Bayesian GMM[*]

Minchul Shin[†]
University of Pennsylvania

This version: November 16, 2014

Job Market Paper
Please download the latest version at
http://goo.gl/Fgkuh9

Abstract

In this paper, I study a semiparametric Bayesian method for over-identified moment condition models. A mixture of parametric distributions with random weights is used to flexibly model an unknown data generating process. The random mixture weights are defined by the exponential tilting projection method to ensure that the joint distribution of the data distribution and the structural parameters are internally consistent with the moment restrictions. In this framework, I make several contributions to Bayesian estimation and inference, as well as model specification. First, I develop simulation-based posterior sampling algorithms based on Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods. Second, I provide a method to compute the marginal likelihood and use it for Bayesian model selection (moment selection) and model averaging. Lastly, I extend the scope of Bayesian analysis for moment condition models. These generalizations include dynamic moment condition models with time-dependent data and moment condition models with exogenous dynamic latent variables.

Keywords: Bayesian inference, sequential Monte Carlo, generalized method of moments, exponential tilting, Euler equation, dynamic latent variable models

JEL codes: C11, C14, E21

[*] I am indebted to my advisors, Frank Diebold and Frank Schorfheide, for their support and guidance at all stages of this research project. I would also like to thank the other dissertation committee members, Xu Cheng and Frank DiTraglia, for their support and valuable comments. I also thank Lorenzo Braccini, Siddhartha Chib, Mike Chirico, Herman van Dijk, Jesús Fernández-Villaverde, Dirk Krueger, Juan Rubio-Ramírez, Felipe Saffie, Anna Simoni, Dongho Song, Matt Taddy, and Molin Zhong for helpful comments and suggestions. All errors are my own.

[†] Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104. E-mail: mincshin@sas.upenn.edu.

1 Introduction

The estimation and testing of econometric models through moment restrictions have been the focus of considerable attention in the literature since the seminal paper by Hansen (1982). These types of models and associated tools have become a major tool for empirical economists due to its generality and flexibility. Many econometric problems, such as instrumental variable regression and quantile regression, can be cast in a moment condition model framework. Moreover, one can perform estimation and testing without fully specifying a model. This is especially important for empirical economists since economic theory does not always fully dictate the probabilistic structure of data.

Despite the popularity and importance of moment condition modeling, existing Bayesian methods have received relatively little attention vis-à-vis the treatment in the frequentist literature. One of the difficulties in Bayesian analysis of moment conditions is that the information contained in moment condition models is insufficient to construct a likelihood function of the model because the moment restrictions characterize only part of the econometric model. Various Bayesian procedures have been proposed to overcome this difficulty, but there are still many gaps in the literature. First, most papers have focused on the problem of parameter inference, omitting other meaningful issues such as model selection and model averaging. Second, most extant Bayesian procedures for moment condition models either assume *i.i.d.* (independently and identically distributed) data, or concentrate out the unknown distribution function of the data generating process and justify their approaches using asymptotic approximations.

As a step toward filling this gap, I develop a semiparametric Bayesian econometric method for moment condition models building on the semiparametric prior proposed by Kitamura and Otsu (2011), and make several contributions to Bayesian estimation and inference, as well as model specification. First, I develop simulation-based algorithms to perform finite-sample posterior analysis based on Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods. Second, I provide a method to compute the marginal likelihood and use it for Bayesian model selection (moment selection) and model averaging. Lastly, I extend the scope of Bayesian analysis for moment condition models to a wider class of data generating processes, such as dynamic moment condition models with time-dependent data as well as moment condition models with exogenous dynamic latent variables.

I flexibly model the unknown data generating process using mixtures of parametric densities. Then, the random mixture weights are restricted so that the data generating process satisfies the relevant moment conditions. Specifically, restricted random mixture weights are obtained by applying exponential tilting projections to distributions over the space of unrestricted random mixing distributions and parameters in the moment functions. As a result, unknown parameters in the moment functions are embedded in the random mixture weights, and the likelihood function can be obtained based on the mixture representation with restricted mixture weights. After specifying suitable prior distributions on model parameters, Bayes' theorem leads to the posterior distribution of the parameters in the moment functions, as well as the probability distribution of the data generating process.

I go on to develop simulation-based approaches that allow me to perform posterior analysis. The algorithms are based on Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods. I also provide a method to compute the marginal likelihood, which is typically challenging for Bayesian semiparametric models. Then, the computed marginal likelihood, in conjunction with the model prior probability, offers a decision-theoretic approach to the moment selection problem (Bayesian model selection). Moreover, it can be used to average a quantity of interest across models rather than analyzing a single selected model (Bayesian model averaging).

I extend the modeling framework and the associated posterior sampling algorithms to cover more complicated data generating processes. Time-dependency in the data is captured by modeling the joint distribution of current and past histories of the data as a mixture of parametric distributions under the assumption that the data generating process follows a p-th order time-homogeneous Markov process. It is also possible to extend the method to models with exogenous dynamic latent variables when its transition law is known up to finite dimensional parameters by modeling the conditional distribution of observables conditioned on latent variables as a mixture of parametric distribution. Then, a similar exponential tilting procedure is applied to the random weights in these mixture models to ensure that the resulting random unknown densities are internally consistent with the moment condition models.

I compare the performance of all three posterior samplers developed in this paper using simulated data. All posterior samplers produce almost identical posterior moment estimates. However, there are differences in their performance in terms of efficiency. Among MCMC-based samplers, the data-augmented version of the sampler improves the plain vanilla version of the sampler (basic sampler). The basic sampler is applicable to all modeling frameworks, while the use of the data-augmentation technique is limited to *i.i.d.* models, as it exploits the particular structure of the likelihood function. Under the simulation design considered, the current version of the SMC algorithm turns out to be less efficient than MCMC-based samplers. However, the SMC algorithm provides an approximation to the marginal likelihood, which is a valuable quantity for posterior analysis, although it is not obvious how to compute it based on the output from MCMC-based samplers, so it may be worth sacrificing some efficiency to achieve this end.

I next illustrate how one can use the marginal likelihood to select a model using simulated data. In the context of moment condition models, different model specifications are defined by different sets of moment conditions on the same dataset. Marginal likelihood computed based on the proposed SMC sampler correctly distinguishes models with invalid moment conditions from the correctly specified moment condition model.

In the empirical section, I use the proposed posterior sampling methods to estimate a risk aversion parameter based on an Euler equation allowing for demographic heterogeneity. Specifically, I use household-level annual food consumption and demographic characteristic data (the number of children) taken from the Panel Study of Income Dynamics (PSID). I impose the proposed modeling framework and apply the SMC sampler to perform posterior analysis. Preliminary results indicate that the risk aversion parameter is around $4.5 \sim 5.6$. Marginal likelihood comparison reveals that Euler equation restrictions are favored by the data, as the marginal likelihoods based on the moment conditions models are higher than those of an unrestricted nonparametric Bayesian model (the Dirichlet mixture model). However, not all Euler equation-based moment restrictions are equally useful. It turns out that the moment restrictions that include the number of children as a set of instruments deteriorate the marginal likelihood compared to other Euler equation models.

Related literature. This paper contributes to first and foremost to the literature on Bayesian approaches to moment condition models. It is most closely related to the work of Kitamura and Otsu (2011), who develop a generic method to construct semiparametric priors using exponential tilting projection and study its frequentist asymptotic properties. However, their actual implementation is limited to *i.i.d.* data. My paper complements theirs by providing a series of posterior samplers that allow one to perform a complete posterior analysis, including model selection and model averaging, for a more general class of models– moment condition models with serially dependent data and these with latent variables. Moreover, I provide conditions under which the proposed posterior samplers converge to the true posterior distribution. Other authors have proposed Bayesian approaches to moment condition models– one of the first attempts to obtain a posterior distribution based on moment conditions without the use of an assumed parametric likelihood function is Zellner's Bayesian method of moments (Zellner and Tobias, 2001, and references therein). However, Zellner's method usually restricts the moment conditions to those restricting first two moments (mean and variance), and the analysis is restricted to linear models, such as linear regression models and simultaneous equations models. More recently, Kim (2002) proposes a limited information likelihood that can be used to construct the posterior distribution based on moment conditions. Chamberlain and Imbens (2003) extend the Bayesian bootstrap to (exactly identified) moment condition models. Lazar (2003) studies posterior distributions based on the empirical likelihood. Schennach (2005) proposes a maximum entropy nonparametric Bayesian procedure. Florens and Simoni (2012) develop a Bayesian approach to GMM based on Gaussian process priors. However, these analyses are all restricted to the *i.i.d.* case except for the limited information likelihood approach of Kim (2002). Kim's pseudo-likelihood is used by Gallant et al. (2014) to estimate moment condition models with latent variables. While Kim's likelihood based method abstracts away from *i.i.d.* environments, his approach is based on asymptotics and does not allow finite sample posterior analysis, as the present work does.

This paper is also related to the literature on Bayesian density estimation and prediction with moment restrictions. The method considered in this paper estimates an unknown distribution in conjunction with moment conditions.[1] Similarly, Choi (2013) considers Bayesian density estimation with moment restrictions where the prior information about the parameters of interest is only available in the form of moment conditions. His focus is on density estimation; this paper considers estimation of both the parametric and nonparametric unknowns. A method to incorporate moment restrictions derived from economic theory into predictive distributions is also proposed in the literature. Robertson et al. (2005) use exponential tilting projection to obtain a refined predictive distribution of macroeconomic variables subject to Taylor rule restrictions. Giacomini and Ragusa (2014) provide formal justification of the method and show that when the moment restrictions are correct the resulting predictive distribution is indeed superior to the original predictive distribution in terms of log-score. The method considered in this paper is different from theirs in that the exponential projection is applied to the prior distribution of the underlying data distribution as opposed to the posterior predictive distribution. Moreover, the underlying data distribution is flexibly modeled and can allow for nonlinearity, while they only consider a linear vector autoregressive model.

This paper utilizes the Dirichlet process mixture model to make inferences about an unknown distribution. After the pioneering work of Ferguson (1974) and Antoniak (1974), there have been much research to develop nonparametric and semiparametric Bayesian methods

[1] Use of extra information in the form of a moment condition is also considered in the frequentist literature. Oryshchenko and Smith (2013) show the efficiency gain in the kernel density estimation when information derived from such moment restrictions is exploited.

under various frameworks.[2] The Dirichlet process mixture modeling approach has first introduced in the econometric literature by Tiwari et al. (1988). Chib and Hamilton (2002) consider a semiparametric panel potential outcomes model where the joint distribution of the treatments and potential outcomes is modeled as a mixture of normals with a random number of components using the Dirichlet process prior. Hirano (2002) extends the random effect autoregressive model to accommodate a flexible distribution for the disturbances using the Dirichlet process mixture model. Griffin and Steel (2004) develop a semiparametric Bayesian method for stochastic frontier models using the Dirichlet process mixture model. Conley et al. (2008) develop a Bayesian semiparametric approach to the linear instrumental variable problem where the joint distribution of the disturbances in the structural and reduced form equations is modeled as the Dirichlet process mixture. Taddy and Kottas (2010) propose a Bayesian nonparametric quantile regression based on the Dirichlet process mixture model. Chib and Greenberg (2010) study a flexible Bayesian analysis of regression models for continuous and categorical outcomes where the regression function is modeled additively by cubic splines and the error distribution is modeled as a Dirichlet process mixture. Applications of the Dirichlet mixture models to stock returns and their volatility are quite an active area of research; see Jensen and Maheu (2014) and references therein. However, none of these consider over-identified moment condition models in a general form: as such.

Other Bayesian density estimation and flexible regression estimation methods are abound. Geweke and Keane (2007), Villani et al. (2009), and Villani et al. (2012) develop a method to estimate a conditional distribution using a finite mixture of normals, allowing the mixing weights to depend on covariates. Norets (2010), Norets and Pelenis (2012, 2013), and Pelenis (2014) provide posterior consistency results for various flexible Bayesian methods to conditional density estimation. These papers are related to mine in the sense that both attempt to model the underlying unknown data distribution in a flexible manner. However, I mostly focus on modeling an unconditional distribution in conjunction with unconditional moment restrictions.

Finally, one of the MCMC-based algorithms proposed in this paper is a modified version of the Blocked Gibbs sampler of Ishwaran and James (2001). The Blocked-Gibbs sampler is a posterior sampling method for the Dirichlet process mixture model. I modify the algorithm to deal with complications induced by the introduction of moment conditions. The SMC-based algorithm in this paper is based on the tempered-likelihood SMC algorithm studied by Herbst and Schorfheide (2014). Their algorithm was developed mainly in the context of DSGE models. In this paper, I study and apply the algorithm in the context of Bayesian

[2] What follows is by no means a complete list. See Dey et al. (1998) for applications in statistics and Griffin et al. (2011) for applications in econometrics.

moment condition models. Different types of SMC methods have also been applied to DPM models (without moment restrictions). For example, Carvalho et al. (2010) apply the SMC algorithm to DPM models in the context of parameter learning and Griffin (2014) develops an adaptive method for truncation order selection in truncated DPM models based on SMC techniques.

Organization of the paper. In Section 2 presents the model for *i.i.d.* data. Section 3 extends model introduced in Section 2 to moment condition models with dependent data and latent variables. Prior specification and other details of the model are discussed in Section 4. Three posterior sampling algorithms are presented and studied in Section 5. The first part of Section 6 presents a comparison of the performance of the proposed algorithms using simulated data under *i.i.d.* environment. The second part illustrates implementation of the posterior algorithms under more complicated setting: an Euler equation model with the time-dependent data and a robust state-space model. Section 7 applies the SMC algorithm to an Euler equation model using actual U.S. household-level data to estimate the risk aversion parameter and investigate whether Euler equation restrictions are favored by the data. Section 8 concludes.

2 Model

2.1 Moment-restricted Dirichlet process model (MR-DPM)

Consider the following moment condition:

$$E_P[g(\beta,x)] = \int g(\beta,x)dP(x) = 0 \tag{1}$$

where $\beta \in B \subseteq R^k$ is a finite-dimensional parameter and x is a $d \times 1$ random vector; E_P is the expectation operator associated with the probability measure P; and the known function $g(\cdot,\cdot)$ maps the parameter β and the realization of x to an $r \times 1$ real-valued vector. r can be larger than k (over-identification). Following Kitamura (2006), I denote $\mathcal{P}(\beta)$ as a set of all probability measures that are compatible with the moment restriction for $\beta \in B$,

$$\mathcal{P}(\beta) = \left\{ P \in M : \int g(\beta,x)dP = 0 \right\}$$

where M is a set of all probability measures on R^d. And the union of $\mathcal{P}(\beta)$ over the parameter space is called a statistical model and is denoted as,

$$\mathcal{P} = \bigcup_{\beta \in B} \mathcal{P}(\beta).$$

The first goal of this paper is to obtain the posterior distribution of β and P (or posterior moments of its functional) given data $\{x_i\}_{i=1}^N$ generated (independently and identically distributed) from the unknown distribution, P which is assumed to be an element of $\mathcal{P}$. To this end, I consider a nonparametric conditional prior where the underlying data density follows a mixture of parametric densities. Specifically, conditional on some $\beta \in B$, the unknown data density is expressed as

$$f_P(x|\beta) = \int k(x;\theta) d\widetilde{G}_\beta(\theta), \tag{2}$$

where $k(\cdot;\theta)$ is called a kernel function and is usually a density of some parametric distribution indexed by θ and the mixing distribution $\widetilde{G}_\beta(\cdot)$ is assumed to be discrete and random, with its realization obtained in two steps. The first step draws a discrete distribution $G(\cdot)$ from the Dirichlet process $DP(\alpha, G_0)$ with concentration parameter α and base measure G_0. The second step solves the following informational projection to obtain the mixing distribution, $\widetilde{G}(\cdot)$:

$$\min_{\widetilde{G}} \int log\left(\frac{d\widetilde{G}}{dG}\right) d\widetilde{G} \quad s.t. \quad \int\int g(\beta, x)k(x;\theta)d\widetilde{G}(\theta)dx = 0. \tag{3}$$

This second procedure is called an "exponential tilting projection" and guarantees that any resulting density function corresponds to a draw from the above-specified nonparametric prior contained in $\mathcal{P}(\beta)$, – that is, it satisfies the moment restrictions at β. The prior specification is completed by imposing a parametric prior distribution for the finite dimensional parameter β.

In the absence of the exponential tilting projection step, the mixture model given by Equation 2 is the Dirichlet process mixture (DPM) model , which is a popular nonparametric Bayesian method for the density estimation problem (e.g., Müller and Quintana, 2004) known to be very flexible and rich. For example, when the base measure is chosen so that it has full support on the real line, the support of the mixing distribution G contains all probability measures (Ghosal, 2010). This model is regarded as nonparametric in the literature, since the number of mixtures is treated as unknown and random.

One major difficulty of using the DPM model in the moment condition model framework is that when one attempts to impose a DPM prior on P in conjunction with a separate independent prior distribution on β, the probability that a draw from the joint prior distribution satisfies the moment restrictions can easily be zero. The exponential tilting projection procedure fixes this problem by projecting probability measures for the mixing distribution, G, onto the space of discrete distributions that satisfy the moment restriction defined in Equation 3. This optimization has a nice dual problem that makes the computation tractable, and the resulting tilted mixing distribution, $\widetilde{G} = \widetilde{G}(\beta, G)$, is given by

$$\frac{d\widetilde{G}}{dG}(\theta) = \frac{\exp\left(\lambda(\beta, G)'\widetilde{g}(\beta,\ \theta)\right)}{\int \exp\left(\lambda(\beta, G)'\widetilde{g}(\beta,\ \theta)\right) dG(\theta)} \tag{4}$$

where

$$\lambda(\beta, G) = \arg\min_{\lambda} \int \exp\left(\lambda'\widetilde{g}(\beta, \theta)\right) dG(\theta). \tag{5}$$

and λ is an $r \times 1$ vector. Throughout the paper, I will refer to $\tilde{g}(\beta, \theta) = \int g(x, \beta)k(x; \theta)dx$ as an integrated moment condition.[3] Note that this minimization problem finds the optima over the finite-dimensional space R^r.

I will refer to the semiparametric model as the moment-restricted Dirichlet process mixture (MR-DPM) model. Under the MR-DPM model, the likelihood function can be expressed as

$$p(x_{1:N}|\beta, G) = \prod_{i=1}^{N} \left(\int k(x_i; \theta)\ d\widetilde{G}(\theta; \beta, G) \right),$$

where a tilted DP draw $\widetilde{G}$ is an implicit function of β and G given by the exponential tilting projection procedure in Equation 4.

Discussion 1 (Exponential tilting projection). The exponential tilting projection in Equation 3 is not the only way to impose restrictions on the mixing distribution $G(\cdot)$. The exponential tilting projection minimizes the Kullback-Leibler (KL) divergence measure between the original mixing distribution $G(\cdot)$ and the restricted mixing distribution $\widetilde{G}(\cdot)$. More generally, one can consider the minimization of f-divergence (Csiszàr, 1967) subject

[3] To obtain this object, I simply change the order of integration in the moment condition:

$$\int\int g(x, \beta)k(x; \theta)d\widetilde{G}(\theta)dx = \int\int g(x, \beta)k(x; \theta)dxd\widetilde{G}(\theta) = \int \widetilde{g}(\beta, \theta)d\widetilde{G}(\theta)$$

to moment conditions,

$$\min_{\widetilde{G}} \int f\left(\frac{d\widetilde{G}}{dG}\right) d\widetilde{G} \quad s.t. \quad \int\int g(\beta, x)k(x;\theta)d\widetilde{G}(\theta)dx = 0. \tag{6}$$

where the function $f(\cdot)$ is strictly convex and satisfies $f(1) = 0$. This class of divergence functions includes well-known divergence measures such as Hellinger and KL divergences (e.g., Kitamura, 2006). The f-divergence minimization problem in Equation 6 also has a dual representation as in Equation 4, thereby rendering computation feasible. However, I will focus on KL divergence in this paper because some nice theoretical properties such as posterior consistency hold under this divergence (Kitamura and Otsu, 2011). Comparison of the posterior distribution resulting from utilizing different divergences is an interesting and open question, both in finite sample and asymptotic analysis.

Discussion 2 (Kitamura and Otsu, 2011). This paper is not the first to apply the exponential projection to a Bayesian moment condition model. The most closely related work is Kitamura and Otsu (2011, hereafter KO), who consider the following problem,

$$\min_{\widetilde{P}} \int log\left(\frac{d\widetilde{P}}{dP}\right) d\widetilde{P} \quad s.t. \quad \int g(\beta, x)d\widetilde{P}(x) = 0, \tag{7}$$

where the exponential projection is used to obtain a tilted probability measure $\widetilde{P}$. In conjunction with the DPM formulation for P, KO call the resulting model the exponentially tilted Dirichlet process mixture (ET-DPM) model. Under the ET-DPM modeling framework, KO show the posterior consistency of the finite dimensional parameter β, and they show that the resulting limit distribution achieves the semiparametric efficiency bound.

KO's ET-DPM modeling framework is slightly different from the projection procedure of this paper (MR-DPM model) in that KO's procedure projects the probability measures of the underlying data generating process, while that in this paper projects mixing distributions over the space of mixture distributions defined in Equation 2. Note, however, that the constraint in both projection problems is identical, and therefore, both generate a semiparametric prior distribution for the moment condition model $\mathcal{P}$.

What makes the approach taken in this paper attractive is its computational tractability and practicality. Under the ET-DPM modeling framework, obtaining the tilted probability measure amounts to evaluating the following integral numerous times:

$$\int \exp(\lambda' g(x, \beta))k(x, \theta)dx,$$

and this can be computationally costly. On the other hand, obtaining the tilted probability measure in the MR-DPM modeling framework amounts to evaluating the term:

$$exp(\lambda' \widetilde{g}(\beta, \theta)) \quad \text{where} \quad \widetilde{g}(\beta, \theta) = \int g(x, \beta)k(x; \theta)dx,$$

where the integral is computed before exponentiation. Computation of the integral in the above term is relatively simpler, at least for the applications considered in this paper. This integral has a closed form for many economic applications, including IV regression, quantile regression, and IV quantile regression. In the Appendix, I provide closed forms and derivations for these models.

In the actual computation of the posterior distribution, KO model the underlying data distribution using a Dirichlet process rather than the Dirichlet process mixture for simplicity and name this the ET-DP model. This leads to,

$$X \sim i.i.d.\ \widetilde{P}, \quad \widetilde{P} \leftarrow P, \quad \text{and} \quad P \sim DP(\alpha, G_0)$$

where $\widetilde{P} \leftarrow P$ denotes the exponential tilting projection in Equation 7. Under this modeling assumption, the optimization problem in Equation 7 is much simpler vis-à-vis the ET-DPM model, which facilitates posterior computation.

Nevertheless, there are a few reasons that one might want to model the unknown data distribution through the Dirichlet process mixture model. First, as mentioned earlier, a draw from the Dirichlet process is discrete with probability one, and therefore, the tilted draw $\widetilde{P}$ inherits this property. If one wants to obtain and analyze a density prediction for a continuous random variable, non-smoothness in the data generating process might be problematic. Second, as will be seen in a later section, the particular choice of the kernel function in the DPM formulation opens the door to Bayesian modeling with moment restrictions under more complicated yet important data structures. Such extensions include dynamic moment condition models with time-dependent data and moment condition models with dynamic latent variables, which are rarely studied in the literature.

The approach taken in this paper is somewhere between the ET-DP and the ET-DPM approach presented in Kitamura and Otsu (2011) in the sense that it keeps computational tractability while sticking with the Dirichlet process mixture formulation.

2.2 Stick-breaking approximation and J-truncated MR-DPM

In practice, solving the minimization problem in Equation 3 requires the actual realization of G from $DP(\alpha, G_0)$. This is infeasible because G can be infinite dimensional. As a work-around, I approximate the DP draw G by a truncated version of the stick-breaking representation of the Dirichlet process. The approximation is based on the stick-breaking representation of Sethuraman (1994), which is defined as

$$G(\cdot) = \sum_{j=1}^{\infty} q_j \delta_{\theta_j}(\cdot) \tag{8}$$

where $\theta_j \sim_{i.i.d.} G_0$ and $\delta_{\theta_j}(\cdot)$ is the Dirac delta function. The weights q_j arise through the stick-breaking construction

$$q_1 = V_1; \quad q_j = V_j \prod_{r=1}^{j-1}(1 - V_r); \quad V_j \sim Beta(1, \alpha). \tag{9}$$

This representation bears out that a realization from the Dirichlet process is a discrete distribution whose support points are randomly assigned based on the base measure, and that its associated weights are constructed using independent Beta random draws V_j. Note that the weights sum to one and are eventually expected to be small as j increases.

The stick-breaking approximation is made tractable by truncating the infinite sum at some finite integer J:

$$G_J(\cdot) = \sum_{j=1}^{J} q_j \delta_{\theta_j}(\cdot), \quad \theta_j \sim_{i.i.d.} G_0 \tag{10}$$

where the weights q_j are constructed in the same way as in Equation 9 for $j = 1, ..., J-1$. The last weight is set to one ($V_J = 1$) so that the sum of the weights totals exactly one. I will denote the J-truncated DP draw as $G_J \sim DP_J(\alpha, G_0)$; note that it can be summarized by a collection of vectors and matrices as $G_J = \{q, \theta\}$ with $q = [q_1, q_2, ..., q_J]$ and $\theta = [\theta_1, \theta_2, ..., \theta_J]$.

With the realization G_J from the truncated Dirichlet process and $\beta \sim p(\beta)$, the exponential tilting projection becomes:

$$\min_{\widetilde{q}} \sum_{j=1}^{J} log\left(\frac{\widetilde{q}_j}{q_j}\right)\widetilde{q}_j \quad s.t. \quad \sum_{j=1}^{J} \widetilde{q}_j \widetilde{g}(\beta, \theta_j) = 0, \quad 0 \leq \widetilde{q}_j \leq 1, \quad \sum_{j=1}^{J} \widetilde{q}_j = 1, \tag{11}$$

and the solution $\widetilde{q} = \widetilde{q}(\beta, G_J)$ is given by

$$\widetilde{q}_j = \frac{\exp\left(\lambda(\beta, G_J)'\widetilde{g}(\beta,\ \theta_j)\right)}{\sum_{j=1}^{J} q_j \exp\left(\lambda(\beta, G_J)'\widetilde{g}(\beta,\ \theta_j)\right)} q_j$$

where

$$\lambda(\beta, G_J) = \arg\min_{\lambda} \sum_{j=1}^{J} q_j \exp\left(\lambda'\widetilde{g}(\beta, \theta_j)\right). \tag{12}$$

and the tilted mixing distribution is composed of the tilted mixture probabilities $\widetilde{q} = [\widetilde{q}_1, \widetilde{q}_2, ..., \widetilde{q}_J]$ and the parameters in the mixture density $\theta = [\theta_1, \theta_2, ..., \theta_J]$, which I will write as $\widetilde{G}_J = \{q, \theta\}$. Note that a vector of tilted mixing weights $\widetilde{q}$ is a function of β and $\{q, \theta\}$; for ease of exposition, I will write this relationship as $\widetilde{q} = \widetilde{q}(\theta, \beta, q)$. The likelihood function of the J-truncated-MR-DPM model is then expressed as

$$p(x_{1:N}|\theta, q, \beta) = \prod_{i=1}^{N} \left(\sum_{j=1}^{J} \widetilde{q}_j(\theta, q, \beta) k(x_i; \theta_j) \right). \tag{13}$$

And the model will be completed below by specifying the kernel function (Section 2.3) and the prior distributions of the unknown parameters (Section 4), as discussed.

The stick-breaking truncation to approximate the DPM model is often used to construct an efficient posterior sampler (e.g., Ishwaran and James, 2001). One can view the truncated MR-DPM model as an approximation to the original MR-DPM model and can thus expect the quality of the approximation to improve as the truncation order increases. In a later section, I will discuss how, instead of simply fixing the truncation order at an arbitrary number, I use an adaptive algorithm to select the truncation order that avoids large approaximation errors. Another view of the model with J-truncation is the following: essentially, the semi-parametric prior with the J-truncated DP imposes a prior distribution over the space of finite mixtures with at most J mixtures, where each element in this set satisfies the moment restrictions. This implies that one assumes that the true data generating process follows the finite mixture model with J^* mixtures, where J^* is treated as unknown but bounded by some known finite integer J. In any case, the implied random densities satisfy the moment restriction regardless of the truncation order because the exponential tilting projection is applied after the truncation.

2.3 Choice of kernel functions

For continuous data, I will make extensive use of the (multivariate) normal density as a kernel function:

$$k(x_i;\ \theta_j) = \frac{1}{\sqrt{2\pi^d |\Sigma_j|}} \exp\left(-\frac{1}{2}(x_i - \mu_j)'\Sigma_j^{-1}(x_i - \mu_j)\right), \tag{14}$$

where $\theta_j = \{\mu_j, \Sigma_j\}$. Natural choices for the base measures are the normal distribution for μ_j and the inverse Wishart distribution for Σ_j, respectively. That is,

$$G_0(\mu, \Sigma) =_d N(\mu;\ m, B)\ IW(\Sigma;\ s, sS)$$

where m, B, s, and S are the associated hyperparameters. Other choices of parametric densities are also possible.

For categorical data, I use the multinomial kernel function in later sections,

$$k(x_i;\ \theta_j) = p_{1,j}^{x_{i,1}} \cdot p_{2,j}^{x_{i,2}} \cdots p_{M,j}^{x_{i,m}}, \quad \sum_{m=1}^{M_x} p_{m,j} = 1, \quad p_{m,j} > 0, \tag{15}$$

where $\theta_j = [p_{1,j}, p_{2,j}, ..., p_{m,j}]$, and M_x is the number of possible outcomes for x_i, and $x_{i,l} = 1$ if $x_i = l$ and 0 otherwise. In this case, a natural choice for the base measure is the Dirichlet distribution,

$$G_0(p) =_d Dir(p; [\alpha_1^p, ..., \alpha_m^p]')$$

where $p = [p_1, ..., p_J]'$ is a vector of multinomial probability parameters and $\alpha^p = [\alpha_1, ..., \alpha_m]'$ is the parameter for the Dirichlet distribution. Negative binomial and Poisson distributions are alternative choices for modeling categorical variables.

It is also possible to model data where continuous ($x_{i,c}$) and categorical ($x_{i,d}$) variables are mixed. Below, I impose the following structure:

$$k(x_{i,c}, x_{i,d};\ \theta_j) = f_N(x_{i,c}; \mu_j, \Sigma_j) f_{MN}(x_{i,d};\ p_j)$$

where f_N is a density function for the normal distribution (Equation 14) and f_{MN} is a probability mass function for the multinomial distribution (Equation 15). Note that even though this kernel function assumes independence between $x_{i,c}$ and $x_{i,d}$ given the mixture parameter θ_j, the resulting random density can ultimately have dependency through mixture probabilities. There are alternative approaches that explicitly model the joint distribution of

$x_{i,c}$ and $x_{i,d}$. One such possible specification is to break down the joint kernel into conditional and marginal kernels:

$$k(x_{i,c}, x_{i,d}; \theta_j) = Pr(x_{i,d}|x_{i,c}, \theta_j)\ k(x_{i,c};\ \theta_j).$$

Then, a linear logistic/probit-type form can be used for the first term when $x_{i,d}$ is binary data and the multivariate normal density can be used for the marginal density for $x_{i,c}$. More discussion can be found in Taddy (2008).

3 Extensions

The MR-DPM modeling framework is flexible enough to cover more complicated data densities. In this section, I introduce two extensions to *i.i.d.* MR-DPM models. The first extension is to introduce time-dependence to the unknown data density. This can be used to estimate stationary time series data with moment restrictions via specific choices of the kernel functions in Equation 2. The second extension is to incorporate exogenous dynamic latent variables.

3.1 MR-DPM model with time-series data

Now suppose that the observation vector $\{x_t\}_{t=1}^T$ exhibits serial dependence with a p-th order time-homogeneous transition density. Moreover, assume the moment conditions dependent on the history of the data through time $t-p$, $E_P[g(x_t, ..., x_{t-p}, \beta)] = 0$. In this case, the mixture model in Equation 2 is not valid as it does not consider the dependence structure of the data. Instead, I consider the following mixture density for the transition density of the underlying data generating process,

$$\begin{aligned} p(x_t|x_{t-1}, ..., x_{t-p}, G) &= \frac{\int k(x_t, x_{t-1}, ..., x_{t-p}; \theta) dG(\theta)}{\int k_M(x_{t-1}, ..., x_{t-p}; \theta) dG(\theta)} \\ G|\alpha, \psi \sim DP(\alpha, G_0), &\quad G_0 = G_0(\cdot|\psi), \end{aligned} \tag{16}$$

where the kernel function in the denominator is obtained by marginalizing x_t out of the joint kernel function $k(x_t, ..., x_{t-p}; \theta)$,

$$k_M(x_{t-1}, ..., x_{t-p}; \theta) = \int k(x_t, x_{t-1}, ..., x_{t-p}; \theta) dx_t.$$

Note that mixture model defined in Equation 2 is a special case of this formulation with $p = 0$.

One can view this nonparametric prior as a variant of the Dirichlet process mixture prior with dependent weights, where the mixing weights are functions of the realizations of observations. To see this, write Equation 16 as:

$$p(x_t|\bar{x}_{t-1}, G) = \sum_{j=1}^{\infty} w_j(\bar{x}_{t-1}; q, \theta)\ \frac{k(x_t, \bar{x}_{t-1}|\theta)}{k_M(\bar{x}_{t-1}|\theta)}, \tag{17}$$

where the mixing weights depend on $\bar{x}_{t-1} = [x_{t-1}, ..., x_{t-p}]$:

$$w_j(\bar{x}_{t-1}) = \frac{q_j k_M(\bar{x}_{t-1}; \theta_j)}{\sum_{m=1}^{\infty} q_m k_M(\bar{x}_{t-1}; \theta_m)}, \tag{18}$$

where $q = \{q_j\}_{j=1}^{\infty}$ and $\theta = \{\theta_j\}_{j=1}^{\infty}$ are the parameters in the stick-breaking representation of the Dirichlet process (Equation 9)

With this formulation the transition and stationary distributions are modeled as nonparametric mixtures, which is an important asset for the exponential tilting procedure since the moment conditions can still be written in the same form as in the *i.i.d.* case:

$$E_P[g(x_t, \bar{x}_{t-1}, \beta)] = \iiint g(x_t, \bar{x}_{t-1}, \beta) k(x_t, \bar{x}_{t-1}; \theta)\ dG(\theta) dx_t d\bar{x}_{t-1}.$$

It is important to note that once the DPM type mixture model is imposed on the unconditional joint distribution of $x_{1:p}$, the exponential tilting procedure in the previous section goes through without any major changes. The likelihood function of the J-truncated MR-DPM model for time-dependent data can be written using the predictive likelihood decomposition:

$$\begin{aligned} p(x_{1:T}|\theta, \beta, q) &= \prod_{t=1}^{T} p(x_t|\bar{x}_{t-1}, \theta, \beta, q) \\ &= \prod_{t=1}^{T} \left(\sum_{j=1}^{J} \frac{\widetilde{q}_j(\theta, \beta, q) k_M(\bar{x}_{t-1}; \theta_j)}{\sum_{m=1}^{J} \widetilde{q}_m(\theta, \beta, q) k_M(\bar{x}_{t-1}; \theta_m)} k(x_t, \bar{x}_{t-1}; \theta_j) \right). \end{aligned} \tag{19}$$

where the tilted probabilities $\widetilde{q}_j$ are obtained by the exponential projection procedure in Equation 12 with the integrated moment condition given by

$$\widetilde{g}(\beta, \theta) = \int g(x_t, \bar{x}_{t-1}, \beta) k(x_t, \bar{x}_{t-1}; \theta) dx_t d\bar{x}_{t-1}.$$

Discussion. Because this modelling approach allows me to estimate the transition density as well as the parameters in the moment restrictions, I can make density predictions in a straightforward manner. The predictive density for x_{T+1} at time T, given unknown parameters, is

$$p(x_{T+1}|x_T, \bar{x}_{T-1}, \theta, \beta, q) = \sum_{j=1}^{J} \frac{\widetilde{q}_j(\theta, \beta, q) k_M(\bar{x}_{T-1}; \theta_j)}{\sum_{m=1}^{J} \widetilde{q}_m(\theta, \beta, q) k_M(\bar{x}_{T-1}; \theta_m)} k(x_T, \bar{x}_{T-1}; \theta_j).$$

The simulation-based approximation to the posterior predictive distribution can be obtained as soon as one can generate draws from the posterior distribution of unknown parameters using the above formula. Density prediction here shares a spirit similar to that in Robertson et al. (2005) and Giacomini and Ragusa (2014), who construct a density prediction by finding a probability distribution that minimizes the Kullback-Leibler divergence between the density prediction based on reduced-form parametric models such as a vector autoregression subject to the moment restrictions. However, their approaches are different from that in this paper in that they use moment restrictions only after estimation, while that in this paper imposes moment restrictions *a priori*. Moreover, they either calibrate or estimate unknown parameters in the moment restrictions separately, while the method in this paper is designed to estimate both the transition density and parameters in the moment restrictions jointly.

This modeling approach to time-dependent data requires users to set the order of dependence, p, *ex ante*. One can select the order based on model selection criteria, such as the density or the predictive score, which are discussed in a later section.

The time-dependent nonparametric prior in Equation 16 is studied by Antoniano-Villalobos and Walker (2014) and Griffin (2014), but neither of them considers this in the context of moment restriction models. There are of course other nonparametric priors for modeling time-dependence. Among these are the generalized polya urn of Caron et al. (2007), the probit stick-breaking process of Rodríguez and Dunson (2011), the order-based dependent Dirichlet process of Griffin and Steel (2006), the autoregressive Beta (BAR) stick-breaking process of Taddy (2010), and the stick-breaking autoregressive process of Griffin and Steel (2011). All of these priors admit a formulation similar to that of the infinite sum in Equation 17 and most of them (save the first) introduce time-dependence through the mixing weights. These priors can also be considered in our framework, but I found that the nonparametric prior in Equation 16 is the most useful, as it directly models the stationary distribution of the time-series, which is important for the exponential tilting procedure.

3.2 MR-DPM model with latent variables

Consider the moment condition model with a latent variable,

$$E_P[g(x_t, z_t, \beta)] = 0, \quad z_t \sim p(z_t|z_{t-1}, \beta_z) \tag{20}$$

where z_t is a latent variable with a transition distribution known up to a finite dimensional parameter β_z. In this framework, the unspecified part of the underlying distribution is a conditional distribution of x_t given z_t and the other unknown paramters β and β_z. I model this using the MR-DPM modeling strategy. That is, the unknown conditional density is modeled as a mixture of parametric densities,

$$p(x_t|z_t, \theta, \beta, \beta_z, q) = \sum_{j=1}^{J} \widetilde{q}_j(\theta, \beta, \beta_z, q) k(x_t; h(z_t, \theta_j))$$

where $h(z_t, \theta_j)$ is a function that returns the parameters in the kernel function $k(\cdot)$ which captures the dependency of the distribution of x_t on z_t and θ_j. The tilted probabilities $\widetilde{q}_j$ are obtained by the exponential projection procedure in Equation 12 with the integrated moment condition,

$$\begin{aligned}\widetilde{g}(\theta_j, \beta, \beta_z) &= \iint g(x_t, z_t, \beta) k(x_t, z_t; \theta_j, \beta_z) dx_t dz_t \\ &= \iint g(x_t, z_t, \beta) k(x_t|z_t; \theta_j) p(z_t; \beta_z) dx_t dz_t.\end{aligned}$$

where $p(z_t; \beta_z)$ is the unconditional probability density of the latent variable which corresponds to the transition density specified in Equation 20.

Econometric approaches to moment condition models with this type of dynamic latent variables are quite new to the literature (e.g., Gallant et al., 2014). Bayesian estimation of the unknown latent variables is relatively easier than frequentist methods because the Bayesian approach treats the unknown latent variable in the same way as unknown finite-dimensional parameters. As can be seen in a later section, the estimation of the model proceeds in the following Gibbs-type iterative algorithm. First, conditional on a sequence of latent variables, estimation of other parameters is the same as previous cases since the conditional likelihood function is,

$$\begin{aligned} p(x_{1:T}|z_{0:T},\theta,\beta,q) &= \prod_{t=1}^{T} p(x_t|z_t,\theta,\beta,q) \\ &= \prod_{t=1}^{T}\left(\sum_{j=1}^{J}\widetilde{q}_j(\theta,\beta,q)k(x_t;h(z_t,\theta_j))\right), \end{aligned} \tag{21}$$

which is essentially the same likelihood function as in the usual MR-DPM model. Second, conditional on the other unknown parameters, estimation of a sequence of latent variables can be done by the single-move Gibbs algorithm (Jacquier et al., 2002). For example, the conditional posterior distribution of z_t is proportional to

$$p(z_t|z_{t-1},\beta_z)p(x_t|z_t,\theta,q,\beta,\beta_z)p(z_{t+1}|z_t,\beta_z), \quad \text{for } 1\leq t\leq T-1,$$

where the first and third terms are given by the transition density of z_t and the second term is given by Equation 21.

Example of latent process and $h(z_t,\theta_j)$. In a later section, I model z_t to be univariate $AR(1)$ with a Gaussian shock,

$$z_t = c_z + \rho_z z_{t-1} + \sigma_z e_{z,t}, \quad e_{z,t}\sim N(0,1),$$

and use the normal density function as a kernel function. The natural choice of the function $h(z_t,\theta_j) = [h_\mu(z_t,\theta_j), h_\Sigma(z_t,\theta_j)]$ is

$$h_\mu(z_t,\theta_j) = \mu_{x,j} + \Sigma_{xz,j}\Sigma_{zz}^{-1}(z_t-\mu_z) \quad \text{and} \quad h_\Sigma(z_t,\theta_j) = \Sigma_{xx,j} - \Sigma_{xz,j}\Sigma_{zz}^{-1}\Sigma'_{xz,j},$$

where $\theta_j = (\mu_{x,j},\Sigma_{xz,j},\Sigma_{xx,j})$, μ_z and Σ_z are the unconditional mean and variance of z_t,

$$\mu_z = \frac{c_z}{1-\rho_z} \quad \text{and} \quad \Sigma_{zz} = \frac{\sigma_z^2}{1-\rho_z^2}.$$

The base measure can be chosen analogously to the *i.i.d.* MR-DPM model,

$$G_0(\theta) = N(\mu_x;\ m_z, V_z)\ IG(\Sigma_{xx};\ s, S)\ N(\Sigma_{xz};\ m_{xz}, V_{mz}).$$

where $(m_z, V_z, s, S, m_{xz}, V_{mz})$ are the associated hyperparameters.

4 Prior specification

In this section, I discuss prior distributions for the unknown parameters in the J-truncated MR-DPM model. The J-truncated MR-DPM model contains the following parameters: $(\theta, q, \beta, \alpha, \psi)$. The model assumes that a collection of θ and q is drawn from the J-truncated Dirichlet process, $G_J = (\theta_{1:J}, q_{1:J}) \sim DP_J(\alpha, G_0(\psi))$ where θ_j is the parameter in the j-th mixture density (kernel function) and q is a vector of pre-exponential-tilting mixture probabilities constructed via the stick-breaking formulation with independent Beta draws $V_{1:J}$ (Equation 9).[4] α is a concentration parameter in the truncated Dirichlet process and ψ is a hyperparameter in the base measure G_0. β is a parameter in the overarching moment condition model. I consider prior distributions in the partially separable form

$$p(\theta|\psi)p(\psi)p(V|\alpha)p(\alpha)p(\beta).$$

This section starts with a discussion of the effects of the exponential tilting projection with a generic prior distribution. Owing to the exponential tilting projection step, the domain of the prior distribution is restricted. I discuss how to analyze resulting prior distributions. Then, I describe how one can form a prior distribution under the multivariate normal kernel function, which I will use extensively in later sections. I also discuss forming priors on hyperparameters such as ψ and α.

4.1 Implied prior distribution

Restricted domain. For each parameter, I begin with some parametric distribution without any restriction on its domain. However, owing to the exponential tilting procedure, the domain of the implied full joint prior distribution will be restricted. In fact, the support of the prior is meaningful only when the solution to the exponential tilting problem in Equation 12 has a solution. Note that the solution exists and is unique when 1) the interior of the convex hull of $\cup_j\{\widetilde{g}(\theta_j, \beta)\}$ contains the origin; and 2) the objective function in Equation 12 is bounded. This leads to the following implied joint prior distribution:

$$p(\theta, \beta, V, \alpha, \psi) \propto p(\theta|\psi)p(\psi)p(V|\alpha)p(\alpha)p(\beta)I(\vec{0} \in H(\theta, \beta)),$$

where $I(\cdot)$ is an indicator function that takes the value 1 if the condition inside of the parentheses holds and 0 otherwise, and the set $H(\theta, \beta)$ denotes the interior of the convex

[4] The posterior sampling algorithm will draw $V_{1:J}$ instead of $q_{1:J}$. Knowning $V_{1:J}$ is equivalent to knowning $q_{1:J}$ because the stick-breaking formulation is deterministic transformation of $V_{1:J}$.

hull of $\cup_j \{\widetilde{g}(\theta_j, \beta)\}$. The indicator function reflects the fact that the resulting joint prior distribution puts positive probability only on the set where the solution to the exponential projection exists. Hence, the marginal prior distribution implied by this distribution is different from the unrestricted beginning prior distribution. I will refer to the former as an "implied prior" and the latter as an "initial prior." For later sections, I will denote $S_{(\theta,\beta)}$ as the domain of the joint prior distribution of θ and β,

$$S_{(\theta,\beta)} = \{\theta, \beta : \vec{0} \in H(\theta, \beta)\} \cap Supp(\text{initial prior for } \theta, \beta).$$

Note that the set $S_{(\theta,\beta)}$ excludes a pair (θ, β) that does not have a solution to the exponential tilting procedure from the joint support of the initial prior distribution for θ and β.

Analyzing the implied prior. It is worth noting that the prior specification for the mixture component parameters θ affects the implied prior for β through the convex hull condition restriction, since not all β can satisfy the convex hull condition given every realization of θ. For example, with the univariate normal kernel function, $k(x_i; \theta_j, \sigma^2)$, and the location moment restriction, $E[x_i - \beta] = 0$, the integrated moment condition can be written as

$$\widetilde{g}(\theta, \beta) = 0 \iff \sum_{j=1}^{J} (\theta_j - \beta)\widetilde{q}_j = 0.$$

Suppose that the prior distribution for θ is chosen so that the realizations of θ_j are all concentrated around some negative number. Then, whenever the realizations of θ_j are negative for all $j = 1, ..., J$, no positive realization of β will satisfy the convex hull condition. This implies that even if the initial prior distribution for β puts large probability in a positive region of β, the implied prior distribution of β will put only small probability in this region. In the extreme, the initial prior for θ_j puts zero probability on positive values and the initial prior for β puts all probability on positive values. In this case, the restricted domain $S_{(\theta,\beta)}$ becomes the empty set.

When the moment conditions are over-identified, the problem gets more complicated and it is hard to analyze the implied prior distributions analytically. However, it is always possible to check the shape and domain of the implied prior distributions by simulating draws from the implied prior distribution. Straightforward method is accept-reject sampling, where the draws that do not satisfy the convex hull condition are discarded. Then, one can analyze the implied prior distribution based on these draws. The following algorithm generates M draws from the implied prior distribution. Since I will refer to this process frequently in subsequent sections, I here formally outline an algorithm to generate M draws from the implied prior distribution.

Algorithm 1. *(Accept/Reject algorithm for the implied prior) Enter the following loop with* $i = 1$.

1. *Draw a full parameter* $(\theta^{(i)}, \beta^{(i)}, V^{(i)}, \alpha^{(i)}, \psi^{(i)})$ *from the prior distribution for the J-truncated MR-DPM model,* $p(\theta|\psi)p(\psi)p(V|\alpha)p(\alpha)p(\beta)$.
2. *If the convex hull of* $\cup_j \{\tilde{g}(\beta^{(i)}, \theta_j^{(i)})\}$ *contains the origin, then keep* $(\theta^{(i)}, \beta^{(i)}, V^{(i)}, \alpha^{(i)}, \psi^{(i)})$ *and set* $i = i + 1$. *Otherwise, discard the current draw.*
3. *If* $i = M$, *then exit the algorithm. Otherwise, go to step 1.*

There are at least two ways of detecting violation of the convex hull condition (step 2 in Algorithm 1). The first method is to compute the convex hull of $\cup_j \{\tilde{g}(\beta, \theta_j)\}$ for each draw and check for containment of the origin. The second method is to perform numerical optimization to solve the exponential tilting projection in Equation 11 with a prior draw in hand. Then, discard a draw if the norm of the moment condition evaluated at the minimizer for that draw is larger than some pre-specified small number.[5] The rationale behind the second method is that if the prior draw violates the convex hull condition, then the corresponding integrated moment condition will never be satisfied with this draw at the optimum obtained by the numerical optimizer. I use the second method in this paper because the first method requires additional computation to compute the convex hull.

4.2 Prior specification with a normal kernel function: *i.i.d.* case

In later sections, I extensively use the (multivariate) normal density as a kernel function for the *i.i.d.* MR-DPM model as well as the dynamic MR-DPM models and models with latent variables. Hence, I describe here how one can choose prior values for such applications. The multivariate normal kernel function is in the following form:

$$k(x_i;\ \theta_j) =_d N(x_i;\ \mu_j, \Sigma_j) \quad \text{for the } j\text{th component,}$$

[5] In this paper, I use a variant of the Newton method for numerical optimization with the maximum number of iterations set to be 50–200 (depending on the application) and the tolerance for the norm of the moment condition to be 10^{-7} . Since the exponential tilting projection in this paper is a convex problem, the convergence is very fast if the solution exists (less than 20 iterations for the applications considered in this paper).

where μ_j is a mean vector and Σ_j is a covariance matrix and parameter θ is a vector that collects all pairs $\theta_j = (\mu_j, \Sigma_j)$ for $j = 1, ..., J$. Then, the natural choice for the base measure G_0 in the Dirichlet mixture process is a multivariate normal for the location component and inverse-Wishart for the scale,

$$G_0(\mu_j, \Sigma_j|\psi) =_d N(\mu_j; m, B) IW(\Sigma_j; s, S)$$

where the parameter ψ collects hyperparameters m, B, s and S. Following the literature, I impose prior distributions on the hyperparameters, m, B, and S,

$$m|B \sim N(m;\ a, B/\kappa) \quad B \sim IW(B;\ \nu, \Lambda), \quad S \sim W(S; q, q^{-1}R). \tag{22}$$

which facilitate computation owing to conjugacy. This additional hierarchical structure provides more flexibility in modelling underlying data density (Müller et al., 1996).

It is desirable that the prior distribution of the location component μ_i be centered around the data while also sufficiently spread out so that it can cover a sizable range of the data. For the rest of paper, I choose a so that μ_j is centered around the mean of the data. Then, I reparameterize the scale matrix Λ in the prior distribution for B as

$$\Lambda = \bar{\lambda} \times diag(cov(X)),$$

where $diag(\cdot)$ is an operator that returns a diagonal matrix with diagonal entries equal to those of the argument, $\bar{\lambda}$ is a positive scalar, and $cov(X)$ is the covariance of the data. Then, the rest of the parameters $(\kappa, \nu, \bar{\lambda})$ are chosen so that the implied prior distribution for the diagonal elements of B/κ can take from a small value (one-tenth of the data variance) to a large value (four times of data variance) with high probability. This ensures that the location components μ_j are distributed around the realized data.[6]

The scale parameter Σ_j plays a role similar to that of bandwidth in frequentists' kernel density estimation, as it governs the smoothness of the underlying density. The larger the diagonal elements in Σ_j, the smoother the density is. Usually, the smoothness of the underlying data density is unknown, and therefore, the prior distribution for Σ_j should cover a wide range of values. As for the location components, it is useful to think about the reasonable range of values in terms of the scale of the data. So, I reparameterize the shape

[6] Such priors are also considered by Ishwaran and James (2002) and Conley et al. (2008) for DPM-based models.

parameter R in the prior distribution for S as

$$R = \bar{r} \times diag(cov(X)).$$

where $\bar{r}$ is a positive real number. Then I choose, s, q and $\bar{r}$ so that diagonal elements of Σ_j are contained between one-tenth and four times the variance of the data with high probability.

The prior distribution for the parameter β in the moment conditions can be flexibly chosen depending on the context. When the moment conditions are implied by economic theory, one might have access to some prior restrictions (For example, the risk aversion parameter in the CRRA utility function has to be positive.).

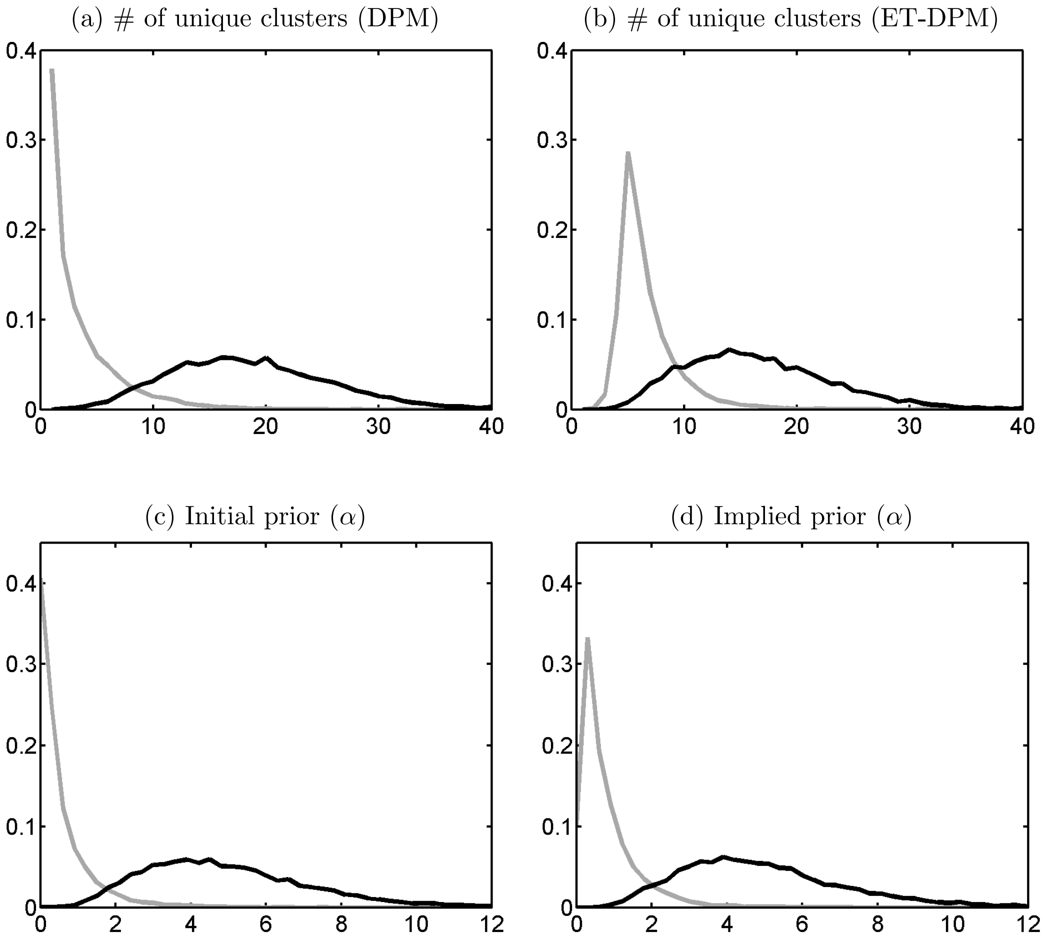
Lastly, the concentration parameter α in the Dirichlet mixture process is assumed to have Gamma density, $\alpha \sim Gamma(a_\alpha, b_\alpha)$. In the usual DPM model, the concentration parameter α is related to the number of unique clusters in the mixture density. Specifically, Antoniak (1974) derived the relationship between α and the number of unique clusters,

$$E[n^*|\alpha] \approx \alpha \log\left(\frac{\alpha + N}{\alpha}\right) \quad \text{and} \quad Var(n^*|\alpha) \approx \alpha \left\{\log\left(\frac{\alpha + N}{\alpha}\right) - 1\right\}.$$

That is, the expected number of unique clusters (n^*) is increasing in α and the number of observations N. This relationship roughly holds for the semiparametric prior considered in this paper as well. However, parameter restrictions given by the convex hull condition complicate the relationship for which there is thus no closed form. Instead, I recommend checking the effect of the prior on α using draws from the implied prior distribution created with Algorithm 1.

Figure 1 illustrates the relationship between α and the number of unique clusters implied by the prior choice for α. I generate 10,000 prior draws from the MR-DPM prior (using Algorithm 1) and the DPM prior (without moment restrictions) with high and low means for the initial prior distribution for α. For the MR-DPM prior, I use instrumental variable regression moment restrictions that will be revisited in the upcoming simulation section. Prior specifications for other parts of the prior are the same as those used later. I focus on the effect of the initial prior distribution for α. Panel (a) shows a histogram of the number of clusters for the DPM prior based on two prior specifications for α with low (grey) and high (black) mean. As expected, all other things equal, the high α leads to more clusters. This is also true for the MR-DPM prior, as can be seen in Panel (b). However, the MR-DPM prior does not put prior probability on small numbers of clusters owning to the convex

Figure 1 Effect of prior distribution for α



Note: The grey line corresponds to a low mean prior for α and the black line corresponds to a high mean α prior. All figures are histograms based on 10,000 draws from the prior distribution described in the main text.

hull condition.[7] This effect also can be seen from the implied prior for α (low mean prior). Compared to the initial prior distribution, the implied prior tends to put less probability on small α.

[7] The semiparametric prior in this paper does not admit very small numbers of clusters because the probability of satisfying the convex hull condition decreases as the number of clusters decreases. For example, when only one cluster is allowed with the location moment restriction, $E[x_i - \beta] = 0$, the prior draw β and θ_1 from the initial prior distribution must coincide to satisfy the convex hull condition, but this is a probability zero event under the initial prior specification described above.

4.3 Prior specification with a normal kernel function: Beyond the *i.i.d.* case

Since the time-series MR-DPM model shares the unknown parameters with the standard *i.i.d.* MR-DPM model, the prior distributions imposed on the unknown parameters are very similar. For models with latent variables, there are two additional unknown parameters vis-à-vis *i.i.d.* MR-DPM models. The first additional unknown is a vector of latent variables, $z_{0:T}$ whose transition probability is known up to finite dimensional parameter β_z. I impose a prior distribution on the initial value for the latent variable, z_0. The second additional unknown is β_z. As this parameter is finite dimensional, one can impose a parametric proper prior distribution.

For example, if the latent variable z_t is assumed to follow the univariate $AR(1)$ with a Gaussian shock,

$$z_t = c_z + \rho_z z_{t-1} + e_{z,t}, \quad e_{z,t} \sim N(0,1),$$

one needs to impose prior distributions on c_z, ρ_z, and z_0. In the subsequent application, I will impose

$$c_z \sim N(m_{c_z}, V_{c_z}), \quad \rho_z \sim N(m_{\rho_z}, V_{\rho_z}), \quad z_0 \sim N\left(0, \frac{1}{1-\rho_z^2}\right).$$

Under the normal kernel function for the conditional likelihood function (Equation 21), I will impose prior distributions on the other parameters, $(\theta, q, \beta, \alpha, \psi)$, *a la* the *i.i.d.* case.

5 Posterior analysis

The goal of this section is to develop a series of methods that allow for the analysis of posterior distributions derived from the J-truncated MR-DPM model presented in Section 2 with priors as specified in Section 4 and the joint posterior distribution of model parameters defined as

$$p(\theta, \beta, V, \alpha, \psi | X) \propto p(X|\theta, \beta, V) p(\theta|\psi) p(V|\alpha) p(\beta) I(\vec{0} \in H(\theta, \beta)) p(\psi) p(\alpha), \tag{23}$$

where $p(X|\theta, \beta, V)$ is the likelihood function given by Equation 13 for an *i.i.d.* model and Equation 19 for a time-dependent model.[8] I denote X as $x_{1:N}$ or $x_{1:T}$ depending on the

[8] The posterior sampling algorithm will draw $V_{1:J}$ instead of $q_{1:J}$. Knowning $V_{1:J}$ is equivalent to knowning $q_{1:J}$ because the stick-breaking formulation is deterministic transformation of $V_{1:J}$.

context. For models with latent variables, I splice β_Z and $z_{0:T}$ into the vector of unknown parameters and consider the posterior distribution for the augmented unknown parameter vector, $p(\theta, \beta, V, \alpha, \psi, \beta_z, z_{0:T}|X)$ based on the likelihood function given in Equation 21.

More specifically, I study three simulation-based posterior samplers that generate samples from the posterior distribution which can be used to approximate the posterior moments of a function of model parameters. This includes posterior moments of a function of parameters in the moment condition,

$$E[h(\beta)|X] = \int h(\beta)\pi(\beta)d\beta,$$

where I denote $\pi(\beta)$ as the marginal posterior distribution of β and $h(\cdot)$ is some function that will be clarified later. For example, if $h(\beta) = \beta$, the above quantity is simply a posterior mean of β. Since the model specifies the underlying data generating process explicitly, it is also possible to approximate posterior moments of functionals of the underlying distribution such as the posterior mean for the data density $f(x_0;\ \widetilde{G}_J)$ at the point x_0,

$$E[f(x_0;\ \widetilde{G}_J)|X] = \int \sum_{j=1}^{J} \widetilde{q}_j(\theta, \beta, V)k(x_0;\ \theta_j)\pi(\theta, \beta, V)d(\theta, \beta, V)$$

where $\pi(\theta, \beta, V)$ is the marginal posterior distribution.

Another important quantity of interest is the marginal likelihood,

$$p(X) = \int p(X|\varphi)p(\varphi)d\varphi,$$

where $\varphi = (\theta, \beta, V, \alpha, \psi)$. The marginal likelihood plays an important role in Bayesian analysis as it can be used to compute the posterior model probability. This, in turn, can be used to obtain the Bayes factor between two competing models for model selection. In addition, the posterior model probability can be used to compute weights for model averaging.

The rest of the section is organized as follows. First, I introduce the basic posterior sampler based on the Metropolis-within-Gibbs algorithm and provide conditions under which this sampler converges to the true posterior distribution as the number of simulations increases. Second, I introduce a modified version of the basic sampler using a data-augmentation method that improves the mixing properties of the posterior sampler, as well as computation time. Third, I discuss sequential Monte Carlo methods that can be used to compute the marginal data density and select the truncation order adaptively.

5.1 Basic sampler and its convergence

The posterior sampler that I introduce in this section will be called the basic sampler. The sampler is based on the Metropolis-within-Gibbs algorithm, which cycles over each parameter of the block $\varphi = (\theta, \beta, V, \alpha, \psi)$ in order; a sequence of draws from this algorithm defines a Markov chain with a transition kernel $K_B(\varphi^*|\varphi^0)$ on the product set $D = S_{(\theta,\beta)} \times (0,1)^J \times R^+ \times supp(\psi)$, where $supp(\psi)$ is the domain of the prior distribution for the hyperparameters.

Algorithm 2. Basic sampler for the J-truncated MR-DPM model. *Enter the following steps with $(\theta^0, \beta^0, V^0, \alpha^0, \psi^0) \in D$ and $i = 1$:*

1. *Draw θ_j^* from $p(\theta_j|\theta_{1:j-1}^*, \theta_{j:J}^0, \beta^0, V^0, \alpha^0, \psi^0, X)$, for $j = 1, ..., J$.*
2. *Draw β^* from $p(\beta|\theta^*, \beta^0, V^0, \alpha^0, \psi^0, X)$*
3. *Draw V_j^* from $p(V_j|\theta^*, \beta^*, V_{1:j-1}^*, V_{j:J}^0, \alpha^0, \psi^0, X)$, for $j = 1, ..., J$.*
4. *Draw α^* from $p(\alpha|\theta^*, \beta^*, V^*, \alpha^0, \psi^0, X)$*
5. *Draw ψ^* from $p(\psi|\theta^*, \beta^*, V^*, \alpha^*, \psi^0, X)$*
6. *Store $(\theta^i, V^i, \beta^i, \alpha^i, \psi^i) = (\theta^*, V^*, \beta^*, \alpha^*, \psi^*)$. Stop if $i = N_s$; otherwise, set*

$$(\theta^0, \beta^0, V^0, \alpha^0, \psi^0) = (\theta^*, \beta^*, V^*, \alpha^*, \psi^*)$$

   and go to step 1 with $i = i + 1$.

Under the multivariate normal kernel and prior specification described in the previous section, a closed-form conditional posterior distribution for α and ψ is possible. However, the conditional posterior distributions for θ, β, and V are not well-known parametric distributions and are complicated. Instead, I use the random-walk Metropolis-Hastings (RWMH) algorithm to draw θ, β, and V from the conditional posteriors. The step for parameters in the mixture kernel function θ depends on the choice of the kernel function and can be further decomposed into smaller blocks. In the case of the multivariate normal kernel function, I decompose θ_j into its mean and variance-covariance matrix, $\theta_j = (\mu_j, \Sigma_j)$ for each $j = 1, ..., J$ and update μ_j and Σ_j separately. The variances of each RWMH proposal densities are adaptively chosen following Atchadé and Rosenthal (2005) so that the resulting acceptance rates are about 30%. A detailed derivation of the posterior sampler is presented in the Appendix.

When there are latent variables in the model, I add RWMH steps for β_z and $z_{0:T}$ to the previous algorithm. The conditional posterior distributions to update these parameters are model-specific and can be different depending on the relationship between the observed data and the latent variables. The following algorithm is based on the example described in Section 2, where z_t follows an $AR(1)$ process and the distribution of x_t depends only on z_t.

Algorithm 3. Basic sampler for the J-truncated MR-DPM model with latent variables. *Enter the following steps with* $(\theta^0, \beta^0, V^0, \alpha^0, \psi^0, \beta_z^0, z_{0:T}^0)$ *and* $i = 1$*:*

1. *Draw* $(\theta^*, \beta^*, V^*, \alpha^*, \psi^*)$ *based on steps 1 – 5 of Algorithm 2 with the likelihood function defined in Equation 21 and prior distributions described in Section 4.*
2. *Draw* β_z^* *from* $p(\beta_z|\theta^*, \beta^*, V^*, \alpha^*, \psi^*, \beta_z^0, z_{0:T}^0, X)$*.*
3. *Draw* z_0^* *using the conditional posterior* $p(z_1^0|z_0, \beta_z^*)p(z_0|\beta_z^*)$*.*
4. *Draw* z_t^* *using the following conditional posterior:*

$$p(z_t|z_{t-1}^*, \beta_z^*)p(x_t|z_t, \theta^*, V^*, \beta^*)p(z_{t+1}^0|z_t, \beta_z^*), \qquad \textit{for } \ 1 \le t \le T-1.$$

5. *Draw* z_T^* *using the conditional posterior* $p(z_T|z_{T-1}^*, \beta_z^*)p(x_T|z_T, \theta^*, V^*, \beta^*)$*.*
6. *Store* $(\theta^i, \beta^i, V^i, \alpha^i, \psi^i, \beta_z^i, Z_{0:T}^i) = (\theta^*, \beta^*, V^*, \alpha^*, \psi^*, \beta_z^*, z_{0:T}^*)$*. Stop if* $i = N_s$*; otherwise, set*

$$(\theta^0, \beta^0, V^0, \alpha^0, \psi^0, \beta_z^0, z_{0:T}^0) = (\theta^*, \beta^*, V^*, \alpha^*, \psi^*, \beta_z^*, z_{0:T}^*)$$

and go to step 1 with $i = i + 1$*.*

Convergence of the basic sampler. Even though the sampler itself is quite standard in the econometrics literature, the convergence of the posterior sampler in the present environment is not straightforward owing to the exponential projection step. The MR-DPM model restricts the parameter space in such a way that the convex hull condition is satisfied and the resulting support is usually different from that of the initial prior distributions for model parameters. Therefore, it is important to identify under what conditions the Markov chain defined by the basic sampler converges to the posterior distribution. To this end, I make the following high-level assumptions.

Assumption 1. ***(Model)*** *Observations,* $x_{1:N}$*, are generated from the* J*-truncated MR-DPM model with a multivariate normal kernel function, and its likelihood has the form in Equation 13.*

Assumption 2. ***(Prior)*** *The joint prior distribution has the form given in section 4 and therefore* $\psi = (m, B, S)$*, and the initial prior distribution for* β *is proper.*

Assumption 3. ***(Support)*** *The restricted support for θ and β implied by the joint prior distribution, $S_{(\theta,\beta)}$, is nonempty, open and arc-connected.*

Assumption 4. ***(Bounded objective function for exponential tilting)*** *The objective function in the exponential tilting projection, Equation 11, is bounded on D.*

Assumptions 1 and 2 have already been discussed in previous sections. Assumption 3 is related to the restricted support ensuing from the exponential projection procedure. This set is required to be open so that the invariant distribution of the transition kernel (posterior distribution) is lower semi-continuous, which in turn guarantees that the transition kernel is aperiodic. Arc-connectedness is used to show that the transition kernel is irreducible. Intuitively, if the domain of the transition kernel of the Markov chain is not connected, then it is possible that the chain visits only a subsection of the support and thus cannot explore the whole domain of the posterior distribution; Assumption 3 rules out this possibility. Assumption 4, in conjunction with the convex hull condition, guarantees that the solution to the dual problem for the exponential tilting projection exists on the domain of the transition kernel. The conditions in Assumption 3 and Assumption 4 depend on the properties of the moment restrictions and can be verified case by case.

Proposition 1 below establishes the convergence of the basic sampler under the L_1-norm. The second part of the proposition shows that posterior moments of the function of the model parameters can be estimated by a Monte Carlo average using draws from the basic sampler (the strong law of large number). Another important implication of this proposition is that one should start the posterior sampler from the set D, which excludes any θ and β that do not satisfy the convex hull condition. In practice, one can draw initial prameters from the prior distribution using the accept-reject sampling described in Algorithm 1. A proof of Proposition 1 is provided in the Appendix.

Proposition 1. *Let $K_B(\varphi^*|\varphi^0)$ denote the transition density of the Markov chain defined by Algorithm 2 without the adaption of the scale parameter in the RWMH proposal density and let $K_B^{(i)}$ denote the i-th iterate of the kernel. If Assumptions 1–4 hold, then for all $\varphi = (\theta, \beta, V, \alpha, m, B, S)$ in $D = S_{(\theta,\beta)} \times (0,1)^J \times R^+ \times R^d \times R_*^{d^2} \times R_*^{d^2}$ (where $R_*^{d^2} = \{d \times d$ positive definite matrices$\}$), as $N_s \to \infty$:*

1. *$|K_B^{(N_s)} - p(\varphi|X)| \to 0$.*
2. *For real-valued, $p(\varphi|X)$-integrable functions $h(\varphi)$,*

$$\frac{1}{N_s}\sum_{i=1}^{N_s} h(\varphi^{(i)}) \to_{a.s.} \int h(\varphi)p(\varphi|X)d\varphi.$$

5.2 Improving mixing through data augmentation

When the underlying density is modeled with the *i.i.d.* MR-DPM model, one can improve the mixing properties of the posterior sampler using a data-augmentation method. Specifically, I introduce the configuration variable $L = (L_1, L_2, ..., L_N)$ as a missing observation. Each element of the configuration variable can take values in $\{1, ..., J\}$ such that all observations with the same configuration arise from the same distribution,

$$x_i|\theta, L_i \sim k(x_i; \theta_{L_i}) \quad \text{for } i = 1, ..., N,$$

and therefore the complete-data likelihood of the J-truncated MR-DPM model now reads:

$$p(X, L|\theta, \beta, V, \alpha, \psi) = \left(\prod_{i=1}^{N} k(x_i; \theta_{L_i})\right) \underbrace{\left(\prod_{j=1}^{J} \widetilde{q}_j^{M_j}\right)}_{=p(L|\theta,\beta,V)} \tag{24}$$

where $M_j = \#\{i : L_i = j\}$ for $j = 1, ..., J$. Note that the configuration vector L breaks down the J-truncated mixture density into individual kernel densities through θ_{L_i}. The last term depends on $\theta, \beta,$ and V due to the exponential tilting procedure. As for the Bayesian estimation of a finite mixture model, one can estimate the augmented parameter $(L, \theta, \beta, V, \alpha, \psi)$ by sampling from the complete-data posterior distribution,

$$p(L, \theta, \beta, V, \alpha, \psi|X) \propto \left(\prod_{i=1}^{N} k(x_i; \theta_{L_i})\right) p(L|\theta, \beta, V)p(\theta|\psi)p(V|\alpha)p(\beta)I(\vec{0} \in H(\theta, \beta))p(\psi)p(\alpha). \tag{25}$$

The posterior sampling method in Algorithm 4 below iterates over each parameter block, $(L, \varphi) = (L, \theta, \beta, V, \alpha, \psi)$ and defines a Markov chain with transition kernel on

$$D = \{perm([1, ..., J]')\} \times S_{(\theta,\beta)} \times (0, 1)^J \times R^+ \times supp(\psi),$$

where $\{perm([1, ..., J])\}$ is a set of all possible permutations of the vector $[1, ..., J]'$.

Algorithm 4. Second posterior sampler for the J-truncated ET-DPM. *Enter the following steps with $(L^0, \theta^0, \beta^0, V^0, \alpha^0, \psi^0) \in D$ and $i = 1$:*

1. *Draw L^* from $p(L|L^0, \theta^0, \beta^0, V^0, \alpha^0, \psi^0, X)$*
2. *Draw θ_j^* from $p(\theta_j|L^*, \theta_{1:j-1}^*, \theta_{j:J}^0, \beta^0, V^0, \alpha^0, \psi^0, X)$, for $j = 1, ..., J$.*
3. *Draw β^* from $p(\beta|L^*, \theta^*, \beta^0, V^0, \alpha^0, \psi^0, X)$*

4. *Draw* V_j^* *from* $p(V_j|L^*, \theta^*, \beta^*, V_{1:j-1}^*, V_{j:J}^0, \alpha^0, \psi^0, X)$, *for* $j = 1, ..., J$.

5. *Draw* α^* *from* $p(\alpha|L^*, \theta^*, \beta^*, V^*, \alpha^0, \psi^0, X)$

6. *Draw* ψ^* *from* $p(\psi|L^*, \theta^*, \beta^*, V^*, \alpha^*, \psi^0, X)$

7. *Store* $(L^i, \theta^i, \beta^i, V^i, \alpha^i, \psi^i) = (L^*, \theta^*, \beta^*, V^*, \alpha^*, \psi^*)$. *Stop if* $i = N_s$, *otherwise set*

$$(L^0, \theta^0, \beta^0, V^0, \alpha^0, \psi^0) = (L^*, \theta^*, \beta^*, V^*, \alpha^*, \psi^*)$$

and go to step 1 with $i = i + 1$.

This algorithm can be viewed as an extension of the Blocked-Gibbs sampler of Ishwaran and James (2001) for the truncated DPM model, which is also based on data augmentation. However, by introducing the moment condition and its related parameters, the full conditional posterior distributions become complicated, and the conditional posteriors for θ, β, and V do not have well-known parametric densities. I utilize the random-walk Metropolis-Hastings algorithm for V and β as in the basic sampler. For θ, I use the independent Metropolis-Hastings algorithm, with which more tailor-made proposal density can be constructed using the structure of the model. It is constructed in such a way that it resembles the conditional posterior distribution without the moment condition. The details of the algorithm are described in the Appendix.

The advantages of this posterior sampler over the basic sampler are the following. First, it reduces computation time by breaking the likelihood into small pieces through the introduction of the configuration variable. Moreover, conditional on the configuration variable, the first term in the complete likelihood function (Equation 24) drops out in many cases, which reduces the total number of evaluations of the kernel function. This is especially beneficial when the number of observations is large, as the computational cost of likelihood function evaluation is linear in the number of observations. Lastly, this algorithm generates less-correlated posterior draws, and therefore, it offers a more efficient approximation of the posterior moments of the object of interest. One of the reasons for this is that the proposal distribution in the Metropolis-Hasting step for θ is constructed in such a way that it resembles the true conditional posterior distribution, while the basic sampler simply employs the random-walk proposal distribution. However, the use of the data-augmentation technique is limited to *i.i.d.* models, as it exploits the particular structure of the likelihood function.

5.3 Posterior analysis via sequential Monte Carlo

In this section, I introduce another type of the posterior sampler based on the sequential Monte Carlo method. The sequential Monte Carlo (or particle filtering) method is a general method to approximate a sequence of multiple distributions of interest by applying importance sampling in an iterative fashion. It has been extensively used to analyze non-linear state-space models (see Doucet and Johansen, 2009, for a review). In addition, Chopin (2002) describes how the sequential Monte Carlo algorithm can be used to obtain the posterior distribution under the static setting where a single such distribution is targeted.

Several researchers have applied the static version of the SMC method (Chopin, 2002) to various models and found that the SMC algorithm can be an attractive alternative to MCMC-based methods. It can perform better than MCMC-based posterior samplers when the posterior distribution exhibits some complicated topography such as multimodality (e.g., Herbst and Schorfheide, 2014). The SMC algorithm is also easily parallelizable vis-à-vis MCMC-based posterior samplers, which reduces computational time significantly (e.g., Durham and Geweke, 2011, 2014).

SMC using the tempered likelihood. The sequential Monte Carlo method applied in this paper approximates a sequence of power posterior distributions indexed by t,

$$\pi_t(\varphi|X) = \frac{1}{C_{\phi_t}(X)}[p(X|\varphi)]^{\phi_t}p(\varphi), \quad t = 1, ..., N_\phi \tag{26}$$

where $\varphi = (\theta, \beta, V, \psi, \alpha)$ and $C_t(X)$ is a normalizing constant for the t-th power posterior. An increasing sequence, $\{\phi_t,\ t = 1, ..., N_\phi\}$, is chosen to satisfy the following

$$0 = \phi_1 < \phi_2 < ... < \phi_{N_\phi - 1} < \phi_{N_\phi} = 1.$$

This sequence is called the tempering or heating schedule. For $t = 1$, the object in Equation 26 is simply a prior distribution and its normalizing constant is one, $C_0(X) = 1$. On the other hand, the power posterior for $t = N_\phi$ is the posterior distribution and its normalizing constant is the marginal likelihood.

In a nutshell, the algorithm recursively obtains the importance sampling approximation to the above power posterior distributions. It starts with draws from the prior distribution of the MR-DPM model and iterates three steps for each t-th power posterior distribution with a set of pairs $\{\varphi_t^i, \widetilde{W}_t^i\}_{i=1}^{N_p}$ that provides a particle approximation[9] to the t-th power posterior distribution $\pi_t(\varphi|X)$. The first step re-weights the particles to reflect the density in iteration t (correction); the second step eliminates degenerated particles by resampling the particles (selection); and the third step propagates the particles forward using a Markov transition kernel $K_t(\varphi_t|\varphi_t'; \zeta_t)$ whose stationary distribution is the t-th intermediate posterior distribution $\pi_t(\varphi)$. I construct the transition kernel based on the MH-within-Gibbs algorithm. Some of the parameter blocks are updated via the random-walk proposal distribution, and parameter vector in the transition kernel ζ_t includes the scaling factor and covariance matrix of those random-walk proposal distributions.

The general form of the algorithm described below is identical to that used in Herbst and Schorfheide (2014). Specifically, I consider the adaptive version of the SMC algorithm where the algorithm computes some tuning parameters during the estimation. The algorithm has two adaptive features. First, the algorithm decides whether to implement re-sampling in the selection step according to effective sample size at every stage. The algorithm performs re-sampling when effective sample size is below $\widehat{\rho} \times N_p$ where $\widehat{\rho} \in [0, 1]$. Second, the algorithm recursively computes the parameters in the transition kernel $\widehat{\zeta}_t = (\widehat{c}_t, \widetilde{\Sigma}_t)$. The scaling factor $\widehat{c}_t$ is computed using the empirical rejection rates from the previous stage to keep the target acceptance rate in the mutation step near some desirable constant. And the covariance matrix of the random-walk proposal distributions $\widetilde{\Sigma}_t$ are computed using the importance sampling approximation to the intermediate distributions at each stage. These adaptive schemes are discussed in detail after I present the algorithm.

Algorithm 5. Simulated tempering SMC for the J-truncated MR-DPM

1. *Initialization.* ($\phi_1 = 0$). *Draw the initial particles from the prior using Algorithm 1,*

$$\varphi_1^i \sim_{i.i.d.} p(\varphi), \quad W_1^i = 1, \quad i = 1, ..., N_p.$$

2. *Recursion. For* $t = 2, ..., N_\phi$,

   (a) *Correction. Reweight the particles from stage* $t-1$ *by defining the incremental and normalized weights*

$$\widetilde{w}_t^i = \left[p(X|\varphi_{t-1}^i)\right]^{\phi_t - \phi_{t-1}}, \quad \widetilde{W}_n^i = \frac{\widetilde{w}_t^i W_{t-1}^i}{\frac{1}{N_p}\sum_{i=1}^{N_p} \widetilde{w}_t^i W_{t-1}^i}, \quad i = 1, ..., N_p$$

[9] A pair of particle systems $\{\varphi_t^i, \widetilde{W}_t^i\}_{i=1}^{N_p}$ approximates the t-th power posterior distribution in the sense that the posterior moments of the unknown parameter can be approximated by $E_{\pi_t}[h(\varphi)] \approx \sum_i^{N_p} h(\varphi_t^i)\widetilde{W}_t^i$ where $\{\widetilde{W}_t^i\}_{i=1}^{N_p}$ serves as importance weights.

(b) *Selection. Compute the effective sample size* $ESS_t = N_p / \left(\frac{1}{N_p} \sum_{i=1}^{N_p} \left(\widetilde{W}_t^i \right)^2 \right)$. *If* $ESS_t < \widehat{\rho} N_p$, *resample the particles via multinomial resampling. Let* $\{\widehat{\varphi}_t^i\}_{i=1}^{N_p}$ *denote* N_p *i.i.d. draws from a multinomial distribution characterized by support points and weights* $\{\varphi_{t-1}^i, \widetilde{W}_t^i\}_{i=1}^{N_p}$ *and set* $W_t^i = 1$. *Otherwise, let* $\widehat{\varphi}_t^i = \varphi_{t-1}^i$ *and* $W_t^i = \widetilde{W}_t^i$, $i = 1, ..., N_p$.

(c) *Mutation. Propagate the particles* $\{\widehat{\varphi}_t^i, W_t^i\}$ *via* M *steps of the MH-within-Gibbs algorithm with transition kernel* $\varphi_t^i \sim K_t(\varphi_t | \widehat{\varphi}_t^i; \zeta_t)$ *whose stationary distribution is* $\pi_t(\varphi)$. *See Algorithm 6 for details.*

3. *For* $t = N_\phi$ $(\phi_{N_\phi} = 1)$ *the final importance sampling approximation of* $E_\pi[h(\varphi)]$ *is given by:*

$$\bar{h}_{N_\phi, N_p} = \sum_{i=1}^{N_p} h(\varphi_{N_\phi}^i) W_{N_\phi}^i. \tag{27}$$

Remark 1 (Markov transition kernel in the mutation step). In every iteration, the mutation step requires a Markov transition kernel. In this paper, I construct the Markov transition kernel based on the MH-within-Gibbs algorithm iterated over the parameter block $[\varphi_\theta, \varphi_\beta, \varphi_V, \varphi_\alpha, \varphi_\psi]$ where $\varphi_{\theta,j} = \theta_j$ for $j = 1, ..., J$; $\varphi_\beta = \beta$; $\varphi_{V,j} = V_j$ for $j = 1, ..., J$; $\varphi_\alpha = \alpha$; $\varphi_\psi = \psi$. The blocks for θ, β, and V involve MH updating and I use RWMH updating with covariance matrix of the form $(\widehat{c}_{b,j} \widetilde{\Sigma}_{b,j})$ for $b \in \{\theta, \beta, V\}$ and $j = 1, ..., J$. The scaling parameters $\widehat{c}_{b,j}$ are computed so that the rejection probabilities of the MH steps stay near 30%. The covariance matrices $\widetilde{\Sigma}_{b,j}$ are adaptively chosen using the importance sampling approximation at every stage. These are summarized below[10].

Algorithm 6. (Adaptive transition kernel) *Enter the algorithm with* $\{\varphi_{t-1}^i, \widetilde{W}_t^i\}_{i=1}^{N_p}$ *and* $\{\widehat{\varphi}_t^i\}_{i=1}^{N_p}$:

1. *Compute importance sampling approximations*

$$\widetilde{\Sigma}_{b,j} = \sum_{i=1}^{N_p} (\varphi_{t-1}^i - \widetilde{\mu}_{b,j})^2 \widetilde{W}_t^i \quad \text{where} \quad \widetilde{\mu}_{b,j} = \sum_{i=1}^{N_p} \varphi_{t-1}^i \widetilde{W}_t^i$$

for $b \in \{\theta, \beta, V\}$ *and* $j = 1, ..., J$.

[10] The algorithm is described for the MR-DPM models without latent variables. When there are latent variables in the model, one can add the random-walk Metropolis-Hastings steps for β_z and $z_{0:T}$ in the above algorithm as in Algorithm 3.

2. Compute the average empirical rejection rate $\widehat{R}_{t-1,b,j}$, based on the mutation step in the previous stage $t-1$ for $b \in \{\theta, \beta, V\}$ and $j = 1, ..., J$.

3. Adjust the scaling factor according to

$$\widehat{c}_{2,b,j} = c^*_{b,j}, \quad \widehat{c}_{t,b,j} = \widehat{c}_{t-1,b,j} f(1 - \widehat{R}_{t-1,b,j}) \quad \textit{for} \quad t \geq 3,$$

for $b \in \{\theta, \beta, V\}$ and $j = 1, ..., J$. And $f(\cdot)$ is given by

$$f(x) = 0.95 + 0.1 \frac{\exp(16(x - 0.30))}{1 + \exp(16(x - 0.30))}.$$

4. For each particle i, run M steps of the following MH-within-Gibbs algorithm. Let $\varphi^i_{t,0} = \widehat{\varphi}^i_t$. For $m = 1$ to M:

(a) Let $\varphi^i_{t,m} = \varphi^i_{t,m-1}$.

(b) For $b \in \{\theta, \beta, V\}$ and $j = 1, ..., J$, generate a proposal draw $\varphi^*_{b,j}$ from

$$\varphi^*_{b,j} \sim N\left(\varphi^i_{t,b,j,m-1}, \; \widehat{c}^2_{t,b,j} \widetilde{\Sigma}_{t,b,j}\right)$$

and define the acceptance probability

$$\alpha(\varphi_{b,j}) = \min\left\{1, \frac{[p(X|\varphi^*_{b,j}, \varphi^i_{t,-(b,j),m})]^{\phi_t} p(\varphi^*_{b,j}, \varphi^i_{t,-(b,j),m})}{[p(X|\varphi^i_{t,b,j,m-1}, \varphi^i_{t,-(b,j),m})]^{\phi_t} p(\varphi_{t,b,j,m-1}, \varphi^i_{t,-(b,j),m-1})}\right\}$$

and let

$$\varphi^i_{t,b,j,m} = \begin{cases} \varphi^*_{b,j} & \textit{with probability } \alpha(\varphi_{b,j}) \\ \varphi^i_{t,b,j,m-1} & \textit{otherwise} \end{cases}$$

(c) Draw $\varphi^*_{t,\alpha,m}$ and $\varphi^*_{t,\psi,m}$ based on the relevant steps in Algorithm 2 with $\varphi^0 = \varphi^i_{t,m}$.

5. Let $\widehat{\varphi}^i_t = \varphi^i_{t,M}$.

Remark 2 (Tuning parameters). In this paper, I employ the following tempering schedule with a scalar parameter $\eta > 0$,

$$\phi_t = \left(\frac{t-1}{N_\phi - 1}\right)^\eta.$$

The parameter η controls the relative similarity of the intermediate power posteriors across stages. For example, for $\eta = 1$, the tempering schedule is linear in t and the similarity of adjacent power posteriors (as measured by the proximity of ϕ_t and ϕ_{t-1}—the smaller $\phi_t - \phi_{t-1}$, the more similar the adjacent power posteriors) is the same across all stages. On the other hand, for $\eta > 1$, two adjacent intermediate posteriors are closer in the initial stages, with the similarity decreasing across stages. This means that a tempering schedule with $\eta > 1$ moves the intermediate power posteriors slowly from the prior distribution at the beginning stages, then transitions faster to the posterior distribution in the later stages. Note that if η is too large, some of the intermediate distributions will be rendered redundant.

For fixed η, the number of stages N_ϕ controls the absolute similarity of intermediate power posteriors over stages. When two adjacent power posteriors are close to each other, more efficient approximation is possible because the SMC sampler utilizes the posterior approximation from the previous stage as an importance sampling distribution. However, a larger number of stages entails more likelihood evaluations, creating a computational trade-off. As mentioned earlier, the algorithm performs the re-sampling if the effective sample size is smaller than $\widehat{\rho} \times N_p$. Tuning parameter M determines the number of transitions in the mutation step.

Remark 3 (SLLN and CLT). The particle approximations to posterior moments such as the one in Equation 27 satisfy the strong law of large numbers (SLLN) and the central limit theorem (CLT) under suitable conditions without any adaptation. The following conditions are provided in Herbst and Schorfheide (2014).

Assumption 5. *Suppose that (i) the prior is proper: $\int p(\varphi)d\varphi < \infty$; (ii) the likelihood function is uniformly bounded: $sup_{\varphi \in D} < M_D < \infty$; and (iii) positive marginal data density:$\int [p(X|\varphi)]^{\phi_2} p(\varphi)d\varphi > 0$.*

Assumption 6. *$\pi_t(\varphi)$ is an invariant distribution associated with the transition kernel, that is: $\int K_t(\varphi|\widehat{\varphi};\zeta_t)\pi_t(\widehat{\varphi})d\widehat{\varphi} = \pi_t(\varphi)$ where $\{\zeta_t\}_{t=1}^{N_\phi}$ is a non-random sequence of transition kernel parameters.*

Under Assumptions 5 and 6 the particle approximations to the posterior moments satisfy the strong law of large numbers. That is,

$$\bar{h}_{N_\phi,N_p} = \sum_{i=1}^{N_p} h(\varphi^i_{N_\phi})W^i_{N_\phi} \rightarrow_{a.s.} E_\pi[h] \quad \text{as } N_p \rightarrow \infty$$

for $h \in \left\{h(\varphi)\middle|\exists\delta > 0 \; s.t. \int |h(\varphi)|^{1+\delta}p(\varphi)d\varphi < \infty\right\}$. This result corresponds to the law of large numbers for the basic sampler in Proposition 1. In addition, the CLT applies to the same quantity as the number particles N_p increases to infinity. The asymptotic variance in the limit distribution depends on many factors, including tuning parameters such as $\{\zeta_t\}_{t=1}^{N_\phi}$.

One of the potentially binding assumptions for these results to hold in the context of the MR-DPM model is the second condition in Assumption 5. For example, when the kernel function is set to the normal density function with heterogeneous location and scale parameters $k(x_i; \mu_j, \sigma_j^2)$, the likelihood function of the J-truncated MR-DPM model is not uniformly bounded. If one of the location parameters takes the same value as the observation ($\mu_j = x_i$), then the likelihood tends to infinity as the corresponding scale parameter gets smaller ($\sigma_j \to 0$). One direct solution to this problem is to set lower bounds on the scale parameters so that the likelihood function does not explode over the parameter space. The current implementation of the algorithm in the simulation and application sections does not restrict the domain of the scale parmaeters. Instead, I monitor whether the posterior simulator pushes the scale parameters toward problematic regions; at least for the application presented, this is not the case.

The other assumptions are implied by Assumptions 1–4. For example, the prior distributions used in this paper are proper (Section 4). Assumption 6 can be verified along the lines of the proof of Proposition 1 because the transition kernels in the SMC sampler (Algorithm 2) and the basic sampler (Algorithm 2) are identical when there is no adaptation (except for the fact that the likelihood function is powered by the positive number ϕ_t).

As pointed out by Herbst and Schorfheide (2014), convergence results with adaptive schemes are harder to show. An alternative SMC algorithm that satisfies the SLLN and the CLT is to run two versions of the SMC algorithms. First, run the adaptive SMC sampler (Algorithm 5) to obtain a sequence of tuning parameters $\{\widehat{c}_t, \widetilde{\Sigma}_t\}_{t=1}^{N_\phi}$. Then, fix these tuning parameters in subsequent runs of the algorithms.

Remark 4 (Marginal likelihood). One attractive feature of this algorithm is that it produces marginal likelihood estimates as a by-product. Using a sequence of normalizing constants of intermediate posteriors, the marginal likelihood can be written in telescoping fashion as

$$p(X) = C_{\phi_N}(X) = \frac{C_{\phi_N}(X)}{C_{\phi_{N-1}}(X)} \times \frac{C_{\phi_{N-1}}(X)}{C_{\phi_{N-2}}(X)} \times ... \times \frac{C_{\phi_1}(X)}{C_{\phi_0}(X)}$$

where $C_{\phi_0}(X) = 1$ by construction. Each ratio can be written as

$$\begin{aligned}\frac{C_{\phi_t}(X)}{C_{\phi_{t-1}}(X)} &= \frac{\int p(X|\varphi)^{\phi_t} p(\varphi) d\varphi}{C_{\phi_{t-1}}(X)} = \int p(X|\varphi)^{\phi_t - \phi_{t-1}} \times \frac{p(X|\varphi)^{\phi_{t-1}} p(\varphi)}{C_{\phi_{t-1}}(X)} d\varphi \\ &= E_{\pi_t}\left[p(X|\varphi)^{(\phi_t - \phi_{t-1})}\right]\end{aligned}$$

and therefore it can be approximated by treating $h(\varphi) = p(X|\varphi)^{(\phi_t - \phi_{t-1})}$ and the marginal likelihood can be computed as:

$$p(X) = \prod_{t=2}^{N_\phi} \left(\frac{1}{N_p} \sum_{i=1}^{N_p} \widetilde{w}_t^i W_{t-1}^i \right).$$

Remark 5 (Truncation order selection through J-posterior tempering). All of the proposed posterior samplers so far assume that the truncation order J is fixed. It turns out that a slight modification[11] of the SMC algorithm introduced in this section offers a natural method to select the truncation order J. The basic idea is to consider the sequence of truncated posterior distributions indexed by $J = J_0, J_1, ...$:

$$\pi_J(\varphi_J|X) \propto p_J(X|\theta_J, \beta, V_J)p(\theta_J|\psi)p(V_J|\alpha)p(\beta)I(\vec{0} \in H(\theta_J, \beta))p(\psi)p(\alpha) \tag{28}$$

where $\varphi_J = (\theta_J, \beta, V_J, \psi, \alpha)$ and $\theta_J = \{\theta_{1,J}, \theta_{2,J}, ..., \theta_{J,J}\}$ and $V_J = \{V_{1,J}, V_{2,J}, ..., V_{J,J}\}$. And the likelihood function is indexed by J. Instead of moving particles through the power posteriors, the algorithm moves particles through the truncation level.

At the arbitrary stage $(s+1)$, the algorithm begins with the particle approximation to the s-th posterior, $\{(\varphi_s^i, \widetilde{W}_s^i)\}_{i=1}^{N_p}$. Then one can propagate particles using the conditional prior distributions of the $(s+1)$-truncated model given φ_s^i for $i = 1, ..., N_p$,

$$V_{s+1,s+1}^i \sim p(V_{s+1,s+1}^i|V_s^i, \alpha^i) \quad \text{and} \quad \theta_{s+1,s+1}^i \sim p(\theta_{s+1,s+1}^i|\theta_s^i, \beta^i, \psi^i) \tag{29}$$

from both of which it is easy to sample under the J-truncated MR-DPM modeling assumptions and the associated prior specifications. Equipped with augmented particles

$$\{(\varphi_s^i, V_{s+1,s+1}^i, \theta_{s+1,s+1}^i, \widetilde{W}_s^i)\}_{i=1}^{N_p},$$

one can update the particle weights to get the particle approximation to the $(s+1)$-th posterior distribution using the following SMC increment

$$\widetilde{w}_{s+1}^i = \frac{\prod_{l=1}^N \left(\sum_{j=1}^{s+1} \widetilde{q}_{j,s+1}(\theta_{s+1}^i, \beta^i, V_{s+1}^i) k(x_l; \theta_{j,s+1}^i) \right)}{\prod_{l=1}^N \left(\sum_{j=1}^{s} \widetilde{q}_{j,s}(\theta_{s}^i, \beta^i, V_{s}^i) k(x_l; \theta_{j,s}^i) \right)},$$

where I define $\theta_{s+1}^i = \{\theta_{1,s}^i, \theta_{2,s}^i, ..., \theta_{s,s}^i, \theta_{s+1,s+1}^i\}$ and $V_{s+1}^i = \{V_{1,s}^i, V_{2,s}^i, ..., V_{s,s}^i, V_{s+1,s+1}^i\}$. Note

[11] This type of modification of the SMC sampler is considered by Griffin (2014) in the context of the DPM model without moment restrictions.

that this SMC increment simplifies to a ratio between the likelihood of the $(s+1)$-truncation model and the s-truncation model because of the choice of the transition distribution in Equation 29.

Unlike the simulated tempring SMC algorithm (Algorithm 5), this algorithm starts from a J_0-truncated model and moves particles forward to a higher-order truncated model. As the truncation order increases, we expect the approximated posterior to be closer to the true posterior, which is achieved when the truncation order is ∞. In practice, the algorithm must stop with a finite truncation order. Griffin (2014) suggests a stopping rule based on the effective sample size (ESS). The rationale behind this idea is that as the truncation order gets larger, two adjacent posterior distributions $\pi_s(\varphi_s|X)$ and $\pi_{s+1}(\varphi_{s+1}|X)$ become very close, since the newly introduced component will not affect the posterior much, and therefore, the SMC increment at stage $(s+1)$ becomes closer to one, i.e., $\widetilde{w}^i_{s+1} \approx 1$ for all i.

The initial posterior distribution for this algorithm can be obtained by Algorithm 5 with some $J > 1$. I close this section with complete instructions for the algorithm.

Algorithm 7. J-selection SMC for the MR-DPM model

1. *Initialization. Run Algorithm 5 with a truncation order J_0 to get the particle approximation to the J_0-truncation posterior, $\{(\varphi^i_s, \widetilde{W}^i_s)\}^{N_p}_{i=1}$, and set $s = 1$.*
2. *Recursion.*
   (a) *Propagation. Propagate particles using the conditional prior distributions of the s-truncated model given φ^i_{s-1} for $i = 1, ..., N_p$,*

   $$V^i_{s,s} \sim p(V^i_{s,s}|V^i_{s-1}, \alpha^i) \quad \textit{and} \quad \theta^i_{s,s} \sim p(\theta^i_{s,s}|\theta^i_{s-1}, \beta^i, \psi^i)$$

   and augment with a set of particle pairs, $\{(\varphi^i_{s-1}, V^i_{s,s}, , \theta^i_{s,s}, \widetilde{W}^i_{s-1})\}^{N_p}_{i=1}$,

   (b) *Correction. Reweight the particles from stage $s-1$ by defining the incremental and normalized weights*

   $$\widetilde{w}^i_s = \frac{\prod^N_{l=1}\left(\sum^s_{j=1} \widetilde{q}_{j,s+1}(\theta^i_s, \beta^i, V^i_s) k(x_l; \theta^i_{j,s})\right)}{\prod^N_{l=1}\left(\sum^{s-1}_{j=1} \widetilde{q}_{j,s-1}(\theta^i_{s-1}, \beta^i, V^i_{s-1}) k(x_l; \theta^i_{j,s-1})\right)}, \quad \widetilde{W}^i_n = \frac{\widetilde{w}^i_s W^i_{s-1}}{\frac{1}{N_p}\sum^{N_p}_{i=1} \widetilde{w}^i_s W^i_{s-1}}$$

   for $i = 1, ..., N_p$.

   (c) *Selection. Compute the effective sample size $ESS_s = N_p / \left(\frac{1}{N_p}\sum^{N_p}_{i=1}\left(\widetilde{W}^i_s\right)^2\right)$. If $ESS_s < \widehat{\rho}N_p$, resample the particles via multinomial resampling. Let $\{\widehat{\varphi}^i\}^{N_p}_{i=1}$*

denote N_p i.i.d. draws from a multinomial distribution characterized by support points and weights $\{\varphi^i_{s-1}, \widetilde{W}^i_s\}_{i=1}^{N_p}$ and set $W^i_s = 1$. Otherwise, let $\widehat{\varphi}^i_s = \varphi^i_{s-1}$ and $W^i_s = \widetilde{W}^i_s$, $i = 1, ..., N_p$.

(d) Mutation. Propagate the particles $\{\widehat{\varphi}^i_s, W^i_s\}$ via M steps of the posterior sampler described in the previous section (Algorithm 6).

(e) Stopping rule. Stop the algorithm if $|ESS_k - ESS_{k-1}| < \epsilon N_p$ for $k = s-2, s-1, s$ and $\epsilon = 10^{-5}$. Otherwise, go to step (a) with $s = s + 1$.

6 Working with simulated data

This section is composed of three subsections designed to illustrate the proposed samplers with simulated data. Simulation designs in each subsection are carefully chosen so that each section delineates a different aspect of the proposed samplers. In the first subsection, data are simulated from an *i.i.d.* instrumental regression model with log-normal shocks. I use this environment to compare the performance of the three algorithms described in the previous section. In the second part of this section, I simulate data based on an Euler equation model. The simulated data are serially correlated, and therefore, I fit the data using the time series MR-DPM model. Here, I focus on the SMC sampler and describe the role of the marginal likelihood in Bayesian moment condition models. Finally, I turn the simulations toward dealing with models with latent variables. In this part, I simulate data from a non-Gaussian state-space model and illustrate how one can perform the MR-DPM model estimation with latent variables.

6.1 IV regression

The model considered in this section is the linear instrumental variable regression

$$\begin{aligned} y_i &= \beta_2 x_i + e_{1,i} \\ x_i &= z_i'(\iota\delta) + e_{2,i} \end{aligned} \tag{30}$$

where y_i and x_i are scalar and z_i is a $r \times 1$ vector. ι is a vector of r ones and δ is a scalar. The parameter of interest is β_2 and it is set to one. I consider two specifications for e_i,

$$\begin{pmatrix} e_{1,i} \\ e_{2,i} \end{pmatrix} = cv;\ v \sim \log N(0,\ 0.6 \cdot \Sigma), \quad \text{where } \Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$

and c is set to 1.25. The z_i are from an independent standard normal distribution. The number of instruments is three, δ is 1, and the sample size is set to 200.

Moment condition. To estimate β, I use the following moment conditions

$$\begin{aligned} E[y_i - \beta_1 - x_i'\beta_2] &= 0 \\ E[z_i(y_i - \beta_1 - x_i'\beta_2)] &= 0. \end{aligned} \tag{31}$$

Note that β_1 is the intercept and β_2 is the parameter loaded on x_i in Equation 30. Under the simulation design, β_1 is non-zero because the shocks to y_i and x_i follow the log-normal distribution. The integrated moment conditions are then

$$\begin{aligned} \mu_Y - \beta_1 - \mu_X'\beta_2 &= 0 \\ \Sigma_{ZY} + \mu_Z\mu_Y - \beta_1\mu_Z - (\Sigma_{ZX} + \mu_Z\mu_X')\beta_2 &= 0, \end{aligned} \tag{32}$$

given some mixture parameter $\theta = (\mu, \Sigma)$, μ_Y denotes the mean parameter for variable Y and Σ_{ZY} denotes covariance parameter for Z and Y.

Kernel function and prior specification. I model the joint distribution of $[y_i', x_i', z_i']'$ based on the *i.i.d.* MR-DPM model in conjunction with moment conditions in Equation 32. The initial prior distribution for β_1 and β_2 follows the uniform distribution

$$\beta_1 \sim Unif[-1, 1] \quad \text{and} \quad \beta_2 \sim Unif[0, 3].$$

and other prior specifications are set according to Section 4.2. To check whether the exponential tilting projection with this prior specification distorts the initial prior distribution, I generate 1,000 draws from the prior distribution using accept/reject sampling (Algorithm 1). Figure 2 shows the scatter plot of prior draws for (β_1, β_2) with their prior means (dashed lines). Visually speaking, it turns out that the introduction of the exponential tilting procedure does not significantly change the initial prior distribution, at least for β – the draws appear uniformly distributed on the domain of the initial distribution.

Tuning of samplers. In this experiment, I compare the three posterior algorithms proposed in Section 5: the basic sampler (Algorithm 2, hereafter B-sampler), the data-augmentation (Algorithm 4, hereafter DA-sampler), and the sequential Monte Carlo sampler (Algorithm 5, hereafter S-sampler). I set the tuning parameters of the SMC sampler as follows: the number of stages $N_\phi = 50$; the number of particles $N_p = 1000$; the number of transitions in

Table 1 Posterior sampler comparison

	Mean	[0.05, 0.95]	SD(Mean)	N_{eff}
B-Sampler				
β_1	0.374	[0.338, 0.995]	0.0028	74.89
β_2	1.021	[0.414, 1.047]	0.0050	12.18
DA-Sampler				
β_1	0.369	[0.329, 0.989]	0.0023	111.03
β_2	1.019	[0.413, 1.048]	0.0010	278.68
S-Sampler				
β_1	0.367	[0.329, 0.995]	0.0057	18.42
β_2	1.023	[0.405, 1.052]	0.0079	4.91

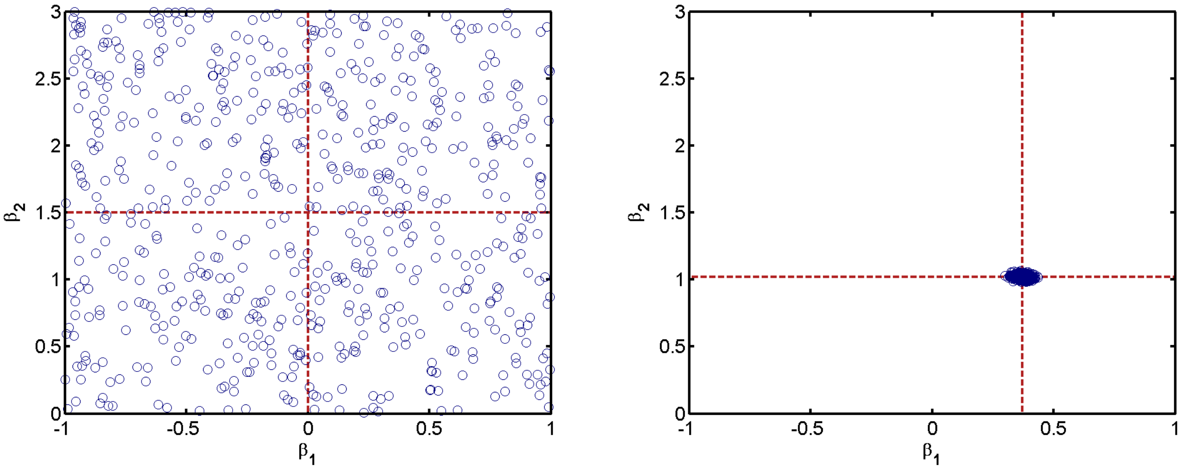
Notes: B-sampler stands for the basic sampler. DA-sampler stands for the modified version of the basic sampler using data augmentation. S-sampler stands for the SMC posterior sampler. Means and standard deviations are over 10 runs for each algorithm. I define $N_{eff} = \widehat{V}(\beta)/SD^2$. Both B- and DA- samplers use 100,000 draws with the first half discarded. The S-samplers use 1,000 particles and 50 stages.

the mutation step $M = 2$; and the bending coefficient $\eta = 1.5$. For the basic sampler and the data augmented sampler, I compute posterior moments based on 100,000 draws, with the first 50,000 draws discarded.

Result: Prior/Posterior. Figure 2 shows the 1,000 draws of β_1 and β_2 from the marginal prior (left panel) and the marginal posterior (right panel) distribution. As one can see, once conditioned on data, the marginal distribution of β_1 and β_2 shrinks and centers around the true value. Figure 3 shows prior and posterior draws of the density of y_i. Prior draws can exhibit various shapes such as skewed and multi-modal densities. Once conditioned on the data, all densities implied by posterior draws have one mode and are distributed around the posterior mean (solid red line). The posterior mean estimate is reasonable in that it is very close to the kernel density estimate implied by Silverman's optimal bandwidth using the same data.

Result: Comparison of proposed samplers. To compare the performance of the algorithms, I ran the three samplers 10 times each with different initial points drawn from the prior distribution. The first two columns of Table 1 show the posterior moments (mean and quantiles) of (β_1, β_2) computed from the three samplers. They are not exactly the same, but are very close to one another. The third column presents the standard deviation of the

Figure 2 Scatter Plot of draws for (β_1, β_2)



(a) Prior Draws

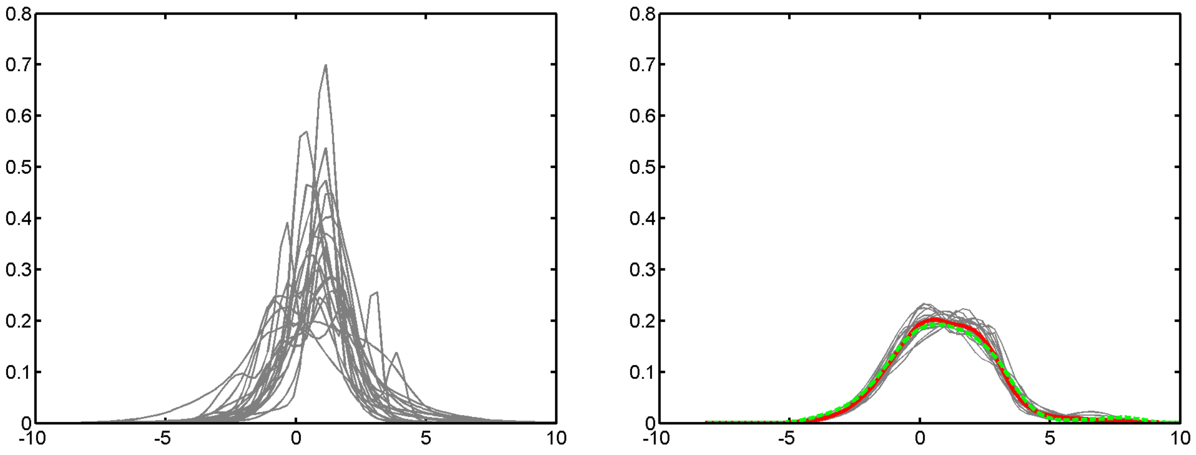
(b) Posterior Draws

Notes: Scatter plot of draws of (β_1, β_2) from prior (left) and posterior (right) distribution. Dashed red lines show prior and posterior means. Figures are generated based on the draws from the S-sampler.

posterior mean estimates from the three samplers. They are computed by taking the sample standard deviation of the posterior mean estimates from 10 runs with different initial values. Several findings emerge. First, the gain from data augmentation is evident. It reduces the standard deviation of the posterior mean estimate by a factor of five for β_2. Second, the SMC sampler produces the least accurate posterior mean estimator given the current choice of the tuning parameters. The same conclusion can be made based on the number of effective samples defined as $N_{eff} = \widehat{V}(\beta)/SD^2$ where $\widehat{V}(\beta)$ is an estimate of the posterior variance of β obtained from the output of the SMC algorithm and the SD^2 is the variance of the posterior mean estimate across the 10 runs of each algorithm. The gain from the data-augmentation is large, but the S-sampler is the least efficient.

The current version of the sequential Monte Carlo sampler needs more investigation. The most immediate exercise is to increase the number of particles to see whether the central limit theorem presented in Section 5 holds under this tuning parameter configuration. Moreover, as Herbst and Schorfheide (2014) point out, other tuning parameters such as the tempering schedule are crucial for the efficient implementation of the SMC algorithm. Currently, I am investigating these issues in detail.

Figure 3 Draws for Density of y_i



(a) Prior Draws

(b) Posterior Draws

Notes: Each draw from the prior/posterior is transformed into a marginal density of y_i and is evaluated at 100 equally spaced grid points from $[-10, 10]$. The left panel shows 50 draws from the prior distribution. The right panel shows 50 draws from the posterior distribution as well as the point-wise posterior mean of the density (thick red line). The dashed green line is the kernel density estimate using the same data with Silverman's optimal bandwidth. Figures are generated based on the draws from the S-sampler.

6.2 Estimating an Euler equation with time series data

In this subsection, I consider the following data generating mechanism,

$$\begin{aligned} r_{t+1} &= -\gamma\omega + \gamma c_{t+1} - \frac{\gamma+1}{2} c_{t+1}^2 - \gamma e_{r,t+1} \\ c_{t+1} &= \rho_c c_t + e_{c,t+1} \end{aligned}$$

where the innovations are generated from the bi-variate normal distribution,

$$\begin{pmatrix} \epsilon_{r,t+1} \\ \epsilon_{c,t+1} \end{pmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 0.2 \end{pmatrix}.$$

This simulation design resembles the representative agent Euler equation model with CRRA utility function and satisfies the following moment conditions,[12]

$$E\left[c_{t+1} - \omega - \frac{1}{\gamma} r_{t+1} - \frac{\gamma+1}{2}(c_{t+1})^2 \middle| I_t \right] = 0 \tag{33}$$

[12] This Euler equation is derived from the household's optimization problem in a life-cycle model based on a second-order Taylor approximation. A similar Euler-equation model is considered by Alan et al. (2012) in the context of heterogeneous agents.

where the information set contains past histories of c_t and r_t, $I_t = \{c_t, r_t, c_{t-1}, r_{t-1}, ...\}$. I simulate 500 observations with $(\omega, \gamma, \rho_c) = (-0.6, 3, 0.6)$.

Moment conditions. To perform the MR-DPM estimation, I transform the above Euler equation into unconditional moment conditions using the instruments z_t,

$$E\left[\left(c_{t+1} - \omega - \frac{1}{\gamma}r_{t+1} - \frac{\gamma+1}{2}(c_{t+1})^2\right)z_t\right] = 0.$$

In this experiment, I consider three model specifications using different moments (instruments, z_t). The first specification uses $\{1, c_t, r_t\}'$ as instruments and is thus the correctly specified model. The second specification uses the same instruments as the first specification but the risk aversion parameter γ is fixed at one (implying log-utility). The last specification adds c_{t+1} to the first instrument set. Note that because of endogeneity, c_{t+1} is an invalid instrument. Note also that the last two approaches are misspecified models and hence the moment conditions are violated.

Kernel function and prior specification. To fit simulated data using the MR-DPM model, I impose the time series version of the MR-DPM modeling assumption on the joint distribution of $(c_t, r_t, c_{t-1}, r_{t-1})$ introduced in Section 3,

$$\begin{pmatrix} c_t \\ r_t \\ c_{t-1} \\ r_{t-1} \end{pmatrix} \sim \sum_{j=1} \widetilde{q}_j(\mu, \Sigma, \beta, V) N\left(\cdot \;\middle|\; \begin{pmatrix} \mu_j \\ \mu_j \end{pmatrix}, \Sigma_j\right)$$

where the multivariate normal density is used for the kernel function; μ_j is 2×1 and Σ_j is 4×4. The prior distribution for $\beta = (\omega, \gamma)$ is set to be a uniform distribution

$$\omega \sim Unif[-3, 0] \quad \text{and} \quad \gamma \sim Unif[0, 5].$$

For the base measure $G_0(\mu, \Sigma)$, I decompose the covariance matrix Σ_j into two pieces, $\Sigma_j = D_j \times R_j \times D_j$ where

$$D_j = \begin{pmatrix} \sigma^2_{1,j} & 0 & 0 & 0 \\ 0 & \sigma^2_{2,j} & 0 & 0 \\ 0 & 0 & \sigma^2_{1,j} & 0 \\ 0 & 0 & 0 & \sigma^2_{2,j} \end{pmatrix} \quad \text{and} \quad R_j = \begin{pmatrix} 1 & \cdot & \cdot & \cdot \\ r_{1,j} & 1 & \cdot & \cdot \\ r_{2,j} & r_{4,j} & 1 & \cdot \\ r_{3,j} & r_{5,j} & r_{1,j} & 1 \end{pmatrix}$$

where the correlation parameters $(r_{1,j},, r_{2,j}, r_{3,j}, r_{4,j}, r_{5,j})$ are assumed to be uniform on $[-1, 1]$ and $\sigma^2_{1,j}$ and $\sigma^2_{2,j}$ are assumed to be Gamma-distributed, $\sigma^2_{1,j} \sim Ga(2, 0.5^2 Var(c_t))$ and $\sigma^2_{2,j} \sim Ga(2, 0.5^2 Var(r_t))$. The prior distribution for μ_j is similar to that presented in Section 4, and a normal distribution is used. Related hyperparameters are fixed so that the prior distribution of μ_j is distributed around the realized data. The prior for the concentration parameter α is implemented as $Ga(10, 2)$.

Tuning of the posterior samplers. In this experiment, I only consider the SMC sampler and focus on computation of the marginal likelihood. I set the tuning parameters as follows: the number of stages $N_\phi = 100$; the number of particles $N_p = 1000$ and 3000; the number of transitions in the mutation step $M = 1$; and the bending coefficient $\eta = 1$.

Result: Marginal likelihood comparison. Table 2 presents the log marginal likelihood estimates (mean and standard deviation) for the time series Euler equation experiment based on the SMC sampler. The first row presents means and standard deviations of log marginal likelihood estimates over the 10 runs for each specification. The marginal likelihood comparison correctly ranks the competing model specifications. Means of the log marginal likelihoods based on the wrong moments ($\mathcal{M}_2$ and $\mathcal{M}_3$) are smaller than that of $\mathcal{M}_1$. The misspecified utility assumption ($\mathcal{M}_2$) is clearly dominated by the other two specifications. However, marginal likelihood comparison between $\mathcal{M}_1$ and $\mathcal{M}_3$ is less sharp because standard deviations of the log marginal likelihood estimates are large and, therefore, require more accurate approximation. As I pointed out in the previous experiment, the current version of the SMC sampler needs more investigation of its tuning setup to improve efficiency. One way to improve the accuracy of the marginal likelihood approximation is to run the sampler with a larger number of particles. To see this, I present log marginal likelihood estimates with 3,000 particles for $\mathcal{M}_1$ in the second row of the table, and as expected, the standard deviation becomes smaller. However, this improvement comes at the cost of computation time.

6.3 Robust estimation of the state-space model

In this section, I illustrate how one can apply the MR-DPM estimation when there are latent variables. I consider the following linear state-space model,

$$\begin{aligned} x_t &= \beta z_t + e_t \\ z_t &= \rho_z z_t + v_t, \quad v_t \sim N(0, 1) \end{aligned}$$

Table 2 Log marginal likelihood estimates

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$N_p = 1,000$	802.16 (9.61)	768.90 (7.97)	798.62 (7.38)
$N_p = 3,000$	807.38 (5.48)	-	-

Notes: This table presents the log marginal likelihood estimates (mean and standard deviation) for the time-series Euler equation experiment based on the SMC sampler. There are three specifications with different sets of moments: $\mathcal{M}_1$ is the correctly specified model ($z_t = \{1,\ c_t,\ r_t\}$); $\mathcal{M}_2$ restricts the risk-aversion parameter to one but uses the same set of moments as $\mathcal{M}_1$ ($z_t = \{1,\ c_t,\ r_t\}$); $\mathcal{M}_3$ contains an invalid moment condition ($z_t = \{1,\ c_t,\ r_t,\ c_{t+1}\}$). Means and standard deviations are computed over 10 runs for each specification. N_p denotes the number of particles used in each run of the SMC samplers.

where z_t is assumed to be unobserved by the econometrician. The measurement error e_t is assumed to follow the two-component mixture of normals,

$$e_t \sim 0.5N(0,\ 1^2) + 0.5N(0, 5^2),$$

so that e_t is mean zero but negatively skewed. Moreover, e_t is assumed to be orthogonal to v_t, and therefore, the model can be written in the same form as Equation 20,

$$E[(x_t - \beta z_t)z_t)] = 0, \quad z_t = \rho_z z_t + v_t, \quad v_t \sim N(0,1).$$

I simulate 500 observations with $(\beta, \rho_z) = (2, 0.8)$.

Kernel function and prior specification. To fit simulated data using the MR-DPM model, I impose the MR-DPM modeling assumption on the conditional distribution of x_t given z_t as illustrated in Section 3.2 with a normal density as the kernel function,

$$p(X_t|Z_t, \theta, \beta, \beta_z, V) = \sum_{j=1}^{J} \widetilde{q}_j(\theta, \beta, \beta_z, V) N(x_t; h_\mu(z_t, \theta_j), h_\Sigma(z_t, \theta_j))$$

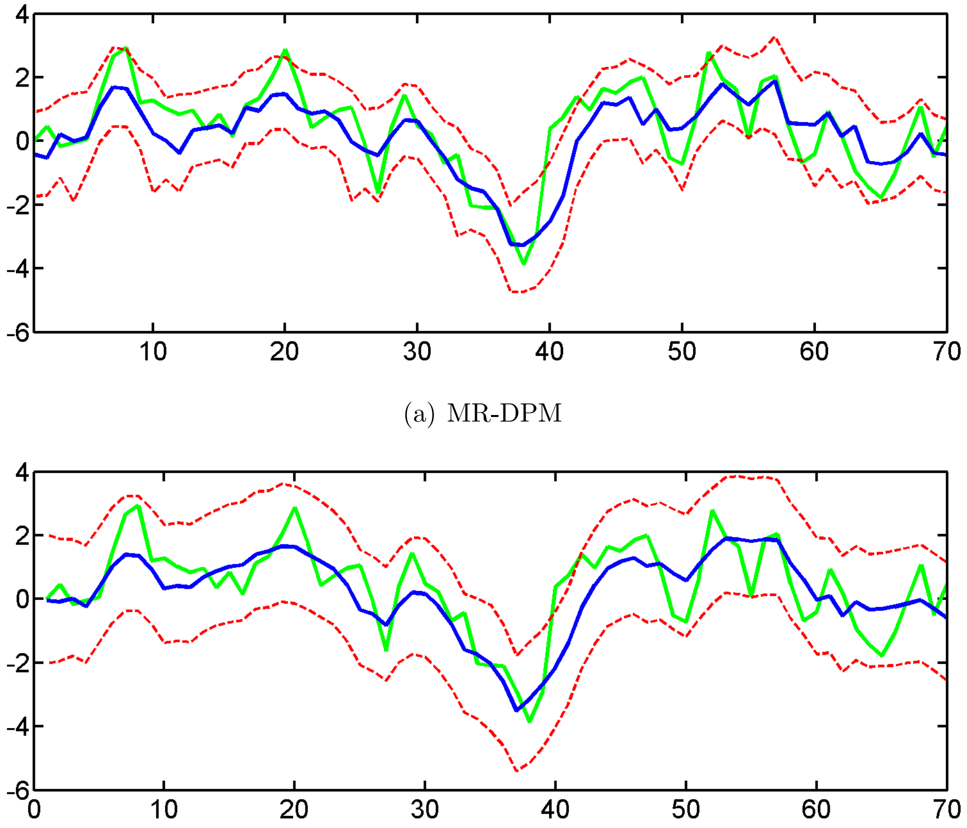
where $\theta_j = (\mu_{x,j}, \rho_{x,j}, \sigma^2_{x,j})$, $\beta_z = \rho_z$, and

$$h_\mu(z_t, \theta_j) = \mu_{x,j} + \rho_{x,j} z_t \quad \text{and} \quad h_\Sigma(z_t, \theta_j) = \sigma^2_{x,j} - \frac{\rho^2_{x,j}}{1-\rho^2_z}.$$

The base measure is set to the following form,

$$G_0(\mu_x, \rho_x, \sigma^2_x) = N(\mu_x; m_\mu, V_\mu) \times N(\rho_x; m_\rho, V_\rho) \times IG(\sigma^2_x; s_\sigma, S_\sigma)$$

Figure 4 Estimated latent variables ((z_t))



(a) MR-DPM

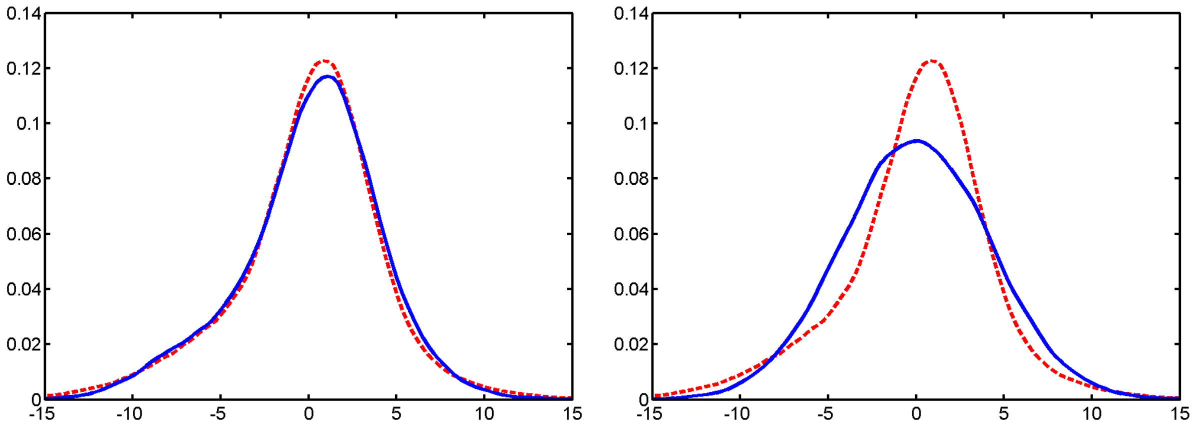
(b) Gaussian measurement assumption

Notes: This figure presents the estimated latent variables for the first 70 among 500 observations. The first figure is based on the MR-DPM estimation, while the second figure is based on the wrong parametric assumption (Gaussian measurement error). The green lines are the true latent series, the blue lines are posterior means, and the dashed red lines are 90% credible bands.

where the hyperparameters are chosen in a way similar to that described in Section 4. The prior for ρ_z is set to be the truncated normal distribution $\rho_z \sim TN(\rho_z; 0.5, 10, [-1, 1])$ and the prior for β is uniform on $[-5, 5]$. The prior for the concentration parameter α is set to be $Gamma(10, 2)$.

Posterior simulators. In this section, I run the basic sampler for the model with latent variables (Algorithm 3). For comparison, I also run the Gibbs sampling algorithm assuming that the measurement error is normally distributed. Apparently, the second posterior simulator is based on the wrong distributional assumption.

Figure 5 Predictive distribution for x_{T+1}



(a) MR-DPM

(b) Gaussian measurement assumption

Notes: This figure presents the estimated 1-step-ahead predictive distributions for x_{T+1} when $x_T = 0$. The left panel is based on the MR-DPM estimation, while the right penal is based on the wrong parametric assumption (Gaussian measurement error). The solid blue lines are estimated 1-step-ahead predictive distributions and the dashed red lines are the true predictive distributions. Owning to skewed measurement error, the predictive distribution is skewed as well.

Results. Figure 4 shows the estimated latent variables for the first 70 observations among the 500. The first figure is based on the MR-DPM estimation, while the second figure is based on the wrong parametric assumption (Gaussian measurement error). Both estimated latent variables (blue lines) are similar and the 90% credible bands (dashed red lines) contain the true latent (green lines) variables during most periods. However, the root mean squared error of state estimates based on the MR-DPM model is 10% smaller than the one computed based on the wrong parametric assumption. The difference in the estimated 1-step-ahead predictive distribution for x_{T+1} based on the two approaches is more stark. The left panel in Figure 5 shows the estimated 1-step-ahead predictive distribution when $x_T = 0$ based on the MR-DPM estimation (solid blue line). It correctly captures the skewness in the true predictive distribution. On the other hand, the estimated predictive distribution based on the wrong parametric assumption is far from the truth as it cannot capture skewness in the data.

7 Empirical application: Estimating an Euler equation

To illustrate the method proposed here, I estimate the risk aversion parameter based on the Euler equation for consumption. More specifically, I take household-level consumption panel data with a large N ($\approx 1,160$ households) and small T ($= 4$ years). Then, I apply the posterior sampling algorithm for the MR-DPM model and estimate both the risk aversion parameter and the joint distribution of household consumption.

Model. I estimate the risk aversion parameter using the Euler equation for consumption, allowing for demographic heterogeneity. Specifically, I consider the following life time optimization problem of a generic household with a CRRA utility function,

$$\max_{\{C^h_{t+j}, A_{t+j+1}\}_{j=0}^{T-t}} E_t \left[\sum_{j=0}^{T-t} \frac{(C^h_{t+j})^{1-\gamma}}{1-\gamma} e^{\zeta X^h_t} \rho^j \right]$$

subject to $A^h_{t+j+1} = (1+R^h_{t+j})A^h_{t+j} + Y^h_{t+j} - C^h_{t+j}$, where C^h_t denotes consumption by household h at time t, X^h_t is the number of children in household h at time t, A^h_t is assets held by household h at time t, Y^h_t is labor income earned by household h at time t, and R^h_t is the rate of interest for household h at time t. Throughout this subsection, I assume that the rate of interest is common and known to all households and that shocks to labor income are independent across households. In addition, I assume that the demographic variable X^h_t is exogenously given to each household and X^h_{t+1} is known to the household at time t. In the within-period utility function, there are three unknown parameters – the discount factor ρ, the risk aversion parameter γ, and the parameter loaded on demographic variables ζ. Note that the level of utility achieved by a given amount of consumption depends on the number of children in the household.

For this specification, the Euler equation for consumption has the form

$$E_t \left[\exp\left(-\gamma \Delta c^h_{t+1} + \zeta \Delta X^h_{t+1} \right) (1 + r_{t+1}) \rho \right] = 1$$

where Δc^h_{t+1} denotes consumption growth and ΔX^h_{t+1} denotes changes in the number of children. In the actual estimation, I approximate the Euler equation and obtain the following expression:

$$\Delta c^h_{t+1} = \omega + \frac{1}{\gamma} \log(1 + r_{t+1}) + \frac{\gamma + 1}{2} \left(\Delta c^h_{t+1} \right)^2 + \frac{\zeta}{\gamma} \Delta X^h_{t+1} + \nu^h_{t+1}, \tag{34}$$

where ω contains the discount rate (ρ) and higher order moments of the consumption growth and the residual ν^h_{t+1} contains the expectation error between t and $t+1$ and an approximation error. I consider past consumption growth and future changes in the number of children as instruments Z_t and therefore,

$$Z_t^h = [1, \Delta c_t^h, \Delta X_{t+1}^h]' \quad \text{and} \quad E[Z_t^h \nu_{t+1}^h] = 0. \tag{35}$$

for all h.

Data. I use the household-level annual food consumption and the demographic characteristic data taken from the Panel Study of Income Dynamics (PSID). The sample covers the period 1980 to 1984. Following Alan et al. (2009), I exclude 1) households that did not report five consecutive years of food expenditure; 2) single-headed households and households whose marital status changed over the sample period; and 3) households that do not have information on savings. The number of remaining observations after imposing the sample selection is 5,800 (1,160 per year). I also assume that all households face the same real interest rate series, computed via the US 3-month Treasury bill rates and the consumer price index. Changes in the number of children are transformed to take one of three values $\{-1, 0, 1\}$. This variable takes a value of -1 if the number of children in the household has decreased, 1 if it has increased, and zero otherwise.

The MR-DPM model. For the MR-DPM model estimation, I collect consumption growth (four years of growth) and changes in the number of children for each household (three years of changes) in one vector,

$$x_h = [x_{c,h}, x_{d,h}]' = (\Delta c_{t+3}^h, \Delta c_{t+2}^h, \Delta c_{t+1}^h, \Delta c_t^h, \Delta X_{t+3}^h, \Delta X_{t+2}^h, \Delta X_{t+1}^h)'.$$

Note that the data contain both continuous ($x_{c,h}$, consumption growth) and discrete ($x_{d,h}$, changes in the number of children) variables. The joint distribution[13] of x_h is modelled as an MR-DPM and

$$x_h \sim i.i.d. \int f_N(x_{c,h};\ \mu, \Sigma) f_M(\Delta X_{t+3}^h;\ p_3) f_M(\Delta X_{t+2}^h;\ p_2) f_M(\Delta X_{t+1}^h;\ p_1) d\widetilde{G}\,(\mu, \Sigma, p_1, p_2, p_3)$$

where f_N is the density function of the multivariate normal distribution and f_M is the probability mass function of the multinomial distribution with one trial and three support points $(-1, 0, 1)$. The parameters p_i in multinomial distributions are 3×1 vectors, each of which

[13] Strictly speaking, estimation of the joint distribution is conditional on the realized path of the interest rate r_{t+1}. However, owing to the assumption I made in the beginning of the section (the interest rate is common and known to all households at time t), the interest rates are treated as exogenous and enter only in the moment condition.

sums up to one. Although continuous and discrete variables are independent given particular component parameters $(\mu, \Sigma, p_1, p_2, p_3)$, their joint distribution can have dependency through the random mixture distribution $\widetilde{G}(\cdot)$. The distribution of the random mixing distributions is assumed to be the tilted Dirichlet process with moment conditions based on the Euler equation in Equation 34 and instruments presented in Equation 35. There are three unknown parameters in the moment condition, (ω, γ, ζ) and I collect them in one vector and write $\beta = (\omega, \gamma, \zeta)$.

In this application, I consider three model specifications based on the set of instruments used in the estimation. The first specification ("Full") refers to the estimation based on instruments $Z_t = [1, \Delta c_t, \Delta X_{t+1}]$ and estimates (ω, γ, ζ). The second specification ("No demographic") refers to the estimation based on instruments without the demographic variables ΔX_{t+1} and estimates only (ω, γ). The last specification ("No moment") estimates only the underlying data generating process without moment restrictions and is a standard DPM model estimation.

Prior Distribution. The initial prior distributions for (ω, γ, ζ) are set to be independent normal distributions with

$$\omega \sim N(-2, 1), \quad \gamma \sim N(4, 4), \quad \text{and} \quad \zeta \sim N(0, 1).$$

The center of the prior for the risk aversion parameter is based on the posterior estimates[14] of Aruoba et al. (2013). The parameter loaded on the demographic variable is centered at zero, reflecting the prior ignorance of the sign of the effect of the demographic variable. The base measure in the Dirichlet process is decomposed as follows:

$$G_0(\mu, \Sigma, p1, p2, p3) =_d N(\mu;\ m, B) IW(\Sigma; s, (sS)^{-1}) \prod_{i=1}^{3} Dir(p_i; [1/3, 1/3, 1/3])$$

where $Dir(p; \alpha_p)$ denotes the Dirichlet distribution with parameter α_p. I set s to be 5 and hyperparameters m, B, and S to have the following prior distributions,

$$m|B \sim N(m;\ a, B/\kappa), B \sim IW(B;\ \nu, \bar{\lambda} \times diag(cov(x_{c,h}))), S \sim W\left(S; q, \frac{r}{q} \times diag(cov(x_{c,h}))\right)$$

where $(a, \kappa, \nu, \bar{\lambda}, r, q) = (mean(x_{c,h}), 10, 7, 1, 0.2, 5)$. The concentration parameter α in the Dirichlet process for G has a Gamma distribution and is independent of all other parameters

[14] Their posterior estimate is based on aggregate macroeconomic time series data, without using aggregate consumption series, while I use panel data on household food consumption.

a priori, $\alpha \sim Ga(10, 2)$. This prior implies that the expected number of clusters (mixtures) is about 4.8 under the DPM model without moment constraints.

Tuning of the SMC sampler. The estimation in this section is based on the SMC sampler. I set tuning parameters of the SMC sampler as follows: the number of stages $N_\phi = 100$; the number of particles $N_p = 3000$; the number of transitions in the mutation step $M = 2$; and the bending coefficient $\eta = 1.5$.

Results: Posterior estimates. Panel (a) in Table 3 presents the mean and quantiles (5% and 95%) of the implied prior distributions as well as the initial prior distribution of ω, γ, and ζ. Unlike the IV regression example in the previous section, the implied distributions are affected by the exponential projection procedure. First, all 90% intervals of the implied prior distributions are shorter than those of the initial prior distributions. Second, for ω and γ, the center of the implied distribution moved slightly to the right, while the center of the implied distribution for ζ stayed the same. Lastly, the support of the implied distribution for γ admits only positive values, even though the initial prior distribution has its support on both negative and positive regions.

Panel (b) in Table 3 presents the posterior moments for ω, γ, and ζ. All posterior intervals are shorter than prior intervals. The posterior mean for ζ is positive (0.84) and its 90% credible set excludes zero, meaning that an increase in the number of children is associated with an increase in the future consumption of the household. The posterior mean for γ is around 5 and is larger than the mean of the initial prior distribution but very close to the mean of the implied prior distribution for γ. However, the posterior interval became shorter compared to that of the implied prior distribution, and therefore, the data are informative about the risk aversion parameter. Figure 6 presents scatter plots of draws for ω, γ, ζ from the implied prior and posterior distributions and shows graphically how the prior belief about these parameters has been updated through the likelihood (data).

The posterior moments of ω and γ based on the second specification (estimation without the instrument for the number of children, "no demographic") are also presented in Table 3. The posterior mean for the risk aversion parameter is smaller than that based on the "full" specification but it is contained in the 90% credible set. Other features of the implied prior and posterior distribution are similar to the results obtained from the "full" specification.

The last row in Table 3 presents the log of the marginal likelihood (ML) from the three specifications. Compared to the non-restricted specification ("no moment"), the other two specifications based on the moment restrictions lead to higher marginal likelihood. This

Table 3 Estimating an Euler Equation

	Initial Prior		Full specification		No demographic		No moment	
(a) Implied Prior	Mean	90%	Mean	90%	Mean	90%	Mean	90%
ω	-2	$[-3.64, -0.36]$	-1.2	$[-2.44, -0.29]$	-1.2	$[-2.51, -0.20]$	-	-
γ	4	$[-2.58, 10.6]$	5.4	$[2.64, 8.24]$	4.9	$[1.99, 7.85]$	-	-
ζ	0	$[-1.64, 1.64]$	0	$[-1.38, 1.37]$	-	-	-	-

	Initial Prior		Full specification		No demographic		No moment	
(b) Posterior	Mean	90%	Mean	90%	Mean	90%	Mean	90%
ω	-2	$[-3.64, -0.36]$	-0.58	$[-0.73, -0.47]$	-0.51	$[-0.67, -0.40]$	-	-
γ	4	$[-2.58, 10.6]$	5.6	$[4.25, 7.15]$	4.5	$[3.01, 6.58]$	-	-
ζ	0	$[-1.64, 1.64]$	0.84	$[0.31, 1.44]$	-	-	-	-

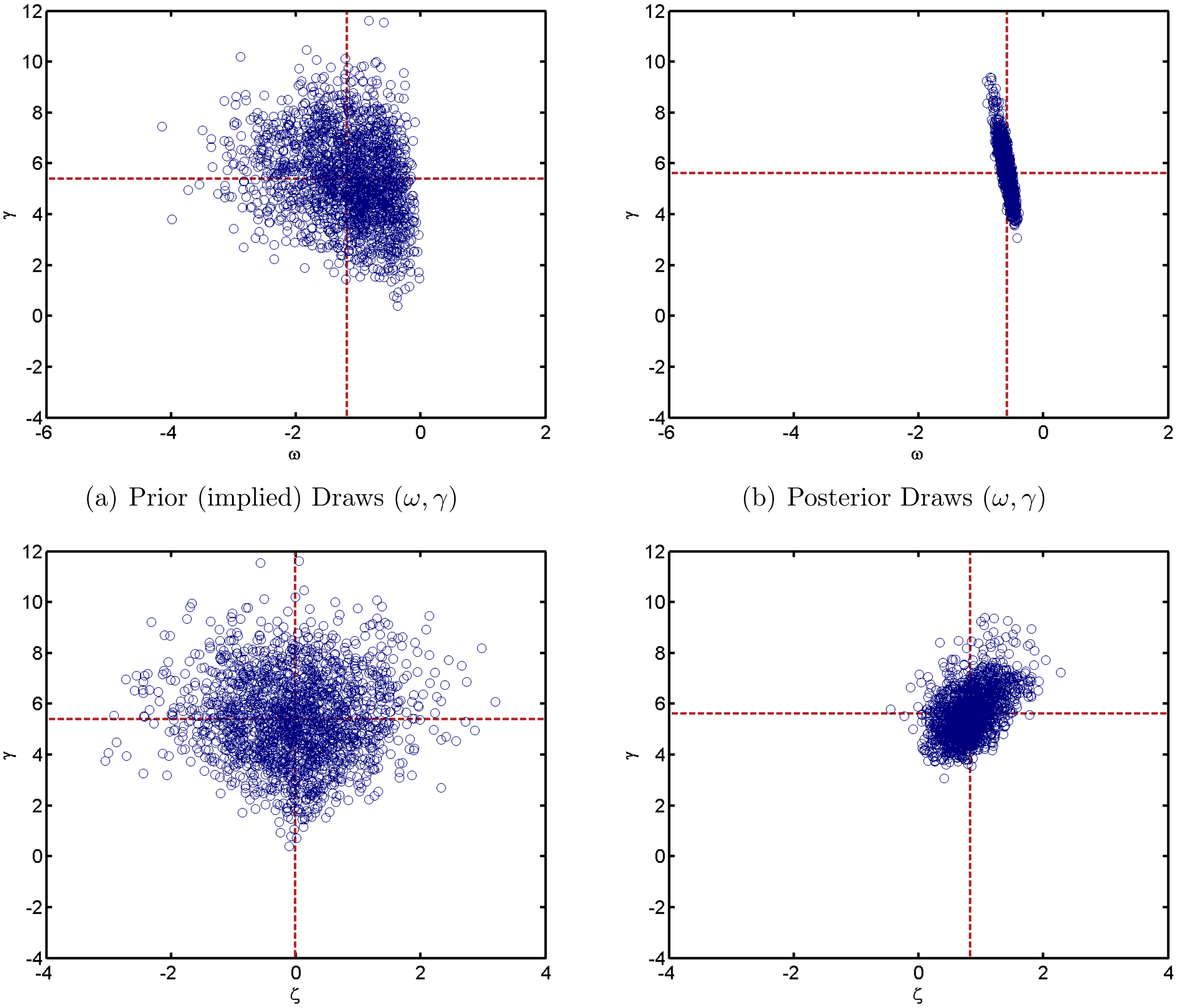
(c) log(ML)	Full	No demographic	No moment
	-2845.31	-2832.63	-2876.50

Notes: This table reports moments (means and 90% equal tail credible sets) of the prior, the implied prior, and the posterior distribution of the parameters in the Euler equation moment conditions based on the SMC sampler. Three model specifications are considered in this table. The first specification ("Full") refers to the estimation based on instruments $Z_t = [1, \Delta c_t, \Delta X_{t+1}]$ and estimates (ω, γ, ζ). The second specification ("No demographic") refers to estimation based on the same instruments without the demographic variable ΔX_{t+1} and estimates only (ω, γ). The last specification ("No moment") estimates only the underlying data generating process without moment restrictions and is a standard DPM model estimation. For all specifications the marginal likelihood (ML) is computed. Tuning parameters for the SMC sampler are the following: the number of stages $N_\phi = 100$; the number of particles $N_p = 3000$; the number of transitions in the mutation step $M = 2$; and the bending coefficient $\eta = 1.5$.

shows that the model fit improves by introducing the Euler equation moment condition. More specifically, since the marginal likelihood can be decomposed into the product of one-step-ahead predictive likelihoods, this finding can be interpreted as the improvement in the model's prediction performance achieved by the Euler equation restriction. Even though the Euler equation restriction improves model fit, not all Euler equation-based moment restrictions are equally useful. Comparing the marginal likelihood between the "full" and "no demographic" specifications, it turns out that the moment restriction based on the number of children instrument reduced log(ML) by about 12.7.

Another important use of the marginal likelihood is for the Bayesian model averaging. Bayesian model averaging provides a coherent decision-theoretic approach to estimation and inference about unknown parameters. More important, it takes account of model uncertainty by taking weighted averages of a quantity of interest. The weights are computed using the posterior model probability which is proportional to a product of the marginal likelihood and the model prior probability. For example, under the quadratic loss function, the optimal risk-

Figure 6 Scatter Plots of draws for (ω, γ, ζ)



(a) Prior (implied) Draws (ω, γ)

(b) Posterior Draws (ω, γ)

(c) Prior (implied) Draws (ζ, γ)

(d) Posterior Draws (ζ, γ)

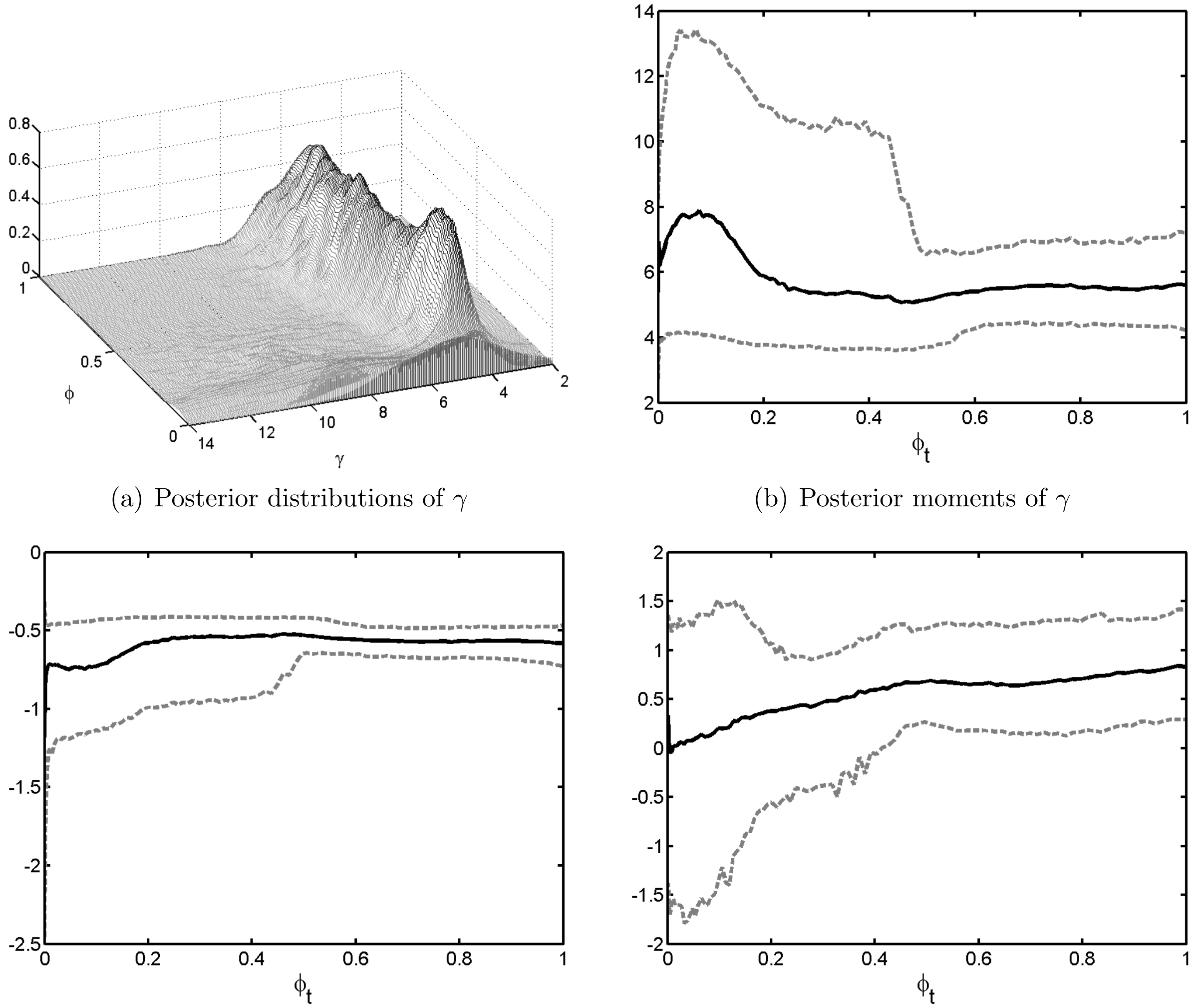
Notes: Scatter plot of draws of (ω, γ, ζ) from implied prior (left) and posterior (right) distributions. The dashed red lines show prior and posterior means. All outputs are based on the SMC sampler and the "full" specification.

aversion parameter estimate is simply a weighted average of the two posterior means of the risk-aversion parameter based on $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$E[\gamma|X] = \underbrace{p(\mathcal{M}_F|X)}_{=0.000003} \times E[\gamma|\mathcal{M}_F, X] + \underbrace{p(\mathcal{M}_N|X)}_{=0.999997} \times E[\gamma|\mathcal{M}_N, X] \approx 4.5.$$

where $\mathcal{M}_F$ is the "Full specification" and $\mathcal{M}_N$ is the "No demographic" model. $p(\mathcal{M}_i|X)$ denotes the model i's posterior model probability. In this example, the Bayesian model averaging estimate for the risk aversion parameter is almost identical to the one obtained

Figure 7 Evolution of power posteriors



(a) Posterior distributions of γ

(b) Posterior moments of γ

(c) Posterior moments of ω

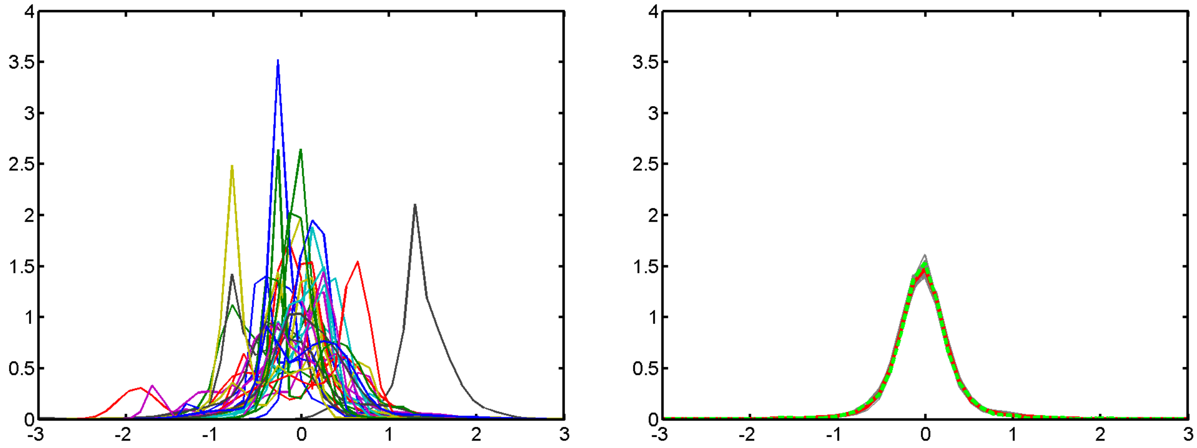
(d) Posterior moments of ζ

Notes: Evolution of the posterior moments for ω, γ, and ζ over the tempering schedule. All outputs are based on the SMC sampler and the "full" specification.

based on $\mathcal{M}_N$ because the marginal likelihood of $\mathcal{M}_N$ far dominates that of $\mathcal{M}_F$.

Figure 7 shows the evolution of the particle approximation to the intermediate posterior distributions from the SMC sampler. Panel (a) presents the evolution of the marginal posterior distribution of γ over the tempering schedule. When $\phi_t = 0$, the marginal power posterior distribution is simply the implied prior distribution of γ and is widely distributed. Then, as ϕ_t increases, the SMC sampler injects the data information gradually and the power posterior distribution becomes narrower and closer to the posterior distribution. Panel (b) conveys a similar information in a different view. This panel shows the evolution of posterior moments (mean and quantiles) over the tempering schedule. At the beginning stages, the

Figure 8 Draws for the density of consumption growth in 1981



(a) Prior Draws

(b) Posterior Draws

Notes: Each draw from the prior/posterior is transformed into a marginal density of y_i and it is evaluated at 100 equally-spaced grid points from $[-3, 3]$. The left panel shows 50 draws from the prior distribution. The right panel shows 50 draws from the posterior distribution as well as the point-wise posterior mean of the density (thick red line). The green dashed line is the kernel density estimate using the same data with Silverman's optimal bandwidth. All outputs are based on the SMC sampler and the "full" specification.

marginal posterior distribution of γ becomes wider with higher mean. A few stages later, the intermediate posterior distributions gradually narrow down until about $\phi_t = 0.45$. Between $\phi_t = 0.45$ and 0.5 there is an abrupt drop in the posterior mean and the standard deviation of the power posterior. Then, its mean gradually increases until the end of the algorithm. Similar abrupt drops in dispersion of the intermediate distributions can be seen from the evolution of the marginal posterior distributions for ω and ζ as well.

After posterior simulation, one can estimate quantities other than moments of the parameters in the moment condition. One important quantity is the density estimates for the underlying data generating process, that is, the density of consumption growth. Figure 8 shows 50 draws of density functions of consumption growth in 1981 from the implied prior (left panel) and the posterior distribution (right panel). Density draws from the prior distribution have various shapes while density draws from the posterior distribution are concentrated around its posterior mean (thick solid line).

8 Concluding Remarks

I have developed practical Bayesian econometric procedures for moment condition models under a far-reaching class of assumptions about the underlying data distribution. Building on the exponentially tilted DPM model of Kitamura and Otsu (2011), I flexibly model the underlying data distribution using a mixture of parametric distributions with random mixture weights restricted by exponential tilting projection. I first show that the baseline *i.i.d.* framework for moment condition models can naturally be extended to more complicated data structures, including models with serially dependent data and laltent variables, through judicious choice of the kernel functions. Then, I developed simulation-based posterior sampling algorithms based on Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods. The SMC algorithm provides a way to computing marginal likelihood and further gives a coherent approach to Bayesian moment selection and Bayesian model averaging.

The proposed posterior sampling algorithms are compared on simulated data. All of the posterior samplers produce almost identical posterior moment estimates. However, the samplers are differentiated by their efficiency, which requires the econometrician to make trade-offs between efficiency and practicality. For example, the current version of the SMC sampler is less efficient than the MCMC-based samplers, but it maintains utility because it produces the marginal likelihood, which is an important object in posterior analysis. It is not obvious how to compute this object based on the output from currently-known MCMC-based samplers.

I also illustrate exactly how one can use the marginal likelihood for posterior analysis based on both simulated and real data. Using simulated data generated from a dynamic Euler equation, the computed marginal likelihood correctly distinguishes the correctly specified moment condition model from models with invalid moment conditions. In application section, I use household-level annual food consumption and the change in the number of children taken from the Panel Study of Income Dynamics (PSID) and find that the Euler equation restrictions are favored by the data.

There are several directions for future research. On the computational side, there is potential room for improvement as regards the efficiency of the SMC sampler adopted in this paper by tweaking its tuning parameters, such as the number of stages N_ϕ or the bending coefficient λ of the tempering schedule. As Herbst and Schorfheide (2014) point out, these tuning parameters are the keys in constructing an efficient SMC algorithm. Currently, I am investigating the possibility of improving the SMC algorithm by explaining the effect of the tuning parameters on the efficiency of the algorithm in the context of the MR-DPM model.

Second, even though this paper mainly focuses on finite-sample posterior analysis for moment condition models, it leads to some interesting and important theoretical research questions such as the asymptotic properties of a Bayes estimator for the finite-dimensional parameter β based on the MR-DPM modeling framework. Kitamura and Otsu (2011) provide conditions for the posterior consistency of β, but their results are limited to the *i.i.d.* environment. Hence, extending the results to the more complicated data structures considered in this paper is an essential task. In addition, it would be interesting to compare the limiting behavior of the moment selection procedure proposed in this paper to that of frequentist moment selection procedures found in the literature (e.g., Andrews, 1999).

Third, the MR-DPM model could be extended to analyze misspecified moment condition models. One plausible approach is to introduce a new hyperparameter τ into the moment condition modeling framework,

$$E[g(X, \beta)] = \tau.$$

The parameter vector τ could then be modeled as an unknown random vector which captures the degree of misspecification. The posterior distribution of τ can be interpreted as updated belief about the degree of misspecification in each of the moment conditions. One of attractive features of this approach is that it can reflect idiosyncratic beliefs about each separate moment condition. This is an important extension, because researchers might be more confident about certain theoretically well-founded moment conditions than about others.

Finally, it would be interesting to extend the implementation of the proposed algorithm to more complex models with latent variables. The illustrated algorithm for the models with latent variables in this paper works when the latent variables are exogenously given. Even though this class of models includes various economic applications, there are some econometric models that require a more complicated relationship between the observables and exogenous variables. One example of such a model is the Dynamic Stochastic General Equilibrium (DSGE) model with endogenous state variables. I plan to pursue several of these lines of research in the near future.

References

Alan, S., K. Atalay, and T. F. Crossley (2012): “Euler Equation Estimation on Micro Data,” *Working paper.*

Alan, S., O. Attanasio, and M. Browning (2009): “Estimating Euler Equations with Noisy Data: Two Exact GMM Estimators,” *Journal of Applied Econometrics*, 24, 309–324.

Andrews, D. (1999): “Consistent Moment Selection Procedures for Generalized Method of Moments Estimation,” *Econometrica*, 67, 543–563.

Antoniak, C. E. (1974): “Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems,” *The Annals of Statistics*, 2, 1152–1174.

Antoniano-Villalobos, I. and S. G. Walker (2014): “A Nonparametric Model for Stationary Time Series,” in *The Contribution of Young Researchers to Bayesian Statistics*, Springer, vol. 63, 3–6.

Aruoba, S. B., L. Bocola, and F. Schorfheide (2013): “Assessing DSGE Model Nonlinearities,” Tech. rep., National Bureau of Economic Research.

Atchadé, Y. F. and J. S. Rosenthal (2005): “On Adaptive Markov Chain Monte Carlo Algorithms,” *Bernoulli*, 11, 815–828.

Caron, F., M. Davy, and A. Doucet (2007): “Generalized Polya Urn for Time-varying Dirichlet Process Mixtures,” *arXiv preprint arXiv:1206.5254.*

Carvalho, C. M., H. F. Lopes, N. G. Polson, and M. Taddy (2010): “Particle Learning for General Mixtures,” *Bayesian Analysis*, 5, 709–740.

Chamberlain, G. and G. Imbens (2003): “Nonparametric Applications of Bayesian Inference,” *Journal of Business & Economic Statistics*, 21, 12–18.

Chernozhukov, V. and C. Hansen (2006): “Instrumental Quantile Regression Inference for Structural and Treatment Effect Models,” *Journal of Econometrics*, 132, 491–525.

Chib, S. and E. Greenberg (1994): “Bayes Inference in Regression Models with $ARMA(p, q)$ Errors,” *Journal of Econometrics*, 64, 183–206.

——— (2010): “Additive Cubic Spline Regression with Dirichlet Process Mixture Errors,” *Journal of Econometrics*, 156, 322–336.

Chib, S. and B. H. Hamilton (2002): “Semiparametric Bayes Analysis of Longitudinal Data Treatment Models,” *Journal of Econometrics*, 110, 67–89.

Choi, H.-s. (2013): “Expert Information and Nonparametric Bayesian Inference of Rare Events,” Working paper.

Chopin, N. (2002): “A Sequential Particle Filter Method for Static Models,” *Biometrika*, 89, 539–552.

Conley, T., C. Hansen, R. McCulloch, and P. Rossi (2008): "A Semi-Parametric Bayesian Approach to the Instrumental Variable Problem," *Journal of Econometrics*, 144, 276–305.

Csiszàr, I. (1967): "On Topological Properties of f-Divergences," *Studia Scientiarum Mathematicarum Hungaria*, 2, 329–339.

Dey, D., P. Müller, and D. Sinha (1998): *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer.

Doucet, A. and A. M. Johansen (2009): "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later," *Handbook of Nonlinear Filtering*, 12, 656–704.

Durham, G. and J. Geweke (2011): "Massively Parallel Sequential Monte Carlo for Bayesian Inference," Working paper.

Durham, G. B. and J. Geweke (2014): "Adaptive Sequential Posterior Simulators for Massively Parallel Computing Environments," in *Advances in Econometrics*, ed. by I. Jeliazkov and D. Poirier, vol. 35, forthcoming.

Ferguson, T. S. (1974): "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 615–629.

Florens, J.-P. and A. Simoni (2012): "Gaussian Processes and Bayesian Moment Estimation," Working paper.

Gallant, A. R., R. Giacomini, and G. Ragusa (2014): "Generalized Method of Moments with Latent Variables," Working Paper.

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

Ghosal, S. (2010): "The Dirichlet Process, Related Priors and Posterior Asymptotics," in *Bayesian Nonparametrics*, ed. by N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, Cambridge University Press, chap. 2, 35–79.

Giacomini, R. and G. Ragusa (2014): "Theory-Coherent Forecasting," *Journal of Econometrics*, 182, 145–155.

Griffin, J. E. (2014): "An Adaptive Truncation Method for Inference in Bayesian Nonparametric Models," *Statistics and Computing*, forthcoming.

Griffin, J. E., F. A. Quintana, and M. F. Steel (2011): "Flexible and Nonparametric Modeling," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. Van Dijk, Oxford University Press, chap. 4, 125–182.

Griffin, J. E. and M. F. Steel (2004): "Semiparametric Bayesian Inference for Stochastic Frontier Models," *Journal of Econometrics*, 123, 121–152.

Griffin, J. E. and M. F. J. Steel (2006): "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

——— (2011): "Stick-Breaking Autoregressive Processes," *Journal of econometrics*, 162, 383–396.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Herbst, E. and F. Schorfheide (2014): "Sequential Monte Carlo Sampling for DSGE Models," *Journal of Applied Econometrics*, forthcoming.

Hirano, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 781–799.

Ishwaran, H. and L. F. James (2001): "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

——— (2002): "Approximate Dirichlet Process Computing in Finite Normal Mixtures," *Journal of Computational and Graphical Statistics*, 11, 508–532.

Jacquier, E., N. G. Polson, and P. E. Rossi (2002): "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business & Economic Statistics*, 20, 69–87.

Jensen, M. J. and J. M. Maheu (2014): "Estimating a Semiparametric Asymmetric Stochastic Volatility Model with a Dirichlet Process Mixture," *Journal of Econometrics*, 178, 523–538.

Kim, J.-Y. (2002): "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, 107, 175 – 193.

Kitamura, Y. (2006): "Empirical Likelihood Methods in Econometrics: Theory and Practice," Yale University, Cowles Foundation for Research in Economics.

Kitamura, Y. and T. Otsu (2011): "Bayesian Analysis of Moment Condition Models using Nonparametric Priors," Working paper.

Lancaster, T. and S. Jun (2010): "Bayesian Quantile Regression Methods," *Journal of Applied Econometrics*, 25, 287–307.

Lazar, N. (2003): "Bayesian Empirical Likelihood," *Biometrika*, 90, 319.

Müller, P., A. Erkanli, and M. West (1996): "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67–79.

Müller, P. and F. A. Quintana (2004): "Nonparametric Bayesian Data Analysis," *Statistical Science*, 19, 95–110.

Norets, A. (2010): "Approximation of Conditional Densities by Smooth Mixtures of Regressions," *The Annals of Statistics*, 38, 1733–1766.

Norets, A. and J. Pelenis (2012): "Bayesian Modeling of Joint and Conditional Distributions," *Journal of Econometrics*, 168, 332–346.

——— (2013): "Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures," *Econometric Theory*, 30, 1–43.

Oryshchenko, V. and R. J. Smith (2013): "Generalized Empirical Likelihood-Based Kernel Density Estimation," Nuffield College Economics Working Paper.

Patel, J. and C. Read (1996): *Handbook of the Normal Distribution*, vol. 150, CRC Press.

Pelenis, J. (2014): "Bayesian Regression with Heteroscedastic Error Density and Parametric Mean Function," *Journal of Econometrics*, 178, 624–638.

Roberts, G. and A. Smith (1994): "Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms," *Stochastic Processes and their Applications*, 49, 207–216.

Robertson, J. C., E. W. Tallman, and C. H. Whiteman (2005): "Forecasting using Relative Entropy," *Journal of Money, Credit and Banking*, 37, 383–401.

Rodríguez, A. and D. B. Dunson (2011): "Nonparametric Bayesian Models Through Probit Stick-Breaking Processes," *Bayesian Analysis*, 6, 145–178.

Schennach, S. (2005): "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika*, 92, 31.

Sethuraman, J. (1994): "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Taddy, M. A. (2008): "Bayesian Nonparametric Analysis of Conditional Distributions and Inference for Poisson Point Processes," Ph.D. thesis, University of California, Santa Cruz.

——— (2010): "Autoregressive Mixture Models for Dynamic Spatial Poisson Processes: Application to Tracking Intensity of Violent Crime," *Journal of the American Statistical Association*, 105, 1403–1417.

Taddy, M. A. and A. Kottas (2010): "A Bayesian Nonparametric Approach to Inference for Quantile Regression," *Journal of Business & Economic Statistics*, 28, 357–369.

Tierney, L. (1994): "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1728.

Tiwari, R. C., S. Jammalamadaka, and S. Chib (1988): "Bayes Prediction Density and Regression Estimation - A Semiparametric Approach," *Empirical Economics*, 13, 209–222.

Villani, M., R. Kohn, and P. Giordani (2009): "Regression Density Estimation using Smooth Adaptive Gaussian Mixtures," *Journal of Econometrics*, 153, 155–173.

Villani, M., R. Kohn, and D. J. Nott (2012): "Generalized Smooth Finite Mixtures," *Journal of Econometrics*, 171, 121–133.

Zellner, A. and J. Tobias (2001): "Further Results on Bayesian Method of Moments Analysis of The Multiple Regression Model," *International Economic Review*, 42, 121–139.

Appendix

A Details of MCMC Samplers

This section describes details of the MCMC-based samplers (Algorithm 2 and 4). The J-truncated MR-DPM model contains the following paramters: $(\theta, \beta, V, \alpha, \psi)$. The posterior distribution has the following form:

$$p(\theta, \beta, V, \alpha, \psi|X) \propto p(X|\theta, \beta, V)p(\theta|\psi)p(V|\alpha)p(\beta)I(\vec{0} \in H(\theta, \beta))p(\psi)p(\alpha),$$

where $p(X|\theta, \beta, V)$ is the likelihood function given by Equation 13 for an *i.i.d.* model and Equation 19 for a time-dependent model. I denote X as $x_{1:N}$ or $x_{1:T}$ depending on the context. Define the likelihood function as

$$f(\mu, \Sigma, \beta, V) = \prod_{i=1}^{N} \left(\sum_{j=1}^{J} \widetilde{q}_j(\theta, \beta, V) k(x_i; \theta_j) \right), \tag{A.1}$$

where the dimension of x_i is $d \times 1$: for notational purposes, I will denote by:

$f_N(\cdot; m, B)$ is a multivariate normal density with mean m and variance B;
$f_{IW}(\cdot; df, S)$ is an inverse Wishart density with degrees of freedom df and scale S;
$f_W(\cdot; df, S)$ is a Wishart density with degrees of freedom df and scale S;
$f_G(\cdot; a, b)$ is a Gamma density with parameters a and b.

Note that the pre-tilting mixture probability q is a function of independent *Beta* draws V,

$$q_1 = V_1;\ q_j = V_j \prod_{r=1}^{j-1}(1 - V_r),\ j = 2, ..., J-1;\ q_J = \prod_{r=1}^{J-1}(1 - V_r), \quad V_l \sim Beta(1, \alpha).$$

For notational convenience, I will denote $q(V)$ as q without its argument. The tilted mixture probability, $\widetilde{q} = \widetilde{q}(\theta, \beta, V)$, is defined by an implicit function induced by the exponential tilting projection:

$$\widetilde{q}_j = \frac{\exp\left(\lambda(\theta, \beta, V)'\widetilde{g}(\beta,\ \theta_j)\right)}{\sum_{j=1}^{J} q_j \exp\left(\lambda(\theta, \beta, V)'\widetilde{g}(\beta,\ \theta_j)\right)} q_j,$$

where

$$\lambda(\theta, \beta, V) = \arg\min_{\lambda} \sum_{j=1}^{J} q_j \exp\left(\lambda'\widetilde{g}(\beta, \theta_j)\right).$$

The first part of this section describes the basic sampler (Algorithm 2) and the second part of this section describes the data-augmented version of the sampler (Algorithm 4).

A.1 Basic sampler

The basic sampler is based on the Metropolis-Hastings-within-Gibbs algorithm iterated over the parameter block $[\theta_1, \theta_2, ..., \theta_J, \beta, V_1, V_2, ..., V_J, \alpha, \psi]$. The full conditionals below are derived under the multivariate normal kernel with mean μ and variance-covariance matrix Σ where $\theta_j = (\mu_j, \Sigma_j)$ and $k(x_i; \theta_j) = f_N(x_i; \mu_j, \Sigma_j)$ in conjunction with conjugate priors for hyperparameters $\psi = (m, B, S)$ presented in section 4.2. Extension to the other cases can be done easily by modifying the algorithm described below.

Updating μ_j for $j = 1, ..., J$. The full conditional posterior density of μ_j is proportional to

$$f(\mu, \Sigma, \beta, V) f_N(\mu_j; m, B) I(\vec{0} \in H(\mu, \Sigma, \beta)),$$

and μ_j is updated via the random-walk Metropolis-Hastings algorithm with the following proposal distribution at the i-th iteration,

$$\mu_j^* = \mu_j^{(i-1)} + e, \quad e \sim N(0, c_{\mu,j}^i \Sigma_{\mu,j}),$$

where $c_{\mu,j}^i$ is a scalar and $\Sigma_{\mu,j}$ is a $d \times d$ matrix. The current implementation of the algorithm sets $\Sigma_{\mu,j}$ to be the identity matrix. The scale parameter $c_{\mu,j}^i$ is adaptively chosen using the following rule (à la Atchadé and Rosenthal, 2005; Griffin, 2014),

$$\log c_{\mu,j}^{i+1} = \log c_{\mu,j}^i + \frac{1}{i^{0.55}}(mh_{\mu,j,i} - 0.3) \tag{A.2}$$

where $mh_{\mu,j,i}$ is the acceptance probability of the i-th MH step for μ_j. Note that this proposal density targets a 30% empirical acceptance rate.

Updating Σ_j for $j = 1, ..., J$. The full conditional posterior density of Σ_j is proportional to

$$f(\mu, \Sigma, \beta, V) f_{IW}(\Sigma_j; s, sS) I(\vec{0} \in H(\mu, \Sigma, \beta)).$$

To perform the random-walk Metropolis-Hastings update step, I first take the Cholesky decomposition of Σ_j:

$$chol(\Sigma_j) = \begin{pmatrix} d_{11,j} & 0 & ... & 0 \\ c_{21,j} & d_{22,j} & ... & 0 \\ ... & ... & ... & 0 \\ c_{p1,j} & c_{p2,j} & ... & d_{pp,j} \end{pmatrix}$$

and then update c and $log(d)$ using proposal distributions at the i-th iteration:

$$c_j^* = c_j^{(i-1)} + e_c, \quad e_c \sim N(0, c_{c,j}^i \Sigma_{c,j})$$
$$\log d_j^* = \log d_j^{(i-1)} + e_d, \quad e_d \sim N(0, c_{d,j}^i \Sigma_{d,j}),$$

where $c_j = [c_{21,j}, ..., c_{d1,j}, ..., c_{d(d-1),j}]'$ and $d_j = [d_{11,j}, ..., d_{dd,j}]'$. $\Sigma_{c,j}$ is set to be the $\frac{1}{2}d(d-1) \times \frac{1}{2}d(d-1)$ identity matrix and $\Sigma_{d,j}$ is set to be the $d \times d$ identity matrix. The scale parameters $c_{c,j}^i$ and $d_{c,j}^j$ are adaptively chosen following the rule in Equation A.2. There are two parameter transformations (Cholesky and log) involved in this step, and Jacobian terms are required to compute the acceptance probability. Note that the determinant of the Jacobian due to the Cholesky decomposition is $2^d \prod_{i=1}^d d_{ii,j}^{d+1-i}$, and the Jacobian of the log transformation is $\prod_{i=1}^d d_{ii,j}$.

Updating β. The full conditional posterior density of β is proportional to

$$f(\mu, \Sigma, \beta, V) p(\beta) I(\vec{0} \in H(\mu, \Sigma, \beta))$$

and β is updated via the random-walk Metropolis-Hastings algorithm with the following proposal distribution at the i-th iteration,

$$\beta^* = \beta^{(i-1)} + e, \quad e \sim N(0, c_\beta^i \Sigma_\beta),$$

where c_β^i is a scalar and Σ_β is a $k \times k$ identity matrix. The scale parameter is adaptively chosen following the rule in Equation A.2.

Updating V_j for $j = 1, ..., J$. The full conditional posterior density of V_j is proportional to

$$f(\mu, \Sigma, \beta, V)(1 - V_j)^{\alpha - 1},$$

where the last term comes from the fact that $V_j \sim Beta(1, \alpha)$. Then V_j is updated via random-walk Metropolis-Hastings algorithm with the following proposal distribution at the

i-th iteration,

$$\Phi^{-1}(V_j^*) = \Phi^{-1}(V_j^i) + e, \quad e \sim N(0, c_{V,j}^i)$$

where $\Phi^{-1}(\cdot)$ is the inverse normal distribution function. The scale parameter $c_{V,j}^i$ is adaptively chosen following the rule in Equation A.2. The Jacobian term due to the inverse normal distribution function is $\phi(V_j)$.

Updating α. The full conditional posterior density of α is of the Gamma family:

$$f_G(\alpha; J + a_\alpha - 1,\ b_\alpha - \log(q_J)), \quad \text{where} \quad \log(q_J) = \log \prod_{j=1}^{J-1} (1 - V_j);$$

α can be drawn directly from this Gamma distribution.

Updating m, B, S. The relevant conditional posterior is,

$$\begin{aligned} p(m, B, S|others, data) &\propto p(\mu, \Sigma|m, B, S)p(m, B, S) \\ &\propto \prod_{j=1}^{J} f_N(\mu_j; m, B) f_N(m; a, B/\kappa) f_{IW}(B; \nu, \Lambda) \\ &\quad \times \prod_{j=1}^{J} f_{IW}(\Sigma_j; s, sS) f_W(S; q, q^{-1}R). \end{aligned}$$

Updating m and B can be done according to the normal-Inverse-Wishart model treating $\mu = (\mu_1, ..., \mu_J)$ as the data. Updating S is done by noting that

$$W_d(\Sigma_j^{-1}|s, (sS)^{-1}) = \frac{1}{2^{(sd)/2}|(sS)^{-1}|^{s/2}\Gamma_d(s/2)} |\Sigma_j^{-1}|^{(s-d-1)/2} \exp(-1/2tr((sS)\Sigma_j^{-1})),$$

and

$$W_d(S|q, q^{-1}R) = \frac{1}{2^{(qd)/2}|q^{-1}R|^{q/2}\Gamma_d(q/2)} |S|^{(q-d-1)/2} \exp(-1/2tr(qR^{-1}S)).$$

The conditional posterior is a Wishart distribution with trace part element:

$$tr(S(s \sum_{j=1}^{J} \Sigma_j^{-1} + qR^{-1}))$$

and determinant part $|S|^{(sJ+q-d-1)/2}$. Therefore, S can be drawn directly from the following Wishart distribution:

$$p(S|others, data) = f_W\left(S; (sJ+q), \left(s\sum_{j=1}^{J}\Sigma_j^{-1} + qR^{-1}\right)^{-1}\right).$$

A.2 Data-augmented sampler

The objective of this section is to derive the full posterior conditionals once configuration variables are introduced and describe posterior sampler in detail. The essence of the sampler is similar to the Blocked-Gibbs sampler for DPM (Ishwaran and James, 2001) but there is an important change due to the exponential tilting procedure. The prior distribution of L differs from the usual truncated DPM case because

$$L_i|\widetilde{q} \sim \sum_{j=1}^{J}\widetilde{q}_j\delta_j(L_i), \quad \text{where } \widetilde{q} = \widetilde{q}(\theta, \beta, V),$$

where $\widetilde{q}$ is a projected mixture probability based on V and $\widetilde{g}(\beta, \theta)$ which can be written (implicitly) as a function of θ, β, V. The probability mass function of the vector L is propositional to

$$p(L|\theta, \beta, V) \propto \prod_{j=1}^{J}\widetilde{q}_j^{M_j}$$

where $M_j = \#|\{i : L_i = j\}|$ for $j = 1, ..., J$.

As in the previous section, the sampler is based on the Metropolis-Hastings-within-Gibbs algorithm and it iterates over the parameter block $[L, \theta_1, \theta_2, ..., \theta_J, \beta, V_1, V_2, ..., V_J, \alpha, \psi]$. The full conditionals below are derived under a multivariate normal kernel with mean μ_j and variance-covariance Σ_j where $\theta_j = (\mu_j, \Sigma_j)$ and $k(x_i; \theta_j) = f_N(x_i; \mu_j, \Sigma_j)$ in conjunction with conjugate priors for hyperparameters $\psi = (m, B, S)$ presented in section 4.2.

Updating for L. First note that,

$$\begin{aligned} p(L|\theta, \beta, V, \alpha, \psi, X) &\propto p(L|\theta, \beta, V, \alpha, \psi)p(X|L, \theta, \beta, V, \alpha, \psi) \\ &\propto p(L|\widetilde{q})p(X|\theta, L). \end{aligned}$$

Each L_i is drawn from a discrete distribution on $\{1, ..., J\}$ with probabilities

$$\widetilde{p}_{ji} \propto \widetilde{q}_j k(x_i; \theta_j),\ j = 1, ..., J.$$

Under the multivariate normal kernel, these probabilities become

$$\widetilde{p}_{ji} \propto \widetilde{q}_j N(x_i;\ \mu_j, \Sigma_j),\ j = 1, ..., J.$$

Updating for θ_j for $j = 1, ..., J$. The conditional posterior distribution of θ can be written as follows (up to a normalizing constant):

$$\begin{aligned} p(\theta|L, \beta, V, \alpha, \psi, X) &\propto p(\theta|L, \beta, V, \alpha, \psi)p(X|\theta, L, \beta, V, \alpha, \psi) \\ &\propto p(L|\theta, \beta, V, \alpha, \psi)p(\theta|\beta, V, \alpha, \psi)p(X|\theta, L, \beta, V, \alpha, \psi), \end{aligned}$$

where $p(\theta|\beta, V, \alpha, \psi) = p(\theta|\psi)$ and $p(X|\theta, L, \beta, V, \alpha, \psi) = p(X|\theta, L)$. One can set the proposal density as

$$p(\theta|\beta, V, \alpha, \psi)p(X|\theta, L, \beta, V, \alpha, \psi) \propto p(\theta|\psi) \prod_{j=1}^{n^*} \prod_{\{i:L_i=L_j^*\}} k(x_i; \theta_{L_j^*}),$$

where n^* is the number of distinct values in the vector L and the L_j^* are these values. Owing to the choice of proposal density, the Metropolis-Hastings weight greatly simplifies as,

$$mh_\theta = \min(r_\theta, 1), \quad \text{where} \quad r_\theta = \frac{p(L^0|\theta^*, \beta^0, V^0)}{p(L^0|\theta^0, \beta^0, V^0)} = \prod_{j=1}^{J} \left(\frac{\widetilde{q}_j^*}{\widetilde{q}_j^0} \right)^{M_j}$$

where θ^* is a proposed draw and $\widetilde{q}^*$ is a tilted mixture probability based on (θ^*, β^0, V^0).

One can also update the θ_j one-by-one in the similar fashion based on the following conditional posterior density:

$$\begin{aligned} p(\theta_j|L, \beta, V, \alpha, \psi, X, \theta_{-j}) &\propto p(\theta_j|\theta_{-j}, L, \beta, V, \alpha, \psi)p(X|\theta, L, \beta, V) \\ &\propto p(L|\theta, \beta, V)p(\theta_j|\theta_{-j}, \beta, V, \alpha, \psi)p(X|\theta, L, \beta, V, \alpha, \psi); \end{aligned}$$

further, $p(\theta_j|\theta_{-j}, \beta, V, \alpha, \psi) = p(\theta_j|\psi)$.

With a multivariate normal kernel with mean μ and variance-covariance matrix Σ in conjunction with conjugate priors for $\psi = (m, B, s)$, the proposal density becomes,

$$p(\theta_j|\psi) \prod_{\{i:L_i=L_j^*\}} k(y_i; \theta_{L_j^*}) = f_N(\mu_j; m, B) f_{IW}(\Sigma_j; s, sS) \prod_{\{i:L_i=L_j^*\}} f_N(x_i; \mu_{L_j^*}, \Sigma_{L_j^*})$$

where its draw $\theta_j = (\mu_j, \Sigma_j)$ can be directly generated from the multivariate normal distribution for μ_j and the Wishart distribution for Σ_j.

Updating β. The full conditional posterior density of β is

$$\begin{aligned} p(\beta|L,\theta,V,\alpha,\psi,X) &\propto p(\beta|L,\theta,V,\alpha,\psi)p(X|\beta,L,\theta,V,\alpha,\psi) \\ &\propto p(L|\theta,\beta,V,\alpha,\psi)p(\beta)\ \underbrace{p(X|\beta,L,\theta,V,\alpha,\psi)}_{\text{not relevant for updating }\beta} \\ &\propto p(L|\theta,\beta,V)p(\beta) \end{aligned}$$

and β is updated using the random-walk Metropolis-Hastings algorithm with the following proposal distribution at the i-th iteration:

$$\beta^* = \beta^{(i-1)} + e, \quad e \sim N(0, c^i_\beta \Sigma_\beta),$$

where c^i_β is a scalar and Σ_β is the $k \times k$ identity matrix. The scale parameter is adaptively chosen following the rule in Equation A.2. Note that the Metropolis-Hastings weight reduces to the following:

$$mh_\beta = \min(r_\beta, 1), \quad \text{where} \quad r_\beta = \frac{p(L^0|\theta^0,\beta^*,V^0)p(\beta^*)}{p(L^0|\theta^0,\beta^0,V^0)p(\beta^0)} = \frac{p(\beta^*)}{p(\beta^0)} \times \prod_{j=1}^{J} \left(\frac{\widetilde{q}^*_j}{\widetilde{q}^0_j}\right)^{M_j}, \tag{A.3}$$

where β^* is the proposed β and $\widetilde{q}^*$ is a tilted mixture probability based on (θ^0, β^*, V^0).

Updating for V_j for $j = 1, ..., J$. The full conditional posterior density of V_j can be written as

$$\begin{aligned} p(V_j|L,\theta,\beta,\alpha,\psi,V_{-j},X) &\propto p(V_j|\theta,L,\beta,\alpha,\psi,V_{-j})\ \underbrace{p(X|L,\theta,\beta,\alpha,\psi,V)}_{\text{not relevant for }V_j\text{ updating}} \\ &\propto \underbrace{p(L|\theta,\beta,V,\alpha,\psi)}_{=p(L|\widetilde{q})}(1-V_j)^{\alpha-1} \end{aligned}$$

where the last term comes from the fact that $V_j \sim Beta(1, \alpha)$. As for the basic sampler, V_j is updated via a random-walk Metropolis-Hastings algorithm with the following proposal distribution at the i-th iteration,

$$\Phi^{-1}(V^*_j) = \Phi^{-1}(V^i_j) + e, \quad e \sim N(0, c^i_{V,j})$$

where $\Phi^{-1}(\cdot)$ is the inverse normal distribution function. The scale parameter $c^i_{V,j}$ is adaptively chosen following the rule in Equation A.2. The Jacobian term due to the inverse normal distribution function is $\phi(V_j)$. Finally, the MH weight is defined in a similar fashion as in Equation A.3.

α and ψ. Updating strategy α and ψ is the same as for the basic sampler.

B Convergence of the Basic Sampler

In this section, I show the convergence of the basic sampler (Algorithm 2). The proof is done by verifying the conditions in Theorem 2 of Roberts and Smith (1994) and Tierney (1994). These are verified in the following two Lemmas.[15] For expositional simplicity, I assume that hyperparameters α and ψ are fixed, that $\alpha \geq 1$, and that the data is univariate with heterogenous location parameters μ_j. The likelihood function for the MR-DPM model is

$$p(X|\mu,\sigma^2,\beta,V) = \prod_{i=1}^{N}\left(\sum_{j=1}^{J}\widetilde{q}_j(\mu,\sigma^2,\beta,V)f_N(x_i;\mu_j,\sigma^2)\right)$$

and the prior distribution is proportional to

$$p(\mu,\sigma^2,\beta,V) \propto \left(\prod_{j=1}^{J} f_N(\mu_j;m,B)\right) f_{IG}(\sigma^2;s,sS)f_N(\beta;m_\beta,V_\beta)f_B(V_j;1,\alpha)I(\vec{0}\in H(\mu,\sigma^2,\beta)),$$

where f_{IG} is an inverse Gamma density function and f_B is a Beta density function. The parameter space is restricted due to the convex hull condition, and I denote it as

$$\varphi \in D = S_{(\mu,\sigma^2,\beta)} \times (0,1)^J$$

where $\varphi = (\mu,\sigma^2,\beta,V)$.

Lemma 1. *Under the stated assumptions, the posterior density of $\varphi = (\mu,\sigma^2,\beta,V)$ defined on the product set $D = S_{(\mu,\sigma^2,\beta)} \times (0,1)^J$ satisfies the following properties:*

1. *$p(\varphi|X)$ is lower semi-continuous at 0. That is, if $p(\varphi'|X) > 0$, there exists an open neighborhood $\varphi \in N_{\varphi'}$ and $\epsilon > 0$ such that, for all $\varphi \in N_{\varphi'}$, $p(\varphi|X) \geq \epsilon > 0$.*

[15] A similar strategy for showing the convergence of the MH-within-Gibbs algorithm is employed by Chib and Greenberg (1994).

2. *$\int p(\varphi|X)d\tau$ is locally bounded for each $\tau \in \{\mu_1, \mu_2, ..., \mu_J, \sigma^2, \beta, V\}$.*
3. *The support D of $p(\varphi|X)$ is arc-connected.*

Proof of Lemma 1.

1. First note that $\widetilde{q}_j(\mu, \sigma^2, \beta, V) > 0$ for some j and $(\mu, \sigma^2, \beta, V) \in S_{(\mu,\sigma^2,\beta)} \times (0,1)^J$. Since $f_N(x_i|\mu_j, \sigma^2)$ is a continuous function on an open set $S_{\mu,\sigma^2,\beta}$ for all j, for any $(\mu', \sigma^{2'}, \beta', V')$ with
$$\sum_{j=1}^{J} \widetilde{q}_j(\mu', \sigma^{2'} \beta', V') f_N(x_i|\mu'_j, \sigma^{2'}) > 0.$$
there exists an open neighborhood of $N_{(\mu',\sigma^{2'},\beta',V')}$ and $\epsilon > 0$ such that, for all $(\mu, \sigma^2, \beta, V) \in N_{(\mu',\sigma^{2'},\beta',V')}$,
$$\sum_{j=1}^{J} \widetilde{q}_j(\mu, \sigma^2, \beta, V) f_N(x_i|\mu_j, \sigma^2) \geq \epsilon > 0.$$
This implies that $\sum_{j=1}^{J} \widetilde{q}_j(\mu, \sigma^2, \beta, V) f_N(x_i|\mu_j, \sigma^2)$ is lower semi-continuous at 0. Moreover, the prior distributions are defined on the product of open sets and continuous densities are imposed on each of these sets. Since the product of functions that are lower semi-continuous at 0 is also lower semi-continuous at 0, $p(\varphi|X)$ is lower semi-continuous at 0.
2. The joint posterior distribution can be written as
$$p(\mu, \sigma^2, \beta, V|X) = \frac{1}{Z} \prod_{i=1}^{N} \left(\sum_{j=1}^{J} \widetilde{q}_j(\mu, \sigma^2, \beta, V) f_N(x_i; \mu_j, \sigma^2) \right)$$
$$\times \underbrace{\left(\prod_{j=1}^{J} f_N(\mu_j; m, B) \right) f_{IG}(\sigma^2; s, sS) p(\beta) f_B(V_j; 1, \alpha) I(\vec{0} \in H(\mu, \sigma^2, \beta)),}_{=prior(\mu,\sigma^2,\beta,V)}$$

where Z is the normalizing constant and $Z < \infty$. First note that

$$\begin{aligned}
p(\mu, \sigma^2, \beta, V|X) &= \frac{1}{Z} \times prior(\mu, \sigma^2, \beta, V) \times \prod_{i=1}^{N} \left(\sum_{j=1}^{J} \widetilde{q}_j(\mu, \sigma^2, \beta, V) f_N(x_i; \mu_j, \sigma^2) \right) \\
&\leq \frac{1}{Z} \times prior(\mu, \sigma^2, \beta, V) \times \prod_{i=1}^{N} \left(\sum_{j=1}^{J} f_N(x_i; \mu_j, \sigma^2) \right), \quad \because \widetilde{q}_j < 1 \\
&\leq \frac{1}{Z} \times prior(\mu, \sigma^2, \beta, V) \times \prod_{i=1}^{N} \left(\sum_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma^2}} \right), \quad \because \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma^2} \right) < 1 \\
&= \underbrace{\frac{1}{Z} \times prior(\mu, \sigma^2, \beta, V) \times \frac{J^N}{\sqrt{2\pi}} \times (\sigma^{-2})^{N/2}}_{=M(\mu, \sigma^2, \beta, V)},
\end{aligned}$$

The quantity obtained by integrating out any of $(\mu_1, ..., \mu_J, \beta, V_1, ..., V_J)$ from $M(\mu, \sigma^2, \beta, V)$ is locally bounded because

$$\begin{aligned}
pior(\mu, \sigma^2, \beta, V) &\leq \left(\frac{1}{\sqrt{2\pi B}} \right)^J \left(\frac{1}{\sqrt{2\pi V_\beta}} \right) \times \prod_{j=1}^{J} \frac{(1 - V_j)^{\alpha - 1}}{B(\alpha, \beta)} \times f_G(\sigma^2; s, sS), \\
&\leq \left(\frac{1}{\sqrt{2\pi B}} \right)^J \left(\frac{1}{\sqrt{2\pi V_\beta}} \right) \times \frac{1}{B(1, \alpha)^J} \times f_G(\sigma^2; s, sS) \quad \because \alpha \geq 1
\end{aligned}$$

where $B(\cdot, \cdot)$ is a Beta function. Moreover, for any even number N,

$$\begin{aligned}
\int M(\mu, \sigma^2, \beta, V) d\sigma^2 &\leq \left(\frac{1}{\sqrt{2\pi B}} \right)^J \left(\frac{1}{\sqrt{2\pi V_\beta}} \right) \times \frac{1}{B(1, \alpha)^J} \int (\sigma^{-2})^{N/2} f_{IG}(\sigma^2; s, sS) d\sigma^2 \\
&= \left(\frac{1}{\sqrt{2\pi B}} \right)^J \left(\frac{1}{\sqrt{2\pi V_\beta}} \right) \times \frac{1}{B(1, \alpha)^J} \int (\sigma^{-2})^{N/2} f_G(\sigma^{-2}; s, (sS)^{-1}) d\sigma^{-2} \\
&\leq \left(\frac{1}{\sqrt{2\pi B}} \right)^J \left(\frac{1}{\sqrt{2\pi V_\beta}} \right) \times \frac{1}{B(1, \alpha)^J} \frac{(sS)^{-s}\,\Gamma(s + N/2)}{\Gamma(s)(sS)^{-(s+N/2)}} \\
&< \infty,
\end{aligned}$$

where $\Gamma(\cdot)$ is a Gamma function.

3. By Assumption 3, $S_{(\mu, \sigma^2, \beta)}$ is arc-connected. Since a set $(0, 1)^J$ is arc-connected, D is also arc-connected.

□

Lemma 2. *The proposal distributions of the Metropolis-Hastings step in the Algorithm 2 for $(\mu_1, ..., \mu_J, \sigma^2, \beta, V_1, ..., V_J)$ satisfy the following. Let $\tau \in \{\mu_1, ..., \mu_J, \sigma^2, \beta, V_1, ..., V_J\}$. For every point $\tau \in Supp(\tau)$ and every $A \subset Supp(\tau)$ with the property $\int_A p(\tau|\varphi_{-\tau}, X) d\tau > 0$, it* *is the case that*

$$\int_A q(\tau^*|\tau^0, \varphi^0_{-\tau}) d\tau^* > 0.$$

Proof of Lemma 2. This holds as long as the random-walk MH proposal density is used, which is true for Algorithm 2. □

C Integrated Moment Conditions and Examples

This section discusses how to compute the moment conditions written in the double integral of Equation 3, which I call an integrated moment condition:

$$\widetilde{g}(\beta, \theta) = \int g(x, \beta) k(x; \theta) dx.$$

To solve the projection problem in Equation 3, one needs to be able to evaluate this function at any point in the support of (β, θ). If the expression does not have a closed form, it can be computed by using numerical methods (such as Gaussain quadrature). Note that a closed form is possible with a multivariate normal kernel in various cases. These include linear instrumental variable (IV) regression, quantile regression, instrumental quantile regression, and the Euler equation for consumption presented in the main text. Here I illustrate how to obtain a closed form of the integrated moment function. Through out the section, I use a (multivariate) normal kernel density,

$$k(x; \theta_j) = f_N(x; \mu_j, \Sigma_j)$$

where $\theta_j = (\mu_j, \Sigma_j)$. The presented examples cover 1) Location model; 2) Linear regression; 3) Linear IV regression; 4) Quantile regression; 5) Quantile IV regression; 6) Euler equation model.

C.1 Location model

Let the moment condition be

$$E[x - \beta] = 0.$$

Then the integrated moment function is

$$\widetilde{g}(\theta, \beta) = \int (x - \beta)\ f_N(x; \mu, \Sigma)\ dx = \mu - \beta.$$

C.2 Linear regression

Consider the following model,

$$y_i = x_i'\beta + u_i, \quad \text{where} \quad E[u_i|x_i] = 0.$$

Let the moment conditions be

$$\begin{aligned} E[y - \alpha - x'\beta] &= 0, \\ E[x(y - \alpha - x'\beta)] &= 0. \end{aligned}$$

Then the first integrated moment function is

$$\mu_y - \alpha - \mu_x'\beta = 0,$$

and the second integrated moment function is

$$\Sigma_{xy} + \mu_x\mu_y - \alpha\mu_x - (\Sigma_{xx} + \mu_x\mu_x')\beta = 0.$$

C.3 IV regression

Consider the following model,

$$y_i = x_i'\beta + u_i, \quad \text{where} \quad E[u_i|z_i] = 0,$$

and $E[x_i z_i'] = 0$. Then the moment conditions are

$$\begin{aligned} E[y - \alpha - x'\beta] &= 0, \\ E[z(y - \alpha - x'\beta)] &= 0. \end{aligned}$$

The kernel function for the MR-DPM model is

$$k([y_i, x_i', z_i']', \theta) = f_N([y_i, x_i', z_i']'; \mu, \Sigma).$$

Hence the first integrated moment function is

$$\mu_y - \alpha - \mu_x'\beta = 0,$$

and the second integrated moment function is

$$\Sigma_{zy} + \mu_z \mu_y - \alpha \mu_z - (\Sigma_{zx} + \mu_z \mu_x')\beta = 0.$$

C.4 Euler equation model

In section 6.2, I consider the following model,[16]

$$\Delta c_{t+1} = \omega + \frac{1}{\gamma} \log(1 + r_{t+1}) + \frac{\gamma + 1}{2} (\Delta c_{t+1})^2 + \nu_{t+1}$$

where

$$E[\nu_{t+1}] = 0, \quad E[\Delta c_t \nu_{t+1}] = 0, \quad \text{and} \quad E[r_t \nu_{t+1}] = 0.$$

The kernel function for the MR-DPM model is then

$$k([\Delta c_{t+1}, r_{t+1}, \Delta c_t, r_t]', \mu, \Sigma) = f_N([\Delta c_{t+1}, r_{t+1}, \Delta c_t, r_t]', \mu, \Sigma)$$

with

$$\mu = \begin{pmatrix} \mu_c \\ \mu_r \\ \mu_c \\ \mu_r \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} V & \Gamma \\ \Gamma & V \end{pmatrix},$$

where V and Γ are 2×2 matrices.

First moment condition. $E[\nu_{t+1}] = 0$ leads to the following integrated moment function:

$$\mu_c - \omega - \frac{1}{\gamma}\mu_r - \frac{\gamma + 1}{2}(V_{cc} + \mu_c^2) = 0.$$

Second moment condition. $E[\Delta c_t \nu_{t+1}] = 0$ leads to the second integrated moment function:

$$(\Gamma_{cc} + \mu_c^2) - \omega\mu_c - \frac{1}{\gamma}(\Gamma_{rc} + \mu_r \mu_c) - \frac{\gamma + 1}{2} E[\Delta c_{t+1}^2 \Delta c_t] = 0$$

where

$$E[\Delta c_{t+1}^2 \Delta c_t] = J_0 \mu_c + J_1 (V_{cc} + \mu_c^2) + J_2 (\mu_c^3 + 3\mu_c V_{cc}^2),$$

[16] The set of integrated moment conditions used in section 7 is almost identical to the one derived in this example.

and

$$\begin{aligned}J_0 &= V_{cc} - \Gamma_{cc}V_{cc}^{-1}\Gamma_{cc} + \mu_c^2 - 2\mu_c^2\Gamma_{cc}V_{cc}^{-1} + (\Gamma_{cc}V_{cc}^{-1})^2\mu_c^2\\ J_1 &= 2\mu_c\Gamma_{cc}V_{cc}^{-1} - 2(\Gamma_{cc}V_{cc}^{-1})^2\mu_c\\ J_2 &= (\Gamma_{cc}V_{cc}^{-1})^2.\end{aligned}$$

Third moment condition. $E[r_t\nu_{t+1}] = 0$ leads to the third integrated moment condition:

$$(\Gamma_{cr} + \mu_c\mu_r) - \omega\mu_r - \frac{1}{\gamma}(\Gamma_{rr} + \mu_r^2) - \frac{\gamma+1}{2}E[(\Delta c_{t+1})^2 r_t] = 0$$

where

$$E[\Delta c_{t+1}^2 r_t] = K_0\mu_r + K_1(V_{rr} + \mu_r^2) + K_2(\mu_r^3 + 3\mu_r V_{rr}^2),$$

and

$$\begin{aligned}K_0 &= V_{cc} - \Gamma_{cr}V_{rr}^{-1}\Gamma_{cr} + \mu_c^2 - 2\mu_r\mu_c\Gamma_{cr}V_{rr}^{-1} + (\Gamma_{cr}V_{rr}^{-1})^2\mu_r^2\\ K_1 &= 2\mu_c\Gamma_{cr}V_{rr}^{-1} - 2(\Gamma_{cr}V_{rr}^{-1})^2\mu_r\\ K_2 &= (\Gamma_{cr}V_{rr}^{-1})^2.\end{aligned}$$

C.5 Quantile regression

Consider a τ-th quantile regression with a single regressor

$$P(y \le \alpha(\tau) + \beta(\tau)x|x) = \tau,$$

which implies unconditional moment conditions

$$\begin{aligned}E\left[1\{y \le \alpha(\tau) + \beta(\tau)x\} - \tau\right] &= 0\\ E\left[x\left(1\{y \le \alpha(\tau) + \beta(\tau)x\} - \tau\right)\right] &= 0.\end{aligned}$$

The kernel function for the MR-DPM model is

$$k([y_i, x_i']', \theta) = f_N([y_i, x_i']'; \mu, \Sigma).$$

First moment condition. First, I examine the left hand side of the first moment condition:

$$\begin{aligned}
E[1\{y \leq \alpha + \beta x\}] &= E\left[\Phi\left(\frac{\alpha + \beta x - \mu_{y|x}}{\sigma_{y|x}}\right)\right] \\
&= E\left[\Phi\left(\frac{\alpha + \beta x - \left(\mu_y + \Sigma_{yx}\Sigma_x^{-1}(x - \mu_x)\right)}{\sigma_{y|x}}\right)\right] \\
&= E\left[\Phi\left(\left(\frac{\alpha - \mu_y + \Sigma_{yx}\Sigma_x^{-1}\mu_x}{\sigma_{y|x}}\right) + \left(\frac{\beta - \Sigma_{yx}\Sigma_x^{-1}}{\sigma_{y|x}}\right)x\right)\right] \\
&= \int \Phi(a + bx)\frac{1}{\sigma_x}\phi\left(\frac{x - \mu_x}{\sigma_x}\right)dx \\
&= \int \Phi(a + b\mu_x + b\sigma_x z)\phi(z)dz \\
&= \Phi\left(\frac{a + b\mu_x}{\sqrt{1 + (b\sigma_x)^2}}\right)
\end{aligned}$$

Second moment condition. The left hand side of the second moment condition is

$$E[x1\{y \leq \alpha + \beta x\}] = E[x\Phi(a + bx)], \quad x \sim \mathcal{N}(\mu_x, \sigma_x^2)$$

and a, b are defined as in the previous derivation. The closed form of the above equation can be obtained in the same way as in the quantile IV regression case in the next subsection.

Special case: Estimating quantiles. The above specification encompasses quantile estimation. Set $\beta = 0$. Then the moment condition is

$$P(Y \leq \alpha) = \tau$$

where α is a quantile at τ. The integrated moment condition is then

$$E\left[\Phi\left(\frac{\alpha - \mu}{\sigma^2}\right) - \tau\right] = 0.$$

C.6 Quantile IV regression

Following Chernozhukov and Hansen (2006) and Lancaster and Jun (2010), I consider the following model

$$y = d'\alpha(u) + x'\beta(u), \quad u|x, z \sim Unif(0, 1)$$

in which d is dependent on u, $d'\alpha(\tau)+x'\beta(\tau)$ is strictly increasing in τ, and z is a set of instrumental variables that are independent of u but dependent on d. Then $d'\alpha(\tau)+x'\beta(\tau)$ is the τ th quantile of y conditional on x, z. That is,

$$P(y \le d'\alpha(\tau)+x'\beta(\tau)|x,z)=\tau$$

I consider following unconditional quantile functions

$$\begin{aligned} E\left[x\big(1\{y \le d'\alpha(\tau)+x'\beta(\tau)\}-\tau\big)\right]&=0\\ E\left[z\big(1\{y \le d'\alpha(\tau)+x'\beta(\tau)\}-\tau\big)\right]&=0. \end{aligned}$$

In addition, following Lancaster and Jun (2010), I consider the case where $x_i=1$. The kernel function for the MR-DPM model is

$$k([y_i,d_i,z_i']',\theta)=f_N([y_i,d_i',z_i']';\mu,\Sigma).$$

First moment condition. Using the conditional argument,

$$E[CDF_{y|d}(d'\alpha(\tau)+\beta(\tau)]=\tau,$$

This becomes

$$\int \Phi\left(\frac{d'\alpha+\beta-\mu_{y|d}}{\sigma_{y|d}}\right)f(d)=\tau \iff \int \Phi(ad+b)\frac{1}{\sigma_d}\phi\left(\frac{d-\mu_d}{\sigma_d}\right)dd=\tau$$

where

$$a=\frac{\alpha-\frac{\sigma_{yd}}{\sigma_d^2}}{\sigma_{y|d}} \quad \text{and} \quad b=\frac{\beta-\left(\mu_y-\frac{\sigma_{yd}}{\sigma_d^2}\mu_d\right)}{\sigma_{y|d}}.$$

because $\mu_{y|d}=\mu_y+\frac{\sigma_{yd}}{\sigma_d^2}(d-\mu_d)$. The reparametrization $c=\frac{d-\mu_d}{\sigma_d}$ gives $d=\mu_d+\sigma_d c$ and

$$\int \Phi(\widetilde{a}c+\widetilde{b})\frac{1}{\sigma_d}\phi(c)\frac{dd}{dc}dc=\tau \iff \int \Phi(\widetilde{a}c+\widetilde{b})\phi(c)dc=\tau$$

where $\widetilde{a}=a\sigma_d$ and $\widetilde{b}=a\mu_d+b$. Hence, we have

$$\Phi\left(\frac{\widetilde{b}}{\sqrt{1+\widetilde{a}^2}}\right)=\tau,$$

Second moment condition. Using the conditional argument

$$E\big[z \times E[1\{y \leq d'\alpha + \beta\}|z]\big] = \tau E[z], \tag{A.4}$$

I first examine the inside expectation on the left hand side:

$$\begin{aligned}
E[1\{y \leq d'\alpha + \beta\}|z] &= E\big[CDF_{y|d,z}(d'\alpha + \beta)|z\big] \\
&= E\left[\Phi\left(\frac{d'\alpha + \beta - \mu_{y|d,z}}{\sigma_{y|d,z}}\right)\Big|z\right] \\
&= E\left[\Phi\left(\frac{\alpha d + \beta - \mu_y - \widetilde{\Sigma}_z(z-\mu_z) + \widetilde{\Sigma}_d\mu_d - \widetilde{\Sigma}_d d}{\sigma_{y|d,z}}\right)\Big|z\right] \\
&= E\left[\Phi\left(\underbrace{\left(\frac{\beta - \mu_y - \widetilde{\Sigma}_z(z-\mu_z) + \widetilde{\Sigma}_d\mu_d}{\sigma_{y|d,z}}\right)}_{=b} + \underbrace{\left(\frac{\alpha - \widetilde{\Sigma}_d}{\sigma_{y|d,z}}\right)}_{=a} d\right)\Bigg|z\right] \\
&= \int_{-\infty}^{\infty} \Phi(b + ad)\frac{1}{\sigma_{d|z}}\phi\left(\frac{d - \mu_{d|z}}{\sigma_{d|z}}\right) dd \\
&= \int_{-\infty}^{\infty} \Phi(b + ad)\frac{1}{\sigma_{d|z}}\phi(c)\frac{dd}{dc}dd \qquad \text{where } d = \frac{d - \mu_{d|z}}{\sigma_{d|z}} \\
&= \int_{-\infty}^{\infty} \Phi(b + a(\mu_{d|z} + \sigma_{d|z}c))\phi(c)dc \\
&= \int_{-\infty}^{\infty} \Phi(b + a\mu_{d|z} + a\sigma_{d|z}c)\phi(c)dc \\
&= \Phi\left(\frac{b + a\mu_{d|z}}{\sqrt{1 + (a\sigma_{d|z})^2}}\right) \\
&= \Phi\left(\frac{b + a\left(\mu_d + \Sigma_{dz}\Sigma_z^{-1}(z - \mu_z)\right)}{\sqrt{1 + (a\sigma_{d|z})^2}}\right) \\
&= \Phi\left(\frac{b + a\mu_d - a\Sigma_{dz}\Sigma_z^{-1}\mu_z + a\Sigma_{dz}\Sigma_z^{-1}z}{\sqrt{1 + (a\sigma_{d|z})^2}}\right) \\
&= \Phi\left(\gamma_0 + \gamma_1 z\right)
\end{aligned}$$

where I write

$$\begin{aligned}\mu_{y|d,z} &= \mu_y + \Sigma_{y,(d,z)}\Sigma_{(d,z)}^{-1}\begin{pmatrix} d-\mu_d \\ z-\mu_z \end{pmatrix} \\ &= \mu_y + \left(\widetilde{\Sigma}_d, \widetilde{\Sigma}_z\right)\begin{pmatrix} d-\mu_d \\ z-\mu_z \end{pmatrix} \\ &= \mu_y + \widetilde{\Sigma}_d(d-\mu_d) + \widetilde{\Sigma}_z(z-\mu_z),\end{aligned}$$

Next, the whole term on the left hand side of Equation A.4 is

$$\begin{aligned} E\big[z \times E[1\{y \leq d'\alpha + \beta\}|z]\big] &= E\big[z \times \Phi\left(\gamma_0 + \gamma_1 z\right)\big] \\ &= \int z\Phi(\gamma_0+\gamma_1 z)\frac{1}{\sigma_z}\phi\left(\frac{z-\mu_z}{\sigma_z}\right)dz \\ &= \int \left(\frac{w-\gamma_0}{\gamma_1}\right)\Phi(w)\frac{1}{\sigma_z}\phi\left(\frac{w-\gamma_0-\gamma_1\mu_z}{\gamma_1\sigma_z}\right)\frac{dz}{dw}dw \\ &= \int \left(\frac{w-\gamma_0}{\gamma_1}\right)\Phi(w)\frac{1}{\gamma_1\sigma_z}\phi\left(\frac{w-(\gamma_0+\gamma_1\mu_z)}{\gamma_1\sigma_z}\right)dw \\ &= k_1 + k_2, \end{aligned}$$

where

$$\begin{aligned} k_1 &= \frac{1}{\gamma_1}\int w\Phi(w)\frac{1}{\gamma_1\sigma_z}\phi\left(\frac{w-(\gamma_0+\gamma_1\mu_z)}{\gamma_1\sigma_z}\right)dw \\ &= \frac{1}{\gamma_1}E\big[w\Phi(w)\big] \qquad \text{where } w \sim N(\gamma_0+\gamma_1\mu_z,\ \gamma_1\sigma_z), \end{aligned}$$

and

$$\begin{aligned} k_2 &= -\frac{\gamma_0}{\gamma_1}\int \Phi(w)\frac{1}{\gamma_1\sigma_z}\phi\left(\frac{w-(\gamma_0+\gamma_1\mu_z)}{\gamma_1\sigma_z}\right)dw \\ &= -\frac{\gamma_0}{\gamma_1}E\big[\Phi(w)\big], \qquad \text{where } w \sim N(\gamma_0+\gamma_1\mu_z,\ \gamma_1\sigma_z). \end{aligned}$$

Note that k_1 and k_2 can be computed using the results in section C.7.

C.7 Useful Results

Let $\Phi(\cdot)$ be the cumulative function of the standard normal distribution and $\phi(\cdot)$ be the probability density function of the standard normal distribution.

Result 1.

$$\int_{-\infty}^{\infty} \Phi(a+by)\phi(y)dy = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right).$$

Result 2.

$$\int_{-\infty}^{\infty} \phi(y)\phi\left(\frac{y-\mu}{\sigma}\right)dy = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\sqrt{\frac{2\pi\sigma^2}{1+\sigma^2}}\exp\left(\frac{\frac{\mu^2}{(1+\sigma^2)^2}-\frac{\mu^2}{(1+\sigma^2)}}{2\left(\frac{\sigma^2}{1+\sigma^2}\right)}\right).$$

Result 3. Let $w \sim N(\mu, \sigma^2)$. Then

$$E[w\Phi(w)] = -\frac{1}{\sigma}\frac{a}{b^2} + \frac{1}{\sigma b^2}\underbrace{\int_{-\infty}^{\infty}\phi(y)\phi(a+by)dy}_{\text{use Results 2}} + \frac{a}{\sigma b^2}\underbrace{\int_{-\infty}^{\infty}\phi(y)\Phi(a+by)dy}_{\text{use Result 1}},$$

where

$$a = -\frac{\mu}{\sigma}, \quad \text{and} \quad b = \frac{1}{\sigma}.$$

Result 4. The following result is taken from Patel and Read (1996),

$$\int X\phi(a+bX)dX = -\frac{1}{b^2}\phi(a+bX) - \frac{a}{b^2}\Phi(a+bX).$$