Calibrated Data Augmentation for Scalable Markov Chain Monte Carlo

Leo L. Duan, James E. Johndrow, David B. Dunson

February 20, 2017

Abstract: Data augmentation is a common technique for building tuning-free Markov chain Monte Carlo algorithms. Although these algorithms are very popular, autocorrelations are often high in large samples, leading to poor computational efficiency. This phenomenon has been attributed to a discrepancy between Gibbs step sizes and the rate of posterior concentration. In this article, we propose a family of calibrated data augmentation algorithms, which adjust for this discrepancy by inflating Gibbs step sizes while adjusting for bias. A Metropolis-Hastings step is included to account for the slight discrepancy between the stationary distribution of the resulting sampler and the exact posterior distribution. The approach is applicable to a broad variety of existing data augmentation algorithms, and we focus on three popular models: probit, logistic and Poisson log-linear. Theoretical support is provided and dramatic gains are shown in applications.

KEY WORDS: Bayesian probit; Bayesian logit; Big n; Data Augmentation; 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 inferences and predictions in these and other settings.

The standard approach to Bayesian posterior computation is Markov chain Monte Carlo (MCMC) and related sampling algorithms. Non-sampling alternatives, such as variational Bayes, tend lack general accuracy guarantees. However, it is well known that 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 error. The latter is related to the properties of the Markov transition kernel; we will refer to this informally as the *mixing properties* of the Markov chain.

In recent years, a substantial literature has developed focusing on decreasing computational cost per iteration (Minsker et al. (2014); Srivastava et al. (2015); Conrad et al. (2015) among others), mainly through accelerating or parallelizing the sampling procedures at each iteration. Moreover, myriad strategies for improving mixing have been described in the literature. For Metropolis-Hastings (M-H) algorithms, improving mixing is usually a matter of constructing a better proposal distribution. An important difference between M-H and Gibbs is that one has direct control over step sizes in M-H through choice of the proposal, while Gibbs step sizes are generally not tunable; on the other hand, finding a good proposal for multi-dimensional parameters in M-H is significantly more challenging compared to Gibbs sampling. Thus, improving mixing for Gibbs has historically focused on decreasing autocorrelation by changing the update rule itself, for example by parameter expansion (PX), marginalization, or slice sampling.[1]

The theory literature on behavior of MCMC for large n and/or p is arguably somewhat limited. Many authors have focused on studying mixing properties by showing an ergodicity condition, such as geometric ergodicity (Roberts et al., 2004; Meyn and Tweedie, 2012). This generally yields bounds on the convergence rate and spectral gap of the Markov chain, but Rajaratnam and Sparks (2015) observe that in many cases, these bounds converge to zero exponentially fast in p or n, so that no meaningful guarantee of performance for large problem sizes is provided by most existing bounds. In the probability literature, a series of papers have developed an analogue of Harris' theorem and ergodic theory for infinite-dimensional state spaces (Hairer et al., 2011). Recent work verifies the existence of MCMC algorithms for computation in differential equation models with dimension-independent spectral gap (Hairer et al., 2014). In this example, the algorithm under consideration is an M-H algorithm, and it is clear that the proposal must be tuned very carefully to achieve dimension independence. Other work has studied the properties of the limiting differential equation that describes infinite-dimensional dynamics of MCMC.

A recent paper (Johndrow et al. (2016)) studies popular data augmentation algorithms for posterior computation in probit (Albert and Chib, 1993) and logistic (Polson et al., 2013) models, showing that the algorithms fail to mix in large sample sizes when the data are imbalanced. An important insight is that the performance can be largely explained by a discrepancy between the rate at which Gibbs step sizes and the width of the high-probability region of the posterior converge to zero as the sample size increases. Thus, since Gibbs step sizes are generally not tunable, slow mixing is likely to occur as the sample size grows unless the order of the step size happens to match the order of the posterior width. This implies that if a way to directly control the step sizes of the Gibbs sampler could be devised, it would be possible to make the mixing

[1] Although strictly speaking, slice sampling is just an alternative approach to sampling from a full conditional distribution, in practice, it is often an alternative to data augmentation, so that using a slice sampling strategy results in the removal of a data augmentation step from an alternative Gibbs sampler.

properties of the sampler insensitive to sample size by scaling the step sizes appropriately. This is similar to the conclusion of Hairer et al. (2014), except in this case, we have growing n instead of growing p.

In this article, we propose a method for tuning Gibbs step sizes by introducing auxiliary parameters that change the variance of full conditional distributions for one or more parameters. Although we focus on data augmentation algorithms for logit, probit, and Poisson log-linear models, in principle the strategy can be applied more generally to align Gibbs step sizes with the size of the space being explored. As these "calibrated" data augmentation algorithms alter the invariant measure, one can use the Gibbs step as a highly efficient M-H proposal, thereby recovering the correct invariant, or view the resulting algorithm as a perturbation of the original Markov chain. In this article, we focus on the former strategy, providing theoretical support and showing very substantial practical gains in computational efficiency attributed to our calibration approach.

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.

We provide brief motivation for our approach, with further theoretical development in Section 3. 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 $g : \Theta \to \mathbb{R}$ at stationarity can be expressed as the

Bayesian fraction of missing information (Papaspiliopoulos et al. (2007), Rubin (2004), Liu (1994b))

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

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

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

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

$$\gamma = \sup_{g \in L^2(\Pi)} \gamma_g = 1 - \inf_{g \in L^2(\Pi)} \frac{\mathbb{E}[\text{var}(g(\theta) \mid z)]}{\text{var}(g(\theta))}$$

is equal to the geometric convergence rate of the data augmentation Gibbs sampler (Liu (1994b)). For $g(\theta) = \theta_j$ 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, b will correspond to a location parameter and r a scale parameter that are tuned to increase $\mathbb{E}[\text{var}(g(\theta) \mid z)]\{\text{var}(g(\theta))\}^{-1}$, although the exact way in which they enter the likelihood and corresponding Gibbs update depend on the application. 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 $Q(\theta; \cdot)$ where

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

for $A \subseteq \Theta$, 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 working parameters during an adaptation phase to minimize the lag-1 autocorrelation for the identity function while maximizing the Metropolis-Hastings acceptance rate, we can select values of the working parameters that yield a computationally efficient algorithm.

2.1 Initial example: probit with intercept only

We first use a toy model to illustrate CDA. Consider an intercept-only probit

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

and improper prior $\pi(\theta) \propto 1$. The basic data augmentation algorithm (Tanner and Wong, 1987; Albert and Chib, 1993) has the update rule

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

$$\theta \mid z, y \sim \text{No}(\sum_i z_i/n, 1/n),$$

where $\text{No}_{[a,b]}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 truncated to the interval $[a, b]$. Johndrow et al. (2016) 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.

As the conditional variance $\text{var}(\theta \mid z, y)$ is independent of z, we introduce a scale parameter r in the update for z, then adjust the conditional mean by a location parameter b. This is equivalent to changing the scale of $z_i \mid \theta, y_i$ from 1 to r and the mean from θ to $\theta + b$. These adjustments yield

$$\text{pr}(y_i = 1 | \theta, r, b) = \int_0^\infty \frac{1}{\sqrt{2\pi r}} \exp\left(-\frac{(z_i - \theta - b)^2}{2r^2}\right) dz_i = \Phi\left(\frac{\theta + b}{\sqrt{r}}\right), \tag{4}$$

leading to the modified data augmentation algorithm

$$z_i \mid \theta, y_i \sim \begin{cases} \text{No}_{[0,\infty)}(\theta + b, r) & \text{if } y_i = 1 \\ \text{No}_{(-\infty,0]}(\theta + b, r) & \text{if } y_i = 0 \end{cases} \qquad i = 1, \ldots, n \tag{5}$$

$$\theta \mid z, y \sim \text{No}(\sum_i (z_i - b)/n, r/n).$$

To achieve step sizes consistent with the width of the high posterior probability region, we need

$$r/n \approx (\log n)^{-1},$$

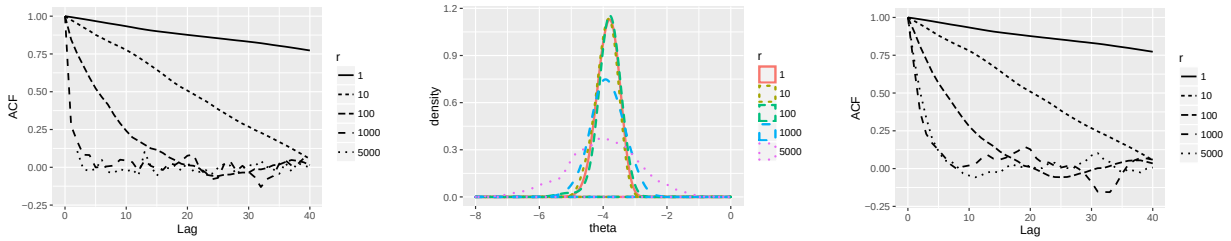
so $r \approx n/\log n$. To preserve the original target, we use (5) to generate an M-H proposal θ^*. The M-H acceptance probability is given by

$$1 \wedge \prod_i \frac{L_{r,b}(\theta; y_i) L(\theta^*; y_i)}{L_{r,b}(\theta^*; y_i) L(\theta; y_i)}, \tag{6}$$

where $L_{r,b}(\theta; y_i) = \Phi\big(\frac{\theta+b}{\sqrt{r}}\big)^{y_i} \Phi\big(-\frac{\theta+b}{\sqrt{r}}\big)^{(1-y_i)}$ and $L(\theta; y_i) = L_{1,0}(\theta; y_i)$. As we show in section 3, M-H algorithms using CDA proposals *always* have M-H acceptance probabilities of this form. Setting $r_i = 1$ and $b_i = 0$ 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 $r = n/\log n$, we then choose the b_i's to increase the acceptance rate in the M-H step. In this simple example, it is easy to compute a "good" value of b_i, since $b_i = -3.7(\sqrt{r} - 1)$ results in $\text{pr}(y_i = 1) = \Phi(-3.7) = n^{-1}\sum_i y_i \approx 10^{-4}$ in the proposal distribution, centering the proposals near the MLE for p_i.

We perform computation for these data with different values of r ranging from $r = 1$ to $r = 5,000$, with $r = 1,000 \approx n/\log n$ corresponding to the theoretically optimal value. Figure 1a plots autocorrelation functions (ACFs) for these different samplers without M-H adjustment. Autocorrelation is very high even at lag 40 for $r = 1$, while increasing r leads to dramatic improvements in mixing. There are no further gains in increasing r from the theoretically optimal value of $r = 1,000$ to $r = 5,000$. Figure 1b shows kernel-smoothed density estimates of the posterior of θ without M-H adjustment for different values of r 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. With M-H adjustment such differences are removed; the M-H step has acceptance probability close to one for $r = 10,100$, is 0.6 for $r = 1,000$, and 0.2 for $r = 5,000$.



(a) ACF for CDA without M-H adjustment.

(b) Density estimates of the posterior without M-H adjustment.

(c) ACF for CDA with M-H adjustment

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

2.2 Probit regression example

We now generalize to a probit regression:

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

with improper prior $\pi(\theta) \propto 1$ and the update rule

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

$$\theta \mid z, x, y \sim \text{No}((X'X)^{-1}X'z, (X'X)^{-1}).$$

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. Since the conditional variance is now a matrix, for flexible tuning, we let r and b vary over index i, yielding 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 modified Bernoulli probability:

$$\text{pr}(y_i = 1|\theta, x_i, r_i, b_i) = \int_0^\infty \frac{1}{\sqrt{2\pi r_i}} \exp\left(-\frac{(z_i - x_i\theta - b_i)^2}{2r_i^2}\right) dz_i = \Phi\left(\frac{x_i\theta + b_i}{\sqrt{r_i}}\right). \tag{8}$$

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 $K_{r,b}((\theta, z); \cdot)$ defined by the Gibbs update rule in (7) would have a unique invariant measure for fixed r, b, which we denote $\pi_{r,b}(\theta, z \mid y)$.

To preserve the original target $\pi_{1,0}(\theta, z \mid y)$, we use (7) to generate an M-H proposal. Specifically, we propose from $Q(\theta^*; \theta) = \int f(\theta^*; z, y)\pi(z; \theta, y)dz$, the θ-marginal of the transition kernel $K_{r,b}((\theta^*, z^*); (\theta, z))$. The acceptance probability is given by

$$1 \wedge \prod_i \frac{L_{r,b}(x_i\theta; y_i)L(x_i\theta^*; y_i)}{L_{r,b}(x_i\theta^*; y_i)L(x_i\theta; y_i)}, \tag{9}$$

where $L_{r,b}(\eta_i; y_i) = \Phi\big(\frac{\eta_i + b}{\sqrt{r}}\big)^{y_i}\Phi\big(-\frac{\eta_i + b}{\sqrt{r}}\big)^{(1-y_i)}$ and we denote $L_{1,0}$ by L.

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

$$\begin{aligned} \text{cov}_{r,b}(\theta_t \mid \theta_{t-1}, X, z, y) &= (X'R^{-1}X)^{-1} + (X'R^{-1}X)^{-1}X'R^{-1}\,\text{cov}(z - b|R)R^{-1}X(X'R^{-1}X)^{-1} \\ &\geq (X'R^{-1}X)^{-1}, \end{aligned} \tag{10}$$

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

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

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

$$\mathbb{E}_{\pi_{r,b}}[\text{var}(\theta_j \mid z)] \geq r_0\mathbb{E}_\pi[\text{var}(\theta_j \mid z)].$$

Of course, this auxiliary Gibbs step is used to generate proposals, so the key to CDA is to choose r, b to make $\mathbb{E}_{\pi_{r,b}}[\text{var}(\theta_j \mid z)]$ close to $\text{var}_{\pi_{r,b}}(\theta_j \mid z)$, while additionally maximizing the M-H acceptance probability. We defer the choice for r, b and their effects to the last subsection.

2.3 Logistic regression example

Calibration was easy to achieve in the probit examples, because $\text{var}(\theta|z,y)$ does not involve the latent variable z. In cases in which the latent variable impacts the variance of the conditional posterior distribution of θ, we propose to increase $\mathbb{E}_z\text{var}(\theta|z,y)$ by modifying the distribution of z. We focus on the logistic regression model with

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

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

$$z_i \sim \text{PG}(1, |x_i\theta|) \quad i = 1,\ldots,n,$$

$$\theta \sim \text{No}\left((X'ZX)^{-1}X'(y-0.5), (X'ZX)^{-1}\right),$$

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

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

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

We rely on replacing $\text{PG}(z_i \mid 1, 0)$ with $\text{PG}(z_i \mid r_i, 0)$ in the step for updating the latent data. Smaller r_i can lead to larger $\mathbb{E}_z\text{var}(\theta|z,y)$, providing a route towards calibration. Applying the bias-adjustment term b_i to the linear predictor $\eta_i = x_i\theta$ leads to

$$\begin{aligned} L_{r,b}(x_i\theta; y_i) =& \int_0^\infty \exp\{(x_i\theta + b_i)(y_i - r_i/2)\}\exp\left\{-\frac{z_i(x_i\theta+b_i)^2}{2}\right\}\text{PG}(z_i \mid r_i, 0)dz_i \\ =& \frac{\exp\{(x_i\theta+b_i)y_i\}}{\{1+\exp(x_i\theta+b_i)\}^{r_i}}, \end{aligned} \tag{11}$$

and the update rule for the proposal:

$$z_i \sim \text{PG}(r_i, |x_i\theta + b_i|) \quad i = 1,\ldots,n,$$

$$\theta^* \sim \text{No}\left((X'ZX)^{-1}X'(y - r/2 - Zb), (X'ZX)^{-1}\right),$$

with acceptance probability:

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

where $L(\theta; y_i) = \frac{\exp(\theta y_i)}{1+\exp(\theta)}$.

To demonstrate why smaller r_i leads to larger $\mathbb{E}_z(X'ZX)^{-1}$, we compute the first negative moment of the Polya-Gamma distribution. Combining Cressie et al. (1981) and Polson et al. (2013), $\mathbb{E}z_i^{-1} = \int_0^\infty \prod_{k=1}^\infty (1 + d_k^{-1}t)^{-r_i}dt$ with $d_k = 2(k-\frac{1}{2})^2\pi^2 + \frac{(x_i\theta+b_i)^2}{2}$, where $(1 + d_k^{-1}t) > 1$.

2.4 Choice of calibration parameters

As illustrated in the previous subsection, efficiency of CDA is dependent on a good choice of the calibration parameters $r = (r_1, \ldots, r_n)$ and $b = (b_1, \ldots, b_n)$. In the intercept-only case for probit, it was possible to analytically calculate appropriate values of these parameters, but in general an empirical approach is needed. We now propose a simple and efficient algorithm for calculating these parameters relying on Fisher information.

In large samples, the inverse of Fisher information evaluated at the posterior mode is a good approximation to the posterior marginal covariance. When the expectation of conditional variance is intractable, we can consider the inverse of expected conditional precision as an approximate. Depending on which is convenient to compute, it is useful to choose r to minimize the one of the distances below:

$$\Delta_{\mathcal{I}^{-1}}(\hat{\theta}_{MAP}) = \|\mathcal{I}^{-1}(\hat{\theta}_{MAP}) - \mathbb{E}_{z|\theta,r,b} \operatorname{var}(\theta^* \mid z)\Big|\hat{\theta}_{MAP}\|_F$$

$$\Delta_{\mathcal{I}}(\hat{\theta}_{MAP}) = \|\mathcal{I}(\hat{\theta}_{MAP}) - \mathbb{E}_{z|\theta,r,b}\big(\operatorname{var}(\theta^* \mid z)^{-1}\big)\Big|\hat{\theta}_{MAP}\|_F$$

where $\hat{\theta}_{MAP}$ is the posterior mode under the target likelihood L, $\mathcal{I}(\hat{\theta}_{MAP}) = E_y\left[\left(\frac{\partial}{\partial\theta}\log L(y;\theta)\right)^2\right]\Big|\hat{\theta}_{MAP}$ with expectation taken over the distribution of the data y under L, and $\|A\|_F$ is the Frobenius norm of A; $\mathbb{E}_{z|\theta,r,b}$ is taken over the latent variable $z \mid \theta$ under $L_{r,b}(\theta; y)$;.

In general, we do not have access to the posterior mode, so we instead use samples from the Markov chain during an adaptation phase to dynamically update r_t, b_t. Specifically, we choose r_{t+1} to minimize $\Delta_{\mathcal{I}}(\theta_t)$ or $\Delta_{\mathcal{I}^{-1}}(\theta_t)$, and set b_{t+1} to minimize the difference between $L_{1,0}(\theta_t; y)$ and $L_{r_{t+1},b_{t+1}}(\theta_t; y)$. Thus, we use r to adjust the conditional variance based on $L_{r,b}$ to match the marginal variance based on L, and b to make $L_{r,b}$ close to $L_{1,0}$ in the neighborhood of θ_t. Intuitively, this will make the target distribution closer to the invariant measure of calibrated Gibbs, and correspondingly increase the MH acceptance rate. For some data augmentation (e.g. Poisson log-normal as discussed below), it is possible that the approximation using Fisher information over-estimates the marginal variance, exhibiting low acceptance rate. When that occurs, we multiply a common constant to the initially estimated r_i to reduce the conditional variance increase. Since smaller variance increase makes $L_{r,b}$ closer to $L_{1,0}$, acceptance rate generally increases.

The proposal kernel we describe above is *adaptive*; that is, we have a collection of proposal kernels $\mathcal{Q} = \{Q_{r,b}\}_{(r,b)\in\mathbb{R}_+\times\mathbb{R}}$, and we choose a different member of $\mathcal{Q}$ at each iteration to create the proposal. The *target* for the resulting transition kernel is $\Pi_{1,0}$ for every $Q_{r,b}$ because of the Metropolis-Hastings rejection step. In general, ergodicity of adaptive algorithms requires a diminishing adaptation condition; a general condition of this sort is given in Roberts and Rosenthal (2007). Although the algorithm we describe is

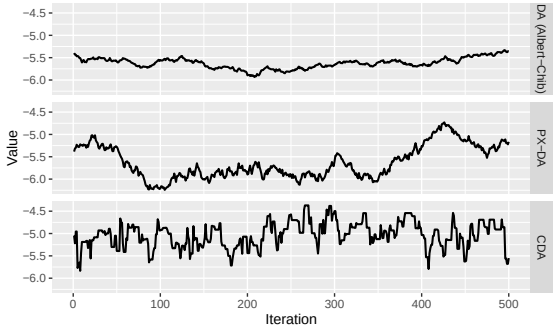
unlikely to satisfy diminishing adaptation if updating of r, b continues indefinitely, the condition is trivially satisfied by any algorithm that stops adaptation after a fixed number of iterations. Thus, for simplicity, we choose a tuning period length, after which we fix r, b at their current values. More sophisticated adaptation schemes could be devised; however, the fixed tuning period works well empirically.

For a concrete illustration, we first return to the first example of probit regression. Putting $\eta_i = x_i\theta$, the inverse Fisher information based on the marginal and the expected conditional variance are:

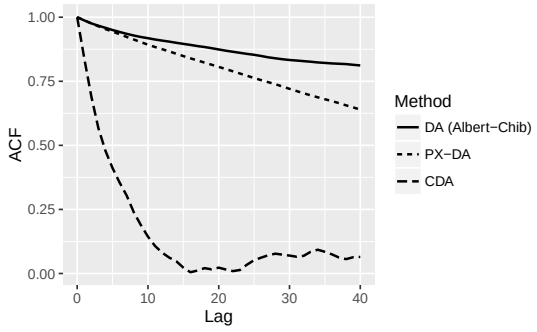
$$\mathcal{I}^{-1}(\theta) = \left[X' \operatorname{diag}\left\{\frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))}\right\}X\right]^{-1}, \qquad \mathbb{E}_z \operatorname{var}_{r,b}(\theta \mid z) = (X'R^{-1}X)^{-1}$$

respectively, where ϕ is the standard normal density. Therefore, setting $r_i = \frac{\Phi(\eta_i)(1-\Phi(\eta_i))}{\phi(\eta_i)^2}$ makes $\Delta_{\mathcal{I}^{-1}}(\theta_t) = 0$. The r_i's can be calculated using this expression at low cost without calculating the full information matrix. We then choose b_t to increase the acceptance rate in the M-H step, $1 \wedge \prod_i \frac{L_{r,b}(\eta_i;y_i)L(\eta_i^*;y_i)}{L_{r,b}(\eta_i^*;y_i)L(\eta_i;y_i)}$. We set $b_i = \eta_i(\sqrt{r_i} - 1)$ to ensure that $|L_{r,b}(\eta_i; y_i) - L(\eta_i; y_i)| = 0$ and proposals near η_i have relatively large acceptance rate.

To illustrate, we consider a 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 $\sum y_i = 20$ among $n = 10,000$. The Albert and Chib (1993) DA algorithm mixes slowly (Figure 2a and 2b). 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. For CDA, we tuned r and b for 100 steps using the Fisher information, reaching an satisfactory acceptance rate of 0.6. Applying CDA afterwards, we obtain 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.

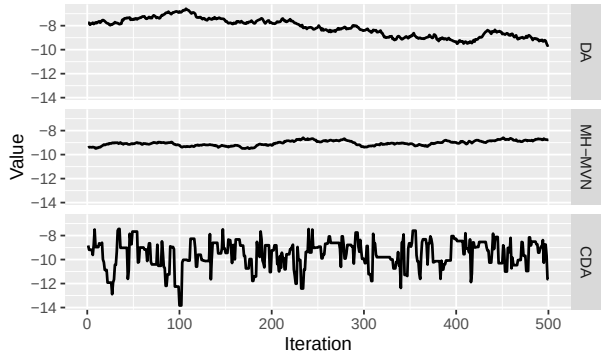
Figure 2: Panel (a) demonstrates in traceplot and panel (b) in autocorrelation the substantial improvement in CDA by correcting the variance mis-match in probit regression with rare event data, compared with the original (Albert and Chib, 1993) and parameter-expanded methods (Liu and Wu, 1999).

For the second example of logistic regression, the Fisher information based on the marginal and the expected conditional precision are:

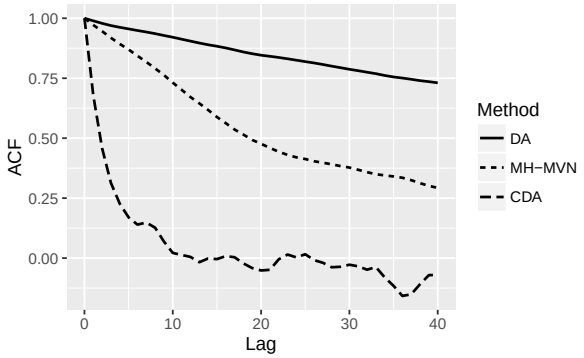
$$\mathcal{I}(\theta) = X' \operatorname{diag}\left\{ \frac{\exp(x_i\theta)}{\{1+\exp(x_i\theta)\}^2} \right\} X, \quad \mathbb{E}_z \operatorname{var}_{r,b}(\theta \mid z)^{-1} = X' \operatorname{diag}\left\{ \frac{r_i}{2|x_i\theta + b_i|} \tanh\left(\frac{|x_i\theta+b_i|}{2}\right) \right\} X$$

Setting $r_i = \frac{\exp(x_i\theta)}{\{1+\exp(x_i\theta)\}^2} 2|x_i\theta + b_i| / \tanh(\frac{|x_i\theta+b_i|}{2})$ makes $\Delta_{\mathcal{I}}(\theta_t) = 0$; and setting $b_i = \log[\{1+\exp(x_i\theta)\}^{1/r_i} - 1] - x_i\theta$ ensures $\{1+\exp(x_i\theta)\} = \{1+\exp(x_i\theta+b_i)\}^{r_i}$.

To illustrate, we use a two parameter intercept-slope model with $x_1 \sim \text{No}(0,1)$ and $\theta = (-9,1)'$. With $n = 10^5$, we obtain rare outcome data with $\sum y_i = 50$. Besides the original DA algorithm (Polson et al., 2013), we also consider an M-H sampler using a multivariate normal proposal $\theta^*|\theta \sim \text{No}(\theta^*|\theta, \mathcal{I}^{-1}(\theta))$ with the inverse Fisher information as the covariance. For CDA we tuned r and b for 100 steps using the Fisher information, reaching an acceptance rate of 0.8. Shown in Figure 3, both DA and M-H with a normal proposal mix slowly, exhibiting strong autocorrelation even at lag 40, while CDA has dramatically better mixing.



(a) Traceplots for DA, CDA and M-H with multivariate normal proposal.



(b) ACF for DA, CDA and M-H with multivariate normal proposal.

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) and the M-H algorithm with multivariate normal proposal (MH-MVN).

3 Theory

In this section, we provide basic theoretical support for CDA algorithms. First, we show that CDA M-H is ergodic. This is basically a consequence of CDA Gibbs being ergodic for fixed r, b and the fact that $\Pi_{r,b}$ and Π are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^p$.

Remark 1 (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* $\Pi_{r,b}(d\theta, dz)$*. Moreover, 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* $\Pi(d\theta)$*.*

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.

$Q_{r,b}(\theta'; \theta)$ 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 [CITE Roberts 1994], CDA M-H is Π-irreducible and aperiodic. □

Having established ergodicity of both CDA Gibbs and CDA M-H under weak assumptions that hold for all of the data augmentation strategies we consider here, we now provide a semi-rigorous argument for why our approach to tuning r and b results in both rapid convergence and closeness of $\Pi_{r,b}$ to Π. Suppose there exists r such that

$$\mathbb{E}_{\Pi_{r,b}}[\mathrm{var}(\theta \mid z)] = \mathrm{var}_{\Pi_{r,b}}(\theta)$$

for any value of b. This is a simplification, since our adaptation strategies only allow us to control the discrepancy between the Fisher information at the current state. By tuning r during the adaptation phase to make the lag-1 autocorrelation for the identity function small, we can numerically approximate the correct value of r.

This is obviously much weaker than minimizing the autocorrelation for worst-case functions. However, for the sake of exposition, we will proceed on the assumption that (1) we can make the lag-1 autocorrelation for the identity function zero by appropriately tuning r and (2) this is sufficient to obtain a Gibbs transition kernel that generates nearly *independent* samples. This makes the rationale for tuning b to increase the Metropolis acceptance probability much clearer. First, we note the form of the Metropolis acceptance ratios, which we have used previously without rigorous justification.

Remark 2. *The CDA M-H acceptance ratio is given by*

$$\frac{L(\theta'; y)\Pi^0(\theta')Q_{r,b}(\theta; \theta')}{L(\theta; y)\Pi^0(\theta)Q_{r,b}(\theta'; \theta)} = \frac{L(\theta'; y)L_{r,b}(\theta; y)}{L(\theta; y)L_{r,b}(\theta'; y)} \tag{12}$$

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

$$\frac{L(\theta'; y)\Pi^0(\theta')Q(\theta; \theta')}{L(\theta; y)\Pi^0(\theta)Q(\theta'; \theta)} = \frac{L(\theta'; y)\Pi^0(\theta')L_{r,b}(\theta; y)\Pi^0(\theta)}{L(\theta; y)\Pi^0(\theta)L_{r,b}(\theta'; y)\Pi^0(\theta')}$$

$$= \frac{L(\theta';y)L_{r,b}(\theta;y)}{L(\theta;y)L_{r,b}(\theta';y)}.$$

□

The expression in (12) will be near 1 at stationarity if

$$\int \log\left(\frac{L(\theta';y)L_{r,b}(\theta;y)}{L(\theta;y)L_{r,b}(\theta';y)}\right) Q_{r,b}(\theta';\theta)\Pi(d\theta) \approx 0.$$

Now, suppose that a Markov chain evolving according to $K_{r,b}$ is rapidly mixing, so that for starting measures satisfying a condition like

$$\sup_A \frac{\nu(A)}{\Pi_{r,b}(A)} < M$$

for M not too large we have

$$\mathrm{KL}\left(\Pi_{r,b} \parallel \int Q_{r,b}(\theta';\theta)\nu(d\theta)\right) \quad \text{small.}$$

Then the symmetric KL is

$$\begin{aligned}
\mathrm{KL}\left(\Pi_{r,b} \parallel \Pi\right) + \mathrm{KL}\left(\Pi \parallel \Pi_{r,b}\right) &= \int \Pi_{r,b}(d\theta) \log \frac{\Pi_{r,b}(\theta)}{\Pi(\theta)} + \int \Pi(d\theta) \log \frac{\Pi(\theta)}{\Pi_{r,b}(\theta)} \\
&= \int \Pi_{r,b}(d\theta) \log \frac{c_{r,b}L_{r,b}(\theta)\Pi_0(\theta)}{cL(\theta)\Pi_0(\theta)} + \int \Pi(d\theta) \log \frac{cL(\theta)\Pi_0(\theta)}{c_{r,b}L_{r,b}(\theta)\Pi_0(\theta)} \\
&\approx \int K_{r,b}(\theta';\theta)\Pi(d\theta) \log \frac{L_{r,b}(\theta')}{L(\theta')} + \int \Pi(d\theta) \log \frac{L(\theta)}{L_{r,b}(\theta)} \\
&= \mathbb{E}\left[\frac{L_{r,b}(\theta')L(\theta)}{L_{r,b}(\theta)L(\theta')}\right],
\end{aligned}$$

for $\theta \sim \Pi$ and $\theta' \mid \theta \sim K_{r,b}(\theta';\theta)$, so that tuning b to make the M-H acceptance ratio larger will tend to make the symmetric KL between $\Pi_{r,b}$ and Π small. This justifies the approach of using the acceptance ratio to tune b. As the acceptance ratio approaches 1, CDA M-H and CDA Gibbs coincide, and the CDA Gibbs invariant measure is identically Π, but the corresponding Gibbs sampler converges rapidly.

4 Co-Browsing Behavior Application

We apply CDA to an online browsing activity dataset. The dataset contains a two-way table of visit count by users who browsed one of 96 client websites of interest, 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 large commercial interest to find out the high-traffic sites with relatively high co-browsing rates, so that ads can be more effectively placed. For the computational

advertising company, it is also useful to understand the co-browsing behavior and predict the traffic pattern of users. We consider two models for these data.

4.1 Hierarchical Binomial Model for Estimating Co-browsing Rates

We initially focus on one client website and analyze co-browsing rates with the high-traffic sites. With the total visit count N_i available for the ith high-traffic site, the count of co-browsing y_i can be considered as the result of a binomial trial, with y_i extremely small relative to N_i (ratio 0.00011 ± 0.00093), the maximum likelihood estimate y_i/N_i can have poor performance. For example, when $y_i = 0$, estimating the rate as exactly 0 is not ideal. Therefore, it is useful to consider a hierarchical model to allow borrowing of information across high-traffic sites:

$$y_i \sim \text{Binomial}\left(N_i, \frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right), \quad \theta_i \overset{iid}{\sim} \text{No}(\theta_0, \sigma_0^2), \quad i = 1 \dots n$$

$$(\theta_0, \sigma_0^2) \sim \pi(\theta_0, \sigma_0^2)$$

Based on expert opinion in quantitative advertising, we use a weakly informative prior $\theta_0 \sim \text{No}(-12, 49)$ and uniform prior on σ_0^2. Similar to the logistic regression, we calibrate the binomial Polya-Gamma augmentation, leading to the proposal likelihood:

$$L_{r,b}(\theta_i : y_i, N_i, r_i, b_i) = \frac{\exp(\theta_i + b_i)_i^y}{\{1+\exp(\theta_i + b_i)\}^{N_i r_i}}$$

Conditioned on the latent Polya-Gamma latent variable z_i, each proposal θ_i^* can be sampled from:

$$z_i \sim \text{PG}\left((N_i r_i), \theta_i + b_i\right)$$

$$\theta_i^* \sim \text{No}\left(\frac{y_i - r_i N_i/2 - z_i b_i + \theta_0/\sigma_0^2}{z_i + 1/\sigma_0^2}, \frac{1}{z_i + 1/\sigma_0^2}\right),$$

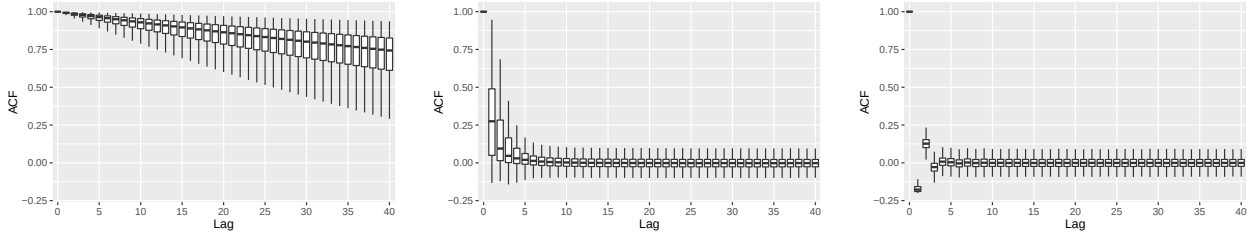
and accepted or rejected using an M-H step. We further require $r_i \geq (y_i - 1)/N_i + \epsilon$ to have a proper $L_{r,b}(\theta_i; y_i, N_i)$ with ϵ a small constant. Similar to logistic regression, the auxiliary parameters are chosen as $r_i = \frac{\exp(\theta_i)}{\{1+\exp(\theta_i)\}^2} / \left(\frac{1}{2|\theta_i + b_i|} \tanh \frac{|\theta_i + b_i|}{2}\right) \vee \left((y_i - 1)/N_i + \epsilon\right)$ and $b_i = \log[\{1+\exp(\theta_i)\}^{1/r_i} - 1] - \theta_i$ during adaptation. Since θ_i's are conditionally independent, the calibrated proposal can be individually accepted with high probability for each i. This leads to a high average acceptance of 0.9, despite the high dimensionality of $59,792$ θ_i's.

After θ_i's are updated, other parameters are sampled from $\theta_0 \sim \text{No}\left((n/\sigma^2 + 1/49)^{-1}(\sum_i \theta_i/\sigma^2 - 12/49), (n/\sigma^2 + 1/49)^{-1}\right)$ and $\sigma_0^2 \sim \text{Inverse-Gamma}(n/2 - 1, \sum_i (\theta_i - \theta_0)^2/2)$.

Figure 4 shows the boxplots of the ACFs for all θ_i's. We compare the result with the original DA (Polson et al., 2013) and Hamiltonian Monte Carlo (HMC) provided by the `STAN` software (Carpenter et al., 2016).

We run DA for $100,000$ steps, HMC for 2000 steps and CDA for $2,000$ steps, so that they have approximately the same effective sample size (calculated with the `CODA` package in `R`). All of the parameters mix poorly in DA; HMC and CDA leads to significant improvement with autocorrelation rapidly decaying to close to zero within 5 lags.

Shown in Table 1, CDA and HMC have very close estimates in posterior means and 95% credible intervals for the parameters; while DA has poor estimates due to critically slow mixing. The difference between HMC and CDA is that, although HMC is slightly more efficient in effective sample size per iteration (T_{eff}/T) for this model, it is much more computationally intensive and generate much less iterations than CDA, within a same budget of computing time. As the result, CDA has the most efficient computing time per effective sample.



(a) ACFs of the rate parameters θ_i using DA.

(b) ACFs of the rate parameters θ_i using CDA.

(c) ACFs of the rate parameters θ_i using HMC.

Figure 4: Boxplots of the ACFs show the mixing of the $59,792$ parameters in the hierarchical binomial model, for the original DA(Polson et al., 2013), CDA and HMC.

	DA	CDA	HMC
$\sum \theta_i/n$	-10.03 (-10.16, -9.87)	-12.05 (-12.09, -12.02)	-12.06 (-12.09, -12.01)
$\sum \theta_i^2/n$	102.25 (98.92, 105.23)	153.04 (152.06, 154.05)	153.17 (152.02, 154.29)
θ_0	-10.03 (-10.17, -9.87)	-12.05 (-12.09,-12.01)	-12.06 (-12.10, -12.01)
σ^2	1.60 (1.36, 1.82)	7.70 (7.49, 7.88)	7.71 (7.51, 7.91)
T_{eff}/T	0.0085 (0.0013 0.0188)	0.5013 (0.1101,1.0084)	0.8404 (0.5149, 1.2470)
Avg Computing Time / T	1.2 sec	1.2 sec	6 sec
Avg Computing Time / T_{eff}	140.4 sec	0.48 sec	1.3 sec

Table 1: Parameter estimates (with 95% credible intervals) and computing speed (ratios among computing time, effective sample sizes T_{eff} and total iterations T) of the DA, CDA and HMC in hierarchical binomial model. CDA provides parameter estimates as accurate as HMC, and is more computationally efficient than HMC.

4.2 Poisson Log-Normal Model for Web Traffic Prediction

The co-browsing on one high-traffic site and one client site is commonly related to the click-through of users from the former to the latter. Therefore, the count of co-browsing is a useful indication of the click-through traffic. For any given client website, predicting the high traffic sites that could generate the most traffic is of high commercial interest. Therefore, we consider a Poisson regression model. We choose the co-browsing

count of one client website as the outcome y_i and the log counts of the other 95 websites as the predictors $x_{ij} = \log(x^*_{ij} + 1)$ for $i = 1 \dots 59,792$ and $j = 1 \dots 95$. A Gaussian random effect is included to account for over-dispersion relative to the Poisson distribution, leading to a Poisson Log-Normal regression model.

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

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 uniform prior.

We first exclude other factors that could contribute to slow mixing. In this case, when β and τ are sampled separately, the random effects $\tau = \{\tau_1, \dots, \tau_n\}$ can cause slow mixing. Instead, we sample β and τ jointly. Using $\tilde{X} = [I_n || X]$ as a $n \times (n+p)$ juxtaposed matrix, and $\eta_i = x_i\beta + \tau_i$ for the linear predictor, the model can be viewed as a linear predictor with $n + p$ coefficients, and $\theta = \{\tau, \beta\}'$ can be sampled jointly in a block. The reason for improved mixing with blocked sampling can be found in Liu (1994a).

We now focus on the mixing behavior due to data augmentation. We first review the the data augmentation for Poisson log-normal model. Zhou et al. (2012) proposed to treat Poisson(η_i) as the limit of the negative binomial NB($\lambda, \frac{\eta_i}{\lambda+\eta_i}$) with $\lambda \to \infty$, and used moderate $\lambda = 1,000$ for approximation. The limit can be simplified as (omitting constant):

$$L(\eta_i; y_i) = \frac{\exp\{y_i\eta_i\}}{\exp\{\exp(\eta_i)\}} = \lim_{\lambda\to\infty} \frac{\exp(y_i\eta_i)}{\{1 + \exp(\eta_i)/\lambda\}^\lambda}. \tag{13}$$

With finite λ approximation, the posterior can be sampled via Polya-Gamma augmented Gibbs sampling:

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

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

However, this approximation-based data augmentation is inherently problematic. For example, setting $\lambda = 1,000$ leads to large approximation error. As in (13), 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 M-H 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 solve this dilemma. 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:

$$z_i \sim \mathrm{PG}(r_i\lambda, \eta_i - \log\lambda + b_i) \quad i = 1 \ldots n$$

$$\theta^* \sim \mathrm{No}\left((\tilde{X}'Z\tilde{X} + \begin{bmatrix} 1/\nu^2 \cdot I_n & 0 \\ 0 & 1/\sigma_\beta^2 \cdot I_p \end{bmatrix})^{-1}\{\tilde{X}'\big(y - r\lambda/2 + Z\log(\lambda - b)\big) + \begin{bmatrix} \tau_0/\nu^2 1_n \\ 0_p \end{bmatrix}\},\right.$$

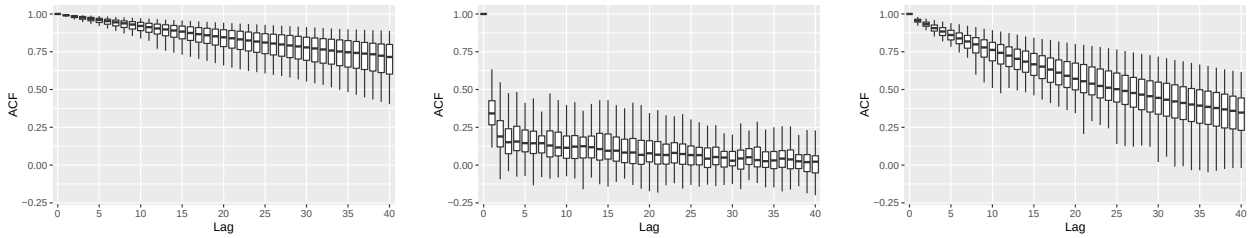
$$\left.(\tilde{X}'Z\tilde{X} + \begin{bmatrix} 1/\nu^2 \cdot I_n & 0 \\ 0 & 1/\sigma_\beta^2 \cdot I_p \end{bmatrix})^{-1}\right)$$

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

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

During the tuning, we set $r_i = 1/c_0\tau_i \exp(\eta_i)/\left(\frac{\lambda}{2|\eta_i + b_i - \log\lambda|}\tanh\frac{|\eta_i + b_i - \log\lambda|}{2}\right) \vee \big((y_i - 1)/\lambda + \epsilon\big)$ based on the Fisher information. We first found the acceptance rate too low with $c_0 = 1$, which indicates an overestimation of the marginal variance; we then reduce $c_0 = 0.1$ to increase the acceptance rate to a satisfactory 0.6. Conditionally on r_i, we used $b_i = \log[\exp\{\exp(\eta_i - \log\lambda - \log r_i)\} - 1] - \eta_i + \log\lambda$. After θ is updated, the other parameters can be sampled via $\tau_0 \sim \mathrm{No}\left((n/\nu^2 + 1/\sigma_\tau^2)^{-1}\sum_i \tau_i/\nu^2, (n/\nu^2 + 1/\sigma_\tau^2)^{-1}\right)$ and $\nu^2 \sim \text{Inverse-Gamma}(n/2 - 1, \sum_i(\tau_i - \tau_0)^2/2)$.

We ran the basic DA with $\lambda = 1,000$ approximation, CDA with $\lambda = 10^9$ and HMC. 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 5 shows the mixings of 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 5: CDA significantly improves the mixing of the parameters in the Poisson log-normal.

To evaluate the prediction 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 log count of $p = 95$ clients $x_i^\dagger = \log(x_i^{\dagger *} + 1)$ to make prediction $\hat{y}_i^\dagger = \mathbb{E}_{\beta,\tau|y,x} y_i^\dagger = \mathbb{E}_{\beta,\tau|y,x} \exp(x_i^\dagger \beta + \tau_i)$ on the client site. The expectation is taken over posterior sample $\beta, \tau \mid y, x$ with training set $\{y, x\}$ 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. 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 Computing Time / T	1.3 sec	1.3 sec	56 sec
Avg Computing 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.

5 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 can slow down mixing when the conditional posterior variance given the augmented data is substantially smaller than the marginal variance. When the sample size is massive, this problem arises when the rates of convergence of the augmented and marginal posterior differ, leading to critical mixing problems. There is a very 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 have limited ability to solve large sample

mixing problems in not addressing the fundamental rate mismatch issue.

To tackle this problem, we propose to calibrate the data augmentation and use a parameter to directly adjust the conditional variance (which is associated with step size). The generated samples are used as a proposal in M-H to obtain the correct posterior. As the original un-calibrated Gibbs sampler is a special case with $r = 1, b = 0$, CDA can be viewed as a generalized class of sampling algorithms with data augmentation. CDA adds a little cost due to the likelihood evaluation, which is often negligible as dominated by the random number generation. In this article, we demonstrate that calibration is generally applicable when $\theta \mid z$ belongs to the location-scale family. We expect it to be extensible to any conditional distribution with a variance or scale.

As both CDA and HMC involve M-H step, we would like to 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.

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.

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *J Stat Softw*, 2016.

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*, ((to appear)), 2015.

Noel Cressie, Anne S Davis, J Leroy Folks, and J Leroy Folks. The moment-generating function and negative integer moments. *The American Statistician*, 35(3):148–150, 1981.

Martin Hairer, Jonathan C Mattingly, and Michael Scheutzow. Asymptotic coupling and a general form of harris theorem with applications to stochastic delay equations. *Probability Theory and Related Fields*, 149 (1-2):223–259, 2011.

Martin Hairer, Andrew M Stuart, Sebastian J Vollmer, et al. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490, 2014.

James E Johndrow, Aaron Smith, Natesh Pillai, and David B Dunson. Inefficiency of data augmentation for large sample imbalanced data. *arXiv preprint arXiv:1605.05798*, 2016.

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.

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

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B Dunson. Robust and scalable bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*, 2014.

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.

Bala Rajaratnam and Doug Sparks. MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. *arXiv preprint arXiv:1508.00947*, 2015.

Gareth O Roberts and Jeffrey S Rosenthal. Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, pages 458–475, 2007.

Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

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

Sanvesh Srivastava, Volkan Cevher, Quoc Tran-Dinh, and David B Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *AISTATS*, 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.

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 Dunson, and Lawrence Carin. Lognormal and gamma mixed negative binomial regression. In *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access, 2012.