Calibrated Data Augmentation for Scalable Markov Chain Monte Carlo

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

January 2, 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 contructing 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:

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

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.

2.1 Initial example: Probit with improper prior

We introduce our calibration approach through a binomial probit model example:

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

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

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

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

where $\mathrm{No}_{[a,b]}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 truncated to the interval $[a, b]$. We propose to make the Gibbs step sizes tunable by introducing an auxiliary parameter r_i multiplying the variance of z_i, while also reducing the bias caused by r_i through adjusting the mean by another auxiliary parameter b_i. These adjustments yield

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

which generalizes $\mathrm{pr}(y_i = 1|\theta, x_i) = \Phi(x_i\theta)$ leading to the modified data augmentation algorithm

$$z_i \mid \theta, x_i, y_i \sim \begin{cases} \mathrm{No}_{[0,\infty)}(x_i\theta + b_i, r_i) & \text{if } y_i = 1 \\ \mathrm{No}_{(-\infty,0]}(x_i\theta + b_i, r_i) & \text{if } y_i = 0 \end{cases} \tag{3}$$

$$\theta \mid z, X \sim \mathrm{No}((X'R^{-1}X)^{-1}X'R^{-1}(z - b), (X'R^{-1}X)^{-1}),$$

where $R = \mathrm{diag}(r_1, \ldots, r_n)$, $b = (b_1, \ldots, b_n)'$, and we let $\pi^*(\theta|y)$ denote the stationary distribution of θ based on repeated samples from (3). This differs fundamentally from the parameter expansion algorithms of Liu and Wu (1999) and Meng and Van Dyk (1999) that rescale θ by $1/\sqrt{r}$, which does not impact the conditional variance of θ and so does not solve the mis-calibration problem.

The update in (3) alters the invariant measure from $\pi(\theta|y)$ to $\pi^*(\theta|y)$, and hence the Gibbs samples for θ will not be exactly from $\pi(\theta|y)$ even after convergence. To adjust for the bias caused by the difference between $\pi(\theta|y)$ and $\pi^*(\theta|y)$, we use (3) as an M-H proposal. Letting $Q(\theta^*; \theta) = \int f(\theta^*|z)\pi(z|\theta)dz$ be the proposal defined by (3) marginalized over z, the proposal is accepted with probability

$$1 \wedge \frac{Q(\theta; \theta^*)\pi(\theta^*) \prod_i L(x_i\theta^*; y_i)}{Q(\theta^*; \theta)\pi(\theta) \prod_i L(x_i\theta; y_i)} = 1 \wedge \frac{\prod_i L_r(x_i\theta; y_i)L(x_i\theta^*; y_i)}{\prod_i L_r(x_i\theta^*; y_i)L(x_i\theta; y_i)}, \tag{4}$$

where $L(\eta_i; y_i) = \Phi(\eta_i)^{y_i}(1-\Phi(\eta_i))^{(1-y_i)}$ and $L_r(\eta_i; y_i) = \Phi(\frac{\eta_i+b_i}{\sqrt{r_i}})^{y_i}(1-\Phi(\frac{\eta_i+b_i}{\sqrt{r_i}}))^{(1-y_i)}$. The second equality holds since $Q(\theta; \theta^*)Q(\theta^*) = Q(\theta; \theta^*)Q(\theta)$ and $Q(\theta) = C\pi(\theta) \prod_i L_r(x_i\theta; y_i)$, which is the posterior density under the altered L_r with C a constant. Setting $r_i = 1$ and $b_i = 0$ leads to acceptance rate of 1, which corresponds to the original Gibbs sampling step.

At the iteration t, when the proposal is accepted $\theta_t = \theta^*$, the covariance:

$$\mathrm{cov}(\theta_t \mid \theta_{t-1}, r, X, z) = (X'R^{-1}X)^{-1} + (X'R^{-1}X)^{-1}X'R^{-1}\,\mathrm{cov}(z - b|R)R^{-1}X(X'R^{-1}X)^{-1},$$

so that the step size is equal to

$$\text{var}(\theta_t \mid \theta_{t-1}, r, X, z) \geq \text{diag}((X'R^{-1}X)^{-1}), \tag{5}$$

with the lower bound a simple function of the r_is. Mis-calibration of the usual data augmentation algorithm, which sets $r_i = 1, b_i = 0$, occurs when the step size in (5) decreases at a faster rate in n and/or p than the posterior $\pi(\theta|y)$ unconditionally on the augmented data z. The key to calibrated data augmentation (CDA) is to choose r, b to minimize or eliminate this mis-calibration while additionally maximizing the M-H acceptance probability, which is similar to minimizing the discrepancy between $\pi^*(\theta|y)$ and $\pi(\theta|y)$. Before describing a general algorithm to choose r, b, we illustrate how CDA can be used to address the problem with DA introduced by Johndrow et al. (2016).

2.2 Imbalanced data intercept only case

In an intercept-only model, the variance is bounded by $\left(\sum_i r_i^{-1}\right)^{-1}$ via (5), which is $1/n$ times the harmonic mean of the r_is. Johndrow et al. (2016) show that when $\sum_i y_i = 1$ and $r_i = 1$, $\text{var}(\theta_t \mid \theta_{t-1})$ 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. To achieve step sizes consistent with the width of the high posterior probability region, we need

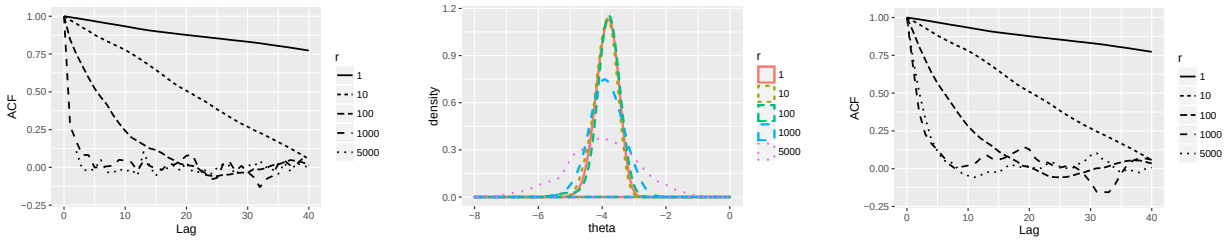
$$\left(\sum_i r_i^{-1}\right)^{-1} \approx (\log n)^{-1},$$

so if $r_i = r$ for all i, $r \approx n/\log n$.

To illustrate the effect of this calibration, consider an intercept only probit model, with $\sum_i y_i = 1$ and $n = 10^4$. Setting $r = 1$ in the proposal corresponds to the original Albert and Chib (1993) Gibbs sampler, which suffers from extremely slow mixing in this case. Letting $r = n/\log n$ to calibrate the sampler, we then choose the b_i's to increase the acceptance rate in the M-H step; as illustration we simply let $b_i = -3.7(\sqrt{r}-1)$ to induce $\text{pr}(y_i = 1) = \Phi(-3.7) = n^{-1}\sum_i y_i = 10^{-4}$ in the proposal distribution.

We ran our CDA Gibbs sampler for these data and different values of r, ranging from $r = 1$ for uncalibrated data augmentation to $r = 5,000$, with $r = 1,000 \approx n/\log n$ corresponding to our recommended default value. Figure 1a plots autocorrelation functions (ACFs) for these different samplers without M-H adjustment. Autocorrelation is very high even at lag 40 for the uncalibrated sampler ($r = 1$), indicating extremely poor mixing. Increasing r leads to dramatic improvements in mixing, but there are no further gains in increasing r from our recommend default value 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.

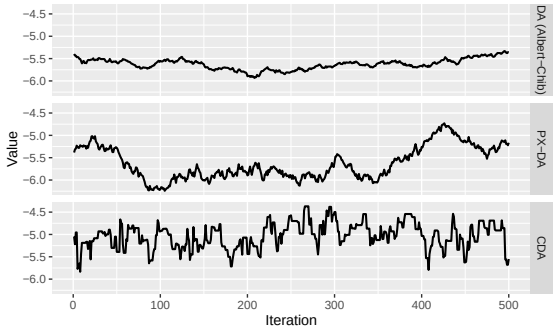
2.3 Choice of calibration parameters

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 this subsection, we propose a simple and efficient algorithm for calculating these parameters relying on Fisher information. We will describe this algorithm in the probit case, but it is straightforward to apply much more broadly. Letting the linear predictor in the probit model be denoted $\eta_i = x_i\theta$, the Fisher information based on the marginal and the conditional posteriors are:

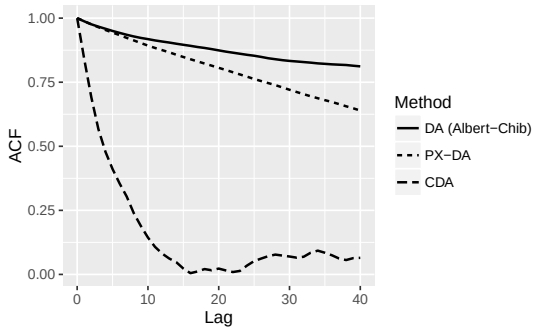
$$X' \operatorname{diag}\left\{\frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))}\right\}X, \qquad X'R^{-1}X$$

respectively, where ϕ is the standard normal density. Therefore, setting $r_i = \frac{\Phi(\eta_i)(1-\Phi(\eta_i))}{\phi(\eta_i)^2}$ completely adjusts for mis-calibration. The r_i's can be calculated using this expression quickly and easily without need to calculate the information matrix. The bias-adjustment parameters b are then chosen to increase the acceptance rate in the M-H step, $1 \wedge \prod_i \frac{L_r(\eta_i;y_i)L(\eta_i^*;y_i)}{L_r(\eta_i^*;y_i)L(\eta_i;y_i)}$. In particular, by setting $b_i = \eta_i(\sqrt{r_i} - 1)$ we ensure that $L_r(\eta_i; y_i) = L(\eta_i; y_i)$.

Since θ and η are not known before sampling, we use a short tuning period to sample them and update r and b at the end of each iteration. After tuning, we stop adaptation and keep r and b fixed. To illustrate, we consider a probit regression with an intercept and two predictors $x_{i,1}, x_{i,2} \sim \text{No}(1, 1)$, 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 variance mismatch problem. In contrast, applying CDA after a tuning period of 100 iterations, 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).

2.4 Logistic regression example: A second calibration approach

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 a different calibration strategy based on inflating the variance of $\pi(z|y)$ targeted towards increasing $\mathbb{E}_z\text{var}(\theta|z, y)$. In developing this second calibration strategy, we focus on the logistic regression model with

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

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

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

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

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

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

Our previous calibration approach would have relied on inflating the variance in the conditional update of θ, but this unfortunately no longer works well due to the occurence of Z in the conditional covariance. Instead, 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

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

$$= \frac{\exp\{(x_i\theta + b_i)y_i\}}{\{1+\exp(x_i\theta+b_i)\}^{r_i}}, \tag{6}$$

and the update rule for the proposal:

$$z_i \sim \mathrm{PG}(r_i, |x_i\theta + b_i|),$$

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

with acceptance probability:

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

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

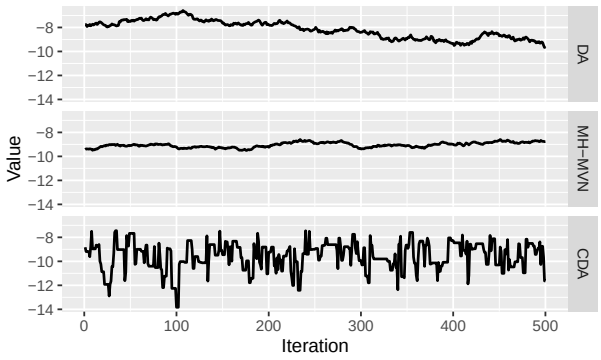
To demonstrate why smaller r_i leads to larger $\mathbb{E}_z(X'ZX)^{-1}$, we compute the first negative moment of the Polya-Gamma distribution. Combining Cressie et al. (1981) and Polson et al. (2013), $\mathbb{E}z_i^{-1} = \int_0^\infty \prod_{k=1}^\infty (1 + d_k^{-1}t)^{-r_i} dt$ with $d_k = 2(k -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

For choosing r during tuning, we compare the two Fisher information matrices based on the marginal and conditional; for the latter, we marginalize out z_i by taking the expectation:

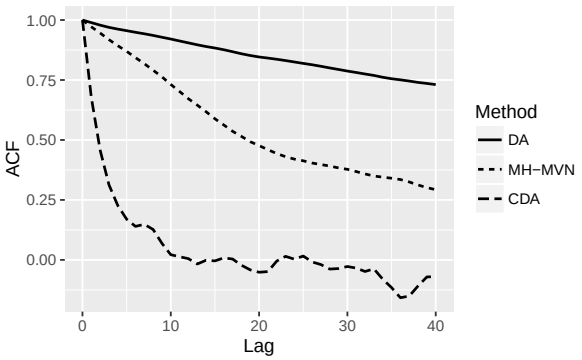
$$X' \operatorname{diag}\left\{\frac{\exp(x_i\theta)}{\{1+\exp(x_i\theta)\}^2}\right\}X, \quad X' \operatorname{diag}\left\{\frac{r_i}{2|x_i\theta+b_i|}\tanh\left(\frac{|x_i\theta+b_i|}{2}\right)\right\}X$$

To correct the difference, we choose r_i to be $\frac{\exp(x_i\theta)}{\{1+\exp(x_i\theta)\}^2} 2|x_i\theta + b_i| / \tanh(\frac{|x_i\theta+b_i|}{2})$. To optimize the acceptance rate, given the value of r_i and $x_i\theta$, setting $\{1 + \exp(x_i\theta)\} = \{1 + \exp(x_i\theta + b_i)\}^{r_i}$ yields $b_i = \log[\{1 + \exp(x_i\theta)\}^{1/r_i} - 1] - x_i\theta$.

As a numerical illustration, we use a two parameter intercept-slope model with $x_1 \sim \mathrm{No}(0, 1)$ and $\theta = (-9, 1)'$. With $n = 10^5$, we obtain rare outcome data with $\sum y_i = 50$. We ran the original DA algorithm (Polson et al., 2013) and an independence chain M-H sampler using a multivariate normal proposal $\theta^*|\theta \sim \mathrm{No}(\theta^*|\theta, \mathcal{I}^{-1}(\theta))$, with $\mathcal{I}(\theta)$ the Fisher information matrix based on the marginal posterior. For CDA we tuned r and b for 100 steps, reaching an acceptance rate of 0.8, and then stopped adaptation and ran an additional burn-in of 100 iterations, with the following 500 samples collected. Shown in Figure 3, both DA and M-H with a normal proposal mix slowly, exhibiting strong autocorrelation even 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).

2.5 General Algorithm

Before providing theoretical support for CDA, we summarize the algorithm in more generality, noting that substantial further generalizations are possible (e.g., to accommodate hierarchical, temporal and spatial dependence). We assume the parameters are multi-dimensional and can be divided into two groups $\{\theta, \tau\}$, with θ and τ sampled separately based on their conditional posterior distributions. For notational ease, we focus on θ and omit the conditioning on τ in the rest of the subsection. Assume θ can be augmented with latent variable z but is susceptible to slow mixing. The augmented likelihood can be expressed as

$$\prod_i L(m_i(\theta); y_i) = \prod_i \int \pi\left(m_i(\theta)|z_i, y_i\right) \pi(z_i|y_i) dz_i, \tag{7}$$

where $m_i(\theta)$ is a continuous and differentiable function $m_i : \mathbb{R}^p \mapsto \mathbb{R}^d$. For example, $m_i(\theta) = x_i\theta$ is the linear predictor in regression. Let the conditional distribution for z and θ be, respectively,

$$z_i \mid \theta, y \sim \pi(z_i; m_i(\theta), y)$$

$$\theta \mid z, y \sim f(\theta; \mu, \Sigma),$$

where $f(\theta; \mu, \Sigma) \propto \pi(\theta) \prod_i \pi\left(m_i(\theta)|z_i, y_i\right)$. To calibrate the variance Σ, we introduce a parameter r_i. When Σ is free from z, we put r_i in each $\pi\left(m_i(\theta)|z_i, y_i\right)$ as in the probit examples above. When Σ involves z, we put r_i in $\pi(z_i|y_i)$ to increase $\mathbb{E}_z\Sigma$ as in the logit examples. Then using another parameter b_i to accommodate the shift in $m_i(\theta)$, we obtain the calibrated data augmentation:

$$\prod_i L_r(m_i(\theta); y_i, r_i, b_i) = \prod_i \int \pi_r\left(m_i(\theta) + b_i|z_i, y_i\right) \pi_r(z_i|y_i) dz_i. \tag{8}$$

With prior $\pi(\theta)$, the proposal can then be sampled from the calibrated distribution:

$$z_i \sim \pi_r(z_i; m_i(\theta) + b_i, y)$$

$$\theta^* \sim f(\theta^*|\mu(r,b),\Sigma(r)).$$

in an M-H step with acceptance probability:

$$1 \wedge \prod_i \frac{L(m_i(\theta^*);y_i)L_r(m_i(\theta);y_i)}{L(m_i(\theta);y_i,r_i,b_i)L_r(m_i(\theta^*);y_i,r_i,b_i)}.$$

In an initial tuning phase, we adaptively update r and b. We choose r at each sampling step to minimize the difference between the Fisher information matrices based on the marginal posterior with the latent data integrated out and the conditional posterior given the latent data. As illustrated in subsection 2.3-2.4 for the two different strategies of augmentation, simple and computationally efficient closed forms are typically available for the r_i's, which avoid potentially expensive calculation of the full information matrix. Then conditional on r_i, we choose b_i to increase the M-H acceptance rate; one choice is the solution to $L_r(m_i(\theta);y_i,r_i,b_i) = L(m_i(\theta);y_i)$.

3 Theory: Mixing Acceleration

We studying the theory behind the acceleration of mixing after calibration focusing on samples collected for fixed values of r, b after the adaptation period. The mixing rate of a Markov chain can be described by the geometric convergence rate. Let $\mathcal{P}(\theta,.)$ be the Markov transition measure, $\pi(.)$ be the target invariant measure and θ be the state in the state space Θ. Starting from the initial state $\theta^{(0)}$, the chain is geometrically ergodic if there exist $M : \Theta \to [0,\infty)$ and $\rho \in [0,1)$ such that $||\mathcal{P}^k(\theta,.) - \pi(.)||_{TV} \leq M(\theta^{(0)})\rho^k$, where $||.||_{TV}$ is the total variation distance $||P_1 - P_2||_{TV} = \sup_{\mathcal{A}\in\mathcal{F}} ||P_1(\mathcal{A}) - P_2(\mathcal{A})||$. As the number of iterations $k \to \infty$, $||\mathcal{P}^k(\theta,.) - \pi(.)||_{TV} \to 0$ leading to convergence to the target. Slow mixing corresponds to ρ being close to 1.

We first utilize another related quantity, the norm of the forward operator $||\mathbf{F}||$, which is defined as $\mathbf{F}s(\theta) = \int \mathcal{P}(\theta,\theta')s(\theta')d\theta' = E\{s(\theta')|\theta\}$. In a Hilbert space $L^2(\pi) = \{s(\theta) : \mathbb{E}s(\theta) = 0, \text{var}\{s(\theta)\} < \infty\}$, the norm is defined as the maximal correlation between two states $||\mathbf{F}|| = \sup_{s(\theta),t(\theta)\in L^2(\pi)} \text{corr}(s(\theta),t(\theta'))$ (Liu, 2008). This norm is related to ρ: when the chain is reversible with detailed balance (e.g. M-H), $\lim_{k\to\infty} ||\mathbf{F}^k||^{1/k} = \rho$; when the chain is non-reversible, $||\mathbf{F}||^2$ is equal to the convergence rate of the reversibilized chain (Fill, 1991).

In each iteration, Gibbs sampling sequentially update all the parameters θ and the latent variable z. We denote the last state as z' and θ' and the new state as z and θ. Then the original DA sample in the order of $\theta' \to z' \to \theta \to z$, where $a \to b \to c$ means that given b, c is conditionally independent of a.

In calculation, omitting the first and last steps do not alter the maximal autocorrelation (Lemma 4 in Liu (1994)), leading to:

$$||\mathbf{F}_{DA}|| = \sup_{s(\theta)\in L^2(\pi)} \frac{\text{var}_{DA}[\mathbb{E}_{DA}\{s(\theta,z)|\theta',z'\}]}{\text{var}_{DA}\{s(\theta,z)\}} = \sup_{s(\theta)\in L^2(\pi)} \frac{\text{var}_{DA}[\mathbb{E}_{DA}\{s(\theta)|z'\}]}{\text{var}_{DA}\{s(\theta)\}}$$

$$= 1 - \inf_{s(\theta)\in L^2(\pi)} \frac{\mathbb{E}_{DA}[\text{var}_{DA}\{s(\theta)|z'\}]}{\text{var}_{DA}\{s(\theta)\}} \tag{9}$$

where the last form without the infimum is known as Bayesian fraction of missing information (Papaspiliopoulos et al., 2007). Slow mixing occurs when $\mathbb{E}_{DA}[\text{var}_{DA}\{s(\theta)|z'\}] \ll \text{var}_{DA}\{s(\theta)\}$.

Calibrated DA samples a little differently: by proposing θ^* in the calibrated sample and using Metropolis-Hastings to accept the new state θ^* with probability $p(\theta^*,\theta')$ or keep the previous state θ, it in fact samples from a two-component mixture:

$$\theta \sim (1-p(\theta^*,\theta'))\delta_{\theta'} + p(\theta^*,\theta')f(\theta^* \mid z').$$

Therefore, the updating sequence is $\theta' \to z'$ and $(\theta', z') \to \theta \to z$. Similarly, omitting the last step of updating z does not alter the maximal autocorrelation, leading to:

$$||\mathbf{F}_{CDA}|| = \sup_{s(\theta)\in L^2(\pi)} \frac{\text{var}_{CDA}[\mathbb{E}_{CDA}\{s(\theta,z)|\theta',z'\}]}{\text{var}_{CDA}\{s(\theta,z)\}} = 1 - \inf_{s(\theta)\in L^2(\pi)} \frac{\mathbb{E}_{CDA}[\text{var}_{CDA}\{s(\theta)|z',\theta'\}]}{\text{var}_{CDA}\{s(\theta)\}} \tag{10}$$

To compare the (9) and (10) directly, we rely on the following lemma.

Lemma 1. *In Metropolis-Hastings step with current state θ' and proposal state θ^* from $f(\theta^*; z')$, if the acceptance probability $p \geq p_0$, the generated state θ satisfies $var_{CDA}\{s(\theta)|z',\theta'\} \geq p_0 \cdot var_{CDA}(s(\theta^*)|z')$.*

Therefore, for a given $s(\theta)$, we can induce an increase in $\mathbb{E}[\text{var}\{s(\theta')|z'\}]$ by γ times over $\mathbb{E}[\text{var}\{s(\theta)|z'\}]$, and obtain the following acceleration:

Theorem 1. *Let $\mathbf{F}_{DA}$ and $\mathbf{F}_{CDA}$ be the forward operators corresponding to the standard DA and the calibrated DA; θ be the random variable from the DA updating rule and θ^* be the one from the CDA proposal. Assume the conditional variance increase in the CDA proposal has $\mathbb{E}[var_{CDA}\{s(\theta^*)|z,y\}] \geq \gamma \cdot \mathbb{E}[var_{DA}\{s(\theta)|z,y\}]$ with the Metropolis-Hastings acceptance probability in (4) greater or equal to $p_0 > 0$. Then if $p_0\gamma \geq 1$,*

$$||\mathbf{F}_{CDA}|| \leq 1 - \gamma p_0 \cdot \inf_{s(\theta)\in L^2(\pi)} \frac{\mathbb{E}_{DA}[var_{DA}\{s(\theta)|z'\}]}{var_{DA}\{s(\theta)\}} \leq ||\mathbf{F}_{DA}||.$$

It is often not tractable to examine every $s(\theta) \in L^2(\pi)$; but in practice, it suffices to check $s(\theta) = \theta_j$ for every element in θ (Yang and Dunson, 2013). Therefore, the ideal r_i would be close to $\frac{\text{var}_{DA}\{\theta\}}{\mathbb{E}_{DA}[\text{var}_{DA}\{\theta|z'\}]}$ so that $||\mathbf{F}_{CDA}|| \approx 1 - p_0$. Both the numerator and the denominator often lack closed-forms for finite sample,

but can be approximately estimated with a term proportional to the Fisher information. As the result, . In our study cases, all of the CDA's have large p_0, which is attributed to the distributional similarity between L_r and L in (4).

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 interests, 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 interests 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 understand the the co-browsing behavior and predict 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 (with 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:

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

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

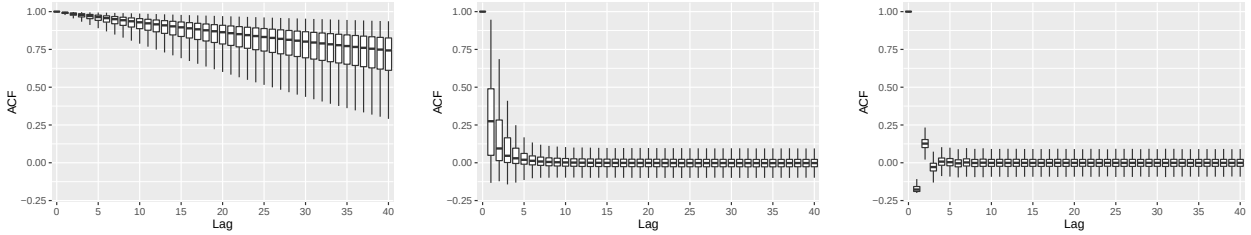
$$z_i \sim \text{PG}\left((N_i r_i), \theta_i + b_i\right)$$

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

and accepted or rejected using an M-H step. Similar to logistic regression, the auxiliary parameters are chosen as $r_i = \frac{\exp(\theta_i)}{\{1+\exp(\theta_i)\}^2} / \left(\frac{1}{2|\theta_i+b_i|}\tanh\frac{|\theta_i+b_i|}{2}\right)$ 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.

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). We run DA for $100,000$ steps and CDA for $2,000$ steps, so that they have approximately the same effective sample size. All of the parameters mix poorly in DA; CDA leads to significant improvement with autocorrelation rapidly decaying to close to zero within 5 lags. Table 1 lists the posterior mean and credible intervals for the parameters, as well as the effective sample size (T_{eff}) per iteration, calculated with the `CODA` package in `R`.

To provide a reference, we ran Hamiltonian Monte Carlo (HMC) provided by the `STAN` software (Carpenter et al., 2016). HMC enjoys very good mixing performance but is computationally intensive due to the numeric leapfrog steps. The parameter estimates from CDA and HMC are remarkably close, while critically slow mixing in the original DA caused poor estimates.



(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 effective sample sizes (T_{eff}) 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 user from the former to the latter. Therefore, the count of the co-browsing is a useful indication of the click-through traffic. For any given client website, being able to predict the high traffic sites that could generate most traffic is certainly of high values. Therefore, we consider a Poisson regression model. We choose the co-browsing count of one client website as the outcome y_i, and the count of other 95 websites as the predictor x^*_{ij} for $i = 1 \dots 59,792$ and $j = 1 \dots 95$. To use them as predictors in Poisson log-normal model, we transform the count onto the log scale with $x_{ij} = \log(x^*_{ij} + 1)$. To allow over-dispersion, we use a random intercept for each i.

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

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.

To focus on data augmentation, we first rule out other factors that can potentially contribute to slow mixing. In this case, random effects τ_i can be problematic for β when they are sampled separately. Therefore, we consider sampling β and τ jointly. Using $\tilde{X} = [I_n||X]$ for the $n \times (n + p)$ juxtaposed projection 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, for which the parameters $\theta = \{\tau, \beta\}'$ can be sampled jointly in a block.

The data augmentation for Poisson log-normal model is less known. Zhou et al. (2012) proposed to utilize 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 method can be simplified as the following:

$$L(x_i\beta, \tau_i; y) = \frac{1}{y!}\frac{\exp(y_i\eta_i\}}{\exp\{\exp(\eta_i)\}} = \frac{1}{y!}\lim_{\lambda\to\infty}\frac{\exp(y_i\eta_i)}{\{1 + \exp(\eta_i)/\lambda\}^\lambda}. \tag{11}$$

With finite λ approximation, it has a Polya-Gamma augmented sampling. Using Gibbs sampler, the approximate posterior can be obtained via:

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

However, this approximate based data augmentation is inherently problematic. For finite λ approximation, setting $1,000$ would cause large approximation error. As in (11), 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, since it would otherwise create extremely large z_i and small conditional covariance for θ.

This is where calibration can 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:

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

and proposal update rule:

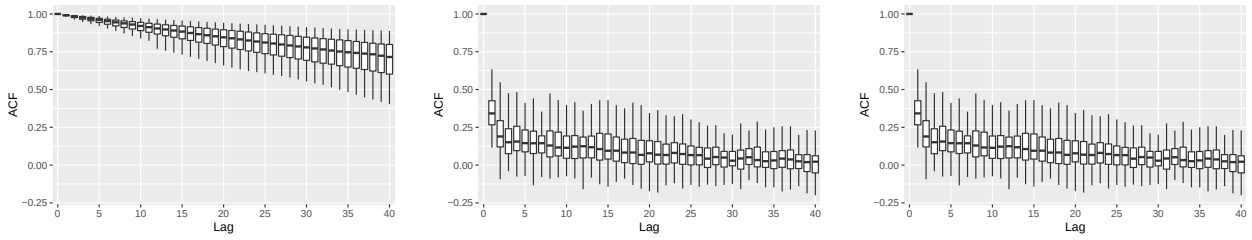
$$\begin{aligned}
&z_i \sim \mathrm{PG}(r_i\lambda, \eta_i - \log\lambda + b_i)\\
&\theta^* \sim \mathrm{No}((\tilde{X}'Z\tilde{X} + \begin{bmatrix} 1/\nu^2\cdot I_n & 0 \\ 0 & 1/\sigma_\beta^2 \cdot I_p \end{bmatrix})^{-1}\{\tilde{X}'\big(y - r\lambda/2 + Z\log(\lambda - b)\big) + \begin{bmatrix} \tau_0/\sigma_\tau^2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

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

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

During the adaptation, we set $r_i = \tau_i \exp(\eta_i)/\left(\frac{\lambda}{2|\eta_i + b_i - \log\lambda|}\tanh\frac{|\eta_i + b_i - \log\lambda|}{2}\right)$ based on the Fisher information and $b_i = \log[\exp\{\exp(\eta_i - \log\lambda - \log r_i)\} - 1] - \eta_i + \log\lambda$.

To compare, we ran approximate DA Gibbs sampler with $\lambda = 1,000$, CDA with M-H proposal and HMC. We ran approximate DA for $100,000$ steps, CDA for $20,000$ steps and HMC for $2,000$ steps so that they have approximately the same effective sample size. For CDA, we used the first $1,000$ steps for adaptating r and b, and obtained an acceptance rate of 0.6. Figure 5 shows the mixings of DA and CDA. Surprisingly, even with small λ, in approximate DA, all of the parameters still mix poorly; CDA substantially improves the mixing.



(a) Autocorrelation of the parameters from DA.

(b) Autocorrelation of the parameters from CDA.

(c) Autocorrelation of the parameters from HMC. **needs new figure**

Figure 5: CDA significantly improves the mixing of the parameters in the Poisson log-normal.

Table 2 compares all three algorithms. Like the binomial case, HMC is the most efficient in mixing and provides the highest effective sample size per iteration, but is quite costly in computation. CDA provides similar mixing performance and is significantly more efficient. Overall, CDA is clearly the best choice to generate the highest effective sample size within the same computing time.

To evaluate the prediction performance, we use another dataset $\{y_{i,new}, x_{i,new}\}$ collected during a different time period for validation. We use the posterior mean $\{\hat{\tau}, \hat{\beta}\}$ for the prediction $\hat{y}_{i,new} = \exp(x_{i,new}\hat{\theta} + \hat{\tau}_i)$; cross-validation root-mean-squared error between $y_{i,new}$ and $\hat{y}_{i,new}$ is computed. CDA and HMC perform quite well and the validation error that is 4 times lower than DA.

	DA	CDA	HMC
$\sum \theta_j/95$	0.072 (0.071, 0.075)	-0.041 (-0.042, -0.038)	-0.040 (-0.042, -0.037)
$\sum \theta_j^2/95$	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	8.18
T_{eff}/T	0.0037 (0.0011 0.0096)	0.3348 (0.0279, 0.699)	???
Avg Computing Time / T	1.3 sec	1.3 sec	??? sec
Avg Computing Time / T_{eff}	346.4 sec	11.5 sec	??? sec

Table 2: Parameter estimates, prediction error and computing speed of the DA, CDA and HMC in Poisson regression model. Compared with HMC, CDA shows similar performance in both parameter estimation and prediction, but is about 5 times faster.

5 Discussion

Data augmentation is a useful technique to sample posterior from the closed-form conditional. It has been realized that this practice could severely stall the mixing, due to the gap between the conditional variance with the augmented data and the marginal one. With data size increases and become complex, it is common for the conditional distribution of the parameter to deviate from the area that has reasonable mixing performance. As we show in the previous examples, this quickly leads to an un-manageable increase in the computational time and poor estimation.

To solve this problem, we propose a general class of method to calibrate the variance conditional on the

latent variable. With a mechanism to adjust the step size, the transition in each iteration is corrected onto the same order of the marginal variance. The generated samples are used as proposal in the Metropolis-Hastings for exact posterior. In this article, we demonstrate that this strategy is applicable when $\theta \mid z$ belongs to the location-scale family. We expect that it can be extensible to any distribution with a variance / scale, possibly with a different bias-reducing mechanism.

There is some similarity between CDA and HMC. Both algorithms excel in seeking proposal with high acceptance rate. The difference is that quite often, the Hamiltonian lacks closed-form solution, and requires multiple steps numeric evaluations of the dynamics for one proposal; whereas CDA only needs one step. Therefore, when the data augmentation exists, CDA is always more efficient in computation.

In this article, we insist on obtaining the exact posterior, to provide an analysis on the mixing property. Without the Metropolis-Hastings step, the sampling strategy in calibrated data augmentation can be used alone to generate approximate posterior. This can be useful when the evaluation of the marginal likelihood is costly.

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.

James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process. *The annals of applied probability*, pages 62–87, 1991.

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, 1994.

Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

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

Yun Yang and David B Dunson. Sequential markov chain monte carlo. *arXiv preprint arXiv:1308.3861*, 2013.

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.

6 Appendix

6.1 Proofs

6.1.1 Lemma 1

As the M-H step in CDA is equivalent to sampling from the mixture that:

$$(1-p)\delta_{\theta'} + pf_{CDA}(\theta^*; z')$$

where p is the acceptance probability in (4) and f_{CDA} is the calibrated proposal distribution. Its conditional variance is:

$$\begin{aligned}
\text{var}_{CDA}\{s(\theta)|z',\theta'\} =&(1-p)s(\theta')^2 + p\mathbb{E}_{CDA}\{s(\theta^*)^2|z'\} - [(1-p)s(\theta') + p\mathbb{E}_{CDA}\{s(\theta^*)|z'\}]^2\\
=&(1-p)[s(\theta')^2 - (1-p)s(\theta')^2 - 2ps(\theta')\mathbb{E}_{CDA}\{s(\theta^*)|z'\} + p\mathbb{E}_f\{s(\theta^*)|z'\}^2]\\
&+ p[\mathbb{E}_{CDA}\{s(\theta^*)^2|z'\} - \mathbb{E}_{CDA}\{s(\theta^*)|z'\}^2]\\
=&(1-p)p[s(\theta') - \mathbb{E}_{CDA}\{s(\theta^*)|z'\}]^2 + p\cdot\text{var}_{CDA}(s(\theta^*)|z')\\
\geq& p\cdot\text{var}_{CDA}(s(\theta^*)|z')\\
\geq& p_0\cdot\text{var}_{CDA}(s(\theta^*)|z')
\end{aligned}$$

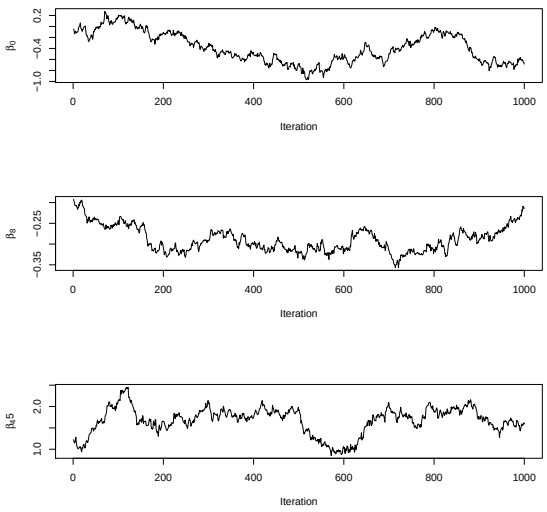
6.1.2 Theorem 1

With Lemma 1,

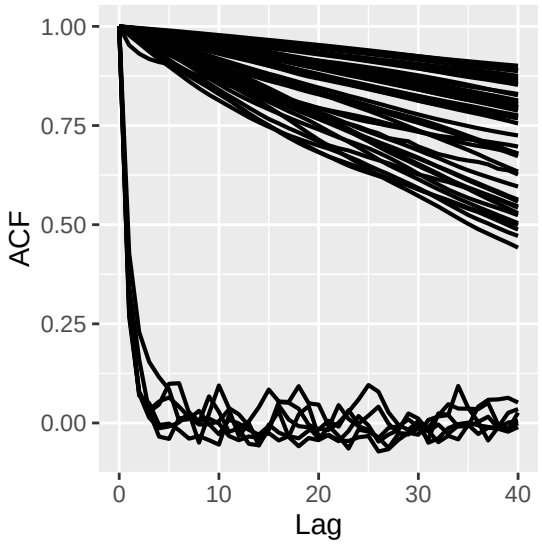
$$\begin{aligned}
\mathbb{E}[\text{var}_{CDA}\{s(\theta)|z',\theta'\}] \geq& p_0\cdot\mathbb{E}[\text{var}_{CDA}(s(\theta^*)|z')]\\
\geq& p_0\gamma\cdot\mathbb{E}[\text{var}_{DA}(s(\theta^*)|z')].
\end{aligned}$$

Since the marginal variances are the same for two algorithms $\text{var}_{DA}\{s(\theta)\} = \text{var}_{CDA}\{s(\theta)\}$. When $p_0\gamma \geq 1$, rearranging terms and taking supremum on both sides complete the proof.

6.2 Mixing of Zero-inflated Poisson without Calibration



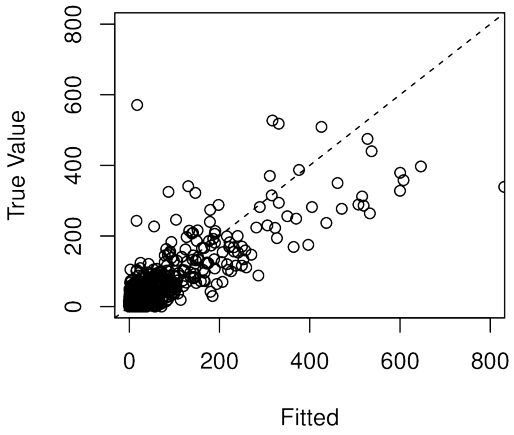
(a) Trace plots of three parameters from DA ZIP model



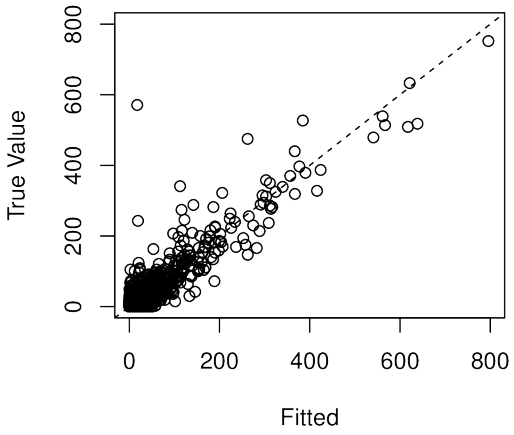
(b) Autocorrelation of all the 96 θ's from DA ZIP model.

Figure 6: The hierarchy in the zero-inflated Poisson model does NOT help reduce the autocorrelation.

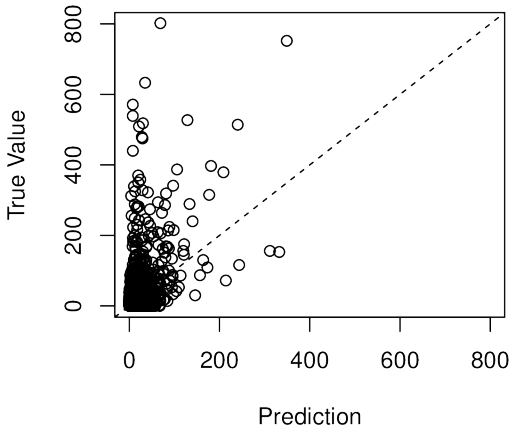
6.3 Goodness-of-Fit and Cross-Validation for Poisson Regression



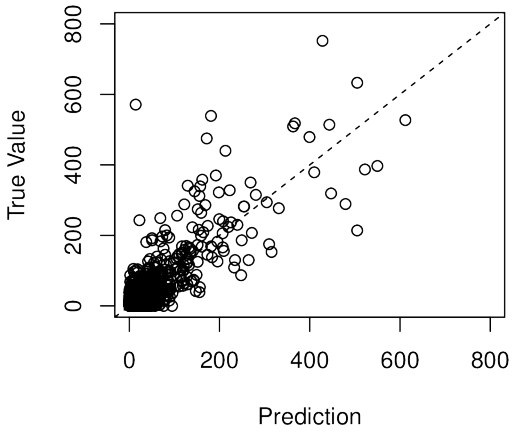
(a) Fitted vs true values using DA



(b) Fitted vs true values using CDA



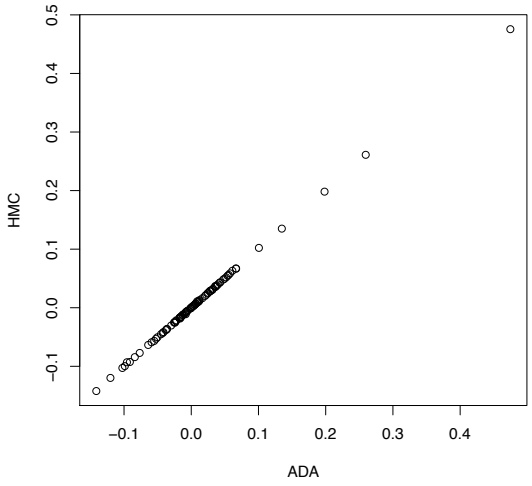
(c) Prediction vs true values using DA



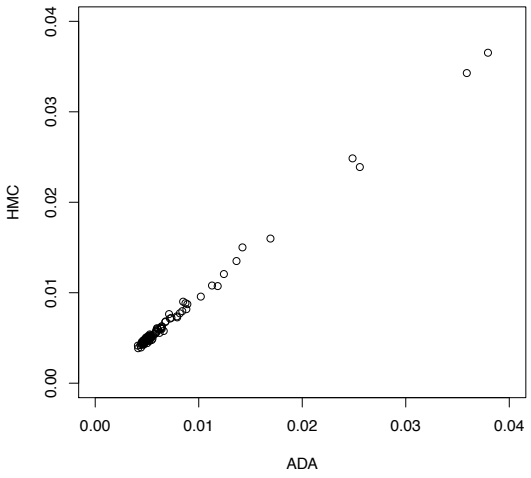
(d) Prediction vs true values using CDA

Figure 7: The posterior estimates produced by CDA is better fitted to the data and have more accurate prediction than DA.

6.4 Comparing posterior samples of CDA with HMC



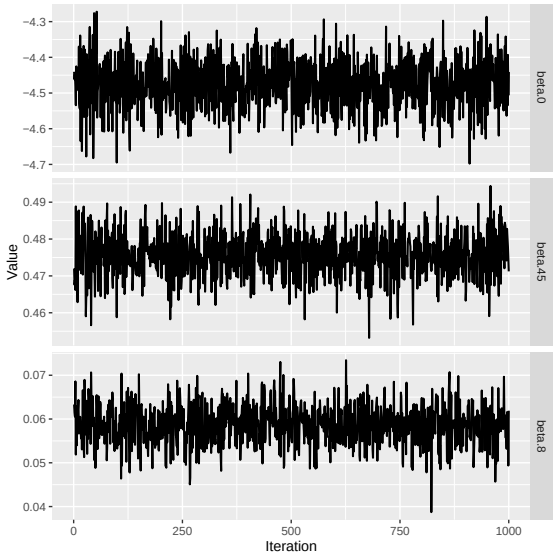
(a) Comparing posterior means for $\theta_1, \ldots \theta_{95}$ from the HMC and CDA. The RMSE between the two is 0.0007.



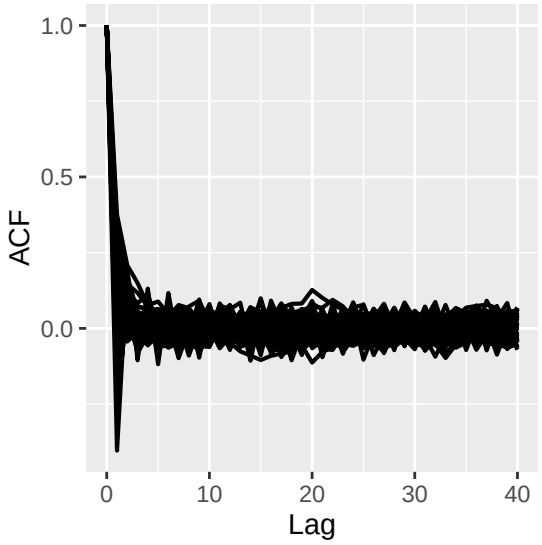
(b) Comparing posterior standard deviation for $\theta_1, \ldots \theta_{95}$ from the HMC and CDA. The RMSE between the two is 0.0004.

Figure 8: The results from CDA and HMC agree very well.

6.5 Mixing of HMC



(a) Traceplots



(b) Autocorrelation

Figure 9: The posterior estimates produced by HMC.