Bayesian Semiparametric Regression

Justinas Pelenis *
pelenis@ihs.ac.at
Institute for Advanced Studies, Vienna

March 14, 2012

Abstract

This paper considers a Bayesian estimation of restricted conditional moment models with linear regression as a particular example. The standard practice in the Bayesian literature for linear regression and other semiparametric models is to use flexible families of distributions for the errors and to assume that the errors are independent from covariates. However, a model with flexible covariate dependent error distributions should be preferred for the following reasons. First, assuming that the error distribution is independent of predictors might lead to inconsistent estimation of the parameters of interest when errors and covariates are dependent. Second, the prediction intervals obtained from a model with predictor dependent error distributions are likely to be superior to the ones obtained assuming a constant error distribution. Third, modeling conditional error density might allow one to obtain a more efficient estimator of the regression coefficients under heteroscedasticity. To address these issues, we develop a Bayesian semiparametric regression 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.

*I am very thankful to Andriy Norets, Bo Honore, Jia Li, Ulrich Müller and Chris Sims as well as seminar participants at Princeton, Royal Holloway, Institute for Advanced Studies, Vienna, 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.

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 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 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 regression coefficients and conditional residual 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), Amewou-Atisso et al. (2003) and Wu and Ghosal (2008) among others. However, estimation of regression coefficients and error densities might be inconsistent if errors and regressors 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 regressor 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 consistent estimation of both regression coefficients and residual 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 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 β 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 β 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 for residual distributions conditional on any predictor value. 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 linear regression model defined by a conditional moment restriction. This flexible approach leads to consistent estimation of both regression coefficients and conditional residual 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 Gaussians and indirectly achieve posterior consistency of regression coefficients.

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 Bayesian risk under linex loss than other alternatives under heteroscedasticity and performs equally well under homoscedasticity. The alternative semiparametric Bayesian estimator based on error density modeled as a mixture of normal distributions performs worse than other methods both under heteroscedasticy 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. Section 5 concludes and suggests directions for future research. The proofs and fine details of posterior computation are contained in the Appendix.

2 Restricted Moment Model

The data consists of n observations of $(Y_n, X_n) = \{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$ 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) satsifies $\mathbb{E}_{F_0}[y|x] = h(x, \theta_0)$ 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

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

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 θ 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 Θ denotes the space of θ and $\mathcal{F}_{\epsilon|x}$ the space of conditional densities with mean zero. That is $\theta \in \Theta \subset \mathbb{R}^p$ and

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

The primary objective is to construct a model to consistently estimate the parameter

of interest θ_0, while consistent estimation of the conditional error densties $f_{0,\epsilon|x}$ 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 θ_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 θ_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 ϵ with mean μ and variance σ^2. Note that by construction a random variable ϵ 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 (η_k, θ) where θ is the parameter of interest and η_k are the nuisance parameters that induce conditional

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

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

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

where $\alpha_j(x_t)$ 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

$$\alpha_j(x_i) = \frac{\exp\left(\rho_j + \gamma_j' x_i\right)}{\sum_{l=1}^{k} \exp\left(\rho_l + \gamma_l' x_i\right)}. \tag{2}$$

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 θ and the nuissance parameters $\eta_k \equiv \left\{\pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}, \rho_j, \gamma_j'\right\}_{j=1}^{k}$. To complete the characterization of this model one would specify a prior Π_θ on Θ and a prior $\Pi_{f_{\epsilon|x}}$ on $\mathcal{F}_{\epsilon|x}$ is induced by a prior Π_η on the parameters η_k. These priors induce a joint prior $\Pi = \Pi_\theta \times \Pi_{f_{\epsilon|x}}$ on Ξ.

In Section 3 we show that for any true DGP F_0 there exists k large enough and

parameters (θ, η_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 θ_0 would be obtained with $k \to \infty$.

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:

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

where η are nuissance 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 (πDDP) proposed by Griffin and Steel (2006), probit stick-breaking process (Chung and Dunson (2009b)), kernel stick-breaking process (Dunson and Park (2008)) and local Dirichlet process (lDP) (Chung and Dunson (2009a)) 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 $j = 1, \ldots$. The covariate dependent mixing weights are

defined as:

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

where $K : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ 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

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

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

where $K : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ is a bounded kernel function, $\left\{\pi_j, \mu_j, \sigma_{j1}^2, \sigma_{j2}^2\right\} \sim G_0$, $\Gamma_j \sim H$, $V_j \sim \text{Beta}(a_j, b_j)$, $\varphi \sim \Pi_\varphi$ and $\theta \sim \Pi_\theta$ where Π_φ and Π_θ are prior distributions for φ and θ. The nuissance 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 nuissance 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 θ. We will show that this is a 'correctly' specified model for the DGP as posterior concentrates on the true parameter θ_0 and on a weak neighborhood of the true conditional densities $f_{0,\epsilon|x}$ and 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) One of the key contributions of this paper is Theorem 2 in Section 3 which proves that kernel stick-breaking processes with exponential kernel can be used to consistently estimate flexible unrestricted conditional densities. Other predictor dependent stick-breaking processes could be considered for the estimation of this model as well, and it might be of interest to do if and when other theoretical results regarding consistent conditional density estimation using those processes are provided.

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 geberating 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 $P^n_{F_0}$ probability as $n \to \infty$ 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

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

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 θ_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 (η_k, θ). Then we show that there exists k large enough and a set of parameters (η_k, θ) 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. *For any (y, x) there exists a hypercube $C(\delta, y, x)$ with side length δ and $y \in C(\delta, y, x)$ such that*

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

Let $p(\cdot|\cdot, \theta, \eta_k)$ be defined as in Equations (1) and (2). Then, for any $\epsilon > 0$ there exists (η_k, θ) 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 the proposal of mixing weights $\alpha(x)$ is 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 pseudotrue values of θ 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 θ_0. Hence, we provide the necessary theoretical foundation for the use of infinite mixture model.

For the infinite mixture model defined in the Equation 1, the priors $G_0, H, \Pi_V, \Pi_\varphi, \Pi_\theta$ and a choice of kernel function K_φ induce a prior Π on Ξ. A conditional density function f_x is said to be in the *KL support* of the prior Π (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 Π 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 Π *be induced by the priors* $G_0, H, \Pi_V, \Pi_\varphi, \Pi_\theta$. *If the priors are such that* θ_0 *is an interior point of support of* Π_θ, $\Pi(\sigma_{j1} < \delta) > 0$ *for any* $\delta > 0$ *and* $\mathcal{X} \subset supp(H)$, *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 Π.

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 θ_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. Let $\gamma \in \{-1, 1\}^D$ and define a quadrant $Q_\gamma = \{z \in \mathbb{R}^d : z_j\gamma_j > 0 \text{ for all } j = 1, \ldots, d$. Furthermore, we require an additional assumption.

Assumption A: For any γ, $F_0(Q_\gamma \backslash \{X : |x_i| < \xi\}) > 0$ for each $i = 1, \ldots, d$ and some $\xi > 0$. This is an assumption used by Wu and Ghosal (2008) and it plays a similar role to the assumption of no multicolinearity in the DGP so that θ_0 can be identified.

Theorem 3. *An (almost) immediate implication of Schwartz (1965). Suppose that* F_0 *satisfy the assumptions of the Theorem 1 and Assumption A and that the prior distributions satisfy the requirements of the Theorem 2. Furthermore, the prior is restricted so that* $\mathbb{E}_f[\epsilon^2|x] < L$ *for all* $x \in \mathcal{X}$ *and all* $f \in supp(\Pi_{f_{\epsilon|x}})$ *and some large* L. *Let* $\mathcal{W} = \mathcal{U}_\delta(f_{0,\epsilon|x}) \times \{\theta : ||\theta - \theta_0|| < \rho\}$ *for some* $\delta > 0$ *and* $\rho > 0$, *then* $\Pi(\mathcal{W}^c|Y_n, X_n) \to 0$ *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. This would require specifying additional assumptions on the function $h(\cdot, \cdot)$ such as Lipshitz continuity and an equivalent version of the Assumption A. These results will be provided in the near future.

In this section we presented novel results of posterior consistency properties for parametric and nonparametric parts of restricted moment moedels with conditional moment constraint. Given the desirable theoretical properties that both parametric and nuissance parts are consistently estimated it achieves the two objectives. Firstly, the estimation of the parameter of interest is consistent. Secondly, consistent estimation of the nuissance parameter, which is 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.

4 Simulation Examples

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

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

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 $a = 0.454\ldots$

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, 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. Second, Bayesian inference based on a normal regression model (NLR), where $\epsilon_i | x_i \sim N(0, h^{-1})$ with priors $\theta \sim N(0, (0.01 I_2)^{-1}), 3h \sim \chi^2_3$. Third, Bayesian inference based on a normal mixture linear regression model (MIX) with $\epsilon_i | x_i, s \sim N(\mu_s, (h h_s) - 1)$ and $P(s = j) = \pi_j, j = 1, 2, 3$ with priors $\theta \sim N(0, (0.01 I_2)^{-1}), 3h \sim \chi^2_3, 3h_j \sim \chi^2_3, (\pi_1, \pi_2, \pi_3) \sim Dirichlet(3, 3, 3)$ and $\mu_j \overset{\text{iid}}{\sim} N(0, (0.4h)^{-1})$. Fourth, Bayesian inference on a linear regression model (TLR) with student-t disturbances with $\epsilon_i | x_i \sim t(0, h^{-1}, \lambda)$. The priors are set to $\theta \sim N(0, (0.01 I_2)^{-1})$, $3h \sim \chi^2_3$ and $\lambda \sim \exp(10)$. Finally, Bayesian inference based on the conditional linear regression model (CLR) proposed in this paper. We consider the finite model with $n = 5$ number of states. The priors are set to $\theta \sim N(0, (0.01 I_2)^{-1})$, $\gamma_j \sim N(0, (0.01 I_2)^{-1}), 3h_{ji} \sim \chi^2_3, \tilde{\mu}_j \sim N(0, 0.25^{-1}), \underline{\pi} = 10$ for all $j = 1, \ldots, n$ and $i = 1, 2$. Posterior computation and full description of the priors are contained in the Appendix.

The parameter of interest is $\beta \in \mathbb{R}$ and we consider four separate criteria for the evaluation of the performance of the proposed estimators. We will compute both root mean squared error (RMSE) and risk under linex loss as suggested by Müller (2010),

where linex loss function at $\alpha = \beta = 0$ is

$$l_n(\theta, a) = \exp[2\sqrt{n}(\beta - a)] - 2\sqrt{n}(\beta - a) - 1$$

with $a \in \mathbb{R}$. 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, linex loss, 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. In Cases (i) and (iii) the best performing models should be OLS and NLR since it is well know that OLS estimator achieves the semi-parametric efficiency under homoscedasticity. Note that both models MIX and TLR perform worse in Cases (iii) and (iv) due to (conditional) assymetries of the error distribution. In Case (iii) this is expected since the pseudotrue value of β is not the true $\beta_0 = 0$ for either of the models. One has to be careful when estimating linear models with flexible unconditional error disturbances such as mixture of normals or symmetric student-t disturbances. As demonstrated in this simulation example estimation of linear models with student-t disturbances (and with other more 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 β might be different from true

Table 1: Simulation results

Method	Criterion	Case (i)	Case (ii)	Case (iii)	Case (iv)
CLR		0.070	0.060	0.072	0.067
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
TLR		0.070	0.067	0.080	0.081
CLR		1	1	1	1
OLS		0.97	1.39	1.05	1.28
NLR	Linex Loss Risk	0.99	1.84	1	1.78
MIX		1.02	1.47	1.54	2.15
TLR		0.98	1.58	1.31	5.44
CLR		0.971	0.948	0.965	0.935
OLS		0.946	0.948	0.949	0.942
NLR	Coverage	0.950	0.805	0.947	0.847
MIX		0.947	0.843	0.905	0.841
TLR		0.948	0.843	0.911	0.836
CLR		0.30	0.23	0.30	0.25
OLS		0.28	0.27	0.28	0.27
NLR	Interval Length	0.28	0.18	0.28	0.20
MIX		0.28	0.20	0.27	0.21
TLR		0.28	0.20	0.27	0.21

Notes: DGPs are in columns and methods of inference in rows. Entries are RMSE, risk under linex loss using the row model divided by the risk of the method (CLR) proposed in this paper, 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, and the risks are estimated from 1000 draws from each DGP. Bayes actions are determined numerically from the posterior distribution.

β_0, for example, when disturbances are asymmetric. As expected, the model proposed in this paper outperforms in the heteroscedastic cases and performs comparably in the homoscedastic cases.

5 Concluding Remarks

A related line of research would be to consider estimating parametric models with conditional median or quantile restraints. Conditional residual distributions could be modeled as a mixture of mixtures of two normal distributions that satisfy the necessary quantile restriction and the flexibility is achieved as the mixture weights vary with predictors. The general idea of this strand of research is to estimate parametric models with extremely flexible varying residual distributions and compare the performance of estimation with alternative misspecified models.

More importantly, it would be good to determine the conditions on the true data generating processes such that asymptotic semiparametric efficiency bound could be achieved by estimating the proposed models. Or alternatively, to provide some conclusive theoretical evidence that efficiency bounds will not be obtained by estimation of the models that are too flexible in the nuissance parameter estimation and provide alternative practical suggestions to achieve semiparametric efficiency bounds.

6 Appendix

6.1 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, $j = 0,1,\ldots,m$, be a partition of $\mathcal{Y}$, where $A_1^m,\ldots,A_m^m$ are adjacent cubes with side length h_m and A_0^m is the rest of set $\mathcal{Y}$. Let B_j^n, $j = 0,1,\ldots,N(n)$, be a partition of $\mathcal{X}$ with $N(n) = n^d$, where $B_1^n,\ldots,B_{N(n)}^n$ are adjacent cubes with side length λ_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 $\lambda_n \to 0$. Second, the area covered by the finer partition has to increase and eventually cover the whole support of $\mathcal{X}$, i.e. $\lambda_n^d N(n) \to \infty$. Furthermore, let x_i^n be the center of B_j^n, $j = 1,\ldots,N(n)$ 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

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

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

We propose mixing probabilities such that

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

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 $c_n \to \infty$ and

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

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 $I_1^n(x, s_n) = \left\{j : ||x_j^n - x||^2 < s_n\right\}$. Using the arguments of the proof of Proposition 4.1. we know that for (n, m) large enough for any $j \in I_1^n(x, s_n)$

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

Assume that $|\partial h(x, \theta_0)/x_j| < T$ for all $x \in \mathcal{X}$. Next, let $\delta_m^* = \delta_m + (d^{1/2}\lambda_n \cdot 1_d)'T$ where $\delta_m \to 0$. Then for $j \in I_1^n(x, s_n)$ and $A_i^m \subset C_{\delta_m^*}(y)$, where $C_\delta(y)$ is a cube with center y and side length δ,

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

Then, inequality above and equation (6) imply

$$
\begin{aligned}
p(y|x,\mathcal{M}_{m,n}) > & \sum_{j\in I_1^n(x,s_n)} \sum_{i:A_i^m\subset C_{\delta_m^*}(y)} F(A_i^m|x_j^n)\frac{\exp\left(-c_n(x_j^{n\prime}x_j^n-2x_j^{n\prime}x)\right)}{\sum_{k=1}^{N(n)}\exp\left(-c_n(x_k^{n\prime}x_k^n-2x_k^{n\prime}x)\right)} \\
& \times \phi\left(y-h(x,\theta);\mu_{ji},\sigma_{ji}^2\right) \\
& \geq \inf_{z\in C_{\delta_m^*}(y),||t-x||^2\leq s_n} f(z|t)\cdot[1-\exp\{-c_n s_n\}/N(n)] \\
& \times \sum_{i:A_i^m\subset C_{\delta_m^*}(y);j\in I_1^n(x,s_n)} \lambda(A_i^m)\phi\left(y-h(x,\theta);\mu_{ji},\sigma_{ji}^2\right).
\end{aligned}
$$

Next, I show that there exists a set of parameters $\left\{\theta,\{\mu_{ji}\}_{j=0,i=0}^{N(n),m}\right\}$ such that

$$
\sum_{i:A_i^m\subset C_{\delta_m^*(y)};j\in I_1^n(x,s_n)} \lambda(A_i^m)\phi\left(y-h(x,\theta);\mu_{ji},\sigma_{ji}^2\right) \tag{8}
$$

approaches 1 as (n,m) increase.

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

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

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 $\mu_{ji} = c_i^m + d_j^m - h(x_j^n,\theta_0)$ where $d_j^m\in[-h_m/2,h_m/2]$ and let μ_{j0} be

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

if $F(A_0^m|x_j^n)>0$ and $\mu_{j0}=0$ otherwise. Then we will show that there exists d_j^m such

that equation (9) is satisfied. Define function $G(d_j^m)$ as

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

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

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

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

Now we can revisit equation (8) and show that it converges to 1. Let $\sigma_{ji} = \sigma_m$ if $i > 0$. For m large enough it is true that $\emptyset \neq \left\{i : A_i^m \subset C_{\delta_m^*(y)}\right\}$, therefore

$$\sum_{i:A_i^m \subset C_{\delta_m^*(y)}; j \in I_1^n(x,s_n)} \lambda(A_i^m)\phi\left(y - h(x,\theta_0); \mu_{ji}, \sigma_m^2\right) =$$

$$\sum_{i:A_i^m \subset C_{\delta_m^*(y)}; j \in I_1^n(x,s_n)} \lambda(A_i^m)\phi\left(y - (h(x,\theta_0) - h(x_j^n,\theta_0)); c_i^m + d_j^m, \sigma_m^2\right).$$

Let $y_j^* = y - (h(x,\theta_0) - h(x_j^n,\theta_0))$, then note that since $j \in I_1^n(x,s_n)$ we have that

$C_{\delta_m}(y_j^*) \subset C_{\delta_m^*}(y)$ and

$$\sum_{i:A_i^m \subset C_{\delta_m^*}(y); j \in I_1^n(x,s_n)} \lambda(A_i^m)\phi\left(y - (h(x,\theta_0) - h(x_j^n,\theta_0)); c_i^m + d_j^m, \sigma_m^2\right)$$
$$\geq \sum_{i:A_i^m \subset C_{\delta_m}(y_j^*); j \in I_1^n(x,s_n)} \lambda(A_i^m)\phi\left(y_j^*; c_i^m + d_j^m, \sigma_m^2\right).$$

By Lemmas 1 and 2 in Norets and Pelenis (2012) (with a minor adjustment in the proofs due to uncentered positions of μ_{ji}) the bound for the sum in equation (8) is

$$\sum_{i:A_i^m \subset C_{\delta_m}(y_j^*)} \lambda(A_i^m)\phi\left(y_j^*; c_i^m + d_j^m, \sigma_m^2\right) \geq 1 - \frac{3h_m}{(2\pi)^{1/2}\sigma_m} - \frac{8\sigma_m}{(2\pi)^{1/2}\delta_m}$$

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

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

Combining all these pieces together we get that

$$p(y|x,\mathcal{M}_{m,n}) > \inf_{z \in C_{\delta_m^*}(y), ||t-x||^2 \leq s_n} f(z|t) \cdot \left[1 - \frac{\exp\{-c_n s_n\}}{N(n)}\right] \cdot \left[1 - \frac{3h_m}{(2\pi)^{1/2}\sigma_m} - \frac{8\sigma_m}{(2\pi)^{1/2}\delta_m}\right]. \tag{10}$$

Given $\epsilon > 0$ there exist (M_1, N_1) large enough such that $\forall\, m > M_1,\ n > N_1$

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

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

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

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

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

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

Integrable upper bound

Let $(y,x) \in \mathcal{Y} \times \mathcal{X}$. Since $\mathcal{Y}$ and $\mathcal{X}$ are bounded sets there exist (M_3, N_3) large enough such that $\forall\, m > M_3$, $n > N_3$ we have that $y \notin A_0^m$ and $x \notin B_0^n$ and equation (10) applies. Therefore, there exist (M_4, N_4) large enough such that $\forall\, m > M_4,\ n > N_4$ we have that

$$p(y|x, \mathcal{M}_{m,n}) > \inf_{z \in C_{\delta_m^*}(y), ||t-x|| \leq \delta} f(z|t)/2$$

as $\delta_m^*/2 < \delta$ for any given x and $\sqrt{s_n} < \delta$ if (m,n) are large enough. Then for any $(m,n) \geq \{\max\{M_3, M_4\}, \max\{N_3, N_4\}\}$

$$\begin{aligned}\log \max\left\{1, \frac{f(y|x)}{p(y|x, \theta_{m,n}, \mathcal{M}_{m,n})}\right\} &\leq \log \max\left\{1, \frac{f(y|x)}{\inf_{z \in C_{\delta_m^*}(y), ||t-x|| \leq \delta} f(z|t)/2}\right\} \\ &= 2\frac{f(y|x)}{\inf_{z \in C_{\delta_m^*}(y), ||t-x|| \leq \delta} f(z|t)}\end{aligned}$$

and the expression above is integrable by Assumption 3. In summary, applying DCT we get that $d_{KL}(f_0(\cdot|\cdot), p(\cdot|\cdot, \mathcal{M})) \to 0$.

Finite model with mixtures of two normal distributions

We start with a finite model as defined in equation (6)

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

Given any j by Lemma 1 below there exists $\left\{p_{ji}, \pi^*_{ji}, \mu^*_{ji}, \sigma^{2*}_{ji}\right\}_{i=0}^{m}$

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

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

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

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

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

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. □

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

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

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 $i = \arg\min_{i \in \{1,\ldots,n\}} \{|\pi_i \mu_i|\}$. 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 $\sigma_{12}^{2*} = \sigma_{11}^{2*} = \sigma_i^2$. If $\mu_i \neq 0$, then pick any $j \neq i$ such that $sign(\mu_i) \neq sign(\mu_j)$. 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 $k = 1, \ldots, j-1, j+1, \ldots, n-1$ 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

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

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 $\sum_{i=1}^{n} p_i^* = 1$. □

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 Thoerem1 there exists a finite number k and a set of parameters (η_k, θ) such that $d_{KL}(f_0(\cdot|\cdot), p(\cdot|\cdot,\theta,\eta_k)) < \epsilon/3$, where $\eta_k = \left\{\pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}, \gamma_j'\right\}_{j=1}^k$. Note that the mixing weights that depend on $\{\rho_j, \gamma_j'\}_{j=1}^k$ can be rewritten as

$$\begin{aligned}
\alpha_j(x) &= \frac{\exp\left(\rho_j + \gamma_j' x\right)}{\sum_{l=1}^k \exp\left(\rho_{l,0} + \gamma_l' x_i\right)} = \frac{\exp\left(\rho_j + \gamma_j'\gamma_j/2 - \gamma_j'\gamma_j/2 + \gamma_j' x - x'x\right)}{\sum_{l=1}^k \exp\left(\rho_l + \gamma_l'\gamma_l/2 - \gamma_l'\gamma_l/2 + \gamma_l' x - x'x\right)} \\
&= \frac{\exp\left((\rho_j + \gamma_j'\gamma_j/2) - 0.5||x-\gamma_j||^2\right)}{\sum_{l=1}^k \exp\left((\rho_l + \gamma_l'\gamma_l/2) - 0.5||x-\gamma_l||^2\right)} \equiv \frac{\alpha_j \exp\left(-\varphi||x-\Gamma_j||^2\right)}{\sum_{l=1}^k \alpha_l \exp\left(-\varphi||x-\Gamma_l||^2\right)} \\
&= \frac{\alpha_j K_\varphi(x,\Gamma_j)}{\sum_{l=1}^k \alpha_l K_\varphi(x,\Gamma_l)}
\end{aligned}$$

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

$$\begin{aligned}
f_{FSMR}(y|x,\theta_k) &= \sum_{j=1}^k p_j(x) f_2(y - h(x,\theta); \pi_j, \mu_j, \sigma_{j1}, \sigma_{j2}) \\
p_j(x) &= \frac{\alpha_j K_\varphi(x,\Gamma_j)}{\sum_{l=1}^k \alpha_l K_\varphi(x,\Gamma_l)}
\end{aligned}$$

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

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

where

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

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

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

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

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

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

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

Given these parameter values of η^* the conditional density induced by the infinite smoothly mixing representation is

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

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

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

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 $\log(f_{ISMR}(y|x,\theta,\eta^*)/f_{TSMR}(y|x,\theta,\eta^N)) < \epsilon/6$. Let the truncated SMR be cut off at a point $N = k*M$ for some M large enough. Then by construction of η^* for any (y,x) the following is true

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

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 $i^*=\arg\max_{j=1,\cdots,k}\{\alpha_j\}$. Since $\mathcal{X}$ is bounded we can find $\overline{K}=\max_{x\in\mathcal{X}}K_\varphi(x,\Gamma_{i^*})>0$. Then

$$\begin{aligned}-\log\left(1-\prod_{1\leq l<k*M}(1-\delta\alpha_l K_\varphi(x,\Gamma_l)\right)&<-\log\left(1-\prod_{1\leq l<M}(1-\delta\alpha_{i^*}\overline{K})\right)\\&=-\log\left(1-(1-\delta\alpha_{i^*}\overline{K})^M\right).\end{aligned}$$

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

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

Hence, for $N>k*M$ we have found η^N such that $\log(f_{ISMR}(y|x,\theta,\eta^*)/f_{TSMR}(y|x,\theta,\eta^N))<\epsilon/6$ for all (y,x) and it follows that

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

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

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

To show this we will show that this integral is (sequentially) continuous in η' at η_N. Let η^l be a sequence of parameter values converging to η_N as $l\to\infty$. Then for every (y,x), we have that $\log(f_{TSMR}(y|x,\theta,\eta_N)/f_{TSMR}(y|x,\theta,\eta^l))\to 1$. 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,

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

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

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

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 Υ of η_N exists.

Finally, given any $\eta' \in \Upsilon$, let $\eta^\infty = (\eta', \eta_{N+1:\infty})$ with $\eta_{N+1:\infty}$ 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 η_N and an open neighborhood Υ of η_N such that for

any $\eta' \in \Upsilon$ and any $\eta^\infty = (\eta', \cdot)$

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

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

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 γ and $j = 1, \ldots, d$

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

By Assumption A 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 $Q_{\gamma,j} = Q_\gamma \backslash \{X : |x_j| < \xi\}$, then by Assumtion A $F_0(Q_{\gamma,j}) = \zeta > 0$. 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 $P_{f_0}(|T_{n,j}| \geq \varepsilon) \to 0$ as $n \to \infty$. Now we just need to show that $\inf_{(f,\theta)\in H_1} P_{f,\theta}(|T_{n,j}| \geq \varepsilon) \to 1$ as $n \to \infty$. 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 $K_{n,j} \to \infty$ as $n \to \infty$, therefore $P_{f,\theta}(|T_{n,j}| \geq \varepsilon) \to 1$. These facts can be used to construct an uniformally 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 $\mathcal{U}_\delta(f_{0,\epsilon|x})$ of conditional error density, let's construct exponentially consistent tests for this hypothesis:

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

for some small $\upsilon > 0$. $\mathcal{U}_\delta(f_{0,\epsilon|x})$ is defined via some bounded uniformly continuous function g, then let Δ_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 θ such that $|\theta - \theta_0| < \upsilon$, where υ 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 ϵ.

By choosing Δ and υ 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 straigtforward application of Schwartz (1965) posterior consistency theorem yields the result that $\Pi(\mathcal{W}^c|Y_n, X_n) \to 0$ a.s. $P_{F_0}^\infty$ for any given $\rho > 0$. □

6.2 Posterior Computation

The finite model with n states is defined as

$$p(y_t|x_t,\beta,\theta_n) = \sum_{j=1}^{n} \alpha_j(x_t) f_2\left(y_t - x_t'\beta; \pi_j, \mu_j, h_{j1}^{-1/2}, h_{j2}^{-1/2}\right)$$
$$\alpha_j(x_t) = \frac{\exp\left(\gamma_{j,0} + \gamma_j' x_t\right)}{\sum_{i=1}^{n} \exp\left(\gamma_{i,0} + \gamma_i' x_t\right)}.$$

We propose Gibbs sampler for the estimation procedure. We introduce latent state variables $s_t = (s_{t1}, s_{t2})$ with $s_{t1} \in \{1,\ldots,n\}$ and $s_{t2} \in \{1,2\}$. Then $p(y_t|s_t,x_t,\theta) = \phi(\cdot, x_t'\beta + \mu_{s_t}, h_{ji}^{-1})$ where $\mu_{j1} = \mu_j$ and $\mu_{j2} = -\frac{\pi_j}{1-\pi_j}\mu_j$ and $P(s_t = (j,i)|x_t,\theta) = \alpha_j(x_t)\pi_{ji}$.

We will use some of the Gibbs sampling techniques used by Geweke and Amisano (2011) to implement the zero mean conditions in Gibbs sampler. To do so consider the following redefinitions. Define 2×1 vector $\mu_j = (\mu_{j1}, \mu_{j2})'$. Let $s_t = (j,i)$, then define d_t as $2n \times 1$ vector with value 1 on $(2(j-1)+i)$'th row and 0 elsewhere. Let π_j be the j'th row of π, then let 2×1 vector C_j be orthonormal compliment of π_j. Define scalar $\tilde{\mu}_j = C_j'\mu_j$. Construct $2n \times n$ matrix $C = Blockdiag[C_1,\ldots,C_n]$ and $n \times 1$ vector $\tilde{\mu} = (\tilde{\mu}_1,\ldots,\tilde{\mu}_n)'$. Then the distribution of observable y_t is

$$(y_t|x_t, s_t = (j,i), \theta) \sim N(x_t'\beta + \tilde{\mu}'C'd_t, h_{ji}^{-1}).$$

Let $\zeta = (\beta', \tilde{\mu}')'$ and $W_t = (X_t', d_t'C)$. Finally,the distribution of observable y_t is

$$(y_t|x_t, s_t = (j,i), \theta) \sim N(W_t\zeta, h_{ji}^{-1})$$

and the prior for ζ is induced by priors on $\tilde{\mu}$ and β. The following priors are used: Gaussian prior for β, γ, $\tilde{\mu}$, inverse Gamma prior for σ^2 and Dirichlet prior for π. The full

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

$$
\begin{aligned}
p(Y_T, S_T, \theta | X_T) \propto & \prod_{t=1}^{T} |h_{s_t}|^{1/2} \exp\left\{-0.5(y_t - W_t\zeta)^2 h_{s_t}\right\} \alpha_{s_{t1}}(x_t)\pi_{s_t} \\
& \cdot |\underline{H}_\zeta|^{1/2} \exp\left\{-0.5(\zeta - \underline{\zeta})' \underline{H}_\zeta (\zeta - \underline{\zeta})\right\} \\
& \cdot \prod_{j=1}^{n} \prod_{i=1}^{2} \pi_{ji}^{\underline{\pi}-1} \\
& \cdot \prod_{j=1}^{n} \prod_{i=1}^{2} |h_{ji}|^{(\underline{\nu}-2)/2} \exp\left\{-0.5\underline{s}^2 h_{ji}\right\} \\
& \cdot \prod_{j=1}^{n} |\underline{H}_\gamma|^{1/2} \exp\left\{-0.5(\gamma_j - \underline{\gamma})' \underline{H}_\gamma (\gamma_j - \underline{\gamma})\right\}
\end{aligned}
$$

where

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

Then Gibbs sampler consists of:

1. *Conditional posterior distribution of* ζ:

$$
\begin{aligned}
p(\zeta | \ldots) &\sim N\left(\overline{\zeta}, \overline{H}_\zeta^{-1}\right) \\
\overline{H}_\zeta &= \underline{H}_\zeta + \sum_{t=1}^{T} W_t' h_{s_t} W_t \\
\overline{\zeta} &= \overline{H}_\zeta^{-1} \left(\underline{H}_\zeta \underline{\zeta} + \sum_{t=1}^{T} W_t' h_{s_t} y_t\right) .
\end{aligned}
$$

2. *Conditional posterior distribution of* h_{ji}:

$$h_{ji} \sim Ga\left(\frac{\underline{\nu}+T_{ji}}{2}, \left[\frac{1}{2}\underline{s}^2+\frac{1}{2}\sum_{t|s_t=(j,i)}(y_t-W_t\zeta)(y_t-W_t\zeta)'\right]^{-1}\right)$$

where $T_{ji}=\sum_t 1_{\{s_t=(j,i)\}}$.

3. *Conditional posterior distribution of* π:

Posterior distribution of π is non-standard and non-conjugate since C is a function of π 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_k for $k=1,\ldots,n$. Note that $\pi_k C_k=0$, then construct unique C_k by first constructing C_k^* as follows: $C_k^*=(\pi_{k1},-\pi_{k1}^2/\pi_{k2})'$. Then construct unique C_k by normalizing column of C_k^* to Euclidean length of 1. Since C_k is a function of π, we will use Metropolis within Gibbs step for each row k of π. Use Dirichlet distribution as a proposal for row π_k

$$p(\pi_k|\ldots)\propto \pi_{k1}^{(\underline{\pi}+T_{k1})}\pi_{k2}^{(\underline{\pi}+T_{k2})}.$$

Construct new $\epsilon_t^*=y_t-W_t^*\zeta$. Then Metropolis acceptance ratio for the new draw of π_k is

$$\frac{\exp\left(-0.5\sum_{t|s_{t1}=k}\epsilon_t^{*'}h_{s_t}\epsilon_t^*\right)}{\exp\left(-0.5\sum_{t|s_{t1}=k}\epsilon_t'h_{s_t}\epsilon_t\right)}$$

where $\epsilon_t=y_t-W_t\zeta$.

4. *Conditional posterior distribution of* γ:

Metropolis-Hastings algorithm is used to sample γ.

5. *Conditional posterior distribution of* s_t:

Gibbs sampler block for s_t has a simple multinomial distribution with

$$p(s_t = (j,i)|\ldots) \propto \alpha_j(x_t)\pi_{ji}|h_{ji}|^{0.5} \exp\Big\{-0.5(y_t - \mu_{ji} - x_t'\beta)'h_{ji}(y_t - \mu_{ji} - x_t'\beta)\Big\}.$$

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*, pages 1–22, 2009a.

Chung, Y. and Dunson, D. B. Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660, 2009b.

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.

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.

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 regresssions. *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.

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.

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.

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.