Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models*

Andriy Norets
Brown University

Current version: August 2018

Abstract

This paper develops a Markov chain Monte Carlo (MCMC) method for a class of models that encompasses finite and countable mixtures of densities and mixtures of experts with a variable number of mixture components. The method is shown to maximize the expected probability of acceptance for cross-dimensional moves and to minimize the asymptotic variance of sample average estimators under certain restrictions. The method can be represented as a retrospective sampling algorithm with an optimal choice of auxiliary priors and as a reversible jump algorithm with optimal proposal distributions. The method is primarily motivated by and applied to a Bayesian nonparametric model for conditional densities based on mixtures of a variable number of experts.

Keywords: Bayes, variable dimension model, reversible jump, MCMC, retrospective sampling, mixtures, mixture of experts, covariate dependent mixture, kernel mixture.

*I am grateful to participants of MCMCSki 2016, ISBA 2016 and seminars at University of Pennsylvania, Chicago, Princeton, and Yale for helpful discussions.

1 Introduction

Models with parameters of variable dimension play an important role in the Bayesian approach to inference. First of all, model comparison can be naturally performed in this framework. Second, many Bayesian non-parametric models, for example those based on varying degree polynomials or mixtures of densities, can be formulated as variable dimension models. The main approaches to MCMC estimation of such models are the reversible jump MCMC (RJMCMC) (Green (1995)), the method of auxiliary prior distributions (Carlin and Chib (1995)), and the birth-death process of Stephens (2000). These approaches require selection of proposal distributions, birth distributions, or auxiliary priors, which is a non-trivial task, especially, in complex models. The literature on choice of efficient proposals for RJMCMC is not very large and the suggested proposals, while quite sensible, appear to be mostly heuristically motivated (see a review in Section 4.1 of a survey by Hastie and Green (2012)).

In this paper, I develop optimal RJMCMC proposals of a certain type for models with a nesting structure. The RJMCMC algorithm under consideration is restricted to move only between the adjacent nested submodels without changing the parameters of the smaller submodel. Under these restrictions, the optimal proposal simulates the parameters present only in the larger submodel from their posterior distribution conditional on the parameters in the smaller submodel. The idea is rather natural and it has appeared in the literature at least in the form of centering the proposal distribution on the conditional posterior mode (see a discussion of the conditional maximization approach in Brooks et al. (2003)). The theoretical contribution of the present paper is to rigorously show that the conditional posterior proposal is optimal in a sense that it maximizes the expected

probability of acceptance for between-submodel moves and minimizes the asymptotic variance of MCMC sample average estimators under additional restrictions.

The proposed algorithm can also be represented as a combination of auxiliary priors approach of Carlin and Chib (1995) and retrospective sampling of Papaspiliopoulos and Roberts (2008) with an optimal choice of auxiliary priors restricted to have a recursive form. The auxiliary priors and retrospective sampling representation of the algorithm was developed before the RJMCMC representation, and, hence, the former is presented before the latter below.

The main motivation and application for the theoretical results described above is a practical MCMC algorithm for estimation of a Bayesian nonparametric model for conditional distributions. The model is a mixture of Gaussian regressions or experts with covariate dependent mixing weights and a variable number of mixture components. Related mixture of experts models with a fixed or a pre-selected number of components demonstrate excellent performance in applications and simulations (Jacobs, Jordan, Nowlan, and Hinton (1991), Jordan and Xu (1995), Peng, Jacobs, and Tanner (1996), Wood, Jiang, and Tanner (2002), Geweke and Keane (2007), Villani et al. (2009)). However, in the context of nonparametric conditional density estimation, the frequentist properties of standard Bayesian model selection procedures applied to choosing the number of components are not understood. Moreover, model averaging, which in this context is equivalent to a model with a varying number of mixture components, is the preferred option from the Bayesian perspective. Norets and Pati (2017) show that under rather standard priors and some regularity assumptions, the posterior in a model with a varying number of experts contracts at an adaptive optimal rate up to a log factor; moreover, the rate is not affected

by the presence of irrelevant covariates in the model. Given these attractive asymptotic guarantees, which do not appear to be currently available for other Bayesian nonparametric models for conditional densities, and excellent performance in applications of the related models, it seems important to develop reliable posterior simulation algorithms for the model with a varying number of components.

RJMCMC proposals based on moment matching (Richardson and Green (1997)) have been used in the literature to estimate mixtures of densities with a variable number of components. However, it is not clear how to implement this approach when the mixing weights depend on covariates. Carlin and Chib (1995) applied their auxiliary priors approach to mixtures of univariate normals with a small number of components where the cross-dimensional moves change the parameters of all the mixture components simultaneously. It is not clear how to implement this approach for the nonparametric conditional density model since the parameter vector is high-dimensional and constructing good auxiliary priors or proposals for the high-dimensional distributions with very complex shapes is a daunting task. Thus, I develop here an algorithm based on the conditional posterior proposals that changes parameters of only one mixture component in a cross-dimensional move.

In the model, the conditional posterior proposals can be evaluated up to normalizing constants that are difficult to compute precisely. Since the normalizing constants are required for computing acceptance probabilities, approximations to conditional posteriors have to be used in the implementation of the algorithm. Posteriors for parameters of one mixture component conditional on the number of components and the rest of the parameters are well behaved; there are no irregularities and label switching issues,

and quadratic approximations to the log of the conditional posterior are adequate. It is straightforward in principle to extend the algorithm from changing parameters of only one mixture component at a time to two or more. However, finding good approximations to conditional posteriors for parameters of two or more mixture components appears infeasible when covariates, especially multivariate ones, are present in the model. Thus, the restriction of changing parameters of only one mixture component in a cross-dimensional move is introduced not for theoretical convenience but rather for feasibility of algorithm implementation.

The resulting "approximately" optimal RJMCMC algorithm provides a feasible posterior simulation method for an attractive Bayesian nonparametric model for conditional densities for which the previous literature does not provide a feasible posterior simulator. The proposed methodology should also be useful for developing posterior simulators for other varying dimension models in which good proposals for the whole parameter vector are difficult to construct.

The rest of the paper is organized as follows. A general model formulation and the mixture of experts example are presented in Section 2. The auxiliary priors representation of the MCMC algorithm is given in Section 3. Section 4 provides the RJMCMC representation. Theoretical results on the algorithm optimality are given in Section 5. An application to the mixture of experts model and simulation results are presented in Section 6. Appendices contain proofs, implementation details, and auxiliary figures.

2 Model description

In this paper, we are concerned with the following class of models. Suppose that for an integer m, $\theta_{1m} = (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta^m = \Theta_1 \times \cdots \times \Theta_m \subset \mathbb{R}^{d_m}$, $\theta_{1\infty} = (\theta_1, \theta_2, \ldots)$, and $Y \in \mathbb{R}^{d_Y}$. Let the observables density satisfy the following restriction

$$p(Y|m, \theta_{1\infty}) = p(Y|m, \theta_{1m}), \tag{1}$$

so that m indexes a sequence of nested models. A prior is specified as follows

$$\Pi(\theta_{1m}|m)\Pi(m), \tag{2}$$

where $\Pi(\theta_{1m}|m)$ is a density with respect to a σ-finite dominating measure $\lambda^m = \lambda_1 \times \cdots \times \lambda_m$ on the Borel σ-field of $\mathbb{R}^{d_m}$. The support of $\Pi(m)$ can be equal to the set of positive integers. This class of models encompasses finite and countable mixtures of densities (McLachlan and Peel (2000), Fruhwirth-Schnatter (2006)) and mixtures of experts (Jacobs et al. (1991), Jordan and Jacobs (1994)) with a variable number of mixture components.

2.1 Application: Mixture of Experts

The main motivation and application for the MCMC algorithm is a nonparametric model for conditional densities from Norets and Pati (2017) based on mixtures of experts. Let $y_i \in \mathbb{R}$ denote a dependent variable and $x_i \in \mathbb{R}^{d_x}$ denote a vector of covariates for observation $i = 1, \ldots, n$. It is assumed that the observations are independently identically distributed. The marginal distribution of covariates is not of interest and, thus, it is not modeled. The conditional density of y_i given x_i is modeled by

$$p(y_i|x_i, \mu, \alpha, h, \nu, m) = \sum_{j=1}^{m} \gamma_j(x_i) \cdot \phi\left(y_i, x_i'\beta_j, (h_y \cdot \nu_{yj})^{-1}\right), \tag{3}$$

$$\gamma_j(x_i) = \frac{\alpha_j \exp\left\{-0.5 \sum_{l=1}^{d_x} h_{xl}\nu_{xjl}(x_{il} - \mu_{jl})^2\right\}}{\sum_{k=1}^{m} \alpha_k \exp\left\{-0.5 \sum_{l=1}^{d_x} h_{xl}\nu_{xkl}(x_{il} - \mu_{kl})^2\right\}},$$

where ϕ is a normal density and a prior on m and parameters completes the model setup.

To fit this model into the formulation (1), let $\theta_j = (\alpha_j, \beta_j, \mu_j, \nu_{yj}, \nu_{xj})$, consider (h_x, h_y) fixed (blocks for (h_x, h_y) can be easily added to the algorithm), and add $(x_1, \ldots, x_n)$ to the conditioning set of all distributions.

The parameters $(h_y, h_x, \nu_{yj}, \nu_{xj},\ j = 1, \ldots, m)$ are not identified in the likelihood of the model conditional on m and covariates, and proper priors must be used for the posterior to be well defined. This specification of scale parameters is justified as follows. The multiplicative part of the scale parameters that is common across all mixture components (h_y, h_x) is introduced so that there is a sufficient prior probability on very small values of scale parameters for all mixture components at the same time, which is required for achieving the optimal posterior contraction rates at smooth data generating densities, see Norets and Pati (2017) for more details. The parts of scale parameters that are specific to mixture components, $(\nu_{yj}, \nu_{xj},\ j = 1, \ldots, m)$, do not affect known asymptotic properties of the model; they are introduced to improve flexibility and small sample performance of the model when m is not large. Similar specifications of scale parameters for univariate mixture models are used in textbooks, see, for example, Geweke (2005).

3 Recursive Auxiliary Priors for Drawing m

In this subsection, let us consider only the algorithm's block for m. For many mixture models, MCMC algorithms for simulating θ_{1m} conditional on m are readily available (see for example, Fruhwirth-Schnatter (2006), Peng et al. (1996), Geweke and Keane (2007),

and Villani et al. (2009)). Section 6 describes the algorithm for simulating θ_{1m} for the model (3) with details relegated to Appendix B.

For $p(Y|m,\theta_{1m})$ and $\Pi(\theta_{1m}|m)\Pi(m)$ in (1), $\theta_{m+1\infty} = (\theta_{m+1}, \theta_{m+2}, \ldots)$, and an arbitrary distribution $\tilde{\Pi}(\theta_{m+1\infty}|m,\theta_{1m},Y)$, let us define a joint distribution

$$p(Y,\theta_{1\infty},m) = \tilde{\Pi}(\theta_{m+1\infty}|m,\theta_{1m},Y)\cdot p(Y|m,\theta_{1m})\cdot\Pi(\theta_{1m}|m)\Pi(m). \tag{4}$$

Importantly, the posterior $\Pi(m,\theta_{1m}|Y)$ implied by this joint distribution is not affected by $\tilde{\Pi}$. Thus, we can design $\tilde{\Pi}$ to facilitate posterior simulation from $\Pi(m,\theta_{1m}|Y)$ by retrospective sampling (Papaspiliopoulos and Roberts (2008)). For densities $\tilde{\pi}_m(\theta_{m+1}|\theta_{1m},Y)$ to be chosen below, let

$$\tilde{\Pi}(\theta_{m+1\infty}|m,\theta_{1m},Y) = \prod_{j=1}^{\infty}\tilde{\pi}_{m+j}(\theta_{m+1+j}|\theta_{1m+j},Y). \tag{5}$$

This recursive definition of $\tilde{\Pi}$ implies a tractable expression for Metropolis-Hastings acceptance probabilities. Specifically, let us consider the Metropolis-within-Gibbs block for $m|Y,\theta_{1\infty}$ with the proposal $Pr(m^* = m+1|m) = Pr(m^* = m-1|m) = 1/2$. For a proposal draw m^* the acceptance probability is equal to $\min\{1,\alpha(m^*,m)\}$, where

$$\begin{aligned}\alpha(m^*,m) = &\frac{p(Y|m^*,\theta_{1m^*})\Pi(\theta_{1m^*}|m^*)\Pi(m^*)}{p(Y|m,\theta_{1m})\Pi(\theta_{1m}|m)\Pi(m)}\\ &\cdot\left(\frac{1\{m^*=m+1\}}{\tilde{\pi}_m(\theta_{m+1}|\theta_{1m},Y)} + 1\{m^*=m-1\}\tilde{\pi}_{m-1}(\theta_m|\theta_{1m-1},Y)\right).\end{aligned} \tag{6}$$

When $m^* = m+1$, θ_{m+1} is simulated retrospectively from $\tilde{\pi}_m(\theta_{m+1}|\theta_{1m},Y)$.

3.1 Choice of Auxiliary Prior

As I show below in Section 5,

$$\tilde{\pi}_m(\theta_{m+1}|\theta_{1m},Y) = p(\theta_{m+1}|Y,m+1,\theta_{1m}) \propto p(Y|m+1,\theta_{1m+1})\Pi(\theta_{1m+1}|m+1) \tag{7}$$

is an optimal choice of $\tilde{\pi}_m$. A simplistic choice of the auxiliary prior with θ_j identically independently distributed for all $j \in \{1, \ldots, \infty\}$ leads to practically zero acceptance rates for m in the application considered in this paper. Thus, the use of an (approximately) optimal $\tilde{\pi}_m$ appears to be crucial for feasibility of the algorithm.

An approximation to $p(\theta_{m+1}|Y, m+1, \theta_{1m})$, with a known normalization constant, which is necessary here, can be given, for example, by a Gaussian distribution with the mean equal to the conditional posterior mode

$$\bar{\theta}_{m+1} = \arg\max_{\theta_{m+1}} p(Y|m+1, \theta_{1m+1})\Pi(\theta_{1m+1}|m+1)$$

and the variance calculated from the Hessian

$$V_{\theta_{m+1}}^{-1} = -\frac{\partial^2}{\partial\theta_{m+1}\partial\theta'_{m+1}} \log[p(Y|m+1, \theta_{1m+1})\Pi(\theta_{1m+1}|m+1)]\Big|_{\theta_{m+1}=\bar{\theta}_{m+1}}. \tag{8}$$

3.2 Previous Literature

Papaspiliopoulos and Roberts (2008) developed retrospective sampling ideas in the context of Dirichlet process mixtures. In those settings, the prior for all components of $\theta_{1\infty}$ does affect the posterior, and, thus, choosing the prior to improve the MCMC performance is not an option, in contrast to the settings considered here.

The birth-death process of Stephens (2000) is somewhat similar to the algorithm developed here and more generally to a RJMCMC that keeps the parameters of the smaller model unchanged when cross-dimensional moves are attempted. Stephens (2000) uses the same prior distribution for all θ_j's as a birth or proposal distribution, and, as I mention above, such proposals produce practically zero acceptance rates in the mixture of experts application.

Carlin and Chib (1995) introduced auxiliary prior distributions in the context of Bayesian model averaging and comparison for a finite number of parametric models. The algorithm proposed here can be thought of as an extension of ideas from Carlin and Chib (1995) to infinite dimensional settings, which also exploits the structure of the problem and more recently developed restrospective sampling ideas. Carlin and Chib apply their algorithm to finite mixture of normals models that can be set up as (1)-(2) with a bounded support for m. However, they treat θ_{1m} and $\theta_{1\tilde{m}}$ with $\tilde{m} \neq m$ as two non-overlapping vectors of parameters and for any given m, they introduce separate auxiliary prior distributions for all $\theta_{1\tilde{m}}$ with $\tilde{m} \neq m$; these auxiliary prior distributions are chosen to approximate $\Pi(\theta_{1\tilde{m}}|Y, \tilde{m})$, where approximations are obtained from a posterior simulator output for $\Pi(\theta_{1\tilde{m}}|Y, \tilde{m})$. In principle, their approach if combined with retrospective sampling could be used for estimation of model in (1)-(2) with an unbounded support for m. However, the posterior for mixture models has a large number of modes and obtaining an approximation for $\Pi(\theta_{1\tilde{m}}|Y, \tilde{m})$ is a challenging problem, especially for larger values of $d \cdot \tilde{m}$. Hence, the need to develop an alternative algorithm for models with large/infinite dimensions, which is addressed here.

4 Reversible Jump Representation

The algorithm for drawing m described in Section 3 can also be formulated as a RJMCMC algorithm (Green (1995)). Let us denote the state space for the RJMCMC by $\mathcal{X} = \cup_{m=1}^{\infty}\{m\} \times \Theta^m$. Let Q be a Markov transition on $\mathcal{X}$ and for $x, x' \in \mathcal{X}$, $f(x, x')$ be a density of $\Pi(dx|Y)Q(x, dx')$ with respect to a symmetric measure on $\mathcal{X} \times \mathcal{X}$ denoted by ϵ.

A RJMCMC update, also called Metropolis-Hastings-Green update because it generalizes the Metropolis-Hastings update to cross-dimensional settings, simulates a proposal $x' \sim Q(x, \cdot)$ that is accepted with probability

$$\min\left\{1, \frac{f(x', x)}{f(x, x')}\right\}.$$

The algorithm in Section 3 is obtained when $Q((\theta_{1m}, m), \cdot)$ draws $(m', \theta'_{1m'})$ as follows: $m' = m - 1$ and $\theta'_{1m'} = \theta_{1m-1}$ with probability 0.5, otherwise $m' = m + 1$ and $\theta'_{1m'} = (\theta_{1m}, \theta'_{m+1})$, where $\theta'_{m+1} \sim \tilde{\pi}_m(\cdot|\theta_{1m}, Y)$ and $\tilde{\pi}_m$ is defined in (5). The dominating measure ϵ is defined by

$$\epsilon(m, A, m', A') = \begin{cases} \int_A \lambda_{m+1}[z \in \Theta_{m+1} : (x, z) \in A']d\lambda^m(x) & \text{if } m' = m + 1 \\ \int_{A'} \lambda_{m'+1}[z \in \Theta_{m'+1} : (x, z) \in A]d\lambda^{m'}(x) & \text{if } m' = m - 1 \\ 0 & \text{if } m' \neq m \pm 1 \end{cases}$$

for Borel measurable $A \subset \Theta^m$ and $A' \subset \Theta^{m'}$. Thus, ϵ is essentially a product of a counting measure on $\{(m, m') : m, m' \in \mathbb{N}, m' = m \pm 1\}$ and a transition kernel $\lambda^{\max\{m,m'\}}$. The density

$$\begin{aligned} f(m, \theta_{1m}, m', \theta'_{1m'}) =& 0.5 \cdot \mathbb{1}\{m' = m + 1, \theta_{1m} = \theta'_{1m}\}\Pi(m, \theta_{1m}|Y)\tilde{\pi}_m(\theta'_{m'}|\theta_{1m}, Y) \\ &+ 0.5 \cdot \mathbb{1}\{m' = m - 1, \theta_{1m-1} = \theta'_{1m-1}\}\Pi(m, \theta_{1m}|Y) \end{aligned}$$

and the acceptance probability is given by (6).

5 Algorithm Optimality

In this section, I consider an optimal choice of $\tilde{\pi}_m$. Since at each MCMC iteration, m can only be changed by 1, one can expect that higher acceptance rates for m^* results in

a more efficient MCMC algorithm. Below, I make this intuition precise. First, I show in Theorem 1 how $\tilde{\pi}_m$ can be chosen to maximize expected acceptance rates for m^*. Then, in Theorem 2, I show that this choice minimizes asymptotic variance for MCMC sample average estimators for a class of functions that depend on (m, θ_{1m-1}).

Let us define the following conditional expected acceptance rates. The expected acceptance rate for $m^* = m + 1$ conditional on (m, θ_{1m}) is

$$\int \min\{1, \alpha(m^*, m)\}\tilde{\pi}_m(\theta_{m+1}|\theta_{1m}, Y)d\lambda_{m+1}(\theta_{m+1}), \tag{9}$$

and for $m^* = m - 1$ conditional on (m, θ_{1m-1}) is

$$\int \min\{1, \alpha(m^*, m)\}p(\theta_m|Y, m, \theta_{1m-1})d\lambda_m(\theta_m). \tag{10}$$

The use of the conditional posterior $p(\theta_m|Y, m, \theta_{1m-1})$ for taking the expectation in (10) is motivated by the fact that the MCMC algorithm converges to the stationary distribution.

Theorem 1. $\tilde{\pi}^{\star}_m(\theta_{m+1}|\theta_{1m}, Y) = p(\theta_{m+1}|Y, m+1, \theta_{1m})$ *maximizes the conditional expected acceptance rates in* (9) *and* (10).

The proof of the theorem is given in Appendix A.1. For $m^* = m + 1$, $\tilde{\pi}^{\star}_m$ tends to produce proposals of θ_{m+1} with high value of the numerator in $\alpha(m^*, m)$, and one would intuitively expect $\tilde{\pi}^{\star}_m$ to work well in this case (this in fact was the original motivation for trying the algorithm out even before its theoretical properties were obtained). The result for $m^* = m - 1$ seems more surprising. The mechanics of the proof are actually the same for $m^* = m + 1$ and $m^* = m - 1$, and they are about making $\alpha(m^*, m)$ as close to 1 as possible on average.

The results in Theorem 1 are of independent interest because for complex models with parameters of variable dimension, it could be hard to construct MCMC algorithms that

produce any accepted draws at all in a reasonable computing time. The theorem also has more formal implications for algorithm optimality.

A standard criterion for MCMC algorithm optimality is the asymptotic variance of sample averages. Let $\mathcal{L} = \{g : \mathcal{X} \to \mathbb{R},\ \int g d\pi = 0,\ \int g^2 d\pi < \infty\}$. For a transition kernel P with the stationary distribution π and $g \in \mathcal{L}$, I define the asymptotic MCMC variance as in Tierney (1998) by

$$v(g, P) = \lim_{n\to\infty} \text{var}_P \left(\sum_{k=1}^{n} g(X_i) \right) / n,$$

where $X_1, X_2, \ldots$ is a Markov chain with the initial distribution π and transition P. A transition kernel P can be called optimal if it minimizes $v(g, P)$ for all $g \in \mathcal{L}$.

Here, I obtain an optimality result under additional restrictions on P and $\mathcal{L}$. The MCMC algorithms I consider are indexed by $\tilde{\pi} = \{\tilde{\pi}_m,\ m = 1, 2, \ldots\}$ and have the following structure

$$P(\tilde{\pi}) = \left(\frac{P_{\theta_{1m-1}}}{2} + \frac{P_{m\theta_{1m}}}{2} \right) P_{\theta_m}, \tag{11}$$

where $P_{m\theta_{1m}}$ denotes the Metropolis-Hastings-Green transition kernel described in Section 4, P_{θ_m} denotes the Gibbs transition kernel for $\theta_m | m, \theta_{1m-1}, Y$, and $P_{\theta_{1m-1}}$ denotes a reversible transition kernel that updates $\theta_{1m-1} | m, \theta_m, Y$, for example, a random sequence scan Gibbs or Metropolis-within-Gibbs sampler for components of θ_{1m-1}. The dependence of $P_{m\theta_{1m}}$ on $\tilde{\pi}$ is not reflected in the notation for brevity.

Theorem 2. *For any $\tilde{\pi}$ and any $g \in \mathcal{L}$ that depends on (m, θ_{1m-1}) but not on θ_m,*

$$v(g, P(\tilde{\pi}^\star)) \leq v(g, P(\tilde{\pi})),$$

where $\tilde{\pi}^\star$ is defined in Theorem 1.

The theorem is proved in Appendix A.2. The proof uses the fact that increasing off diagonal transition probabilities of a reversible transition kernel with a fixed stationary distribution decreases the asymptotic variance for any $g \in \mathcal{L}$ (this result, due to Peskun (1973) and Tierney (1998), is formally presented in Appendix A.4).

The maximization of the expected acceptance rates for m^*, as in Theorem 1, actually reduces the off diagonal transition probabilities of $P(\tilde{\pi})$ when m stays the same (even though other off diagonal probabilities increase). There appears to be no obvious way to alter and/or combine $(P_{\theta_m}, P_{\theta_{1m-1}}, P_{m\theta_{1m}})$ that would lead to increased probabilities of all off diagonal transitions. Nevertheless, it is still possible to exploit the increased off diagonal transition probabilities of events that involve a change in m. The key observation here is that $P(\tilde{\pi})$ in (11) induces a Markov chain for (m, θ_{1m-1}) (with θ_m excluded). For this chain, all the off diagonal transition probabilities are maximized by $\tilde{\pi}^{\star}_m$ from Theorem 1. Moreover, the induced chain for (m, θ_{1m-1}) is reversible and, thus, the claim of Theorem 2 holds.

An ideal optimality result would hold for functions that can depend not only on (m, θ_{1m-1}) but on θ_m as well, and it would not depend on a particular combination and order of MCMC blocks in (11). Such a result appears to be difficult to obtain. Nevertheless, the demonstrated optimality results provide useful guidelines for constructing MCMC algorithms and deliver an explanation for the good practical performance of the approximate version of the algorithm implemented for the mixture of experts model. This is especially the case if we take into account that results on MCMC optimality are scarce and mostly restricted to discrete settings (see Chen (2013) for a survey).

6 Application to Mixture of Experts

In this section, I apply the algorithm to model (3). In what follows, I discuss prior specification, details of algorithm implementation, and tests for correctness of the implemented algorithm. The last two subsections evaluate the algorithm performance on simulated and real data.

Of course, it would be desirable to compare the algorithm with some benchmark methods. Unfortunately, the literature does not seem to provide other feasible methods for the mixture of a variable number of experts model. Specifically, the use of priors as proposals as in the retrospective sampling (Papaspiliopoulos and Roberts (2008)) or birth-death process (Stephens (2000)) does not deliver any accepted cross-dimensional moves. It is obvious that using a good approximation to the posterior of the whole parameter vector as a proposal would deliver a more efficient MCMC algorithm (Carlin and Chib (1995) do that for a simpler and smaller model). However, it is not at all clear how one could construct approximations to the complex shape posterior of the whole parameter vector for the model considered here.

6.1 Prior Specification

The prior is specified as follows. For $j = 1, \ldots, m$,

$$
\begin{aligned}
&\beta_j \overset{iid}{\sim} N(\underline{\beta}, \underline{H}_{\beta}^{-1}),\ \mu_j \overset{iid}{\sim} N(\underline{\mu}, \underline{H}_{\mu}^{-1}),\\
&\nu_{yj} \overset{iid}{\sim} G(\underline{A}_{\nu y}, \underline{B}_{\nu y}),\ \nu_{xlj} \overset{iid}{\sim} G(\underline{A}_{\nu xl}, \underline{B}_{\nu xl}),\ l = 1, \ldots, d_x,\\
&(h_y)^{1/2} \overset{iid}{\sim} G(\underline{A}_{hy}, \underline{B}_{hy}),\ (h_{xl})^{1/2} \overset{iid}{\sim} G(\underline{A}_{hxl}, \underline{B}_{hxl}),\ l = 1, \ldots, d_x,\\
&\alpha_j \overset{iid}{\sim} G(\underline{a}/m, 1),
\end{aligned}
\tag{12}
$$

$$\Pi\big(m=k\big) \propto e^{-\underline{A}_m \cdot k(\log k)^\tau},\ \tau \geq 0, \underline{A}_m > 0,$$

where $G(A,B)$ stands for a Gamma distribution with shape A and rate B. Some of these prior functional form assumptions are made so that asymptotic results in Norets and Pati (2017) apply. Specifically, a gamma prior for (h_{xl}, h_y) would not put sufficient mass in the tails for the asymptotic results, and hence, the square of a gamma prior is used. The division by m in $G(\underline{a}/m, 1)$ prior for α_j is also required. The tail of the prior for m also has to be essentially of the assumed form.

6.2 MCMC Algorithm

This subsection presents a general discussion of the MCMC algorithm for the conditional density model. A detailed description of the algorithm is provided in Appendix B. As is common in the literature on MCMC for finite mixture models, I introduce latent mixture allocation variables $(s_1, \ldots, s_n)$ (Diebolt and Robert (1994)) to facilitate the simulation from blocks of the Metropolis-within-Gibbs for given m: $y_i|x_i, s_i, m, \theta_{1m}, h_x, h_y \sim N(x_i'\beta_{s_i}, (h_y\nu_{ys_i})^{-1})$ and $\Pi(s_i = j) = \gamma_j(x_i)$, where $\gamma_j(x_i)$ is defined below (3). Then, Gibbs sampler blocks for (s_i, β_j, ν_{yj}) have standard distributions and are simulated directly. The rest of the parameters are simulated by the Metropolis-within-Gibbs algorithm. The Metropolis-within-Gibbs block for m described in Section 3 does not condition on the latent mixture allocation variables. Therefore, the block for the mixture allocation variables needs to be placed right after the block for m.

When the algorithm attempts to jump from m to $m-1$ the m^{th} component that would be deleted in case of a successful jump is selected randomly from all the current

mixture components. This is essentially a random label switching that does not affect the stationary distribution of the chain and helps the chain not to get stuck when the m^{th} component is important for explaining the data.

6.3 Tests for Correctness of the Algorithm Design and Implementation

The simulator is implemented in Matlab. To check that the simulator is designed and implemented correctly, I conduct the joint distribution tests proposed in Geweke (2004). The tests are based on a comparison of the prior distribution and the output of a successive conditional simulator that simulates both data and parameters as follows. On each iteration, the parameters are updated by the the posterior simulator given the current data draw and then the new data draw is obtained from the likelihood conditional on the current parameter draw. The resulting algorithm is a hybrid MCMC algorithm (or just a Gibbs sampler if direct simulation rather than MCMC is used for posterior simulator) for exploring the joint prior distribution of parameters and data. If the data and posterior simulators are correct then draws from the successive conditional simulator should be consistent with the prior distribution, which can be checked by standard mean equality tests. Table 1 presents the t-statistics from the mean equality tests for the parameters and their squares. As can be seen from the table the hypotheses of mean equality are not rejected at conventional significance levels for all but one parameter, which indicates that there are no errors in simulator design and implementation (the tests did help to find and correct a few errors at the development stage).

Table 1: Joint Distribution Tests

Parameter	t-stat	Parameter	t-stat	Parameter	t-stat
β_{11}	-2.19	μ_{11}	-1.17	h_{x1}	-0.17
β_{11}^2	1.99	μ_{11}^2	0.57	h_{x1}^2	-0.28
β_{12}	-0.04	μ_{12}	-0.56	h_{x2}	0.90
β_{12}^2	0.03	μ_{12}^2	0.85	h_{x2}^2	0.85
β_{13}	-1.60	μ_{13}	-1.73	h_{x3}	-0.37
β_{13}^2	1.72	μ_{13}^2	1.20	h_{x3}^2	-0.38
β_{14}	1.73	μ_{14}	-0.02	h_{x4}	-0.95
β_{14}^2	1.75	μ_{14}^2	-0.03	h_{x4}^2	-1.41
β_{15}	-0.05	ν_{x11}	-1.01	m	-0.76
β_{15}^2	-0.11	ν_{x11}^2	-1.39	m^2	-0.81
h_y	-0.95	ν_{x12}	0.37	$1\{m = 1\}$	0.67
h_y^2	-0.80	ν_{x12}^2	0.65	$1\{m = 2)$	-0.53
ν_{y1}	-0.30	ν_{x13}	-0.14	$1\{m = 3)$	-0.55
ν_{y1}^2	-0.07	ν_{x13}^2	-0.08	$1\{m = 4)$	-0.36
$\sum_{j=1}^m \alpha_j$	0.49	ν_{x14}	-1.64	$1\{m = 5)$	-0.71
$(\sum_{j=1}^m \alpha_j)^2$	0.28	ν_{x14}^2	-1.64	$1\{m = 6)$	-0.52

Figure 5 in Appendix C compares the exact prior probability mass function for m and the probability mass function obtained from the successive conditional simulator. Figure 6 presents a trace plot of m.

6.4 Experiments on Simulated Data

This subsection describes the performance of the MCMC algorithm on simulated data with different dimension of covariates. For a given d_x, the covariates are generated from a uniform distribution, $x_i = (x_{i1}, \ldots, x_{id_x})' \sim U[0,1]^{d_x}$. The conditional distribution of the outcome is a mixture of two normal distributions with nonlinear means, variances, and mixing probabilities.

$$y_i|x_i \sim e^{-\sqrt{x_{i1}}}\phi\left(\cdot; \Phi\left(\psi_1(x_i)\right), 0.5\psi_1(x_i)\right) + (1 - e^{-\sqrt{x_{i1}}})\phi\left(\cdot; \Phi\left(-\psi_1(x_i)\right), 0.1\psi_2(x_i)\right), \quad (13)$$

where $\psi_1(x_i) = \sum_{k=1}^{d_x} x_{ik}/k^4$, $\psi_2(x_i) = \sum_{k=1}^{d_x} x_{ik}^{2+k}/d_x$, $\phi(\cdot; \mu, \sigma)$ is a normal density with mean μ and standard deviation σ, and Φ is standard normal cumulative distribution function. The number of observations in each simulated dataset is 2000. The simulated data for $d_x = 1$ are shown in Figure 7 in Appendix C.

The average acceptance rates for m calculated from 100,000 MCMC iterations are presented in Table 2. The corresponding MCMC trace plots are shown in Figure 8.

Table 2: Acceptance rates

d_x	$\dim(\theta_m)$	Acceptance Rate, %
1	6	0.38
4	15	0.19
7	24	0.25
10	33	0.04

As can be seen from the table, the acceptance rates tend to decline as the dimension of θ_m increases. Nevertheless, the algorithm seems to provide reasonable descriptions of the posterior distributions for m. A trace plot of the log likelihood evaluated at MCMC draws

of the parameters is shown in Figure 9.

6.5 Engel Curve Estimation

An Engel curve is a relationship between the fraction of income spent on a particular good (or a category of goods) and the total income of a consumer (Lewbel (2008)). In empirical economics, Engel curves are often assumed to be linear or quadratic up to an additive error term. In this section, I estimate the density of the fraction of food expenditure conditional on total income using data from Battistin and Nadai (2015). The data consists of 2311 observations on individuals. Possible measurement errors and instrumental variables specifications, which are considered in the literature on Engel curve estimation, are ignored here. In this context, I evaluate performance of the MCMC algorithm and also compare out-of-sample predictive performance of the Bayesian mixture of experts, linear and quadratic normal regressions, and a cross-validated kernel estimator.

The prior hyperparameters in (12) in this estimation exercise are selected in an empirical Bayes fashion as follows. First, all the variables in the dataset are standardized to have zero mean and unit variance. The prior mean for β_j is set to the ordinary least squares (OLS) estimate and the prior variance to the variance of the OLS estimator under homoskedasticity multiplied by 10^3. The prior mean and variance for μ_j are set to $(0, 1)$ to match the sample mean and variance of the standardized covariates. To limit the variation of component specific scale parameters and help mitigate their lack of likelihood identification the hyperparameters of the Gamma priors for (ν_{yj}, ν_{xjk}) in (12) are chosen so that they have mean 1 and variance 0.1 . The hyperparameters of the Gamma prior for $h_y^{1/2}$ are chosen so that it has the variance equal to 10 and the mean equal to the in-

verse of the standard error of regression in the OLS. The hyperparameters of the Gamma prior for $h_{xk}^{1/2}$ are chosen so that it has the variance equal to 10 and the mean equal to 1 (or, more generally, the inverse of the sample standard deviation of the corresponding covariate). Finally, $\underline{a} = 8$, $\underline{A}_m = 1$, and $\tau = 0$. In the out of sample prediction exercises, only the estimation part of the data (but not the prediction part) is used to compute the hyperparameters as described above. The estimation results are not very sensitive to moderate variations in prior hyperparameters around the values suggested by the empirical Bayes procedure; however, moving prior means away from corresponding data analogs and reducing the prior variances can result in estimation results that are dominated by such a strong prior.

Figure 1 shows the raw data and the estimated posterior means of the conditional densities. Figure 2 shows the prior and the estimated posterior probability mass functions for m. Figures 3 and 4 show MCMC trace plots of m and the log likelihood evaluated at the parameter draws. The trace plot suggests that the algorithm converges. The average acceptance rate for m in this MCMC run is 2.9% and the effective sample size is 330. A desktop with a 3.5GHz processor and 32GB RAM takes about 5.4 seconds to perform 100 MCMC iterations.

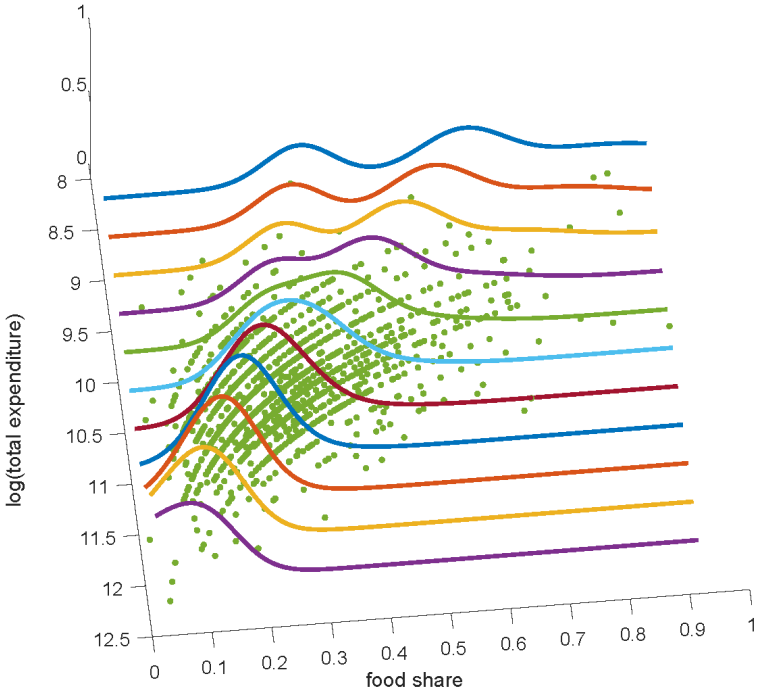

Figure 1: Data and estimated densities

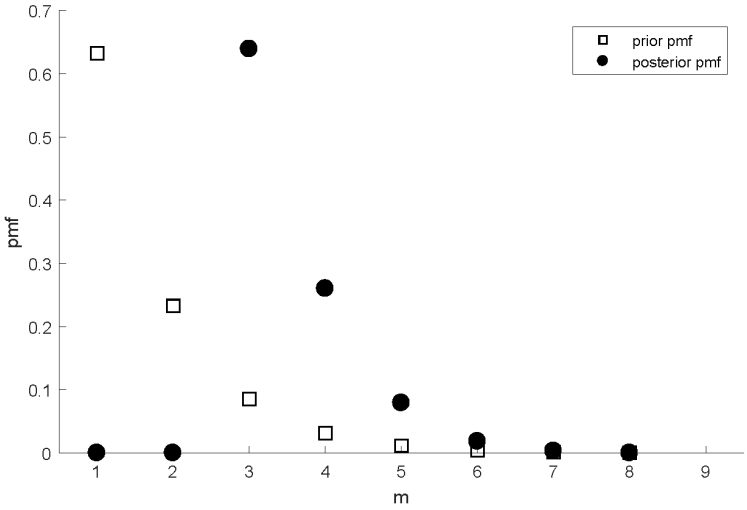

Figure 2: Prior and posterior for m

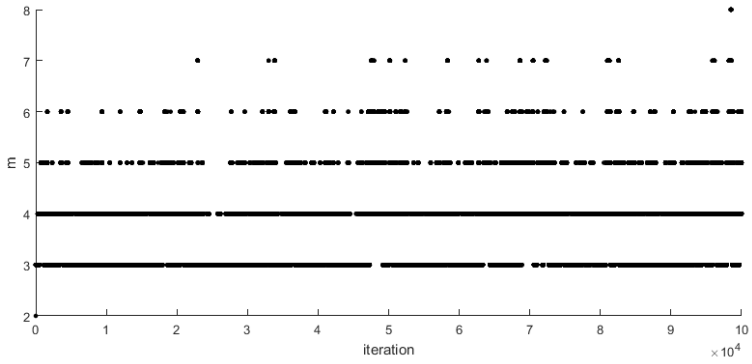

Figure 3: MCMC trace plot for m

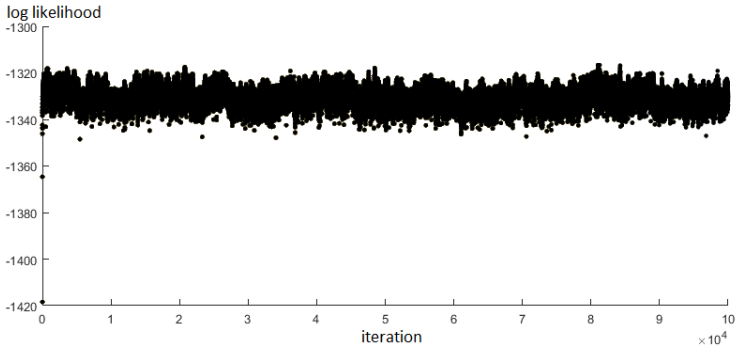

Figure 4: Trace plot of log likelihood

To compare the quality of out-of-sample predictions of the Bayesian mixture of experts and classical parametric and kernel estimators, I conduct a Monte Carlo exercise. On each iteration of the exercise, all the models are estimated on a randomly selected half of the observations and the predictive densities implied by the estimated models are evaluated on the observations not used in estimation. The log predictive densities for each model averaged over 30 iterations are reported in Table 3.

Table 3: Predictive performance

Method	Average of log predictive density
NP Bayes	-1375
Kernel	-1398
Linear	-1444
Quadratic	-1444

The length of MCMC runs for estimation of mixtures of experts in the Monte Carlo experiment is 5000. The acceptance rates for these MCMC runs were between 2% and 7%. The kernel conditional density estimation with cross-validated bandwidth selection (Hall et al. (2004)) was performed by R package *np* (Hayfield and Racine (2008)). The nonparametric Bayesian model (NP Bayes in the table) outperforms the kernel estimator, which in turn outperforms the linear and quadratic normal regressions. In line with the asymptotic results of Norets and Pati (2017), this comparison of predictive performance suggests that a mixture of variable number of experts is an attractive model for nonparametric estimation of conditional densities. The MCMC algorithm proposed in this paper makes Bayesian estimation of this model practical.

7 Conclusion

The main objective of this research project is to develop a feasible posterior simulator for a theoretically attractive Bayesian nonparametric model for conditional densities based on mixtures of variable number of experts (Norets and Pati (2017)). After extensive experimentation with different proposals and methods, I have not managed to come up

with an alternative MCMC algorithm for this model with non-zero acceptance rates for cross-dimensional moves. The success of the method in simulation experiments stimulated my interest in its theoretical properties. The painstaking analysis presented in the paper indeed shows that the method is an approximation to an RJMCMC with a proposal distribution that is optimal under the restriction of keeping the parameter values in smaller submodel unchanged when the cross-dimensional moves are attempted. It is worth emphasizing that the restrictions under which the proposals are optimal and the use of approximations to the optimal proposals are dictated by the feasibility of the method implementation for the mixture of experts model. The proposed methodology should also be useful for developing posterior simulators for other varying dimension models with a nesting structure in which good proposals for the whole parameter vector are difficult to construct.

A Appendix. Proofs and Auxiliary Results

A.1 Proof of Theorem 1

First, observe that the problem of finding $\tilde{\pi}_m$ that maximizes the conditional acceptance rates can be reformulated as follows

$$\max_{g} \int \min\left\{1, \frac{c \cdot f(z)}{g(z)}\right\} g(z) d\lambda(z), \tag{14}$$

where $c \geq 0$, and g is restricted to be a density with respect to measure λ. For $m^* = m+1$, $f(\cdot)$ denotes $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ as a function of θ_{m+1}, $g(\cdot)$ denotes $\tilde{\pi}_m(\theta_{m+1}|\theta_{1m}, Y)$ as a function of θ_{m+1}, λ denotes λ_{m+1}, and $c = p(Y, m+1, \theta_{1m})/p(Y, m, \theta_{1m})$. For $m^* = m-1$,

$f(\cdot)$ denotes $p(\theta_m|Y, m, \theta_{1m-1})$ as a function of θ_m, $g(\cdot)$ denotes $\tilde{\pi}_{m-1}(\theta_m|\theta_{1m-1}, Y)$ as a function of θ_m, λ denotes λ_m, and $c = p(Y, m, \theta_{1m-1})/p(Y, m-1, \theta_{1m-1})$.

Since $\min\{g(z), cf(z)\} = (g(z) + cf(z))/2 - |g(z) - cf(z)|/2$ and $\int g(z)d\lambda(z) = 1$, the problem in (14) is equivalent to

$$\min_g \int |g(z) - cf(z)|d\lambda(z). \tag{15}$$

For $c > 1$, any g^* with $g^*(z) \leq cf(z)$ for (λ almost surely) all z solves (15). To see this formally, consider g such that $g(z) > cf(z)$ on Z^+, $\lambda(Z^+) > 0$, and $g(z) \leq cf(z)$ on $Z^- = Z \setminus Z^+$, where Z is the domain for f and g. Let us define $g'(z) = cf(z)$ on Z^+ and $g'(z) = g(z) + r \cdot (cf(z) - g(z))$ on Z^-, where

$$r = \int_{Z^+} (g(z) - cf(z))d\lambda(z) \Big/ \int_{Z^-} (cf(z) - g(z))d\lambda(z).$$

Note that g' is a density and $r \in (0, 1)$ because $\int_{Z^+}(g(z) - cf(z))d\lambda(z) - \int_{Z^-}(cf(z) - g(z))d\lambda(z) = 1 - c < 0$. Also, $|g(z) - cf(z)| \geq |g'(z) - cf(z)|$ with a strict inequality on a set of λ positive measure. Thus, g is dominated by $g' \leq cf$. For any $g^* \leq cf$, $\int |g^*(z) - cf(z)|d\lambda(z) = c - 1$.

For $c < 1$, an analogous argument with $g'(z) = cf(z)$ on Z^-, $g'(z) = cf(z) + r \cdot (g(z) - cf(z))$ on Z^+, and $r = \int(g(z) - cf(z))d\lambda(z) / \int_{Z^+}(g(z) - cf(z))d\lambda(z)$, shows that any g^* with $g^*(z) \geq cf(z)$ for (λ almost surely) all z solves (15). For $c = 1$, $g^* = f$ is obviously the solution. Thus, $g^* = f$ solves (14), and it is a unique solution that does not depend on c.

A.2 Proof of Theorem 2

Let us define a transition kernel Q on $\cup_{m=1}^{\infty}\{m\}\times\Theta^{m-1}$ by

$$Q((m,\theta_{1m-1}),m'\times A'_{m'-1}) = P((m,\theta_{1m-1}),m'\times A'_{m'-1}\times\Theta_{m'}), \tag{16}$$

where $A'_{m'-1}$ is a measurable subset of $\Theta^{m'-1}$ and P is defined in (11) with dependence on $\tilde{\pi}$ not reflected in the notation for brevity ($Q(\tilde{\pi})$ is used below whenever explicit dependence on $\tilde{\pi}$ is convenient). Note that P does not depend on θ_m as it starts from redrawing $\theta_m|m,\theta_{1m-1},Y$, and Q is indeed a well defined transition kernel on $\cup_{m=1}^{\infty}\{m\}\times\Theta^{m-1}$. Note also that Q can be expressed as $Q((m,\theta_{1m-1}),m'\times A'_{m'-1}) = P_{\theta_m}\cdot P((m,\theta_{1m-1}),m'\times A'_{m'-1}\times\Theta_{m'})$ as the multiplication by P_{θ_m} from the left does not affect the transition for (m,θ_{1m-1}). $P_{\theta_m}\cdot P$ is a palindromic combination of reversible kernels and, thus, reversible (see Section A.3). Therefore, Lemma 1 applies and Q is a reversible transition kernel.

Next, let us show that $Q(\tilde{\pi}^*)\succeq Q(\tilde{\pi})$, where "domination off diagonal" relation, "$\succeq$", is defined in Section A.4. Since $P_{\theta_{1m-1}}P_{\theta_m}$ does not depend on $\tilde{\pi}$, we can consider only $Q_1 = P_{m\theta_{1m}}P_{\theta_m}$, and it suffices to show that for any measurable sets $A_j\subset\Theta^j$, $j\in\{m-2,m\}$, $Q_1((m,\theta_{1m-1}),\{j+1\}\times A_j)$ is maximized when $\tilde{\pi}=\tilde{\pi}^\star$ (for any measurable set $A_{m-1}\subset\Theta^{m-1}$, $Q_1((m,\theta_{1m-1}),\{m\}\times A_{m-1}\setminus\{\theta_{1m-1}\})=0$ and $j=m-1$ does not need to be considered). For $j=m$, $Q_1((m,\theta_{1m-1}),\{m+1\}\times A_m)$ is equal

$$\frac{1}{2}\int_{A_m(\theta_{1m-1})}\Pr(m+1\text{ is accepted}|m,\theta_{1m})\Pi(d\theta_m|m,\theta_{1m-1},Y),$$

where $A_m(\theta_{1m-1}) = \{\theta_m\in\Theta_m:(\theta_{1m-1},\theta_m)\in A_m\}$ and $\Pr(m+1\text{ is accepted}|m,\theta_{1m})$ is given by (9). By Theorem 1, (9) is maximized at $\pi^\star$, and, thus, $Q_1((m,\theta_{1m-1}),\{m+1\}\times A_m)$ is maximized at $\pi^\star$ as well. For $j=m-2$, $Q_1((m,\theta_{1m-1}),\{m-1\}\times A_{m-2})$ is equal

$$\mathbb{1}_{A_{m-2}}(\theta_{1m-2})\frac{1}{2}\int_{A_m(\theta_{1m-1})}\Pr(m-1\text{ is accepted}|m,\theta_{1m})\Pi(d\theta_m|m,\theta_{1m-1},Y),$$

where the integral is equal to (10). By Theorem 1, (10) is maximized at $\pi^\star$, and, thus, $Q_1((m, \theta_{1m-1}), \{m-1\} \times A_{m-2})$ is maximized at $\pi^\star$ as well.

Since $Q(\tilde{\pi})$ is reversible for any $\tilde{\pi}$ and $Q(\tilde{\pi}^*) \succeq Q(\tilde{\pi})$, Peskun-Tierney theorem from Section A.4 delivers $v(g, Q(\tilde{\pi}^\star)) \leq v(g, Q(\tilde{\pi}))$ for any $g \in \mathcal{L}$ that depends on (m, θ_{1m-1}) but not θ_m. The claim of the theorem follows since $v(g, Q(\tilde{\pi})) = v(g, P(\tilde{\pi}))$ for such g.

A.3 Standard Facts About Reversibility

Transition kernel P is reversible with respect to π if $\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$. The following elementary MCMC updates are reversible: a Metropolis-Hastings update on a part of the parameter vector, a Gibbs sampler block, and a Metropolis-Hastings-Green update. A mixture of reversible transition kernels is reversible. A palindromic combination of reversible transition kernels is reversible, for example, $P_1P_2P_1$ is reversible when P_1 and P_2 are reversible. Combinations of reversible transition kernels such as a Gibbs sampler with a fixed order of blocks are not reversible in general. A random sequence scan Gibbs or Metropolis-within-Gibbs sampler is reversible. A detailed presentation of these facts can be found in Geyer (2005).

A.4 Peskun-Tierney Theorem

The earliest fundamental result in the literature on optimal MCMC is due to Peskun (1973), who shows that increasing the off-diagonal elements in a reversible Markov transition matrix with a fixed stationary distribution reduces $v(g, P)$. Tierney (1998) extends this result to Markov chains on general state space. For transition kernels P_1 and P_2 with invariant distribution π, P_1 is said to dominate P_2 off the diagonal, $P_1 \succeq P_2$, if

$P_1(x, A \setminus \{x\}) \geq P_2(x, A \setminus \{x\})$ for any measurable A and π almost all x. Theorem 4 in Tierney (1998): When P_1 and P_2 are reversible, $P_1 \succeq P_2$ implies $v(g, P_1) \leq v(g, P_2)$.

A.5 Auxiliary Results

Lemma 1. *If a transition kernel P on $\cup_{m=1}^{\infty}\{m\} \times \Theta^m$ is reversible with respect to some π and P does not depend on θ_m, then $Q((m, \theta_{1m-1}), m' \times A'_{m'-1}) = P((m, \theta_{1m-1}), m' \times A'_{m'-1} \times \Theta_m)$ is a reversible transition kernel on $\cup_{m=1}^{\infty}\{m\} \times \Theta^{m-1}$ with respect to $\pi(m, \theta_{1m-1}) = \int \pi(m, \theta_{1m-1}, d\theta_m)$.*

Proof. The reversibility of P is equivalent to

$$\int_{\{m\}\times A} P((m, \theta_{1m-1}), m' \times A')d\pi(m, \theta_{1m}) = \int_{\{m'\}\times A'} P((m', \theta'_{1m'-1}), m \times A)d\pi(m', \theta'_{1m'}).$$

Setting $A = A_{m-1} \times \Theta_m$ and $A' = A'_{m'-1} \times \Theta_{m'}$ immediately implies the reversibility of Q. □

B Appendix. MCMC Algorithm

In the following Metropolis-within-Gibbs blocks, the rest of the parameters and data in the conditioning sets are denoted by '...' as in $\beta_j|\ldots$.

- $Pr(s_i = j|\ldots) \propto \gamma_j(x_i) \cdot \phi\left(y_i, x_i'\beta_j, (h_y \cdot \nu_{yj})^{-1}\right)$, $i = 1, \ldots, n$.

- $\beta_j|\ldots \sim N(\bar{\beta}_j, \bar{H}_{\beta j}^{-1})$, $j = 1, \ldots, m$,

  where $\bar{H}_{\beta j} = \underline{H}_\beta + h_y\nu_{yj}\sum_{i:\, s_i=j} x_i x_i'$ and $\bar{\beta}_j = \bar{H}_{\beta j}^{-1}(\underline{H}_\beta\underline{\beta} + h_y\nu_{yj}\sum_{i:\, s_i=j} x_i y_i)$.

- $\nu_{yj}|\ldots \sim G(\bar{A}_{\nu y}, \bar{B}_{\nu y})$, $j = 1, \ldots, m$,

  where $\bar{A}_{\nu y} = \underline{A}_{\nu y} + 0.5\sum_i 1\{s_i = j\}$ and $\bar{B}_{\nu y} = \underline{B}_{\nu y} + 0.5h_y\sum_{i:\, s_i=j}(y_i - x_i'\beta_j)^2$.

- $h_y|\ldots$ is simulated by the Metropolis-Hastings algorithm with proposal $G(\bar{A}_{hy}, \bar{B}_{hy})$, where $\bar{A}_{hy} = 0.5\underline{A}_{\nu y} + 0.5\sum_i 1\{s_i = j\}$ and $\bar{B}_{hy} = 0.5\sum_i (y_i - x_i'\beta_{s_i})^2 \nu_{ys_i}$.

- $\tilde{\alpha} = (\alpha_1/\sum_{j=1}^m \alpha_j, \ldots, \alpha_{m-1}/\sum_{j=1}^m \alpha_j)|\ldots$ is simulated by the Metropolis-Hastings algorithm with proposal $N(\bar{\tilde{\alpha}}, \bar{H}_{\tilde{\alpha}}^{-1})$. Two alternative procedures with different proposal parameters are implemented. The first one is a random walk with precision that is a function of the current parameter value: $\bar{\tilde{\alpha}} = \tilde{\alpha}$ and $\bar{H}_{\tilde{\alpha}} = -\frac{\partial^2}{\partial\tilde{\alpha}\partial\tilde{\alpha}'} \log[p(Y|X, s, m, \theta_{1m}, h_y, h_x)\Pi(\tilde{\alpha}|m)]$. The second one is an independence chain with $\bar{\tilde{\alpha}}$ equal to the conditional posterior mode (obtained by the Newton method) and the precision equal to the negative of the Hessian as in the first alternative but evaluated at the mode. Note that the acceptance probabilities have different expressions for the two alternative procedures. The random walk procedure is much faster and it does not lead to any noticeable increase in serial correlation of MCMC draws in simulations.

- $\sum_{j=1}^m \alpha_j|\ldots \sim G(\underline{a}, 1)$. This quantity is independent of data since the likelihood conditional on m depends only on $\tilde{\alpha}$ defined above. It is determined by the prior distribution only. It is used with $\tilde{\alpha}$ for computing α, which is required for the block for m.

- Blocks $h_x|\ldots$, $\nu_{xjk}|\ldots$, $\mu_j|\ldots$ are handled analogously to $\tilde{\alpha}|\ldots$.

- Random label switching: simulate j_1 from a uniform distribution on $\{1, \ldots, m\}$ and set $\theta_{temp} = \theta_m$, $\theta_m = \theta_{j_1}$, and $\theta_{j_1} = \theta_{temp}$.

- Block for $m|\ldots$ is implemented following the general description in Section 3 with the following simplifications. To increase the speed of computation and

avoid calculations of cross-derivatives, μ_m, β_m, ν_{ym}, ν_{xmk}, and α_m are assumed independent in the proposal. A Newton method is used to find the conditional posterior mode $\bar{\theta}_m = (\bar{\mu}_m, \bar{\beta}_m, \bar{\nu}_{ym}, \bar{\nu}_{xmk}, \bar{\alpha}_m)$. The μ_m's part of the proposal is a multivariate normal with the mean set to $\bar{\mu}_m$ and the variance set to the negative inverse of the Hessian with respect to μ_m evaluated at $\bar{\theta}_m$, $-\left[\frac{\partial^2}{\partial\mu_m\partial\mu_m'}\log[p(Y|m,\theta_{1m})\Pi(\theta_{1m}|m)]\Big|_{\theta_m=\bar{\theta}_m}\right]^{-1}$; the β_m's part of the proposal is constructed in the same fashion. The ν_{ym}'s part of the proposal is a Gamma distribution with the shape and rate parameters selected so that the mean and the variance of the Gamma distribution match correspondingly the conditional posterior mode, $\bar{\nu}_{ym}$, and $-\left[\frac{\partial^2}{\partial\nu_{ym}\partial\nu_{ym}'}\log[p(Y|m,\theta_{1m})\Pi(\theta_{1m}|m)]\Big|_{\theta_m=\bar{\theta}_m}\right]^{-1}$. The parts of the proposal for $\bar{\alpha}_m$ and $\bar{\nu}_{xmk}$, $k = 1,\ldots,d_x$ are constructed in the same fashion. Setting the mode rather than the mean of Gamma proposals to the conditional posterior mode or using a truncated normal instead of a Gamma lead to a slightly worse algorithm performance.

As mentioned in Section 6.2, mixture allocation variables s_i, $i = 1,\ldots,n$ are marginalized out and not present in the conditioning set of block $m|\ldots$.

C Appendix. Additional Figures

C.1 Figures for Joint Distribution tests

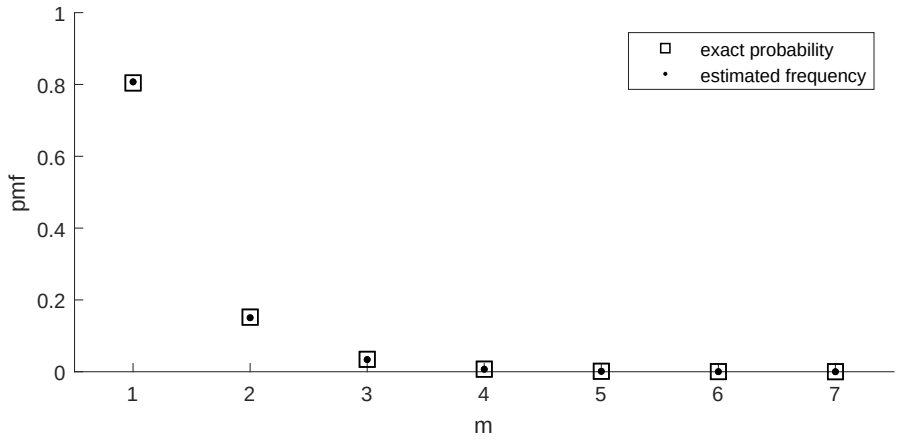

Figure 5: Probability Mass Function for m

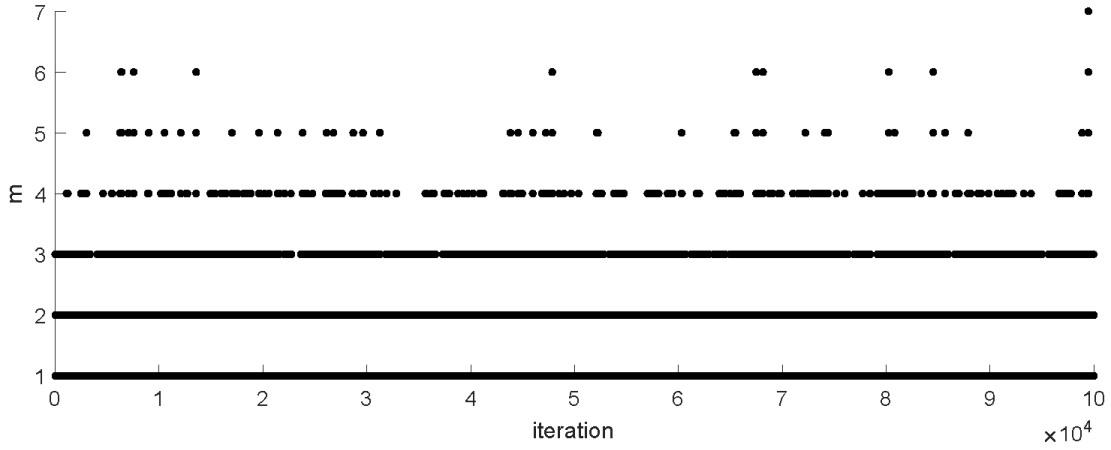

Figure 6: Trace Plot for m

C.2 Figures for Simulated Data Experiments

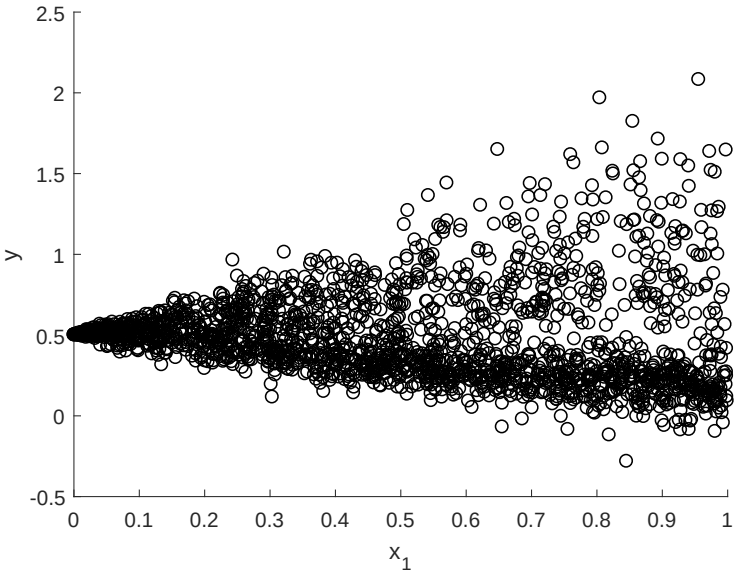

Figure 7: Simulated data, $d_x = 1$

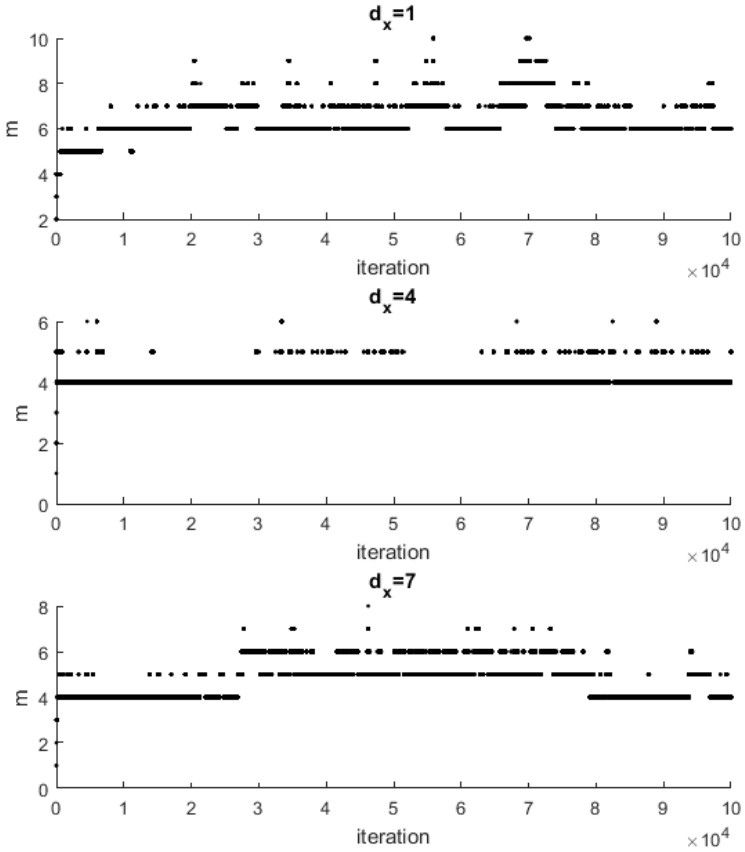

Figure 8: Trace plots for m, simulated data

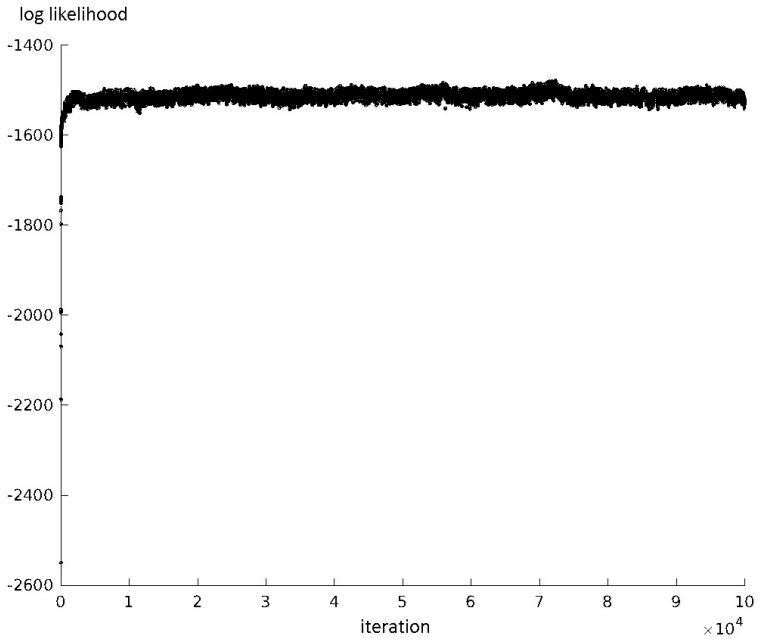

Figure 9: Trace plots for log likelihood, simulated data, $d_x = 1$

References

Battistin, E. and M. D. Nadai (2015): "Identification and Estimation of Engel Curves with Endogenous and Unobserved Expenditures," *Journal of Applied Econometrics*, 30, 487–508.

Brooks, S. P., P. Giudici, and G. O. Roberts (2003): "Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 3–39.

Carlin, B. P. and S. Chib (1995): "Bayesian model choice via Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society. Series B (Methodological)*, 473–484.

Chen, T.-L. (2013): "Optimal Markov chain Monte Carlo sampling," *Wiley Interdisciplinary Reviews: Computational Statistics*, 5, 341–348.

Diebolt, J. and C. P. Robert (1994): "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 363–375.

Fruhwirth-Schnatter, S. (2006): *Finite Mixture and Markov Switching Models (Springer Series in Statistics)*, Springer, 1 ed.

Geweke, J. (2004): "Getting it Right: Joint Distribution Tests of Posterior Simulators," *Journal of the American Statistical Association*, 99, 799–804.

——— (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience.

Geweke, J. and M. Keane (2007): "Smoothly mixing regressions," *Journal of Econometrics*, 138, 252–290.

Geyer, J. C. (2005): "Markov Chain Monte Carlo Lecture Notes," Unpublished.

Green, P. J. (1995): "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Hall, P., J. Racine, and Q. Li (2004): "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hastie, D. I. and P. J. Green (2012): "Model choice using reversible jump Markov chain Monte Carlo," *Statistica Neerlandica*, 66, 309–338.

Hayfield, T. and J. S. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991): "Adaptive mixtures of local experts," *Neural Computation*, 3, 79–87.

Jordan, M. and L. Xu (1995): "Convergence results for the EM approach to mixtures of experts architectures," *Neural Networks*, 8, 1409 – 1431.

Jordan, M. I. and R. A. Jacobs (1994): "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, 6, 181–214.

Lewbel, A. (2008): "Engel curve," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf and L. E. Blume, Basingstoke: Palgrave Macmillan.

McLachlan, G. and D. Peel (2000): *Finite Mixture Models*, John Wiley & Sons, Inc.

Norets, A. and D. Pati (2017): "Adaptive Bayesian Estimation of Conditional Densities," *Econometric Theory*, 33, 9801012.

Papaspiliopoulos, O. and G. O. Roberts (2008): "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95, 169–186.

Peng, F., R. A. Jacobs, and M. A. Tanner (1996): "Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With an Application to Speech Recognition," *Journal of the American Statistical Association*, 91, 953–960.

Peskun, P. H. (1973): "Optimum Monte-Carlo Sampling Using Markov Chains," *Biometrika*, 60, 607–612.

Richardson, S. and P. J. Green (1997): "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 731–792.

Stephens, M. (2000): "Bayesian analysis of mixture models with an unknown number of componentsan alternative to reversible jump methods," *The Annals of Statistics*, 28, 40–74.

Tierney, L. (1998): "A note on Metropolis-Hastings kernels for general state spaces," *The Annals of Applied Probability*, 8, 1–9.

Villani, M., R. Kohn, and P. Giordani (2009): "Regression density estimation using smooth adaptive Gaussian mixtures," *Journal of Econometrics*, 153, 155 – 173.

Wood, S., W. Jiang, and M. Tanner (2002): "Bayesian mixture of splines for spatially adaptive nonparametric regression," *Biometrika*, 89, 513–528.