

Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models[†]

Andriy Norets

Brown University, Providence, RI, USA.

E-mail: andriy.norets@brown.edu

Summary. 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 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.

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 two main approaches to MCMC estimation of such models are the reversible jump MCMC (RJMCMC) ([Green \(1995\)](#)) and the method of auxiliary prior distributions ([Carlin and Chib \(1995\)](#)). These general approaches require selection of proposal distributions or auxiliary priors, which is a non-trivial task, especially, in complex models. The literature on choice of efficient proposals for RJMCMC or auxiliary priors 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 \(2012\)](#)).

[†]Preliminary. Please, do not circulate.

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. [Norets and Pati \(2014\)](#) show that under rather standard priors and some regularity assumptions, the posterior in this model 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, it seems important to develop reliable posterior simulation algorithms for this model. 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. This approach seems difficult to implement though when mixing weights depend on covariates. Thus, I develop an algorithm based on the conditional posterior proposals described above. 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, Gaussian approximations for conditional posteriors are used in the implementation of the algorithm. The mean of the Gaussian approximation is set to the conditional posterior mode obtained by the Newton method (first and second order derivatives of unnormalized conditional posteriors are available in closed form) and the variance

is set to the inverse of the Hessian of the logarithm of the conditional posterior. The resulting “approximately” optimal RJMCMC algorithm provides a feasible estimation method in a wide range of settings and specifications. In simulation experiments, the acceptance rate for between-submodel moves falls in 0.1-10% range.

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. The theoretical results on the algorithm optimality are given in Section 5. The application to the mixture of experts model and simulation results are presented in Section 6. Appendix contains proofs and implementation details.

2. Model description

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

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

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

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

where $\Pi(\theta_{1m}|m)$ is a density with respect to a σ -finite dominating measure $\lambda^m = \lambda_1 \times \dots \times \lambda_m$ 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. Example: Nonparametric Model for Conditional Densities

The original motivation for the present paper is the following nonparametric model for conditional densities:

$$p(y_i|x_i, \mu, \alpha, h, \nu, m) = \sum_{j=1}^m \gamma_j(x_i) \cdot \phi(y_i, x'_i \beta_j, (h_y \cdot \nu_{y_j})^{-1}), \quad i = 1, \dots, n \quad (3)$$

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

where ϕ is a normal density. 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, \dots, x_n) to the conditioning set of all distributions.

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 the model (3).

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

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

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

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

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

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

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,

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

is an optimal choice of $\tilde{\pi}_m$. A simplistic choice of the auxiliary prior with θ_j identically independently distributed for all $j \in \{1, \dots, \infty\}$ 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 good performance 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

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

Alternatively, if a fast simulator for the conditional posterior $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ is available, then one could build a normal approximation to $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ from simulated draws (to make the proposal deterministic, one could use the same fixed initial draw and random generator seed).

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.

[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 retrospective 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

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

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, it seems plausible that higher acceptance rates for m^* result in a more efficient MCMC algorithm. Below, I first 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}) only.

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

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

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

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

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

The proof of the theorem is given in Appendix A.1. For $m^* = m + 1$, $\tilde{\pi}_m^*$ tends to produce proposals of θ_{m+1} with high value of the numerator in $\alpha(m^*, m)$, and one would intuitively expect $\tilde{\pi}_m^*$ 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 $\mathcal{L} = \{g : \mathcal{X} \rightarrow \mathbb{R}, \int g d\pi = 0, \int g^2 d\pi < \infty\}$. 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

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

where X_1, X_2, \dots 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 $\tilde{\pi} = \{\tilde{\pi}_m, m = 1, 2, \dots\}$ and have the following structure

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

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 ,

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

where $\tilde{\pi}^*$ 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 $\tilde{\pi}_m^*$ 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. However, such a result appears to be difficult to obtain. Nevertheless, the demonstrated optimality for functions of (m, θ_{1m-1}) is a useful justification for the proposed algorithm, especially, 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 a Covariate Dependent Mixture Model

6.1. Prior Specification

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

$$\begin{aligned} \beta_j &\stackrel{iid}{\sim} N(\underline{\beta}, \underline{H}_\beta^{-1}), \quad \mu_j \stackrel{iid}{\sim} N(\underline{\mu}, \underline{H}_\mu^{-1}), \\ \nu_{yj} &\stackrel{iid}{\sim} G(\underline{A}_{\nu y}, \underline{B}_{\nu y}), \quad \nu_{xlj} \stackrel{iid}{\sim} G(\underline{A}_{\nu xl}, \underline{B}_{\nu xl}), \quad l = 1, \dots, d_x, \\ (h_y)^{1/2} &\stackrel{iid}{\sim} G(\underline{A}_{hy}, \underline{B}_{hy}), \quad (h_{xl})^{1/2} \stackrel{iid}{\sim} G(\underline{A}_{hxl}, \underline{B}_{hxl}), \quad l = 1, \dots, d_x, \\ \alpha_j &\stackrel{iid}{\sim} G(\underline{a}/m, 1), \\ \Pi(m = k) &= (e^{\underline{A}_m} - 1)e^{-\underline{A}_m \cdot k}, \end{aligned}$$

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 (2014)

apply. Specifically, a gamma prior for (h_{x_l}, 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. A slightly more general prior for m (for example, $\Pi(m = k) \propto e^{-\underline{A}_m \cdot k \log(k)}$) would also work.

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, \dots, s_n) (Diebolt and Robert (1994)) to facilitate the simulation from blocks of the Metropolis-within-Gibbs for given m : $y_i | x_i, s_i, m, \theta_{1m}, h_x, h_y \sim N(x_i' \beta_{s_i}, (h_y \nu_{y s_i})^{-1})$ 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 . Random label switching on every MCMC iteration considerably increases acceptance rate for m .

The simulator is implemented in Matlab. It passes the joint distribution tests for correctness of the algorithm design and implementation from Geweke (2004). The objects of interest are (functionals of) the predictive conditional density, which is label invariant. Convergence test performed on label invariant statistics (Geweke (2007)) indicate that the algorithm converges.

6.3. Simulations and Applications

The algorithm implementation is currently being tested on simulated data. The data generating processes (DGP) are taken from Norets and Pelenis (2014) and Villani et al. (2009) (their heteroskedastic DGP). For 1-5 dimensional covariates and 500-10000 observations in the sample, acceptance rates for m are in 0.1-10% range and acceptance rates for other parameters exceed 60%. For some DGPs and sample sizes, the algorithm started at $m = 1$ can visit $m = 15$.

The preliminary results suggest that the proposed algorithm could provide a reliable posterior simulator for Bayesian nonparametric models of conditional densities. The algorithm can also be useful for estimation of simpler models such as countable mixtures of (multivariate) normals.

Simulation results and applications to real data will be added below ... One set of simulation experiments (a Monte Carlo comparing a mixture of experts to a cross-validated kernel estimator) can be found in Section 7 of Norets and Pati (2014).

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

$$\max_g \int \min \left\{ 1, \frac{c \cdot f(z)}{g(z)} \right\} g(z) d\lambda(z), \quad (12)$$

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 (12) is equivalent to

$$\min_g \int |g(z) - cf(z)| d\lambda(z). \quad (13)$$

For $c > 1$, any g^* with $g^*(z) \leq cf(z)$ for (λ almost surely) all z solves (13). 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

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

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 (13). For $c = 1$, $g^* = f$ is obviously the solution. Thus, $g^* = f$ solves (12), and it is a unique solution that does not depend on c . \square

A.2. Proof of Theorem 2

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

$$Q((m, \theta_{1m-1}), m' \times A'_{m'-1}) = P((m, \theta_{1m-1}), m' \times A'_{m'-1} \times \Theta_{m'}), \quad (14)$$

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 $\tilde{\pi} = \tilde{\pi}^*$ (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 π^* , and, thus, $Q_1((m, \theta_{1m-1}), \{m+1\} \times A_m)$ is maximized at π^* 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 π^* , and, thus, $Q_1((m, \theta_{1m-1}), \{m-1\} \times A_{m-2})$ is maximized at π^* 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 $v(g, Q(\tilde{\pi}^*)) \leq v(g, Q(\tilde{\pi}))$ 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 . \square

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 .

B. Appendix. MCMC Algorithm for Conditional Density Model

Metropolis-within-Gibbs blocks:

- $Pr(s_i = j | \dots) \propto \gamma_j(x_i) \cdot \phi(y_i, x'_i \beta_j, (h_y \cdot \nu_{yj})^{-1})$, $i = 1, \dots, n$.
- $\beta_j | \dots \sim N(\bar{\beta}_j, \bar{H}_{\beta_j}^{-1})$, $j = 1, \dots, m$,
where $\bar{H}_{\beta_j} = \underline{H}_{\beta_j} + h_y \nu_{yj} \sum_{i: s_i=j} x_i x'_i$ and $\bar{\beta}_j = \bar{H}_{\beta_j}^{-1} (\underline{H}_{\beta_j} \beta_j + h_y \nu_{yj} \sum_{i: s_i=j} x_i y_i)$.
- $\nu_{yj} | \dots \sim G(\bar{A}_{\nu y}, \bar{B}_{\nu y})$, $j = 1, \dots, m$,
where $\bar{A}_{\nu y} = \underline{A}_{\nu y} + 0.5 \sum_i 1\{s_i = j\}$ and $\bar{B}_{\nu y} = \underline{B}_{\nu y} + 0.5 h_y \sum_{i: s_i=j} (y_i - x'_i \beta_j)^2$.
- $h_y | \dots$ is simulated by the Metropolis-Hastings algorithm with proposal $G(\bar{A}_{h_y}, \bar{B}_{h_y})$,
where $\bar{A}_{h_y} = 0.5 \underline{A}_{h_y} + 0.5 \sum_i 1\{s_i = j\}$ and $\bar{B}_{h_y} = 0.5 \sum_i (y_i - x'_i \beta_{s_i})^2 \nu_{ys_i}$.
- $\tilde{\alpha} = (\alpha_1 / \sum_{j=1}^m \alpha_j, \dots, \alpha_{m-1} / \sum_{j=1}^m \alpha_j) | \dots$ 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 $\bar{\alpha}$ 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. Of course, the acceptance probabilities have different expressions for the two alternative procedures.

- $\sum_{j=1}^m \alpha_j | \dots \sim G(\underline{a}, 1)$ (this quantity is independent of data and 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 $h_x | \dots, \nu_{xjk} | \dots, \mu_j | \dots$ are handled analogously to $\tilde{\alpha} | \dots$
- Random label switching: simulate j_1 and j_2 randomly from a uniform distribution on $\{1, \dots, m\}$ and set $\theta_{temp} = \theta_{j_1}$, $\theta_{j_1} = \theta_{j_2}$, and $\theta_{j_2} = \theta_{temp}$.
- Block for $m | \dots$ is as described in Section 3, except instead of (8) a block diagonal approximation is used, where cross derivatives such as $\frac{\partial^2}{\partial \beta_j \partial \mu_j'}$ (but not off diagonal components in $\frac{\partial^2}{\partial \beta_j \partial \beta_j'}$) are set to zero to simplify calculations.

References

- 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.
- (2007): “Interpretation and inference in mixture models: Simple MCMC works,” *Computational Statistics and Data Analysis*, 51, 3529 – 3550.

- 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.
- HASTIE, D. I. AND P. J. GREEN (2012): “Model choice using reversible jump Markov chain Monte Carlo,” *Statistica Neerlandica*, 66, 309–338.
- 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. I. AND R. A. JACOBS (1994): “Hierarchical mixtures of experts and the EM algorithm,” *Neural Computation*, 6, 181–214.
- MCLACHLAN, G. AND D. PEEL (2000): *Finite Mixture Models*, John Wiley & Sons, Inc.
- NORETS, A. AND D. PATI (2014): “Adaptive Bayesian Estimation of Conditional Densities,” Forthcoming in *Econometric Theory*.
- NORETS, A. AND J. PELENIS (2014): “Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures,” *Econometric Theory*, 30, 606–646.
- 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.
- 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.