

IHS Economics Series  
Working Paper 285  
April 2012

# Bayesian Semiparametric Regression

Justinas Pelenis



## Impressum

---

**Author(s):**

Justinas Pelenis

**Title:**

Bayesian Semiparametric Regression

**ISSN: Unspecified**

**2012 Institut für Höhere Studien - Institute for Advanced Studies  
(IHS)**

Josefstädter Straße 39, A-1080 Wien

[E-Mail: office@ihs.ac.at](mailto:office@ihs.ac.at)

Web: [www.ihs.ac.at](http://www.ihs.ac.at)

All IHS Working Papers are available online:

[http://irihs.ihs.ac.at/view/ihs\\_series/](http://irihs.ihs.ac.at/view/ihs_series/)

This paper is available for download without charge at:

<https://irihs.ihs.ac.at/id/eprint/2129/>

285

Reihe Ökonomie  
Economics Series

# Bayesian Semiparametric Regression

Justinas Pelenis



INSTITUT FÜR HÖHERE STUDIEN  
INSTITUTE FOR ADVANCED STUDIES  
Vienna



285

Reihe Ökonomie  
Economics Series

# Bayesian Semiparametric Regression

Justinas Pelenis

April 2012

Institut für Höhere Studien (IHS), Wien  
Institute for Advanced Studies, Vienna

**Contact:**

Justinas Pelenis  
Department of Economics and Finance  
Institute for Advanced Studies  
Stumpergasse 56  
1060 Vienna, Austria  
☎: +43/1/599 91-143  
email: pelenis@ihs.ac.at

---

Founded in 1963 by two prominent Austrians living in exile – the sociologist Paul F. Lazarsfeld and the economist Oskar Morgenstern – with the financial support from the Ford Foundation, the Austrian Federal Ministry of Education and the City of Vienna, the Institute for Advanced Studies (IHS) is the first institution for postgraduate education and research in economics and the social sciences in Austria. The **Economics Series** presents research done at the Department of Economics and Finance and aims to share “work in progress” in a timely way before formal publication. As usual, authors bear full responsibility for the content of their contributions.

Das Institut für Höhere Studien (IHS) wurde im Jahr 1963 von zwei prominenten Exilösterreichern – dem Soziologen Paul F. Lazarsfeld und dem Ökonomen Oskar Morgenstern – mit Hilfe der Ford-Stiftung, des Österreichischen Bundesministeriums für Unterricht und der Stadt Wien gegründet und ist somit die erste nachuniversitäre Lehr- und Forschungsstätte für die Sozial- und Wirtschaftswissenschaften in Österreich. Die **Reihe Ökonomie** bietet Einblick in die Forschungsarbeit der Abteilung für Ökonomie und Finanzwirtschaft und verfolgt das Ziel, abteilungsinterne Diskussionsbeiträge einer breiteren fachinternen Öffentlichkeit zugänglich zu machen. Die inhaltliche Verantwortung für die veröffentlichten Beiträge liegt bei den Autoren und Autorinnen.

## **Abstract**

We consider Bayesian estimation of restricted conditional moment models with linear regression as a particular example. The standard practice in the Bayesian literature for semiparametric models is to use flexible families of distributions for the errors and assume that the errors are independent from covariates. However, a model with flexible covariate dependent error distributions should be preferred for the following reasons: consistent estimation of the parameters of interest even if errors and covariates are dependent; possibly superior prediction intervals and more efficient estimation of the parameters under heteroscedasticity. To address these issues, we develop a Bayesian semiparametric model with flexible predictor dependent error densities and with mean restricted by a conditional moment condition. Sufficient conditions to achieve posterior consistency of the regression parameters and conditional error densities are provided. In experiments, the proposed method compares favorably with classical and alternative Bayesian estimation methods for the estimation of the regression coefficients.

### **Keywords**

Bayesian semiparametrics, Bayesian conditional density estimation, heteroscedastic linear regression, posterior consistency

### **JEL Classification**

C11, C14

**Comments**

I am very thankful to Andriy Norets, Bo Honore, Sylvia Frühwirth-Schnatter, Jia Li, Ulrich Müller, and Chris Sims as well as seminar participants at Princeton, Royal Holloway, Institute for Advanced Studies, Vienna, Vienna University of Economics and Business, Seminar on Bayesian Inference in Econometrics and Statistics (SBIES), and Cowles summer conferences for helpful discussions and multiple suggestions that greatly contributed to improve the quality of the manuscript. All remaining errors are mine.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Restricted Moment Model</b>	<b>5</b>
	2.1 Finite Smoothly Mixing Regression .....	6
	2.2 Infinite Smoothly Mixing Regression .....	8
<b>3</b>	<b>Consistency Properties</b>	<b>10</b>
<b>4</b>	<b>Simulation Examples</b>	<b>15</b>
<b>5</b>	<b>Appendix</b>	<b>18</b>
	5.1 Proofs .....	18
	5.2 Posterior Computation .....	35
	<b>References</b>	<b>38</b>



# 1 Introduction

Estimation of regression coefficients in linear regression models can be consistent but inefficient if heteroscedasticity is ignored. Furthermore, the regression curve only provides a summary of the mean effects but does not provide any information regarding conditional error distributions which might be of interest to the decision maker. Estimation of conditional error distributions is useful in settings where forecasting and out of sample predictions are the object of interest. In this paper I propose a novel Bayesian method for consistent estimation of both linear regression coefficients and conditional residual distributions when data generating process satisfies a linear conditional moment restriction  $\mathbb{E}[y|x] = x'\beta$  or a more general restricted conditional moment condition of  $\mathbb{E}[y|x] = h(x, \theta)$  for some known function  $h$ . The contribution of this proposal is that the model is correctly specified for a large class of true data generating processes without imposing specific restrictions on the conditional error distributions and hence consistent and efficient estimation of the parameters of interest might be expected.

The most widely used method to estimate the mean of a continuous response variable as a function of predictors is, without doubt, the linear regression model. Often the models considered impose the assumptions of constant variance and/or symmetric and unimodal error distributions and such restrictions are often inappropriate for real-life datasets where conditional variability, skewness and asymmetry might hold. The prediction intervals obtained using models with constant variance and/or symmetric error distributions are likely to be inferior to the prediction intervals obtained from models with predictor dependent residual densities. To achieve full inference of parameters of interest and conditional error densities I propose a semiparametric Bayesian model for simultaneous estimation of regression coefficients and predictor dependent error densities. A Bayesian approach might be more effective in small samples as it enables exact inference given observed data instead of relying on asymptotic approximations.

Most of the semiparametric Bayesian literature focuses on constructing nonparametric priors for error distribution. The common assumption is that the errors are generated independently from regressors  $x$  and usually satisfy either a median or quantile restriction. Estimation and consistency of such models is discussed in [Kottas and Gelfand \(2001\)](#), [Hirano \(2002\)](#), [Amewou-Atisso et al. \(2003\)](#), [Conley et al. \(2008\)](#) and [Wu and Ghosal \(2008\)](#) among others. However, estimation of the parameters and error densities might be inconsistent if errors and covariates are dependent. For example, under heteroscedasticity or conditional asymmetry of error distributions the pseudo-true values of regression coefficients in a linear model with errors generated by covariate independent mixtures of normals are not generally equal to the true parameter values. One of the contributions of this paper is to show that the model proposed in this manuscript that incorporates predictor dependent residual densities is flexible and leads to a consistent estimation of both parameters of interest  $\theta$  and conditional error densities. Other Bayesian proposals that incorporate predictor dependent residual density modeling into parametric models are by [Pati and Dunson \(2009\)](#) where residual density is restricted to be symmetric, by [Kottas and Krnjajic \(2009\)](#) for quantile regression but without accompanying consistency theorems and by [Leslie et al. \(2007\)](#) who accommodate heteroscedasticity by multiplying the error term by a predictor dependent factor. However, none of these papers address the issue of conditional error asymmetry, and the estimation of regression coefficients by these methods might be inconsistent in the presence of residual asymmetry as the proposed models are misspecified.

Flexible models with covariate dependent error densities might lead to a more efficient estimator of the regression coefficients. For a linear regression problem, often only the regression coefficient  $\beta$  is of interest. It is a well known fact that if the conditional moment restriction holds then the weighted least squares estimator is more efficient than ordinary least squares estimator under heteroscedasticity. It is known that in parametric

models, by assertion of Le Cam’s parametric Bernstein-von Mises theorem, the posterior behaves as if one has observed normally distributed maximum likelihood estimator with variance equal to the inverse of Fisher information, see [van der Vaart \(1998\)](#). Semiparametric versions of Bernstein-von Mises theorem have been obtained by [Shen \(2002\)](#) and [Kleijn and Bickel \(2010\)](#), but the conditions are hard to verify. Nonetheless there is an expectation that posterior distribution of  $\beta$  is normal and centered at the true value in correctly specified semiparametric models if the priors are carefully chosen. Since the most popular frequentist approach of using OLS with heteroscedasticity robust covariance matrix ([White \(1982\)](#)) is suboptimal in a linear regression model with conditional moment restriction, one should expect to achieve a more efficient estimator by estimating a correctly specified model proposed here. Simulation results presented in Section 4 support the hypothesis that the proposed model gives a more efficient estimator of regression coefficients under heteroscedasticity.

The defining feature of the proposed model is that we impose a zero mean restriction on conditional error densities conditional on any predictor value. Imposition of the conditional restriction on the error distributions can be expected to be of benefit as a more efficient estimation of the parameters of interest might be expected. We model residual distributions flexibly as a finite or infinite mixtures of a base kernel. The base kernel for residual density is a mixture of two normal distributions with a joint mean of 0.

The probability weights in both finite and infinite mixtures are predictor dependent and vary smoothly with changes in predictor values. We consider a finite smoothly mixing regression model similar to the ones considered by [Geweke and Keane \(2007\)](#) and [Norets \(2010\)](#) and show that estimation would be consistent if the number of mixtures is allowed to increase. In such models, an appropriate number of mixtures needs to be selected which presents an additional complication. To avoid such complications, an alternative is to estimate a fully nonparametric model (i.e. infinite mixture). We consider the kernel

stick breaking process as a fully non-parametric approach to inference in a restricted moment model defined by a conditional moment restriction. This flexible approach leads to consistent estimation of both parameters of interest and conditional error densities.

Another contribution of this paper is to provide weak posterior consistency theorems for conditional density estimation in a Bayesian framework for a large class of true data generating processes using kernel stick breaking process (KSBP) with an exponential kernel proposed by [Dunson and Park \(2008\)](#). There are two alternative approaches for conditional density estimation in the Bayesian literature. The first general approach is to use dependent Dirichlet processes ([MacEachern \(1999\)](#), [De Iorio et al. \(2004\)](#), [Griffin and Steel \(2006\)](#) and others) to model conditional density directly. The second approach is to model joint unconditional distributions ([Muller et al. \(1996\)](#), [Norets and Pelenis \(2012\)](#) and others) and extract conditional densities of interest from joint distribution of observables. Even though many varying approaches for direct modeling of conditional distributions have been considered, consistency properties have been largely unstudied and only recent studies of [Tokdar et al. \(2010\)](#), [Norets and Pelenis \(2011\)](#) and [Pati et al. \(2011\)](#) address this question using different setups. We provide a set of sufficient conditions to ensure weak posterior consistency of conditional densities using KSBP with an exponential kernel and mixtures of Gaussian distributions and indirectly achieve posterior consistency of the parametric part.

In [Section 4](#), we conduct a Monte Carlo evaluation of the proposed method and compare it to a selection of alternative Bayesian and classical approaches for estimating regression coefficients. The proposed semiparametric estimator has smaller RMSE and better posterior coverage properties than other alternatives under heteroscedasticity and performs equally well under homoscedasticity. The alternative semiparametric Bayesian estimator based on an error density modeled as a mixture of normal distributions performs worse than other methods both under heteroscedasticity and conditional asymmetry of

error distributions. This is unsurprising as the pseudo-true values of regression coefficients in this misspecified alternative Bayesian semiparametric model are not equal to the true parameter values.

The outline of the paper is as follows: Section 2 introduces the finite and infinite models for estimation of a semiparametric linear regression with a conditional moment constraint. Section 3 provides theoretical results regarding the posterior consistency of both the parametric and nonparametric components of the model. Section 4 contains small sample simulation results. The proofs and details of posterior computation are contained in the Appendix.

## 2 Restricted Moment Model

The data consists of  $N$  observations of  $(Y_N, X_N) = \{(y_1, x_1), (y_2, x_2), \dots, (y_N, x_N)\}$  where  $y_i \in \mathcal{Y} \subseteq \mathbb{R}$  is a response variable and  $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$  are the covariates. The observations are independently and identically distributed  $(y_i, x_i) \sim F_0$  under the assumption that the data generating process (DGP) satisfies  $\mathbb{E}_{F_0}[y|x] = h(x, \theta_0)$  for all  $x \in \mathcal{X}$  for some known function  $h : \mathcal{X} \times \Theta \mapsto \mathcal{Y}$ . Alternatively, the restricted moment model can be written as

$$y_i = h(x_i, \theta_0) + \epsilon_i, (y_i, x_i) \sim F_0, i = 1, \dots, n.$$

with  $\mathbb{E}_{F_0}[\epsilon|x] = 0$  for all  $x \in \mathcal{X}$ .

The unknown parameters of this semiparametric model would be  $(\theta, f_{\epsilon|x})$ , where  $\theta$  is the finite dimensional parameter of interest and  $f_{\epsilon|x}$  is the infinite dimensional parameter. Let  $\Xi = \mathcal{F}_{\epsilon|x} \times \Theta$  be the parameter space, where  $\Theta$  denotes the space of  $\theta$  and  $\mathcal{F}_{\epsilon|x}$  the space of conditional densities with mean zero. That is  $\theta \in \Theta \subset \mathbb{R}^p$  and

$$\mathcal{F}_{\epsilon|x} = \left\{ f_{\epsilon|x} : \mathbb{R} \times \mathcal{X} \mapsto [0, \infty) : \int_{\mathbb{R}} f_{\epsilon|x}(\epsilon, x) d\epsilon = 1, \int_{\mathbb{R}} \epsilon f_{\epsilon|x}(\epsilon, x) d\epsilon = 0 \quad \forall x \in \mathcal{X} \right\}.$$

The primary objective is to construct a model to consistently estimate the parameter of interest  $\theta_0$ , while consistent estimation of the conditional error densities  $f_{0,\epsilon|x}$  is of secondary interest. This joint objective is achieved by proposing a flexible predictor dependent model for residual densities that allows the residual density to vary with predictors  $x \in \mathcal{X}$ . The model is correctly specified under weak restrictions on  $\mathcal{F}_{\epsilon|x}$  and leads to consistent estimation of both  $\theta_0$  and conditional error densities. Furthermore, the simulation results in Section 4 show that this flexible approach might be helpful to achieve a more efficient estimates of the parameter of interest  $\theta_0$ .

## 2.1 Finite Smoothly Mixing Regression

First, we define a density  $f_2(\cdot)$  which is a mixture of two normal distributions with a joint mean of zero. That is density of  $f_2$  given parameters  $\{\pi, \mu, \sigma_1, \sigma_2\}$  is defined as

$$f_2(\epsilon; \pi, \mu, \sigma_1, \sigma_2) = \pi\phi(\epsilon; \mu, \sigma_1^2) + (1 - \pi)\phi(\epsilon; -\mu\frac{\pi}{1 - \pi}, \sigma_2^2)$$

where  $\phi(\epsilon; \mu, \sigma^2)$  is a standard normal density evaluated at  $\epsilon$  with mean  $\mu$  and variance  $\sigma^2$ . Note that by construction a random variable  $\epsilon$  with a probability density function  $f_2$  has an expected value 0 as desired. In Section 3 we show that any density belonging to a large class of densities with mean 0 can be approximated by a countable collection of mixtures of  $f_2$ .

The proposed finite smoothly mixing regression model that imposes a conditional moment restriction is a special case of a mixtures of experts as introduced by [Jacobs et al. \(1991\)](#). Let the proposed model  $\mathcal{M}_k$  be defined by a set of parameters  $(\eta_k, \theta)$  where  $\theta$  is the parameter of interest and  $\eta_k$  are the nuisance parameters that induce conditional



densities  $f_{\epsilon|x}$ . The density of observable  $y_i$  is modeled as:

$$p(y_i|x_i, \theta, \eta_k) = \sum_{j=1}^k \alpha_j(x_i) f_2(y_i - h(x_i, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \quad (1)$$

$$\sum_{j=1}^k \alpha_j(x_i) = 1, \quad \forall x_i \in \mathcal{X}$$

where  $\alpha_j(x_i)$  is a regressor dependent smoothly varying probability weight. Note that by construction  $\mathbb{E}_p[y|x] = h(x, \theta)$  as desired. The conditional distribution of residuals is modeled as a flexible countable mixture of densities  $f_2$  with predictor dependent mixing weights.

Modeling of  $\alpha_j(x)$  is the choice of the econometrician and there are few available alternatives. We will use a linear logit regression considered by [Norets \(2010\)](#) as it has desirable theoretical properties. Mixing probabilities  $\alpha_j(x_i)$  are modeled as

$$\alpha_j(x_i) = \frac{\exp(\rho_j + \gamma'_j x_i)}{\sum_{l=1}^k \exp(\rho_l + \gamma'_l x_i)}. \quad (2)$$

The linear logit regression is not a unique choice as [Geweke and Keane \(2007\)](#) considered a multinomial probit model for  $\alpha_j(x)$ , and a multiple number of alternative possibilities have been considered in predictor-dependent stick breaking process literature. Generally, this finite mixture model can be considered as a special case of smoothly mixing regression model for conditional density estimation that imposes a linear mean but leaves residual densities unconstrained.

The full finite mixture model is characterized by the parameter of interest  $\theta$  and the nuisance parameters  $\eta_k \equiv \{\pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}, \rho_j, \gamma'_j\}_{j=1}^k$ . To complete the characterization of this model one would specify a prior  $\Pi_\theta$  on  $\Theta$  and a prior  $\Pi_\eta$  on the parameters  $\eta_k$  that induces a prior  $\Pi_{f_{\epsilon|x}}$  on  $\mathcal{F}_{\epsilon|x}$ . These priors induce a joint prior  $\Pi = \Pi_\theta \times \Pi_{f_{\epsilon|x}}$  on  $\Xi$ .

In [Section 3](#) we show that for any true DGP  $F_0$  there exists  $k$  large enough and

parameters  $(\theta, \eta_k)$  such that the proposed model is arbitrarily close in KL distance to the true DGP. This property can be used to show that a consistent estimation of  $\theta_0$  would be obtained with  $k \rightarrow \infty$ .

## 2.2 Infinite Smoothly Mixing Regression

Estimation of a finite mixture model introduces an additional complication of having to estimate the number of mixture components  $k$ . An alternative solution would be to consider an infinite smoothly mixing regression. The conditional density of the observable  $y_i$  is modeled as:

$$p(y_i|x_i, \theta, \eta) = \sum_{j=1}^{\infty} p_j(x_i) f_2(y_i - h(x_i, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2})$$

where  $\eta$  are nuisance parameters to be specified later,  $p_j(x_i)$  is a predictor dependent probability weight and  $\sum_{j=1}^{\infty} p_j(x) = 1$  a.s. for all  $x \in \mathcal{X}$ . To construct this infinite mixture model we will employ predictor-dependent stick breaking processes.

Similarly to the choice of  $\alpha_j(x)$  in the finite smoothly mixing regressions, various constructions of  $p_j(x)$  have been considered in the literature. Those methods include order based dependent Dirichlet processes ( $\pi$ DDP) proposed by [Griffin and Steel \(2006\)](#), probit stick-breaking process ([Chung and Dunson \(2009\)](#)), kernel stick-breaking process ([Dunson and Park \(2008\)](#)) and local Dirichlet process (IDP) ([Chung and Dunson \(2011\)](#)) which is a special case of kernel stick-breaking processes. We will be employing a kernel stick-breaking process introduced by [Dunson and Park \(2008\)](#). It is defined using a countable sequence of mutually independent random components  $V_j \sim \text{Beta}(a_j, b_j)$  and  $\Gamma_j \sim H$  independently for each  $j = 1, \dots$ . The covariate dependent mixing weights are

defined as:

$$p_j(x) = V_j K_\varphi(x, \Gamma_j) \prod_{l < j} (1 - V_l K_\varphi(x, \Gamma_l)), \text{ for all } x \in \mathcal{X}$$

where  $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]$  is any bounded kernel function. Kernel functions that have been considered in practice are  $K_\varphi(x, \Gamma_j) = \exp(-\varphi \|x - \Gamma_j\|^2)$  and  $K_\varphi(x, \Gamma_j) = \mathbf{1}(\|x - \Gamma_j\| < \varphi)$ , where  $\|\cdot\|$  is the Euclidean distance.

Jointly the conditional density of  $y_i$  conditional on covariate  $x_i$  is defined as

$$p(y_i | x_i, \theta, \eta) = \sum_{j=1}^{\infty} p_j(x_i) f_2(y_i - h(x_i, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \quad (3)$$

$$p_j(x) = V_j K_\varphi(x, \Gamma_j) \prod_{l < j} (1 - V_l K_\varphi(x, \Gamma_l))$$

For Bayesian analysis the parameters are endowed with these priors:  $\{\pi_j, \mu_j, \sigma_{j1}^2, \sigma_{j2}^2\} \sim G_0$ ,  $\Gamma_j \sim H$ ,  $V_j \sim \text{Beta}(a_j, b_j)$ ,  $\varphi \sim \Pi_\varphi$  and  $\theta \sim \Pi_\theta$ . The nuisance parameter is  $\eta = \{\varphi, \{\pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}, V_j, \Gamma_j\}_{j=1}^{\infty}\}$  and jointly these priors on the nuisance parameters induce a prior  $\Pi_{f_{\epsilon|x}}$  on  $\mathcal{F}_{\epsilon|x}$ .

This is a very flexible model for predictor dependent conditional densities, however it also imposes the desired property that conditional error densities have a mean of zero in order to identify parameter of interest  $\theta$ . We will show that this is a ‘correctly’ specified model for the DGP as posterior concentrates on the true parameter  $\theta_0$  and on a weak neighborhood of the true conditional densities  $f_{0, \epsilon|x}$  using a particular choice of a kernel function. Exponential kernel is chosen as it is closely related to the linear logit regression used in finite mixture model. Therefore, we will use kernel stick-breaking processes with exponential kernel as our choice to construct  $p_j(x)$ .

Even though practical suggestions have been plentiful, theoretical results regarding the consistency properties of predictor dependent stick-breaking processes are scarce.

Related theoretical results are presented by [Tokdar et al. \(2010\)](#), [Pati et al. \(2011\)](#) and [Norets and Pelenis \(2011\)](#). One of the key contributions of this paper are Theorem 2 and Corollary 1 in Section 3 that show that kernel stick-breaking processes with exponential kernel can be used to consistently estimate flexible unrestricted conditional densities and the parameters of interest. Hopefully, consistent estimation of the error densities could lead to a more efficient estimation of the parameters of interest as compared to the other methods that do not directly impose a conditional moment restriction.

### 3 Consistency Properties

We provide general sufficient conditions on the true data generating process that lead to posterior consistency in estimating regression parameters and conditional residual densities. I show that residual densities induced by the proposed models can be chosen to be arbitrarily close in Kullback-Leibler distance to true conditional densities that satisfy the conditional moment restriction. That is the Kullback-Leibler (KL) closure of proposed models in Section 2 include all true data generating distributions that satisfy a set of general conditions outlined below.

Let  $p(y|x, \mathcal{M})$  be the conditional density of  $y$  given  $x$  implied by some model  $\mathcal{M}$ . The models considered in this paper were presented in Sections 2.1 and 2.2. Let the true data generating joint density of  $(y, x)$  be  $f_0(y|x)f_0(x)$ , then the joint marginal density induced by the model  $\mathcal{M}$  is  $p(y|x, \mathcal{M})f_0(x)$ . Note that in the models considered in Section 2 we modeled only conditional error density and left the data generating density  $f_0(x)$  of  $x \in \mathcal{X}$  unspecified. The KL distance between  $f_0(y|x)f_0(x)$  and  $p(y|x, \mathcal{M})f_0(x)$  is defined as

$$d_{KL}(f_0, p_{\mathcal{M}}) = \int \log \frac{f_0(y|x)f_0(x)}{p(y|x, \mathcal{M})f_0(x)} F_0(dy, dx) = \int \log \frac{f_0(y|x)}{p(y|x, \mathcal{M})} F_0(dy, dx). \quad (4)$$

Given the true conditional data generating density  $f_0(y|x)$ , define  $f_{0,\epsilon|x}$  as  $f_{0,\epsilon|x}(\epsilon|x) = f_0(\epsilon+h(x, \theta_0)|x)$ . We say that posterior is consistent for estimating  $(f_{0,\epsilon|x}, \theta_0)$  if  $\Pi(\mathcal{W}|Y_n, X_n)$  converges to 1 with  $P_{F_0}^n$  probability as  $n \rightarrow \infty$  for any neighborhood  $\mathcal{W}$  of  $(f_{0,\epsilon|x}, \theta_0)$  when the true data generating distribution is  $F_0$ . We define a weak neighborhood  $\mathcal{U}_\delta(f_{\epsilon|x})$  as

$$\mathcal{U}_\delta(f_{\epsilon|x}) = \left\{ f_{|x} : f_{|x} \in \mathcal{F}_{\epsilon|x}, \left| \int_{\mathbb{R} \times \mathcal{X}} g f_{|x}(\epsilon, x) f_0(x) d\epsilon dx - \int_{\mathbb{R} \times \mathcal{X}} g f_{\epsilon|x}(\epsilon, x) f_0(x) d\epsilon dx \right| < \delta, \right. \\ \left. g : \mathbb{R} \times \mathcal{X} \mapsto \mathbb{R} \text{ is bounded and uniformly continuous} \right\}.$$

Then we consider neighborhoods  $\mathcal{W}$  of  $(f_{0,\epsilon|x}, \theta_0)$  of the form  $\mathcal{U}_\delta(f_{0,\epsilon|x}) \times \{\theta : \|\theta - \theta_0\| < \rho\}$  for any  $\delta > 0$  and  $\rho > 0$ . Since our primary objective is consistent estimation of  $\theta_0$  it suffices to consider only the weak neighborhoods of conditional densities.

First, we will consider the case of the finite model described in Section 2.1. Let the proposed model  $\mathcal{M}_k$  be defined by the parameters  $(\eta_k, \theta)$ . Then we show that there exists  $k$  large enough and a set of parameters  $(\eta_k, \theta)$  such that KL distance between true conditional densities and the ones implied by the finite model is arbitrarily close to 0.

**Theorem 1.** *Assume that*

1.  $f_0(y|x)$  is continuous in  $(y, x)$  a.s.  $F_0$ .
2.  $\mathcal{X}$  has bounded support and  $\mathbb{E}_{F_0}[y^2|x] < \infty$  for all  $x \in \mathcal{X}$ .
3.  $h$  is Lipschitz continuous in  $x$ .
4. There exists  $\delta > 0$  such that

$$\int \log \frac{f_0(y|x)}{\inf_{\|y-z\| < \delta, \|x-t\| < \delta} f_0(z|t)} F_0(dy, dx) < \infty \quad (5)$$

Let  $p(\cdot|\cdot, \theta, \eta_k)$  be defined as in Equations (1) and (2). Then, for any  $\epsilon > 0$  there exists

$(\eta_k, \theta)$  such that

$$d_{KL}(f_0(\cdot|\cdot), p(\cdot|\cdot, \theta, \eta_k)) < \epsilon.$$

Theorem 1 is proved rigorously in the appendix. The basic idea is that any unconditional density with mean 0 can be approximated by a finite mixture of  $f_2$  densities. To approximate conditional densities we show that mixing weights  $\alpha(x)$  are flexible enough so that for any  $x \in \mathcal{X}$  most of the mass on the neighborhood of  $x$  induced by a subset of particular mixing weights approaches 1. Then only unconditional density with mean 0 at that particular  $x \in \mathcal{X}$  has to be approximated and that is feasible.

The results above imply the existence of a large number  $k$  of mixture components such that induced conditional densities are close to the true values of the DGP. However, this does not provide a direct method of estimating  $k$ , the number of mixtures, to be used in applications. Furthermore, one can show that any finite model could have pseudo-true values of  $\theta$  different from true values for some data generating distributions that belong to the general class  $\mathcal{F}$  of DGPs. Such concerns do not play a role if an infinite smoothly regression model induced by a predictor dependent stick breaking process prior is used for inference. Below we show that estimation of infinite mixture model would lead to posterior consistent estimation of  $f_{0,\epsilon|x}$  and  $\theta_0$ . Hence, we provide the necessary theoretical foundation for the use of infinite mixture model.

For the infinite mixture model defined in the Equation 3, the priors  $G_0, H, \Pi_V, \Pi_\varphi, \Pi_\theta$  and a choice of kernel function  $K_\varphi$  induce a prior  $\Pi$  on  $\Xi$ . A conditional density function  $f_x$  is said to be in the *KL support* of the prior  $\Pi$  (i.e.  $f_x \in KL(\Pi)$ ), if for all  $\epsilon > 0$ ,  $\Pi(K_\epsilon(f_x)) > 0$ , where  $K_\epsilon(f_x) \equiv \{(\theta, \eta) : d_{KL}(f_x(\cdot|\cdot), p(\cdot|\cdot, \theta, \eta)) < \epsilon\}$  and  $d_{KL}(\cdot, \cdot)$  is defined in the Equation 4. The next theorem shows that if a true data generating distribution  $F_0$  satisfies the assumptions of the Theorem 1, then  $f_0$  belongs to the KL support of  $\Pi$  under general conditions on the prior distributions and for a particular kernel function.

**Theorem 2.** *Assume  $F_0$  satisfies assumptions of Theorem 1 and  $f_0(\cdot|\cdot)$  are covariate dependent conditional densities of  $y \in \mathcal{Y}$  induced by  $F_0$ . Let  $K_\varphi(x, \Gamma) = \exp(-\varphi\|x - \Gamma\|^2)$  and let the prior  $\Pi$  be induced by the priors  $G_0, H, \Pi_V, \Pi_\varphi, \Pi_\theta$ . If the priors are such that  $\theta_0$  is an interior point of support of  $\Pi_\theta$ ,  $\Pi(\sigma_{j_1} < \delta) > 0$  for any  $\delta > 0$  and  $\mathcal{X} \subset \text{supp}(H)$ , then  $f_0 \in KL(\Pi)$ .*

The full proof of the theorem is provided in the Appendix, while the intuition is provided below. The proof is constructing by showing that there exists a particular set of parameters of infinite smoothly mixing regression and an open neighborhood of this particular set of parameters that are arbitrarily close in KL sense to the finite smoothly mixing regression that is close to the DGP. Hence the true data generating conditional densities belong to the KL support of the prior  $\Pi$ .

Once the KL support property is established we hope to proceed to use Schwartz's posterior consistency theorem (Schwartz (1965)) to show that posterior is weakly consistent at  $f_{0,\epsilon|x}$  and  $\theta_0$ . First, we will consider the case of the linear regression with  $h(x, \theta) = x'\theta$  as an illustrative example of the additional assumptions that are necessary to achieve posterior consistency. This is an assumption used by Wu and Ghosal (2008) and it plays a similar role to the assumption of no multicollinearity in the DGP so that  $\theta_0$  can be identified. Let  $\gamma \in \{-1, 1\}^d$  and define a quadrant  $Q_\gamma = \{z \in \mathbb{R}^d : z_j \gamma_j > 0 \text{ for all } j = 1, \dots, d\}$ .

**Theorem 3.** *An (almost) immediate implication of Schwartz (1965). Suppose that  $F_0$  satisfy the assumptions of the Theorem 1 and that the prior distributions satisfy the requirements of the Theorem 2 and that  $E_{F_0}[y|x] = x'\theta_0$ . Suppose that for any  $\gamma$ ,  $F_0(Q_\gamma \setminus \{X : |x_i| < \xi\}) > 0$  for each  $i = 1, \dots, d$  and some  $\xi > 0$ . Furthermore, the prior is restricted so there exists a large  $L$  such that  $\mathbb{E}_f[\epsilon^2|x] < L$  for all  $x \in \mathcal{X}$  and all  $f \in \text{supp}(\Pi_{f_{\epsilon|x}})$ . Let  $\mathcal{W} = \mathcal{U}_\delta(f_{0,\epsilon|x}) \times \{\theta : \|\theta - \theta_0\| < \rho\}$  for some  $\delta > 0$  and  $\rho > 0$ , then  $\Pi(\mathcal{W}^c|Y_N, X_N) \rightarrow 0$  a.s.  $P_{F_0}^\infty$ .*

The theorem is proved rigorously in the Appendix. It consists of the construction

of exponentially consistent tests for testing  $H_0 : (f_{\epsilon|x}, \theta) = (f_{0,\epsilon|x}, \theta_0)$  against alternative hypothesis  $H_1 : (f_{\epsilon|x}, \theta) \in \mathcal{W}^c$ . Once that is accomplished it is a straightforward application of Schwartz's posterior consistency theorem as KL-property is already proved in Theorem 2.

Theorem 3 can be extended to other restricted moment models beyond linear regression case with an additional general assumption. The proof of the corollary below is presented in the Appendix, but it is a fairly straightforward extension of the construction of the exponentially consistent tests in a more general than linear regression setting.

**Corollary 1.** *Suppose that  $F_0$  satisfy the assumptions of the Theorem 1 and that the prior distributions satisfy the requirements of the Theorem 2. Additionally, assume that this identification restriction is satisfied:  $E_{F_0}[|h(x, \theta) - h(x, \theta_0)|] \geq \xi \|\theta - \theta_0\|$  for some  $\xi > 0$ . Furthermore, the prior is restricted so that there exists a large  $L$  such that  $\mathbb{E}_f[\epsilon^2|x] < L$  for all  $x \in \mathcal{X}$  and all  $f \in \text{supp}(\Pi_{f_{\epsilon|x}})$ . Let  $\mathcal{W} = \mathcal{U}_\delta(f_{0,\epsilon|x}) \times \{\theta : \|\theta - \theta_0\| < \rho\}$  for some  $\delta > 0$  and  $\rho > 0$ , then  $\Pi(\mathcal{W}^c|Y_N, X_N) \rightarrow 0$  a.s.  $P_{F_0}^\infty$ .*

This corollary establishes that consistent estimation of parameter of interest and conditional error densities will be achieved if the true data generating process satisfies a conditional moment restriction. Given the desirable theoretical properties that both parametric and nuisance parts are consistently estimated it achieves the two objectives. Firstly, the estimation of the parameter of interest is consistent. Secondly, consistent estimation of the nuisance parameter, which are conditional error densities in this case, might lead to a more efficient estimation of the parameter of interest which would be a justification for the estimation of the full semiparametric model as opposed to some alternative simplified model.



## 4 Simulation Examples

A number of simulation examples is considered to assess the performance of the method proposed in this paper. Consider a linear regression model with

$$y_i = \alpha + x_i' \beta + \epsilon_i, \quad (y_i, x_i) \stackrel{\text{iid}}{\sim} F_0, \quad i = 1, \dots, n.$$

and  $\mathbb{E}_{F_0}[\epsilon_i | x_i] = 0$  and  $x$  is one-dimensional. We consider four alternative data generating processes (DGPs), with first three suggested by Müller (2010).

1. Case (i):  $y_i = 0 + 0 \cdot x_i + \epsilon_i$ ,  $\epsilon_i \sim N(0, 1)$ .
2. Case (ii):  $y_i = 0 + 0 \cdot x_i + \epsilon_i$ ,  $\epsilon_i | x_i \sim N(0, a^2(|x_i| + 0.5)^2)$ , where  $a = 0.454\dots$
3. Case (iii):  $y_i = 0 + 0 \cdot x_i + \epsilon_i$ ,  $\epsilon_i | x_i, s \sim N([1 - 2 \cdot \mathbf{1}(x_i < 0)]\mu_s, \sigma_s^2)$ , where  $P(s = 1) = 0.8$ ,  $P(s = 2) = 0.2$ ,  $\mu_1 = -0.25$ ,  $\sigma_1 = 0.75$ ,  $\mu_2 = 1$  and  $\sigma_2 = \sqrt{1.5}$ .
4. Case (iv):  $y_i = 0 + 0 \cdot x_i + \epsilon_i$ ,  $\epsilon_i | x_i \sim N(x_i \mu_s, 0.5^2)$ , where  $P(s = 1) = P(s = 2) = 0.5$  and  $\mu_1 = -\mu_2 = 0.5$ .

All four DGPs are such that  $\mathbb{E}[(x_i \epsilon_i)^2] = 1$  and  $x_i \sim N(0, 1)$ .

Inference is based on the following methods. First, inference based on an infeasible generalized least squares (GLS) with a correct covariance matrix specification. Second, Bayesian inference based on the artificial sandwich posterior (OLS) as proposed by Müller (2010). Let  $\theta = (\alpha, \beta)'$ , then  $\theta \sim N(\hat{\theta}, \hat{\Sigma})$  where  $\hat{\theta}$  is the ordinary least squares coefficient and  $\hat{\Sigma}$  is the “sandwich” covariance matrix. Note that inference based on this sandwich posterior is asymptotically equivalent to inference using Bayesian bootstrap (Lancaster (2003)) so there is a Bayesian alternative to this frequentist inspired procedure when the regression coefficient is the object of interest. Third, Bayesian inference based on a normal regression model (NLR), where  $\epsilon_i | x_i \sim N(0, h^{-1})$  with priors

$\theta \sim N(0, (0.01I_2)^{-1})$ ,  $3h \sim \chi_3^2$ . Fourth, Bayesian inference based on a normal mixture linear regression model (MIX) with  $\epsilon_i|x_i, s \sim N(\mu_s, (hh_s) - 1)$  and  $P(s = j) = \pi_j$ ,  $j = 1, 2, 3$  with priors  $\theta \sim N(0, (0.01I_2)^{-1})$ ,  $3h \sim \chi_3^2$ ,  $3h_j \sim \chi_3^2$ ,  $(\pi_1, \pi_2, \pi_3) \sim \text{Dirichlet}(3, 3, 3)$  and  $\mu_j \stackrel{\text{iid}}{\sim} N(0, (0.4h)^{-1})$ . Fifth, inference based on a [Robinson \(1987\)](#) asymptotically efficient uniform weight k-NN estimator with  $k_n = n^{4/5}$ . Finally, Bayesian inference based on the conditional linear regression model (CLR) proposed in this paper. We consider the finite model with  $k = 5$  number of states. The priors are set to  $\theta \sim N(0, (0.01I_2)^{-1})$ ,  $\gamma_j \sim N(0, (0.01I_2)^{-1})$ ,  $3h_{ji} \sim \chi_3^2$ ,  $\tilde{\mu}_j \sim N(0, 0.25^{-1})$ ,  $\underline{\pi} = 10$  for all  $j = 1, \dots, n$  and  $i = 1, 2$ . Posterior computation and full description of the priors are contained in the [Appendix 5.2](#). Posterior simulation for the infinite mixture model could be accomplished using retrospective or slice sampling methods proposed by [Papaspiliopoulos and Roberts \(2008\)](#), [Walker \(2007\)](#) and [Kalli et al. \(2011\)](#).

The parameter of interest is  $\beta \in \mathbb{R}$  and we consider three separate criteria for the evaluation of the performance of the proposed estimators. First, we will compute root mean squared error (RMSE). While Bayesian credibility regions are different from confidence intervals in practice one can still expect some similarity even in moderate samples. Therefore, for practical purposes we construct 95% intervals using 0.025 and 0.975 quantiles of the posterior distribution and report coverage probabilities. Furthermore, we consider the lengths of these credibility regions as another indicator of the performance of the estimator. Similar approaches for evaluating performance have been considered by [Conley et al. \(2008\)](#).

We repeat simulation exercise 1000 times for each DGP. The results are displayed in [Table 1](#). Relative performance of the methods is similar whether RMSE, coverage, or interval length is used as an evaluation criterion. The results show that the conditional linear regression model proposed in this paper performs better than alternatives in Cases (ii) and (iv) in the presence of heteroscedasticity and performs comparably in other cases.

Table 1: Simulation results

Method	Criterion	Case (i)	Case (ii)	Case (iii)	Case (iv)
GLS		0.070	0.056	0.071	0.060
OLS		0.070	0.069	0.071	0.071
NLR	RMSE	0.070	0.069	0.070	0.071
MIX		0.071	0.067	0.083	0.075
k-NN		0.070	0.063	0.072	0.066
CLR		0.070	0.060	0.072	0.067
GLS		0.949	0.946	0.950	0.940
OLS		0.946	0.948	0.949	0.942
NLR	95 %	0.950	0.805	0.947	0.847
MIX	Coverage	0.947	0.843	0.905	0.841
k-NN		0.952	0.853	0.943	0.870
CLR		0.971	0.948	0.965	0.935
GLS		0.28	0.22	0.28	0.24
OLS		0.28	0.27	0.28	0.27
NLR	Interval	0.28	0.18	0.28	0.20
MIX	Length	0.28	0.20	0.27	0.21
k-NN		0.28	0.18	0.28	0.20
CLR		0.30	0.23	0.30	0.25

Notes: DGPs are in columns and methods of inference in rows. Entries are RMSE, Coverage of 95% Bayesian credibility region and interval length of the Bayesian credibility region. Bayesian inference in each method is implemented by a Gibbs sampler with 8000 draws and first 2000 discarded as burn-in for 1000 draws from each DGP.

In Cases (i) and (iii) the best performing models should be OLS and NLR since it is well known that OLS estimator achieves the semi-parametric efficiency under homoscedasticity. Note that model MIX performs worse in Cases (iii) and (iv) due to (conditional) asymmetries of the error distribution. In Case (iii) this is expected since the pseudo-true value of  $\beta$  is not the true  $\beta_0 = 0$  for the MIX model. As demonstrated in this simulation example estimation of linear models with unconditional error densities modeled as mixtures of normals or with flexible symmetric residual densities as proposed in [Pati and Dunson \(2009\)](#) might be misguided if the regression coefficients are the object of interest. The reason being that the pseudo-true values of  $\beta$  might be different from the true  $\beta_0$ , for example, when disturbances are asymmetric. As expected, the model CLR proposed in

this paper outperforms other alternatives in the heteroscedastic cases and performs comparably in the homoscedastic cases even when compared to the asymptotically efficient k-NN estimator.

## 5 Appendix

### 5.1 Proofs

*Proof.* (Theorem 1)

Note that  $d_{KL}$  is always non-negative, hence for any model  $\mathcal{M}_{m,n}$

$$0 \leq \int \log \frac{f_0(y|x)}{p(y|x, \theta_{m,n}, \mathcal{M}_{m,n})} F_0(dy, dx) \leq \int \log \max \left\{ 1, \frac{f_0(y|x)}{p(y|x, \theta_{m,n}, \mathcal{M}_{m,n})} \right\} F_0(dy, dx).$$

Therefore, it would suffice to show that the last integral in the inequality converges to 0 as  $(m, n)$  increase. Dominated convergence theorem will be used for that. In the first part we will show pointwise convergence to 0 for any given  $(y, x)$  a.s.  $F$ . Then we will present conditions for the existence of an integrable upper bound on the integrand.

#### *Pointwise Convergence*

Let  $A_j^m$ ,  $j = 0, 1, \dots, m$ , be a partition of  $\mathcal{Y}$ , where  $A_1^m, \dots, A_m^m$  are adjacent cubes with side length  $h_m$  and  $A_0^m$  is the rest of set  $\mathcal{Y}$ . Let  $B_j^n$ ,  $j = 0, 1, \dots, N(n)$ , be a partition of  $\mathcal{X}$  with  $N(n) = n^d$ , where  $B_1^n, \dots, B_{N(n)}^n$  are adjacent cubes with side length  $\lambda_n$  and  $B_0^n$  is the rest of  $\mathcal{X}$ . This partition has to satisfy two conditions. First, the partition becomes finer as  $n$  increases with  $\lambda_n \rightarrow 0$ . Second, the area covered by the finer partition has to increase and eventually cover the whole support of  $\mathcal{X}$ , i.e.  $\lambda_n^d N(n) \rightarrow \infty$ . Furthermore, let  $x_j^n$  be the center of  $B_j^n$ ,  $j = 1, \dots, N(n)$  and  $x_0^n \in B_0^n$  be such that  $\left\{ \|x_0^n - x\|^2 > s_n : \forall x \in \bigcup_{i=1}^{N(n)} B_i^n \right\}$  where  $s_n$  is the squared diagonal of  $B_i^n$ . Let's consider

a model  $\mathcal{M}_{m,n}$  such that

$$p(y|x, \mathcal{M}_{m,n}) = \sum_{j=0}^{N(n)} \sum_{i=0}^m \alpha_{ji}^{nm}(x) \phi(y - h(x, \theta); \mu_{ji}, \sigma_{ji}^2) \quad (6)$$

$$\sum_{j=0}^{N(n)} \sum_{i=0}^m \alpha_{ji}^{nm}(x) \mu_{ji} = 0 \quad \text{for all } x \in \mathcal{X}.$$

We propose mixing probabilities such that

$$\begin{aligned} \alpha_{ji}^{nm}(x) &= \pi_{ji} \alpha_j(x) \\ \pi_{ji} &= F_0(A_i^m | x_j^n) \\ \alpha_j(x) &= \frac{\exp(-c_n(x_j^{n'} x_j^n - 2x_j^{n'} x))}{\sum_{i=0}^{N(n)} \exp(-c_n(x_i^{n'} x_i^n - 2x_i^{n'} x))}. \end{aligned}$$

Under appropriate conditions for  $c_n$ , we can show that some collection of  $\alpha_j(x)$  approximates  $\mathbf{1}_{B_j^n(x)}$ . All that is required is that  $c_n$  is such that  $c_n \rightarrow \infty$  and

$$\exp\{-c_n s_n\} / N(n) \rightarrow 0, \quad \text{where } s_n = \lambda_n^2 d$$

*i.e.*  $s_n$  is the squared diagonal of  $B_i^n$ . Such a sequence  $c_n$  always exists, for example all the necessary conditions would be satisfied for  $\lambda_n = N(n)^{-d} n^{-1/2} = n^{-1/2}$  and  $c_n = s_n^{-2}$ . Following the proof of Proposition 4.1. in [Norets \(2010\)](#) define  $I^n(x, s_n) = \{j : \|x_j^n - x\|^2 < s_n\}$ . Using the arguments of the proof of Proposition 4.1. we know that for  $(n, m)$  large enough for any  $j \in I^n(x, s_n)$

$$\sum_{j \in I_1^n(x, s_n)} \alpha_j(x) \geq 1 - \exp\{-c_n s_n\} / N(n). \quad (7)$$

Since  $h$  is Lipschitz continuous by Assumption 3, hence there exists  $L$  large enough such that  $|h(x, \theta_0) - h(x_j^n, \theta_0)| \leq L \|x - x_j^n\|$ . Let  $\delta_m^* = \delta_m + L s_n^{1/2}$  where  $\delta_m \rightarrow 0$ . Then

for any  $j \in I_1^n(x, s_n)$  and  $A_i^m \subset C_{\delta_m^*}(y)$ , where  $C_\delta(y)$  is an interval centered at  $y$  with length  $\delta$ ,

$$F_0(A_i^m | x_j^n) \geq \lambda(A_i^m) \inf_{z \in C_{\delta_m^*}(y), \|t-x\|^2 \leq s_n} f_0(z|t). \quad (8)$$

For each  $x_j^n$  the parameters  $\{\mu_{ji}\}_{i=0}^m$  must satisfy

$$\sum_{i=0}^m \pi_{ji} \mu_{ji} = \sum_{i=0}^m F_0(A_i^m | x_j^n) \mu_{ji} = 0. \quad (9)$$

Let  $\theta = \theta_0$ , let  $c_i^m$  be the center of the cube  $A_i^m$  if  $i \neq 0$ , then for  $i \neq 0$  let  $\mu_{ji} = c_i^m + d_j^n - h(x_j^n, \theta_0)$  where  $d_j^n \in [-h_m/2, h_m/2]$  and let  $\mu_{j0}$  be

$$\mu_{j0} = \frac{\int_{A_0^m} f_0(y|x_j^n)(y - h(x_j^n, \theta_0))dy}{F_0(A_0^m | x_j^n)}$$

if  $F_0(A_0^m | x_j^n) > 0$  and  $\mu_{j0} = 0$  otherwise. We show that there exists  $d_j^n$  such that equation (9) is satisfied. Define function  $G(d_j^n | x_j^n)$  as

$$\begin{aligned} G(d_j^n | x_j^n) &= \sum_{i=0}^m F_0(A_i^m | x_j^n) \mu_{ji} \\ &= \sum_{i=1}^m \int_{A_i^m} f_0(y|x_j^n)(c_i^m + d_j^n - h(x_j^n, \theta_0))dy + \int_{A_0^m} f_0(y|x_j^n)(y - h(x_j^n, \theta_0))dy. \end{aligned}$$

Clearly, the function  $G(d_j^n | x_j^n)$  is linear in  $d_j^n$  and therefore continuous in  $d_j^n$ . Note that

$$\begin{aligned} G(h_m/2 | x_j^n) &= \sum_{i=1}^m \int_{A_i^m} f_0(y|x_j^n)(c_i^m + h_m/2 - h(x_j^n, \theta_0))dy + \int_{A_0^m} f_0(y|x_j^n)(y - h(x_j^n, \theta_0))dy \\ &\geq \sum_{i=1}^m \int_{A_i^m} f_0(y|x_j^n)(y - h(x_j^n, \theta_0))dy + \int_{A_0^m} f_0(y|x_j^n)(y - h(x_j^n, \theta_0))dy \\ &\geq \sum_{i=0}^m \int_{A_i^m} f_0(y|x_j^n)(y - h(x_j^n, \theta_0))dy = 0 \end{aligned}$$

since  $\mathbb{E}_{F_0}[y|x] = h(x, \theta_0)$  for all  $x \in \mathcal{X}$  and hence  $G(h_m/2|x_j^n) \geq 0$ . By the same argument it follows that  $0 \geq G(-h_m/2|x_j^n)$ . As we have mentioned earlier  $G(\cdot|x_j^n)$  is a continuous function, therefore  $\exists d_j^n \in [-h_m/2, h_m/2]$  such that  $G(d_j^n|x_j^n) = 0$  and equivalently equation (9) is satisfied.

For any  $j \in I^n(x, s_n)$  let  $y_j^* = y - (h(x, \theta_0) - h(x_j^n, \theta_0))$ , then  $C_{\delta_m}(y_j^*) \subset C_{\delta_m^*}(y)$  by definition of  $\delta_m$  and  $\delta_m^*$ . Let  $\sigma_{ji} = \sigma_m$  if  $i > 0$  and  $\sigma_{j0} = \sigma_0$  to be chosen later. Then for  $m$  large enough such that  $\emptyset \neq \{i : A_i^m \subset C_{\delta_m}(y_j^*)\}$

$$\begin{aligned}
& \sum_{i: A_i^m \subset C_{\delta_m^*}(y)} \lambda(A_i^m) \phi(y - h(x, \theta_0); \mu_{ji}, \sigma_m^2) = \\
& \sum_{i: A_i^m \subset C_{\delta_m^*}(y)} \lambda(A_i^m) \phi(y - (h(x, \theta_0) - h(x_j^n, \theta_0)); c_i^m + d_j^n, \sigma_m^2) \\
& \geq \sum_{i: A_i^m \subset C_{\delta_m}(y_j^*)} \lambda(A_i^m) \phi(y_j^*; c_i^m + d_j^n, \sigma_m^2) \\
& \geq 1 - \frac{3h_m}{(2\pi)^{1/2}\sigma_m} - \frac{8\sigma_m}{(2\pi)^{1/2}\delta_m}. \tag{10}
\end{aligned}$$

with last inequality derived from Lemmas 1 and 2 in [Norets and Pelenis \(2012\)](#) (with a minor adjustment in the proofs due to uncentered positions of  $\mu_{ji}$ ).

Then equation (6) and inequalities (7), (8) and (10) can be combined to show that for any given  $(x, y)$  there exist  $(m, n)$  large enough such that

$$\begin{aligned}
p(y|x, \mathcal{M}_{m,n}) & > \sum_{j \in I_1^n(x, s_n)} \sum_{i: A_i^m \subset C_{\delta_m^*}(y)} F_0(A_i^m|x_j^n) \alpha_j(x) \phi(y - h(x, \theta); \mu_{ji}, \sigma_m^2) \\
& \geq \inf_{z \in C_{\delta_m^*}(y), \|t-x\|^2 \leq s_n} f_0(z|t) \sum_{j \in I_1^n(x, s_n)} \alpha_j(x) \sum_{i: A_i^m \subset C_{\delta_m^*}(y)} \lambda(A_i^m) \phi(y - h(x, \theta); \mu_{ji}, \sigma_m^2) \\
& \geq \inf_{z \in C_{\delta_m^*}(y), \|t-x\|^2 \leq s_n} f_0(z|t) \left( 1 - \frac{3h_m}{(2\pi)^{1/2}\sigma_m} - \frac{8\sigma_m}{(2\pi)^{1/2}\delta_m} \right) \sum_{j \in I_1^n(x, s_n)} \alpha_j(x) \\
& \geq \inf_{z \in C_{\delta_m^*}(y), \|t-x\|^2 \leq s_n} f_0(z|t) \left( 1 - \frac{3h_m}{(2\pi)^{1/2}\sigma_m} - \frac{8\sigma_m}{(2\pi)^{1/2}\delta_m} \right) \left( 1 - \frac{\exp\{-c_n s_n\}}{N(n)} \right)
\end{aligned}$$

Let  $\{\delta_m, \sigma_m, h_m, c_n, s_n\}$  satisfy the following:

$$\delta_m \rightarrow 0, \sigma_m/\delta_m \rightarrow 0, h_m/\sigma_m \rightarrow 0, c_n \rightarrow \infty, s_n \rightarrow 0, \exp\{-c_n s_n\}/N(n) \rightarrow 0.$$

Hence for any given  $(x, y)$  and a given  $\epsilon > 0$  there exist  $(M_1, N_1)$  large enough such that  $\forall m > M_1, n > N_1$

$$p(y|x, \mathcal{M}_{m,n}) > \inf_{z \in C_{\delta_m^*}^*(y), \|t-x\|^2 \leq s_n} f_0(z|t) \cdot (1 - \epsilon).$$

By Assumption 1  $f_0(y|x)$  is continuous in  $(y, x)$  and if  $f_0(y|x) > 0$ , then there exist  $(M_2, N_2)$  large enough such that  $\forall m > M_2, n > N_2$

$$\frac{f_0(y|x)}{\inf_{z \in C_{\delta_m^*}^*(y), \|t-x\|^2 \leq s_n} f_0(z|t)} \leq 1 + \epsilon$$

since  $s_n \rightarrow 0$  and  $\delta_m^* \rightarrow 0$ . Then for any  $(m, n) \geq \{\max\{M_1, M_2\}, \max\{N_1, N_2\}\}$

$$1 \leq \frac{f_0(y|x)}{p(y|x, \mathcal{M}_{m,n})} \leq \frac{f_0(y|x)}{\inf_{z \in C_{\delta_m^*}^*(y), \|t-x\|^2 \leq s_n} f_0(z|t)(1 - \epsilon)} \leq \frac{1 + \epsilon}{1 - \epsilon}.$$

Henceforth,  $\log \max\{1, f_0(y|x)/p(y|x, \mathcal{M}_{m,n})\} \rightarrow 0$  a.s.  $F_0$  as long as  $f_0(y|x)$  is continuous in  $(y, x)$  a.s.  $F_0$ . This result establishes pointwise convergence.

#### *Integrable upper bound*

Now we will establish an integrable upper bound for the application of the DCT. We know from equation (7) that for any  $x \in \mathcal{X}$  there exists  $n$  large enough such that  $\alpha_j(x) > 1/2$  for  $j \in I^n(x, s_n)$ . Similarly from equation (10) that for any  $(y, x)$  the bound for the Riemann sum is bounded below by  $1/2$  for  $m$  large enough. These facts will be used in deriving an integrable upper bound.

Before we proceed one additional change to the model  $\mathcal{M}_{m,n}$  has to be made. Note that in the previous part we defined  $\pi_{j0} = F_0(A_0^m | X_j^n)$  and some particular  $\mu_{j0}$  for each



$j$ . We define a new model  $\mathcal{M}_{m,n}^*$  by introducing  $\pi_{j00} = 0.5F_0(A_0^m|X_j^n)$  and  $\mu_{j00} = 0$  and  $\pi_{j01} = 0.5F_0(A_0^m|X_j^n)$  and  $\mu_{j01} = 2\mu_{j0}$ . Then for  $j \in I^n(x, s_n)$  and  $n, m$  large enough

$$\begin{aligned}
p(y|x, \mathcal{M}_{m,n}^*) &= \sum_{j=1}^{N(n)} \alpha_j(x) \sum_{m=1}^M F_0(A_i^m|x_j^n) \phi(y - h(x, \theta); \mu_{ji}, \sigma_m^2) \\
&\quad + 0.5F_0(A_0^m|X_j^n) \phi(y - h(x, \theta); 2\mu_{j00}, \sigma_0^2) + 0.5F_0(A_0^m|X_j^n) \phi(y - h(x, \theta); 0, \sigma_0^2) \\
&> \alpha_j(x) \left( \sum_{m=1}^M F_0(A_i^m|x_j^n) \phi(y - h(x, \theta); \mu_{ji}, \sigma_m^2) + 0.5F_0(A_0^m|X_j^n) \phi(y - h(x, \theta); 0, \sigma_0^2) \right) \\
&> 0.5[1 - 1_{A_0^m}(y)] \inf_{z \in C_{\delta_m^*}(y), \|t-x\|^2 \leq \delta} f_0(z|t) \sum_{i: A_i^m \subset C_{\delta_m^*}(y)} \lambda(A_i^m) \phi(y - h(x, \theta); \mu_{ji}, \sigma_m^2) \\
&\quad + 0.25 \cdot 1_{A_0^m}(y) F_0(A_0^m|X_j^n) \phi(y - h(x, \theta); 0, \sigma_0^2) \\
&> 0.25[1 - 1_{A_0^m}(y)] \inf_{z \in C_{\delta_m^*}(y), \|t-x\|^2 \leq \delta} f_0(z|t) \\
&\quad + 0.25 \cdot 1_{A_0^m}(y) \inf_{z \in C_{\delta}(y), \|t-x\|^2 \leq \delta} f_0(z|t) \cdot \lambda(C_{\delta}(y)) \phi(y - h(x, \theta); 0, \sigma_0^2) \\
&> 0.25[1 - 1_{A_0^m}(y)] \inf_{z \in C_{\delta}(y), \|t-x\|^2 \leq \delta} f_0(z|t) \\
&\quad + 0.25 \cdot 1_{A_0^m}(y) \inf_{z \in C_{\delta}(y), \|t-x\|^2 \leq \delta} f_0(z|t) \cdot \delta \cdot \phi(y - h(x, \theta); 0, \sigma_0^2) \\
&> 0.25 \inf_{z \in C_{\delta}(y), \|t-x\|^2 \leq \delta} f_0(z|t) \cdot \delta \cdot \phi(y - h(x, \theta); 0, \sigma_0^2)
\end{aligned}$$

for  $m, n$  large enough such that  $\delta_m^* < \delta$ ,  $s_n < \delta$  and  $\sigma_0^2$  chosen such that  $\delta \phi(y - h(x, \theta); 0, \sigma_0^2) < 1$ . Then

1. Then

$$\begin{aligned}
&\log \max \left\{ 1, \frac{f_0(y|x)}{p(y|x, \theta, \mathcal{M}_{m,n}^*)} \right\} \leq \log \max \left\{ 1, \frac{f_0(y|x)}{0.25\delta \phi(y - h(x, \theta); 0, \sigma_0^2) \inf_{z \in C_{\delta}(y), \|t-x\|^2 \leq \delta} f_0(z|t)} \right\} \\
&= \log \left\{ \frac{1}{0.25\delta \phi(y - h(x, \theta); 0, \sigma_0^2)} \cdot \frac{f_0(y|x)}{\inf_{z \in C_{\delta}(y), \|t-x\|^2 \leq \delta} f_0(z|t)} \right\} \\
&= -\log \{0.25\delta \phi(y - h(x, \theta); 0, \sigma_0^2)\} + \log \frac{f_0(y|x)}{\inf_{z \in C_{\delta}(y), \|t-x\|^2 \leq \delta} f_0(z|t)}
\end{aligned}$$

and the first logarithm above is integrable by Assumption 2 while the second logarithm

is integrable by Assumption 4.

In summary, applying DCT we get that  $d_{KL}(f_0(\cdot|\cdot), p(\cdot|\cdot, \theta, \mathcal{M}_{m,n})) \rightarrow 0$ .

*Finite model with mixtures of two normal distributions*

The final component of the proof is to show that the model  $\mathcal{M}_{m,n}$  can be rewritten as a finite regressor dependent mixture of  $f_2$  rather than just Gaussian densities. We start with a finite model as defined in equation (6)

$$p(y|x, \mathcal{M}_{m,n}) = \sum_{j=0}^n \alpha_j(x) \sum_{i=0}^m \pi_{ji} \phi(y - h(x, \theta); \mu_{ji}, \sigma_{ji}^2).$$

Given any  $j$  by Lemma 1 below there exists  $\{p_{ji}, \pi_{ji}^*, \mu_{ji}^*, \sigma_{ji}^{2*}\}_{i=0}^m$

$$\sum_{i=0}^m \pi_{ji} \phi(y - h(x, \theta); \mu_{ji}, \sigma_{ji}^2) = \sum_{i=0}^m p_{ji} \sum_{l=1}^2 \pi_{jil}^* \phi(y - h(x, \theta); \mu_{jil}^*, \sigma_{jil}^{2*})$$

such that  $\sum_{i=0}^m p_{ji} = 1$ ,  $\sum_{l=1}^2 \pi_{jil}^* = 1$  and  $\sum_{l=1}^2 \pi_{jil}^* \mu_{jil}^* = 0$  for all  $i$ . Note that predictor dependent weights can be expressed as

$$\alpha_j(x) = \frac{\exp(-c_n(x_j^{n'} x_j^n - 2x_j^{n'} x))}{\sum_{i=0}^n \exp(-c_n(x_i^{n'} x_i^n - 2x_i^{n'} x))} = \frac{\exp(\phi_{j,0} + \phi_{j,-0}x)}{\sum_{i=0}^n \exp(\phi_{i,0} + \phi_{i,-0}x)}$$

Define  $\phi_{jk,0}^* = \log(p_{jk}) + \phi_{j,0}$  and  $\phi_{jk,-0}^* = \phi_{j,-0}$  where  $p_{jk}$  are the probability weights

constructed using Lemma 1 for each  $j$ . Then

$$\begin{aligned}
p(y|x, \mathcal{M}_{m,n}) &= \sum_{j=0}^n \alpha_j(x) \sum_{i=0}^m \pi_{ji} \phi(y - h(x, \theta); \mu_{ji}, \sigma_{ji}^2) \\
&= \sum_{j=0}^n \frac{\exp(\phi_{j,0} + \phi_{j,-0}x)}{\sum_{k=0}^n \exp(\phi_{k,0} + \phi_{k,-0}x)} \sum_{i=0}^m \pi_{ji} \sum_{l=1}^2 \pi_{jil}^* \phi(y - h(x, \theta); \mu_{jil}^*, \sigma_{jil}^{2*}) \\
&= \sum_{j=0}^n \sum_{i=0}^m \frac{\exp(\phi_{ji,0}^* + \phi_{ji,-0}^*x)}{\sum_{k=0}^n \exp(\phi_{k,0}^* + \phi_{k,-0}^*x)} \sum_{l=1}^2 \pi_{jil}^* \phi(y - h(x, \theta); \mu_{jil}^*, \sigma_{jil}^{2*}) \\
&= \sum_{j=0}^{m \times n} \frac{\exp(\phi_{j,0}^* + \phi_{j,-0}^*x)}{\sum_{k=0}^{n \times m} \exp(\phi_{k,0}^* + \phi_{k,-0}^*x)} \sum_{l=1}^2 \pi_{jl}^* \phi(y - h(x, \theta); \mu_{jl}^*, \sigma_{jl}^{2*}).
\end{aligned}$$

This shows that model  $\mathcal{M}_{m,n}$  can be represented using a finite predictor-dependent mixture of 2-component mixture models with mean zero.  $\square$

**Lemma 1.** Let  $\theta_n = \{\pi_i, \mu_i, \sigma_i^2\}_{i=1}^n$  be such that

$$p(y|\theta_n) = \sum_{i=1}^n \pi_i \phi(y; \mu_i, \sigma_i^2)$$

such that  $\sum_{i=1}^n \pi_i = 1$  and  $\sum_{i=1}^n \pi_i \mu_i = 0$ . Then there exists a set of parameters  $\theta_n^* = \{p_i^*, \pi_{i1}^*, \mu_{i1}^*, \sigma_{i1}^{2*}, \pi_{i2}^*, \mu_{i2}^*, \sigma_{i2}^{2*}\}_{i=1}^n$  such that

$$p(y|\theta_n) = p(y|\theta_n^*) = \sum_{i=1}^{n-1} p_i^* \sum_{l=1}^2 \pi_{il}^* \phi(y; \mu_{il}^*, \sigma_{il}^{2*})$$

such that  $\sum_{i=1}^{n-1} p_i^* = 1$ ,  $\sum_{l=1}^2 \pi_{il}^* = 1$  and  $\sum_{l=1}^2 \pi_{il}^* \mu_{il}^* = 0$  for each  $i$ .

*Proof.* (Lemma 1)

Find  $i = \arg \min_{i \in \{1, \dots, n\}} \{|\pi_i \mu_i|\}$ . Let  $i = n$  without loss of generality. If  $\mu_i = 0$  then let  $p_1^* = \pi_i$  and  $\pi_{11}^* = \pi_{12}^* = 1/2$ ,  $\mu_{11}^* = \mu_{12}^* = \mu_i$  and  $\sigma_{11}^{2*} = \sigma_{12}^{2*} = \sigma_i^2$ . If  $\mu_i \neq 0$ , then pick any  $j \neq i$  such that  $\text{sign}(\mu_i) \neq \text{sign}(\mu_j)$ . Then let  $\pi_{11}^* = (p_1^*)^{-1} \pi_i$  and  $\pi_{12}^* = (p_1^*)^{-1} \pi_i |\mu_i| / |\mu_j|$  where  $p_1^*$  is the normalizing constant to get  $\pi_{11}^* + \pi_{12}^* = 1$ . Let  $\mu_{11}^* = \mu_i$ ,  $\mu_{12}^* = \mu_j$ ,  $\sigma_{11}^{2*} = \sigma_i^2$

and  $\sigma_{12}^{2*} = \sigma_j^2$ . Then  $\sum_{l=1}^2 \pi_{il}^* = 1$  and  $\sum_{l=1}^2 \pi_{il}^* \mu_{il}^* = 0$  for  $i = 1$ .

Let  $\tilde{\pi}_k = \pi_k$  for all  $k = 1, \dots, j-1, j+1, \dots, n-1$  and let  $\tilde{\pi}_j = \pi_j - \pi_n |\mu_n| / |\mu_j|$ .

Then

$$\sum_{i=1}^n \pi_i \phi(y; \mu_i, \sigma_i^2) = \sum_{i=1}^{n-1} \tilde{\pi}_i \phi(y; \mu_i, \sigma_i^2) + p_1^* \sum_{l=1}^2 \pi_{1l}^* \phi(y; \mu_{1l}^*, \sigma_{1l}^{2*}).$$

By induction

$$\sum_{i=1}^n \pi_i \phi(y; \mu_i, \sigma_i^2) = \sum_{i=1}^{n-1} p_i^* \sum_{l=1}^2 \pi_{il}^* \phi(y; \mu_{il}^*, \sigma_{il}^{2*})$$

where  $\sum_{l=1}^2 \pi_{il}^* = 1$  and  $\sum_{l=1}^2 \pi_{il}^* \mu_{il}^* = 0$  for each  $i$ . Note that  $\sum_{i=1}^{n-1} p_i^* = 1$  since integral of the LHS w.r.t  $y$  is 1 and integral of RHS w.r.t  $y$  is 1 iff  $\sum_{i=1}^n p_i^* = 1$ .  $\square$

*Proof.* (Theorem 2)

We want to show that  $f_0 \in KL(\Pi)$ , that is  $\Pi(\{(\theta, \eta) : d_{KL}(f_0(\cdot, \cdot), p(\cdot | \cdot, \theta, \eta)) < \epsilon\}) > 0$ .

Let  $\epsilon > 0$  be given. By Theorem 1 there exists a finite number  $k$  and a set of parameters  $(\eta_k, \theta)$  such that  $d_{KL}(f_0(\cdot | \cdot), p(\cdot | \cdot, \theta, \eta_k)) < \epsilon/3$ , where  $\eta_k = \{\pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}, \gamma'_j\}_{j=1}^k$ . Note that the mixing weights that depend on  $\{\rho_j, \gamma'_j\}_{j=1}^k$  can be rewritten as

$$\begin{aligned} \alpha_j(x) &= \frac{\exp(\rho_j + \gamma'_j x)}{\sum_{l=1}^k \exp(\rho_l + \gamma'_l x)} = \frac{\exp(\rho_j + \gamma'_j \gamma_j / 2 - \gamma'_j \gamma_j / 2 + \gamma'_j x - x'x)}{\sum_{l=1}^k \exp(\rho_l + \gamma'_l \gamma_l / 2 - \gamma'_l \gamma_l / 2 + \gamma'_l x - x'x)} \\ &= \frac{\exp((\rho_j + \gamma'_j \gamma_j / 2) - 0.5 \|x - \gamma_j\|^2)}{\sum_{l=1}^k \exp((\rho_l + \gamma'_l \gamma_l / 2) - 0.5 \|x - \gamma_l\|^2)} \equiv \frac{\alpha_j \exp(-\varphi \|x - \Gamma_j\|^2)}{\sum_{l=1}^k \alpha_l \exp(-\varphi \|x - \Gamma_l\|^2)} \\ &= \frac{\alpha_j K_\varphi(x, \Gamma_j)}{\sum_{l=1}^k \alpha_l K_\varphi(x, \Gamma_l)} \end{aligned}$$

with a set of parameters  $\{\varphi, \alpha_j, \Gamma_j\}_{j=1}^k$ . In this particular construction  $\varphi = 0.5$ , however any other positive constant could have been used.

Let  $f_{FSMR}(\cdot|\cdot, \theta, \eta_k)$  be constructed as

$$f_{FSMR}(y|x, \theta_k) = \sum_{j=1}^k p_j(x) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2})$$

$$p_j(x) = \frac{\alpha_j K_\varphi(x, \Gamma_j)}{\sum_{l=1}^k \alpha_l K_\varphi(x, \Gamma_l)}$$

and we know that  $d_{KL}(f_0(\cdot, \cdot), f_{FSMR}(\cdot|\cdot, \theta, \eta_k)) < \epsilon/3$  for some particular parameters  $(\theta, \eta_k)$ . Now, we will show that there exists a truncated at some large  $K$  infinite smoothly mixing regression  $f_{TSMR}$  such that

$$\int \log \frac{f_{FSMR}(y|x, \theta, \eta_k)}{f_{TSMR}(y|x, \theta, \eta_K)} dF_0(y, x) < \frac{\epsilon}{3}$$

where

$$f_{TSMR}(y_i|x_i, \theta, \eta_K) = \sum_{j=1}^K p_j(x_i) f_2(y_i - h(x_i, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2})$$

$$p_j(x) = V_j K_\varphi(x, \Gamma_j) \prod_{l < j} (1 - V_l K_\varphi(x, \Gamma_l)).$$

Let's construct an infinite smoothly mixing regression with parameters  $(\theta^*, \eta^*)$  based on the parameters  $(\theta, \eta_k)$  of  $f_{FSMR}$ . Let  $\theta^* = \theta$ , and  $\eta^* \equiv \{\pi_j^*, \mu_j^*, \sigma_{j1}^*, \sigma_{j2}^*, V_j^*, \Gamma_j^*\}_{j=1}^k$  be defined as

$$(\pi_h^*, \mu_h^*, \sigma_{h1}^*, \sigma_{h2}^*) = (\pi^*, \mu_1^*, \sigma_1^*, \sigma_2^*)_{(h \bmod k)} = (\pi_j, \mu_{j2}, \sigma_{j1}, \sigma_{j2})$$

$$K_\varphi(x, \Gamma_h^*) = K_\varphi(x, \Gamma_{(h \bmod k)}^*) = K_\varphi(x, \Gamma_j)$$

$$V_h^* = V_{(h \bmod k)}^* = \alpha_j \cdot \delta$$

where  $j = (h \bmod k)$  and for some small  $\delta$  with  $\max \{\alpha_j\}^{-1} > \delta > 0$  and any  $\varphi > 0$ .

Given these parameter values of  $\eta^*$  the conditional density induced by the infinite

smoothly mixing representation is

$$\begin{aligned}
f_{ISMR}(y|x, \theta^*, \eta^*) &= \sum_{j=1}^k \delta\alpha_j K_\varphi(x, \Gamma_j) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \prod_{0 < l < j} (1 - \delta\alpha_l K_\varphi(x, x_l)) \\
&+ \sum_{j=k+1}^{2 \cdot k} \delta\alpha_j K_\varphi(x, x_j) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \prod_{k < l < j} (1 - \delta\alpha_l K_\varphi(x, x_l)) \\
&\cdot \prod_{0 < i \leq k} (1 - \delta\alpha_i K_\varphi(x, x_i)) \\
&+ \sum_{j=2 \cdot k + 1}^{3 \cdot k} \delta\alpha_j K_\varphi(x, x_j) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \prod_{2 \cdot k < l < j} (1 - \delta\alpha_l K_\varphi(x, x_l)) \\
&\cdot \prod_{0 < i \leq 2 \cdot k} (1 - \delta\alpha_i K_\varphi(x, x_i)) \\
&+ \dots
\end{aligned}$$

and it combines to

$$f_{ISMR}(y|x, \theta^*, \eta^*) = \frac{\sum_{j=1}^k \delta\alpha_j K_\varphi(x, x_j) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \cdot \prod_{l < j} (1 - \delta\alpha_l K_\varphi(x, x_l))}{\sum_{j=1}^k \delta\alpha_j K_\varphi(x, x_j) \prod_{l < j} (1 - \delta\alpha_l K_\varphi(x, x_l))}.$$

It is almost immediate that  $f_{ISMR}(y|x)$  induced by infinite representation approaches  $f_{SMR}(y|x, \theta, \eta_k)$  for all values of  $(y, x)$  as  $\delta \rightarrow 0$ . To make this statement precise note that

$$\begin{aligned}
\frac{f_{SMR}(y|x, \theta, \eta_k)}{f_{ISMR}(y|x, \theta^*, \eta^*)} &= \frac{\sum_{j=1}^k \alpha_j K_\varphi(x, x_j) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2})}{\sum_{j=1}^k \alpha_j K_\varphi(x, x_j) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \prod_{l < j} (1 - \delta\alpha_l K_\varphi(x, x_l))} \\
&\cdot \frac{\sum_{j=1}^k \alpha_j K_\varphi(x, x_j) \prod_{l < j} (1 - \delta\alpha_l K_\varphi(x, x_l))}{\sum_{j=1}^k \alpha_j K_\varphi(x, x_j)} \\
&< \frac{\sum_{j=1}^k \alpha_j K_\varphi(x, x_j) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2})}{\sum_{j=1}^k \alpha_j K_\varphi(x, x_l) f_2(y - h(x, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \cdot (1 - \delta \max \alpha_l)^k} \cdot 1 \\
&= \frac{1}{(1 - \delta \max \alpha_l)^k}.
\end{aligned}$$

Then if we pick  $\delta < (1 - \exp(-\epsilon/(6k)))/\max\{\alpha_j\}$  it immediately implies that  $\log(f_{SMR}(y|x, \theta, \eta_k)/f_{ISMR}(y|x, \theta, \eta^*)) < \epsilon/6$  for all  $(y, x)$ . Now we want to show that there exists  $f_{TSMR}$  such that  $\log(f_{ISMR}(y|x, \theta, \eta^*)/f_{TSMR}(y|x, \theta, \eta^K)) < \epsilon/6$ . Let the truncated SMR be cut off at a point  $K = k * M$  for some  $M$  large enough. Then by construction of  $\eta^*$  for any  $(y, x)$  the following is true

$$f_{ISMR}(y|x, \theta, \eta^*) \left(1 - \prod_{1 \leq l < k * M} (1 - \delta \alpha_l K_\varphi(x, \Gamma_l))\right) = f_{TSMR}(y|x, \theta, \eta^K)$$

$$\frac{f_{ISMR}(y|x, \theta, \eta^*)}{f_{TSMR}(y|x, \theta, \eta^K)} = \left(1 - \prod_{1 \leq l < k * M} (1 - \delta \alpha_l K_\varphi(x, \Gamma_l))\right)^{-1}$$

The objective is to show that  $M$  large enough exists such that

$$-\log \left(1 - \prod_{1 \leq l < k * M} (1 - \delta \alpha_l K_\varphi(x, \Gamma_l))\right) < \epsilon/6.$$

This is achieved by considering let  $i^* = \arg \max_{j=1, \dots, k} \{\alpha_j\}$ . Since  $\mathcal{X}$  is bounded we can find  $\bar{K} = \max_{x \in \mathcal{X}} K_\varphi(x, \Gamma_{i^*}) > 0$ . Then

$$-\log \left(1 - \prod_{1 \leq l < k * M} (1 - \delta \alpha_l K_\varphi(x, \Gamma_l))\right) < -\log \left(1 - \prod_{1 \leq l < M} (1 - \delta \alpha_{i^*} \bar{K})\right)$$

$$= -\log (1 - (1 - \delta \alpha_{i^*} \bar{K})^M).$$

Then for  $M > \frac{\log(1 - e^{-\epsilon/6})}{\log(1 - \delta \alpha_{i^*} \bar{K})}$  this inequality is true

$$-\log (1 - (1 - \delta \alpha_{i^*} \bar{K})^M) < \epsilon/6.$$

Hence, for  $K > k * M$  we have found  $\eta^K$  such that  $\log(f_{ISMR}(y|x, \theta, \eta^*)/f_{TSMR}(y|x, \theta, \eta^K)) <$

$\epsilon/6$  for all  $(y, x)$  and it follows that

$$\int \log \frac{f_{SMR}(y|x, \theta, \eta_k)}{f_{TSMR}(y|x, \theta, \eta_K)} dF_0(y, x) < \frac{\epsilon}{3}.$$

Next we will show that there exists an open neighborhood  $\Upsilon$  of  $\eta_K$  such that for any  $\eta' \in \Upsilon$

$$\int \log \frac{f_{TSMR}(y|x, \theta, \eta_K)}{f_{TSMR}(y|x, \theta, \eta')} dF_0(y, x) < \frac{\epsilon}{3}.$$

To show this we will show that this integral is (sequentially) continuous in  $\eta'$  at  $\eta_K$ . Let  $\eta^l$  be a sequence of parameter values converging to  $\eta_K$  as  $l \rightarrow \infty$ . Then for every  $(y, x)$ , we have that  $\log(f_{TSMR}(y|x, \theta, \eta_K)/f_{TSMR}(y|x, \theta, \eta^l)) \rightarrow 1$ . To show that the integral is continuous we will use the dominated convergence theorem. We need to show that there exist integrable with respect to  $F_0$  lower and upper bounds for  $-\log(f_{TSMR}(y|x, \theta, \eta^l))$ . Since,

$$f_{TSMR}(y_i|x_i, \theta, \eta^l) = \sum_{j=1}^K p_j(x_i) f_2(y_i - h(x_i, \theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2})$$

As  $\eta^l \rightarrow \eta_K$ , therefore for  $l$  large enough and for some finite  $\bar{\mu} > \underline{\mu}$ ,  $\bar{\sigma} > \underline{\sigma}$  and  $\bar{\pi} > \underline{\pi}$  we will find that  $\pi_j^l \in (\underline{\pi}, \bar{\pi})$ ,  $\mu_j^l \in (\underline{\mu}, \bar{\mu})$ ,  $-\mu_j^l \frac{\pi_j^l}{1-\pi_j^l} \in (\underline{\mu}, \bar{\mu})$  and  $\sigma_{j1}^l, \sigma_{j2}^l \in (\underline{\sigma}, \bar{\sigma})$ . Then

$$\begin{aligned} \phi(0; 0, \underline{\sigma}) &\geq f_{TSMR}(y_i|x_i, \theta, \eta^l) \\ &\geq \frac{1_{(-\infty, \underline{\mu})} \exp\left(-\frac{(y-\bar{\mu})^2}{2\underline{\sigma}^2}\right) + 1_{(\underline{\mu}, \bar{\mu})} \exp\left(-\frac{(\mu-\bar{\mu})^2}{2\underline{\sigma}^2}\right) + 1_{(\bar{\mu}, \infty)} \exp\left(-\frac{(y-\mu)^2}{2\underline{\sigma}^2}\right)}{\sqrt{2\pi\bar{\sigma}^2}}. \end{aligned}$$

The logarithm of the upper bound is constant and finite, hence integrable. The logarithm of the lower bound is integrable by the Assumption 2 of the Theorem 1 as the conditional second moments of  $y$  are finite under  $F_0$ . Hence the integral is continuous and an open



neighborhood  $\Upsilon$  of  $\eta_K$  exists.

Finally, given any  $\eta' \in \Upsilon$ , let  $\eta^\infty = (\eta', \eta_{K+1:\infty})$  with  $\eta_{K+1:\infty}$  unrestricted. Then

$$\log \frac{f_{TSMR}(y|x, \theta, \eta')}{f_{ISMR}(y|x, \theta, \eta^\infty)} < 0$$

for any  $(y, x)$  by definition.

In conclusion, then there exists  $\eta_K$  and an open neighborhood  $\Upsilon$  of  $\eta_K$  such that for any  $\eta' \in \Upsilon$  and any  $\eta^\infty = (\eta', \cdot)$

$$\begin{aligned} & \int \log \frac{f_0(y|x)}{f_{ISMR}(y|x, \theta, \eta^\infty)} dF_0(y, x) \\ &= \int \log \frac{f_0(y|x)}{f_{SMR}(y|x, \theta, \eta_k)} dF_0(y, x) + \int \log \frac{f_{SMR}(y|x, \theta, \eta_k)}{f_{TSMR}(y|x, \theta, \eta_k)} dF_0(y, x) \\ &+ \int \log \frac{f_{TSMR}(y|x, \theta, \eta_k)}{f_{TSMR}(y|x, \theta, \eta')} dF_0(y, x) + \int \log \frac{f_{TSMR}(y|x, \theta, \eta')}{f_{ISMR}(y|x, \theta, \eta^\infty)} dF_0(y, x) \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} + 0 < \epsilon. \end{aligned}$$

Hence, if  $\eta_k$  is in the support of the prior and the priors assign a positive mass for some neighborhood of any  $\eta_k$ , then  $\Pi(\Upsilon) > 0$  and  $f_0$  is in the KL support of  $\Pi$ .  $\square$

*Proof.* (Theorem 3)

First, we would like to construct an unbiased test for the hypothesis:

$$H_0 : (f, \theta) = (f_0, \theta_0) \text{ against } H_1 : (f, \theta) \in \mathcal{V} \times \{\theta : \|\theta - \theta_0\| \geq \rho\}$$

where  $\mathcal{V} = \text{supp}(\Pi_{f_{\epsilon|x}})$ . To do so we will consider a finite set of alternative hypothesis such that their union would be a superset of  $H_1$ . Therefore, consider for some small  $\Delta > 0$  a group of hypothesis for each  $\gamma$  and  $j = 1, \dots, d$

$$H_0 : (f, \theta) = (f_0, \theta_0) \text{ against } H_1 : (f, \theta) \in \mathcal{W} \times \{\theta : (\theta - \theta_0) \in Q_\gamma, \gamma_j(\theta_j - \theta_{0,j}) > \Delta\}.$$

For any  $j$  consider only  $x$  observations such that  $x_j \gamma_j > \xi$  where  $x_j$  is the  $j$ -th coordinate of  $x$  and  $x \in Q_\gamma$ , then by construction  $x'\theta - x'\theta_0 > \xi\Delta$ . Let  $Q_{\gamma,j} = Q_\gamma \setminus \{X : |x_j| < \xi\}$ , then by the assumption of the theorem  $F_0(Q_{\gamma,j}) = \zeta > 0$ . Then we will use Chebyshev's inequality to construct strictly unbiased test for  $H_0$  against  $H_1$ . By assumptions 1 and 2 of Theorem 1 there exists  $M$  such that  $M > \sup_{x \in \mathcal{X}} \mathbb{E}_{F_0}[(y - x'\theta_0)^2 | x]$ . For any  $n$  let  $K_{n,j} = \sum_{i=1}^n 1\{x_i \in Q_{\gamma,j}\}$ . Then let  $T_{n,j} = K_{n,j}^{-1} \sum_{i=1}^n 1\{x_i \in Q_{\gamma,j}\} (y_i - x_i \theta_0)$ . Then by Chebyshev's inequality  $P_{f_0}(|T_{n,j}| \geq \varepsilon) \leq \frac{M^2}{K_{n,j} \varepsilon^2}$  and since  $\zeta > 0$  it is immediate that  $P_{f_0}(|T_{n,j}| \geq \varepsilon) \rightarrow 0$  as  $n \rightarrow \infty$ . Now we just need to show that  $\inf_{(f,\theta) \in H_1} P_{f,\theta}(|T_{n,j}| \geq \varepsilon) \rightarrow 1$  as  $n \rightarrow \infty$ . Then note that for any  $\theta \in H_1$ , we have that  $x'\theta - x'\theta_0 > \xi\Delta$ . So let  $\varepsilon = \xi\Delta/3$  and consider a statistic  $\tilde{T}_{n,j} = K_{n,j}^{-1} \sum_{i=1}^n 1\{x_i \in Q_{\gamma,j}\} (y_i - x_i \theta)$ . Then

$$P_{f,\theta}(|T_{n,j}| \geq \varepsilon) \geq P_{f,\theta}(|\tilde{T}_{n,j}| \leq \varepsilon) \geq 1 - \frac{L^2}{K_{n,j} \varepsilon^2}$$

where  $L > \sup_{x \in \mathcal{X}, f \in \mathcal{W}, \theta \in \Theta} \mathbb{E}_{f,\theta}[(y - x'\theta)^2 | x]$  for some  $L$  by conditions on the prior support. Since  $K_{n,j} \rightarrow \infty$  as  $n \rightarrow \infty$ , therefore  $P_{f,\theta}(|T_{n,j}| \geq \varepsilon) \rightarrow 1$ . These facts can be used to construct an uniformly consistent sequence of tests

$$\phi_n(X_n, Y_n) = 1\{|T_{n,j}| \geq \xi\Delta/3\}$$

and by Proposition 4.4.1 by [Ghosh and Ramamoorthi \(2003\)](#) this implies the existence of exponentially consistent tests. Note that by choosing  $\Delta > 0$  sufficiently small the union of these sets of alternative hypothesis will contain the set  $\{(f, \theta) \in \mathcal{V} \times \{\theta : \|\theta - \theta_0\| \geq \rho\}\}$  which is the original alternative hypothesis.

Given a weak neighborhood of  $\mathcal{U}_\delta(f_{0,\varepsilon|x})$  of conditional error density, let's construct exponentially consistent tests for this hypothesis:

$$H_0 : (f_{\varepsilon|x}, \theta) = (f_{0,\varepsilon|x}, \theta_0) \text{ against } H_1 : (f_{\varepsilon|x}, \theta) \in \mathcal{U}_\delta(f_{0,\varepsilon|x})^c \times \{\theta : \|\theta - \theta_0\| < v\}$$

for some small  $v > 0$ .  $\mathcal{U}_\delta(f_{0,\epsilon|x})$  is defined via some bounded uniformly continuous function  $g$ , then let  $\Delta_2$  be such that if  $|\epsilon_1 - \epsilon_2| < \Delta_2$ , then  $|g(\epsilon_1, x) - g(\epsilon_2, x)| < \delta/2$ . Then for  $f_{\epsilon|x} \in \mathcal{U}_\delta(f_{0,\epsilon|x})^c$  and any  $\theta$  such that  $|\theta - \theta_0| < v$ , where  $v$  is such that  $|x'(\theta - \theta_0)| < \Delta_2$  as  $\mathcal{X}$  is bounded, define  $f_{y|x,\theta} = f_{\epsilon|x}(\epsilon + h(x, \theta), x)$ . Then  $f_{y|x,\theta} \in \mathcal{U}_{\delta/2}(f_{0,y|x})^c$  where  $\mathcal{U}_{\delta/2}(f_{0,y|x})$  is a weak neighborhood of conditional densities  $f_0(y|x)$ . Exponentially consistent tests for such weak neighborhoods always exist, while we could not have constructed exponentially consistent tests based on unobservable  $\epsilon$ .

By choosing  $\Delta$  and  $v$  small enough the union of the sets of alternative hypothesis would contain  $\{(f_{\epsilon|x}, \theta) \in \mathcal{U}_\delta(f_{0,\epsilon|x})^c \times \{\theta : \|\theta - \theta_0\| \geq \rho\}\}$ . Then exponentially consistent tests for  $H_0 : (f, \theta) = (f_0, \theta_0)$  against  $H_1 : (f, \theta) \in (\mathcal{U}_\delta(f_{0,\epsilon|x}) \times \{\theta : \|\theta - \theta_0\| < \rho\})^c$  exist, and since  $f_0 \in KL(\Pi)$  by Theorem 2, then a straightforward application of [Schwartz \(1965\)](#) posterior consistency theorem yields the result that  $\Pi(\mathcal{W}^c | Y_n, X_n) \rightarrow 0$  a.s.  $P_{F_0}^\infty$  for any given  $\rho > 0$ .  $\square$

*Proof.* (Corollary 1)

First, we would like to construct an unbiased test for the hypothesis:

$$H_0 : (f, \theta) = (f_0, \theta_0) \text{ against } H_1 : (f, \theta) \in \mathcal{V} \times \{\theta : \|\theta - \theta_0\| \geq \rho\}$$

where  $\mathcal{V} = \text{supp}(\Pi_{f_{\epsilon|x}})$ .

By the assumption 2 of Theorem 1 there exists  $M$  such that  $M > \sup_{x \in \mathcal{X}} \mathbb{E}_{F_0}[(y - h(x, \theta_0))^2 | x]$ . Let  $T_n = n^{-1} \sum_{i=1}^n y_i - h(x_i, \theta_0)$ . Then by Chebyshev's inequality  $P_{f_0}(|T_n| \geq \varepsilon) \leq \frac{M^2}{n\varepsilon^2}$  and it is immediate that  $P_{f_0}(|T_n| \geq \varepsilon) \rightarrow 0$  as  $n \rightarrow \infty$ . Now we just need to show that  $\inf_{(f,\theta) \in H_1} P_{f,\theta}(|T_n| \geq \varepsilon) \rightarrow 1$  as  $n \rightarrow \infty$ . Then note that for any  $\theta \in H_1$ , we have that  $E_{F_0}[|h(x, \theta) - h(x, \theta_0)|] > \xi\rho$ . So let  $\varepsilon = \xi\rho/3$  and consider a statistic

$\tilde{T}_n = n^{-1} \sum_{i=1}^n y_i - h(x_i, \theta)$ . Then

$$P_{f,\theta}(|T_n| \geq \varepsilon) \geq P_{f,\theta}(|\tilde{T}_n| \leq \varepsilon) \geq 1 - \frac{L^2}{n\varepsilon^2}$$

where  $L > \sup_{x \in \mathcal{X}, f \in \mathcal{W}, \theta \in \Theta} \mathbb{E}_{f,\theta}[(y - h(x, \theta))^2 | x]$  for some  $L$  by conditions on the prior support. Since  $n \rightarrow \infty$ , therefore  $P_{f,\theta}(|T_n| \geq \varepsilon) \rightarrow 1$ . These facts can be used to construct an uniformly consistent sequence of tests

$$\phi_n(X_n, Y_n) = 1\{|T_n| \geq \xi\rho/3\}$$

By Proposition 4.4.1 by [Ghosh and Ramamoorthi \(2003\)](#) this implies the existence of exponentially consistent tests. The rest of the proof is identical to the Theorem 3.  $\square$

## 5.2 Posterior Computation

The finite model with  $k$  states is defined as

$$p(y_i|x_i, \beta, \eta_k) = \sum_{j=1}^k \alpha_j(x_i) f_2\left(y_i - x_i'\beta; \pi_j, \mu_j, h_{j1}^{-1/2}, h_{j2}^{-1/2}\right)$$

$$\alpha_j(x_i) = \frac{\exp(\rho_j + \gamma_j'x_i)}{\sum_{l=1}^k \exp(\rho_l + \gamma_l'x_i)}.$$

We propose Gibbs sampler for the estimation procedure. We introduce latent state variables  $s_i = (s_{i1}, s_{i2})$  with  $s_{i1} \in \{1, \dots, k\}$  and  $s_{i2} \in \{1, 2\}$ . Then  $p(y_i|s_i, x_i, \beta, \eta_k) = \phi(\cdot, x_i'\beta + \mu_{s_i}, h_{s_i}^{-1})$ , with  $\mu_{j1} = \mu_j$  and  $\mu_{j2} = -\frac{\pi_j}{1-\pi_j}\mu_j$  and  $P(s_i = (j, l)|x_i, \beta, \eta_k) = \alpha_j(x_i)\pi_{jl}$ .

Gibbs sampling techniques by [Geweke and Amisano \(2011\)](#) are employed to implement the zero mean conditions in Gibbs sampler. Let  $s_i = (j, l)$ , then define  $d_i$  as  $2k \times 1$  vector with value 1 on  $(2(j-1) + l)$ 'th row and 0 elsewhere. Let  $\pi_j$  be the  $j$ 'th row of  $\pi$ , then let  $2 \times 1$  vector  $C_j$  be the orthonormal complement of  $\pi_j$ . Define scalar  $\tilde{\mu}_j = C_j'(\mu_{j1}, \mu_{j2})'$ . Construct  $2k \times k$  matrix  $C = \text{Blockdiag}[C_1, \dots, C_k]$  and  $k \times 1$  vector  $\tilde{\mu} = (\tilde{\mu}_1, \dots, \tilde{\mu}_k)'$ . Then the distribution of observable  $y_i$  is

$$(y_i|x_i, s_i, \beta, \eta_k) \sim N(x_i'\beta + \tilde{\mu}'C'd_i, h_{s_i}^{-1}).$$

Let  $\zeta = (\beta', \tilde{\mu}')'$  and  $W_i = (X_i', d_i'C)$ . Finally, the distribution of observable  $y_i$  is

$$(y_i|x_i, s_i, \beta, \eta_k) \sim N(W_i\zeta, h_{s_i}^{-1})$$

and the prior for  $\zeta$  is induced by priors on  $\tilde{\mu}$  and  $\beta$ . The following priors are used: Gaussian priors for  $\beta, \rho, \gamma, \tilde{\mu}$ , a Gamma prior for  $h^2$  and a Dirichlet prior for  $\pi$ . The full

posterior is proportional to the joint distribution of observables and unobservables:

$$\begin{aligned}
p(Y_N, S_N, \beta, \eta_k | X_N) &\propto \prod_{i=1}^N |h_{s_i}|^{1/2} \exp \{ -0.5(y_i - W_i \zeta)^2 h_{s_i} \} \alpha_{s_{i1}}(x_i) \pi_{s_i} \\
&\cdot |\underline{H}_\zeta|^{1/2} \exp \{ -0.5(\zeta - \underline{\zeta})' \underline{H}_\zeta (\zeta - \underline{\zeta}) \} \\
&\cdot \prod_{j=1}^k \prod_{l=1}^2 \pi_{jl}^{\pi-1} \\
&\cdot \prod_{j=1}^k \prod_{l=1}^2 |h_{jl}|^{(\nu-2)/2} \exp \{ -0.5 \underline{s}^2 h_{jl} \} \\
&\cdot \prod_{j=1}^k |\underline{H}_\gamma|^{1/2} \exp \{ -0.5(\gamma_j - \underline{\gamma})' \underline{H}_\gamma (\gamma_j - \underline{\gamma}) \} \\
&\cdot \prod_{j=1}^k |\underline{H}_\rho|^{1/2} \exp \{ -0.5(\rho_j - \underline{\rho})' \underline{H}_\rho (\rho_j - \underline{\rho}) \}
\end{aligned}$$

where

$$\begin{aligned}
\underline{\zeta} &= [\underline{\beta}', 0']', \\
\underline{H}_\mu &= \underline{h}_\mu \cdot I_k, \\
\underline{H}_\zeta &= \begin{bmatrix} \underline{H}_\beta & 0 \\ 0 & \underline{H}_\mu \end{bmatrix}.
\end{aligned}$$

Then Gibbs sampler consists of:

1. *Conditional posterior distribution of  $\zeta$ :*

$$\begin{aligned}
p(\zeta | \dots) &\sim N \left( \bar{\zeta}, \bar{H}_\zeta^{-1} \right) \\
\bar{H}_\zeta &= \underline{H}_\zeta + \sum_{i=1}^N W_i' h_{s_i} W_i \\
\bar{\zeta} &= \bar{H}_\zeta^{-1} \left( \underline{H}_\zeta \underline{\zeta} + \sum_{i=1}^N W_i' h_{s_i} y_i \right).
\end{aligned}$$

2. *Conditional posterior distribution of  $h_{jl}$ :*

$$h_{jl} \sim Ga \left( \frac{\nu + N_{jl}}{2}, \left[ \frac{1}{2} s^2 + \frac{1}{2} \sum_{i|s_i=(j,l)} (y_i - W_i \zeta)(y_i - W_i \zeta)' \right]^{-1} \right)$$

where  $N_{jl} = \sum_{i=1}^N 1_{\{s_i=(j,l)\}}$ .

3. *Conditional posterior distribution of  $\pi$ :*

Posterior distribution of  $\pi$  is non-standard and non-conjugate since  $C$  is a function of  $\pi$  and we must account for that in our posterior simulator. Firstly, we will follow [Geweke and Amisano \(2011\)](#) in producing a unique representation of  $C_j$  for  $j = 1, \dots, k$ . Note that  $\pi_j C_j = 0$ , then construct unique  $C_j$  by first constructing  $C_j^*$  as follows:  $C_j^* = (\pi_{j1}, -\pi_{j1}^2/\pi_{j2})'$ . Then construct unique  $C_j$  by normalizing column of  $C_j^*$  to Euclidean length of 1. Since  $C_j$  is a function of  $\pi$ , we will use Metropolis within Gibbs step for each row  $j$  of  $\pi$ . Use Dirichlet distribution as a proposal for row  $\pi_j$

$$p(\pi_j | \dots) \propto \pi_{j1}^{(\pi_{j1} + N_{j1})} \pi_{j2}^{(\pi_{j2} + N_{j2})}.$$

Construct new  $\epsilon_i^* = y_i - W_i^* \zeta$ . Then Metropolis acceptance ratio for the new draw of  $\pi_j$  is

$$\frac{\exp \left( -0.5 \sum_{i|s_{i1}=j} \epsilon_i^{*'} h_{s_i} \epsilon_i^* \right)}{\exp \left( -0.5 \sum_{i|s_{i1}=j} \epsilon_i' h_{s_i} \epsilon_i \right)}$$

where  $\epsilon_i = y_i - W_i \zeta$ .

4. *Conditional posterior distribution of  $(\rho, \gamma)$ :*

Metropolis-Hastings algorithm is used to sample  $(\rho, \gamma)$ .

5. *Conditional posterior distribution of  $s_i$ :*

Gibbs sampler block for  $s_i$  has a simple multinomial distribution with

$$p(s_i = (j, l) | \dots) \propto \alpha_j(x_i) \pi_{jl} |h_{jl}|^{0.5} \exp \{-0.5(y_i - \mu_{jl} - x_i' \beta)' h_{jl} (y_i - \mu_{jl} - x_i' \beta)\}.$$

## References

- Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9(2):291–312, 2003.
- Chung, Y. and Dunson, D. The local dirichlet process. *Annals of the Institute of Statistical Mathematics*, 63(1):59–80, 2011.
- Chung, Y. and Dunson, D. B. Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660, 2009.
- Conley, T. G., Hansen, C. B., McCulloch, R. E., and Rossi, P. E. A semi-parametric bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276 – 305, 2008.
- De Iorio, M., Mller, P., Rosner, G. L., and MacEachern, S. N. An anova model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215, 2004.
- Dunson, D. B. and Park, J.-H. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.
- Geweke, J. and Amisano, G. Hierarchical markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, 26(1):1–29, 2011.



- Geweke, J. and Keane, M. Smoothly mixing regressions. *Journal of Econometrics*, 138(1):252–290, 2007.
- Ghosh, J. K. and Ramamoorthi, R. V. *Bayesian Nonparametrics*. Springer, New York, 2003.
- Griffin, J. E. and Steel, M. F. J. Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.
- Hirano, K. Semiparametric bayesian inference in autoregressive panel data models. *Econometrica*, 70(2):781–799, 2002. ISSN 00129682.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.
- Kalli, M., Griffin, J. E., and Walker, S. G. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.
- Kleijn, B. J. K. and Bickel, P. J. The semiparametric bernstein-von mises theorem. *Working paper*, 2010.
- Kottas, A. and Gelfand, A. E. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96(456):1458–1468, 2001.
- Kottas, A. and Krnjajic, M. Bayesian nonparametric modeling in quantile regression. *Scandinavian Journal of Statistics*, 36(2):297–319, 2009.
- Lancaster, T. A note on bootstraps and robustness. *Working paper*, 2003.
- Leslie, D. S., Kohn, R., and Nott, D. J. A general approach to heteroscedastic linear regression. *Statistics and Computing*, 17(2):131–146, 2007.
- MacEachern, S. N. Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, 1999.

- Muller, P., Erkanli, A., and West, M. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79, 1996.
- Müller, U. K. Risk of bayesian inference in misspecified models and the sandwich covariance matrix. *Working paper*, 2010.
- Norets, A. Approximation of conditional densities by smooth mixtures of regressions. *Annals of Statistics*, 38(3):1733–1766, 2010.
- Norets, A. and Pelenis, J. Posterior consistency in conditional density estimation by covariate dependent mixtures. *Working paper*, 2011.
- Norets, A. and Pelenis, J. Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, forthcoming, 2012.
- Papaspiliopoulos, O. and Roberts, G. O. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.
- Pati, D., Dunson, D., and Tokdar, S. Posterior consistency in conditional distribution estimation. *Working paper*, 2011.
- Pati, D. and Dunson, D. B. Bayesian nonparametric regression with varying residual density. *Working paper*, 2009.
- Robinson, P. M. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica*, 55(4):875–891, 1987.
- Schwartz, L. On bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4:10–26, 1965.
- Shen, X. Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, 97(457):222–235, 2002.
- Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian Analysis*, 5(2):319–344, 2010.

van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, 1998.

Walker, S. G. Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007.

White, H. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Wu, Y. and Ghosal, S. Posterior consistency for some semi-parametric problems. *Sankhya : The Indian Journal of Statistics*, 70(3):0–46, 2008.



---

Author: Justinas Pelenis

Title: Bayesian Semiparametric Regression

Reihe Ökonomie / Economics Series 285

Editor: Robert M. Kunst (Econometrics)

Associate Editors: Walter Fisher (Macroeconomics), Klaus Ritzberger (Microeconomics)

ISSN: 1605-7996

© 2012 by the Department of Economics and Finance, Institute for Advanced Studies (IHS),  
Stumpergasse 56, A-1060 Vienna • ☎ +43 1 59991-0 • Fax +43 1 59991-555 • <http://www.ihs.ac.at>

---

