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

Andriy Norets

Brown University

Current version: December 2019

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. The mixture of experts model outperforms standard parametric and nonparametric alternatives in out of sample performance comparisons in an application to Engel curve estimation. The proposed MCMC algorithm makes estimation of this model practical.

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

*I thank participants of MCMCSki conferences and seminars at University of Pennsylvania, Chicago, Princeton, and Yale for helpful discussions. I thank editors and anonymous referees for comments that helped to improve the manuscript. The support from the NSF Award SES-1851796 is gratefully acknowledged.

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 & 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 & Green (2012)).

This paper develops 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 the 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 the auxiliary priors approach of Carlin & Chib (1995) and retrospective sampling of Papaspiliopoulos & 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, & Hinton (1991), Jordan & Xu (1995), Peng, Jacobs, & Tanner (1996), Wood, Jiang, & Tanner (2002), Geweke & 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 & 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 & 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 & 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, an algorithm based on the conditional posterior proposals that changes parameters of only one mixture component in a cross-dimensional move is developed here.

In the mixture of experts 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 for this model. Posteriors for parameters of one mixture component conditional on the number of components and the rest of the parameters are much better behaved than posteriors for parameters of all or few mixture components as label switching is not an issue. Quadratic approximations to the log of the conditional posterior appear to be adequate in the considered applications. 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 practical performance of the proposed posterior simulator and the model are evaluated in simulations and in an application to estimation of the Engel curve between food expenditure share and total income. In simulations, the proposed MCMC algorithm reliably explores the posterior distribution of the number of mixture components even when the dimension of the parameter vector for each mixture component reaches 15. In the application to the food expenditure and income data, the estimated mixture of experts model outperforms standard parametric and nonparametric alternatives in out of sample performance comparisons.

The proposed methodology also appears promising for developing posterior simulators for other varying dimension models in which good proposals for the whole parameter vector are difficult to construct. An important class of such models are Bayesian nonparametric models based on nonlinear transformations of polynomials or other basis expansions with a prior on the number of basis functions.

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. A simple illustration of the algorithm on a normal regression with a nonparametric prior on the conditional mean is given in Section 6. An application to the mixture of experts model and simulation results are presented in Section 7. 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 & Peel (2000), Fruhwirth-Schnatter (2006)) and mixtures of experts (Jacobs et al. (1991), Jordan & Jacobs (1994)) with a varying number of mixture components and models based on polynomials or other basis expansions with a prior on the degree of polynomials (or more generally number of terms in the basis expansions).

2.1 Main Application: Mixture of Experts

The main motivation and application for the MCMC algorithm is a nonparametric model for conditional densities from Norets & Pati (2017) based on mixtures of experts. Let $y_i \in \mathbb{R}$ denote a dependent variable and $x_i = (1, x_{i1}, \ldots, x_{id_x})' \in \mathbb{R}^{d_x+1}$ 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 a mixture of normal linear regressions with the mixing weights that depend on covariates

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

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

where ϕ denotes a normal density and θ_{1m} includes $h_x \in \mathbb{R}_+^{d_x}$ and $h_y \in \mathbb{R}_+$ as a part of θ_1 and the sequences $\beta_j \in \mathbb{R}^{d_x+1}$, $\alpha_j \in \mathbb{R}_+$, $\mu_j \in \mathbb{R}^{d_x}$, $\nu_{xj} \in \mathbb{R}_+^{d_x}$, $\nu_{yj} \in \mathbb{R}_+$, $j = 1, 2, \ldots, m$. A prior $\Pi(m)\Pi(\theta_{1m}|m)$ specified in Section 7 completes the model setup.

The scale 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 & 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 & Keane (2007), and Villani et al. (2009)). Section 7 overviews 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 the same as the one implied by $p(Y|m, \theta_{1m})\Pi(\theta_{1m}|m)\Pi(m)$ and it is not affected by $\tilde{\Pi}$, which can be established by integrating out $\theta_{m+1\infty}$ from (4). The auxiliary prior $\tilde{\Pi}(\theta_{m+1\infty}|m, \theta_{1m}, Y)$ can only affect a posterior simulator for $\Pi(m, \theta_{1m}|Y)$. In what follows, we design $\tilde{\Pi}$ to facilitate posterior simulation from $\Pi(m, \theta_{1m}|Y)$ by retrospective sampling (Papaspiliopoulos & 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

A simplistic choice of the auxiliary prior with θ_j identically independently distributed for all $j \in \{1, \dots, \infty\}$ leads to practically zero acceptance rates for m in the mixture of experts application considered in this paper. Thus, the choice of $\tilde{\pi}_m$ appears to be crucial for feasibility of the algorithm. As shown 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 $\tilde{\pi}_m$. The idea of using the conditional posterior $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ as a proposal for cross-dimensional MCMC moves is rather natural and it has appeared in the literature in the form of centering a proposal for an RJMCMC algorithm on the conditional posterior mode (see a discussion of the conditional maximization approach in Brooks et al. (2003)). In Section 5, it is shown that the conditional posterior proposal maximizes the expected probability of acceptance for between-submodel moves and minimizes the asymptotic variance of MCMC sample average estimators under additional restrictions.

In some models, the conditional posterior $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ can be available in closed form as in the normal regression with a nonparametric prior on the conditional mean considered in Section 6. However, in general and in our main application to the mixture of experts model, direct draws from $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ and, especially, the normalization constant can be difficult to obtain. In this case, it is necessary to construct approximations to $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ with known normalization constants and from which fast simulation is possible.

When θ_m is one dimensional, as could be the case in models based on polynomials, suitable approximations to conditional posteriors/optimal proposals can be constructed from piece-wise linear approximations to the log of the unnormalized conditional posterior on a grid, similarly to adaptive rejection Metropolis sampling within Gibbs algorithm from Gilks et al. (1995). The resulting proposals would be piece-wise exponential with analytical expressions for normalization constants and it is straightforward to simulate from them.

When θ_m is multidimensional, an approximation to $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ can be given 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

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

This approximation can be motivated by the Bernstein-von Mises (BVM) theorem on the asymptotic normality of the posterior in well behaved or regular models, see, Chernozhukov & Hong (2003) and Kleijn & van der Vaart (2012) for versions of the theorem under misspecification that are more relevant here. The BVM theorem may fail to hold in some scenarios, see, for example, Chen et al. (2014). Even when it fails, the approximations centered at the mode of the target distribution seem reasonable and I am not aware of other possible approximations that would be feasible to obtain on every iteration of a long MCMC run. Thus, this approach is used in the mixture of experts application with a couple of simplifications. First, some of the cross derivatives in the Hessian are set to zero to simplify the derivations and speed up computation. Second, for parameters restricted to be positive, such as precision parameters, the corresponding cross-derivatives are set to zero in V_{m+1}^{-1} and the proposal is given by a Gamma distribution with the shape and rate parameters selected so that the mode and the variance of the Gamma distribution match the corresponding components in $(\bar{\theta}_{m+1}, V_{m+1})$. Appendix B provides more details on the algorithm implementation for the mixture of experts model.

3.2 Previous Literature

Papaspiliopoulos & 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 an 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 such proposals produce practically zero acceptance rates in the mixture of experts application.

Carlin & 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 & 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 the 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 reversible jump MCMC (Green (1995)) is the most popular approach to simulation from posterior for variable dimension models. In this section, it is shown that the algorithm for drawing m described in Section 3 can also be formulated as an RJMCMC algorithm. This RJMCMC algorithm is restricted to move only between the adjacent nested submodels without changing the parameters of the smaller submodel and it uses $\tilde{\pi}_m$ from Section 3 as a part of the proposal distribution for such cross-dimensional moves. Thus, the results on the optimal choice of $\tilde{\pi}_m$ presented in Section 5 below can be interpreted as the results on the optimal proposal distribution for a class of RJMCMC algorithms. This is noteworthy as the existing literature on choice of proposals for RJMCMC does not seem to contain rigorous optimality results (see, for example, Brooks et al. (2003) and Hastie & Green (2012)).

The RJMCMC representation of the algorithm requires additional notation. 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. An 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, an optimal choice of $\tilde{\pi}_m$ is considered. Since at each MCMC iteration, m can only be changed by 1, one can expect that higher acceptance rates for m^* result in a more efficient MCMC algorithm. Below, this intuition is made precise. First, Theorem 1 shows how $\tilde{\pi}_m$ can be chosen to maximize expected acceptance rates for m^*. Then, Theorem 2 shows 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 $g \in \mathcal{L}$, let us define the asymptotic MCMC variance as in Tierney (1998) by

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

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, an optimality result is obtained under additional restrictions on P and $\mathcal{L}$. The considered MCMC algorithms 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 Simple Illustration in Normal Regression

This section presents a very simple illustration of the proposed algorithm on a problem of flexible estimation of the conditional mean. Let us model the distribution of a univariate

y_i conditional on a covariate x_i by a normal distribution

$$y_i|x_i, \theta_{1m}, m \sim N(X_i^{1m}\theta_{1m}, 1),$$

where $\theta_{1m} \in \mathbb{R}^m$ contains linear coefficients, $X_i^{1m} = (\psi_1(x_i), \ldots, \psi_m(x_i))$, and $\psi_j(x_i)$ is a Legendre polynomial of order $j-1$. While univariate Legendre polynomials are used in this illustration, more generally, any basis and a multivariate x_i can be accommodated in the same fashion. It is also possible to model the variance of y_i and correspondingly extend the algorithm below, however, the variance is set to 1 here for simplicity. The prior on the degree of the polynomial is

$$\pi(m = j) = [\exp(\underline{A}_m) - 1] \cdot \exp(-\underline{A}_m j) \cdot 1\{m \geq 1\}. \tag{12}$$

Let us use a conditionally conjugate prior for the linear coefficients

$$\pi(\theta_{1m}|m) = \phi(\theta_{1m}; \underline{\theta}_{1m}, \underline{H}_{1m}^{-1}),$$

where ϕ is a multivariate normal density, $\underline{\theta}_{1m}$ is a prior mean, and $\underline{H}_{1m}$ is a diagonal precision matrix with a j's diagonal element denoted by $\underline{H}_j$. For this model, the posterior distribution for (m, θ_{1m}) can be obtained in a closed form. While MCMC is not necessary for estimation of this model, it does provide convenient settings for illustrating and checking the algorithm as the simulation results can be compared with the closed form expressions.

Let us denote the data by $y = (y_1, \ldots, y_n)'$ and let us implicitly condition on covariates hereinafter. Under a conditionally conjugate prior for θ_{1m}, all the relevant distributions are available in closed form. The Gibbs sampler block for linear coefficients is normal,

$$\pi(\theta_{1m}|m, y) = \phi(\theta_{1m}; \bar{\theta}_{1m}, \bar{H}_{1m}^{-1}), \tag{13}$$

where $\bar{H}_{1m} = \underline{H}_{1m} + (X^{1m})'X^{1m}$, $X^{1m} = ((X_1^{1m})', \ldots, (X_n^{1m})')'$, and

$$\bar{\theta}_{1m} = \bar{H}_{1m}^{-1}[\underline{H}_{1m}\underline{\theta}_{1m} + (X^{1m})'y].$$

The conditional posterior (optimal proposal) is also normal,

$$\pi(\theta_{m+1}|m+1, \theta_{1m}, y) = \phi(\theta_{m+1}; \bar{\theta}_{m+1}, \bar{H}_{m+1}^{-1}), \tag{14}$$

where $\bar{H}_{m+1} = \underline{H}_{m+1} + (X^{m+1})'X^{m+1}$, X^{m+1} is the last column of X^{1m+1}, and

$$\bar{\theta}_{m+1} = \bar{H}_{m+1}^{-1}[\underline{H}_{m+1}\underline{\theta}_{m+1} + (X^{m+1})'(y - X^{1m}\theta_{1m})].$$

Finally, let us note that the likelihood is given by

$$p(y|m, \theta_{1m}) = (2\pi)^{-n/2} \exp[-0.5(y - X^{1m}\theta_{1m})'(y - X^{1m}\theta_{1m})] \tag{15}$$

and that the marginal likelihood for m has a closed form expression

$$p(y|m) = \frac{p(y|m, \theta_{1m})\pi(\theta_{1m}|m)}{\pi(\theta_{1m}|m, y)},$$

where the closed forms of the likelihood, prior, and posterior for θ_{1m} conditional on m are specified above.

MCMC algorithm:[1]

1. Simulate θ_{1m} from $\pi(\theta_{1m}|m, y)$ defined in (13).

2. Simulate the proposed m^* from distribution $(0.5, 0.5)$ on $(m+1, m-1)$.

   (a) If $m^* = m+1$, simulate θ_{m+1}^* from $\pi(\theta_{m+1}|m+1, \theta_{1m}, y)$ defined in (14) and U from a uniform on $[0, 1]$. If

$$U \leq \frac{\exp(-\underline{A}_m)p(y|m+1, \theta_{1m}, \theta_{m+1}^*)\phi(\theta_{m+1}^*; \underline{\theta}_{m+1}, \underline{H}_{m+1})}{p(y|m, \theta_{1m})\phi(\theta_{m+1}^*; \bar{\theta}_{m+1}, \bar{H}_{m+1}^{-1})}$$

[1] The Matlab code for the algorithm is available at https://www.brown.edu/Departments/Economics/Faculty/Andriy_Norets/papers/code_poly_regression.zip

then accept the proposal (change the current state of the Markov chain from (θ_{1m}, m) to $((\theta_{1m}, \theta^*_{m+1}), m^*))$. Return to step 1.

(b) If $m^* = m - 1$, simulate U from a uniform on $[0, 1]$. If

$$U \leq \frac{p(y|m-1, \theta_{1m-1})\phi(\theta_m; \bar{\theta}_m, \bar{H}_m^{-1}) \cdot 1\{m \geq 2\}}{\exp(-\underline{A}_m)p(y|m, \theta_{1m})\phi(\theta_m; \underline{\theta}_m, \underline{H}_m^{-1})}$$

then accept the proposal (change the current state of the Markov chain from (θ_{1m}, m) to $(\theta_{1m-1}, m - 1))$. Return to step 1.

The performance of the estimation algorithm is illustrated on the following data: $y_i|x_i \sim N(2 \cdot x_i^2 \exp(x_i), 1)$, $x_i = -1 + 2(i-1)/(n-1)$, $i = 1, \ldots, n$, $n = 1000$, and the prior hyper-parameters: $\underline{A}_m = 1$, $\underline{\theta}_j = 0$, $\underline{H}_j = 1$ for all j. The simulated data, the true conditional mean, and the posterior conditional mean are presented in Figure 1.

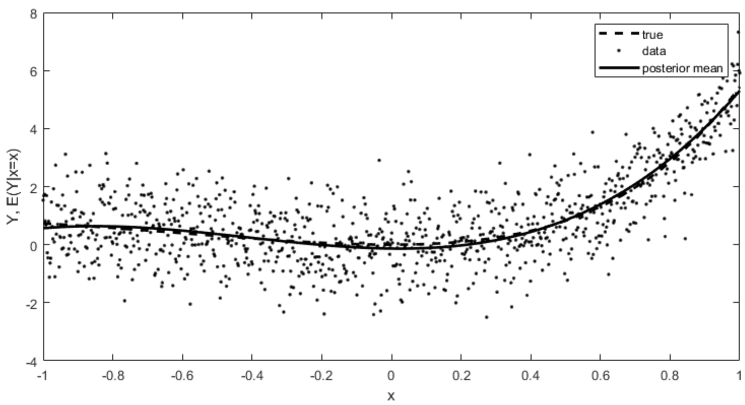

Figure 1: Simulated data (y_i, x_i), $i = 1, \ldots, 1000$, with the true conditional mean $E(y_i|x_i) = 2 \cdot x_i^2 \exp(x_i)$ (dashed line) and the estimated conditional mean (solid line).

The trace plot of MCMC draws of m in Figure 2 suggests that the Markov chain for exploring the posterior mixes well. The acceptance rate for m is 20%.

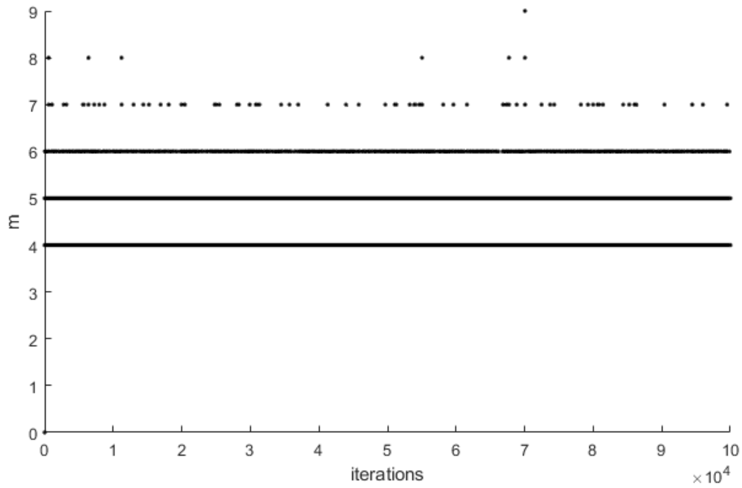

Figure 2: MCMC draws of m from the posterior of the normal regression model with a nonparametric prior on the conditional mean.

Figure 3 shows the prior for m, the posterior for m obtained from the MCMC draws, and the posterior obtained from the closed form expression for the marginal likelihood. The latter two probability mass functions essentially coincide, which confirms that the algorithm is implemented correctly and convergence is attained in this application.

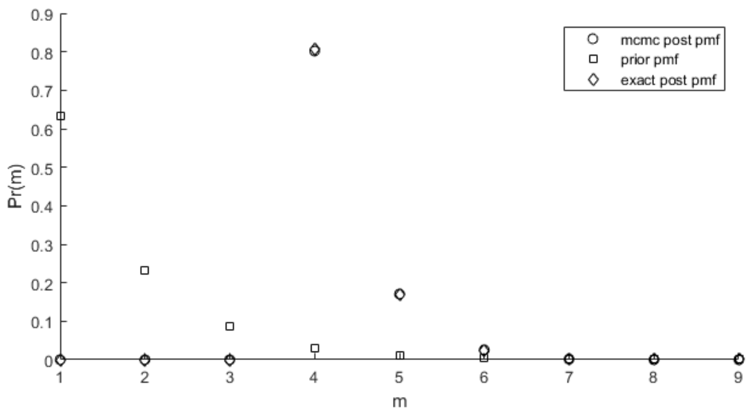

Figure 3: The prior and posterior of m for the normal regression model with a nonparametric prior on the conditional mean.

The following section presents a more involved application of the proposed algorithm in settings where other approaches do not seem to be available in the literature.

7 Application to Mixture of Experts

In this section, the algorithm is applied to the mixture of experts model described in Section 2.1. A discussion of prior specification, details of algorithm implementation, and tests for correctness of the implemented algorithm are presented below. 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 & 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 & 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.

7.1 Prior Specification

The prior $\Pi(\theta_{1m}|m)\Pi(m)$ is specified as follows. For $j = 1, \ldots, m$,

$$\begin{aligned}
&\beta_j \overset{iid}{\sim} N(\underline{\beta}, \underline{H}_\beta^{-1}),\ \mu_j \overset{iid}{\sim} N(\underline{\mu}, \underline{H}_\mu^{-1}),\\
&\nu_{yj} \overset{iid}{\sim} G(\underline{A}_{\nu y}, \underline{B}_{\nu y}),\ \nu_{xlj} \overset{iid}{\sim} G(\underline{A}_{\nu xl}, \underline{B}_{\nu xl}),\ l = 1, \ldots, d_x,\\
&(h_y)^{1/2} \overset{iid}{\sim} G(\underline{A}_{hy}, \underline{B}_{hy}),\ (h_{xl})^{1/2} \overset{iid}{\sim} G(\underline{A}_{hxl}, \underline{B}_{hxl}),\ l = 1, \ldots, d_x, \qquad (16)\\
&\alpha_j \overset{iid}{\sim} G(\underline{a}/m, 1),
\end{aligned}$$

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

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 & 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.

7.2 Overview of MCMC Algorithm

This subsection presents an overview of the MCMC algorithm for the conditional density model. A complete description of the algorithm is provided in Appendix B. As is common in the literature on MCMC for finite mixture models (Diebolt & Robert (1994)), latent mixture allocation variables $(s_1, \ldots, s_n)$ are added to facilitate the simulation from blocks of the Metropolis-within-Gibbs for given m: $y_i|x_i, s_i, m, \theta_{1m} \sim N((1; x_i)'\beta_{s_i}, (h_y \nu_{ys_i})^{-1})$ and $\Pi(s_i = j) = \gamma_j(x_i; m, \theta_{1m})$, where $\gamma_j(x_i; m, \theta_{1m})$ 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. More details are provided in Appendix B.[2]

7.3 Tests for Correctness of the Algorithm Design and Implementation

The simulator is implemented in Matlab. The joint distribution tests proposed in Geweke (2004) are used to check that the simulator is designed and implemented correctly. 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 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 the 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

[2]The Matlab code for the algorithm is available at `https://www.brown.edu/Departments/Economics/Faculty/Andriy_Norets/papers/code_mixture_experts.zip`

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 9 in Appendix C compares the exact prior probability mass function for m and the probability mass function obtained from the successive conditional simulator. Figure 10 presents a trace plot of m.

7.4 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.

$$y_i|x_i \sim e^{-\sqrt{x_{i1}}}\phi\left(\cdot; \Phi\left(\psi_1(x_i)\right), 0.5\psi_1(x_i)\right) + (1 - e^{-\sqrt{x_{i1}}})\phi\left(\cdot; \Phi\left(-\psi_1(x_i)\right), 0.1\psi_2(x_i)\right), \quad (17)$$

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 11 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 12.

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 13.

7.5 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 or a household (Lewbel (2008)). In empirical economics, Engel curves are often assumed to be linear or quadratic up to an additive error term. In this section, the density of the fraction of food expenditure conditional on the total income is estimated on the data from Battistin & Nadai (2015). Battistin & Nadai (2015) use data from the 2010 wave of the Bank of Italy's Survey on Households Income and Wealth to create a sample of households in which the male is between 25 and 56 years old, with the sample size of 2311 households. Battistin & Nadai (2015) focus on possible measurement errors and instrumental variables specifications. These issues are sidestepped in the present analysis. Instead, the assumptions of linear and quadratic model specifications are evaluated in a comparison of out of sample predictive performance of the Bayesian mixture of experts, linear and quadratic normal regressions, and a cross-validated kernel estimator. The performance of the MCMC algorithm for the mixture of experts is evaluated as well.

The prior hyperparameters in (16) for the mixture of experts model 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 (16) 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 inverse 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 4 shows the raw data and the estimated posterior means of the conditional densities and the conditional expectations. As can be seen from the figure, the distributions of the food expenditure share conditional on the total expenditure are noticeably more spread out and possibly multimodal for lower total expenditure households (such information would not be readily available from just regression estimation). Also, it appears that linear and quadratic specifications for the regression functions could be inferior to

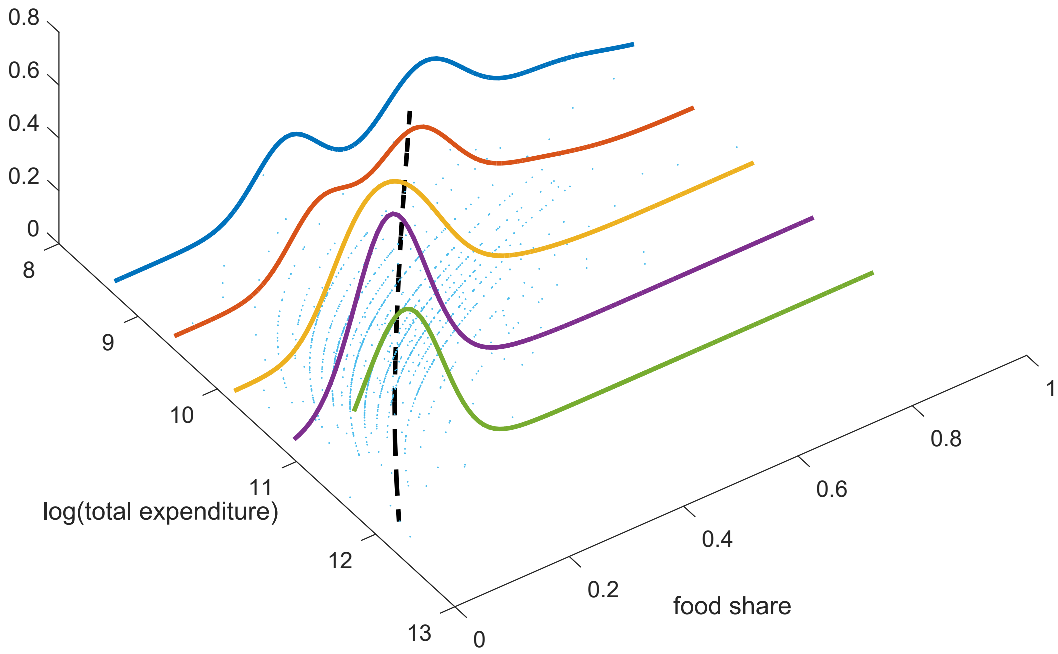

Figure 4: Observations on the food expenditure shares, y_i, and the logarithm of income, x_i, $i = 1, \ldots, 2311$; predictive conditional densities $p(y|x, y_1, x_1, \ldots, y_n, x_n)$ for $x \in \{8.6, 9.3, 10.1, 10.9, 11.6\}$ (solid colored lines) and the conditional expectation $E(y|x, y_1, x_1, \ldots, y_n, x_n)$ (dashed black line).

Figure 5 shows the prior and the estimated posterior probability mass functions for m. The posterior assigns the highest probability to $m = 3$, which suggests that just a few mixture components provide adequate fit to the data.

Figures 6 shows a trace plot with the MCMC draws of m. Figure 7 shows a trace plot of the log likelihood evaluated at the MCMC parameter draws. The log likelihood is a label invariant function of all the parameters and also a measure of model fit, thus, it is a convenient statistic for monitoring convergence. The trace plots in both figures suggest 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.

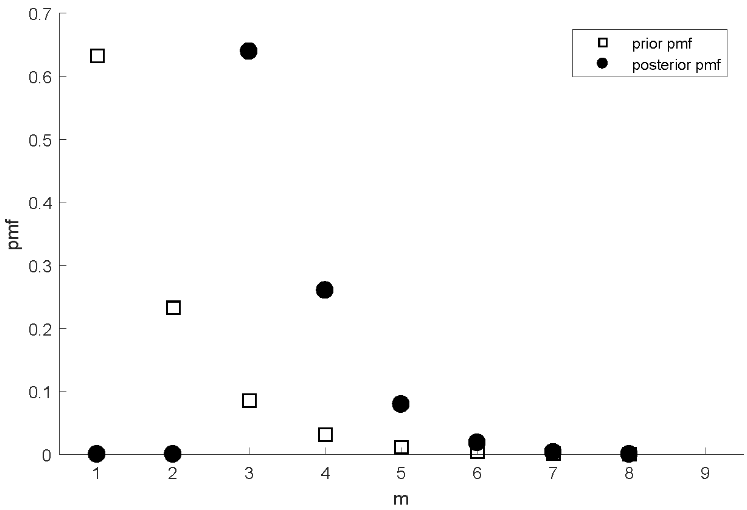

Figure 5: The prior and posterior distributions of the number of mixture components, m, for the mixture of experts model estimated on food expenditure shares and total expenditure data.

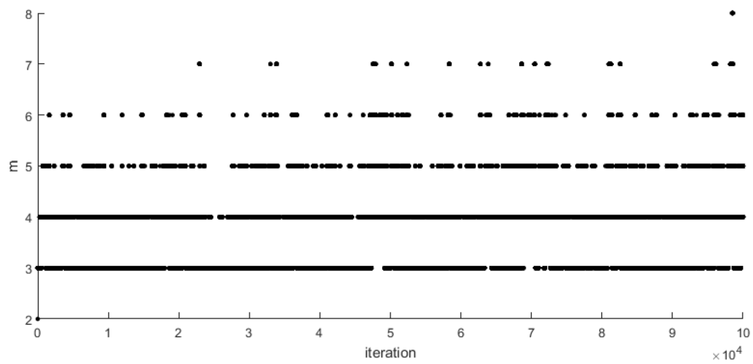

Figure 6: MCMC draws of m for the mixture of experts model estimated on food expenditure shares and income data.

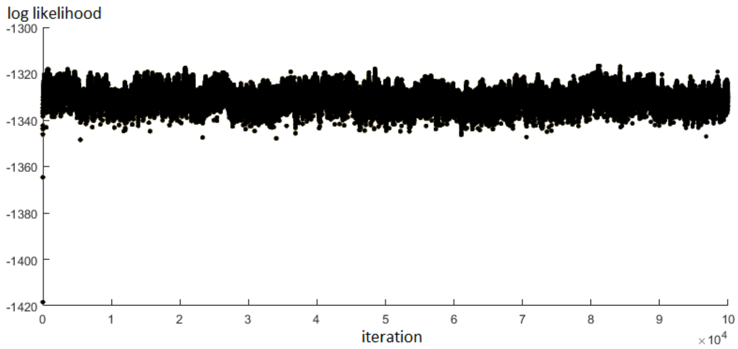

Figure 7: The log likelihood evaluated at MCMC draws of (m, θ_{1m}) for the mixture of experts model estimated on food expenditure shares and income data.

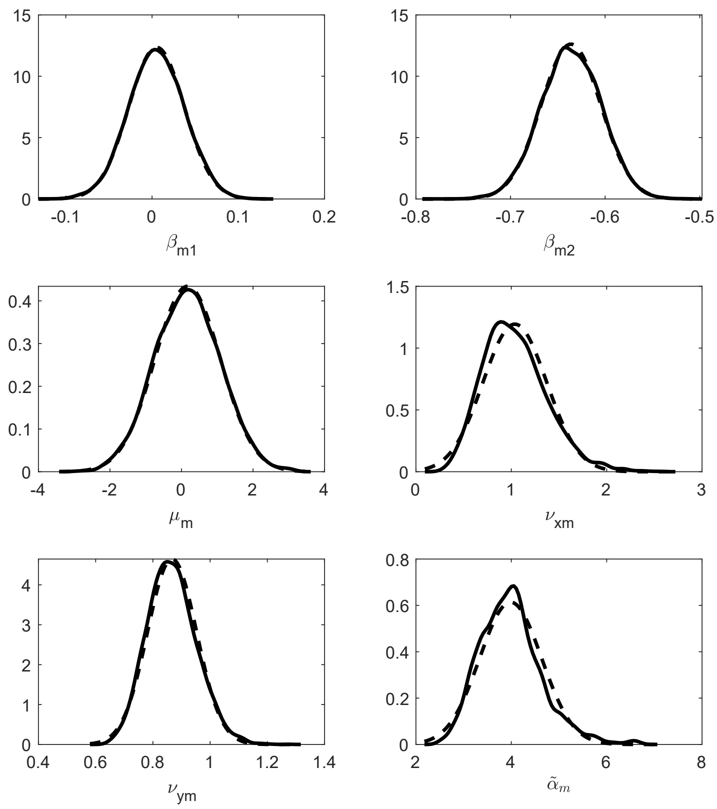

Figure 8: Marginal distributions for $p(\theta_m|Y, m, \theta_{1m-1})$ estimated from MCMC draws (solid lines) and their normal approximations (dashed lines) for a fixed (m, θ_{1m-1}).

As discussed in Section 3.1, the implementation of the MCMC algorithm uses approximations to the optimal proposal $p(\theta_m|Y, m, \theta_{1m-1})$. Figure 8 shows the marginals of the optimal proposal estimated by kernel smoothing (with default settings of ksdensity function in Matlab) on draws of θ_m from a version of the MCMC algorithm that keeps (m, θ_{1m-1}) fixed, and, thus, explores $p(\theta_m|Y, m, \theta_{1m-1})$. The figure also shows the corresponding normal approximations in dashed lines. The approximations appear to be adequate in this application. The fixed value of (m, θ_{1m-1}) is obtained as the last draw from a 5000 iteration posterior simulator run, and, thus, it is approximately simulated from $p(m, \theta_{1m-1}|Y)$.

The out-of-sample predictions of the Bayesian mixture of experts and classical parametric and kernel estimators are evaluated in 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 the 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 & 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 & 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.

8 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 numbers of experts (Norets & 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 theoretical results developed in the paper indeed show 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 the 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 implemented approximately optimal

MCMC method reliably explores the posterior distribution for the mixture of a variable number of experts model and the model outperforms standard parametric and nonparametric alternatives in out of sample performance comparisons in the application to Engel curve estimation.

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.

References

Battistin, E. & M.D. Nadai (2015) Identification and estimation of engel curves with endogenous and unobserved expenditures. *Journal of Applied Econometrics* 30(3), 487–508.

Brooks, S.P., P. Giudici, & 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(1), 3–39.

Carlin, B.P. & 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(5), 341–348.

Chen, X., M. Ponomareva, & E. Tamer (2014) Likelihood inference in some finite mixture models. *Journal of Econometrics* 182(1), 87 – 99.

Chernozhukov, V. & H. Hong (2003) An mcmc approach to classical estimation. *Journal of Econometrics* 115(2), 293–346.

Diebolt, J. & C.P. Robert (1994) Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society Series B (Methodological)* 56(2), 363–375.

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

Geweke, J. (2004) Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association* 99, 799–804.

Geweke, J. (2005) *Contemporary Bayesian Econometrics and Statistics.* Wiley-Interscience.

Geweke, J. (2007) Interpretation and inference in mixture models: Simple mcmc works. *Computational Statistics and Data Analysis* 51(7), 3529 – 3550.

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

Geyer, J.C. (2005) Markov chain monte carlo lecture notes. Unpublished.

Gilks, W.R., N.G. Best, & K.K.C. Tan (1995) Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 44(4), 455–472.

Green, P.J. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* 82(4), 711–732.

Hall, P., J. Racine, & Q. Li (2004) Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99(468), 1015–1026.

Hastie, D.I. & P.J. Green (2012) Model choice using reversible jump markov chain monte carlo. *Statistica Neerlandica* 66(3), 309–338.

Hayfield, T. & J.S. Racine (2008) Nonparametric econometrics: The np package. *Journal of Statistical Software* 27(5), 1–32.

Jacobs, R.A., M.I. Jordan, S.J. Nowlan, & G.E. Hinton (1991) Adaptive mixtures of local experts. *Neural Computation* 3(1), 79–87.

Jordan, M. & L. Xu (1995) Convergence results for the em approach to mixtures of experts architectures. *Neural Networks* 8(9), 1409 – 1431.

Jordan, M.I. & R.A. Jacobs (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6(2), 181–214.

Kleijn, B. & A. van der Vaart (2012) The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics* 6, 354–381.

Lewbel, A. (2008) Engel curve. In S.N. Durlauf & L.E. Blume (eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke.

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

Norets, A. & D. Pati (2017) Adaptive bayesian estimation of conditional densities. *Econometric Theory* 33(4), 980–1012.

Papaspiliopoulos, O. & G.O. Roberts (2008) Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika* 95(1), 169–186.

Peng, F., R.A. Jacobs, & 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(435), 953–960.

Peskun, P.H. (1973) Optimum monte-carlo sampling using markov chains. *Biometrika* 60(3), 607–612.

Richardson, S. & 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(4), 731–792.

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics* 28(1), 40–74.

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

Villani, M., R. Kohn, & P. Giordani (2009) Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics* 153(2), 155 – 173.

Wood, S., W. Jiang, & M. Tanner (2002) Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* 89(3), 513–528.

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

$$\max_g \int \min\left\{1, \frac{c \cdot f(z)}{g(z)}\right\} g(z) d\lambda(z), \tag{18}$$

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 (18) is equivalent to

$$\min_g \int |g(z) - cf(z)| d\lambda(z). \tag{19}$$

For $c > 1$, any g^* with $g^*(z) \leq cf(z)$ for (λ almost surely) all z solves (19). 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

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

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, the integral in (19) evaluated at g is strictly larger than the integral evaluated at $g' \leq cf$. 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 (19). For $c = 1$, $g^* = f$ is obviously the solution. Thus, $g^* = f$ solves (18), 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

$$Q((m, \theta_{1m-1}), m' \times A'_{m'-1}) = P((m, \theta_{1m-1}), m' \times A'_{m'-1} \times \Theta_{m'}), \tag{20}$$

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 $Q(\tilde{\pi}^*) \succeq Q(\tilde{\pi})$, the Peskun-Tierney theorem from Section A.4 delivers $v(g, Q(\tilde{\pi}^\star)) \le v(g, Q(\tilde{\pi}))$ 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, $P_1 P_2 P_1$ 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. $\square$

B Appendix. MCMC Algorithm for Mixture of Experts Model

This Appendix presents a detailed description of the MCMC algorithm for the mixture of experts model introduced in Section 2.1. The likelihood function for this model is given by

$$p(Y|X; m, \theta_{1m}) = \prod_{i=1}^{n} p(y_i|x_i; m, \theta_{1m})$$

where $p(y_i|x_i; m, \theta_{1m})$ is defined in (3). As mentioned in the algorithm overview in Section 7.2, it is convenient to introduce latent mixture allocation variables $s = (s_1, \ldots, s_n)$ with

$$\begin{aligned} p(Y, s|X; m, \theta_{1m}) &= \prod_{i=1}^{n} p(y_i|s_i, x_i; m, \theta_{1m}) p(s_i|x_i; m, \theta_{1m}) \\ &= \prod_{i=1}^{n} \phi\left(y_i, x_i'\beta_{s_i}, (h_y \cdot \nu_{ys_i})^{-1}\right) \gamma_{s_i}(x_i; m, \theta_{1m}). \end{aligned} \tag{21}$$

The prior distribution $\Pi(\theta_{1m}|m)\Pi(m)$ is specified in Section 7.1. To pin down the parameterization and transformations of the Gamma distributions note that the prior density of ν_{yj} is proportional to

$$\nu_{yj}^{\underline{A}_{\nu y}-1} \exp\{-\underline{B}_{\nu y}\nu_{yj}\}$$

and the prior density of h_y is proportional to

$$h_y^{\underline{A}_{hy}/2-1} \exp\{-\underline{B}_{hy} h_y^{1/2}\}.$$

The Metropolis-within-Gibbs blocks for all the parameters and latent variables are presented next.

Metropolis-within-Gibbs blocks

Block for s_i

For each $i = 1, \ldots, n$, the latent mixture allocation variable s_i has a multinomial distribution on $(1, \ldots, m)$ with the probabilities proportional to

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

Block for β_j

For $j = 1, \ldots, m$,

$$p\left(\beta_j|Y, s, X, m, \theta_{1m} \setminus \{\beta_j\}\right) \propto p(Y, s|X; m, \theta_{1m})\Pi(\theta_{1m}|m)$$

and with the normal conditionally conjugate prior it is also a normal distribution $N(\bar{\beta}_j, \bar{H}_{\beta j}^{-1})$, with precision $\bar{H}_{\beta j} = \underline{H}_\beta + h_y \nu_{yj} \sum_{i:\, s_i=j} x_i x_i'$ and mean $\bar{\beta}_j = \bar{H}_{\beta j}^{-1}(\underline{H}_\beta \underline{\beta} + h_y \nu_{yj} \sum_{i:\, s_i=j} x_i y_i)$.

Block for ν_{yj}

For $j = 1, \ldots, m$,

$$p\left(\nu_{yj}|Y, s, X, m, \theta_{1m} \setminus \{\nu_{yj}\}\right) \propto p(Y, s|X; m, \theta_{1m})\Pi(\theta_{1m}|m)$$

and with the conditionally conjugate gamma prior, it is a gamma distribution $G(\bar{A}_{\nu yj}, \bar{B}_{\nu yj})$, where $\bar{A}_{\nu yj} = \underline{A}_{\nu y} + 0.5 \sum_i 1\{s_i = j\}$ and $\bar{B}_{\nu yj} = \underline{B}_{\nu y} + 0.5 h_y \sum_{i:\, s_i=j}(y_i - x_i'\beta_j)^2$.

Block for h_y

The density of the Gibbs sampler block for h_y satisfies

$$
\begin{aligned}
&p\left(h_y|Y,s,X,m,\theta_{1m}\setminus\{h_y\}\right)\propto p(Y,s|X;m,\theta_{1m})\cdot\Pi(\theta_{1m}|m)\\
&\propto h_y^{n/2}\exp\{-0.5h_y\sum_{i=1}^{n}(y_i-x_i'\beta_{s_i})^2\nu_{ys_i}\}\cdot h_y^{\underline{A}_{hy}/2-1}\exp\{-h_y^{1/2}\underline{B}_{hy}\}.
\end{aligned}
\tag{22}
$$

As described in Section 7.1, the prior for h_y ($h_y^{1/2}\sim G(\underline{A}_{hy},\underline{B}_{hy})$) is not conditionally conjugate, which is required for good asymptotic properties. Hence, a Metropolis-within-Gibbs step is used for this block. Specifically, the Metropolis-Hastings proposal distribution is $G(\bar{A}_{hy},\bar{B}_{hy})$, with $\bar{A}_{hy}=(\underline{A}_{\nu y}+n)/2$ and $\bar{B}_{hy}=0.5\sum_i(y_i-x_i'\beta_{s_i})^2\nu_{ys_i}$, motivated by the functional form of the likelihood part and the polynomial part of the prior in (22). A proposed draw h_y^* is accepted with probability $\min\{1,\exp(-\underline{B}_{hy}((h_y^*)^{1/2}-(h_y)^{1/2}))\}$.

Blocks for h_x,ν_{xj},μ_j

Gibbs sampler block distributions for h_x, ν_{xj}, and μ_j do not have a known form and can only be evaluated up to a normalization constant. Hence, random walk Metropolis-Hastings within Gibbs steps are used for each of these blocks. For a block index b, $\theta_{1m}=(\theta_{1m,b},\theta_{1m,-b})$ and $\theta_{1m,b}\in\{h_x,\nu_{xj},\mu_j,\ j=1,\ldots,m\}$. Then, for the current parameter draw θ_{1m} and latent variables s the proposal distribution for $\theta_{1m,b}^*$ is $N(\theta_{1m,b},H(\theta_{1m,b})^{-1})$ centered at the current value $\theta_{1m,b}$ and the precision

$$
H(\theta_{1m,b})=-\frac{\partial^2}{\partial\theta_{1m,b}\partial\theta_{1m,b}'}\log[p(Y,s|X,m,\theta_{1m})\Pi(\theta_{1m}|m)],
$$

which is motivated by asymptotic normal approximations to the conditional posterior. The acceptance probability is

$$
\min\left\{1,\frac{p(Y,s|X,m,\theta_{1m,b}^*,\theta_{1m,-b})\Pi(\theta_{1m,b}^*,\theta_{1m,-b}|m)\phi(\theta_{1m,b};\theta_{1m,b}^*,H(\theta_{1m,b}^*)^{-1})}{p(Y,s|X,m,\theta_{1m})\Pi(\theta_{1m}|m)\phi(\theta_{1m,b}^*;\theta_{1m,b},H(\theta_{1m,b})^{-1})}\right\},
$$

where ϕ is the multivariate normal density. Note that the proposal precision in the numerator, $H(\theta^*_{1m,b})$, needs to be recomputed at the proposed value $\theta^*_{1m,b}$. If $H(\theta_{1m,b})$ is not positive definite then it is replaced by a diagonal matrix with the absolute values of the second derivatives on the diagonal; this does not happen often in simulations.

If a proposed draw $\theta^*_{1m,b}$ is outside of the parameter support (a negative component in h_x, for example), then it is automatically rejected as the prior density is zero at such values; this does not happen often in simulations.

An alternative independence chain algorithm for this type of blocks with the normal proposal centered at the conditional posterior mode found by a Newton method takes more time but does not lead to any substantial reduction in serial correlation of MCMC draws. Therefore, the described random walk approach is used in simulations.

Block for α_j

The simulation of $(\alpha_1, \ldots, \alpha_m)$ is performed in two steps: (i) simulate $\tilde{\alpha} = (\alpha_1 / \sum_{j=1}^m \alpha_j, \ldots, \alpha_{m-1} / \sum_{j=1}^m \alpha_j)$ and (ii) simulate $\sum_{j=1}^m \alpha_j$. The reason for this is that the posterior of $\sum_{j=1}^m \alpha_j$ conditional on $\tilde{\alpha}$ is equal to its prior, $G(\underline{a}, 1)$, and the likelihood of the model depends on $(\alpha_1, \ldots, \alpha_m)$ only through its normalized version $\tilde{\alpha}$. Therefore, to simulate $(\alpha_1, \ldots, \alpha_m)$, $\tilde{\alpha}$ is simulated first from the random walk within the Gibbs step as described above for (h_x, ν_{xj}, μ_j), and then $\sum_{j=1}^m \alpha_j$ is simulated from $G(\underline{a}, 1)$. Note that the implied prior for $\tilde{\alpha}$ that is used in the random walk within the Gibbs algorithm is equal to the Dirichlet distribution with the parameter $(\underline{a}/m, \ldots, \underline{a}/m)$.

Block for 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 $\theta_{j_1} = \theta_{temp}$.

The purpose of this block is to ameliorate the well known problem of insufficient label switching in MCMC algorithms for mixture models (Geweke (2007)) and how it interacts with jumps from m to $m-1$ here. For a fixed m, an MCMC algorithm can get stuck in one of the $m!$ symmetric posterior modes and produce little if any label switching. If the m^{th} mixture component happens to be a very important one then the algorithm is very unlikely to ever jump from m to $m-1$. Random label switching simply resolves this issue.

Note that a random permutation of the labels is a Markov transition that preserves the target posterior distribution as the likelihood and prior are invariant to label switching. If this permutation is performed so that the resulting Markov transition is reversible (e.g., as described above, or more generally the distribution on permutations should be such that any permutation has the same probability as its inverse), then, the theoretical results presented in the paper are not affected in any way if such a reversible Markov transition is added to the algorithm in a reversible fashion.

Block for m

As mentioned in Section 7.2, mixture allocation variables s_i, $i = 1, \ldots, n$ are marginalized out and not present in the conditioning set of the block for m.

The proposed draw m^* is equal to $m+1$ or $m-1$ with probabilities $(1/2, 1/2)$. For $m^* = m+1$, θ_{m+1} is simulated from $\tilde{\pi}_m(\theta_{m+1}|\theta_{1m}, Y)$, which is described precisely below, and m^* is accepted with probability $\min\{1, \alpha(m+1, m)\}$, where

$$\alpha(m+1, m) = \frac{p(Y|m+1, \theta_{1m+1})\Pi(\theta_{1m+1}|m+1)\Pi(m+1)}{p(Y|m, \theta_{1m})\Pi(\theta_{1m}|m)\Pi(m)\tilde{\pi}_m(\theta_{m+1}|\theta_{1m}, Y)}.$$

Proposed $m^* = m - 1$ is accepted with probability $\min\{1, \alpha(m-1, m)\}$, where

$$\alpha(m-1, m) = \frac{p(Y|m-1, \theta_{1m-1})\Pi(\theta_{1m-1}|m-1)\Pi(m-1)\tilde{\pi}_{m-1}(\theta_m|\theta_{1m-1}, Y)}{p(Y|m, \theta_{1m})\Pi(\theta_{1m}|m)\Pi(m)}.$$

When $m = 1$, $m^* = m-1$ is immediately rejected as the prior probability of $m^* = 0$ is zero and it enters the numerator of the acceptance probability. The approximately optimal proposal, $\tilde{\pi}_m(\theta_{m+1}|\theta_{1m}, Y)$, is constructed as follows. First, the mode of the conditional posterior, $\bar{\theta}_{m+1} = (\bar{\mu}_{m+1}, \bar{\beta}_{m+1}, \bar{\nu}_{ym+1}, \bar{\nu}_{xm+1}, \bar{\alpha}_{m+1})$, is obtained

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

The maximization is performed by a Newton method.

To increase the computation speed and avoid calculations of cross-derivatives, the parameter subvectors μ_{m+1}, β_{m+1}, ν_{ym+1}, ν_{xm+1k}, and α_{m+1} are set to be independent in the proposal. The proposal variance for each parameter subvector $\theta_{m+1,b} \in \{\mu_{m+1}, \beta_{m+1}, \nu_{ym+1}, \bar{\alpha}_{m+1}, \bar{\nu}_{xm+1k},\ k = 1, \ldots, d_x\}$ is obtained from the Hessians

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

For $\theta_{m+1,b} \in \{\mu_{m+1}, \beta_{m+1}\}$, the proposal is normal with mean $\bar{\theta}_{m+1,b}$ and variance $V_{m+1,b}$. For $\theta_{m+1,b} \in \{\nu_{ym+1}, \bar{\alpha}_{m+1}, \bar{\nu}_{xm+1k},\ k = 1, \ldots, d_x\}$, the proposal is Gamma with mode $\bar{\theta}_{m+1,b}$ and variance $V_{m+1,b}$ (using a truncated normal instead of a Gamma proposal for these parameters leads to a slightly worse algorithm performance).

The Newton method for finding the mode $\bar{\theta}_{m+1}$ sets cross subvector derivatives $\partial^2/\partial\theta_{m+1,b1}\partial\theta'_{m+1,b2}$ to zero for $b1 \neq b2$, so that the Hessian is block diagonal.

C Appendix. Additional Figures

C.1 Figures for Joint Distribution tests in Section 7.3

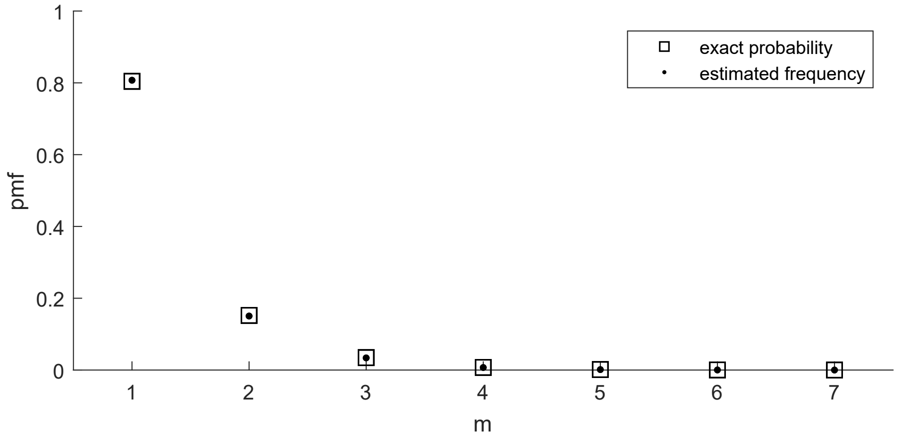

Figure 9: The exact prior probability mass function for m and the one obtained from the successive conditional simulator of the joint distribution test.

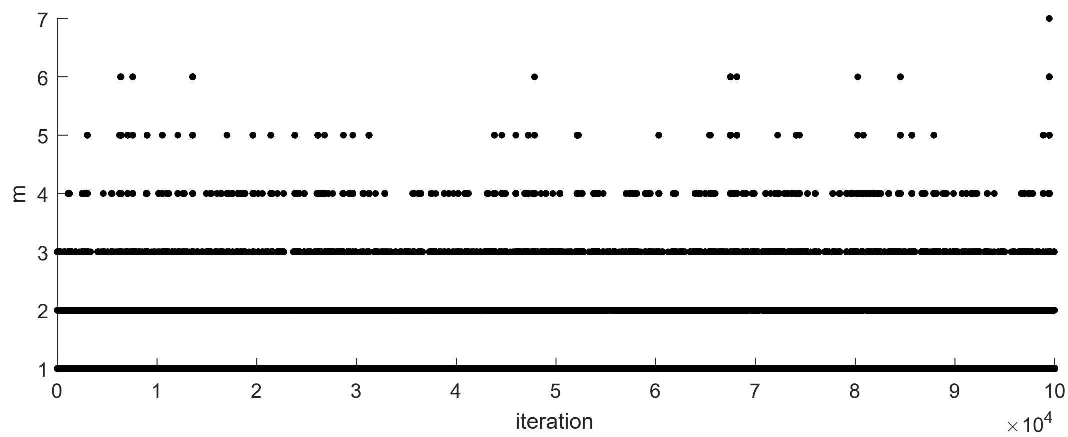

Figure 10: MCMC draws of m from the successive conditional simulator of the joint distribution test.

C.2 Figures for Simulated Data Experiments in Section 7.4

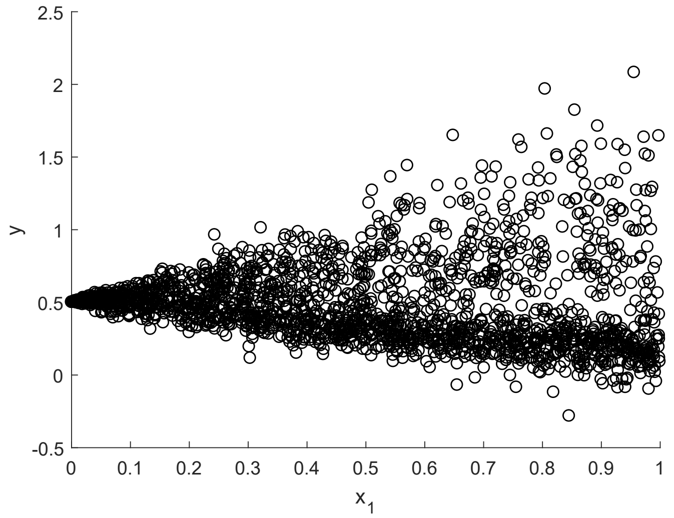

Figure 11: Simulated data for experiments, $d_x = 1$, $n = 2000$.

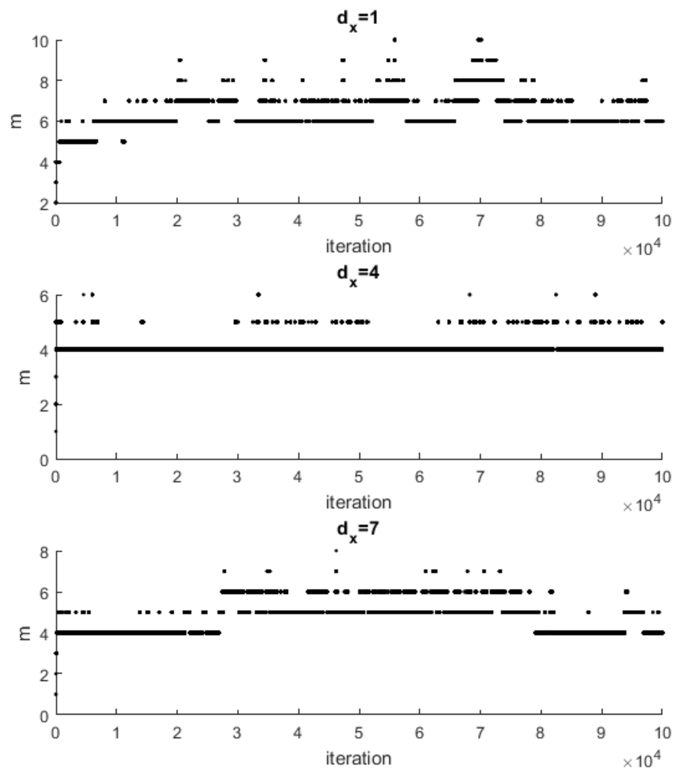

Figure 12: MCMC draws of m for simulated datasets with different dimension of covariates, $d_x \in \{1, 4, 7\}$

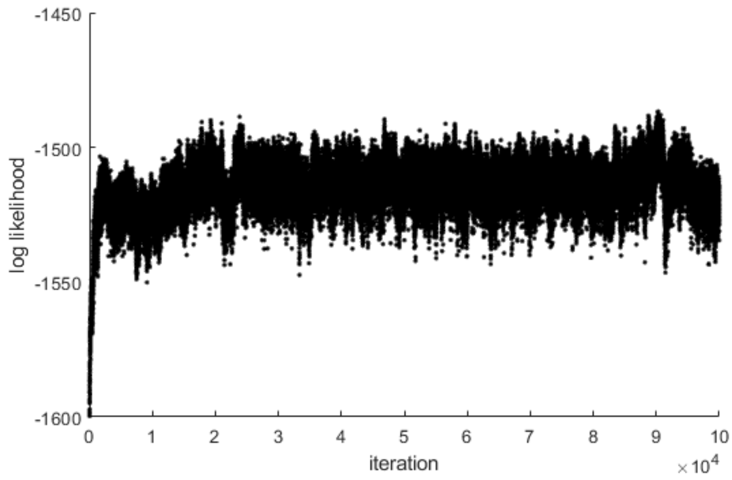

Figure 13: The log likelihood evaluated at MCMC draws of parameters, simulated data, $d_x = 1$.