.16}$$

The hyper-parameters a and b are assumed to be known.

5.2.1 Bayesian Analysis of the inverse latent regression problem

The Bayesian analysis of the estimation problem at the two stages of the inference, the forward stage and the inverse stage, is given below:

Forward stage:

At the forward stage of inference the latent variable $\mathbf{Z}$ and the unknown parameter $\theta = \kappa$ is inferred through their posterior distribution given the data $(\mathbf{y}, \mathbf{x})$. The joint posterior distribution of $\mathbf{Z}$ and $\theta = \kappa$ given $(\mathbf{y}, \mathbf{x})$ is defined by Bayes' law as:

$$P(\mathbf{Z}, \theta|\mathbf{y}, \mathbf{x}) = \frac{P(\mathbf{y}|\mathbf{x}, \mathbf{Z}, \theta)P(\mathbf{Z}, \theta)}{\int_{\mathbf{Z},\theta} P(\mathbf{y}|\mathbf{x}, \mathbf{Z}, \theta)P(\mathbf{Z}, \theta)\mathrm{d}\mathbf{Z}\mathrm{d}\theta}, \tag{5.17}$$

where $P(\mathbf{y}|\mathbf{x}, \mathbf{Z}, \theta)$ is the likelihood of $\mathbf{Z}$ and θ given $\mathbf{y}$ and $\mathbf{x}$. The term $P(\mathbf{Z}, \theta) = P(\mathbf{Z}|\theta)P(\theta)$ is the prior distribution assumed over $\mathbf{Z}$ and θ.

Inverse stage: At the inverse stage of inference the knowledge of $\mathbf{Z}$ and θ from the forward stage is used to predict the unknown explanatory variable $\mathrm{X_{new}}$ of a latent regression given data $(\mathrm{y_{new}}, \mathbf{y}, \mathbf{x})$. The posterior distribution of $\mathrm{X_{new}}$ given $(\mathrm{y_{new}}, \mathbf{y}, \mathbf{x})$ is given in Eq. 5.6. The prior distribution assumed over $\mathrm{X_{new}}$ is an improper prior:

$$P(\mathrm{X_{new}}) \propto 1; \mathrm{X_{new}} \in \{1, \ldots, p\}. \tag{5.18}$$

5.2.2 VB approximation to the inference problem

The VB approximation to the inference problem is described at the two stages separately:

VB approximation at the forward stage

$$P(\mathbf{Z}, \theta|\mathbf{y}, \mathbf{x}) \approx q(\mathbf{Z}, \theta|\mathbf{y}, \mathbf{x}), \tag{5.19}$$

$$q(\mathbf{Z}, \theta|\mathbf{y}, \mathbf{x}) = q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, \mathbf{x})q_{\theta}(\theta|\mathbf{y}, \mathbf{x}), \tag{5.20}$$

The VB marginal $q_\kappa(\kappa|\mathbf{y},\mathbf{x})$ is obtained as:

$$\begin{aligned} q_\kappa(\kappa|\mathbf{y},\mathbf{x}) &\propto \exp\left[\mathbb{E}_{q_\mathbf{Z}(\mathbf{Z})}\log P(\mathbf{y},\mathbf{Z},\kappa|\mathbf{x})\right], && (5.21)\\ &\propto \exp\left[\log P(\kappa)+\mathbb{E}_{q_\mathbf{Z}(\mathbf{Z})}\log P(\mathbf{Z}|\kappa)\right], && (5.22)\\ &= Gamma(a^\star,b^\star), && (5.23)\\ a^\star &= a+0.5p, && (5.24)\\ b^\star &= b+0.5\left[2\sum_{i=1}^{p}\mathbb{E}_q(\mathrm{Z}_i^2)-\mathbb{E}_q(\mathrm{Z}_p^2)-2\sum_{i\neq m}^{p}\mathbb{E}_q(\mathrm{Z}_i\mathrm{Z}_m)\right]. && (5.25)\end{aligned}$$

The VB marginal $q_\mathbf{Z}(\mathbf{Z}|\mathbf{y},\mathbf{x})$ is given as follows:

$$\begin{aligned} q_\mathbf{Z}(\mathbf{Z}|\mathbf{y},\mathbf{x}) &\propto \exp\left[\mathbb{E}_{q_\kappa(\kappa)}\log P(\mathbf{y},\mathbf{Z},\kappa|\mathbf{x})\right], && (5.26)\\ &\propto \exp\left[\mathbb{E}_{q_\kappa(\kappa)}\log\left(P(\mathbf{y}|\mathbf{Z},\mathbf{x})P(\mathbf{Z}|\kappa)\right)\right], && (5.27)\\ &\propto \exp\left[\log P(\mathbf{y}|\mathbf{Z},\mathbf{x})+\mathbb{E}_{q_\kappa(\kappa)}\log P(\mathbf{Z}|\kappa)\right]. && (5.28)\end{aligned}$$

The VB marginal $q_\mathbf{Z}(\mathbf{Z}|\mathbf{y},\mathbf{x})$ is not a standard distribution. As $\mathbf{Z}$ is a high dimensional vector, the computation of the VB marginal by a numerical integration would be not be possible. It is needed as a known standard distribution at the inverse stage of inference for a tractable inverse prediction of the explanatory variables $\mathbf{X}_{\text{new}}$ given a set of new counts $\mathbf{y}_{\text{new}}$. A Gaussian approximation (as considered to approximate intractable VB marginals in the previous chapter for the non-conjugate-exponential non-latent problem) could be applied here to approximate $q_\mathbf{Z}(\mathbf{Z}|\mathbf{y},\mathbf{x})$ into a standard normal distribution.

Gaussian approximation of $q_\mathbf{Z}(\mathbf{Z}|\mathbf{y},\mathbf{x})$:

The likelihood $P(\mathbf{y}|\mathbf{Z},\mathbf{x})$ is an exponential function of $\mathbf{Z}$ but its prior distribution $P(\mathbf{Z}|\kappa)$ is not conjugate to the likelihood. Therefore for a standard approximation of $q_\mathbf{Z}(\mathbf{Z}|\mathbf{y},\mathbf{x})$, a possible suggestion is to approximate $\log P(\mathbf{y}|\mathbf{Z},\mathbf{x})$ as a quadratic function of $\mathbf{Z}$ that also makes the $P(\mathbf{Z}|\kappa)$ conjugate to the $P(\mathbf{y}|\mathbf{Z},\mathbf{x})$, hence giving a tractable VB approximation.

Only the term $\exp(\mathbf{Z}(\mathrm{x}_j))$; $\forall j$ in the log-likelihood $\log P(\mathbf{y}|\mathbf{Z},\mathbf{x})$ is not quadratic (or linear) in $\mathbf{Z}$. A second order Taylor's expansion of $\exp(\mathbf{Z}(\mathrm{x}_j))$; $\forall j$ is considered

around $\mathbf{Z}(\mathrm{x}_j) = \mathbf{Z}^m(\mathrm{x}_j)$, where $\mathbf{Z}^m(\mathrm{x}_j)$ is the posterior mode of $\mathbf{Z}$. A Gaussian approximation is:

$$q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y},\mathbf{x}) \approx q^g_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y},\mathbf{x}), \tag{5.29}$$

$$q^g_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y},\mathbf{x}) = N_p(\mu^\star_{\mathbf{Z}}, {Q^\star_{\mathbf{Z}}}^{-1}), \tag{5.30}$$

where,

$$Q^\star_{\mathbf{Z}} = \mathbb{E}_q(\kappa)R + diag(V_{\mathbf{Z}}), \tag{5.31}$$

$$\mu^\star_{\mathbf{Z}} = B^T_{\mathbf{Z}} {Q^\star_{\mathbf{Z}}}^{-1}, \tag{5.32}$$

$$B_{\mathbf{Z}} = [A_{\mathrm{Z}_i};\ i = 1:p] + \mathbf{Z}^m V_{\mathbf{Z}}, \tag{5.33}$$

$$V_{\mathbf{Z}} = diag[V_{\mathrm{Z}_i}\ i = 1:p], \tag{5.34}$$

$$V_{\mathrm{Z}_i} = -\frac{\partial^2}{\partial \mathrm{Z}_i^2} \sum_{\substack{j=1 \\ \mathrm{x}_j = i}}^{n} \log P(\mathrm{y}_j|\mathbf{Z}(\mathrm{x}_j))\Big|_{\mathrm{Z}_i = \mathrm{Z}_i^m}$$

$$= \sum_{\substack{j=1 \\ \mathrm{x}_j = i}}^{n} \exp(\mathbf{Z}^m(\mathrm{x}_j)), \tag{5.35}$$

$$A_{\mathrm{Z}_i} = \frac{\partial}{\partial \mathrm{Z}_i} \sum_{\substack{j=1 \\ \mathrm{x}_j = i}}^{n} \log P(\mathrm{y}_j|\mathbf{Z}(\mathrm{x}_j))\Big|_{\mathrm{Z}_i = \mathrm{Z}_i^m}$$

$$= \sum_{\substack{j=1 \\ \mathrm{x}_j = i}}^{n} \left[\exp(\mathbf{Z}^m(\mathrm{x}_j)) - \mathrm{y}_j\right], \tag{5.36}$$

where $\mathbf{Z}^m = [\mathrm{Z}_i^m;\ i = 1:p]^T$ is the posterior mode of $\mathbf{Z}$ to be found. Gradient descent methods for finding mode are usually time-consuming but Rue & Held (2005) give an iterative algorithm:

1. set the initial value of the mode $\mathbf{Z}^m$,
2. find the values of $B^T_{\mathbf{Z}}$ and $V_{\mathbf{Z}}$ depending on $\mathbf{Z}^m$,
3. compute $Q^\star_{\mathbf{Z}}$ and $\mu^\star_{\mathbf{Z}}$ as defined in the Eq. 5.31 and 5.32 respectively,
4. since, the mode of a Gaussian distribution is equal to its mean, find a new

value of the mode $\mathbf{Z}^m$ as:

$$\mathbf{Z}^m = \mu^{\star}_{\mathbf{Z}},$$

5. repeat the step 2 to 4 until the convergence.

To save computational time, the modes of the Gaussian approximations are computed only once at each VB iteration.

If the posterior distribution of $\mathbf{Z}$ is multi-modal, the global mode of the posterior distribution should be considered for a Gaussian approximation. However, the latent models we have considered in the chapter lead to a uni-modal posterior distribution of $\mathbf{Z}$. Therefore, the multi-modality situation can be avoided.

VB approximation at the inverse stage:

The VB approximation of the posterior distribution of $\mathrm{X}_{\mathrm{new}}$ at the inverse stage is given as:

$$\begin{aligned}
P(\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \theta|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) &\approx q(\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \theta|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}), && (5.37)\\
q(\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \theta|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) &= q_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) q_{\mathbf{Z}}(\mathbf{Z}|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) \\
&\quad \times q_{\theta}(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}), && (5.38)\\
&\approx q_{\mathrm{X}_{\mathrm{new}}}(\mathbf{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{x}, \mathbf{y}) q_{\theta}(\theta|\mathbf{x}, \mathbf{y}). && (5.39)
\end{aligned}$$

The VB marginal of $\mathrm{X}_{\mathrm{new}}$ $q_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y})$ is restricted VB approximation as the VB marginals $q_{\mathbf{Z}}(\mathbf{Z}|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y})$ and $q_{\theta}(\theta|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y})$ are restricted to a known distribution. The restricted VB marginal $q_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y})$ is given as:

$$\begin{aligned}
\log q_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) &\approx \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{x},\mathbf{y}) q_{\theta}(\theta|\mathbf{x},\mathbf{y})} \Bigg[\log \Big\{ P(\mathrm{y}_{\mathrm{new}}|\mathbf{Z}(\mathrm{X}_{\mathrm{new}})) \\
&\quad \times P(\mathbf{Z}|\mathrm{X}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) P(\mathrm{X}_{\mathrm{new}}) \Big\} \Bigg], && (5.40)\\
\log q(\mathrm{X}_{\mathrm{new}} = i|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) &\approx -\mathbb{E}_q \left[\exp(\mathbf{Z}(\mathrm{X}_{\mathrm{new}}))\right] + \mathrm{y}_{\mathrm{new}} \mathbb{E}_q(\mathbf{Z}(\mathrm{X}_{\mathrm{new}})) - \log(\mathrm{y}_{\mathrm{new}}!) \\
&\quad -0.5\, diag\left[Var_{\mathbf{Z}}\right]^T diag\left[Q^{\star}_{\mathbf{Z}}\right] \\
&\quad + \sum_{i \neq j} Var_{\mathbf{Z}}(i,j) Q^{\star}_{\mathbf{Z}}(i,j). && (5.41)
\end{aligned}$$

where $Var_{\mathbf{Z}}$ is the posterior variance of $\mathbf{Z}$. The terms $Var_{\mathbf{Z}}(i,j)$ and $Q_{\mathbf{Z}}(i,j)$ are the $(i,j)^{th}$ entries of the posterior variance and posterior precision of $\mathbf{Z}$ respectively.

5.2.3 Comparison of the VB approximation with approximation by INLA

The VB results must be compared with the results from other methods for its accuracy. The Poisson latent model is high-dimensional in the latent variable $\mathbf{Z}$. Therefore, a numerical integration method for the true result would be very expensive. The MCMC method for the estimation problem for the model will be slow and have issues with convergence and mixing. INLA is an alternative approach that has growing popularity for latent models for its accuracy and quickness. The VB method is simple to apply and a quick method. Therefore, the VB approximations for the Poisson latent model are compared with the approximation by INLA. The direct VB approximation is compared with the results from INLA at the forward stage of inference. Whereas at the inverse stage of inverse, compared are the restricted VB result and the results from the Laplace approximation, which uses results from INLA, obtained at the forward stage are used.

Another accuracy check for inverse estimation To check the accuracy of the approximation at the inverse stage, a 95% HPD region is computed. The percentage of true values of new explanatory variable that lie within the 95% HPD region gives an accuracy measure of the approximation to inverse inference. Salter-Townshend (2009) described a method to compute HPD bounds for discrete distributions that we use here.

5.2.4 Result

An example with simulated data from the Poisson latent model is used. A set of one hundred values, from one to fifty, on the explanatory variable are drawn from a discrete uniform distribution. Fifty values of the latent variable (given over a grid defined by the explanatory variable) are generated from a GMRF distribution with a

Gamma variate precision parameter. Corresponding to the true values of the latent variable and the data on the explanatory variable, a set of one hundred counts on the response variable are generated from a Poisson latent model. The data on the response variable and the explanatory variable are used for the estimation of the latent variable at the forward stage.

A set of one hundred values on the explanatory variable and the response variables are generated (as mentioned above) for the inverse estimation of the unknown explanatory variable. The true values of explanatory variable are used to check the accuracy of the approximation of the inverse estimation via a cross validation technique described with a 95% HPD region.

Results of the VB approximation and its comparison with the results from INLA for the Poisson latent model, are shown in Fig. 5.4, 5.5 and 5.6. The R-INLA package is used to implement INLA for the model. In Fig. 5.4, The VB means of the latent variable $\mathbf{Z}$ (defined over a grid) are compared with the mean of the approximation by INLA. The figure shows that the VB approximation and the approximation by INLA. They are well fitting the true value as they stay close to the true values of the latent variable at the grid locations. The 95% HPD regions from both approximations (VB and INLA) is narrow at the locations where the data are available in abundance, whereas for less informative or less amount of data the region is large. The approximations are not very smooth as they are expected. The reason behind the non-smoothness of the approximation may be the insufficient data at some of the grid locations.

Fig. 5.5 and 5.6 present the multi-modal density of a new unknown explanatory variable $\mathrm{X}_{\mathrm{new}}$ given data on the response variable $\mathbf{Y}_{\mathrm{new}} = 1$ to show insufficient data and $\mathbf{Y}_{\mathrm{new}} = 124$ to show strong information in data respectively. It can be seen from the figures that the VB approximation of the inverse estimation matches with the result by INLA. Given non-informative data $\mathbf{Y}_{\mathrm{new}} = 1$, the approximations (by VB and INLA) is multi-modal and the coverage of the approximation is large. Whereas, given strong information $\mathbf{Y}_{\mathrm{new}} = 124$ the VB approximation and the approximation by INLA overlap and they estimate the true value quite accurately. The accuracy of the approximation by INLA with 95% HPD region is only 88% and that of by

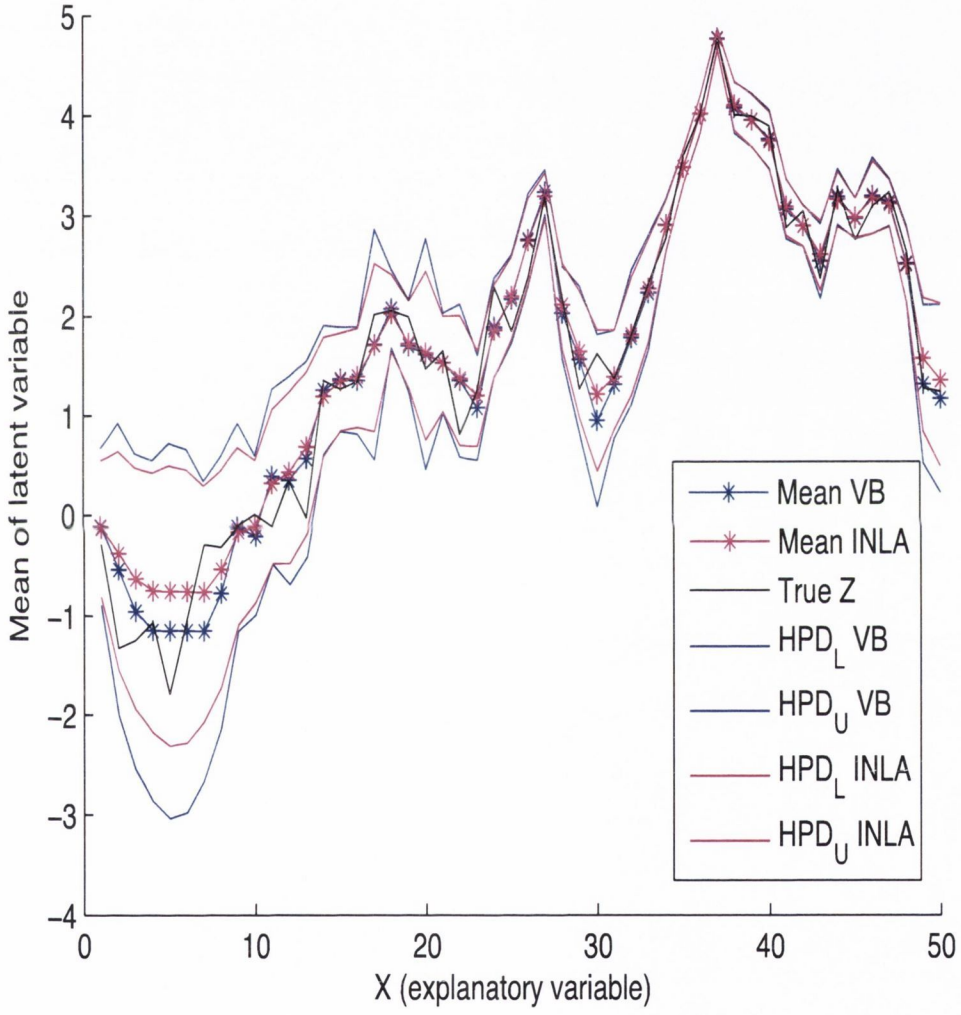

Fig. 5.4: The comparison of true value (black), VB-mean(blue), mean (estimated by INLA) (magenta) of the latent variables $\mathbf{Z}$ corresponding to the response variable $\mathbf{Y}$ for a Poisson latent model.

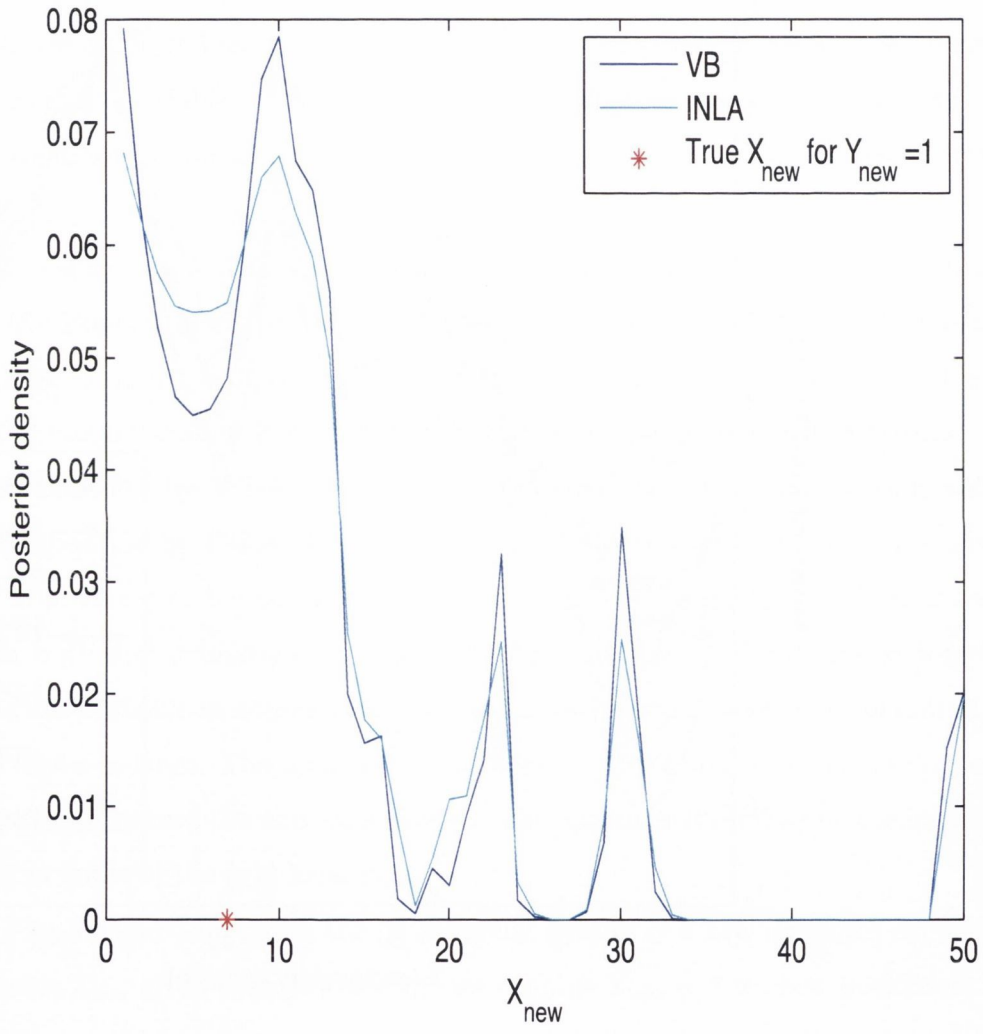

Fig. 5.5: The comparison of the approximation by the VB method and by the INLA of true posterior distribution of X_{new} given $Y_{new} = 1$ for a Poisson latent model.

	Forward stage	Inverse stage
VB	0.143 sec	0.053 sec
INLA	2.05 sec	0.107 sec

Table 5.1: The comparison of the computational time of VB method and INLA at both stages for Poisson latent model. At the inverse stage the computation time shows the time taken by the method for the inverse estimation for all test data.

the VB method is only 84%. Hence it can be concluded that the methods require informative data for a good approximation as the prior distribution assumed over X_{new} is not proper.

Computational time:

The computational time is noted to check if the VB method provides a better trade off between accuracy and computation than INLA. The comparison of the computation time of both methods at both stages are displayed in Table 5.1. It can be concluded that the VB method takes less time in comparison to INLA and provides similar results.

5.3 VB approximation for the inverse latent random effects regression models

The prior distributions over the unknowns of the inverse latent with random effects regression problem are assumed as follows:

1. The prior distribution over the latent variable $\mathbf{Z}$ is assumed to be a GMRF as given below:

$$P(\mathbf{Z}) = \prod_{k}^{K} GMRF_p(\mathbf{Z}_k; \underline{0}, Q_{\mathbf{Z}_k}^{-1}), \tag{5.42}$$

$$Q_{\mathbf{Z}_k} = \kappa_k R_{p\times p}. \tag{5.43}$$

where $\mathbf{Z}$ is K dimensional latent variable. The independent GMRF is assumed over each $\mathbf{Z}_k$; $\forall k$ wit mean vector zero and precision $Q_{\mathbf{Z}_k}$. The term $\kappa = \{\kappa_k; k = 1 : K\}$ is a set of unknown smoothing parameters in the precision

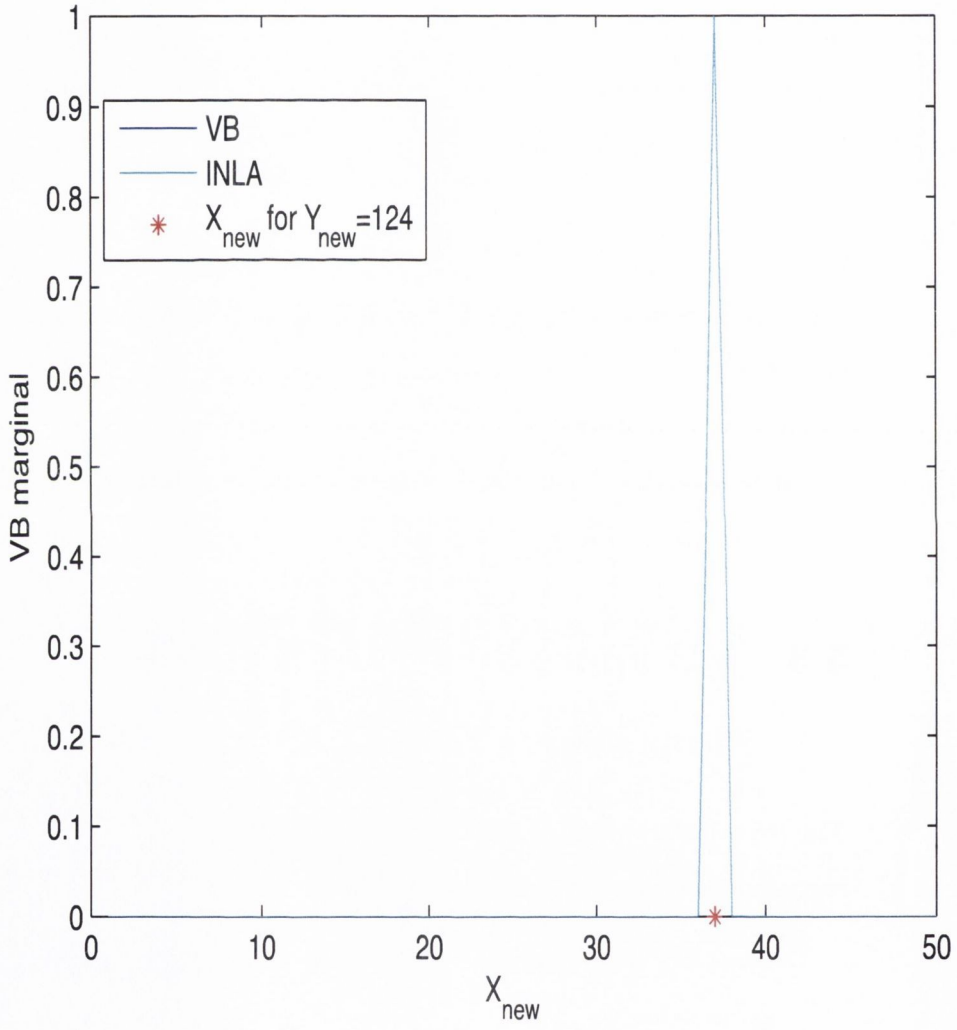

Fig. 5.6: The comparison of the approximation by the VB method and by the INLA of true posterior distribution of X_{new} given $Y_{new} = 124$ for a Poisson latent model.

parameter and the matrix R is same as defined for the Poisson latent regression problem.

2. The prior distribution considered on $\mathbf{U}$ is a multivariate Gaussian given as follows:

$$P(\mathbf{U}) = \prod_{j=1}^{n} MVN_K(\mathbf{U}_j; \underline{0}, Q_{\mathbf{U}}^{-1}), \tag{5.44}$$

where $\underline{0}$ is a zero vector mean and $Q_{\mathbf{U}}$ is unknown precision common to every $\mathbf{U}_j;\ j = 1 : n$.

3. The prior distribution over κ_k is assumed as a Gamma distribution:

$$P(\kappa_k) = Gamma(\kappa_k; a_k, b_k);\ \forall k. \tag{5.45}$$

4. The prior distribution assumed over $Q_{\mathbf{U}}$ is given as:

$$P(Q_{\mathbf{U}}) = Wishart_K(Q_{\mathbf{U}}; SS, df); \tag{5.46}$$

where df is the degree of freedom in the distribution and SS is a $K \times K$ positive-definite matrix.

5. The prior distribution over the unknown new explanatory variable $\mathrm{X}_{\mathrm{new}}$ is assumed as an improper prior:

$$P(\mathrm{X}_{\mathrm{new}}) \propto 1; \mathrm{X}_{\mathrm{new}} \in \{1, \ldots, p\}. \tag{5.47}$$

The hyper-parameters a_k, b_k, df and SS are assumed to be known.

5.3.1 Bayesian analysis of the inference problem

Bayesian estimation of the latent variable, random effects and hyper-parameters and the Bayesian inverse prediction of unknown explanatory variable are discussed at the two stages separately.

Forward stage:

At the forward stage, the posterior distributions of $\mathbf{Z}$, $\mathbf{U}$ and other parameters $\theta = \{\kappa, Q_{\mathbf{U}}\}$, are to be inferred given the data-set $\mathbf{y}, \mathbf{x}$. The joint distribution of $\mathbf{Z}$, $\mathbf{U}$ and θ given the training data set $(\mathbf{y}, \mathbf{x})$, can be defined by Bayes' law as:

$$P(\mathbf{Z}, \mathbf{U}, \theta|\mathbf{y}, \mathbf{x}) = \frac{P(\mathbf{Z}, \mathbf{U}, \theta|\mathbf{y}, \mathbf{x})P(\mathbf{Z}|\mathbf{x}, \theta)P(\mathbf{U}|\theta)P(\theta)}{\int_{\mathbf{Z},\mathbf{U},\theta} P(\mathbf{Z}, \mathbf{U}, \theta|\mathbf{y}, \mathbf{x})P(\mathbf{Z}|\mathbf{x}, \theta)P(\mathbf{U}|\theta)P(\theta)\mathrm{d}\mathbf{Z}\mathrm{d}\mathbf{U}\mathrm{d}\theta}. \tag{5.48}$$

In the above expression for the joint posterior distribution, the term $P(\mathbf{Z}, \mathbf{U}, \theta|\mathbf{y}, \mathbf{x})$ is the likelihood of unknowns given data:

$$P(\mathbf{Z}, \mathbf{U}, \theta|\mathbf{y}, \mathbf{x}) = \prod_{k=1}^{K}\prod_{j=1}^{n} P(\mathrm{y}_{kj}|\mathrm{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj}, \theta). \tag{5.49}$$

For the prediction of $\mathrm{X}_{\mathrm{new}}$ at the inverse stage, the interest lies in the computation of the marginal posterior distribution over the unknowns that can be computed as:

$$P(\mathbf{Z}|\mathbf{y}, \mathbf{x}) = \int_{\mathbf{U},\theta} P(\mathbf{Z}, \mathbf{U}, \theta|\mathbf{y}, \mathbf{x})\mathrm{d}\mathbf{U}\mathrm{d}\theta, \tag{5.50}$$

$$P(\theta|\mathbf{y}, \mathbf{x}) = \int_{\mathbf{U},\mathbf{Z}} P(\mathbf{Z}, \mathbf{U}, \theta|\mathbf{y}, \mathbf{x})\mathrm{d}\mathbf{U}\mathrm{d}\mathbf{Z}. \tag{5.51}$$

The marginal posterior distribution over the components of θ can further be computed by integrating out other unknowns from the joint posterior distribution.

Inverse stage:

The aim of the inverse stage is to infer $\mathrm{X}_{\mathrm{new}}$ corresponding to the new observation on the response variable $\mathbf{y}_{\mathrm{new}}$, having learnt about the model at the forward stage. The posterior distribution of $\mathrm{X}_{\mathrm{new}}$ can be obtained by integrating all the other unknowns

from the joint posterior distribution given the data:

$$
\begin{aligned}
P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{y}_{\mathrm{new}},\mathbf{x}) &= \int P(\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta|\mathbf{y}_{\mathrm{new}},\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta, \\
&= \int \frac{P(\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta,\mathbf{y}_{\mathrm{new}}|\mathbf{y},\mathbf{x})}{P(\mathbf{y}_{\mathrm{new}}|\mathbf{y},\mathbf{x})}\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta, \\
&\propto \int P(\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta,\mathbf{y}_{\mathrm{new}}|\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta, \\
&= \int P(\mathbf{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta)P(\mathbf{U}_{\mathrm{new}}|\theta)P(\mathbf{Z},\theta|\mathrm{X}_{\mathrm{new}},\mathbf{y},\mathbf{x}) \\
&\qquad \times P(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta, \\
&\approx \int P(\mathbf{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta)P(\mathbf{U}_{\mathrm{new}}|\theta)P(\mathbf{Z},\theta|\mathrm{X}_{\mathrm{new}},\mathbf{y},\mathbf{x}) \\
&\qquad \times P(\mathrm{X}_{\mathrm{new}})\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta, \\
&= \int P(\mathbf{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta)P(\mathbf{U}_{\mathrm{new}}|\theta)P(\mathbf{Z}|\mathrm{X}_{\mathrm{new}},\mathbf{y},\mathbf{x},\theta) \\
&\qquad \times P(\theta|\mathrm{X}_{\mathrm{new}},\mathbf{y},\mathbf{x})P(\mathrm{X}_{\mathrm{new}})\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta, \\
&= \int P(\mathbf{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta)P(\mathbf{U}_{\mathrm{new}}|\theta)P(\mathbf{Z}|\mathrm{X}_{\mathrm{new}},\mathbf{y},\mathbf{x},\theta)P(\theta|\mathbf{y},\mathbf{x}) \\
&\qquad \times P(\mathrm{X}_{\mathrm{new}})\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta, \\
&\approx \int P(\mathbf{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\mathbf{Z},\mathbf{U}_{\mathrm{new}},\theta)P(\mathbf{U}_{\mathrm{new}}|\theta)P(\mathbf{Z}|\mathrm{X}_{\mathrm{new}},\mathbf{y},\theta)P(\theta|\mathbf{y},\mathbf{x}) \\
&\qquad \times P(\mathrm{X}_{\mathrm{new}})\mathrm{d}\mathbf{Z}\ \mathrm{d}\mathbf{U}_{\mathrm{new}}\ \mathrm{d}\theta. \qquad (5.52)
\end{aligned}
$$

The approximations used in computation of $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}},\mathbf{y},\mathbf{x})$ are described as follows:

- As given in above the expression $P(\mathbf{y}_{\mathrm{new}}|\theta,\mathbf{Z},\mathbf{U}_{\mathrm{new}},\mathbf{y},\mathbf{x}) = P(\mathbf{y}_{\mathrm{new}}|\theta,\mathbf{Z},\mathbf{U}_{\mathrm{new}})$, the new observation $\mathbf{y}_{\mathrm{new}}$ is independent of the data $(\mathbf{y},\mathbf{x})$ used at the forward stage given θ, $\mathbf{Z}$, $\mathrm{X}_{\mathrm{new}}$ and $\mathbf{U}_{\mathrm{new}}$, i.e

  $$\mathbf{y}_{\mathrm{new}} \perp \{\mathbf{y},\mathbf{x}\}|\{\theta,\mathbf{Z},\mathrm{X}_{\mathrm{new}},\mathbf{U}_{\mathrm{new}}\}.$$

  It is also evident from the DAGs shown in Fig 5.2 and 5.3 that $\mathbf{y}_{\mathrm{new}}$ is independent of $(\mathbf{y},\mathbf{x})$ given $\mathbf{Z},\theta$ and $\mathbf{U}_{\mathrm{new}}$ by the property of d-separation (discussed in Chapter 2).

- The term $\mathbf{U}_{\mathrm{new}}$ denotes the random effects corresponding to $\mathbf{y}_{\mathrm{new}}$, is indepen-

dent of $\mathbf{y}, \mathbf{x}$ and $\mathbf{Z}$ as given in the expression $P(\mathbf{U}_{\text{new}}|\theta, \mathbf{y}, \mathbf{x}, \mathbf{Z}) = P(\mathbf{U}_{\text{new}}|\theta)$. This can also be experienced via the DAGs shown in Fig 5.2 and 5.3 which is a result of the d-separation property. That is

$$\mathbf{U}_{\text{new}} \perp \{\mathbf{Z}, \mathbf{y}, \mathbf{x}\}|\theta.$$

- The prior distribution of X_{new}, $P(\mathrm{X}_{\text{new}}|\mathbf{y}, \mathbf{x})$, if not experienced from past study or past data, may simply be approximated to an unconditional prior $P(\mathrm{X}_{\text{new}})$ given in the expression as $P(\mathrm{X}_{\text{new}}|\mathbf{y}, \mathbf{x}) \approx P(\mathrm{X}_{\text{new}})$. That is

$$\mathrm{X}_{\text{new}} \perp \{\mathbf{y}, \mathbf{x}\}.$$

- The parameters θ are non-spatial hence, they are independent of X_{new} as used in the above expression: $P(\theta|\mathrm{X}_{\text{new}}, \mathbf{y}, \mathbf{x}) = P(\theta|\mathbf{y}, \mathbf{x})$. That is

$$\theta \perp \mathrm{X}_{\text{new}}|\{\mathbf{y}, \mathbf{x}\}.$$

- It is mentioned previously that the explanatory variable is taken on a finite grid and so, X_{new} and $\mathbf{x}$ belong to the same set of discrete grid values. Therefore, $\mathbf{x}$ is ignored from the posterior distribution of $\mathbf{Z}$, $P(\mathbf{Z}|\mathrm{X}_{\text{new}}, \mathbf{x}, \mathbf{y}, \theta)$, to make the computational problem simpler or low-dimensional.

$$P(\mathbf{Z}|\mathrm{X}_{\text{new}}, \mathbf{y}, \mathbf{x}, \theta) \approx P(\mathbf{Z}|\mathrm{X}_{\text{new}}, \mathbf{y}, \theta).$$

After this approximation, $P(\mathbf{Z}|\mathrm{X}_{\text{new}}, \mathbf{x}, \mathbf{y}, \theta)$ becomes a function of X_{new} only.

The posterior distribution is high dimensional and its closed form solution is not available. A Variational Bayes approximation to the Bayesian analysis of the inverse inference problem for the latent random effect model is presented in the next section.

5.3.2 VB approximation to the Bayesian analysis of the problem

The aim of this section is to provide a VB approximation to the inference problems at both the stages separately. The complexity of the approximation and its tractable solution are discussed in detail.

VB approximation at the forward stage:

The VB method is applied to approximate the intractable joint posterior distribution over $\mathbf{Z}$, $\mathbf{U}$ and $\theta = \{\kappa, Q_{\mathrm{u}}\}$ of Eq. 5.48, given the data set $(\mathbf{y}, \mathbf{x})$ presented as:

$$P(\mathbf{Z}, \kappa, \mathbf{U}, Q_{\mathrm{u}}|\mathbf{y}, \mathbf{x}) \approx q(\mathbf{Z}, \kappa, \mathbf{U}, Q_{\mathrm{u}}|\mathbf{y}, \mathbf{x}), \tag{5.53}$$

$$= \left[\prod_{k=1}^{K} q_{\mathbf{Z}_k}(\mathbf{Z}_k) q_{\kappa_k}(\kappa_k)\right] \left[\prod_{j=1}^{n} q_{\mathbf{U}_j}(\mathbf{U}_j)\right] q_{Q_{\mathrm{u}}}(Q_{\mathrm{u}}). \tag{5.54}$$

The terms $\mathbf{y}, \mathbf{x}$ in the VB marginals $q(\cdot)$, have been dropped for the simplicity of the notation.

VB marginal over κ and Q_{u}:

The hyper-parameters κ and Q_{u} are independent of data in their VB approximations given $\mathbf{Z}$ and $\mathbf{U}$ respectively. Their tractable VB marginals are given as follows:

$$q_{\kappa_k}(\kappa_k) \equiv \mathrm{Gamma}(\kappa_k; a_k^\star, b_k^\star), \ \forall k, \tag{5.55}$$

$$q_{Q_{\mathrm{u}}}(Q_{\mathrm{u}}) = \mathrm{Wishart}_K(SS^\star, df^\star). \tag{5.56}$$

The form of the VB-parameters $a_k^\star$, $b_k^\star$, $SS^\star$ and $df^\star$ are defined for the Poisson latent random effect and ZI-Poisson latent random effect models in the next section.

The VB marginals of $\mathbf{Z}_k$ and $\mathbf{U}_j$ are not of any standard form and are difficult

to compute:

$$\begin{aligned}
\log q_{\mathbf{Z}_k}(\mathbf{Z}_k) &\approx \left[\mathbb{E}_{q_{\kappa_k}(\kappa_k)}\{\log P(\mathbf{Z}_k|\kappa_k)\} + \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})}\left\{\log P(\mathbf{y}_k|\mathbf{Z}_k, \mathbf{U}_k)\right\}\right], \\
&\approx \left[-\frac{1}{2}\mathbf{Z}_k^T \mathbb{E}_q(\kappa_k) R \mathbf{Z}_k + \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})}\left\{\log P(\mathbf{y}_k|\mathbf{Z}_k, \mathbf{U}_k)\right\}\right], \qquad (5.57) \\
\log q_{\mathbf{U}_j}(\mathbf{U}_j) &\approx \left[\mathbb{E}_{q_{Q_{\mathrm{u}}}(Q_{\mathrm{u}})}\{\log P(\mathbf{U}_j|Q_{\mathrm{u}})\} + \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})}\left\{\log P(\mathbf{y}_j|\mathbf{Z}(\mathrm{x}_j), \mathbf{U}_j, \hat{\alpha})\right\}\right], \\
&\approx \left[-\frac{1}{2}\mathbf{U}_j^T \mathbb{E}_q(Q_{\mathrm{u}}) \mathbf{U}_j + \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})}\left\{\log P(\mathbf{y}_j|\mathbf{Z}(\mathrm{x}_j), \mathbf{U}_j)\right\}\right]. \qquad (5.58)
\end{aligned}$$

In the R.H.S of the expressions of VB marginals, the likelihood term (the second term in the expression) is a non-Gaussian function. The expectations of the likelihood with respect to VB marginals of $\mathbf{Z}$ and $\mathbf{U}$, are not in a form for a standard VB approximation. A simplification would be to substitute the VB-mode of $\mathbf{Z}$ and $\mathbf{U}$ in both Equations 5.57 and 5.58, but even this is not enough to produce a computationally tractable expression for both VB marginals.

A method of Gaussian approximation (Rue & Held, 2005) for these intractable VB marginals, is explained below.

Gaussian approximation:

A Gaussian approximation to the VB marginal of $\mathbf{Z}_k$ may be obtained by a quadratic approximation to the log-density around its mode. An optimization method of finding mode may be quite slow and may add to the computational time. Rue & Held (2005) describes a quick way to find a Gaussian distribution with a Gaussian prior.

For example, in Eq. 5.57 and 5.58, the posterior $q_{\mathbf{Z}_k}(\mathbf{Z}_k)$ and $q_{\mathbf{U}_j}(\mathbf{U}_j)$ are the intractable VB marginals. In both equations, there are two terms. First is the log of the Gaussian priors, quadratic in unknowns ($\mathbf{Z}_k$ and $\mathbf{U}_j$ respectively). Second is the non-Gaussian log-likelihood. Therefore, a quadratic approximation is only needed for the second term.

For example, consider a second-order Taylor expansion of the log-likelihood $\mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})}\left\{\log P(\mathbf{y}_k|\mathbf{Z}_k, \mathbf{U}_k)\right\}$ around the posterior (VB) mode $\mathbf{Z}_k^m$ of $\mathbf{Z}_k$ as:

$$\mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})}\left\{\log P(\mathbf{y}_k|\mathbf{Z}_k, \mathbf{U}_k)\right\} \approx A_{\mathbf{Z}_k} + B_{\mathbf{Z}_k}\mathbf{Z}_k - 0.5\mathbf{Z}_k^T V_{\mathbf{Z}_k}\mathbf{Z}_k, \qquad (5.59)$$

where $A_{\mathbf{Z}_k}$ is the collection of the constant terms (independent of $\mathbf{Z}$) in the approximation, $B_{\mathbf{Z}_k}$ and $V_{\mathbf{Z}_k}$ are some functions of the first and second derivatives of the log-likelihood $\mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})}\{\log P(\mathbf{y}_k^t|\mathbf{Z}_k, \mathbf{U}_k)\}$ around $\mathbf{Z}_k^m$:

$$B_{\mathbf{Z}_k} = \left\{\mathbf{Z}_k V_{\mathbf{Z}_k} + \frac{\partial}{\partial \mathbf{Z}_k}\mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})}\{\log P(\mathbf{y}_k|\mathbf{Z}_k, \mathbf{U}_k)\}\right\}\Big|_{\mathbf{Z}_k=\mathbf{Z}_k^m}, \tag{5.60}$$

$$V_{\mathbf{Z}_k} = -\frac{\partial^2}{\partial \mathbf{Z}_k^2}\mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})}\{\log P(\mathbf{y}_k|\mathbf{Z}_k, \mathbf{U}_k)\}\Big|_{\mathbf{Z}_k=\mathbf{Z}_k^m}. \tag{5.61}$$

In Eq. 5.60 and 5.61, we require the expectation of ZI-Poisson likelihood with respect to the VB marginal of $\mathbf{U}$, but this is unfortunately not available in a closed form. For a Gaussian approximation, a VB-mode of $\mathbf{U}$ is plugged-in in the likelihood instead.

A Gaussian approximation of $q(\mathbf{Z}|\mathbf{y})$ is obtained as

$$\log q^g(\mathbf{Z}_k|\mathbf{y}) \approx -\frac{1}{2}\mathbf{Z}_k^T Q^\star_{\mathbf{Z}_k}\mathbf{Z}_k + B^T_{\mathbf{Z}_k}\mathbf{Z}_k, \tag{5.62}$$

$$Q^\star_{\mathbf{Z}_k} = \mathbb{E}_q(\kappa_k)R + V_{\mathbf{Z}_k}, \tag{5.63}$$

where the posterior (VB) precision $Q^\star_{\mathbf{Z}_k}$ is a function of prior precision $\mathbb{E}_q(\kappa)R$ and the precision $V_{\mathbf{Z}_k}$ obtained from the likelihood term.

The mean of the approximation can be obtained as a function of the posterior precision $Q^\star_{\mathbf{Z}}$ and the vector $B_{\mathbf{Z}_k}$:

$$\mu_{\mathbf{Z}_k} = {Q^\star_{\mathbf{Z}_k}}^{-1} B^T_{\mathbf{Z}_k}. \tag{5.64}$$

Being a function of $B_{\mathbf{Z}_k}$ and $V_{\mathbf{Z}_k}$, the mean $\mu_{\mathbf{Z}_k}$ and the precision $Q^\star_{\mathbf{Z}_k}$ also depend on $\mathbf{Z}_k^m$. The approximation remains intractable unless the value of $\mathbf{Z}_k^m$ is found. A quick way to find a mode is described already in the previous section for the inverse latent non random model and may here also be applied. A Gaussian approximation of the VB marginal of $\mathbf{U}$ can be found in the same manner.

It might be the case that the posterior distributions are multi-modal. Then it would be difficult to choose a single mode for a Gaussian approximation. The

latent random effect models used in the chapter do not lead to multi-modal posterior distribution of $\mathbf{Z}_k$ and $\mathbf{U}_j$. Hence the problem of multi-modality can be avoided.

Since the Gaussian approximations of the VB marginals of $\mathbf{U}_j$ and $\mathbf{Z}_k$ interact through their posterior modes, ignorance of any of the modes (of $\mathbf{U}_j$ and $\mathbf{Z}_k$) may leave the approximation intractable. Hence it requires an algorithm to find these Gaussian approximations. To save computational time, the modes of the Gaussian approximations are computed only once at each VB iteration. The VB approximation proceeds as follows:

Specify initial parameters (or moments) and modes of the VB marginals,
At each VB-iteration $m = 1 : M$,

1. find the mode of the Gaussian approximation $q^g_{\mathbf{Z}_k}(\mathbf{Z}_k)$ of the intractable $q_{\mathbf{Z}_k}(\mathbf{Z}_k)$ given the posterior modes of $\mathbf{U}$ and $\mathbf{Z}$ evaluated at previous VB-iteration.
2. compute the mean and variance of $q^g_{\mathbf{Z}_k}(\mathbf{Z}_k)$.
3. find the mode of the Gaussian approximation $q^g_{\mathbf{U}_j}(\mathbf{U}_j)$ of $q_{\mathbf{U}_j}(\mathbf{U}_j)$ given the posterior mode of $\mathbf{Z}$ evaluated at the previous step and that of $\mathbf{U}_j$ found at the previous VB-iteration.
4. compute the mean and variance of $q^g_{\mathbf{U}_j}(\mathbf{U}_j)$.
5. compute the VB marginals of other unknowns.

Set the number of the VB-iterations M to a sufficiently large value so that all the VB-parameters converge.

VB approximation at the inverse stage

The VB approximation of the posterior distribution of $\mathrm{X}_{\mathrm{new}}$ is given as follows:

$$
\begin{aligned}
P(\mathrm{X}_{\mathrm{new}}, \mathbf{U}_{\mathrm{new}}, \mathbf{Z}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) &\approx q(\mathrm{X}_{\mathrm{new}}, \mathbf{U}_{\mathrm{new}}, \mathbf{Z}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}), && (5.65)\\
&= q_{\mathrm{X}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) q_{\mathbf{U}_{\mathrm{new}}}(\mathbf{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) \\
&\quad \times q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) && (5.66)
\end{aligned}
$$

Since the VB marginal $q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x})$ is not available, it is fixed to the VB marginal obtained at the forward stage:

$$q(\mathrm{X}_{\text{new}}, \mathbf{U}_{\text{new}}, \mathbf{Z}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) \approx q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) q_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, \mathbf{x}). \tag{5.67}$$

The restricted VB marginal of X_{new} can be computed as given below:

$$\begin{aligned} \log q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{x}, \mathbf{y}) \approx \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{x},\mathbf{y}) q_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{x},\mathbf{y})} \Bigg[\log \Bigg\{ & P(\mathrm{y}_{\text{new}}|\mathbf{Z}(\mathrm{X}_{\text{new}}), \mathbf{U}_{\text{new}}) \\ & \times P(\mathbf{Z}|\mathrm{X}_{\text{new}}, \mathbf{xy}) P(\mathrm{X}_{\text{new}}) \Bigg\} \Bigg], \end{aligned} \tag{5.68}$$

where $\mathbf{U}_{\text{new}}$ is random effect corresponding to the new $\mathbf{y}_{\text{new}}$. The VB marginal of $\mathbf{U}_{\text{new}}$ given data $\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}$ is not found at the forward stage. Therefore, it needs to be computed at the inverse stage with the VB marginal of $\mathbf{X}_{\text{new}}$.

$$\begin{aligned} \log q_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{x}, \mathbf{y}) \approx \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{x},\mathbf{y}) q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{x},\mathbf{y}) q_\theta(\theta|\mathbf{x},\mathbf{y})} \Bigg[\log \Bigg\{ & P(\mathbf{U}_{\text{new}}|\theta) \\ & \times P(\mathrm{y}_{\text{new}}|\mathbf{Z}(\mathrm{X}_{\text{new}}), \mathbf{U}_{\text{new}}) \Bigg\} \Bigg]. \end{aligned} \tag{5.69}$$

Just like at the forward stage, the VB marginal $q_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathrm{y}_{\text{new}}, \mathbf{x}, \mathbf{y})$ is not a tractable distribution due to the non-conjugacy of prior to likelihood or the complex form of the likelihood. The Gaussian approximation (as described earlier in the section) of $q_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x})$ is given as:

$$q_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) \approx q^g_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}), \tag{5.70}$$

$$q^g_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) = N_p(\mu^\star_{\mathbf{U}_{\text{new}}}, Q^{\star\,-1}_{\mathbf{U}_{\text{new}}}), \tag{5.71}$$

where the term $q^g_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x})$ is the Gaussian with mean $\mu^\star_{\mathbf{U}_{\text{new}}}$ and precision

$Q^{\star}_{\mathrm{U_{new}}}$ given as follows:

$$Q^{\star}_{\mathrm{U_{new}}} = \mathbb{E}_q(Q_{\mathbf{U}}) + diag(V_{\mathrm{U_{new}}}), \tag{5.72}$$

$$\mu^{\star}_{\mathrm{U_{new}}} = B^T_{\mathrm{U_{new}}} Q^{\star\,-1}_{\mathrm{U_{new}}}, \tag{5.73}$$

$$B_{\mathrm{U_{new}}} = [A_{\mathrm{U_{new}}\,k};\ k = 1 : K]^T + \mathbf{U}^m_{\mathrm{new}} V_{\mathrm{U_{new}}}, \tag{5.74}$$

$$V_{\mathrm{U_{new}}} = diag[V_{\mathrm{U_{new}}\,k}\ k = 1 : K], \tag{5.75}$$

$$V_{\mathrm{U_{new}}\,k} = -\frac{\partial^2}{\partial \mathrm{U}^2_{\mathrm{new}\,k}} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z}) q_{\mathrm{X_{new}}}(\mathrm{X_{new}}|\mathbf{y}_{\mathrm{new}},\mathbf{y},\mathbf{x})} \log P(\mathrm{y_{new}}|\mathbf{Z}_k(\mathrm{X_{new}}), \mathrm{U_{new}}\,_k)\Big|_{\mathrm{U_{new}}\,_k = \mathrm{U}^m_{\mathrm{new}\,k}} \tag{5.76}$$

$$A_{\mathrm{U_{new}}\,k} = \frac{\partial}{\partial \mathrm{U_{new}}\,_k} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z}) q_{\mathrm{X_{new}}}(\mathrm{X_{new}}|\mathbf{y}_{\mathrm{new}},\mathbf{y},\mathbf{x})} \log P(\mathrm{y_{new}}|\mathbf{Z}_k(\mathrm{X_{new}}), \mathrm{U_{new}}\,_k)\Big|_{\mathrm{U_{new}}\,_k = \mathrm{U}^m_{\mathrm{new}\,k}} \tag{5.77}$$

The form of the parameters of the posterior distributions described above are explained for the Poisson and ZI-Poisson latent random effect models in next section.

5.3.3 Comparison of the VB approximation with approximation by INLA

INLA will be used to compare the accuracy and the computation time of the VB method for inverse estimation for the latent random effect models discussed in the next sections. At the forward stage the VB approximation is compared with the results from INLA. INLA might be inappropriate for the inverse estimation, hence the results from the VB method are compared with the results from the Laplace approximation which uses results from INLA obtained at the forward stage.

Another accuracy check for inverse estimation

To check the accuracy of the approximation at the inverse stage, a 95% HPD region is computed. The percentage of true values of the new explanatory variables that lie within the 95% HPD region gives an accuracy measure of the inverse inference.

The general framework for the VB approximation for the inverse latent random effect problem has been presented in this section. In the next section, the VB approximations for the inverse Poisson latent random effect and the zero-inflated Poisson latent random effect problem are described based on the approximations explained in this section.

5.3.4 VB approximation for the inverse Poisson latent random effect regression model

VB approximations at the forward stage and the inverse stage for the inverse Poisson latent with random effects regression model are presented below:

VB solution at Forward stage:

1. The VB marginal $q_{\kappa_k}(\kappa|\mathbf{y},\mathbf{x})$; $\forall k$ is obtained as:

$$
\begin{aligned}
q_{\kappa_k}(\kappa_k|\mathbf{y},\mathbf{x}) &= Gamma(\kappa_k; a_k^\star, b_k^\star), && (5.78)\\
a_k^\star &= a_k + 0.5p, && (5.79)\\
b_k^\star &= b_k + 0.5\Bigg[2\sum_{i=1}^{p}\mathbb{E}_q(\mathrm{Z}_{ki}^2) - \mathbb{E}_q(\mathrm{Z}_{kp}^2) \\
&\qquad -2\sum_{i\neq m}^{p}\sum\mathbb{E}_q(\mathrm{Z}_{ki}\mathrm{Z}_{km})\Bigg]. && (5.80)
\end{aligned}
$$

2. The VB marginal $q_{Q_\mathbf{U}}(Q_\mathbf{U}|\mathbf{y},\mathbf{x})$ is given as:

$$
\begin{aligned}
q_{Q_\mathbf{U}}(Q_\mathbf{U}|\mathbf{y},\mathbf{x}) &= Wishart(Q_\mathbf{U}; SS^\star, df^\star), && (5.81)\\
df^\star &= df + 0.5N, && (5.82)\\
SS^\star &= \left[SS^{-1} + \sum_{j=1}^{N}\sum_{k,l=1}^{K}\mathbb{E}_q(\mathrm{U}_{kj}\mathrm{U}lj)\right]. && (5.83)
\end{aligned}
$$

3. The VB marginal $q_\mathbf{Z}(\mathbf{Z}|\mathbf{y},\mathbf{x})$ is given as follows:

$$
\begin{aligned}
q_{\mathbf{Z}_k}(\mathbf{Z}_k) &\approx q^g_{\mathbf{Z}_k}(\mathbf{Z}_k), && (5.84)\\
q^g_{\mathbf{Z}_k}(\mathbf{Z}_k) &= N_p(\mu^\star_{\mathbf{Z}_k}, {Q^\star_{\mathbf{Z}_k}}^{-1}), && (5.85)
\end{aligned}
$$

where the term $q^g_{\mathbf{Z}_k}(\mathbf{Z}_k)$ denotes the Gaussian approximation of $q_{\mathbf{Z}_k}(\mathbf{Z}_k)$ with

mean $\mu^{\star}_{\mathbf{Z}_k}$ and precision $Q^{\star}_{\mathbf{Z}_k}$ given as follows:

$$
\begin{aligned}
Q^{\star}_{\mathbf{Z}_k} &= \mathbb{E}_q(\kappa_k)R + diag(V_{\mathbf{Z}_k}), && (5.86)\\
\mu^{\star}_{\mathbf{Z}_k} &= B^T_{\mathbf{Z}_k} Q^{\star\,-1}_{\mathbf{Z}_k}, && (5.87)\\
B_{\mathbf{Z}_k} &= [A_{\mathrm{Z}_{ki}};\ i = 1:p]^T + \mathbf{Z}^m_k V_{\mathbf{Z}_k}, && (5.88)\\
V_{\mathbf{Z}_k} &= diag[V_{\mathrm{Z}_{ki}}\ i = 1:p]^T, && (5.89)\\
V_{\mathrm{Z}_{ki}} &= -\frac{\partial^2}{\partial \mathrm{Z}^2_{ki}} \sum_{\substack{j=1\\ \mathrm{x}_j = i}}^{n} \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{Z}_{ki}=\mathrm{Z}^m_{ki}}, && (5.90)\\
&= \sum_{\substack{j=1\\ \mathrm{x}_j = i}}^{n} \exp(\mathbf{Z}^m_k(\mathrm{x}_j))\mathbb{E}_q(\exp(\mathrm{U}_{kj})), && (5.91)\\
A_{\mathrm{Z}_{ki}} &= \frac{\partial}{\partial \mathrm{Z}_{ki}} \sum_{\substack{j=1\\ \mathrm{x}_j = i}}^{n} \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{Z}_{ki}=\mathrm{Z}^m_{ki}}, && (5.92)\\
&= \sum_{\substack{j=1\\ \mathrm{x}_j = i}}^{n} \left[\exp(\mathbf{Z}^m_k(\mathrm{x}_j))\mathbb{E}_q(\exp(\mathrm{U}_{kj})) - \mathrm{y}_{kj}\right], && (5.93)
\end{aligned}
$$

where $\mathbf{Z}^m_k = [\mathrm{Z}^m_{ki};\ i = 1:p]^T$ is the posterior mode of $\mathbf{Z}_k$.

4. The VB marginal $q_{\mathbf{U}_j}(\mathbf{U}_j|\mathbf{y}, \mathbf{x})$ is given as follows:

$$
\begin{aligned}
q_{\mathbf{U}_j}(\mathbf{U}_j) &\approx q^g_{\mathbf{U}_j}(\mathbf{U}_j), && (5.94)\\
q^g_{\mathbf{U}_j}(\mathbf{U}_j) &= N_p(\mu^{\star}_{\mathbf{U}_j}, Q^{\star\,-1}_{\mathbf{U}_j}), && (5.95)
\end{aligned}
$$

where the term $q^g_{\mathbf{U}_j}(\mathbf{U}_j)$ denotes the Gaussian approximation of $q_{\mathbf{U}_j}(\mathbf{U}_j)$ with

mean $\mu^{\star}_{\mathbf{U}_j}$ and precision $Q^{\star}_{\mathbf{U}_j}$ given as follows:

$$
\begin{aligned}
Q^{\star}_{\mathbf{U}_j} &= \mathbb{E}_q(Q_{\mathbf{U}}) + diag(V_{\mathbf{U}_j}), && (5.96)\\
\mu^{\star}_{\mathbf{U}_j} &= B^T_{\mathbf{U}_j} Q^{\star\,-1}_{\mathbf{U}_j}, && (5.97)\\
B_{\mathbf{U}_j} &= [A_{\mathrm{U}_{kj}};\ k = 1:K]^T + \mathbf{U}^m_j V_{\mathbf{U}_j}, && (5.98)\\
V_{\mathbf{U}_j} &= diag[V_{\mathrm{U}_{kj}}\ k = 1:K]^T, && (5.99)\\
V_{\mathrm{U}_{kj}} &= -\frac{\partial^2}{\partial \mathrm{U}^2_{kj}} \sum_{j=1}^{n} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{U}_{kj}=\mathrm{U}^m_{kj}}, && (5.100)\\
&= \exp(\mathrm{U}^m_{kj})\mathbb{E}_q(\exp(\mathbf{Z}_k(\mathrm{x}_j))), && (5.101)\\
A_{\mathrm{U}_{kj}} &= \frac{\partial}{\partial \mathrm{U}_{kj}} \sum_{j=1}^{n} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{U}_{kj}=\mathrm{U}^m_{kj}}, && (5.102)\\
&= \exp(\mathrm{U}^m_{kj})\mathbb{E}_q(\exp(\mathbf{Z}_k(\mathrm{x}_j))) - \mathrm{y}_{kj}, && (5.103)
\end{aligned}
$$

where $\mathbf{U}^m_j = [\mathrm{U}^m_{kj};\ i = 1:p]^T$ is the posterior mode of $\mathbf{U}_j$.

VB approximation at the inverse stage:

The restricted VB marginal of $\mathrm{X}_{\mathrm{new}}$ is obtained as:

$$
\begin{aligned}
\log q(\mathrm{X}_{\mathrm{new}} = i|\mathrm{y}_{\mathrm{new}}, \mathbf{x}, \mathbf{y}) \approx \sum_{k=1}^{K} \Big[& -\mathbb{E}_q(\exp(\mathbf{Z}_k(i)))\mathbb{E}_q(\exp(\mathrm{U}_{\mathrm{new}\,k})) + \mathrm{y}_{\mathrm{new}\,k}(\mathbb{E}_q(\mathbf{Z}(\mathrm{X}_{\mathrm{new}})) \\
& + \mathbb{E}_q(\mathrm{U}_{\mathrm{new}\,k}))\Big] + \sum_{k=1}^{K} [-\log(\mathrm{y}_{\mathrm{new}\,k}!)] \\
& - 0.5\Big[\sum_{i\neq j}^{p} Var_{\mathbf{Z}_k}(i,j) Q^{\star}_{\mathbf{Z}_k}(i,j) \\
& + diag[Var_{\mathbf{Z}_k}]^T diag[Q^{\star}_{\mathbf{Z}_k}] + \Big]. \qquad (5.104)
\end{aligned}
$$

where $Var_{\mathbf{Z}_k}$ is the VB-variance of $\mathbf{Z}_k$. The term $Var_{\mathbf{Z}_k}(i,j)$ and $Q_{\mathbf{Z}_k}(i,j)$ are $(i,j)^{th}$ entry for $i,j = 1:p$ of the VB variance and VB-precision of $\mathbf{Z}_k$.

The Gaussian approximation of (the intractable) VB marginal $q_{\mathbf{U}_{\mathrm{new}}}(\mathbf{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})$

is given as:

$$q_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{y},\mathbf{x}) \approx q^{g}_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{y},\mathbf{x}), \tag{5.105}$$

$$q^{g}_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{y},\mathbf{x}) = N_p(\mu^{\star}_{\mathbf{U}_{\text{new}}}, Q^{\star\,-1}_{\mathbf{U}_{\text{new}}}), \tag{5.106}$$

where the term $q^{g}_{\mathbf{U}_{\text{new}}}(\mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{y},\mathbf{x})$ is the Gaussian approximation with mean $\mu^{\star}_{\mathbf{U}_{\text{new}}}$ and precision $Q^{\star}_{\mathbf{U}_{\text{new}}}$ given as follows:

$$Q^{\star}_{\mathbf{U}_{\text{new}}} = \mathbb{E}_q(Q_{\mathbf{U}}) + diag(V_{\mathbf{U}_{\text{new}}}), \tag{5.107}$$

$$\mu^{\star}_{\mathbf{U}_{\text{new}}} = B^{T}_{\mathbf{U}_{\text{new}}} Q^{\star\,-1}_{\mathbf{U}_{\text{new}}}, \tag{5.108}$$

$$B_{\mathbf{U}_{\text{new}}} = [A_{\mathrm{U}_{\text{new}}\,k};\ k = 1:K]^{T} + \mathbf{U}^{m}_{\text{new}} V_{\mathbf{U}_{\text{new}}}, \tag{5.109}$$

$$V_{\mathbf{U}_{\text{new}}} = diag[V_{\mathrm{U}_{\text{new}}\,k}\ k = 1:K]^{T}, \tag{5.110}$$

$$V_{\mathrm{U}_{\text{new}}\,k} = -\frac{\partial^2}{\partial \mathrm{U}^2_{\text{new}\,k}} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z}) q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{y},\mathbf{x})} \log P(\mathrm{y}_{\text{new}}|\mathbf{Z}_k(\mathrm{X}_{\text{new}}), \mathrm{U}_{\text{new}\,k}) \Big|_{\mathrm{U}_{\text{new}\,k}=\mathrm{U}^{m}_{\text{new}\,k}} \tag{5.111}$$

$$= \exp(\mathrm{U}^{m}_{\text{new}\,k}) \sum_{i=1}^{p} \mathbb{E}_q(\exp(\mathbf{Z}_k(\mathrm{X}_{\text{new}}))) q(\mathrm{X}_{\text{new}} = i|\mathrm{y}_{\text{new}},\mathbf{x},\mathbf{y}), \tag{5.112}$$

$$A_{\mathrm{U}_{\text{new}}\,k} = \frac{\partial}{\partial \mathrm{U}_{\text{new}\,k}} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z}) q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}}|\mathbf{y}_{\text{new}},\mathbf{y},\mathbf{x})} \log P(\mathrm{y}_{\text{new}}|\mathbf{Z}_k(\mathrm{X}_{\text{new}}), \mathrm{U}_{\text{new}\,k}) \Big|_{\mathrm{U}_{\text{new}\,k}=\mathrm{U}^{m}_{\text{new}\,k}} \tag{5.113}$$

$$= \exp(\mathrm{U}^{m}_{\text{new}\,k}) \sum_{i=1}^{p} \mathbb{E}_q(\exp(\mathbf{Z}_k(\mathrm{X}_{\text{new}}))) q_{\mathrm{X}_{\text{new}}}(\mathrm{X}_{\text{new}} = i|\mathbf{y}_{\text{new}},\mathbf{y},\mathbf{x}) - \mathrm{y}_{\text{new}\,k} \tag{5.114}$$

Result

To show the VB approximation for the Poisson latent random effect regression model an example of simulated data is considered. A set of one hundred discrete values, from one to twenty five, on explanatory variable $\mathbf{x}$ are drawn from a discrete uniform distribution. Defined over twenty five equi-distant discrete locations, latent variables $\mathbf{Z}$ are generated from independent GMRF distributions with Gamma variate precision parameters. one hundred values of tri-variate random effects $\mathbf{U}$ are generated from independent multivariate normal distributions with zero means and a common precision following a Wishart distribution. Corresponding to these values, a set of hundred values on response variable $\mathbf{y}$ are simulated from a Poisson latent random effect model. These data on $\mathbf{x}$ and $\mathbf{y}$ are used at forward stage to

approximate unknown (assumed) $\mathbf{Z}$ and $\mathbf{U}$.

A set of fifty new observations on $\mathbf{y}_{\text{new}}$ and $\mathbf{X}_{\text{new}}$ following the same mechanism. The true values of explanatory variable $\mathbf{X}_{\text{new}}$ are used for the accuracy of the approximation of inverse estimation of $\mathbf{X}_{\text{new}}$ via cross validation technique with 95% HPD region. Fig. 5.7 and 5.8 represent the VB results for the Poisson latent random effect model. Fig. 5.7 shows the comparison between the mean of approximations of the true posterior mean of the latent variables $\mathbf{Z}$ by the VB method and by INLA. The VB method and INLA provides quite similar approximations to the true values of $\mathbf{Z}$. None of them approximate $\mathbf{Z}$ very accurately. At some grid locations, the true values of $\mathbf{Z}$ lie outside the 95% HPD region. It should be noted that the data provided at the forward stage is not very informative and includes many zeros which results in bad approximation by both methods.

In Fig. 5.8 approximations (by VB and INLA) of posterior distribution of a new unknown explanatory variable X_{new} given data on the response variable $\mathbf{Y}_{\text{new}}$ are shown. The VB approximation is close to the approximation by INLA. The approximations are non-smooth multi-modal densities as a result of the bad approximation by both methods at the forward stage. The accuracy of the approximation by INLA with 95% HPD region is only 78% and that of the VB approximation with 95% HPD region is only 74%.

The count data, to show the VB results for inverse estimation for Poisson latent random effect model, are zero abundant. The Poisson model might be inappropriate for modeling count data with excess of zeros. In the next section, a zero-inflated Poisson model is discussed which is expected to model excess of zeros accurately.

Computational time:

The computational time taken by the VB method and INLA are compared and displayed in Table 5.2. It clear from the table that the VB method provides similar results in very less time as compared to INLA.

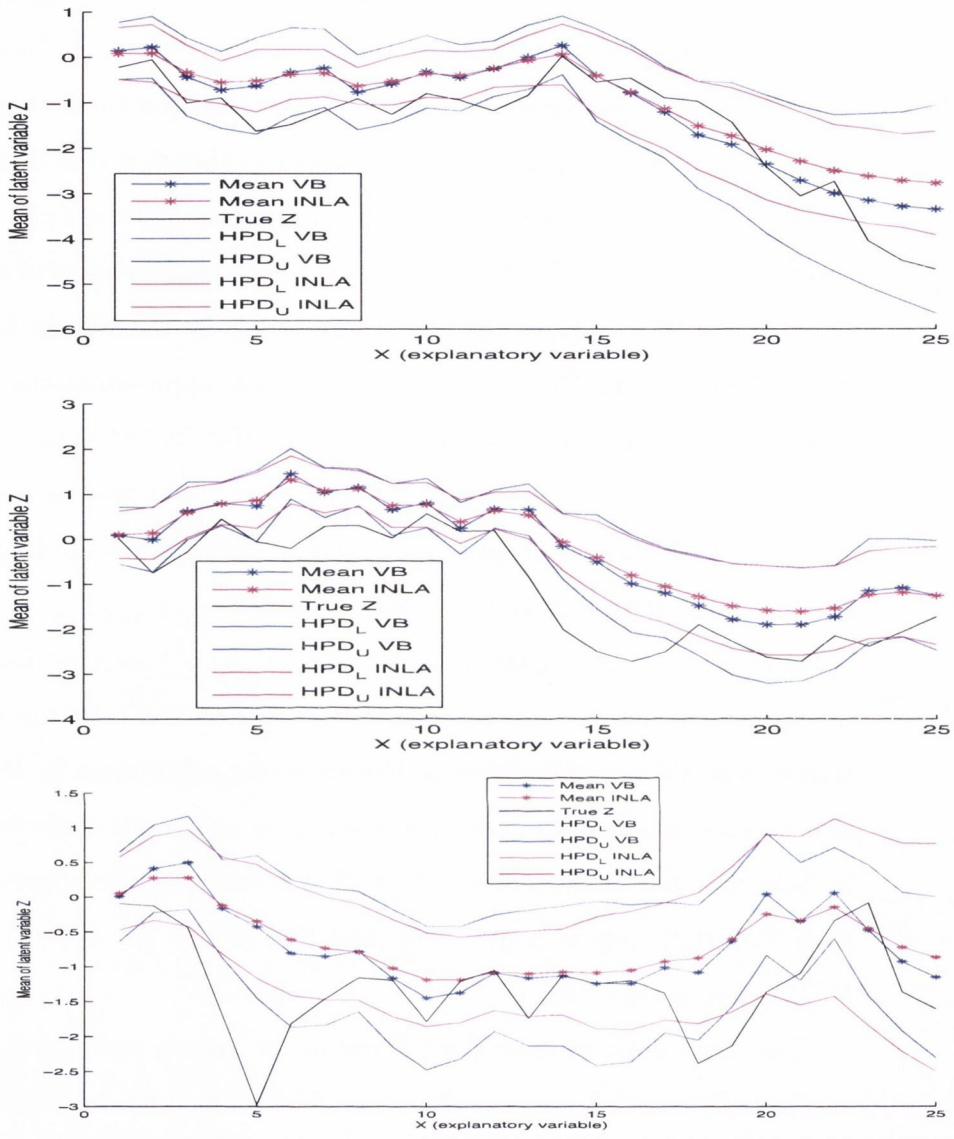

Fig. 5.7: The comparison of true value (black), VB-mean(blue), mean (estimated by INLA) (magenta) of the latent variables $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ corresponding to the response variable $\mathbf{Y}_1$ (the upper one), $\mathbf{Y}_2$ (the middle one) and $\mathbf{Y}_3$ (the bottom one) respectively of a Poisson latent random effect model.

	Forward stage	Inverse stage
VB	0.74 sec	4.4 sec
INLA	9.27 sec	10.25 sec

Table 5.2: The comparison of the computational time of VB method and INLA at both stages for Poisson latent random effect model. At the inverse stage the computation time shows the time taken by the method for the inverse estimation for all test data.

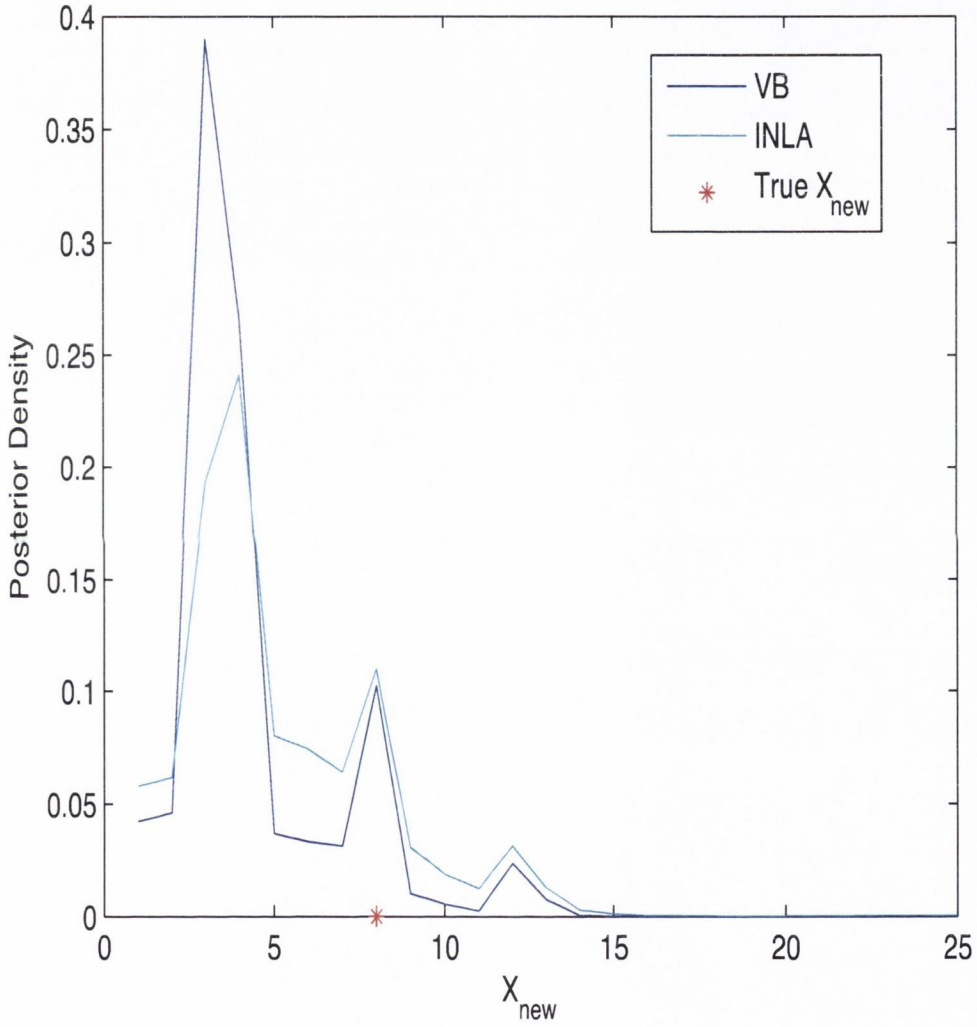

Fig. 5.8: The comparison of the approximation by the VB method (blue) and by the INLA (cyan) of true posterior distribution of X_{new} for a Poisson latent random effect model.

5.3.5 VB approximation for the inverse zero-inflated Poisson latent with random effects regression model

The VB method is applied to approximate the intractable joint posterior distribution over $\mathbf{Z}$, $\mathbf{U}$ and $\theta = \{\kappa, Q_{\mathrm{u}}\}$, given the training data set $(\mathbf{y}, \mathbf{x})$. A VB approximation of the posterior distribution of power index (zero-inflated) parameter α is not carried out because assuming α to be unknown presents difficulties for VB since the log-joint likelihood can no longer be factorized in a way that facilitates VB. Here, α is estimated by numerical optimization of $\log P(\alpha|\mathbf{y}, \mathbf{x}, \hat{\mathbf{Z}}, \hat{\mathbf{U}})$, and inference then is conducted on its modal value $\hat{\alpha}$, where $\hat{\mathbf{Z}}$ and $\hat{\mathbf{U}}$ are the posterior (VB) modes of $\mathbf{Z}$ and $\mathbf{U}$ respectively.

A conjugate prior distribution is assumed over α

$$P(\alpha) = \prod_{k=1}^{K} log\ Normal(\alpha_k;\ \mu_{\alpha_k}, \sigma^2_{\alpha_k});\ \alpha_k \in (0, \infty). \tag{5.115}$$

The hyper-parameters in the above prior distributions are assumed to be known.

VB approximation at the forward stage:

The VB marginals obtained at the forward stage are described below:

1. The VB marginal $q_{\kappa_k}(\kappa|\mathbf{y}, \mathbf{x}, \hat{\alpha});\ \forall k$ is obtained as:

$$\begin{aligned} q_{\kappa_k}(\kappa_k|\mathbf{y}, \mathbf{x}, \hat{\alpha}) &= Gamma(\kappa_k; a^\star_k, b^\star_k), && (5.116)\\ a^\star_k &= a_k + 0.5p, && (5.117)\\ b^\star_k &= b_k + 0.5\Big[2\sum_{i=1}^{p}\mathbb{E}_q(\mathrm{Z}^2_{ki}) - \mathbb{E}_q(\mathrm{Z}^2_{kp}) \\ &\qquad -2\sum_{i\neq m}^{p}\sum\mathbb{E}_q(\mathrm{Z}_{ki}\mathrm{Z}_{km})\Big]. && (5.118) \end{aligned}$$

2. The VB marginal $q_{Q_{\mathbf{U}}}(Q_{\mathbf{U}}|\mathbf{y}, \mathbf{x}, \hat{\alpha})$ is given as:

$$q_{Q_{\mathbf{U}}}(Q_{\mathbf{U}}|\mathbf{y}, \mathbf{x}, \hat{\alpha}) = Wishart(Q_{\mathbf{U}}; SS^{\star}, df^{\star}), \tag{5.119}$$

$$df^{\star} = df + 0.5N, \tag{5.120}$$

$$SS^{\star} = \left[SS^{-1} + \sum_{j=1}^{N} \sum_{k,l=1}^{K} \mathbb{E}_q(\mathrm{U}_{kj}\mathrm{U}_{lj}) \right]. \tag{5.121}$$

3. The VB marginal $q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, \mathbf{x}, \hat{\alpha})$ is given as follows:

$$q_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y}, \mathbf{x}, \hat{\alpha}) \approx q^g_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y}, \mathbf{x}, \hat{\alpha}), \tag{5.122}$$

$$q^g_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y}, \mathbf{x}, \hat{\alpha}) = N_p(\mu^{\star}_{\mathbf{Z}_k}, Q^{\star\,-1}_{\mathbf{Z}_k}), \tag{5.123}$$

where the term $q^g_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y}, \mathbf{x}, \hat{\alpha})$ denotes the Gaussian approximation of $q_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y}, \mathbf{x}, \hat{\alpha})$ with mean $\mu^{\star}_{\mathbf{Z}_k}$ and precision $Q^{\star}_{\mathbf{Z}_k}$ given as follows:

$$Q^{\star}_{\mathbf{Z}_k} = \mathbb{E}_q(\kappa_k)R + diag(V_{\mathbf{Z}_k}), \tag{5.124}$$

$$\mu^{\star}_{\mathbf{Z}_k} = B^T_{\mathbf{Z}_k} Q^{\star\,-1}_{\mathbf{Z}_k}, \tag{5.125}$$

$$B_{\mathbf{Z}_k} = [A_{\mathrm{Z}_{ki}};\ i = 1:p]^T + \mathbf{Z}^m_k V_{\mathbf{Z}_k}, \tag{5.126}$$

$$V_{\mathbf{Z}_k} = diag[V_{\mathrm{Z}_{ki}}\ i = 1:p]^T, \tag{5.127}$$

$$V_{\mathrm{Z}_{ki}} = -\frac{\partial^2}{\partial \mathrm{Z}^2_{ki}} \sum_{\substack{j=1 \\ \mathrm{x}_j = i}}^{n} \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj}) \Big|_{\mathrm{Z}_{ki}=\mathrm{Z}^m_{ki}}, \tag{5.128}$$

$$A_{\mathrm{Z}_{ki}} = \frac{\partial}{\partial \mathrm{Z}_{ki}} \sum_{\substack{j=1 \\ \mathrm{x}_j = i}}^{n} \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj}) \Big|_{\mathrm{Z}_{ki}=\mathrm{Z}^m_{ki}}, \tag{5.129}$$

where $\mathbf{Z}^m_k = [\mathrm{Z}^m_{ki};\ i = 1:p]^T$ is the posterior mode of $\mathbf{Z}_k$. The terms $V_{\mathrm{Z}_{ki}}$ and $A_{\mathrm{Z}_{ki}}$ are given in the Appendix.

4. The VB marginal $q_{\mathbf{U}_j}(\mathbf{U}_j|\mathbf{y}, \mathbf{x})$ is given as follows:

$$q_{\mathbf{U}_j}(\mathbf{U}_j|\mathbf{y}, \mathbf{x}, \hat{\alpha}) \approx q^g_{\mathbf{U}_j}(\mathbf{U}_j|\mathbf{y}, \mathbf{x}, \hat{\alpha}), \tag{5.130}$$

$$q^g_{\mathbf{U}_j}(\mathbf{U}_j|\mathbf{y}, \mathbf{x}, \hat{\alpha}) = N_p(\mu^{\star}_{\mathbf{U}_j}, Q^{\star\,-1}_{\mathbf{U}_j}), \tag{5.131}$$

where the term $q^g_{\mathrm{U}_j}(\mathbf{U}_j)$ denotes the Gaussian approximation of $q_{\mathbf{U}_j}(\mathbf{U}_j)$ with mean $\mu^\star_{\mathbf{U}_j}$ and precision $Q^\star_{\mathbf{U}_j}$ given as follows:

$$
\begin{aligned}
Q^\star_{\mathbf{U}_j} &= \mathbb{E}_q(Q_{\mathbf{U}}) + diag(V_{\mathbf{U}_j}), && (5.132)\\
\mu^\star_{\mathbf{U}_j} &= B^T_{\mathbf{U}_j} Q^{\star\,-1}_{\mathbf{U}_j}, && (5.133)\\
B_{\mathbf{U}_j} &= [A_{\mathrm{U}_{kj}};\ k = 1:K]^T + \mathbf{U}^m_j V_{\mathbf{U}_j}, && (5.134)\\
V_{\mathbf{U}_j} &= diag[V_{\mathrm{U}_{kj}}\ k = 1:K]^T, && (5.135)\\
V_{\mathrm{U}_{kj}} &= -\frac{\partial^2}{\partial \mathrm{U}^2_{kj}} \sum_{j=1}^{n} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{U}_{kj}=\mathrm{U}^m_{kj}}, && (5.136)\\
A_{\mathrm{U}_{kj}} &= \frac{\partial}{\partial \mathrm{U}_{kj}} \sum_{j=1}^{n} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{U}_{kj}=\mathrm{U}^m_{kj}}, && (5.137)
\end{aligned}
$$

where $\mathbf{U}^m_j = [\mathrm{U}^m_{kj};\ i = 1:p]^T$ is the posterior mode of $\mathbf{U}_j$. Due to the complexity of the functions $V_{\mathrm{U}_{kj}}$ and $A_{\mathrm{U}_{kj}}$ to be explained in the chapter, they are described in the appendix chapter.

VB approximation at the inverse stage:

As explained in the previous chapter, the restrictive VB approximation to $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}},$ might not be tractable with a non-exponential likelihood. It might require other approximations, e.g. a Gaussian approximation. It should be noted that the explanatory variable is assumed to be a discrete variable and is the index of the latent field $\mathbf{Z}$. A Gaussian approximation to a discrete distribution of an index type variable may be inappropriate. Also, the VB marginals depend on the VB-moments but moments of indexes may not exists. Keeping all this in mind, the VB method is avoided to approximate the posterior distribution of climate.

A Laplace type approximation, as suggested by Rue. et al. (2009), is applied. The marginal posterior distribution $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})$ can be redefined as follows:

$$
\begin{aligned}
P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}) &\approx \frac{P(\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \mathbf{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})}{P_G(\mathbf{Z}, \mathbf{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}, \mathrm{X}_{\mathrm{new}})}\Bigg|_{\substack{\mathbf{Z}=\hat{\mathbf{Z}}(\mathrm{X}_{\mathrm{new}})\\ \mathbf{U}_{\mathrm{new}}=\hat{\mathbf{U}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}})}},\\
&\propto \frac{P(\mathrm{X}_{\mathrm{new}}, \mathbf{Z}, \mathbf{U}_{\mathrm{new}}, \mathbf{y}_{\mathrm{new}}|\mathbf{y}, \mathbf{x})}{P_G(\mathbf{Z}, \mathbf{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}, \mathrm{X}_{\mathrm{new}})}\Bigg|_{\substack{\mathbf{Z}=\hat{\mathbf{Z}}(\mathrm{X}_{\mathrm{new}})\\ \mathbf{U}_{\mathrm{new}}=\hat{\mathbf{U}_{\mathrm{new}}}(\mathrm{X}_{\mathrm{new}})}} \qquad (5.138)
\end{aligned}
$$

where $P_G(\mathbf{Z}, \mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}, \mathrm{X}_{\text{new}})$ is a Gaussian approximation to be defined later and the terms $\hat{\mathbf{Z}}(\mathrm{X}_{\text{new}})$ and $\hat{\mathbf{U}_{\text{new}}}(\mathrm{X}_{\text{new}})$ are the posterior modes of the Gaussian approximation.

Numerator of the approximation:

The numerator of the R.H.S of Eq. 5.138, can be further decomposed as:

$$\begin{aligned}
P(\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \mathbf{y}_{\text{new}}|\mathbf{y}, \mathbf{x}) &= P(\mathbf{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \mathbf{y}, \mathbf{x})P(\mathbf{Z}, \mathbf{U}_{\text{new}}, \mathrm{X}_{\text{new}}|\mathbf{y}, \mathbf{x}), \\
&= P(\mathbf{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \mathbf{y}, \mathbf{x})P(\mathbf{Z}|\mathbf{y}, \mathbf{x}, \mathrm{X}_{\text{new}}, \mathbf{U}_{\text{new}}) \\
&\quad \times P(\mathbf{U}_{\text{new}}|\mathrm{X}_{\text{new}}, \mathbf{y}, \mathbf{x})P(\mathrm{X}_{\text{new}}|\mathbf{y}, \mathbf{x}), \\
&\approx P(\mathbf{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \mathbf{y}, \mathbf{x})P(\mathbf{Z}|\mathbf{y}, \mathrm{X}_{\text{new}}) \\
&\quad \times P(\mathbf{U}_{\text{new}}|\mathbf{y}, \mathbf{x})P(\mathbf{C}_{\text{new}}), \\
&= P(\mathbf{Z}|\mathbf{y}, \mathrm{X}_{\text{new}})P(\mathrm{X}_{\text{new}})\left[\int_{\alpha} P(\mathbf{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \alpha)\right. \\
&\quad \left.\times P(\alpha|\mathbf{y}, \mathbf{x})\mathrm{d}\alpha\right]\left[\int_{Q_{\mathrm{u}}} P(\mathbf{U}_{\text{new}}|Q_{\mathrm{u}})P(Q_{\mathrm{u}}|\mathbf{y}, \mathbf{x})\mathrm{d}Q_{\mathrm{u}}\right], \\
&\approx P(\mathbf{Z}|\mathbf{y}, \mathrm{X}_{\text{new}})P(\mathrm{X}_{\text{new}})P(\mathbf{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \hat{\alpha}) \\
&\quad \times \left[\int_{Q_{\mathrm{u}}} P(\mathbf{U}_{\text{new}}|Q_{\mathrm{u}})P(Q_{\mathrm{u}}|\mathbf{y}, \mathbf{x})\mathrm{d}Q_{\mathrm{u}}\right]. \qquad (5.139)
\end{aligned}$$

The marginal posterior distribution of α, $P(\alpha|\mathbf{y}, \mathbf{x})$, is not computed explicitly. Its posterior mode is plugged-in the likelihood, as in Eq. 5.62. The conditional posterior distribution of $\mathbf{Z}$ is independent of $\mathbf{U}_{\text{new}}$ at the forward stage. Given X_{new}, the posterior distribution is independent of $\mathbf{x}$, hence $P(\mathbf{Z}|\mathbf{y}, \mathbf{x}, \mathrm{X}_{\text{new}}, \mathbf{U}_{\text{new}}) \approx P(\mathbf{Z}|\mathbf{y}, \mathrm{X}_{\text{new}})$. VB approximations of marginal posteriors $P(\mathbf{Z}|\mathbf{y}, \mathrm{X}_{\text{new}})$ and $P(Q_{\mathrm{u}}|\mathbf{y}, \mathbf{x})$ are found at the forward stage and can be used straight into the computation. The integration with respect to Q_{u} is tractable.

$$\begin{aligned}
P(\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \mathbf{y}_{\text{new}}|\mathbf{y}, \mathbf{x}) &\approx q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, \mathrm{X}_{\text{new}})P(\mathrm{X}_{\text{new}})P(\mathbf{y}_{\text{new}}|\mathrm{X}_{\text{new}}, \mathbf{Z}, \mathbf{U}_{\text{new}}, \hat{\alpha}) \\
&\quad \times \left[\int_{Q_{\mathrm{u}}} P(\mathbf{U}_{\text{new}}|Q_{\mathrm{u}})q_{Q_{\mathrm{u}}}(Q_{\mathrm{u}}|\mathbf{y}, \mathbf{x})\mathrm{d}Q_{\mathrm{u}}\right] \qquad (5.140)
\end{aligned}$$

Denominator of the approximation:

The denominator of the R.H.S of Eq. 5.138 is a Gaussian approximation to the joint

posterior distribution of $\mathbf{Z}$ and $\mathrm{U}_{\mathrm{new}}$ given $\mathrm{X}_{\mathrm{new}}$.

$$\begin{aligned} P_G(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}, \mathrm{X}_{\mathrm{new}}) &= \text{Gaussian approx.}\left[P(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}, \mathrm{X}_{\mathrm{new}})\right], \\ \text{and } P(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}, \mathrm{X}_{\mathrm{new}}) &\propto P(\mathbf{y}_{\mathrm{new}}|\mathbf{Z}, \mathrm{U}_{\mathrm{new}}, \mathrm{X}_{\mathrm{new}})P(\mathbf{Z}|\mathbf{y}, \mathbf{x}, \mathrm{X}_{\mathrm{new}})P(\mathrm{U}_{\mathrm{new}}|\mathbf{y}, \mathbf{x}) \end{aligned}$$

Given $\mathrm{X}_{\mathrm{new}}$, $\mathbf{x}$ can be dropped:

$$P(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x}, \mathrm{X}_{\mathrm{new}}) \propto P(\mathbf{y}_{\mathrm{new}}|\mathbf{Z}, \mathrm{U}_{\mathrm{new}}, \mathrm{X}_{\mathrm{new}})P(\mathbf{Z}|\mathbf{y}, \mathrm{X}_{\mathrm{new}})P(\mathrm{U}_{\mathrm{new}}|\mathbf{y}, \mathbf{x}), \tag{5.141}$$

where the marginal posterior distribution $P(\mathrm{U}_{\mathrm{new}}|\mathbf{y}, \mathbf{x}) = \int_{Q_{\mathrm{u}}} P(\mathrm{U}_{\mathrm{new}}|Q_{\mathrm{u}})q_{Q_{\mathrm{u}}}(Q_{\mathrm{u}}|\mathbf{y}, \mathbf{x})\mathrm{d}$ is a complicated function of $\mathrm{U}_{\mathrm{new}}$. To reduce the functional complexity, the VB method is applied to approximate the posterior distribution of $\mathbf{Z}$, $\mathrm{U}_{\mathrm{new}}$ and Q_{u}:

$$\begin{aligned} P(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}, Q_{\mathrm{u}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}}) &\approx q(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}, Q_{\mathrm{u}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}}), \\ &= q_{Q_{\mathrm{u}}}(Q_{\mathrm{u}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}}) \\ &\quad \times q(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}}). \end{aligned} \tag{5.142}$$

The joint VB marginal of $\mathbf{Z}$ and $\mathrm{U}_{\mathrm{new}}$, $q(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}})$, is not in a closed form. A Gaussian approximation is computed (as described in Section 5.3.2), denoted as $q^g(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}})$ that can be further used as an approximation to $P_G(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}})$.

Laplace approximation of the posterior distribution of $\mathrm{X}_{\mathrm{new}}$:

Combining all the approximations from Eq. 5.140 and 5.142, the marginal posterior distribution $P(\mathrm{X}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathbf{x})$ at a given value $\mathrm{X}_{\mathrm{new}} = i$, is defined as:

$$P(\mathrm{X}_{\mathrm{new}} = i|\mathbf{y}_{\mathrm{new}}, \mathbf{y}) \propto \left.\frac{P(\mathbf{y}_{\mathrm{new}}|\mathbf{Z}, \mathrm{U}_{\mathrm{new}}, i)q^g_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, i)P(\mathrm{U}_{\mathrm{new}}|\mathbf{y}, \mathbf{x})}{q^g(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}} = i)}\right|_{\substack{\mathbf{Z}=\mathbf{Z}^\star(\mathrm{X}_{\mathrm{new}}) \\ \mathrm{U}_{\mathrm{new}}=\mathrm{U}_{\mathrm{new}}\star(\mathrm{X}_{\mathrm{new}})}} ; \; i = \tag{5.143}$$

In the numerator of the above expression, the VB marginal $q^g_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, i)$ is found at the forward stage of the inference problem. The joint VB marginal, $q^g(\mathbf{Z}, \mathrm{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, \mathrm{X}_{\mathrm{new}} = i)$, is a low dimensional distribution, and therefore, can be computed very quickly

for each value i of X_{new}.

Result

The results of the VB approximation are shown with a set of simulated data from a zero-inflated Poisson latent random effect model. To simulate values on the tri-variate response variable of the model, a set of five hundred values of the (uni-variate) explanatory variable, from one to fifty, are drawn from a discrete uniform distribution. The true value of the latent variables (corresponding to the tri-variate response variable) are generated from independent GMRF distributions with Gamma variate precision parameters. Five hundred values of tri-variate random effects are generated from independent multivariate normal distributions with zero means and a common precision following a Wishart distribution. A set of five hundred counts on tri-variate response variable are generated from the independent ZI-Poisson latent random effect model (described in the chapter) with mean parameters as an exponential function of random effects and latent variables. This simulated set of data contribute to the estimation of latent variables and the random effects at the forward stage of the inference.

For the inverse estimation of the unknown explanatory variable corresponding to the new values on the response variable, a set of one hundred values on the explanatory variable and the same number of counts on the response variables are generated from the models described above. The true values of explanatory variable are used for the accuracy of the approximation via cross validation technique with 95% HPD region of the unknown explanatory variable.

Results for the zero-inflated Poisson latent random effect model, are shown in Fig. 5.9, 5.10 and 5.11. In Fig. 5.9, the mean of approximations of the true posterior mean of the latent variables $\mathbf{Z}$ by the VB method is compared with the approximation by INLA. It can be seen the figure that INLA smoothes the mean values of the latent variable and the mean values and the 95% HPD region of the approximation are continuous over the grid values defined by the discrete values on the explanatory variable $\mathbf{x}$. But comparatively, the VB-means of latent variables $\mathbf{Z}$ stay close to their true values. The 95% VB HPD region is tight at the locations

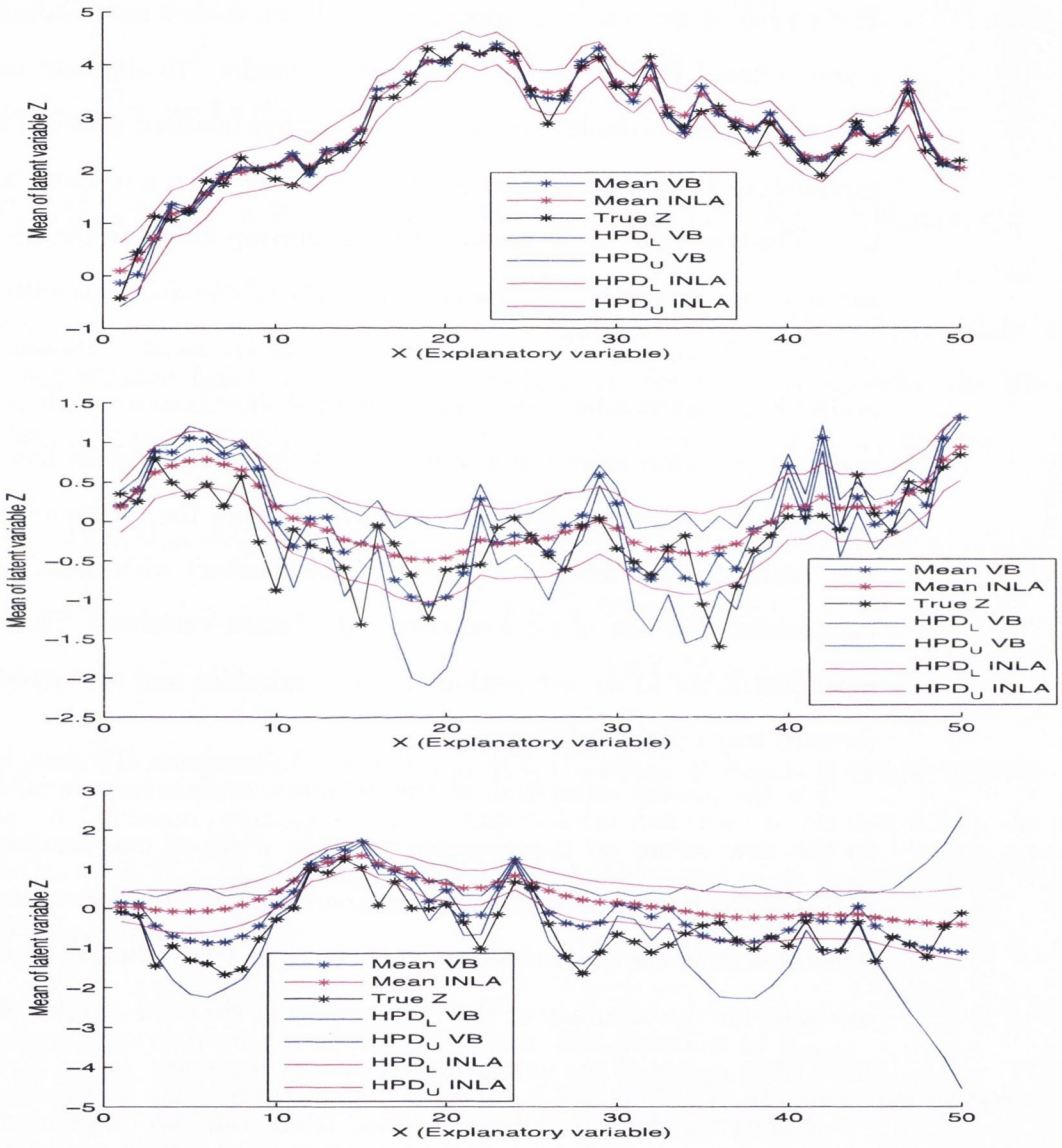

Fig. 5.9: The comparison of true value (black), VB-mean(blue), mean (estimated by INLA) (magenta) of the latent variables $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ (defined over the discrete values of the explanatory variable $\mathbf{x}$) corresponding to the response variable $\mathbf{Y}_1$ (the upper one), $\mathbf{Y}_2$ (the middle one) and $\mathbf{Y}_3$ (the bottom one) respectively for a zero-inflated Poisson latent random effect model.

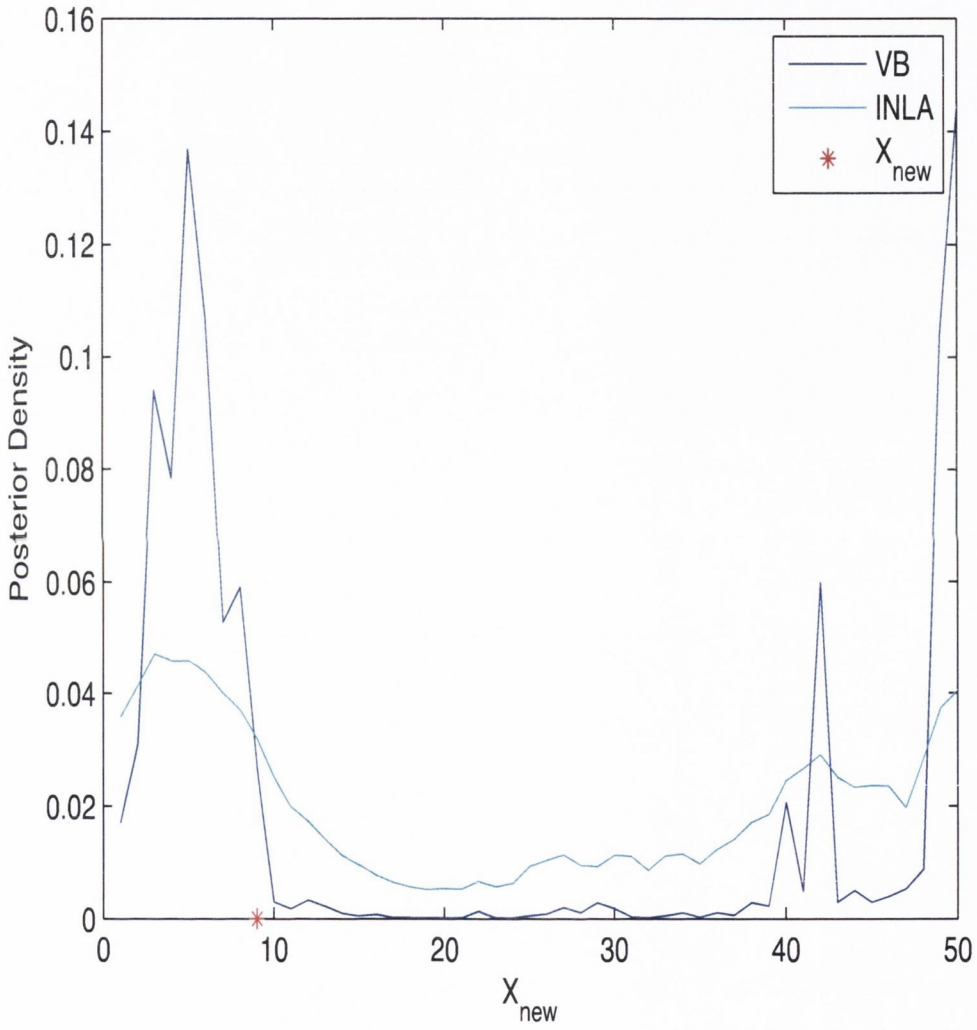

Fig. 5.10: The comparison of the approximation by the VB method and by the INLA of true posterior distribution of X_{new} given $\mathbf{Y}_{new} = (0, 6, 0)$ of a zero-inflated Poisson latent random effect model.

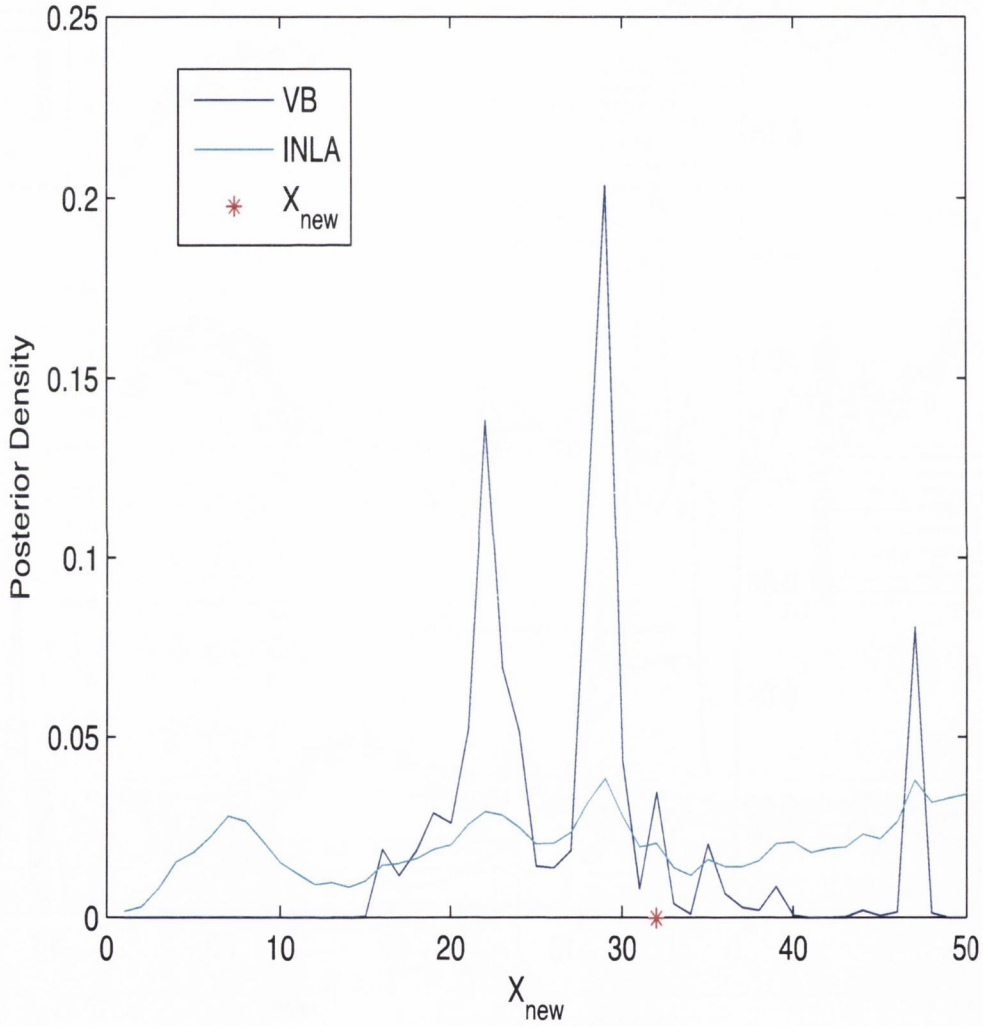

Fig. 5.11: The comparison of the approximation by the VB method and by the INLA of true posterior distribution of X_{new} given $\mathbf{Y}_{new} = (116, 2, 0)$ of a zero-inflated Poisson latent random effect model.

where the data are available in abundance, whereas for less informative or less amount of data the region is large. The variance of the VB approximation is small compared to the variance of the approximation by INLA. The under-estimation of the VB-variance and the non-smooth behaviour of the VB mean are due to the independence assumption of the VB method. If data is less informative, a VB approximation may depend on the choice of prior distribution. Since the data drawn from the zero-inflated Poisson latent random effect model is less informative (with excess of zero counts), the VB approximation of the mean of the latent variables are close to the prior mean, As the prior assumptions are similar to the true model, the VB means of responses are close to their true values, as evident from Fig. 5.9.

Fig. 5.10 and 5.11 presents the multi-modal density of a new unknown explanatory variable $\mathrm{X}_{\mathrm{new}}$ given a test data $(0, 6, 0)$ and $(116, 2, 0)$ of the response variable $\mathbf{Y}_{\mathrm{new}}$ respectively. The first set of test data on $\mathbf{Y}_{\mathrm{new}}$ represent the lack of information in data. The VB posterior distribution of $\mathrm{X}_{\mathrm{new}}$ is rather irregular because of the lack of smoothness in the VB marginal-means of the latent variable $\mathbf{Z}$. A loss of accuracy is due to independence assumption of the VB method and other approximations applied in the process of reconstruction. The posterior distribution given less informative data (in shown Fig. 5.10) is flat compared to that given some informative data (shown in Fig. 5.11). The approximations of the posterior distribution by INLA (given the two sets of data) are very flat. Also, the accuracy of the approximation by INLA with 95% HPD region is 100% which indicates that the method needs more information through data. Whereas, the accuracy of the VB approximation by a cross-validation technique with 95% HPD region is only 96%. Hence it can be said that for this zero-inflated Poisson latent random effect model, the VB method outperforms INLA in the inverse estimation.

As there are many approximations used at both stages of inference, it is an important question that which approximation among others is the most responsible for non-smooth approximation (inverse estimation). The standard result of the VB approximation is that it under-estimates the variance, hence the VB approximations of the latent variables are not smooth. The VB approximation of the inverse estimation uses a Gaussian approximation and the MAP estimates of power index of

	Forward stage	Inverse stage
VB	27.748 sec	4.6 sec
INLA	231.015 sec	3.339 sec

Table 5.3: The comparison of the computational time of VB method and INLA at both stages for zero-inflated Poisson latent random effect model. At the inverse stage the computation time shows the time taken by the method for the inverse estimation for all test data.

ZI-Poisson likelihood which results in over-fitting of the model. The approximation is bad specially for zero counts since a Gaussian approximation given 'no or little information' (zero counts) may be inaccurate. This requires to replace the Gaussian approximation with a suitable approximation for near-zero counts and to approximate the power index parameter for a better approximation for inverse estimation.

Computational time:

Table 5.3 shows the computational time taken by INLA and the VB method at both stages. Just as for other two models, the VB method provides good results in very less time as compared to INLA.

5.4 Discussion

The chapter has shown that the VB method is useful for fast implementation of Bayesian inference for multidimensional inverse latent regression problems. The accuracy and the tractability of VB approximations depend on the model assumptions and the nature of data available for the study. VB approximations can depend on prior assumptions for models with noisy and less informative data. It might be intractable for non-conjugate and non-exponential distributions. The intractability issue of the method for complex models can be solved with the use of further approximations such as Gaussian.

The VB method uses the posterior independence assumption and provides quick solutions to estimation problems. The simplicity and low computational cost of the method makes it desirable to use for multi-dimensional inverse latent regression problems. Three different types of model have been used with increasing complex-

ity for a successful application of the VB method: Poisson latent regression model, Poisson latent random effect regression model and zero-inflated Poisson latent random effect model. The assumption of random effects in the models makes them less complex though it increases the dimension of the problem. Therefore, all three models are good examples of complex non-conjugate exponential or non-conjugate non-exponential latent models which present difficulty against the VB method for tractable VB approximations.

To solve the intractability issue of the VB method, further approximations e.g. a Gaussian approximation or plugging-in of posterior modes, are applied. These approximations seem to work well for non-zero counts. But, they perform poorly for zero or close-to-zero counts. It is shown that the method provides good approximation if the data is very informative. For less informative data, as the approximation may depart from Gaussianity, the approximation is bad. The bad approximation may also be a result of the complex structure of the model or the independence assumption of the method to model such complexity or choice of prior for less informative data. It is shown that the performance of the VB method in terms of accuracy and computation time is good in comparison to INLA for the complex models used in the chapter. It is worth checking if choice of prior has an important effect on the VB approximation.

Due to the unsuitability of the VB method for index type variables, the Laplace approximation is used for the inverse estimation for the zero-inflated Poisson latent random effect model. For prediction of the non-index type of unknowns, it may be suggested to use the VB method for quick solutions.

It is shown that the VB approximation and results from INLA depend on the structure of model and availability of data which is also experienced by Salter-Townshend (2009) for the use of INLA for inverse estimation for latent regression models discussed in the chapter.

Chapter 6

VB approximation for Palaeoclimate Reconstruction problem

The chapter describes a VB approximation for inverse latent regression through the palaeoclimate reconstruction problem. The reconstruction problem provides a motivating example of a complex inverse latent regression problems to check the usefulness and accuracy of the VB method. In the previous chapter, the intractability issue of the method was discussed for non-conjugate inverse latent regression models. This chapter proceeds with the same intractability issue of the method and describes how to deal with the problem for a successful VB approximation to multi-dimensional, complex inverse latent regression models.

6.1 Introduction

The palaeoclimate reconstruction problem (Haslett et al., 2006) is the motivating example of an inverse latent regression problem for this thesis. In the reconstruction problem, past climate is inferred from a proxy, such as data on relative abundances of different types of fossil pollen that are believed to respond differently to climate. The pollen respond smoothly to climate, hence it is a useful proxy for climate. A convenient model for pollen data as a function of climate makes use of a smooth

response surface. Responses of the pollen to climate are latent in nature. Therefore, in order to infer past climate, they should be modeled through pollen data as a smooth curve of climate. Each taxon has its own response surface to be modeled at all the locations in the climate space. Then the combined information on the response surfaces (of all pollen) is used to estimate past climate and the uncertainty in the estimate.

The palaeoclimate reconstruction problem can be modeled through inverse latent regression. If $\mathbf{y}$ represents a set of modern data of pollen abundance, $\mathbf{Z}$ the corresponding latent responses and $\mathbf{c}$ is a set of known values of climate, the model can be defined mathematically as:

$$\mathbf{y} = f(\mathbf{Z}(\mathbf{c}); \theta). \tag{6.1}$$

The aim of the reconstruction is to predict past (or future) climate, denoted as $\mathbf{C}^a$, given past and current data $(\mathbf{y}^a, \mathbf{y}, \mathbf{c})$ through the knowledge of $\mathbf{Z}$ and θ.

A VB approximation of the reconstruction problem is discussed in Section 6.2. Vatsa & Wilson (2010) has made an attempt in this. The authors explained the application of the VB method for the palaeoclimate reconstruction in a similar way that we described later in the chapter. They carry out a simpler VB approximation to the reconstruction of climate and to the model fitting in only one stage. However, their method may make the reconstruction process very slow since it may require to carry out the whole process of reconstruction every time a prediction is needed. Whereas, the method of splitting the prediction problem in two stages can be very useful. Once a forward model is studied, it can be saved and may be used further for many future predictions. The authors also suggested to use a more informative data to increase the accuracy of the VB approximation.

Apart from this paper, no other literature is found for the application of the VB method for palaeoclimate reconstruction problems. Haslett et al. (2006) and Salter-Townshend (2009) used inverse problems as a method of reconstruction of climate. Haslett et al. (2006) commented on the application of the MCMC method in the climate reconstruction that it suffers from mixing and convergence. Salter-

Townshend (2009) applied the Integrated Nested Laplace approximation (INLA) for the reconstruct of climate.

6.1.1 Description of Palaeoclimate Model

The RS10 data set is taken from Allen et al. (2000) which describes the nature of the palaeoclimate data on pollen and climate. The data set consists of two types of data: modern and fossil or ancient. The modern data set provides data on current climate at 7742 locations around the world, as well as observations of pollen counts of different taxa at those locations. This allows us to fit forward inference and infer the relationship between pollen and climate. The ancient data has only count data on the fossil pollen (extracted from lake sediment cores) and the corresponding ancient climate is missing. It is natural to build a regression model for the response of pollen to climate, and use this by inverse regression to infer ancient climate from pollen. Thus, it is an interesting problem to describe through an inverse regression.

The aim of the chapter is to explain the VB approximation for inverse latent regression with the help of the palaeoclimate problem. It does not attempt to provide real reconstruction of ancient climate, hence it neglects the fossil data from the study. For the study of VB approximation to the problem, the modern data is divided in two parts, training data and test data. The training data set is used to learn the palaeoclimate model and test data on pollen are used reconstruct the corresponding climate. The performance of the reconstruction of climate is checked using test data on climate.

Description of palaeoclimate data

There may be several possible climate variables to study climate behaviour. Haslett et al. (2006) considered mainly two climate variables: MTCO (Mean temperature of coldest month) and GDD5 (Growing degree days above $5°C$). There are 28 pollen types considered in palaeoclimatology. The data on the different taxa at the various locations in the climate space, are typically in counts. Due to huge variation in the tolerance limit of the taxa to different climate at different locations, the count data on taxa are highly over-dispersed with an abundance of zero counts. The zero-

inflated behaviour of the count data can be understood with diagrams shown in Fig. 6.1. The diagram represents the histograms of the count data on the taxa, Abies and Carnipus, that shows that the most of the data points are zero (about 66% for Abies and 91% for Carnipus).

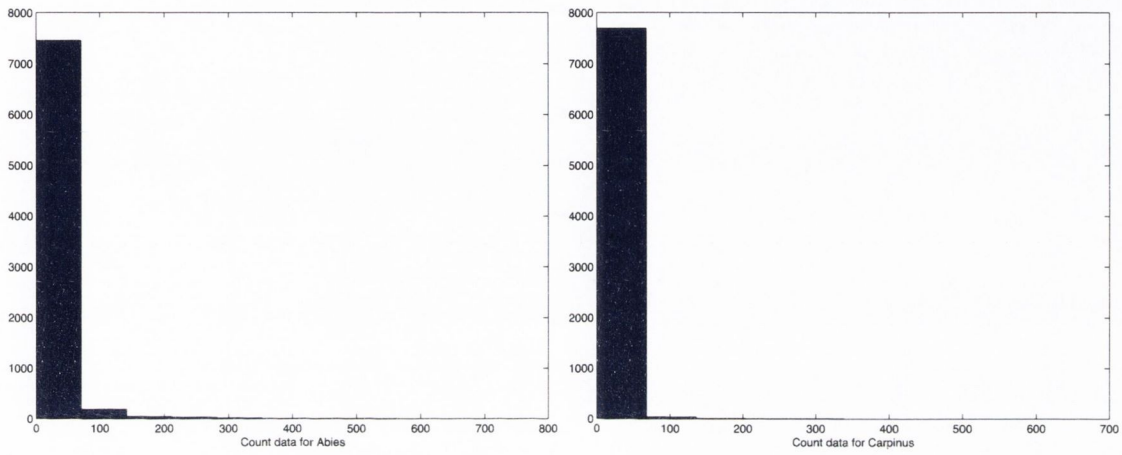

Fig. 6.1: Histogram of the count data for (left) Abies and (right) Carnipus.

Generally, counts are modelled with a Poisson distribution. If zero-inflated data are modelled with a Poisson distribution, then the mean of the distribution will be underestimated and so the variance. Salter-Townshend (2009) suggested to use a zero-inflated Poisson distribution (Ridout et al., 1998) to model the data with excess of zeros.

Representation of the model via a DAG

The climate model considered in Haslett et al. (2006) is presented via a DAG in Fig. 6.2. The variables used in the DAG to represent the palaeoclimate model stand for

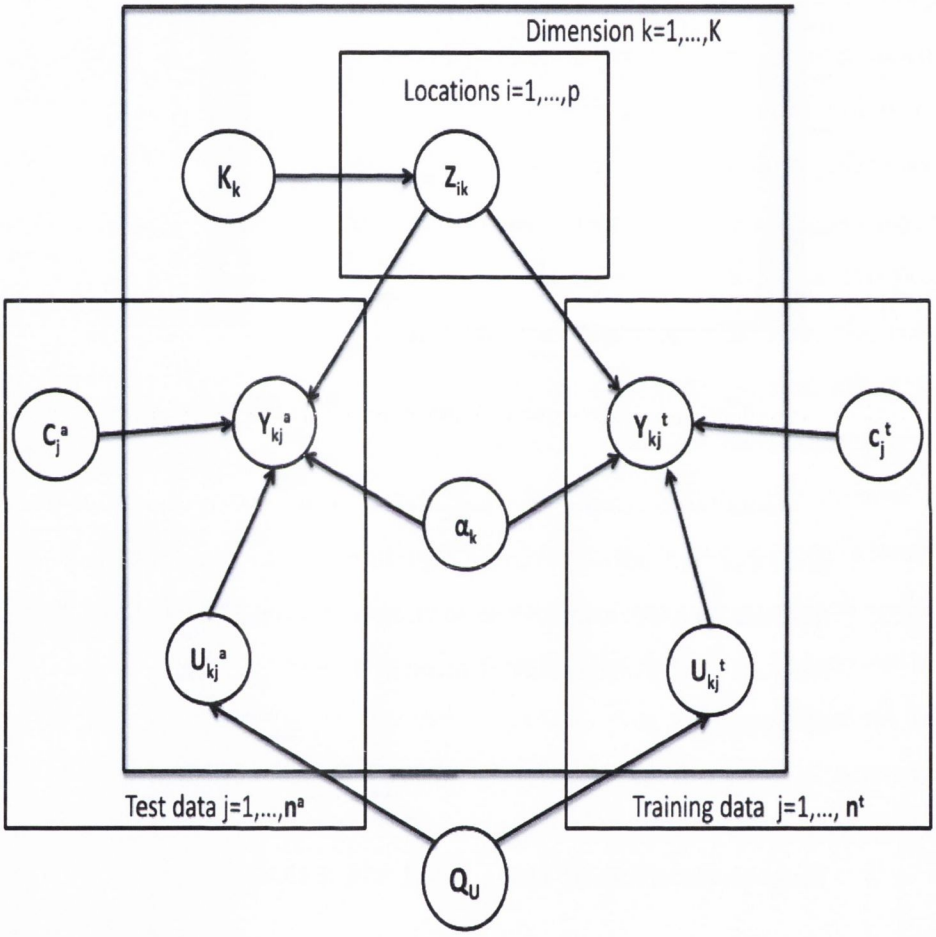

Fig. 6.2: A DAG representing the palaeoclimate model of Haslett et al. (2006).

following:

p : Number of discretized climate locations in the study,

K : Number of pollen taxa considered in the study,

n^t : Number of sample points in training data,

n^a : Number of sample points in test data,

Y^t_{kj} : j^{th} training count data on k^{th} taxa for which the climate is known,

Y^a_{kj} : j^{th} test count data on k^{th} taxa for which the climate is to be reconstructed,

c^t_j : data on climate corresponding to j^{th} training data $\mathbf{Y}^t_j$,

C^a_j : unknown climate corresponding to j^{th} test data $\mathbf{Y}^a_j$,

Z_{ik} : latent response of k^{th} taxa climate with discrete value equal to $i = 1, \ldots, p$,

U^t_{kj} : Random effects (to simplify the model) corresponding to j^{th} training data on k^{th} taxa,

U^a_{kj} : Random effects (to simplify the model) corresponding to j^{th} test data on k^{th} taxa,

κ_k : smoothing parameter in the GMRF prior over $\mathbf{Z}_k$ corresponding to k^{th} taxa,

α_k : Power law index in the zero-inflated Poisson likelihood of data on K^{th} taxa,

$Q_{\mathbf{U}}$: precision parameter in the prior over $\mathbf{U}_j$ $\forall j$.

The superscripts t and a stand for training and test respectively. Notations without any superscript stand for both the types of data, training and test.

Reconstruction of climate is highly dimensional and a complex latent regression problem. To simplify the problem there are some adjustment done in the original climate model:

- To control the size of the reconstruction problem, climate locations are discretized and limited to the size of p. Hence, the problem is now defined with multi-observations at one climate location instead of one observation per climate location.
- Data on climate are discretized and assumed to be on a regular grid of size p.
- Latent responses $\mathbf{Z}$ are assumed to be defined on the grid, hence they are

indexed by discretized climate, that is the response of the k^{th} taxon in the j^{th} observation is $Z_{c_j k}$ e.g. climate location is c_j.

The term $Z_{c_j k}$ models the response of j^{th} observation of k^{th} taxon to the climate $\mathbf{C}$ under independence, while U_{jk}, the random effect corresponding to j^{th} observation of k^{th} taxon, induces dependence between taxa.

The aim of the reconstruction problem is to infer unknown climate C^a (discretized) for each test data $\mathbf{Y}^a$ at inverse stage borrowing the knowledge of unknown variables inferred at forward stage of inference.

Again, we note that the inverse regression is inferring the index C^a of the latent process as in Chapter 5.

Likelihood

One climate variable model is assumed. An initial model assumes the counts are independent given a latent response, and that the latent variables are indexed by climate and are also independent. Dependence between counts for different taxa is introduced by adding a random effect into the model. Random effects also capture over-dispersion in the count data. A zero-inflated Poisson distribution to model count data is defined as:

$$P(\mathbf{y}|\mathbf{Z}, \mathbf{U}, \mathbf{C}) = \prod_{k=1}^{K} \prod_{j=1}^{n} ZIP(y_{kj};\ \lambda_{kj}, q_{kj}), \tag{6.2}$$

$$\text{ZIP}(y_{kj};\ \lambda_{kj}, q_{kj}) = \begin{cases} 1 - q_{kj} + q_{kj}\, e^{-\lambda_{kj}}, & \text{if } y_{kj} = 0; \\ q_{kj}\text{Poiss}(y_{kj};\ \lambda_{kj}), & \text{if } y_{kj} > 0 \end{cases}$$

where y_{kj} is the j^{th} count for k^{th} taxon. The term $(1 - q_{kj})$ is the probability of observing essential zero counts. Salter-Townshend (2009) used a power law functional relationship to define the probability q_{kj} as in terms of λ_{kj}:

$$q_{kj} = \left(\frac{\lambda_{kj}}{1 + \lambda_{kj}}\right)^{\alpha_k}, \ \forall k,\ j, \tag{6.3}$$

where $\lambda_{kj} = \exp(Z_{C_j k} + U_{kj})$ is mean of the Poisson term in the likelihood. The

term α in the ZI-Poisson likelihood is called the power law index. The properties of α is already defined in Chapter 5 in Section 5.1 in the definition of zero-inflated latent random effect model.

Prior distribution

There are three set of unknowns to be inferred in the forward problem: responses $\mathbf{Z}$, random effects $\mathbf{U}$ corresponding to the training data set and unknown parameters θ. The prior distribution assumed over unknowns are as follows:

- Latent responses $\mathbf{Z}$ are assumed over a regular grid. Hence, an independent GMRF prior by Rue & Held (2005), can be a suitable prior distribution of $\mathbf{Z}$:

$$P(\mathbf{Z}|\kappa) = \prod_{k=1}^{K} P(\mathbf{Z}_k|\kappa_k), \tag{6.4}$$

$$= \prod_{k=1}^{K} GMRF(\mathbf{Z}_k; \underline{0}, Q_{\mathbf{Z}_k^{-1}}), \tag{6.5}$$

where the latent response variable, $\mathbf{Z}_k$; $k = 1 : K$, (indexed by 1-dimensional climate variable) is a GMRF with mean vector zero and precision $Q_{\mathbf{Z}_k}$; $k = 1 : K$. The precision $Q_{\mathbf{Z}_k}$ is defined as:

$$Q_{\mathbf{Z}_k} = \kappa_k R \, ; k = 1 : K, \tag{6.6}$$

$$R = \begin{pmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix},$$

where κ_k is an unknown precision parameter. Precision $Q_{\mathbf{Z}_k}$ (defined for one dimensional climate) is structured such that the GMRF density of $\mathbf{Z}_k$ is proper. Salter-Townshend (2009) assumes an intrinsic GMRF (improper) prior for re-

sponses. But, it is safer to assume a proper prior distribution to avoid any chance of an improper VB approximation of posterior distribution, if given data is less informative. The sparse behavior of $Q_{\mathbf{Z}_k}$ is assumed for ease of computation.

- The prior distribution over $\mathbf{U}$ is assumed to be independent multivariate Gaussians:

$$P(\mathbf{U}^t|Q_{\mathrm{U}}) = \prod_{j=1}^{n} \mathrm{MVN}_K\left(\mathbf{U}_j; \underline{0}, Q_{\mathrm{U}}^{-1}\right). \tag{6.7}$$

  The term Q_{U} is the unknown precision in the prior.

- The parameter θ is a set of unknown: zero-inflated (power index) parameters α_k; $k = 1:K$, hyper-parameters κ_k; $k = 1:K$ and precision Q_{U}. For ease of computation, conjugate prior distributions are assumed over components of θ:

$$P(\alpha) = \prod_{k=1}^{K} \mathit{Log\text{-}Normal}(\alpha_k;\ \mu_{\alpha_k}, \sigma^2_{\alpha_k}), \tag{6.8}$$

$$P(\kappa) = \prod_{k=1}^{K} \mathit{Gamma}(\kappa_k;\ a_k, b_k), \tag{6.9}$$

$$P(Q_{\mathrm{U}}) = \mathrm{Wishart}\,(df, V). \tag{6.10}$$

  The hyper-parameters in the above prior distributions are assumed to be known.

- At the inverse stage, climate $\mathbf{C}_j^a$ is to be predicted for a new (test) pollen data $\mathbf{y}_j^a$. To account for a lack of information about climate behaviour, a uniform prior is assumed for $\mathbf{C}_j^a$:

$$P(\mathbf{C}_j^a) \propto 1\ \forall j. \tag{6.11}$$

6.2 VB approximation to the palaeoclimate reconstruction problem

The palaeoclimate model is a real example of a zero-inflated Poisson latent random effect model which has already been studied for inverse estimation of unknown explanatory variable in Chapter 5. The climate variable of the palaeoclimate model is same as the explanatory variable of the zero-inflated Poisson latent random effect model of Chapter 5. VB approximation to inverse estimation problem for the model is already discussed in the chapter. In this chapter the result of the reconstruction problem with real and simulated will directly be presented following the discussion of Bayesian analysis of the problem presented already in Chapter 5.

6.2.1 Evaluation of the VB approximation

The approximations applied with the VB method for a tractable VB approximation (discussed in Chapter 5), may affect the accuracy of the result. An accuracy check should be performed at both stages of inference.

Evaluation of the approximation at the forward stage:
Usually, the predictive distribution is used to check accuracy of the model fitting. The accuracy check can be performed with a 95% predictive interval (PI) in terms of coverage in a cross validation experiment. The predictive distribution may often be computationally intensive or intractable. Also, the predictive interval of multidimensional predictive distribution may be very complicated. To avoid such

computational complexities, z-scores of each pollen count is computed:

$$\begin{aligned} \text{z-score} &= \frac{\text{Observed} - \text{Predicted}}{\text{Standard deviation of Predicted}}, \\ &= \frac{\mathrm{y} - E(\mathrm{y})}{\sigma(\mathrm{y})}, \\ &\sim N(0,1) \text{ asymptotically}, \end{aligned} \tag{6.12}$$

$$\text{where } E(\mathrm{y}) \approx \hat{q}\hat{\lambda}, \tag{6.13}$$

$$\sigma(\mathrm{y}) \approx \sqrt{\hat{q}\hat{\lambda}(1+(1-\hat{q})\hat{\lambda})}, \tag{6.14}$$

$$\hat{q} \approx \left(\frac{\hat{\lambda}}{1+\hat{\lambda}}\right)^{\hat{\alpha}}, \tag{6.15}$$

$$\hat{\lambda} \approx \exp(\hat{\mathrm{Z}} + \hat{\mathrm{U}}). \tag{6.16}$$

Though the behaviour of z-scores might not be standard-normal (symmetric) for small counts, it is assumed that the scores are asymptotically normal. This makes the problem of finding 95% HPD interval, as for a standard-Gaussian variable the interval is defined between -2 to $+2$. A measure of the accuracy of the model fitting is presented via the percentage of the z-scores (standardized counts), lie within 95% acceptance region or between -2 to $+2$.

Evaluation of the approximation at the inverse stage:

To check the accuracy of the approximation at the inverse stage, 95% HPD region is computed as discueed in Chapter 5.

6.2.2 Results with simulated examples

The aim of considering simulated data to show the results for VB approximation to reconstruction problem is to check whether the approximation is affected by the choice of prior and the model assumption. To show the effect of the model assumption on the approximation, two models with one climate variable and three taxa are considered with following type of latent responses:

- one with linear responses and,
- another with irregular responses.

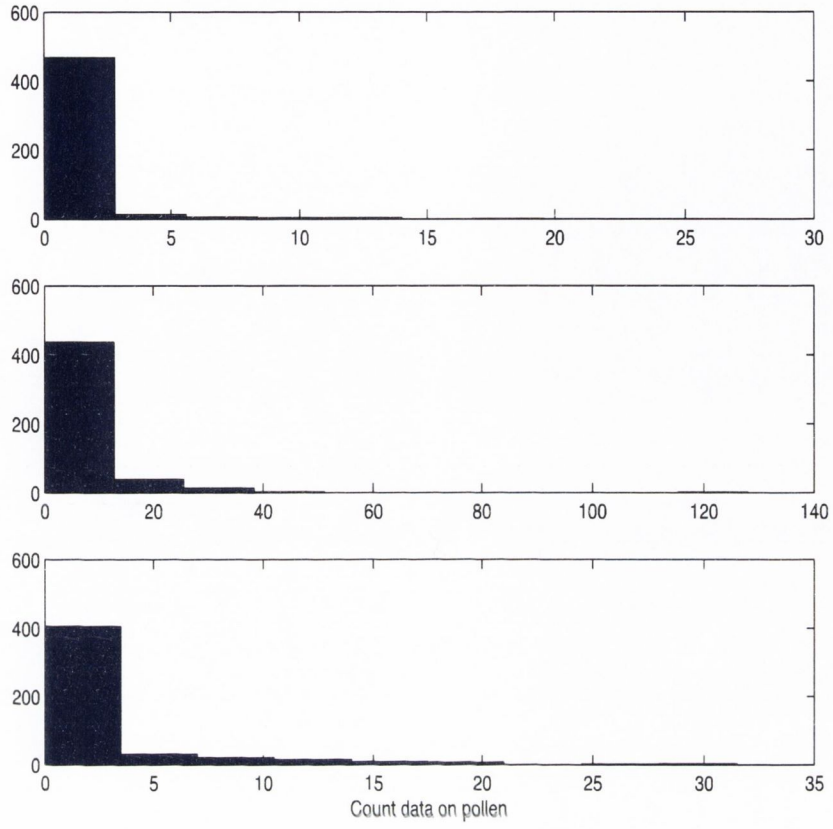

Fig. 6.3: Histogram of simulated training data on pollen generated from a ZI-Poisson model with linear responses.

A set of five hundred values of climate, from one to fifty, are drawn from a discrete uniform distribution. For the model with linear responses, fifty regular spaced values are assumed on responses of three taxa. The same number of values of (irregular) responses of the taxa are generated from independent GMRF distributions with Gamma variate precision parameters. Five hundred values of tri-variate random effects are generated from independent multivariate normal distributions with zero mean vectors and a common precision following a Wishart distribution. Two sets of five hundred counts on three taxa (corresponding to the two models) are generated from three independent ZI-Poisson distributions with mean parameters as an exponential function of random effects and response surfaces (linear and irregular respectively). These data on climate and pollen are presented as training data.

For a test data set, one hundred values on climate and the same number of pollen counts are generated from the models described above.

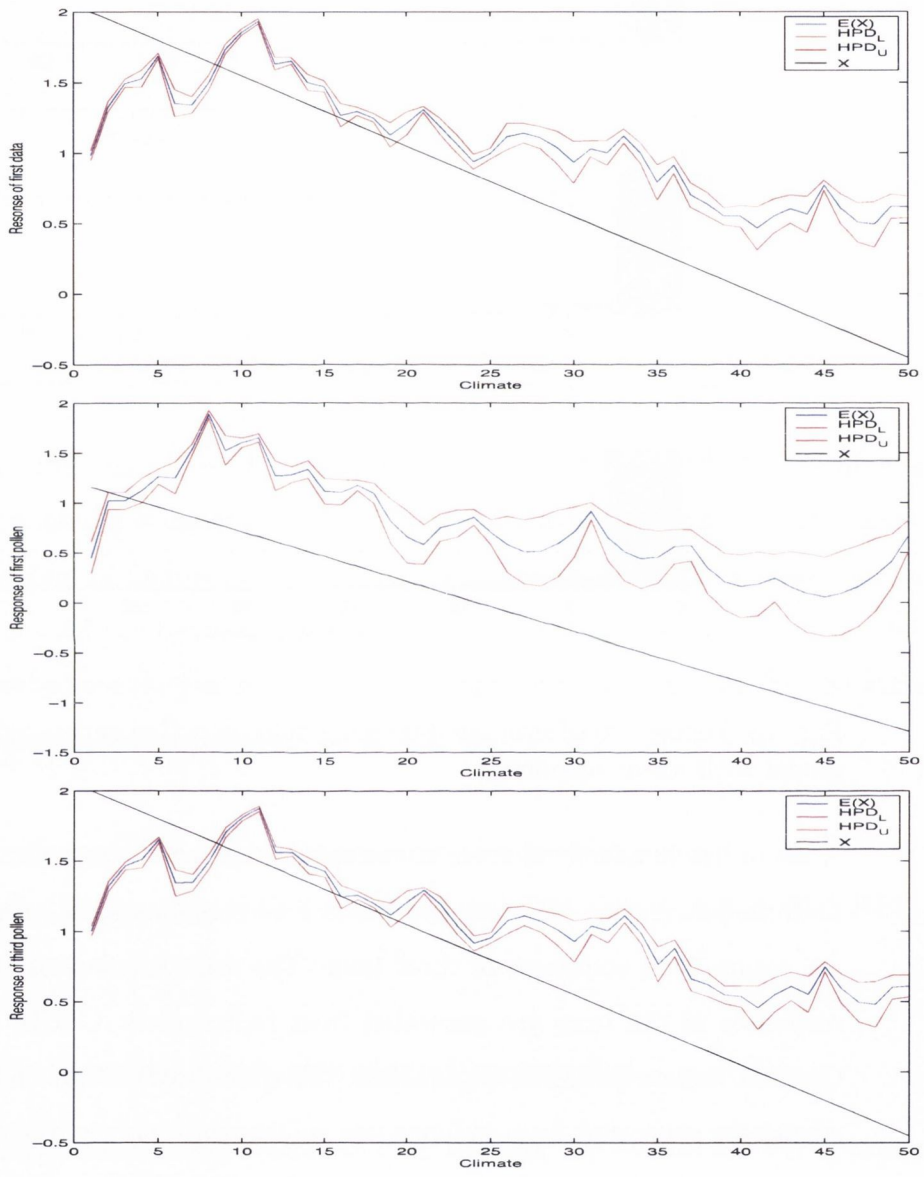

Fig. 6.4: The true (black) and the VB-mean (blue) of the responses for the model with linear responses, are compared. The uncertainty of the fitting of the responses are shown with 95% HPD region (red).

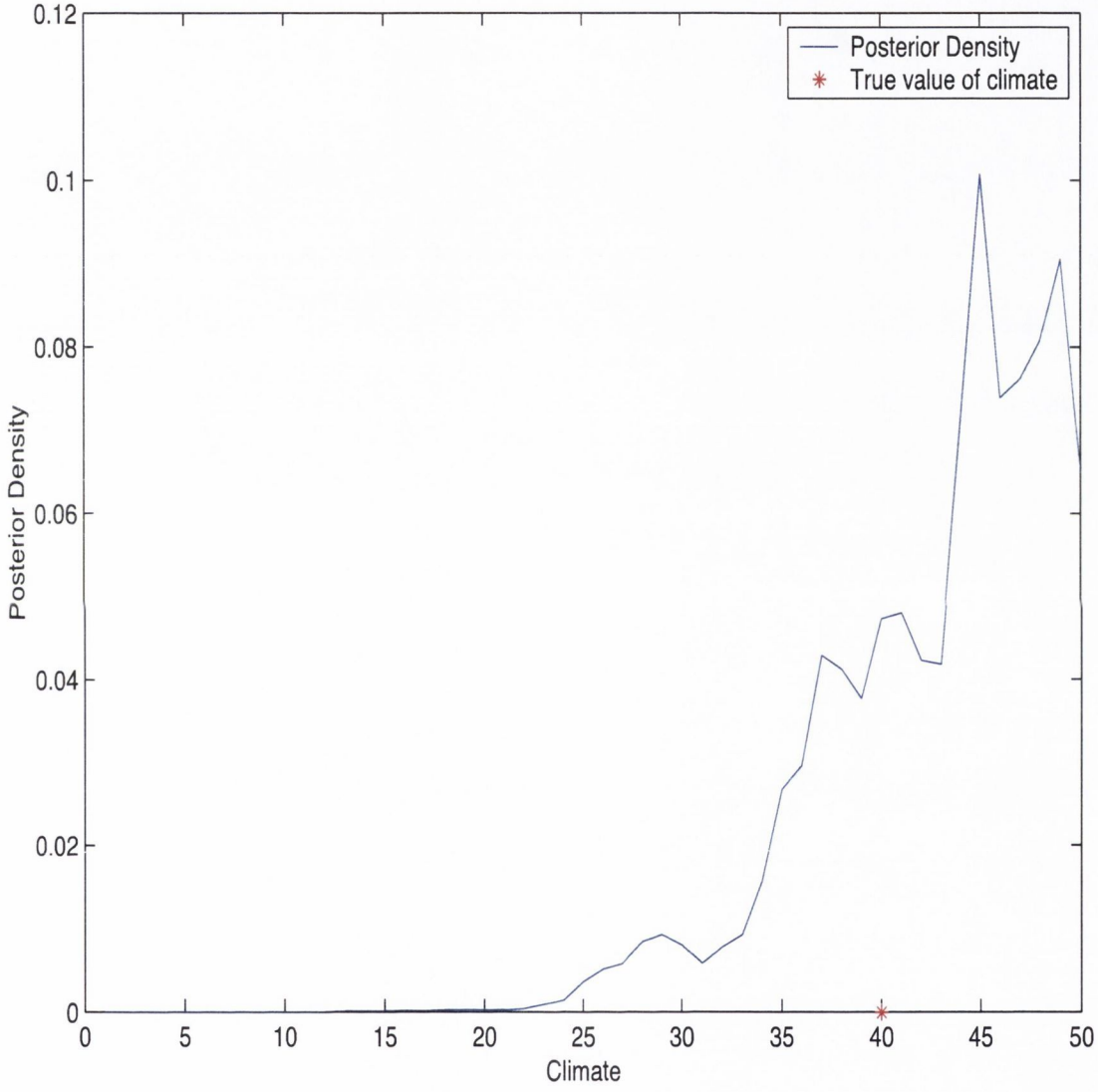

Fig. 6.5: Reconstruction at the inverse stage: Inverse prediction (reconstruction) of climate given a test data on three pollen (0, 65, 6) for the model with linear responses, is shown by means of the VB marginal of climate. The true value of climate, (∗), is displayed to compare with the posterior mode of the climate.

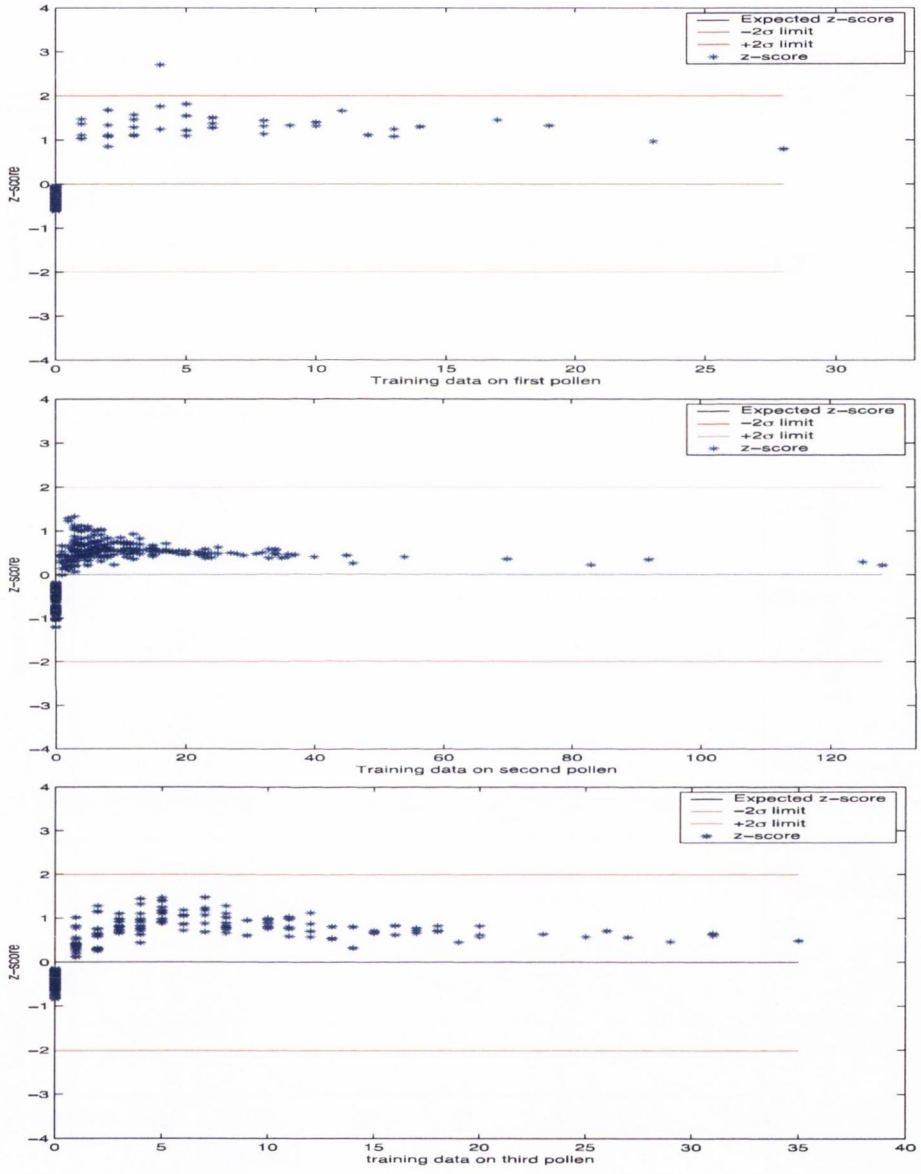

Fig. 6.6: Accuracy at the forward stage: The z-score of pollen data ($*$, standard error on y-axis) for the model with linear responses, are displayed against counts on taxa, with their true mean values, zero (black), and the true 95% confidence interval (red). Since the responses are over-estimated in Fig. 5.4, the z-scores are positive for most of the count data. The large variance of count data is due to fact that the variance of ZI-Poisson is greater than its mean.

Prior dist. of κ		VB-mean of κ of responses			Coverage at Forward stage			Coverage
Scale	Shape	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	Inverse stage
10	0.8	15.0411	22.1683	16.7692	99.8	100	100	92
8	0.5	18.8396	22.6923	20.7391	99.8	98	100	92
4	0.9	8.6007	14.3982	10.7124	99.8	100	100	91
1.5	1	6.1402	11.1061	8.3466	99.8	100	100	91

Table 6.1: The VB-mean of precision parameters of responses, κ, and the accuracy of the approximation at the forward stage and the inverse stage are displayed for the model with linear responses with different of values hyper-parameters of the precision parameters .

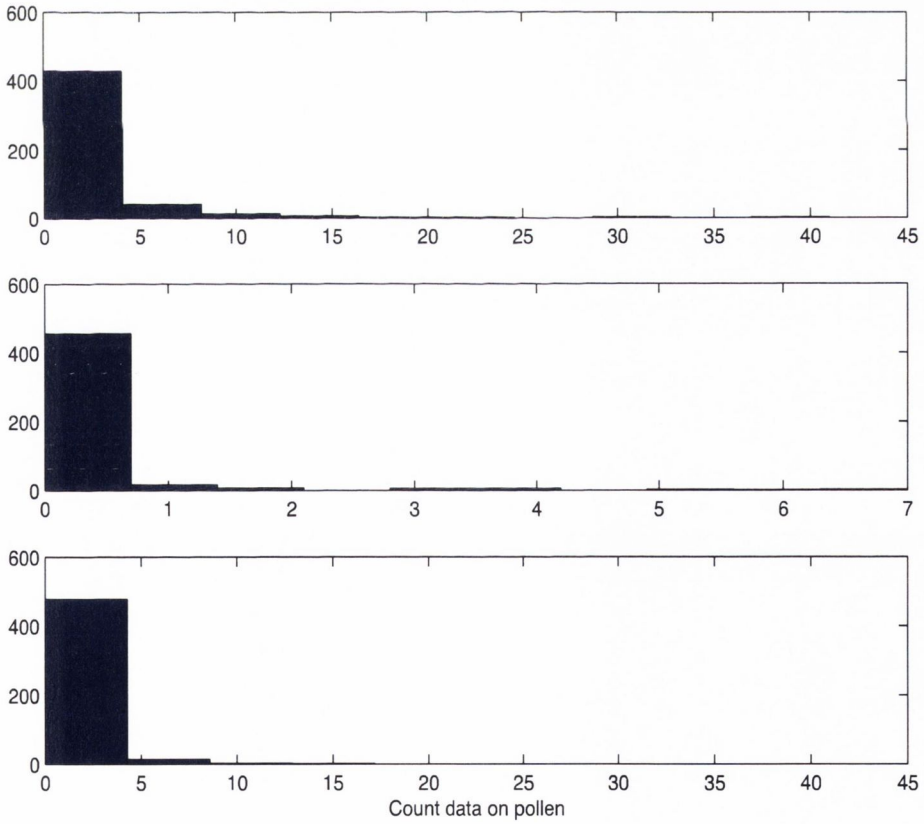

Fig. 6.7: Histogram of simulated training data on pollen generated from a ZI-Poisson model with irregular responses.

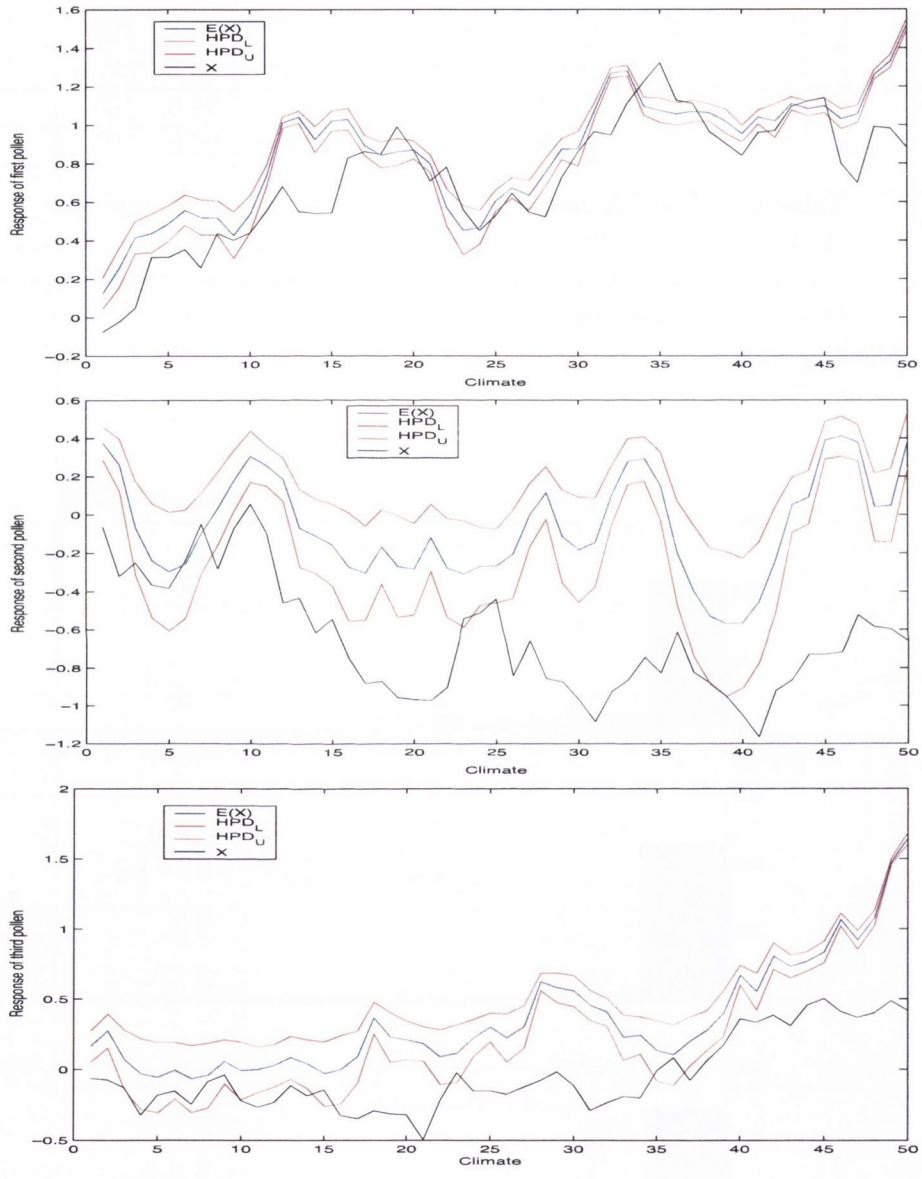

Fig. 6.8: The true (black) and the VB-mean (blue) of the responses for the model with irregular responses, are compared. The uncertainty of the fitting of the responses are shown with 95% HPD region (red).

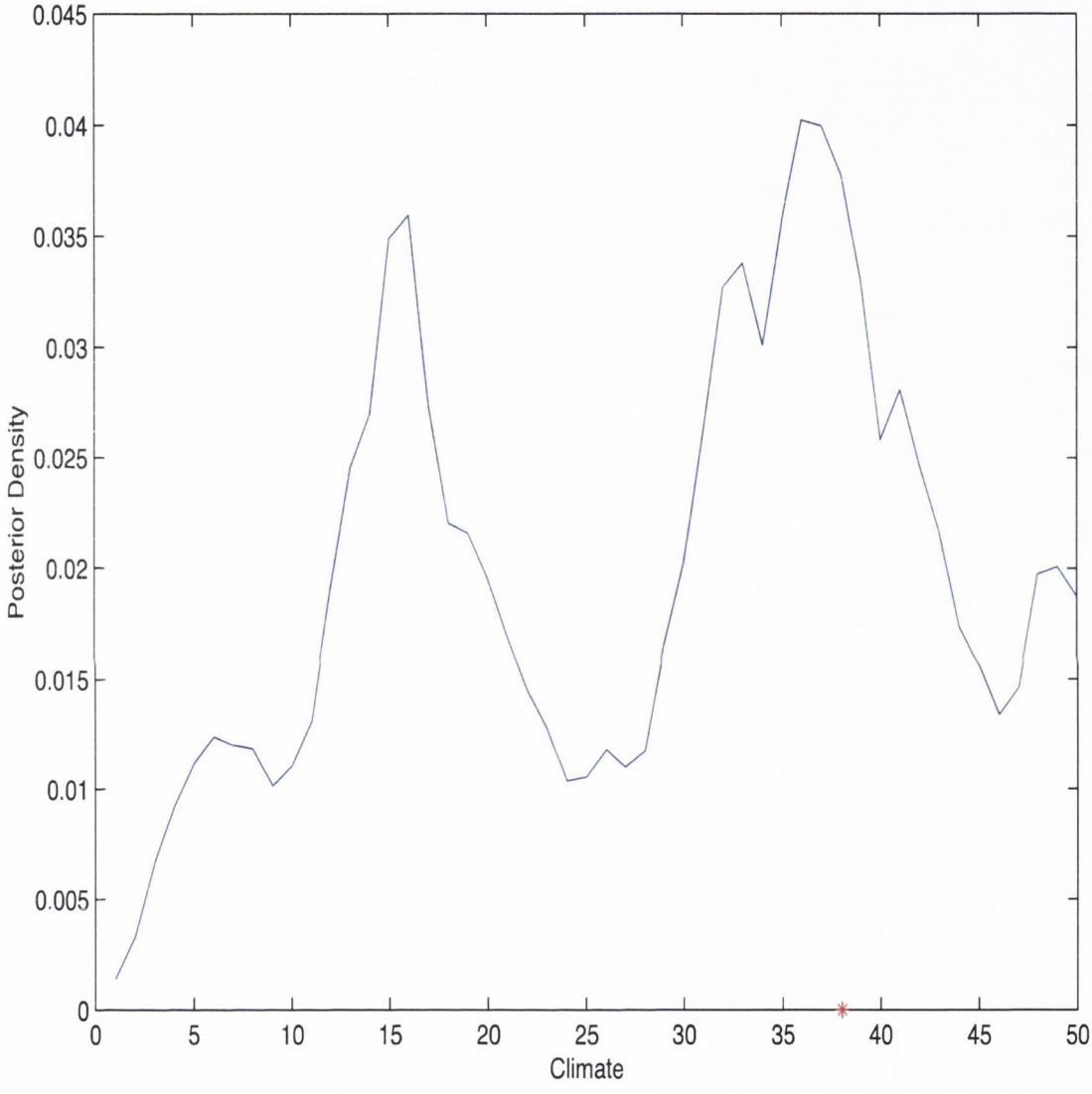

Fig. 6.9: Reconstruction at the inverse stage: Inverse prediction (reconstruction) of climate given a test data on three pollen (21,0,0) for the model with irregular responses, is shown by means of the VB marginal of climate. The true value of climate, (∗), is displayed to compare with the posterior mode of the climate. Since the responses are over-estimated in Fig. 5.8, the z-scores are positive for most of the count data. Also,the variance of ZI-Poisson is greater than its mean, therefore 95% HPD regon is very large.

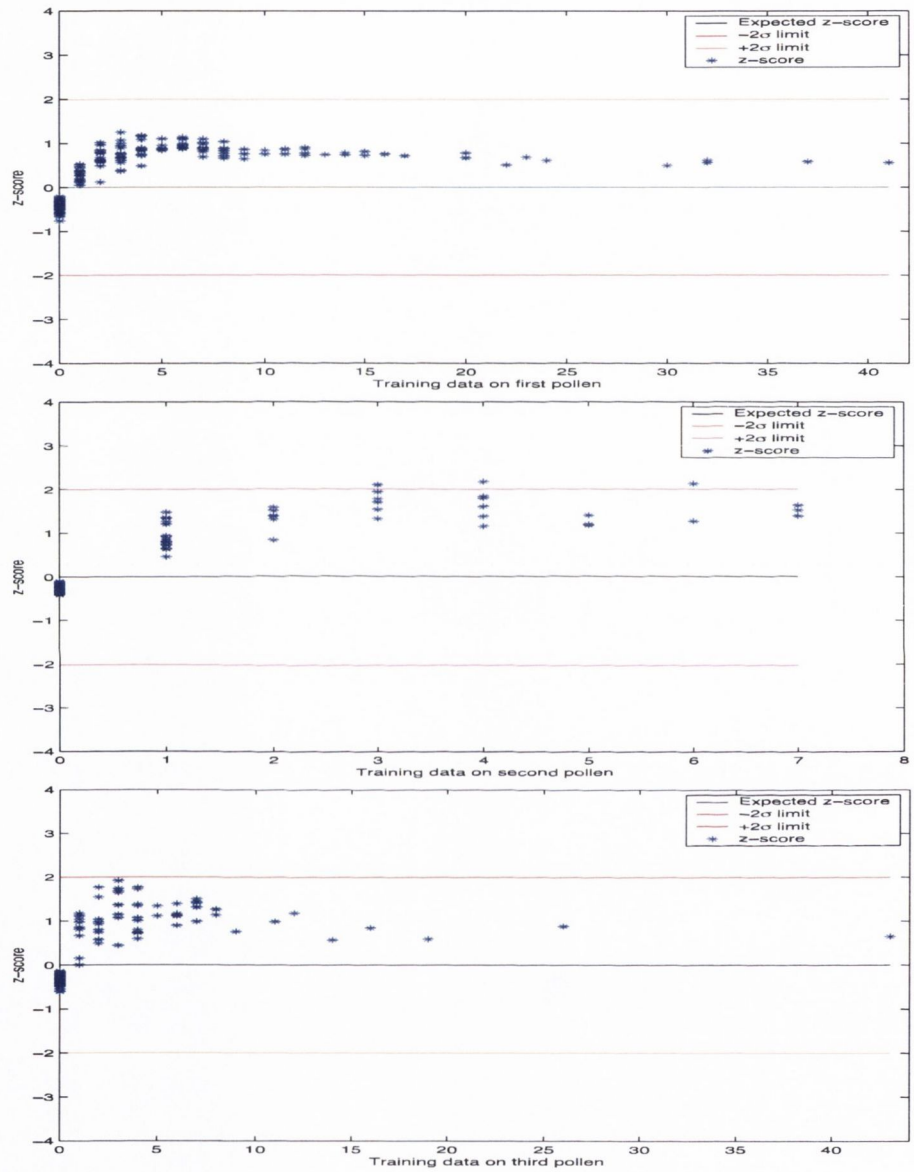

Fig. 6.10: Accuracy at the forward stage: The z-score of pollen data ($*$, standard error on y-axis) for the model with irregular responses, are displayed against counts on taxa, with their true mean values, zero (black), and the true 95% confidence interval (red).

Prior dist. of κ		VB-mean of κ of responses			Coverage at Forward stage			Coverage
Scale	Shape	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	Inverse stage
10	0.8	21.8645	15.8151	17.8225	100	100	100	89
8	0.5	28.648	21.1749	23.5094	100	99.2	100	90
4	0.9	14.3571	8.7509	10.688	100	99.4	100	90
1.5	1	11.3319	6.1162	7.8736	100	99.4	100	90

Table 6.2: The VB-mean of the precision parameters of the responses, κ, and the accuracy of the approximation at the forward stage and the inverse stage are displayed for the model with irregular responses with different of values hyper-parameters of the precision parameters.

Results for the model with linear response surfaces, are shown in Fig. 6.3–6.6 and Table 6.1. Figures 6.7–6.10 and Table 6.2 present the results for the model with irregular response surfaces. Fig. 6.3 and 6.7 show that the data on pollen are zero-inflated and highly over-dispersed. Effect of choice of prior is shown in Table 6.1 and 6.2

The VB approximation of the marginal posterior distributions of responses, as discussed at the forward stage, are shown in Fig. 6.4 and 6.8 for the two types the simulated data respectively. It can be seen that the VB-means of response surfaces are not as smooth as expected and the VB-variances are under-estimated. This might be due to the independence assumption of the VB method. Next, since zero-inflated data are less informative, the VB marginals of unknowns may largely depend on the choice of the prior distribution. For the model with irregular response surface, the prior assumptions are similar to the true response model. Hence, the VB means of responses are close to their true values, as evident from Fig. 6.8. Whereas the true and the prior model of linear response surfaces differ, hence, the responses are not well fitted.

The accuracy of the approximation at the forward stage is shown in Fig. 6.6 and 6.10 by means of z-scores of the pollen data for the models with linear and irregular responses respectively. The results shown in the figures indicates a 100% coverage of the model fitting. Since there are many approximations used in the model fitting, there should be some loss of accuracy in the model fitting. The z-

scores of pollen data are approximated by plugging-in the posterior modes of the unknowns (including the power index α of ZI-Poisson likelihood), and so are over-estimated which may lead to an accuracy of approximation showing over-fitting of the model for non-zero counts.

There is clearly an un-modelled trend in the z-scores which is due to over-fitting of the model, plugging-in the modes of the power index (zero-inflation) parameter of the ZI-Poisson likelihood and not modelling the zero counts well enough with a Gaussian approximation . The predicted value of a count for a ZI-Poisson model is approximated by:

$$\begin{aligned} \mathbb{E}(\mathrm{y}) &\approx \hat{q}\hat{\lambda}, \\ \hat{\lambda} &= \exp(\hat{\mathrm{Z}} + \hat{\mathrm{U}}). \end{aligned}$$

For an accurate prediction of a zero count, the posterior mode of the random effects $\hat{\mathrm{U}}$ should be a very small negative value (if $\hat{\mathrm{Z}}$ is comparatively large). The VB approximation of the random effects are carried out for each data point on taxa independently using a Gaussian approximation. The Gaussian approximation approximation given non-informative (zero counts) data may not be very accurate leading to a bad approximation of random effects for counts near zero.

A Gaussian approximation for large counts should lead to a very accurate approximation of the responses and the random effects which should result in an accurate approximation of the z-score for large counts. The z-score for non-zero counts (Fig. 6.6 and 6.10) show a positive bias which is a result of plugging-in the modes (or mean a Gaussian density) of the responses and the random effects in the definition (see Eq. 6.13–6.16). Since by Jensen's inequality,

$$h(\mathbb{E}(\mathrm{y})) \leq \mathbb{E}(h(\mathrm{y})), \tag{6.17}$$

where h is a convex function, resulting $\hat{\lambda}$ and $\hat{q}$ under-estimated (given in Eq. 6.15 and 6.16), hence over-estimating the z-scores.

Fig. 6.5 and 6.9 present the multi-modal climate density for a test data of three pollen (0,65,6) and (21,0,0) for both the models respectively. These test counts show

that some pollen are intolerant to extreme climate, while some could exist even in adverse climatic situation. A multi-modal climate distribution is desirable due to the multi-response behaviour of taxa. The posterior distribution of climate is rather irregular because of the lack of smoothness in the VB marginal means of the response surfaces. The accuracy of the reconstruction by a cross-validation technique with 95% HPD region is only 90%. The accuracy of reconstruction of climate is shown with 95% HPD region. A loss of accuracy is due to independence assumption of the VB method and other approximations applied in the process of reconstruction.

Table 6.1 and 6.2 show the effect of the prior on the accuracy of the VB-solution. Different experiments with the different values of hyper-parameters of the precision parameters of responses, show different results. When compared for the two models, the reconstruction of climate is more accurate for the linear response surfaces. Data on pollen for the linear response surfaces are less noisy, hence they are considered an easier case to reconstruct climate.

As there are many approximations used at both stages of inference, it is an important question that which approximation among others is the most responsible for the bad approximation. The standard result of the VB approximation is that it under-estimates the variance, hence the VB approximation of the responses are not smooth (see Wang & Titterington (2005)). The VB approximation of the reconstruction problem uses a Gaussian approximation and the MAP estimates of power index of ZI-Poisson likelihood which results in over-fitting of the model. The approximation is particularly inaccurate for zero counts since a Gaussian approximation given ńo or little information' (zero counts) may be inaccurate. This requires to replace the Gaussian approximation with a suitable approximation for near-zero counts, and to approximate the power index parameter, if inference for the reconstruction problem is improved. This could be found by running the same VB-experiment many times with different values of the power-law index and picking the most suitable value for the zero-inflated data. A good replacement for the Gaussian approximation for zero-counts needs to be investigated.

Prior dist. of κ		VB-mean of κ of responses			Accuracy at Forward stage			Accuracy
Scale	Shape	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	Inverse stage
10	0.5	12.5981	16.3055	7.7826	98.0857	96.0571	95.0429	83.5
10	2	7.1385	7.8698	4.9609	98.0857	95.9429	95.0714	85
4	0.5	9.4738	11.943	5.5091	98.0857	95.9571	95.0429	83.5
4	2	6.4153	5.9478	3.7664	99.3571	95.8	95.0571	89

Table 6.3: VB-mean of smoothing parameters of responses of the (true) palaeoclimate model, κ, and accuracy of approximation at forward stage and inverse stage are displayed for the model (with random true responses) with different of values hyper-parameters of the smoothing parameters.

6.2.3 VB solution with real palaeoclimate data

To demonstrate the VB approximation to the palaeoclimate reconstruction problem, three taxa among twenty eight, Alnus, Abies and Corylus, and a climate variable, GDD5 of the palaeoclimate model, are considered. Among 7742 modern data on taxa, 7000 are randomly chosen as a training data set and 200 random points from the rest are taken for the illustration of approximation to inverse estimation. The data on the climate variable, GDD5, are discretized and considered on a regular grid of size 50.

In Fig. 6.11–6.13, the zero-inflated behaviour of the data are shown by their histograms and scatter plots. The results with the real data set are shown in Fig. 6.14–6.16 and in Table 6.3. The fitting of the responses of the taxa by the VB method is shown in Fig. 6.12. The mean response surfaces are not very smooth, as reflected from the figure. Also, the HPD bounds of the response surfaces are very tight. There are two obvious reasons behind the under-estimation of the variance: the posterior independence assumption of the VB method and the number of data available on each grid point on the climate variable. The VB-variances of responses are big for large values of climate. It is clear from the scatter plots in Fig. 6.11 that there are fewer data for big values of climate. If the grid size is small, then there are many non-zero data points related to each of the grid points which make the inference very certain. Therefore, the variance of the responses at climate values with less informative data are larger than those with more informative data. The

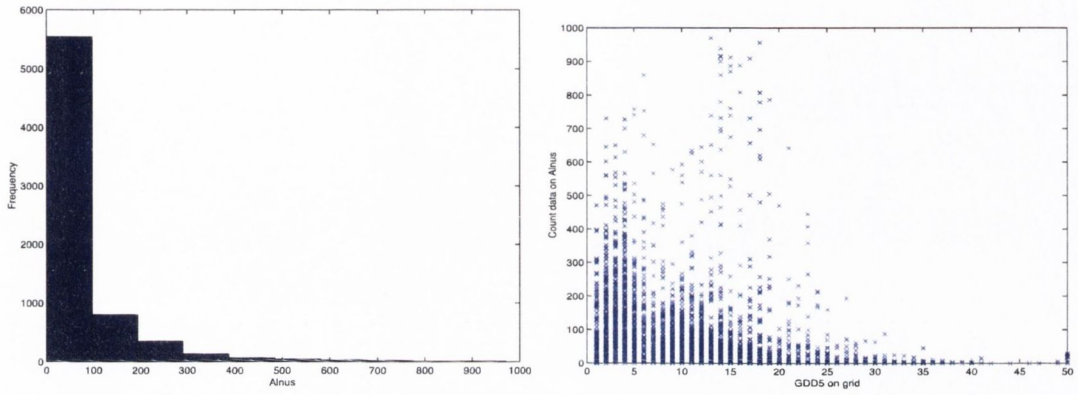

Fig. 6.11: Histogram of Alnus (left) and scatter plot of GDD5 and Alnus (right)

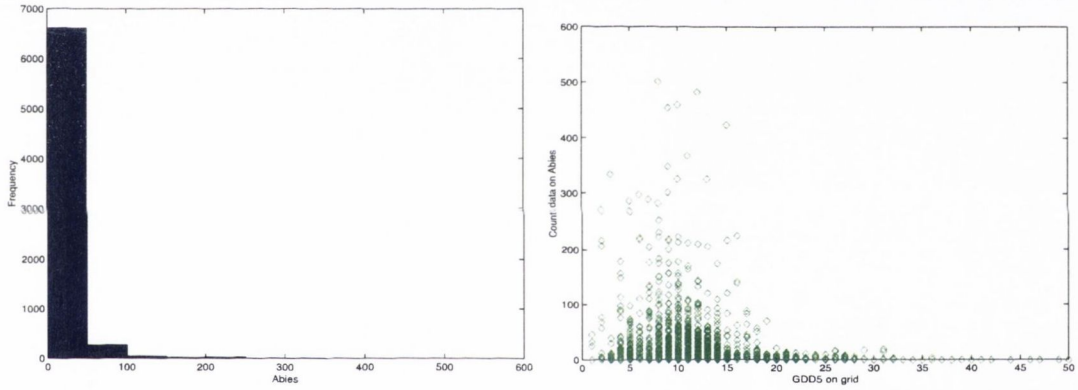

Fig. 6.12: Histogram of Abies (left) and scatter plot of GDD5 and Abies (right)

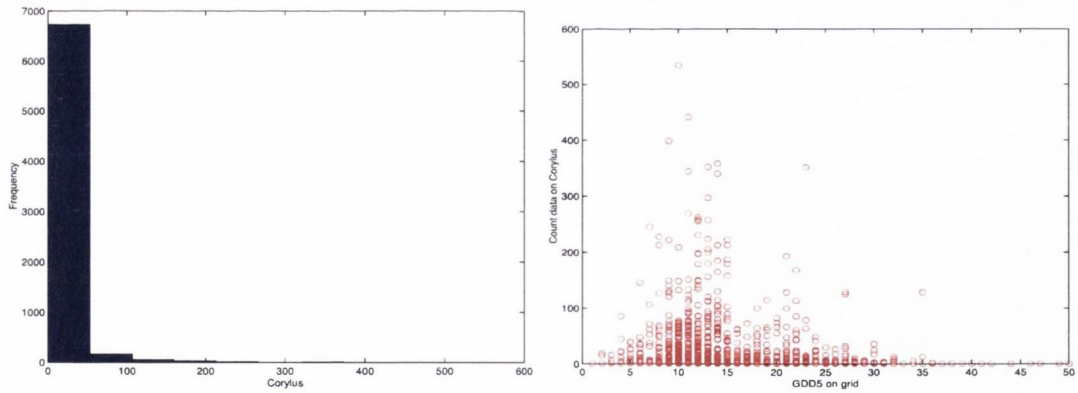

Fig. 6.13: Histogram of Corylus and scatter plot of GDD5 and Corylus

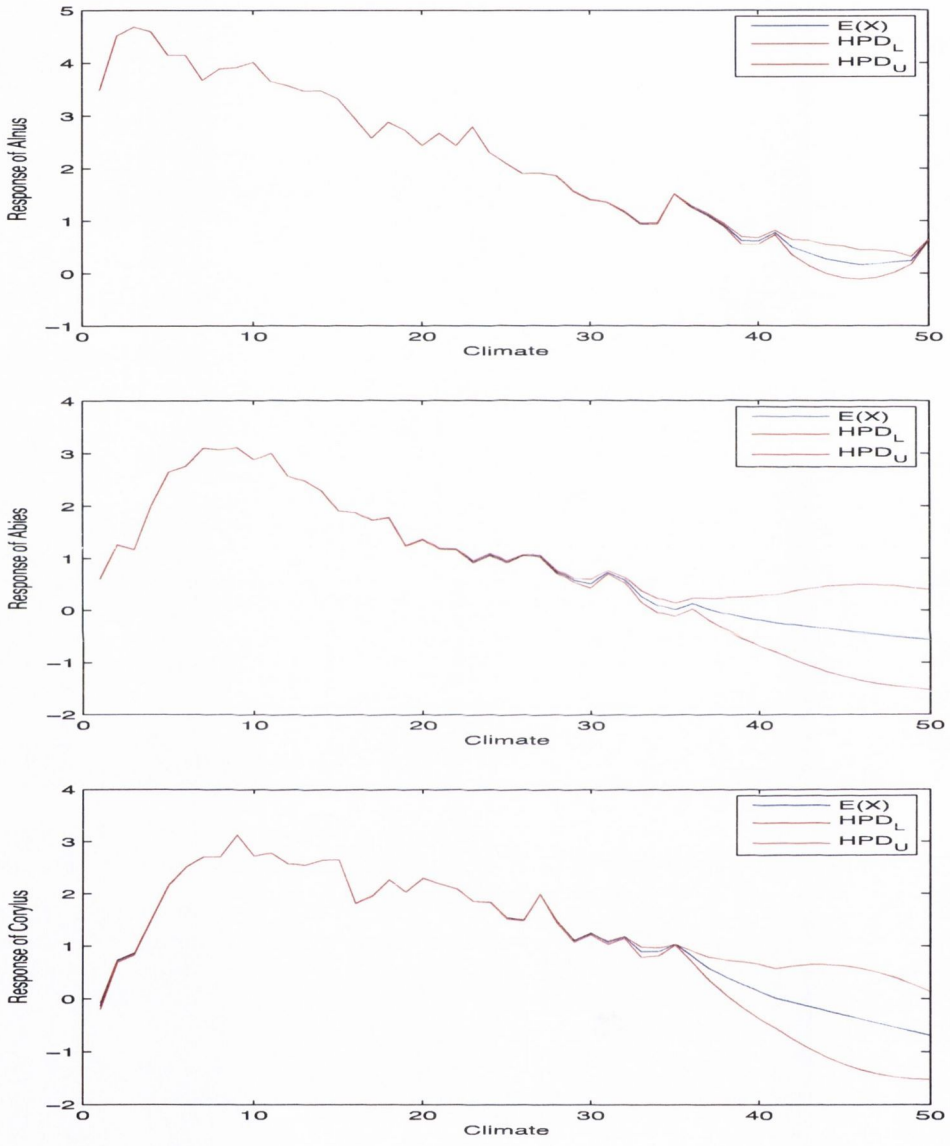

Fig. 6.14: Responses of Alnus (upper most), Abies (middle) and Corylus (lowest one) are shown against climate locations. The VB-mean (blue) of responses of the taxa are presented. Uncertainty of fitting of the responses are shown with 95% HPD region (red).

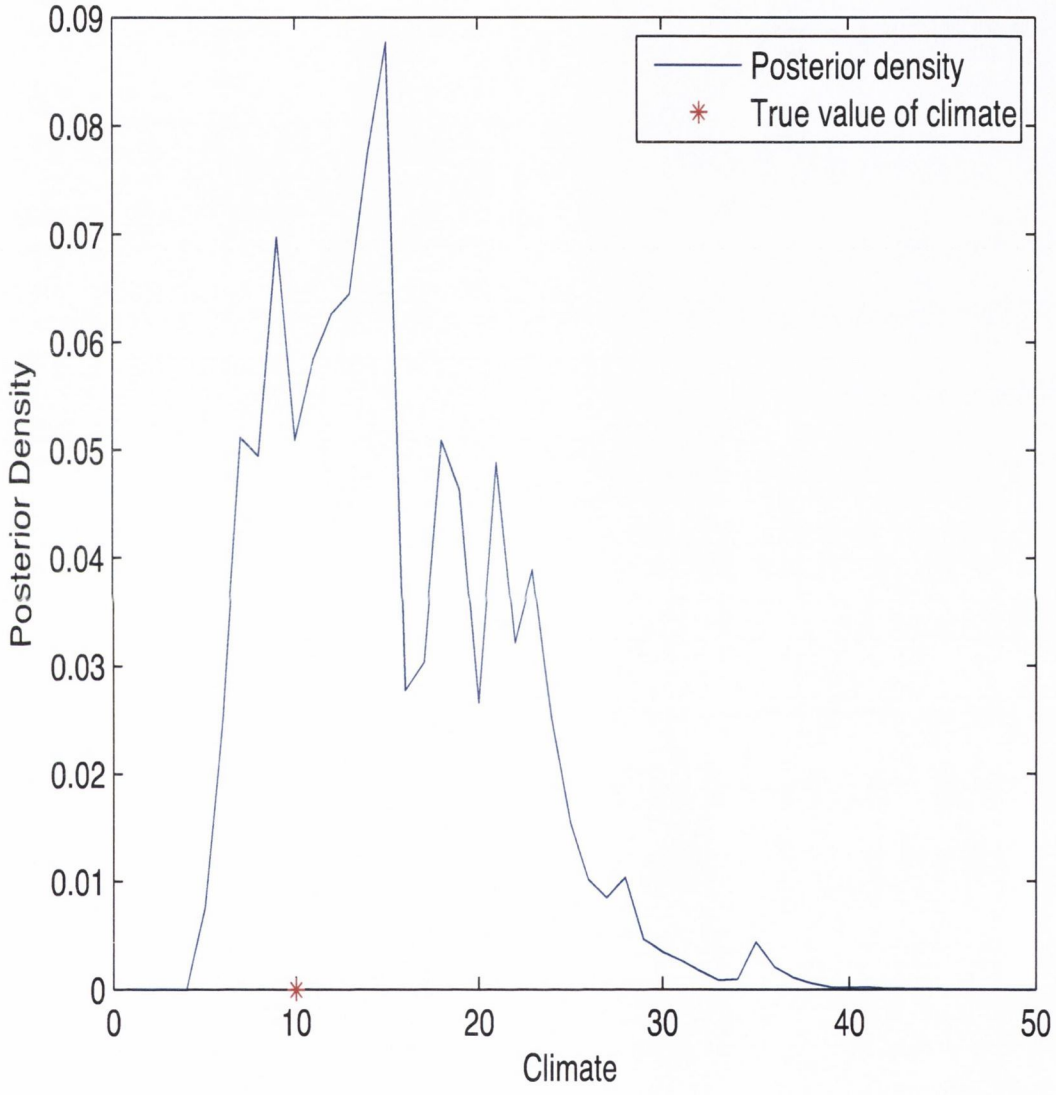

Fig. 6.15: Inverse prediction (reconstruction) of climate given a test data-set on three pollen (77, 0, 48) for the model with real data on pollen, is shown by means of its posterior density of climate. True value of climate, $(*)$, is displayed to compare with the posterior mode of the climate.

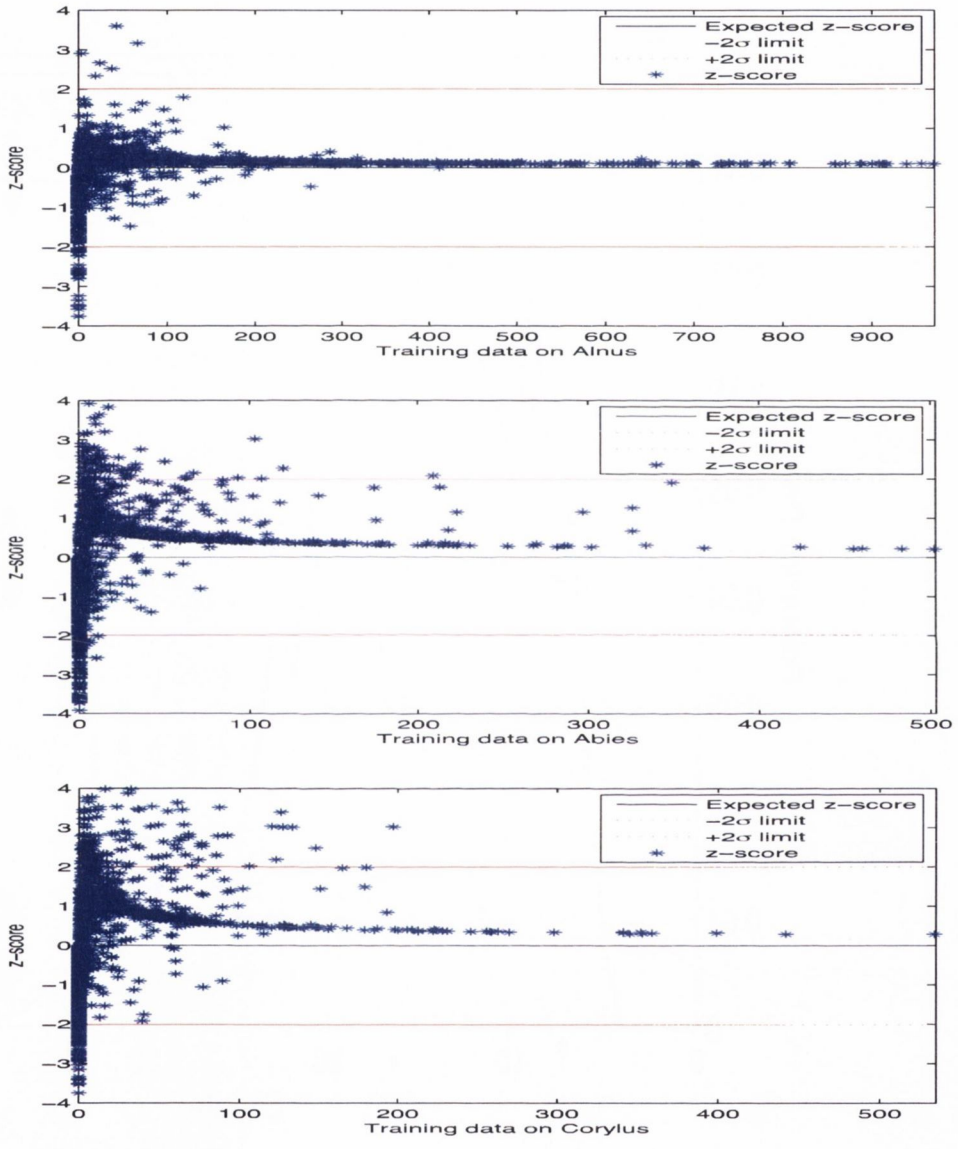

Fig. 6.16: z-scores of real count data on pollen (on y-axiz), (∗), of the palaeoclimate model, are displayed against real counts on Alnus (upper most), on Abies (middle) and on Corylus (lowest) on x-axis with their true mean, zero (black), and true 95% confidence interval (red).

uncertainty in the inference can be increased by taking a large grid size though it may lead to a computational burden. A larger grid size may also increase non-smoothness in the fitting of the responses (a coarser grid tend to have fewer non-zero counts at each grid point of climate may lead to non-smoothness of the responses).

It is noted that very similar results are obtained from other runs of the inference with different samples of data from the same model and different setting for the various numerical approximation, e.g. starting values for optimization. So there is some empirical evidence that the results in this chapter are representative.

The accuracy of the approximation at the forward stage is displayed in Fig. 6.14 by means of the z-scores of pollen data. It is clear from the figure that the z-scores for zero counts are not fitted well by the method. Also, the z-score are over-estimated for large non-zero counts. These results are similar to those for simulated data. The only difference is that the accuracy of the model fitting is now reduced to about 95%. This may be due to the independent model assumed for the pollen data, the independent assumption of the VB method and the bad Gaussian approximation for zero counts.

The VB approximation of climate reconstruction by means of the posterior distribution of climate for a test data on pollen $(77, 0, 48)$, is shown in Fig. 6.13. The climate density is multi-modal. As described earlier for simulated data, the multi-modal behaviour of the density is due to the non-smoothness of the mean response surfaces. The multi-modal behaviour of climate is also experienced in previous reconstructions e.g. in Haslett et al. (2006). The accuracy of the reconstruction of climate by a cross validation technique is about 85%. The loss of accuracy is because of independence assumption of the VB method (used at for ward and inverse stages) and not using all the palaeoclimate data which provide important information about climate.

Table 6.3 shows that the choice of values of hyper-parameters affects the VB solution. Hence, the accuracy of the reconstruction may be improved with more informative prior.

Computational time: The results of the VB approximation of the reconstruction

problem are carried out with MATLAB 6.1. The VB approximation for simulated data of the palaeoclimate reconstruction problem takes only few minutes to converge (locally). A Gaussian approximation of random effects for each data count leads to a slow approximation. The real data is large and contains several zero-counts, hence the VB approximation takes about an hour or more to converge to a local result. The computation time of the approximation of the reconstruction problem at the forward is order $O(Kr^2p + K^3n)$, where K is the number of taxa, p the number of grid points, r is the bandwidth of the sparse precision of latent responses and n is the sample size. The first term in $O(\cdot)$ comes from the matrix computation for K responses. Since the precision of responses are sparse, it takes only r^2p flops for the matrix computation (inverse operation to compute posterior variance). The second term in the function is due to the matrix operation for n random effects. The inverse stage uses a Gaussian approximation which requires $(2K)^3$ flops for the matrix operation for each test data. Hence the computation time at the inverse stage is of order $O(K^3)$.

6.2.4 Discussion

The palaeoclimate reconstruction provides a motivating example of a multi-dimensional complex inverse latent regression problem to explore the usefulness and the limitations of the VB method. With the assumption of posterior independence, the method is capable of handling the multi-dimensional reconstruction problem though at a cost of some accuracy loss in the inverse estimation of climate.

It is shown in the results that the performance of VB approximation in palaeoclimate reconstruction depends on the choice of prior. It suggests that the method may not work very well with noisy or less informative data. The VB approximation for zero counts of real data are not accurate. Since the approximation also depends on the model assumption, the VB method performs very poorly for zero counts modelled with ZI-Poisson model. It might be tempting to consider other models e.g. a negative Binomial distribution to fit zero counts. But, this distribution may present a more challenging model for VB for its tractable solution and might demand many approximations in the estimation process.

The approximation to the model-fitting and the reconstruction of climate with real data not only depends on the VB approximation and other approximations used in the inference process, but it is also conditioned on the model assumption and the amount of information used. The full data structure (28 taxa and 7742 counts per taxon) is correlated and provides important information about climate. There are only 3 taxa used among 28, also an independent ZI-Poisson model is assumed for a successful VB approximation. Haslett et al. (2006) have modelled the correlation structure of data by a compound multinomial likelihood. Salter-Townshend (2009) suggests to use a nesting structure model to account for the correlation in data.

It is experienced that the VB method can also be slow, if prior assumptions are not appropriate in the case of less informative data. VB approximation for the inverse latent regression using palaeoclimate model can be compared with the results of Salter-Townshend (2009). The author used similar (ZI-Poisson) palaeoclimate model. The VB method used with many other approximations and the posterior independence assumption, predicts a climate variable with 85% of accuracy which is less than the result by INLA (for two climate variable) of Salter-Townshend (2009).

The accuracy of VB approximation for inverse latent regression problems with the palaeoclimate model, can be increased to some extent with more informative data. Allowing the dependence between responses and random effects in the joint posterior may boost in the accuracy of the approximation (responses and random effects are highly correlated in the palaeoclimate model). However for more than three taxa, it may lead to an impractical computational burden.

Chapter 7

Conclusion and Future Work

7.1 Conclusions

In this thesis, the VB approximation to Bayesian estimation of inverse regression problems has been classified into two categories: inverse latent regression and inverse non-latent regression. The VB method is very useful for fast approximations for multi-dimensional inverse regression problems. It provides simple and quick approximations at some cost of accuracy loss in the approximation due to the conditional independence assumption. In Chapter 4, the inverse non-latent regression has been described through the simple linear regression, quadratic regression and Poisson regression mixture of Poisson regression and zero-inflated Poisson regression problems. The inverse simple linear regression and the inverse quadratic regression problems present a good example of conjugate-exponential (CE) models, for which the VB method provides tractable approximations (Beal, 2003; Šmídl & Quinn, 2006). It was observed for inverse quadratic regression problems that the method is capable of providing a multi-modal approximation to a multi-modal posterior distribution if the multi-modality is not induced due to some latent variables. The accuracy and the tractability of the method was explored for non-CE regression models through the inverse Poisson regression, mixture of Poisson regression and zero-inflated Poisson regression problems. It was attempted to make the method amenable to non-CE regression models by applying it with further approximations, such as a Gaussian approximation. In Chapter 5, the VB approximation for latent

regression problems has been described with the help of three models: Poisson latent model, Poisson latent random effect model and zero-inflated Poisson random effect model. The intractability issue of the method was further explored for latent regression models. It was shown that a Gaussian approximation (to solve the intractability of the method) might lead to inaccurate VB solutions where the result is sensitive to model assumption and nature of data under study. The use of a Gaussian approximation to approximate the intractable VB approximation of the new explanatory variable (of index type) of the zero-inflated Poisson model was avoided, since the moments of an index-type variable, though might exit, do not make sense. Moreover as the VB approximations relate to the VB-moments, the VB method is limited to the models for which the moments can be defined. However, Vatsa & Wilson (2010) have presented a VB approximation for the climate (index type) variable of the palaeoclimate model which depends on VB-expectations of the latent response surface with respect to the VB approximations of climate, but does not require the VB-moments of unknown climate in the algorithm. In Chapter 6, the VB approximation for inverse latent regression models has been studied via the palaeoclimate reconstruction problem. It was suggested to use a more informative prior when the data are noisy and less informative, since the VB solution might depend on the choice of prior and the initial settings of the VB-parameters. However with sufficient data, the method provides accurate estimation.

In short, the VB method can be very useful for fast approximations to Bayesian inference in inverse regression problems. As far as the intractability issue of the method is concerned, it can be solved by applying it with further approximations. Extra care should be taken in the prior elicitation, as the VB approximation may hugely depend on prior specification in case of insufficient data.

7.2 Future Work

The work on the VB method for inverse regression problems has given rise to many questions for further study:

1. To carry out tractable VB approximation for inverse non-CE regression mod-

els, a Gaussian approximation is applied with the VB method. The VB method needs to be explored further for its intractability for the models where a Gaussian approximation may be inappropriate. Other variational methods, e.g. variational tangent approach may be applied in this case but it also has limited applicability.

2. In the latent random effect regression model of Chapter 5, the index type behaviour of explanatory variable restricts the application of the VB method to the inverse stage of inference. It needs to explore more latent models where the VB method can be successfully applied for inverse estimation of non-index type variables.

3. The uncertainty of the power index of the ZI-Poisson latent random effect model of Chapter 5, could not be studied due to the intractability issue of the VB method. Further approximation is needed to carry out a VB approximation for the model with the VB-estimation of the power-law index parameter. Other models e.g mixture of Poisson model may be considered which avoids power-law index parameter, but it may not fit the zero-inflated data very well degrading the accuracy of the approximation.

4. The assumption of conditional independence may lead to a bad approximation for multi-dimensional models where the components of parameters are highly correlated and the data counts are also correlated. A structured VB approximation for a nested palaeoclimate model of Salter-Townshend (2009) may increase the accuracy by allowing dependency between the responses and the random effects of the model such that it should not lead to a huge computational burden on the VB-estimation. It may also solve the problem of the selection of the priors (to some extent), as for a nested model it will no longer need to approximate the random effects for each count data (including zero counts) independently. However, due to the basic definition of the VB method, with KL-divergence a non-negative quantity which requires the support the VB approximation smaller than the true posterior distribution, the under-estimation of posterior variance by VB method cannot be avoided even

with a fully structured VB approximation.

5. It would be of a great interest to find a measure of under-estimation of VB-variance which may lead us to increase the accuracy of the VB-approximation. Wang & Titterington (2005) suggested to use the inverse of the Fisher's information for the VB-variance to increase the accuracy of the VB-interval estimates. This idea of replacing the VB-variance is based on the asymptotic normality of the VB approximation (Wang & Titterington, 2004). It may be useful to explore the idea of Wang & Titterington (2005) to increase the accuracy of the VB interval-estimates for models with a large data set.

6. A method to monitor which of the approximations applied in the process on inverse estimation for ZI-Poisson latent random effect model are the most responsible for a bad approximation, should also be found in order to improve the VB approximation for the model.

Chapter 8

Appendix

8.1 VB approximations for zero-inflated-Poisson non-latent model of Chapter 4

This section defines the VB marginals for the ZI-Poisson non-latent model of Chapter 4.

1. The Gaussian approximation to the VB marginals of β_0 ,$q(\beta_0|\mathbf{y},\mathbf{x})$ is obtained as:

$$
\begin{aligned}
q_g(\beta_0|\mathbf{y},\mathbf{x}) &= N(\mu^{\star}_{\beta_0}, S^{2\,\star}_{\beta_0}), && (8.1)\\
S^{2\,\star}_{\beta_0} &= \left[\frac{1}{S^2_{\beta_0}} - f_2\right], && (8.2)\\
\mu^{\star}_{\beta_0} &= \left(\frac{\mu_{\beta_0}}{S^2_{\beta_0}} + f_1 - \beta_0^m f_2\right), && (8.3)\\
f_1 &= \frac{\partial}{\partial\beta_0}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, && (8.4)\\
f_2 &= \frac{\partial^2}{\partial\beta_0^2}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, && (8.5)
\end{aligned}
$$

where the functions f_1 and f_2 are defined as:

$$f_1 = \sum_{i=1}^{n}\left[(1-\mathrm{z}_i)(1-Po_{0i}^m-B_{0i}^m+\mathrm{y}_i)-\mathrm{z}_i Po_{0i}^m\frac{(F_{0i}^m D_{0i}^m+B_{0i}^m C_{0i}^m)}{(1-Po_{0i}^m D_{0i}^m)}\right], \quad (8.6)$$

$$f_2 = -\sum_{i=1}^{n}\left[(1-\mathrm{z}_i)(B_{0i}^m+Po_{0i}^m F_{0i}^m)\right]-\sum_{i=1}^{m}\left[\mathrm{z}_i {Po_{0i}^m}^2\frac{(F_{0i}^m D_{0i}^m-B_{0i}^m C_{0i}^m)^2}{(1-Po\_0i^m D_{0i}^m)^2}\right]$$
$$-\sum_{i=1}^{n}\left[\mathrm{z}_i B_{0i}^m C_{0i}^m\frac{(Po_{0i}^m F_{0i}^m+Po_{0i}^m(1+1/B_0^m)-B_{0i}^m)}{(1-Po_{0i}^m D_{0i}^m)}\right], \quad (8.7)$$

where,

$$B_{0i}^m = \exp(\beta_0^m+\beta_1^m \mathrm{x}_i), \quad (8.8)$$

$$C_{0i}^m = \exp(-B_{0i}^m), \quad (8.9)$$

$$D_{0i}^m = (1-C_{0i}^m), \quad (8.10)$$

$$F_{0i}^m = (1+B_{0i}^m)^{-1}, \quad (8.11)$$

$$Po_{0i}^m = B_{0i}^m F_{0i}^m. \quad (8.12)$$

2. The Gaussian approximation to the VB marginals of β_1 ,$q(\beta_1|\mathbf{y},\mathbf{x})$ is obtained as:

$$q_g(\beta_1|\mathbf{y},\mathbf{x}) = N(\mu_{\beta_1}^\star, S_{\beta_1}^{2\star}), \quad (8.13)$$

$$S_{\beta_1}^{2\star} = \left[\frac{1}{S_{\beta_1}^2}-g_2\right], \quad (8.14)$$

$$\mu_{\beta_1}^\star = \left(\frac{\mu_{\beta_1}}{S_{\beta_1}^2}+g_1-\beta_1^m g_2\right), \quad (8.15)$$

$$g_1 = \frac{\partial}{\partial\beta_1}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, \quad (8.16)$$

$$g_2 = \frac{\partial^2}{\partial\beta_1^2}\log P(\mathbf{y}|\mathbf{x},\beta_0,\beta_1)\Big|_{\beta_0^m,\beta_1^m}, \quad (8.17)$$

where g_1 and g_2 are defined as:

$$g_1 = \sum_{i=1}^{n}\left[\mathrm{x}_i(1-\mathrm{z}_i)(1-Po_{1i}^m-B_{1i}^m+\mathrm{y}_i)\right] - \sum_{i=1}^{n}\left[\mathrm{x}_i\mathrm{z}_i Po_{1i}^m\frac{(F_{1i}^m D_{1i}^m+B_{1i}^m C_{1i}^m)}{(1-Po_{1i}^m D_{1i}^m)}\right], \tag{8.18}$$

$$g_2 = -\sum_{i=1}^{n}\left[\mathrm{x}_i^2(1-\mathrm{z}_i)(B_{1i}^m+Po_{1i}^m F_{1i}^m)\right] - \sum_{i=1}^{m}\left[\mathrm{x}_i^2\mathrm{z}_i {Po_{1i}^m}^2\frac{(F_{1i}^m D_{1i}^m-B_{1i}^m C_{1i}^m)^2}{(1-Po_{1i}^m D_{1i}^m)^2}\right] - \sum_{i=1}^{n}\left[\mathrm{x}_i^2\mathrm{z}_i B_{1i}^m C_{1i}^m\frac{(Po_{1i}^m F_{1i}^m+Po_{1i}^m(1+1/B_{1i}^m)-B_{1i}^m)}{(1-Po_{1i}^m D_{1i}^m)}\right], \tag{8.19}$$

where,

$$B_{1i}^m = \exp(\beta_0^m+\beta_1^m\mathrm{x}_i), \tag{8.20}$$

$$C_{1i}^m = \exp(-B_{1i}^m), \tag{8.21}$$

$$D_{1i}^m = (1-C_{1i}^m), \tag{8.22}$$

$$F_{1i}^m = (1+B_{1i}^m)^{-1}, \tag{8.23}$$

$$Po_{1i}^m = B_{1i}^m F_{1i}^m. \tag{8.24}$$

3. The Gaussian approximation to the intractable VB approximation of $\mathrm{X}_{\mathrm{new}}$ is found as:

$$q_g(\mathrm{X}_{\mathrm{new}}|\mathbf{y},\mathbf{x}) = N(\mu_{\mathrm{X}}^\star, S_{\mathrm{X}}^{2\,\star}), \tag{8.25}$$

$$S_{\mathrm{X}}^{2\,\star} = \left[\frac{1}{S_{\mathrm{X}}^2}-h_2\right], \tag{8.26}$$

$$\mu_{\mathrm{X}}^\star = \left(\frac{\mu_{\mathrm{X}}}{S_{\mathrm{X}}^2}+h_1-\mathrm{X}_{\mathrm{new}}^m h_2\right), \tag{8.27}$$

$$h_1 = \frac{\partial}{\partial \mathrm{X}_{\mathrm{new}}}\log P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1)\Big|_{\mathrm{X}_{\mathrm{new}}^m,\beta_0^m,\beta_1^m}, \tag{8.28}$$

$$h_2 = \frac{\partial^2}{\partial \mathrm{X}_{\mathrm{new}}^2}\log P(\mathrm{y}_{\mathrm{new}}|\mathrm{X}_{\mathrm{new}},\beta_0,\beta_1)\Big|_{\mathrm{X}_{\mathrm{new}}^m,\beta_0^m,\beta_1^m}, \tag{8.29}$$

where h_1 and h_2 are defined as:

$$h_1 = \left[(1-\mathrm{z}_{\mathrm{new}})(1-Po_x^m-B_x^m+\mathrm{y}_{\mathrm{new}})\right] - \left[\mathrm{z}_{\mathrm{new}}\beta_1^m Po_x^m \frac{(F_x^m D_x^m + B_x^m C_x^m)}{(1-Po_x^m D_x^m)}\right], \quad (8.30)$$

$$h_2 = \beta_1^{m\,2}\left[(1-\mathrm{z}_{\mathrm{new}})(B_x^m+Po_x^m F_x^m)\right] - \beta_1^{m\,2}\left[\mathrm{z}_{\mathrm{new}} Po_x^{m\,2}\frac{(F_x^m D_x^m - B_x^m C_x^m)^2}{(1-Po_x^m D_x^m)^2}\right] - \beta_1^{m\,2}\left[\mathrm{z}_{\mathrm{new}} B_x^m C_x^m \frac{(Po_x^m F_x^m + Po_x^m(1+1/B_x^m) - B_x^m)}{(1-Po_x^m D_x^m)}\right] \quad (8.31)$$

where,

$$B_x^m = \exp(\beta_0^m + \beta_1^m \mathrm{X}_{\mathrm{new}}^m), \quad (8.32)$$

$$C_x^m = \exp(-B_x^m), \quad (8.33)$$

$$D_x^m = (1-C_x^m), \quad (8.34)$$

$$F_x^m = (1+B_x^m)^{-1}, \quad (8.35)$$

$$Po_x^m = B_x^m F_x^m. \quad (8.36)$$

8.2 VB approximations for ZI-Poisson latent random effect model of Chapter 5

This section presents the hyper-parameters of the VB marginals of unknowns at the forward stage of inference for ZI-Poisson model.

8.2.1 Gaussian approximation of the VB marginal of $\mathbf{Z}_k$

The VB marginal $q_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y},\mathbf{x},\hat{\alpha})$ is given as follows:

For all $k = 1:K$,

$$q_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y},\mathbf{x},\hat{\alpha}) \approx q^g_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y},\mathbf{x},\hat{\alpha}), \quad (8.37)$$

$$q^g_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y},\mathbf{x},\hat{\alpha}) = N_p(\mu^\star_{\mathbf{Z}_k}, Q^{\star\,-1}_{\mathbf{Z}_k}), \quad (8.38)$$

where the term $q^g_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y}, \mathbf{x}, \hat{\alpha})$ denotes the Gaussian approximation of $q_{\mathbf{Z}_k}(\mathbf{Z}_k|\mathbf{y}, \mathbf{x}, \hat{\alpha})$ with mean $\mu^\star_{\mathbf{Z}_k}$ and precision $Q^\star_{\mathbf{Z}_k}$ given as follows:

$$
\begin{aligned}
Q^\star_{\mathbf{Z}_k} &= \mathbb{E}_q(\kappa_k)R + diag(V_{\mathbf{Z}_k}), && (8.39)\\
\mu^\star_{\mathbf{Z}_k} &= Q^{\star\,-1}_{\mathbf{Z}_k} B_{\mathbf{Z}_k}, && (8.40)\\
B_{\mathbf{Z}_k} &= [A_{\mathrm{Z}_{ki}};\ i = 1:p]^T + V_{\mathbf{Z}_k}\mathbf{Z}^m_k, && (8.41)\\
V_{\mathbf{Z}_k} &= diag[V_{\mathrm{Z}_{ki}}\ i = 1:p]^T, && (8.42)\\
V_{\mathrm{Z}_{ki}} &= -\frac{\partial^2}{\partial \mathrm{Z}^2_{ki}} \sum_{\substack{j=1\\ \mathrm{x}_j=i}}^{n} \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{Z}_{ki}=\mathrm{Z}^m_{ki}}, && (8.43)\\
A_{\mathrm{Z}_{ki}} &= \frac{\partial}{\partial \mathrm{Z}_{ki}} \sum_{\substack{j=1\\ \mathrm{x}_j=i}}^{n} \mathbb{E}_{q_{\mathbf{U}}(\mathbf{U})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{Z}_{ki}=\mathrm{Z}^m_{ki}}, && (8.44)
\end{aligned}
$$

where $\mathbf{Z}^m_k = [\mathrm{Z}^m_{ki};\ i = 1:p]^T$ is the posterior mode of $\mathbf{Z}_k$. It is already mentioned that the expectation of the likelihood w.r.t $q_{\mathbf{U}}(\mathbf{U})$ is not in tractable, therefore the modes of $q^g_{\mathbf{U}}(\mathbf{U})$ are plugged-in the likelihood instead.)

For all $i = 1:p$, define

$$
\begin{aligned}
F_{1i} &= \exp\left(\sum_{\substack{j=1\\ \mathrm{x}_j=i}}^{n} \mathrm{U}^m_{kj} + \mathrm{Z}^m_{ki}\right), && (8.45)\\
F_{2i} &= \exp(-F_{1i}), && (8.46)\\
F_{3i} &= \frac{F_{1i}}{1+F_{1i}}, && (8.47)\\
F_{4i} &= (F_{3i})^{\hat{\alpha}_k}. && (8.48)
\end{aligned}
$$

(where U_{jk}^m and Z_{ki}^m are the posterior modes of U_{jk} and Z_{ki} respectively) then

$$A_{\mathrm{Z}_{ki}} = h_{1i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) + h_{2i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big), \tag{8.49}$$

$$h_{1i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) = \sum_{\substack{j=1 \\ \mathrm{x}_j=i}}^{n} I(\mathrm{y}_{kj} \neq 0)\Big[- F_{1k} + \mathrm{y}_{kj} + \hat{\alpha}_k\,(1 - F_{3k})\Big], \tag{8.50}$$

$$h_{2i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) = \sum_{\substack{j=1 \\ \mathrm{x}_j=i}}^{n} I(\mathrm{y}_{kj} = 0)\Big[- F_{2i}F_{4i}F_{1i} - F_{4i}(1 - F_{2i})(-F_{3i} + 1)\hat{\alpha}_k\Big] \times\big[1 - F_{4i}(1 - F_{2i})\big]^{-1}, \tag{8.51}$$

and

$$V_{\mathrm{Z}_{ki}} = m_{1i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) + m_{2i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) + m_{3i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big), \tag{8.52}$$

$$m_{1i}\big(\mathbf{y}_k, \mathbf{U}_{k0}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) = \sum_{\substack{j=1 \\ \mathrm{x}_j=i}}^{n} I(\mathrm{y}_{kj} \neq 0)\Big[- F_{1i} + \hat{\alpha}_k(F_{3i} - 1)F_{3i}\Big], \tag{8.53}$$

$$m_{2i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) = -\sum_{\substack{j=1 \\ \mathrm{x}_j=i}}^{n} I(\mathrm{y}_{kj} = 0)\Big[F_{2i}F_{4i}F_{1i} - F_{4i}(1 - F_{2i})(-F_{3i} + 1)\hat{\alpha}_k\Big]^2 \times\big[1 - F_{4i}(1 - F_{2i})\big]^{-2}, \tag{8.54}$$

$$m_{3i}\big(\mathbf{y}_k, \mathbf{U}_k^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) = \sum_{\substack{j=1 \\ \mathrm{x}_j=i}}^{n} I(\mathrm{y}_{kj} = 0)\Big[F_{2i}F_{4i}(F_{1i})^2 - F_{4i}(1 - F_{2i})(2(F_{3i})^2 - 3F_{3i} + 1)\hat{\alpha}_k - F_{4i}(1 - F_{2i})(-F_{3i} + 1)^2(-1 + \hat{\alpha}_k)\hat{\alpha}_k - 2F_{2i}F_{4i}(-F_{3i} + 1)F_{1i}\hat{\alpha}_k\Big]\Big[1 - F_{4i}(1 - F_{2i})\Big]^{-1}. \tag{8.55}$$

8.2.2 Gaussian approximation of the VB marginal of U_j

For all $j = 1 : n$, the VB marginal $q_{\mathbf{U}_j}(\mathbf{U}_j|\mathbf{y}, \mathbf{x})$ is given as follows:

$$q_{\mathbf{U}_j}(\mathbf{U}_j|\mathbf{y}, \mathbf{x}, \hat{\alpha}) \approx q_{\mathbf{U}_j}^g(\mathbf{U}_j|\mathbf{y}, \mathbf{x}, \hat{\alpha}), \tag{8.56}$$

$$q_{\mathbf{U}_j}^g(\mathbf{U}_j|\mathbf{y}, \mathbf{x}, \hat{\alpha}) = N_p(\mu_{\mathbf{U}_j}^{\star}, Q_{\mathbf{U}_j}^{\star\,-1}), \tag{8.57}$$

where the term $q^g_{\mathbf{U}_j}(\mathbf{U}_j)$ denotes the Gaussian approximation of $q_{\mathbf{U}_j}(\mathbf{U}_j)$ with mean $\mu^\star_{\mathbf{U}_j}$ and precision $Q^\star_{\mathbf{U}_j}$ given as follows:

$$
\begin{aligned}
Q^\star_{\mathbf{U}_j} &= \mathbb{E}_q(Q_{\mathbf{U}}) + diag(V_{\mathbf{U}_j}), && (8.58)\\
\mu^\star_{\mathbf{U}_j} &= Q^{\star\,-1}_{\mathbf{U}_j} B_{\mathbf{U}_j}, && (8.59)\\
B_{\mathbf{U}_j} &= [A_{\mathrm{U}_{kj}};\ k = 1:K]^T + V_{\mathbf{U}_j}\mathbf{U}^m_j, && (8.60)\\
V_{\mathbf{U}_j} &= diag[V_{\mathrm{U}_{kj}}\ k = 1:K]^T, && (8.61)\\
V_{\mathrm{U}_{kj}} &= -\frac{\partial^2}{\partial \mathrm{U}^2_{kj}} \sum_{j=1}^{n} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{U}_{kj}=\mathrm{U}^m_{kj}}, && (8.62)\\
A_{\mathrm{U}_{kj}} &= \frac{\partial}{\partial \mathrm{U}_{kj}} \sum_{j=1}^{n} \mathbb{E}_{q_{\mathbf{Z}}(\mathbf{Z})} \log P(\mathrm{y}_{kj}|\mathbf{Z}_k(\mathrm{x}_j), \mathrm{U}_{kj})\Big|_{\mathrm{U}_{kj}=\mathrm{U}^m_{kj}}, && (8.63)
\end{aligned}
$$

where $\mathbf{U}^m_j = [\mathrm{U}^m_{kj};\ i = 1:p]^T$ is the posterior mode of $\mathbf{U}_j$.. (Since $\mathbf{U}_{jk}$ appear independently in the ZI-Poisson likelihood, there is no covariance term in $V_{\mathbf{U}_j}$, the second derivative of the likelihood w.r.t $\mathbf{U}_j$. The expectation of the likelihood w.r.t $q_{\mathbf{Z}}(\mathbf{Z})$ is not in tractable, therefore the modes of $q^g_{\mathbf{Z}}(\mathbf{Z})$ are plugged-in the likelihood instead.)

For all $k = 1:K$, define

$$
\begin{aligned}
E_{1k} &= \exp\left(\mathrm{U}^m_{kj} + \mathrm{Z}^m_k(\mathrm{x}_j)\right), && (8.64)\\
E_{2k} &= \exp(-E_{1k}), && (8.65)\\
E_{3k} &= \frac{E_{1k}}{1+E_{1k}}, && (8.66)\\
E_{4k} &= (E_{3k})^{\hat{\alpha}_k}. && (8.67)
\end{aligned}
$$

then

$$
\begin{aligned}
A_{\mathrm{U}_{kj}} &= f_{1k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) + f_{2k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big), && (8.68)\\
f_{1k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) &= I(\mathrm{y}_{kj}^m \neq 0)\Big[- E_{1k} + \mathrm{y}_{kj}^m + \hat{\alpha}_k\,(1 - E_{3k})\Big], && (8.69)\\
f_{2k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) &= I(\mathrm{y}_{kj} = 0)\Big[- E_{2k}E_{4k}E_{1k} - E_{4k}(1 - E_{2k}) \\
&\qquad \times(-E_{3k} + 1)\hat{\alpha}_k\Big] \times \big[1 - E_{4k}(1 - E_{2k})\big]^{-1}, && (8.70)\\
V_{\mathrm{U}_{kj}} &= g_{1k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) + g_{2k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) \\
&\qquad + g_{3k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big), && (8.71)\\
g_{1k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) &= I(\mathrm{y}_{kj} \neq 0)\Big[- E_{1k} + \hat{\alpha}_k(E_{3k} - 1)E_{3k}\Big], && (8.72)\\
g_{2k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) &= -I(\mathrm{y}_{kj} = 0)\Big[E_{2k}E_{4k}E_{1k} - E_{4k}(1 - E_{2k}) \\
&\qquad \times(-E_{3k} + 1)\hat{\alpha}_k\Big]^2 \\
&\qquad \times\big[1 - E_{4k}(1 - E_{2k})\big]^{-1}, && (8.73)\\
g_{3k}\big(\mathrm{y}_{kj}, \mathrm{U}_{kj}^m, \mathbf{Z}_k^m, \hat{\alpha}_k\big) &= I(\mathrm{y}_{kj} = 0)\Big[E_{2k}E_{4k}(E_{1k})^2 - E_{4k}(1 - E_{2k})(2(E_{3k})^2 \\
&\qquad -3E_{3k} + 1)\hat{\alpha}_k - E_{4k}(1 - E_{2k})(-E_{3k} + 1)^2 \\
&\qquad \times(-1 + \hat{\alpha}_k)\hat{\alpha}_k - 2E_{2k}E_{4k}(-E_{3k} + 1) \\
&\qquad \times E_{1k}\hat{\alpha}_k\Big]\Big[1 - E_{4k}(1 - E_{2k})\Big]^{-1}. && (8.74)
\end{aligned}
$$

8.2.3 Posterior distribution of α

The posterior distribution of α_k is defined as:

$$
P(\alpha|\mathbf{y}, \mathbf{x}) = \int P(\alpha, \mathbf{Z}, \mathbf{U}, Q_{\mathrm{U}}, \kappa|\mathbf{y}, \mathbf{x})\mathrm{d}\mathbf{Z}\mathrm{d}\mathbf{U}\mathrm{d}Q_{\mathrm{U}}\mathrm{d}\kappa, \tag{8.75}
$$

which is not in a closed form. To avoid complexity of Bayesian computation, the posterior mode of α is used in the VB approximation of other unknowns. To find the mode of α, it requires to define the first and second derivatives of its posterior distribution log-likelihood w.r.t to α. Since an independent log-normal prior distribution is assumed and α_k $\forall k$ appear independently in the ZI-Poisson likelihood, the

modes of α_k $\forall k$ are found independently from its conditional posterior distribution $P(\alpha_k|\mathbf{y}, \mathbf{x}, \mathbf{Z}_k^m, \mathbf{U}_k^m)$. The first and the second derivatives of $\log P(\alpha_k|\mathbf{y}, \mathbf{x}, \mathbf{Z}_k^m, \mathbf{U}_k^m)$ are described as:

$$\frac{\partial}{\partial \alpha_k}\{\log P(\alpha_k|\mathbf{y}, \mathbf{x}, \mathbf{Z}_k^m, \mathbf{U}_k^m)\} = r_{1k} + r_{2k} + r_{3k}, \tag{8.76}$$

$$\frac{\partial^2}{\partial \alpha_k^2}\{\log P(\alpha_k|\mathbf{y}, \mathbf{x}, \mathbf{Z}_k^m, \mathbf{U}_k^m)\} = s_{1k} + s_{2k} + s_{3k}, \tag{8.77}$$

$$r_{1k} = -\frac{1}{\alpha_k} - \left[-\mu_{\alpha_k} + \frac{\log \alpha_k}{\alpha_k \sigma_{\alpha_k}^2}\right], \tag{8.78}$$

$$r_{2k} = \sum_{j=1}^{n} I(\mathrm{y}_{kj} = 0)\left[H_{4jk}(1 - H_{2jk}) \log H_{3jk}\right] \times \left[1 - H_{4jk}(1 - H_{2jk})\right]^{-1}, \tag{8.79}$$

$$r_{3k} = \sum_{j=1}^{n} I(\mathrm{y}_{kj} \neq 0)\left[\mathrm{Z}_k^m(\mathrm{x}_j) + \mathrm{U}_{kj}^m - \log(1 + H_{1jk})\right], \tag{8.80}$$

$$s_{1k} = \left[\sigma_{\alpha_k}^2 - 1 - \mu_{\alpha_k} + \log \alpha_k\right]\left[\alpha_k^2 \sigma_{\alpha_k}^2\right]^{-1}, \tag{8.81}$$

$$s_{2k} = -\sum_{j=1}^{n} I(\mathrm{y}_{kj} = 0)\left[H_{4jk}(1 - H_{2jk}) \log H_{3jk}\right]^2 \times \left[1 - H_{4jk}(1 - H_{2jk})\right]^{-2}, \tag{8.82}$$

$$s_{3k} = -\frac{1}{2}(s_{2k})^2, \tag{8.83}$$

$$H_{1jk} = \exp(\mathrm{Z}_k^m(\mathrm{x}_j) + \mathrm{U}_{kj}^m), \tag{8.84}$$

$$H_{2jk} = \exp(-H_{1jk}), \tag{8.85}$$

$$H_{3jk} = \frac{H_{1jk}}{1 + H_{1jk}}, \tag{8.86}$$

$$H_{4jk} = (H_{3jk})^{\alpha_k}. \tag{8.87}$$

8.2.4 VB approximation to the posterior distribution of $\mathrm{X}_{\mathrm{new}}$

In this section, the Laplace approximation of the posterior distribution of $\mathrm{X}_{\mathrm{new}}$ (of ZI-Poisson model) is defined as:

for $i = 1 : p$

$$\log P(\mathrm{X}_{\mathrm{new}} = i|\mathbf{y}_{\mathrm{new}}, \mathbf{y}) \approx \Big[\log P(\mathbf{y}_{\mathrm{new}}|\mathbf{Z}, \mathbf{U}_{\mathrm{new}}, i) + \log q^g_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, i) + \log P(\mathbf{U}_{\mathrm{new}}|\mathbf{y}, \mathbf{x}) - \log q^g(\mathbf{Z}, \mathbf{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, i)\Big]\Big|_{\substack{\mathbf{Z}=\mathbf{Z}^\star(\mathrm{X}_{\mathrm{new}}) \\ \mathbf{U}_{\mathrm{new}}=\mathbf{U}^\star_{\mathrm{new}}(\mathrm{X}_{\mathrm{new}})}}, \tag{8.88}$$

$$\log P(\mathbf{y}_{\mathrm{new}}|\mathbf{Z}, \mathbf{U}_{\mathrm{new}}, i) = \sum_{k=1}^{K} I(\mathrm{y}_{\mathrm{new}\,k} \neq 0)\Big[\log D^a_k - A^a_k + \mathrm{y}_{\mathrm{new}\,k} \log A^a_k + \log(\mathrm{y}_{\mathrm{new}\,k}!)\Big] + I(\mathrm{y}_{\mathrm{new}\,k} = 0)\Big[\log(1 - D^a_k(1 - B_k)^a)\Big], \tag{8.89}$$

$$\log q^g_{\mathbf{Z}}(\mathbf{Z}|\mathbf{y}, i) \approx -0.5 \sum_{k=1}^{K} \left[\log V_{k\,i,i} + \frac{1}{V_{k\,i,i}}\Big(\mathbf{Z}_k(\mathrm{X}_{\mathrm{new}} = i) - \mu_{k\,i}\Big)^2\right],$$

$$\log P(\mathbf{U}|\mathbf{y}, \mathbf{x}) = 0.5 \log\left[det\left(SS^{\star\,-1} + \mathbf{U}^T_{\mathrm{new}}\mathbf{U}_{\mathrm{new}}\right)^{-1}\right] + 0.5\Big[(df^\star + 1) - df^\star \log det\, SS^\star\Big], \tag{8.90}$$

$$A^a_k = \exp\Big(\mathbf{Z}_k + \mathbf{U}_{\mathrm{new}\,k}\Big)\ \forall k, \tag{8.91}$$

$$B^a_k = \exp(-A^a_k)\ \forall k, \tag{8.92}$$

$$D^a_k = \frac{A^a_k}{1 + A^a_k}\ \forall k, \tag{8.93}$$

$$\tag{8.94}$$

where $V_k = Q^{\star\,-1}_{\mathbf{Z}_k}$ and $\mu_k = \mu_{\mathbf{Z}_k}$ are the VB-variance and the VB-mean of $q^g_{\mathbf{Z}_k}(\mathbf{Z}_k)$, $SS^\star$ and $df^\star$ are the hyper-parameters in the VB marginal of $q_{Q_{\mathrm{U}}}(Q_{\mathrm{U}})$, $\mathbf{Z}^\star(\mathrm{X}_{\mathrm{new}})$ and $\mathbf{U}^\star_{\mathrm{new}}(\mathrm{X}_{\mathrm{new}})$ are VB-modes of the VB marginal $q^g(\mathbf{Z}, \mathbf{U}_{\mathrm{new}}|\mathbf{y}_{\mathrm{new}}, \mathbf{y}, i)$ (defined in a same way as $q^g(\mathbf{Z}_k|\mathbf{y}, \mathbf{x})$ and $q^g(\mathbf{U}_j|\mathbf{y}, \mathbf{x})$) corresponding to $\mathbf{U}_{\mathrm{new}}$ and $\mathbf{Z}$ respectively.

$$q^g(\mathbf{Z}, \mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, i) = N\Big([\mathbf{Z}(\mathrm{X}_{\text{new}})^T \ \mathbf{U}_{\text{new}}(\mathrm{X}_{\text{new}})]^T; \mu^{\star}_{\text{new}}, Q^{\star\,-1}_{\text{new}} \Big), \tag{8.95}$$

$$Q^{\star}_{\text{new}} = \begin{pmatrix} diag[V_{k\,i,i};\ k = 1:K] & D_0 \\ D_0 & \mathbb{E}_q(Q^{\text{new}}_{\text{U}}) \end{pmatrix}_{2K\times 2K} \tag{8.96}$$

$$- \begin{pmatrix} D^a_{1\,k=1:K} & D^a_{2\,k=1:K} \\ D^a_{3\,k=1:K} & D^a_{4\,k=1:K} \end{pmatrix}_{2K\times 2K}, \tag{8.97}$$

$$\mu^{\star}_{\text{new}} = Q^{\star\,-1}_{\text{new}} B^{\star\,T}_{\text{new}}, \tag{8.98}$$

$$B^{\star\,T}_{\text{new}} = [B^a_{1\,k=1:K}\ B^a_{2\,k=1:K}]^T, \tag{8.99}$$

where the term D_0 is a $K \times K$ zero matrix, $Q^{\star}_{\text{new}}$ is the precision parameter of $q^g(\mathbf{Z}, \mathbf{U}_{\text{new}}|\mathbf{y}_{\text{new}}, \mathbf{y}, i)$.

$$\mathbb{E}_q(Q^{\text{new}}_{\text{u}}) = \int Q_{\text{U}} q_{Q_{\text{U}}}(Q_{\text{U}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) \mathrm{d}Q_{\text{U}}, \tag{8.100}$$

$$= df^{\star}_{\text{new}} SS^{\star}_{\text{new}}, \tag{8.101}$$

$$q_{Q_{\text{U}}}(Q_{\text{U}}|\mathbf{y}_{\text{new}}, \mathbf{y}, \mathbf{x}) = \text{Wishart}_K(SS^{\star}_{\text{new}}, df^{\star}_{\text{new}}), \tag{8.102}$$

$$\text{where } df^{\star}_{\text{new}} = df^{\star} + 1, \tag{8.103}$$

$$SS^{\star}_{\text{new}} = \left(SS^{\star\,-1} + \sum_{j=1}^{n_{\text{new}}} \mathbb{E}_q(\mathbf{U}_{\text{new}} \mathbf{U}^T_{\text{new}}) \right)^{-1}. \tag{8.104}$$

For $k = 1 : K$ define,

$$D_k^{a\,2} = -\frac{\partial^2}{\partial \mathbf{Z}_k(\mathrm{X}_{\mathrm{new}}^2)} \left\{\log P(\mathbf{y}_{\mathrm{new}\,k}|\mathbf{Z}_k(\mathrm{X}_{\mathrm{new}}), \mathbf{U}_{\mathrm{new}\,k}, \hat{\alpha}_k)\right\} \Big|_{\substack{\mathbf{Z}=\mathbf{Z}^\star(\mathrm{X}_{\mathrm{new}}) \\ \mathbf{U}_{\mathrm{new}}=\mathbf{U}^\star_{\mathrm{new}}(\mathrm{X}_{\mathrm{new}})}} \tag{8.105}$$

$$D_k^{a\,1} = \frac{\partial}{\partial \mathbf{Z}_k(\mathrm{X}_{\mathrm{new}})} \left\{\log P(\mathbf{y}_{\mathrm{new}\,k}|\mathbf{Z}_k(\mathrm{X}_{\mathrm{new}}), \mathbf{U}_{\mathrm{new}\,k}, \hat{\alpha}_k)\right\} \Big|_{\substack{\mathbf{Z}=\mathbf{Z}^\star(\mathrm{X}_{\mathrm{new}}) \\ \mathbf{U}_{\mathrm{new}}=\mathbf{U}^\star_{\mathrm{new}}(\mathrm{X}_{\mathrm{new}})}} \tag{8.106}$$

$$Bx_k = \frac{\mu_{\mathbf{Z}_k\,i}}{V_{\mathbf{Z}_k\,i,i}}, \tag{8.107}$$

$$D_{1\,k}^a = D_k^{a\,2}, \tag{8.108}$$

$$D_{2\,k}^a = D_{1\,k}^a, \tag{8.109}$$

$$D_{3\,k}^a = D_{1\,k}^a, \tag{8.110}$$

$$D_{4\,k}^a = D_{1\,k}^a, \tag{8.111}$$

$$\tag{8.112}$$

For all $k = 1 : K$, define

$$E_{1k}^a = \exp\left(\mathbf{U}^\star_{\mathrm{new}\,k}(\mathrm{X}_{\mathrm{new}}) + \mathbf{Z}^\star_k(\mathrm{X}_{\mathrm{new}})\right), \tag{8.113}$$

$$E_{2k}^a = \exp(-E_{1k}^a), \tag{8.114}$$

$$E_{3k}^a = \frac{E_{1k}^a}{1 + E_{1k}^a}, \tag{8.115}$$

$$E_{4k}^a = (E_{3k}^a)^{\hat{\alpha}_k}. \tag{8.116}$$

then if

$$f^a_{1k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) = I(\mathrm{y}_{\text{new }k} \neq 0)\Big[\mathrm{y}_{\text{new }k} - E^a_{1k}\Big] \\ I(\mathrm{y}_{\text{new }k} \neq 0)\Big[\hat{\alpha}_k\,(1 - E^a_{3k})(8.117$$

$$f^a_{2k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) = I(\mathrm{y}^a_k = 0)\Big[- E^a_{2k}E^a_{4k}E^a_{1k} - E^a_{4k}(1 - E^a_{2k}) \\ \times(-E^a_{3k} + 1)\hat{\alpha}_k\Big]\big[1 - E^a_{4k}(1 - E^a_{2k})\big]^{-1}(8.118$$

$$g^a_{1k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) = I(\mathrm{y}_{\text{new }k} \neq 0)\Big[- E^a_{1k} + \hat{\alpha}_k(E^a_{3k} - 1)E^a_{3k}\Big](8.119$$

$$g^a_{2k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) = -I(\mathrm{y}_{\text{new }k} = 0)\Big[E^a_{2k}E^a_{4k}E^a_{1k} - E^a_{4k}(1 - E^a_{2k}) \\ \times(-E^a_{3k} + 1)\hat{\alpha}_k\Big]^2\big[1 - E^a_{4k}(1 - E^a_{2k})\big]^{-1}(8.120$$

$$g^a_{3k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) = I(\mathrm{y}_{\text{new }k} = 0)\Big[1 - E^a_{4k}(1 - E^a_{2k})\Big]^{-1} \\ \times\Big[E^a_{2k}E^a_{4k}(E^a_{1k})^2 - E^a_{4k}(1 - E^a_{2k})(2(E^a_{3k})^2 \\ -3E^a_{3k} + 1)\hat{\alpha}_k - E^a_{4k}(1 - E^a_{2k})(-E^a_{3k} + 1)^2 \\ \times(-1 + \hat{\alpha}_k)\hat{\alpha}_k - 2E^a_{2k}E^a_{4k}(-E^a_{3k} + 1) \\ \times E^a_{1k}\hat{\alpha}_k\Big]. \qquad (8.121$$

$$D^{a\,1}_k = f^a_{1k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) + f^a_{2k}\big(\mathbf{y}_k, \mathbf{U}^m_{kj}, \mathbf{Z}^m_k, \hat{\alpha}_k\big)(8.122)$$

$$D^{a\,2}_k = g^a_{1k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) \\ + g^a_{2k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) \\ + g^a_{3k}\big(\mathbf{y}_{\text{new }k}, \mathbf{U}^\star_{\text{new }k}(\mathrm{X}_{\text{new}}), \mathbf{Z}^\star_k(\mathrm{X}_{\text{new}}), \hat{\alpha}_k\big) \qquad (8.123)$$

then

$$B^a_{2\,k=1:K} = D^{a\,1}_{k=1:K} - \mathbf{U}^{\star}_{\text{new}}(\mathrm{X}_{\text{new}}) \begin{pmatrix} D^a_{1\,k=1:K} & D^a_{2\,k=1:K} \\ D^a_{3\,k=1:K} & D^a_{4\,k=1:K} \end{pmatrix}_{2K\times 2K}, \tag{8.124}$$

$$B^a_{1\,k=1:K} = Bx_{k=1:K} + D^{a\,1}_{k=1:K} - \mathbf{Z}^{\star}(\mathrm{X}_{\text{new}}) \begin{pmatrix} D^a_{1\,k=1:K} & D^a_{2\,k=1:K} \\ D^a_{3\,k=1:K} & D^a_{4\,k=1:K} \end{pmatrix}_{2K\times 2K} \tag{8.125}$$

Thus

$$\log q^g(\mathbf{Z}, \mathbf{U}_{\text{new}} | \mathbf{y}_{\text{new}}, \mathbf{y}, i)\Big|_{\substack{\mathbf{Z}=\mathbf{Z}^{\star}(\mathrm{X}_{\text{new}}) \\ \mathbf{U}_{\text{new}}=\mathbf{U}^{\star}_{\text{new}}(\mathrm{X}_{\text{new}})}} = \frac{1}{2} \log\ det\ Q^{\star}_{\text{new}}. \tag{8.126}$$

Bibliography

Allen, R. M. J., Watts, W. A., & Huntley, B. (2000). Weichselian palynostratigraphy, palaeovegetation and palaeoenvironment; the record from lago grande di monticchio, southern italy. *Quaternary International, 73–74*, 91–110.

Attias, H. (1999). Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *UAI*, (pp. 21–30).

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory* (1ed ed.). Wiley.

Billingsley, P. (1986). *Probability and Measure (2nd ed.)*.

Box, G. E. P. & Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis (Wiley Classics Library)*. Wiley–Interscience.

Casella, G. & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician, 46*(3), 167–174.

Chib, S. & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician, 49*(4), 327–335.

Dale, A. I. (1999). *A History of Inverse Probability: From Thomas Bayes to Karl Pearson* (2nd Revised edition ed.). Sources and Studies in the History of Mathematics and Physical Sciences. Springer–Verlag New York Inc.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*(412), 972–985.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (Second ed.). Chapman & Hall/CRC.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov Chain Monte Carlo in Practice*: *Interdisciplinary Statistics* (1 ed. ed.). Chapman & Hall/CRC Interdisciplinary Statistics.

Hall, P., Ormerod, J. T., & Wand, M. P. (2011). Theory of Gaussian Variational Approximation for a Poisson Mixed Model. *Statistica Sinica*, *21*, 369–389.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B., & Mitchell, F. J. G. (2006). Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, *169*(3), 395–438.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hinton, G. E. & van Camp, D. (1993). Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the sixth annual conference on Computational learning theory*, (pp. 5–13)., New York, NY, USA. ACM.

Hoadley, B. (1970). A Bayesian look at inverse linear regression. *Journal of American Statistical Association*, *65*(329), 356–369.

Hunter, W. G. & Lamboy, W. F. (1981). A Bayesian Analysis of the Linear Calibration Problem. *Technometrics*, *23*, 323–350.

Isserlis, L. (1936). Inverse Probability. *Journal of the Royal Statistical Society, 99*(1), 130–137.

Jaakkola, T. S. (2000). Tutorial on variational approximation methods.

Jaakkola, T. S. & Jordon, M. I. (1999). Improving the mean field approximation via the use of mixture distributions, 163–173.

Jordan, M. I. (1999). An introduction to variational methods for graphical models. In *Machine Learning*, (pp. 183–233). MIT Press.

Krutchkoff, R. G. (1967). Classical and Inverse Regression Methods of Calibration. *Technometrics, 9*(3), 425–439.

Kullback, S. (1997). *Information Theory and Statistics.* Dover Publications Inc.

Kullback, S. & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Lee, P. M. (2004). *Bayesian Statistics: An Introduction* (Third ed.). A Hodder Arnold Publication.

MacKay, D. J. C. (1998). Ensemble Learning for Hidden Markov Models. In Jordan, M. I. & Solla, M. J. K. S. A. (Eds.), *Advances in Neural Information Processing Systems 10.* The MIT Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21*(6), 1087–1092.

Minka, T. P. (2001). Expectation Propagation for approximate Bayesian inference. In Breese, J. S. & Koller, D. (Eds.), *UAI*, (pp. 362–369). Morgan Kaufmann.

Murphy, K. P., Weiss, Y., & Jordan, M. I. (1999). Loopy Belief Propagation for Approximate Inference: An Empirical Study. (pp. 467–475).

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization* (2nd ed.). Springer.

Ormerod, J. T. & Wand, M. P. (2010). Explaining Variational Approximations. *The American Statistician*, *64*, 140–153.

Racine-Poon, A. (1988). A Bayesian Approach to Nonlinear Calibration Problems. *Journal of the American Statistical Association*, *83*(403), 650–656.

Ridout, M., Demetrio, C. G. B., & Hinde, J. (1998). Models for Count Data with Many Zeros. In *Proceedings of the XIXth International Biometric Conference*, Invited Papers, (pp. 179–192).

Robert, P. C. & Casella, G. (1999). *Monte Carlo Statistical Methods*, volume 1. Springer–Verlag.

Rue, H. & Held, L. (2005). *Gaussian Markov random fields. Theory and applications (Monographs on Statistics and Applied Probability)* (1 ed.)., volume 104. Chapman & Hall/CRC.

Rue., H., Martino, S., & Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society Series B*, *71(2)*, 1–35.

Salter-Townshend, M. (2009). *Fast Approximate Inverse Bayesian Inference in non-parametric Multivariate Regression with application to palaeoclimate reconstruction.* PhD thesis, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland.

Saul, L. K., Jaakkola, T., & Jordan, M. I. (1996). Mean Field Theory for Sigmoid Belief Networks. *Journal of Artificial Intelligence Research*, *4*, 61–76.

Stigler, S. M. (1986). Laplace's 1774 Memoir on Inverse Probability. *Statistical Science*, *1*(3), 359–363.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*, 1701–1762.

Tierney, L. & Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, *81*(393), 82–86.

Vatsa, R. & Wilson, S. (2010). The Variational Bayes Method for Inverse Regression Problems with an Application to the Palaeoclimate Reconstruction. *Journal of Combinatorics, Information & System Sciences*, *35*(1–2), 221–248.

Šmídl, V. & Quinn, A. (2006). *The Variational Bayes Method in Signal Processing.* Signals and Communication Technology. Springer.

Wang, B. & Titterington, D. M. (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. *Artificial Intelligence*, *158*, 97–106.

Wang, B. & Titterington, D. M. (2005). Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations. In *AISTATS05 , Society for Artificial Intelligence and Statistics*, (pp. 373–380).

Waterhouse, S. & Robinson, D. M. . T. (1996). Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems 8.* MIT Press.

Winn, J. & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, *6*, 661–694.

Zabell, S. (1989). R. A. Fisher on the History of Inverse Probability. *Statistical Science*, *4*(3), 247–256.