

Particle Swarm Optimization Assisted Metropolis Hastings Algorithms

Matthew Simpson¹

Abstract

Fitting dependent data models is often a challenging endeavor and often requires some form of dimension reduction or customized estimation algorithm. Particle swarm optimization (PSO) refers to a class of heuristic optimization algorithms that exploit analogies with animal flocking behavior in order to obtain optima without strong conditions on the objective function. We introduce two new classes of PSO algorithms, termed adaptively tuned PSO (AT-PSO) and adaptively tuned bare bones PSO (AT-BBPSO). In both algorithms we add a dynamically tuned parameter to previously existing PSO algorithms. We propose using these PSO algorithms to approximate Bayesian posterior distributions in order to construct efficient proposals for independent Metropolis-Hastings and independent Metropolis-Hastings within Gibbs algorithms. For the latter, we propose using a global approximation to the posterior in order to construct an approximation to the conditional distribution which requires a Metropolis step, thereby requiring the optimization algorithm only once rather than every iteration of the Markov chain Monte Carlo (MCMC) algorithm. In order to illustrate our method and compare it to alternatives, we provide a simulation study and apply it to constructing MCMC algorithms to estimate reduced rank spatial models of American Community Survey (ACS) 5-year estimates of county populations in the United States.

KEY WORDS: Bayesian estimation; Markov chain Monte Carlo; Official statistics; Optimization; Spatial data

¹(to whom correspondence should be addressed) Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100, themattsimpson@gmail.com

1 Introduction

[REWRITE THIS PARAGRAPH - IS TALL DATA WORTH MENTIONING?] One common type of big data problem may be called “tall” data — that is many observations of a relatively small number of variables. Bayesian estimation of tall data models can be particularly challenging in the dependent data and non-Gaussian settings due to the need to estimate or integrate out a high dimensional latent process, e.g., with dimension equal to the size of the dataset. Sometimes the latent process can be written as a function of a relatively small number of latent random variables, but, even in this case, standard Markov chain Monte Carlo (MCMC) techniques to approximate the posterior are often still slow. A common alternative is to leverage normal approximations to the posterior to obtain approximations to the marginal distributions of each parameter, e.g., INLA (Rue et al., 2009). This approximation is often very good and much faster than MCMC, though neither advantage is universally true, particularly when features of the joint posterior distribution rather than just the marginals are desired or when the parameter space is too large (e.g., see Taylor and Diggle (2014)). [CHECK ON THE JOINT VS MARGINAL THING AND OTHER INLA LIMITATIONS — CITATIONS!!!] Another common strategy is to use Hamiltonian Monte Carlo (HMC) (Neal et al., 2011), particularly the No-U-Turn sampler (NUTS) (Homan and Gelman, 2014) which is implemented in the Stan software (Carpenter et al., 2015). However HMC requires many log posterior evaluations per iteration and in the tall data setting these are typically expensive.

We propose using old MCMC technology improved by new optimization techniques. It is well known that a Bayesian posterior distributions for a fixed set of parameters tends asymptotically to a normal distribution centered at the posterior mode as the sample size increases; see Schervish (1997, Chapter 7.4). This normal approximation, often called a Laplace approximation, is used by INLA but is also frequently used to construct independent

Metropolis-Hastings (IMH) samplers (Metropolis et al., 1953; Hastings, 1970) — typically using a t proposal instead of a normal. IMH samplers based on a good approximation to the posterior distribution typically have high acceptance rates along with fast mixing and convergence, but it is often impractical to find an adequate approximation. Even in cases where the normal approximation is appropriate the posterior may be too high dimensional for this approach to be practical. In larger models numerical methods are usually required to find the posterior mode and sometimes also to integrate out a latent process, but standard methods for doing so are usually prohibitively slow or fail outright. However, new heuristic optimization algorithms tend to work reasonably well over a wider class of objective functions and in larger parameter spaces, for example, genetic algorithms (Goldberg and Holland, 1988) and particle swarm optimization (PSO) (Clerc and Kennedy, 2002; Blum and Li, 2008; Clerc, 2010). We propose using PSO to obtain the posterior mode so that effective independent Metropolis samplers can be constructed in a wider range of models. Other heuristic optimization algorithms may be useful for the same task, but we do not explore them here. We also develop several novel PSO algorithms that help obtain good estimates of the posterior mode and that may also have utility in other optimization contexts. We illustrate our method on two different modeling applications. The first application considers a group of reduced rank spatial models of American Community Survey (ACS) 5-year period estimates of county population estimates for the U.S. in 2014. In the second application, we adapt a hierarchical model for predicting the outcome of the 1988 presidential election from Gelman and Hill (2006). We fit these models using our proposed independent Metropolis-Hastings algorithm, an independent Metropolis within Gibbs algorithm based on the same Laplace approximation, and a variety of other MCMC algorithms. We compare the computational cost for each of the algorithms and discuss their ease of use.

[REWRITE THIS PARAGRAPH - SECTIONS HAVE CHANGED]

2 Particle swarm optimization

We briefly describe PSO here; refer to Blum and Li (2008) for an excellent introduction and Clerc (2010) for more detail. Suppose that we wish to optimize some objective function $Q(\boldsymbol{\theta}) : \mathbb{R}^D \rightarrow \mathbb{R}$ — without loss of generality we will assume the goal is maximization. Let $i = 1, 2, \dots, n$ index a set of particles over time, $t = 1, 2, \dots, T$, where in every period each particle consists of a location $\boldsymbol{\theta}_i(t) \in \mathbb{R}^D$, a velocity $\mathbf{v}_i(t) \in \mathbb{R}^D$, a personal best location $\mathbf{p}_i(t) \in \mathbb{R}^D$, and a group best location $\mathbf{g}_i(t) \in \mathbb{R}^D$. Here we mean “best” in the sense of maximizing Q , so $Q(\mathbf{p}_i(t)) \geq Q(\boldsymbol{\theta}_i(s))$ for any $s \leq t$. The group best location is defined with respect to some neighborhood \mathcal{N}_i of particle i ; that is, $\mathbf{g}_i(t) = \arg \max_{\{\mathbf{p}_j | j \in \mathcal{N}_i\}} Q(\mathbf{p}_j(t))$. In the simplest case where the entire swarm is the neighborhood of each particle, $\mathbf{g}_i(t) \equiv \mathbf{g}(t) = \arg \max_{j \in 1:n} Q(\mathbf{p}_j(t))$. The generic PSO algorithm updates as follows:

$$\begin{aligned}
\mathbf{v}_i(t+1) &= \omega \mathbf{v}_i(t) + \phi_1 \mathbf{r}_{1i}(t) \circ \{\mathbf{p}_i(t) - \boldsymbol{\theta}_i(t)\} + \phi_2 \mathbf{r}_{2i}(t) \circ \{\mathbf{g}_i(t) - \boldsymbol{\theta}_i(t)\}, \\
\boldsymbol{\theta}_i(t+1) &= \boldsymbol{\theta}_i(t) + \mathbf{v}_i(t+1), \\
\mathbf{p}_i(t+1) &= \begin{cases} \mathbf{p}_i(t) & \text{if } Q(\mathbf{p}_i(t)) \geq Q(\boldsymbol{\theta}_i(t+1)) \\ \boldsymbol{\theta}_i(t+1) & \text{otherwise,} \end{cases} \\
\mathbf{g}_i(t+1) &= \arg \max_{\{\mathbf{p}_j(t+1) | j \in \mathcal{N}_i\}} Q(\mathbf{p}_j(t+1)), \tag{1}
\end{aligned}$$

where \circ denotes the Hadamard product (element-wise product), $\mathbf{r}_{1i}(t)$ and $\mathbf{r}_{2i}(t)$ are each vectors of D random variates independently generated from the $U(0, 1)$ distribution, and $\omega > 0$, $\phi_1 > 0$, and $\phi_2 > 0$ are user-defined parameters. The term $\omega \mathbf{v}_i(t)$ controls the particle’s tendency to keep moving in the direction it is already going, so ω is called the inertia parameter. For $\omega < 1$ velocities tend to decrease over time, while for $\omega > 1$ they tend to increase over time. Similarly $\phi_1 \mathbf{r}_{1i}(t) \circ (\mathbf{p}_i(t) - \boldsymbol{\theta}_i(t))$ controls the particle’s tendency to move towards its personal best location while $\phi_2 \mathbf{r}_{2i}(t) \circ (\mathbf{g}_i(t) - \boldsymbol{\theta}_i(t))$ controls its tendency to move toward the group’s best location, so ϕ_1 and ϕ_2 are called the cognitive correction

factor and social correction factor, respectively (Blum and Li, 2008). This version of PSO is equivalent to Clerc and Kennedy (2002)’s constriction type I particle swarm. There are many variants of the PSO algorithm, often obtained through choosing (ω, ϕ_1, ϕ_2) in special ways or sometimes even dynamically. A default version of the algorithm sets $\omega = 0.7298$ and $\phi_1 = \phi_2 = 1.496$; see Clerc and Kennedy (2002) and Blum and Li (2008) for justification of these choices. Even when $\omega < 1$, if ϕ_1 and ϕ_2 are set high enough the velocities of the particles can continually increase and cause the swarm to make jumps that are much too large. A heavy handed way to solve this problem is by setting an upper bound on the velocity of any particle in any given direction, called velocity clamping. However, the default parameter values suggested by Clerc and Kennedy (2002) are also designed to prevent exactly this sort of velocity explosion.

Any PSO variant can also be combined with various neighborhood topologies that control how the particles communicate to each other. The default global topology allows each particle to see each other particle’s previous best location for the social components of their respective velocity updates, but this can cause inadequate exploration and premature convergence. Alternative neighborhood topologies limit how many other particles each particle can communicate with. For example, particle 5 may only look at itself and particles 4 and 6 when determining what its group best location is. This allows information about high value locations in the domain of the objective function to eventually reach every particle in the swarm, but much more slowly so that each particle has an opportunity to explore the space more fully first. Appendix A contains a short description of the ring topologies, but there are many alternatives in the literature.[CITATION]

A variant of PSO, the bare bones PSO algorithm (BBPSO) is a PSO algorithm introduced by Kennedy (2003) that strips away the velocity term and removes the need for the user to choose parameters outside of the swarm size. Let $\theta_{ij}(t)$ denote the j th coordinate of the position for the i th particle in period t , and similarly for $p_{ij}(t)$ and $g_{ij}(t)$. Then the BBPSO

algorithm obtains a new position coordinate θ_{ij} via

$$\theta_{ij}(t+1) \sim N\left(\frac{p_{ij}(t) + g_{ij}(t)}{2}, |p_{ij}(t) - g_{ij}(t)|^2\right) \quad (2)$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and standard deviation σ^2 . The updates of $\mathbf{p}_i(t)$ and $\mathbf{g}_i(t)$ are the same as in (1). There are several variants of this algorithm proposed including using distributions from different location-scale families — e.g., using the t -distribution and modifying the location or the scale parameters, for example Krohling et al. (2009), Hsieh and Lee (2010), Richer and Blackwell (2006), and Campos et al. (2014). Appendix A contains more detail on the BBPSO variants in the literature which we employ.

2.1 Adaptively tuned BBPSO

BBPSO adapts the size effective search space of the swarm over time through the variance term: $|p_{ij}(t) - g_{ij}(t)|^2$. As the personal best locations of the swarm move closer together, these variances decrease and the swarm tries locations which are closer to known high value areas in the space. This behavior is desirable, but the adaptation is forced to occur through one channel: the personal and group best locations. If the personal best locations of the swarm are arranged in a rough ring around the global optimum, smaller variances are desirable so that the new particle locations have a tendency to be in the center of the ring. On the other hand, if the personal best locations of the swarm are all to one side of the optimum and fairly far away, larger variances are desirable so that the particles can quickly approach the neighborhood of the optimum. BBPSO cannot distinguish between these two cases. [CAN ADD A GRAPHIC ILLUSTRATING THIS. WORTH IT?]

In order to allow it to adapt in a more flexible manner, we modify the BBPSO variance to $\sigma^2(t)|p_{ij}(t) - g_{ij}(t)|^2$ and tune $\sigma^2(t)$ in a manner similar to adaptive random walk Metropolis MCMC algorithms (Andrieu and Thoms, 2008). Define the improvement rate of the swarm in period t as $R(t) = \#\{i : Q(\mathbf{p}_i(t)) > Q(\mathbf{p}_i(t-1))\}/n$ where if A is a set then $\#A$ is

the number of members of that set, and let R^* denote the target improvement rate. Then what we term adaptively tuned BBPSO (AT-BBPSO) updates personal best and group best locations as in (1), then updates particle locations as follows:

$$\begin{aligned}\theta_{ij}(t+1) &\sim t_{df} \left(\frac{p_{ij}(t) + g_{ij}(t)}{2}, \sigma^2(t) |p_{ij}(t) - g_{ij}(t)|^2 \right), \\ \log \sigma^2(t+1) &= \log \sigma^2(t) + c \times \{R(t+1) - R^*\},\end{aligned}\tag{3}$$

where df is a user chosen degrees of freedom parameter and c is another user chosen parameter that controls the speed of adaptation. We use a t kernel instead of a Gaussian in order to allow for more flexibility, and smaller values of df appear to combine well with adaptively tuning $\sigma^2(t)$ — in particular $df = 1$. Since the target rate in AT-BBPSO is similar to the target rate in an adaptive random walk Metropolis algorithm, a priori values around 0.25 and in particular below 0.5 seem reasonable (Gelman et al., 1996). In practice we find that values from $R^* = 0.3$ to $R^* = 0.5$ tend to yield good AT-BBPSO algorithms. The parameter c controls the speed of adaptation so that larger values of c mean the algorithm adapts $\sigma^2(t)$ faster. We find that $c = 0.1$ to be a good value, though anything within an order of magnitude often yields similar results. The initial value $\sigma^2(0)$ also needs to be chosen, though this does not have much impact on the algorithm as long as c is not too small. We use $\sigma^2(0) = 1$ to initialize the algorithm at the standard BBPSO algorithm.

Both using a t kernel and adding a fixed scale parameter have been discussed in the BBPSO literature, but as Kennedy (2003) notes, something about setting $\sigma = 1$ is special that causes the algorithm to work well. Another similar BBPSO algorithm in the literature comes from Hsieh and Lee (2010). They propose a modified version of BBPSO with

$$\theta_{ij}(t+1) \sim N \left(\omega \frac{p_{ij}(t) + g_{ij}(t)}{2}, \sigma^2 |p_{ij}(t) - g_{ij}(t)|^2 \right),$$

where $\omega \leq 1$ and $\sigma^2 \leq 1$ are constriction parameters that are eventually both set to one after enough iterations of the algorithm. The authors suggest dynamically adjusting the

constriction parameters in the early stage of the algorithm before they are set to one, but give no suggestion for how to do this. The default BBPSO algorithm essentially sets $\sigma^2(t) = 1$ for all t and Hsieh and Lee (2010)’s algorithm deterministically adjusts $\sigma^2(t)$, but our AT-BBPSO algorithm is able to adapt its value on the fly based on local knowledge about the objective function. If too much of the swarm is failing to find new personal best locations, AT-BBPSO proposes new locations closer to known high value areas. If too much of the swarm is improving, AT-BBPSO proposes bolder locations in an effort to make larger improvements. This ability to adapt to local information about the objective function allows AT-BBPSO to more quickly traverse the search space towards the global optimum, though by using local information AT-BBPSO does risk premature convergence to a local optimum. The adaptively tuned component of AT-BBPSO can also be combined with most BBPSO variants, some of which are outlined in Appendix A. In Appendix B we conduct a simulation study on several test functions that compares AT-BBPSO variants to other PSO and BBPSO variants in order to justify the parameter settings discussed above and demonstrate that making a BBPSO variant into an AT-BBPSO variant will tend to improve it.

2.2 Adaptively tuned PSO

In AT-BBPSO variants, the parameter $\sigma(t)$ partially controls the effective size of the swarm’s search area, and we increase or decrease $\sigma(t)$ and consequently the search area depending on how much of the swarm is finding new personal best locations. In standard PSO there is no direct analogue to $\sigma(t)$, though the inertia parameter, ω in (1), is related. It controls the effective size of the swarm’s search area by controlling how the magnitude of the velocities evolve over time — larger values of ω allow for larger magnitude velocities in future periods. In AT-BBPSO we use an analogy with tuning a random walk Metropolis-Hastings MCMC algorithm in order to build intuition about how to tune $\sigma(t)$. The analogy is much weaker

in this case; nonetheless, we allow $\omega(t)$ to be time-varying and use the same mechanism in order to tune it as we did for $\sigma(t)$.

The idea of time-varying $\omega(t)$ has been in the PSO literature for some time. An early suggestion was to set $\omega(0) = 0.9$ and deterministically decrease it until it reaches $\omega(T) = 0.4$ after the maximum number of iterations allowed (Eberhart and Shi, 2000). In particular, Tuppadung and Kurutach (2011) suggest defining $\omega(t)$ via the parameterized inertia weight function

$$\omega(t) = \frac{1}{1 + \left(\frac{t}{\alpha}\right)^\beta} \quad (4)$$

where α and β are user-defined parameters. Roughly, α controls how low $\omega(t)$ can go and β controls how fast it gets there, so α and β can be thought of as intercept and slope parameters respectively. The suggestion in Tuppadung and Kurutach (2011) is to set α to a small fraction of the total amount of iterations in which the algorithm is allowed to run (e.g., 10% or 20%), and set β between one and four.

This approach tends to improve on standard PSO if $\omega(t)$'s progression is set appropriately, but it invariably makes using PSO more difficult for the average user. Additionally, depending on the problem, it may be more useful to let the swarm explore the space for more or less iterations, necessitating different progressions of $\omega(t)$. A priori it may not be clear exactly which approach is best for any given problem, so an automatic method is desirable. Adaptively tuned PSO (AT-PSO) is just that — it provides an automatic method to adjust the value of $\omega(t)$ depending on local information obtained by the particle swarm. Formally, AT-PSO updates personal and group best locations as in (1) and updates $\omega(t)$ and particle

locations as follows:

$$\begin{aligned}
\mathbf{v}_i(t+1) &= \omega(t+1)\mathbf{v}_i(t) + \phi_1\mathbf{r}_{1i}(t) \circ \{\mathbf{p}_i(t) - \boldsymbol{\theta}_i(t)\} + \phi_2\mathbf{r}_{2i}(t) \circ \{\mathbf{g}_i(t) - \boldsymbol{\theta}_i(t)\}, \\
\boldsymbol{\theta}_i(t+1) &= \boldsymbol{\theta}_i(t) + \mathbf{v}_i(t+1), \\
\log \omega(t+1) &= \log \omega(t) + c \times \{R(t+1) - R^*\},
\end{aligned} \tag{5}$$

where $R(t)$ is the improvement rate of the swarm in iteration t , R^* is the target improvement rate, and c controls how much $R(t)$ changes on a per iteration basis. For AT-BBPSO we used an analogy with random walk Metropolis-Hastings algorithms to suggest that a good value for the target improvement rate is smaller than 0.5 but not too small and this turned out to be correct. The analogy does not apply as well here, though we still find in Appendix B that $R^* = 0.3$ or 0.5 still seems to work well for AT-PSO. The value of c controls the speed of adaptation, and in particular if c is small and $\omega(0)$ is large, AT-PSO can mimic DI-PSO to some extent. This turns out to produce poor AT-PSO algorithms, however. We use $c = 0.1$ as a default value and in simulations not reported here, we find that the gains from optimizing c appear to be small. However, very small values like those suggested by an attempt to mimic DI-PSO turn out to cause the algorithm to perform very poorly.

A major strength of AT-PSO relative to DI-PSO and standard PSO is that AT-PSO can increase $\omega(t)$ when information from the swarm suggests there is an unexplored high value region of the space — when too much of the swarm is improving on their personal best locations AT-PSO increases $\omega(t)$ until velocities start increasing, the swarm starts exploring a larger amount of the nearby space, and more of the particles fail to find improvements on their personal best. Just like in AT-BBPSO, this mechanism provides a way for the swarm to adapt its behavior on the fly based on local conditions but it can also cause premature convergence to a local optimum. While DI-PSO monotonically decreases $\omega(t)$ toward some minimum value, AT-PSO typically oscillates $\omega(t)$ rapidly so that the algorithm alternates between favoring exploration more than standard PSO and favoring exploitation more. [CAN

ADD A PLOT OF INERTIA OVER TIME HERE. WORTH IT?] Appendix B contains an extended simulation study comparing a variety of these PSO and BBPSO algorithms on a suite of test functions that motivates some of the recommendations detailed above and demonstrates that AT-PSO tends to be consistently one of the best performing PSO or BBPSO algorithms and often is the best.

3 Laplace Approximations and Independent Metropolis-Hastings

Ultimately we will use PSO algorithms to help construct MCMC algorithms for posterior sampling, in particular independent Metropolis-Hastings (IMH) algorithms based on Laplace approximations. To do this, first we introduce the Laplace approximation. Let $\boldsymbol{\theta}$ be the model parameter and $[\boldsymbol{\theta}|\mathbf{Y}_{1:n}]$ be its posterior distribution, available up to a normalizing constant. We use the generic notation $[X|Y]$ to denote the density or mass function of X given Y . Further let $\boldsymbol{\theta}_n^*$ denote the posterior mode of $[\boldsymbol{\theta}|\mathbf{Y}_{1:n}]$ and let $\mathbf{H}_n(\boldsymbol{\theta}_n^*)$ denote the Hessian matrix of $\log[\boldsymbol{\theta}|\mathbf{Y}_{1:n}]$ evaluated at $\boldsymbol{\theta}_n^*$. Then under suitable regularity conditions (Schervish, 1997, Sections 7.4.2 and 7.4.3) $\boldsymbol{\theta}$'s posterior distribution is asymptotically normal so that for a fixed but large value of n , $\boldsymbol{\theta}|\mathbf{Y}_{1:n} \stackrel{a}{\sim} N(\boldsymbol{\theta}_n^*, -\mathbf{H}_n^{-1}(\boldsymbol{\theta}_n^*))$ where the notation $\stackrel{a}{\sim}$ means “approximately distributed as.” We do not discuss the precise technical conditions necessary for this result here, but intuitively for the approximation to be good each element of $\boldsymbol{\theta}$ must have enough “observations” for asymptotics to kick in. We put “observations” in scare quotes because often rather than data, it is other elements of $\boldsymbol{\theta}$ are directly informing on another element. In order to see this more clearly we add structure to the model by considering a class of generalized linear mixed models (GLMMs) (Stroup, 2012), but by no means is this the only class of models where the Laplace approximation works well.

3.1 Laplace approximations for GLMMs

Generalized linear mixed models with latent Gaussian processes (LGP) provide a plethora of examples where PSO assisted IMH algorithms are attractive for MCMC. That latent parameters are Gaussian distributed often increases the quality of the normal approximation to the posterior, but it is not a necessary feature of the model for the approximation to work well. We will conceptualize our model class using the strategy of Berliner (1996) and Wikle et al. (2003), that is hierarchically with a data model conditional on parameters and a latent process, a process model conditional on parameters, and finally a parameter model. Let $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$ denote a vector of response variables. We assume a conditionally independent data model given a vector of location parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$ and a common dispersion parameter ϕ , i.e. $Z_i \stackrel{ind}{\sim} [Z|\mu_i, \phi]$ where $[z|\mu, \phi]$ is some known family of density functions indexed by (μ, ϕ) , often an exponential family. The location parameters are a known function g of a LGP, so $g^{-1}(\boldsymbol{\mu}) = \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\delta}$. Here $\mathbf{X}'\boldsymbol{\beta}$ represents fixed effects and $\mathbf{S}'\boldsymbol{\delta}$ represents random effects, where \mathbf{X} is a known $n \times p$ matrix, \mathbf{S} is a known $n \times r$ matrix, $\boldsymbol{\beta}$ is an unknown p -dimensional vector and $\boldsymbol{\delta}$ is an unknown r -dimensional vector. Often $r \ll n$, but in some cases $r = n$. $\boldsymbol{\delta}$ is further modeled as Gaussian with mean zero and a covariance matrix $\boldsymbol{\Sigma}$ with a structure appropriate to the specific problem. Finally a prior on $(\phi, \boldsymbol{\beta}, \mathbf{\Sigma})$ serves as the parameter model where typically $\boldsymbol{\beta}$ is assumed normally distributed a priori so that $g^{-1}(\boldsymbol{\mu})$ is an LGP. The model can be generalized to multivariate Z_i and multivariate Y_i , but we omit this complication here.

Three key factors affect the quality of the Laplace approximation for GLMMs: the sample size, how close the data model is to Gaussian at reasonable parameter values for the given dataset, and the structure of $\boldsymbol{\Sigma}$. Due to the asymptotic nature of the Laplace approximation, the larger the sample size the better the approximation is for $\boldsymbol{\beta}$ and ϕ . To the extent that the data model is already approximately normal, smaller sample sizes will still

yield reasonably good approximations. [I CAN ILLUSTRATE THIS WITH SOME SIMULATIONS IN A GLMM WITH A BETA DATA MODEL. NECESSARY?] The last factor can best be illustrated by considering two extremes: iid random effects where $\Sigma = \sigma^2 \mathbf{I}_r$, and fully correlated random effects where Σ is fully parameterized and unknown. In the iid case when r is moderately large, e.g. $r = 10$, σ^2 or more likely $\log \sigma^2$ is often approximately normal. The “observations” directly informing on σ^2 are the estimated random effects in δ rather than data, but the Laplace approximation works all the same. In the fully correlated case, the only “observation” relevant to estimating Σ is δ , and there is only one observation to estimate $r(r + 1)/2$ unique parameters. In this case it is unrealistic to expect Σ or any commonly used transformation of Σ to be approximately normal in the posterior unless there is strong information in the prior suggesting otherwise.

When $r = n$ it is more important that the data model is already approximately normal and that the random effects are a priori modeled as Gaussian because there are only n data points informing on n latent random effects. Since the dimension of δ is growing at the same rate as the sample size in that case, asymptotic arguments about δ ’s posterior distribution do not apply. Additionally, when n is large and $r = n$ it may be difficult to find the posterior mode in a reasonable amount of time using any optimization algorithm. Many GLMMs feature e.g. group specific random effects, so r is often much smaller than n and these concerns do not apply. The same principles apply to other models outside of the GLMM framework — each model parameter and random effect needs enough effective observations informing it so that the asymptotics can take hold or else other features of the model need to ensure that the parameters are approximately Gaussian in the posterior.

3.2 Laplace approximations for IMH algorithms

Given any reasonable approximation to the posterior distribution, we can construct an IMH MCMC sampler using that approximation (Robert and Casella, 2013, Chapter 7.4), though typically a t_{df} proposal is used instead with df set small enough so that its tails dominate the posterior's in order to ensure that the Markov chain is uniformly ergodic — see Robert and Casella (2013, Theorem 7.8). As long as the conditions mentioned in Section 3.1 are satisfied, the Laplace approximation is likely to be good and the IMH algorithm is likely to yield efficient MCMC. In particular, when the approximation is good, IMH has high acceptance rates, converges quickly, and mixes well. This is well known, but the problem lies in finding good approximations. The Laplace approximation is often poor even for the class of GLMMs we discussed in Section 3.1 because of the random effect covariance matrix. Often there are a small number of random effects or they are allowed to have less structured covariance matrices, so the asymptotics never kick in. Yet the random effects themselves and the fixed effect parameter are often still approximately normal in their conditional posterior. Further, the dispersion parameter and the random effect covariance matrix often have easy conditional posteriors which are easy to sample from.

This suggests an obvious alternative: construct a Metropolis within Gibbs MCMC algorithm (Robert and Casella, 2013, Chapter 10.3) with an independent Metropolis step for $(\boldsymbol{\beta}, \boldsymbol{\delta})$, i.e an independent Metropolis-Hastings within Gibbs (IMHwG) algorithm. This algorithm consists of three steps: 1) sample $\boldsymbol{\Sigma}$ directly from its full conditional distribution, 2) sample ϕ directly from its full conditional distribution, and 3) sample $(\boldsymbol{\beta}, \boldsymbol{\delta})$ from its full conditional distribution using an independent Metropolis-Hastings step with the Laplace approximation as a proposal. Typically steps 1) and 2) are independent of each other, so the algorithm only consists of two Gibbs steps.

The naive way to implement this algorithm is to construct the Laplace approximation to

$[\boldsymbol{\beta}, \boldsymbol{\delta} | \boldsymbol{\Sigma}, \phi, \mathbf{Y}, \mathbf{X}]$ every iteration of the MCMC algorithm to use it as the IMH step proposal. When the optimization is costly, this cost get multiplied by the number of MCMC iterations and can therefore be prohibitive. We call this approximation strategy the conditional Laplace approximation. Alternatively, the global Laplace approximation to the full joint posterior implies a conditional distribution for $\boldsymbol{\beta}, \boldsymbol{\delta} | \boldsymbol{\Sigma}, \phi$. Let $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\delta})$ and $\boldsymbol{\theta}_2 = (\phi, \boldsymbol{\Sigma})$, or rather that $\boldsymbol{\theta}_2$ is a vector containing ϕ and the unique elements of $\boldsymbol{\Sigma}$. Then suppose we have the global Laplace approximation to the posterior of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ given by $\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X} \stackrel{a}{\sim} N(\boldsymbol{\theta}^*, (-\mathbf{H}^*)^{-1})$ where $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)$ and

$$\boldsymbol{\Omega}^* = (-\mathbf{H}^*)^{-1} = \begin{bmatrix} \boldsymbol{\Omega}_{11}^* & \boldsymbol{\Omega}_{12}^* \\ \boldsymbol{\Omega}_{21}^* & \boldsymbol{\Omega}_{22}^* \end{bmatrix}.$$

Then the approximation to the conditional posterior of $\boldsymbol{\theta}_1$ implied by the global Laplace approximation is $\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \mathbf{Y}, \mathbf{X} \stackrel{a}{\sim} N(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\Omega}}_{11})$, where $\tilde{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1^* + \boldsymbol{\Omega}_{12}^* (\boldsymbol{\Omega}_{22}^*)^{-1} (\boldsymbol{\theta}_2^{(t+1)} - \boldsymbol{\theta}_2^*)$ and $\tilde{\boldsymbol{\Omega}}_{11} = \boldsymbol{\Omega}_{11}^* - \boldsymbol{\Omega}_{12}^* (\boldsymbol{\Omega}_{22}^*)^{-1} \boldsymbol{\Omega}_{21}^*$. We call this approximation strategy the global Laplace approximation.

The global Laplace approximation may be worse than the conditional approximation since the conditional Laplace approximation directly approximates the target conditional posterior, but the computation savings realized from only having to run the optimization algorithm once are typically significant and make it worthwhile. By moving from an IMH algorithm to an IMHwG algorithm the mixing and convergence properties of the chain do deteriorate to the extent that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are dependent in the posterior. In practice this is often not an issue, or if it is it can typically be solved by reparameterizing the model, e.g. using a non-centered parameterization (Gelfand et al., 1995; Roberts and Sahu, 1997; Van Dyk and Meng, 2001; Bernardo et al., 2003).

4 Spatially modeling county population estimates

The American Community Survey (ACS) provides 5-year period estimates of county populations as recently as 2014. In 2014 there were 3,142 counties in the United States, including the District of Columbia and counties in Alaska and Hawaii. We use two separate data models in order to illustrate when the normal approximation works well. The first is a Poisson data model, i.e., $Z_i \sim \text{Poisson}(\lambda_i)$ where $\lambda_i = \exp(Y_i)$. Through visual inspection of a histogram, it was determined that, on the log scale, county populations look approximately normally distributed. So, our second data model is $\log Z_i \sim N(\mu_i, \phi^2)$ where $\mu_i = Y_i$.

The process model in both cases is a reduced rank spatial model $Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\delta}$, where $\mathbf{X}\boldsymbol{\beta}$ represents the process mean at each county and $\mathbf{S}\boldsymbol{\delta}$ implies the spatial correlation across counties. The spatial correlation term consists of a set of r basis functions evaluated at each of the $n = 3,142$ counties, denoted by the $n \times r$ matrix \mathbf{S} , and a common random effect $\boldsymbol{\delta}$. We assume that $\boldsymbol{\delta}$ is r -dimensional with $r \ll n$ so that the model is reduced rank. Any set of spatial basis functions could be used for \mathbf{S} but we use the Moran's I (MI) basis set, described below (see Hughes and Haran (2013), Porter et al. (2015), Bradley et al. (2015) and references therein for additional discussion). Another possibility is to define a reduced rank model for a point-level spatial process using a basis function expansion and compute the implied set of basis functions for each of the census tracts by integrating the point level basis functions appropriately. See Sections 2.1, 3.1, and 4 of Bradley et al. (2016) for details.

The MI basis functions are defined through the orthogonal projection matrix $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Let \mathbf{A} denote the binary adjacency matrix with $a_{ij} = 1$ if counties i and j are neighbors, $a_{ij} = 0$ otherwise, and $a_{ii} = 0$ along the diagonal, and define the MI operator \mathbf{G} as

$$\mathbf{G} = (\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{A}(\mathbf{I}_n - \mathbf{P}_\mathbf{X})$$

where \mathbf{I}_n is the $n \times n$ identity matrix. The spectral decomposition of \mathbf{G} is $\mathbf{G} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}'$. To use

a reduced rank version of the MI basis functions we truncate the basis function expansion and take \mathbf{S} to be the $n \times r$ matrix formed by the r columns of $\mathbf{\Phi}$ corresponding to the largest r eigenvalues of \mathbf{G} . The random effect $\boldsymbol{\delta}$ is further modeled as $\boldsymbol{\delta} \sim N(\mathbf{0}_r, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ where $\mathbf{0}_r$ denotes an r -dimensional vector of zeroes and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is an unknown covariance matrix. The covariance matrix of $\mathbf{S}\boldsymbol{\delta}$ is then $\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{S}'$.

The process model depends on choices for \mathbf{X} and r . In practice r can be chosen using a sensitivity analysis. Since our goal is to illustrate computational methods, we will elide choosing r in a principled way and instead use several values for r in order to illustrate when the parameter space becomes too high dimensional for our method to be advantageous. For simplicity we choose an intercept only model, but all derivations for the model will assume that \mathbf{X} is $n \times p$.

Finally, we consider two distinct parameterizations of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ — the iid parameterization, and the full parameterization. In the iid parameterization the prior for $\boldsymbol{\Sigma}$ is $\sigma^2 \sim IG(a_\sigma, b_\sigma)$ where $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_r$, while in the full parameterization we assume that $\boldsymbol{\Sigma} \sim IW(d, \mathbf{E})$. The prior for $\boldsymbol{\beta}$ in all models is $\boldsymbol{\beta} \sim N(\mathbf{b}, v^2 \mathbf{I}_p)$, and the lognormal models the prior for ϕ^2 is $\phi^2 \sim IG(a_\phi, b_\phi)$. We assume that the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and when applicable ϕ^2 are mutually independent in the prior. For our examples we assume that $a_\sigma = a_\phi = b_\sigma = b_\phi = 1$, $\mathbf{b} = \mathbf{0}_p$, $v = 10$, $d = r + 1$ and $\mathbf{E} = \mathbf{I}_p$. Often a more complicated prior is appropriate on variance or covariance matrix parameters so that the marginal posterior is not sensitive to arbitrary choices in the prior. We use these conditionally conjugate priors because they allow for a fair comparison between MCMC algorithms — most alternatives will complicate Gibbs samplers with extra steps or necessitate Metropolis steps making IMH or IMHwG relatively more attractive.

Between the two possible parameterizations of $\boldsymbol{\Sigma}$ and the two choices for the data model — Poisson versus Lognormal — we consider four possible classes of models. Then for the

models with iid random effects the posterior distributions can be written as

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, \phi^2 | \mathbf{z}, \mathbf{X}, \mathbf{S}) \propto (\phi^2)^{-n/2-a_\phi-1} \exp \left[-\frac{1}{\phi^2} \left\{ \frac{(\log \mathbf{z} - \mathbf{y})'(\log \mathbf{z} - \mathbf{y})}{2} + b_\phi \right\} \right] \\ \times (\sigma^2)^{-\frac{r}{2}-a_\sigma-1} \exp \left\{ -\frac{1}{\sigma^2} \left(\frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{2} + b_\sigma \right) \right\} \exp \left\{ -\frac{(\boldsymbol{\beta} - \mathbf{b})'(\boldsymbol{\beta} - \mathbf{b})}{2v} \right\} \quad (\text{iid lognormal}), \quad (6)$$

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} | \mathbf{z}, \mathbf{X}, \mathbf{S}) \propto \prod_{i=1}^n \frac{\exp\{-\exp(y_i)\} \exp(y_i z_i)}{z_i!} \\ \times (\sigma^2)^{-\frac{r}{2}-a_\sigma-1} \exp \left\{ -\frac{1}{\sigma^2} \left(\frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{2} + b_\sigma \right) \right\} \exp \left\{ -\frac{(\boldsymbol{\beta} - \mathbf{b})'(\boldsymbol{\beta} - \mathbf{b})}{2v} \right\} \quad (\text{iid Poisson}). \quad (7)$$

For fully parameterized $\boldsymbol{\Sigma}$ we write the posteriors in terms of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, yielding

$$p(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\delta}, \phi^2 | \mathbf{z}, \mathbf{X}, \mathbf{S}) \propto (\phi^2)^{-n/2-a_\phi-1} \exp \left\{ -\frac{1}{\phi^2} \left(\frac{(\log \mathbf{z} - \mathbf{y})'(\log \mathbf{z} - \mathbf{y})}{2} + b_\phi \right) \right\} \\ \times |\boldsymbol{\Omega}|^{(d-r)/2} \exp \left\{ -\frac{1}{2} \left(\boldsymbol{\delta}'\boldsymbol{\Omega}\boldsymbol{\delta} + \frac{(\boldsymbol{\beta} - \mathbf{b})'(\boldsymbol{\beta} - \mathbf{b})}{v} + \text{tr}(\mathbf{E}\boldsymbol{\Omega}) \right) \right\} \quad (\text{full lognormal}), \quad (8)$$

$$p(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\delta} | \mathbf{z}, \mathbf{X}, \mathbf{S}) \propto \prod_{i=1}^n \frac{\exp\{-\exp(y_i)\} \exp(y_i z_i)}{z_i!} \\ \times |\boldsymbol{\Omega}|^{(d-r)/2} \exp \left\{ -\frac{1}{2} \left(\boldsymbol{\delta}'\boldsymbol{\Omega}\boldsymbol{\delta} + \frac{(\boldsymbol{\beta} - \mathbf{b})'(\boldsymbol{\beta} - \mathbf{b})}{v} + \text{tr}(\mathbf{E}\boldsymbol{\Omega}) \right) \right\} \quad (\text{full Poisson}) \quad (9)$$

where $\text{tr}(\cdot)$ denotes the trace operator. For some MCMC algorithms it will be easier to work with the Cholesky decomposition of $\boldsymbol{\Omega}$ given by $\mathbf{L}\mathbf{L}' = \boldsymbol{\Omega}$ where \mathbf{L} is lower triangular. In practice it is often more convenient to put the prior distribution directly on \mathbf{L} rather than on $\boldsymbol{\Omega}$ or $\boldsymbol{\Omega}^{-1}$ and solving for the Jacobian, but this depends in part on which MCMC algorithm is used to fit the model. So while we use the prior distribution on \mathbf{L} implied by a Wishart prior on $\boldsymbol{\Omega}$, in practice it is advantageous to use one of the priors suggested by Chen and Dunson (2003) or Frühwirth-Schnatter and Tüchler (2008). We allow the diagonal entries of \mathbf{L} to be negative in the independent Metropolis-Hastings algorithms in order to facilitate MCMC, so \mathbf{L} is not strictly speaking a Cholesky decomposition. The determinant

of the Jacobian is the same in both cases up to a proportionality constant. The signs of the elements of \mathbf{L} are not identified, therefore care needs to be taken when interpreting the results of MCMC. Transforming back to the precision matrix in a post processing step is sufficient. Let ℓ_{ij} denote the (i, j) th element of \mathbf{L} . Then the Jacobian of $\mathbf{\Omega} \rightarrow \mathbf{L}$ is given by

$$|J(\mathbf{\Omega} \rightarrow \mathbf{L})| \propto \prod_{k=1}^r |\ell_{kk}|^{r+1-k}$$

where $\mathbf{\Omega}$ is $r \times r$. Under this parameterization the full posteriors can be written as

$$p(\boldsymbol{\beta}, \mathbf{L}, \boldsymbol{\delta}, \phi^2 | \mathbf{z}, \mathbf{X}, \mathbf{S}) \propto (\phi^2)^{-n/2 - a_\phi - 1} \exp \left[-\frac{1}{\phi^2} \left\{ \frac{(\log \mathbf{z} - \mathbf{y})'(\log \mathbf{z} - \mathbf{y})}{2} + b_\phi \right\} \right] \prod_{k=1}^r (\ell_{kk}^2)^{(d-k+1)/2} \\ \times \exp \left[-\frac{1}{2} \left\{ \boldsymbol{\delta}' \mathbf{L} \mathbf{L}' \boldsymbol{\delta} + \frac{(\boldsymbol{\beta} - \mathbf{b})'(\boldsymbol{\beta} - \mathbf{b})}{v} + \text{tr}(\mathbf{E} \mathbf{L} \mathbf{L}') \right\} \right] \quad (\text{full lognormal}), \quad (10)$$

$$p(\boldsymbol{\beta}, \mathbf{L}, \boldsymbol{\delta} | \mathbf{z}, \mathbf{X}, \mathbf{S}) \propto \prod_{i=1}^n \frac{\exp\{-\exp(y_i)\} \exp(y_i z_i)}{z_i!} \times \prod_{k=1}^r (\ell_{kk}^2)^{(d-k+1)/2} \\ \times \exp \left[-\frac{1}{2} \left\{ \boldsymbol{\delta}' \mathbf{L} \mathbf{L}' \boldsymbol{\delta} + \frac{(\boldsymbol{\beta} - \mathbf{b})'(\boldsymbol{\beta} - \mathbf{b})}{v} + \text{tr}(\mathbf{E} \mathbf{L} \mathbf{L}') \right\} \right] \quad (\text{full Poisson}). \quad (11)$$

In Appendix C we derive the Hessian for the fully parameterized Poisson model. The other models are analogous, though the variances in the iid models and in the lognormal models should be transformed to the log scale first.

In Section 6 we consider several Gibbs sampling algorithms which draw the covariance matrix parameter $\boldsymbol{\theta}$ from its full conditional distribution. When the covariance matrix is fully parameterized the full conditional distribution of the precision matrix $\mathbf{\Omega}$ is Wishart, that is $\mathbf{\Omega} \sim W(\tilde{d}, \tilde{\mathbf{E}})$. This draw is usually accomplished via the Bartlett decomposition (Smith and Hocking, 1972). Let $\tilde{\mathbf{C}}$ be the lower triangular Cholesky decomposition of $\tilde{\mathbf{E}}$. Then let \mathbf{A} be an $r \times r$ random lower triangular matrix with independent elements $\{a_{ij} : 0 < i \leq j \leq r\}$ where $a_{ii} \sim \sqrt{\chi_{\tilde{d}-i+1}^2}$ for $i = 1, 2, \dots, r$ and $a_{ij} \sim N(0, 1)$ for $0 < i < j \leq r$. Then $\mathbf{\Omega} = \mathbf{L} \mathbf{L}' \sim W(\tilde{d}, \tilde{\mathbf{E}})$ where $\mathbf{L} = \tilde{\mathbf{C}} \mathbf{A}$. In the process of drawing $\mathbf{\Omega}$ we must first draw \mathbf{L} , so we construct our Gibbs samplers in terms of \mathbf{L} instead of $\mathbf{\Omega}$.

4.1 Predicting the 1988 presidential election

Gelman and Hill (2006, Chapter 14) describes a model used to predict state-level opinions about the 1988 presidential candidates from national polls in order to predict the outcome of the election. They model the responses to a series of seven polls conducted by CBS News during the week before the 1988 presidential election. The variable of interest is binary: $Z_i = 1$ if the i th respondent said they supported the Republican candidate and $Z_i = 0$ if they said they supported the Democratic candidate, with undecideds being excluded. Focusing on the last poll, they ultimately estimate a logistic regression model with fixed effects for race (whether the respondent was African American or not), sex, and race \times sex, and random effects for four age categories, four education categories, and 16 age \times education categories, as well as for the respondent's state of residence (including District of Columbia). The mean of the state random effect distribution is one of five region random effects plus the proportion of the state that voted republican in the last election times a slope coefficient.

We reduce the size of this model somewhat by omitting the age and education random effects, but keeping the age \times education interaction terms — so the age \times education random effects now represent the random effects for each age \times education category. The single poll data model is

$$P(Z_i = 1) = \theta_i, \quad \theta_i = \exp(Y_i) / \{1 + \exp(Y_i)\},$$

$$Y_i = \beta_0 + f_i\beta_f + b_i\beta_b + f_ib_i\beta_{fb} + \alpha_{ae}[ae_i] + \alpha_s[s_i] \quad (\text{single poll data model}), \quad (12)$$

where f_i indicates whether respondent i identified as female, b_i indicates whether respondent i identified as African American, ae_i indicates respondent i 's age \times education category, and s_i indicates respondent i 's state of residence. Here we use $\alpha_s[k]$ to denote the k th element of the vector α_s , so α_{ae} contains 16 elements, and α_s contains 51 elements (50 states plus

the District of Columbia). The single poll process model is

$$\begin{aligned}
\alpha_s[k] &\overset{ind}{\sim} N(\alpha_r[r_k] + prev_k \beta_{prev}, \sigma_s^2) \text{ for } k = 1, 2, \dots, 51, \\
\alpha_{ae}[k] &\overset{iid}{\sim} N(0, \sigma_{ae}^2) \text{ for } k = 1, 2, \dots, 16, \\
\alpha_r[k] &\overset{iid}{\sim} N(0, \sigma_r^2) \text{ for } k = 1, 2, \dots, 5 \quad (\text{single poll process model}),
\end{aligned} \tag{13}$$

where $\alpha_r[r_k]$ denotes the region containing state k and $prev_k$ denotes the average vote share for the Republicans in the previous three presidential elections. This model expands the class of models discussed at the beginning of this section by allowing the mean of δ to depend on random effects that are further modeled. Adding a level to the hierarchy does not fundamentally change the applicability of PSO assisted MCMC algorithms, so long as the parameter space is still not too large for PSO to be feasible and the normal approximation is reasonable for the additional parameters.

The last poll had 2,015 respondents, but together all seven polls have 11,566 respondents. Using each poll with a minimal number of additional parameters to account for poll to poll variability should increase the quality of the model and result in a posterior with a better Laplace approximation. We analyze a model for all of the polls using the following data model

$$\begin{aligned}
P(Z_i = 1) &= \theta_i, \quad \theta_i = \exp(Y_i) / \{1 + \exp(Y_i)\}, \\
Y_i &= \beta_0 + f_i \beta_f + b_i \beta_b + f_i b_i \beta_{fb} + \alpha_{ae}[ae_i] + \alpha_s[s_i] + \alpha_p[p_i] \quad (\text{all polls data model}),
\end{aligned} \tag{14}$$

where p_i denotes which poll respondent i was surveyed in. The process model is given by

$$\begin{aligned}
\alpha_s[k] &\overset{ind}{\sim} N(\alpha_r[r_k] + prev_k \beta_{prev}, \sigma_s^2) \text{ for } k = 1, 2, \dots, 51, \\
\alpha_{ae}[k] &\overset{iid}{\sim} N(0, \sigma_{ae}^2) \text{ for } k = 1, 2, \dots, 16, \\
\alpha_r[k] &\overset{iid}{\sim} N(0, \sigma_r^2) \text{ for } k = 1, 2, \dots, 5, \\
\alpha_p[k] &\overset{iid}{\sim} N(0, \sigma_p^2) \text{ for } k = 1, 2, \dots, 7 \quad (\text{all polls process model}).
\end{aligned} \tag{15}$$

In both models we assume each of the β s have independent $N(0, 1000)$ priors, and each of the σ^2 s have $IG(1, 1)$ priors. Including the random effects the single poll model contains 80 parameters while the all polls model contains 89 parameters, so both models are large enough to be challenging for PSO and other optimization algorithms. Writing down the log posteriors and deriving the Hessians is straightforward but tedious for these models, so we omit these steps, though note that in both the PSO and IMH algorithms the variances should be transformed to the log scale.

5 PSO results for finding posterior modes

[I HAVE RESULTS USING 10 RANDOM EFFECTS iid MODELS AND 5 IN full; WORTH FINDING SPACE FOR?]

We conduct a simulation study using R (R Development Core Team, 2008) to compare various PSO algorithms at finding the posterior mode in each of the example models from Sections 4 and 4.1. Based on the results of Appendix B, we limit the study to 7 PSO algorithms: standard PSO algorithm using parameter values suggested by Blum and Li (2008) and Clerc and Kennedy (2002) (PSO in Figures 1, 2, and [REFERENCE TO POLL BOX PLOT FIGURE]), the standard BBPSOxp-MC algorithm (BBPSO), DI-PSO with $\alpha = 0.2 \times n_{iter} = 200$ and $\beta = 1$ (DI-PSO), AT-PSO with $c = 0.1$ and either $R^* = 0.3$ or 0.5 (AT-PSO-0.3 and AT-PSO-0.5), and AT-BBPSOxp-MC with $df = 1$, $c = 0.1$, and either $R^* = 0.3$ or 0.5 (AT-BBPSO-0.3 and AT-BBPSO-0.5). Each algorithm was tried with one of two initializations. In the “BFGS” initialization, we first ran the BFGS algorithm using R’s `optim` function (R Development Core Team, 2008) until convergence using default settings to obtain an initial guess of the argmax, $\hat{\theta}$, then initialized the swarm with one particle at this initial guess and the rest uniformly in a length 2 hypercube centered on $\hat{\theta}$, i.e. $\theta_1 = \hat{\theta}$ and $\theta_{ij} = \hat{\theta}_j + U(-1, 1)$ for $i = 2, 3, \dots, 50$ and $j = 1, 2, \dots, n_{par}$ where n_{par} is the number of

parameters in the model and the $U(-1, 1)$ random variates are drawn independently. In the “no BFGS” initialization, each particle was initialized uniformly in a length 200 hypercube centered at zero, i.e. $\theta_{ij} \stackrel{iid}{\sim} U(-100, 100)$ for $i = 1, 2, \dots, 50$ and $j = 1, 2, \dots, n_{par}$. In addition, each algorithm was run using both the ring-1 and ring-3 neighborhood topologies, for a total of four combinations of initializations and neighborhoods, each for 20 replications of 1,000 iterations using 50 particles.

Figure 1 contains boxplots of the results of the simulations for the Poisson county population models of Section 4. Models with iid random effects had 30 random effects while models with fully correlated random effects had 15. Each box plot was created using the 10, 25, 50, 75, and 90 percentiles of the maximum value of the log posterior found in all 20 replications after 1,000 iterations. Figure 2 is similar, except for the lognormal models. Across all models we see that the ring-3 topology and BFGS initialization both improve each of the algorithms, especially in combination. In Appendix B the ring-3 neighborhood performed the best on our suite of test functions, and the same seems to be true here. The BFGS initialization is cheap, taking essentially no time to compute, yet seems to drastically improve the quality of every PSO algorithms. In Appendix B we speculated that the AT-PSO algorithms would benefit significantly from some sort of stage one optimization, and our results confirm this for all of the algorithms. Another lesson from these boxplots is that the AT-PSO and AT-BBPSO algorithms tend to do the best, especially when $R^* = 0.5$ in both cases.

[PARAGRAPH OR TWO AND PLOTS FOR THE ELECTION MODELS]

For our purposes, the PSO algorithms are useful only insofar as they allow us to construct IMH and IMHwG algorithms. So we conduct another simulation study to see when each algorithm gets close enough to the posterior mode for the IMH and IMHwG algorithms to have high acceptance rates. To this end we conduct another simulations study using the same 7 PSO algorithms from the previous study, except only using the ring-3 neighborhood

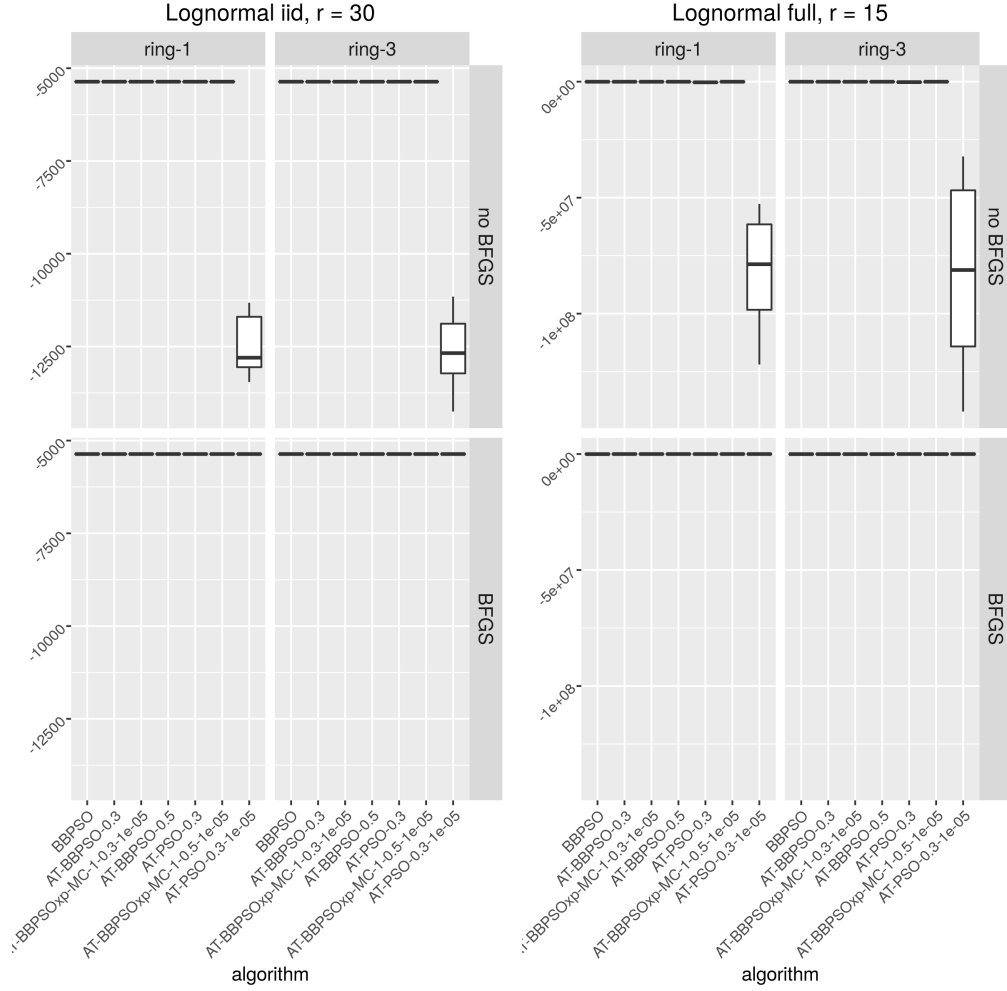


Figure 2: Boxplots of the maximum value of the log posterior found for lognormal models by random effect type, number of random effects, neighborhood topology, optimization initialization, and algorithm. Each box plot was created using the 10, 25, 50, 75, and 90 percentiles of the maximum found from 20 replications of 1,000 iterations with 50 particles for each factor combination.

algorithm we still run the BFGS initialization, so this serves as a control to see if running the PSO algorithm is necessary to get reasonable acceptance rates. Additionally, we run each IMH and IMHwG algorithm with different choices for the degrees of freedom parameter in the Laplace approximation.

IMH applies straightforwardly to each of our models, though see Appendix C for a detailed derivation of the Hessian for the county population models with a fully parameterized covariance matrix associated, i.e. from equations (10) and (11). To apply IMHwG in the county population models, we draw $(\boldsymbol{\beta}, \boldsymbol{\delta})$ in the Metropolis step and either σ^2 or \mathbf{L} in the conditionally conjugate step, depending on the model. The full conditional distribution of σ^2 in (6) and (7) is $IG(a_\sigma + r/2, b_\sigma + \boldsymbol{\delta}'\boldsymbol{\delta}/2)$. The full conditional distribution of $\boldsymbol{\Omega}$ in (8) and (9) is $W(r + 1, (\mathbf{E} + \boldsymbol{\delta}\boldsymbol{\delta}')^{-1})$. Then a draw from the full conditional distribution of \mathbf{L} can be obtained via the Bartlett decomposition as described at the end of Section 4. In the lognormal models the full conditional distribution of ϕ^2 is $IG(a_\phi + n/2, b_\phi + (\log \mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S}\boldsymbol{\delta})'(\log \mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S}\boldsymbol{\delta})/2)$.

In single poll model of Section 4.1, we draw $(\beta_0, \beta_f, \beta_b, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_{ae})$ in the Metropolis step, while in the all polls model we draw all of those parameters and additionally $\boldsymbol{\alpha}_p$ in the Metropolis step. Then for both models there are two additional Gibbs steps — one where each random effect variance is drawn, and one where $(\beta_{prev}, \boldsymbol{\alpha}_r)$ is drawn. These full conditionals are straightforward to derive, so we do not reproduce them here. Likewise, the Hessian is easy though tedious to derive, so we omit it as well.

[SIMULATION STUDY TO BE COMPLETED AND PUT HERE - WON'T TAKE VERY LONG, 1 DAY]

6 Comparison of MCMC techniques

Next we move to comparing various MCMC algorithms for both classes of models. The other MCMC algorithms we consider in this section are single move random walk Metropolis within Gibbs (RWwG), block random walk Metropolis within Gibbs (B-RWwG), Hamiltonian Monte Carlo as implemented by the Stan software (Carpenter et al., 2015) using the No U-turn Sampler (Homan and Gelman, 2014) and for the lognormal model for county pop-

ulations, a two step Gibbs sampler (Gibbs). The last algorithm is only considered for the lognormal model because only in that case are all of the full conditionals tractable. We use standard settings to fit each of the models in Stan, though minimal tweaking was required in some cases. In all of the Gibbs algorithms, the blocks are the same as in the IMHwG algorithms detailed in the previous section. Both the RWwG and B-RWwG algorithms are tuned during the burn-in using an adaptive method. Appendix D explains the details of the two random walk algorithms we use including how the adaptation was performed.

We compare these algorithms in Table [TO BE CONSTRUCTED] in terms of two measures: the minimum estimated effective sample size for all parameters in the model (n_{eff} , see Robert and Casella (2013, Section 12.3.5)), and time in seconds per n_{eff} . The effective sample size is the size of an iid sample which yields the same standard error for estimating the mean of some function of the parameters in our MCMC simulations, so we take the minimum estimated n_{eff} among all elements of the parameter vector and latent process.

[SIMULATIONS TO BE CONDUCTED - SHOULD TAKE A COUPLE DAYS ONCE THEY'RE RUNNING. FOR THE IMH AND IMHwG ALGORITHMS, THESE WILL ONLY USE ONE PSO ALGORITHM]

7 Discussion

[WILL NEED TO BE REWRITTEN IN LIGHT OF FORTHCOMING RESULTS]

In practice PSO-assisted Metropolis-Hastings algorithms are useful to the extent that the problem falls in something of a “sweet spot”. The model must be complex enough that the posterior mode cannot be found analytically and that a fully conjugate Gibbs sampler is not available. The combined parameter space and latent space must be large enough that traditional numerical methods fail to find the posterior mode, but small enough that the heuristic PSO algorithms have some chance without being too costly. Finally, both model

parameters and latent random variables must be approximately Gaussian in the posterior, though there is some leeway for model parameters using the IMHwG algorithm. Under these conditions, using PSO to obtain the Laplace approximation for use as a proposal for IMH or IMH within Gibbs improves on other more commonly used approaches for MCMC. The comparison to Stan is also relevant — even though our PSO assisted IMH algorithms beat Stan in terms of total computational time for a desired level of precision for some of the models we considered, the difference will be small enough in many settings for most users to stick with the more general algorithm. [SOMETHING ABOUT THE ADVANTAGE BEING IN BIG / TALL DATA SITUATIONS - STAN REQUIRES A LOT OF LOG POSTERIOR EVALUATIONS]

Our approach is fairly hands off, so long as the conditions listed above appear to be satisfied. In that case the standard PSO algorithm with default parameter values and the ring-1 neighborhood usually does a great job of finding the posterior mode quickly relative to the most competitive alternatives. Alternatives such as the adaptively tuned BBPSO variants we introduced only seem to do better when neither algorithm has converged on the mode yet or when the Laplace approximation is poor. So in practice we recommend using standard PSO first.

The AT-BBPSOxp-MC algorithms we introduced still require a larger swarm size or more iterations or both to do as well as standard PSO when standard PSO works well. With the same swarm size and number of iterations AT-BBPSOxp-MC does only slightly worse than standard PSO in terms of maximizing the objective function, but that slight difference often amounts to a large difference in the Metropolis acceptance rate for the resulting MCMC algorithms. When AT-BBPSOxp-MC does do better than standard PSO, the resulting MCMC algorithms do not always have high acceptance rates, at least not without running AT-BBPSOxp-MC with a high swarm size for many iterations. So in practice we recommend using standard PSO first. If the resulting MCMC algorithms have poor acceptance rates or

if it seems likely a priori that PSO might have trouble, then it can be useful to explore the options provided by AT-BBPSOxp-MC with the caveat that a large swarm size and number of iterations may be required.

References

- Andrieu, C. and Thoms, J. (2008). “A tutorial on adaptive MCMC.” *Statistics and Computing*, 18, 4, 343–373.
- Berliner, L. M. (1996). “Hierarchical Bayesian time series models.” In *Maximum entropy and Bayesian methods*, 15–22. Springer.
- Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). “Non-centered Parameterisations for Hierarchical Models and Data Augmentation.” In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, 307–326. London: Oxford University Press.
- Blum, C. and Li, X. (2008). “Swarm Intelligence in Optimization.” In *Swarm Intelligence: Introduction and Applications*, eds. C. Blum and D. Merkle. Springer.
- Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). “Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics.” *Annals of Applied Statistics*, 9, 4, 1761–1791.
- Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016). “Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Campos, M., Krohling, R. A., and Enriquez, I. (2014). “Bare bones particle swarm optimization with scale matrix adaptation.” *Cybernetics, IEEE Transactions on*, 44, 9, 1567–1578.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2015). “Stan: a probabilistic programming language.” *Journal of Statistical Software*.

- Chen, Z. and Dunson, D. B. (2003). “Random effects selection in linear mixed models.” *Biometrics*, 59, 4, 762–769.
- Clerc, M. (2010). *Particle swarm optimization*. John Wiley & Sons.
- Clerc, M. and Kennedy, J. (2002). “The particle swarm-explosion, stability, and convergence in a multidimensional complex space.” *Evolutionary Computation, IEEE Transactions on*, 6, 1, 58–73.
- Eberhart, R. C. and Shi, Y. (2000). “Comparing inertia weights and constriction factors in particle swarm optimization.” In *Evolutionary Computation, 2000. Proceedings of the 2000 Congress on*, vol. 1, 84–88. IEEE.
- Frühwirth-Schnatter, S. and Tüchler, R. (2008). “Bayesian parsimonious covariance estimation for hierarchical linear mixed models.” *Statistics and Computing*, 18, 1, 1–13.
- Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). “Efficient Parametrisations for Normal Linear Mixed Models.” *Biometrika*, 82, 3, 479–488.
- Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Gelman, A., Roberts, G., and Gilks, W. (1996). “Efficient Metropolis jumping rules.” *Bayesian statistics*, 5, 599–608, 42.
- Goldberg, D. E. and Holland, J. H. (1988). “Genetic algorithms and machine learning.” *Machine learning*, 3, 2, 95–99.
- Hastings, W. K. (1970). “Monte Carlo sampling methods using Markov chains and their applications.” *Biometrika*, 57, 1, 97–109.

- Homan, M. D. and Gelman, A. (2014). “The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.” *The Journal of Machine Learning Research*, 15, 1, 1593–1623.
- Hsieh, H.-I. and Lee, T.-S. (2010). “A modified algorithm of bare bones particle swarm optimization.” *International Journal of Computer Science Issues*, 7, 11.
- Hughes, J. and Haran, M. (2013). “Dimension reduction and alleviation of confounding for spatial generalized linear mixed models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 1, 139–159.
- Kennedy, J. (2003). “Bare bones particle swarms.” In *Swarm Intelligence Symposium, 2003. SIS’03. Proceedings of the 2003 IEEE*, 80–87. IEEE.
- Krohling, R., Mendel, E., et al. (2009). “Bare bones particle swarm optimization with Gaussian or Cauchy jumps.” In *Evolutionary Computation, 2009. CEC’09. IEEE Congress on*, 3285–3291. IEEE.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). “Equation of state calculations by fast computing machines.” *The journal of chemical physics*, 21, 6, 1087–1092.
- Neal, R. M. et al. (2011). “MCMC using Hamiltonian dynamics.” *Handbook of Markov Chain Monte Carlo*, 2, 113–162.
- Porter, A. T., Holan, S. H., and Wikle, C. K. (2015). “Bayesian semiparametric hierarchical empirical likelihood spatial models.” *Journal of Statistical Planning and Inference*, 165, 78–90.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

- Richer, T. J. and Blackwell, T. M. (2006). “The Lévy particle swarm.” In *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, 808–815. IEEE.
- Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. 2nd ed. Springer Science & Business Media.
- Roberts, G. O. and Sahu, S. K. (1997). “Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 2, 291–317.
- Rue, H., Martino, S., and Chopin, N. (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.” *Journal of the royal statistical society: Series B (statistical methodology)*, 71, 2, 319–392.
- Schervish, M. J. (1997). *Theory of statistics*. Springer Science & Business Media.
- Smith, W. and Hocking, R. (1972). “Algorithm AS 53: Wishart variate generator.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21, 3, 341–345.
- Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.
- Taylor, B. M. and Diggle, P. J. (2014). “INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes.” *Journal of Statistical Computation and Simulation*, 84, 10, 2266–2284.
- Tuppadung, Y. and Kurutach, W. (2011). “Comparing nonlinear inertia weights and constriction factors in particle swarm optimization.” *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15, 2, 65–70.
- Van Dyk, D. and Meng, X.-L. (2001). “The Art of Data Augmentation.” *Journal of Computational and Graphical Statistics*, 10, 1, 1–50.

Wikle, C. K. et al. (2003). “Hierarchical Models in Environmental Science.” *International Statistical Review*, 71, 2, 181–199.