VARIATIONAL APPROXIMATIONS FOR DENSITY DECONVOLUTION

Yue Chang

VARIATIONAL APPROXIMATIONS FOR DENSITY DECONVOLUTION

A Dissertation Presented

by

YUE CHANG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2018

Department of Mathematics and Statistics

VARIATIONAL APPROXIMATIONS FOR DENSITY DECONVOLUTION

A Dissertation Presented

by

YUE CHANG

Approved as to style and content by:

John Staudenmayer, Chair

Anna Liu, Member

Krista Gile, Member

Leontine Alkema, Member

Nathaniel Whitaker, Department Head
Mathematics and Statistics

DEDICATION

To my parents.

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my advisor Prof. John Staudenmayer for his guidance in the past five years. I have benefited from his patience, encouragement and immense knowledge. I am lucky to have him as my advisor.

I would like to thank the other members in my dissertation committee, Prof. Anna Liu, Prof. Krista Gile and Prof. Leontine Alkema, for their insightful comments and questions.

I would also like to thank many faculty members who have been my instructors or mentors at University of Science and Technology of China and University of Massachusetts Amherst.

Thanks to my fellow graduate students in mathematics and statistics department for their companions during many seminars and workshops.

Last, I am forever grateful to my parents for their love and support.

ABSTRACT

VARIATIONAL APPROXIMATIONS FOR DENSITY DECONVOLUTION

SEPTEMBER 2018

YUE CHANG

B.S., UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

PH.D. UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor John Staudenmayer

This thesis considers the problem of density estimation when the variables of interest are subject to measurement error. The measurement error is assumed to be additive and homoscedastic. We specify the density of interest by a Dirichlet Process Mixture Model and establish variational approximation approaches to the density deconvolution problem. Gaussian and Laplacian error distributions are considered, which are representatives of supersmooth and ordinary smooth distributions, respectively. We develop two variational approximation algorithms for Gaussian error deconvolution and one variational approximation algorithm for Laplacian error deconvolution. Their performances are compared to deconvoluting kernels and Monte Carlo Markov Chain method by simulation experiments. A conjecture based on hidden variables categorization is proposed to explain why two variational approximation algorithms for Gaussian error deconvolution perform differently. We establish a stochastic variational approximation algorithm for Gaussian error deconvolution, which improves the performance of variational approximation algorithm and performs as well as MCMC method at faster speed. The stochastic variational approximation algorithm is applied to simulation experiments and an example of physical activity measurements.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This thesis considers the problem of density estimation when the variables from the density of interest cannot be observed directly. Observations are measurements of the variable of interest and subject to additive measurement error. The problem can be formulated as a one-way random effects model,

$$y_{ij} = x_i + u_{ij},\ i = 1, \cdots, n,\ j = 1, \cdots, m_i \tag{1.1}$$

where y_{ij}s are observations, x_is are unobserved variables from an unknown density f_x, u_{ij}s are measurement errors from a known density f_u. Furthermore, we assume $x_i \overset{\text{i.i.d.}}{\sim} f_x(x)$, $u_{ij} \overset{\text{i.i.d.}}{\sim} f_u(u)$, u_{ij} are independent from x_i. This problem arises whenever estimation of f_x is desired. The problem of estimating f_x from observations y_{ij}s is called measurement error deconvolution. An application of deconvolution is to estimate the distribution of long-term mean sedentary time in a population. In this case, the individual's long-term mean sedentary time cannot be observed directly. Repeated device-based measurements are subject to measurement errors.

There are a lot of literature studying the deconvolution problem. We review two important nonparametric deconvolution approaches: deconvoluting kernel (DK) method and Bayesian nonparametric deconvolution method. The deconvolution kernel method proposed in [Stefanski and Carroll, 1990] is based on deconvoluting a kernel estimator of the observed data. The shortcoming of DK methods is that rate of convergence is very low for Gaussian measurement error. [Carroll and Hall, 1988] shows that the best possible rate of convergence of the mean integrated squared error (MISE) of a deconvoluting kernel estimator is only $(\log n)^{-\frac{k}{2}}$ if the density of interest has kth bounded derivatives and errors are Gaussian. [Sarkar et al., 2014] proposes the Bayesian non-parametric model to solve the deconvolution problem, where the density of interest f_x is specified by a Dirichlet Process Mixture Model (DPMM). The Markov Chain Monte Carlo (MCMC) sampling of the Bayesian non-parametric model is computational extensive and has low computing speed. In this thesis, we develop mean-field variational approximation (VA) type approaches to the Bayesian nonparametric model. We approximate the DPMM by a finite mixture model with symmetric Dirichlet prior. Two types of measurement error are considered for the deconvolution problem: Gaussian and Laplacian error. We establish a stochastic variational approximation (SVA) approach for Gaussian error deconvolution which performs better than DK method and achieves comparable accuracy with MCMC method. We investigate VA approaches for Laplacian error and explain the reasons why VA for Laplacian error performs worse than VA for Gaussian error.

The rest of the thesis is organized as follows. In Chapter 2, we review deconvoluting kernels and MCMC sampling for the Bayesian nonparametric deconvolution model, which are the benchmark methods. In Chapter 3, we establish two VA algorithms for Gaussian error deconvolution. Algorithm A includes $\mathbf{x}$ as latent variables in the posterior distribution, while algorithm B excludes $\mathbf{x}$ by integrating out $\mathbf{x}$. The performances of the two algorithms are compared through simulation experiments. We propose a conjecture to explain why algorithm B outperforms algorithm A and improve VA algorithm B by stochastic optimization. In Chapter

4, we develop a VA algorithm for Laplacian error and compare VA with benchmark methods through a simulation study. Chapter 5 applies the SVA algorithm to a physical activity dataset. Chapter 6 concludes the thesis and discusses extensions.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews the state-of-art deconvolution methods. Section 2.1 reviews deconvoluting kernels, which are nonparametric approaches based on a transformation of the kernel density estimator. Section 2.2 presents a Bayesian nonparametric deconvolution model built on Dirichlet Process Mixture Model (DPMM) and reviews its Markov Chain Monte Carlo (MCMC) algorithm. Section 2.3 talks about other existing deconvolution methods.

2.1 Deconvoluting Kernels

The deconvolution problem is originally formulated as

$$y_i = x_i + u_i,\ i = 1, \cdots, n \tag{2.1}$$

where y_is are observations, x_is are independent variables from the unknown distribution f_x, u_is are independent measurement errors from a known distribution f_u. The problem is to estimate the density of interest f_x. The deconvoluting kernels can be derived from the inverse Fourier transform,

$$f_x(x) = \frac{1}{2\pi}\int e^{-itx}\psi_x(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{\psi_y(t)}{\psi_u(t)}dt$$

where ψ_x, ψ_y, ψ_u denote the characteristic functions of the variable x, y, u respectively, that is, $\psi_\eta(t) = \int e^{it\eta} f_\eta(\eta) d\eta$ for $\eta \in \{x, y, u\}$. A function $K(x)$ can be used as a kernel if $K(x)$ satisfies $\int K(x)dx = 1$ and is symmetric about origin. Replacing $f_y(y)$ in $\psi_y(t)$ by its kernel estimator $(nh)^{-1} \sum_{i=1}^n K((y - y_i)/h)$ gives deconvolution kernel density estimator

$$\hat{f}_x(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K_h^* \left(\frac{x - y_i}{h}; h \right) \tag{2.2}$$

where

$$K_h^*(z; h) = \frac{1}{2\pi} \int e^{-itz} \frac{\psi_K(t)}{\psi_u(t/h)} dt \tag{2.3}$$

is called the deconvolution kernel [Stefanski and Carroll, 1990]. Given a bandwidth value h the MISE is defined by

$$MISE(h) = \mathrm{E} \int \left\{ \hat{f}_x(x; h) - f_x(x) \right\}^2 dx \tag{2.4}$$

From the calculation in [Wand, 1998] the MISE can be derived as

$$\begin{aligned} MISE(h) = &(2\pi nh)^{-1} \int \psi_K(t)^2 \left| \psi_u(t/h) \right|^{-2} dt \\ &+ (2\pi)^{-1} \int \left\{ (1 - n^{-1}) \psi_K(ht)^2 - 2\psi_K(ht) + 1 \right\} \left| \psi_x(t) \right|^2 dt \end{aligned} \tag{2.5}$$

The issue of how to choose bandwidth has been discussed in [Stefanski and Carroll, 1990] [Fan, 1991] [Fan, 1992][Hesse, 1999] [Delaigle and Gijbels, 2004a] [Delaigle and Gijbels, 2004b]. The R package **fDKDE** by [Delaigle and Wang, 2015] provides two options for choosing bandwidth : (1)two-stage plug-in bandwidth as in [Delaigle and Gijbels, 2002] (2)cross-validated bandwidth as in [Stefanski and Carroll, 1990]. [Delaigle and Gijbels, 2004b] compares the plug-in bandwidth selectors with bootstrap and cross-validated bandwidth selectors and concludes plug-in and bootstrap bandwidth selectors perform similarly, and both outperform cross-validated bandwidth selector. Two-stage plug-in bandwidth is used for DK method in simulation experiments of this thesis. For

Gaussian error the characteristic function ψ_K with compact support leads to a finite integral $K_h^*(z;h)$. The choice of kernel function in **fDKDE** is

$$K(x) = \frac{48\cos x}{\pi x^4}\left(1-\frac{15}{x^2}\right)-\frac{144\sin x}{\pi x^5}\left(2-\frac{5}{x^2}\right) \tag{2.6}$$

Its corresponding characteristic function is

$$\psi_K(t) = (1-t^2)^3, \quad |t|<1 \tag{2.7}$$

[Delaigle and Meister, 2008] extends the deconvoluting kernel method to models with replicated measurements where the density of measurement error is known. Since $\{y_{1\cdot},\cdots,y_{n\cdot}\}$ is a sufficient statistic for f_x, there is no loss of information when applying the estimator (2.2) to $\bar{y}_{i\cdot} = x_i + \bar{u}_{i\cdot}$, where $\overline{y}_{i\cdot} = \sum_{j=1}^{m_i} y_{ij}/m_i$, $\overline{u}_{i\cdot} = \sum_{j=1}^{m_i} u_{ij}/m_i$. Replacing $\psi_u(t)$ by $\psi_{\bar{u}_{i\cdot}}(t) = \psi_u(t/\sqrt{m_i})$ in (2.2) gives an estimator for the model with replications. If the density of u is unknown and the model has replicated measurements, $\psi_u(t)$ can be estimated by

$$\hat{\psi}_u(t) = \left| \frac{2}{\sum_{i=1}^n m_i(m_i-1)} \sum_{i=1}^n \sum_{1\le j_1<j_2\le m_i} \cos\left(t(y_{ij_1}-y_{ij_2})\right) \right|^{\frac{1}{2}} \tag{2.8}$$

under the assumption that the f_u is symmetric. The asymptotic properties and convergence rate of the estimator of f_x using (2.8) are discussed in [Delaigle et al., 2008].

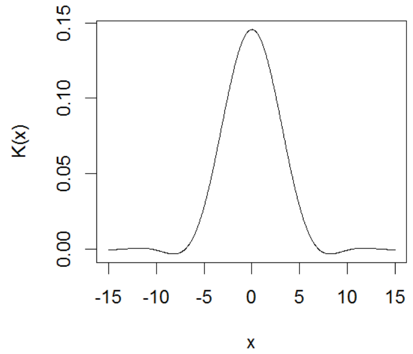

Figure 2.1: The kernel function used in fDKDE

2.2 Bayesian Approaches

2.2.1 Dirichlet Process Mixture Models

Reconsider the deconvolution problem

$$y_i = x_i + u_i,\ i = 1, \cdots, n \tag{2.9}$$

where y_is are observations, $x_i \overset{\text{i.i.d.}}{\sim} f_x$, $u_i \overset{\text{i.i.d.}}{\sim} f_u$, f_u is known, f_x is the density to be estimated. Dirichlet process [Ferguson, 1973] mixture model is used for density deconvolution problems in [Sarkar et al., 2014]. The density of interest f_x is specified as a mixture of normals. Let $\mathcal{N}(\,\cdot\,|\,\mu,\,\sigma^2)$ denote a normal distribution with mean μ and variance σ^2, $\mathcal{IG}(\gamma, \beta)$ denote an inverse gamma prior with shape parameter γ and scale parameter β. If $\mu|\sigma^2, \mu_0, \lambda_0 \sim \mathcal{N}(\mu_0, \sigma^2/\lambda_0),\ \sigma^2|\gamma_0, \beta_0 \sim \mathcal{IG}(\gamma_0, \beta_0)$, then (μ, σ^2) has a normal-inverse-gamma distribution, denoted by $(\mu,\, \sigma^2) \sim \mathcal{NIG}(\mu_0,\, \lambda_0,\, \gamma_0,\, \beta_0)$. [Ishwaran and Zarepour, 2002] proves that the DPMM can be obtained by taking the limit as K goes to infinity in the following mixture model

$$\begin{aligned}
y_i \,|\, x_i &\overset{\text{i.i.d.}}{\sim} f_u(y_i - x_i),\ i = 1, \cdots, n \\
x_i \,|\, c_i,\, \boldsymbol{\phi} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_{c_i},\, \sigma^2_{c_i}\right) \\
c_i \,|\, \boldsymbol{\pi} &\overset{\text{i.i.d.}}{\sim} \text{Categorical}\left(\pi_1,\, \pi_2, \cdots, \pi_K\right) \\
\boldsymbol{\pi} &\sim \text{Dirichlet}\left(\alpha/K, \alpha/K, \cdots, \alpha/K\right) \\
\phi_c &\overset{\text{i.i.d.}}{\sim} \mathcal{NIG}\left(\mu_0,\, \lambda_0,\, \gamma_0,\, \beta_0\right)
\end{aligned} \tag{2.10}$$

where c_i denotes the latent cluster associated with x_i. For each cluster c, the variable $\phi_c = (\,\mu_c,\, \sigma^2_c\,)$ determines the distribution of the associated xs. The collection of all $(\,\mu_c,\, \sigma^2_c\,)$ is denoted by $\boldsymbol{\phi}$.

Integrating out the probabilities vector $\boldsymbol{\pi}$ and then taking K to infinity, we can write $p(\mathbf{c}|\alpha)$ as the product of conditional probabilities of the following forms

[Neal, 2000]:

$$
\begin{aligned}
p(c_i = k \ and \ k = c_j \ for \ some \ j < i | c_1, \ldots, c_{i-1}) &= \frac{\sum_{j=1}^{i-1} \delta_k(c_j)}{i-1+\alpha} \\
p(c_i \neq c_j \ for \ all \ j < i | c_1, \ldots, c_{i-1}) &= \frac{\alpha}{i-1+\alpha}
\end{aligned}
$$

where $\delta_k(c) = 1$ if $c = k$, $\delta_k(c) = 0$ otherwise. The posterior $p(\mathbf{x}, \mathbf{c}, \boldsymbol{\phi} | \mathbf{y}, \alpha, \mu_0, \lambda_0, \gamma_0, \beta_0)$ is proportional to

$$
\begin{aligned}
& p(\mathbf{y}, \, \mathbf{x}, \, \mathbf{c}, \, \boldsymbol{\phi} | \alpha, \mu_0, \lambda_0, \gamma_0, \beta_0) \\
& \propto f_u(\mathbf{y} - \mathbf{x}) \times f_x(\mathbf{x} | \mathbf{c}, \boldsymbol{\phi}) \times p(\mathbf{c} | \alpha) \times f_\phi(\boldsymbol{\phi} | \mu_0, \lambda_0, \gamma_0, \beta_0) \\
& \propto f_u(\mathbf{y} - \mathbf{x}) \times \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right\} \right)^{\delta_k(c_i)} \times \frac{\alpha^K \prod_{k \in \mathcal{K}} (n_k - 1)!}{\alpha(1+\alpha) \cdots (n-1+\alpha)} \\
& \quad \times \prod_{k \in \mathcal{K}} \frac{1}{\sigma_k^{2(\gamma_0+1)}} \exp\left\{ -\frac{\beta_0}{\sigma_k^2} \right\} \times \sqrt{\frac{\lambda_0}{\sigma_k^2}} \exp\left\{ -\frac{\lambda_0 (\mu_k - \mu_0)^2}{2\sigma_k^2} \right\} \qquad (2.11)
\end{aligned}
$$

where $n_k = \sum_{i=1}^{n} \delta_k(c_i)$, $\mathcal{K} = \bigcup_{i=1}^{n} \{c_i\}$, and K denotes the cardinality of $\mathcal{K}$.

[Ishwaran and Zarepour, 2002] points that a finite mixture with symmetric Dirichlet prior, which can be obtained by fixing K in (2.10), strongly approximates a Dirichlet process. We call (2.10) a truncated DPMM if K is fixed. The posterior of the truncated DPMM is proportional to

$$
\begin{aligned}
& p(\mathbf{y}, \, \mathbf{x}, \, \mathbf{c}, \, \boldsymbol{\pi}, \, \boldsymbol{\phi} | \alpha, \mu_0, \lambda_0, \gamma_0, \beta_0) \\
& \propto f_u(\mathbf{y} - \mathbf{x}) \times f(\mathbf{x} | \mathbf{c}, \boldsymbol{\phi}) \times f(\mathbf{c} | \boldsymbol{\pi}) \times f(\boldsymbol{\pi} | \alpha) \times f(\boldsymbol{\phi} | \mu_0, \lambda_0, \gamma_0, \beta_0) \\
& \propto f_u(\mathbf{y} - \mathbf{x}) \times \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right\} \right)^{\delta_k(c_i)} \times \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{\delta_k(c_i)} \\
& \quad \times \frac{\Gamma(\alpha)}{\Gamma(\frac{\alpha}{K})^K} \prod_{k=1}^{K} \pi_k^{\frac{\alpha}{K}-1} \times \prod_{k=1}^{K} \frac{1}{\sigma_k^{2(\gamma_0+1)}} \exp\left\{ -\frac{\beta_0}{\sigma_k^2} \right\} \times \sqrt{\frac{\lambda_0}{\sigma_k^2}} \exp\left\{ -\frac{\lambda_0 (\mu_k - \mu_0)^2}{2\sigma_k^2} \right\}
\end{aligned}
$$

(2.12)

2.2.2 MCMC Algorithms

We consider the Gaussian and Laplacian distributions for measurement error, which are two important cases having super-smoothness and ordinary-smoothness, respectively. The density of interest $f_x(x)$ can be estimated by Markov Chain Monte Carlo (MCMC) method. From the joint posterior distribution (2.11), we draw samples from posterior distribution using Gibbs sampler and Metropolis-Hasting algorithm. Specifically, each MCMC iteration contains the following steps.

Algorithm 1. *1.* ***Updating the distribution of x using Gibbs sampler.*** *([Neal, 2000] Algorithm 2)*

The full conditional distribution of c_i is given by

$$p(c_i = k,\, k \in \mathbf{c}_{-i} | \mathbf{x},\, \mathbf{c}_{-i},\, \boldsymbol{\phi},\, \alpha) = l \frac{n_{-i,k}}{n-1+\alpha} \mathcal{N}(x_i \,|\, \mu_k, \sigma_k^2)$$

$$p(c_i \notin \mathbf{c}_{-i} | \mathbf{x},\, \mathbf{c}_{-i},\, \boldsymbol{\phi},\, \alpha) = l \frac{\alpha}{n-1+\alpha} \mathcal{T}_{2\gamma_0}\left(\frac{x_i - \mu_0}{N}\right) \times \frac{1}{N}$$

where $M = \sqrt{\beta_0(\lambda_0+1)/(\lambda_0\gamma_0)}$, $n_{-i,k} = \sum_{j \neq i} \delta_k(c_j)$, $\mathcal{T}_\nu(\cdot)$ denotes a student's t-distribution with ν degree of freedom. The constant l is chosen such that $\sum_{k \in \mathbf{c}_{-i}} p(c_i = k | \mathbf{x},\, \mathbf{c}_{-i},\, \boldsymbol{\phi},\, \alpha) + p(c_i \notin \mathbf{c}_{-i} | \mathbf{x},\, \mathbf{c}_{-i},\, \boldsymbol{\phi},\, \alpha) = 1$.

For $k \in \mathbf{c}$, the joint full conditional distribution of $(\mu_k,\, \sigma_k^2)$ is $\mathcal{NIG}(\mu_{nk},\, \lambda_{nk},\, \gamma_{nk},\, \beta_{nk})$, where $\lambda_{nk} = \lambda_0 + n_k$, $\gamma_{nk} = \gamma_0 + n_k/2$, $\mu_{nk} = (\lambda_0\mu_0 + \sum_{i \in I_k} x_i)/(\lambda_0 + n_k)$ and $\beta_{nk} = \beta_0 + (\sum_{i \in I_k} x_i^2 + \lambda_0\mu_0^2 - \lambda_{nk}\mu_{nk}^2)/2$.

2. ***Updating x***

(a) Update x_i by Gibbs sampler if $u_i \sim \mathcal{N}(0, \sigma_u^2)$, where σ_u^2 is known. The full conditional distribution of x_i is given by $p(x_i \,|\, \mathbf{y},\, \mathbf{c},\, \boldsymbol{\phi},\, \alpha) = p(x_i \,|\, y_i,\, (\mu_{c_i},\, \sigma_{c_i}^2)) \propto$ $f_x(x_i \,|\, (\mu_{c_i},\, \sigma_{c_i}^2)) \times f_u(y_i - x_i)$

$\sim \mathcal{N}\left(x_i \,|\, (\sigma_u^2\mu_{c_i} + \sigma_{c_i}^2 y_i)/(\sigma_u^2 + \sigma_{c_i}^2), \sigma_{c_i}^2\sigma_u^2/(\sigma_{c_i}^2 + \sigma_u^2)\right)$.

(b) Update x_i by random walk Metropolis algorithm if $u_i \sim \text{Laplace}(0, b)$, where $b = \sqrt{\sigma_u^2/2}$ and σ_u^2 is known. The full conditional distribution of x_i given by $p(x_i \,|\, \mathbf{y},\, \mathbf{c},\, \boldsymbol{\phi},\, \alpha) = p(x_i \,|\, y_i,\, (\mu_{c_i},\, \sigma_{c_i}^2)) \propto f_x(x_i \,|\, (\mu_{c_i},\, \sigma_{c_i}^2)) \times f_u(y_i - x_i) \propto \exp\{-\frac{1}{2\sigma_{c_i}^2}(x_i - \mu_{c_i})^2 - \frac{|y_i - x_i|}{b}\}$. For $i = 1, \cdots, n$, we propose a new value for x_i with proposal distribution $q(x_i^*|x_i) = (2\pi\sigma_p^2)^{-\frac{1}{2}} \exp\{-(x_i^* - x_i)^2/(2\sigma_p^2)\}$. According to pre-run results, setting $\sigma_p^2 = \hat{\text{var}}(y)$ produces an acceptance rate between 25% and 60%, We update x_i to x_i^* with probability

$$\begin{aligned}
&\min\left\{1,\ \frac{f_u(y_i - x_i^*) f_x(x_i^*|\mu_{c_i}, \sigma_{c_i}^2)}{f_u(y_i - x_i) f_x(x_i|\mu_{c_i}, \sigma_{c_i}^2)}\right\} \\
= &\min\left\{1,\ \exp\left\{-\frac{1}{2\sigma_{c_i}^2}(x_i^* - \mu_{c_i})^2 - \frac{|y_i - x_i^*|}{b} + \frac{1}{2\sigma_{c_i}^2}(x_i - \mu_{c_i})^2 + \frac{|y_i - x_i|}{b}\right\}\right\}
\end{aligned}$$

From [Escobar and West, 1995], a Monte Carlo estimate of $f_x(\cdot|\mathbf{y})$ based on S samples from the posterior is given by

$$\hat{f}_x(x|\mathbf{y}) = \frac{1}{S} \sum_{s=1}^{S} \left[\sum_{i=1}^{n} \frac{1}{n + \alpha} \mathcal{N}(x|\mu_{c_i}^{(s)}, \sigma_{c_i}^{2(s)}) + \frac{\alpha}{n + \alpha} \frac{1}{M} \mathcal{T}_{2\gamma_0}\left(\frac{x - \mu_0}{M}\right) \right] \tag{2.13}$$

2.3 Other Existing Deconvolution Methods

Besides the deconvoluting kernel and DPMM methods, other nonparametric deconvolution methods have been studied by researchers. [Carroll and Hall, 2004] proposed an orthogonal series method. It expresses f_x in an orthogonal expansion with estimable coefficients and the functions in the orthogonal series may be polynomials or trigonometric functions. [Pensky et al., 1999] [Fan and Koo, 2002] [Donoho et al., 1996] discussed deconvolution by wavelets, which is a variety of orthogonal series method.

If one has a parametric assumption for f_x, e.g., normal, skew-normal, gamma, the SNP (seminonparametric) family [Zhang and Davidian, 2001], likelihood methods can be used to estimate the unknown parameters and obtain the density estimate for x. [Carroll et al., 2006] gives a review of parametric deconvolution methods in chapter 12.

CHAPTER 3

VARIATIONAL APPROXIMATION APPROACHES FOR GAUSSIAN ERROR DECONVOLUTION

In Chapter 3, we establish VA approaches for Gaussian error deconvolution and compare our approaches with other nonparametric methods (DK and MCMC) through simulation experiments. Section 3.1 reviews mean-field variational approximation. Section 3.2 develops VA algorithm A for Gaussian error deconvolution which includes $\mathbf{x}$ as latent variables in the posterior distribution. Section 3.3 establishes algorithm B which excludes $\mathbf{x}$ by integrating out $\mathbf{x}$. The performances of the two algorithms are compared through simulation experiments in section 3.4. We propose a conjecture to explain why algorithm B outperforms algorithm A in section 3.5. Section 3.6 develops a stochastic variational approximation (SVA) approach, which applies stochastic optimization to VA algorithm B. Section 3.7 compares SVA to VA algorithm B on simulated datasets and shows that SVA improves VA algorithm B. Section 3.8 adds SVA and MCMC to the deconvolution problems in [Wand, 1998] and shows that SVA outperforms DK and performs similarly with MCMC at a faster speed.

3.1 Introduction to Variational Approximation

Variational approximations (VA) are a class of alternatives to MCMC for approximating marginal likelihood and posterior densities. Variational approximations tend to be more computationally efficient than MCMC, but statistical properties of variational approximations are less studied than MCMC. [Blei et al., 2016] gives a review of statistical research on variational inferences. [Hall et al., 2011] [You et al., 2014][Wang et al., 2006] develop theory regarding consistency and asymptotic properties of point estimator of variational approximations for particular models. [Wang and Blei, 2017] investigates the frequentist consistency and asymptotic properties of variational bayes estimators. [Blei and Jordan, 2006] presents a mean-field variational inference algorithm for DPMM based on the truncated stick-breaking representation of a DPMM. [Kurihara et al., 2007] experimentally shows that there is little difference between the variational inference in the truncated stick-breaking representation and the finite mixture model with symmetric Dirichlet priors. In this paper we develop VA algorithms for the deconvolution problem.

The basic idea of variational approximations is to approximate a posterior or marginal distribution by a family of distributions and to solve it by optimization. In this paper, we apply the most common class of variational approaches known as mean-field approximation [Parisi, 1988]. Let $\boldsymbol{\theta}$ represent the collection of latent variables in the model. Suppose the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is approximated by a variational distribution $q(\boldsymbol{\theta})$, which is a family of distributions on $\boldsymbol{\theta}$, the mean-field approximation aims to minimize the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] from $q(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\mathbf{y})$. The KL divergence from P to Q

is defined as $D_{KL}(P(x)||Q(x)) = \int P(x)\text{log}P(x)dx - \int P(x)\text{log}Q(x)dx$, then

$$D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) = \int q(\boldsymbol{\theta})\log q(\boldsymbol{\theta})d\boldsymbol{\theta} - \int q(\boldsymbol{\theta})\log p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \tag{3.1}$$

It follows from the properties of KL divergence that $D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \geq 0$ with equality holding if and only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$ almost everywhere.

The logarithm of the marginal likelihood satisfies:

$$\begin{aligned} \log p(\mathbf{y}) &= \int q(\boldsymbol{\theta})\log p(\mathbf{y})d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta})\log\left(\frac{p(\mathbf{y},\boldsymbol{\theta})/q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})/q(\boldsymbol{\theta})}\right)d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta})\log\left(\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)d\boldsymbol{\theta} + D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \\ &\geq \int q(\boldsymbol{\theta})\log\left(\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)d\boldsymbol{\theta} \end{aligned} \tag{3.2}$$

Define the lower bound of the marginal likelihood (or lower bound of the evidence, ELBO)

$$\underline{p}(\mathbf{y};q) = \exp\int q(\boldsymbol{\theta})\log\left(\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)d\boldsymbol{\theta} \tag{3.3}$$

Minimizing $D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y}))$ is equivalent to maximizing $\log\underline{p}(\mathbf{y};q)$. To make the mean-field approximation more tractable, $q(\boldsymbol{\theta})$ is restricted to factorize into $q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_{\theta_j}(\boldsymbol{\theta}_j)$ at the cost of degrading the dependency among $\boldsymbol{\theta}_j$, $j = 1, \cdots, J$, where $\{\boldsymbol{\theta}_j\}_{j=1}^{J}$ is a disjoint vector partition of $\boldsymbol{\theta}$. The parameters of the variational distributions $q_{\theta_j}(\boldsymbol{\theta}_j)$ are called variational parameters. We optimize $\log\underline{p}(\mathbf{y};q)$ by coordinate ascent method, iteratively maximizing $\log\underline{p}(\mathbf{y};q)$ with respect to each variational distribution $q_{\theta_j}(\boldsymbol{\theta}_j)$ while holding other variational distributions fixed. In particular, optimization of $\log\underline{p}(\mathbf{y};q)$ with respect to $q_{\theta_j}(\boldsymbol{\theta}_j)$ can be achieved by

$$q^*_{\theta_j}(\boldsymbol{\theta}_j) \propto \exp\left\{\mathrm{E}_{-\theta_j}\log p(\mathbf{y},\boldsymbol{\theta})\right\} \tag{3.4}$$

where $\mathrm{E}_{-\theta_j}$ denotes the expectation with respect to the density $\prod_{l\neq j} q_{\theta_l}(\boldsymbol{\theta}_l)$. The formula (3.4) can be derived as follows:

$$\begin{aligned}
\log \underline{p}(\mathbf{y};q) &= \int \left\{ \mathrm{log}p(\mathbf{y},\boldsymbol{\theta}) - \sum_l \mathrm{log}q_{\theta_l}(\boldsymbol{\theta}_l) \right\} \prod_l q_{\theta_l}(\boldsymbol{\theta}_l) \prod_l d\boldsymbol{\theta}_l \\
&= \int \left\{ \int \mathrm{log}p(\mathbf{y},\boldsymbol{\theta}) \prod_{l\neq j} q_{\theta_l}(\boldsymbol{\theta}_l) \prod_{l\neq j} d\boldsymbol{\theta}_l \right\} q_{\theta_j}(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j \\
&\quad - \int q_{\theta_j}(\boldsymbol{\theta}_j)\mathrm{log}q_{\theta_j}(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j + Q(\boldsymbol{\theta}_{-j}) \qquad (3.5)
\end{aligned}$$

where $\boldsymbol{\theta}_{-j} = \boldsymbol{\theta}\backslash\boldsymbol{\theta}_j$, $Q(\boldsymbol{\theta}_{-j})$ only depends on $\boldsymbol{\theta}_{-j}$. Define a new posterior density $\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y})$ by

$$\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y}) = \frac{\exp \int \mathrm{log}p(\mathbf{y},\boldsymbol{\theta}) \prod_{l\neq j} q_{\theta_l}(\boldsymbol{\theta}_l) \prod_{l\neq j} d\boldsymbol{\theta}_l}{\int \left\{ \exp \int \mathrm{log}p(\mathbf{y},\boldsymbol{\theta}) \prod_{l\neq j} q_{\theta_l}(\boldsymbol{\theta}_l) \prod_{l\neq j} d\boldsymbol{\theta}_l \right\} d\boldsymbol{\theta}_j} \qquad (3.6)$$

then equation (3.5) can be written as

$$\begin{aligned}
\log \underline{p}(\mathbf{y};q) &= \int q_{\theta_j}(\boldsymbol{\theta}_j)\mathrm{log}\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y}) d\boldsymbol{\theta}_j - \int q_{\theta_j}(\boldsymbol{\theta}_j)\mathrm{log}q_{\theta_j}(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j + \tilde{Q}(\boldsymbol{\theta}_{-j}) \\
&= -D_{KL}(q_{\theta_j}(\boldsymbol{\theta}_j)||\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y})) + \tilde{Q}(\boldsymbol{\theta}_{-j}) \qquad (3.7)
\end{aligned}$$

Therefore, the optimal $q_{\theta_j}(\boldsymbol{\theta}_j)$ is $q^*_{\theta_j}(\boldsymbol{\theta}_j) = \tilde{p}(\boldsymbol{\theta}_j|\mathbf{y}) \propto \exp\left\{\mathrm{E}_{-\theta_j}\log p(\mathbf{y},\boldsymbol{\theta})\right\}$.

Furthermore, if we assume that the prior $p(\boldsymbol{\theta}_j)$ and the full conditional distribution $p(\boldsymbol{\theta}_j|\,\mathbf{y},\boldsymbol{\theta}_{-j})$ are in the conjugate exponential family, the conditional distribution $p(\boldsymbol{\theta}_j|\,\mathbf{y},\boldsymbol{\theta}_{-j})$ can be written as:

$$p(\boldsymbol{\theta}_j|\,\mathbf{y},\boldsymbol{\theta}_{-j}) = h(\boldsymbol{\theta}_j)\mathrm{exp}\left\{\boldsymbol{\zeta}(\boldsymbol{\theta}_{-j},\mathbf{y})^T\mathbf{T}(\boldsymbol{\theta}_j) - A(\boldsymbol{\zeta}(\boldsymbol{\theta}_{-j},\mathbf{y}))\right\} \qquad (3.8)$$

From the formula (3.4) we have

$$\begin{aligned}
\log q^*_{\theta_j}(\boldsymbol{\theta}_j) &= \mathrm{E}_{-\theta_j} \log p(\mathbf{y}, \boldsymbol{\theta}) + C_1 \\
&= \mathrm{E}_{-\theta_j} \{\log p(\boldsymbol{\theta}_j | \, \mathbf{y}, \boldsymbol{\theta}_{-j}) + \log p(\boldsymbol{\theta}_{-j} | \, \mathbf{y})\} + C_1 \\
&= \log h(\boldsymbol{\theta}_j) + \mathrm{E}_{-\theta_j} \left\{\boldsymbol{\zeta}(\boldsymbol{\theta}_{-j}, \mathbf{y})^T\right\} \mathbf{T}(\boldsymbol{\theta}_j) - \mathrm{E}_{-\theta_j} \{A(\boldsymbol{\zeta}(\boldsymbol{\theta}_{-j}, \mathbf{y})) + \log p(\boldsymbol{\theta}_{-j} | \, \mathbf{y})\} + C_1 \\
&= \log h(\boldsymbol{\theta}_j) + \mathrm{E}_{-\theta_j} \left\{\boldsymbol{\zeta}(\boldsymbol{\theta}_{-j}, \mathbf{y})^T\right\} \mathbf{T}(\boldsymbol{\theta}_j) + C_2 \qquad (3.9)
\end{aligned}$$

where C_1 and C_2 are constants. Therefore, the optimal density $q^*_{\theta_j}(\boldsymbol{\theta}_j)$ has a close form which belongs to the same exponential family as $p(\boldsymbol{\theta}_j | \, \mathbf{y}, \boldsymbol{\theta}_{-j})$. We can see the relation between Gibbs sampling and mean-field VA for conjugate exponential families from (3.8) and (3.9): Gibbs sampling iterates by drawing $\boldsymbol{\theta}_j$ from $p(\boldsymbol{\theta}_j | \, \mathbf{y}, \boldsymbol{\theta}_{-j})$, while mean-field VA iterates by evaluating the expection of the natural parameters in $p(\boldsymbol{\theta}_j | \, \mathbf{y}, \boldsymbol{\theta}_{-j})$ for $j = 1, \cdots, J$.

3.2 VA Algorithm A

In this section, we develop a variational approximation algorithm for the Gaussian error deconvolution problem. This algorithm allows repeated measurements for the variable of interest. The deconvolution problem is formulated as

$$y_{ij} = x_i + u_{ij}, \ i = 1, \cdots, n, \ j = 1, \cdots, m_i \qquad (3.10)$$

where $\{y_{ij}\}_{j=1}^{m_i}$ are repeated measurements for x_i, $x_i \overset{\text{i.i.d.}}{\sim} f_x(x)$, $u_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2)$, σ_u^2 is known and $f_x(x)$ is unknown. The algorithm aims to estimate $f_x(x)$.

3.2.1 Model specification

For the deconvolution problem (6.1), we consider the truncated representation of model (2.10),

$$\begin{aligned} y_{ij}\,|\,x_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(x_i,\sigma_u^2),\ i=1,\cdots,n,\ j=1,\cdots,m_i \\ x_i\,|\,c_i,\,\boldsymbol{\phi} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_{c_i},\,\sigma_{c_i}^2\right) \\ c_i\,|\,\boldsymbol{\pi} &\overset{\text{i.i.d.}}{\sim} \text{Categorical}\left(\pi_1,\,\pi_2,\cdots,\pi_K\right) \\ \boldsymbol{\pi} &\sim \ \text{Dirichlet}\left(\alpha/K,\alpha/K,\cdots,\alpha/K\right) \\ \phi_c &\overset{\text{i.i.d.}}{\sim} \mathcal{NIG}\left(\mu_0,\,\lambda_0,\,\gamma_0,\,\beta_0\right) \end{aligned} \tag{3.11}$$

The collection of all latent variables is $\boldsymbol{\theta}=[\mathbf{x},\mathbf{c},\boldsymbol{\pi},\boldsymbol{\phi}]$. The partition $p(\mathbf{x},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\pi}|\mathbf{y})\approx q_x(\mathbf{x})q_c(\mathbf{c})q_\pi(\boldsymbol{\pi})q_\phi(\boldsymbol{\phi})$ gives closed form for the full conditionals by 3.4. The variational distributions have the following form: $q_x^*(\mathbf{x})=\prod_{i=1}^n\mathcal{N}(x_i;\mu_{q(x_i)},\sigma^2_{q(x_i)})$,
$q_c^*(\mathbf{c})=\prod_{i=1}^n\text{Categorical}(c_i;\omega_{i1},\omega_{i2},\cdots,\omega_{iK})$,
$q_\pi^*(\boldsymbol{\pi})=\text{Dirichlet}(\boldsymbol{\pi};\alpha_{q(\pi),1},\alpha_{q(\pi),2},\cdots,\alpha_{q(\pi),K})$,
$q_\phi^*(\boldsymbol{\phi})=\prod_{k=1}^K\mathcal{NIG}(\phi_k;\,\mu_{q(\phi_k)},\lambda_{q(\phi_k)},A_{q(\phi_k)},B_{q(\phi_k)})$. The variational parameters are updated by the following algorithm.

3.2.2 Estimation method

Algorithm 2 (VA algorithm A for Gaussian error and known error variance)**.** *Initialize:* $\mu_{q(x_i)},\mu_{q(\phi_k)}\in \text{R}$ *and* $\sigma^2_{q(x_i)},\lambda_{q(\phi_k)},A_{q(\phi_k)},B_{q(\phi_k)},\omega_{ik}>0$ *for* $k=1,\ldots,K,\ i=1,\cdots,n$ *such that* $\sum_{k=1}^K\omega_{ik}=1$ *for all* i*. Repeat the following steps until the in-*

crease in $log\underline{p}(\mathbf{y};q)$ *is negligible: For* $i=1,\cdots,n$*:*

$$\sigma^2_{q(x_i)} \leftarrow \left(\frac{m_i}{\sigma^2_u}+\sum_{k=1}^{K}\frac{A_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}\right)^{-1}$$

$$\mu_{q(x_i)} \leftarrow \sigma^2_{q(x_i)}\left(\frac{y_{i\cdot}}{\sigma^2_u}+\sum_{k=1}^{K}\frac{A_{q(\phi_k)}\mu_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}\right)$$

For $i=1,\ldots,n$ *and* $k=1,\ldots,K$*:*

$$\nu_{ik} \leftarrow -\frac{1}{2}logB_{q(\phi_k)}+\frac{1}{2}\Psi(A_{q(\phi_k^2)})-\frac{1}{2}\frac{A_{q(\phi_k)}}{B_{q(\phi_k)}}\left((\mu_{q(x_i)}-\mu_{q(\phi_k)})^2+\sigma^2_{q(x_i)}\right)-\frac{1}{2\lambda_{q(\phi_k)}}+\Psi(\alpha_{q(\pi),k})$$

$$\omega_{ik} \leftarrow \frac{\exp(\nu_{ik})}{\sum_{l=1}^{K}\exp(\nu_{il})}$$

For $k=1,\ldots,K$*:*

$$\omega_{\cdot k} \leftarrow \sum_{i=1}^{n}\omega_{ik}$$

$$\mu_{q(\phi_k)} \leftarrow \frac{\sum_{i=1}^{n}\mu_{q(x_i)}\omega_{ik}+\lambda_0\mu_0}{\omega_{\cdot k+\lambda_0}}$$

$$\lambda_{q(\phi_k)} \leftarrow \omega_{\cdot k}+\lambda_0$$

$$A_{q(\phi_k)} \leftarrow \frac{\omega_{\cdot k}}{2}+\gamma_0$$

$$B_{q(\phi_k)} \leftarrow \beta_0+\frac{1}{2}\sum_{i=1}^{n}\omega_{ik}\left(\mu^2_{q(x_i)}+\sigma^2_{q(x_i)}\right)+\frac{1}{2}\lambda_0\mu_0^2-\frac{1}{2}\left(\omega_{\cdot k}+\lambda_0\right)\mu^2_{q(\phi_k)}$$

$$\alpha_{q(\pi),k} \leftarrow \omega_{\cdot k}+\frac{\alpha}{K}$$

Unknown measurement error variance

If σ_u^2 is unknown and there are replicate measures $y_{i1},\cdots,y_{im_i}$ on subject i,

$$y_{ij}=x_i+u_{ij},\ x_i\overset{\text{i.i.d.}}{\sim}f_x,\ u_{ij}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma_u^2),\ i=1,\cdots,n,\ j=1,\cdots,m_i \qquad (3.12)$$

the model is identifiable and σ_u^2 can be estimated from the model. Let N denote the total number of observations, i.e. $N=\sum_{i=1}^{n}m_i$. We put inverse gamma priors on σ_u^2, $\sigma_u^2 \sim \mathrm{IG}(\gamma_{\sigma_\mathrm{u}^2}, \beta_{\sigma_\mathrm{u}^2})$. Conjugacy of prior for σ_u^2 leads to: $q^*_{\sigma_u^2}(\sigma_u^2) = \mathrm{IG}(\sigma_\mathrm{u}^2; \mathrm{A}_{\mathrm{q}(\sigma_\mathrm{u}^2)}, \mathrm{B}_{\mathrm{q}(\sigma_\mathrm{u}^2)})$. The step of updating $q^*_{\sigma_u^2}(\sigma_u^2)$ is

$$A_{q(\sigma_u^2)} \leftarrow \gamma_{\sigma_u^2} + \frac{N}{2}$$
$$B_{q(\sigma_u^2)} \leftarrow \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}^2 - 2\mu_{q(x_i)}y_{ij} + \mu^2_{q(x_i)} + \sigma^2_{q(x_i)}) + \beta_{\sigma_u^2}$$

The step of updating $(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$ is modified by replacing $1/\sigma_u^2$ by its expectation under $q_{\sigma_u^2}(\sigma_u^2)$, which is $\frac{A_{q(\sigma_u^2)}}{B_{q(\sigma_u^2)}}$,

$$\sigma^2_{q(x_i)} \leftarrow \left(\frac{m_i A_{q(\sigma_u^2)}}{B_{q(\sigma_u^2)}} + \sum_{k=1}^{K}\frac{A_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}\right)^{-1}$$
$$\mu_{q(x_i)} \leftarrow \sigma^2_{q(x_i)}\left(\frac{A_{q(\sigma_u^2)}y_{i\cdot}}{B_{q(\sigma_u^2)}} + \sum_{k=1}^{K}\frac{A_{q(\phi_k)}\mu_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}\right)$$

3.2.3 Density prediction

Next we derive the density estimator based on the variational distributions. The predictive distribution for x_{n+1} conditional on observations $\mathbf{y}$ is given by:

$$p(x_{n+1}|\mathbf{y}) = \int \sum_{k=1}^{K} \pi_k\, p(x_{n+1}|\phi_k) dP^*(\pi_k, \phi_k|\mathbf{y}) \tag{3.13}$$

Under the factorized variational approximation to the posterior, the predictive distribution can be written as a product of expectations:

$$p(x_{n+1}|\mathbf{y}) = \sum_{k=1}^{K} \mathrm{E}_{q^*(\pi)}[\pi_k]\, \mathrm{E}_{q^*(\phi_k)}[p(x_{n+1}|\phi_k)] \tag{3.14}$$

where $q^*(\cdot)$ denote the optimal densities. The explicit form of $\mathrm{E}_{q^*(\phi_k)}[p(x_{n+1}|\phi_k)]$ can be derived as follows:

$$p(\phi_k|x_{n+1}) = \frac{1}{\mathrm{E}_{q^*(\phi_k)}[p(x_{n+1}|\phi_k)]} p(x_{n+1}|\phi_k) q^*(\phi_k) \tag{3.15}$$

where

$$\begin{aligned}p(\phi_k|x_{n+1}) =& \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\tilde{\lambda}_{q(\phi_k)}}}{\sqrt{\sigma_k^2}}\exp\left\{-\frac{\tilde{\lambda}_{q(\phi_k)}(\mu_k-\tilde{\mu}_{q(\phi_k)})^2}{2\sigma_k^2}\right\}\\ &\times\frac{\tilde{B}_{q(\phi_k)}^{\tilde{A}_{q(\phi_k)}}}{\Gamma(\tilde{A}_{q(\phi_k)})}\sigma_k^{2(-\tilde{A}_{q(\phi_k)}-1)}\exp\left\{-\frac{\tilde{B}_{q(\phi_k)}}{\sigma_k^2}\right\}\end{aligned} \tag{3.16}$$

$$\tilde{\lambda}_{q(\phi_k)} = \lambda^*_{q(\phi_k)} + 1 \tag{3.17}$$

$$\tilde{\mu}_{q(\phi_k)} = \frac{\lambda^*_{q(\phi_k)}\mu^*_{q(\phi_k)} + x_{n+1}}{\lambda^*_{q(\phi_k)} + 1} \tag{3.18}$$

$$\tilde{A}_{q(\phi_k)} = A^*_{q(\phi_k)} + \frac{1}{2} \tag{3.19}$$

$$\tilde{B}_{q(\phi_k)} = B^*_{q(\phi_k)} + \frac{\lambda^*_{q(\phi_k)}(x_{n+1}-\mu^*_{q(\phi_k)})^2}{2(\lambda^*_{q(\phi_k)}+1)} \tag{3.20}$$

and

$$\begin{aligned}p(x_{n+1}|\phi_k)q^*(\phi_k) =& \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left\{-\frac{(x_{n+1}-\mu_k)^2}{2\sigma_k^2}\right\}\frac{1}{\sqrt{2\pi}}\frac{\sqrt{\lambda^*_{q(\phi_k)}}}{\sqrt{\sigma_k^2}}\\ &\times\exp\left\{-\frac{\lambda^*_{q(\phi_k)}(\mu_k-\mu^*_{q(\phi_k)})^2}{2\sigma_k^2}\right\}\frac{B^{*A^*_{q(\phi_k)}}_{q(\phi_k)}}{\Gamma(A^*_{q(\phi_k)})}\sigma_k^{2(-A^*_{q(\phi_k)}-1)}\exp\left\{-\frac{B^*_{q(\phi_k)}}{\sigma_k^2}\right\}\end{aligned} \tag{3.21}$$

Comparing the factor terms in (3.15) which do not include μ_k or σ_k^2 gives the expression for $\mathrm{E}_{q^*(\phi_k)}[p(x_{n+1}|\phi_k)]$:

$$\mathrm{E}_{q^*(\phi_k)}[p(x_{n+1}|\phi_k)] = \frac{1}{\sqrt{2\pi}}\left(\frac{\lambda^*_{q(\phi_k)}}{\tilde{\lambda}_{q(\phi_k)}}\right)^{\frac{1}{2}}\frac{\Gamma(\tilde{A}_{q(\phi_k)})}{\Gamma(A^*_{q(\phi_k)})}\frac{B^{*A^*_{q(\phi_k)}}_{q(\phi_k)}}{\tilde{B}^{\tilde{A}_{q(\phi_k)}}_{q(\phi_k)}} \tag{3.22}$$

Plugging (3.22) into (3.14) gives

$$\begin{aligned}p(x_{n+1}|\mathbf{y}) &= \sum_{k=1}^{K}\frac{\alpha^*_{q(\pi_k)}}{\alpha+n}\frac{1}{\sqrt{2\pi}}\left(\frac{\lambda^*_{q(\phi_k)}}{\tilde{\lambda}_{q(\phi_k)}}\right)^{\frac{1}{2}}\frac{\Gamma(\tilde{A}_{q(\phi_k)})}{\Gamma(A^*_{q(\phi_k)})}\frac{B^{*A^*_{q(\phi_k)}}_{q(\phi_k)}}{\tilde{B}^{\tilde{A}_{q(\phi_k)}}_{q(\phi_k)}}\\ &\sim \sum_{k=1}^{K}\frac{\alpha^*_{q(\pi_k)}}{\alpha+n}\mathcal{T}\left(\frac{\sqrt{A^*_{q(\phi_k)}}(x_{n+1}-\mu^*_{q(\phi_k)})}{\sqrt{B^*_{q(\phi_k)}}};2A^*_{q(\phi_k)}\right)\end{aligned} \tag{3.23}$$

where $\mathcal{T}(y;\nu)$ denotes that the random variable y is from the student's t-distribution with degrees of freedom ν. If n goes to infinity, (3.23) can be approximated by

$$p(x_{n+1}|\mathbf{y}) \sim \sum_{k=1}^{K} \frac{\alpha^*_{q(\pi_k)}}{\alpha+n} \mathcal{N}\left(x_{n+1}; \mu^*_{q(\phi_k)}, \frac{B^*_{q(\phi_k)}}{A^*_{q(\phi_k)}}\right) \tag{3.24}$$

3.3 VA Algorithm B

In this section, we establish another VA algorithm for Gaussian error deconvolution problem. This algorithm assumes that there is no repeated measurement for x_i. The problem is formulated as

$$y_i = x_i + u_i,\ i = 1, \cdots, n \tag{3.25}$$

where y_i is a measurement for x_i, $x_i \overset{\text{i.i.d.}}{\sim} f_x(x)$, $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2)$, σ_u^2 is known and $f_x(x)$ is to be estimated.

We integrate out x_i from model (3.11). Let t_k denote the percentage of the measurement error variance in the variance of observations from kth cluster, that is, $t_{\phi,k} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_{\phi,k}^2}$. The model can be re-parameterized as follows:

$$\begin{aligned}
y_i \,|\, c_i, \mathbf{t}_\phi, \boldsymbol{\mu}_\phi, \sigma_u^2 &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{\phi,c_i}, \sigma_u^2/t_{\phi,c_i}),\ i = 1, \cdots, n \\
c_i \,|\, \boldsymbol{\pi} &\overset{\text{i.i.d.}}{\sim} \text{Categorical}\,(\pi_1,\, \pi_2, \cdots, \pi_K) \\
\boldsymbol{\pi} &\sim \text{Dirichlet}\,(\alpha/K, \alpha/K, \cdots, \alpha/K) \\
\mu_{\phi,k}|t_{\phi,k}, \sigma_u^2, \mu_0, \lambda_0 &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_0, \sigma_u^2/(\lambda_0 t_{\phi,k})\right) \\
t_{\phi,k}|a_0, c_0 &\overset{\text{i.i.d.}}{\sim} \mathcal{TG}((0,1]; a_0, c_0)
\end{aligned} \tag{3.26}$$

where $\mathcal{TG}((l,u]; a, c)$ represent a truncated gamma distribution on $(l,u]$ with shape parameter a and rate parameter c. In model (3.26) the parameters $\sigma_u^2, \alpha, \mu_0, \lambda_0, a_0, c_0$ are known, K is fixed.

Let $\tilde{\phi}_k = (t_{\phi,k}, \mu_{\phi,k})$ and $\tilde{\boldsymbol{\phi}}$ be the collection of $\tilde{\phi}_k$, $k = 1, \cdots, K$. The posterior is partitioned by $p(\mathbf{c}, \tilde{\boldsymbol{\phi}}, \boldsymbol{\pi}|\mathbf{y}) \approx q_c(\mathbf{c}) q_{\tilde{\phi}}(\tilde{\boldsymbol{\phi}}) q_\pi(\boldsymbol{\pi})$. Since $p(\boldsymbol{\pi}|\alpha)$, $p(\mathbf{c}|\boldsymbol{\pi})$ and $p(\tilde{\boldsymbol{\phi}}|\mu_0, \lambda_0, a_0, c_0)$ are conjugate priors, the optimal densities $q^*_{\mathbf{c}}(\mathbf{c})$, $q^*_\pi(\boldsymbol{\pi})$ and $q^*_{\tilde{\phi}}(\tilde{\boldsymbol{\phi}})$ can be derived by $q^*_{\theta_j}(\boldsymbol{\theta}_j) \propto \exp\{\mathrm{E}_{-\theta_j} \log p(\mathbf{y}, \boldsymbol{\theta})\}$. For concision of the notations, we omit the known parameters $\{\sigma^2_u, \alpha, \mu_0, \lambda_0, a_0, c_0\}$ in the derivation. Assume X is a truncated Gamma random variable $X \sim \mathcal{TG}((0,1]; a, c),\ a > 0, c > 0$, the following results of X will be used in the estimation method:

$$E[X|a,c] = \int_0^1 \frac{c^a}{\Gamma(a) F_g(1; a, c)} t^a \exp(-ct)\, dt = \frac{a F_g(1; a+1, c)}{c F_g(1; a, c)} \tag{3.27}$$

$$\begin{aligned} E[\log X|a,c] &= \int_0^1 \frac{c^a}{\Gamma(a) F_g(1; a, c)} \log(t) t^{a-1} \exp(-ct)\, dt && (3.28)\\ &= \frac{c^a}{\Gamma(a) F_g(1; a, c)} \frac{\partial}{\partial a} \int_0^1 t^{a-1} \exp(-ct)\, dt \\ &= \frac{c^a}{\Gamma(a) F_g(1; a, c)} \lim_{\delta \to 0} \frac{1}{\delta} \left(\frac{\Gamma(a+\delta) F_g(1; a+\delta, c)}{c^{a+\delta}} - \frac{\Gamma(a) F_g(1; a, c)}{c^a} \right) \\ & && (3.29) \end{aligned}$$

where $F_g(x; a, c)$ represent the cumulative distribution function of gamma random variable $X \sim \mathrm{G}(a, c)$:

$$F_g(x; a, c) = \int_0^x \frac{c^a}{\Gamma(a)} t^{a-1} \exp(-ct)\, dt,\ x > 0 \tag{3.30}$$

3.3.1 Estimation method

Algorithm 3 (VA algorithm B for Gaussian error and known error variance)**.** *Initialize:* $\mu_{q(\phi_k)} \in \mathrm{R}$ *and* $\lambda_{q(\phi_k)}, A_{q(\phi_k)}, C_{q(\phi_k)}, \omega_{ik} > 0$ *for* $k = 1, \ldots, K,\ i = 1, \ldots, n$ *such that* $\sum_{k=1}^K \omega_{ik} = 1$ *for all* i*. Repeat the following steps until the increase in*

$log\underline{p}(\mathbf{y};q)$ is negligible: For $i = 1, \ldots, n$, $k = 1, \cdots, K$,

$$\begin{aligned}\nu_{ik} \leftarrow & \frac{1}{2}E[log(t_{\phi,k})|A_{q(\phi_k)}, C_{q(\phi_k)}] - \frac{1}{2\sigma_u^2}E[t_{\phi,k}|A_{q(\phi_k)}, C_{q(\phi_k)}](y_i - \mu_{q(\phi_k)})^2 \\ & - \frac{1}{2\lambda_{q(\phi_k)}} + \Psi(\alpha_{q(\pi),k}) \end{aligned} \tag{3.31}$$

$$\omega_{ik} \leftarrow \frac{\exp(\nu_{ik})}{\sum_{l=1}^{K}\exp(\nu_{il})} \tag{3.32}$$

For $k = 1, \ldots, K$:

$$\mu_{q(\phi_k)} \leftarrow \frac{\sum_{i=1}^{n} y_i\omega_{ik} + \lambda_0\mu_0}{\omega_{\cdot k+\lambda_0}} \tag{3.33}$$

$$\lambda_{q(\phi_k)} \leftarrow \omega_{\cdot k} + \lambda_0 \tag{3.34}$$

$$A_{q(\phi_k)} \leftarrow \frac{\omega_{\cdot k}}{2} + a_0 \tag{3.35}$$

$$C_{q(\phi_k)} \leftarrow c_0 + \frac{1}{2\sigma_u^2}\left(\sum_{i=1}^{n}\omega_{ik}y_i^2 + \lambda_0\mu_0^2 - (\omega_{\cdot k} + \lambda_0)\,\mu_{q(\phi_k)}^2\right) \tag{3.36}$$

$$\alpha_{q(\pi_k)} \leftarrow \omega_{\cdot k} + \frac{\alpha}{K} \tag{3.37}$$

3.3.2 Density prediction

Under the factorized variational approximation to the posterior, the predictive distribution can be written as

$$p(x_{n+1}|\mathbf{y}) = \sum_{k=1}^{K} \mathrm{E}_{q^*(\pi)}[\pi_k]\,\mathrm{E}_{q^*(\tilde{\phi}_k)}[p(x_{n+1}|\tilde{\phi}_k)] \tag{3.38}$$

The expectation terms are given by

$$\mathrm{E}_{q^*(\pi)}[\pi_k] = \frac{\alpha^*_{q(\pi_k)}}{\alpha + n} \tag{3.39}$$

$$\mathrm{E}_{q^*(\tilde{\phi}_k)}[p(x_{n+1}|\tilde{\phi}_k)] = \int_0^1 \int_{-\infty}^{\infty} p(x_{n+1}|\mu_{\phi,k}, t_{\phi,k}) q^*(\mu_{\phi,k}|t_{\phi,k}) q^*(t_{\phi,k}) d\mu_{\phi,k} dt_{\phi,k}$$

$$= \int_0^1 \int_{-\infty}^{\infty} \frac{\sqrt{t_{\phi,k}}}{\sqrt{2\pi\sigma_u^2(1-t_{\phi,k})}} \exp\left\{-\frac{t_{\phi,k}(x_{n+1}-\mu_{\phi,k})^2}{2\sigma_u^2(1-t_{\phi,k})}\right\}$$

$$\times \frac{\sqrt{t_{\phi,k}\lambda^*_{q(\phi_k)}}}{\sqrt{2\pi\sigma_u^2}} \exp\left\{-\frac{\lambda^*_{q(\phi_k)} t_{\phi,k}(\mu_{\phi,k}-\mu_0)^2}{2\sigma_u^2}\right\}$$

$$\times \frac{C^{*A^*_{q(\phi_k)}}_{q(\phi_k)} t_{\phi,k}^{A^*_{q(\phi_k)}-1}}{\Gamma(A^*_{q(\phi_k)}) F_g(1; A^*_{q(\phi_k)}, C^*_{q(\phi_k)})} \exp\left\{-C^*_{q(\phi_k)} t_{\phi,k}\right\} d\mu_{\phi,k} dt_{\phi,k}$$

$$= \int_0^1 \frac{\sqrt{\lambda^*_{q(\phi_k)} t_{\phi,k}}}{\sqrt{2\pi\sigma_u^2(\lambda^*_{q(\phi_k)} - \lambda^*_{q(\phi_k)} t_{\phi,k} + 1)}} \exp\left\{-\frac{\lambda^*_{q(\phi_k)} t_{\phi,k}(x_{n+1}-\mu_0)^2}{2\sigma_u^2(\lambda^*_{q(\phi_k)} - \lambda^*_{q(\phi_k)} t_{\phi,k} + 1)}\right\}$$

$$\times \frac{C^{*A^*_{q(\phi_k)}}_{q(\phi_k)} t_{\phi,k}^{A^*_{q(\phi_k)}-1}}{\Gamma(A^*_{q(\phi_k)}) F_g(1; A^*_{q(\phi_k)}, C^*_{q(\phi_k)})} \tag{3.40}$$

$$\exp\left\{-C^*_{q(\phi_k)} t_{\phi,k}\right\} dt_{\phi,k} \tag{3.41}$$

The integral (3.41) can be obtained by numerical methods.

3.4 Simulation 1: Comparison between VA algorithm A, VA algorithm B, MCMC and DK

This section compares VA algorithm A and B with DK method and MCMC method on simulated data. We consider a target density which is a two-components normal mixture with $\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 0, \mu_2 = 1.5, \sigma_1 = 1, \sigma_2 = 0.2$. The variance of Gaussian error is set to 0.25, which corresponds to reliability 81%. We generated 100 simulated dataset of size $n = 1000, m_i = 1$ for $i = 1, \cdots, n$ and applied the four methods on each dataset. Numerical integrated squared error

(ISE) of the density estimate from dth dataset was calculated by $ISE(\hat{f}_x^{(d)}) = \sum_{t=1}^{T}\{f_x(x_t^*) - \hat{f}_x^{(d)}(x_t^*)\}^2\Delta_t$, where $x_a = x_1^* \leq \cdots \leq x_T^* = x_b$ is an evenly spaced grid points and $\Delta_t = x_{t+1}^* - x_t^*$. The endpoints were chosen to be $x_a = -8, x_b = 8$, which cover the support of target density.

The DK method was implemented by R package "fDKDE". This package provides the a plug-in (PI) bandwidth selector by [Delaigle and Gijbels, 2004b] and a cross-validated (CV) bandwidth selector by [Stefanski and Carroll, 1990]. We used the PI bandwidth in all simulation experiments in this thesis since [Delaigle and Gijbels, 2004b] shows that it outperforms the CV bandwidth through simulation experiments. The VA and MCMC methods were also programmed in R. VA algorithms used $K = 10$ clusters for the truncated DPMM. We chose hyper-parameters $\gamma_0 = \beta_0 = \lambda_0 = 0.1, \mu_0 = \overline{y}.$ for MCMC method and VA algorithm A, $a_0 = c_0 = \lambda_0 = 0.1, \mu_0 = \overline{y}.$ for VA algorithm B. We set the concentration parameter $\alpha = 0.1$ for both VA and MCMC methods. For VA the loop continues to iterate until the increase in the lower bound of log marginal likelihood $\log \underline{p}(\mathbf{y}; q)$ is less than 10^{-4}. For MCMC each chain was run for 6000 iterations with the first 1000 discarded as burn-in.

Table 3.1: Simulation 1: ISE and speed comparison of DK, VA and MCMC method.

	Quartiles of 100×ISE			
	25%	50%	75%	Computing time for running one dataset
DK	9.39	10.59	11.82	40 s
VA algorithm A	18.64	19.49	20.21	2 s
VA algorithm B	0.93	2.88	4.75	5 s
MCMC	1.32	2.05	3.37	10 min

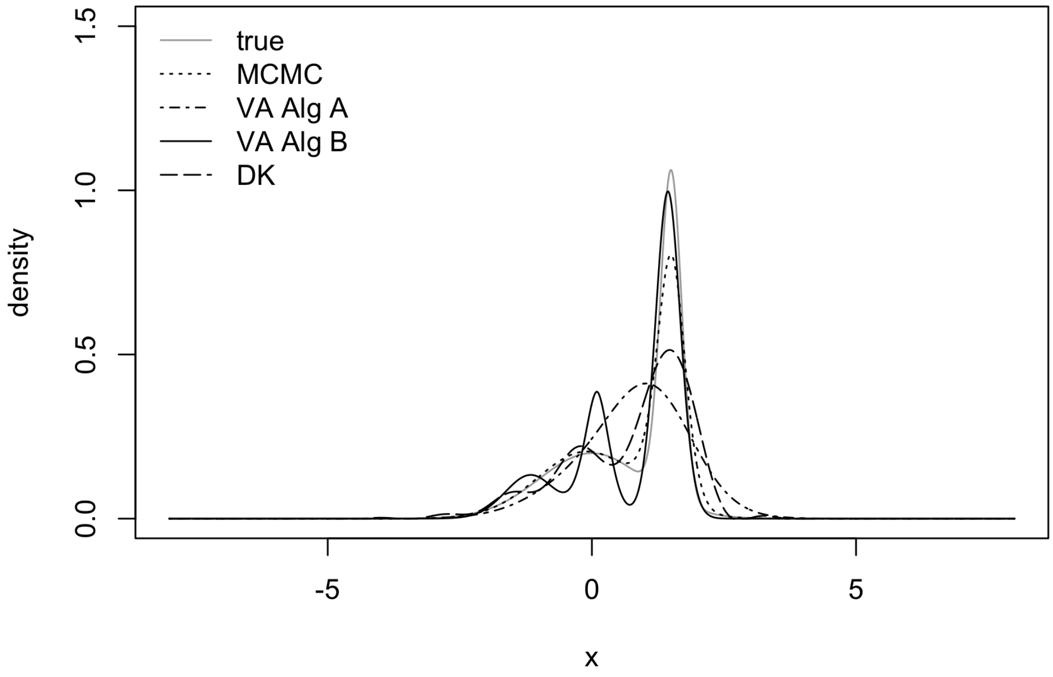

Figure 3.1: Simulation 1: the density estimate of the median ISE

The density estimate corresponding to the median ISE by MCMC, VA algorithm A, algorithm B and DK are presented in figure 3.1 by dotted, dot-dashed, solid and dashed line, respectively. VA algorithm B outperforms both VA algorithm A and DK method; the density estimate of median ISE by VA algorithm A cannot catch the two modes of the density of interest.

3.5 Exploration on Performance of Algorithm A and B

The simulation experiment in section 3.4 shows that VA algorithm B outperforms algorithm A in estimating the density of the two-mode mixture of normals.

In this section, we investigate the reasons behind it. This section is organized as follows. In subsection 3.5.1, we find different modes of the objective function of algorithm A by deterministic annealing. We see that a higher bound of marginal likelihood does not necessarily give a better density estimate for algorithm A. In subsection 3.5.2, we propose a conjecture to explain the drawbacks of algorithm A. Subsection 3.5.3 shows that VA algorithm B can also get stuck at a local optimum. The ability of jumping out of a local optimum needs to be improved for VA algorithm B. This section we reconsider the following deconvolution problem.

$$y_i = x_i + u_i,\ i = 1, \cdots, n \tag{3.42}$$

where y_i is a measurement for x_i, $x_i \overset{\text{i.i.d.}}{\sim} f_x(x)$, $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2)$, σ_u^2 is known and $f_x(x)$ is to be estimated.

3.5.1 Multiple modes of algorithm A

To find different modes of the log marginal likelihood lower bound, we apply deterministic annealing to VA algorithm A. Deterministic annealing (DA), which is variant of simulated annealing [Kirkpatrick et al., 1983], uses a time-dependent temperature parameter to deterministically deform the objective function. It can find a different local optima from optimizing a fixed objective function. Deterministic annealing was originally established for clustering in [Rose et al., 1990]. Later, [Ueda and Nakano, 1995] developed deterministic annealing variant of the EM algorithms for estimating maximum likelihood parameters; [Katahira et al., 2008] [Abrol et al., 2014] [Mandt et al., 2016] applied deterministic annealing to variational inference.

To apply deterministic annealing to VA algorithm A, we introduce temperature parameters $T > 0$ and define the tempered lower bound of marginal likelihood as

$$\log \underline{p}_T(\mathbf{y}|q) = \mathrm{E}_q[\log p(\mathbf{y}|\mathbf{x})] + \frac{1}{T}\mathrm{E}_q[\log p(\mathbf{x}|\mathbf{c},\boldsymbol{\phi})] + \mathrm{E}_q[\log p(\mathbf{c}|\boldsymbol{\pi})] + \mathrm{E}_q[\log p(\boldsymbol{\pi})]$$
$$+\mathrm{E}_q[\log p(\boldsymbol{\phi})] - \log C(T) - \mathrm{E}_q[\log q_x(\mathbf{x})] - \mathrm{E}_q[\log q_c(\mathbf{c})] - \mathrm{E}_q[\log q_\pi(\boldsymbol{\pi})] - \mathrm{E}_q[\log q_\phi(\boldsymbol{\phi})] \tag{3.43}$$

where

$$C(T) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\mathbf{c},\boldsymbol{\phi})^{\frac{1}{T}}p(\mathbf{c}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\phi})d\mathbf{y}d\mathbf{x}d\mathbf{c}d\boldsymbol{\pi}d\boldsymbol{\phi} \tag{3.44}$$

When $T = 1$, the sum of first six terms in (3.43) is the expected logarithm of the joint distribution of hidden variables and observed data. It favors q to put high probability on values of hidden variables that can best fit the observed data. The rest terms are entropies of the variational distributions. The entropies work like regularization and induce smoothness to q. To achieve a density that better explains the data, we initialize T at $T_0 = 0.7$ and increase T by 10^{-3} at each iteration until $T = 1$. Following is the deterministic annealing algorithm A.

Algorithm 4 (DAVA algorithm A)**.** *Initialize:* $\mu_{q(x_i)}, \mu_{q(\phi_k)} \in \mathrm{R}$ *and* $\sigma^2_{q(x_i)}, \lambda_{q(\phi_k)}, A_{q(\phi_k)}, B_{q(\phi_k)}, \omega_{ik} > 0$ *for* $k = 1, \ldots, K,\ i = 1, \ldots, n$ *such that* $\sum_{k=1}^K \omega_{ik} = 1$ *for all* i*. Choose* $T_0 < 1$*. Repeat the following steps until the increase in* $log\underline{p}(\mathbf{y}; q)$ *is negligible: For* $i = 1, \ldots, n$*:*

$$\sigma^2_{q(x_i)} \leftarrow \left(\frac{m_i}{\sigma^2_u} + \sum_{k=1}^K \frac{A_{q(\phi_k)}\omega_{ik}}{TB_{q(\phi_k)}}\right)^{-1}$$

$$\mu_{q(x_i)} \leftarrow \sigma^2_{q(x_i)}\left(\frac{y_{i\cdot}}{\sigma^2_u} + \sum_{k=1}^K \frac{A_{q(\phi_k)}\mu_{q(\phi_k)}\omega_{ik}}{TB_{q(\phi_k)}}\right)$$

For $i = 1, \ldots, n$ *and* $k = 1, \ldots, K$*:*

$$\nu_{ik} \leftarrow \frac{1}{T}\left(-\frac{1}{2}logB_{q(\phi_k)} + \frac{1}{2}\Psi(A_{q(\phi_k^2)}) - \frac{1}{2}\frac{A_{q(\phi_k)}}{B_{q(\phi_k)}}\left((\mu_{q(x_i)} - \mu_{q(\phi_k)})^2 + \sigma^2_{q(x_i)}\right) - \frac{1}{2\lambda_{q(\phi_k)}}\right)$$
$$+ \Psi(\alpha_{q(\pi),k})$$

$$\omega_{ik} \leftarrow \frac{\exp(\nu_{ik})}{\sum_{l=1}^K \exp(\nu_{il})}$$

```
For $k = 1, \dots, K$:
    $\omega_{\cdot k} \leftarrow \sum_{i=1}^{n} \omega_{ik}$
    $\mu_{q(\phi_k)} \leftarrow \frac{\sum_{i=1}^{n} \mu_{q(x_i)} w_{ik} + T\lambda_0\mu_0}{\omega_{\cdot k} + T\lambda_0}$
    $\lambda_{q(\phi_k)} \leftarrow \frac{\omega_{\cdot k}}{T} + \lambda_0$
    $A_{q(\phi_k)} \leftarrow \frac{\omega_{\cdot k}}{2T} + \gamma_0$
    $B_{q(\phi_k)} \leftarrow \beta_0 + \frac{1}{2T} \sum_{i=1}^{n} \omega_{ik} \left(\mu_{q(x_i)}^2 + \sigma_{q(x_i)}^2\right) + \frac{1}{2}\lambda_0\mu_0^2 - \frac{1}{2T}\left(\omega_{\cdot k} + \lambda_0\right)\mu_{q(\phi_k)}^2$
    $\alpha_{q(\pi),k} \leftarrow \omega_{\cdot k} + \frac{\alpha}{K}$
Increase $T$ if $T < 1$.
```

We generated a dataset from the model used in section 3.4, estimated variational parameters from the posteriors obtained by MCMC, and then chose the estimated variational parameters as initial values for algorithm A. If the global maximum of the lower bound of marginal likelihood can be obtained when the variational distributions are the true posterior distributions, iterations starting with those initial values will quickly converge to the global optima and give a similar density estimate as MCMC. The variational parameters were estimated as follows. We applied the MCMC method to the dataset and kept the samples drawn after the burn-in period. Let S be the number of samples drawn from Markov chain, the mean and variance posterior of x_i can be estimated by $\hat{\mu}_{x_i} = \frac{1}{S}\sum_{s=1}^{S} x_i^{(s)}$ and $\hat{\sigma}_{x_i}^2 = \frac{1}{S-1}\sum_{s=1}^{S}(x_i^{(s)} - \hat{\mu}_{x_i})^2$. At *s-th* iteration after burn-in, the two largest clusters were kept and ordered by the cluster mean. They were indexed by $(\mu_{nk}^{(s)}, \lambda_{nk}^{(s)}, \gamma_{nk}^{(s)}, \beta_{nk}^{(s)})$, $k = 1, 2$. We initialized the variational parameters in VA algorithm A and DAVA algorithm A by setting

$\mu_{q(x_i)} = \hat{\mu}_{x_i}$, $\sigma_{q(x_i)}^2 = \hat{\sigma}_{x_i}^2$, for $i = 1, \cdots, n$;

$\alpha_{q(\pi),k} = \frac{1}{S}\sum_{s=1}^{S} \frac{n_k^{(s)}}{n+\alpha}$ for $k = 1, 2$, $\alpha_{q(\pi),k} = 0.1$ for $k = 3, \cdots, K$;

$\mu_{q(\phi_k)} = \frac{1}{S}\sum_{s=1}^{S} \mu_{nk}^{(s)}$ for $k = 1, 2$, $\mu_{q(\phi_k)} = \bar{y}. = \frac{1}{n}\sum_{i=1}^{n} y_i$ for $k = 3, \cdots, K$;

$\lambda_{q(\phi_k)} = \frac{1}{S}\sum_{s=1}^{S} \lambda_{nk}^{(s)}$ for $k = 1, 2$, $\lambda_{q(\phi_k)} = 0.1$ for $k = 3, \cdots, K$;

$A_{q(\phi_k)} = \frac{1}{S}\sum_{s=1}^{S} \gamma_{nk}^{(s)}$ for $k = 1, 2$, $A_{q(\phi_k)} = 0.1$ for $k = 3, \cdots, K$;

$B_{q(\phi_k)} = \frac{1}{S}\sum_{s=1}^{S} \beta_{nk}^{(s)}$ for $k = 1, 2$. $B_{q(\phi_k)} = 0.1$ for $k = 3, \cdots, K$.

Figure 3.2 shows that the two algorithms found different local optimum. Although the initial values come from posterior means, VA algorithm A did not catch the two modes of the target density. Density estimate of DAVA algorithm A is closer to the target density than VA algorithm A, while the ELBO of DAVA algorithm A is lower than VA algorithm A.

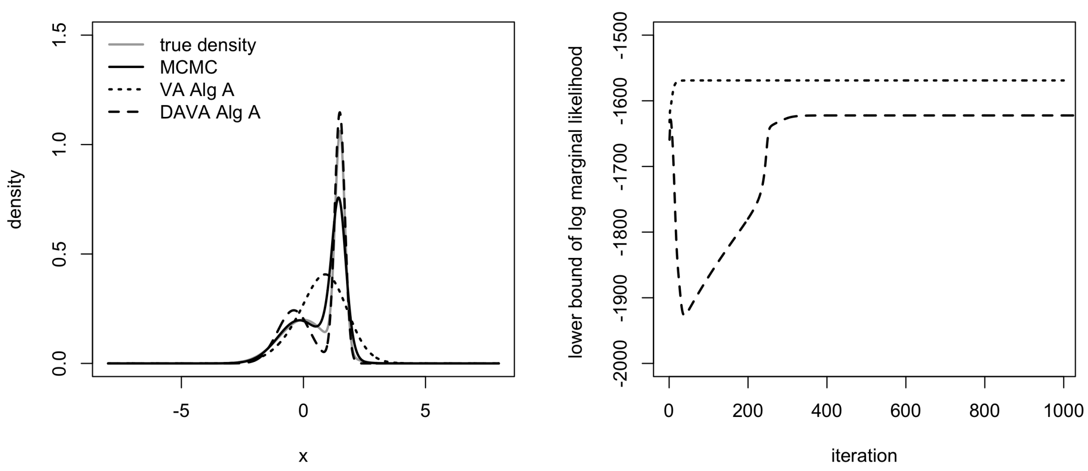

Figure 3.2: Multiple modes of VA algorithm A

3.5.2 A conjecture based on the simulation results

This section starts with comparing the Bayesian networks of the two algorithms and proposes a conjecture to explain why algorithm B outperforms algorithm A.

In model 3.25, the observed variable is $\mathbf{y}$; the hidden variables are $\mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi}$ for algorithm A, and $\mathbf{c}, \boldsymbol{\pi}, \tilde{\boldsymbol{\phi}}$ for algorithm B. We categorize the hidden variables into global hidden variables and local hidden variables, which are terminologies defined in [Hoffman et al., 2013]. Hidden variables whose sizes grow linearly with the size of observed variables are local hidden variables; hidden variables whose sizes do not depend on the size of observed variables are global hidden variables. In algorithm A, local hidden variables are $\{\mathbf{x}, \mathbf{c}\}$, global hidden variables are $\{\boldsymbol{\pi}, \boldsymbol{\phi}\}$. In algorithm B, local hidden variables are $\{\mathbf{c}\}$, global hidden variables are $\{\boldsymbol{\pi}, \tilde{\boldsymbol{\phi}}\}$. The dependency between variables are shown in Bayesian networks 3.3 and 3.4.

Algorithm A approximates the posterior by $p(\mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi}|\mathbf{y}) \approx q_x(\mathbf{x})q_c(\mathbf{c})q_\pi(\boldsymbol{\pi})q_\phi(\boldsymbol{\phi})$, which breaks the dependancy between local variables $\mathbf{x}$ and $\mathbf{c}$, that is, $p(x_i, c_i|y_i, \boldsymbol{\pi}, \boldsymbol{\phi}) \neq p(x_i|y_i, \boldsymbol{\pi}, \boldsymbol{\phi})p(c_i|y_i, \boldsymbol{\pi}, \boldsymbol{\phi})$. Algorithm B approximates the posterior by $p(\mathbf{c}, \boldsymbol{\pi}, \tilde{\boldsymbol{\phi}}|\mathbf{y}) \approx q_c(\mathbf{c})q_\pi(\boldsymbol{\pi})q_{\tilde{\phi}}(\tilde{\boldsymbol{\phi}})$ and does not lose dependency between local hidden variables. Based on the performances of algorithm A and B, we have a conjecture that breaking dependency between local hidden variables leads to loss of accuracy in variational inference.

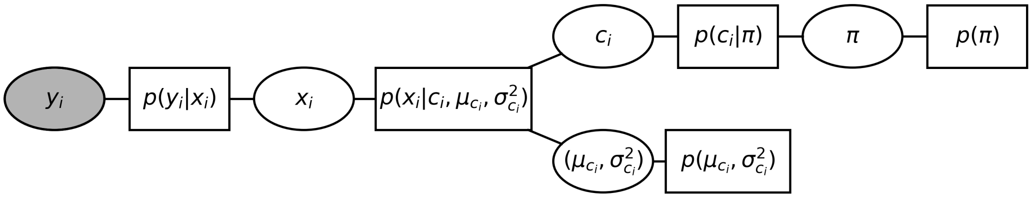
y_i
$p(y_i|x_i)$
x_i
$p(x_i|c_i, \mu_{c_i}, \sigma^2_{c_i})$
c_i
$p(c_i|\pi)$
π
$p(\pi)$
$(\mu_{c_i}, \sigma^2_{c_i})$
$p(\mu_{c_i}, \sigma^2_{c_i})$

Figure 3.3: Bayesian network of VA algorithm A

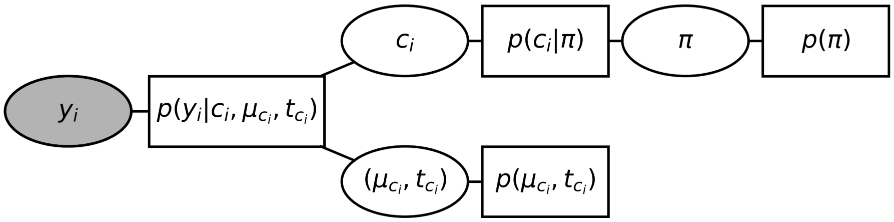

Figure 3.4: Bayesian network of VA algorithm B

3.5.3 Multiple modes of algorithm B

VA algorithm B may also get stuck at a local optimum. To find a second local optimum of algorithm B, we apply DA to VA algorithm B by defining the tempered lower bound of marginal likelihood as

$$\begin{aligned}\text{log}\underline{p}_T(\mathbf{y}|q) = \frac{1}{T}\Big\{&\text{E}_q[\text{log}p(\mathbf{y}|\mathbf{c}, \tilde{\boldsymbol{\phi}})] + \text{E}_q[\text{log}p(\mathbf{c}|\boldsymbol{\pi})] + \text{E}_q[\text{log}p(\boldsymbol{\pi})] + \text{E}_q[\text{log}p(\tilde{\boldsymbol{\phi}})]\Big\} \\ &- \text{log}C(T) - \text{E}_q[\text{log}q_c(\mathbf{c})] - \text{E}_q[\text{log}q_\pi(\boldsymbol{\pi})] - \text{E}_q[\text{log}q_{\tilde{\phi}}(\tilde{\boldsymbol{\phi}})]\end{aligned} \tag{3.45}$$

where

$$C(T) = \int p(\mathbf{y}|\mathbf{c}, \tilde{\boldsymbol{\phi}})^{\frac{1}{T}} p(\mathbf{c}|\boldsymbol{\pi})^{\frac{1}{T}} p(\boldsymbol{\pi})^{\frac{1}{T}} p(\tilde{\boldsymbol{\phi}})^{\frac{1}{T}} d\mathbf{y} d\mathbf{c} d\boldsymbol{\pi} d\tilde{\boldsymbol{\phi}} \tag{3.46}$$

The algorithm started with $T = 4$ and gradually reduced T to 1, it favored q to have high entropy and smoothness at start then allowed q to put higher probability on hidden variables that can better fit the observed data. Deterministic annealing VA algorithm B is given as follows.

Algorithm 5 (DAVA algorithm B). *Initialize:* $\mu_{q(\phi_k)} \in \text{R}$ *and* $\lambda_{q(\phi_k)}, A_{q(\phi_k)}, C_{q(\phi_k)}, \omega_{ik} > 0$ *for* $k = 1, \ldots, K,\ \ i = 1, \ldots, n$ *such that* $\sum_{k=1}^{K} \omega_{ik} = 1$

for all i. Set $T_0 > 1$. *Repeat the following steps until the increase in* $log\underline{p}(\mathbf{y};q)$ *is negligible: For* $i = 1,\ldots,n,\ k = 1,\cdots,K$,

$$\nu_{ik} \leftarrow \frac{1}{T}\left\{\frac{1}{2}E[log(t_{\phi,k})|A_{q(\phi_k)},C_{q(\phi_k)}] - \frac{1}{2\sigma_u^2}E[t_{\phi,k}|A_{q(\phi_k)},C_{q(\phi_k)}](y_i - \mu_{q(\phi_k)})^2 \right.$$
$$\left. -\frac{1}{2\lambda_{q(\phi_k)}} + \Psi(\alpha_{q(\pi),k})\right\} \tag{3.47}$$

$$\omega_{ik} \leftarrow \frac{\exp(\nu_{ik})}{\sum_{l=1}^{K}\exp(\nu_{il})} \tag{3.48}$$

For $k = 1,\ldots,K$*:*

$$\mu_{q(\phi_k)} \leftarrow \frac{\sum_{i=1}^{n} y_i\omega_{ik} + \lambda_0\mu_0}{\omega_{\cdot k+\lambda_0}} \tag{3.49}$$

$$\lambda_{q(\phi_k)} \leftarrow \frac{1}{T}\left(\omega_{\cdot k} + \lambda_0\right) \tag{3.50}$$

$$A_{q(\phi_k)} \leftarrow \frac{\omega_{\cdot k}}{2T} + \frac{a_0}{T} \tag{3.51}$$

$$C_{q(\phi_k)} \leftarrow \frac{c_0}{T} + \frac{1}{2T\sigma_u^2}\left(\sum_{i=1}^{n}\omega_{ik}y_i^2 + \lambda_0\mu_0^2 - \left(\omega_{\cdot k} + \lambda_0\right)\mu_{q(\phi_k)}^2\right) \tag{3.52}$$

$$\alpha_{q(\pi_k)} \leftarrow \frac{\omega_{\cdot k}}{T} + \frac{\alpha}{TK} \tag{3.53}$$

Decrease T *if* $T > 1$.

We chose a dataset from the 100 simulated datasets of section 3.4 and applied both VA algorithm B and DAVA algorithm B using same initial values for variational parameters. The modes of marginal likelihood lower bound found by DAVA and VA are almost equally high, while their density estimates are apparently different as shown in 3.5.

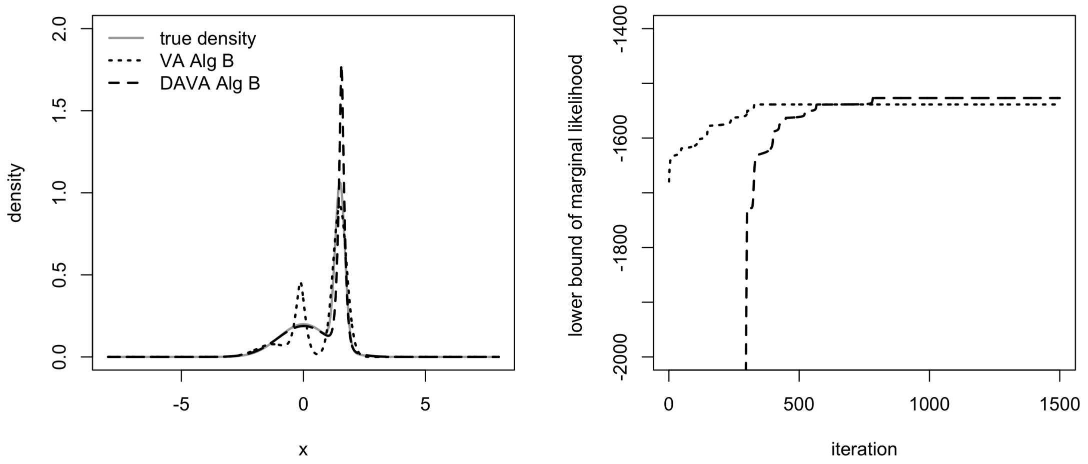

Figure 3.5: Multiple modes of VA algorithm B

3.5.4 Limitations of VA algorithm B

VA algorithm B considers the deconvolution problem with balanced replications and Gaussian measurement error. It has not been able to handle the following situations.

1. Unbalanced replications

   If the model has balanced replications, that is $m_i = m$ for all i, then algorithm B can be implemented by considering $\bar{y}_{i\cdot} \mid c_i, \mathbf{t}_\phi, \boldsymbol{\mu}_\phi, \sigma_u^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{\phi,c_i}, \tilde{\sigma}_u^2/t_{\phi,c_i})$, $\tilde{\sigma}_u^2 = \frac{\sigma_u^2}{m}, t_{\phi,k} = \frac{\tilde{\sigma}_u^2}{\tilde{\sigma}_u^2+\sigma_{\phi,k}^2}$. If the model has unbalanced repeated measurements, the fraction $\frac{\sigma_u^2/m_i}{\sigma_u^2/m_i+\sigma_{\phi,k}^2}$ does not only depend on k but it also depends on i. The parameterization of algorithm B does not have a closed form for the variational distribution $q_{\tilde{\phi}}(\tilde{\boldsymbol{\phi}})$. If $(\mu_{\phi,k}, \sigma_{\phi,k}^2)$ is restricted to be a normal-inverse gamma distribution, variational parameters in $q_\phi(\mu_{\phi,k}, \sigma_{\phi,k}^2)$ need to be updated by derivative-based optimization methods at lower speed.

2. Non-Gaussian distributed errors

   The latent variable $\mathbf{x}$ may not be integrated out if the measurement error is non-Gaussian distributed.

Since integrating out $\mathbf{x}$ improves the performance of VA, it is desired to seek for approaches to extend algorithm B to the above situations.

3.6 Stochastic Variational Approximation

In this section, we apply stochastic optimization to improve performance of algorithm B. Stochastic optimization [Robbins and Monro, 1951] allows randomness in the optimization process and may enable the algorithm to escape a local optimum. Stochastic optimization has been used with Expectation-Maximization (EM) algorithm in [Cappé and Moulines, 2009] [Nielsen et al., 2000] [Diebolt and Ip, 1996]. An EM algorithm can be expressed as a mean-field [Neal and Hinton, 1998]. Stochastic optimization is also applied to traditional variational approximation by [Hoffman et al., 2013] [Kiciman et al., 2008]. Rather than iterating between re-analyzing each data in the whole dataset and re-estimating its hidden structure, stochastic variational approximation (SVA) iterates between randomly sampling a subset of the whole dataset and estimating the hidden structure based only on the subset. In this section, SVA for algorithm B is established based on the idea of [Hoffman et al., 2013].

Algorithm B includes observations $y_{i=1\cdots,n}$, global hidden variables $\boldsymbol{\beta} = \{\mathbf{t}, \boldsymbol{\mu}, \boldsymbol{\pi}\}$ and local hidden variables $c_{i=1\cdots,n}$. In model (3.26), both $\boldsymbol{\beta}$ and $\mathbf{c}$ have conjugate priors which belong to exponential family. Their full conditionals can be expressed in the following form,

$$\begin{aligned} p(c_i|\mathbf{y}, \mathbf{c}_{-i}, \boldsymbol{\beta}, \sigma_u^2) &= h_c(c_i)\text{exp}\left\{\boldsymbol{\eta}_c(y_i, \boldsymbol{\beta}, \sigma_u^2)^T \mathbf{T}_c(c_i) - \xi_c(\boldsymbol{\eta}_c(y_i, \boldsymbol{\beta}, \sigma_u^2))\right\} \\ p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{c}, \sigma_u^2) &= h_\beta(\boldsymbol{\beta})\text{exp}\left\{\boldsymbol{\eta}_\beta(\mathbf{y}, \mathbf{c}, \sigma_u^2)^T \mathbf{T}_\beta(\boldsymbol{\beta}) - \xi_\beta(\boldsymbol{\eta}_\beta(\mathbf{y}, \mathbf{c}, \sigma_u^2))\right\} \end{aligned} \tag{3.54}$$

where $\boldsymbol{\eta}_c(y_i, \boldsymbol{\beta}, \sigma_u^2)$ or $\boldsymbol{\eta}_\beta(\mathbf{y}, \mathbf{c}, \sigma_u^2)$ is called natural parameter, $\mathbf{T}_c(c_i)$ or $\mathbf{T}_\beta(\boldsymbol{\beta})$ is sufficient statistic.

Under the restriction $q(\boldsymbol{\beta}, \mathbf{c}) \approx q_\beta(\boldsymbol{\beta})q_c(\mathbf{c})$ and by (3.9), the optimal variational distribution q_c and q_β belong to the same exponential family as $p(c_i|\mathbf{y}, \mathbf{c}_{-i}, \boldsymbol{\beta}, \sigma_u^2)$ and $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{c}, \sigma_u^2)$, respectively.

$$\begin{aligned} &q_c(c_i) = h_c(c_i)\text{exp}\left\{\boldsymbol{\rho}_i^T \mathbf{T}_c(c_i) - \xi_c(\boldsymbol{\rho}_i)\right\} \\ &q_\beta(\boldsymbol{\beta}) = h_\beta(\boldsymbol{\beta})\text{exp}\left\{\boldsymbol{\zeta}^T \mathbf{T}_\beta(\boldsymbol{\beta}) - \xi_\beta(\boldsymbol{\zeta})\right\} \end{aligned} \tag{3.55}$$

The coordinate decent algorithm iterates between updating $\boldsymbol{\zeta}$ and $\boldsymbol{\rho}_i$ for $i = 1, \cdots, n$. It updates the variational parameters to the expectation of the natural parameters with respect to the current variational distributions. In the following algorithm local parameters refer to the variational parameters in q_c, global parameter refer to the variational parameter in q_β.

Algorithm 6 (Coordinate decent algorithm for variational approximation)**.**

1. *Initialize* $\boldsymbol{\zeta}$.
2. *Repeat:*
3. *Update* $\boldsymbol{\rho}_i \leftarrow \text{E}_\zeta\left\{\boldsymbol{\eta}_c(y_i, \boldsymbol{\beta}, \sigma_u^2)\right\}$ *for* $i = 1, \cdots, n$.
4. *Update* $\boldsymbol{\zeta} \leftarrow \text{E}_\rho\left\{\boldsymbol{\eta}_\beta(\mathbf{y}, \mathbf{c}, \sigma_u^2)\right\}$
5. *Until the increase in log*$\underline{p}(\mathbf{y}; q)$ *is negligible.*

Step 3 evaluates local parameters for each local hidden variable, then step 4 updates global parameter based on all local hidden structures. The algorithm 6 can find a local optima as it converges. The stochastic variation approximation (SVA) allows randomness in the process of optimization. Instead of evaluating the hidden structure for all local hidden variables, SVA randomly samples a batch of data from the whole dataset and only evaluates the hidden structures for the subsample. Rather than update the global parameter using all local parameters, SVA estimates intermediate global parameter based on the hidden structures of the subsample and then updates the global parameter to a weighted sum of the former estimate and the intermediate estimate.

Algorithm 7 (Stochastic variational approximation algorithm)**.**

1. Initialize $\boldsymbol{\zeta}^{(0)}$.

2. For $t = 1, 2, \cdots, \infty$, *set step-size* δ_t *appropriately, repeat:*

3. Sample $\{y_{i_1}, y_{i_2}, \cdots, y_{i_s}\}$ *randomly from the whole dataset, update* $\boldsymbol{\rho}_{i_l}^{(t)} \leftarrow \mathrm{E}_{\zeta^{(t-1)}}\{\boldsymbol{\eta}_c(y_{i_l}, \boldsymbol{\beta}, \sigma_u^2)\}$ *for* $l = 1, \cdots, s$.

4. Evaluate $\tilde{\boldsymbol{\zeta}} \leftarrow \mathrm{E}_{\rho^{(t)}}\left\{\boldsymbol{\eta}_\beta(\mathbf{y}_{i_l, l=1,\cdots,s}, \mathbf{c}, \sigma_u^2)\right\}$.

5. Update $\boldsymbol{\zeta}^{(t)} \leftarrow (1-\delta_t)\boldsymbol{\zeta}^{(t-1)} + \delta_t\tilde{\boldsymbol{\zeta}}$.

Applying the SVA algorithm to algorithm B, we have the following algorithm.

Algorithm 8 (SVA for Gaussian error deconvolution)**.**

Initialize: $\mu_{q(\phi_k)}^{(0)} \in \mathrm{R}$ *and* $\lambda_{q(\phi_k)}^{(0)}, A_{q(\phi_k)}^{(0)}, C_{q(\phi_k)}^{(0)}, \omega_{ik}^{(0)} > 0$ *for* $k = 1, \ldots, K$, $i = 1, \ldots, n$ *such that* $\sum_{k=1}^{K} \omega_{ik}^{(0)} = 1$ *for all* i. *Let* $D_k^{(0)} = \lambda_{q(\phi_k)}^{(0)} \mu_{q(\phi_k)}^{(0)}$. *Set step-size* δ_t *appropriately, repeat the following steps for* $t = 1, 2, \cdots, T$*: Sample* $\{y_{i_1}, y_{i_2}, \cdots, y_{i_s}\}$

randomly from the whole dataset. For $l = 1, 2, \cdots, s$*,*

$$\begin{aligned}\nu_{i_l k} \leftarrow & \frac{1}{2} E[log(t^{(t)}_{\phi,k} | A^{(t)}_{q(\phi_k)}, C^{(t)}_{q(\phi_k)}] - \frac{1}{2\sigma_u^2} E[t_{\phi,k} | A^{(t)}_{q(\phi_k)}, C^{(t)}_{q(\phi_k)}](y_{i_l} - \mu^{(t)}_{q(\phi_k)})^2 \\ & - \frac{1}{2\lambda^{(t)}_{q(\phi_k)}} + \Psi(\alpha^{(t)}_{q(\pi),k}) \end{aligned} \tag{3.56}$$

$$\omega^{(t)}_{i_l k} \leftarrow \frac{\exp(\nu_{i_l k})}{\sum_{g=1}^{K} \exp(\nu_{i_l g})} \tag{3.57}$$

For $k = 1, \ldots, K$*:*

$$\tilde{D}_k \leftarrow \sum_{l=1}^{s} y_{i_l} \omega^{(t)}_{i_l k} + \lambda_0 \mu_0 \tag{3.58}$$

$$D^{(t)}_k \leftarrow (1 - \delta_t) D^{(t-1)}_k + \delta_t \tilde{D}_k \tag{3.59}$$

$$\tilde{\lambda}_{q(\phi_k)} \leftarrow \omega^{(t)}_{\cdot k} + \lambda_0 \tag{3.60}$$

$$\lambda^{(t)}_{q(\phi_k)} \leftarrow (1 - \delta_t) \lambda^{(t-1)}_{q(\phi_k)} + \delta_t \tilde{\lambda}_{q(\phi_k)} \tag{3.61}$$

$$\mu^{(t)}_{q(\phi_k)} \leftarrow \frac{D_k}{\lambda^{(t)}_{q(\phi_k)}} \tag{3.62}$$

$$\tilde{A}_{q(\phi_k)} \leftarrow \frac{\omega^{(t)}_{\cdot k}}{2} + a_0 \tag{3.63}$$

$$A^{(t)}_{q(\phi_k)} \leftarrow (1 - \delta_t) A^{(t-1)}_{q(\phi_k)} + \delta_t \tilde{A}_{q(\phi_k)} \tag{3.64}$$

$$\tilde{C}_{q(\phi_k)} \leftarrow c_0 + \frac{1}{2\sigma_u^2} \left(\sum_{l=1}^{s} \omega^{(t)}_{i_l k} y^2_{i_l} + \lambda_0 \mu_0^2 - \left(\omega^{(t)}_{\cdot k} + \lambda_0 \right) \mu^{2(t)}_{q(\phi_k)} \right) \tag{3.65}$$

$$C^{(t)}_{q(\phi_k)} \leftarrow (1 - \delta_t) C^{(t-1)}_{q(\phi_k)} + \delta_t \tilde{C}_{q(\phi_k)} \tag{3.66}$$

$$\tilde{\alpha}_{q(\pi_k)} \leftarrow \omega^{(t)}_{\cdot k} + \frac{\alpha}{K} \tag{3.67}$$

$$\alpha^{(t)}_{q(\pi_k)} \leftarrow (1 - \delta_t) \alpha^{(t-1)}_{q(\pi_k)} + \delta_t \tilde{\alpha}_{q(\pi_k)} \tag{3.68}$$

A variation of algorithm 8 can solve the Gaussian measurement error deconvolution with repeated measurements.

Algorithm 9 (SVA for Gaussian error deconvolution with replicated measurements)**.**

Initialize: $\mu^{(0)}_{q(\phi_k)} \in \mathrm{R}$ *and* $\lambda^{(0)}_{q(\phi_k)}, A^{(0)}_{q(\phi_k)}, C^{(0)}_{q(\phi_k)}, \omega^{(0)}_{ik} > 0$ *for* $k = 1, \ldots, K$, $i = 1, \ldots, n$ *such that* $\sum_{k=1}^{K} \omega^{(0)}_{ik} = 1$ *for all* i*. Let* $D^{(0)}_k = \lambda^{(0)}_{q(\phi_k)} \mu^{(0)}_{q(\phi_k)}$*. Let* $m_{min} = \min_{i=1,\cdots,n}\{m_i\}$*. Set step-size* δ_t *appropriately, repeat the following steps for* $t = 1, 2, \cdots, T$*: Sample* $y_{i,t_1}, y_{i,t_2}, \cdots, y_{i,t_{m_{min}}}$ *randomly from* $y_{i,1}, \cdots, y_{i,m_i}$ *for* $i = 1, 2, \cdots, n$ *and let* $\overline{y}^*_{i\cdot} = \sum_{j=1}^{m_{min}} y_{i,t_j}/m_{min}$*,*

$$\begin{aligned}\nu_{ik} \leftarrow & \frac{1}{2} E[log(t^{(t)}_{\phi,k} | A^{(t)}_{q(\phi_k)}, C^{(t)}_{q(\phi_k)}] - \frac{m_{min}}{2\sigma_u^2} E[t_{\phi,k} | A^{(t)}_{q(\phi_k)}, C^{(t)}_{q(\phi_k)}](\overline{y}^*_{i\cdot} - \mu^{(t)}_{q(\phi_k)})^2 \\ & - \frac{1}{2\lambda^{(t)}_{q(\phi_k)}} + \Psi(\alpha^{(t)}_{q(\pi),k})\end{aligned} \tag{3.69}$$

$$\omega^{(t)}_{ik} \leftarrow \frac{\exp(\nu_{ik})}{\sum_{l=1}^{K} \exp(\nu_{il})} \tag{3.70}$$

For $k = 1, \ldots, K$*:*

$$\tilde{D}_k \leftarrow \sum_{i=1}^{n} \overline{y}_{i\cdot} \omega^{(t)}_{ik} + \lambda_0 \mu_0 \tag{3.71}$$

$$D^{(t)}_k \leftarrow (1 - \delta_t) D^{(t-1)}_k + \delta_t \tilde{D}_k \tag{3.72}$$

$$\tilde{\lambda}_{q(\phi_k)} \leftarrow \omega^{(t)}_{\cdot k} + \lambda_0 \tag{3.73}$$

$$\lambda^{(t)}_{q(\phi_k)} \leftarrow (1 - \delta_t) \lambda^{(t-1)}_{q(\phi_k)} + \delta_t \tilde{\lambda}_{q(\phi_k)} \tag{3.74}$$

$$\mu^{(t)}_{q(\phi_k)} \leftarrow \frac{D_k}{\lambda^{(t)}_{q(\phi_k)}} \tag{3.75}$$

$$\tilde{A}_{q(\phi_k)} \leftarrow \frac{\omega^{(t)}_{\cdot k}}{2} + a_0 \tag{3.76}$$

$$A^{(t)}_{q(\phi_k)} \leftarrow (1 - \delta_t) A^{(t-1)}_{q(\phi_k)} + \delta_t \tilde{A}_{q(\phi_k)} \tag{3.77}$$

$$\tilde{C}_{q(\phi_k)} \leftarrow c_0 + \frac{1}{2\sigma_u^2} \left(\sum_{i=1}^{n} \omega^{(t)}_{ik} \overline{y}^{*2}_{i\cdot} + \lambda_0 \mu_0^2 - \left(\omega^{(t)}_{\cdot k} + \lambda_0 \right) \mu^{2(t)}_{q(\phi_k)} \right) \tag{3.78}$$

$$C^{(t)}_{q(\phi_k)} \leftarrow (1 - \delta_t) C^{(t-1)}_{q(\phi_k)} + \delta_t \tilde{C}_{q(\phi_k)} \tag{3.79}$$

$$\tilde{\alpha}_{q(\pi_k)} \leftarrow \omega^{(t)}_{\cdot k} + \frac{\alpha}{K} \tag{3.80}$$

$$\alpha^{(t)}_{q(\pi_k)} \leftarrow (1 - \delta_t) \alpha^{(t-1)}_{q(\pi_k)} + \delta_t \tilde{\alpha}_{q(\pi_k)} \tag{3.81}$$

In algorithm 9, the variance of measurement error σ_u^2 is assumed to be known.

If σ_u^2 is unknown, it can be estimated by $\hat{\sigma}_u^2 = \frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2$, where $\overline{y}_{i\cdot} = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$, then algorithm 9 can be applied.

3.7 Simulation 2: Comparison between SVA, VA Algorithm B, MCMC and DK

3.7.1 Non-repeated measurements

The SVA algorithm 9 was applied to the 100 datasets simulated in section 3.4. The step-size function was chosen as $\delta_t = t^{-0.7}$ and 2000 iterations were run for each dataset. We compare the results of SVA with VA algorithm B, MCMC and DK. Table 3.2 shows median ISE of SVA is about half of the median ISE of VA algorithm B, the overall performance of SVA is better than VA algorithm B.

Table 3.2: Simulation 2 - non-repeated measurements: ISE and speed comparison of SVA, DK, VA and MCMC method.

	Quartiles of 100×ISE			
	25%	50%	75%	Computing time for running one dataset
DK	9.39	10.59	11.82	40 s
VA algorithm B	0.93	2.88	4.75	5 s
SVA	0.75	1.46	2.36	10 s
MCMC	1.32	2.05	3.37	10 min

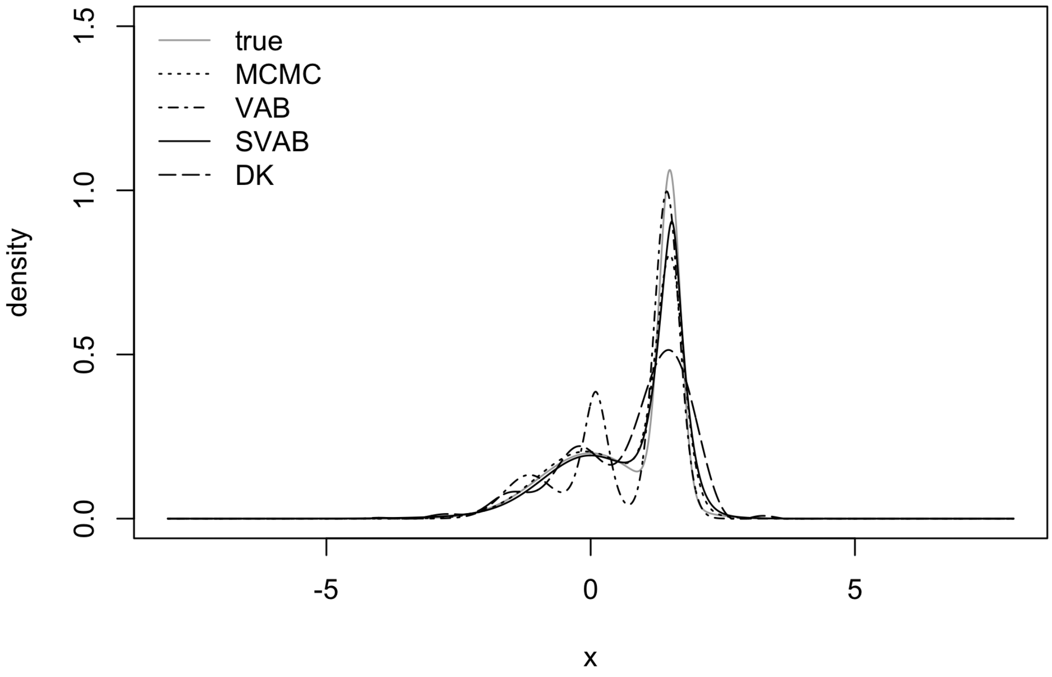

Figure 3.6: Simulation 2 - non-repeated measurements: the density estimate of the median ISE

3.7.2 Repeated measurements

We choose the target density $f(x) = 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(1.5, 0.2^2)$ and error variance $\sigma_u^2 = 0.25$, which are the same as the last experiment. The sample size is $n = 240$, the number of subjects is 60 for each number of repeated measurements $m_i \in \{1, 2, 3, 4\}$. The DK method for repeated measurements was implemented by R package "fDKDEheterosc". For SVA approach, the step-function was chosen as $\delta_t = t^{-\frac{1}{2}}$ and 3000 iterations were run for each dataset. Table 3.3 compares the ISEs of density estimates by SVA, DK and MCMC method and their computing speed. Figure 3.7 shows the density estimates corresponding to the median ISEs.

SVA achieves more accuracy in terms of ISE than DK and has comparable speed with DK. Density estimate by MCMC is a little more precise than SVA in this simulation experiment.

Table 3.3: Simulation 2 - repeated measurements: ISE of density estimates and computing speed.

	Quartiles of 100×ISE			
	25%	50%	75%	Computing time for running one dataset
DK	4.99	6.33	7.41	20 s
SVA	2.08	3.74	5.74	20 s
MCMC	1.60	2.46	3.49	5 min

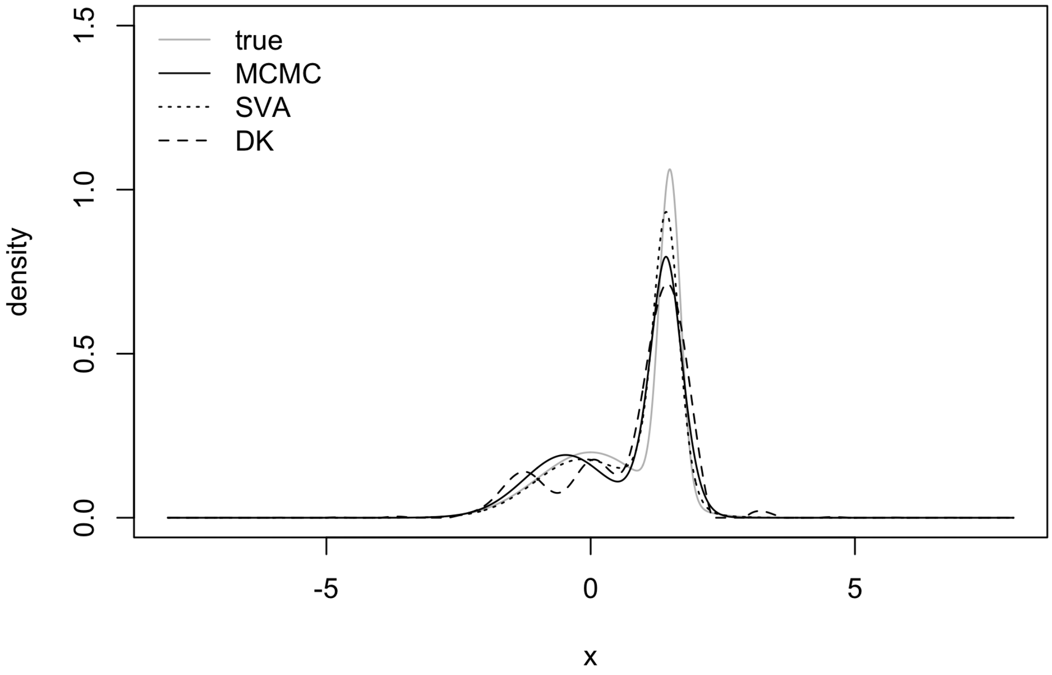

Figure 3.7: Simulation 2 - repeated measurements: the density estimate of the median ISE

3.8 Simulation 3: Add MCMC and SVA to [Wand, 1998]

This section presents the simulation results for four target densities which were used in [Wand, 1998] to represent important density shapes in practice. These target densities are: (i)the standard normal density $\mathcal{N}(0,1)$, (ii)a two-components normal mixture $\frac{2}{3}\mathcal{N}(0,1)+\frac{1}{3}\mathcal{N}(0,\frac{1}{5^2})$, (iii)a gamma density $\mathcal{G}(4,1)$, (iv)a two-component gamma mixture density $\frac{2}{5}\mathcal{G}(5,1)+\frac{3}{5}\mathcal{G}(13,1)$. For each target density, we choose different error variances by varying the percentage of variance of the observed data which is due to measurement error $p = \text{var}(u)/(\text{var}(x)+\text{var}(u))$. Let p vary from $p = 0$ to $p = 50\%$, corresponding to reliability from 100% to 50%. For each model, we generated 100 simulated datasets of size 250 observations and applied the naive kernel density estimation, DK, SVA and MCMC method on each dataset. Numerical ISE of the density estimate from dth dataset is calculated by $ISE(\hat{f}_x^{(d)}) = \sum_{t=1}^{T}\{f_x(x_t^*) - \hat{f}_x^{(d)}(x_t^*)\}^2\Delta_t$, where $x_a = x_1^* \leq \cdots \leq x_T^* = x_b$ is an evenly spaced grid points and $\Delta_t = x_{t+1}^* - x_t^*$. We choose $x_a = (0.0005 \text{ quantile of } f_x - 3.29\sigma_x)$ and $x_b = (0.9995 \text{ quantile of } f_x - 3.29\sigma_x)$ as the endpoints so that $[x_a, x_b]$ can cover an observation with probability $\geq 99.8\%$ for $p = 0.1, 0.2, 0.3, 0.4, 0.5$. The naive kernel density estimation was performed using binned kernel density estimate function in R package "KernSmooth". We used default options including standard normal kernel and the 'oversmoothed bandwidth selector' of [Wand and Jones, 1994]. The DK method was implemented by R package "fDKDE" and the plug-in (PI) bandwidth of [Delaigle and Gijbels, 2004b] was used. SVA approach uses $K = 10$ clusters for the truncated DPMM. We set the concentration parameter $\alpha = 0.1$ for both SVA and MCMC method. We choose the hyper-prior parameters $a_0 = c_0 = 0.1 = \lambda_0 = 0.1, \mu_0 = \overline{y}_.$ for SVA approach, $\gamma_0 = \beta_0 = \lambda_0 = 0.1, \mu_0 = \overline{y}_.$ for MCMC method. The step-size function was chosen as $\delta_t = t^{-0.7}$ for SVA and 2000 iterations were run for each dataset. For MCMC each chain was run for 6000 iterations with the first 1000 discarded as burn-in.

Figure 3.9 shows the 95% confidence interval of MISE of the density estimates. The distribution of the logarithm of ISE is approximately normal, so we calculate the 95% confidence limits for the MISE by $\exp\left\{\overline{\log(ISE(\hat{f}_x))} \pm t_{99,0.025}\hat{\sigma}_{\log(ISE(\hat{f}_x))}/\sqrt{100}\right\}$, where $\overline{\log(ISE(\hat{f}_x))}$ and $\hat{\sigma}_{\overline{\log(ISE(\hat{f}_x))}}$ are the sample mean and standard deviation of $\log(ISE(\hat{f}_x))$ respectively, $t_{99,0.025}$ denotes the upper tail 2.5% critical point of t distribution with 99 degrees of freedom. We can see from figure 3.9 that SVA and MCMC outperform DK for target density (1)(2)(3) and and have similar performance with DK for target density (4). SVA and MCMC have comparable accuracy of density estimate. For a dataset of sample size $n = 250$, it takes about 1 sec, 20 sec, 3 sec and 5 mins for Naive, DK, SVA and MCMC to perform density estimation, respectively.

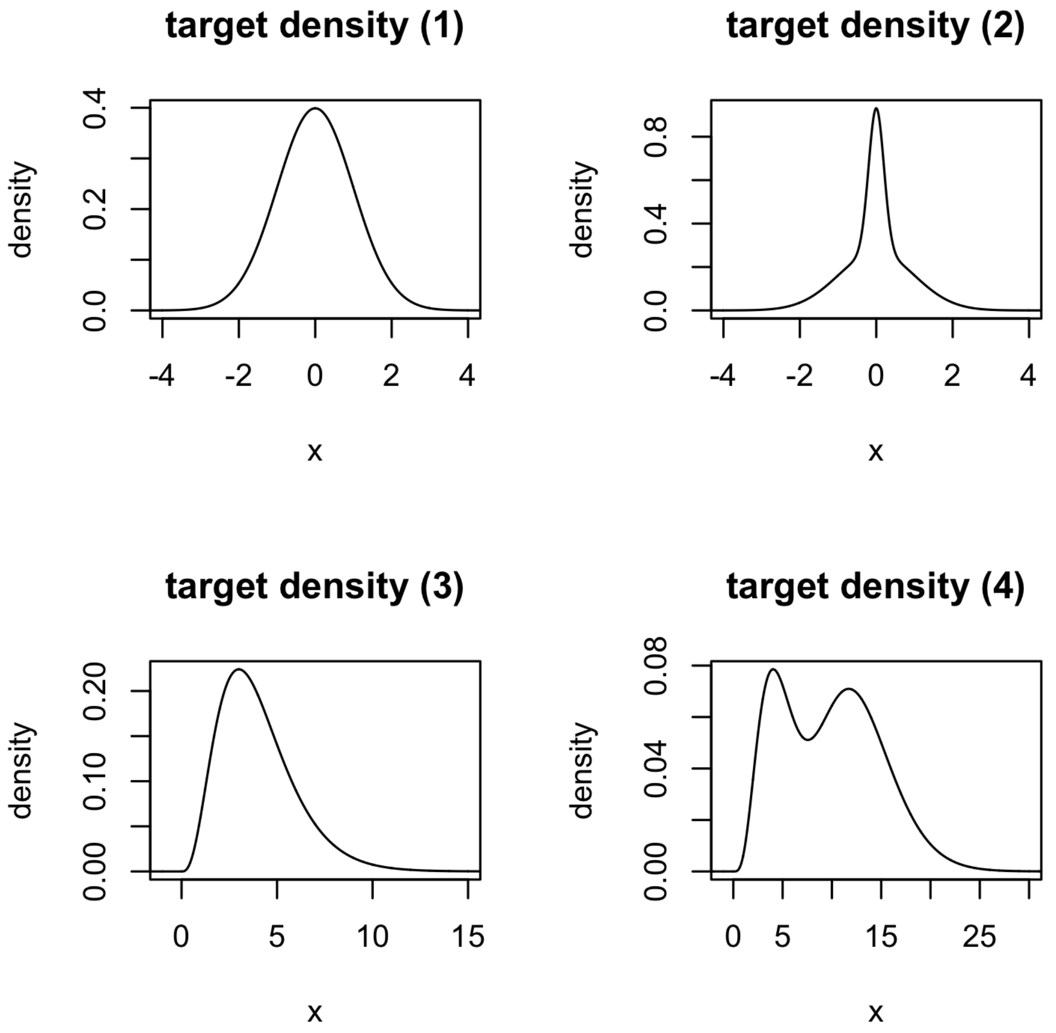

Figure 3.8: Simulation 3: target densities

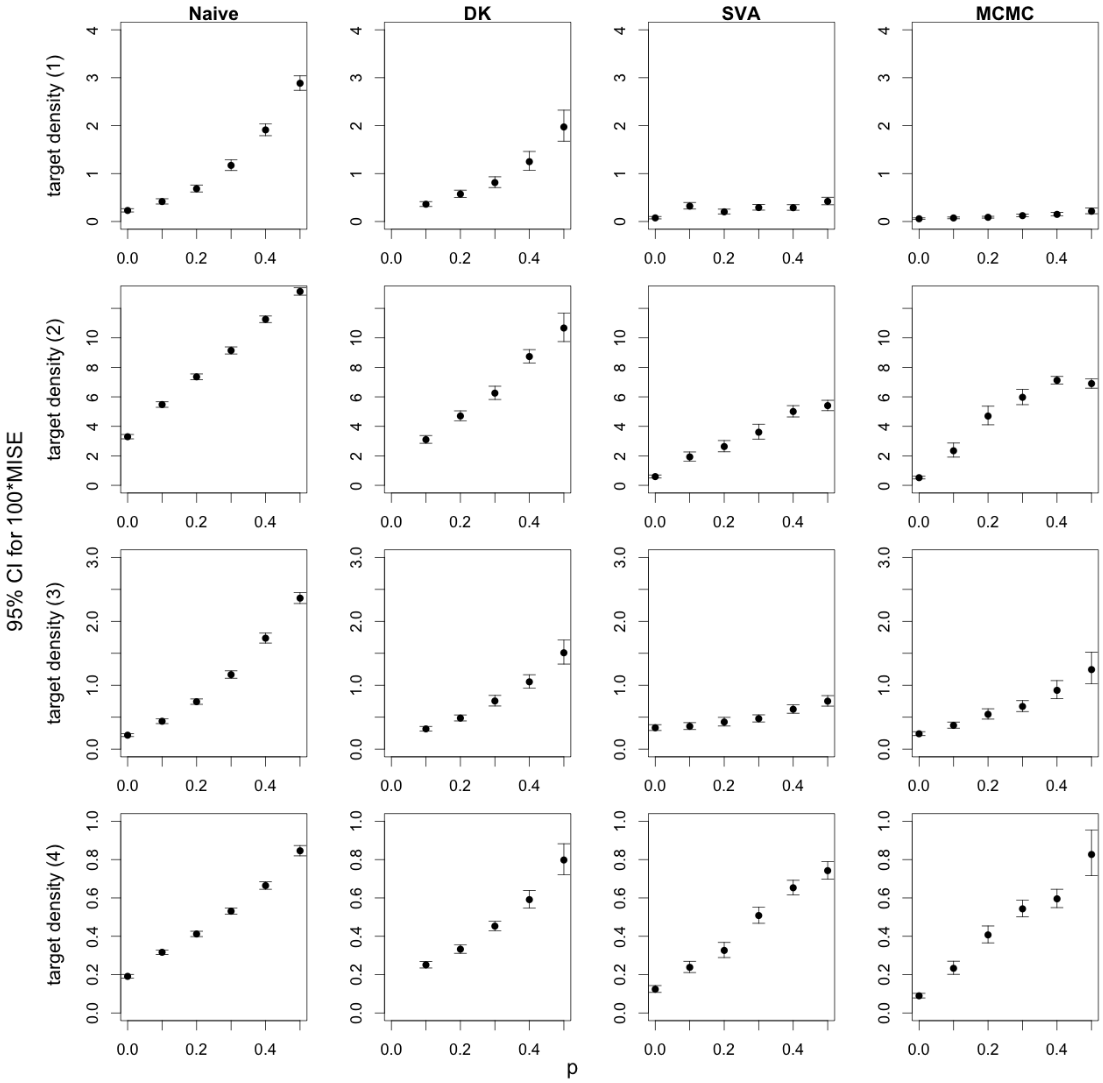

Figure 3.9: Simulation 3: 95 % CI of 100×MISE

CHAPTER 4

VARIATIONAL APPROXIMATION APPROACHES FOR LAPLACIAN MEASUREMENT ERROR

In this chapter, section 4.1 develops a variational approximation algorithm for the Laplacian error deconvolution problem. This algorithm allows repeated measurements for the variable of interest. The deconvolution problem is formulated as

$$y_{ij} = x_i + u_{ij},\ i = 1, \cdots, n,\ j = 1, \cdots, m_i \tag{4.1}$$

where $\{y_{ij}\}_{j=1}^{m_i}$ are repeated measurements for x_i, $x_i \overset{\text{i.i.d.}}{\sim} f_x(x)$, $u_{ij} \overset{\text{i.i.d.}}{\sim} \text{Laplacian}(0, b)$, b is known and $f_x(x)$ is unknown. The algorithm aims to estimate $f_x(x)$. Section 4.2 compares performances of DK, VA and MCMC method through simulation experiments.

4.1 VA Algorithm

4.1.1 Model specification

We model the density of interest $f_x(x)$ by the truncated DPMM,

$$
\begin{aligned}
y_{ij}\,|\,x_i &\overset{\text{i.i.d.}}{\sim} \text{Laplacian}(x_i, \sigma_u/\sqrt{2}), i = 1, \cdots, n,\; j = 1, \cdots, m_i \\
x_i\,|\,c_i,\, \boldsymbol{\phi} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_{c_i},\, \sigma^2_{c_i}\right) \\
c_i\,|\,\boldsymbol{\pi} &\overset{\text{i.i.d.}}{\sim} \text{Multinomial}\left(\pi_1,\, \pi_2, \cdots, \pi_K\right) \\
\boldsymbol{\pi} &\sim \text{Dirichlet}\left(\alpha/K, \alpha/K, \cdots, \alpha/K\right) \\
\phi_c &\overset{\text{i.i.d.}}{\sim} \mathcal{NIG}\left(\mu_0,\, \lambda_0,\, \gamma_0,\, \beta_0\right)
\end{aligned}
\tag{4.2}
$$

We approximate the posterior by $p(\mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi}|\mathbf{y}) \approx q_x(\mathbf{x})q_c(\mathbf{c})q_\pi(\boldsymbol{\pi})q_\phi(\boldsymbol{\phi})$.The prior for $\mathbf{x}$ is not conjugate and therefore optimal $q_x^*(\mathbf{x})$ does not have a close form. We restrict $q(x_i)$ to be a normal distribution $\mathcal{N}(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$ and update its parameters by Newton-Raphson method.

4.1.2 Estimation method

Algorithm 10 (VA approach for Laplacian error and known error variance)**.** *Initialize:* $\mu_{q(x_i)}, \mu_{q(\phi_k)} \in \mathrm{R}$ *and* $\sigma^2_{q(x_i)}, \lambda_{q(\phi_k)}, A_{q(\phi_k)}, B_{q(\phi_k)}, \omega_{ik} > 0$ *for* $k = 1, \ldots, K,\ i = 1, \ldots, n$ *such that* $\sum_{k=1}^K \omega_{ik} = 1$ *for all* i*. Repeat the following steps until the increase in* $log\underline{p}(\mathbf{y}; q)$ *is negligible: For* $i = 1, \ldots, n$ *: find the maximum point* $(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$ *of the following function with other parameters of* $q(\boldsymbol{\theta})$ *fixed:*

$$
\begin{aligned}
log\underline{p}(\mathbf{y}; q) &= \mathrm{E}\{logf(\mathbf{y}|\mathbf{x})\} + \mathrm{E}\{logf(\mathbf{x}|\mathbf{c}, \boldsymbol{\phi})\} + \text{Entropy}\{q(\mathbf{x})\} + C \\
&= -\frac{1}{b}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \sigma_{q(x_i)}\left(z_{ij}\left(2\Phi(z_{ij}) - 1\right) + 2\phi(z_{ij})\right) + \frac{1}{2}\sum_{i=1}^{n} log\sigma^2_{q(x_i)} \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{\omega_{ik}A_{q(\phi_k)}}{B_{q(\phi_k)}}\left(\sigma^2_{q(x_i)} + \mu^2_{q(x_i)} - 2\mu_{q(x_i)}\mu_{q(\phi_k)}\right) + C'
\end{aligned}
\tag{4.3}
$$

where C, C' are functions not containing $\mu_{q(x_i)}$ or $\sigma^2_{q(x_i)}$, $z_{ij} = \frac{y_{ij}-\mu_{q(x_i)}}{\sigma_{q(x_i)}}$, $\Phi(\cdot)$ and ϕ denote the PCF and PDF of the standard normal distribution respectively. Replacing $\sigma^2_{q(x_i)}$ by $\exp(l_{q(x_i)})$ in (4.3) guarantees its differentiability on R^2, where $l_{q(x_i)} = log(\sigma^2_{q(x_i)})$ The partial derivatives of (4.3) with respect to $(\mu_{q(x_i)}, l^2_{q(x_i)})$ are given by

$$\begin{aligned}\frac{\partial log\, \underline{p}(\mathbf{y};q)}{\partial \mu_{q(x_i)}} &= -\frac{1}{b}\sum_{j=1}^{m_i}\left(1-2\Phi(z_{ij})-2z_{ij}\phi(z_{ij})-2\phi'(z_{ij})\right) \\ &\quad -\sum_{k=1}^{K}\frac{A_{q(\phi_k^2)}}{B_{q(\phi_k^2)}}\omega_{ik}(\mu_{q(x_i)}-\mu_{q(\phi_k)}) \qquad (4.4)\end{aligned}$$

$$\begin{aligned}\frac{\partial log\, \underline{p}(\mathbf{y};q)}{\partial l_{q(x_i)}} &= \left(-\frac{1}{b}\sum_{j=1}^{m_i}\left(\phi(z_{ij})-z_{ij}^2\phi(z_{ij})-2z_{ij}\phi'(z_{ij})\right)-\frac{1}{2}\sum_{k=1}^{K}\frac{A_{q(\phi_k)}}{B_{q(\phi_k)}}\omega_{ik}\right) \\ &\quad \times \exp(l_{q(x_i)})+\frac{1}{2} \qquad (4.5)\end{aligned}$$

For $i = 1, \ldots, n$ and $k = 1, \ldots, K$:

$$\begin{aligned}\nu_{ik} \leftarrow &-\frac{1}{2}logB_{q(\phi_k)}+\frac{1}{2}\Psi(A_{q(\phi_k^2)})-\frac{1}{2}\frac{A_{q(\phi_k)}}{B_{q(\phi_k)}}\left((\mu_{q(x_i)}-\mu_{q(\phi_k)})^2+\sigma^2_{q(x_i)}\right) \\ &-\frac{1}{2\lambda_{q(\phi_k)}}+\Psi(\alpha_{q(\pi),k})\end{aligned}$$

$$\omega_{ik} \leftarrow \frac{\exp(\nu_{ik})}{\sum_{l=1}^{K}\exp(\nu_{il})}$$

For $k = 1, \ldots, K$:

$$\omega_{\cdot k} \leftarrow \sum_{i=1}^{n}\omega_{ik}$$

$$\mu_{q(\phi_k)} \leftarrow \frac{\sum_{i=1}^{n}\mu_{q(x_i)}\omega_{ik}+\lambda_0\mu_0}{\omega_{\cdot k+\lambda_0}}$$

$$\lambda_{q(\phi_k)} \leftarrow \omega_{\cdot k}+\lambda_0$$

$$A_{q(\phi_k)} \leftarrow \frac{\omega_{\cdot k}}{2}+\gamma_0$$

$$B_{q(\phi_k)} \leftarrow \beta_0+\frac{1}{2}\sum_{i=1}^{n}\omega_{ik}\left(\mu^2_{q(x_i)}+\sigma^2_{q(x_i)}\right)+\frac{1}{2}\lambda_0\mu_0^2-\frac{1}{2}\left(\omega_{\cdot k}+\lambda_0\right)\mu^2_{q(\phi_k)}$$

$$\alpha_{q(\pi),k} \leftarrow \omega_{\cdot k}+\frac{\alpha}{K}$$

Unknown measurement error variance

If σ_u^2 is unknown and there are replicate measures $y_{i1}, \cdots, y_{im_i}$ on subject i, the parameter σ_u^2 can be estimated from the model

$$y_{ij} = x_i + u_{ij},\ x_i \overset{\text{i.i.d.}}{\sim} f_x,\ u_{ij} \overset{\text{i.i.d.}}{\sim} \text{Laplacian}(0, b),\ \sigma_u^2 = 2b^2 \tag{4.6}$$

We put conjugate inverse gamma priors on b, $b \sim \text{IG}(\gamma_\text{b}, \beta_\text{b})$ and approximate $p(\mathbf{x}, \mathbf{c}, \boldsymbol{\phi}, \boldsymbol{\pi}, b|\mathbf{y})$ by $q_x(\mathbf{x})q_c(\mathbf{c})q_\pi(\boldsymbol{\pi})q_\phi(\boldsymbol{\phi})q_b(b)$. Since $p(b)$ is a conjugate prior, the variational distribution $q_b(b)$ can be updated by (3.4) $q_b^*(b) = \text{IG}(\text{b}; \text{A}^*_{\text{q(b)}}, \text{B}^*_{\text{q(b)}})$, where

$$\begin{aligned} A^*_{q(b)} &\leftarrow \gamma_b + N \\ B^*_{q(b)} &\leftarrow \sum_{i=1}^{n}\sum_{j=1}^{m_i} \sigma_{q(x_i)} \left(z_{ij}\left(2\Phi(z_{ij}) - 1\right) + 2\phi(z_{ij})\right) + \beta_b \end{aligned}$$

The step of updating $(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$ is changed by replacing $1/b$ by its expectation under $q_b(b)$, which is $\frac{A_{q(b)}}{B_{q(b)}}$. We update $(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$ to $(\mu^*_{q(x_i)}, \sigma^{2*}_{q(x_i)})$ which maximizes $\log \underline{p}(\mathbf{y}; q)$ with other parameters of $q(\boldsymbol{\theta})$ fixed:

$$\begin{aligned} \log \underline{p}(\mathbf{y}; q) = &-\frac{A_{q(b)}}{B_{q(b)}} \sum_{i=1}^{n}\sum_{j=1}^{m_i} \sigma_{q(x_i)} \left(z_{ij}\left(2\Phi(z_{ij}) - 1\right) + 2\phi(z_{ij})\right) + \frac{1}{2}\sum_{i=1}^{n} \log\sigma^2_{q(x_i)} \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{\omega_{ik} A_{q(\phi_k)}}{B_{q(\phi_k)}} \left(\sigma^2_{q(x_i)} + \mu^2_{q(x_i)} - 2\mu_{q(x_i)}\mu_{q(\phi_k)}\right) + C' \end{aligned} \tag{4.7}$$

where C' is a function not containing $(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$ and $z_{ij} = \frac{y_{ij} - \mu_{q(x_i)}}{\sigma_{q(x_i)}}$.

4.2 Simulation 4: Compare VA to MCMC and DK for Laplacian measurement error

This section compares the performance of VA approach with DK method and MCMC method on simulated data. We present results for two target densities which are shown in figure 4.1: (i) $0.5\mathcal{N}(0,1)+0.5\mathcal{N}(1.5,0.2^2)$ (ii)$\frac{2}{3}\mathcal{N}(0,1)+\frac{1}{3}\mathcal{N}(0,0.2^2)$. Density (i) is left skewed and has two modes; density (ii) is symmetric and with a kurtosis coefficient about 2.23 times that of the normal. For each density, we choose different measurement error variance levels, $p=0,0.1,0.2,0.3,0.4,0.5$, where $p=\text{var}(u)/(\text{var}(x)+\text{var}(u))$.

We generated 100 simulated dataset of size $n=250, m_i=1$ for $i=1,\cdots,n$ and applied four approaches on each dataset, naive kernel density estimation, DK method with PI bandwidth, VA algorithm 10 and MCMC. We chose concentration parameter $\alpha=0.1$ and hyper-prior parameters $\gamma_0=\beta_0=\lambda_0=0.1, \mu_0=\overline{y}_.$ for both VA and MCMC method. The number of clusters for truncated DPMM of VA was set at $K=10$. Iterations of VA stop if the increase in $\log\underline{p}(\mathbf{y};q)$ is less than 10^{-4}. For MCMC each chain was run for 6000 iterations with the first 1000 discarded as burn-in. Figure 4.2 shows the ISEs of density estimates by the four approaches. For a dataset of sample size $n=250$, the average computing speed of Naive, DK, VA and MCMC method are 1 sec, 20 sec, 20 sec and 5 mins, respectively.

MCMC achieves comparable accuracy with DK but has lower speed; VA performs worse than MCMC. VA algorithm 10 is developed based on the factorization $p(\mathbf{x},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\pi}|\mathbf{y})\approx q_x(\mathbf{x})q_c(\mathbf{c})q_\pi(\boldsymbol{\pi})q_\phi(\boldsymbol{\phi})$, where the dependence of local hidden variables $\mathbf{x}$ and $\mathbf{c}$ is broken. It can explain why VA algorithm 10 performs worse than MCMC by the conjecture proposed in section 3.5.2. Since the $p(y_i|x_i,b)$ is a Laplacian density and $p(x_i|c_i,\boldsymbol{\phi})$ is a Gaussian density, integrating out x_i in model 4.2 does not give a closed form for $p(\mathbf{y}|\boldsymbol{\phi},\mathbf{c})$ or $p(\boldsymbol{\phi}|\mathbf{y},\mathbf{c})$, therefore integrating out $\mathbf{x}$ for Laplacian error cannot achieve same computation efficiency as VA algorithm B for Gaussian error. Given that DK can achieve comparable accuracy with MCMC, we will not further explore VA algorithms for Laplacian error deconvolution.

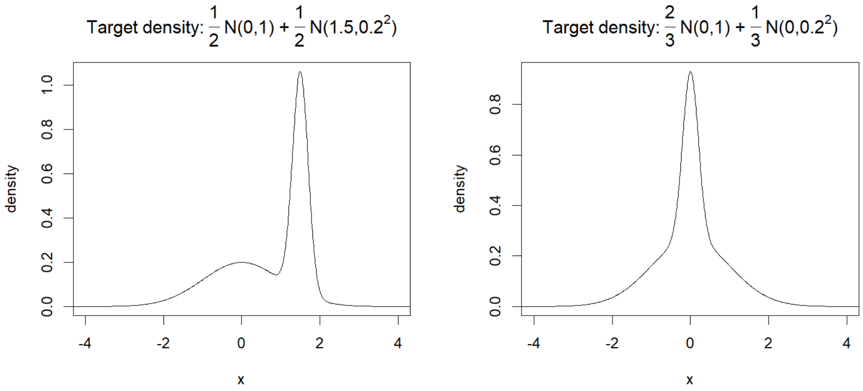

Figure 4.1: Simulation 4: target densities

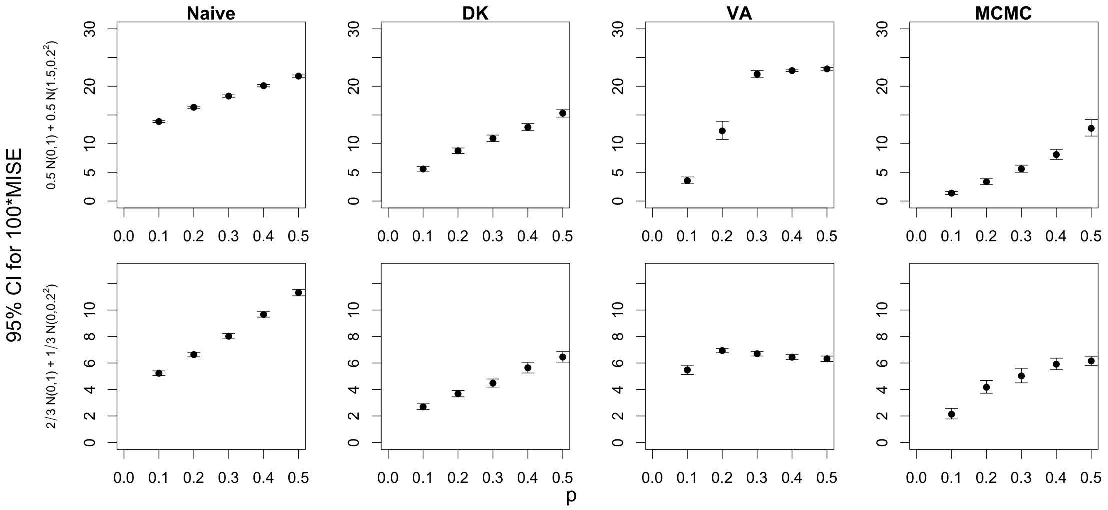

Figure 4.2: Simulation 4: 95% CI of 100×MISE

CHAPTER 5

ANALYZING PHYSICAL ACTIVITY DATA

In Chapter 5, we apply deconvolution methods to analyze physical activity data. Section 5.1 describes the dataset and develops a deconvolution model for the activPAL sedentary behavior data. Section 5.2 solves the deconvolution model and checks the fitness of the model. Specifically, subsection 5.2.1 and 5.2.2 apply SVA and MCMC method to estimate the density of the average daily sedentary time, respectively. Subsection 5.2.3 assesses the fitness of models by posterior predictive methods. Subsection 5.2.4 discusses possibility of heteroscedasticity in measurement error.

5.1 Data Description and Model Specification

Physical inactivity or excessive time spent on sedentary behavior has been identified as a risk factor for mortality and many adverse health conditions [Matthews et al., 2012] [Thorp et al., 2011]. Physical activity researchers are interested in the distribution of long-term sedentary time in populations. This information is important for public health surveillance and examining associations between physical activity and health outcomes.

Measurement of individuals' physical activity time is subject to measurement error. ActivPAL is one type of physical activity monitors. It is worn on the mid-right thigh and uses information about thigh position to determine the time period spent on sitting/lying or standing/stepping. [Grant et al., 2006] reported that the activPAL accuracy for measuring posture and motion is $95\% - 100\%$, therefore the activPAL is regarded as a reliable and unbiased activity monitor. We illustrate the deconvolution methods using the datasets of the active and sedentary behavior study conducted by [Matthews et al., 2013]. During the study, 201 participants wore an activity monitor during waking hours. Records with wear time greater than 10 hours are considered as valid and included in data analysis. We rescale the data such that the wear time is 10 hours. We use a linear mixed model to analyze the activPAL data,

$$y_{ij} = x_i + d_{ij}, \quad i = 1, \cdots, n, j = 1, \cdots, m_i.\ n = 201,\ 1 \leq m_i \leq 9 \tag{5.1}$$

where x_i represents the average daily sedentary time of ith subject over a long time period, i.e., usual daily sedentary time of subject i. The term d_{ij} includes the deviation of the sedentary time of subject i on day j from the usual sedentary time of subject i and the measurement error of activPAL. Furthermore, we assume $x_i \overset{\text{i.i.d.}}{\sim} f_x$, $d_{ij} \overset{\text{i.i.d.}}{\sim} f_d$, x_i and d_{ij} are independent.

5.2 Estimation Methods

5.2.1 Stochastic variational approximation

We apply SVA approach Algorithm 9 to estimate the distribution of x_i under the assumption that $d_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_d^2)$ and estimate σ_d^2 by $\hat{\sigma}_d^2 = \frac{1}{N-n}(y_{ij} - \bar{y}_{i\cdot})^2 = 0.9485$. For the truncated DPMM, we chose concentration parameter $\alpha = 0.1$, number of cluster $K = 10$ and hyper-prior parameters $\lambda_0 = a_0 = c_0 = 0.1$, $\mu_0 = \frac{1}{n}\sum_{i=1}^{n} \overline{y}_{i\cdot}$. Step-function $\delta_t = t^{-1}$ was applied and 3000 iterations were run for SVA.

Figure 5.1 shows the density estimate for average daily sedentary hours by SVA in solid line. It is left-tailed. The naive density estimate was obtained by applying kernel density estimation to $\overline{y}_{i\cdot}$ and shown in dashed line.

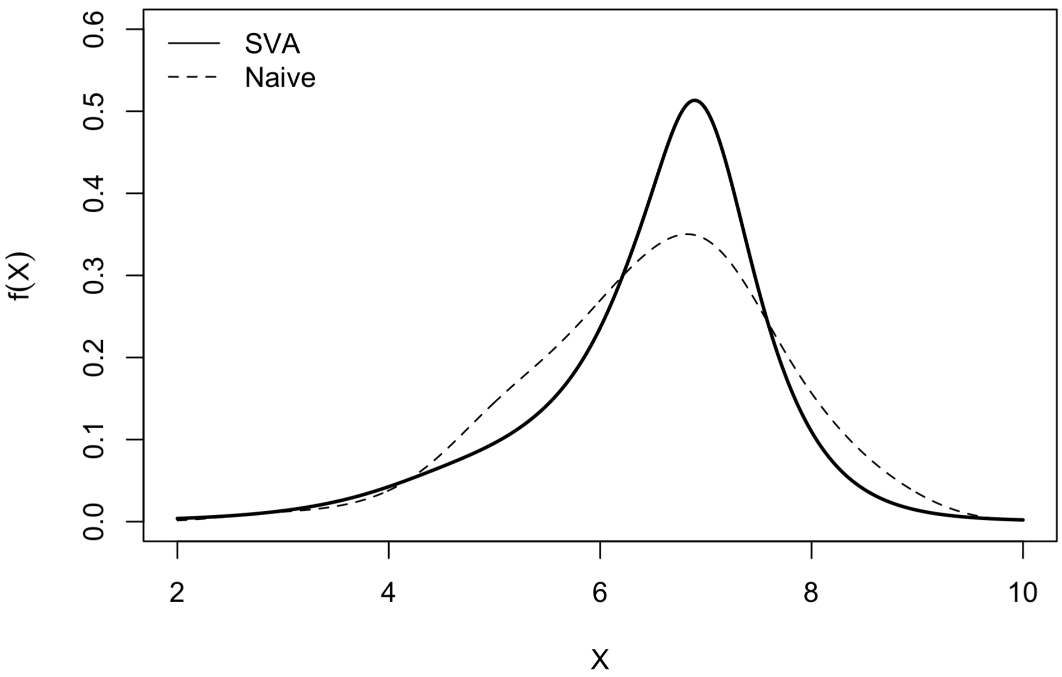

Figure 5.1: Sedentary time density estimation by SVA and naive method

Figure 5.2 shows the posterior mean estimate of $f(x)$ by SVA in solid line and 95% confidence interval of $f(x)$. The 95% confidence interval of $f(x)$ is obtained by Monte Carlo method as follows. Let $\boldsymbol{\theta} = \left\{\mathbf{c}, \pi, \tilde{\boldsymbol{\phi}}\right\}$ be the collection of all hidden variables in the model. We drew $S = 1000$ samples $\boldsymbol{\theta}^{(s)}_{s=1,\cdots,S}$ from variational distributions, i.e., $\mathbf{c}^{(s)} \overset{\text{i.i.d.}}{\sim} q_c^*(\mathbf{c})$, $\boldsymbol{\pi}^{(s)} \overset{\text{i.i.d.}}{\sim} q_\pi^*(\boldsymbol{\pi})$, $\tilde{\boldsymbol{\phi}}^{(s)} \overset{\text{i.i.d.}}{\sim} q_{\tilde{\phi}}^*(\tilde{\boldsymbol{\phi}})$. The density estimate based on $\boldsymbol{\theta}^{(s)}$ is given by $\hat{f}(x)^{(s)} = \sum_{k=1}^{K} \pi_k^{(s)} \mathcal{N}(x; \mu_{\phi,k}^{(s)}, \hat{\sigma}_d^2/t_{\phi,k}^{(s)} - \hat{\sigma}_d^2)$. The 95%

confidence limits of f_x are estimated by $\overline{\hat{f}(x)} \pm z_{0.025} \left(\frac{1}{S-1} \sum_{s=1}^{S} (\hat{f}(x)^{(s)} - \overline{\hat{f}(x)})^2 \right)^{\frac{1}{2}}$, where $\overline{\hat{f}(x)} = \frac{1}{S} \sum_{s=1}^{S} \hat{f}(x)^{(s)}$ and $z_{0.025}$ denotes the 2.5% quantile of the standard normal distribution.

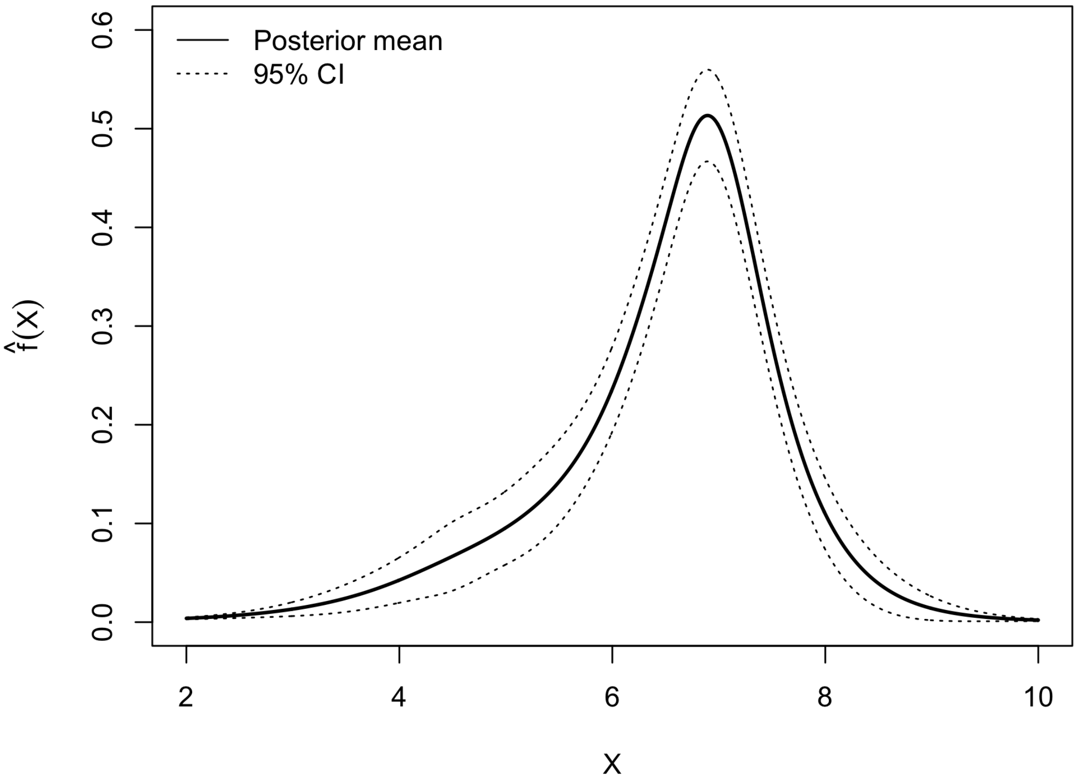

Figure 5.2: Posterior mean and 95% CI of sedentary time density estimated by SVA

5.2.2 MCMC

Figure 5.3 shows the posterior mean estimate of $f(x)$ by MCMC. The density $f(x)$ is modeled by a truncated DPMM with $K = 10$ clusters. A inverse gamma prior was put on σ_d^2, $\sigma_d^2 \sim \mathcal{IG}(0.1, 0.1)$. We chose concentration parameter $\alpha = 0.1$ and hyper-prior parameters $\lambda_0 = \gamma_0 = \beta_0 = 0.1$, $\mu_0 = \frac{1}{n} \sum_{i=1}^{n} \overline{y}_{i\cdot}$. Six thousand iterations were run for MCMC with the first 1000 discarded as burin-in.

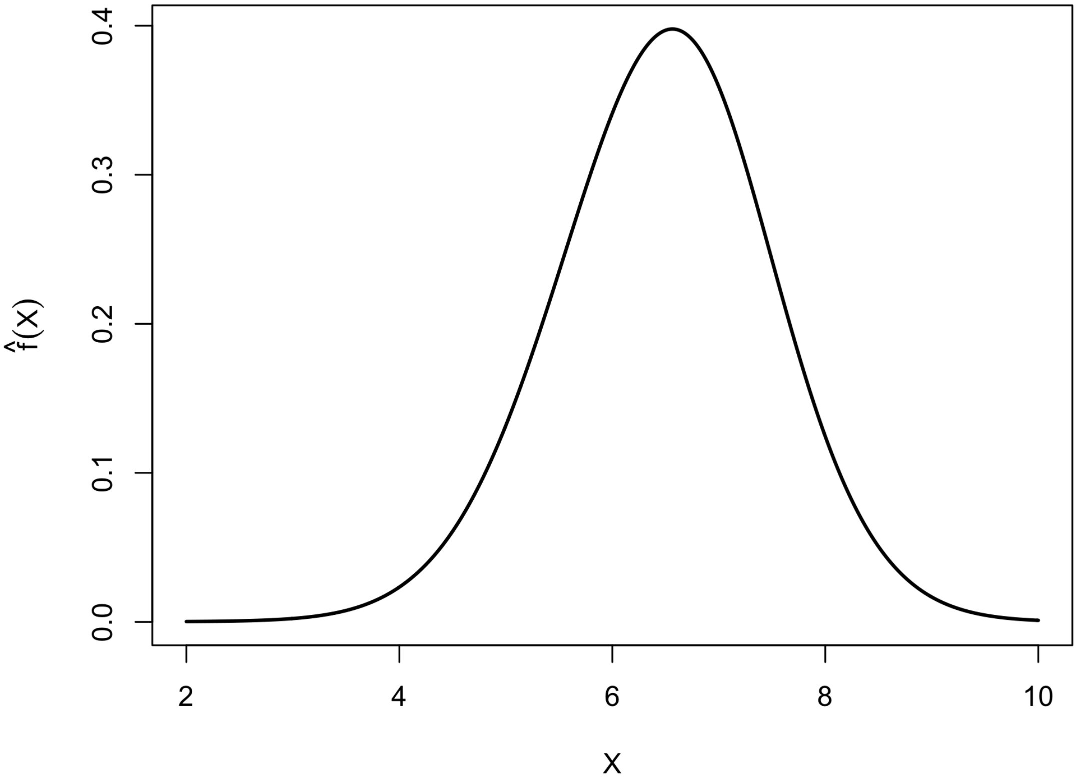

Figure 5.3: Sedentary time density estimation by MCMC

5.2.3 Posterior predictive assessment

We use posterior predictive assessment methods of [Gelman et al., 1996] to check the fitness of model specified in section 5.1. In the framework of posterior predictive assessment, we select a discrepancy statistic denoted by $D(\mathbf{y}; \boldsymbol{\theta})$, which measures the discrepancy of observed data and a posited model M with parameters $\boldsymbol{\theta}$. Define $\mathbf{y}^{rep}$ as the replicated data that would appear if the experiment that produced $\mathbf{y}$ were replicated with the same model M. The posterior predictive distribution of discrepancy is derived from the joint posterior distribution of $\mathbf{y}^{rep}$

and $\boldsymbol{\theta}$,

$$P(\mathbf{y}^{rep}, \boldsymbol{\theta}|M, \mathbf{y}) = P(\mathbf{y}^{rep}|H, \boldsymbol{\theta})P(\boldsymbol{\theta}|M, \mathbf{y})$$

Under the posterior predictive distribution of discrepancy, the posterior predictive p-value is defined as

$$p - value(\mathbf{y}) = P(D(\mathbf{y}^{rep}; \boldsymbol{\theta}) \geq D(\mathbf{y}; \boldsymbol{\theta})|M, \mathbf{y})$$

The discrepancy is chosen as

$$D(\mathbf{y}; \boldsymbol{\theta}) = -2\text{log}p(\mathbf{y}|\boldsymbol{\theta}) = -2\sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\overline{y}_{i\cdot}; \mu_{\phi,k}, \hat{\sigma}_d^2(1/t_{\phi,k} + 1/m_i - 1))\right\}$$

The calculation of posterior predictive p-value is implemented by Monte Carlo simulation. Given $S = 500$ draws $\boldsymbol{\theta}^{(s)}_{s=1,\cdots,S}$ from the variational distributions, we drew a simulated replicated data $\mathbf{y}^{rep,s}$ from $P(\mathbf{y}^{rep}|M, \boldsymbol{\theta}^{(s)})$ for each s and calculated $D(\mathbf{y}; \boldsymbol{\theta}^{(s)})$, $D(\mathbf{y}^{rep,s}; \boldsymbol{\theta}^{(s)})$. Figure 5.4 shows the scatter plots of $D(\mathbf{y}^{rep,s}; \boldsymbol{\theta}^{(s)})$ versus $D(\mathbf{y}; \boldsymbol{\theta}^{(s)})$. The predictive posterior p-value is estimated by the proportion of points above the 45% line, p-value=0.654. We obtained no evidence of lack of fit of the homoscedastic Gaussian error deconvolution model.

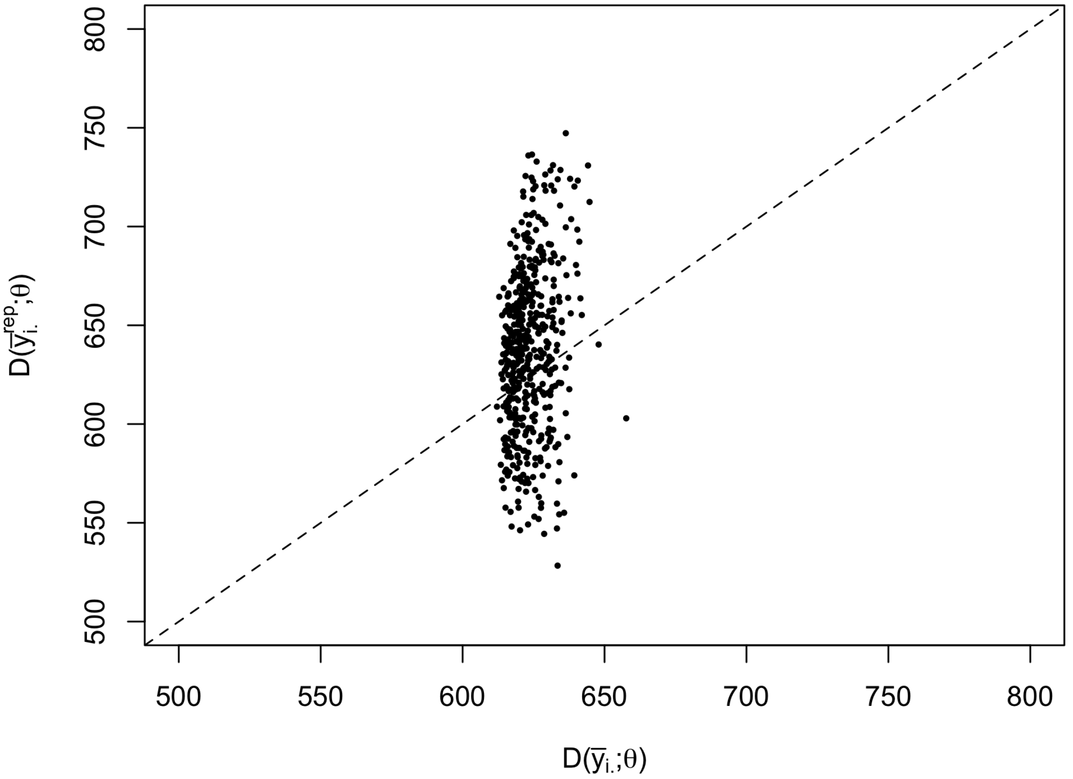

Figure 5.4: Scatter plot of predictive vs realized discrepancies for the homoscedastic Gaussian error deconvolution model

5.2.4 Heteroscedasticity assumption

This subsection discusses the possibility of heteroscedastic measurement error. Figure 5.5 shows the scatter plot of within-subject variance s_i^2 versus subject mean $\overline{y}_{i\cdot}$, where $\overline{y}_{i\cdot} = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ and $s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2$. Figure 5.5 shows no linear or curvilinear trend as $\overline{y}_{i\cdot}$ gets larger, although the points with large s_i^2 scatter in $\overline{y}_{i\cdot} < 7$.

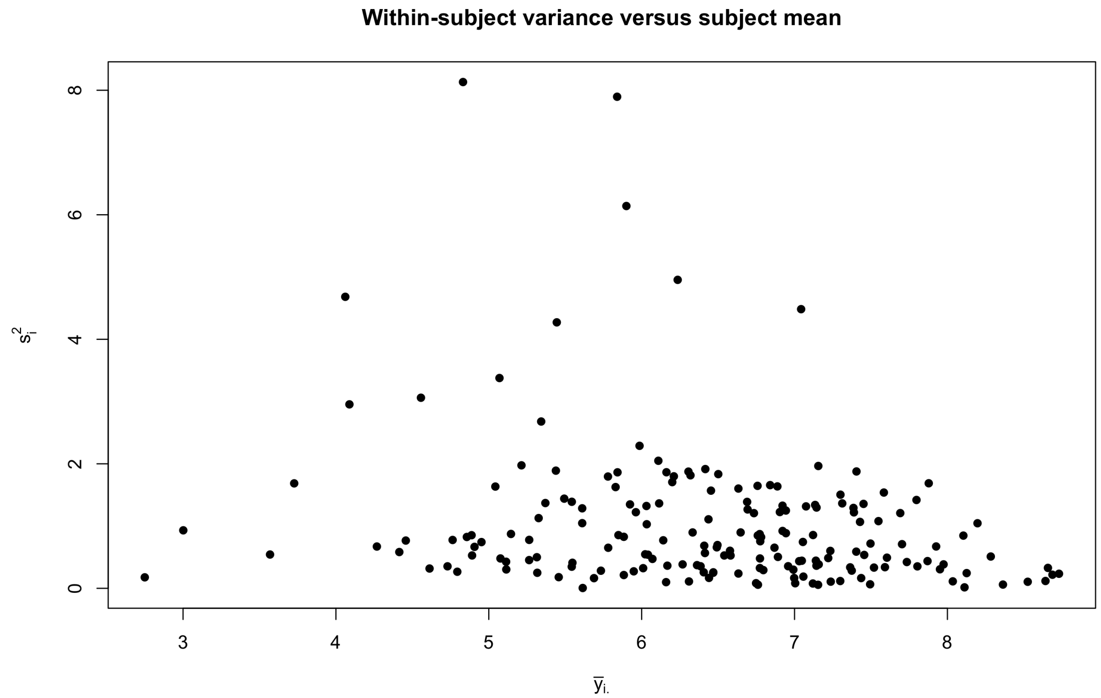

Figure 5.5: Scatter plot of within-subject variance versus subject mean

Under the heteroscedasticity assumption, we model within-subject variance as a function of subject mean, i.e. $var(d_{ij}) = \exp\{g(x_i)\}$. The variance function $g(x_i)$ is specified as a linear mixed model,

$$g(x) = \gamma_{0,\eta} + \gamma_{1,\eta}x + \sum_{k=1}^{K_\eta} \rho_{k,\eta}(x - \kappa_{k,\eta})_+, \quad \boldsymbol{\rho}_\eta \sim \text{MVN}(\mathbf{0}, \sigma^2_{\rho_\eta}\mathbf{I})$$

We put normal prior on $\boldsymbol{\gamma}_\eta$ and inverse gamma prior on $\sigma^2_{\rho_\eta}$: $\boldsymbol{\gamma}_\eta \sim \text{MVN}(\mathbf{0}, \sigma^2_{\gamma_\eta}\mathbf{I})$, $\sigma^2_{\rho_\eta} \sim \text{IG}(a_{\rho_\eta}, b_{\rho_\eta})$.

In figure 5.6, the first plot shows 95% CI interval and posterior mean estimation for $g(x)$; the second figure shows the posterior mean estimate of $f(x)$ by MCMC. The first plot shows that the log variance function $g(x)$ decreases as x increases. The spline function is based on 35 evenly spaced knots on the range of y_{ij}. Latent variables are updated by Metropolis-Hasting algorithm and Gibbs sampling. We run 60000 iterations for MCMC with the first 10000 discarded as burin-in.

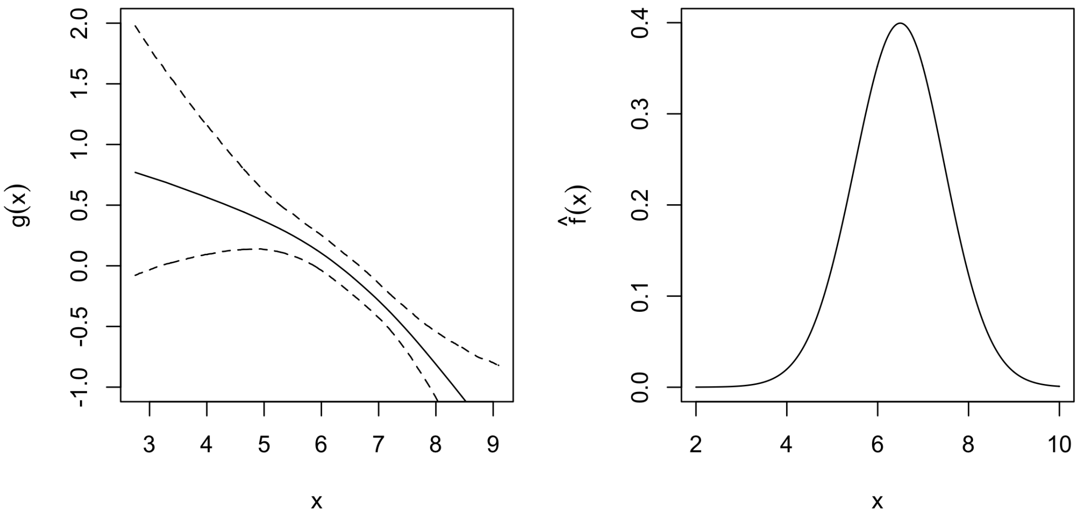

Figure 5.6: Sedentary time density and variance function estimated by MCMC under heteroscedasticity assumption

CHAPTER 6

SUMMARY AND EXTENSIONS

In this chapter, section 6.1 discusses an extension of the deconvolution problem which relaxes the parametric assumption of measurement error; section 6.2 summarizes the thesis.

6.1 Extension to Nonparametric Measurement error

This section considers an extension of the deconvolution problem, where the distribution of measurement error is unknown. The problem is formulated as

$$y_{ij} = x_i + u_{ij},\ i = 1, \cdots, n,\ j = 1, \cdots, m_i \tag{6.1}$$

where $\{y_{ij}\}_{j=1}^{m_i}$ are repeated measurements for x_i, $x_i \overset{\text{i.i.d.}}{\sim} f_x(x)$, $u_{ij} \overset{\text{i.i.d.}}{\sim} f_u(u)$. Both f_x and f_u are unknown, the density of interest is f_x.

6.1.1 Model specification

$$y_{ij} = x_i + u_{ij},\ i = 1, 2, \ldots, n,\ j = 1, \ldots, m_i$$

where $x_i \overset{\text{i.i.d.}}{\sim} f_x(\cdot)$, $u_{ij} \overset{\text{i.i.d.}}{\sim} f_u(\cdot)$, $f_u(\cdot)$ has mean zero. We specify f_x and f_u by mixture of normals: $f_x(x) = \sum_{k=1}^{K_x} \pi_{x,k}\mathcal{N}(x; \mu_{x,k}, \sigma_{x,k}^2)$, $f_u(u) = \sum_{k=1}^{K_u} \pi_{u,k}\mathcal{N}(u; \mu_{u,k}, \sigma_{u,k}^2)$.

We use the method in [Bao and Hanson, 2016] to impose the mean zero constraint on f_u. Let $\zeta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\zeta^2)$ and choose L to be a $K_u \times (K_u - 1)$ matrix such that the $K_u - 1$ columns of L spans the space orthogonal to the vector of all ones $\mathbf{1}_{K_u}$, then $\mu_{u,k} = \pi_{u,k}^{-1} l_k^T \boldsymbol{\zeta}$ satisfy $\sum_{k=1}^{K_u} \pi_{u,k}\mu_{u,k} = 0$, where l_k^T denotes the kth row of L and $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \cdots, \zeta_{K_u-1})$. The Bayesian hierarchical model can be specified as follows:

$$\begin{aligned}
y_{ij} \,|\, x_i, c_{u_{ij}}, \boldsymbol{\zeta}, \boldsymbol{\pi}_u, \boldsymbol{\sigma}_u^2 &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pi_{u,c_{u_{ij}}}^{-1} l_{c_{u_{ij}}}^T \boldsymbol{\zeta},\, \sigma_{u,c_{u_{ij}}}^2),\; i = 1, 2, \ldots, n,\; j = 1, \ldots, m_i \\
x_i \,|\, c_{x_i},\, \boldsymbol{\phi}_x &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_{x,c_{x_i}},\, \sigma_{x,c_{x_i}}^2\right) \\
c_{x_i} \,|\, \boldsymbol{\pi}_x &\overset{\text{i.i.d.}}{\sim} \text{Categorical}\left(\pi_{x,1},\, \pi_{x,2}, \cdots, \pi_{x,K_x}\right) \\
\boldsymbol{\pi}_x &\sim \text{Dirichlet}\left(\alpha_x/K_x, \alpha_x/K_x, \cdots, \alpha_x/K_x\right) \\
\phi_{x,c} &\overset{\text{i.i.d.}}{\sim} \mathcal{NIG}\left(\mu_{x,0},\, \lambda_{x,0},\, \gamma_{x,0},\, \beta_{x,0}\right) \\
c_{u_{ij}} \,|\, \boldsymbol{\pi}_u &\overset{\text{i.i.d.}}{\sim} \text{Categorical}\left(\pi_{u,1},\, \pi_{u,2}, \cdots, \pi_{u,K_u}\right) \\
\boldsymbol{\pi}_u &\sim \text{Dirichlet}\left(\alpha_u/K_u, \alpha_u/K_u, \cdots, \alpha_u/K_u\right) \\
\zeta_k &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\zeta^2\right) \\
\sigma_{u,k}^2 &\overset{\text{i.i.d.}}{\sim} \mathcal{IG}(\gamma_{u,0}, \beta_{u,0})
\end{aligned} \tag{6.2}$$

where c_η (η may be x_i *or* u_{ij}) denotes the cluster from which η is drawn. $\phi_{x,c} = (\mu_{x,c}, \sigma_{x,c}^2)$.

6.1.2 Variational approximation approach

Let $\boldsymbol{\theta} = \{\mathbf{x}, \mathbf{c}_x, \boldsymbol{\phi}_x, \boldsymbol{\pi}_x, \mathbf{c}_u, \boldsymbol{\zeta}, \boldsymbol{\pi}_u, \boldsymbol{\sigma}_u^2\}$ denote the collection of all hidden variables in the model and assume $q(\boldsymbol{\theta})$ can be factorized into
$q(\boldsymbol{\theta}) = q(\mathbf{x})q(\mathbf{c}_x)q(\boldsymbol{\phi}_x)q(\boldsymbol{\pi}_x)q(\mathbf{c}_u)q(\boldsymbol{\zeta})q(\boldsymbol{\pi}_u)q(\boldsymbol{\sigma}_u^2)$. The conditional posterior $p(\theta_j|\mathbf{y}, \boldsymbol{\theta}_{-j})$ has close form for all $\theta_j \in \boldsymbol{\theta} \backslash \boldsymbol{\pi}_u$, therefore $q_{\theta_j}^*(\theta_j)$ can be updated by formula (3.4)

for all $\theta_j \in \boldsymbol{\theta}\backslash\boldsymbol{\pi}_u$. For $q(\boldsymbol{\pi}_u)$, we restrict it to be a Dirichlet density

$$q_{\pi_u}(\boldsymbol{\pi}_u) = \frac{\Gamma(\tilde{\alpha}_{q(\pi_u)})}{\prod_{k=1}^{K_u}\Gamma(\alpha_{q(\pi_{u,k})})}\pi_{u,k}^{\alpha_{q(\pi_{u,k})}-1},\ \textit{where}\ \tilde{\alpha}_{q(\pi_u)} = \sum_{k=1}^{K_u}\alpha_{q(\pi_{u,k})} \tag{6.3}$$

and update $\{\alpha_{q(\pi_{u,k})}\}_{k=1}^{K_u}$ by Newton-Ralphon's method:

$$\boldsymbol{\alpha}_{q(\pi_u)} \leftarrow \text{argmax}\ \boldsymbol{\alpha}_{q(\pi_u):\alpha_{q(\pi_{u,k})}\in(2,+\infty)}\log\underline{p}(\mathbf{y};q) \tag{6.4}$$

Algorithm 11 (Variational approximation algorithm for nonparametric error)**.**

Initialize: $\mu_{q(x_i)}, \mu_{q(\phi_{x,k})} \in \mathrm{R}, \sigma^2_{q(x_i)}, \lambda_{q(\phi_{x,k})}, A_{q(\phi_{x,k})}, B_{q(\phi_{x,k})}, \omega_{ik} > 0$ *for* $k = 1,\ldots,K_x,\ i = 1,\ldots,n,\ j = 1,\ldots,m_i$ *such that* $\sum_{k=1}^{K}\omega_{ik} = 1$ *for all* i*;* $\boldsymbol{\zeta} \in \mathrm{R}^{K_u-1}$ *and* $A_{q(\sigma^2_{u,k})}, B_{q(\sigma^2_{u,k})}, \rho_{ij,k}$ *for* $k = 1,\ldots,K_u,\ i = 1,\ldots,n,\ j = 1,\ldots,m_i$ *such that* $\sum_{k=1}^{K}\rho_{ij,k} = 1$ *for all* i,j*.*

Repeat the following steps until the increase in $log\underline{p}(\mathbf{y};q)$ *is negligible:*

$$\Sigma_\zeta \leftarrow \left(\sum_{k=1}^{K_u}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}}\rho_{\cdot\cdot,k}l_k l_k^T + \frac{I_{K_u-1}}{\sigma^2_\zeta}\right)^{-1}$$

$$\boldsymbol{\mu}_\zeta \leftarrow \Sigma^*_\zeta\left\{\sum_{k=1}^{K_u}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\mu_{q(x_i)})\rho_{ij,k}\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}}l_k\right\}$$

For $i = 1,\ldots,n$*:*

$$\sigma^2_{q(x_i)} \leftarrow \left\{\sum_{k=1}^{K_u}\frac{A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u.k})}}\rho_{i\cdot,k} + \sum_{k=1}^{K_x}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\omega_{ik}\right\}^{-1}$$

$$\mu_{q(x_i)} \leftarrow \sigma^2_{q(x_i)}\left\{\sum_{k=1}^{K_u}\frac{A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u.k})}}\sum_{j=1}^{m_i}\rho_{ij,k}\left(y_{ij}-\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}l_k^T\boldsymbol{\mu}_\zeta\right) + \sum_{k=1}^{K_x}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\omega_{ik}\mu_{q(\phi_{x,i})}\right\}$$

For $i = 1,\ldots,n$ *and* $k = 1,\ldots,K_x$*:*

$$\nu_{ik} \leftarrow -\frac{1}{2}logB_{q(\phi_{x,k})} + \frac{1}{2}\Psi(A_{q(\phi^2_{x,k})}) - \frac{1}{2}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\left((\mu_{q(x_i)}-\mu_{q(\phi_{x,k})})^2 + \sigma^2_{q(x_i)}\right)$$
$$-\frac{1}{2\lambda_{q(\phi_{x,k})}} + \Psi(\alpha_{q(\pi_{x,k})})$$

$$\omega_{ik} \leftarrow \frac{\exp(\tau_{ij,k})}{\sum_{l=1}^{K}\exp(\tau_{ij,l})}$$

For $i = 1, \ldots, n,\ j = 1, \ldots, m_i$ *and* $k = 1, \ldots, K_u$*:*

$$\begin{aligned}\tau_{ij,k} \leftarrow{}& \Psi(\alpha_{q(\pi_{u,k})}) - \frac{1}{2} log B_{q(\sigma^2_{u,k})} + \frac{1}{2}\Psi(A_{q(\sigma^2_{u,k})}) - \frac{1}{2}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}}\Bigg(y_{ij}^2 - 2y_{ij}\mu_{q(x_i)} + \mu^2_{q(x_i)} + \sigma^2_{q(x_i)} \\ & -2(y_{ij}-\mu_{q(x_i)})\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1} l_k^T \boldsymbol{\mu}_\zeta + \frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)} l_k^T(\boldsymbol{\mu}_\zeta\boldsymbol{\mu}_\zeta^T + \Sigma_\zeta) l_k \Bigg)\end{aligned}$$

$$\rho_{ij,k} \leftarrow \frac{\exp(\tau_{ij,k})}{\sum_{l=1}^{K} \exp(\tau_{ij,l})}$$

For $k = 1, \ldots, K_x$*:*

$$\omega_{\cdot k} \leftarrow \sum_{i=1}^{n} \omega_{ik}$$

$$\mu_{q(\phi_{x,k})} \leftarrow \frac{\sum_{i=1}^{n} \mu_{q(x_i)}\omega_{ik} + \lambda_{x,0}\mu_{x,0}}{\omega_{\cdot k} + \lambda_{x,0}}$$

$$\lambda_{q(\phi_{x,k})} \leftarrow \omega_{\cdot k} + \lambda_{x,0}$$

$$A_{q(\phi_{x,k})} \leftarrow \frac{\omega_{\cdot k}}{2} + \gamma_{x,0}$$

$$B_{q(\phi_{x,k})} \leftarrow \beta_{x,0} + \frac{1}{2}\sum_{i=1}^{n} \omega_{ik}\left(\mu^2_{q(x_i)} + \sigma^2_{q(x_i)}\right) + \frac{1}{2}\lambda_{x,0}\mu^2_{x,0} - \frac{1}{2}\left(\omega_{\cdot k} + \lambda_{x,0}\right)\mu^2_{q(\phi_{x,k})}$$

$$\alpha_{q(\pi_{x,k})} \leftarrow \omega_{\cdot k} + \frac{\alpha_x}{K_x}$$

For $k = 1, \ldots, K_u$*:*

$$A_{q(\sigma^2_{u,k})} \leftarrow \frac{\rho_{\cdot\cdot,k}}{2} + \gamma_{u,0}$$

$$\begin{aligned}B_{q(\sigma^2_{u,k})} \leftarrow{}& \beta_{u,0} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(y_{ij}^2 - 2y_{ij}\mu_{q(x_i)} + \mu^2_{q(x_i)} + \sigma^2_{q(x_i)} - 2\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1} l_k^T \mu_\zeta (y_{ij} - \mu_{q(x_i)})\right)\rho_{ij,k} \\ & + \frac{1}{2}\rho_{\cdot\cdot,k}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)} l_k^T(\boldsymbol{\mu}_\zeta\boldsymbol{\mu}_\zeta^T + \Sigma_\zeta) l_k\end{aligned}$$

Update $\boldsymbol{\alpha}_{q(\pi_u)}$ to $\boldsymbol{\alpha}^{*}_{q(\pi_u)}$ which maximizes $log\underline{p}(\mathbf{y};q)$ with other parameters fixed:

$$\begin{aligned}
log\underline{p}(\mathbf{y};q) = &-\frac{1}{2}\sum_{k=1}^{K_u}\frac{A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u,k})}}\Bigg\{-\sum_{i=1}^{n}\sum_{j=1}^{m_i}2(y_{ij}-\mu_{q(x_i)})\rho_{ij,k}\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}l_k^T\mu_\zeta \\
&\qquad + \rho_{\cdot\cdot k}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)}l_k^T(\mu_\zeta\mu_\zeta^T+\Sigma_\zeta)l_k\Bigg\} \\
&+\sum_{k=1}^{K_u}\left(\rho_{\cdot\cdot,k}+\frac{\alpha_u}{K_u}-1\right)\Psi(\alpha_{q(\pi_{u,k})})-\left(N+\alpha_u-\tilde{\alpha}_{q(\pi_u)}\right)\Psi(\tilde{\alpha}_{q(\pi_u)}) \\
&+\sum_{k=1}^{K_u}log\Gamma(\alpha_{q(\pi_{u,k})})-log\Gamma(\tilde{\alpha}_{q(\pi_u)})+C
\end{aligned}$$

where C is a function not containing $\alpha_{q(\pi_{u,k})}$ for all $k=1,\ldots,K_u$. $\alpha_{q(\pi_{u,k})}>2$ for all $k=1,\ldots,K_u$.

Partial derivative of $log\underline{p}(\mathbf{y};q)$ with respect to $\alpha_{q(\pi_{u,k})}$ is given by:

$$\begin{aligned}
\frac{\partial log\underline{p}(\mathbf{y};q)}{\partial\alpha_{q(\pi_{u,k})}} = &-\frac{1}{2}\sum_{t=1}^{K_u}\frac{A_{q(\sigma^2_{u,t})}}{B_{q(\sigma^2_{u,t})}}\Bigg\{-\sum_{i=1}^{n}\sum_{j=1}^{m_i}2(y_{ij}-\mu_{q(x_i)})\rho_{ij,t}\frac{1}{\alpha_{q(\pi_{u,t})}-1}l_t^T\mu_\zeta \\
&\qquad + \rho_{\cdot\cdot,t}\frac{2\tilde{\alpha}_{q(\pi_u)}-3}{(\alpha_{q(\pi_{u,t})}-1)(\alpha_{q(\pi_{u,t})}-2)}l_t^T(\mu_\zeta\mu_\zeta^T+\Sigma_\zeta)l_t\Bigg\} \\
&-\frac{1}{2}\frac{A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u,k})}}\Bigg\{\sum_{i=1}^{n}\sum_{j=1}^{m_i}2(y_{ij}-\mu_{q(x_i)})\rho_{ij,k}\frac{\tilde{\alpha}_{q(\pi_{u,k})}-1}{(\alpha_{q(\pi_{u,k})}-1)^2}l_k^T\mu_\zeta \\
&\qquad - \rho_{\cdot\cdot,k}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)(2\alpha_{q(\pi_{u,k})}-3)}{(\alpha_{q(\pi_{u,k})}-1)^2(\alpha_{q(\pi_{u,k})}-2)^2}l_k^T(\mu_\zeta\mu_\zeta^T+\Sigma_\zeta)l_k\Bigg\} \\
&+\left(\rho_{\cdot\cdot,k}+\frac{\alpha_u}{K_u}-\alpha_{q(\pi_{u,k})}\right)\Psi'(\alpha_{q(\pi_{u,k})})-\left(n+\alpha_u-\tilde{\alpha}_{q(\pi_u)}\right)\Psi'(\tilde{\alpha}_{q(\pi_u)})+C'
\end{aligned}$$

where C' is a function which does not depend on $\alpha_{q(\pi_{u,k})}$.

6.1.3 A simulation example

We consider a target density which is a two-components normal mixture with $\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 0, \mu_2 = 1.5, \sigma_1 = 1, \sigma_2 = 0.2$. The measurement errors are from a two-components normal mixture with $\pi_1 = \pi_2 = 0.5, \mu_1 = -1, \mu_2 = 1, \sigma_1 = \sigma_2 = 0.5$. We generated 100 simulated dataset of size $n = 1000, m_i = 5$ and applied VA algorithm 11 on each dataset. The number of clusters was set at $K_x = 10$ for f_x and $K_u = 5$ for f_u. We chose concentration parameter $\alpha_x = \alpha_u = .1$ and hyper-prior parameters $\gamma_{x,0} = \lambda_{x,0} = \beta_{x,0} = 0.1$, $\mu_{x,0} = \sum_{i=1}^{n} \bar{y}_{i\cdot}/m$, $\gamma_{u,0} = 3$, $\beta_{u,0} = \hat{\sigma}_u^2/(\gamma_{u,0} - 1)$, where $\hat{\sigma}_u^2 = \sum_{i=1}^{n} \sum_{j_1=1}^{m_i} \sum_{j_2=1}^{m_i} \frac{(y_{ij_1} - y_{ij_2})^2}{2m_i(m_i-1)}$. For VA algorithm 11, the loop continues to iterate until the increase in the lower bound of log marginal likelihood $\log \underline{p}(\mathbf{y}; q)$ is less than 10^{-4}.

In addition, we applied SVA algorithm 9 to the datasets by neglecting the non-Gaussian pattern of the measurement error and assuming $\bar{y}_{i\cdot}|x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(x_i, 1.25/5)$. Figure 6.1 shows the density estimates by algorithm 11 and 9 corresponding to $1st, 2nd$ and $3rd$ quartile of ISEs, respectively. VA algorithm 11 cannot catch the two-mode shape and work worse than SVA algorithm 9.

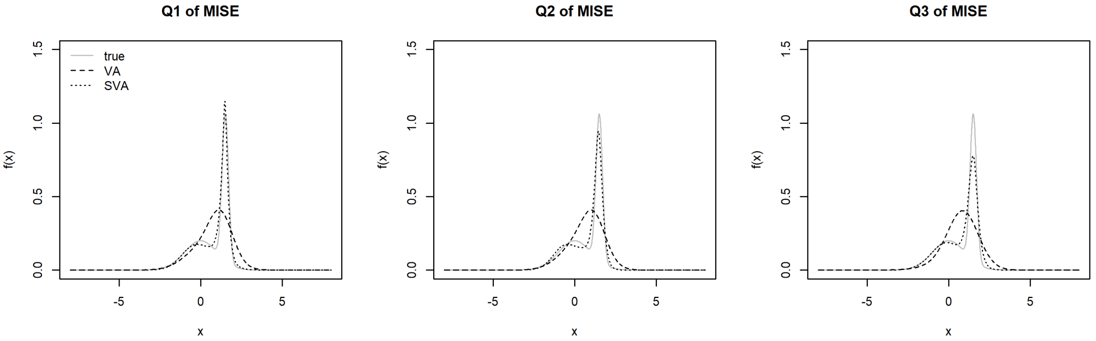

Figure 6.1: Density estimates which correspond to quantiles of MISE

In VA algorithm 11, local hidden variables are $\{\mathbf{x}, \mathbf{c}_x, \mathbf{c}_u\}$. Approximating the posterior by $p(\boldsymbol{\theta}|\mathbf{y}) \approx q(\mathbf{x})q(\mathbf{c}_x)q(\boldsymbol{\phi}_x)q(\boldsymbol{\pi}_x)q(\mathbf{c}_u)q(\boldsymbol{\zeta})q(\boldsymbol{\pi}_u)q(\boldsymbol{\sigma}_u^2)$ leads to loss the dependencies between $\mathbf{x}$ and $\mathbf{c}_x$, $\mathbf{x}$ and $\mathbf{c}_u$, which are supposed to be reason for loss of accuracy.

6.2 Summary

This thesis considered the problem of density estimation when the observations are contaminated with measurement error and developed VA-type approaches to this problem. This thesis had three achievements. First, it developed two variational approximation algorithms for Gaussian error deconvolution. Their performances were compared to deconvoluting kernels and Monte Carlo Markov Chain method by simulation experiments. A conjecture was proposed to explain why two variational approximation algorithms for Gaussian error deconvolution perform differently. Secondly, the thesis established a stochastic variational approximation (SVA) approach to the Bayesian nonparametric model for Gaussian error deconvolution. The SVA approach outperforms DK method and performs comparably well with MCMC at faster speed. The SVA approach for Gaussian error deconvolution was illustrated through simulation experiments and data from a physical activity study.

Thirdly, this thesis also investigated VA approach for Laplacian error deconvolution and extended VA to nonparametric error deconvolution. Simulation experiments showed that DK method performs comparably well with MCMC method for Laplacian error deconvolution. Simulation experiments suggested that breaking dependencies between local hidden variables leads to loss to accuracy in variational inference. It provided an explanation for the fact that the accuracy of VA approach for Gaussian error deconvolution was not achieved for Laplacian or nonparametric error deconvolution. Potential extensions of this thesis would be investigating sufficient and necessary conditions in theory to obtain consistency of VA algorithms for the deconvolution problem.

APPENDIX A

DERIVATION OF VARIATIONAL APPROXIMATION ALGORITHMS

In this chapter, we derived VA algorithm A and B for Gaussian error deconvolution, VA algorithms for Laplacian and Nonparametric error.

Gaussian or Laplacian measurement error with known variance.

VA Algorithm A

According to model 3.2, the joint distribution of $(\mathbf{y}, \mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi})$ is given by

$$\begin{aligned}
&\log p(\mathbf{y}, \mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi}|\alpha, \mu_0, \lambda_0, \gamma_0, \beta_0) \\
&= \log p(\mathbf{y}|\mathbf{x}, \sigma_u^2) + \log p(\mathbf{x}|\mathbf{c}, \boldsymbol{\phi}) + \log p(\mathbf{c}|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi}|\alpha) + \log p(\boldsymbol{\phi}|\mu_0, \lambda_0, \gamma_0, \beta_0) \\
&= \log p(\mathbf{y}|\mathbf{x}, \sigma_u^2) + \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{-\frac{1}{2}\log(2\pi\sigma_k^2) - \frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right\}\delta_k(c_i) + \sum_{i=1}^{n}\sum_{k=1}^{K}\delta_k(c_i)\log(\pi_k) \\
&\quad + \log\Gamma(\alpha) - K\log\Gamma(\frac{\alpha}{K}) + (\frac{\alpha}{K}-1)\sum_{k=1}^{K}\log\pi_k + K\gamma_0\log\beta_0 - K\log\Gamma(\gamma_0) \\
&\quad + \sum_{k=1}^{K}\left\{-(\gamma_0+1)\log(\sigma_k^2) - \frac{\beta_0}{\sigma_k^2} - \frac{1}{2}\log(2\pi\sigma_k^2/\lambda_0) - \frac{\lambda_0(\mu_k-\mu_0)^2}{2\sigma_k^2}\right\} \qquad \text{(A.1)}
\end{aligned}$$

If f_u is a Gaussian distribution, the optimal density $q^*_{\theta_j}(\theta_j) \propto \exp\{\mathrm{E}_{-\theta_j}\log p(\mathbf{y}, \boldsymbol{\theta})\}$, $\theta_j \in \{\mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi}\}$ are in the same exponential family as the prior of θ_j due to conjugacy of the model. Following is the derivation for $q^*_{\theta_j}(\theta_j)$.

Table A.1: Conditional expectations to be evaluated in $q_\eta(\eta)^*$

η	prior	$q^*_\eta(\eta)$	conditional expectations in $q^*_\eta(\eta)$
x_i	$\mathcal{N}(x;\, \mu_{c_i}, \sigma^2_{c_i})$	$\mathcal{N}(x;\, \mu^*_{q(x_i)}, \sigma^{2*}_{q(x_i)})$	$\mathrm{E}_{-\mathbf{x}}\frac{1}{\sigma^2_k}, \mathrm{E}_{-\mathbf{x}}\delta_k(c_i)$
c_i	$\mathrm{Cat}(c_i;\, \pi_1, \pi_2, \cdots, \pi_K)$	$\mathrm{Cat}(c_i; \omega^*_{i1}, \omega^*_{i2}, \cdots, \omega^*_{iK})$	$\mathrm{E}_{-\mathbf{c}}\log\sigma^2_k, \mathrm{E}_{-\mathbf{c}}\frac{(x_i-\mu_k)^2}{\sigma^2_k}, \mathrm{E}_{-\mathbf{c}}\log\pi_k$
$\boldsymbol{\pi}$	$\mathrm{Dir}(\boldsymbol{\pi};\, \frac{\alpha}{K}, \frac{\alpha}{K}, \cdots, \frac{\alpha}{K})$	$\mathrm{Dir}(\boldsymbol{\pi}; \alpha^*_{q(\pi),1}, \alpha^*_{q(\pi),2}, \cdots, \alpha^*_{q(\pi),K})$	$\mathrm{E}_{-\pi}\delta_k(c_i)$
ϕ_k	$\mathcal{NIG}(\phi_k;\, \mu_0, \lambda_0, \gamma_0, \beta_0)$	$\mathcal{NIG}(\phi_k;\, \mu^*_{q(\phi_k)}, \lambda^*_{q(\phi_k)}, \gamma^*_{q(\phi_k)}, \beta^*_{q(\phi_k)})$	$\mathrm{E}_{-\phi}\log\sigma^2_k, \mathrm{E}_{-\phi}\frac{(x_i-\mu_k)^2}{\sigma^2_k}, \mathrm{E}_{-\phi}\delta_k(c_i)$

Table A.2: Values of conditional expectations

Conditional Expectation	Value
$E_{-\mathbf{x}}\frac{1}{\sigma^2_k}$	$A_{q(\phi_k)}/B_{q(\phi_k)}$
$\mathrm{E}_{-\mathbf{c}}\log\sigma^2_k$, $\mathrm{E}_{-\phi}\log\sigma^2_k$	$\log B_{q(\phi_k)} - \Psi(A_{q(\phi_k)})$
$\mathrm{E}_{-\mathbf{c}}\frac{(x_i-\mu_k)^2}{\sigma^2_k}$	$\left((\mu_{q(x_i)} - \mu_{q(\phi_k)})^2 + \sigma^2_{q(x_i)}\right) A_{q(\phi_k)}/B_{q(\phi_k)} + 1/\lambda_{q(\phi_k)}$
$\mathrm{E}_{-\mathbf{c}}\log\pi_k$	$\Psi(\alpha_{q(\pi),k}) - \Psi(\sum_{k=1}^K \alpha_{q(\pi),k})$
$\mathrm{E}_{-\pi}\delta_k(c_i)$, $\mathrm{E}_{-\phi}\delta_k(c_i)$	ω_{ik}
$\mathrm{E}_{-\phi}\frac{(x_i-\mu_k)^2}{\sigma^2_k}$	$(\sigma^2_{q(x_i)} + \mu^2_{q(x_i)} - 2\mu_{q(x_i)}\mu_k + \mu^2_k)/\sigma^2_k$

- Derivation of $q^*_x(\mathbf{x})$

$$
\begin{aligned}
q^*_x(\mathbf{x}) &\propto \exp\left\{\mathrm{E}_{-x}\log p(\mathbf{y}, \mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi})\right\} \\
&\propto \exp\left\{\mathrm{E}_{-x}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(-\frac{1}{2}\log(2\pi\sigma^2_u) - \frac{(y_{ij}-x_i)^2}{2\sigma^2_u}\right)\right. \\
&\left. +\sum_{i=1}^{n}\sum_{k=1}^{K}\left(-\frac{1}{2}\log(2\pi\sigma^2_k) - \frac{(x_i-\mu_k)^2}{2\sigma^2_k}\right)\delta_k(c_i)\right\} \\
&\propto \exp\left\{\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{1}{\sigma^2_u} + \frac{A_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}\right)\left(x_i - \frac{\frac{y_i}{\sigma^2_u} + \sum_{k=1}^K \frac{A_{q(\phi_k)}\mu_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}}{\frac{1}{\sigma^2_u} + \sum_{k=1}^K \frac{A_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}}\right)^2\right\} \\
&\propto \prod_{i=1}^{n}\mathcal{N}(x_i; \mu_{q(x_i)}, \sigma^2_{q(x_i)}) \qquad \text{(A.2)}
\end{aligned}
$$

where

$$\sigma^2_{q(x_i)} = \left(\frac{m_i}{\sigma^2_u} + \sum_{k=1}^{K} \frac{A_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}} \right)^{-1}, \quad \mu_{q(x_i)} = \frac{\frac{y_{i\cdot}}{\sigma^2_u} + \sum_{k=1}^{K} \frac{A_{q(\phi_k)}\mu_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}}{\frac{m_i}{\sigma^2_u} + \sum_{k=1}^{K} \frac{A_{q(\phi_k)}\omega_{ik}}{B_{q(\phi_k)}}}$$

- Derivation of $q^*_\phi(\boldsymbol{\phi})$

$$\begin{aligned}
q^*_\phi(\boldsymbol{\phi}) &\propto \exp\left\{\mathrm{E}_{-\phi}\log p(\mathbf{y}, \mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi})\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\phi} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \left(-\frac{1}{2}\log\sigma^2_k - \frac{(x_i-\mu_k)^2}{2\sigma^2_k} \right) \delta_k(c_i) \right.\right. \\
&\qquad \left.\left. -\frac{1}{2}\log\sigma^2_k - \frac{\lambda_0(\mu_k-\mu_0)^2}{2\sigma^2_k} - \frac{\beta_0}{\sigma^2_k} - (\gamma_0+1)\log\sigma^2_k \right) \right\} \\
&\propto \exp\left\{ \sum_{k=1}^{K} -\frac{\omega_{\cdot k}+\lambda_0}{2\sigma^2_k} \left(\mu_k - \frac{\sum_{i=1}^{n}\mu_{q(x_i)}\omega_{ik} + \lambda_0\mu_0}{\omega_{\cdot k+\lambda_0}} \right)^2 - \left(\frac{\omega_{\cdot k}+1}{2} + \gamma_0 + 1 \right) \log\sigma^2_k \right. \\
&\quad \left. - \left(\beta_0 + \frac{1}{2}\sum_{i=1}^{n} \omega_{ik} \left(\mu^2_{q(x_i)} + \sigma^2_{q(x_i)} \right) + \frac{1}{2}\lambda_0\mu_0^2 - \frac{1}{2}\left(\omega_{\cdot k}+\lambda_0\right)\mu^2_{q(\phi_k)} \right) \frac{1}{\sigma^2_k} \right\} \qquad \text{(A.3)} \\
&\propto \prod_{k=1}^{K} \mathcal{NIG}(\phi_k; \mu^*_{q(\phi_k)}, \lambda^*_{q(\phi_k)}, A^*_{q(\phi_k)}, B^*_{q(\phi_k)}) \qquad \text{(A.4)}
\end{aligned}$$

where

$$\begin{aligned}
\mu^*_{q(\phi_k)} &= \frac{\sum_{i=1}^{n}\mu_{q(x_i)}\omega_{ik} + \lambda_0\mu_0}{\omega_{\cdot k+\lambda_0}} \\
\lambda^*_{q(\phi_k)} &= \omega_{\cdot k} + \lambda_0 \\
A^*_{q(\phi_k)} &= \frac{\omega_{\cdot k}}{2} + \gamma_0 \\
B^*_{q(\phi_k)} &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n} \omega_{ik} \left(\mu^2_{q(x_i)} + \sigma^2_{q(x_i)} \right) + \frac{1}{2}\lambda_0\mu_0^2 - \frac{1}{2}\left(\omega_{\cdot k}+\lambda_0\right)\mu^2_{q(\phi_k)}
\end{aligned}$$

- Derivation of $q^*_{\mathbf{c}}(\mathbf{c})$

$$\begin{aligned}
q^*_{\mathbf{c}}(\mathbf{c}) &\propto \exp\left\{\mathrm{E}_{-\mathbf{c}}\log p(\mathbf{y}, \mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi})\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\mathbf{c}}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(-\frac{1}{2}\log\sigma_k^2 - \frac{1}{2}\frac{(x_i-\mu_k)^2}{\sigma_k^2} + \log\pi_k\right)\delta_k(c_i)\right\} \\
&\propto \prod_{i=1}^{n}\exp\left\{\sum_{k=1}^{K}\left(-\frac{1}{2}\log B_{q(\phi_k)} + \frac{1}{2}\Psi(A_{q(\phi_k)})\right.\right. \\
&\left.\left. -\frac{1}{2}\frac{A_{q(\phi_k)}}{B_{q(\phi_k)}}\left((\mu_{q(x_i)} - \mu_{q(\phi_k)})^2 + \sigma^2_{q(x_i)}\right) - \frac{1}{2\lambda_{q(\phi_k)}} + \Psi(\alpha_{q(\pi),k})\right)\delta_k(c_i)\right\} \\
&\propto \prod_{i=1}^{n}\text{Categorical}\left(c_i; \omega^*_{i1}, \omega^*_{i2}, \ldots, \omega^*_{iK}\right) \qquad \text{(A.5)}
\end{aligned}$$

where

$$\begin{aligned}
\omega^*_{ik} &= \frac{\exp(\nu_{ik})}{\sum_{l=1}^{K}\exp(\nu_{il})} \\
\nu_{ik} &= -\frac{1}{2}\log B_{q(\phi_k)} + \frac{1}{2}\Psi(A_{q(\phi_k^2)}) - \frac{1}{2}\frac{A_{q(\phi_k)}}{B_{q(\phi_k)}}\left((\mu_{q(x_i)} - \mu_{q(\phi_k)})^2 + \sigma^2_{q(x_i)}\right) \\
&\quad - \frac{1}{2\lambda_{q(\phi_k)}} + \Psi(\alpha_{q(\pi),k})
\end{aligned}$$

for $i = 1, 2 \ldots, n$ and $k = 1, 2 \ldots, K$. Digamma function is denoted by Ψ.

- Derivation of $q^*_{\pi}(\boldsymbol{\pi})$

$$\begin{aligned}
q^*_{\pi}(\boldsymbol{\pi}) &\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\pi}}\log p(\mathbf{y}, \mathbf{x}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\phi})\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\pi}}\sum_{k=1}^{K}\left(\left(\sum_{i=1}^{n}\delta_k(c_i)\right) + \frac{\alpha}{K} - 1\right)\log\pi_k\right\} \\
&\propto \prod_{k=1}^{K}\pi_k^{\omega_{\cdot k} + \frac{\alpha}{K} - 1} \\
&\propto \text{Dirichlet}\left(\boldsymbol{\pi}; \alpha^*_{q(\pi),1}, \alpha^*_{q(\pi),2}, \ldots, \alpha^*_{q(\pi),k}\right) \qquad \text{(A.6)}
\end{aligned}$$

where $\alpha^*_{q(\pi),k} = \omega_{\cdot k} + \frac{\alpha}{K}$ for $k = 1, 2, \ldots, K$.

If f_u is Laplacian density, the optimal density $q_x(x) \propto \exp\{\mathrm{E}_{-x}\log p(\mathbf{y}, \boldsymbol{\theta})\}$ does not have closed form. We restrict $q_{x_i}(x_i)$ to be normal distributions with mean $\mu_{q(x_i)}$ and standard deviation $\sigma_{q(x_i)}$ and update $(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$ by Newton-Ralphson's method:

$$(\mu^*_{q(x_i)}, \sigma^{2*}_{q(x_i)}) \leftarrow \operatorname{argmax}_{\substack{\mu_{q(x_i)} \in \mathrm{R} \\ \sigma^2_{q(x_i)} \in \mathrm{R}^+}} \log \underline{p}(\mathbf{y}; q) \tag{A.7}$$

The coordinate decent algorithm cyclically iterates to update each variational parameter until the increase of the Evidence Lower Bound (ELBO) is less than the threshold. The ELBO can be found by

$$\begin{aligned}\log\underline{p}(\mathbf{y}; q) = {} & \text{Entropy}\{q(\mathbf{x})\} + \text{Entropy}\{q(\boldsymbol{\phi})\} + \text{Entropy}\{q(\mathbf{c})\} + \text{Entropy}\{q(\boldsymbol{\pi})\} \\ & + \mathrm{E}\{\log p(\mathbf{y}|\mathbf{x}, \sigma_u^2)\} + \mathrm{E}\{\log p(\mathbf{x}|\mathbf{c}, \boldsymbol{\phi})\} + \mathrm{E}\{\log p(\mathbf{c}|\boldsymbol{\pi})\} + \mathrm{E}\{\log p(\boldsymbol{\pi}|\alpha)\} \\ & + \mathrm{E}\{\log p(\boldsymbol{\phi}|\mu_0, \lambda_0, \gamma_0, \beta_0)\}\end{aligned} \tag{A.8}$$

The explicit expression for each term is given by

$$\text{Entropy}\{q(\mathbf{x})\} = \frac{n}{2}\log(2\pi) + \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} \log(\sigma^2_{q(x_i)}) \tag{A.9}$$

$$\begin{aligned}\text{Entropy}\{q(\boldsymbol{\phi})\} = {} & \frac{K}{2}\log(2\pi) + \frac{K}{2} + \sum_{k=1}^{K} \Big\{ -\frac{1}{2}\log(\lambda_{q(\phi_k)}) + A_{q(\phi_k)} + \frac{3}{2}\log(B_{q(\phi_k)}) \\ & + \log\Gamma(A_{q(\phi_k)}) - (A_{q(\phi_k)} + \frac{3}{2})\Psi(A_{q(\phi_k)}) \Big\}\end{aligned} \tag{A.10}$$

$$\text{Entropy}\{q(\mathbf{c})\} = -\sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik}\log\omega_{ik} \tag{A.11}$$

$$\begin{aligned}\text{Entropy}\{q(\boldsymbol{\pi})\} = {} & -\log\Gamma\left(\sum_{k=1}^{K} \alpha_{q(\pi_k)}\right) + \sum_{k=1}^{K} \big(\log\Gamma(\alpha_{q(\pi_k)}) \\ & -(\alpha_{q(\pi_k)} - 1)\left(\Psi(\alpha_{q(\pi_k)}) - \Psi\left(\sum_{k=1}^{K} \alpha_{q(\pi_k)}\right)\right)\Big)\end{aligned} \tag{A.12}$$

$$\mathrm{E}\{\log p(\mathbf{y}|\mathbf{x},\sigma_u^2)\} = -\frac{N}{2}\log(2\pi\sigma_u^2) - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left((y_{ij}-\mu_{q(x_i)})^2+\sigma^2_{q(x_i)}\right) \tag{A.13}$$

where $N=\sum_{i=1}^{n} m_i$, if f_u is a Gaussian distribution.

$$\mathrm{E}\{\log p(\mathbf{y}|\mathbf{x},\sigma_u^2)\} = -\frac{N\log(2b)}{2} - \frac{1}{b}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sigma_{q(x_i)}z_{ij}\left(2\Phi(z_{ij})-1\right)+2\phi(z_{ij}) \tag{A.14}$$

where $b=\sigma_u/\sqrt{2}$ and $z_{ij}=\frac{y_{ij}-\mu_{q(x_i)}}{\sigma_{q(x_i)}}$, if f_u is a Laplacian distribution.

$$\begin{aligned}\mathrm{E}\{\log p(\mathbf{x}|\mathbf{c},\boldsymbol{\phi})\} = &-\frac{n}{2}\log(2\pi) - \sum_{k=1}^{K}\frac{\omega_{\cdot k}}{2}\left(\log(B_{q(\phi_k)}) - \Psi(A_{q(\phi_k)}) + \frac{1}{\lambda_{q(\phi_k)}}\right)\\ &- \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\omega_{ik}A_{q(\phi_k)}}{2B_{q(\phi_k)}}\left((\mu_{q(x_i)}-\mu_{q(\phi_k)})^2+\sigma^2_{q(x_i)}\right)\end{aligned} \tag{A.15}$$

$$\mathrm{E}\{\log p(\mathbf{c}|\boldsymbol{\pi})\} = \sum_{k=1}^{K}\omega_{\cdot k}\Psi(\alpha_{q(\pi_k)}) - n\Psi(\alpha+n) \tag{A.16}$$

$$\mathrm{E}\{\log p(\boldsymbol{\pi}|\alpha)\} = \log\Gamma\left(\alpha\right) - K\log\Gamma\left(\frac{\alpha}{K}\right) + \left(K-\alpha\right)\Psi\left(\alpha+m\right) + \sum_{k=1}^{K}\left(\frac{\alpha}{K}-1\right)\Psi(\alpha_{q(\pi_k)}) \tag{A.17}$$

$$\begin{aligned}\mathrm{E}\{\log p(\boldsymbol{\phi}|\mu_0,\lambda_0,\gamma_0,\beta_0)\} = &-\frac{K}{2}\log(2\pi) + \frac{K}{2}\log(\lambda_0) + K\gamma_0\log(\beta_0) - K\log(\Gamma(\gamma_0))\\ &- \sum_{k=1}^{K}\left\{\frac{\lambda_0}{2\lambda_{q(\phi_k)}} + \frac{\lambda_0 A_{q(\phi_k)}}{2B_{q(\phi_k)}}\left((\mu_{q(\phi_k)}-\mu_0)^2+\frac{2\beta_0}{\lambda_0}\right)\right.\\ &\left. + \left(\frac{3}{2}+\gamma_0\right)\left(\log(B_{q(\sigma_k^2)}) - \Psi(A_{q(\sigma_k^2)})\right)\right\}\end{aligned} \tag{A.18}$$

VA Algorithm B

The logarithm of the joint distribution of $\mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\mu}_\phi, \mathbf{t}_\phi$ and $\mathbf{y}$ is,

$$
\begin{aligned}
&\log p(\mathbf{y}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\mu}_\phi, \mathbf{t}_\phi | \alpha, \mu_0, \lambda_0, a_0, b_0, \sigma_u^2) \\
&= \log p(\mathbf{y}|\mathbf{c}, \boldsymbol{\mu}_\phi, \mathbf{t}_\phi, \sigma_u^2) + \log p(\mathbf{c}|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi}|\alpha) + \log p(\boldsymbol{\mu}_\phi|\mathbf{t}, \mu_0, \lambda_0, \sigma_u^2) + \log p(\mathbf{t}|a_0, c_0) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ -\frac{1}{2}\log(2\pi\sigma_u^2/t_{\phi,k}) - \frac{t_{\phi,k}(y_i - \mu_{\phi,k})^2}{2\sigma_u^2} \right\} \delta_k(c_i) + \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_k(c_i)\log(\pi_k) \\
&\quad + \log\Gamma(\alpha) - K\log\Gamma(\frac{\alpha}{K}) + (\frac{\alpha}{K} - 1)\sum_{k=1}^{K} \log\pi_k + Ka_0\log c_0 - K\log\Gamma(a_0) \\
&- K\log(F_g(1; a_0, c_0)) + \sum_{k=1}^{K} \{(a_0 - 1)\log(t_{\phi,k}) - c_0 t_{\phi,k} - \\
&\frac{1}{2}\log\left(\frac{2\pi\sigma_u^2}{t_{\phi,k}\lambda_0}\right) - \frac{\lambda_0 t_{\phi,k}(\mu_{\phi,k} - \mu_0)^2}{2\sigma_u^2} \Bigg\} \qquad \text{(A.19)}
\end{aligned}
$$

where $F_g(x; a, c)$ represent the cumulative distribution function of gamma random variable $X \sim \mathrm{G}(a, c)$:

$$
F_g(x; a, c) = \int_0^x \frac{c^a}{\Gamma(a)} t^{a-1} \exp\left(-ct\right) dt, \;\; x > 0 \qquad \text{(A.20)}
$$

Since $p(\boldsymbol{\pi}|\alpha)$, $p(\mathbf{c}|\boldsymbol{\pi})$ and $p(\boldsymbol{\mu}_\phi, \mathbf{t}_\phi|\mu_0, \lambda_0, a_0, c_0)$ are conjugate priors, the optimal density $q_{\mathbf{c}}^*(\mathbf{c})$, $q_\pi^*(\boldsymbol{\pi})$ and $q_\phi^*(\boldsymbol{\phi})$ can be derived by $q_{\theta_j}^*(\theta_j) \propto \exp\{\mathrm{E}_{-\theta_j} \log p(\mathbf{y}, \boldsymbol{\theta})\}$. For concision of the notations, we omit the known parameters $\{\sigma_u^2, \alpha, \mu_0, \lambda_0, a_0, c_0\}$ in the derivation. Assume X is a truncated Gamma random variable $X \sim \mathcal{TG}((0, 1]; a, c)$,$a >$

$0, c > 0$, the following results of X will be used in the derivation:

$$E[X|a,c] = \int_0^1 \frac{c^a}{\Gamma(a)F_g(1;a,c)} t^a \mathrm{exp}\,(-ct)\,dt = \frac{aF_g(1;a+1,c)}{cF_g(1;a,c)} \tag{A.21}$$

$$E[\mathrm{log}X|a,c] = \int_0^1 \frac{c^a}{\Gamma(a)F_g(1;a,c)} \mathrm{log}(t) t^{a-1} \mathrm{exp}\,(-ct)\,dt \tag{A.22}$$

$$\begin{aligned} &= \frac{c^a}{\Gamma(a)F_g(1;a,c)} \frac{\partial}{\partial a} \int_0^1 t^{a-1} \mathrm{exp}\,(-ct)\,dt \\ &= \frac{c^a}{\Gamma(a)F_g(1;a,c)} \lim_{\delta \to 0} \frac{1}{\delta} \left(\frac{\Gamma(a+\delta)F_g(1;a+\delta,c)}{c^{a+\delta}} - \frac{\Gamma(a)F_g(1;a,c)}{c^a} \right) \end{aligned} \tag{A.23}$$

- Derivation of $q^*_\phi(\boldsymbol{\phi})$

$$\begin{aligned} q^*_\phi(\boldsymbol{\phi}) &\propto \mathrm{exp}\left\{ \mathrm{E}_{-\phi} \mathrm{log}\, p(\mathbf{y}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\mu}_\phi, \mathbf{t}_\phi) \right\} \\ &\propto \mathrm{exp}\left\{ \mathrm{E}_{-\phi} \sum_{k=1}^K \left(\sum_{i=1}^n \left(\frac{1}{2}\mathrm{log}(t_{\phi,k}) - \frac{t_{\phi,k}(y_i - \mu_{\phi,k})^2}{2\sigma_u^2} \right) \delta_k(c_i) \right. \right. \\ &\qquad \left. \left. +(a_0 - 1)\mathrm{log}(t_{\phi,k}) - c_0 t_{\phi,k} - \frac{1}{2}\mathrm{log}\left(\frac{2\pi\sigma_u^2}{\lambda_0 t_{\phi,k}} \right) - \frac{\lambda_0 t_{\phi,k}(\mu_{\phi,k} - \mu_0)^2}{2\sigma_u^2} \right) \right\} \\ &\propto \mathrm{exp}\left\{ \sum_{k=1}^K - \frac{t_{\phi,k}(\omega_{\cdot k} + \lambda_0)}{2\sigma_u^2} \left(\mu_{\phi,k} - \frac{\sum_{i=1}^n y_i \omega_{ik} + \lambda_0 \mu_0}{\omega_{\cdot k + \lambda_0}} \right)^2 \right. \\ &\qquad - \left(\frac{\omega_{\cdot k} + 1}{2} + a_0 - 1 \right) \mathrm{log}(t_{\phi,k}) \\ &\left. - \left(c_0 + \frac{1}{2\sigma_u^2} \left(\sum_{i=1}^n \omega_{ik} y_i^2 + \lambda_0 \mu_0^2 - (\omega_{\cdot k} + \lambda_0)\, \mu^2_{q(\phi_k)} \right) \right) t_{\phi,k} \right\} \\ &\propto \prod_{k=1}^K \mathcal{N}\left(\mu_{\phi,k}; \mu^*_{q(\phi_k)}, \frac{\sigma_u^2}{\lambda^*_{q(\phi_k)} t_{\phi,k}} \right) \times \mathcal{TG}(t_{\phi,k}; (0,1]; A^*_{q(\phi_k)}, C^*_{q(\phi_k)}) \end{aligned} \tag{A.24}$$

where

$$
\begin{aligned}
\mu^*_{q(\phi_k)} &= \frac{\sum_{i=1}^n y_i\omega_{ik} + \lambda_0\mu_0}{\omega_{\cdot k+\lambda_0}} \\
\lambda^*_{q(\phi_k)} &= \omega_{\cdot k} + \lambda_0 \\
A^*_{q(\phi_k)} &= \frac{\omega_{\cdot k}}{2} + a_0 \\
C^*_{q(\phi_k)} &= c_0 + \frac{1}{2\sigma_u^2}\left(\sum_{i=1}^n \omega_{ik}y_i^2 + \lambda_0\mu_0^2 - (\omega_{\cdot k} + \lambda_0)\,\mu^2_{q(\phi_k)}\right)
\end{aligned}
$$

- Derivation of $q^*_{\mathbf{c}}(\mathbf{c})$

$$
\begin{aligned}
q^*_{\mathbf{c}}(\mathbf{c}) &\propto \exp\left\{\mathrm{E}_{-\mathbf{c}}\log p(\mathbf{y}, \mathbf{c}, \boldsymbol{\pi}, \boldsymbol{\mu}_\phi, \mathbf{t}_\phi)\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\mathbf{c}}\sum_{i=1}^n\sum_{k=1}^K\left(\frac{1}{2}\log(t_{\phi,k}) - \frac{t_{\phi,k}(y_i-\mu_{\phi,k})^2}{2\sigma_u^2} + \log\pi_k\right)\delta_k(c_i)\right\} \\
&\propto \prod_{i=1}^n \exp\left\{\sum_{k=1}^K\left(\frac{1}{2}E[\log(t_{\phi,k})|A_{q(\phi_k)}, C_{q(\phi_k)}] - \frac{1}{2\sigma_u^2}E[t_{\phi,k}|A_{q(\phi_k)}, C_{q(\phi_k)}](y_i - \mu_{q(\phi_k)})^2\right.\right. \\
&\qquad\qquad \left.\left. - \frac{1}{2\lambda_{q(\phi_k)}} + \Psi(\alpha_{q(\pi),k})\right)\delta_k(c_i)\right\} \\
&\propto \prod_{i=1}^n \text{Categorical}\left(c_i; \omega^*_{i1}, \omega^*_{i2}, \ldots, \omega^*_{iK}\right) \qquad \text{(A.25)}
\end{aligned}
$$

where

$$
\begin{aligned}
\omega^*_{ik} &= \frac{\exp(\nu_{ik})}{\sum_{l=1}^K \exp(\nu_{il})} \\
\nu^*_{ik} &= \frac{1}{2}E[\log(t_{\phi,k})|A_{q(\phi_k)}, C_{q(\phi_k)}] - \frac{1}{2\sigma_u^2}E[t_{\phi,k}|A_{q(\phi_k)}, C_{q(\phi_k)}](y_i - \mu_{q(\phi_k)})^2 \\
&\quad - \frac{1}{2\lambda_{q(\phi_k)}} + \Psi(\alpha_{q(\pi),k})
\end{aligned}
$$

for $i = 1, 2\ldots, n$ and $k = 1, 2\ldots, K$.

- Derivation of $q^*_\pi(\boldsymbol{\pi})$

$$\begin{aligned}
q^*_\pi(\boldsymbol{\pi}) &\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\pi}}\log p(\mathbf{y},\mathbf{x},\mathbf{c},\boldsymbol{\pi},\boldsymbol{\phi})\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\pi}}\sum_{k=1}^{K}\left(\left(\sum_{i=1}^{n}\delta_k(c_i)\right)+\frac{\alpha}{K}-1\right)\log\pi_k\right\} \\
&\propto \prod_{k=1}^{K}\pi_k^{\omega_{\cdot k}+\frac{\alpha}{K}-1} \\
&\propto \text{Dirichlet}\left(\boldsymbol{\pi};\alpha^*_{q(\pi),1},\alpha^*_{q(\pi),2},\ldots,\alpha^*_{q(\pi),k}\right) \qquad \text{(A.26)}
\end{aligned}$$

where $\alpha^*_{q(\pi),k}=\omega_{\cdot k}+\frac{\alpha}{K}$ for $k=1,2,\ldots,K$.

The lower bound on the log marginal likelihood is given by

$$\begin{aligned}
\log\underline{p}(\mathbf{y};q) =&\text{Entropy}\{q_\phi(\boldsymbol{\phi})\}+\text{Entropy}\{q(\mathbf{c})\}+\text{Entropy}\{q(\boldsymbol{\pi})\}+\mathrm{E}\{\log p(\mathbf{y}|\mathbf{c},\boldsymbol{\mu}_\phi,\mathbf{t})\} \\
&+\mathrm{E}\{\log p(\mathbf{c}|\boldsymbol{\pi})\}+\mathrm{E}\{\log p(\boldsymbol{\pi}|\alpha)\}+\mathrm{E}\{\log p(\boldsymbol{\phi}|\mu_0,\lambda_0,a_0,c_0)\} \qquad \text{(A.27)}
\end{aligned}$$

The explicit forms of the terms in (A.27) are given by

$$\begin{aligned}
\mathrm{E}\{\log p(\mathbf{y}|\mathbf{c},\boldsymbol{\mu}_\phi,\mathbf{t})\} =& -\frac{N}{2}\log(2\pi\sigma_u^2)+\frac{1}{2}\sum_{k=1}^{K}\omega_{\cdot k}E[\log(t_{\phi,k})|A_{q(\phi_k)},C_{q(\phi_k)}] \\
&-\frac{E[t_{\phi,k}|A_{q(\phi_k)},C_{q(\phi_k)}]}{2\sigma_u^2}\sum_{i=1}^{n}\sum_{k=1}^{K}\left((y_i-\mu_{q(\phi_k)})^2\omega_{ik}-\frac{\omega_{\cdot k}}{2\lambda_{q(\phi_k)}}\right)
\end{aligned}$$

(A.28)

$$\mathrm{E}\{\log p(\mathbf{c}|\boldsymbol{\pi})\}=\sum_{k=1}^{K}\omega_{\cdot k}\Psi(\alpha_{q(\pi_k)})-n\Psi(\alpha+n) \qquad \text{(A.29)}$$

$$\mathrm{E}\{\log p(\boldsymbol{\pi}|\alpha)\} =\log\Gamma\left(\alpha\right)-K\log\Gamma\left(\frac{\alpha}{K}\right)+\left(K-\alpha\right)\Psi\left(\alpha+n\right)+\sum_{k=1}^{K}\left(\frac{\alpha}{K}-1\right)\Psi(\alpha_{q(\pi_k)})$$

(A.30)

$$\begin{aligned}\mathrm{E}\{\log p(\boldsymbol{\phi}|\mu_0,\lambda_0,a_0,c_0)\} &= -\frac{K}{2}\log(2\pi\sigma_u^2/\lambda_0) + Ka_0\log(c_0) - K\log\Gamma(a_0) - KF_g(1;a_0,c_0)\\ &+ \frac{1}{2}\sum_{k=1}^{K}\left\{E[\log(t_{\phi,k})|A_{q(\phi_k)},C_{q(\phi_k)}] - \frac{\lambda_0}{\sigma_u^2}E[t_{\phi,k}|A_{q(\phi_k)},C_{q(\phi_k)}](\mu_{q(\phi_k)}-\mu_0)^2 - \frac{\lambda_0}{\lambda_{q(\phi_k)}}\right\}\\ &+ (a_0-1)\sum_{k=1}^{K}E[\log(t_{\phi,k})|A_{q(\phi_k)},C_{q(\phi_k)}] - \sum_{k=1}^{K}c_0E[t_{\phi,k}|A_{q(\phi_k)},C_{q(\phi_k)}]\end{aligned} \tag{A.31}$$

$$\begin{aligned}\mathrm{Entropy}\{q(\boldsymbol{\phi})\} = \frac{K}{2}\log(2\pi\sigma_u^2) + \frac{K}{2} + \sum_{k=1}^{K}\Bigg\{&-\frac{1}{2}\log(\lambda_{q(\phi_k)}) + F_g(1;A_{q(\phi_k)},C_{q(\phi_k)})\\ &- A^{*}_{q(\phi_k)}\log(C_{q(\phi_k)}) + \log\Gamma(A_{q(\phi_k)}) + C_{q(\phi_k)}E[t_{\phi,k}|A_{q(\phi_k)},C_{q(\phi_k)}]\\ &- \left(A_{q(\phi_k)} - \frac{1}{2}\right)E[\log(t_{\phi,k})|A_{q(\phi_k)},C_{q(\phi_k)}]\Bigg\}\end{aligned} \tag{A.32}$$

$$\mathrm{Entropy}\{q(\mathbf{c})\} = -\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}\log\omega_{ik} \tag{A.33}$$

$$\mathrm{Entropy}\{q(\boldsymbol{\pi})\} = -\log\Gamma(\alpha+n) + \sum_{k=1}^{K}\big(\log\Gamma(\alpha_{q(\pi_k)}) - (\alpha_{q(\pi_k)}-1)(\Psi(\alpha_{q(\pi_k)}) - \Psi(\alpha+n))\big) \tag{A.34}$$

Gaussian or Laplacian measurement error with unknown variance.

- Derivation for $q^*_{\sigma_u^2}(\sigma_u^2)$ if measurement error is Gaussian.

$$\begin{aligned}
&q^*_{\sigma_u^2}(\sigma_u^2)\\
&\propto \exp\left\{\mathrm{E}_{-\sigma_u^2}\log p(\mathbf{y},\mathbf{x},\mathbf{c},\boldsymbol{\pi},\boldsymbol{\phi},\sigma_u^2)\right\}\\
&\propto \exp\left\{\mathrm{E}_{-\sigma_u^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(-\frac{1}{2}\log(2\pi\sigma_u^2)-\frac{(y_{ij}-x_i)^2}{2\sigma_u^2}\right)-\frac{\beta_{\sigma_u^2}}{\sigma_u^2}-(\gamma_{\sigma_u^2}+1)\log\sigma_u^2\right)\right\}\\
&\propto \exp\left\{-\left(\gamma_{\sigma_u^2}+\frac{N}{2}+1\right)\log(\sigma_u^2)\right.\\
&\quad\left.-\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}^2-2\mu_{q(x_i)}y_{ij}+\mu^2_{q(x_i)}+\sigma^2_{q(x_i)})+\beta_{\sigma_u^2}\right)\frac{1}{\sigma_u^2}\right\}\\
&\propto \mathcal{IG}(\sigma_u^2;A^*_{q(\sigma_u^2)},B^*_{q(\sigma_u^2)}) \qquad\qquad (A.35)
\end{aligned}$$

where

$$\begin{aligned}
A^*_{q(\sigma_u^2)} &= \gamma_{\sigma_u^2}+\frac{N}{2}\\
B^*_{q(\sigma_u^2)} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}^2-2\mu_{q(x_i)}y_{ij}+\mu^2_{q(x_i)}+\sigma^2_{q(x_i)})+\beta_{\sigma_u^2}
\end{aligned}$$

- Derivation for $q^*_b(b)$ if measurement error is Laplacian.

$$\begin{aligned}
&q^*_b(b)\\
&\propto \exp\left\{\mathrm{E}_{-b}\log p(\mathbf{y},\mathbf{x},\mathbf{c},\boldsymbol{\pi},\boldsymbol{\phi},b)\right\}\\
&\propto \exp\left\{\mathrm{E}_{-b}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(-\log(2b)-\frac{|y_{ij}-x_i|}{b}\right)-\frac{\beta_b}{b}-(\gamma_b+1)\log b\right)\right\}\\
&\propto \exp\left\{-(\gamma_b+n+1)\log(b)-\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sigma_{q(x_i)}\left(z_{ij}\left(2\Phi(z_{ij})-1\right)+2\phi(z_{ij})\right)+\beta_b\right)\frac{1}{b}\right\}\\
&\propto \mathcal{IG}(b;A^*_{q(b)},B^*_{q(b)}) \qquad\qquad (A.36)
\end{aligned}$$

where $z_{ij} = \frac{y_{ij} - \mu_{q(x_i)}}{\sigma_{q(x_i)}}$,

$$
\begin{aligned}
A^*_{q(b)} &= \gamma_b + N \\
B^*_{q(b)} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} \sigma_{q(x_i)} \left(z_{ij} \left(2\Phi(z_{ij}) - 1 \right) + 2\phi(z_{ij}) \right) + \beta_b
\end{aligned}
$$

Nonparametric measurement error

According to the model specification in section 4.1, the joint distribution of $(\mathbf{y}, \mathbf{x}, \mathbf{c}_x, \boldsymbol{\pi}_x, \boldsymbol{\phi}_x, \mathbf{c}_u, \boldsymbol{\pi}_u, \boldsymbol{\zeta}, \boldsymbol{\sigma}^2_u)$ is

$$
\begin{aligned}
&\log p(\mathbf{y}, \mathbf{x}, \mathbf{c}_x, \boldsymbol{\pi}_x, \boldsymbol{\phi}_x, \mathbf{c}_u, \boldsymbol{\pi}_u, \boldsymbol{\zeta}, \boldsymbol{\sigma}^2_u | \alpha_x, \mu_{x,0}, \lambda_{x,0}, \gamma_{x,0}, \beta_{x,0}, \alpha_u, \gamma_{u,0}, \beta_{u,0}, \sigma^2_\zeta) \\
&= \log p(\mathbf{y}|\mathbf{x}, \mathbf{c}_u, \boldsymbol{\pi}_u, \boldsymbol{\zeta}, \boldsymbol{\sigma}^2_u) + \log p(\mathbf{x}|\mathbf{c}_x, \boldsymbol{\phi}_x) + \log p(\mathbf{c}_x|\boldsymbol{\pi}_x) + \log p(\boldsymbol{\pi}_x|\alpha_x) \\
&\quad + \log p(\boldsymbol{\phi}_x|\mu_{x,0}, \lambda_{x,0}, \gamma_{x,0}, \beta_{x,0}) + \log p(\mathbf{c}_u|\boldsymbol{\pi}_u) + \log p(\boldsymbol{\pi}_u|\alpha_u) + \log p(\boldsymbol{\zeta}|\sigma^2_\zeta) \\
&\quad + \log p(\boldsymbol{\sigma}^2_u|\gamma_{u,0}, \beta_{u,0}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{K_u} \left\{ -\frac{1}{2}\log(2\pi\sigma^2_{u,k}) - \frac{(y_{ij} - x_i - \pi^{-1}_{u,c_{u_{ij}}} l_k^T \boldsymbol{\zeta})^2}{2\sigma^2_{u,k}} \right\} \delta_k(c_{u_{ij}}) \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{K_x} \left\{ -\frac{1}{2}\log(2\pi\sigma^2_{x,k}) - \frac{(x_i - \mu_{x,k})^2}{2\sigma^2_{x,k}} \right\} \delta_k(c_{x_i}) + \sum_{i=1}^{n}\sum_{k=1}^{K_x} \delta_k(c_{x_i}) \log(\pi_{x,k}) \\
&\quad + \log\Gamma(\alpha_x) - K_x \log\left(\frac{\alpha_x}{K_x} \right) + \left(\frac{\alpha_x}{K_x} - 1 \right) \sum_{k=1}^{K_x} \log\pi_{x,k} + K_x \gamma_{x,0} \log\beta_{x,0} - K_x \log\Gamma(\gamma_{x,0}) \\
&\quad + \sum_{k=1}^{K_x} \left\{ -(\gamma_{x,0} + 1)\log\sigma^2_{x,k} - \frac{\beta_{x,0}}{\sigma^2_{x,k}} - \frac{1}{2}\log\left(\frac{2\pi\sigma^2_{x,k}}{\lambda_{x,0}} \right) - \frac{\lambda_{x,0}(\mu_{x,k} - \mu_{x,0})^2}{2\sigma^2_{x,k}} \right\} \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{K_u} \delta_k(c_{u_{ij}}) \log(\pi_{u,k}) + \log\Gamma(\alpha_u) - K_u \log\Gamma(\frac{\alpha_u}{K_u}) \\
&\quad + \left(\frac{\alpha_u}{K_u} - 1 \right) \sum_{k=1}^{K_u} \log(\pi_{u,k}) + K_u \gamma_{u,0} \log\beta_{u,0} - K_u \log\Gamma(\lambda_{u,0}) - \frac{K_u}{2}\log(2\pi\sigma^2_\zeta) \\
&- \sum_{k=1}^{K_u} \frac{\zeta_k^2}{2\sigma^2_\zeta} - \sum_{k=1}^{K_u} \left\{ \frac{\beta_{u,0}}{\sigma^2_{u,k}} + (\gamma_{u,0} + 1)\log\sigma^2_{u,k} \right\}
\end{aligned} \tag{A.37}
$$

The optimal densitiy $q^*_{\theta_j}(\theta_j) \propto \exp\{\mathrm{E}_{-\theta_j}\log p(\mathbf{y},\boldsymbol{\theta})\}$ for $\theta_j \in \boldsymbol{\theta}\backslash\boldsymbol{\pi}_u$ can be derived as follows.

- Derivation of $q^*_{\mathbf{x}}(\mathbf{x})$

$$\begin{aligned}
q_{\mathbf{x}}(\mathbf{x}) &\propto \exp\left\{\mathrm{E}_{-x}\log p(\mathbf{y},\boldsymbol{\theta})\right\}\\
&\propto \exp\left\{\mathrm{E}_{-x}\sum_{i=1}^{n}\left\{\sum_{j=1}^{m_i}\sum_{k=1}^{K_u}-\frac{(y_{ij}-x_i-\pi_{u,k}^{-1}l_k^T\boldsymbol{\zeta})^2}{2\sigma^2_{u,k}}\delta_k(c_{u_{ij}})\right.\right.\\
&\left.\left.+\sum_{k=1}^{K_x}-\frac{(x_i-\mu_{x,k})^2}{2\sigma^2_{x,k}}\delta_k(c_{x_i})\right\}\right\}\\
&\propto \exp\left\{\sum_{i=1}^{n}\left\{\sum_{j=1}^{m_i}\sum_{k=1}^{K_u}-\frac{A_{q(\sigma^2_{u,k})}}{2B_{q(\sigma^2_{u.k})}}\rho_{ij,k}\left(x_i^2-2x_i\left(y_{ij}-\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}l_k^T\boldsymbol{\mu}_\zeta\right)\right)\right.\right.\\
&\qquad\left.\left.\sum_{k=1}^{K_x}-\frac{A_{q(\phi_{x,k})}}{2B_{q(\phi_{x,k})}}(x_i^2-2x_i\mu_{q(\phi_{x,k})})\omega_{i,k}\right\}\right\}\\
&\propto \prod_{i=1}^{n}\mathcal{N}(x_i;\mu_{q(x_i)},\sigma^2_{q(x_i)}) \qquad\qquad (\mathrm{A.38})
\end{aligned}$$

where

$$\sigma^2_{q(x_i)} = \left\{\sum_{k=1}^{K_u}\frac{A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u.k})}}\rho_{i\cdot,k}+\sum_{k=1}^{K_x}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\omega_{ik}\right\}^{-1}$$

$$\mu_{q(x_i)} = \sigma^2_{q(x_i)}\left\{\sum_{k=1}^{K_u}\frac{A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u.k})}}\sum_{j=1}^{m_i}\rho_{ij,k}\left(y_{ij}-\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}l_k^T\boldsymbol{\mu}_\zeta\right)+\sum_{k=1}^{K_x}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\omega_{ik}\mu_{q(\phi_{x,i})}\right\}$$

- Derivation of $q^*_{\mathbf{c}_x}(\mathbf{c}_x)$

$$\begin{aligned}
q^*_{\mathbf{c}_x}(\mathbf{c}_x) &\propto \exp\left\{\mathrm{E}_{-\mathbf{c}_x}\log p(\mathbf{y},\boldsymbol{\theta})\right\}\\
&\propto \exp\left\{\mathrm{E}_{-\mathbf{c_x}}\sum_{i=1}^{n}\sum_{k=1}^{K_x}\left(-\frac{1}{2}\log\sigma^2_{x,k}-\frac{1}{2}\frac{(x_i-\mu_{x,k})^2}{\sigma^2_{x,k}}+\log\pi_{x,k}\right)\delta_k(c_{x_i})\right\}\\
&\propto \prod_{i=1}^{n}\exp\left\{\sum_{k=1}^{K_x}\left(-\frac{1}{2}\log B_{q(\phi_{x,k})}+\frac{1}{2}\Psi(A_{q(\phi_{x,k})})\right.\right.\\
&\left.\left.-\frac{1}{2}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\left((\mu_{q(x_i)}-\mu_{q(\phi_{x,k})})^2+\sigma^2_{q(x_i)}\right)-\frac{1}{2\lambda_{q(\phi_{x,k})}}+\Psi(\alpha_{q(\pi_{x,k})})\right)\delta_k(c_{x_i})\right\}\\
&\propto \prod_{i=1}^{n}\mathrm{Categorical}\left(c_i;\omega^*_{i1},\omega^*_{i2},\ldots,\omega^*_{iK_x}\right) \qquad\qquad (\mathrm{A.39})
\end{aligned}$$

where

$$
\begin{aligned}
\omega_{ik}^{*} &= \frac{\exp(\nu_{ik})}{\sum_{l=1}^{K_x}\exp(\nu_{il})} \\
\nu_{ik} &= -\frac{1}{2}\log B_{q(\phi_{x,k})} + \frac{1}{2}\Psi(A_{q(\phi_{x,k}^2)}) - \frac{1}{2}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\left((\mu_{q(x_i)} - \mu_{q(\phi_{x,k})})^2 + \sigma_{q(x_i)}^2\right) \\
&\quad - \frac{1}{2\lambda_{q(\phi_{x,k})}} + \Psi(\alpha_{q(\pi_{x,k})})
\end{aligned}
$$

- Derivation of $q_{\pi_x}^{*}(\boldsymbol{\pi}_x)$

$$
\begin{aligned}
q_{\pi_x}^{*}(\boldsymbol{\pi}_x) &\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\pi}_x}\log p(\mathbf{y},\boldsymbol{\theta})\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\pi}_x}\sum_{k=1}^{K_x}\left(\left(\sum_{i=1}^{n}\delta_k(c_{x_i})\right) + \frac{\alpha_x}{K_x} - 1\right)\log\pi_{x,k}\right\} \\
&\propto \prod_{k=1}^{K_x}\pi_{x,k}^{\omega_{\cdot k}+\frac{\alpha_x}{K_x}-1} \\
&\propto \text{Dirichlet}\left(\boldsymbol{\pi}_x; \alpha_{q(\pi_{x,1})}^{*}, \alpha_{q(\pi_{x,2})}^{*}, \ldots, \alpha_{q(\pi_{x,k})}^{*}\right) \qquad \text{(A.40)}
\end{aligned}
$$

where $\alpha_{q(\pi_{x,k})}^{*} = \omega_{\cdot k} + \frac{\alpha_x}{K_x}$ for $k = 1, 2, \ldots, K_x$.

- Derivation of $q^*_{\phi_x}(\boldsymbol{\phi}_x)$

$$\begin{aligned}
&q^*_{\phi_x}(\boldsymbol{\phi}_x) \propto \exp\left\{\mathrm{E}_{-\phi_x}\log p(\mathbf{y},\boldsymbol{\theta})\right\}\\
&\propto \exp\Bigg\{\mathrm{E}_{-\phi_x}\sum_{k=1}^{K_x}\Bigg(\sum_{i=1}^{n}\Bigg(-\frac{1}{2}\log\sigma^2_{x,k}-\frac{(x_i-\mu_{x,k})^2}{2\sigma^2_{x,k}}\Bigg)\delta_k(c_{x_i})\\
&-\frac{1}{2}\log\sigma^2_{x,k}-\frac{\lambda_{x,0}(\mu_{x,k}-\mu_{x,0})^2}{2\sigma^2_{x,k}}-\frac{\beta_{x,0}}{\sigma^2_{x,k}}-(\gamma_{x,0}+1)\log\sigma^2_{x,k}\Bigg)\Bigg\}\\
&\propto \exp\Bigg\{\sum_{k=1}^{K_x}-\frac{\omega_{\cdot k}+\lambda_{x,0}}{2\sigma^2_{x,k}}\left(\mu_{x,k}-\frac{\sum_{i=1}^{n}\mu_{q(x_i)}\omega_{ik}+\lambda_{x,0}\mu_{x,0}}{\omega_{\cdot k+\lambda_{x,0}}}\right)^2\\
&-\left(\frac{\omega_{\cdot k}+1}{2}+\gamma_{x,0}+1\right)\log\sigma^2_{x,k}\\
&-\left(\beta_{x,0}+\frac{1}{2}\sum_{i=1}^{n}\omega_{ik}\left(\mu^2_{q(x_i)}+\sigma^2_{q(x_i)}\right)+\frac{1}{2}\lambda_{x,0}\mu^2_{x,0}-\frac{1}{2}\left(\omega_{\cdot k}+\lambda_0\right)\mu^2_{q(\phi_k)}\right)\frac{1}{\sigma^2_k}\Bigg\}\\
&\propto \prod_{k=1}^{K_x}\mathcal{NIG}(\phi_{x,k};\mu^*_{q(\phi_{x,k})},\lambda^*_{q(\phi_{x,k})},A^*_{q(\phi_{x,k})},B^*_{q(\phi_{x,k})}) \qquad \text{(A.41)}
\end{aligned}$$

where

$$\begin{aligned}
\mu^*_{q(\phi_{x,k})} &= \frac{\sum_{i=1}^{n}\mu_{q(x_i)}\omega_{ik}+\lambda_{x,0}\mu_{x,0}}{\omega_{\cdot k+\lambda_{x,0}}}\\
\lambda^*_{q(\phi_{x,k})} &= \omega_{\cdot k}+\lambda_{x,0}\\
A^*_{q(\phi_{x,k})} &= \frac{\omega_{\cdot k}}{2}+\gamma_{x,0}\\
B^*_{q(\phi_{x,k})} &= \beta_{x,0}+\frac{1}{2}\sum_{i=1}^{n}\omega_{ik}\left(\mu^2_{q(x_i)}+\sigma^2_{q(x_i)}\right)+\frac{1}{2}\lambda_{x,0}\mu^2_{x,0}-\frac{1}{2}\left(\omega_{\cdot k}+\lambda_{x,0}\right)\mu^2_{q(\phi_{x,k})}
\end{aligned}$$

- Derivation of $q^*_{\mathbf{c}_u}(\mathbf{c}_u)$

$$
\begin{aligned}
&q^*_{\mathbf{c}_u}(\mathbf{c}_u) \propto \exp\left\{\mathrm{E}_{-\mathbf{c}_u} \log p(\mathbf{y}, \boldsymbol{\theta})\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\mathbf{c}_u} \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{K_u}\left(-\frac{1}{2}\log\sigma^2_{u,k} - \frac{1}{2}\frac{(y_{ij} - x_i - \pi^{-1}_{u,k} l_k^T \boldsymbol{\zeta})^2}{2\sigma^2_{u,k}} + \log\pi_{u,k}\right)\delta_k(c_{u_{ij}})\right\} \\
&\propto \prod_{i=1}^{n}\prod_{j=1}^{n_j}\exp\Bigg\{\sum_{k=1}^{K_u}\Bigg(\Psi(\alpha_{q(\pi_{u,k})}) - \frac{1}{2}\log B_{q(\sigma^2_{u,k})} + \frac{1}{2}\Psi(A_{q(\sigma^2_{u,k})}) \\
&-\frac{1}{2}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}}\Bigg(y_{ij}^2 - 2y_{ij}\mu_{q(x_i)} + \mu^2_{q(x_i)} + \sigma^2_{q(x_i)} - 2(y_{ij} - \mu_{q(x_i)})\frac{\tilde{\alpha}_{q(\pi_u)} - 1}{\alpha_{q(\pi_{u,k})} - 1} l_k^T \boldsymbol{\mu}_\zeta \\
&+\frac{(\tilde{\alpha}_{q(\pi_u)} - 1)(\tilde{\alpha}_{q(\pi_u)} - 2)}{(\alpha_{q(\pi_{u,k})} - 1)(\alpha_{q(\pi_{u,k})} - 2)} l_k^T(\boldsymbol{\mu}_\zeta \boldsymbol{\mu}_\zeta^T + \Sigma_\zeta) l_k\Bigg)\Bigg)\delta_k(c_{u_{ij}})\Bigg\} \\
&\propto \prod_{i=1}^{n}\prod_{j=1}^{m_i} \text{Categorical}\left(c_{u_{ij}}; \rho^*_{ij,1}, \rho^*_{ij,2}, \ldots, \rho^*_{ij,K_u}\right) \qquad \text{(A.42)}
\end{aligned}
$$

$$
\begin{aligned}
\rho^*_{ik} &= \frac{\exp(\tau_{ij,k})}{\sum_{l=1}^{K}\exp(\tau_{ij,l})} \\
\tau_{ij,k} &= \Psi(\alpha_{q(\pi_{u,k})}) - \frac{1}{2}\log B_{q(\sigma^2_{u,k})} + \frac{1}{2}\Psi(A_{q(\sigma^2_{u,k})}) \\
&\quad - \frac{1}{2}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}}\Bigg(y_{ij}^2 - 2y_{ij}\mu_{q(x_i)} + \mu^2_{q(x_i)} + \sigma^2_{q(x_i)} \\
&\quad -2(y_{ij} - \mu_{q(x_i)})\frac{\tilde{\alpha}_{q(\pi_u)} - 1}{\alpha_{q(\pi_{u,k})} - 1} l_k^T \boldsymbol{\mu}_\zeta + \frac{(\tilde{\alpha}_{q(\pi_u)} - 1)(\tilde{\alpha}_{q(\pi_u)} - 2)}{(\alpha_{q(\pi_{u,k})} - 1)(\alpha_{q(\pi_{u,k})} - 2)} l_k^T(\boldsymbol{\mu}_\zeta \boldsymbol{\mu}_\zeta^T + \Sigma_\zeta) l_k\Bigg)
\end{aligned}
$$

- Derivation of $q^*_\zeta(\boldsymbol{\zeta})$

$$\begin{aligned}
q^*_\zeta(\boldsymbol{\zeta}) &\propto \exp\Big\{\mathrm{E}_{-\boldsymbol{\zeta}} \log p(\mathbf{y},\boldsymbol{\theta})\Big\} \\
&\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\zeta}} \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{K_u} -\frac{(y_{ij}-x_i-\pi_{u,k}^{-1} l_k^T\boldsymbol{\zeta})^2}{2\sigma^2_{u,k}}\delta_k(c_{u_{ij}}) - \frac{1}{2}\sum_{k=1}^{K_u-1}\frac{\zeta_k^2}{\sigma^2_\zeta}\right\} \\
&\propto \exp\left\{\mathrm{E}_{-\boldsymbol{\zeta}} - \frac{1}{2}\boldsymbol{\zeta}^T\left(\sum_{k=1}^{K_u}\frac{\rho_{\cdot\cdot,k} l_k l_k^T}{\pi^2_{u,k}\sigma^2_{u,k}} + \frac{I_{K_u-1}}{\sigma^2_\zeta}\right)\boldsymbol{\zeta} + \sum_{k=1}^{K_u}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\mu_{q(x_i)})\rho_{ij,k}\right)\frac{\pi^{-1}_{u,k}}{\sigma^2_{u,k}} l_k^T\boldsymbol{\zeta}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\zeta}^T\left(\sum_{k=1}^{K_u}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}}\rho_{\cdot\cdot,k} l_k l_k^T + \frac{I_{K_u-1}}{\sigma^2_\zeta}\right)\boldsymbol{\zeta}\right. \\
&\qquad\left. + \sum_{k=1}^{K_u}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\mu_{q(x_i)})\rho_{ij,k}\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}} l_k^T\right)\boldsymbol{\zeta}\right\} \\
&\propto \mathcal{MVN}(\boldsymbol{\zeta};\, \boldsymbol{\mu}^*_\zeta, \Sigma^*_\zeta) \qquad \text{(A.43)}
\end{aligned}$$

where

$$\Sigma^*_\zeta = \left(\sum_{k=1}^{K_u}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}}\rho_{\cdot\cdot,k} l_k l_k^T + \frac{I_{K_u-1}}{\sigma^2_\zeta}\right)^{-1}$$

$$\boldsymbol{\mu}^*_\zeta = \Sigma^*_\zeta\left\{\sum_{k=1}^{K_u}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\mu_{q(x_i)})\rho_{ij,k}\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}\frac{A_{q(\sigma_{u,k})}}{B_{q(\sigma_{u,k})}} l_k\right\}$$

- Derivation of $q^*_{\sigma^2_u}(\boldsymbol{\sigma}^2_u)$

$$\begin{aligned}
q^*_{\sigma^2_u}(\boldsymbol{\sigma}^2_u) &\propto \exp\left\{\mathrm{E}_{-\sigma^2_u}\log p(\mathbf{y},\boldsymbol{\theta})\right\}\\
&\propto \exp\left\{\mathrm{E}_{-\sigma^2_{u,k}}\sum_{k=1}^{K_u}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(-\frac{1}{2}\log\sigma^2_{u,k}-\frac{(y_{ij}-x_i-\pi^{-1}_{u,k}l_k^T\boldsymbol{\zeta})^2}{2\sigma^2_{x,k}}\right)\delta_k(c_{x_i})\right.\\
&\qquad\left.-\frac{\beta_{u,0}}{\sigma^2_{u,k}}-(\gamma_{u,0}+1)\log\sigma^2_{u,k}\right\}\\
&\propto \prod_{k=1}^{K_u}\exp\left\{-\left(\frac{\rho_{\cdot\cdot,k}}{2}+\gamma_{u,0}+1\right)\log\sigma^2_{u,k}-\frac{1}{\sigma^2_{u,k}}\left(\beta_{u,0}\right.\right.\\
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(y^2_{ij}-2y_{ij}\mu_{q(x_i)}+\mu^2_{q(x_i)}+\sigma^2_{q(x_i)}-2\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}l_k^T\mu_\zeta(y_{ij}-\mu_{q(x_i)})\right)\rho_{ij,k}\\
&\left.\left.+\frac{1}{2}\rho_{\cdot\cdot,k}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)}l_k^T(\boldsymbol{\mu}_\zeta\boldsymbol{\mu}_\zeta^T+\Sigma_\zeta)l_k\right)\right\}\\
&\propto \prod_{k=1}^{K_u}\mathcal{IG}(\sigma^2_{u,k};\ A^*_{q(\sigma^2_{u,k})},B^*_{q(\sigma^2_{u,k})})
\end{aligned}\tag{A.44}$$

where

$$\begin{aligned}
A^*_{q(\sigma^2_{u,k})} &= \frac{\rho_{\cdot\cdot,k}}{2}+\gamma_{u,0}\\
B^*_{q(\sigma^2_{u,k})} &= \beta_{u,0}+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(y^2_{ij}-2y_{ij}\mu_{q(x_i)}+\mu^2_{q(x_i)}+\sigma^2_{q(x_i)}\right.\\
&\qquad\left.-2\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}l_k^T\mu_\zeta(y_{ij}-\mu_{q(x_i)})\right)\rho_{ij,k}\\
&\qquad+\frac{1}{2}\rho_{\cdot\cdot,k}\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)}l_k^T(\boldsymbol{\mu}_\zeta\boldsymbol{\mu}_\zeta^T+\Sigma_\zeta)l_k
\end{aligned}$$

APPENDIX B

CONVERGENCE OF VARIATIONAL APPROXIMATION ALGORITHMS

This chapter proves that the variational approximation algorithms developed in this thesis converge to local optima.

Convergence of the variational approximation algorithms for Gaussian and Laplacian error distributions

In this section we show that the variational approximation method for Laplacian error with unknown variance is convergent. The proofs of the convergence results of variational approximation algorithms A and B for Gaussian error are similar to the one we give here.

The objective function is

$$
\begin{aligned}
&\log\underline{p}(\mathbf{y};q) = \text{Entropy}\{q(\mathbf{x})\} + \text{Entropy}\{q(\boldsymbol{\phi})\} + \text{Entropy}\{q(\mathbf{c})\} + \text{Entropy}\{q(\boldsymbol{\pi})\} \\
&\ + \text{Entropy}\{q(b)\} + \text{E}\{\log f_u(\mathbf{y}-\mathbf{x}|b)\} + \text{E}\{\log p(\mathbf{x}|\mathbf{c},\boldsymbol{\phi})\} + \text{E}\{\log p(\mathbf{c}|\boldsymbol{\pi})\} + \text{E}\{\log p(\boldsymbol{\pi}|\alpha)\} \\
&\ + \text{E}\{\log p(\boldsymbol{\phi}|\mu_0,\lambda_0,\gamma_0,\beta_0)\} \\
&= -\frac{\log(2)N}{2} + \frac{N}{2} + \frac{K}{2} - (N+\gamma_b+1)\left(\log(B_{q(b)}) - \psi(A_{q(b)})\right) \\
&\ - \frac{A_{q(b)}}{B_{q(b)}}\left\{\beta_0 + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sigma_{q(x_i)}z_{ij}\left(2\Phi(z_{ij})-1\right) + 2\phi(z_{ij})\right\} \\
&\ + \sum_{k=1}^{K}\left\{-\frac{1}{2}\sum_{i=1}^{n}\omega_{ik}\left(\sigma^2_{q(x_i)} + (\mu_{q(x_i)}-\mu_{q(\phi_k)})^2\right) - \frac{\lambda_0}{2}(\mu_{q(\phi_k)}-\mu_0)^2 - \beta_0 + B_{q(\phi_k)}\right\}\frac{A_{q(\phi_k)}}{B_{q(\phi_k)}} \\
&\ + \sum_{k=1}^{K}\left\{\left(-\frac{\omega_{\cdot k}}{2} - \gamma_0\right)\log(B_{q(\phi_k)}) + \left(\frac{\omega_{\cdot k}}{2} + \gamma_0 - A_{q(\phi_k)}\right)\Psi(A_{q(\phi_k)})\right\} \\
&\ + \sum_{k=1}^{K}\left\{\left(\omega_{\cdot k} + \frac{\alpha}{K} - \alpha_{q(\pi_k)}\right)\Psi(\alpha_{q(\pi_k)}) + \log\Gamma(A_{q(\phi_k)}) + \log\Gamma(\alpha_{q(\pi_k)}) - \frac{\omega_{\cdot k}+\lambda_0}{2\lambda_{q(\phi_k)}} - \frac{1}{2}\log(\lambda_{q(\phi_k)})\right\} \\
&\ + \frac{1}{2}\sum_{i=1}^{n}\log\sigma^2_{q(x_i)} - \sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}\log\omega_{ik} + A_{q(b)} + \log(B_{q(b)}) + \log\Gamma(A_{q(b)}) - (1+A_{q(b)})\psi(A_{q(b)}) \\
&\ - K\log\Gamma\left(\frac{\alpha}{K}\right) + \log\Gamma(\alpha) + \frac{K}{2}\log\lambda_0 + K\gamma_0\log\beta_0 - K\log\Gamma(\gamma_0) - \log\Gamma(\alpha+n) \\
&\ + \gamma_b\log\beta_b - \log\Gamma(\gamma_b)
\end{aligned}
\tag{B.1}
$$

subject to the constraint $\sum_{k=1}^{K}\omega_{ik} = 1$ for $i = 1, \cdots, n$. The parameters updated in each iteration are $\boldsymbol{\xi} = \left(\left\{(\mu_{q(x_i)}, \sigma^2_{q(x_i)})\right\}_{i=1}^{n}, \left\{(\mu_{q(\phi_k)}, \lambda_{q(\phi_k)}, A_{q(\phi_k)}, B_{q(\phi_k)})\right\}_{k=1}^{K}, \{\omega_{ik}\}_{i=1,\ k=1}^{n,\ K},\right.$ $\left.\left\{\alpha_{q(\pi_k)}\right\}_{k=1}^{K}, (A_{q(b)}, B_{q(b)})\right)$, if we divide them into N_B=3 blocks $\xi_1 = \left\{(\mu_{q(x_i)}, \sigma^2_{q(x_i)})\right\}_{i=1}^{n}$, $\xi_2 = \{\omega_{ik}\}_{i=1,\ k=1}^{n,\ K}$, $\xi_3 = \left(\left\{(\mu_{q(\phi_k)}, \lambda_{q(\phi_k)}, A_{q(\phi_k)}, B_{q(\phi_k)})\right\}_{k=1}^{K}, \left\{\alpha_{q(\pi_k)}\right\}_{k=1}^{K}, (A_{q(b)}, B_{q(b)})\right)$, then parameters in the same block can be updated simultaneously in each iteration. Our maximization methods can be rewritten as the following iteration:

$$
\xi_i^{t+1} = arg \max_{\tau\in\text{domain}(\xi_i)} \log\underline{p}(\mathbf{y}; q, \xi_1^{t+1}, \cdots, \xi_{i-1}^{t+1}, \tau, \xi_{i+1}^{t}, \cdots, \xi_{N_B}^{t}) \tag{B.2}
$$

[Grippo and Sciandrone, 2000] shows that the sequence $\left\{(\xi_1^t, \cdots, \xi_{N_B}^t)\right\}$ generated by the method is convergent to a local maxima if the objective function is componentwise strictly quasiconcave with respect to $N_B - 2$ components. To prove the convergence results for our variational approximation method, we only need to show that $\log\underline{p}(\mathbf{y}; q)$ is strictly quasiconcave with respect to ξ_2. With the constraint $\sum_{k=1}^{K} \omega_{ik} = 1$ for $i = 1, \cdots, n$, there are $n(K-1)$ free parameters in ξ_2, denoted by $\tilde{\xi}_2 = \{\omega_{ik}\}_{i=1,\ k=1}^{n,\ K-1}$. Since $\frac{\partial^2 \log\underline{p}(\mathbf{y};q)}{\partial\omega_{i_1k_1}\partial\omega_{i_2k_2}} = 0$ for $i_1 \neq i_2$, the objection function $\log\underline{p}(\mathbf{y}; q)$ is strictly concave with respect to $\tilde{\xi}_2$ if and only if it is strictly concave with respect to $\tilde{\xi}_{2(i)}$, where $\tilde{\xi}_{2(i)} = \{\omega_{ik}\}_{k=1}^{K-1}$. Next we show that $\frac{\partial^2 \log\underline{p}(\mathbf{y};q)}{\partial\tilde{\xi}_{2(i)}^2}$ is a negative definite matrix. Noting that $\frac{\partial^2 \log\underline{p}(\mathbf{y};q)}{\partial\omega_{ik}^2} = -\frac{1}{\omega_{ik}} - \frac{1}{1-\sum_{k=1}^{K-1}\omega_{ik}}$, $\frac{\partial^2 \log\underline{p}(\mathbf{y};q)}{\partial\omega_{ik}\partial\omega_{il}} = -\frac{1}{1-\sum_{k=1}^{K-1}\omega_{ik}}$, for $1 \leq k, l \leq K-1$ and $k \neq l$, so $\frac{\partial^2 \log\underline{p}(\mathbf{y};q)}{\partial\tilde{\xi}_{2(i)}^2} = -\text{diag}(\omega_{i1}, \cdots, \omega_{i,K-1}) - \frac{1}{1-\sum_{k=1}^{K-1}\omega_{ik}} J_{K-1}$, where J_{K-1} is a $(K-1) \times (K-1)$ matrix of ones. For any nonzero vector $\mathbf{v}$ with length $K-1$, we have $\mathbf{v}^T \frac{\partial^2 \log\underline{p}(\mathbf{y};q)}{\partial\tilde{\xi}_{2(i)}^2} \mathbf{v} = -\sum_{k=1}^{K-1} \omega_{ik} v_k^2 - \frac{1}{1-\sum_{k=1}^{K-1}\omega_{ik}} (\sum_{k=1}^{K} v_k)^2 < 0$ when $\omega_{ik} > 0$ for all $k = 1, \cdots, K$, which indicates that $\frac{\partial^2 \log\underline{p}(\mathbf{y};q)}{\partial\tilde{\xi}_{2(i)}^2}$ is negative definite.

Convergence of the variational approximation algorithms for nonparametric error distributions

The objective function is

$$\begin{aligned}
\log\underline{p}(\mathbf{y}; q) = {} & \text{Entropy}\{q(\mathbf{x})\} + \text{Entropy}\{q(\boldsymbol{\phi}_x)\} + \text{Entropy}\{q(\mathbf{c}_x)\} + \text{Entropy}\{q(\boldsymbol{\pi}_x)\} \\
& + \text{Entropy}\{q(\boldsymbol{\zeta})\} + \text{Entropy}\{q(\boldsymbol{\sigma}_\zeta^2)\} + \text{Entropy}\{q(\mathbf{c}_u)\} + \text{Entropy}\{q(\boldsymbol{\pi}_u)\} \\
& + \text{E}\{\log p(\mathbf{y}|\mathbf{x}, \mathbf{c}_u, \boldsymbol{\pi}_u, \boldsymbol{\zeta}, \boldsymbol{\sigma}_u^2)\} + \text{E}\{\log p(\mathbf{x}|\mathbf{c}_x, \boldsymbol{\phi}_x)\} + \text{E}\{\log p(\boldsymbol{\phi}_x|\mu_{x,0}, \lambda_{x,0}, \gamma_{x,0}, \beta_{x,0})\} \\
& + \text{E}\{\log p(\mathbf{c}_x|\boldsymbol{\pi}_x)\} + \text{E}\{\log p(\boldsymbol{\pi}_x|\alpha_x)\} + \text{E}\{\log p(\mathbf{c}_u|\boldsymbol{\pi}_u)\} + \text{E}\{\log p(\boldsymbol{\pi}_u|\alpha_u)\}
\end{aligned}$$

$$
\begin{aligned}
&+\mathrm{E}\{\log p(\boldsymbol{\zeta}|\sigma_\zeta^2)\}+\mathrm{E}\{\log p(\boldsymbol{\sigma}_u^2|\gamma_{u,0},\beta_{u,0})\}\\
&=-\frac{N}{2}\log(2\pi)+\frac{n+K_x+K_u-1}{2}+\sum_{k=1}^{K_x}\left(-\frac{\omega_{\cdot k}}{2}-\gamma_{x,0}\right)\log(B_{q(\phi_{x,k})})\\
&+\sum_{k=1}^{K_u}\left(-\frac{\rho_{\cdot\cdot k}}{2}-\gamma_{u,0}\right)\log(B_{q(\sigma^2_{u,k})})+\sum_{k=1}^{K_x}\left(\frac{\omega_{\cdot k}}{2}+\gamma_{x,0}-A_{q(\phi_{x,k})}\right)\Psi(A_{q(\phi_{x,k})})\\
&+\sum_{k=1}^{K_u}\left(\frac{\rho_{\cdot\cdot k}}{2}+\gamma_{u,0}-A_{q(\sigma^2_{u,k})}\right)\Psi(A_{q(\sigma^2_{u,k})})+\sum_{k=1}^{K_u}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u,k})}}\Bigg\{B_{q(\sigma^2_{u,k})}-\beta_{u,0}\\
&-\frac{\rho_{ij,k}}{2}\Bigg(y_{ij}^2-2y_{ij}\mu_{q(x_i)}+\mu^2_{q(x_i)}+\sigma^2_{q(x_i)}-2(y_{ij}-\mu_{q(x_i)})\frac{\tilde{\alpha}_{q(\pi_u)}-1}{\alpha_{q(\pi_{u,k})}-1}l_k^T\boldsymbol{\mu}_\zeta\\
&\qquad+\frac{(\tilde{\alpha}_{q(\pi_u)}-1)(\tilde{\alpha}_{q(\pi_u)}-2)}{(\alpha_{q(\pi_{u,k})}-1)(\alpha_{q(\pi_{u,k})}-2)}l_k^T(\boldsymbol{\mu}_\zeta\boldsymbol{\mu}_\zeta^T+\Sigma_\zeta)l_k\Bigg)\Bigg\}+\sum_{k=1}^{K_x}\frac{A_{q(\phi_{x,k})}}{B_{q(\phi_{x,k})}}\Bigg\{B_{q(\phi_{x,k})}\\
&-\beta_{x,0}-\frac{1}{2}\sum_{i=1}^{n}\omega_{ik}\left(\sigma^2_{q(x_i)}+(\mu_{q(x_i)}-\mu_{q(\phi_{x,k})})^2\right)-\frac{\lambda_{x,0}}{2}(\mu_{q(\phi_{x,k})}-\mu_{x,0})^2\Bigg\}\\
&+\sum_{k=1}^{K_x}\left\{\left(\omega_{\cdot k}+\frac{\alpha_x}{K_x}-\alpha_{q(\pi_{x,k})}\right)\Psi(\alpha_{q(\pi_{x,k})})+\log\Gamma(\alpha_{q(\pi_{x,k})})+\log\Gamma(A_{q(\phi_{x,k})})\right\}\\
&+\sum_{k=1}^{K_u}\left\{\left(\rho_{\cdot\cdot k}+\frac{\alpha_u}{K_u}-\alpha_{q(\pi_{u,k})}\right)\Psi(\alpha_{q(\pi_{u,k})})+\log\Gamma(\alpha_{q(\pi_{u,k})})+\log\Gamma(A_{q(\phi_{u,k})})\right\}\\
&+(\tilde{\alpha}_{q(\pi_x)}-n+\alpha_x)\Psi(\tilde{\alpha}_{q(\pi_x)})+(\tilde{\alpha}_{q(\pi_u)}-N+\alpha_u)\Psi(\tilde{\alpha}_{q(\pi_u)})-\log\Gamma(\tilde{\alpha}_{q(\pi_x)})\\
&-\log\Gamma(\tilde{\alpha}_{q(\pi_u)})-\sum_{i=1}^{n}\sum_{k=1}^{K_x}\omega_{ik}\log\omega_{ik}-\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{K_u}\rho_{ij,k}\log\rho_{ij,k}+\frac{1}{2}\sum_{i=1}^{n}\log\sigma^2_{q(x_i)}\\
&-\frac{K_u-1}{2}\log\sigma^2_\zeta-\frac{1}{2\sigma^2_\zeta}(\boldsymbol{\mu}_\zeta^T\boldsymbol{\mu}_\zeta+tr\Sigma_\zeta)+\frac{K_x}{2}\log\lambda_{x,0}+\frac{1}{2}\log(det(\Sigma_\zeta))\\
&-\sum_{k=1}^{K_x}\left\{\frac{\lambda_{x,0}+\omega_{\cdot k}}{2\lambda_{q(\phi_{x,k})}}+\frac{1}{2}\log\lambda_{q(\phi_{x,k})}\right\}\\
&+\log\Gamma(\alpha_x)-K_x\log\left(\frac{\alpha_x}{K_x}\right)+\log\Gamma(\alpha_u)-K_u\log\left(\frac{\alpha_u}{K_u}\right)+K_x\gamma_{x,0}\log\beta_{x,0}\\
&-K_x\log\Gamma(\gamma_{x,0})+K_u\gamma_{u,0}\log\beta_{u,0}-K_u\log\Gamma(\gamma_{u,0})
\end{aligned}\tag{B.3}
$$

Let $\boldsymbol{\xi}$ denote the collection of all parameters contained in (B.3) and partition $\boldsymbol{\xi}$ into $N_B = 5$ blocks: $\xi_1 = \Big\{\mu_{q(x_i)}, \sigma^2_{q(x_i)}\Big\}_{i=1}^{n}$, $\xi_2 = \Big\{\{\omega_{ik}\}_{i=1,k=1}^{n\ \ K_x-1}, \{\rho_{ij,k}\}_{i=1,j=1,k=1}^{n\ \ m_i\ \ K_u-1}\Big\}$, $\xi_3 = \Big\{\boldsymbol{\mu}_\zeta, \Sigma_\zeta\Big\}$,

$\xi_4 = \Big\{(A_{q(\sigma^2_{u,k})}, B_{q(\sigma^2_{u,k})})_{k=1}^{K_u}, (\mu_{q(\phi_{x,k})}, \lambda_{q(\phi_{x,k})}, A_{q(\phi_{x,k})}, B_{q(\phi_{x,k})})_{k=1}^{K_x}\Big\}, \ \xi_5 = \Big\{\boldsymbol{\pi}_x, \boldsymbol{\pi}_u\Big\}$. According to our estimation algorithm, the parameters in the same block can be updated simultaneously in each iteration and the iterations can be represented by

$$\xi_i^{t+1} = arg \max_{\tau\in\text{domain}(\xi_i)} \log\underline{p}(\mathbf{y}; q, \xi_1^{t+1}, \cdots, \xi_{i-1}^{t+1}, \tau, \xi_{i+1}^{t}, \cdots, \xi_{N_B}^{t}),\ N_B = 5 \tag{B.4}$$

According to the results of [Grippo and Sciandrone, 2000], convergence of the estimation algorithm can be established if $\log\underline{p}(\mathbf{y}; q)$ is componentwise strictly quasiconcave with respect to $N_B - 2 = 3$ blocks. For simplicity of notation, we let f represent $\log\underline{p}(\mathbf{y}; q)$ in the following proofs.

Lemma 1. *The function $log\underline{p}(\mathbf{y}; q)$ is strictly concave with respect to ξ_1.*

Proof: Since

$$\begin{aligned}
&\frac{\partial^2 f}{\partial\mu^2_{q(x_i)}} = -\sum_{k=1}^{K_u}\frac{m_i A_{q(\sigma^2_{u,k})}}{B_{q(\sigma^2_{u,k})}} - \sum_{k=1}^{K_x}\frac{\omega_{ik} A_{q(\sigma^2_{x,k})}}{B_{q(\sigma^2_{x,k})}} < 0\\
&\frac{\partial^2 f}{\partial\mu_{q(x_i)}\partial\sigma^2_{q(x_i)}} = 0\\
&\frac{\partial^2 f}{\partial{\sigma^2_{q(x_i)}}^2} = -\frac{1}{2\sigma^2_{q(x_i)}} < 0
\end{aligned}$$

the function f is strictly concave with respect to $(\mu_{q(x_i)}, \sigma^2_{q(x_i)})$. Noting that $\frac{\partial^2 f}{\partial\mu_{q(x_i)}\partial\sigma^2_{q(x_j)}} = \frac{\partial^2 f}{\partial\mu_{q(x_i)}\partial\mu^2_{q(x_j)}} = \frac{\partial^2 f}{\partial\sigma^2_{q(x_i)}\partial\sigma^2_{q(x_j)}} = 0$ for all $i \neq j$, we conclude that f is strictly concave with respect to ξ_1.

Lemma 2. *The function $log\underline{p}(\mathbf{y}; q)$ is strictly concave with respect to ξ_2.*

Proof: Let $\boldsymbol{\omega}_i = [\omega_{i1}, \omega_{i2}, \cdots, \omega_{i,K_x-1}]$ for $i = 1, \cdots, n$ and $\boldsymbol{\rho}_{ij} = [\rho_{ij,1}, \rho_{ij,2}, \cdots, \rho_{ij,K_u-1}]$ for $i = 1, \cdots, n, \ j = 1, \cdots, m_i$. According to (B.3) we have $\frac{\partial^2 f}{\partial\omega^2_{ik}} = -\frac{1}{\omega_{ik}} - \frac{1}{1-\sum_{k=1}^{K_x-1}\omega_{ik}}$, $\frac{\partial^2 f}{\partial\omega_{ik}\partial\omega_{il}} = -\frac{1}{1-\sum_{k=1}^{K_x-1}\omega_{ik}}$, for $1 \leq k, l \leq K_x - 1$ and $k \neq l$, so $\frac{\partial^2 f}{\partial\xi^2_{2(i)}} = -\text{diag}(\omega_{i1}, \cdots, \omega_{i,K_x-1}) - \frac{1}{1-\sum_{k=1}^{K_x-1}\omega_{ik}} J_{K_x-1}$, where J_{K_x-1} is a $(K_x - 1) \times (K_x - 1)$

matrix of ones. For any nonzero vector $\mathbf{v}$ with length $K_x - 1$, we have $\mathbf{v}T\frac{\partial^2 f}{\partial \boldsymbol{\omega}_i^2}\mathbf{v} = -\sum_{k=1}^{K_x-1}\omega_{ik}v_k^2 - \frac{1}{1-\sum_{k=1}^{K-1}\omega_{ik}}(\sum_{k=1}^{K} v_k)^2 < 0$ when $\omega_{ik} > 0$ for all $k = 1, \cdots, K_x$. Therefore $\frac{\partial^2 f}{\partial \boldsymbol{\omega}_i^2}$ is negative definite. Similarly we can derive that $\frac{\partial^2 f}{\partial \boldsymbol{\rho}_i^2}$ is also negative definite. So f is strictly concave with respect to $\boldsymbol{\omega}_i$ and $\boldsymbol{\rho}_{ij}$ for all $i = 1, \cdots, n$ and $j = 1, \cdots, m_i$. Noting the facts that $\frac{\partial^2 f}{\partial \omega_{i_1 k_1}\partial \omega_{i_2 k_2}} = 0$ for $i_1 \neq i_2$, $\frac{\partial^2 f}{\partial \rho_{i_1 j_1, k_1}\partial \rho_{i_2 j_2, k_2}} = 0$ for $(i_1, j_1) \neq (i_2, j_2)$ and $\frac{\partial^2 f}{\partial \omega_{i_1 k_1}\partial \rho_{i_2 j_2, k_2}} = 0$ for all i_1, i_2, j_2, k_1, k_2, we can conclude that f is strictly concave with respect to ξ_2.

<u>Lemma</u> 3. *Let S_+^K denote the set of all symmetric and positive definite matrices of size K, then $g : S_+^K \to R$ give by $g(\Sigma) = log(det(\Sigma))$ is a strictly concave function.*

Proof:

$$\begin{aligned}\tilde{g}(t) &= \log(det(\Sigma + tX)) = \Sigma^{-\frac{1}{2}} + \log(det(I + t\Sigma^{-\frac{1}{2}}X\Sigma^{-\frac{1}{2}})) \\ &= \Sigma^{-\frac{1}{2}} + \sum_{k=1}^{K}\log(1 + tw_k)\end{aligned}$$

where w_k are the eigenvalues of $\Sigma^{-\frac{1}{2}}X\Sigma^{-\frac{1}{2}}$. For any nonzero symmetric matrix X such that $\Sigma + X \in S_+^K$, $[w_k]_{k=1}^K$ is a nonzero vector, it follows that $\tilde{g}(t)$ is strictly concave, thus g is strictly concave.

<u>Lemma</u> 4. *The function $log\underline{p}(\mathbf{y}; q)$ is strictly concave with respect to ξ_3.*

Proof: Since

$$\frac{\partial^2 f}{\partial \boldsymbol{\mu}_\zeta^2} = -\sum_{k=1}^{K_u}\frac{\rho_{\cdot\cdot,k}A_{q(\sigma_{u,k}^2)}}{B_{q(\sigma_{u,k}^2)}}\frac{(\tilde{\alpha}_{q(\pi_u)} - 1)(\tilde{\alpha}_{q(\pi_u)} - 1)}{(\tilde{\alpha}_{q(\pi_u)} - 1)(\tilde{\alpha}_{q(\pi_u)} - 1)}l_k l_k^T - \frac{1}{\sigma_\zeta^2}I_{K_u-1}$$

where I_{K_u-1} is a identity matrix of size n. Since $\mathbf{v}^T\frac{\partial^2 f}{\partial \boldsymbol{\mu}_\zeta^2}\mathbf{v} < 0$ for all nonzero vector $\mathbf{v}$ with length $K_u - 1$, f is strictly concave with respect to $\boldsymbol{\mu}_\zeta$.

According to (B.3), we can write f as $f = \frac{1}{2}\log(det\Sigma_\zeta) + \tilde{f}_{\Sigma_\zeta}$, where $\tilde{f}_{\Sigma_\zeta}$ is a linear function of Σ_ζ. By Lemma 3, f is strictly concave with respect to Σ_ζ.

Noting that $\frac{\partial f}{\partial \boldsymbol{\mu}_\zeta}$ does not depend on Σ_ζ and f is componentwise strictly concave with respect to μ_ζ, Σ_ζ, we conclude that f is strictly concave with respect to ξ_3. By the results of Lemma 1,2 and 4, the estimation algorithm is convergent to at least a local maximum.

BIBLIOGRAPHY

[Abrol et al., 2014] Abrol, F., Mandt, S., Ranganath, R., and Blei, D. (2014). Deterministic annealing for stochastic variational inference. *stat*, 1050:7.

[Bao and Hanson, 2016] Bao, J. and Hanson, T. E. (2016). A mean-constrained finite mixture of normals model. *Statistics & Probability Letters*, 117:93–99.

[Blei and Jordan, 2006] Blei, D. M. and Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–144.

[Blei et al., 2016] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*.

[Cappé and Moulines, 2009] Cappé, O. and Moulines, E. (2009). On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

[Carroll and Hall, 1988] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186.

[Carroll and Hall, 2004] Carroll, R. J. and Hall, P. (2004). Low order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):31–46.

[Carroll et al., 2006] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

[Delaigle and Gijbels, 2002] Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):869–886.

[Delaigle and Gijbels, 2004a] Delaigle, A. and Gijbels, I. (2004a). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics*, 56(1):19–47.

[Delaigle and Gijbels, 2004b] Delaigle, A. and Gijbels, I. (2004b). Practical bandwidth selection in deconvolution kernel density estimation. *Computational statistics & data analysis*, 45(2):249–267.

[Delaigle et al., 2008] Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, pages 665–685.

[Delaigle and Meister, 2008] Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, pages 562–579.

[Delaigle and Wang, 2015] Delaigle, A. and Wang, T. (2015). R package for deconvolution kernel estimator.

[Diebolt and Ip, 1996] Diebolt, J. and Ip, E. H. (1996). Stochastic em: method and application. In *Markov chain Monte Carlo in practice*, pages 259–273. Springer.

[Donoho et al., 1996] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539.

[Escobar and West, 1995] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

[Fan, 1991] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272.

[Fan, 1992] Fan, J. (1992). Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20(2):155–169.

[Fan and Koo, 2002] Fan, J. and Koo, J.-Y. (2002). Wavelet deconvolution. *IEEE transactions on information theory*, 48(3):734–747.

[Ferguson, 1973] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

[Gelman et al., 1996] Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760.

[Grant et al., 2006] Grant, P. M., Ryan, C. G., Tigbe, W. W., and Granat, M. H. (2006). The validation of a novel activity monitor in the measurement of posture and motion during everyday activities. *British journal of sports medicine*, 40(12):992–997.

[Grippo and Sciandrone, 2000] Grippo, L. and Sciandrone, M. (2000). On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136.

[Hall et al., 2011] Hall, P., Pham, T., Wand, M. P., Wang, S. S., et al. (2011). Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics*, 39(5):2502–2532.

[Hesse, 1999] Hesse, C. H. (1999). Data-driven deconvolution. *Journal of Nonparametric Statistics*, 10(4):343–373.

[Hoffman et al., 2013] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

[Ishwaran and Zarepour, 2002] Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283.

[Katahira et al., 2008] Katahira, K., Watanabe, K., and Okada, M. (2008). Deterministic annealing variant of variational bayes method. In *Journal of Physics: Conference Series*, volume 95, page 012015. IOP Publishing.

[Kiciman et al., 2008] Kiciman, E., Maltz, D., and Platt, J. C. (2008). Fast variational inference for large-scale internet diagnosis. In *Advances in Neural Information Processing Systems*, pages 1169–1176.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., et al. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

[Kurihara et al., 2007] Kurihara, K., Welling, M., and Teh, Y. W. (2007). Collapsed variational dirichlet process mixture models. In *IJCAI*, volume 7, pages 2796–2801.

[Mandt et al., 2016] Mandt, S., McInerney, J., Abrol, F., Ranganath, R., and Blei, D. (2016). Variational tempering. In *Artificial Intelligence and Statistics*, pages 704–712.

[Matthews et al., 2012] Matthews, C. E., George, S. M., Moore, S. C., Bowles, H. R., Blair, A., Park, Y., Troiano, R. P., Hollenbeck, A., and Schatzkin, A. (2012). Amount of time spent in sedentary behaviors and cause-specific mortality in us adults–. *The American journal of clinical nutrition*, 95(2):437–445.

[Matthews et al., 2013] Matthews, C. E., Keadle, S. K., Sampson, J., Lyden, K., Bowles, H. R., Moore, S. C., Libertine, A., Freedson, P. S., and Fowke, J. H. (2013). Validation of a previous-day recall measure of active and sedentary behaviors. *Medicine and science in sports and exercise*, 45(8):1629.

[Neal, 2000] Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

[Neal and Hinton, 1998] Neal, R. M. and Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

[Nielsen et al., 2000] Nielsen, S. F. et al. (2000). The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489.

[Parisi, 1988] Parisi, G. (1988). *Statistical field theory*. Addison-Wesley.

[Pensky et al., 1999] Pensky, M., Vidakovic, B., et al. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *The Annals of Statistics*, 27(6):2033–2053.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

[Rose et al., 1990] Rose, K., Gurewitz, E., and Fox, G. (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594.

[Sarkar et al., 2014] Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D., and Carroll, R. J. (2014). Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics*, 23(4):1101–1125.

[Stefanski and Carroll, 1990] Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics*, 21(2):169–184.

[Thorp et al., 2011] Thorp, A. A., Owen, N., Neuhaus, M., and Dunstan, D. W. (2011). Sedentary behaviors and subsequent health outcomes in adults: a systematic review of longitudinal studies, 1996–2011. *American journal of preventive medicine*, 41(2):207–215.

[Ueda and Nakano, 1995] Ueda, N. and Nakano, R. (1995). Deterministic annealing variant of the em algorithm. In *Advances in neural information processing systems*, pages 545–552.

[Wand, 1998] Wand, M. P. (1998). Finite sample performance of deconvolving density estimators. *Statistics & Probability Letters*, 37(2):131–139.

[Wand and Jones, 1994] Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Crc Press.

[Wang et al., 2006] Wang, B., Titterington, D., et al. (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650.

[Wang and Blei, 2017] Wang, Y. and Blei, D. M. (2017). Frequentist consistency of variational bayes. *arXiv preprint arXiv:1705.03439*.

[You et al., 2014] You, C., Ormerod, J. T., and Müller, S. (2014). On variational bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1):73–87.

[Zhang and Davidian, 2001] Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802.