Journal of Machine Learning Research 19 (2018) 1-34 Submitted 9/17; Revised 8/18; Published 10/18

Scaling up Data Augmentation MCMC via Calibration

Leo L. Duan LI.DUAN@UFL.EDU
Department of Statistics
University of Florida
Gainesville, FL

James E. Johndrow JOHNDROW@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA

David B. Dunson DUNSON@DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC

Editor: Ryan Adams

Abstract

There has been considerable interest in making Bayesian inference more scalable. In big data settings, most of the focus has been on reducing the computing time per iteration rather than reducing the number of iterations needed in Markov chain Monte Carlo (MCMC). This article considers data augmentation MCMC (DA-MCMC), a widely used technique. DA-MCMC samples tend to become highly autocorrelated in large samples, due to a mis-calibration problem in which conditional posterior distributions given augmented data are too concentrated. This makes it necessary to collect very long MCMC paths to obtain acceptably low MC error. To combat this inefficiency, we propose a family of calibrated data augmentation algorithms, which appropriately adjust the variance of conditional posterior distributions. A Metropolis-Hastings step is used to eliminate bias in the stationary distribution of the resulting sampler. Compared to existing alternatives, this approach can dramatically reduce MC error by reducing autocorrelation and increasing the effective number of DA-MCMC samples per unit of computing time. The approach is simple and applicable to a broad variety of existing data augmentation algorithms. We focus on three popular generalized linear models: probit, logistic and Poisson log-linear. Dramatic gains in computational efficiency are shown in applications.

Keywords: Bayesian Probit, Biased subsampling, Big n, Data augmentation, Log-linear model, Logistic regression, Maximal correlation, Polya-Gamma

1. Introduction

With the deluge of data in many modern application areas, there is pressing need for scalable computational algorithms for inference from such data, including uncertainty quantification (UQ). Somewhat surprisingly, even as the volume of data increases, uncertainty often remains sizable. Examples in which this phenomenon occurs include financial fraud detection (Ngai et al., 2011), disease mapping (Wakefield, 2007) and online click-through tracking (Wang et al., 2010). Bayesian approaches provide a useful paradigm for quantifying uncertainty in these and other settings.

The standard approach to Bayesian posterior computation is Markov chain Monte Carlo (MCMC) and related sampling algorithms. However, conventional MCMC algorithms often scale poorly in problem size and complexity. Due to its sequential nature, the computational cost of MCMC is the product of two factors: the evaluation cost at each sampling iteration and the total number of iterations needed to obtain an acceptably low Monte Carlo (MC) error. While a substantial literature has developed focusing on decreasing computational cost per iteration in "big data" (large sample) settings (Minsker et al. (2017); Maclaurin and Adams (2015); Srivastava et al. (2015); Conrad et al. (2016) among others), there has been less focus on reducing the required number of MCMC iterations. This contrasts with a historical focus in the statistics and probability literature on improving mixing and convergence of MCMC in more traditional small to moderate sample size problems, and suggests the opportunity for improved performance in big data settings through a renewed focus on improving mixing.

A major concern in applying MCMC algorithms in big data problems is that the level of autocorrelation in the MCMC path may increase with the size of the data. Markov chains with high autocorrelation have low *effective sample size (ESS)* per unit computational time, which we refer to informally as the *slow mixing* problem. The ESS compares the asymptotic variance of the MCMC time averaging estimate to a gold standard Monte Carlo algorithm that collects independent samples. For example, if the number of effective samples in 1,000 MCMC iterations is only 10, then the MCMC algorithm will need to be run 100 times as long as an ordinary MC algorithm to obtain the same MC error for time averages. Such a scenario is not unusual in big data problems, leading MCMC algorithms to face a *double burden*, with the time per iteration increasing and it becoming necessary to collect more iterations as sample size increases.

This double burden has led many members of the machine learning community to abandon MCMC in favor of more easily scalable alternatives, such as variational approximations. Unfortunately, these approaches typically lack theoretical guarantees and often badly underestimate posterior uncertainty. Hence, there has been substantial interest in recent years in designing scalable MCMC algorithms. The focus of this paper is a popular and broad class of Data Augmentation (DA)-MCMC algorithms. DA-MCMC algorithms are used routinely in many classes of models, with the algorithms of Albert and Chib (1993) for probit models and Polson et al. (2013) for logistic models being particularly popular. Our focus is on improving the performance of such algorithms in big data settings in which poor scalability occurs both because of high cost per iteration and deterioration of mixing as sample size increases. We focus here on the slow mixing problem.

Johndrow et al. (2018) demonstrate that popular DA-MCMC algorithms have small effective sample sizes in large data settings involving imbalanced data. For example, data may be binary with a high proportion of zeros. A key insight is that this problem results from a discrepancy in the rates at which Gibbs step sizes and the width of the high-probability region of the posterior converge to zero as n increases. In particular, the conditional posterior given the augmented data may simply be too concentrated relative to the marginal posterior, with this problem amplified as the data sample size increases. There is a rich literature on methods for accelerating mixing in DA-MCMC algorithms using tricks ranging from reparameterization to parameter-expansion (Liu and Wu, 1999; Meng and Van Dyk, 1999; Papaspiliopoulos et al., 2007). However, we find that such approaches fail to address the miscalibration problem and have no impact on the worsening mixing rate with increasing data sample size n.

This article proposes a general new class of algorithms that addresses the miscalibration of step sizes in DA. The idea underlying these *calibrated* DA (CDA) algorithms is to introduce auxiliary parameters that change the variance of full conditional distributions for one or more parameters. These auxiliary parameters can adapt with the data sample size n to correct the typical step sizes of the CDA algorithm to match the rate at which the high probability region of the posterior contracts as n increases. In general, the invariant measure of CDA-MCMC – which typically does exist and is unique – differs from the posterior of interest. Thus, CDA-MCMC is a computationally more efficient perturbation of the original Markov chain, and the bias can be eliminated using Metropolis-Hastings. Compared to other adaptive Metropolis-Hastings algorithms, which often require carefully chosen multivariate proposals and complicated adaptation with multiple chains (Tran et al., 2016), CDA-MCMC only requires a simple modification to Gibbs sampling steps to generate proposals. We show the auxiliary parameters can be efficiently adapted for each type of data augmentation via minimizing the difference between Fisher information of conditional and marginal distributions.

2. Calibrated Data Augmentation

Data augmentation Gibbs samplers alternate between sampling latent data z from their conditional posterior distribution given model parameters θ and observed data y, and sampling parameters θ given z and y; either of these steps can be further broken down into a series of full conditional sampling steps but we focus for simplicity on algorithms of the form:

$$\begin{aligned} z \mid \theta, y &\sim \pi(z; \theta, y) \\ \theta \mid z, y &\sim f(\theta; z, y), \end{aligned} \tag{1}$$

where f belongs to a location-scale family, such as the Gaussian. Popular data augmentation algorithms are designed so that both of these sampling steps can be conducted easily and efficiently; e.g., sampling the latent data for each subject independently and then drawing θ simultaneously (or at least in blocks) from a multivariate Gaussian or other standard distribution. This effectively avoids the need for tuning, which is a major issue for Metropolis-Hastings algorithms, particularly when θ is high-dimensional. Data augmentation algorithms are particularly common for generalized linear models (GLMs), with

$\mathbb{E}(y_i \mid x_i, \theta) = g^{-1}(x_i\theta)$ and a conditionally Gaussian prior distribution chosen for θ. We focus in particular on Poisson log-linear, binomial logistic, and binomial probit as motivating examples.

Consider a Markov kernel $K((\theta, z); \cdot)$ with invariant measure Π and update rule of the form (1), and a Markov chain (θ_t, z_t) on a state space $\Theta \times \mathcal{Z}$ evolving according to K. We will abuse notation in writing $\Pi(d\theta) = \int_{z\in\mathcal{Z}} \Pi(d\theta, dz)$. The lag-1 autocorrelation for a function $h : \Theta \to \mathbb{R}$ at stationarity can be expressed as the Bayesian fraction of missing information (Papaspiliopoulos et al. (2007), Rubin (2004), Liu (1994b))

$$\gamma_g = 1 - \frac{\mathbb{E}[\text{var}(h(\theta) \mid z)]}{\text{var}(h(\theta))}, \tag{2}$$

where the integrals in the numerator are with respect to $\Pi(d\theta, dz)$ and in the denominator with respect to $\Pi(d\theta)$. Let

$$L_2(\Pi) = \left\{ h : \Theta \to \mathbb{R}, \int_{\theta\in\Theta} \{h(\theta)\}^2 \Pi(d\theta) < \infty \right\}$$

be the set of real-valued, Π square-integrable functions. The *maximal autocorrelation*

$$\gamma = \sup_{h\in L_2(\Pi)} \gamma_h = 1 - \inf_{h\in L_2(\Pi)} \frac{\mathbb{E}[\text{var}(h(\theta) \mid z)]}{\text{var}(h(\theta))}$$

is equal to the geometric convergence rate of the data augmentation Gibbs sampler (Liu (1994b)). For $h(\theta) = \theta_j$ a coordinate projection, the numerator of the last term of (2) is, informally, the average squared step size for the augmentation algorithm at stationarity in direction j, while the denominator is the squared width of the bulk of the posterior in direction j. Consequently, γ will be close to 1 whenever the average step size at stationarity is small relative to the width of the bulk of the posterior.

The purpose of CDA is to introduce additional parameters that allow us to control the step size relative to the posterior width – roughly speaking, the ratio in (2) – with greater flexibility than reparametrization or parameter expansion. The flexibility gains are achieved by allowing the invariant measure to change as a result of the introduced parameters. The additional parameters, which we denote (r, b), correspond to a collection of reparametrizations, each of which defines a proper (but distinct) likelihood $L_{r,b}(\theta; y)$, and for which there exists a Gibbs update rule of the form (1). In general, r is a vector of scale parameters that are tuned to increase $\mathbb{E}[\text{var}(h(\theta) \mid z)]\{\text{var}(h(\theta))\}^{-1}$ – usually for coordinate projections $h(\theta) = \theta_j$ – although the exact way in which they enter the likelihood and corresponding Gibbs update depend on the application; b are location parameters that shift the high posterior region of $L_{r,b}(\theta; y)$ to better approximate $L(\theta; y)$. The reparametrization also has the property that $L_{1,0}(\theta; y) = L(\theta; y)$, the original likelihood. The resulting Gibbs sampler, which we refer to as CDA Gibbs, has θ-marginal invariant measure $\Pi_{r,b}(\theta; y) \propto L_{r,b}(\theta; y)\Pi^0(\theta)$, where $\Pi^0(\theta)$ is the prior. Ultimately, we are interested in $\Pi_{1,0}(\theta; y)$, so we use CDA Gibbs as an efficient proposal for Metropolis-Hastings. That is, we propose θ^* from $q(\theta^*; \theta)$ with

$$q(\theta^*; \theta) = \int_{z\in\mathcal{Z}} \pi_{r,b}(z; \theta, y) f_{r,b}(\theta^*; z, y) dz, \tag{3}$$

where $\pi_{r,b}$ and $f_{r,b}$ denote the conditional densities of z and θ in the Gibbs sampler with invariant measure $\Pi_{r,b}$. By tuning (r, b) during an adaptation phase to reduce the autocorrelations and increase the Metropolis-Hastings acceptance rate, we can obtain a computationally efficient algorithm. Tuning is facilitated by the fact that the MH acceptance ratios using this proposal kernel have a convenient form, which is a nice feature of using Gibbs to generate MH proposals.

Remark 1 *The CDA MH acceptance ratio is given by*

$$\begin{aligned}\alpha(\theta, \theta^*) &= \min\left\{1, \frac{L(\theta^*; y)\Pi^0(\theta^*)q(\theta; \theta^*)}{L(\theta; y)\Pi^0(\theta)q(\theta^*; \theta)}\right\} \\ &= \min\left\{1, \frac{L(\theta^*; y)L_{r,b}(\theta; y)}{L(\theta; y)L_{r,b}(\theta^*; y)}\right\}. \end{aligned} \tag{4}$$

A general strategy for tuning is given in Section 4.

We give a basic convergence guarantee that holds for CDA MH under weak assumptions on $L_{r,b}$, which is based on Roberts and Smith (1994). Basically, one needs $\Pi(\cdot) \ll \Pi_{r,b}(\cdot)$ for all r, b, where for two probability measures μ, ν, $\mu(\cdot) \ll \nu(\cdot)$ means μ is absolutely continuous with respect to ν.

Remark 2 (Ergodicity) *Assume that $\Pi(d\theta)$ and $\Pi_{r,b}(d\theta)$ have densities with respect to Lebesgue measure on $\mathbb{R}^p$, and that $K_{r,b}((\theta, z); (\theta', z')) > 0 \,\forall\, ((\theta, z), (\theta', z')) \in (\Theta \times \mathcal{Z}) \times (\Theta \times \mathcal{Z})$. Then,*

- *For fixed r, b, CDA Gibbs is ergodic with invariant measure $\Pi_{r,b}(d\theta, dz)$.*
- *A Metropolis-Hastings algorithm with proposal kernel $q_{r,b}(\theta'; \theta)$ as defined in (3) with fixed r, b is ergodic with invariant measure $\Pi(d\theta)$.*

Proofs are located in the Appendix.

2.1. Initial Example: Probit with Intercept Only

To illustrate the CDA algorithm, we first present a toy example of the probit regression with intercept only.

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = \Phi(\theta) \quad i = 1, \ldots, n$$

and improper prior $\Pi^0(\theta) \propto 1$. The data augmentation algorithm (Tanner and Wong, 1987; Albert and Chib, 1993) is based on the following integral

$$L(y_i; \theta) = \begin{cases} \int_0^\infty f(z_i; \theta, 1)dz_i & \text{if } y_i = 1 \\ \int_{-\infty}^0 f(z_i; \theta, 1)dz_i & \text{if } y_i = 0 \end{cases} \qquad i = 1, \ldots, n$$

where $f(z; \mu, \sigma^2)$ is the density for normal distribution $\text{No}(\mu, \sigma^2)$.

This leads to the update rule

$$z_i \mid \theta, y_i \sim \begin{cases} \text{No}_{[0,\infty)}(\theta, 1) & \text{if } y_i = 1 \\ \text{No}_{(-\infty,0]}(\theta, 1) & \text{if } y_i = 0 \end{cases} \qquad i = 1, \ldots, n$$

$$\theta \mid z, y \sim \text{No}\left(n^{-1}\sum_i z_i, n^{-1}\right),$$

where the subscript in $\text{No}_{[a,b]}(\mu, \sigma^2)$ denotes the truncation to the interval $[a, b]$. Johndrow et al. (2018) show that when $\sum_i y_i = 1$, $\text{var}(\theta_t \mid \theta_{t-1})$ is approximately $n^{-1}\log n$, while the width of the high probability region of the posterior is order $(\log n)^{-1}$, leading to slow mixing.

We introduce a scale parameter r_i in the update for z_i, and adjust the conditional mean by a location parameter b_i. This is equivalent to changing the scale of $z_i \mid \theta, y_i$ from 1 to r_i and the mean from θ to $\theta + b_i$. These adjustments yield

$$\begin{aligned}\Pr(y_i = 1|\theta, r_i, b_i) &= \int_0^\infty \frac{1}{\sqrt{2\pi r_i}} \exp\left(-\frac{(z_i - \theta - b_i)^2}{2r_i^2}\right) dz_i \\ &= \Phi\left(\frac{\theta + b_i}{\sqrt{r_i}}\right). \end{aligned} \tag{5}$$

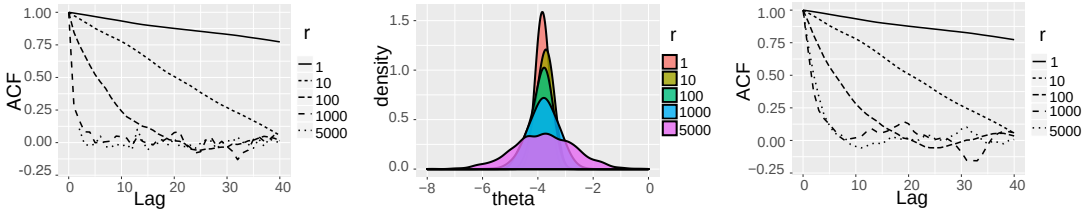
In this simple example, we set the tuning parameters to be all the same: $r_i = r_0$ and $b_i = b_0$ over $i = 1, \ldots, n$, with r_0 and b_0 two scalars. This leads to the modified data augmentation algorithm

$$\begin{aligned} z_i \mid \theta, y_i &\sim \begin{cases} \text{No}_{[0,\infty)}(\theta + b_0, r_0) & \text{if } y_i = 1 \\ \text{No}_{(-\infty,0]}(\theta + b_0, r_0) & \text{if } y_i = 0 \end{cases} \qquad i = 1, \ldots, n \\ \theta \mid z, y &\sim \text{No}\left(n^{-1}\sum_i (z_i - b_0), n^{-1} r_0\right). \end{aligned} \tag{6}$$

To achieve step sizes consistent with the width of the high posterior probability region, we need $n^{-1}r_0 \approx (\log n)^{-1}$, so $r_0 \approx n/\log n$. To preserve the original target, we use (6) to generate an MH proposal θ^*. By Remark 1, the MH acceptance probability is given by (4) with $L_{r,b}(\theta; y_i) = \Phi\big((\theta + b_0){r_0}^{-1/2}\big)^{y_i}\Phi\big(-(\theta + b_0){r_0}^{-1/2}\big)^{(1-y_i)}$ and $L(\theta; y_i) = L_{1,0}(\theta; y_i)$. Setting $r_0 = 1$ and $b_0 = 0$ leads to acceptance rate of 1, which corresponds to the original Gibbs sampler.

To illustrate, we consider $\sum_i y_i = 1$ and $n = 10^4$. Letting $r_0 = n/\log n$, we then choose b_0 to increase the acceptance rate in the MH step. In this simple example, it is easy to compute a "good" value of b_0, since $b_0 = -3.7(\sqrt{r} - 1)$ results in $\Pr(y_i = 1) = \Phi(-3.7) = n^{-1}\sum_i y_i \approx 10^{-4}$ in the proposal distribution, centering the proposals near the MLE for p_i.

We perform computation for these data with different values of r_0 ranging from $r_0 = 1$ to $r_0 = 5,000$, with $r_0 = 1,000 \approx n/\log n$ corresponding to the theoretically optimal value. Figure 1(a) plots autocorrelation functions (ACFs) for these different samplers without MH adjustment. Autocorrelation is very high even at lag 40 for $r_0 = 1$, while increasing r_0 leads to dramatic improvements in mixing. There are no further gains in increasing r_0 from the theoretically optimal value of $r_0 = 1,000$ to $r_0 = 5,000$. Figure 1(b) shows kernel-smoothed density estimates of the posterior of θ without MH adjustment for different values of r_0 and based on long chains to minimize the impact of Monte Carlo error; the posteriors are all centered on the same values but with variance increasing somewhat with r_0. With MH adjustment such differences are removed; the MH step has acceptance probability close to one for $r_0 = 10$ and $r_0 = 100$, about 0.6 for $r_0 = 1,000$, and 0.2 for $r_0 = 5,000$.



(a) ACF for CDA without MH adjustment. (b) Posterior density estimates without MH adjustment. (c) ACF for CDA with MH adjustment

Figure 1: Autocorrelation functions (ACFs) and kernel-smoothed density estimates for different CDA samplers in intercept-only probit model.

We also study a common hierarchical Gaussian example in appendix C.

3. Specific Algorithms

In this section, we describe CDA algorithms for general probit and logistic regression.

3.1. Probit Regression

Consider the probit regression:

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = \Phi(x_i\theta) \quad i = 1, \ldots, n$$

with improper prior $\Pi^0(\theta) \propto 1$. The data augmentation sampler (Tanner and Wong, 1987; Albert and Chib, 1993) has the update rule

$$z_i \mid \theta, x_i, y_i \sim \begin{cases} \text{No}_{[0,\infty)}(x_i\theta, 1) & \text{if } y_i = 1 \\ \text{No}_{(-\infty,0]}(x_i\theta, 1) & \text{if } y_i = 0 \end{cases} \quad i = 1, \ldots, n$$
$$\theta \mid z, x, y \sim \text{No}((X'X)^{-1}X'z, (X'X)^{-1}).$$

Liu and Wu (1999) and Meng and Van Dyk (1999), among others, previously studied this algorithm and proposed to rescale θ through parameter expansion. However, this modification does not impact the conditional variance of θ and thus does not directly increase typical step sizes.

Our approach is fundamentally different, since we directly adjust the conditional variance. Similar to the intercept only model, we modify $\text{var}(\theta|z)$ by changing the scale of each z_i. This yields the update rule

$$z_i \mid \theta, x_i, y_i \sim \begin{cases} \text{No}_{[0,\infty)}(x_i\theta + b_i, r_i) & \text{if } y_i = 1 \\ \text{No}_{(-\infty,0]}(x_i\theta + b_i, r_i) & \text{if } y_i = 0 \end{cases} \quad i = 1, \ldots, n \tag{7}$$
$$\theta \mid z, X \sim \text{No}((X'R^{-1}X)^{-1}X'R^{-1}(z-b), (X'R^{-1}X)^{-1}),$$

where $R = \text{diag}(r_1, \ldots, r_n)$, $b = (b_1, \ldots, b_n)'$. Under the Bernoulli likelihood, we have

$$\Pr(y_i = 1|\theta, x_i, r_i, b_i) = \int_0^\infty \frac{1}{\sqrt{2\pi r_i}} \exp\left(-\frac{(z_i - x_i\theta - b_i)^2}{2r_i}\right) dz_i$$

$$= \Phi\left(\frac{x_i\theta + b_i}{\sqrt{r_i}}\right). \tag{8}$$

For fixed $r = (r_1, \ldots, r_n)$ and $b = (b_1, \ldots, b_n)$, (8) defines a proper Bernoulli likelihood for y_i conditional on parameters, and therefore the transition kernel $Q_{r,b}((\theta, z); \cdot)$ defined by the Gibbs update rule in (7) would have a unique invariant measure for fixed r, b, which we denote $\Pi_{r,b}(\theta, z \mid y)$.

For insight into the relationship between r and step size, consider the θ-marginal autocovariance in a Gibbs sampler evolving according to $K_{r,b}$

$$\begin{aligned}
&\mathrm{cov}_{r,b}(\theta_t \mid \theta_{t-1}, X, z, y) \\
=&(X'R^{-1}X)^{-1} + (X'R^{-1}X)^{-1}X'R^{-1}\mathrm{cov}(z-b|R)R^{-1}X(X'R^{-1}X)^{-1} \\
\geq&(X'R^{-1}X)^{-1}.
\end{aligned}$$

In the special case where $r_i = r_0$ for all i, we have

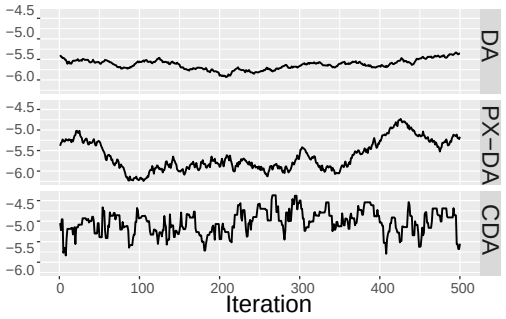
$$\mathrm{cov}_{r,b}(\theta_t \mid \theta_{t-1}, X, z, y) \geq r_0(X'X)^{-1},$$

so that all of the conditional variances are increased by at least a factor of r_0. This holds uniformly over the entire state space, so it follows that

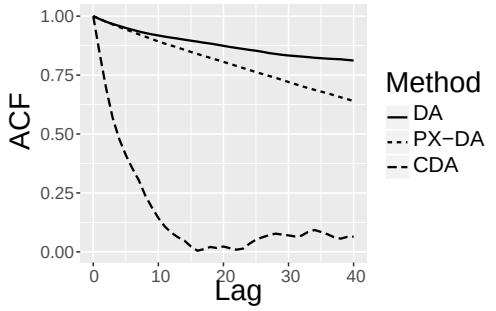
$$\mathbb{E}_{\Pi_{r,b}}[\mathrm{var}(\theta_j \mid z)] \geq r_0\mathbb{E}_{\Pi}[\mathrm{var}(\theta_j \mid z)].$$

The key to CDA is to choose r, b to make $\mathbb{E}_{\Pi_{r,b}}[\mathrm{var}(\theta_j \mid z)]$ close to $\mathrm{var}_{\Pi_{r,b}}(\theta_j \mid z)$, while additionally maximizing the MH acceptance probability. We defer the details of tuning algorithm for r, b to the next section.

For illustration, we consider a simulation study for probit regression with an intercept and two predictors $x_{i,1}, x_{i,2} \sim \mathrm{No}(1, 1)$, with $\theta = (-5, 1, -1)'$, generating $\sum_i y_i \approx 20$ among $n = 10,000$. The Albert and Chib (1993) DA algorithm mixes slowly (Figure 2(a) and 2(b)). We also show the results of the parameter expansion algorithm (PX-DA) proposed by Liu and Wu (1999). PX-DA only mildly reduces the correlation, as it does not solve the small step size problem. After tuning, CDA reaches a satisfactory acceptance rate of 0.6 and leads to dramatically better mixing.



(a) Traceplot for the original DA, parameter expanded DA and CDA algorithms.



(b) ACF for original DA, parameter expanded DA and CDA algorithms.

Figure 2: Panel (a) demonstrates in traceplot and panel (b) in autocorrelation the substantial improvement in CDA by correcting the variance mis-match in probit regression with rare event data, compared with the original (Albert and Chib, 1993) and parameter-expanded methods (Liu and Wu, 1999).

3.2. Logistic Regression

In the second example, we focus on the logistic regression model with

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = \frac{\exp(x_i\theta)}{1+\exp(x_i\theta)} \quad i = 1, \ldots, n \tag{9}$$

and improper prior $\Pi^0(\theta) \propto 1$. For this model, Polson et al. (2013) proposed Polya-Gamma data augmentation:

$$\begin{aligned} z_i &\sim \text{PG}(1, |x_i\theta|) \quad i = 1, \ldots, n, \\ \theta &\sim \text{No}\left((X'ZX)^{-1}X'(y-0.5), (X'ZX)^{-1}\right), \end{aligned}$$

where $Z = \text{diag}(z_1, \ldots, z_n)$. This algorithm relies on expressing the logistic regression likelihood as

$$L(x_i\theta; y_i) = \int_0^\infty \exp\{x_i\theta(y_i - 1/2)\} \exp\left\{ -\frac{z_i(x_i\theta)^2}{2}\right\} \text{PG}(z_i \mid 1, 0) dz_i,$$

where $\text{PG}(a_1, a_2)$ denotes the density of the Polya-Gamma distribution with parameters a_1, a_2, with $\mathbb{E}z_i = a_1/(2a_2)\tanh(a_2/2)$.

Since our goal is to increase the conditional variance $(X'ZX)^{-1}$, we can achieve this stochastically by reducing the mean $\mathbb{E}z_i$. We replace $\text{PG}(z_i \mid 1, 0)$ with $\text{PG}(z_i \mid r_i, 0)$ in the step for updating the latent data. Adding the location term b_i to the linear predictor $\eta_i = x_i\theta$ leads to

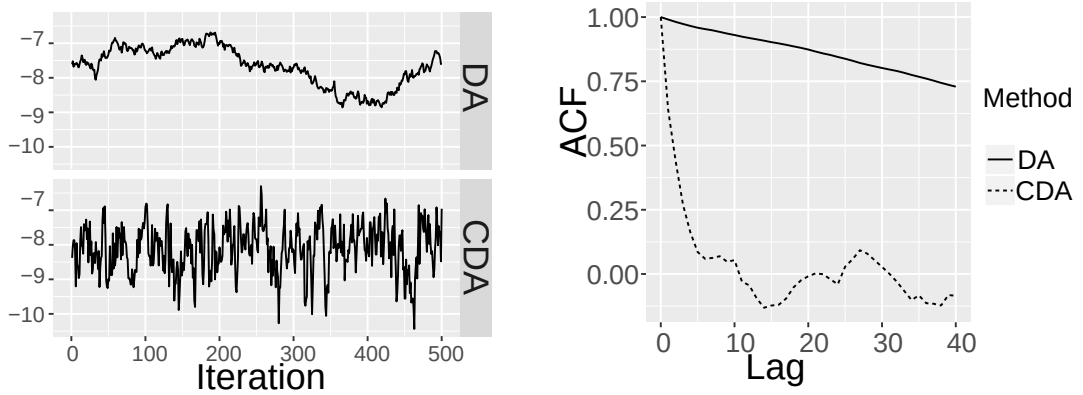
$$\begin{aligned} L_{r,b}(x_i\theta; y_i) &= \int_0^\infty \exp\{(x_i\theta + b_i)(y_i - r_i/2)\} \exp\left\{ -\frac{z_i(x_i\theta + b_i)^2}{2}\right\} \text{PG}(z_i \mid r_i, 0) dz_i \\ &= \frac{\exp\{(x_i\theta + b_i)y_i\}}{\{1+\exp(x_i\theta + b_i)\}^{r_i}}, \end{aligned} \tag{10}$$

and the update rule for the CDA Gibbs sampler is then

$$z_i \sim \mathrm{PG}(r_i, |x_i\theta + b_i|) \quad i = 1, \ldots, n,$$
$$\theta' \sim \mathrm{No}\left((X'ZX)^{-1}X'(y - r/2 - Zb), (X'ZX)^{-1}\right),$$

where $r = (r_1, \ldots, r_n)'$ and $b = (b_1, \ldots, b_n)'$. We again defer the tuning details for r and b to the next section.

For illustration, we use a two-parameter intercept-slope model with $x_{i1} \overset{iid}{\sim} \mathrm{No}(0, 1)$ and $\theta = (-8, 1)'$. With $n = 10^5$, we obtain rare outcome data with $\sum y_i \approx 50$. In CDA, after tuning, it reaches an acceptance rate of 0.8. Shown in Figure 3, DA mixes slowly, exhibiting strong autocorrelation even at lag 40, while CDA has dramatically better mixing.



(a) Traceplots for DA and CDA.

(b) ACF for DA and CDA.

Figure 3: Panel (a) demonstrates in traceplot and panel (b) in autocorrelation the substantial improvement of CDA in logistic regression with rare event data, compared with the original DA (Polson et al., 2013).

4. Automatic Tuning of Calibration Parameters

As illustrated in the previous subsection, efficiency of CDA is dependent on good choices of the calibration parameters r and b. We propose a simple and efficient algorithm for calculating "good" values of these parameters utilizing the Fisher information and empirical MH acceptance rate. Although our choice of calibration parameters relies on large sample approximations, we find that this calibration approach also works well for modest sample size.

Our goal is to adjust the conditional variance under calibration of (r, b) to approximately match the marginal variance under the exact target distribution, while maintaining a reasonable MH acceptance rate.

To approximate the marginal variance, we use the inverse of the observed Fisher information (Efron and Hinkley, 1978):

$$\begin{aligned}&\mathrm{var}(\theta \mid y) \approx \mathcal{I}^{-1}(\hat{\theta})\\ &\left(\mathcal{I}(\hat{\theta})\right)_{i,j} = \left(\frac{\partial}{\partial \theta_i} \log L(\theta; y)\right)\left(\frac{\partial}{\partial \theta_j} \log L(\theta; y)\right)\Big|_{\theta=\hat{\theta}}\end{aligned}$$

for $i = 1, \ldots, p$, $j = 1, \ldots, p$, where $\hat{\theta}$ is the Maximum a Posteriori (MAP) estimate of θ.

Recall that the CDA proposal has density

$$q(\theta^*; \theta) = \int f_{r,b}(\theta^*; z, y)\pi_{r,b}(z; \theta, y)dz,$$

and the conditional variance has lower bound $\mathrm{var}(\theta^* \mid \theta) \geq \mathbb{E}_{z|\theta}\mathrm{var}(\theta^* \mid z)$. We use the inverse of the observed Fisher information to approximate $\mathrm{var}(\theta^* \mid z)$ via

$$\begin{aligned}&\mathbb{E}_{z|\theta}\mathrm{var}(\theta^* \mid z) \approx \mathbb{E}_{z|\theta}\mathcal{I}^{-1}(\hat{\theta}; r, b, z) \approx \mathcal{I}^{-1}(\hat{\theta}; r, b, \tilde{z}(\hat{\theta}))\\ &\left(\mathcal{I}(\hat{\theta}; r, b, z)\right)_{i,j} = \left(\frac{\partial}{\partial \theta_i^*} \log f_{r,b}(\theta^*; y, z)\right)\left(\frac{\partial}{\partial \theta_j^*} \log f_{r,b}(\theta^*; y, z)\right)\Big|_{\theta^*=\hat{\theta}}.\end{aligned}$$

Since $\mathbb{E}_{z|\theta}\mathcal{I}^{-1}(\hat{\theta}; r, b, z)$ is often intractable or cumbersome to compute, we instead use the second approximation, the Fisher information evaluated at $\tilde{z}(\hat{\theta})$, the conditional mean or mode of $\pi_{r,b}(z; \hat{\theta}, y)$. The choice between mean and mode depends on which has a closed-form expression.

One can now adjust r, b to reduce the distance

$$d_1(r, b) = \mathrm{Dist}\left[\mathcal{I}^{-1}(\hat{\theta}), \mathcal{I}^{-1}(\hat{\theta}; r, b, \tilde{z}(\hat{\theta}))\right], \tag{11}$$

where $\mathrm{Dist}(M_1, M_2)$ is a distance between two matrices, such as $\|M_1 - M_2\|_F$ or $\|M_1^{-1} - M_2^{-1}\|_F$, the Frobenius norm of the difference.

However, the increase in proposal variance only results in an increase in the variance of the Metropolis-Hastings transition density so long as the acceptance probability is not substantially depressed (relative to the DA Gibbs sampler, which has acceptance probability one). Therefore, one also needs to adjust (r, b) both to optimize the acceptance rate $\alpha(\theta, \theta^*)$ and the proposal variance. Considering the average acceptance rate (on the negative-log scale), with expectation over proposal density $q(\theta^*; \theta)$ and posterior $\pi(\theta \mid y)$

$$\begin{aligned}&\mathbb{E}_{\theta|y}\mathbb{E}_{\theta^*|\theta}\big[-\log \alpha(\theta, \theta^*)\big]\\ =&\mathbb{E}_{\theta|y}\mathbb{E}_{\theta^*|z}\mathbb{E}_{z|\theta} \max\left[-\log \frac{L(\theta^*; y)}{L_{r,b}(\theta^*; y)} + \log \frac{L(\theta; y)}{L_{r,b}(\theta; y)}, 0\right].\end{aligned}$$

To provide tractable computation, we again use the functions evaluated at the conditional mean or mode to approximate the three expectations. This yields

$$d_2(r, b) = \max\left[-\log \frac{L(\tilde{\theta}^*(\tilde{z}(\hat{\theta})); y)}{L_{r,b}(\tilde{\theta}^*(\tilde{z}(\hat{\theta})); y)} + \log \frac{L(\hat{\theta}; y)}{L_{r,b}(\hat{\theta}; y)}, 0\right], \tag{12}$$

with $\hat{\theta}$ the MAP of θ, $\tilde{z}(\theta)$ the mean or mode of $\pi_{r,b}(z;\theta,y)$, $\tilde{\theta}^*$ the mean or mode of $f_{r,b}(\theta^*;z,y)$.

Combining (11) and (12), this yields the optimal tuning parameters under those two criteria:

$$(\hat{r},\hat{b}) = \min_{r,b} \big[d_1(r,b) + \lambda d_2(r,b)\big]. \tag{13}$$

The optional parameter $\lambda > 0$ allows for differential weighting of the acceptance rate and variance, although the default $\lambda = 1$ works well for all of our applications.

To facilitate automatic tuning in generic cases, we exploit the automatic differentiation and optimization software, TensorFlow, to compute the Fisher information and optimize for $(\hat{r},\hat{b})$. One only needs to provide the densities $L_{r,b}(\theta;y)$, $f_{r,b}(\theta^*;z,y)$, $\pi_{r,b}(z;\theta,y)$ and two conditional estimators $\tilde{z}(\theta)$ and $\tilde{\theta}^*(z)$.

We now provide the tuning details for probit and logistic regression. The likelihood and update densities $L_{r,b}(\theta;y)$, $f_{r,b}(\theta^*;z,y)$, $\pi_{r,b}(z;\theta,y)$ are already given, we present the conditional estimators. For probit regression, the two conditional modes for $\pi_{r,b}(z;\theta,y)$, $\tilde{\theta}^*$ and $f_{r,b}(\theta^*;z,y)$ are available in closed form, viz

$$\begin{aligned}
\tilde{z}_i(\theta) =& \begin{cases} x_i\theta + b_i & \text{if } (y_i - 0.5)(x_i\theta + b_i) > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1,\dots,n, \\
\tilde{\theta}^*(z) =& (X'R^{-1}X)^{-1}X'R^{-1}(z-b).
\end{aligned}$$

For logistic regression, the conditional means for $\pi_{r,b}(z;\theta,y)$, $\tilde{\theta}^*$ and $f_{r,b}(\theta^*;z,y)$ all have closed-form expressions given by

$$\begin{aligned}
\tilde{z}_i(\theta) =& \frac{r_i}{2|x_i\theta_i + b_i|}\tanh(\frac{|x_i\theta_i + b_i|}{2}) \quad i = 1,\dots,n, \\
\tilde{\theta}^*(z) =& (X'ZX)^{-1}X'(y - r/2 - Zb).
\end{aligned}$$

5. Geometric convergence rates for CDA-MH and CDA-Gibbs

Although Remark 2 gives a basic guarantee of convergence of the usual time-averaging estimators commonly used in MCMC, the goal of CDA-MH is to improve upon the convergence rate of the usual DA Gibbs. Motivation for CDA is provided by the results of Johndrow et al. (2018), which studied the special case of intercept-only logistic and probit regression when $y = 1$ and $n \to \infty$ so that the data grow increasingly imbalanced as the sample size increases. Johndrow et al. (2018) showed that in this setting, the spectral gap of DA converges to zero at least as fast as $n^{-1/2}(\log n)^k$ for $k \leq 5.5$, while random-walk Metropolis has spectral gap order $(\log n)^3$ or larger. This suggests the superiority of Metropolis algorithms in the large sample imbalanced data setting. However, to implement Metropolis effectively with moderate to large numbers of covariates, one needs an efficient way to construct proposals, which is the goal of CDA.

We now give a result on the convergence rates of CDA and CDA-MH for imbalanced intercept-only logistic regression. The result shows that the spectral gap is larger than that for DA (as a function of n), and comparable to MH with optimally tuned proposals when y grows no faster than $\log n$. While this is a special case, we note that the result in Johndrow et al. (2018) is given only for fixed $y = 1$, and thus our result is more general. The difficulty

of obtaining quantitative estimates of the rate at which the spectral gap converges to zero as n grows is underscored by the length and complexity of the arguments in Johndrow et al. (2018).

Consider intercept-only logistic regression from (9) with $x_i = 1$ for $i = 1, \ldots, n$ and prior $\theta \sim \text{No}(0, \sigma^2)$. As all p_i's are the same, we use a single scalar $r_i = r$ and $b_i = b$ for all i. With r, b fixed, the update rule for CDA-Gibbs is

$$\pi_{r,b}(z \mid \theta) = \text{PG}(nr, |\theta + b|)$$
$$f_{r,b}(\theta' \mid z) = \text{No}(m, \Lambda)$$

where $\Lambda = (z + 1/\sigma^2)^{-1}$, $m = \Lambda a - b$ and $a = \sum_i y_i - nr/2 + b/\sigma^2$.

Theorem 3 *Consider intercept-only logistic regression with n observations. Then*

1. *CDA-Gibbs is uniformly ergodic*

2. *CDA-MH is uniformly ergodic*

3. *If $\sum_i y_i = o(\log n)$, there exist choices for r, b such that CDA-MH has spectral gap*

$$\kappa = \mathcal{O}(n^{-\frac{2.5+2\log 2}{\sigma^2}}).$$

Thus, for $\sigma^2 > 5 + 4\log 2 \approx 7.77$, the spectral gap of CDA-MH goes to zero more slowly with increasing n than DA-Gibbs. Moreover, if we choose the prior $\sigma^2 = \log n$, the spectral gap of CDA-MH is independent of n. It follows that CDA-MH mixes rapidly as $n \to \infty$ in the large-sample imbalanced setting, unlike DA-Gibbs, which has spectral gap converging to zero at rate $n^{-1/2}$ or faster (ignoring logarithmic factors).

To show that the result is borne out empirically, we conduct simulations as in Johndrow et al. (2018), with fixed $\sum_i y_i = 1$ and increasing n from 10^1 to a massive 10^{14}. Figure 4 compares the effective sample size per $1,000$ steps using DA and CDA. The deterioration of DA shows up as early as $n = 10^2$; its slow-down becomes critical at $n = 10^4$ with effective sample size close to 0. CDA performs exceptionally well, even at $n = 10^{14}$ (we stop at 10^{14} as $1/n$ reaches the limit of floating point accuracy).

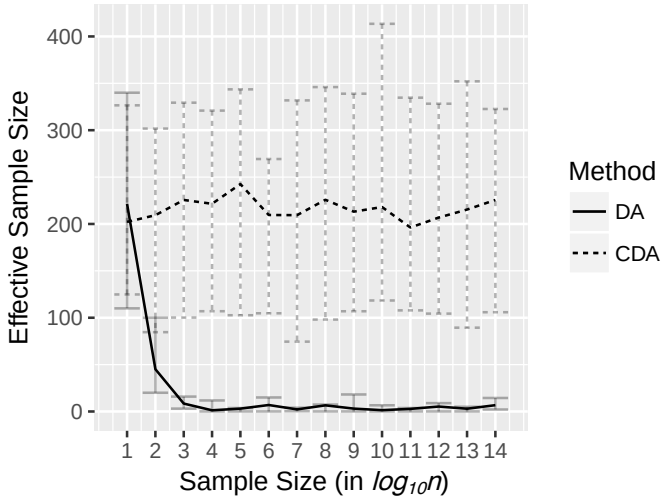


Figure 4: Effective sample size (with 95% pointwise confidence interval) per $1,000$ steps with different sample size n from 10 to 10^{14}, using logistic regression model with intercept.

6. Simulation Study

In this section, we compare the performance of CDA against popular alternative algorithms.

6.1. Comparison with Downsampling Algorithm

As motivated in the introduction, two factors are necessary for MCMC to be practically useful: a low computing cost in each iteration and a high effective sample size within a small number of iterations.

One potential issue for data augmentation in general is the large number of latent variables to sample in each iteration. A common strategy is to avoid sampling latent variables for every observation by approximating the Markov transition kernel using subsamples (Korattikara et al., 2014; Quiroz et al., 2018; Bardenet et al., 2017). Unlike other alternative algorithms we consider here, this changes the invariant measure. Finding a suitable subsample size while controlling the approximation error is challenging and usually problem-specific (Johndrow et al., 2017; Rudolf et al., 2018), and we do not consider it here. Instead, our goal is to show sub-sampling alone does not address the low ESS of DA; whereas one can trivially combine our proposed CDA strategy with subsampling to scale DA-MCMC up to enormous data sample sizes. We illustrate this strategy here.

We consider the same two-parameter intercept-slope model in logistic regression as described above, except we now vary data sample size from $n = 10^5$ to 10^8. We simulate Bernoulli outcomes $y_i \sim \text{Bernoulli}\big((1+\exp(-x_i\theta))^{-1}\big)$ with $x_i = (1, w_i)$ for $w_i \overset{iid}{\sim} \text{No}(0,1)$ and $\theta = (-\theta_0, 1)'$. We vary θ_0 to obtain $\sum y_i \approx 10$ for each n. We utilize the minibatch Polya-Gamma algorithm described by Johndrow et al. (2017), and apply CDA to calibrate

the variance discrepancy. Since y is highly imbalanced, we apply biased sampling by including all data with $y_i = 1$, while sub-sampling 1% of data with $y_i = 0$.

Denoting the set of all data with $y_i = 1$ as V_1 and a random subset with $y_i = 0$ as V_0, we adjust the likelihood contribution from $y_i = 0$ via a power of $(n - |V_1|)/|V_0|$ to compensate for the downsampling, leading to an approximate likelihood

$$L(\theta; y) = \prod_{i \in V_1} \frac{\exp(x_i\theta)}{1 + \exp(x_i\theta)} \left(\prod_{i \in V_0} \frac{1}{1 + \exp(x_i\theta)} \right)^{\frac{n-|V_1|}{|V_0|}}.$$

The number of latent variables is reduced to $n_0 \equiv |V_0| + |V_1|$; since n_0 is still large, slow mixing remains a problem and calibration is needed. The algorithmic details are presented in the appendix.

Figure 5 compares the performance of the two approximating algorithms, one combining CDA and sub-sampling, and one using sub-sampling alone. Clearly, sub-sampling alone still results in very small effective sample size, while using CDA and sub-sampling together can produce excellent computational performance.

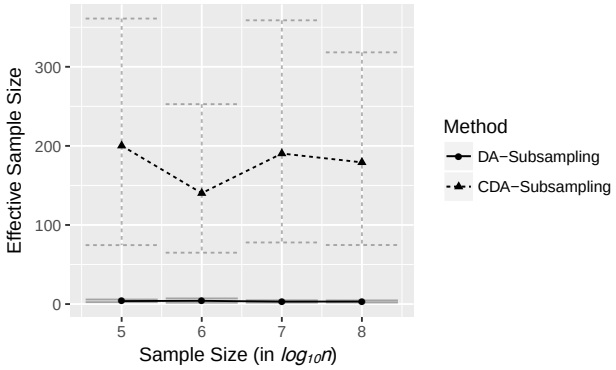


Figure 5: Comparing the performance of CDA and DA, coupled with sub-sampling approximation to reduce the number of sampled latent variables.

6.2. Comparison with Independence Metropolis-Hastings

In the CDA-MH algorithm, we utilize the marginal quantities $\hat{\theta}$ and $\mathcal{I}(\hat{\theta})$ to tune the (r, b) parameters. We compare the performance of the CDA proposal against alternative MH proposal with access to the same information. Specifically, we analyze MH using independent multivariate t proposals with mean $\hat{\theta}$ and variance $\mathcal{I}^{-1}(\hat{\theta})$. We show that this algorithm has very low acceptance rate relative to CDA-MH.

The general form of the MH acceptance rate is given by

$$\alpha(\theta, \theta^*) = \min \left\{ 1, \frac{L(\theta^*; y) q(\theta; \theta^*)}{q(\theta^*; \theta) L(\theta; y)} \frac{\Pi^0(\theta^*)}{\Pi^0(\theta)} \right\}.$$

Assuming the prior $\Pi^0(\theta)$ has negligible impact when n is large, the key to a high acceptance rate is to have $L(\theta; y)/q(\theta; \theta^*)$ close to a constant in the high posterior density region of the

parameter space. However, for computational convenience, one often has to use a proposal that is easy to sample. The density mis-match between $L(\theta; y)$ and $q(\theta; \theta^*)$ can cause the ratio to decrease rapidly moving away from the posterior mode of θ, resulting in a high rejection rate.

To illustrate, we consider the independent multivariate t-distribution proposal for logistic regression:

$$q(\theta^*, \theta) = q(\theta) = t_\nu \left\{ \theta; \hat{\theta}, (\nu - 2)\nu^{-1}\mathcal{I}^{-1}(\hat{\theta}) \right\},$$

where $\nu > 2$ and $\mathcal{I}(\hat{\theta}) = X^T \text{diag}[\exp(x_i\hat{\theta})\{1 + \exp(x_i\hat{\theta})\}^{-2}]X$. The second parameter is set to have $\text{var}(\theta) = \mathcal{I}^{-1}(\hat{\theta})$ exactly. We choose $\nu = 3$ to induce a tail heavier than the target likelihood, which is a necessary condition for geometric ergodicity of MH with independent proposals (Mengersen et al., 1996).

The density ratio is

$$\begin{aligned} \frac{L(\theta; y)}{q(\theta)} &= c_1 \left\{ \prod_i \frac{\exp(y_i x_i \theta)}{1 + \exp(x_i\theta)} \right\} \left\{ 1 + \frac{1}{\nu - 2}(\theta - \hat{\theta})^{\mathrm{T}} \mathcal{I}(\hat{\theta})(\theta - \hat{\theta}) \right\}^{(\nu+p)/2} \\ &= c_1 \frac{\exp(\sum_i y_i x_i \theta)}{\prod_i \{1 + \exp(x_i\theta)\}} \left[1 + \sum_i \frac{1}{\nu - 2}(x_i\theta - x_i\hat{\theta})^2 \exp(x_i\hat{\theta})\{1 + \exp(x_i\hat{\theta})\}^{-2} \right]^{(\nu+p)/2}, \end{aligned} \tag{14}$$

where c_1 denotes the constant part. We give an approximation of the acceptance ratio.

We focus on the case $\sum y_i \ll n$, where the mixing is slow for DA-Gibbs. This results in $\exp(x_i\hat{\theta}) \approx 0$ for most i. Assuming the high posterior density region is a neighborhood $\{\theta : |x_i\theta - x_i\hat{\theta}| < \eta \text{ for all } i\}$, where η is a bounded constant, the second term in (14) is close to a constant, while the first term is approximately equal to its numerator. The acceptance ratio is thus approximately

$$\frac{L(\theta^*; y)q(\theta)}{q(\theta^*)L(\theta; y)} \approx \exp\left\{ \sum_i y_i x_i (\theta^* - \theta) \right\},$$

which decreases exponentially away from the current state.

In contrast, since the CDA proposal density is similar to the target, with calibration the density ratio can be made close to a constant in the neighborhood of the mode. Consider the density ratio in the logistic CDA proposal:

$$\frac{L(\theta; y)}{L_{r,b}(\theta; y)} = c_2 \prod_i \frac{\{1 + \exp(x_i\theta + b_i)\}^{r_i}}{1 + \exp(x_i\theta)},$$

where c_2 is a constant. Minimizing the Fisher information distance gives approximately $r_i \approx \exp(x_i\hat{\theta})$ and $b_i \approx -x_i\hat{\theta}$, so the density ratio is approximately c_2. Thus the acceptance ratio

$$\frac{L(\theta^*; y)L_{r,b}(\theta; y)}{L_{r,b}(\theta^*; y)L(\theta; y)} \approx 1.$$

We compare the performance of MH algorithms with t_3 and CDA proposals, using the two-parameter intercept-slope example described in Section 6.1. Figure 6 shows the acceptance ratio at different intercept values θ_0, which is approximately the average of $x_i\theta$.

The acceptance rate drops rapidly to 0 for the t_3 proposal, and is close to one for the CDA proposal.

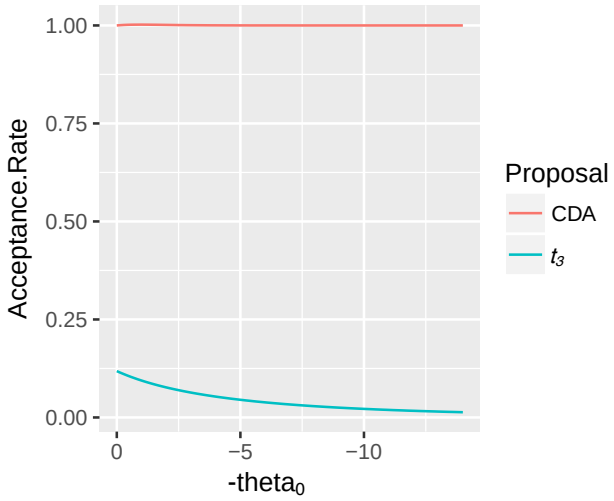


Figure 6: Comparing the acceptance ratios using the multivariate t-distribution and CDA proposals in logistic regression, with variance fixed at the inverse Fisher information. CDA has a much higher acceptance ratio than the multivariate t proposal.

7. Data Applications

7.1. Bernoulli Latent Factor Model with Group Intercepts for Network Modeling

We now apply CDA to accelerate estimation of group intercepts in a latent factor model. The dataset is a large sparse network from the Human Connectome Project (Marcus et al., 2011). The network data under consideration is an adjacency matrix representing the connectivity among $V = 1015$ macroscopic regions of one human brain. The matrix $\{A_{ij}\}_{(i,j)\in\{1\ldots V\}^2}$ is binary and symmetric. For $i \neq j$, $A_{ij} = 1$ if regions i and j are connected, $A_{ij} = 0$ otherwise; A_{ii} are missing as self-connections are ignored. Therefore, there are effectively $n = V(V-1)/2 = 514,605$ observed binary outcomes.

There are 507 regions located in the left ($\mathcal{L}$) and 508 in the right hemisphere ($\mathcal{R}$). There are many more connections within each hemisphere ($\sum A_{i\in\mathcal{L},j\in\mathcal{L},i>j} = 2,280$, $\sum A_{i\in\mathcal{R},j\in\mathcal{R},i>j} = 2,443$), than across hemispheres ($\sum A_{i\in\mathcal{L},j\in\mathcal{R}} = 23$). To quantify this phenomenon, we use two intercepts β_0 and β_1 to represent the within- and across-hemisphere fixed effects

within the following Bernoulli probit latent factor model

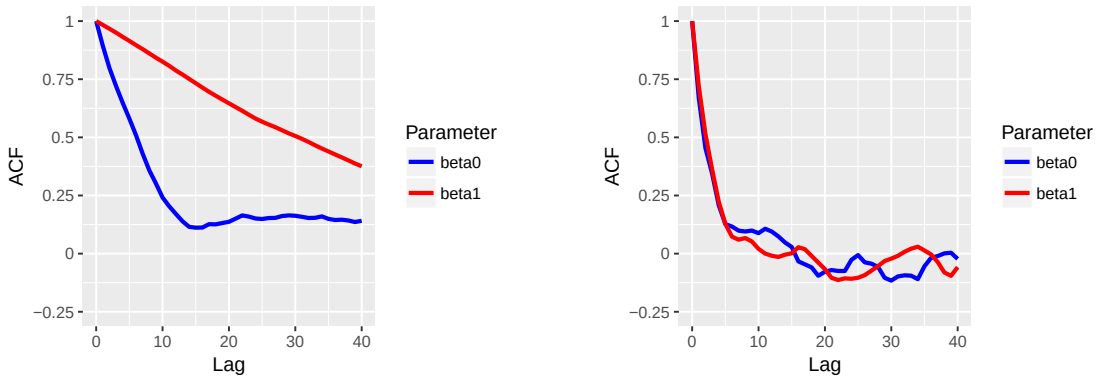
$$
\begin{aligned}
&A_{ij} \sim \text{Bernoulli}(p_{ij}), \qquad p_{ij} = \Phi(\psi_{ij}),\\
&\psi_{ij} = \sum_{r=1}^{d} u_{ir} v_r u_{jr} + \beta_0 w_{ij} + \beta_1 (1 - w_{ij}) \quad \text{for } i = 2 \dots V, j < i,\\
&\pi(U) \propto 1, \qquad U^{\mathrm{T}} U = I_d,\\
&v_r \sim \text{No}_{(0,\infty)}(0, \sigma^2) \qquad \text{for } r = 1, \dots, d,\\
&\beta_0 \sim \text{No}(0, 100), \quad \beta_1 \sim \text{No}(0, 100), \quad \sigma^2 \sim \text{Inverse-Gamma}(2, 1),
\end{aligned}
$$

where $w_{ij} = 0$ if $i \in \mathcal{L}$ and $j \in \mathcal{R}$, otherwise $w_{ij} = 1$; $U = \{u_{ir}\}$ is a V-by-d matrix of latent factors. Following Hoff (2009), we assign U a uniform prior on Stiefel manifold $\mathbb{S}^{V \times d} = \{U : U^{\mathrm{T}} U = I_d\}$, and set the latent dimension at $d = 10$. The latent variable updates in the probit data augmentation algorithm are given by

$$
\begin{aligned}
&z_{ij} \sim \begin{cases} \text{No}_{(0,\infty)}(\psi_{ij}, 1) & \text{if } A_{ij} = 1 \\ \text{No}_{(-\infty,0)}(\psi_{ij}, 1) & \text{if } A_{ij} = 0 \end{cases} \qquad \text{for } i = 2 \dots V, j < i,\\
&z_{ji} = z_{ij}.
\end{aligned}
$$

Because the connection data are highly imbalanced – fewer than 5,000 connections out of a possible 514,605 – the intercepts β_0 and β_1 mix slowly in an ordinary DA Gibbs algorithm (Figure 7(a)). Without using DA, efficient MH proposals are difficult to develop due to the restriction that $U \in \mathbb{S}^{V \times d}$. The DA-Gibbs relies on the full conditional distribution

$$
U \mid \beta_0, \beta_1, Z \sim \text{Bingham}(\{z_{ij} - \beta_0 w_{ij} - \beta_1 (1 - w_{ij})\}, \text{diag}\{v_r/2\}).
$$



(a) ACFs of the parameters β_0 and β_1 using DA.

(b) ACFs of the parameters β_0 and β_1 using CDA.

Figure 7: ACFs show the mixing performance of β_0 and β_1 in modeling average sparsity in network connectivity of a brain.

We use CDA to calibrate the updates of β_0 and β_1, while keeping the other Gibbs sampling steps unchanged, i.e.

$$z_{ij}^* \sim \begin{cases} \mathrm{No}_{(0,\infty)}(\psi_{ij} + b_{ij}, r_{ij}) & \text{if } A_{ij} = 1 \\ \mathrm{No}_{(-\infty,0)}(\psi_{ij} + b_{ij}, r_{ij}) & \text{if } A_{ij} = 0 \end{cases} \quad \text{for } i = 2 \dots V, j < i,$$

$$\beta_0^* \sim \mathrm{No}\Big([\sum_{j<i} \frac{w_{ij}}{r_{ij}}]^{-1} \sum_{j<i} [\frac{w_{ij}}{r_{ij}} (z_{ij}^* - b_{ij} - \sum_{r=1}^{d} u_{ir} v_r u_{jr})], [\sum_{j<i} \frac{w_{ij}}{r_{ij}}]^{-1} \Big),$$

$$\beta_1^* \sim \mathrm{No}\Big([\sum_{j<i} \frac{1-w_{ij}}{r_{ij}}]^{-1} \sum_{j<i} [\frac{1-w_{ij}}{r_{ij}} (z_{ij}^* - b_{ij} - \sum_{r=1}^{d} u_{ir} v_r u_{jr})], [\sum_{j<i} \frac{1-w_{ij}}{r_{ij}}]^{-1} \Big).$$

Then β_0^* and β_1^* are accepted via MH step with calibrated conditional density

$$L_{r,b}(\beta_0, \beta_1 \mid U, V, A) = \prod_{j<i} \Phi(\psi_{ij})^{A_{ij}} [1 - \Phi(\psi_{ij})]^{(1-A_{ij})}$$

$$\psi_{ij} = r_{ij}^{-1} [\sum_{r=1}^{d} u_{ir} v_r u_{jr} + \beta_0 w_{ij} + \beta_1 (1 - w_{ij}) + b_{ij}]$$

The tuning parameters are optimized using the approach described in Section 4, except with $x_i\theta$ replaced by $\sum_{r=1}^{d} u_{ir} v_r u_{jr} + \beta_0 w_{ij} + \beta_1 (1 - w_{ij})$.

	DA	CDA
β_0	-2.09 (-2.10, -2.08)	-2.09 (-2.10, -2.08)
β_1	-3.68 (-3.72, -3.64)	-3.75 (-3.86, -3.66)
Fitted AUC	90.5%	92.1%
T_{eff}/T	0.008	0.142
Avg Computing Time / T	2.0 sec	2.0 sec
Avg Computing Time / T_{eff}	251 sec	14.1 sec

Table 1: Parameter estimates and computing speed of DA and CDA in Bernoulli latent factor modeling of a brain network.

We run DA for $30,000$ steps and CDA for $2,000$ steps, so that they have approximately the same effective sample size (calculated with the `CODA` package in `R`). Both algorithms are initialized at the MAP estimates. CDA leads to significant reduction in autocorrelation (Figure 7(b)) and about 18 times lower computing time per effective sample size. We also compare the in-sample fitted AUCs, computed based on A_{ij} and the posterior mean of p_{ij}. The CDA estimates clearly have a better fit to the data.

7.2. Poisson Log-Normal Model for Web Traffic Prediction

As a second application, we apply CDA to an online browsing activity dataset obtained from a computational advertising company. The dataset contains a two-way table of visit count by users who browsed one of 96 websites belonging to clients of the computational advertising agency, and one of the $n = 59,792$ high-traffic sites during the same browsing

session. We refer to visiting more than one site during the same session as co-browsing. For each of the client websites, it is of commercial interest to identify the high-traffic sites with relatively high co-browsing rates, so that ads can be more effectively placed. In computational advertising, it is also valuable to understand the co-browsing behavior and predict the traffic pattern of users.

We consider a Poisson regression model for co-browsing. We use the co-browsing count of a single client website as the outcome y_i and the log of one plus the co-browsing count of the other 95 websites as the predictors, i.e. $x_{ij} = \log(x^*_{ij} + 1)$ for $i = 1, \ldots, 59792$ and $j = 1, \ldots, 95$, where x^* is the raw co-browsing count for high-traffic site i and client site j. A Gaussian random effect is included to account for over-dispersion relative to the Poisson distribution, leading to a Poisson log-normal regression model:

$$\begin{aligned} y_i \sim \text{Poisson}\left(\exp(x_i\beta + \tau_i)\right), \quad \tau_i \overset{iid}{\sim} \text{No}(\tau_0, \nu^2), \quad i = 1 \ldots n \\ \beta \sim \text{No}(0, I\sigma^2_\beta), \quad \tau_0 \sim \text{No}(0, \sigma^2_\tau) \quad \nu^2 \sim \pi(\nu^2). \end{aligned}$$

We assign a weakly informative prior for β and τ_0 with $\sigma^2_\beta = \sigma^2_\tau = 100$. For the over-dispersion parameter ν^2, we assign a non-informative flat prior on $(0, \infty)$.

When β and τ are sampled separately, the random effects $\tau = \{\tau_1, \ldots, \tau_n\}$ mix slowly. Instead, we sample β and τ jointly. Letting $\tilde{X}$ be the $n \times (n+p)$ matrix given by $\tilde{X} = [I_n\, X]$, and $\eta_i = x_i\beta + \tau_i$ the linear predictor, $\theta = \{\tau, \beta\}'$ can be sampled jointly in a block. An explanation of improved mixing with blocked sampling can be found in Liu (1994a).

We now focus on the mixing behavior of data augmentation. We first review data augmentation for the Poisson log-normal model. Zhou et al. (2012) proposed to treat Poisson(η_i) as the limit of the negative binomial $\text{NB}\big(\lambda, \eta_i/(\lambda + \eta_i)\big)$ with $\lambda \to \infty$, and used moderate $\lambda = 1,000$ for approximation. The limit relationship, omitting constants, is given by

$$L(\eta_i; y_i) = \frac{\exp(y_i\eta_i\}}{\exp\{\exp(\eta_i)\}} = \lim_{\lambda\to\infty} \frac{\exp(y_i\eta_i)}{\{1 + \exp(\eta_i)/\lambda\}^\lambda}. \tag{15}$$

With finite λ approximation, the posterior can be sampled using Polya-Gamma data augmentation

$$\begin{aligned} z_i \mid \eta_i &\sim \text{PG}(\lambda, \eta_i - \log\lambda) \quad i = 1 \ldots n \\ \theta \mid z, y &\sim \text{No}\Bigg(\Big(\tilde{X}'Z\tilde{X} + \begin{bmatrix} 1/\nu^2 \cdot I_n & 0 \\ 0 & 1/\sigma^2_\beta \cdot I_p \end{bmatrix}\Big)^{-1} \\ &\qquad \Big\{\tilde{X}'\big(y - \lambda/2 + Z\log\lambda\big) + \begin{bmatrix} \tau_0/\nu^2 1_n \\ 0_p \end{bmatrix}\Big\}, \\ &\qquad \Big(\tilde{X}'Z\tilde{X} + \begin{bmatrix} 1/\nu^2 \cdot I_n & 0 \\ 0 & 1/\sigma^2_\beta \cdot I_p \end{bmatrix}\Big)^{-1}\Bigg), \end{aligned}$$

where $Z = \text{diag}\{z_1, \ldots, z_n\}$, $1_n = \{1, \ldots, 1\}'$ and $0_p = \{0, \ldots 0\}'$.

However, this approximation-based data augmentation is inherently problematic. For example, setting $\lambda = 1,000$ leads to large approximation error. As in (15), the approximating denominator has $(1 + \exp(\eta_i)/\lambda)^\lambda = \exp\{\exp(\eta_i) + \mathcal{O}(\exp(2\eta_i)/\lambda)\}$; for moderately large $\eta_i \approx 10$, λ needs to be at least 10^9 to make $\exp(2\eta_i)/\lambda$ close to 0. This large error

cannot be corrected with an additional MH step, since the acceptance rate would be too low. On the other hand, it is not practical to use a large λ in a Gibbs sampler, as it would create extremely large z_i (associated with small conditional covariance for θ), resulting in slow mixing.

We use CDA to circumvent this issue. We first choose a very large λ (10^9) to control the approximation error, then use a small fractional r_i multiplying to λ for calibration. This leads to a proposal likelihood similar to the logistic CDA:

$$L_{r,b}(x_i\theta; y_i) = \frac{\exp(\eta_i - \log\lambda + b_i)^{y_i}}{\{1+\exp(\eta_i - \log\lambda + b_i)\}^{r_i\lambda}},$$

with $r_i \geq (y_i - 1)/\lambda + \epsilon$ for proper likelihood, and proposal update rule:

$$\begin{aligned}
z_i &\sim \mathrm{PG}(r_i\lambda, \eta_i - \log\lambda + b_i) \quad i = 1 \dots n \\
\theta^* &\sim \mathrm{No}\Bigg(\Big(\tilde{X}' Z \tilde{X} + \begin{bmatrix} 1/\nu^2 \cdot I_n & 0 \\ 0 & 1/\sigma_\beta^2 \cdot I_p \end{bmatrix}\Big)^{-1} \\
&\qquad \Big\{\tilde{X}'\big(y - r\lambda/2 + Z\log(\lambda - b)\big) + \begin{bmatrix} \tau_0/\nu^2 1_n \\ 0_p \end{bmatrix}\Big\}, \\
&\qquad \Big(\tilde{X}' Z \tilde{X} + \begin{bmatrix} 1/\nu^2 \cdot I_n & 0 \\ 0 & 1/\sigma_\beta^2 \cdot I_p \end{bmatrix}\Big)^{-1}\Bigg)
\end{aligned}$$

Letting $\eta_i^* = \tilde{X}\theta^*$, the proposal is accepted with probability (based on the Poisson density and the approximation $L_{r,b}(x_i\theta; y_i)$):

$$\min\left\{1, \prod_i \frac{\exp\{\exp(\eta_i)\}}{\exp\{\exp(\eta_i^*)\}} \frac{\{1+\exp(\eta_i^* - \log\lambda + b_i)\}^{r_i\lambda}}{\{1+\exp(\eta_i - \log\lambda + b_i)\}^{r_i\lambda}}\right\}.$$

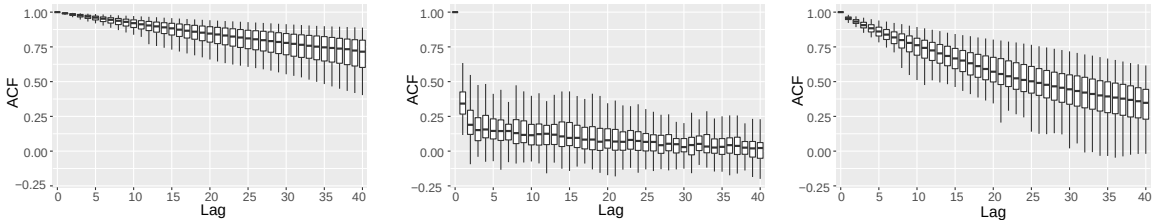
The tuning parameters are then optimized as described in Section 4, using

$$\begin{aligned}
\tilde{z}_i(\theta) =& \frac{\lambda r_i}{2|\eta_i - \log\lambda + b_i|} \tanh\left(\frac{|\eta_i - \log\lambda + b_i|}{2}\right) \quad i = 1, \dots, n, \\
\tilde{\theta}^*(z) =& \Big(\tilde{X}' Z \tilde{X} + \begin{bmatrix} 1/\nu^2 \cdot I_n & 0 \\ 0 & 1/\sigma_\beta^2 \cdot I_p \end{bmatrix}\Big)^{-1} \\
& \Big\{\tilde{X}'\big(y - r\lambda/2 + Z\log(\lambda - b)\big) + \begin{bmatrix} \tau_0/\nu^2 1_n \\ 0_p \end{bmatrix}\Big\}.
\end{aligned}$$

After θ is updated, the other parameters can be sampled via $\tau_0 \sim \mathrm{No}\big((n/\nu^2 + 1/\sigma_\tau^2)^{-1} \sum_i \tau_i/\nu^2, (n/\nu^2 + 1/\sigma_\tau^2)^{-1}\big)$ and $\nu^2 \sim \text{Inverse-Gamma}(n/2 - 1, \sum_i(\tau_i - \tau_0)^2/2)$.

We ran the ordinary DA algorithm with $\lambda = 1,000$, CDA with $\lambda = 10^9$ and Hamiltonian Monte Carlo with No-U-Turn sampler under the default tuning setting (as implemented in `STAN 2.17`). All algorithms are initialized at the MAP. We ran DA for $200,000$ steps, CDA for $2,000$ steps and HMC for $20,000$ steps so that they have approximately the same effective sample size. For CDA, we used the first $1,000$ steps for adapting r and b. Figure 8 shows empirical autocorrelations for DA, CDA and HMC. Even with small $\lambda = 1,000$ in DA, all of the parameters mix poorly; HMC seemed to be affected by the presence of random

effects, and most of parameters remain highly correlated within 40 lags; CDA substantially improves the mixing. Table 2 compares all three algorithms. CDA has the most efficient computing time per effective sample, and is about $30 - 300$ times more efficient than the other two algorithms.



(a) Autocorrelation of the parameters from DA.

(b) Autocorrelation of the parameters from CDA.

(c) Autocorrelation of the parameters from HMC.

Figure 8: CDA significantly improves the mixing of the parameters in the Poisson log-normal.

To evaluate predictive performance, we use another co-browsing count table for the same high traffic and client sites, collected during a different time period. We use the high traffic co-browsing count $x_{ij}^{\dagger *}$ and their log transform $x_{ij}^{\dagger} = \log(x_{ij}^{\dagger *} + 1)$ for the $j = 1, \ldots, 95$ clients to predict the count for the client of interest $y_i^{\dagger}$, over the high traffic site $i = 1, \ldots, 59792$. We predict using $\hat{y}_i^{\dagger} = \mathbb{E}_{\beta,\tau|y,x} y_i^{\dagger} = \mathbb{E}_{\beta,\tau|y,x} \exp(x_i^{\dagger}\beta + \tau_i)$ on the client site. The expectation is approximated using the MCMC sample path for $\beta, \tau \mid y, x$ obtained using training set $\{y, x\}$, as discussed above. Cross-validation root-mean-squared error $\left(\sum_i (\hat{y}_i^{\dagger} - y_i^{\dagger})^2/n\right)^{1/2}$ between the prediction and actual count $y_i^{\dagger}$'s is computed. As shown in Table 2, slow mixing in DA and HMC cause poor estimation of the parameters and high prediction error, while CDA has significantly lower error.

	DA	CDA	HMC
$\sum \beta_j/95$	0.072 (0.071, 0.075)	-0.041 (-0.042, -0.038)	-0.010 (-0.042, -0.037)
$\sum \beta_j^2/95$	0.0034 (0.0033, 0.0035)	0.231 (0.219 0.244)	0.232 (0.216 0.244)
$\sum \tau_i/n$	-0.405 (-0.642, -0.155)	-1.292 (-2.351, -0.446)	-1.297 (-2.354, -0.451)
$\sum \tau_i^2/n$	1.126 (0.968, 1.339)	3.608 (0.696, 7.928)	3.589 (0.678, 8.011)
Prediction RMSE	33.21	8.52	13.18
T_{eff}/T	0.0037 (0.0011 0.0096)	0.3348 (0.0279, 0.699)	0.0173 (0.0065, 0.0655)
Avg Comp. Time / T	1.3 sec	1.3 sec	56 sec
Avg Comp. Time / T_{eff}	346.4 sec	11.5 sec	3240.6 sec

Table 2: Parameter estimates, prediction error and computing speed of the DA, CDA and HMC in Poisson regression model.

8. Discussion

Data augmentation (DA) is a technique routinely used to enable implementation of simple Gibbs samplers, avoiding the need for expensive and complex tuning of Metropolis-Hastings algorithms. Despite the convenience, DA mixes slowly when the conditional posterior variance given the augmented data is substantially smaller than the marginal variance. When the data sample size is massive, this problem arises when the rates of convergence of the augmented and marginal posterior differ. There is a rich literature on strategies for improving mixing rates of Gibbs samplers, with centered or non-centered re-parameterizations (Papaspiliopoulos et al., 2007) and parameter-expansion (Liu and Wu, 1999) leading to some improvements. However, existing approaches do not solve large sample mixing problems because they do not address the fundamental rate mismatch issue.

To tackle this problem, we propose to calibrate data augmentation by directly adjusting the conditional variance (which is associated with step size). CDA adds a small cost for likelihood evaluation, which is often negligible when compared to the random number generation required at each iteration of DA. In this article, we demonstrate that calibration is generally applicable when $\theta \mid z$ belongs to a location-scale family. We expect it to be also useful outside of location-scale families, but have not pursued that here.

As both CDA and HMC involve MH steps, we draw some further comparison between the two. Both methods rely on finding a good proposal by searching a region far from the current state. One key difference lies in the computing efficiency. Although HMC is more generally applicable beyond data augmentation, it is computationally intensive since Hamiltonian dynamics often requires multiple numeric steps. CDA only requires one step of calibrated Gibbs sampling, which is often much more efficient, leveraging on existing data augmentation algorithms. The idea of using an auxiliary Gibbs chain to generate MH proposals seems generally promising (Tran et al., 2016), yet has received little attention in the literature.

In this work, we focused on cases when the sample size n is large, with the parameter dimension p moderate. One limitation of CDA-MH is that when p grows, in order to maintain a reasonable acceptance rate, the range to increase the conditional variance has to decrease. This is a common problem for general MH algorithms. Therefore, solutions to high dimensionality require further study.

Appendix A. Proof of Remark 1

Proof Since $q_{r,b}(\theta;\theta')$ is the θ marginal of a Gibbs transition kernel, and Gibbs is reversible on its margins, we have

$$q(\theta;\theta^*)\Pi_{r,b}(\theta^*) = q(\theta^*;\theta)\Pi_{r,b}(\theta),$$

and so

$$\begin{aligned}\frac{L(\theta^*;y)\Pi^0(\theta^*)q(\theta;\theta^*)}{L(\theta;y)\Pi^0(\theta)q(\theta^*;\theta)} &= \frac{L(\theta^*;y)\Pi^0(\theta^*)L_{r,b}(\theta;y)\Pi^0(\theta)}{L(\theta;y)\Pi^0(\theta)L_{r,b}(\theta^*;y)\Pi^0(\theta^*)} \\ &= \frac{L(\theta^*;y)L_{r,b}(\theta;y)}{L(\theta;y)L_{r,b}(\theta^*;y)}.\end{aligned}$$

■

Appendix B. Proof of Remark 2

Proof For any r, b, the conditionals $\Pi_{r,b}(z \mid \theta)$ and $\Pi_{r,b}(\theta \mid z)$ are well-defined for all $z \in \mathcal{Z}, \theta \in \Theta$, and therefore the Gibbs transition kernel $K_{r,b}((\theta, z); \cdot)$ and corresponding marginal kernels $Q_{r,b}(\theta; \cdot)$ are well-defined. Moreover, for any $(z, \theta) \in \mathcal{Z} \times \Theta$, we have $\mathbb{P}[(\theta', z') \in A \mid (\theta, z)] > 0$ by assumption. Thus $K_{r,b}$ is aperiodic and $\Pi_{r,b}$-irreducible (see the discussion following Corollary 1 in Roberts and Smith (1994)).

$Q_{r,b}(\theta'; \theta)$ is aperiodic and $\Pi_{r,b}(\theta)$-irreducible, since it is the θ marginal transition kernel induced by $K_{r,b}((\theta, z); \cdot)$. Thus, it is also $\Pi(\theta)$-irreducible so long as $\Pi \gg \Pi_{r,b}$, where for two measure μ, ν, $\mu \gg \nu$ indicates absolute continuity. Since $\Pi, \Pi_{r,b}$ have densities with respect to Lebesgue measure, $\Pi_{r,b}$-irreducibility implies Π irreducibility. Moreover, $q(\theta; \theta') > 0$ for all $\theta \in \Theta$. Thus, by Theorem 3 of Roberts and Smith (1994), CDA MH is Π-irreducible and aperiodic. ■

Appendix C. Toy example: Hierarchical Normal

To demonstrate the effects of r, b, we use a toy example commonly used in the data augmentation literature (Liu and Wu, 1999). Consider a marginal Normal model

$$y_i \sim \mathrm{No}(\theta, \sigma^2 + 1) \qquad i = 1, \ldots, n$$

with σ^2 known and improper prior $\pi(\theta) \propto 1$. This can be considered as a hierarchical model

$$y_i \sim \mathrm{No}(z_i, \sigma^2), \quad z_i \sim \mathrm{No}(\theta, 1), \qquad i = 1, \ldots, n, \tag{16}$$

where $z = \{z_1, \ldots, z_n\}$ are augmented data. The standard data augmentation algorithm has the update rule

$$\begin{aligned} z_i \mid y, \theta &\sim \mathrm{No}\left(\frac{y_i \sigma^{-2} + \theta}{\sigma^{-2} + 1}, \frac{1}{\sigma^{-2} + 1}\right) \qquad i = 1, \ldots, n \\ \theta \mid z &\sim \mathrm{No}(n^{-1} \sum_i z_i, n^{-1}). \end{aligned}$$

Thanks to the simple form, it is straightforward to compute the marginal variance of θ, $\mathrm{var}(\theta \mid y) = n^{-1}(1 + \sigma^2)$. Clearly, this is larger than the conditional variance $\mathbb{E}_z \mathrm{var}(\theta' \mid z) = n^{-1}$, when σ^2 is large.

To be able to adjust the conditional variance, we consider an alternative hierarchical model

$$y_i \sim \mathrm{No}(z_i, \sigma^2), \quad z_i \sim \mathrm{No}(\theta + b_0, r_0), \qquad i = 1, \ldots, n,$$

with update rule

$$\begin{aligned} z_i \mid y, \theta &\sim \mathrm{No}\left(\frac{y_i\sigma^{-2} + (\theta + b_0)r_0^{-1}}{\sigma^{-2} + r_0^{-1}}, \frac{1}{\sigma^{-2} + r_0^{-1}}\right) \qquad i = 1, \ldots, n \\ \theta^* \mid z &\sim \mathrm{No}(n^{-1}\sum_i z_i - b_0, n^{-1}r_0). \end{aligned} \tag{17}$$

To correct the deviation caused by the alternative model, we treat θ^* as a proposal to the target model (16), using M-H as in Remark 1 with $L_{r,b}(\theta; y) = (r_0 + \sigma^2)^{-1/2}\phi[(r_0 + \sigma^2)^{-1/2}(y_i - \theta - b_0)]$ and ϕ the standard normal density. We can choose r_0 so that the proposal variance equals to the target marginal variance $\mathrm{var}_{r,b}(\theta' \mid z) = \mathrm{var}(\theta \mid y)$; this yields

$$r_0 = 1 + \sigma^2.$$

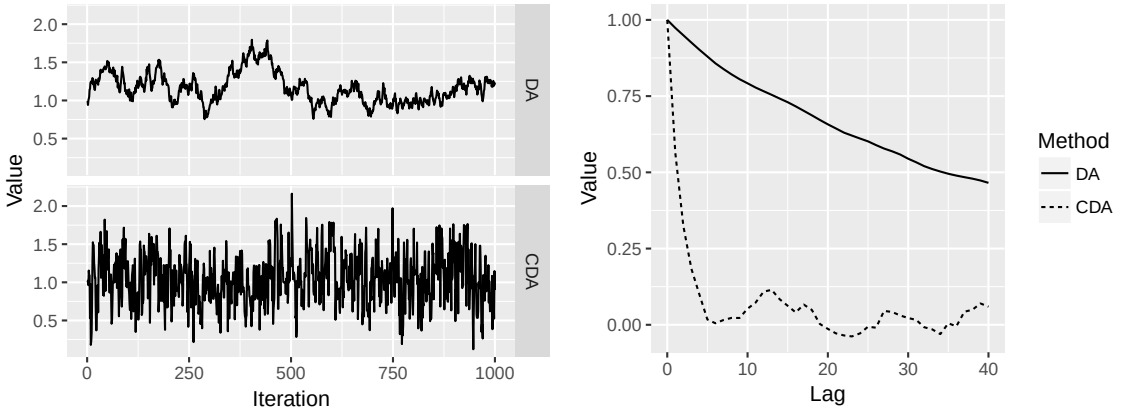
Note the proposal mean has

$$\mathbb{E}(\theta^* \mid \theta) = \mathbb{E}_{z|\theta}\mathbb{E}_{\theta^*|z}(\theta^* \mid z) = \theta + \frac{n^{-1}\sum_i y_i r_0}{\sigma^2 + r_0} - \frac{(b_0 + \theta)r_0}{\sigma^2 + r_0}$$

Intuitively, one way to improve the acceptance rate is to have the proposal centered at the current θ in the high posterior density region. That is, $\mathbb{E}(\theta^* \mid \theta) \approx \theta$ for θ near the MAP $\hat{\theta} = n^{-1}\sum_i y_i$. This yields one choice for b_0

$$\frac{n^{-1}\sum_i y_i r_0}{\sigma^2 + r_0} - \frac{(b_0 + \hat{\theta})r_0}{\sigma^2 + r_0} = 0 \qquad \Rightarrow b_0 = 0$$

We use $\sigma^2 = 100$, $\theta = 1$ to simulate $n = 1000$ data. Figure 9 compares the mixing performance, in terms of traceplots and autocorrelation plots (ACF) for the original DA and calibrated DA. Each algorithm was initiated at the MAP $\hat{\theta} = n^{-1}\sum_i y_i$. CDA significantly improves the mixing performance, with acceptance rate approximately 0.9.



(a) Traceplot for DA and CDA.

(b) ACF for DA and CDA.

Figure 9: Trace and autocorrelation plots for DA and CDA in hierarchical normal model.

Note in this special example, instead of relying on $\text{var}_{r,b}(\theta^* \mid z)$, one could directly adjust $\text{var}_{r,b}(\theta^* \mid \theta) = [r_0^2 + 2r_0\sigma^2]/[n(r_0 + \sigma_0^2)]$ to match $\text{var}(\theta \mid y)$. However, in general non-Gaussian cases, $\text{var}_{r,b}(\theta^* \mid \theta)$ is intractable, so we expect adjusting $\text{var}(\theta^* \mid z)$ to be more useful.

Appendix D. Calibrated Polya-Gamma Algorithm with Sub-sampling

Adapting based on Johndrow et al. (2017), we first randomly sample a subset of indices V of size $|V|$. This algorithm generates proposals from

$$
\begin{aligned}
&V = V_1 \cup V_0, \quad V_1 = \{i \in \{1, \ldots, n\} : y_i = 1\}, \quad V_0 \sim \text{Subset}(|V|, \{i \in \{1, \ldots, n\} : y_i = 0\}) \\
&z_i \sim \text{PG}(k_i r_i, |x_i\theta + b_i|) \quad i \in V, \\
&\theta^* \sim \text{No}\left((X_V' Z_V X_V)^{-1} X_V'(y_V - k_V r_V/2 - Z_V b_V), (X_V' Z_V X_V)^{-1}\right),
\end{aligned}
$$

where subscript $._V$ indicates the sub-matrix or sub-vector corresponding to the sub-sample; $k_i = 1$ if $y_i = 1$, and $k_i = (n - |V_1|)/|V_0|$. We accept θ^* in an MH step using calibrated likelihood

$$
L_{r,b}(\theta; y) = \prod_{i \in V_1} \frac{\exp(x_i\theta + b_i)}{\{1 + \exp(x_i\theta + b_i)\}^{r_i}} \Big(\prod_{i \in V_0} \frac{1}{\{1 + \exp(x_i\theta + b_i)\}^{r_i}}\Big)^{\frac{n - |V_1|}{|V_0|}},
$$

with target approximate likelihood $L_{1,0}(\theta; y)$.

Appendix E. Proof of Theorem 1

Proof Let Q be the proposal kernel for CDA-MH, which is identically the transition kernel for CDA-Gibbs, and let $\mathcal{P}$ be the Markov transition semigroup of CDA-MH.

Both have densities with respect to Lebesgue measure given by

$$
\begin{aligned}
q(\theta, \theta') &= \int_{\mathcal{Z}} f_{r,b}(\theta' \mid z, y) \pi_{r,b}(z \mid \theta, y) dz \\
p(\theta, \theta') &= \alpha(\theta, \theta') q(\theta, \theta') + \delta_\theta(\theta') \left(1 - \int \alpha(\theta, \tilde{\theta}) q(\theta, \tilde{\theta}) d\tilde{\theta}\right),
\end{aligned}
$$

respectively.

We seek a constant $c > 0$ and a density g such that

$$
\inf_{\theta \in \Theta} p(\theta, \theta') > c g(\theta')
$$

We proceed in the following steps:

1. Show that there exists a constant $c_1 > 0$ and a density g such that $\int_\Theta g(\theta) d\theta = 1$ for which

$$
\inf_{\theta \in \Theta} q(\theta, \theta') \geq c_1 g(\theta');
$$

conclude that CDA-Gibbs is uniformly ergodic.

2. Show that there exists $S \subset \Theta$ and a constant $c_2 > 0$ such that

$$\inf_{\theta \in \Theta, \theta' \in S} \alpha(\theta, \theta') > c_2.$$

3. Combine 1 and 2 to show $p(\theta, \theta') \geq \kappa g_S(\theta')$, where $\kappa = c_1 c_2 c_3$ with

$$c_3 = \int_S g(\theta) d\theta, \quad g_S(\theta) = c_3^{-1} g(\theta) \mathbf{1}\{\theta \in S\}$$

the restriction of g to S. Conclude that CDA-MH is uniformly ergodic with spectral gap κ.

4. Find values (r_0, b_0, S_0) of the tuning parameters r, b, S so that κ goes to zero slowly as $n \to \infty$.

1. Show that $q(\theta, \theta') \geq c_1 g(\theta')$. First we bound $\pi_{r,b}(\theta' \mid z)$ by a constant times a function depending on z

$$\begin{aligned}
\pi_{r,b}(\theta' \mid z) &= (2\pi)^{-1/2}(z + 1/\sigma^2)^{1/2} \exp\left[-\frac{1}{2}(\theta' - m)(z + 1/\sigma^2)(\theta' - m) \right] \\
&= (2\pi)^{-1/2}(z + 1/\sigma^2)^{1/2} \exp\left[-\frac{1}{2}\left\{ (\theta' + b)(z + 1/\sigma^2)(\theta' + b) - 2a(\theta' + b) + \frac{a^2}{z + 1/\sigma^2} \right\} \right] \\
&> (2\pi)^{-1/2}(1/\sigma^2)^{1/2} \exp\left[-\frac{1}{2}\left\{ (\theta' + b)(z + 1/\sigma^2)(\theta' + b) - 2a(\theta' + b) + \frac{a^2}{1/\sigma^2} \right\} \right] \\
&= (2\pi)^{-1/2}\sigma^{-1} \exp\left[-\frac{1}{2}\left\{ (\theta' + b)^2/\sigma^2 - 2a(\theta' + b) + a^2\sigma^2 \right\} \right] \exp\left[-\frac{1}{2}\left\{ (\theta' + b)^2 z \right\} \right]
\end{aligned} \tag{18}$$

in which the inequality holds since $z > 0$.

Using the Laplace transform of $\omega \sim \mathrm{PG}(\alpha, \beta)$

$$\mathbb{E}[\exp(-\omega t)] = \frac{\cosh^\alpha(\beta/2)}{\cosh^\alpha(\sqrt{(\beta^2/2 + t)/2})},$$

we proceed to bound the expectation of (18) with respect to z

$$\begin{aligned}
\int_0^\infty \exp\left[-\frac{1}{2}\left\{ (\theta' + b)^2 z \right\} \right] \pi_{r,b}(z \mid \theta) dz &= \cosh^{nr}\left(\frac{|\theta + b|}{2}\right) \cosh^{-nr}\left(\frac{\sqrt{((\theta + b)^2 + (\theta' + b)^2)}}{2}\right) \\
&\geq \cosh^{nr}\left(\frac{|\theta + b|}{2}\right) \cosh^{-nr}\left(\frac{(|\theta + b| + |\theta' + b|)}{2}\right) \\
&\geq 2^{-nr} \cosh^{nr}\left(\frac{|\theta + b|}{2}\right) \cosh^{-nr}\left(\frac{|\theta + b|}{2}\right) \cosh^{-nr}\left(\frac{|\theta' + b|}{2}\right) \\
&= 2^{-nr} \cosh^{-nr}\left(\frac{|\theta' + b|}{2}\right) \\
&\geq 2^{-nr} \exp\left[-\frac{nr|\theta' + b|}{2} \right]
\end{aligned}$$

$$\geq 2^{-nr} \exp\left[- \frac{nr[(\theta' + b)^2 + 1]}{4} \right]$$

where the first inequality uses $a^2 + b^2 \leq (|a| + |b|)^2$; the second uses Lemma 3.2 of Choi and Hobert (2013); the third uses the property of cosh; and the fourth uses $|a| \leq (1 + a^2)/2$. We combine to obtain $q(\theta, \theta') > c_1 g(\theta')$, viz

$$\begin{aligned}
q(\theta, \theta') &= \int_0^\infty \pi_{r,b}(\theta' \mid z, y_{1:n}) \pi_{r,b}(z \mid \theta, y_{1:n}) dz \\
&> (2\pi)^{-1/2} \sigma^{-1} \exp\left[- \frac{1}{2} \left\{ (\theta' + b)^2/\sigma^2 - 2a(\theta' + b) + a^2\sigma^2 \right\} \right] 2^{-nr} \exp\left[- \frac{nr[(\theta' + b)^2 + 1]}{4} \right] \\
&= (2\pi)^{-1/2} \sigma^{-1} 2^{-nr} \exp\left[- \frac{1}{2} \left\{ (\theta' + b)^2 (\frac{1}{\sigma^2} + \frac{nr}{2}) - 2a(\theta' + b) \right\} \right] \exp\left[- \frac{nr}{4} - \frac{a^2\sigma^2}{2} \right] \\
&= \sigma^{-1} 2^{-nr} (\frac{1}{\sigma^2} + \frac{nr}{2})^{-1/2} \exp\left[- \frac{nr}{4} - \frac{a^2\sigma^2}{2} \right] \exp\left[\frac{1}{2} a^2 (\frac{1}{\sigma^2} + \frac{nr}{2})^{-1} \right] \\
&\quad (2\pi)^{-1/2} (\frac{1}{\sigma^2} + \frac{nr}{2})^{1/2} \exp\left[- \frac{1}{2} \left(\theta' + b - (\frac{1}{\sigma^2} + \frac{nr}{2})^{-1} a \right)^2 (\frac{1}{\sigma^2} + \frac{nr}{2}) \right] \\
&= c_1 g(\theta')
\end{aligned}$$

where

$$\begin{aligned}
c_1 &= \sigma^{-1} 2^{-nr} (\frac{1}{\sigma^2} + \frac{nr}{2})^{-1/2} \exp\left[- \frac{nr}{4} - \frac{a^2\sigma^2}{2} \right] \exp\left[\frac{1}{2} a^2 (\frac{1}{\sigma^2} + \frac{nr}{2})^{-1} \right] \\
g(\theta') &= \mathrm{No}\left[\theta' \mid (\frac{1}{\sigma^2} + \frac{nr}{2})^{-1} a - b, (\frac{1}{\sigma^2} + \frac{nr}{2})^{-1} \right].
\end{aligned}$$

This completes the first part.

2. Show that $\inf_{\theta \in \Theta, \theta' \in S} \alpha(\theta, \theta') > c_2$ for some set S.

The acceptance ratio is

$$\alpha(\theta, \theta') = \min\left(\left[\frac{\alpha_0(\theta')}{\alpha_0(\theta)} \right]^n, 1 \right), \qquad \alpha_0(x) = \frac{\{1 + \exp(x + b)\}^r}{1 + \exp(x)}$$

Differentiating with respect to x we obtain

$$\frac{\partial \alpha_0(x)}{\partial x} = \frac{e^x \left(e^{b+x} + 1 \right)^{r-1} \left((r-1) e^b e^x + e^b r - 1 \right)}{\left(e^x + 1 \right)^2},$$

Assuming that $r < 1$ and $e^b r > 1$, since

$$\frac{e^x \left(e^{b+x} + 1 \right)^{r-1}}{\left(e^x + 1 \right)^2} > 0$$

and there is only one root on $(-\infty, \infty)$ for

$$(r-1) e^b e^x + e^b r - 1 = 0 \quad \Longrightarrow \quad x = \log(\frac{e^b r - 1}{1 - r}) - b \equiv \hat{\theta}$$

$$(r-1)e^b e^x + e^b r - 1 < 0 \quad \Longrightarrow \quad x > \hat{\theta}$$
$$(r-1)e^b e^x + e^b r - 1 > 0 \quad \Longrightarrow \quad x < \hat{\theta}$$

Therefore, $\hat{\theta}$ is the unique mode of α_0, and α_0 is (1) monotonically increasing for $x < \hat{\theta}$ and monotonically decreasing for $x > \hat{\theta}$.

For convenience, we write $b = -\log(r) + \xi$ with $\xi > 0$, so that $1 + \exp(\hat{\theta} + b) = (e^{\xi} - r)/(1 - r)$. Now set $S = (s_1, s_2)$. We now show that $\alpha(\theta, \theta') > c_2$ for $\theta' \in S$. We proceed in two cases.

1. **Case 1**: $\theta \leq \hat{\theta}$. We have three subcases

   (a) If $\theta < \theta' \leq \hat{\theta}$, then $\alpha_0(\theta') \geq \alpha_0(\theta)$, $\alpha(\theta, \theta') = 1$.

   (b) If $s_1 < \theta' \leq \theta \leq \hat{\theta}$ then

$$\begin{aligned}\alpha(\theta, \theta') =& \left(\frac{1 + \exp(\theta)}{1 + \exp(\theta')}\right)^n \left(\frac{1 + \exp(\theta' + b)}{1 + \exp(\theta + b)}\right)^{rn} \\ \geq& 1 \times \quad \left(\frac{1}{1 + \exp(\hat{\theta} + b)}\right)^{rn} = \left(\frac{1 - r}{e^{\xi} - r}\right)^{rn}\end{aligned}$$

   (c) If $\theta \leq \hat{\theta} < \theta' < s_2$,

$$\begin{aligned}\alpha(\theta, \theta') =& \left(\frac{1 + \exp(\theta)}{1 + \exp(\theta')}\right)^n \left(\frac{1 + \exp(\theta' + b)}{1 + \exp(\theta + b)}\right)^{rn} \\ \geq& \left(\frac{1}{1 + \exp(\theta')}\right)^n \times 1 \geq \left(\frac{1}{1 + \exp(s_2)}\right)^n\end{aligned}$$

   where we used that $\theta' > \theta$ so the second term is bounded below by 1, and that $1 + e^{\theta} > 1$. If $s_2 \leq \hat{\theta}$, then $\theta' < \hat{\theta}$, we only need to consider the condition (a) and (b).

2. **Case 2:** $\theta > \hat{\theta}$,

   (a) If $\hat{\theta} < \theta' \leq \theta$, then $\alpha_0(\theta') \geq \alpha_0(\theta)$, $\alpha(\theta, \theta') = 1$.

   (b) If $s_1 < \theta' \leq \hat{\theta} < \theta$, because α_0 is monotone nondecreasing on $(-\infty, \hat{\theta})$, we have:

$$\alpha_0(\theta') = \frac{\{1 + \exp(\theta' + b)\}^r}{1 + \exp(\theta')} \geq \lim_{\theta' \to -\infty} \alpha_0(\theta') = 1,$$

   Further, because α_0 is monotone nonincreasing on $(\hat{\theta}, \infty)$ we have

$$\frac{1}{\alpha_0(\theta)} = \frac{1 + \exp(\theta)}{\{1 + \exp(\theta + b)\}^r} \geq \frac{1}{\alpha_0(\hat{\theta})}$$
$$\alpha(\theta, \theta') = \alpha_0(\theta') \frac{1}{\alpha_0(\theta)}$$

$$\geq 1 \times \frac{1}{\alpha_0(\hat{\theta})} = \frac{\{1+\exp(\hat{\theta})\}^n}{\{1+\exp(\hat{\theta}+b)\}^{rn}}$$
$$\geq \frac{1}{\{1+\exp(\hat{\theta}+b)\}^{rn}} = \left(\frac{1-r}{e^{\xi}-r}\right)^{rn}$$

(c) If $\hat{\theta} < \theta < \theta' < s_2$,

$$\alpha(\theta,\theta') = \left(\frac{1+\exp(\theta)}{1+\exp(\theta')}\right)^n \left(\frac{1+\exp(\theta'+b)}{1+\exp(\theta+b)}\right)^{rn}$$
$$\geq \left(\frac{1}{1+\exp(\theta')}\right)^n \times 1 = \left(\frac{1}{1+\exp(s_2)}\right)^n$$

If $s_2 \leq \hat{\theta}$, then $\theta' < \hat{\theta}$ and we only need to consider the condition (b).

Combining (1) and (2), even when $s_2 \leq \hat{\theta}$, the lower bound still has:

$$\left(\frac{1-r}{e^{\xi}-r}\right)^{rn} \geq \min\left\{\left(\frac{1-r}{e^{\xi}-r}\right)^{rn}, \left(\frac{1}{1+\exp(s_2)}\right)^n\right\}$$

Therefore we have the common lower bound:

$$\alpha(\theta,\theta') \geq c_2, \quad c_2 = \min\left\{\left(\frac{1-r}{e^{\xi}-r}\right)^{rn}, \left(\frac{1}{1+\exp(s_2)}\right)^n\right\}$$

for $\theta' \in (s_1, s_2)$. Since this does not depend on s_1, we take $s_1 = -\infty$.

3. Combine to show $p(\theta,\theta') \geq c_1 c_2 c_3 g_S(\theta')$

Since

$$p(\theta,\theta') = \alpha(\theta,\theta')q(\theta,\theta') + \delta_\theta(\theta')\left(1 - \int \alpha(\theta,\tilde{\theta})q(\theta,\tilde{\theta})d\tilde{\theta}\right),$$
$$\geq \alpha(\theta,\theta')q(\theta,\theta'),$$

parts (1) and (2) establish the bound

$$\inf_{\theta\in\Theta} p(\theta,\theta') \geq c_1 c_2 g(\theta')\mathbf{1}\{\theta' \in S\}$$
$$= c_1 c_2 c_3 c_3^{-1} g(\theta')\mathbf{1}\{\theta' \in S\},$$

where

$$c_3 = \int g(\theta')\mathbf{1}\{\theta' \in S\},$$

so that $g_S(\theta') = c_3^{-1} g(\theta')\mathbf{1}\{\theta' \in S\}$ is a density. Specifically we have

$$g_S(\theta') = c_3^{-1}\left(\frac{1}{\sigma^2} + \frac{nr}{2}\right)^{1/2} \phi\left(\frac{\theta' - \left((\frac{1}{\sigma^2} + \frac{nr}{2})^{-1}a - b\right)}{(\frac{1}{\sigma^2} + \frac{nr}{2})^{-1/2}}\right)\mathbf{1}\{\theta' \in S\}$$

for $\phi(\cdot)$ the standard Gaussian density, where

$$c_3 = \Phi\left\{\left(\frac{1}{\sigma^2}+\frac{nr}{2}\right)^{1/2}\left[s_2-(\frac{1}{\sigma^2}+\frac{nr}{2})^{-1}a+b\right]\right\}$$
$$= \Phi\left[\left(\frac{1}{\sigma^2}+\frac{rn}{2}\right)^{1/2}(s_2+b)-\left(\frac{1}{\sigma^2}+\frac{rn}{2}\right)^{-1/2}a\right].$$

It follows that $\mathcal{P}$ is uniformly ergodic with spectral gap at least $\kappa = c_1c_2c_3$.

4. Tune constants so that $\kappa \to 0$ slowly as $n \to \infty$

We now may choose r, b, S in such a way as to minimize the rate at which the spectal gap goes to zero, subject to the constraints on r, b from part (2) and

$$\kappa(r,b,S) = c_1c_2c_3$$
$$= \sigma^{-1}2^{-nr}(\frac{1}{\sigma^2}+\frac{nr}{2})^{-1/2}\exp\left[-\frac{nr}{4}-\frac{a^2\sigma^2}{2}\right]\exp\left[\frac{1}{2}a^2\left(\frac{1}{\sigma^2}+\frac{nr}{2}\right)^{-1}\right]$$
$$\times \min\left\{\left(\frac{1-r}{e^{\xi}-r}\right)^{rn},\left(\frac{1}{1+\exp(s_2)}\right)^{n}\right\}$$
$$\times \Phi\left[\left(\frac{1}{\sigma^2}+\frac{rn}{2}\right)^{1/2}(s_2+b)-\left(\frac{1}{\sigma^2}+\frac{rn}{2}\right)^{-1/2}a\right]$$

First, we note that because $b = \xi - \log r$, tuning of r, ξ is equivalent to tuning of r, b, so we elect to do the former.

First, to reduce the effect of n, we set $r = w/n$, with $0 < w < n$. Noting $\exp(-a^2\sigma^2/2)$ decreases rapidly in a and recalling $a = \sum_i y_i - nr/2 + b/\sigma^2$ and $b = \xi - \log r$, we solve for w to make $a = 0$

$$\sum_i y_i - w/2 + (-\log(w)+\log(n)+\xi)/\sigma^2 = 0$$
$$\frac{\log w}{\sigma^2}+\frac{w}{2} = \sum_i y_i + \frac{\log n}{\sigma^2}+\frac{\xi}{\sigma^2}$$

assuming $\sum y_i + \xi/\sigma^2 = o(\log(n))$, we have

$$w = 2\log(n)/\sigma^2 + o(\log(n)).$$

Second, we make c_3 a constant independent of y, n, by choosing s_2 such that

$$(\frac{1}{\sigma^2}+\frac{rn}{2})^{1/2}(s_2+b)-(\frac{1}{\sigma^2}+\frac{rn}{2})^{-1/2}a = 0$$
$$(\frac{1}{\sigma^2}+\frac{rn}{2})^{1/2}(s_2+b) = 0$$
$$s_2 = -b = \log(w)-\log(n)-\xi.$$

which yields $c_3 = 0.5$.

Third, choose ξ so that

$$\xi \leq \log\left\{\left(1-\frac{w}{n}\right)e+\frac{w}{n}\right\} \Longrightarrow \log\left(\frac{e^{\xi}-w/n}{1-w/n}\right) \leq 1$$

meaning

$$\left(\frac{1-r}{e^{\xi}-r}\right)^{rn} = \left(\frac{1-w/n}{e^{\xi}-w/n}\right)^{w} = \exp\left(-w\log\left(\frac{e^{\xi}-w/n}{1-w/n}\right)\right) \geq e^{-w}$$

and

$$\left(\frac{1}{1+\exp(s_2)}\right)^{n} = \left(\frac{1}{1+we^{-\xi}/n}\right)^{n} \geq \exp(-we^{-\xi}) \geq e^{-w}$$

We have

$$c_2 = \min\left\{\left(\frac{1-r}{e^{\xi}-r}\right)^{rn}, \left(\frac{1}{1+\exp(s_2)}\right)^{n}\right\} \geq \exp(-w).$$

Combining results and choosing $r = r_0 = w/n$, $b = b_0 = -\log(w) + \log(n) + \xi$, $S = S_0 = (-\infty, \log(w) - \log(n) - \xi)$, with $(w, \xi) : \sum_i y_i - w/2 + (-\log(w) + \log(n) + \xi)/\sigma^2 = 0, \xi \leq \log[(1-w/n)e + w/n]$, we have

$$\begin{aligned}
\kappa(r_0, b_0, S_0) =& \sigma^{-1}2^{-w-1}(\frac{1}{\sigma^2}+\frac{w}{2})^{-1/2}\exp\left[-\frac{w}{4}\right]\exp(-w) \\
=& \mathcal{O}\bigg(\exp\big[-(\frac{5}{4}+\log 2)w\big]\bigg) \\
=& \mathcal{O}\bigg(n^{-\frac{5/2+2\log 2}{\sigma^2}}\bigg).
\end{aligned}$$

■

References

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

Patrick R Conrad, Youssef M Marzouk, Natesh S Pillai, and Aaron Smith. Accelerating asymptotically exact MCMC for computationally intensive models via local approximations. *Journal of the American Statistical Association*, 111(516):1591–1607, 2016.

Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, 65(3):457–483, 1978.

Peter D Hoff. Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009.

James E Johndrow, Jonathan C Mattingly, Sayan Mukherjee, and David B Dunson. Optimal approximating Markov chains for Bayesian inference. *arXiv preprint arXiv:1508.03387*, 2017.

James E Johndrow, Aaron Smith, Natesh Pillai, and David B Dunson. MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, (in press):1–44, 2018.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning*, pages 181–189, 2014.

Jun S Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427): 958–966, 1994a.

Jun S Liu. The fraction of missing information and convergence rate for data augmentation. *Computing Science and Statistics*, pages 490–490, 1994b.

Jun S Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

Dougal Maclaurin and Ryan P Adams. Firefly Monte Carlo: exact MCMC with subsets of data. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 543–552, 2015.

Daniel Marcus, John Harwell, Timothy Olsen, Michael Hodge, Matthew Glasser, Fred Prior, Mark Jenkinson, Timothy Laumann, Sandra Curtiss, and David Van Essen. Informatics and data mining tools and strategies for the human connectome project. *Frontiers in neuroinformatics*, 5:4, 2011.

Xiao-Li Meng and David A Van Dyk. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320, 1999.

Kerrie L Mengersen, Richard L Tweedie, et al. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B Dunson. Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18(1):4488–4527, 2017.

EWT Ngai, Yong Hu, YH Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011.

Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, (in press): 1–35, 2018.

Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and Their Applications*, 49(2):207–216, 1994.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

Daniel Rudolf, Nikolaus Schweizer, et al. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018.

Sanvesh Srivastava, Volkan Cevher, Quoc Tran-Dinh, and David B Dunson. WASP: scalable Bayes via Barycenters of subset posteriors. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 912–920, 2015.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

Minh-Ngoc Tran, Michael K Pitt, and Robert Kohn. Adaptive Metropolis–Hastings sampling using reversible dependent mixture proposals. *Statistics and Computing*, 26(1-2): 361–381, 2016.

Jon Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2): 158–183, 2007.

Xuerui Wang, Wei Li, Ying Cui, Ruofei Zhang, and Jianchang Mao. Click-through rate estimation for rare events in online advertising. *Online Multimedia Advertising: Techniques and Technologies*, pages 1–12, 2010.

Mingyuan Zhou, Lingbo Li, David B Dunson, and Lawrence Carin. Lognormal and Gamma mixed negative Binomial regression. In *Proceedings of the International Conference on Machine Learning*, volume 2012, page 1343, 2012.