Particle EM for Variable Selection

Revision

Veronika Ročková*

June 18, 2017

Abstract

Despite its long history of success, the EM algorithm has been vulnerable to local entrapment when the posterior/likelihood is multi-modal. This is particularly pronounced in spike-and-slab posterior distributions for Bayesian variable selection. The main thrust of this article is to introduce the Particle EM algorithm, a new population-based optimization strategy that harvests multiple modes in search spaces that present many local maxima. Motivated by non-parametric variational Bayes strategies, Particle EM achieves this goal by deploying an ensemble of interactive repulsive particles. These particles are geared towards uncharted areas of the posterior, providing a more comprehensive summary of its topography than simple parallel EM deployments. A sequential Monte Carlo variant of Particle EM is also proposed that explores a sequence of annealed posteriors by sampling from a set of mutually avoiding particles. Particle EM outputs a deterministic reconstruction of the posterior distribution for approximate fully Bayes inference by capturing its essential modes and mode weights. This reconstruction reflects model selection uncertainty and is supported by asymptotic considerations, which indicate that the requisite number of particles need not be large in the presence of sparsity (when $p > n$).

1 Introduction

The practical costs of Bayesian variable selection can be formidable within the scope of modern analyses. Beyond MCMC, deterministic search approaches (such as the EM algorithm; Dempster et al. (1977)) have further unleashed its practical potential, providing a venue for fast posterior exploration in demanding problems (Ročková and George, 2014a; Ormerod et al., 2014). These methods have primarily focused on finding a global mode that identifies as the "best model". However, the report of a single model will be a misleading reflection of the model uncertainty in a highly multimodal posterior. We move beyond this limitation by proposing a Particle EM ensemble optimization approach to identify a collection of representative models.

*Veronika.Rockova@ChicagoBooth.edu; Veronika Ročková is assistant professor in Econometrics and Statistics at the Booth School of Business of the University of Chicago.

Approaches for Bayesian variable selection and inference stem from the general Bayesian formalism for model uncertainty where a space of models, indexed by binary strings $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, is considered for modeling data $\boldsymbol{Y} \in \mathbb{R}^n$ with $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$, an $n \times p$ matrix of p potential standardized predictors. We assume a classical linear model

$$\pi(\boldsymbol{Y} \mid \alpha, \boldsymbol{\beta}) = \mathcal{N}_n(\mathbf{1}_n\alpha + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n), \tag{1.1}$$

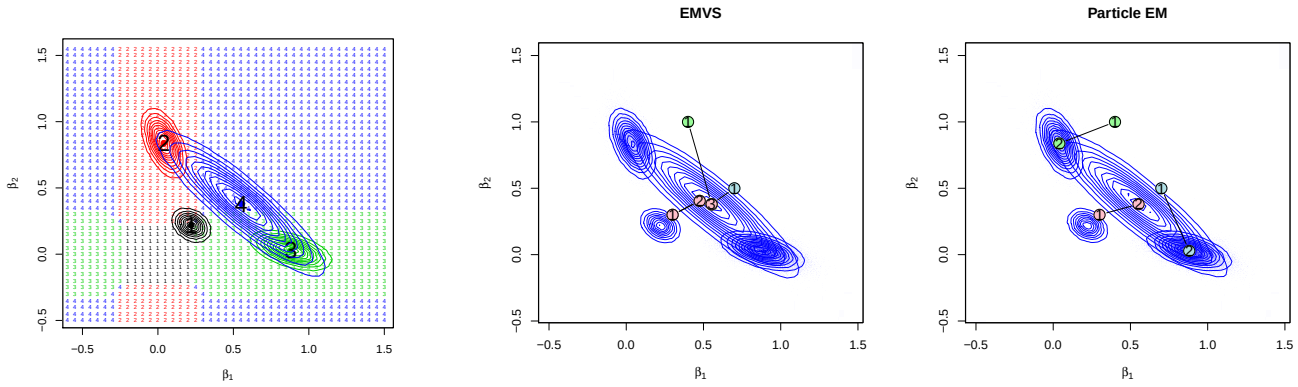
where $\mathbf{1}_n$ is an $n \times 1$ vector of 1's, α is an unknown scalar intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and σ^2 is an unknown positive scalar. We tacitly assume that $\boldsymbol{Y}$ has been centered and thereby omit the intercept α. We endow this model with one of the most popular Bayesian variable selection priors (George and McCulloch, 1993). Namely, as a prior $\pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$, we consider an independent product of Gaussian mixtures

$$\pi(\beta_i \mid \gamma_i) = \gamma_i \phi(\beta_i \mid v_1) + (1 - \gamma_i)\phi(\beta_i \mid v_0), \quad 1 \leq i \leq p, \tag{1.2}$$

where $\phi(\beta \mid v)$ is a Gaussian density centered at zero with variance v. We assume $0 < v_0 < v_1$ so that $\gamma_i = 1$ indicates those $\beta_i's$ which are likely to be the largest.

The continous spike-and-slab prior (1.2) is a predecessor to the more popular point-mass mixtures. Recently, there has been a resurrection of interest in such *continuous* spike-and-slab constructions (Narisetty and He, 2014; Ishwaran and Rao, 2005, 2003, 2011; Ročková and George, 2014a, 2016). Due to the continuity, these priors are more "fluid" for posterior exploration, both with MCMC and optimization. The Gaussian mixture prior has a particular appeal for our proposed optimization due to its conditional conjugacy, where similar implementations with other mixtures are less obvious. Furthermore, continuous mixtures can absorb spurious posterior peaks by performing annealing with dynamic exploration. By varying the spike variance, one can explore an entire path of sliding posteriors as the prior approaches the point-mass limit. Beyond computational advantages, there is theoretical evidence that continuous mixture priors can achieve optimal posterior behavior. These results include model selection consistency in high-dimensional settings (Narisetty and He, 2014), the oracle property of the posterior mean (Ishwaran and Rao, 2005, 2011) as well as optimal posterior concentration (Ročková, 2017; Ročková and George, 2016).

For the prior on σ^2, we consider the inverse gamma prior distribution $\pi(\sigma^2) \sim IG(\eta/2, \eta\nu/2)$ (as in Ročková and George (2014a)). However, here we deploy it in a non-conjugate way (as in George and McCulloch (1993)), where σ^2 enters only the likelihood, not the prior on $\boldsymbol{\beta}$. As a prior



(a) EMVS Convergence Regions (b) EMVS (c) Particle EM

Figure 1: Posterior distribution $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y}, \sigma^2 = 1)$ with $p = 2$, $\boldsymbol{\beta}_0 = (0, 1)'$ and $\texttt{corr}(X_1, X_2) = 0.9$. Left: Convergence regions of EMVS; Middle: Three trajectories of EMVS (independent initialization); Right: Particle EM trajectories

$\pi(\boldsymbol{\gamma})$ we deploy the hierarchical beta-Bernoulli prior $\pi(\gamma_i \mid \theta) \overset{\text{iid}}{\sim} \text{Bern}(\theta)$ with $\theta \sim \mathcal{B}(a, b)$. Post-data uncertainty about model parameters and models is encapsulated in the posterior distributions $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ and $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$, which can be distilled in various ways to solve a variety of problems (Hahn and Carvalho, 2015). The practical implementation of these strategies has continued to rely on some form of posterior simulation.

Sampling from the posteriors $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ and/or $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ is rather straightforward, at least conceptually. George and McCulloch (1997) provide an early overview of various strategies. Since then, numerous Markov Chain Monte Carlo (MCMC) implementations have mushroomed (Carlin and Chib (1995); Madigan et al. (1995); Clyde et al. (2011); Ghosh and Clyde (2011); Griffin et al. (2014) to name just a few). Although success stories abound, there have been persistent difficulties such as local entrapment of single chains in steep local posterior peaks (if they are well-separated), and/or an inability to discover relevant but isolated regions of the model space. These difficulties have been attenuated with the assistance of population-based MCMC (Jasra et al., 2007; Bottolo and Richardson, 2010; Hans et al., 2007; Liang and Wong, 2000) including sequential Monte Carlo implementations (Schafer and Chopin, 2013; Shi and Dunson, 2011; Ma, 2015). However, posterior simulation slows down dramatically for the kind of very high-dimensional problems that are of so much interest today.

Motivated by the success of population-based stochastic search, we introduce a new framework for conquering multi-modal spike-and-slab posteriors. This framework is also population-based but

entirely deterministic. A springboard for our development has been the EMVS procedure of Ročková and George (2014a). EMVS is a single-mode EM hunting strategy and, as such, it has some limitations. First, the report of a single model is a misleading reflection of model uncertainty. Second, the hunt for the global mode can be hampered by local entrapment. These caveats are illustrated in Figure 1. The plots display a posterior landscape $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y}, \sigma^2 = 1)$, assuming (1.1) and (1.2) with $p = 2$ collinear variables where $\boldsymbol{\beta}_0 = (1, 0)'$ and $\sigma^2 = 1$. The posterior terrain is comprised of 4 modes, one for each model, where the global mode (number 3) is associated with the true model $\boldsymbol{\gamma}_0 = (1, 0)'$. Figure 1(a) delineates a tessellation of $\mathbb{R}^2$ into basins of attraction, where initializations from within each box arrive at the same mode. Figure 1(b) illustrates this phenomenon with three EMVS trajectories. Initialized in the same basin, these trajectories gravitate towards the same suboptimal mode. This suggests an intriguing possibility: What would happen if we let the trajectories communicate? In this paper, we introduce a formal framework that allows for repelling interactions. Our proposed Particle EM method generates a multitude of trajectories that are mutually avoiding, thus enhancing the chances of finding a global mode by discovering a more comprehensive set of posterior modes. The success of this strategy is illustrated in Figure 1(c), where Particle EM indeed discovers three different modes, including the global one.

Our main contributions can be summarized as follows:

(a) As a precursor to Particle EM, we propose the Reversed EMVS algorithm, a discrete optimization approach for spike-and-slab variable selection. While the EMVS procedure (Ročková and George, 2014a) treats the binary strings $\boldsymbol{\gamma}$ as missing data and $\boldsymbol{\beta}$ as parameters of interest, Reversed EMVS switches their roles by treating $\boldsymbol{\gamma}$ as parameters of interest and $\boldsymbol{\beta}$ as missing data. This strategy requires only simple closed form updates, targeting directly the discrete model space without thresholding.

(b) Going further, we introduce the Particle EM algorithm, a population-based optimization approach that exhibits both individual and social behavior. Particle EM is motivated as a variational Bayes approach for obtaining the best multi-point approximation to the posterior. As such, Particle EM can be viewed as an elaboration of EM modal estimation, which seeks only single-point approximations. Particles share a common goal (finding essential posterior modes) and realize it by exploring the posterior environment while mutually interacting. The social behavior is mediated through entropy, which serves as a diversifying penalty. This is mani-

fested by local repulsion between particles that embark on the same mode. Particle EM outputs a weighted discrete posterior reconstruction that can be used for approximate fully Bayes inference (model averaging, uncertainty quantification etc.). We provide asymptotic arguments (Supplemental material) that, in the presence of sparsity, the requisite number of particles need not be large.

(c) Bayesian variable selection with spike-and-slab priors has traditionally involved a single posterior distribution of interest (with fixed hyper-parameters). However, an entire trajectory of evolving posteriors is far more informative for variable selection than a single snapshot, particularly in the absence of oracle hyper-parameter tuning (Ročková and George, 2014a). Adopting this perspective, we extend Particle EM to dynamic scenarios by allowing the proliferation of particles through a sequence of sliding distributions as v_0 (the spike variance) is varied.

(d) By forging connections between Particle EM and sequential Monte Carlo (SMC) procedures, we propose a stochastic Particle EM variant which treats particles as random samples rather than posterior modes. Instead of endowing the static spike-and-slab model with an artificial dynamic structure (Schafer and Chopin, 2013; Shi and Dunson, 2011), we consider a sequence of *annealed* posteriors indexed by v_0. In addition, our sequential Monte Carlo variant fosters diversity among particles through a repulsive transition kernel. This aspect provides a promising new route towards improving the mixing of SMC.

(e) Spike-and-slab posteriors can be regarded as non-concave penalized likelihood surfaces, for which regularization plots are used to track the location of a single mode (Ročková and George, 2016). Particle EM moves a step further by tracking multiple modes at the same time. We encapsulate this nice property within a new visualization tool based on these multiple trajectories.

The outline of the article is as follows. Section 2 unveils the variational Bayes motivation of Particle EM. Section 3 outlines the algorithmic machinery underpinning Particle EM, including its dynamic deployment. Section 4 highlights connections to existing procedures. Section 5 proposes the sequential Monte Carlo variant of Particle EM with repulsive particles. Section 6 provides illustrations on simulated data. Section 7 wraps up with a discussion.

2 Variational Bayes Motivation

As a prelude to the Particle EM algorithm, we first unravel the Particle EM learning criterion by revisiting the well-known connection between variational Bayes and EM learning. The EM algorithm, though originally conceived for maximum likelihood estimation, can be viewed as a variational Bayes approach for obtaining the best *single-point* approximation to the posterior. In a similar vein, Particle EM will be motivated as a variational Bayes approach for obtaining the best *multi-point* approximation.

We start our exposition with the simplest singleton case. Assume that one wishes to approximate $\pi(\boldsymbol{\gamma}, \sigma^2 \,|\, \boldsymbol{Y})$ with a Dirac measure $q(\boldsymbol{\gamma}, \sigma^2 \,|\, \boldsymbol{\gamma}_1, \sigma_1^2) = q(\boldsymbol{\gamma} \,|\, \boldsymbol{\gamma}_1) q_\sigma(\sigma^2 \,|\, \sigma_1^2) = \mathbb{I}(\boldsymbol{\gamma} = \boldsymbol{\gamma}_1)\mathbb{I}(\sigma^2 = \sigma_1^2)$ centered at an unknown binary vector $\boldsymbol{\gamma}_1$ and variance σ_1^2. Variational Bayes learning provides the closest approximation in a Kullback-Leibler (KL) sense by finding the binary string $\boldsymbol{\gamma}_1$ and scalar σ_1^2 that maximize the evidence lower bound

$$\log \pi(\boldsymbol{Y}) = \log \sum_{\boldsymbol{\gamma}} \int_{\sigma^2} \pi(\boldsymbol{Y}, \boldsymbol{\gamma}, \sigma^2) \geq \mathsf{E}_{q,q_\sigma} \left[\log \frac{\pi(\boldsymbol{Y}, \boldsymbol{\gamma}, \sigma^2)}{q(\boldsymbol{\gamma} \,|\, \boldsymbol{\gamma}_1) q_\sigma(\sigma^2 \,|\, \sigma_1^2)} \right]. \tag{2.1}$$

The variational lower bound (2.1) can be viewed as the height of the posterior $\pi(\boldsymbol{\gamma}, \sigma^2 \,|\, \boldsymbol{Y})$ at $(\boldsymbol{\gamma}_1, \sigma_1^2)$. It is thus not surprising that the best possible placement of the single atom $(\widehat{\boldsymbol{\gamma}}_1, \widehat{\sigma}_1^2)$ is at the global mode, i.e.

$$(\widehat{\boldsymbol{\gamma}}_1, \widehat{\sigma}_1^2) = \arg \max_{\boldsymbol{\gamma}_1 \in 2^p; \sigma_1^2 \in \mathbb{R}^+} \mathsf{E}_{q,q_\sigma} \left[\log \frac{\pi(\boldsymbol{Y}, \boldsymbol{\gamma}, \sigma^2)}{q(\boldsymbol{\gamma} \,|\, \boldsymbol{\gamma}_1) q_\sigma(\sigma^2 \,|\, \sigma_1^2)} \right] = \arg \max_{\boldsymbol{\gamma}_1 \in 2^p; \sigma_1^2 \in \mathbb{R}^+} \log \pi(\boldsymbol{\gamma}_1, \sigma_1^2 \,|\, \boldsymbol{Y}). \tag{2.2}$$

The optimization (2.2) is well-suited for EM-like algorithms (as will be shown in Section 3.1). However, certifiable global mode detection with EM is not guaranteed due to its proclivity for local maxima. Moreover, compressing the posterior into a single-mode summary will often be a deceptive understatement of model uncertainty. We will instead regard the EM learning criterion (2.2) as part of a larger perspective on modal estimation, moving forward towards new multi-point approximations.

Assume now that one wishes to find the closest approximation to $\pi(\boldsymbol{\gamma} \,|\, \boldsymbol{Y})$ using a mixture of atoms in the model space

$$q_{\text{PEM}}(\boldsymbol{\gamma} \,|\, \boldsymbol{\Gamma}, \boldsymbol{w}) = \sum_{k=1}^{K} w_k \mathbb{I}(\boldsymbol{\gamma} = \boldsymbol{\gamma}_k). \tag{2.3}$$

This "histogram" approximation has an intuitive appeal for polymodal posteriors, as it can capture and weave together their multiple peaks. We shall refer to (2.3) as the Particle EM approximation, since it can be obtained with a multi-point extension of the EM algorithm (as shown in Section

3.2). The free parameters of this Particle EM approximation are (a) the K binary string vectors $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]$ further referred to as *particles*, and (b) the *importance weights* $\boldsymbol{w} = (w_1, \ldots, w_K)'$ where $\sum_{k=1}^{K} w_k = 1$ and $0 \leq w_k \leq 1$. Similarity should be noted between (2.3) and the weighted point-mass posterior approximations used by particle Monte Carlo methods (Doucet et al., 2001). Such Monte Carlo deployments treat particles as stochastic samples, whose weighted aggregation provides an iid reconstruction of a target posterior. In contrast, Particle EM aligns more closely with an optimization point of view by treating particles as modes, capturing the essence of a multimodal posterior. We further elaborate on the conceptual similarities between particle MCMC and Particle EM later in Section 5.

The Particle EM learning criterion is obtained with the evidence lower bound (2.1) by replacing $q(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_1)$ with $q_{\text{PEM}}(\boldsymbol{\gamma} \mid \boldsymbol{\Gamma})$. This insertion yields

$$\mathcal{F}_\lambda(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma_1^2) \equiv \mathsf{E}_{q_{\text{PEM}} q_\sigma} \left[\log \frac{\pi(\boldsymbol{Y}, \boldsymbol{\gamma}, \sigma^2)}{q_{\text{PEM}}(\boldsymbol{\gamma} \mid \boldsymbol{\Gamma}, \boldsymbol{w}) q_\sigma(\sigma^2 \mid \sigma_1^2)} \right] = \sum_{k=1}^{K} w_k \, \log \pi(\boldsymbol{\gamma}_k, \sigma_1^2 \mid \boldsymbol{Y}) + \lambda \, H(\boldsymbol{\Gamma}, \boldsymbol{w}), \quad (2.4)$$

with $\lambda = 1$. This objective function is subject to maximization w.r.t. $(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma_1^2)$. Seen as a negative loss function over the entire particle system $(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma_1^2)$, (2.4) achieves a healthy balance between (a) optimal allocation and (b) diversity. The first summand in $\mathcal{F}_\lambda(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma_1^2)$ is the cumulative height of the weighted log-posterior at the particle locations $[\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]$, gearing particles towards areas with high posterior mass. Alone, this term would be maximized if all the particles were positioned at the global mode. Such an allocation, however, is counteracted by the second term

$$H(\boldsymbol{\Gamma}, \boldsymbol{w}) = -\mathsf{E}_{q_{\text{PEM}}} \left[\log q_{\text{PEM}}(\boldsymbol{\gamma} \mid \boldsymbol{\Gamma}, \boldsymbol{w}) \right],$$

the entropy of the distribution $q_{\text{PEM}}(\boldsymbol{\gamma} \mid \boldsymbol{\Gamma}, \boldsymbol{w})$. Reflecting the information content of the particle system, this entropy serves as a diversifying penalty by discouraging particles from gathering at overcrowded modes. As will be formalized in Section 3.2, this is manifested as mutual repulsion between overlapping particles. The strength of this repulsive effect is regulated by the parameter $\lambda \geq 0$. The default value obtained directly from variational calculus is $\lambda = 1$. Our framework allows for a broader continuum between no diversification (when $\lambda = 0$) and strong diversification (when $\lambda > 1$). However, with $\lambda \neq 1$, (2.4) is not necessarily a lower bound to the marginal likelihood. We would also like to draw attention to the fact that the notion of diversity here arises as an important decision theoretic consideration. Rather than finding the one best location K times, a more valuable strategy is to find the K top locations only once.

The Particle EM algorithm (proposed in Section 3) is devised towards maximizing $\mathcal{F}_\lambda(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma_1^2)$, which entails finding $\widehat{\boldsymbol{\Gamma}} = [\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_K]$, $\widehat{\boldsymbol{w}}$ and $\widehat{\sigma}_1$ that jointly satisfy

$$(\widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{w}}, \widehat{\sigma}_1^2) = \arg\max_{\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma_1^2} \mathcal{F}_\lambda(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma_1^2), \quad \text{subject to} \quad \sum_{k=1}^{K} w_k = 1, \quad 0 \leq w_k \leq 1. \tag{2.5}$$

It is worthwhile to note that the solution $\widehat{\boldsymbol{\Gamma}}$ will generally not yield K *unique* particles. More important modes will attract and accumulate particles more easily. The gravitational pull, and inherently the ability of each mode to retain particles, is affected by its weight w_k. In Section 3.3 we motivate these weights as inverse temperature parameters, having an opposite effect than in parallel tempering.

Because the particle ensemble $\boldsymbol{\Gamma}$ can contain copies, we need to refine our notation. Throughout the paper, we denote by $\boldsymbol{\Gamma}^\star = [\boldsymbol{\gamma}_1^\star, \ldots, \boldsymbol{\gamma}_{K^\star}^\star]$ the $K^\star$ *unique* particles contained within $\boldsymbol{\Gamma}$. We then denote by $p_l^\star$ the cumulative importance weight associated with each unique particle $\boldsymbol{\gamma}_l^\star$, i.e.

$$p_l^\star = \sum_{k=1}^{K} w_k \mathbb{I}(\gamma_k = \gamma_l^\star). \tag{2.6}$$

With this notation, we can express the entropy in the more familiar form

$$H(\boldsymbol{\Gamma}, \boldsymbol{w}) = -\sum_{l=1}^{K^\star} p_l^\star \log p_l^\star. \tag{2.7}$$

The entropy term (2.7) alone is maximized when the particles $\boldsymbol{\Gamma}$ are all unique, i.e. $K^\star = K$ and $p_l^\star = w_l$ for $l = 1, \ldots, K$. This provides an explanation for how the loss function (2.4) balances fit and diversity. The entropy penalty neutralizes the strong attraction of promising modes and diversifies the solution in the presence of model uncertainty. The Particle EM computational approach underpinning the diversification will be outlined in the next section.

3 Ensemble Optimization with Particle EM

What makes Particle EM unique relative to existing EM deployments is the opportunity it affords to (a) create multiple solution paths and, more importantly, (b) let them collaborate. The key idea behind Particle EM is to traverse the model space with an ensemble of repulsive particles, creating hill-climbing trajectories that are mutually aware. In this section, we will describe the machinery underpinning this intuitive description. We will assemble the Particle EM algorithm from single manageable pieces, starting with a simple variant with just one particle ($K = 1$) and gradually unveiling the full-blown ensemble approach ($K > 1$).

As was alluded to in Section 2, with $\lambda = 0$ the particles are left free with no collaborative responsibility. We will refer to this special case as the Parallel EM algorithm, since it amounts to running EM independently for each particle. However, allowing the particles to interact (with $\lambda > 0$) has distinctive performance advantages, as will be shown later in Section 6.

Conceptually, Particle EM pertains closely to existing particle filtering methods which combine importance sampling with MCMC in order to explore a dynamic sequence of distributions of interest. While Particle EM is of independent interest in static scenarios (with a fixed value v_0), we will take advantage of its computational speed and extend Particle EM to dynamic scenarios by allowing the proliferation of particles throughout a sequence of sliding posteriors (as v_0 is varied). We refer to this deployment as Dynamic Particle EM (Section 3.3).

3.1 Reversed EMVS

Before describing the Particle EM algorithm in its entirety, we begin by assuming $K = 1$. As noted in Section 2, Particle EM then seeks the best possible location for a single particle. This amounts to maximum a-posteriori model detection

$$(\widehat{\boldsymbol{\gamma}}, \widehat{\sigma}^2) = \arg\max_{\boldsymbol{\gamma}, \sigma^2} \ \log \pi(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{Y}). \tag{3.1}$$

This discrete optimization problem can be tackled with an EM data augmentation strategy. Instead of maximizing (3.1) directly, this strategy proceeds iteratively by optimizing a converging sequence of surrogate objective functions.

Suppose $(\boldsymbol{\beta}, \theta)$ are treated as missing data and let $\boldsymbol{\gamma}^{(m)}$ and $\sigma^{(m)}$ be the state of the particle and the variance at the m^{th} iteration. The following complete-data surrogate objective function

$$Q(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}) = \mathsf{E}_{\boldsymbol{\beta}, \theta \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}} \ \log \pi(\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\beta}, \theta \mid \boldsymbol{Y}), \tag{3.2}$$

then serves as a minorant to (3.1) and will be iteratively optimized using an EM-like algorithm. The surrogate objective function (3.2) can be written conveniently as

$$\begin{aligned} Q(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}) =& C + \frac{1}{2} \sum_{i=1}^{p} \gamma_i \log \left(\frac{v_0}{v_1} \right) - \frac{1}{2} \sum_{i=1}^{p} \left(\frac{\gamma_i}{v_1} + \frac{1 - \gamma_i}{v_0} \right) \mathsf{E} \left[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)} \right] \\ &+ \sum_{i=1}^{p} \gamma_i \mathsf{E} \left[\log \left(\frac{\theta}{1 - \theta} \right) \Big| \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)} \right] \\ &- \frac{n + \eta}{2} \log \sigma^2 - \frac{\eta \nu + \mathsf{E} \left[||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \Big| \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)} \right]}{2\sigma^2}. \end{aligned} \tag{3.3}$$

The constant C above absorbs all the summands in $Q(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{\gamma}^{(m)}, \sigma^{(m)})$ that do not depend on the parameters of interest $(\boldsymbol{\gamma}, \sigma)$. The iterative procedure now proceeds with two steps: the E-step, computing the expectations in (3.2), and the M-step which maximizes (3.2) with respect to $(\boldsymbol{\gamma}, \sigma)$.

The task of finding the single highest mode $(\widehat{\boldsymbol{\gamma}}, \widehat{\sigma})$ can also be approached indirectly with the EMVS procedure of Ročková and George (2014a). EMVS targets modes in the parameter space $\pi(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{Y})$ and associates them with modes in the model domain $\pi(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{Y})$ through thresholding. Here, we propose a different strategy which optimizes within the binary space, outputting directly a high-probability model. This strategy is very much in contrast with the EMVS method, which treats $\boldsymbol{\gamma}$ as missing data and $(\boldsymbol{\beta}, \theta)$ as parameters of interest. Here, the role of $\boldsymbol{\gamma}$ and $(\boldsymbol{\beta}, \theta)$ is reversed. Thereby we call this new procedure Reversed EMVS. [1]

3.1.1 The E-step

The E-step is available in closed form. First, we need to compute the second moments $\mathsf{E}(\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)})$ for each $i = 1, \ldots, p$. Denote by $\boldsymbol{D}(\boldsymbol{\gamma}) = \mathrm{diag}\{\gamma_i \frac{1}{v_1} + (1-\gamma_i)\frac{1}{v_0}\}_{i=1}^p$. From the conditional conjugacy property of the Gaussian mixture prior it follows that $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{\gamma}, \sigma^2) = \mathcal{N}_p\left[\mu(\boldsymbol{\gamma}, \sigma), \Sigma(\boldsymbol{\gamma}, \sigma)\right]$, where $\Sigma(\boldsymbol{\gamma}, \sigma) = \sigma^2[\boldsymbol{X}'\boldsymbol{X} + \sigma^2 \boldsymbol{D}(\boldsymbol{\gamma})]^{-1}$ and $\boldsymbol{\mu}(\boldsymbol{\gamma}, \sigma) = \frac{1}{\sigma^2}\Sigma(\boldsymbol{\gamma}, \sigma)\boldsymbol{X}'\boldsymbol{Y}$. Thus

$$\mathsf{E}[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}] = \mu_i(\boldsymbol{\gamma}^{(m)}, \sigma^{(m)})^2 + \Sigma_{ii}(\boldsymbol{\gamma}^{(m)}, \sigma^{(m)}). \tag{3.4}$$

The E-step is then completed with the conditional expectation $\mathsf{E}\left[\log\left(\frac{\theta}{1-\theta}\right) \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right]$. It is useful to note that given $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(m)}$, the parameter θ is independent of $\boldsymbol{Y}$ and σ. With $\theta \sim \mathcal{B}(a, b)$, we have $\pi(\theta \mid \boldsymbol{\gamma}) \sim \mathcal{B}(a + |\boldsymbol{\gamma}|, b + p - |\boldsymbol{\gamma}|)$, and thus the conditional expectation can be obtained as follows:

$$\mathsf{E}\left[\log\left(\frac{\theta}{1-\theta}\right) \,\middle|\, \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right] = \psi(a + |\boldsymbol{\gamma}^{(m)}|) - \psi(b + p - |\boldsymbol{\gamma}^{(m)}|), \tag{3.5}$$

where $\psi(\cdot)$ denotes the digamma function (a logarithmic derivative of the Gamma function).

3.1.2 The M-step

The M-step entails finding the binary vector $\boldsymbol{\gamma}^{(m+1)}$ and variance $\sigma^{(m+1)}$ which maximizes (3.2). Since the entries in $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ are conditionally independent of each other, given the missing data, we can find $\boldsymbol{\gamma}^{(m+1)}$ by updating separately each individual coordinate $\gamma_i^{(m+1)}$. Focusing on the

[1] We wish to draw attention to concurrent and independent work of Wang et al. (2016) who has devised a very similar EM algorithm to Reversed EMVS. Their version, however, treats θ as a parameter of interest rather than as missing data.

i^{th} coordinate, we can write

$$\gamma_i^{(m+1)} = \arg\max_{\gamma\in\{0,1\}} \left\{ \frac{\gamma}{2}\log\left(\frac{v_0}{v_1}\right) - \frac{\gamma}{2}\left(\frac{1}{v_1}-\frac{1}{v_0}\right)\mathsf{E}\left[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right] + \gamma\,\mathsf{E}\left[\log\left(\frac{\theta}{1-\theta}\right)\Big|\, \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right]\right\}. \tag{3.6}$$

The maximization (3.6) is trivialized by noting that γ_i is binary. The objective function (3.6) can be regarded as the log-likelihood of a Bernoulli trial with an inclusion probability

$$\pi_i = \left(1+\exp\left\{\mathsf{E}\left[\log\left(\frac{1-\theta}{\theta}\right)|\boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right]\right\}\frac{\phi\left(\sqrt{\mathsf{E}\left[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right]}, v_0\right)}{\phi\left(\sqrt{\mathsf{E}\left[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right]}, v_1\right)}\right)^{-1}. \tag{3.7}$$

The success probability π_i increases with the expected prior odds of entering the model. It is also concordant with the slab vs. spike likelihood ratio associated with the expected size of the i^{th}coefficient. The next update of $\gamma_i^{(m+1)}$ is then obtained by simply thresholding π_i, i.e.

$$\gamma_i^{(m+1)} = 1 \quad \text{if and only if} \quad \pi_i > 0.5. \tag{3.8}$$

The entire M-step is completed with a single sweep of independent updates (3.8) for each $i = 1, \ldots, p$.

It is worthwhile to point out the similarity between π_i and $\mathsf{P}(\gamma_i = 1 \mid \beta_i, \theta) = \left(1+\frac{1-\theta}{\theta}\frac{\phi(\beta_i, v_0)}{\phi(\beta_i, v_1)}\right)^{-1}$ used by the EMVS procedure. EMVS "continuizes" the binary model space by updating the continuous inclusion probabilities $\mathsf{P}(\gamma_i = 1 \mid \beta_i, \theta)$. Reversed EMVS, on the other hand, updates the binary vectors by dichotomizing a variant of these probabilities. The lack of continuity may render Reversed EMVS more susceptible to the entrapment in suboptimal modes, in particular when $p > n$. In the next section, we show how to boost the performance of Reversed EMVS with multiple interacting particles. Finally, the M-step has one extra update for obtaining $\sigma^{(m+1)}$ from

$$\begin{aligned}\sigma^{(m+1)2} = &\frac{1}{n+\eta}\Big\{\eta\nu + \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\mu}(\boldsymbol{\gamma}^{(m)}, \sigma^{(m)})'\boldsymbol{X}'\boldsymbol{Y} + \\ &\operatorname{tr}\left[\boldsymbol{X}'\boldsymbol{X}\left(\Sigma(\boldsymbol{\gamma}^{(m)}, \sigma^{(m)}) + \boldsymbol{\mu}\left(\boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right)\boldsymbol{\mu}\left(\boldsymbol{\gamma}^{(m)}, \sigma^{(m)}\right)'\right)\right]\Big\}.\end{aligned}$$

3.2 Particle EM

The Particle EM algorithm is an ensemble extension of Reversed EMVS designed to solve the constrained optimization problem (2.5) when $K > 1$. Algorithmically speaking, Particle EM alternates between collaborative updating of the particle locations $[\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]$, redefining the importance weights $(w_1, \ldots, w_K)'$ and the variance σ^2. Conditionally on the particles, the importance weights are available in closed form (as shown later in Section 3.2.2). To refresh the particle locations, however, we need to step outside the variational framework with an extra data augmentation step. To this

end, we devise a nested EM-like strategy for maximizing (2.5) that is implicitly embedded within the variational Bayes apparatus.

3.2.1 Updating Particle Locations $\boldsymbol{\Gamma}^{(m+1)}$

Denote by $\boldsymbol{\Gamma}^{(m)}$ the state of the particle system at the m^{th} iteration and by $\boldsymbol{w}^{(m)}$ the associated weights. First, we describe how to obtain new particle locations $\boldsymbol{\Gamma}^{(m+1)}$, given $\boldsymbol{w}^{(m)}$ and $\sigma^{(m)}$. The vehicle for this update is the expected data augmented analog of (2.4).

Instead of directly optimizing (2.4), we proceed with a "minorize-maximize strategy" focusing on a surrogate objective

$$Q(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma^2 \mid \boldsymbol{Y}, \boldsymbol{\Gamma}^{(m)}, \sigma^{(m)}) = \sum_{k=1}^{K} w_k \, Q(\boldsymbol{\gamma}_k, \sigma^2 \mid \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}) + \lambda \, H(\boldsymbol{\Gamma}, \boldsymbol{w}), \tag{3.9}$$

where $Q(\boldsymbol{\gamma}_k, \sigma^2 \mid \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)})$ was defined earlier in (3.2) and where $H(\boldsymbol{\Gamma}, \boldsymbol{w})$ is the entropy defined in (2.7). This surrogate function is obtained after minorizing each summand $\log \pi(\boldsymbol{\gamma}_k, \sigma^2 \mid \boldsymbol{Y}, \sigma^{(m)})$ in (2.4) by the minorant $\mathsf{E}_{\boldsymbol{\beta},\theta \mid \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}} \log \pi(\boldsymbol{\gamma}_k, \boldsymbol{\beta}, \theta, \sigma^2 \mid \boldsymbol{Y})$. After inserting (3.1), we can rewrite (3.9) as follows:

$$\begin{aligned} Q(\boldsymbol{\Gamma}, \boldsymbol{w}, \sigma^2 \mid \boldsymbol{Y}, \boldsymbol{\Gamma}^{(m)}, \sigma^{(m)}) =& \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{p} w_k \gamma_{ik} \log\left(\frac{v_0}{v_1}\right) + \lambda \, H(\boldsymbol{\Gamma}, \boldsymbol{w}) + C_1 \\ & - \sum_{k=1}^{K} w_k \left[\sum_{i=1}^{p} \mathsf{E}(\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}) \left(\frac{\gamma_i}{v_1} + \frac{1-\gamma_{ik}}{v_0} \right) \right] \\ & + \sum_{k=1}^{K} w_k \sum_{i=1}^{p} \gamma_{ik} \mathsf{E}\left[\log\left(\frac{\theta}{1-\theta}\right) \,\middle|\, \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)} \right] \\ & - \frac{n+\eta}{2} \log \sigma^2 - \frac{\eta\nu + \sum_{k=1}^{K} w_k \mathsf{E}\left[||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \,\middle|\, \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)} \right]}{2\sigma^2}, \end{aligned} \tag{3.10}$$

where C_1 is a constant that does not depend on $\boldsymbol{\Gamma} = \{\gamma_{ik}\}_{i,k=1}^{p,K}$, σ^2 or $\boldsymbol{w}$. The conditional expectations $\mathsf{E}(\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)})$ and $\mathsf{E}\left[\log\left(\frac{\theta}{1-\theta}\right) \,\middle|\, \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)} \right]$ are evaluated in the E-step. These are the exact same calculations as in Section 3.1.1. We would like to draw attention to the fact that the E-step evaluation *only* requires calculating (3.4) and (3.5) for each of the $K^\star$ *unique* particles in $\boldsymbol{\Gamma}^{(m)}$. These can be carried out in parallel.

Once the E-step is completed, $Q(\boldsymbol{\Gamma}, \boldsymbol{w}^{(m)}, \sigma^{(m)2} \mid \boldsymbol{Y}, \boldsymbol{\Gamma}^{(m)}, \sigma^{(m)})$ serves as a surrogate minorant to (2.5) whose (local) maximum improves on the previous guess $\boldsymbol{\Gamma}^{(m)}$. We now outline the M-step of the nested EM algorithm which outputs the improvement, denoted by $\boldsymbol{\Gamma}^{(m+1)}$. Before proceeding, let us introduce auxiliary notation. Denote by $\boldsymbol{\Gamma}_{\setminus ik}$ all but the $(i,k)^{th}$ entry γ_{ik} in the matrix $\boldsymbol{\Gamma}$. Moreover,

let $H(\gamma_{ik}, \mathbf{\Gamma}_{\setminus ik}, \boldsymbol{w})$ denote the entropy term $H(\mathbf{\Gamma}, \boldsymbol{w})$, explicitly spelling out the value of the $(i,k)^{th}$ coordinate γ_{ik}.

With $\lambda = 0$, the M-step would be equivalent to applying (3.8) independently for each unique particle, the Parallel EM strategy. However, with $\lambda > 0$ the particles are intertwined through the entropy penalty. To obtain $\mathbf{\Gamma}^{(m+1)}$, we proceed sequentially in coordinate-wise fashion, cycling over one-site updates of $\gamma_{ik}^{(m+1)}$ conditionally on the most recent value of $\mathbf{\Gamma}_{\setminus ik}^{(m+1)}$. To this end, we introduce the one-site analog of (3.10) in the $(i,k)^{th}$ direction

$$Q_{ik}(\gamma_{ik}, \mathbf{\Gamma}_{\setminus ik}^{(m+1)}, \boldsymbol{w}^{(m)}, \sigma^{(m)2} \mid \boldsymbol{Y}, \mathbf{\Gamma}^{(m)}, \sigma^{(m)}) = \frac{\gamma_{ik}}{2}\left\{\log\left(\frac{v_0}{v_1}\right) - \left(\frac{1}{v_1} - \frac{1}{v_0}\right)\mathsf{E}\left[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}\right]\right\}$$
$$+ \gamma_{ik}\,\mathsf{E}\left[\log\left(\frac{\theta}{1-\theta}\right)\,\middle|\,\boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}\right] + \lambda/w_k^{(m)}\, H(\gamma_{ik}, \mathbf{\Gamma}_{\setminus ik}^{(m+1)}, \boldsymbol{w}^{(m)}). \tag{3.11}$$

This quantity captures the portion of $Q(\mathbf{\Gamma}, \boldsymbol{w}^{(m)}, \sigma^{(m)2} \mid \boldsymbol{Y}, \mathbf{\Gamma}^{(m)}, \sigma^{(m)})$ that depends on γ_{ik} while keeping $\mathbf{\Gamma}_{\setminus ik}$ fixed at $\mathbf{\Gamma}_{\setminus ik}^{(m+1)}$. The new one-site update of γ_{ik} satisfies

$$\gamma_{ik}^{(m+1)} = \arg\max_{\gamma_{ik}\in\{0,1\}} Q_{ik}(\gamma_{ik}, \mathbf{\Gamma}_{\setminus ik}^{(m+1)}, \boldsymbol{w}^{(m)}, \sigma^{(m)2} \mid \boldsymbol{Y}, \mathbf{\Gamma}^{(m)}, \sigma^{(m)}).$$

Similarly as in Section 3.1.2, (3.11) can take only two values, depending on the status of γ_{ik}. We can regard (3.11) as the log-likelihood of a Bernoulli random variable with an inclusion probability

$$\pi_{ik} = \left(1 + \exp\left\{\mathsf{E}\left[\log\left(\frac{1-\theta}{\theta}\right) \mid \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}\right]\right\} \frac{\phi\left(\sqrt{\mathsf{E}\left[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}\right]}; v_0\right)}{\phi\left(\sqrt{\mathsf{E}\left[\beta_i^2 \mid \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}\right]}; v_1\right)} Rep\left(\mathbf{\Gamma}_{\setminus ik}^{(m+1)}, \boldsymbol{w}^{(m)}\right)\right)^{-1}. \tag{3.12}$$

where

$$Rep\left(\mathbf{\Gamma}_{\setminus ik}^{(m+1)}, \boldsymbol{w}^{(m)}\right) = \frac{\exp\left\{\lambda/w_k^{(m)}\, H(0, \mathbf{\Gamma}_{\setminus ik}^{(m+1)}, \boldsymbol{w}^{(m)})\right\}}{\exp\left\{\lambda/w_k^{(m)}\, H(1, \mathbf{\Gamma}_{\setminus ik}^{(m+1)}, \boldsymbol{w}^{(m)})\right\}} \tag{3.13}$$

We pause now to appreciate the difference between (3.7) and (3.12). The crucial aspect of (3.12), which is absent from (3.7), is the term (3.13), an entropy of two particle systems which differ only by the status of γ_{ik}. This term provides a fundamental basis for understanding how Particle EM exerts repulsion among particles. By encouraging binary flips that increase entropy, this term gears the one-site update away from a solution that would be redundant in light of other particle locations. This effect is balanced by the evidence in the data, where the repulsive term takes over only if the data are ambiguous as to whether γ_{ik} should be switched on/off.

Algorithm:	*Particle EM Algorithm for Variable Selection*
Initialize with $\mathbf{\Gamma}^{(0)}$ and $\sigma^{(0)}$	
Repeat the E-step and M-step until convergence*	
The E-Step	
	For each $k \in \{1, \ldots, K\}$
Regression Coefficients	Compute
	$\Sigma_k = \sigma^{(m)2}[\boldsymbol{X}'\boldsymbol{X} + \sigma^{(m)2}\boldsymbol{D}(\boldsymbol{\gamma}_k)]^{-1}$
	$\boldsymbol{\mu}_k = \frac{1}{\sigma^{(m)2}}\Sigma_k \boldsymbol{X}'\boldsymbol{Y}$
	Update $\mathsf{E}[\beta_{jk}^2 \,\big|\, \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}]$ from (3.4) for $j = 1, \ldots, p$
Prior Odds Ratio	Update $\mathsf{E}\left[\log\left(\frac{\theta}{1-\theta}\right) \,\Big|\, \boldsymbol{Y}, \boldsymbol{\gamma}_k^{(m)}, \sigma^{(m)}\right]$ from (3.5)
The M-Step	
Binary Matrix	Repeat the following steps until convergence*
	For $j = 1, \ldots, p$ and $k = 1, \ldots, K$
	Update π_{ik} from (3.12)
	Set $\gamma_{ik}^{(m+1)} = 1$ if $\pi_{ik} > 0.5$ and zero otherwise
Weights	Update $w_k^{(m+1)}$ from (3.18)
Variance	Update $\sigma^{(m+1)2}$ from (3.19)

Table 1: Particle EM algorithm for variable selection; *convergence claimed when no binary switch in $\mathbf{\Gamma}$ occurs in two consecutive iterations

Conditionally on the most recent values $\mathbf{\Gamma}_{\setminus ik}^{(m+1)}$, Particle EM cycles over one-site updates

$$\gamma_{ik}^{(m+1)} = 1 \quad \text{if and only if} \quad \pi_{ik} > 0.5 \tag{3.14}$$

for each (i, k) until an equilibrium is reached (i.e. no flipping of states is observed in two consecutive iterations). This is an indication that the particle system $\mathbf{\Gamma}^{(m+1)}$ is ready to be transmitted to the next step of the algorithm. It is worthwhile to note that a single cycle would be enough to improve upon $\mathbf{\Gamma}^{(m)}$. However, multiple cycles can amplify the improvement.

3.2.2 Updating Weights $\boldsymbol{w}^{(m+1)}$

Once we obtain the new particle system $\mathbf{\Gamma}^{(m+1)} = [\boldsymbol{\gamma}_1^{(m+1)}, \ldots, \boldsymbol{\gamma}_K^{(m+1)}]$, we can refresh the importance weights $\boldsymbol{w}^{(m+1)}$, given $\sigma^{(m)}$. While nested EM augmentation was needed for the update $\mathbf{\Gamma}^{(m+1)}$, no such step is necessary for the weights. Thereby we can work directly with the *original* variational lower bound (2.4). The new weights are seen to satisfy

$$\boldsymbol{w}^{(m+1)} = \arg\max_{\boldsymbol{w}} \left\{ \sum_{k=1}^{K} w_k \, \log \pi(\boldsymbol{\gamma}_k^{(m+1)}, \sigma^{(m)2} \mid \boldsymbol{Y}) + \lambda \, H(\mathbf{\Gamma}^{(m+1)}, \boldsymbol{w}) \right\}, \tag{3.15}$$

where $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$. To facilitate this update, it will be convenient to work instead with the aggregated weights $p_l^\star$ (introduced earlier in Section 2). Denote by $\boldsymbol{p}^{\star(m+1)}$ the vector of aggregated weights associated with $\boldsymbol{w}^{(m+1)}$ according to (2.6) with $\mathbf{\Gamma} = \mathbf{\Gamma}^{(m+1)}$. Then (3.15) implies

that $\boldsymbol{p}^{\star(m+1)}$ satisfy

$$\boldsymbol{p}^{\star(m+1)} = \arg\max_{\boldsymbol{p}^\star} \left\{ \sum_{l=1}^{K^\star} p_l^\star \, \log \pi(\boldsymbol{\gamma}_l^{\star(m+1)}, \sigma^{(m)2} \mid \boldsymbol{Y}) - \lambda \sum_{l=1}^{K^\star} p_l^\star \log p_l^\star \right\}, \tag{3.16}$$

where $0 \leq p_l^\star \leq 1$ and $\sum_{l=1}^{K\star} p_l^\star = 1$. Taking the derivative of (3.16) with respect to $p_l^\star$, we can easily see that $p_l^{\star(m+1)} \propto \pi(\boldsymbol{\gamma}_l^{\star(m+1)}, \sigma^{(m)2} \mid \boldsymbol{Y})$. Thus, the optimal weight $p_l^{\star(m+1)}$ (associated with $\boldsymbol{\gamma}_l^{\star(m+1)}$) is proportional to the posterior probability $\pi(\boldsymbol{\gamma}_l^{\star(m+1)}, \sigma^{(m)2} \,|\boldsymbol{Y})$. This is reassuring because we would like to regard the Particle EM approximation (2.3) as a normalized restriction of $\pi(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{Y})$. The posterior model probabilities can be computed in closed form, up to a constant, according to (George and McCulloch, 1997)

$$\pi(\boldsymbol{\gamma}, \sigma^{(m)2} \,|\boldsymbol{Y}) \propto \frac{\pi(\boldsymbol{\gamma})\pi(\sigma^{(m)2})|\boldsymbol{D}(\boldsymbol{\gamma})|^{\frac{1}{2}}}{|\boldsymbol{X}'\boldsymbol{X}/\sigma^{(m)2} + \boldsymbol{D}(\boldsymbol{\gamma})|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2\sigma^{(m)2}} \left[\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \sigma^{(m)2}\boldsymbol{D}(\boldsymbol{\gamma}))^{-1}\boldsymbol{X}'\boldsymbol{Y} \right] \right\}. \tag{3.17}$$

The original weights $\boldsymbol{w}^{(m+1)}$ can be recovered from $p^{\star(m+1)}$ as follows:

$$w_k^{(m+1)} \propto \frac{\sum_{l=1}^{K^\star} p_l^\star \mathbb{I}(\boldsymbol{\gamma}_k^{(m+1)} = \boldsymbol{\gamma}_l^{\star(m+1)})}{\sum_{l=1}^{K} \mathbb{I}(\boldsymbol{\gamma}_k^{(m+1)} = \boldsymbol{\gamma}_l^{(m+1)})} \propto \frac{\pi(\boldsymbol{\gamma}_k^{(m+1)} \mid \boldsymbol{Y}, \sigma^{(m)})}{n_k}, \tag{3.18}$$

where n_k is the number of copies of $\boldsymbol{\gamma}_k^{(m+1)}$ in $\boldsymbol{\Gamma}^{(m+1)}$. In other words, the original weight w_k is again proportional to a posterior model probability (through the aggregated weight $p_l^\star$). However, in case there are multiple copies of the particle $\boldsymbol{\gamma}_k$ the weight $p_l^\star$ gets divided equally among the multiple copies. We could also divide the weights unequally. This would not change the quality of the particle approximation to the posterior.

3.2.3 Updating the Variance $\sigma^{(m+1)2}$

Finally, the M-step has one extra update for obtaining $\sigma^{(m+1)2}$, conditionally on $\boldsymbol{\Gamma}^{(m+1)}$ and $\boldsymbol{w}^{(m+1)}$,

$$\sigma^{(m+1)2} = \sum_{k=1}^{K} \frac{w_k^{(m+1)}}{n+\eta} \Big\{ \eta\nu + \mathsf{E}[||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \mid \boldsymbol{\gamma}_k^{(m)}, \boldsymbol{Y}, \sigma^{(m)}] \Big\}. \tag{3.19}$$

With $K = 1$, the point-mass approximating density $q_\sigma(\sigma^2 \,|\sigma_1^2)$ aims at the MAP estimate of $\pi(\boldsymbol{\gamma}, \sigma^2 \,|\boldsymbol{Y})$. With $K > 1$, the update (3.19) aims at a global "model averaged" estimate of σ^2, where an improved approximation could be obtained with a mixture of point-masses, one atom for each model.

3.3 Exploring Multiple Posteriors with Dynamic PEM

Particle EM can be expanded to dynamic scenarios for exploring an entire path of posteriors rather than just a single snapshot. For variable selection, such dynamic approaches entail a sequence of

tentative posteriors $\pi_t(\boldsymbol{\gamma} \mid \boldsymbol{Y})$, constructed either by adding data points in a sequential fashion (Shi and Dunson, 2011), or artificially by forming a path towards a target posterior $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ (e.g. via a geometric bridge (Schafer and Chopin, 2013)). Here, we adopt a new different approach. We regard the spike variance v_0 as the tempering parameter indexing a continuum of posteriors between "friendly landscapes" (with v_0 large) and spiky posteriors (with v_0 small). Dynamic Particle EM deployment sequentially reinitializes the particles with warm starts over a grid of diminishing values v_0. The hope is that the backward proliferation of particles from *large to small* v_0 improves particle placement over the static scenario when v_0 is small. With large v_0, dominant posterior modes are few and far between, easy to track down by particles. As v_0 gets smaller, the dominant posterior modes melt down and split, having their particles scattered due to their mutual repulsion. The Dynamic PEM warm start strategy can be regarded as an ensemble extension to deterministic annealing EM for global mode detection (Ueda and Nakano, 1998; Yoshida and West, 2010; Ročková and George, 2014a).

4 Links to Existing Work

Before proceeding, we pause and highlight the connections between Particle EM and related approaches for multimodal posterior exploration that have occurred in the literature. The closest relatives to Particle EM are parallel tempering sampling methods for Bayesian variable selection (Strens, 2003; Bottolo and Richardson, 2010) and non-parametric variational Bayes (Jaakkola and Jordan, 1998; Gershman et al., 2012). More parallels with sequential MCMC will be drawn in Section 5. Both in Section 4 and Section 5, we will assume that σ^2 is either known or integrated out, where the object of interest is the marginal posterior $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ rather than the joint posterior $\pi(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{Y})$.

4.1 Connection to Parallel Tempering

The variational lower bound (2.4) has an appeal also purely from a Monte Carlo point of view. This idea is formalized by associating (2.4) with a joint posterior

$$\Pi(\boldsymbol{\Gamma} \mid \boldsymbol{Y}) \propto \prod_{k=1}^{K} \pi(\boldsymbol{\gamma}_k \mid \boldsymbol{Y})^{w_k} \tag{4.1}$$

targeted by parallel tempering MCMC algorithms (Geyer, 1991; Bottolo and Richardson, 2010; Liang and Wong, 2000). More precisely, the logarithm of the joint target (4.1) is mathematically equivalent to (2.4) with one important difference, the entropy term is missing. Thus, the default parallel tem-

pering population MCMC methods do not foster diversity across their MCMC chains. In contrast, the Particle EM approach exerts diversification (when $\lambda > 0$). After drawing this useful link, it is natural to inquire whether diversification might enhance the effectiveness of population MCMC. We will elaborate on this possibility in Section 5.

While parallel tempering runs MCMC on (4.1), generating a population of interacting chains $(\boldsymbol{\gamma}_1^{(t)}, \ldots, \boldsymbol{\gamma}_K^{(t)})'$ with local (within chain) and global (between chain) swaps, Particle EM generates a population of interacting hill-climbing trajectories towards maximizing (2.4). There is yet another interesting link between these two approaches. It concerns the importance weights $(w_1, \ldots, w_K)'$. As seen from (4.1), these weights can be regarded as inverse temperatures (Jasra et al., 2007), flattening the peaks of the posterior distribution $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$. In parallel tempering, the higher the temperature $1/w_k$, the easier it is for the Markov chain to escape from spurious local peaks. In Particle EM, however, it is more desirable that more important modes attract and retain more particles. This is why the temperatures are actually smaller for more important models (with larger w_k). Moreover, Particle EM weights are not fixed (as in parallel tempering) but subject to estimation.

4.2 Connection to Non-parametric Variational Bayes

Besides its close ties with population MCMC, Particle EM also has immediate connections with deterministic approaches for coping with posterior multi-modality. One such pioneering strategy was proposed by Jaakkola and Jordan (1998), who extended the naive mean-field approach (independent product approximations) to mixture approximating forms. Going further, Gershman et al. (2012) developed a "non-parametric variational Bayes" approach, focusing on Gaussian mixtures. Motivated by these developments, Ročková et al. (2016) pursued the non-parametric variational Bayes strategy for Bayesian spike-and-slab variable selection. Their focus was on approximating the posterior $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ in the *parameter* domain $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ with a Gaussian mixture. This variational optimization turns out to be rather challenging since the evidence lower bound involves intractable terms such as the entropy of a Gaussian mixture. By shifting our focus to the discrete space $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$, we instead only have to cope with the entropy of a discrete mixture, which is extremely feasible. And as a further benefit, Particle EM does not have to use fixed weights w_k. Another related stream of research includes relaxed mean field approximations with elaborate distributions (Saul and Jordan, 1996; Wand et al., 2011).

Algorithm: *Repulsive Particle Filter for Variable Selection*	
Initialization	
Spike Temperatures	$\{v_0^0 > v_0^1 > \cdots > v_0^M\}$
Particle Locations	For each $k \in \{1, \ldots, K\}$ sample
	$\boldsymbol{\gamma}_k(v_0^0) \sim \pi_{v_0^0}(\boldsymbol{\gamma} \mid \boldsymbol{Y})$
Weights	Set $w_k(v_0^0) = 1/K$
Particle Filtering	
For $m \in \{1, \ldots, M\}$	Cycle through
Reweighting	Update $w_k(v_0^m)$ from (5.5)
Proliferation	Set $\boldsymbol{\gamma}_k(v_0^m) = \boldsymbol{\gamma}_k(v_0^{m-1})$
	If $\mathsf{Var}\,(\boldsymbol{w}) > \varepsilon$
	For $k \in \{1, \ldots, K\}$
Resample	Select $\boldsymbol{\gamma}_k(v_0^m)$ from $\{\boldsymbol{\gamma}_k(v_0^m)\}_{k=1}^K$ with weights $\{w_k(v_0^m)\}_{k=1}^K$
	Set $w_k(v_0^m) = 1/K$
Rejuvenate	Resample $\boldsymbol{\gamma}_k(v_0^m)$ from (5.1) using (5.3)

Table 2: Sequential Monte Carlo for Bayesian variable selection with repulsive particles

5 Population MCMC with Repulsive Particles

Despite being inherently deterministic, Particle EM also provides a promising new avenue towards population stochastic search. One of the attractive features of Particle EM is its ability to diversify particles. This feature has an appeal also in the Monte Carlo domain, sequential Monte Carlo in particular, where it can help improve mixing (Gilks and Berzuini, 2001). In this section, we propose a stochastic counterpart to Particle EM, capitalizing on existing developments in static (Robert and Mengersen, 2003) and dynamic (Schafer and Chopin, 2013) stochastic particle systems.

5.1 Repulsive Particle Stochastic Search (Fixed v_0)

Our starting point is the pinball sampler of Robert and Mengersen (2003), a particle filtering approach that simultaneously processes a vector of continuous random variables towards a simulation from a *static* reference distribution. This is achieved via a random walk with a correction to avoid the immediate vicinity of other particles. Here, we adapt their approach for the *static* discrete reference distribution $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ (for a fixed value $v_0 > 0$) using a set of K bit-string particles $[\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]$ and a new repulsive proposal distribution based on the entropy of the particle system $H(\boldsymbol{\Gamma}, \boldsymbol{w})$ defined in (2.7). Throughout this section, the regression vector $\boldsymbol{\beta}$ will be implicitly integrated out and thereby not subject to posterior sampling (as in e.g. Shi and Dunson (2011)).

Given an initialization $\boldsymbol{\Gamma}^{(0)} = [\boldsymbol{\gamma}_1^{(0)}, \ldots, \boldsymbol{\gamma}_K^{(0)}]$, we set out to simulate an assemblage of chains indexed by $t = 1, \ldots, T$ via

$$\boldsymbol{\gamma}_k^{(t+1)} \sim K_k(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_1^{(t+1)}, \ldots, \boldsymbol{\gamma}_{k-1}^{(t+1)}, \boldsymbol{\gamma}_k^{(t)}, \boldsymbol{\gamma}_{k+1}^{(t)}, \ldots, \boldsymbol{\gamma}_K^{(t)}), \quad \text{for} \quad k = 1, \ldots, K, \tag{5.1}$$

where $K_k(\cdot)$ is a conditional transition kernel, which depends on the current state of the entire particle system. This strategy (under standard irreducibility conditions on the kernel) yields an approximate iid sample from $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ at any slice in time t after burn-in (according to Theorem 1 of Robert and Mengersen (2003)). Our transition kernel $K_k(\boldsymbol{\gamma} \mid \cdot)$ has a point mass at the current state $\boldsymbol{\gamma}_k^{(t)}$ and allows for a transition onto a new state $\boldsymbol{\gamma}_k^{new}$ chosen from a one-site proposal mechanism that reflects the location of other particles. This is achieved by first proposing $\boldsymbol{\gamma}_k^{new}$ towards sampling from a tilted target distribution

$$q_k(\boldsymbol{\gamma} \mid \boldsymbol{\Gamma}_{\setminus k}^{(t+1)}) \propto \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \, \exp\left\{\lambda \, H\left(\boldsymbol{\Gamma}_k^{(t+1)}(\boldsymbol{\gamma}), \boldsymbol{w}\right)\right\}, \tag{5.2}$$

where $\boldsymbol{\Gamma}_k^{(t+1)}(\boldsymbol{\gamma}) \equiv [\boldsymbol{\gamma}_1^{(t+1)}, \ldots, \boldsymbol{\gamma}_{k-1}^{(t+1)}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_{k+1}^{(t)}, \ldots, \boldsymbol{\gamma}_K^{(t)}]$ and $\boldsymbol{\Gamma}_{\setminus k}^{(t+1)}$ contains all but the k^{th} column in $\boldsymbol{\Gamma}_k^{(t+1)}(\boldsymbol{\gamma})$. For simplicity, we shall assume that each particle is assigned an equal weight in the entropy term, i.e. $\boldsymbol{w} = (1/K, \ldots, 1/K)'$. The transition onto the new state is then guided by the following scheme

$$\boldsymbol{\gamma}_k^{(t+1)} = \begin{cases} \boldsymbol{\gamma}_k^{(t)} & \text{with probability} \quad 1-\alpha \\ \boldsymbol{\gamma}_k^{new} & \text{with probability} \quad \alpha, \end{cases} \tag{5.3}$$

where

$$\alpha = \left(1 \wedge \frac{q_k(\boldsymbol{\gamma}^{new} \mid \boldsymbol{\Gamma}_{\setminus k}^{(t+1)})}{q_k(\boldsymbol{\gamma}_k^{(t)} \mid \boldsymbol{\Gamma}_{\setminus k}^{(t+1)})}\right) \left(1 \wedge \frac{\exp\left\{\lambda \, H\left(\boldsymbol{\Gamma}_k^{(t+1)}(\boldsymbol{\gamma}_k^{(t)}), \boldsymbol{w}\right)\right\}}{\exp\left\{\lambda \, H\left(\boldsymbol{\Gamma}_k^{(t+1)}(\boldsymbol{\gamma}_k^{new}), \boldsymbol{w}\right)\right\}}\right).$$

The acceptance probability α can be decoupled into two consecutive Metropolis-Hasting steps. The first step is devised towards sampling from a wrong target distribution $q_k(\boldsymbol{\gamma} \mid \boldsymbol{\Gamma}_{\setminus k}^{(m+1)})$. The second step corrects for this mistake by tilting the acceptance probability towards sampling from the correct target. By Lemma 1 of Robert and Mengersen (2003), this scheme guarantees sampling from the correct stationary distribution $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$.

5.2 Repulsive Particle Filtering (Sequence of v_0 Values)

Going further, we show how the particle stochastic search from Section 5.1 can be expanded to dynamic settings by serving as a repulsive transition kernel for particle filtering. Particle filtering methods for variable selection (Ma, 2015; Schafer and Chopin, 2013; Shi and Dunson, 2011) rely on a discrete approximation to a sequence of moving posterior distributions $\pi_t(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ indexed by t, using a system of particles $[\boldsymbol{\gamma}_1^{(t)}, \ldots, \boldsymbol{\gamma}_K^{(t)}]$ and weights $\boldsymbol{w}^{(t)}$. Similarly as in Section 3.3, we consider a continuum of posteriors $\pi_{v_0}(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ indexed by the spike variance v_0.

Our sequential Monte Carlo counterpart to Particle EM generates the following sequence of posterior approximations

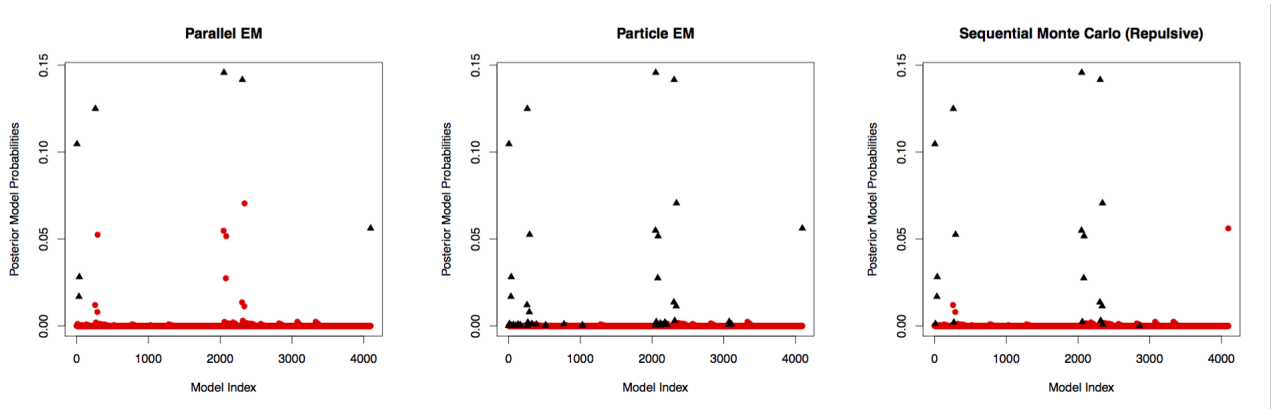
$$\widehat{\pi}_{v_0}(\boldsymbol{\gamma} \mid \boldsymbol{Y}) = \sum_{k=1}^{K} w_k(v_0) \mathbb{I}[\boldsymbol{\gamma} = \boldsymbol{\gamma}_k(v_0)] \tag{5.4}$$

indexed by $v_0 \in \{v_0^1 > \cdots > v_0^M\}$, using an ensemble of *stochastic* particles $\boldsymbol{\Gamma}(v_0)$ and importance weights $\boldsymbol{w}(v_0)$. As with Dynamic Particle EM, the particles are proliferated backwards from large v_0^1 to small (nonzero) v_0^M. However, the particle allocations are randomized and the weights are constructed according to an importance sampling routine.

Our vanilla particle filtering approach outputs a sequence of approximations (5.4) by propagating particles $\boldsymbol{\gamma}_k(v_0^m) = \boldsymbol{\gamma}_k(v_0^{m-1})$ and by updating the importance weights

$$w_k(v_0^m) \propto \frac{\pi_{v_0^m}\left[\boldsymbol{\gamma}_k(v_0^m) \mid \boldsymbol{Y}\right]}{\pi_{v_0^{m-1}}\left[\boldsymbol{\gamma}_k(v_0^{m-1}) \mid \boldsymbol{Y}\right]} w_k(v_0^{m-1}). \tag{5.5}$$

The updating scheme (5.5) is similar to the one of Shi and Dunson (2011) who used a temporal sequence of posteriors by sequentially adding observations. Also note that the posterior model probabilities $\pi_{v_0}\left[\boldsymbol{\gamma} \mid \boldsymbol{Y}\right]$ in (5.5) are available in closed form (George and McCulloch, 1997), one of the appeals of the Gaussian mixture prior. For rapidly changing posteriors (as v_0 progresses), updating just the weights without refreshing the particles may lead to increasingly less representative approximations (Gilks and Berzuini, 2001). To mitigate such particle quality decay, it is customary to rejuvenate the particles when the discrepancy between the weights is large (i.e. when most of the mass is accumulated on just a few particles). A standard strategy is to remove irrelevant particles by weighted subsampling from $[\boldsymbol{\gamma}_1(v_0^{m-1}), \ldots, \boldsymbol{\gamma}_K(v_0^{m-1})]$, thus multiplying promising particles. This strategy is often followed by a resampling step which replaces the particles by a draw from a Markov kernel with invariant measure $\pi_{v_0}(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ (Gilks and Berzuini, 2001). To this end, we can apply our repulsive proposal kernel introduced in Section 5.1. The hope is that with mutual repulsion, the particles can spread out more freely to uncharted areas of the posterior. The stochastic PEM algorithm, seen as a variant of resample-move sampler of Gilks and Berzuini (2001), is summarized in Table 2. The first step consists of standard MCMC sampling from a very smooth posterior obtained with a very large variance v_0^0 and assigning equal weights $1/K$. The next step is propagating the particles and updating the weights according to (5.5). In case the variance of the weights is too large, the particles are multiplied (through sampling with replacement) and then diversified through the repulsive Markov kernel.



(a) 7 models; 61.78%; $\lambda = 0$ (b) 57 models; 95.05%; $\lambda = 1$ (c) 19 models; 85.01%; $\lambda = 1$

Figure 2: Posterior exploration with PEM (low-dimensional example). Black triangles designate visited models and the red dots are false negatives. Left: Parallel EM ($\lambda = 0$); Middle: Particle EM ($\lambda = 1$); Right: Sequential Monte Carlo with repulsive particles ($\lambda = 1$).

6 Diversified Ensemble Learning: Illustrations

High collinearity among predictors is inevitably manifested by multi-modal spike-and-slab posteriors which can be hostile for exploration. In this section, we present a series of numerical experiments demonstrating the potential of Particle EM for learning such posterior distributions, both in low and high-dimensional settings. We begin with an intentionally simple example, which enables the exact enumeration of the model space. Later, we showcase the benefits of Particle EM when $p > n$ using MCMC approximated model probabilities as a benchmark for comparison. For meaningful comparisons between Particle EM, exhaustive model search and MCMC, we focus on the case when σ^2 is known[2]. In the Supplemental material, we also provide a further illustration on a real dataset.

6.1 Low-dimensional Case

The purpose of the first experiment is to demonstrate that Particle EM acquires many more of the important modes, compared to Parallel EM, and thereby serves as a better vehicle for posterior reconstruction. The setup for our experiment is as follows. We consider $n = 50$ observations on $p = 12$ predictors that are clustered within 4 independent blocks, where the collinearity within each block is very high. Namely, $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$, where $\boldsymbol{\Sigma} = \mathrm{bdiag}(\boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_1})$ and $\boldsymbol{\Sigma_1} = (\sigma_{ij})_{i,j=1}^{3,3}$ where $\sigma_{ij} = 0.9$ for $i \neq j$ and $\sigma_{ii} = 1$. The true vector of regression coefficients

[2]For unknown σ^2, MCMC targets the marginal posterior $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ rather than the joint $\pi(\boldsymbol{\gamma}, \sigma^2 \mid \boldsymbol{Y})$.

includes one active predictor for each block, i.e. $\boldsymbol{\beta}_0 = (1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0)'$. The responses are then generated from $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathrm{I}_n)$. Throughout this section, we assume $\theta \sim \mathcal{B}(1, p)$, $v_1 = 100$ and we consider either a single value v_0 or a sequence of v_0 values.

6.1.1 Fixed v_0

We begin our exploration in a static setting, where v_0 is fixed at 0.1. The topography of the posterior $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ for $v_0 = 0.1$ is depicted in Figure 2. The figure shows all 2^{12} posterior model probabilities, ordered according to their binary numbers, and reveals the presence of a few dominant peaks. The goal of our analysis is to identify all dominant peaks besides just the global optimum.

We initialize Particle EM with a binary matrix $\boldsymbol{\Gamma}^{(0)} = [\boldsymbol{\gamma}_1^{(0)}, \ldots, \boldsymbol{\gamma}_K^{(0)}]$ with $K = 100$ particles, sampled from independent Bernoulli trials with a success probability 0.1. To begin, we first set $\lambda = 0$. This corresponds to the Parallel EM algorithm initialized independently from the $K = 100$ random locations. By not allowing the particles to communicate, this strategy collects only 7 unique modes explaining merely 61.78% of the posterior probability. This is a rather inefficient allocation of $K = 100$ particles, many of which converged to the same destination failing to discover other relevant areas of the posterior. Next, with the same set of starting vectors, we run PEM with $\lambda = 1$, i.e. allowing the particles to interact. We observe a dramatic improvement. Ensemble PEM discovers 57 unique modes which account for 95.05% of the posterior probability. With entropy diversification, particles have been allocated far more efficiently.

The plot of the discovered posterior modes is depicted in Figure 2. The black triangles designate models visited by PEM, where red dots are the false negatives. Figure 2(a) illustrates how Parallel EM (with $\lambda = 0$) misses the large portion of significant modes. Figures 2(b) showcases the benefits of letting the particles interact when $\lambda = 1$. It is worthwhile to note that a simplified variant of PEM with weights fixed at $1/K$ also succeeds in finding the majority of dominant posterior modes (28 unique models capturing 93.33% of posterior probability).

We also compare Particle EM with its stochastic variant proposed in Section 5.2. Using the same number of particles ($K = 100$), we run the algorithm in Table 2 (with $\epsilon = 10^{-4}$) using again $v_0 \in \{0.01 + k \times 0.01; k = 0, \ldots, 50\}$. Starting with $v_0 = 0.51$ ($k = 50$) and stopping the sequential procedure at $v_0 = 0.1$, particle filtering acquires 19 unique modes that account for 85% of posterior probability (Figure 2(c)). In order to mimic the Particle EM performance, more particles are needed for the stochastic variant.

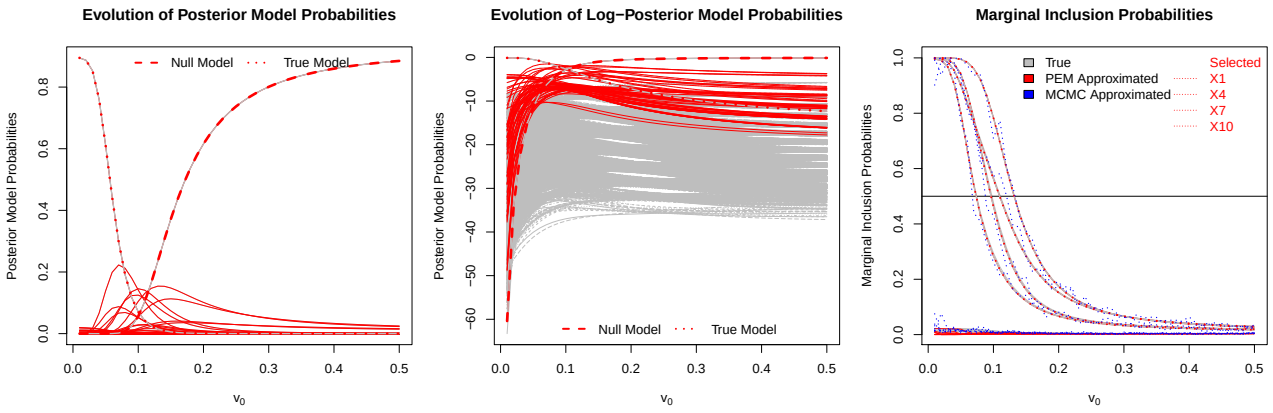
$v_0 = 0.1$	K=10				K=50				K=100				MCMC	MCMC
λ	0	1	2	3	0	1	2	3	0	1	2	3	$T = 100$	$T = 1\,000$
#Modes	4.22	8.98	9.97	10	9.26	33.51	46.93	48.77	12.03	59.33	89.16	95.7	17.68	56.09
% Posterior	0.57	0.77	0.7	0.66	0.76	0.94	0.93	0.9	0.82	0.97	0.97	0.96	0.88	0.97
Global Mode	88	97	95	91	99	100	100	100	99	100	100	100	100	100

Table 3: Simulation results (the low-dimensional case): given the 100 repetitions # stands for the average number of unique models found, % stands for the average percentage of posterior mass captured, "Global Mode" stands for the number of times a global mode was found. The last two columns correspond to the SSVS sampler with T iterations, the remaining columns are PEM with various degrees of repulsion and K.

We repeat the experiment 100 times, generating new responses $\boldsymbol{Y}$ and starting values $\boldsymbol{\Gamma}^{(0)}$ for each repetition. We consider different numbers of particles K as well as different degrees of repulsion λ, assuming $v_0 = 0.1$. The results are summarized in Table 3. Particle EM with $\lambda = 1$ outperforms Parallel EM ($\lambda = 0$) by a large margin, both in terms of the amount of probability mass explained with unique particles and in its success at finding the global mode. It is also interesting to note that higher degrees of repulsion ($\lambda > 1$) here worsen the performance. The study suggests that $\lambda = 1$, which follows from the variational calculus, here yields the right balance between fit and diversity. We compared Particle EM to a benchmark MCMC analysis with the SSVS sampler of George and McCulloch (1993). Initializing at origin, the Markov chain needed to run for around $1\,000$ iterations to find a similar set of dominant modes as PEM with $K = 100$ particles. As shown in the Supplemental material, one iteration of PEM is comparable to K iterations of SSVS. PEM took 0.096 seconds with $K = 100$.

6.1.2 Sequence of v_0 values

Because the implementation of PEM is very efficient, we can afford to explore an entire path of sliding posteriors when v_0 is varied. We demonstrate this dynamic extension on the same dataset as in Section 6.1.1 (Figure 2). For the following sequence $v_0 \in \{0.01 + k \times 0.01; k = 0, \ldots, 50\}$, we propagate the particles (i.e. reinitialize PEM at the output from a each previous run) from larger to smaller v_0 values. The dynamic plots of posterior probabilities (on absolute and log scales) is depicted in Figure 4. As the resolution gets higher (i.e. v_0 gets smaller), the posterior depreciates the null model and attributes increasingly more probability to the true model. The red curves that overlay the grey correspond to models visited by Particle EM. On the absolute scale (Figure 4(a)), we see that none of



(a) Model probabilities: log scale (b) Model probabilities: log scale (c) Variable inclusion probabilities

Figure 3: Posterior exploration with Particle EM (PEM). Red lines designate visited models (on the left and in the middle). Marginal inclusion probabilities (exact and PEM approximated) on the right.

the important models were missed by PEM. On the log-scale (Figure 4(b)), we see that Particle EM largely concentrates on very good models. However, smaller values v_0 yield spiky posteriors that are less friendly for PEM (as would be the case for any other algorithm, stochastic or deterministic).

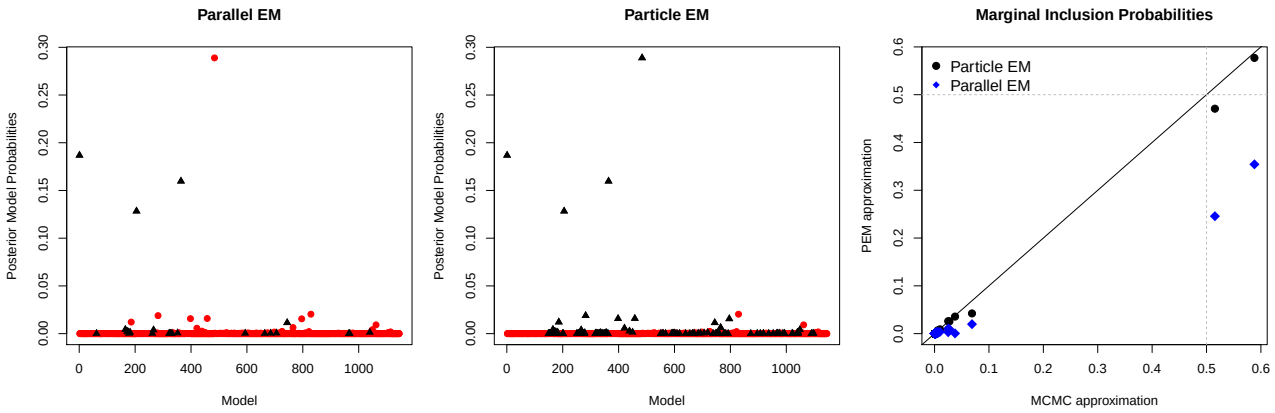
Figure 4(c) then shows the marginal inclusion probabilities. These probabilities can be obtained from the PEM approximation in (2.3) as follows

$$\widehat{P}(\gamma_i = 1 \,|\, \boldsymbol{Y}) = \sum_{k=1}^{K} \widehat{w}_k \, \widehat{\gamma}_{ki}, \tag{6.1}$$

where $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \ldots, \widehat{w}_K)'$ are the normalized posterior model probabilities at the estimated particle locations $\widehat{\boldsymbol{\gamma}}_k = (\widehat{\gamma}_{k1}, \ldots, \widehat{\gamma}_{kp})'$, $k = 1, \ldots, K$. As the resolution gets higher, the four true variables are rewarded with increasing posterior inclusion probability. With the median probability model rule, one would select the true model as long as v_0 is sufficiently small. We also performed a benchmark MCMC analysis (using the SSVS sampler of George and McCulloch (1993)) using $1\,000$ samples from the individual posteriors. The MCMC approximated marginal inclusion probabilities are plotted in dotted lines in Figure 4(c). Longer Markov chains would be needed to obtain a smoother estimate of these probabilities.

6.2 High-dimensional Case

We now confirm the potential of Particle EM on a more challenging example with $p > n$. We assume $p = 200, n = 100$ and $\boldsymbol{\Sigma} = \text{bdiag}\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_1\}$ with $\boldsymbol{\Sigma}_1 = \{\sigma_{ij}\}_{i,j=1}^{10}$ and $\sigma_{ij} = 0.99$ when $i \neq j$



(a) 27 models; 50.39%; $\lambda = 0$ (b) 173 models; 90.47%; $\lambda = 1$ (c) Marginal Inclusion Probabilities

Figure 4: Posterior exploration with PEM, the high-dimensional example. Black triangles designate visited models, red dots are false negatives. Left: Parallel EM ($\lambda = 0$); Middle: Particle EM ($\lambda = 1$); Right: Marginal inclusion probabilities.

and $\sigma_{ii} = 1$. There are 20 blocks of 10 nearly perfectly collinear variables, creating nontrivial multimodality issues. The true vector $\boldsymbol{\beta}_0$ has $q = 4$ nonzero entries at locations $(1, 11, 21, 31)$, where the magnitude of the nonzero effects is $(1.5, 2, 2.5, 3)$. To obtain benchmark model probabilities, we run $T = 100\,000$ iterations of the SSVS sampler of George and McCulloch (1993) to approximate the posterior distribution $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$. As before, we first assume a static scenario with a single value $v_0 = 0.08$ and $v_1 = 100$. Later, we consider a dynamic variant, sliding the posterior over a ladder of v_0 values. Throughout this section, we again assume $\theta \sim \mathcal{B}(1, p)$.

6.2.1 Fixed v_0

Out of the 2^{200} possible models, the MCMC trajectory with fixed $v_0 = 0.08$ visits $1\,187$ unique models whose estimated frequencies (used as benchmark model probabilities) are depicted in Figure 4. The most frequently visited model here consisted of the two variables $\{X_{21}, X_{31}\}$. With this benchmark approximation to $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$, we turn to the challenge of finding posterior modes. First, we apply the EMVS procedure of Rockova and George (2014) with a single value $v_0 = 0.08$ and a deterministic annealing starting vector. EMVS outputs the null model, which constitutes 18.68% of the total posterior mass (based on the benchmark MCMC approximation). Hoping to recover a larger set of dominant modes, we run Particle EM with $K = 200$. We initialize the computation with a matrix of zeroes $\boldsymbol{\Gamma}^{(0)} = \mathbf{0}_{p,K}$, reflecting our anticipation of sparsity. Setting $\lambda = 1$, we find 173

unique models explaining 90.47% of the posterior mass (as seen from Figure 4(b)). Similar success is achieved when the weights w_k are fixed to $1/K$. In contrast, we were not able to replicate this remarkable performance with the parallel EM version (with $\lambda = 0$). Initializing at a random binary matrix, with around 1% of ones[3], Parallel EM captures merely 50.39% of the posterior mass with 27 unique models, missing the global mode (Figure 4(a)). This is further evidence that interaction among particles improves performance. We also compare PEM to a shorter MCMC run of the SSVS sampler (initialized at a zero vector) to see how many MCMC iterations are needed to find similar dominant modes. The Markov chain had to run for around $T = 1\,000$ iterations in order to encounter modes consisting of around 90% posterior probability. While PEM with $K = 200$ took around 7 seconds, this MCMC run was about 5 times slower. In the Supplemental material, we discusses computational times and complexity in more detail.

With the Particle EM output, we can compute estimates of marginal quantities such as marginal inclusion probabilities (MIP) for each variable (according to (6.1)). Figure 4(c) shows the MIP estimates obtained from Particle EM ($\lambda = 1$; black dots) against the MCMC approximated MIPs. Recall that the MCMC benchmark approximation was obtained with $T = 100\,000$ posterior samples. PEM, on the other hand, used merely $K = 200$ particles. However, both of these posterior approximations are comparable in the sense that they yield very similar MIP estimates. Parallel EM, on the other hand, underestimated inclusion probabilities of important variables by having missed important modes (blue diamonds in Figure 4(c)).

We again repeated the experiment 100 times, generating new responses and new starting vectors for Parallel EM for each repetition. The results are summarized in Table 4. Again, we observe a marked improvement of Particle EM ($\lambda > 0$) over Parallel EM ($\lambda = 0$). Even with only $K = 50$ particles, Particle EM (with $\lambda = 1$) succeeded in capturing (on average) 86.15% of the posterior mass, finding the most frequently visited model 96 times out of the 100 repetitions. In contrast, Parallel EM finds the global mode only 49 times, explaining (on average) only 39.46% of the probability.

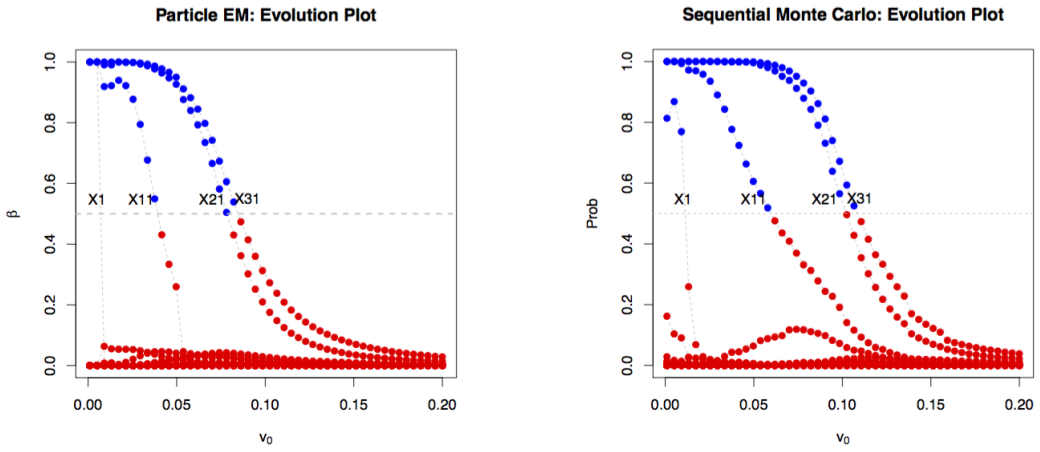
6.2.2 Sequence of v_0 values

With $v_0 = 0.08$, only two variables $\{X_{21}, X_{31}\}$ pass the selection threshold (according to the median probability model rule). Rather than conditioning on a single v_0 value, we explore the evolution of inclusion probabilities as v_0 is varied. To this end, we perform Dynamic Particle EM mode detection

[3]Initializations at denser matrices performed no better.

$v_0 = 0.08$	K=50				K=100				K=200			
λ	0	1	2	3	0	1	2	3	0	1	2	3
#Modes	10.13	42.72	48.57	50	16.47	85.85	97.98	100	24.11	171.9	167.48	154.95
% Posterior	39.46	86.15	61.54	24.45	45.17	89.01	70.73	61.42	54.64	90.52	75.85	61.76
"Global Mode"	49	96	68	29	50	97	78	68	62	98	83	68

Table 4: Simulation results (the high-dimensional case): given the 100 repetitions # stands for the average number of unique models found, % stands for the average percentage of posterior mass captured, "Global Mode" stands for the number of times the most frequently visited model by MCMC was found.



(a) Particle EM ($\lambda = 1$)

(b) Sequential Monte Carlo ($\lambda = 1$)

Figure 5: Evolution plots of marginal inclusion probabilities. Left: Repulsive particle filtering. Right: Repulsive Particle EM

by considering a sequence of values $v_0 \in \{0.001+k\times 0.004; k = 0, \ldots, 50\}$, proliferating the particles sequentially from large values to small values v_0. The evolution plot of inclusion probabilities (Figure 5(a)) shows that with higher resolution (small v_0), the posterior singles out the 4 true predictors.

As a final step, we run the sequential Monte Carlo procedure with repulsive particles (according to the scheme in Table 2). With $K = 5\,000$ and $\lambda = 1$, we proliferate particles with an occasional rejuvenation step using again $v_0 \in \{0.001+k\times 0.004; k = 0, \ldots, 50\}$. The evolution plot of marginal inclusion probabilities is displayed in Figure 5(b). We can clearly see agreement between Particle EM and sequential Monte Carlo. They have both been effective. Nevertheless, Particle EM needed only $K = 200$ particles to obtain this output.

7 Closing Remarks

We have proposed the Particle EM algorithm, a deterministic counterpart to particle Monte Carlo for Bayesian variable selection. Particle EM can be viewed as a variational Bayes approach for obtaining the best multi-point approximation to a multi-modal posterior.

We have focused on the problem of variable selection with spike-and-slab priors, for which multi-modal posteriors are commonplace. However, the ideas behind Particle EM reach far beyond this framework, opening up new directions for ensemble mode discovery. Notable efforts along these lines have already been made (Hans et al., 2007; Zhang, 2010; Gershman et al., 2012). In particular, Zhang (2010) proposed the Plus algorithm for tracking multiple local maximizers in non-concave regularization. In contrast, Particle EM capitalizes on the mixture representation of the posterior by tracking modes in the underlying latent variable space. As such, Particle EM can be generalized to more generic mixture problems.

A crucial new aspect of Particle EM is the mutual repulsion among the particles. In this way, our paper contributes to the growing literature on probabilistic models for diversification (Kulesza and Taskar, 2012). An interesting future direction is exploring the usefulness of determinantal point process priors for our setup. The appealing aspect of determinantal priors is their ability to account for an actual distance between particles, not only whether or not they are the same. First steps in this direction were taken by Ročková et al. (2016), who proposed determinantal penalty functions for ensemble regularization. However, the implementation of determinantal point processes requires costly matrix inversions. Our entropy-based diversification is far more practical.

The need for diversification among particles in sequential Monte Carlo has long been recognized. The idea of incorporating diversifying priors into a proposal distribution in static setups was pioneered by Robert and Mengersen (2003). We have expanded on this idea by introducing a repulsive entropy-based proposal kernel for sequential Monte Carlo implementations.

There are two aspects of our Gausian mixture product prior that make Particle EM computationally attractive: (1) continuity (no point mass at zero) and (2) independence correlation structure. Ročková and George (2014b) propose EMVS extensions for the point-mass mixture prior as well as the g-prior. While these modifications required an approximate E-step, Particle EM would require an approximate M-step. There is no closed form M-step for the point-mass mixture prior even in the simplest case when $K = 1$, where a joint update of $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ is required. The continuous

independent product prior, on the other hand, makes the M-step calculations far more feasible by separating the indicators $\boldsymbol{\gamma}$ from the likelihood, where each γ_i can be updated separately. Extensions to other continuous mixture priors (such as the Spike-and-Slab LASSO prior of Ročková (2017)) are in principle possible with an approximate E-step. Our Gaussian mixture prior, on the other hand, is conditionally conjugate and thereby yields the E-step in closed form.

References

Bottolo, L. and Richardson, S. (2010), "Evolutionary stochastic search for Bayesian model exploration," *Bayesian Analysis*, 5, 583–618.

Carlin, B. P. and Chib, S. (1995), "Bayesian model choice via Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society. Series B*, 57, 473–484.

Clyde, M., Ghosh, J., and Littman, M. (2011), "Bayesian adaptive sampling for variable selection and model averaging," *Journal of Computational and Graphical Statistics*, 20, 80–101.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, 39, 1–38.

Doucet, A., De Freitas, N., and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, Springer-Verlag.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

George, E. I. and McCulloch, R. E. (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–373.

Gershman, S., Hoffman, M., and Blei, D. (2012), "Nonparametric variational inference," in "Proceedings of the 29th International Conference on Machine Learning," .

Geyer, C. (1991), "Markov chain Monte Carlo maximum likelihood," in "In Computing Science and Statistics: Proceedings of the 23^{rd} Symposium on the Interface," .

Ghosh, J. and Clyde, M. (2011), "Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach," *Journal of the American Statistical Association*, 106, 1041–1052.

Gilks, W. and Berzuini, C. (2001), "Following a moving target - Monte Carlo inference for dynamic Bayesian models," *Journal of the Royal Statistical Society. Series B*, 63, 127–146.

Griffin, J., Latuszynski, K., and Steel, M. (2014), *Individual adaptation: an adaptive MCMC scheme for variable selection problems*, Technical report, School of Mathematics, Statistics and Actuarial Science, University of Kent.

Hahn, R. and Carvalho, C. (2015), "Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective," *Journal of the American Statistical Association*, 110, 435–448.

Hans, C., Dobra, A., and West, M. (2007), "Shotgun stochastic search for "large p" regression," *Journal of the American Statistical Association*, 102, 507–516.

Ishwaran, H. and Rao, J. S. (2003), "Detecting differentially expressed genes in microarrays using Bayesian model selection," *Journal of the American Statistical Association*, 98, 438–455.

Ishwaran, H. and Rao, J. S. (2005), "Spike and slab variable selection: frequentist and Bayesian strategies," *The Annals of Statistics*, 33, 730–773.

Ishwaran, H. and Rao, J. S. (2011), "Consistency of spike and slab regression," *Statistics and Probability Letters*, 81, 1920–1928.

Jaakkola, T. and Jordan, M. (1998), "Improving the mean field approximation via the use of mixture distributions," in "Learning in Graphical Models," volume Volume 89 of the series NATO ASI Series pp 163-173.

Jasra, A., Stephens, D., and Holmes, C. (2007), "On population-based simulation for static inference," *Statistics and Computing*, 17, 263–279.

Kulesza, A. and Taskar, B. (2012), "Determinantal point processes for machine learning," *Foundations and Trends in Machine Learning*, 5, 1–120.

Liang, F. and Wong, W. (2000), "Evolutionary Monte Carlo: Applications to c_p model sampling and change point problem," *Statistica Sinica*, 10, 317–342.

Ma, L. (2015), "Scalable Bayesian model averaging through local information propagation," *Journal of the American Statistical Association*, 110, 795–809.

Madigan, D., York, J., and Allard, D. (1995), "Bayesian graphical models for discrete data," *International Statistical Review / Revue Internationale de Statistique*, 63, 215–232.

Narisetty, N. and He, X. (2014), "Bayesian variable selection with shrinking and diffusing priors," *The Annals of Statistics*, 42, 789–817.

Ormerod, J. T., You, C., and Müller, S. (2014), "A variational Bayes approach to variable selection," *Manuscript*.

Robert, C. and Mengersen, K. (2003), "IID sampling using self-avoiding population monte carlo: the pinball sampler," *Bayesian Statistics*, 7, 277–292.

Ročková, V. (2017), "Bayesian estimation of sparse signals with a continuous spike-and-slab prior," *The Annals of Statistics (to appear)*.

Ročková, V. and George, E. (2014a), "EMVS: The EM approach to Bayesian variable selection," *Journal of the American Statistical Association*, 109, 828–846.

Ročková, V. and George, E. (2014b), "Negotiating multicollinearity with spike-and-slab priors," *Metron*, 72, 217–229.

Ročková, V. and George, E. (2016), "The Spike-and-Slab LASSO," *Journal of the American Statistical Association*.

Ročková, V., Moran, G. E., and George, E. (2016), "Determinantal regularization for ensemble variable selection," in "Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS)," .

Saul, K. and Jordan, M. (1996), "Exploiting tractable substructures in intractable networks," *Advances in Neural Information Processing Systems*, pages 435–442.

Schafer, C. and Chopin, N. (2013), "Sequential Monte Carlo on large binary sampling spaces," *Statistics and Computing*, 23, 163–184.

Shi, M. and Dunson, D. (2011), "Bayesian variable selection via particle stochastic search," *Statistics & Probability Letters*, 81, 283–291.

Strens, M. (2003), "Evolutionary MCMC sampling and optimization in discrete spaces," in "Proceedings of the 20^{th} International Conference on Machine Learning (ICML)," .

Ueda, N. and Nakano, R. (1998), "Deterministic annealing EM algorithm," *Neural Networks*, 11, 271–282.

Wand, M., Ormerod, J., Paroan, S., and Frühwirth, R. (2011), "Mean field variational Bayes for elaborate distributions," *Bayesian Analysis*, 6, 847–900.

Wang, J., Liang, F., and Ji, Y. (2016), "An ensemble EM algorithm for Bayesian variable selection," *Manuscript*.

Yoshida, R. and West, M. (2010), "Bayesian learning in sparse graphical factor models via variational mean-field annealing," *Journal of Machine Learning Research*, 11, 1771–1798.

Zhang, C. H. (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38, 894–942.