이학박사 학위논문

Bayesian curve fitting for discontinuous functions using overcomplete system with multiple kernels

다중커널 과완비 체계를 이용한 불연속 함수의 베이즈 함수 추정

2018년 2월

서울대학교 대학원

통계학과

이영선

Bayesian curve fitting for discontinuous functions using overcomplete system with multiple kernels

지도교수 이재용

이 논문을 이학박사 학위논문으로 제출함

2017년 12월

서울대학교 대학원

통계학과

이 영 선

이영선의 이학박사 학위논문을 인준함

2017년 12월

위 원 장	오 희 석	(인)
부 위 원 장	이 재 용	(인)
위 원	장 원 철	(인)
위 원	임 채 영	(인)
위 원	최 태 련	(인)

Bayesian curve fitting for discontinuous functions using overcomplete system with multiple kernels

by

Youngseon Lee

A Thesis
submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
Feburary, 2018

Abstract

We propose a Bayesian model for estimating functions that may have jump discontinuities, and variational method for inference. The proposed model is an extension of the LARK model, which enables functions to be represented by the small number of elements from an overcomplete system composing of multiple kernels. The location of jumps, the number of elements, and even the smoothness of functions are automatically determined by the Levy random measure, there is no need for model selection. A simulation study and a real data analysis illustrate that the proposed model performs better than the standard nonparametric models for the estimation of discontinuous functions and show the suggested variational method significantly reduces the computation time than the conventional inference method, reversible jump Markov chain Monte Carlo. Finally, we prove prior positivity of the model and show that the prior has sufficiently large support including discontinuous functions with finite number of jumps.

Keywords: Bayesian nonparametric regression, overcomplete system, multiple kernel, Levy random measure, Poisson random measure, variational method, simulated annealing

Student Number: 2011-30896

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Nonparametric Bayesian regression model

A nonparametric model is a model in which the dimension of at least one parameter is infinite. An infinite-dimensional parameter is usually related to a function in the model.

In many cases, we are interested in the relationship between covariates $x \in \mathcal{X}$ and outcomes $Y \in \mathbb{R}$, which the model for this relationship is called the regression model. The relationship between variables is fully determined by data in nonparameteric regression. Data must offer a model structure as well as model estimates, so that nonparametric regression models require a large sample size. That is the reason that nonparametric models can flexibly express the relationship between variables.

The nonparametric regression model can be understood in two ways (Gray et al., 2016). A first approach assumes that covariates are fixed. A conditional distribution $f(Y|x)$ is directly modeled. The ways of modeling $f(Y|x)$ are sub-

divided into two methods: the semiparametric model and the fully nonparametric model. Some semiparametric regression model assume that a conditional distribution is modeled as $f(Y|x) = N(\eta(x), \sigma^2)$, where η is a real-valued function. In Bayesian approach, a prior distribution is assigned for unknown mean functions, η. In fully nonparametric method, $f(Y|x) = \int_\Theta f(Y|x, \theta) P_x(d\theta)$ is often assumed to be a conditional distribution which is determined by unknown mixing distribution, P_x. From Bayesian point of view, a prior distribution is defined on a family of probability distributions $(P_x)_{x \in \mathcal{X}}$. The dependent Dirichlet process (MacEachern, 1999), which derives many variations (Dunson et al., 2007; Griffin and Steel, 2006; Caron et al., 2007; Dunson and Park, 2008) are popularly used for the prior of P_x.

An alternative approach considers covariates as random variables (Muller et al, 1996). In this approach, regression problem changes into problem of the density estimation since the conditional mean function is obtained from the estimated joint density. The Dirichlet process mixture (DPM) is widely used for estimating the joint density in Bayesian field. Based on DPM, various models have been proposed (Shahbaba and Neal, 2009; Hannah et al., 2001; Wade el al., 2014).

Since we are interested in directly estimating the mean function, we consider a semiparametric model with fixed covariates. First, some methodologies for function estimation will be reviewed in the next section.

1.2 Literature review of nonparametric function estimation

Suppose we observe $\{(x_i, Y_i)\}_{i=1,\cdots,n}$, where $x_i \in \mathbb{R}, Y_i \in \mathbb{R}$, and n is the number of observations. We postulate that

$$Y_i = \eta(X_i) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, 2, \cdots, n, \tag{1.1}$$

where η is a real-valued function from $\mathbb{R}$. In this setting, estimating mean function is considered as curve fitting. There are many methodologies for nonparametric curve fitting: using basis expansion, using smoothing, using a Gaussian process, using overcomplete system.

A traditional way to approximate a function is to represent it as a linear combination of basis functions. Let $\{\phi_j\}_{j\in J}$ be the basis set of a function space to which a mean function belongs. Then, η can be expressed as

$$\eta(\cdot) = \sum_{j\in J} \beta_j \phi_j(\cdot),$$

where $\beta_j \in \mathbb{R}$ is coefficient of the basis function ϕ_j, and $J \in \mathbb{N} \cup \infty$. To estimate η, we need to estimate β_js. Polynomial basis, spline basis, Fourier basis, and wavelet basis are generally used in this methodology.

A spline basis was suggested to avoid the degree of polynomial greatly increasing when using a polynomial basis in function estimation. As an alternative to increasing the order of polynomial, the covariate space is divided into small intervals and a simple function is fitted by using observations in each interval. The basis composing this simple functions is called the spline basis.

There are many types of spline basis. A truncated power basis of the degree D is $\{1, x, x^2, \cdots, x^D, (x-\xi_1)^D, \cdots, (x-\xi_K)^D\}$, where the selected K points ξ_j are called knots. If selected knots are very close each other, values of associated truncated power basis functions will be similar in all observations, so that it results in multi-collinearity. Multi-collinearity brings unstable numerical calculation. This problem can be avoided by using other basis which produces the same range of curves. Among equivalent basis sets, the most popular one is a B-spline basis. This basis does not cause multi-collinearity, so it makes an algorithm stable.

A wavelet basis is based on wavelet transform. Wavelet transform is related to Fourier transform that represents a signal as a function of frequency. Eventhough Fourier transform is useful in understanding signals in frequency domain, information in time domain is lost. To overcome this, continuous wavelet transform (CWT) was proposed in 1984 (Grossmann et al.). However, CWT has to pay expensive cost because it has too much information. Fortunately, low frequencies will not change very quickly, we can understand a signal despite frequencies are not measured frequently. A good way to select frequencies is to use a dyadic scheme. This means Wavelet transform (WT) is done only at the positions sampled with a power of 2 in time domain. This type of WT is called discrete wavelet transform (DWT). Typically, a wavelet basis is called as a basis set composed by DWT. In curve fitting using wavelet basis, coefficients which has small values are thresholded and inverse DWT is taken with these thresholding coefficients.

A methodology based on smoothing such as spline smoothing and kernel smoothing is also popular in nonparametric curve fitting. Spline smoothing is

close to ridge regression. Every points of covariate are considered as knots, coefficients of spline basis functions are shrinked by regularization to avoid overfitting. A cubic spline and a B-spline are typically used for spline smoothing, which regularized form are a natural cubic spline and a P-spline, respectively. Such smoothing methods do not require selection of knots, but they suffer from other kind of model selection: selection of tunning parameter for determining the degree of smoothness.

Kernel smoothing is a smoothing method which averaging observations in a neighborhood at each grid point to evaluate a function value. As a neighborhood is defined by a bandwidth of kernel, according the selection of bandwidth may greatly change the shape of function. Local polynomial regression can be understood as a kind of kernel smoothing. Unlike assuming underlying functions locally constant, local polynomial function is postulated.

From Bayesian point of view, Gaussian processes can be used for function estimation, in which the mean function η is assumed

$$\eta(\cdot) \sim GP(\mu(\cdot), K(\cdot, \cdot)),$$

where μ is a mean function of Gaussian process and K is a covariance function. When using Gaussian processes, $n \times n$ dimensional matrix inversions are required repeatedly in order to obtain the estimate for η. This operation is numerically heavy when the number of observations is large. Fortunately, inverse operation can be circumvented by using Karhunen-Loeve representation of a function.

Karhunen-Loeve theorem states that a stochastic process can be represented as a linear combination of an infinitely number of orthogonal functions,

in which coefficients are random variables and a basis is determined by a covariance function that determines a process (Ghanem and Spanos, 2003). Suppose that a centered random process, $\eta(x), x \in [a, b]$, has a continuous covariance function $K(\cdot, \cdot)$. There is a linear operator T_K corresponding to this covariance function, i.e.,

$$[T_K \eta](x) = \int_a^b K(x, s)\eta(s)ds, \forall \eta \in L^2[a, b].$$

Denote the eigenfunction of T_K as ϕ_j, and the eigenvalue corresponding to ϕ_j is denoted as σ_j. According to mercer theorem, eigenfunctions of T_K form a orthonormal basis of $L^2[a, b]$ and K is expressed as

$$K(s, t) = \sum_{j=1}^{\infty} \sigma_j \phi_j(s)\phi_j(t).$$

A process η which has the covariance function K can be represented as

$$\eta(x) = \sum_{j=1}^{\infty} Z_j \phi_j(x),$$

where the uncorrelated random variable Z_j has mean of 0 and variance of σ_j. If η is a Gaussian process, coefficients are also Gaussian and independent. However, finding eigenfunctions and eigenvalues of a linear operator induced by a covariance function is not easy.

A covariance function of a Gaussian process can be considered as a reproducing kernel of reproducing kernel Hilbert space (RKHS). Let $\mathcal{X}$ be an arbitrary set, and H be a Hilbert space of real-valued functions defined on $\mathcal{X}$. When $L_x : \eta \mapsto \eta(x)$, a linear operator defined in H for $x \in \mathcal{X}$, is continuous

for all $\eta \in H$, we call H as RKHS. η can be represented as

$$\eta(x) = \sum_{j=1}^{\infty} a_j K(x_j, x),$$

where $a_j \in \mathbb{R}$ and $K(x_j, x)$ is a kernel that determines corresponding reproducing kernel Hilbert space.

Representing a function by using an overcomplete system is another approach for curve fitting. A collection of reproducing kernels is a kind of overcomplete system. A subset of the vectors $\{\phi_i\}_{i \in J}$ of Banach space $\mathcal{F}$ is called complete system if for all $f \in \mathcal{F}$,

$$||f - \sum_{i \in J} \beta_i \phi_i|| < \epsilon, \forall \epsilon > 0$$

where $\beta_i \in \mathbb{R}, J \in \mathbb{N} \cup \infty$. A complete system is called as overcomplete if removal of a ϕ_j from the system still results in a complete system. Because of the inherent redundancy in the overcomplete system, the representation using the overcomplete system can be more flexible and parsimonious than those with a complete system. However, representation would not be no longer unique due to the redundancy.

The Levy adaptive regression kernel (LARK) model, proposed by Tu (2006) is a general model which utilizes overcomplete systems in Bayesian function estimation. Pillai (2007, 2008) proved the relationship between a LARK model and a reproducing kernel Hilbert space, and he showed the posterior consistency in a LARK model. Wolpert et al. (2011) summarized the LARK model and showed convergence properties in the function space induced by the LARK

model.

1.3 Literature review of nonparametric function estimation for functions with jumps

In the nonparametric function estimation, discontinuities are common. When a important event occurs at a certain point in time, time series data may have a jump at that point. For example, on the Friday after Brexit referendum both of S&P 500 and the Dow had a big jump at that time. In x-ray data or well log data, jumps occur where different materials meet. In sensor data, data points may be suddenly shifted when sensor captures certain events.

However, most well-known methods are suitable only for continuous, especially smooth functions, so they have poor performance for estimating function with jumps. For example, in basis expansion method using a B-spline it needs repeated knots at the location of jumps, but finding these knots is practically impossible. In wavelet regression, it is known that even using continuous type of wavelet basis can approximate all functions of $L_2(\mathbb{R})$ including discontinuous functions. However, it needs an infinitely many basis functions to express a jump or it suffers to Gibbs phenomenon near a jump. Contrary, when using discontinuous type of wavelet basis such as Haar, smooth regions would not be fitted well with a finite number of basis functions.

In addition, since smoothing method and Gaussian process regression were originally devised in order to fit smoothing functions, they are not proper to estimate discontinuous functions.

Even the LARK model which utilizing an overcomplete system, same phe-

nomenon appears like the wavelet regression. In words, it requires infinite elements in order to fit jump discontinuities when using continuous type of function as the element of an overcomplete system. This phenomenon destroys identity, called parsimony, of the model which uses an overcomplete system.

By these reasons, nonparametric curve fitting methods for discontinuous functions have been studied. In the Bayesian side using basis expansion, Denison et al. (1998) used piecewise polynomial basis having free knots. Dimatteo et al. (2001) investigated free-knot selection method using B-spline basis, which is called Bayesian adaptive regression spline (BARS). In these models the number of knots can vary but the degree of smoothness of functions should be predetermined.

Many techniques have been proposed to estimate functions with discontinuities in smoothing method. Qiu (2003) and Gijbels et al. (2007) considered discontinuous function estimation in local polynomial smoothing. They assumed that every design point could be a jump. The mean function is estimated by using local polynomial smoothing with right and left side kernels at each design point. Kernel-based techniques have been studied by Muller (1992) and Kang et al. (2000), and Gijbels and Goderniaux (2004). These methods are related to selection of kernel in different regions. Spline-based methods have been considered by Koo (1997) and Spiriti et al. (2013). They studied about knot selection.

However, these methods suffer from model selection problem. It means that users have to decide tunning parameters related to something such as bandwidth, the number of jumps and the degree of smoothness. Unfortunately, performance of models heavily depends on the choice of these tunning param-

eters.

The models using an overcomplete system may mitigate model selection problem. In the LARK model, every parameters which compose a function automatically determined in inference process, so that we may bypass model selection problem. We expected that this property would be preserved even if the LARK model extends. This is one reason that we focused on the model using an overcomplete system, especially the LARK model.

Chapter 2

Bayesian curve fitting for discontinuous functions using overcomplete system with multiple kernels

2.1 Introduction

In this paper, we pay attention to the method using an overcomplete system in order to mitigate the model selection problem. In particular, focusing on the LARK model, we take the Bayesian approach for the inference of model (1.1). The original version of the LARK model utilizes only one type of kernel for composing an overcomplete system. We propose an extension of LARK model with an overcomplete system which consists of many types of kernel. We call the proposed model the Levy adaptive regression with multiple kernels

(LARMuK).

The paper is organized in the following order. In section 2, the LARK model is introduced, and needs for extension is also discussed. In section 3, we explore the structure of the proposed model and present a theorem that says that the proposed prior has large support. In section 4, posterior inference for the model using reversible jump Markov chain Monte Carlo (RJMCMC) method is described. In section 5, data analysis is given. We demonstrate the proposed method estimates diverse shapes of functions with the small number of parameters. MSE are small enough comparing to other models. In the final section, conclusions and problems for further research are discussed.

2.2 The LARK model

Let L is an infinitely divisible valued random measure defined on a complete separable metric space Ω. For L there exists the triple of sigma-finite measures (δ, Σ, ν) consisting of a signed measure $\delta(dw)$, a positive measure $\Sigma(dw)$ on Ω and a positive measure $\nu(d\beta, dw)$ on $\mathbb{R} \times \Omega$ which satisfies $\nu(\{0\}, \Omega) = 0$ and

$$\int\int_{\mathbb{R}\times A} (1 \wedge \beta^2)\nu(d\beta, dw) < \infty \quad \text{(L2 integrability condition)}$$

for each compact set $A \subset \Omega$ such that

$$\mathbb{E}[e^{itL(A)}] = \exp\left\{it\delta(A) - \frac{1}{2}t^2\Sigma(A) + \int\int_{\mathbb{R}\times A}(e^{it\beta} - 1 - ith_0(\beta))\nu(d\beta, dw)\right\} \quad (2.1)$$

where $h_0(\beta) \equiv \beta I_{[-1,1]}(\beta)$. h_0 can be replaced by any bounded measurable function h satisfying

$$h(\beta) = h_0(\beta) + O(\beta^2), \quad \beta \approx 0,$$

and $\delta(dw)$ may be replaced with

$$\delta_h(dw) = \delta(dw) + \int_{\mathbb{R}} (h(\beta) - h_0(\beta))\nu(d\beta, dw)$$

correspondingly. Then, equation (2.1) is changed into

$$\mathbb{E}[e^{itL(A)}] = \exp\left\{it\delta_h(A) - \frac{1}{2}t^2\Sigma(A) + \int\int_{\mathbb{R}\times A}(e^{it\beta} - 1 - ith(\beta))\nu(d\beta, dw)\right\}.$$

Removing Gaussain part, above equation is written by

$$\mathbb{E}[e^{itL(A)}] = \exp\left\{\int\int_{\mathbb{R}\times A}(e^{it\beta} - 1 - ith(\beta))\nu(d\beta, dw)\right\}. \tag{2.2}$$

The random measure L which has characteristic function (2.2) is called a Levy random measure and ν is called a Levy measure.

When ν satisfies L_1 integrability condition:

$$\int\int_{\mathbb{R}\times A}(1 \wedge |\beta|)\nu(d\beta, dw) < \infty, \text{ for all compact } A \subset \Omega, \tag{2.3}$$

the characteristic function of $L(A)$ becomes

$$\mathbb{E}[e^{itL(A)}] = \exp\left\{\int\int_{\mathbb{R}\times A}(e^{it\beta} - 1)\nu(d\beta, dw)\right\}, \text{ for all } A \subset \Omega. \tag{2.4}$$

Let $g(x, w)$ be a real-valued function defined on $\mathcal{X} \times \Omega$. By integrating g with respect to a Levy random measure L, we can define a real-valued function defined on $\mathcal{X}$:

$$\eta(x) \equiv \int_{\Omega} g(x, w) L(dw). \tag{2.5}$$

If ν satisfies L_1 integrability condition, $L(dw) = \int_{\mathbb{R}} \beta N(d\beta, dw)$, where N is the Poisson process on $\mathbb{R} \times \Omega$ with mean measure ν. In this case, the integral in (2.5) can be expressed as

$$\int_{\Omega} g(x, w) L(dw) = \int_{\Omega} \int_{\mathbb{R}} g(x, w) \beta N(d\beta, dw).$$

Furthermore, if $\nu(\mathbb{R} \times \Omega) = M < \infty$, $N(d\beta, dw) = \sum_{j \leq J} \beta_j \delta_{w_j}(dw)$, where $J \sim Poisson(M)$ and $(\beta_j, w_j) \overset{iid}{\sim} \nu/M$, $j = 1, 2, \ldots, J$. Thus, in this case,

$$\eta(x) = \sum_{j=1}^{J} g(x, w_j) \beta_j. \tag{2.6}$$

If g is uniformly bounded and ν satisfies L_1 integrability condition, the integral in (2.5) exists with probability 1. We call g the generating function.

Possible choices of generating functions for $\mathcal{X} = \mathbb{R}$ include symmetric kernel function, such as the Haar kernel,

$$g(x, w) \equiv I_{\left\{|\frac{x-\chi}{\lambda}| \leq 1\right\}}(x),$$

Gaussian kernel,

$$g(x, w) \equiv \exp\left\{-\frac{(x-\chi)^2}{2\lambda^2}\right\},$$

or the Laplacian kernel,

$$g(x, w) \equiv \exp\left\{-\frac{|x-\chi|}{\lambda}\right\}$$

with $w \equiv (\chi, \lambda) \in \mathcal{X} \times \mathbb{R}^+ \equiv \Omega$. For the asymmetry, one-sided exponential kernel

$$g(x, w) \equiv \exp\left\{-\frac{x-\chi}{\lambda}\right\} I_{\{x>\chi\}}(x)$$

can be used.

With likelihood (1.1), the LARK model is defined as

$$\begin{aligned} \eta(x) &\equiv \int_\Omega g(x, w) L(dw) \\ L|\theta &\sim \text{Levy}(\nu) \\ \theta &\sim \pi_\theta(d\theta), \end{aligned}$$

where Levy(ν) denotes the generating process for the Levy random measure L having the characteristic function (2.4) and $\nu(d\beta, dw)$ is a Levy measure satisfying L1 integrability condition. The conditional distribution for Y has a hyperparmeter vector θ, and π_θ denotes the probability distribution of θ. In all of examples using the LARK model, a product measure $\nu(d\beta, dw) = \nu_\beta(\beta)d\beta|\Omega|\pi_w(dw)$ is used, with $\pi_w(\cdot)$ a probability measure on Ω, $|\Omega|$ a measure of the volume of Ω, and $\nu_\beta(\cdot) > 0$ a nonnegative function on $\mathbb{R}$. Gamma, symmetric Gamma, and symmetric α-stable Levy measures are used for examples. In these examples, $\nu(\mathbb{R} \times \Omega) < \infty$ condition is not satisfied, so they adopt a truncation method in order to approximate the Levy measure to a finite Levy measure. Truncated - finite version of Levy measure is used for

practical inference process in the LARK model.

The LARK model is a novel Bayesian function estimation model with an overcomplete system. However, the LARK model could not parsimoniously represent smooth functions having discontinuities although it has overcompleteness. This is because that the LARK model uses an overcomplete system composed by single type of kernel. To extend the LARK model, we propose a model using an overcomplete with many types of kernel.

2.3 Levy adaptive regression with mutiple kernels (LARMuK)

2.3.1 Structure of proposed model

In this section, we describe the propose model, the nonparametric regression model whose mean function is expressed by an integral of a Levy random measure with multiple kernels. We consider three types of kernels, Haar, Laplacian, and Gaussian, for composing an overcomplete system. The Gaussian, Laplacian and Haar kernels are for smooth part, sharp peaks and jumps of the function, respectively.

The LARMuK model is extended from the LARK model by combining the type configuration c. Let $\boldsymbol{\theta} \equiv (w, c) \in \Omega' \equiv \Omega \times \{0, 1, 2\}$ and the generating function is denoted by

$$g(x, \boldsymbol{\theta}) = g_c(x, w) \tag{2.7}$$

where g_c represents Haar, Laplacian or Gaussian kernel depending on the value

of $c = 0, 1, 2$. Let

$$g_0(x, w) \equiv I_{\left\{|\frac{x-\chi}{\lambda}|\leq 1\right\}}(x), \tag{2.8}$$

$$g_1(x, w) \equiv \exp\left\{-\frac{|x-\chi|}{\lambda}\right\}, \tag{2.9}$$

$$g_2(x, w) \equiv \exp\left\{-\frac{(x-\chi)^2}{2\lambda^2}\right\}, \tag{2.10}$$

where $w = (\chi, \lambda)$ is the parameters of the generating function, and χ and λ are the center and scale parameter, respectively.

Generating functions of the LARMuK model also can be represented by a linear combination of three kernels whose coefficients follow the multinomial distribution with parameters (p_0, p_1, p_2), i.e.,

$$g(x, \boldsymbol{\theta}^*) = z_0 g_0(x, w) + z_1 g_1(x, w) + z_2 g_2(x, w), \tag{2.11}$$

$$(z_0, z_1, z_2) \sim \text{Multi}(1, (p_0, p_1, p_2)),$$

where $\boldsymbol{\theta}^* \equiv (w, z) \in \Omega^* \equiv \Omega \times \Omega_1$, Ω_1 is support of a multinomial distribution with the number of trial 1. By this reason, we usually call the generating function of the LARMuK model as the multiple kernel.

The mean function is defined as

$$\eta(x) \equiv \int_{\Omega'} g(x, \boldsymbol{\theta}) L(d\boldsymbol{\theta}),$$

and the randomness of a mean function is determined from the Levy random measure,

$$L \sim \text{Levy}(\nu(d\beta, d\boldsymbol{\theta}))$$

where $\nu(d\beta, d\boldsymbol{\theta})$ is a Levy measure satisfying $M \equiv \nu(\mathbb{R} \times \Omega') < \infty$. The mean function can be represented as a random finite sum,

$$\eta(x) = \sum_{j=1}^{J} g_{c_j}(x, w_j)\beta_j,$$

where $J \sim \text{Poi}(M)$ and $(\beta_j, \boldsymbol{\theta}_j) \overset{iid}{\sim} \pi(d\beta, d\boldsymbol{\theta}) \equiv \nu(d\beta, d\boldsymbol{\theta})/M$. In this paper, we consider

$$\pi(d\beta, d\boldsymbol{\theta}) = \text{N}(\beta; 0, \sigma_\beta^2)d\beta \cdot \text{Unif}(\chi; \mathcal{X})d\chi \cdot \text{Ga}(\lambda; a_\lambda, b_\lambda)d\lambda \cdot \text{Cat}(c; p_0, p_1, p_2).$$

While the LARK model considered Levy measures whose the total mass may not be finite, we consider only finite Levy measures in the LARMuK model. The total mass of Levy measure acts as a device of regularization, finite total mass may prevent overfitting by controlling the number of parameters.

Below we summarize the LARMuK model:

$$
\begin{aligned}
Y_i|x_i &\overset{ind}{\sim} \mathrm{N}(\eta(x_i), \sigma^2) \quad i = 1, 2, \cdots, n, \\
\eta(x) &= \beta_0 + \sum_{j=1}^{J} g_{c_j}(x, w_j)\beta_j, \\
J &\sim \mathrm{Poi}(M), \\
\beta_j &\overset{iid}{\sim} \mathrm{N}(0, \sigma_\beta^2), \quad j = 1, 2, \cdots, J, \\
w_j \equiv (\chi_j, \lambda_j) &\overset{iid}{\sim} \mathrm{Unif}(\mathcal{X}) \cdot \mathrm{Ga}(a_\lambda, b_\lambda), \quad j = 1, 2, \cdots, J, \\
c_j &\overset{iid}{\sim} \mathrm{Cat}(p_0, p_1, p_2), \quad j = 1, 2, \cdots, J, \\
\sigma^2 &\sim \mathrm{IG}\left(\frac{r}{2}, \frac{rR}{2}\right), \\
M &\sim \mathrm{Ga}(a_\gamma, b_\gamma),
\end{aligned}
\tag{2.12}
$$

and for the simplicity, we set $\beta_0 = \bar{Y}$.

2.3.2 Prior

We use the term "feature" for parameters $(\beta, \chi, \lambda, c)$ which compose the function. The number of features J follows Poisson distribution whose mean is the total mass of Levy measure, and the total mass follows gamma distribution with shape parameter a_γ and rate parameter b_γ. By the gamma-Poisson mixture, $J \sim \mathrm{NB}(a_\gamma, 1/(b_\gamma + 1))$. With the negative binomial distribution for J, the variance of J becomes larger than that of Poisson distribution.

The parameter β representing the coefficient depends on a difference between the maximum and the minimum value of the outcomes. i.e.,

$$\beta \sim \mathrm{N}(0, \sigma_\beta^2), \sigma_\beta \equiv (\max(Y_i) - \min(Y_i))/2.$$

The parameter χ which denotes center of a generating function is assumed to follow a uniform distribution over the covariate space. In words,

$$\chi \sim \text{Unif}(\mathcal{X}).$$

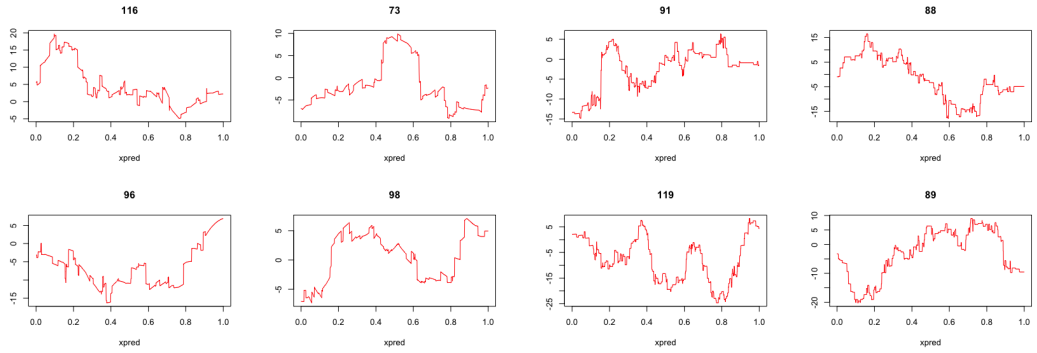
The scale parameter of a generating function λ follows a gamma distribution with the scale parameter a_λ and the rate parameter b_λ. i.e.,

$$\lambda \sim \text{Ga}(a_\lambda, b_\lambda).$$

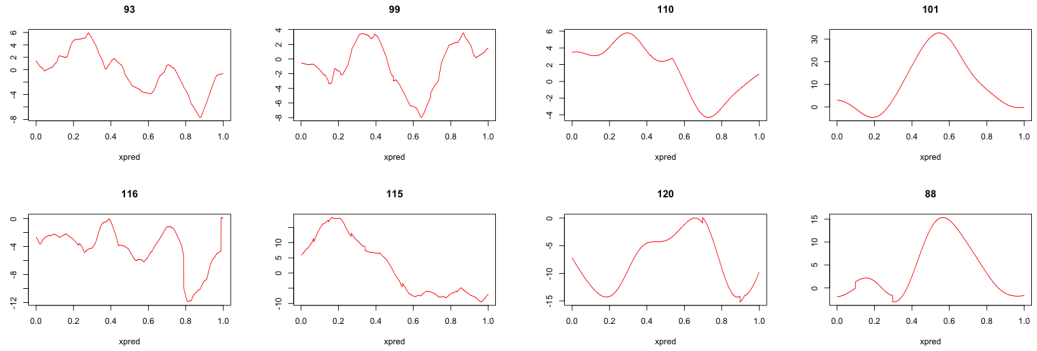
Since a_λ and b_λ control features determining the smoothness of $\eta(x)$, they play a similar role as a bandwidth in kernel smoothing. c is a configuration parameter indicating a type of kernel. c is assumed to follow a categorical distribution with probability (p_0, p_1, p_2). When probability (p_0, p_1, p_2) are fixed at $(1, 0, 0)$, the LARMuK model is same as the LARK model using Haar kernel only, $(0, 1, 0)$ and $(0, 0, 1)$ cases are equivalent to the LARK model using Laplacian kernel and Gaussain kernel only, respectively. We may control this probability in order for a mean function to posses a larger number of certain types of kernel. Examples of the LARMuK prior realizations with different cases for probability are shown in Figure 2.1. The number on top of each figure denotes the total number of features used for a realization.

Prior positivity

In this section, we will prove the LARMuK model has sufficiently large support including discontinuous function with finite jumps. Recall ν is a finite measure on $\mathbb{R} \times \Omega'$, $L \sim \text{Levy}(\nu)$ and the generating function is defined as (2.7). For



(a) $p_0 = p_1 = p_2 = \frac{1}{3}$ (b) $p_0 = \frac{98}{100}, p_1 = p_2 = \frac{1}{100}$



(c) $p_1 = \frac{98}{100}, p_0 = p_2 = \frac{1}{100}$ (d) $p_2 = \frac{98}{100}, p_0 = p_1 = \frac{1}{100}$

Figure 2.1: Examples of LARMuK prior realization with different configuration probability. The number on the top of each figure denotes the number of features used for a realization.

the simplicity, suppose $\mathcal{X} = [0, 1]$. Define

$$\Theta = \{\eta \in D[0,1] : \eta(x) \equiv \sum_{j=1}^{J} g_{c_j}(x, w_j)\beta_j, J \in \mathbb{N}, \beta_j \in \mathbb{R}, (w_j, c_j) \in \Omega'\}$$

where $D[0, 1]$ is the space of cadlag functions on $[0, 1]$, right continuous functions with left limits. Let $\bar{\Theta}$ be the closure of Θ and $\mathcal{B}(\bar{\Theta})$ be the associated

Borel σ-field. For $\eta_0 \in \Theta$, denote the ball of radius δ of η_0 by

$$B_\delta(\eta_0) \equiv \{\eta \in \bar{\Theta} : ||\eta - \eta_0||_* < \delta\}$$

where $||\cdot||_*$ is a norm defined by Skorokhod metric.

The following theorem shows that the LARMuK model has full support in the sense of Skorokhod metric.

Theorem 2.3.1. *Let ν be a finite measure with support $\mathbb{R} \times \Omega'$ and $L \sim$ Levy(ν). Let Π be a probability measure on $(\bar{\Theta}, \mathcal{B}(\bar{\Theta}))$ induced by the LARMuK model. Then for all $\delta > 0$, $\Pi(B_\delta(\eta_0)) > 0$ for every $\eta_0 \in \Theta$.*

Proof. Without loss of generality, we assume $D[0,1]$ is complete with respect to Skorokhod metric since there exists a topologically equivalent metric which $D[0,1]$ is complete. Then, $\bar{\Theta}$ is a complete set because $\bar{\Theta}$ is a subset of $D[0,1]$ and closed, and Θ is dense in $\bar{\Theta}$.

Let fix $\delta > 0$. For $\eta_0 \in \bar{\Theta}$, there exists $J_\delta \in \mathbb{N}$ and $\{\beta_j^*, {w_j}^*, c_j^*\}_{j=1}^{J_\delta}$ satisfying

$$||\eta_0(x) - \sum_{j=1}^{J_\delta} g_{c_j^*}(x, w_j^*)\beta_j^*|| < \delta/2.$$

Let $M_0 \equiv \sum_{j=1}^{J_\delta} |\beta_j^*| < \infty$ and $\kappa \equiv \sup |g_c(x,w)| < \infty$. Define $\epsilon = \frac{\delta}{2(\kappa + M_0)}$. In Skorokhod topology, there exists ϵ' s.t.

$$||w - w'|| < \epsilon' \Rightarrow \max_l ||g_l(x,w) - g_l(x,w')||_* < \epsilon$$

for all $x \in [0,1]$.

Now, define

$$B'_\delta(\eta_0) \equiv \{\eta : \eta(x) = \sum_{j=1}^{J_\delta} g_{c_j}(x, w_j)\beta_j, \sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| < \epsilon, ||w_j - w_j^*|| < \epsilon', c_j^* = c_j, \forall j\}.$$

For the proof of this theorem, we need next lemmas.

Lemma 2.3.2.

$$B'_\delta(\eta_0) \subset B_\delta(\eta_0).$$

Proof. For any $\eta \in B'_\delta(\eta_0)$, $\eta = \sum_{j=1}^{J_\delta} g_{c_j^*}(x, w_j)\beta_j$.

$$\begin{aligned}
& ||\eta(x) - \sum_{j=1}^{J_\delta} g_{c_j^*}(x, w_j^*)\beta_j^*||_* \\
\leq & \sum_{j=1}^{J_\delta} ||g_{c_j^*}(x, w_j)\beta_j - g_{c_j^*}(x, w_j^*)\beta_j^*||_* \\
\leq & \sum_{j=1}^{J_\delta} ||g_{c_j^*}(x, w_j)\beta_j - g_{c_j^*}(x, w_j)\beta_j^*||_* + \sum_{j=1}^{J_\delta} ||g_{c_j^*}(x, w_j)\beta_j^* - g_{c_j^*}(x, w_j^*)\beta_j^*||_* \\
\leq & \sum_{j=1}^{J_\delta} ||g_{c_j^*}(x, w_j)||_* \cdot |\beta_j - \beta_j^*| + \sum_{j=1}^{J_\delta} ||g_{c_j^*}(x, w_j) - g_{c_j^*}(x, w_j^*)||_* \cdot |\beta_j^*| \\
\leq & \kappa \sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| + \sum_{j=1}^{J_\delta} \max_l ||g_l(x, w_j) - g_l(x, w_j^*)||_* \cdot |\beta_j^*| \\
\leq & \kappa \sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| + \epsilon \sum_{j=1}^{J_\delta} |\beta_j^*| \\
\leq & \kappa\epsilon + \epsilon M_0 = (\kappa + M_0)\epsilon = \delta/2.
\end{aligned}$$

By triangular inequality,

$$||\eta - \eta_0||_* \leq ||\eta - \sum_{j=1}^{J_\delta} g_{c_j^*}(x, w_j^*)\beta_j^*||_* + ||\sum_{j=1}^{J_\delta} g_{c_j^*}(x, w_j^*)\beta_j^* - \eta_0||_* < \delta$$

holds. □

Lemma 2.3.3.

$$\Pi(B'_\delta(\eta_0)) > 0.$$

Proof.

$$\begin{aligned}
\Pi(\eta \in B'_\delta(\eta_0)) &= \Pi\left(\int\int\int_{\mathbb{R}\times\Omega\times\{0,1,2\}} g_c(x,w)\beta N(d\beta, dw, dc) \in B'_\delta(\eta_0)\right) \\
&= \mathbb{P}\left[\sum_{j=1}^{J} g_{c_j}(x, w_j)\beta_j \in B'_\delta(\eta_0)\right] \\
&= \mathbb{P}\left[\sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| < \epsilon, ||w_j - w_j^*|| < \epsilon', c_j = c_j^*, J = J_\delta\right] \\
&= \mathbb{P}\left[\sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| < \epsilon, ||w_j - w_j^*|| < \epsilon', c_j = c_j^*, J = J_\delta | J = J_\delta\right] \\
&\quad \times \mathbb{P}\left[J = J_\delta\right] \\
&> \mathbb{P}\left[|\beta_j - \beta_j^*| < \epsilon/J_\delta, ||w_j - w_j^*|| < \epsilon', c_j = c_j^* \ \ (j = 1, \cdots, J_\delta) | J = J_\delta\right] \\
&\quad \times \mathbb{P}[J = J_\delta] \\
&= \prod_{j=1}^{J_\delta}\left(\int_{|\beta-\beta_j^*|<\epsilon/J_\delta} \pi_\beta(\beta)d\beta \cdot \int_{||w-w_j^*||<\epsilon'} \pi_w(w)dw \cdot p(c = c_j^*)\right) \\
&\quad \times \frac{M^{J_\delta} e^M}{J_\delta!}.
\end{aligned}$$

Since $\pi_\beta(\beta) = \mathrm{N}(\beta; 0, \sigma_\beta^2), \pi_w(w) = \mathrm{Unif}(\chi; \mathcal{X}) \cdot \mathrm{Ga}(\lambda; a_\lambda, b_\lambda), c \sim \mathrm{Cat}(p_0, p_1, p_2)$

in the LARMuK model, above expression is always bigger than 0. □

Therefore,

$$\Pi(B_\delta(\eta_0)) \geq \Pi(B'_\delta(\eta_0)) > 0$$

and the proof of Theorem 2.3.1 is done. □

The next theorem says that discontinuous functions having a finite number of jumps on $[0, 1]$ are included in the support of the LARMuK model.

Theorem 2.3.4. *A prior distribution* Π *defined by the LARMuK model has positive probability on the neighborhood of discontinuous functions having finite jumps.*

Proof. By Theorem 2.3.1, it suffices to show a collection of discontinuous functions having finite jumps is included in $\bar{\Theta}$.

Let

$$\begin{aligned}
\mathcal{F}_L &= \{f \in C[0,1] : f(x) = \sum_{j=1}^{J} g_1(x, w_j)\beta_j, J \in \mathbb{N}, \beta_j \in \mathbb{R}, w_j \in \Omega\}, \\
\mathcal{F}_G &= \{f \in C[0,1] : f(x) = \sum_{j=1}^{J} g_2(x, w_j)\beta_j, J \in \mathbb{N}, \beta_j \in \mathbb{R}, w_j \in \Omega\},
\end{aligned}$$

and $\mathcal{F} \equiv \mathcal{F}_L \bigoplus \mathcal{F}_G$.

Suppose $\mathcal{X} = [0, 1]$ and a discontinuous function on $[0, 1]$ having p number of jumps can be defined as

$$\eta(x) = f(x) + \sum_{j=1}^{p} u_j I(s_j < x \leq 1)$$

where $u_j \in \mathbb{R}$, $s_j \in [0,1)$, and $f \in \mathcal{F}$. Define

$$\Theta_p = \{\eta \in D[0,1] : \eta(x) = f(x) + \sum_{j=1}^{p} u_j I(s_j < x \leq 1), f \in \mathcal{F}, u_j \in \mathbb{R}, s_j \in [0,1)\}.$$

When support of $\nu(dw)$ includes $[\frac{1}{2}, 1) \times (0, \frac{1}{2}]$, for all $s_j \in [0,1)$ $I(s_j < x \leq 1)$ is equivalent to Haar kernel with $\chi_j = \frac{1+s_j}{2}, \lambda_j = \frac{1-s_j}{2}$. Also, $f \in \mathcal{F}$ can be represented by finite sum of Laplacian and Gaussian kernel, all elements of Θ_p can be represented by elements defininig $\bar{\Theta}$. This means $\Theta_p \subset \bar{\Theta}$, therefore the proof is done.

□

2.4 Algorithm

The posterior distribution of the model (2.12) is as follows.

$$
\begin{aligned}
& [\boldsymbol{\beta}, \mathbf{w}, \mathbf{c}, J, M, \sigma^2 | \mathbf{Y}] \qquad (2.13)\\
\propto\ & [\mathbf{Y}|\eta, \sigma^2] \times [\boldsymbol{\beta}, \mathbf{w}, \mathbf{c}|J] \times [J|M] \times [M] \times [\sigma^2] \\
\propto\ & \left[(\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{J} g_{c_j}(x_i, w_j)\beta_j)^2 \right\} \right] \\
& \times \left[\exp\left\{ -\frac{1}{2\sigma_\beta^2} \sum_{j=1}^{J} \beta_j^2 \right\} \right] \times \left[\frac{1}{|\mathcal{X}|^J} \prod_{j=1}^{J} I(\chi_j \in \mathcal{X}) \right] \\
& \times \left[\left(\prod_{j=1}^{J} \lambda_j \right)^{a_\lambda - 1} \exp\left\{ -b_\lambda \sum_{j=1}^{J} \lambda_j \right\} \right] \times \left[\left(\prod_{j=1}^{J} p_{c_j} \right) \right] \\
& \times \left[\frac{M^J}{J!} \exp\{-M\} \right] \times \left[M^{a_\gamma - 1} \exp\{-b_\gamma M\} \right] \\
& \times \left[(\sigma^2)^{-\frac{r}{2}+1} \exp\left\{ -\frac{rR}{2\sigma^2} \right\} \right].
\end{aligned}
$$

Since features may have the varying dimension in the LARMuK model, we use reversible jump Markov chain Monte Carlo (RJMCMC) method for posterior computation (Green (1995)). Denote $\boldsymbol{\xi} \equiv \{\xi_j\}_{j=1,\cdots,J}$ where $\xi_j = (\beta_j, \chi_j, \lambda_j, c_j)$. A proposal distribution moves the number of features J to one of cases among $J-1$, J and $J+1$, which are called the death, walk or birth step, respectively. A form of proposal distribution q is

$$
q(\boldsymbol{\xi}'|\boldsymbol{\xi}) = p_B \cdot q_B(\boldsymbol{\xi}'|\boldsymbol{\xi}) + p_D \cdot q_D(\boldsymbol{\xi}'|\boldsymbol{\xi}) + p_W \cdot q_W(\boldsymbol{\xi}'|\boldsymbol{\xi}),
$$

where p_B, p_D, and p_W are the probabilities of choosing birth, death, and walk

step, respectively. When current $\boldsymbol{\xi}$ have J number of features, the proposal distribution for each step is defined by

$$\begin{aligned} q_B(\boldsymbol{\xi}'|\boldsymbol{\xi}) &= b(\xi_{J+1}) \times \frac{1}{J+1}, \\ q_D(\boldsymbol{\xi}'|\boldsymbol{\xi}) &= \frac{1}{J}, \\ q_W(\boldsymbol{\xi}'|\boldsymbol{\xi}) &= q_W(\xi_r'|\xi_r) \text{ for some } r, \end{aligned} \tag{2.14}$$

where $b(\xi)$ is a distribution generating a new feature. In birth step, we assume the new feature is located in the last order. In death and walk step, randomly chosen r-th feature is deleted and changed, respectively.

Since q is a mixture of three proposal distributions, it is enough to get the acceptance ratio at each step. Jacobian is 1 in all cases. Recall $J \sim \text{Poi}(M)$, and donote $\prod_{j=1}^{J} \pi(d\beta_j, d\boldsymbol{\theta}_j)$ as $\Pi(\boldsymbol{\xi})$.

Birth step It is selected with probability p_B. If birth step is accepted, change J into $J+1$ and place a new feature ξ_{J+1} in the last order. ξ_{J+1} is generated from $b(\xi)$. The acceptance ratio is $\min\left\{1, \frac{L(\mathbf{Y}|\boldsymbol{\xi}')\Pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})\Pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})}\right\}$, where

$$\frac{L(\mathbf{Y}|\boldsymbol{\xi}')\Pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})\Pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})} = \frac{L(\mathbf{Y}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})} \times \frac{\pi(\xi_{J+1})M}{J} \times \frac{p_D}{p_B \times b(\xi_{J+1})}.$$

Death step It is selected with probability p_D. If death step is accepted, randomly select an index r of one of J indices and delete the corresponding feature ξ_r. The acceptance ratio is $\min\left\{1, \frac{L(\mathbf{Y}|\boldsymbol{\xi}')\Pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})\Pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})}\right\}$, where

$$\frac{L(\mathbf{Y}|\boldsymbol{\xi}')\Pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})\Pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})} = \frac{L(\mathbf{Y}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})} \times \frac{J}{\pi(\xi_r)M} \times \frac{p_B \times b(\xi_r)}{p_D}.$$

Walk step It is selected with probability p_W. If walk step is accepted, randomly select one index r and update corresponding feature ξ_r. The dimension of features are maintained. Updating procedure in this step is totally equivalent to general Metropolis-Hasting (MH) updating. The acceptance ratio is $\min\left\{1, \frac{L(\mathbf{Y}|\boldsymbol{\xi}')\Pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})\Pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})}\right\}$, where

$$\frac{L(\mathbf{Y}|\boldsymbol{\xi}')\Pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{L(\mathbf{Y}|\boldsymbol{\xi})\Pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})} = \frac{L(\mathbf{Y}|\boldsymbol{\xi}')\Pi(\boldsymbol{\xi}')q_W(\xi_r|\xi_r')}{L(\mathbf{Y}|\boldsymbol{\xi})\Pi(\boldsymbol{\xi})q_W(\xi_r'|\xi_r)}.$$

When using independent proposal distribution for $q_W(\xi'|\xi)$, and set $q_W(\xi') = \pi(\xi_r)$, acceptance ratio is equivalent to likelihood ratio.

In the birth and death step, setting $b(\xi) = \pi(\xi)$ makes acceptance ratio much simpler because some factors - the likelihood ratio, the number of features in current state, mean of the number of features, and p_B or p_D - would determine the acceptance ratio. In addition, when $p_B = p_D$, the acceptance ratio is determined by the likelihood ratio and some values related to the number of features. However, it is better to set $b(\xi) \neq \pi(\xi)$ for mixing in practice.

In walk step, Gibbs sampling is used. χ and λ are updated by MH algorithm, while β and c are directely sampled from their conditional posterior distribution. Both σ^2 and M are also able to be sampled from their conditional posterior distribution by using conjugacy. Details are the followings.

(Sampling β)

$$
\begin{aligned}
& [\beta_k|\beta_{-k}, \text{others}, \mathbf{Y}] \\
\propto\ & \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{J} g_{c_j}(x_i, w_j)\beta_j\right)^2\right\} \times \exp\left\{-\frac{1}{2\sigma_\beta^2}\beta_k^2\right\} \\
=\ & \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\beta_k g_{c_k}(x_i, w_k) - (Y_i - \beta_0 - \sum_{j\neq k} g_{c_j}(x_i, w_j)\beta_j)\right)^2 - \frac{1}{2\sigma_\beta^2}\beta_k^2\right\} \\
=\ & \exp\left\{-\frac{1}{2\sigma^2}\left(\beta_k^2\sum_{i=1}^{n} g_{c_k}(x_i, w_k)^2 - 2\beta_k\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j\neq k} g_{c_j}(x_i, w_j)\beta_j\right)\cdot g_{c_k}(x_i, w_k)\right)\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma_\beta^2}\beta_k^2\right\} \\
=\ & \exp\left\{-\frac{1}{2}\left(\frac{\sum_{i=1}^{n} g_{c_k}(x_i, w_k)^2}{\sigma^2} + \frac{1}{\sigma_\beta^2}\right)\beta_k^2\right\} \\
& \times \exp\left\{\left(\frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j\neq k} g_{c_j}(x_i, w_j)\beta_j)\cdot g_{c_k}(x_i, w_k)}{\sigma^2}\right)\beta_k\right\}.
\end{aligned}
$$

Therefore, posterior distribution of β_k is

$$\beta_k \sim \mathrm{N}(\mu_{\beta_k}, \sigma_{\beta_k}^2)$$

with

$$
\begin{aligned}
\frac{1}{\sigma_{\beta_k}^2} &= \frac{\sum_{i=1}^{n} g_{c_k}(x_i, w_k)^2}{\sigma^2} + \frac{1}{\sigma_\beta^2}, \\
\mu_{\beta_k} &= \sigma_{\beta_k}^2 \times \frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j\neq k} g_{c_j}(x_i, w_j)\beta_j)\cdot g_{c_k}(x_i, w_k)}{\sigma^2}.
\end{aligned}
$$

(Sampling σ^2)

$$[\sigma^2|\text{others}, \mathbf{Y}] \quad \propto \quad (\sigma^2)^{-(n+r)/2+1} \exp\left\{ -\frac{1}{2\sigma^2} \left[\left(\sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{J} g_{c_j}(x_i, w_j)\beta_j \right)^2 + rR \right] \right\}.$$

Therefore, posterior distribution of σ^2 is

$$\sigma^2 \sim \text{IG}\left(\frac{r_0}{2}, \frac{r_0 R_0}{2} \right)$$

with

$$\begin{aligned} r_0 &= r + n, \\ R_0 &= \frac{\sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{J} g_{c_j}(x_i, w_j)\beta_j)^2 + rR}{r_0}. \end{aligned}$$

(Sampling M)

$$[M|\text{others}] \propto M^{J+a_\gamma - 1} \exp\{-(1 + b_\gamma)M\}.$$

Therefore, posterior distribution of M is

$$M \sim \text{Ga}(a_{\gamma 0}, b_{\gamma 0})$$

with

$$\begin{aligned} a_{\gamma 0} &= a_\gamma + J, \\ b_{\gamma 0} &= b_\gamma + 1. \end{aligned}$$

(Sampling c)

$$[c_k = l|c_{-k}, \text{others}, \mathbf{Y}] \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j\neq k} g_{c_j}(x_i, w_j)\beta_j - g_{c_k=l}(x_i, w_k)\beta_k\right)^2\right\} \times p_l.$$

Therefore, posterior distribution of c_k is

$$c_k \sim \text{Cat}(p_{k0}, p_{k1}, p_{k2}),$$

for each $l = 0, 1, 2$, posterior probability $p_{kl} = \mathbb{P}[c_k = l]$ can be obtained by normarizing above expressions.

2.5 Data analysis

The models for comparison to the LARMuK model denote as SP-# (B-spline basis regression with # knots), K (Nadaraya–Watson kernel smoothing with a Gaussian kernel), WT-S (wavelet basis regression using DWT with soft thresholding), WT-H (wavelet basis regression using DWT with hard thresholding), BPP-#1-#2 (Bayesian curve fitting using piecewise polynomial with degree l =#1, l_0 =#2 ; Denison et al.,1998), BARS-# (Bayeian adaptive regresson spline with order #; DiMatteo et al.,2001), and LK-H, LK-L, LK-G (the LARK model using Haar, Laplacian, Gaussian kernel, respectively). Functions are fitted in R using packages with default options. For the thresholding of wavelet basis regression, empirical Bayes thresholding is used. The LARK model could be considered as a special case of LARMuK model in which one of configuration

probability p_0, p_1, p_2 equal 1. The LARMuK model is denoted as LMK.

The comparison of the LARMuK model with other competing methods in the simulation and real data analysis shows that the LARMuK model has flexibility and parsimony. It is flexible for it can estimate discontinuous functions as well as continuous functions and parameters need not to be controlled depending on the shape of functions. It is parsimonious for the fitted curve has small number of features. The fact that the LARMuK model adaptively select features affects flexibility and parsimony of the model.

2.5.1 Simulation data analysis

In simulation study, the mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\eta(x_i) - \hat{\eta}(x_i))^2$$

is used to compare with the performance of competing models

The MCMC samples are generated from the posterior with 5 chains of 200,000 iterations. For each chain, the first 100,000 samples were discarded as burnin and for the remaining samples one in every 100 iteration is selected, resulting in 1,000 samples for each chain. The average of the posterior curves is used for the estimate of the mean function. The performance of each model was measured by average of the MSE.

We examined how the estimates of 5 chains are different from the estimate of one chain. Since the estimated mean functions were not much different from each others, we could say that the identifiability problem which is inherent in overcomplete representation, is not so controversal in estimating mean func-

tions. Thus, posterior samples of first chain were actually used for estimating function in data analysis.

Functions with same type of elements: Bumps, Blocks, Doppler

Bumps, Blocks and Doppler test functions of Donoho and Johnstone (1994) were used for examples. Each mean function may be assumed to consist of same-shaped elements. For example, Bumps test function is composed by spike-shaped elements. Data were generated from each test function by adding Gaussian random errors at $n = 128$ equally spaced points on $\mathcal{X} = [0, 1]$. The original data was standardized before estimation. The signal-to-noise ratio (SNR) is set at 5 and 10.

When $p_0 = p_1 = p_2 = \frac{1}{3}$ is set for configuration probability, fitted curves of LARMuK model are in Figure 2.2. Fitted curves of LARMuK model look similar to those of LARK model which select specific type of kernel in each case. This results show that the LARMuK model is flexible and can be used as off-the-shelf method, for the user does not need to select a type of kernel in advance.

Figure 2.3 shows the fitted curves of LARMuK model and those of BARS-1 in each test function. The shape of curves are very similar to each other, but the (posterior mean of) number of features used for fitting curves in the LARMuK model are much less than the (posterior mean of) number of knots used in BARS-1 (Table 2.1). Nevertheless, MSE of LARMuK model are consistently smaller than the other models (Table 2.2). These results indicate inherent parsimony of the LARMuK model.

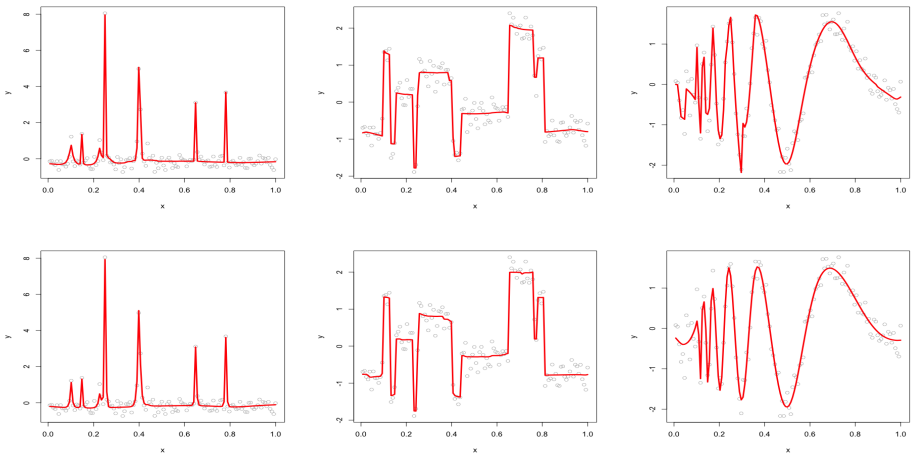


Figure 2.2: First row: estimated three test functions (Bumps, Blocks, and Doppler functions) using LMK with configuration probability, $p_0 = p_1 = p_2 = \frac{1}{3}$. Second row: estimated three test functions using LK with Haar, Laplacian, Gaussian kernel, respectively. SNR=5 is set.

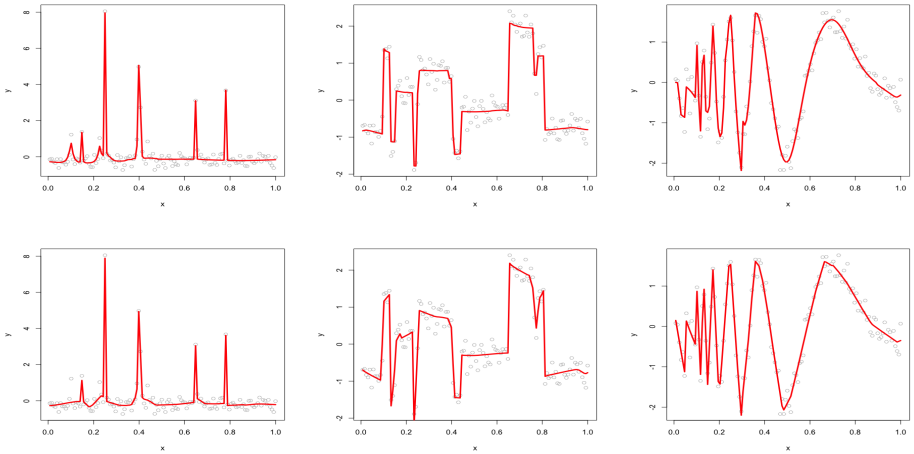


Figure 2.3: First row: estimated three test functions (Bumps, Blocks, and Doppler functions) using LMK with configuration probability, $p_0 = p_1 = p_2 = \frac{1}{3}$. Second row: estimated three test functions using BARS-1. SNR=5 is set.

Table 2.1: The (posterior mean of) number of features or knots of fitted curves using LMK and BARS-1.

Model	Bumps		Blocks		Doppler	
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
LMK	6	13	14	19	16	31
BARS-1	22	31	24	26	22	32

Table 2.2: MSE for the estimated mean function of each model

Model	Bumps		Blocks		Doppler	
	SNR=5	SNR=10	SNR=5	SNR=10	SNR=5	SNR=10
LMK	0.033	0.012	0.013	**0.003**	**0.029**	**0.008**
LK	**0.031**	0.047	**0.005**	0.003	0.036	0.019
BPP-1-0	0.135	0.110	0.040	0.024	0.043	0.422
BPP-2-0	0.091	0.074	0.038	0.010	0.032	0.021
BPP-2-1	0.288	0.313	0.066	0.063	0.027	0.015
BPP-2-2	0.971	0.961	0.132	0.137	0.083	0.068
BARS-1	0.041	**0.005**	0.035	0.003	0.029	0.009
BARS-2	0.991	0.991	0.103	0.078	0.042	0.036
BARS-3	0.992	0.991	0.312	0.226	0.073	0.075
WT-S	0.066	0.043	0.033	0.027	0.032	0.018
WT-H	0.150	0.045	0.013	0.049	0.029	0.013
SP	0.767	0.787	0.073	0.557	0.045	0.027
K	0.478	0.061	0.065	0.127	0.035	0.126

Functions with the different types of elements : Blip, Multi, Heavisine

In the second example, each mean function is assumed be a mixture of different type of elements. For example, Blip data is generated from a function consist of lines, curves, jumps, and constants (Antoniadis et al. (2001)). In this paper, we call the function having property that derivatives are much different along the covariate space as "multi-scale function". Multi data is generated from a smooth but multi-scale function. This kind of function is known to be difficult to fit by nonparametric regression models. Multi data do not have jumps, but we considered this example to illustrate that how the LARMuK model fit arbitrary function well.

Heavisine data can be easily found in digital modulation. Phase shift keying is one of the methods used when transmitting digital signals over analog channels. Quadrature phase shift keying (QPSK) is well-known method for digital modulation. This modulation scheme produces a signal with a binary or quadrature signal added to the carrier. Carrier is a periodic function with a specific frequency, and this periodic function is changed by a binary or quadrature signal. The shape of the modulated signal is mixture of binary functions and periodic functions. In fact, since the carrier and signal types used in QPSK are known in advance, the LARMuK model may not have good performance compared to the other models which are optimized for this type of data. However, in order to see how well functions with jumps are estimated, Heavysine data was used for example.

The data were generated from each function by adding Gaussian random errors at $n = 128$ equally spaced points on $\mathcal{X} = [0, 1]$. Original data was

not standardized. Signal-to-noise ratio (SNR) set 5, meanwhile SNR set 10 in Heavisine data in order to discriminate jump signal and noise well. Formulas of the three functions - Blip, Multi and Heavisine are:

Blip

$$f(x) = \begin{cases} 0 & x = 0 \\ 0.32 + 0.6x + 0.3e^{-100(x-0.3)^2} & 0 < x < 0.8 \\ -0.28 + 0.6x + 0.3e^{-100(x-1.3)^2} & 0.8 \leq x \leq 1 \end{cases}$$

Multi

$$f(x) = 3\left\{\sin(x) + \frac{\sin(\pi(x-5))}{\pi(x-5)} + \frac{\sin(5\pi(x-2))}{5\pi(x-2)} + 1\right\}$$

Heavisine

$$f(x) = 4\sin(4\pi x) - 4sgn(x-0.1) - 2sgn(x-0.3) + 4sgn(x-0.5) + 2sgn(x-0.7),$$

where $sgn(x)$ is a function with a value of 1 if x is greater than or equal to 0, or a value of -1 if less than 0. True functions are in Figure 2.4.

In Figure 2.5, we find out that the LARMuK model estimates Blip and Heavisine function well, but the LARK model can not estimate these functions properly no matter what type of kernel is selected. The LARK model could not fit jump regions when it use a continuous type of kernel (Gaussian kernel), and it may not estimate continuous regions with a discontinuous type of kernel (Haar kernel).

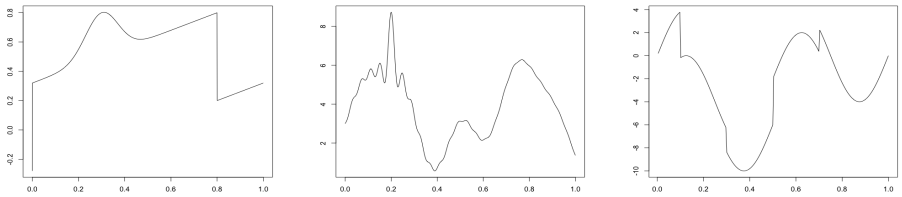


Figure 2.4: From left to right, figures are Blip, Multi, and Heavsine functions, respectively.

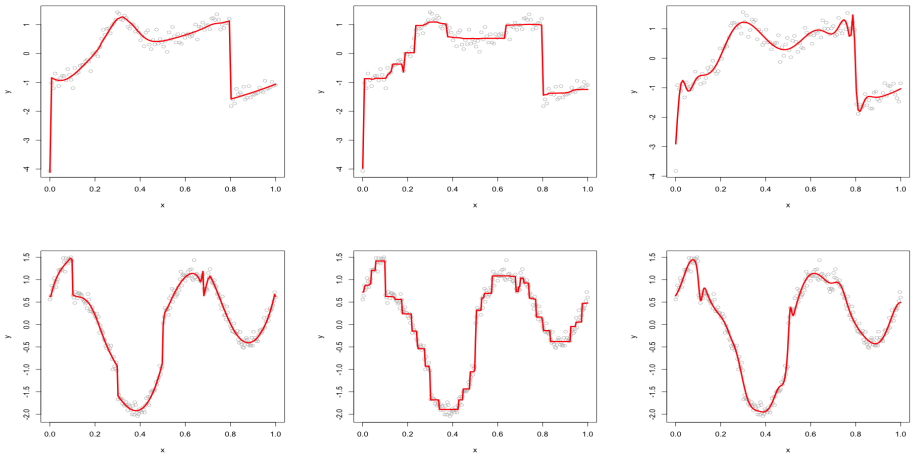


Figure 2.5: First row: estimated Blip functions of LMK and LK-H, LK-G, respectively. Second row: estimated Heavisine functions of LMK and LK-H, LK-G, respectively

Table 2.3: MSE for the estimated mean function of each model

Model	Blip	Multi	Heavisine
LMK	**0.005**	**0.008**	**0.002**
LK-G	0.035	0.049	0.012
BPP-1-0	0.019	0.043	0.010
BPP-2-0	0.034	0.038	0.010
BPP-2-1	0.015	0.019	0.006
BPP-2-2	0.046	0.022	0.010
BARS-1	0.005	0.016	0.007
BARS-2	0.032	0.013	0.013
BARS-3	0.030	0.013	0.012
WT-S	0.038	0.057	0.007
WT-H	0.030	0.019	0.011
SP-20	0.061	0.044	0.010
SP-50	0.032	0.019	0.009

Figure 2.5 and Table 2.3 show that the performance of the LARMuK model is superior comparing with other models. The estimated function of LARMuK model captures jumps as well as smooth regions, and the MSE are much small than other competing models. Surprisingly, $p_0 = p_1 = p_2 = \frac{1}{3}$ were used for configuration probability in all examples, Blip, Multi and Heavysine data. This means that functions would be estimated without controlling model specifications in the LARMuK model. This represents the flexibility of the LARMuK model.

Figure 2.6 illustrates the adaptiveness of LARMuK model. This type of data is known not to be fitted well by conventional models. Fitted curves of Multi data using the LARMuK model are compared with those using B-spline basis regression with 20 and 50 knots. In Figure 2.6, we find out that the selected features are adaptively scattered in the LARMuK model. This

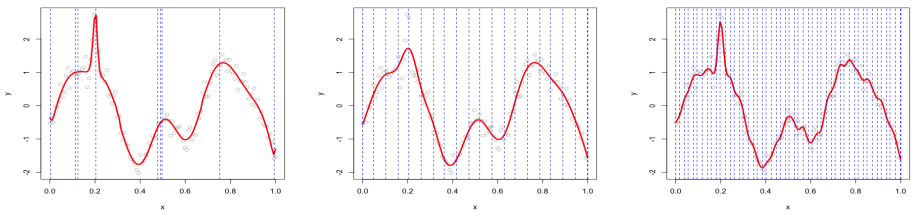


Figure 2.6: From left to right, figures are fitted curves using LMK and SP-20, SP-50, respectively. Blue dotted vertical lines denote the center features or knots.

property makes the LARMuK model be parsimonious and flexible.

2.5.2 Real data analysis

To further illustrate the LARMuK model, we analyzed a sensor data. Sensor data refers to signal data obtained from sensors responding to factors like humidity, temperature, and motion. Signals may not move continuously when sudden changes occur. In the case of sensor that responds to motion, jumps may occur when some movements are detected. Or a signal related to temperature is suddenly changed when fire starts. It is important to detect and examine these signals because we may not see what actually happened at that time. We are able to sense the occurrences of events only through the signal patterns. There are models for special types of signal, but few models are applicable for general types of signal.

Figure 2.7 is a record of the temperature signal from one sensor. We can easily find out there are at least two jumps by eyes. The total number of observations is $n = 512$ and randomly chosen 20%, 50% of dataset are used as test sets for validation.

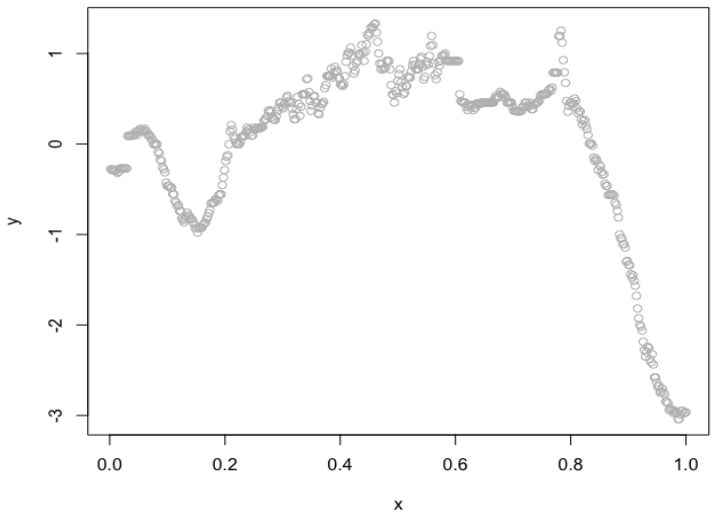


Figure 2.7: Temperature signal from one sensor

Figure 2.8 and 2.9 show the predicted curves of the LARMuK model and those of the LARK model using each of the three types of kernel. In the LARK model using the Haar kernel, the locations of the jumps are captured well but the the continuous regions are fitted as piecewise constant functions. On the other hand, for the case of LARK model with the Laplacian or Gaussian kernel, the continuous regions are reasonably fitted, but jumps are not estimated at all. These results would give us distorted interpretation of signal. When the signal is fitted with continuous curve only, some important information, such as when the signal jumped, what factor caused the signal to suddenly move, may lose. However, the LARMuK model could capture discontinuities and preserve them in predicted curve, thus we could understand the signal without losing information.

From the Table 2.4, the LARMuK model have more small MSE than others.

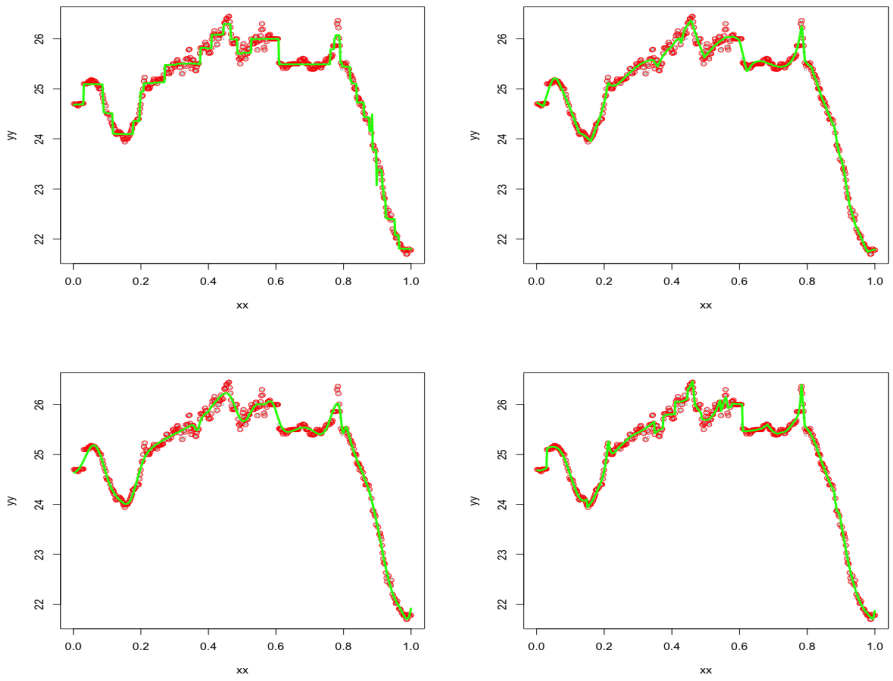


Figure 2.8: From top left to bottom right, using 80% of dataset as traing set, curves are predicted by LK-H, LK-L, LK-G and LMK. Grey circles indicate original dataset, red circles denote training set, green line is the predicted curve.

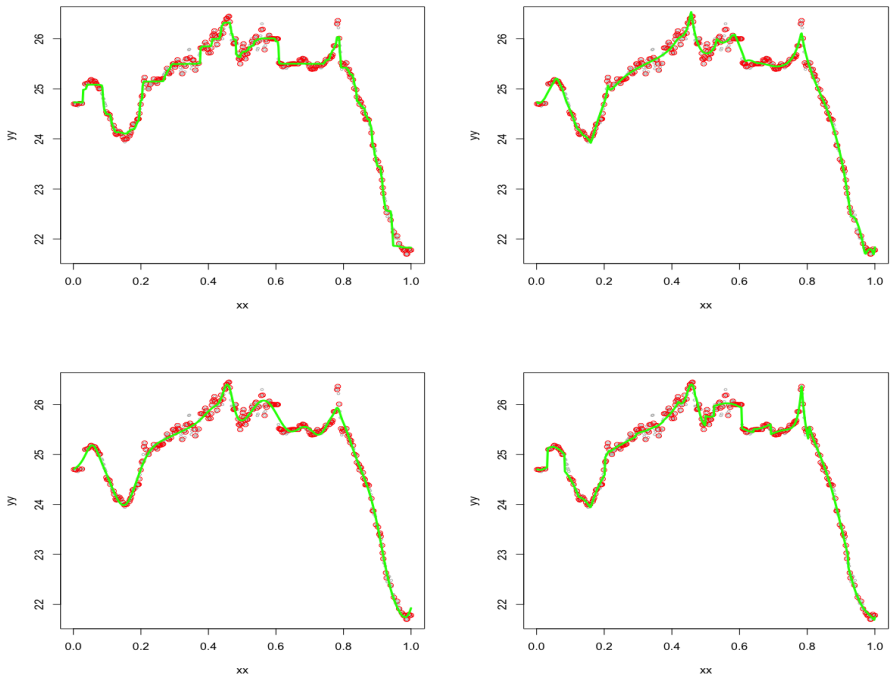


Figure 2.9: From top left to bottom right, using 50% of dataset as traing set, curves are predicted by LK-H, LK-L, LK-G and LMK. Grey circles indicate original dataset, red circles denote training set, green line is the predicted curve.

Table 2.4: MSE for the predicted mean function of each model.

Model	LMK	LK-H	LK-L	LK-G
20%	**0.0074**	0.0207	0.0078	0.0089
50%	**0.0092**	0.0107	0.0094	0.0103

Table 2.5: The (posterior mean of) number of features or knots of predicted curves using each method. 20% denotes that twenty percent of the total dataset is used for the test set.

Model	LMK	LK-H	LK-L	LK-G	BARS-1	BARS-2	BARS-3
20%	**23**	24	34	**25**	54	24	18
50%	**15**	21	17	**14**	48	20	16

Even if the mean function have jump discontinuities, the number of features used in the LARMuK model is as small as those of the LARK model using Gaussian kernel (Table 2.5). This means that the LARMuK model selects features in considerably adaptive way.

In simulation example, BARS-1 and BPP-2-0 have also small MSE thus they may be considered appropriate for estimating functions having discontinuities. However, BARS-1 and BPP-2-0 make predicted curves to be wiggling. This is because selection procedure of knots heavily depends on the number of observations. They tend to choose many parameters than LARMuK model when the number of observation increases. Figure 2.10 shows these phenomenons. Additionally, they are meaningless in real data examples with jumps because they can not adaptively select the smoothness of the curve along the covariate space. Once the user determines the degree of smoothness in advance, it must be used in the entire covariate space. Furthermore, users can not really know how smooth the underlying functions are. This causes model selection

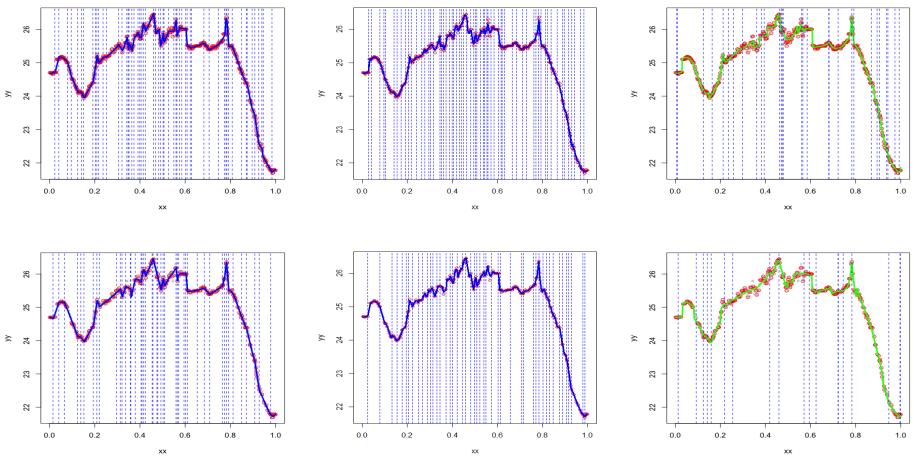


Figure 2.10: From left to right, in each training set (80%, 50% of dataset), figures are predicted curves of BARS-1, BPP-2-0 and LMK, respectively. Grey circles indicate original dataset, red circles denote training set, line is predicted curve of each model. Blue dotted vertical lines denote knots or center features.

problem.

2.6 Discussion

We proposed the LARMuK model which is an extension of the LARK model. The LARMuK model uses an overcomplete system with multiple kernels in order to parsimoniously represent functions, especially discontinuous functions. In particular, we used three types of kernel; Haar, Laplacian, and Gaussian kernels are used to compose an overcomplete system.

Theories and data analysis confirm that LARMuK model can aptly estimate a large range of functions including continuous functions and discontinuous functions. The LARK model with one kernel sometimes suffers difficulty of parsimoniously representing functions with regionally different smoothness

properties. By using multiple kernels, the LARMuK model however can parsimoniously represent such functions. In addition, since the estimated functions are composed by the small number of kernels, fitted curves can be understood in a natural way. Nevertheless, the LARMuK model is superior in performance than any other model.

One drawback of the LARMuK model is that RJMCMC used for inference. Sampling procedure from the posterior distribution is slow in mixing and is time-consuming. We need to improve the inference process by using other approaches such as optimization instead of sampling. Variational method is a promising computation method.

Asymptotic theory such as posterior consistency has not been proved yet. Future task will be to prove theoretical basis such as posterior consistency and posterior convergence rate as well as some connection with other mathematical theories like connection with RKHS.

Chapter 3

Stochastic variational inference for the LARMuK model

3.1 Introduction

Variational method is often useful in nonparametric Bayesian model, including Dirichlet Process (Blei and Jordan, 2004) and Gaussian Process (Winther, 2000). Variational method may change the sampling-based inference into deterministic problem, inference for models would be involved optimization.

Eventhough the LARMuK model is attractive for function estimation, sampling-based inference is challenging. For this reason, we suggest an alternative method to infer the LARMuK model using variational method.

There are some problems when we bring variational method for the LARMuK model. First, it is difficult to update some variational distributions for variables lying inside generating functions. Second, since the number of features is also random in the LARMuK model, optimizing features induced by

the number of features are intractable. In this chapeter we will show that variational EM method adding probabilistic procedure can mitigate these problems, In particular, simulated annealing algorithm is used for probabilistic procedure. We call the proposed variational method as stochastic variational method.

The paper is organized in the following order. In section 2 general variational method is introduced, and relationship between variational method and EM method is discussed. In section 3, we explain simulated annealing algorithm. In section 4, stochastic variational method for the LARMuK model is illustrated. A simulation study and a real data analysis are given in section 5. In the final section, conclusions and problems for further researches are discussed.

3.2 Variational method in general

Variational method approximates the true posterior via a simpler distribution and changes sampling-based inference into a deterministic optimization problem. In variational method, the closest distribution to the posterior distribution is used as a proxy for the true posterior. "Close" means that the Kulback-Leibler divergence between a proxy and the true posterior distribution is as small as possible. Typically, some classes of distributions are considered for the set of candidates of a proxy. We call this as the class of variational distributions, and the closest member to the posterior distribution is called the optimal variational distribution.

If we assume all variables are independent of each other in the class of joint variational distributions, a joint variational distribution can be expressed as

the product of marginal variational distributions. This is called factorization. Factorization usually makes inference procedure be tractable, while correlations between variables could not be captured anymore.

Let assume that there are two variables, θ and z. Observations are denoted by x. Denote the joint posterior distribution as $p(\theta, z|x)$ and a joint variational distribution as $q(\theta, z)$. The optimal joint variational distribution is obtained from the following optimization procedure:

$$\hat{q}(\theta, z) = \arg\min_{q} \mathrm{KL}(q||p(\theta, z|x)) = \int\int \log \frac{q(\theta, z)}{p(\theta, z|x)} q(\theta, z) d\theta dz.$$

The best $\hat{q}$ is the posterior distribution. However, in most of cases the form of posterior distribution is intractable. In order to make problem simple, a specific class of q would be considered in variational method.

Let define

$$J(q) \equiv KL(q||p(\theta, z, x)) = KL(q||p(\theta, z|x)) - \log p(x).$$

Since $\log p(x)$ is constant with respect to q, finding $\hat{q}$ minimizes $KL(q||p(\theta, z|x))$ is equivalanet to finding $\hat{q}$ which minimizes $J(q)$. We can easily find out that

$$\log p(x) + J(q) = KL(q||p(\theta, z|x)) \geq 0$$

holds because Kulback-Liebler divergence is always equal or greater than 0. This means $-J(q)$ is a lower bound of $\log p(x)$, the ELBO (evidence lower

bound) is defined as

$$L(q) \equiv -J(q) = -KL(q||p(\theta, z, x)) = E_q \log p(\theta, z, x) - E_q \log q(\theta, z).$$

The joint variational distribution is optimized untill the ELBO converges as tight as possible.

In factorization, the joint variational distribution of θ and z is expressed as $q(\theta, z) = q(\theta)q(z)$. Variational method using factorization is called mean-field variational method. In mean-field variational method, the ELBO is expressed as

$$L(q) = \int\int \log p(\theta, z, x) q(\theta) q(z) d\theta dz + H_q(\theta) + H_q(z) \tag{3.1}$$

where $H_q(\cdot) \equiv -E_{q(\cdot)} \log q(\cdot)$ denotes entropy of $q(\cdot)$, therefore, finding $\hat{q}(\theta, z)$ changes into the problem of finding $\hat{q}(\theta)$ and $\hat{q}(z)$, respectively. It means that it is enough to find

$$\hat{q}(\theta) = \arg\min_{q(\theta)} \left[\int \left(\int \log p(\theta, z, x) q(z) d\theta dz \right) q(\theta) d\theta + H_q(\theta) \right], \tag{3.2}$$

and

$$\hat{q}(z) = \arg\min_{q(z)} \left[\int \left(\int \log p(\theta, z, x) q(\theta) dz d\theta \right) q(z) dz + H_q(z) \right]. \tag{3.3}$$

For the case of θ, (3.2) is represented as

$$\hat{q}(\theta) = \arg\min_{q(\theta)} KL(q(\theta)||f(\theta)) = \int \log \frac{q(\theta)}{f(\theta)} q(\theta) d\theta,$$

where $\log f(\theta) \equiv \int \log p(\theta, z, x) q(z) d\theta$, thus the optimal variational distribution will be $\hat{q}(\theta) \propto f(\theta)$.

So far we do not specify a class of q. When the prior and the likelihood are conjugate, we can take the form of candidate distribution as the same as the form of prior distribution. We could easily obtain a variational distribution in this situation. But how we specify a class of variational distributions in nonconjugate cases?

Variational method for nonconjugate models has been studied by Wang et al.(2013). In this study, two approximation methods were introduced: Laplace approximation and delta approximation. These methods approximate conditional posterior distributions for some variables, which can not be expressed in particular forms, into Gaussian distributions. They considered the generic model with observations x and variables θ, z,

$$p(\theta, z, x) = p(x|z)p(z|\theta)p(\theta),$$

and dealt with nonconjugate models defined as follows.

- $p(\theta)$ is twice differentiable with respect to θ. If $\theta > \theta_0$ is required, a distribution over $\log(\theta - \theta_0)$ could be defined.
- $p(z|\theta)$ is in the exponential family.
- $p(x|z)$ is in the exponential family which makes the distribution of z is conjugate to the conditional distribution of x; the conditional $p(z|\theta, x)$ is in the same family as $p(z|\theta)$.

This means that θ is a nonconjugate variable, z is a conjugate variable.

Laplace approximation method is based on a Taylor expansion of $f(\theta)$ at the maximum a posteriori (MAP) estimate, $\hat{\theta}^{MAP}$. Assuming that $q(\theta)$ is a Gaussian distribution, the MAP estimate of $f(\theta)$ is related to the parameters of a variational distribution for θ, i.e.,

$$q(\theta) \approx N(\hat{\theta}^{MAP}, -\nabla^2 f(\hat{\theta}^{MAP})^{-1}).$$

In Laplace method, original ELBO is used as the objective function and a class of variational distributions is assumed to be a Gaussian.

On the other hand, delta approximation method modifies the ELBO by using a Taylor expansion at $\hat{\theta}$. In this case, a class of $q(\theta)$ is specified as a Gaussian in advance, variational parameters of $q(\theta)$ are numerically determined. That is, $q(\theta) = N(\mu, \Sigma)$ is assumed no matter what the true posterior distribution of θ is, so that the ELBO is

$$L(q) = \mathbb{E}_q f(\theta) + \frac{1}{2}\log|\Sigma|.$$

Then, the ELBO is approximated as

$$\begin{aligned} L(q) \approx & \left[f(\hat{\theta}) + \nabla f(\hat{\theta})^T(\mu - \hat{\theta}) + \frac{1}{2}(\mu - \hat{\theta})^T \nabla^2 f(\hat{\theta})(\mu - \hat{\theta}) + \frac{1}{2} Tr\{\nabla^2 f(\hat{\theta})\Sigma\} \right] \\ & + \frac{1}{2}\log|\Sigma|. \end{aligned}$$

Three options are available for $\hat{\theta}$. The first one is the MAP estimate of $f(\theta)$, the second choice is the value of previous iteration. The third choice is $\hat{\theta} = \mu$, which could simplify the objective function.

In fact, variational method is not appropriate for infering the LARMuK

model. First, the dimension of variables depends on the number of features J, however, the dimension of variables is assumed to be fixed in variational methods for even nonparametric Bayesian model which has possibility of varying dimension of variables (Doshi et al., 2009). Second, some variables are lying inside generating functions. It is difficult to get expectations involving these implicit variables, so that we can not take good advantage of variational method. Third, some variables are not proper to assume certain family of distributions for marginal variational distributions. Although the LARMuK model does not satisfies the conditions suggested by Wang et al. (2013), the idea of approximation methods used in variational method for nonconjugate model could be available. However, general approximation method used in Wang et al. (2013) is difficult to be adopted since some variables would be defined on unstructured support. For example, χ is defined on support which covariates defined on.

In summary, the most important part in order to use variational method for the LARMuK model is how to specify a class of variational distributions for features to evaluate expectations easily. And also, we need to consider the method effectively updating J.

One idea is to assume Dirac delta function as candidate distribution of χ_k, λ_k and J. In other words, $q(\chi_k)$ is assumed to be a degenerated function at one point. Then, the problem of finding the optimal variational distributions of χ_k, λ_k and J become the problem of finding the point that maximizes the ELBO or finding the point that minimizes Kulback-Leibler divergence.

In the LARMuK model, variational distributions depend on $\mathbb{E}_q g_{c_k}(x_i; \chi_k, \lambda_k)$. Since χ_k and λ_k are implicit variables, it is hard to get an expectation of $g_{c_k}(x_i; \chi_k, \lambda_k)$ without numerical methods. However, if we assume that vari-

ational distributions of χ_k and λ_k are Dirac delta functions, $\mathbb{E}_q g_{c_k}(x_i; \chi_k, \lambda_k)$ may be computed in a simple way. In addition, finding the optimal $q(\chi_k)$ maximizes $f(\chi_k)$ becomes equivalent to finding $q(\chi_k)$ maximizes $L(q(\chi_k))$ since the entropy of a delta distribution function is constant. We do not have to worry about which objective function to be considered.

When the posterior distribution have complex form, this approach may have some drawbacks related to optimization such as finding a local optimum. Furthermore, someone may insist that using delta distribution functions be too restrictive. We will discuss about these concerns later.

Now, let's go back to (3.1) and consider the case where variational distributions for some variables are delta functions. Suppose $q(\theta)$ is a delta function, i.e.,

$$q(\theta) = \delta_\alpha(\theta) = \begin{cases} 1, & \theta = \alpha \\ 0, & \theta \neq \alpha \end{cases}.$$

Denote the optimal α as θ_0. Let a class of variational distributions for z, $q(z)$, could have free form. Under these assumptions, it is enough to find θ_0 satisfying

$$\begin{aligned} \theta_0 &= \arg\max_\alpha L(q) \\ &= \arg\max_\alpha \left[\mathbb{E}_{q(z)} \log p(\alpha, z, x) + H_q(\theta)\right] \\ &= \arg\max_\alpha \mathbb{E}_{q(z)} \log p(\alpha, z, x). \end{aligned}$$

On the other hands, the optimal variational distribution of z could be obtained

by

$$\begin{aligned}
\hat{q}(z) &= \arg\max_{q(z)} L(q) \\
&= \arg\max_{q(z)} \left[\mathbb{E}_{q(\theta)} \log p(\theta, z, x) + H_q(z)\right] \\
&= \arg\max_{q(z)} \left[\log p(\theta_0, z, x) + H_q(z)\right] \\
&= \arg\min_{q(z)} KL(q(z)||p(\theta_0, z, x)) \\
&\propto \exp\{\log p(\theta_0, z, x)\}.
\end{aligned}$$

Now, it is time to discuss about one concern mentioned before. *Is it proper to set the class of variational distributions for some variables as Dirac delta functions?* The answer is yes. Because using delta distribution functions as variational distributions leads generalized EM method.

3.2.1 The relationship with EM method

The goal of EM algorithm is to find the maximum likelihood estimator in the case of model has latent variables. Suppose θ is a parameter to be estimated, z is a latent variable and x is set of observations. Likelihood denotes $p(x|\theta)$. The maximum likelihood estimator of θ is obtained by

$$\hat{\theta}^{ML} = \arg\max_{\theta} \log p(x|\theta).$$

We can easily find out that

$$\log p(x|\theta) = L(q(z), \theta) + KL(q(z)||p(z|x, \theta))$$

holds, where

$$L(q(z), \theta) = -KL(q(z)||p(x, z|\theta))$$

In EM method, procedure of obtaining the maximum likelihood estimator has two steps.

1. Estimate $q(z)$ when $\hat{\theta}$ is given:

$$\hat{q}(z) = p(z|x, \hat{\theta}) = \arg\max_{q(z)} L(q(z), \hat{\theta}) = \arg\max_{q(z)} \left[\mathbb{E}_{q(z)} \log p(x, z|\hat{\theta}) + H_q(z)\right].$$

2. Estimate θ when $\hat{q}(z)$ is given:

$$\hat{\theta} = \arg\max_{\theta} L(\hat{q}(z), \theta) = \arg\max_{\theta} \mathbb{E}_{\hat{q}(z)} \log p(x, z|\theta).$$

It is well known that iteration of this procedure makes $\hat{\theta}$ go to the maximum likelihood estimator. Instead of using $\hat{q}(z)$ as $p(z|x, \hat{\theta})$, we could specify a class of $q(z)$. This modifying procedure is called variational EM method.

On the other hands, the goal of variational method is to obtain the closest $q(\theta, z)$ to the posterior distribution $p(\theta, z|x)$, i.e.,

$$\hat{q}(\theta, z) = \arg\min_{q(\theta,z)} KL(q(\theta, z)||p(\theta, z|x)).$$

In mean-field variational method, we know that

$$\log p(x) = L(q(\theta), q(z)) + KL(q(\theta)q(z)||p(z, \theta|x))$$

holds, where

$$L(q(\theta), q(z)) = -KL(q(\theta)q(z)||p(x, \theta, z)).$$

Therefore, the procedure of obtaining variational distibutions is as follows.

1. Estimate $q(z)$ when $\hat{q}(\theta)$ is given:

$$\hat{q}(z) = \arg\max_{q(z)} L(\hat{q}(\theta), q(z)) = \arg\max_{q(z)} \left[\mathbb{E}_{q(z)}\mathbb{E}_{\hat{q}(\theta)} \log p(\theta, z, x) + H_q(z)\right].$$

2. Estimate $q(\theta)$ when $\hat{q}(z)$ is given:

$$\hat{q}(\theta) = \arg\max_{q(\theta)} L(q(\theta), \hat{q}(z)) = \arg\max_{q(\theta)} \left[\mathbb{E}_{q(\theta)}\mathbb{E}_{\hat{q}(z)} \log p(\theta, z, x) + H_q(\theta)\right].$$

If we specify $q(\theta)$ as a delta distribution function, above procedure would be modified.

1. Estimate $q(z)$ when $\hat{\theta}$ is given:

$$\hat{q}(z) = \arg\max_{q(z)} \left[\mathbb{E}_{q(z)} \log p(\hat{\theta}, z, x) + H_q(z)\right].$$

2. Estimate θ when $\hat{q}(z)$ is given:

$$\hat{\theta} = \arg\max_{\theta} \mathbb{E}_{\hat{q}(z)} \log p(\theta, z, x).$$

The main difference between variational EM method and mean-field variational method using a class of delta distribution functions is the form of objective function of θ. In variational EM method, the objective function of θ is $\mathbb{E}_{\hat{q}(z)} \log p(x, z|\theta)$, while the objective function in mean-field variational method using delta distribution of θ is $\mathbb{E}_{\hat{q}(z)} \log p(x, z, \theta) = \left[\mathbb{E}_{\hat{q}(z)} \log p(x, z|\theta) + \log p(\theta)\right]$.

θ is assumed to be random and prior distribution of θ is considered in mean-field variational method.

Since there are many theoretical supports for EM method, it is reasonable to use delta distribution functions as variational distributions from the view of generalization of EM method. Furthermore, what's even better is that the estimate of θ would be more robust than variational EM method because prior distribution is considered in mean-field variational method.

3.3 Simulated annealing

To mitigate a problem finding local optima and to change the number of features in variational method, stochastic procedure is needed. In particular, we use simulated annealing method.

Annealing is referred to as tempering materials by heating and cooling in order to make materials good quality. The simulation with a procedure of annealing is called simulated annealing (SA). This is a kind of algorithm to find the global optimum using probabilistic technique, but it is heuristic to approximate global optimum in a large space. Sometimes, finding an approximate global optimum is more important than finding the precise local optimum, SA may be preferable to alternatives in some optimization problems.

SA is composed of two stochastic procedure: one for the generation of a candidate solution and the other for the acceptance of a candidate. T is a control parameter called temperature, which controls the size of perturbations of the energy function E. The probability of a state change is determined by a energy difference of two states, $P = e^{-\frac{\Delta E}{T}}$. Details of an algorithm of SA is

as follows.

1. Randomize $s(0)$.
2. Initialize T with a lage value.
3. Repeat:
   (a) Add a random perturbation to the state $s^* = s(i) + \epsilon$.
   (b) Evaluate $\Delta E = E(s^*) - E(s(i))$:
      - If $\Delta E > 0, s(i+1) = s^*$;
      - otherwise, accept the new state $s(i+1) = s^*$ with probability $P = e^{-\frac{\Delta E}{T}}$.
   (c) Set $T = T - \Delta T$.

   Until T is small enough.

If T decreases too rapidly, SA may find a local minimum. On the other hands, if it is reduced too slow, SA converges very slowly. This means that SA explores parameter space at high temperatures, while it restricts exploration at lower temperatures. In practice, T is usually applied as a follwing schedule (Geman and Geman, 1984):

$$T(t) = \alpha T(t-1) \quad \text{with} \quad 0.85 \leq \alpha \leq 0.96.$$

3.4 Stochastic variational method for the LARMuK model

3.4.1 The ELBO

For the varitional method of the LARMuK model, we consider the following variational distributions for β, χ, and λ.

$$\begin{aligned} q(\beta_k) &\equiv \mathrm{N}(\mu_{0k}, \sigma_{0k}), \\ q(\chi_k) &\equiv \delta_{\chi_{0k}}(\chi_k), \\ q(\lambda_k) &\equiv \delta_{\lambda_{0k}}(\lambda_k). \end{aligned}$$

We set variational distributions for other variables as

$$\begin{aligned} q(c_k) &\equiv \mathrm{Cat}(\nu_{k0}, \nu_{k1}, \nu_{k2}), \\ q(\sigma^2) &\equiv \mathrm{IG}\left(\frac{r_0}{2}, \frac{r_0 R_0}{2}\right), \end{aligned}$$

and

$$q(J) \equiv \delta_{J_0}(J).$$

The dimension of variables may be changed in the LARMuK model, we need to consider the number of features J as another variable, thus the ELBO

of LARMuK model is defined as

$$\begin{aligned} L(q) &\equiv \mathbb{E}_q \log p(\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J, \mathbf{Y}) + H_q(\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J) \\ &= \mathbb{E}_q \log p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J) + \mathbb{E}_q \log p(\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J) \\ &+ H_q(\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J). \end{aligned} \tag{3.4}$$

The first term of right side is proportional to the following expression.

$$\begin{aligned} \mathbb{E}_q \log p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J) &\propto \mathbb{E}_q \left[-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{g_i}^T \boldsymbol{\beta})^2 \right] \\ &\propto \frac{n}{2} \mathbb{E}_q \log \frac{1}{\sigma^2} - \frac{1}{2} \mathbb{E}_q \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbb{E}_q (y_i - \mathbf{g_i}^T \boldsymbol{\beta})^2 \\ &\propto \frac{n}{2} \left(\Psi\left(\frac{r_0}{2}\right) - \log \frac{r_0 R_0}{2} \right) \\ &\quad - \frac{1}{2R_0} \sum_{i=1}^{n} \mathbb{E}_q (y_i - \mathbf{g_i}^T \boldsymbol{\beta})^2, \end{aligned}$$

where Ψ is the digamma function and $\mathbf{g_i}$ denotes $(g_{c_1}(x_i, w_1), \cdots, g_{c_J}(x_i, w_J))^T$.

For the second term of right side, since variables $\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}$, and σ^2 are determined independently for given J,

$$\begin{aligned} & \mathbb{E}_q \log p(\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J) \\ = \; & \mathbb{E}_q \log p(\boldsymbol{\beta}|J) + \mathbb{E}_q \log p(\boldsymbol{\chi}|J) + \mathbb{E}_q \log p(\boldsymbol{\lambda}|J) + \mathbb{E}_q \log p(\mathbf{c}|J) + \mathbb{E}_q \log p(\sigma^2) \\ & + \mathbb{E}_q \log p(J). \end{aligned}$$

Note that $\mathbb{E}_q \log p$ depends on parameters of underlying model p and parameters of the variational distribution q. The details of all terms are as follows.

$$\begin{aligned}\mathbb{E}_q \log p(\boldsymbol{\beta}|J) &= \mathbb{E}_q \left[-\frac{J_0}{2} \log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2} \sum_{j=1}^{J_0} (\beta_j - 0)^2 \right] \\ &\propto \frac{J_0}{2} \log \frac{1}{\sigma_\beta^2} - \frac{1}{2\sigma_\beta^2} \sum_{j=1}^{J_0} (\sigma_{0j}^2 + \mu_{0j}^2).\end{aligned} \tag{3.5}$$

$$\begin{aligned}\mathbb{E}_q \log p(\boldsymbol{\chi}|J) &= \sum_{j=1}^{J_0} \mathbb{E}_q \log p(\chi_j) \\ &= \sum_{j=1}^{J_0} \mathbb{E}_q [\log I(0 \leq \chi_j \leq 1)] \\ &= \sum_{j=1}^{J_0} \log I(0 \leq \chi_{0j} \leq 1) \\ &= 0.\end{aligned} \tag{3.6}$$

$$\begin{aligned}\mathbb{E}_q \log p(\boldsymbol{\lambda}|J) &= \sum_{j=1}^{J_0} \mathbb{E}_q \log p(\lambda_j) \\ &= \sum_{j=1}^{J_0} \mathbb{E}_q [\log b_\lambda^{a_\lambda} - \log \Gamma(a_\lambda) + (a_\lambda - 1) \log \lambda_j - b_\lambda \lambda_j] \\ &\propto \sum_{j=1}^{J_0} [(a_\lambda - 1) \mathbb{E}_q \log \lambda_j - b_\lambda \mathbb{E}_q \lambda_j] \\ &= \sum_{j=1}^{J_0} [(a_\lambda - 1) \log \lambda_{0j} - b_\lambda \lambda_{0j}] \\ &= (a_\lambda - 1) \sum_{j=1}^{J_0} \log \lambda_{0j} - b_\lambda \sum_{j=1}^{J_0} \lambda_{0j}.\end{aligned} \tag{3.7}$$

$$
\begin{aligned}
\mathbb{E}_q \log p(\mathbf{c}|J) &= \sum_{j=1}^{J_0} \mathbb{E}_q \log p(c_j) \\
&= \sum_{j=1}^{J_0} [I(c_j = 0) \cdot \log p_0 + I(c_j = 1) \cdot \log p_1 + I(c_j = 2) \cdot \log p_2] \\
&\propto \sum_{j=1}^{J_0} [\nu_{j0} \log p_0 + \nu_{j1} \log p_1 + \nu_{j2} \log p_2] \\
&= \log p_0 \sum_{j=1}^{J_0} \nu_{j0} + \log p_1 \sum_{j=1}^{J_0} \nu_{j1} + \log p_2 \sum_{j=1}^{J_0} \nu_{j2}.
\end{aligned}
\tag{3.8}
$$

$$
\begin{aligned}
\mathbb{E}_q \log p(\sigma^2) &= \mathbb{E}_q \left[\log \frac{rR^{r/2}}{2} - \log \Gamma \left(\frac{r}{2} \right) + \left(\frac{r}{2} - 1 \right) \log \frac{1}{\sigma^2} - \frac{rR}{2\sigma^2} \right] \\
&\propto \left(\frac{r}{2} - 1 \right) \mathbb{E}_q \log \frac{1}{\sigma^2} - \frac{rR}{2} \mathbb{E}_q \frac{1}{\sigma^2} \\
&\propto \left(\frac{r}{2} - 1 \right) \left(\Psi \left(\frac{r_0}{2} \right) - \log \frac{r_0 R_0}{2} \right) - \frac{rR}{2R_0}.
\end{aligned}
\tag{3.9}
$$

$$
\begin{aligned}
\mathbb{E}_q \log p(J) &= \mathbb{E}_q [J \log M - M - \log J!] \\
&\propto J_0 \log M - \log J_0!.
\end{aligned}
\tag{3.10}
$$

The third term of right side of equation (3.4) is the entropy of distribution q. Since the entropy is defined as $H_q(\theta) \equiv -\int \log q(\theta) q(\theta) d\theta$, it depens on the parameters which are involved in the variational distribution q. In mean-field variational method for the LARMuK model, we need conditional entropies for $\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}$, and $\mathbf{c}$.

For the random variables X and Y, the conditional entropy $H_q(X|Y)$ is defined as

$$H_q(X|Y) \equiv \int \int \log q(x|y) q(x,y) dx dy.$$

With the simple calculation, we can show that $H_q(X|Y) = H_q(X,Y) - H_q(Y)$ where

$$H_q(X,Y) \equiv \int \int \log q(x,y) q(x,y) dx dy$$

and

$$H_q(Y) \equiv \int \log q(y) q(y) dy.$$

Therefore, the conditional entropy for each variable of the LARMuK model are the followings.

$$\begin{aligned}
H_q(\boldsymbol{\beta}|J) &= -\mathbb{E}_q \log q(\boldsymbol{\beta}|J) = -\mathbb{E}_{q(J)} \mathbb{E}_{q(\boldsymbol{\beta}|J)} \log q(\boldsymbol{\beta}|J) \\
&= -\mathbb{E}_{q(J)} \sum_{j=1}^{J} \left[-\frac{1}{2} \log(2\pi\sigma_{0j}^2) - \frac{1}{2\sigma_{0j}^2} \mathbb{E}_q (\beta_j - \mu_{0j})^2 \right] \\
&= \sum_{j=1}^{J_0} \left[\frac{1}{2} \log(2\pi\sigma_{0j}^2) + \frac{1}{2\sigma_{0j}^2} \mathbb{E}_q (\beta_j - \mu_{0j})^2 \right] \\
&= \sum_{j=1}^{J_0} \left[\frac{1}{2} \log(2\pi\sigma_{0j}^2) + \frac{1}{2} \right] \\
&\propto \frac{1}{2} \sum_{j=1}^{J_0} \log \sigma_{0j}^2,
\end{aligned}$$

$$H_q(\boldsymbol{\chi}|J) = -\mathbb{E}_{q(J)}\mathbb{E}_{q(\boldsymbol{\chi}|J)}(\log q(\boldsymbol{\chi}|J)) = -\sum_{j=1}^{J_0} \log I(\chi_j = \chi_{0j}) = 0,$$

$$H_q(\boldsymbol{\lambda}|J) = -\mathbb{E}_{q(J)}\mathbb{E}_{q(\boldsymbol{\lambda}|J)}(\log q(\boldsymbol{\lambda}|J)) = -\sum_{j=1}^{J_0} \log I(\lambda_j = \lambda_{0j}) = 0,$$

$$\begin{aligned} H_q(\mathbf{c}|J) &= -\mathbb{E}_q \log q(\mathbf{c}|J) = -\mathbb{E}_{q(J)}\mathbb{E}_{q(\mathbf{c}|J)} \log q(\mathbf{c}|J) \\ &= -\sum_{j=1}^{J_0} \mathbb{E}_q(\log q(c_j)) \\ &= -\sum_{j=1}^{J_0} [\nu_{j0} \log \nu_{j0} + \nu_{j1} \log \nu_{j1} + \nu_{j2} \log \nu_{j2}], \end{aligned}$$

$$\begin{aligned} H_q(\sigma^2) &= -\mathbb{E}_q \left[\log \left(\frac{r_0 R_0}{2} \right)^{\frac{r_0}{2}} - \log \Gamma \left(\frac{r_0}{2} \right) + \left(\frac{r_0}{2} - 1 \right) \log \frac{1}{\sigma^2} - \frac{r_0 R_0}{2\sigma^2} \right] \\ &= \log \Gamma \left(\frac{r_0}{2} \right) - \frac{r_0}{2} \log \frac{r_0 R_0}{2} - \left(\frac{r_0}{2} - 1 \right) \mathbb{E}_q \log \frac{1}{\sigma^2} + \frac{r_0 R_0}{2} \mathbb{E}_q \frac{1}{\sigma^2} \\ &= \log \Gamma \left(\frac{r_0}{2} \right) - \frac{r_0}{2} \log \frac{r_0 R_0}{2} - \left(\frac{r_0}{2} - 1 \right) \left(\Psi \left(\frac{r_0}{2} \right) - \log \frac{r_0 R_0}{2} \right) + \frac{r_0}{2} \\ &= \log \Gamma \left(\frac{r_0}{2} \right) - \log \frac{r_0 R_0}{2} - \left(\frac{r_0}{2} - 1 \right) \Psi \left(\frac{r_0}{2} \right) + \frac{r_0}{2}, \end{aligned}$$

$$H_q(J) = -\mathbb{E}_q \log q(J) = -\log I(J = J_0) = 0.$$

For the simple expression for the ELBO, we may assume configuration of generating function follows a multinomial distribution with trial 1. This is

equivalent to c_j following a categorical distribution, i.e., we uses expression,

$$\mathbf{c_j} \equiv (c_{j0}, c_{j1}, c_{j2})^T \sim \text{Multi}(1, \nu_{j0}, \nu_{j1}, \nu_{j2})$$

instead of $c_j \sim \text{Cat}(\nu_{j0}, \nu_{j1}, \nu_{j2})$. Then

$$\mathbb{E}[c_{jk}^2] = \nu_{jk}, \mathbb{E}[c_{jk_1} c_{jk_2}] = 0 (k_1 \neq k_2)$$

hold since

$$\mathbb{E}[c_{jk}] = \nu_{jk}, \text{Var}[c_{jk}] = \nu_{jk}(1 - \nu_{jk}), \text{Cov}[c_{jk_1}, c_{jk_2}] = -\nu_{jk_1}\nu_{jk_2} (k_1 \neq k_2).$$

When assuming delta distribution functions as variational distributions for χ_j and λ_j, for all l,

$$\mathbb{E}_q g_l^2(x_i, w_j) = \int g_l^2(x_i; \chi_j, \lambda_j) q(\chi_j) q(\lambda_j) d\chi_j d\lambda_j = g_l^2(x_i; \chi_{0j}, \lambda_{0j}),$$

where $\mathbb{E}_q$ means expectation with respect to $q(\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J)$.

For the convenience, for $l = 0, 1, 2$, let

$$g_{ijl} \equiv g_l(x_i; \chi_j, \lambda_j), \tag{3.11}$$

$$g_{0ijl} \equiv \mathbb{E}_q[g_{ijl}] = g_l(x_i; \chi_{0j}, \lambda_{0j}), \tag{3.12}$$

$$g_{ij} \equiv c_{j0} \cdot g_{ij0} + c_{j1} \cdot g_{ij1} + c_{j2} \cdot g_{ij2}. \tag{3.13}$$

In addition, let $\mathbb{E}_q[g_{ij}]$ and $\mathbb{E}_q[g_{ij}^2]$ be g_{0ij} and g_{0ij}^2, respectively. Then, they are represented as

$$
\begin{aligned}
g_{0ij} &\equiv \mathbb{E}_q[g_{ij}] \\
&= \mathbb{E}_q[c_{j0} \cdot g_{ij0} + c_{j1} \cdot g_{ij1} + c_{j2} \cdot g_{ij2}] \\
&= \nu_{j0} \cdot g_{0ij0} + \nu_{j1} \cdot g_{0ij1} + \nu_{j2} \cdot g_{0ij2},
\end{aligned}
$$

$$
\begin{aligned}
g^2_{0ij} &\equiv \mathbb{E}_q[g^2_{ij}] \\
&= \mathbb{E}_q[c_{j0} \cdot g_{ij0} + c_{j1} \cdot g_{ij1} + c_{j2} \cdot g_{ij2}]^2 \\
&= \mathbb{E}_q[c^2_{j0} g^2_{ij0} + c^2_{j1} g^2_{ij1} + c^2_{j2} g^2_{ij2} + 2(c_{j0} c_{j1} g_{ij0} g_{ij1} + c_{j1} c_{j2} g_{ij1} g_{ij2} + c_{j0} c_{j2} g_{ij0} g_{ij2})] \\
&= \mathbb{E}_q[c^2_{j0}] \cdot \mathbb{E}_q[g^2_{ij0}] + \mathbb{E}_q[c^2_{j1}] \cdot \mathbb{E}_q[g^2_{ij1}] + \mathbb{E}_q[c^2_{j2}] \cdot \mathbb{E}_q[g^2_{ij2}] \\
&+ 2(\mathbb{E}_q[c_{j0} c_{j1}] \cdot \mathbb{E}_q[g_{ij0} g_{ij1}] + \mathbb{E}_q[c_{j1} c_{j2}] \cdot \mathbb{E}_q[g_{ij1} g_{ij2}] + \mathbb{E}_q[c_{j0} c_{j2}] \cdot \mathbb{E}_q[g_{ij0} g_{ij2}]) \\
&= \nu_{j0} \cdot g^2_{0ij0} + \nu_{j1} \cdot g^2_{0ij1} + \nu_{j2} \cdot g^2_{0ij2}
\end{aligned}
$$

Furthermore, when $j \neq l$, all of the elements of generating functions are totally independent,

$$
\begin{aligned}
\mathbb{E}_q[g_{ij} g_{il}] &= \mathbb{E}_q[(c_{j0} g_{ij0} + c_{j1} g_{ij1} + c_{j2} g_{ij2})(c_{l0} g_{il0} + c_{l1} g_{il1} + c_{l2} g_{il2})] \\
&= \mathbb{E}_q[g_{ij}] \mathbb{E}_q[g_{il}] \\
&= g_{0ij} g_{0il}.
\end{aligned}
$$

Finally, we get

$$\mathbb{E}_q[\mathbf{g_i}^T\boldsymbol{\beta}] = \mathbb{E}_q[\mathbf{g_i}]^T\mathbb{E}_q[\boldsymbol{\beta}] = \sum_{j=1}^{J_0}\mu_{0j}g_{0ij}, \tag{3.14}$$

$$\begin{aligned}
\mathbb{E}_q[(\mathbf{g_i}^T\boldsymbol{\beta})^2] &= \sum_{j=1}^{J_0}\mathbb{E}_q[\beta_j^2 g_{ij}^2] + \sum_{l,j,l\neq j}\mathbb{E}_q[\beta_j g_{ij}\beta_l g_{il}] \\
&= \sum_{j=1}^{J_0}\mathbb{E}_q[\beta_j^2]\mathbb{E}_q[g_{ij}^2] + \sum_{l,j,l\neq j}\mathbb{E}_q[\beta_j\beta_l]\mathbb{E}_q[g_{ij}g_{il}] \\
&= \sum_{j=1}^{J_0}[\sigma_{0j}^2+\mu_{0j}^2]\mathbb{E}_q[g_{ij}^2] + \sum_{l,j,l\neq j}[\mu_{0j}\mu_{0l}]\mathbb{E}_q[g_{ij}g_{il}] \\
&= \sum_{j=1}^{J_0}[\sigma_{0j}^2+\mu_{0j}^2]g_{0ij}^2 + \sum_{l,j,l\neq j}[\mu_{0j}\mu_{0l}]g_{0ij}g_{0il},
\end{aligned} \tag{3.15}$$

and we evaluate the expectation,

$$\begin{aligned}
& \mathbb{E}_q\left[\sum_{i=1}^{n}(y_i - \mathbf{g_i}^T\boldsymbol{\beta})^2\right] \\
= \;& \mathbb{E}_q\left[\sum_{i=1}^{n}y_i^2 - 2\sum_{i=1}^{n}y_i(\mathbf{g_i}^T\boldsymbol{\beta}) + \sum_{i=1}^{n}(\mathbf{g_i}^T\boldsymbol{\beta})^2\right] \\
= \;& \sum_{i=1}^{n}y_i^2 - 2\sum_{i=1}^{n}y_i\mathbb{E}_q[\mathbf{g_i}^T\boldsymbol{\beta}] + \sum_{i=1}^{n}\mathbb{E}_q[(\mathbf{g_i}^T\boldsymbol{\beta})^2] \\
= \;& \sum_{i=1}^{n}y_i^2 - 2\sum_{i=1}^{n}y_i\left[\sum_{j=0}^{J_0}\mu_{0j}g_{0ij}\right] + \sum_{i=1}^{n}\left[\sum_{j=1}^{J_0}[\sigma_{0j}^2 g_{0ij}^2 + \mu_{0j}^2 g_{0ij}^2] + \sum_{l,j,l\neq j}[\mu_{0j}\mu_{0l}g_{0ij}g_{0il}]\right] \\
= \;& \sum_{i=1}^{n}[y_i^2 - 2y_i(\mathbf{g_{0i}}^T\boldsymbol{\mu}_0) + (\mathbf{g_{0i}}^T\boldsymbol{\mu}_0)^2 + \sum_{j=1}^{J_0}\sigma_{0j}^2 g_{0ij}^2] \\
= \;& \sum_{i=1}^{n}(y_i - \mathbf{g_{0i}}^T\boldsymbol{\mu}_0)^2 + \sum_{j=1}^{J_0}\sum_{i=1}^{n}\sigma_{0j}^2 g_{0ij}^2.
\end{aligned}$$

Here $\mathbf{g_{0i}}$ denote $(g_{0i1}, g_{0i2}, \cdots, g_{0iJ_0})^T$ and $\boldsymbol{\mu}_0 = (\mu_{01}, \cdots, \mu_{0J_0})^T$.

In summary, the ELBO is

$$
\begin{aligned}
L(q) \;=\; & \left[\frac{n}{2}\left(-\log 2\pi + \Psi\left(\frac{r_0}{2}\right) - \log\frac{r_0 R_0}{2}\right) - \frac{1}{2R_0}\left(\sum_{i=1}^{n}(y_i - \mathbf{g_{0i}}^T\boldsymbol{\mu}_0)^2 + \sum_{j=1}^{J_0}\sum_{i=1}^{n}\sigma_{0j}^2 g_{0ij}^2\right)\right] \\
& - \left[\frac{J_0}{2}\log 2\pi\sigma_\beta^2 + \frac{1}{2\sigma_\beta^2}\sum_{j=1}^{J_0}(\mu_{0j}^2 + \sigma_{0j}^2)\right] \\
& + \left[J_0(a_\lambda \log b_\lambda - \log\Gamma(a_\lambda)) + (a_\lambda - 1)\sum_{j=1}^{J_0}\log\lambda_{0j} - b_\lambda\sum_{j=1}^{J_0}\lambda_{0j}\right] \\
& + \left[\log p_0\sum_{j=1}^{J_0}\nu_{j0} + \log p_1\sum_{j=1}^{J_0}\nu_{j1} + \log p_2\sum_{j=1}^{J_0}\nu_{j2}\right] \\
& + \left[\frac{r}{2}\log\frac{rR}{2} - \log\Gamma\left(\frac{r}{2}\right) - \left(\frac{r}{2} - 1\right)\left(\Psi\left(\frac{r_0}{2}\right) - \log\frac{r_0 R_0}{2}\right) - \frac{rR}{2R_0}\right] \\
& + [J_0\log M - \log J_0! - M] \\
& + \sum_{j=1}^{J_0}\left[\frac{1}{2}\log 2\pi\sigma_{0j}^2 + \frac{1}{2}\right] \\
& - \sum_{j=1}^{J_0}(\nu_{j0}\log\nu_{j0} + \nu_{j1}\log\nu_{j1} + \nu_{j2}\log\nu_{j2}) \\
& + \log\Gamma\left(\frac{r_0}{2}\right) - \log\frac{r_0 R_0}{2} - \left(\frac{r_0}{2} - 1\right)\Psi\left(\frac{r_0}{2}\right) + \frac{r_0}{2}.
\end{aligned}
$$

3.4.2 Updating variational parameters

For $\boldsymbol{\chi}$ and $\boldsymbol{\lambda}$, which are located inside generating functions, updating variational parameters is intractable eventhough variational distributions of these variables are assumed as delta distribution functions. For the case of χ_k, the

function to be maximized is

$$\begin{aligned} & \mathbb{E}_{-q(\chi_k)} \log p(x|\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{c}, \sigma^2, J) + \log p(\chi_k) \\ \propto \quad & \mathbb{E}_{q(\boldsymbol{\chi}_{-k}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{c}, \sigma^2, J)} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} g_{c_j}(x_i; \chi_j, \lambda_j) \beta_j \right)^2 \right], \chi_k \in (0,1), \end{aligned}$$

Therefore, the objective function is

$$\text{objF}(q(\chi_k)) \quad = \quad -\frac{1}{2R_0} \left[(\sigma_{0k}^2 + \mu_{0k}^2) \sum_{i=1}^{n} g_{0ik*}^2 - 2\mu_{0k} \sum_{i=1}^{n} g_{0ik*} (y_i - \sum_{j \neq k} \mu_{0j} g_{0ij}) + \sum_{i=1}^{n} y_i^2 \right],$$

where $g_{0ik*} = \nu_{k0} \cdot g_0(x_i; \chi_k, \lambda_{0k}) + \nu_{k1} \cdot g_1(x_i; \chi_k, \lambda_{0k}) + \nu_{k2} \cdot g_2(x_i; \chi_k, \lambda_{0k})$.

Likewise, the object function for λ_k is

$$\begin{aligned} & \text{objF}(q(\lambda_k)) \\ = \quad & \mathbb{E}_{-q(\lambda_k)} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{g_i}^T \boldsymbol{\beta})^2 + (a_\lambda - 1) \log \lambda_k - b_\lambda \lambda_k \right] \\ = \quad & -\frac{1}{2R_0} \left[(\sigma_{0k}^2 + \mu_{0k}^2) \sum_{i=1}^{n} g_{0ik*}^2 - 2\mu_{0k} \sum_{i=1}^{n} g_{0ik*} (y_i - \sum_{j \neq k} \mu_{0j} g_{0ij}) + \sum_{i=1}^{n} y_i^2 \right] \\ & + (a_\lambda - 1) \log \lambda_k - b_\lambda \lambda_k, \end{aligned}$$

where $g_{0ik*} = \nu_{k0} \cdot g_0(x_i; \chi_{0k}, \lambda_k) + \nu_{k1} \cdot g_1(x_i; \chi_{0k}, \lambda_k) + \nu_{k2} \cdot g_2(x_i; \chi_{0k}, \lambda_k)$. Since we assume that the variational distribution for J is also a delta distribution function, the expection of J may not be considered at all in these objective functions. It can be treated as a fixed value.

The optimal χ_{0k}, λ_{0k} will be obtained as the maximizer of each objective function. It is difficult to use conventional optimization methods because a form of objective function is hard to handle. Objective functions of χ_k and λ_k

may have discontinuities. The simple way is to discretize supports of χ_k, λ_k and to find the approximates for the optimal χ_{0k}, λ_{0k}.

For $\mathbf{c}$, since we assumed the variational distribution of c_k as $\mathrm{Cat}(\nu_{k0}, \nu_{k1}, \nu_{k2})$, the optimal variational parameters can be obtained by

$$\begin{aligned}
\log Q(c_k = l) &\propto \mathbb{E}_{-q(c_k)}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{g_i}^T\boldsymbol{\beta})^2\right] + \log p_l \\
&\propto -\frac{1}{2R_0}\mathbb{E}_{-q(c_k)}\left[\sum_{i=1}^{n}(y_i - \mathbf{g_i}^T\boldsymbol{\beta})^2\right] + \log p_l \\
&\propto -\frac{1}{2R_0}\left[\sum_{i=1}^{n}(y_i - \mathbf{g_{0i}^k}^T\boldsymbol{\mu}_0)^2 + \sum_{j\neq k}\sum_{i=1}^{n}\sigma_{0j}^2 g_{0ij}^2 + \sum_{i=1}^{n}\sigma_{0k}^2 g_{0ikl}^2\right] \\
&\quad + \log p_l,
\end{aligned}$$

where $\mathbf{g_{0i}^k}$ denotes that k-th element in $\mathbf{g_{0i}}$ has been changed to g_{0ikl}. Note that unlike other elements of $\mathbf{g_{0i}^k}$, k-th element g_{0ikl} is not a multiple kernel but a single kernel, i.e.,

$$\begin{aligned}
g_{0ikl} &= g_l(x_i; \chi_{0k}, \lambda_{0k}) \\
&\neq \nu_{k0}\cdot g_0(x_i; \chi_{0k}, \lambda_{0k}) + \nu_{k1}\cdot g_1(x_i; \chi_{0k}, \lambda_{0k}) + \nu_{k2}\cdot g_2(x_i; \chi_{0k}, \lambda_{0k}), \\
g_{0ikl}^2 &= g_l^2(x_i; \chi_{0k}, \lambda_{0k}) \\
&\neq \nu_{k0}\cdot g_0^2(x_i; \chi_{0k}, \lambda_{0k}) + \nu_{k1}\cdot g_1^2(x_i; \chi_{0k}, \lambda_{0k}) + \nu_{k2}\cdot g_2^2(x_i; \chi_{0k}, \lambda_{0k}).
\end{aligned}$$

For the case of β_k and σ^2, variational parameters could be easily obtained by using conventional way. Variational parameters of β_k is obtained by the

following procedure.

$$
\begin{aligned}
\log q(\beta_k) &\propto \mathbb{E}_{-q(\beta_k)}\left[\sum_{i=1}^{n}\log p(y_i|x_i,\beta)+\log p(\beta_k)\right] \\
&\propto \mathbb{E}_{-q(\beta_k)}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mathbf{g_i}^T\boldsymbol{\beta})^2\right]-\frac{1}{2\sigma_\beta^2}\beta_k^2 \\
&= \mathbb{E}_{-q(\beta_k)}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-g_{ik}\beta_k-\mathbf{g_{i,-k}}^T\boldsymbol{\beta}_{-k})^2-\frac{1}{2\sigma_\beta^2}\beta_k^2\right] \\
&\propto \mathbb{E}_{-q(\beta_k)}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(g_{ik}^2\beta_k^2-2(y_i-\mathbf{g_{i,-k}}^T\boldsymbol{\beta}_{-k})g_{ik}\beta_k)-\frac{1}{2\sigma_\beta^2}\beta_k^2\right] \\
&= -\mathbb{E}_{-q(\beta_k)}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{n}(g_{ik}^2\beta_k^2-2(y_i-\mathbf{g_{i,-k}}^T\boldsymbol{\beta}_{-k})g_{ik}\beta_k)\right]-\frac{1}{2\sigma_\beta^2}\beta_k^2 \\
&= -\frac{1}{2R_0}\sum_{i=1}^{n}(\beta_k^2\mathbb{E}_{-q(\beta_k)}[g_{ik}^2]-2(y_i-\mathbb{E}_{-q(\beta_k)}[\mathbf{g_{i,-k}}^T\boldsymbol{\beta}_{-k}])\beta_k\mathbb{E}_{-q(\beta_k)}[g_{ik}])-\frac{1}{2\sigma_\beta^2}\beta_k^2 \\
&\propto -\frac{1}{2R_0}(\beta_k^2\sum_i g_{0ik}^2-2\beta_k\sum_i(y_i-\mathbf{g_{0i,-k}}^T\boldsymbol{\mu}_{0,-k})g_{0ik})-\frac{1}{2\sigma_\beta^2}\beta_k^2 \\
&= -\frac{1}{2}\left(\beta_k^2\left(\frac{\sum_{i=1}^n g_{0ik}^2}{R_0}+\frac{1}{\sigma_\beta^2}\right)-2\beta_k\frac{\sum_{i=1}^n(y_i-\mathbf{g_{0i,-k}}^T\boldsymbol{\mu}_{0,-k})g_{0ik}}{R_0}\right),
\end{aligned}
$$

variational parameters (μ_{0j},σ_{0j}^2) for β_k are

$$\frac{1}{\sigma_{0k}^2}=\frac{1}{\sigma_\beta^2}+\frac{1}{R_0}\sum_i g_{0ik}^2$$

$$\mu_{0k}=\sigma_{0k}^2\cdot\frac{\sum_{i=1}^n(y_i-\mathbf{g_{0i,-k}}^T\boldsymbol{\mu}_{0,-k})g_{0ik}}{R_0},$$

where subscript $-k$ denote that k-th element is eliminated. For σ^2, since

$$\begin{aligned}
\log q(\sigma^2) &\propto \mathbb{E}_{-q(\sigma^2)}\left[\sum_{i=1}^{n}\log p(y_i|x_i,\sigma^2)+\log p(\sigma^2)\right] \\
&\propto \frac{n}{2}\log\frac{1}{\sigma^2}-\frac{1}{2\sigma^2}\mathbb{E}_{-q(\sigma^2)}\left[\sum_{i=1}^{n}(y_i-\mathbf{g_i}^T\boldsymbol{\beta})^2\right]+\left(\frac{r}{2}-1\right)\log\frac{1}{\sigma^2}-\frac{rR}{2\sigma^2} \\
&= \left(\frac{n+r}{2}-1\right)\log\frac{1}{\sigma^2}-\frac{1}{\sigma^2}\frac{\left(\sum_{i=1}^{n}\mathbb{E}_{-q(\sigma^2)}[(y_i-\mathbf{g_i}^T\boldsymbol{\beta})^2]+rR\right)}{2},
\end{aligned}$$

variational parameters r_0, R_0 for σ^2 are

$$r_0 = n + r,$$

$$R_0 = \frac{\sum_{i=1}^{n}(y_i-\mathbf{g_{0i}}^T\boldsymbol{\mu}_0)^2+\sum_{j=0}^{J_0}\sum_{i=1}^{n}\sigma_{0j}^2 g_{0ij}^2+rR}{r_0}.$$

The number of features J would be obtained in a different way from other variables. Most of variational methods for nonparametric Bayesian models involving J consider J as a fixed value, however, we suggest a method for inferring J in this paper. Denote $\boldsymbol{\theta}$ as all the variables except J. When the prior distribution of J is Poi(M) and the variational distribution of J is assumed to

be $\delta_{J_0}(J)$, the objective function of J is

$$
\begin{aligned}
\text{objF}(q(J)) &\propto \mathbb{E}_q \log[p(x|\boldsymbol{\theta})p(\boldsymbol{\theta}|J)p(J)] + H_q(\boldsymbol{\theta}|J) + H_q(J) \\
&\propto -\frac{1}{2R_0}\left[\sum_{i=1}^{n}(y_i - \mathbf{g_{0i}}^T\boldsymbol{\mu}_0)^2 + \sum_{j=1}^{J_0}\sum_{i=1}^{n}\sigma_{0j}^2 g_{0ij}^2\right] \\
&\quad - \left[\frac{J_0}{2}\log 2\pi\sigma_\beta^2 + \frac{1}{2\sigma_\beta^2}\sum_{j=1}^{J_0}(\mu_{0j}^2 + \sigma_{0j}^2)\right] \\
&\quad + \left[J_0(a_\lambda \log b_\lambda - \log\Gamma(a_\lambda)) + (a_\lambda - 1)\sum_{j=1}^{J_0}\log\lambda_{0j} - b_\lambda\sum_{j=1}^{J_0}\lambda_{0j}\right] \\
&\quad + \left[\log p_0 \sum_{j=1}^{J_0}\nu_{j0} + \log p_1 \sum_{j=1}^{J_0}\nu_{j1} + \log p_2 \sum_{j=1}^{J_0}\nu_{j2}\right] \\
&\quad + [J_0 \log M - \log J_0!] \\
&\quad + \left[\sum_{j=1}^{J_0}\frac{1}{2}\log(2\pi\sigma_{0j}^2) + \frac{J_0}{2}\right] \\
&\quad - \sum_{j=1}^{J_0}(\nu_{j0}\log\nu_{j0} + \nu_{j1}\log\nu_{j1} + \nu_{j2}\log\nu_{j2}).
\end{aligned}
$$

As one optimizing J in this objective function, it tends to select a constant value of J since variational parameters related to other variables except J were already optimized in the setting of given current J. To improve this, we use simulated annealing method which has a possibility to move even if it is not optimal. The energe function E in SA is objF$(q(J))$, and a random neighborhood of J which we consider is $(J-1, J, J+1)$. Initial temperature is set to depend on the number of observations.

When the number of features increases, the expectations for new features are needed. New expectations for χ could be randomly selected on the covariate space $\mathcal{X}$. For the new expectations of λ we use a small value such as 0.01. There

are two reasons to use this setting. First, it can capture some signals occured in narrow regions. Second, we want a mean function which is obtained with new J not to be significantly different from before J changes. By using these setting, we expect that a mean function converges eventhough the dimension of features changes. We set the new expectations of μ that involves β following Gaussian distribution with mean 0 and variance $4 \times R_0$. This means that a new signal tends to be greater than estimated error.

3.5 Data analysis

In this section, we mainly compare the results of stochastic variational method in LARMuK model with the results of RJMCMC method. We denote stochastic variational method for LARMuK model as V-LMK, and RJMCMC method for LARMuK model is denoted by RJ-LMK. The analysis in simulation data and real data shows the efficiency of stochastic variational method for LARMuK model. Efficiency is considered from the view of computing time until convergence as well as MSE.

There are many local optima for the objective function of mean function, but stochastic variational method may not guarantee convergence to the global optimum. In order to avoid finding a local optimum if possible, we used heuristic way in which we tried repeating an algorithm several times. Randomly selected initial values are used in each repetition, then we choose the best result as the optimizer among the estimates, which make the ELBO large. The repetition can be implemented by using GPU since the repetition does not involved each other. This is the reason that stochastic variational method is

Table 3.1: Computing times (second) until obtaining the estimates in each method.

Method	Bumps	Blocks	Doppler	Blip	Multi	Heavisine
V-LMK	271.54	535.88	296.90	448.51	256.63	514.37
RJ-LMK	10108.28	10997.20	10391.96	10566.16	10342.76	17114.04

more attractive than sampling-based method eventhough stochastic variational method seems to time-consuming.

3.5.1 Simulation data analysis

Bumps, Blocks, Doppler, Blip, Multi, and Heavysine function are used for simulation study. In each example except Heavysine, $n = 128$ and SNR $= 5$ is set. SNR $= 10$ is set in Heavysine data as in the previous chapter. Typically, values of the ELBO depend on the number of observations, so that we set $10 \times n$ as the initial temperature T for simulated annealing. For the cooling rate α we use 0.1.

The computing times until obtaining the estimates are listed in Table 3.1. While the computing times of stochastic variational method are significantly small comparing to times of RJMCMC, performance of stochastic variational method is similar to that of RJMCMC method for the LARMuK model and the LARK model (Table 3.2 and Figure 3.1). Therefore, we can say that the proposed method is competible enough comparing conventional inference method.

Note that the proposed method is not good at analysis for Bumps data. Since variational method is a kind of approximation method, sharp peaks are hard to be captured and the estimates for peaks tend to be underestimated.

Table 3.2: MSE for the estimated mean function of each method. H, L, G are denoted as Haar, Laplacian and Gaussian kernel, respectively.

Method	Bumps	Blocks	Doppler	Blip	Multi	Heavisine
V-LMK	0.249	0.033	0.053	0.013	0.023	0.012
RJ-LMK	0.033	0.013	0.012	0.005	0.008	0.002
RJ-LK	0.031(L)	0.005(H)	0.036(G)	0.035(G)	0.049(G)	0.012(G)

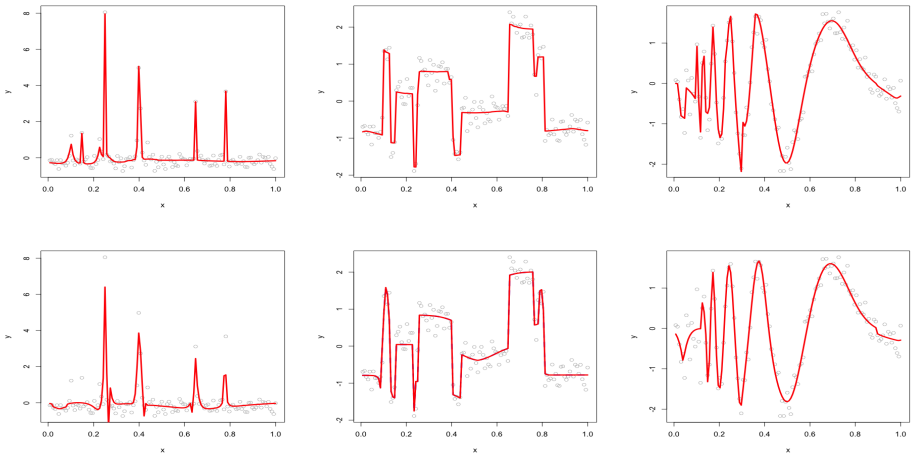


Figure 3.1: First row: estimated Bumps, Blocks, and Doppler function of LMK using RJMCMC method. Second row: estimated Bumps, Blocks, and Doppler function of LMK using stochastic variational method.

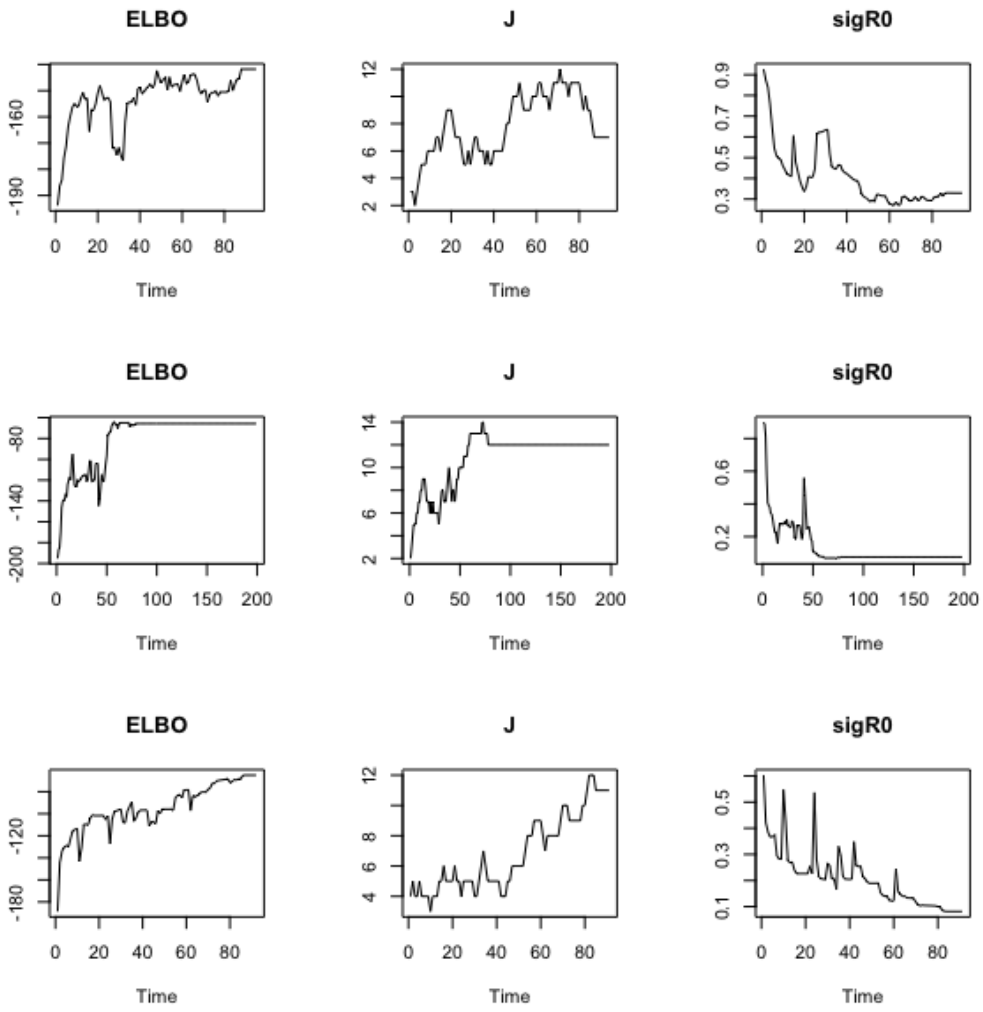


Figure 3.2: First row to third row: the ELBO, the number of features, and σ^2 estimated from Bumps, Blocks, and Doppler data, respectively.

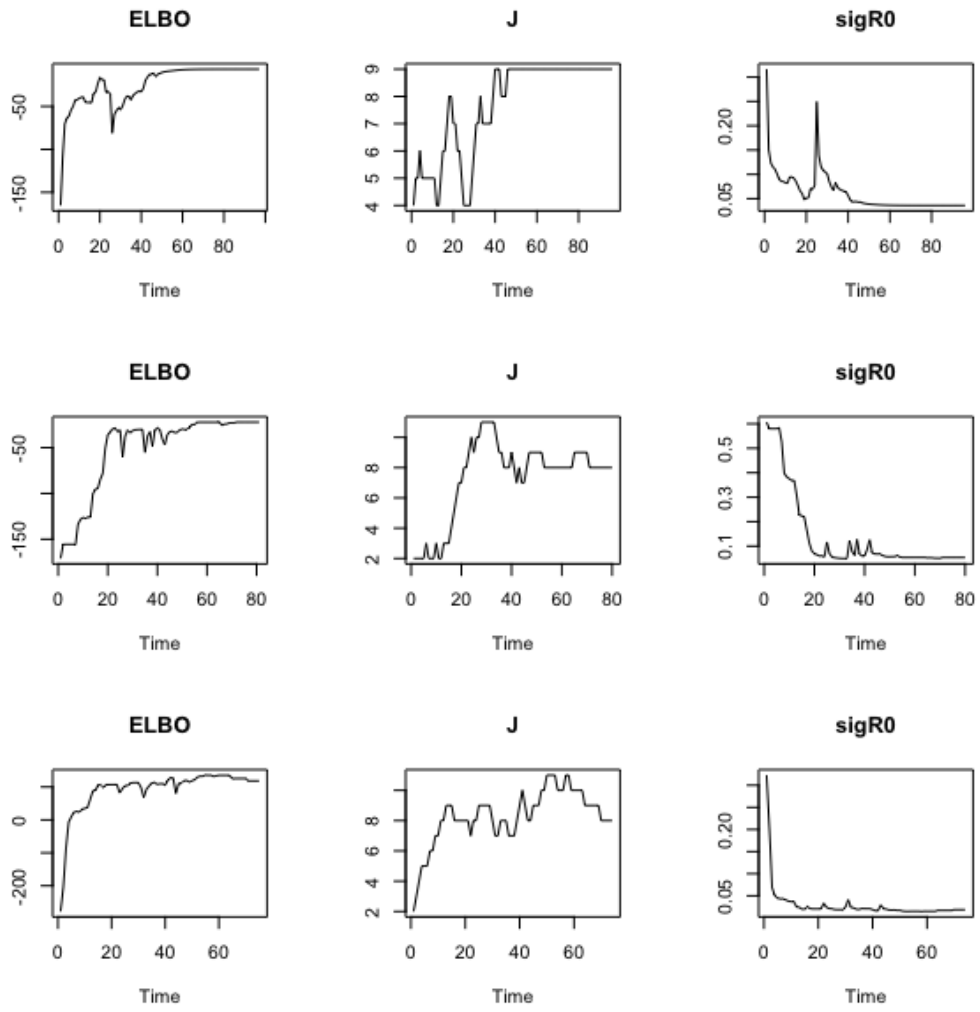


Figure 3.3: First row to third row: the ELBO, the number of features, and σ^2 estimated from Blip, Multi, and Heavysine data, respectively.

Table 3.3: The esimated σ in each example. Stochastic variational method and RJMCMC method are used.

Method	Bumps	Blocks	Doppler	Blip	Multi	Heavisine
True σ	0.20	0.20	0.20	0.20	0.20	0.10
$\hat{\sigma}$ of V-LMK	0.57	0.27	0.25	0.19	0.23	0.16
$\hat{\sigma}$ of RJ-LMK	0.28	0.23	0.25	0.23	0.20	0.10

Figure 3.2 and 3.3 are time series plots for the ELBO, the number of features, and the estimates of σ^2. The ELBO tends to increase and $\hat{\sigma}^2$ converges to the true value of σ^2. True value of σ is 0.2 except Heavysine data. In Heavysine data, the true value of σ is 0.1 corresponding $SNR = 10$. Each estimated σ is listed in Table 3.3. The results are similar to the results of RJMCMC method, which uses exact posterior distribution. It indicates that performance of stochastic variational method is good enough despite it uses approximation of posterior distribution.

3.5.2 Real data analysis

In real data example, we used signal data which we already analyzed in the previous chapter. Figure 3.4 shows that the predicted curve of stochastic variational method are as similar as the one of RJMCMC method. Furthermore, MSE of stochastic variational method are small enough comparing to RJMCMC method.

The ability to capture jumps is important property of the LARMuK model, so that we need to check that stochastic variational method of the LARMuK model can capture jumps well. From Figure 3.4 we find out that stochastic variational method captures the locations of jump well. When the number of

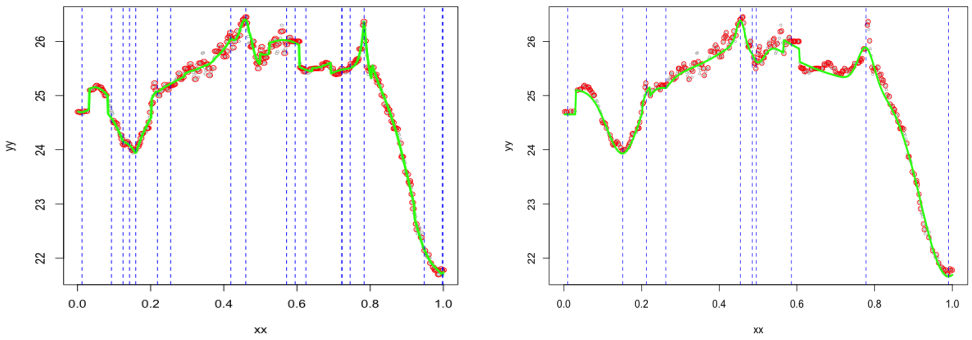


Figure 3.4: Left: using 50% of dataset as test set for validation, curves are predicted using RJMCMC in LARMuK model. Right: curves are predicted using stochastic variational method in LARMuK model. Grey circles indicate original dataset, red circles denote training set, green line is predicted curve. Blue dotted vertical lines denote center of features.

Table 3.4: MSE for the predicted mean function of each method, stochastic variational method and RJMCMC method, respectively. 80% means that 80% of dataset($n = 410$) used for training data and 20% of dataset($n = 102$) used for validation.

Methods	80% ($n = 410$)	50% ($n = 256$)	20% ($n = 102$)
V-LMK	0.0139	0.0111	0.0285
RJ-LMK	0.0074	0.0092	0.0390

observations increases and data are piled near jumps, stochastic variational method will capture jumps more correctly.

To verify the performance of suggested method, we have set up three test sets: randomly chosen 20, 50, 80 percent of dataset. Automatically, the rest is used for training set. MSE was obtaind in each test set. Of course, MSE tends to increase as the percentage of dataset for test set increases. It is remarkable that MSE of stochastic variational method is not much different from MSE of

RJMCMC (Table 3.4).

3.6 Discussion

In this chapter, we proposed stochastic variational method for the LARMuK model. This method can evaluate intractable expectations involving generating functions by using Dirac delta functions as variational distributions of some variables lying inside generation functions. Furthermore, using simulated annealing method as probabilistic procedure in conventional optimization process for variational method, so that we are able to infer the number of features and hidden variables effectively even when the dimension of features changes.

From the data analysis, we found out that the results are similar to the results based on sampling, which uses exact posterior distribution for inference, while computation time is much shorter than that of RJMCMC method.

However, there is a drawback that the speed of computation is not so fast because heuristic method like discretization is used. The methodology for improving the heuristic method should be studied for further researches.

Bibliography

[1] Anestis Antoniadis, Jeremie Bigot, and Theofanis Sapatinas. Wavelet estimators in nonparametric regression: a comparative simulation study. *Journal of Statistical Software*, 6:pp–1, 2001.

[2] David M Blei and Michael I Jordan. Variational methods for the Dirichlet process. In *Proceedings of the twenty-first international conference on Machine learning*, page 12. ACM, 2004.

[3] François Caron, Manuel Davy, and Arnaud Doucet. Generalized Polya urn for time-varying Dirichlet process mixtures. *arXiv preprint arXiv:1206.5254*, 2012.

[4] DGT Denison, BK Mallick, and AFM Smith. Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):333–350, 1998.

[5] Ilaria DiMatteo, Christopher R Genovese, and Robert E Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

[6] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[7] Finale Doshi, Kurt Miller, Jurgen V Gael, and Yee W Teh. Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 137–144, 2009.

[8] David B Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

[9] David B Dunson, Natesh Pillai, and Ju-Hyun Park. Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183, 2007.

[10] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[11] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

[12] Irene Gijbels and Anne-Cécile Goderniaux. Bandwidth selection for changepoint estimation in nonparametric regression. *Technometrics*, 46(1):76–86, 2004.

[13] Irene Gijbels, Alexandre Lambert, and Peihua Qiu. Jump-preserving regression and smoothing using local linear fitting: a compromise. *Annals of the Institute of Statistical Mathematics*, 59(2):235, 2007.

[14] Tristan Gray-Davies, Chris C Holmes, François Caron, et al. Scalable Bayesian nonparametric regression via a Plackett-Luce model for conditional ranks. *Electronic Journal of Statistics*, 10(2):1807–1828, 2016.

[15] Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[16] Jim E Griffin and MF J Steel. Order-based dependent Dirichlet processes. *Journal of the American statistical Association*, 101(473):179–194, 2006.

[17] Alexander Grossmann and Jean Morlet. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis*, 15(4):723–736, 1984.

[18] Lauren A Hannah, David M Blei, and Warren B Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12(Jun):1923–1953, 2011.

[19] Kee-Hoon Kang, Ja-Yong Koo, and Cheol-Woo Park. Kernel estimation of discontinuous regression functions. *Statistics & probability letters*, 47(3):277–285, 2000.

[20] Ja-Yong Koo. Spline estimation of discontinuous regression functions. *Journal of Computational and Graphical Statistics*, 6(3):266–284, 1997.

[21] Steven N MacEachern. Dependent Dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, pages 1–40, 2000.

[22] Hans-Georg Muller. Change-points in nonparametric regression analysis. *The Annals of Statistics*, pages 737–761, 1992.

[23] Peter Muller, Alaattin Erkanli, and Mike West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, pages 67–79, 1996.

[24] Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684, 2000.

[25] Natesh S Pillai. *Lévy random measures: Posterior consistency and applications*. Duke University, 2008.

[26] Natesh S Pillai, Qiang Wu, Feng Liang, Sayan Mukherjee, and Robert L Wolpert. Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research*, 8(Aug):1769–1797, 2007.

[27] Peihua Qiu. A jump-preserving curve fitting procedure based on local piecewise-linear kernel estimation. *Journal of Nonparametric Statistics*, 15(4-5):437–453, 2003.

[28] Babak Shahbaba and Radford Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10(Aug):1829–1850, 2009.

[29] Steven Spiriti, Randall Eubank, Philip W Smith, and Dennis Young. Knot selection for least-squares and penalized splines. *Journal of Statistical Computation and Simulation*, 83(6):1020–1036, 2013.

[30] Chong Tu. *Bayesian nonparametric modeling using Levy process priors with applications for function estimation, time series modeling and spatio-temporal modeling*. PhD thesis, Duke University, 2006.

[31] Sara Wade, Stephen G Walker, and Sonia Petrone. A predictive study of Dirichlet process mixture models for curve fitting. *Scandinavian Journal of Statistics*, 41(3):580–605, 2014.

[32] Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.

[33] Robert L Wolpert, Merlise A Clyde, and Chong Tu. Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics*, pages 1916–1962, 2011.

국문초록

본 논문에서는 불연속 점이 있을 수도 있는 임의의 함수를 추정하기 위한 베이지안 모형과, 모형의 추론을 위한 변분 방법을 제안한다. 제안한 모형은 LARK모형을 확장한 것으로서, 과완비 체계의 요소로 다중커널을 사용하여 함수를 적은 개수의 요소들로 표현하는 것이 가능하다. 점프의 위치, 구성요소의 개수, 부드러움의 정도 등의 함수를 구성하는 모든 요소들이 레비임의측도에 의해서 자동적으로 결정되기 때문에, 이 모형은 모형선택의 문제를 갖지 않는다. 시뮬레이션과 실제 자료분석을 통해서 제안된 모형이 불연속 함수를 추정하는 다른 비모수 모형들에 비해 성능이 우수함을 입증하였으며, 모형의 추론을 위해 제안된 확률적 변분 방법이 모수의 추출에 의존하는 가역 점프 마르코프 체인 몬테 카를로에 비해 계산 시간을 크게 줄일 수 있음을 확인하였다. 또한, 제안한 모형이 불연속이 있는 함수를 포함한 상당히 넓은 함수공간을 받침으로 갖는다는 사실을 증명했다.

주요어: 베이즈 함수 추정, 과완비 체계, 다중 커널, 레비 임의측도, 포아송 임의측도, 변분 방법, 모의 담금질

학번: 2011-30896