Calibrated Data Augmentation for Scalable Markov Chain Monte Carlo

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

December 7, 2016

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 large samples. In this article, we propose a family of calibrated data augmentation algorithms, which adjust for this discrepancy by inflating Gibbs step sizes with an auxiliary parameter. The bias introduced by the scale parameter can be eliminated through a Metropolis-Hastings step. 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 a general 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

Our method is developed primarily in the context of data augmentation Gibbs samplers, which are based on the integral $\pi(\theta|y) = \int f(\theta|z,y)\pi(z|y)dz$, where θ are model parameters, y denotes observed data, and z denotes latent data included for computational purposes. Data augmentation Gibbs samplers alternate between sampling the latent data z from their conditional posterior distribution given θ and y, and sampling parameters θ given the latent z and observed data 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 $R = \mathrm{diag}(r_1, \ldots, r_n)$, $b = (b_1, \ldots, b_n)'$. 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

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

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 estimate 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

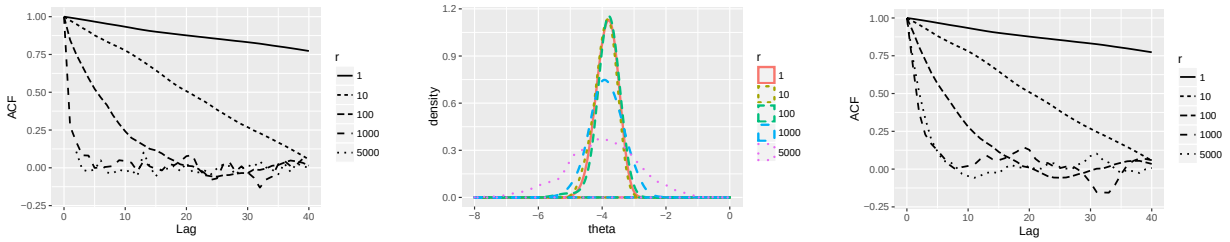
$$\left(\sum_i r_i^{-1}\right)^{-1} \approx (\log n)^{-1},$$

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. Later we will propose a method for estimating the r_is and b_is.

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 on differences in the estimates; it is apparent that the density estimates change with r. Therefore, for the target distribution with $r = 1$, the M-H adjustment that proposes form one with increased r, small increase in r (e.g. $r = 10,100$) retains a close to 1 acceptance rate but little improvement

in mixing, large r results in smaller acceptance rate (0.6 for $r = 1,000$, 0.2 for $r = 5,000$) but significant improvement in mixing (Figure 1c).



(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 r and b using Asymptotic Approximation

As illustrated in the previous subsection, efficiency of CDA is dependent on a good choice of the auxiliary parameters $r = (r_1, \ldots, r_n)$ and $b = (b_1, \ldots, b_n)$. In general, it is not straightforward to analyze the exact difference between the conditional and marginal variances, however, one can compute the two Fisher information matrices based on conditional and marginal posteriors, and use their inverse as the asymptotic approximates.

Continuing on the probit regression example with the linear predictor $\eta_i = x_i\theta$, the Fisher information based on the marginal and the conditional posteriors given z_i are:

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

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 calibrates the difference between the two.

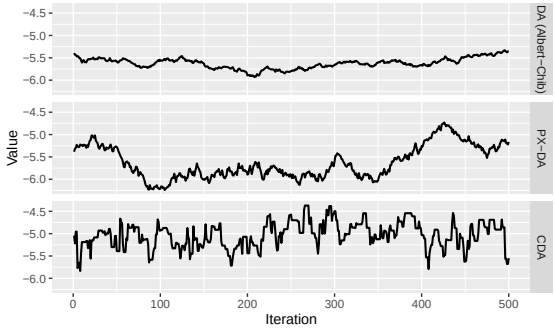
Then we change b to increase the acceptance rate. Since the acceptance probability at each step is $1 \wedge \prod_i \frac{L_r(\eta_i;y_i)L(\eta_i^*;y_i)}{L_r(\eta_i^*;y_i)L(;y_i)}$. Given the values of r_i and η_i, one choice for b_i is the analytical solution that makes $L_r(\eta_i; y_i) = L(\eta_i; y_i)$. In the probit example, this is $b_i = \eta_i(\sqrt{r_i} - 1)$.

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. The acceptance rate can be monitored as the number of acceptance divided by the total tuning iterations. After the rate reaches reaches satisfactory level (e.g. 0.3), we stop the adaption and keep r and b fixed. Then the algorithm goes through burn-in and posterior sampling like ordinary MCMC.

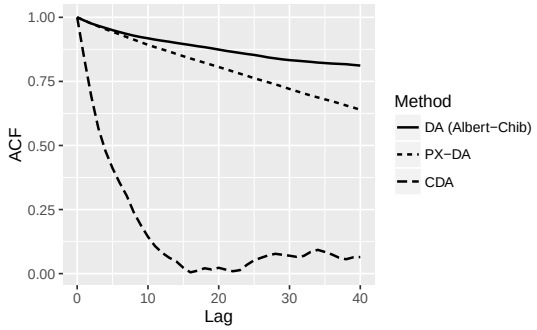
To illustrate, we consider a probit regression with an intercept and two predictors $x_{i,1}, x_{i,2} \sim \text{No}(1, 1)$,

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). To compare with an early method, we test 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.

We started CDA with the adaptive algorithm for 100 iterations, obtaining an acceptance rate of 0.43, then ran 100 steps as burn-in and collected the posterior sample for 500 steps. The calibrated algorithm solves the mixing problem (Figure 2a and 2b).



(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 about 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 Calibrating Random Conditional Variance via Latent Variable

It is easy to obtain good acceptance rate in CDA, because the increase in conditional variance does not affect the existence of marginal closed-form, and the calibrated marginal is close to the original one.

In the previous examples, as $\text{var}(\theta|z, y)$ does not involve the random latent variable z, one can directly use r on it to increase the conditional variance. On the other hand, when $\text{var}(\theta|z, y)$ depends on the random value of z, including r in it is effective and could cause difficulty in integration $\pi(\theta|y) = \int f(\theta|z, y)\pi(z|y)dz$. Therefore, in this section we show a different strategy that modifies $\pi(z|y)$ to increase $\mathbb{E}_z\text{var}(\theta|z, y)$.

To illustrate, consider the logistic regression:

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

and improper prior $\pi(\theta) = 1$. The Polya-Gamma data augmentation has the update rule (Polson et al., 2013):

$$z_i \sim \text{PG}(1, |x_i\theta|),$$

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

where $Z=\mathrm{diag}(z_1,\ldots,z_n)$. This algorithm is derived based on $L(y_i \mid x_i\theta)=\int \exp\{x_i\theta(y_i-1/2)\}\exp(-\frac{z_i(x_i\theta)^2}{2})\,\mathrm{PG}(z_i \mid 1,0)dz_i$. As described above, multiplying r_i on z_i is ineffective and makes the integration intractable.

Instead, observing the value of z can be influenced by the first parameter in $\mathrm{PG}(a_1,a_2)$, $\mathbb{E}z_i=\frac{a_1}{2a_2}\tanh(\frac{a_2}{2})$, we replace $\mathrm{PG}(z_i \mid 1,0)$ with $\mathrm{PG}(z_i \mid r_i,0)$. Smaller r_i can lead to larger $\mathbb{E}_z\mathrm{var}(\theta|z,y)$.

It can be verified the integration is still tractable, applying the mean adjusting term b_i on $x_i\theta$, leading to:

$$\begin{aligned}L_r(x_i\theta;y_i)=&\int_0^\infty \exp\{(x_i\theta+b_i)(y_i-r_i/2)\}\exp(-\frac{z_i(x_i\theta+b_i)^2}{2})\,\mathrm{PG}(z_i \mid r_i,0)dz_i\\ =&\frac{\exp\{(x_i\theta+b_i)y_i\}}{\{1+\exp(x_i\theta+b_i)\}^{r_i}},\end{aligned} \tag{6}$$

and the update rule for the proposal:

$$\begin{aligned}&z_i\sim \mathrm{PG}(r_i,|x_i\theta+b_i|),\\ &\theta^*\sim \mathrm{No}\left((X'ZX)^{-1}X'(y-r_i/2-Zb),(X'ZX)^{-1}\right),\end{aligned}$$

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

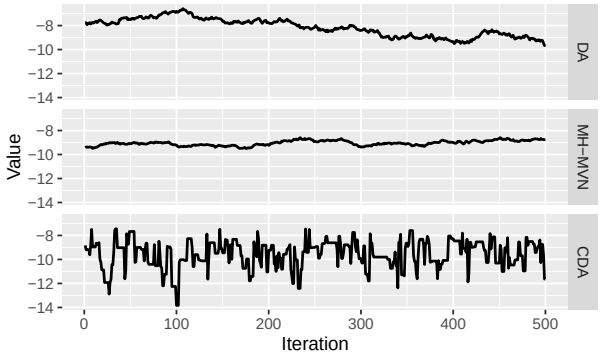
To see exactly why the smaller r_i leads to larger $\mathbb{E}_z(X'ZX)^{-1}$, we compute the first negative moment of the Polya-Gamma distribution. Combining Cressie et al. (1981) and Polson et al. (2013), $\mathbb{E}z_i^{-1}=\int_0^\infty\prod_{k=1}^\infty(1+d_k^{-1}t)^{-r_i}dt$ with $d_k=2(k-\frac{1}{2})^2\pi^2+\frac{(x_i\theta+b_i)^2}{2}$.

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

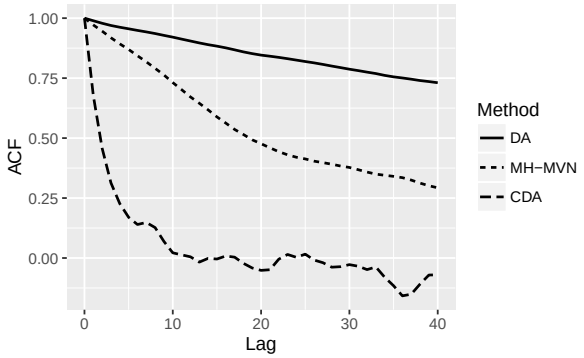
$$X'\,\mathrm{diag}\{\frac{\exp(x_i\theta)}{\{1+\exp(x_i\theta)\}^2}\}X,\quad X'\,\mathrm{diag}\{\frac{r_i}{2|x_i\theta+b_i|}\tanh(\frac{|x_i\theta+b_i|}{2})\}X$$

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 an analytical choice for b_i as $\log[\{1+\exp(x_i\theta)\}^{1/r_i}-1]-x_i\theta$ at the end of each tuning iteration.

As a numerical illustration, we use a two parameter intercept-slope model with $x_1 \sim \text{No}(0, 1)$ and $\theta = \{-9, 1\}$. With $n = 10^5$, it leads to rare positive outcome $\sum y_i = 50$. To compare, we ran the original DA (Polson et al., 2013) and the M-H with simple multivariate normal proposal $\theta^*|\theta \sim \text{No}(\theta^*|\theta, \mathcal{I}^{-1}(\theta))$, with $\mathcal{I}(\theta)$ being the Fisher information matrix based on the marginal posterior. For CDA we tuned the r and b for 100 steps, reaching an acceptance rate of 0.8; then stop adaption and run 100 and 500 steps for burn-in and the posterior collecting. Shown in Figure 3, both DA and M-H with normal proposal mix slowly, exhibiting strong autocorrelation even after 40 lags; whereas CDA substantially accelerates the mixing.



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



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

Figure 3: Panel (a) demonstrates in traceplot and panel (b) in autocorrelation about the substantial improvement in CDA by correcting the variance mis-match 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 proceeding into theory, we summarize the general algorithm for CDA. We assume the parameters are multi-dimensional and can be divided into two groups $\{\theta, \tau\}$. We sample the parameters in $\eta \mid \theta, y$ and the ones in $\theta \mid \tau, y$ alternatively. For notational ease, we now focus on θ and omit the conditioning on τ in the the rest of section.

Assume θ can be augmented with latent variable z but is susceptible to slow mixing issue. The total augmented likelihood is the product:

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

where $m_i(\theta)$ is a continuous and differentiable function $m_i : \mathbb{R}^p \mapsto \mathbb{R}^d$, conditioning on which, each integral on the right hand side is independent. For example, $m_i(\theta) = x_i\theta$ is the linear predictor in regression; C_i is the constant free from θ. Let the the conditional distribution for z and θ be:

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

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

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 $\pi\left(m_i(\theta)|z_i, y_i\right)$. When Σ involves z, we put r_i in $\pi(z_i|y_i)$ to increase $\mathbb{E}_z\Sigma$. Then using another parameter b_i to accomodate 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:

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

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

in a M-H step with accepting probability:

$$1 \wedge \prod_i \frac{L(m_i(\theta^*); y_i) L_r(m_i(\theta); y_i)}{L(m_i(\theta); y_i, r_i, b_i) L_r(m_i(\theta^*); y_i, r_i, b_i)}. \tag{9}$$

We initially run the algorithm by adaptively estimating r and b. To choose the value for r, we first compute the Fisher information matrix based on the marginal posterior $\pi(\theta|y) = C_L \pi(\theta) \prod_i L(m_i(\theta); y_i)$:

$$\begin{aligned}
\mathcal{I}(\theta) =& \left[\mathbb{E}_y \left(\frac{\partial \log \pi(\theta) \prod_i L(m_i(\theta); y_i)}{\partial \theta_{j_1}}\right) \left(\frac{\partial \log \pi(\theta) \prod_i L(m_i(\theta); y_i)}{\partial \theta_{j_2}}\right)\right]_{j_1, j_2} \\
=& J' \operatorname{Diag} \left[\mathbb{E}_y \left(\frac{\partial \log L(m_i(\theta); y_i)}{\partial m_i(\theta)}\right) \left(\frac{\partial \log L(m_i(\theta); y_i)}{\partial m_i(\theta)}\right)\right]_i J + \left[\left(\frac{\partial \log \pi(\theta)}{\partial \theta_{j_1}}\right) \left(\frac{\partial \log \pi(\theta)}{\partial \theta_{j_2}}\right)\right]_{j_1, j_2},
\end{aligned}$$

where $J = [\frac{\partial m_i(\theta)}{\partial \theta_j}]_{i,j}$, and Diag is the block diagonal matrix (or simple diagonal matrix when all $m_i(\theta)$'s are scalars). Then the Fisher information matrix based on the conditional $f(\theta|\mu(r, b), \Sigma(r)) = C_f \pi(\theta) \prod_i \pi_r\left(m_i(\theta)|z_i, y_i\right)$, marginalized over z:

$$\begin{aligned}
\mathbb{E}_z \mathcal{I}(\theta|z) =& \left[\mathbb{E}_y \left(\frac{\partial \log \pi(\theta) \prod_i \pi_r\left(m_i(\theta)|z_i, y_i\right)}{\partial \theta_{j_1}}\right) \left(\frac{\partial \pi(\theta) \prod_i \pi_r\left(m_i(\theta)|z_i, y_i\right)}{\partial \theta_{j_2}}\right)\right]_{j_1, j_2} \\
=& J' \operatorname{Diag} \left[\mathbb{E}_z \mathbb{E}_y \left(\frac{\partial \log \pi_r\left(m_i(\theta)|z_i, y_i\right)}{\partial m_i(\theta)}\right) \left(\frac{\partial \log \pi_r\left(m_i(\theta)|z_i, y_i\right)}{\partial m_i(\theta)}\right)\right]_i J + \left[\left(\frac{\partial \log \pi(\theta)}{\partial \theta_{j_1}}\right) \left(\frac{\partial \log \pi(\theta)}{\partial \theta_{j_2}}\right)\right]_{j_1, j_2},
\end{aligned}$$

Since r_i is in $\pi_r\left(m_i(\theta)|z_i, y_i\right)$, we use it to minimize the difference between $\mathbb{E}_y \left(\frac{\partial \log L(m_i(\theta); y_i)}{\partial m_i(\theta)}\right) \left(\frac{\partial \log L(m_i(\theta); y_i)}{\partial m_i(\theta)}\right)$ and $\mathbb{E}_z \mathbb{E}_y \left(\frac{\partial \log \pi_r(m_i(\theta)|z_i, y_i)}{\partial m_i(\theta)}\right) \left(\frac{\partial \log \pi_r(m_i(\theta)|z_i, y_i)}{\partial m_i(\theta)}\right)$ for all i. Due to the conditional independence induced by

$m_i(\theta)$, computing r_i is often straightforward; and often their difference can be completely removed given θ. Then conditional on r_i, we choose b_i to optimize the acceptance rate, one choice is the solution to $L_r(m_i(\theta); y_i, r_i, b_i) = L(m_i(\theta); y_i)$. After the acceptance rate reaches satisfactory value, we stop tuning r and b, and continue in burn-in and posterior sample collecting.

3 Theory: Mixing Acceleration

We now study the theory behind acceleration of the mixing after the calibration. Since the posteriors are collected after we stop adaptation, we focus on the period that the parameters r and b are fixed.

The mixing of Markov chain can be described by the geometric convergence rate. Let $\mathcal{P}(\theta, .)$ be the the Markov transition measure and $\pi(.)$ 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. The slow mixing is attributed to ρ being too close to 1. As shown by (Scott et al., 2016), the ρ approaches 1 as n increases, leading to a complete break-down of algorithm.

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, the original DA samples in the sequence of $\theta' \to z' \to \theta \to z$, where $a \to b \to c$ means that given b, c is conditionally independent of a. Omitting y for simpler notation, by Lemma 4 in Liu (1994):

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

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

in which, slow mixing occurs when $\mathbb{E}_{DA}[\text{var}_{DA}\{s(\theta)|z'\}] \ll \text{var}_{DA}\{s(\theta)\}$.

The calibrated DA samples a little differently: by proposing θ^* in the calibrated sample and use Metropolis-Hastings to accept the new state θ^* or keep the previous state θ, it in fact samples in the sequence of $(\theta', z') \to \theta \to z$, similarly we obtain:

$$||\mathbf{F}_{CDA}|| = \sup_{s(\theta)\in L^2(\pi)} \frac{\text{var}_{CDA}[\mathbb{E}_{CDA}\{s(\theta,z)|\theta^{'},z'\}]}{\text{var}_{CDA}\{s(\theta,z)\}} = 1 - \inf_{s(\theta)\in L^2(\pi)} \frac{\mathbb{E}_{CDA}[\text{var}_{CDA}\{s(\theta)|z',\theta'\}]}{\text{var}_{CDA}\{s(\theta)\}} \quad (11)$$

To compare the (10) and (11) 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, 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 (9) greater or equal to $p_0 > 0$. Then if $p_0\gamma \geq 1$,*

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

Since r_i allows us to freely increase γ, we choose it to be close to $\frac{\text{var}_{DA}\{s(\theta)\}}{\mathbb{E}_{DA}[\text{var}_{DA}\{s(\theta)|z'\}]}$. Both the numerator and the denominator often lack closed-forms for finite sample, but can be approximately estimated with Fisher information. As the result, $||\mathbf{F}_{CDA}|| \approx 1 - p_0$. Then it suffices to verify $p_0 > 0$. In our study cases, all of the CDA's have large p_0, which is attributed to the similarity between L_r and L in (9). For example, in the logit CDA. The acceptance probability is $1 \wedge \prod_i \frac{\{1+\exp(x_i\theta)\}\{1+\exp(x_i\theta^*+b_i)\}^{r_i}}{\{1+\exp(x_i\theta^*)\}\{1+\exp(x_i\theta+b_i)\}^{r_i}}$. When $x_i\theta$ is negative (corresponding to large variance gap that causes slow mixing), the adapted $b_i = \log\{1/r_i + \mathcal{O}(\frac{\exp(x_i\theta)}{r_i})\} \approx -\log r_i$. Then $\{1+\exp(x_i\theta - \log r_i)\}^{r_i} = 1 + \exp(x_i\theta) + \mathcal{O}(\frac{\exp(2x_i\theta)}{r_i})$.

4 Real Data Application: Poisson Regression for Online Advertisement Tracking

We now apply CDA to a real data application in online advertisement tracking. The advertisement is displayed on $n = 59,792$ originating websites, pointing to 96 different targets. The count of click-throughs is recorded for each combination. The counts contain many zeros (95.5%), as not all 96 advertisements are shown on all the websites. For commercial interests, it is useful to predict the traffic of the new advertisements using the existing ones. Therefore, we use the data of 95 advertisements as predictors x_i and the one left

as the outcome y_i for a count regression. We use training data collected from a two-week period, and a validation data set collected during another two-week window.

One common practice to handle the large proportion of zeros is to use zero-inflated Poisson. However, for predictive modeling, this is suboptimal as it would require another set of coefficients to predict the latent binary event, e.g. $y_i \sim p\left(g(x_i\eta)\right)\delta_0 + \{1-p\left(g(x_i\eta)\right)\}Poisson\{\exp(x_i\theta)\}$, while does not solve the slow mixing issue (see appendix). Instead, it is rather useful to consider a simpler model $y_i \sim Poisson\{\exp(\theta_0 + \sum_j x_{i,j}\theta_j)\}$ with a quite negative intercept θ_0.

It is known that the posterior sampling for Poisson is hindered by slow mixing, which is especially worse with large amount of zeros. The traditional M-H lacks a good strategy to propose multidimensional variables ($p = 96$ in this case). There is a Gibbs sampling strategy, first discovered by (Zhou et al., 2012) with negative binomial approximation. We further simplify and present the algorithm.

The Poisson density can be viewed as a limit:

$$L(x_i\theta; y) = \frac{\exp(y_i x_i\theta)}{\exp\{\exp(x_i\theta)\}y!} = \lim_{\lambda\to\infty} \frac{\exp(y_i x_i\theta)}{\{1+\exp(x_i\theta)/\lambda\}^{\lambda} y!}.$$

With large but finite λ (e.g. $10,000$), one can sample from the approximate posterior:

$$z_i \sim \mathrm{PG}\left(\lambda, x_i\theta - \log\lambda\right)$$

$$\theta \sim \mathrm{No}\left(\left[(X'ZX)^{-1}X'\big(y - \lambda/2 + z\log\lambda\big), (X'ZX)^{-1}\right)\right.$$

The problem with this DA is that as λ increases, the accuracy is improved but the large z quickly reduces the conditional variance for θ, creating mixing bottleneck. It inevitably becomes a dilemma to trade between accuracy or mixing rate in choosing λ.

As λ control the magnitude of the latent z, the calibration is straightforward by replacing λ with small r_i and $-\log\lambda$ with b_i, giving the calibrated likelihood:

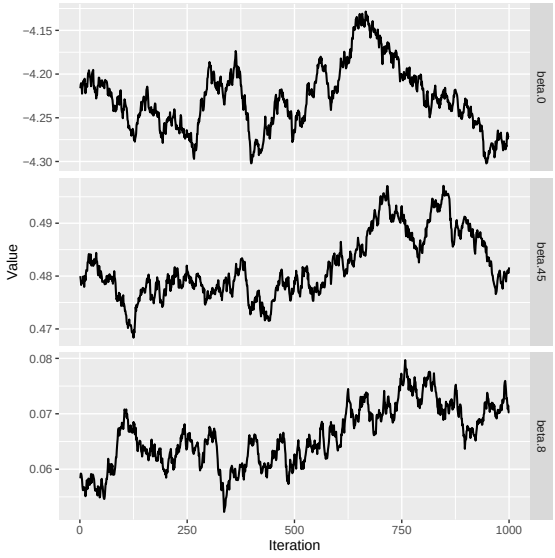
$$L_r(x_i\theta; y) = \frac{\exp\{y_i(x_i\theta + b_i)\}}{\{1+\exp(x_i\theta + b_i)\}^{r_i}},$$

which generates the same sampling algorithm as the CDA in logit regression, except that $r_i > y_i$ to ensure the posterior properiety.

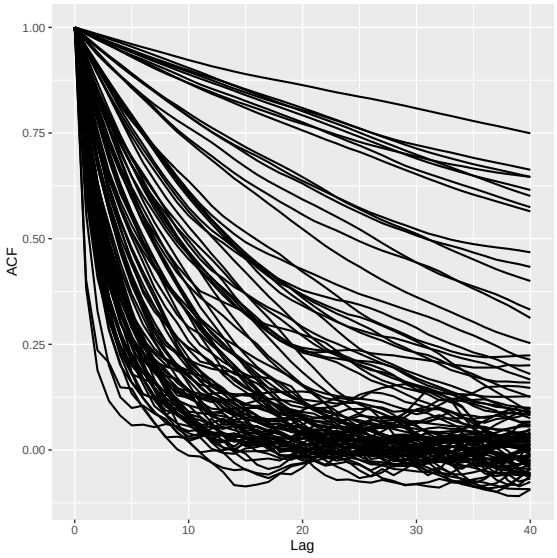
We run the approximate DA with large $\lambda = 10,000$ and the exact CDA for posterior computation. For CDA, we adapt for 100 iterations and reaches an acceptance rate of 0.6. We then run each algorithm for $4,000$ steps and use the last $1,000$ as the posterior sample.

The mixing of DA and CDA is compared in traceplots and autocorrelation plots in Figure 4. DA shows slow mixing for several parameters (Figure4b), including the important intercept estimate θ_0 (first plot

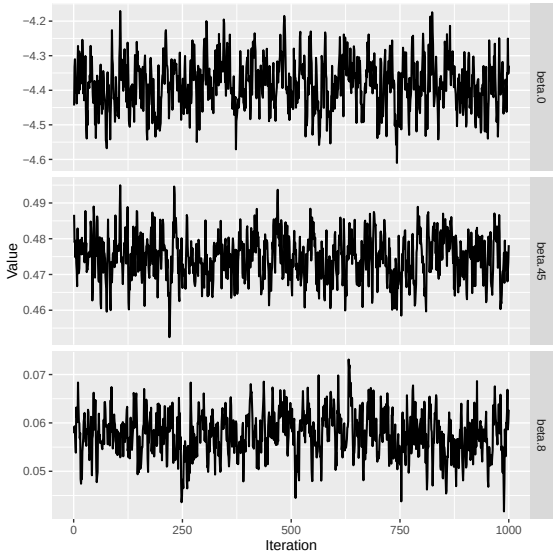
in Figure4a). After calibration, the slow mixing problem is solved: *all* of 96 parameters show very low autocorrelation.



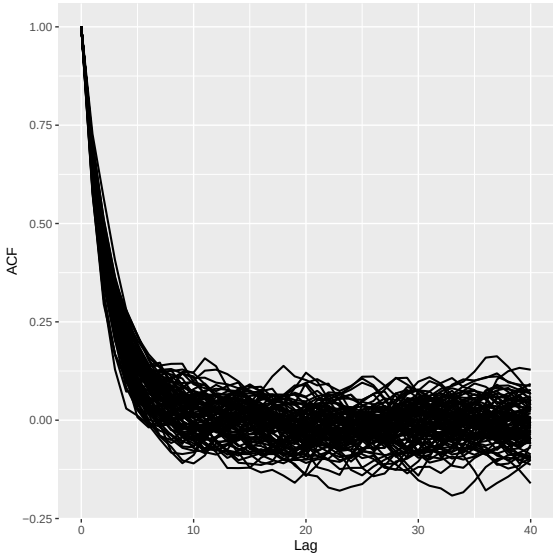
(a) Trace plots of three parameters from DA.



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



(c) Trace plots of three parameters from CDA.



(d) Autocorrelation of all the 96 θ's from CDA.

Figure 4: Panels (c) and (d) show significant improvement of the mixing in Poisson data augmentation. Panel (d) show the CDA reduces the autocorrelation for all of the parameters.

To empirically evaluate the accuracy of the estimates, we also run Hamiltonian Monte Carlo (HMC) as the reference. HMC is known for its good mixing properties, despite of its costly evaluation. We list the parameter estimates and fit statistics in Table 1. For simplicity, we include the posterior mean and standard deviation for the intercept θ_0 and the norm of the coefficients $\sum_{j=0}^{95} |\theta_j|$. For goodness-of-fit, we compute root-mean-squared error $RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \mu_i)^2/n}$ and the deviance $D = 2\sum_{i=1}^{n}\{y_i \log(y_i/\mu_i) - (y_i - \mu_i)\}$,

with $\mu_i = \exp(x_i\hat{\theta})$ and $\hat{\theta}$ as the posterior mean. For prediction performance, we use the testing dataset and $\hat{y}_{i,new} = \exp(x_{i,new}\hat{\theta})$ as the estimator. We evaluate the cross-validation RMSE between $y_{i,new}$ and $\hat{y}_{i,new}$.

As expected, the estimates for θ_0 from all models are quite negative. However, the DA severely underestimates the variance of the intercept. The coefficient norm also differs greatly from CDA and HMC. Obviously, the poor mixing causes the Markov chain in DA to be trapped in a suboptimal state. After calibration, CDA performs exceptionally well in fit statistics and the validation error that is almost 4 times lower. The results of CDA and HMC are nearly identical (see appendix for more comparison).

Lastly, it is worth to compare the computing time needed for the three methods. The CDA operates at almost the same cost as the original DA in each iteration. After the extra adaptation period (about 100 iterations), CDA quickly converges to the target in the first few iterations; whereas DA seems to be stuck even after $2,000$ iterations. HMC is computationally intensive as it requires evaluation of the gradient and multiple Hamiltonian steps in generating a proposal, therefore the speed is about 10 times slower; and the adaptation in both step size and step number is also time-consuming. In conclusion, CDA is most computationally efficient.

	DA	CDA	HMC
θ_0	-4.21 (0.042)	-4.38 (0.075)	-4.47 (0.071)
$\sum_{j=0}^{95} \lvert\theta_j\rvert$	12.24 (0.10)	8.58 (0.11)	8.68 (0.11)
RMSE	32.86	5.06	4.88
D	182127.7	107076.9	106791.3
CV-RMSE	32.01	8.61	8.28
Steps for Adaptation & Burn-in	2000	100	500
Computing Speed (per 1,000 steps)	25 mins	26 mins	300 mins

Table 1: Performance of DA, CDA and HMC in Poisson log-linear regression with online advertisement tracking data. Posterior estimates for the intercept and the norm of the coefficients are shown. The CDA shows much improved fit statistics such as root-mean-squared error (RMSE) and deviance (D). In cross-validation (CV-RMSE), the CDA outperforms DA in nearly 4 times lower in error. The CDA converges much more rapidly than DA. CDA agrees with the HMC very well but takes significantly less time and the adaptation is simpler.

5 Discussion

In posterior sampling, when the parameters lack closed-form in the marginal distribution, data augmentation is a useful technique. 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. On the other hand, it is not feasible to directly use the marginals with Metropolis-Hastings, when the parameters are in multi-dimensions, since it is challenging finding a proposal with the right correlation structure.

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 machineary.

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

In this article, we insist on obtaining the exact posterior, to provide a rigorous 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.

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.

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.

Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

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.

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 (9) and f_{CDA} is the calibrated proposal distribution. Its conditional variance is:

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

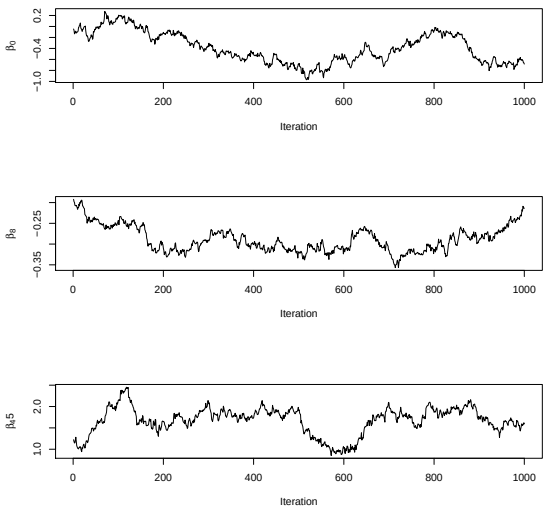
6.1.2 Theorem 1

With Lemma 1,

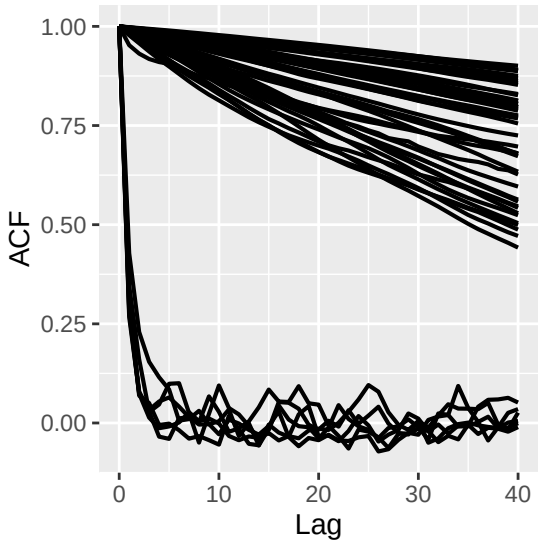
$$\begin{aligned}
\mathbb{E}[\mathrm{var}_{CDA}\{s(\theta)|z',\theta'\}] \geq& p_0\cdot\mathbb{E}[\mathrm{var}_{CDA}(s(\theta^*)|z')] \\
\geq& p_0\gamma\cdot\mathbb{E}[\mathrm{var}_{DA}(s(\theta^*)|z')].
\end{aligned}$$

Since the marginal variances are the same for two algorithms $\mathrm{var}_{DA}\{s(\theta)\} = \mathrm{var}_{CDA}\{s(\theta)\}$. 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



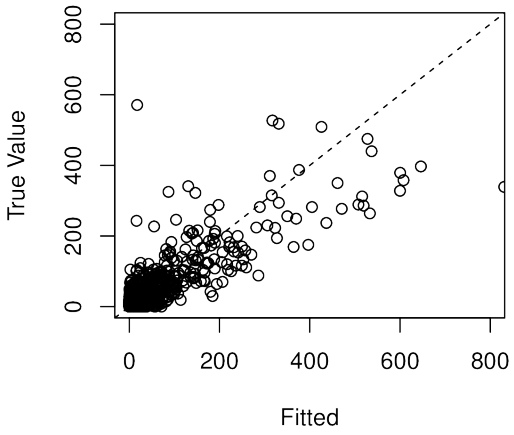
(a) Trace plots of three parameters from DA ZIP model



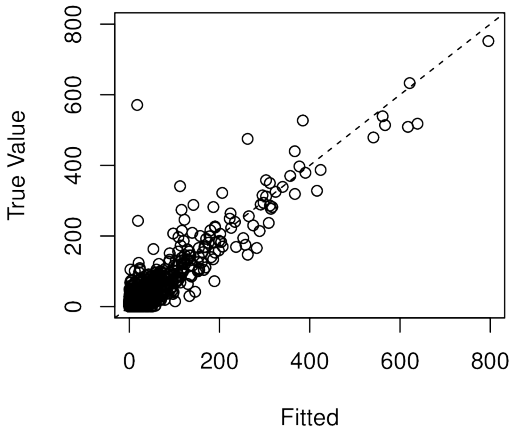
(b) Autocorrelation of all the 96 θ's from DA ZIP model.

Figure 5: The hierarchy in the zero-inflated Poisson model does NOT help reduce the autocorrelation.

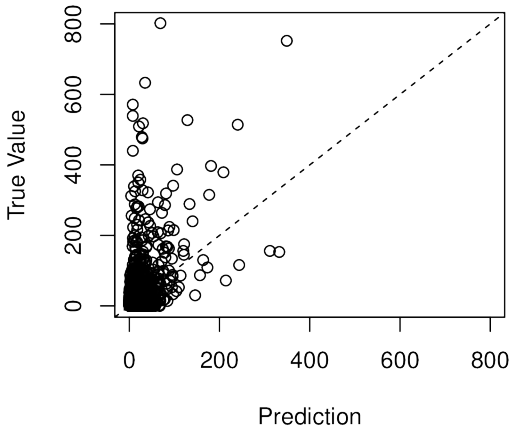
6.3 Goodness-of-Fit and Cross-Validation for Poisson Regression



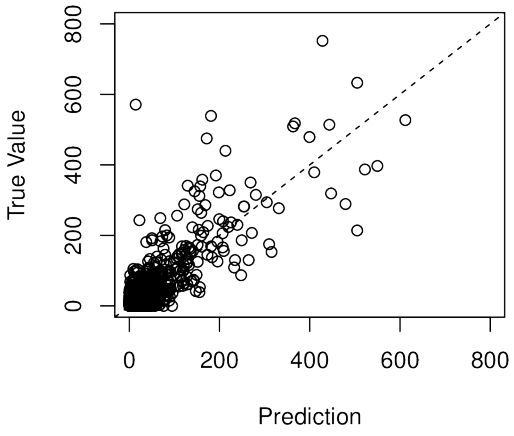
(a) Fitted vs true values using DA



(b) Fitted vs true values using CDA



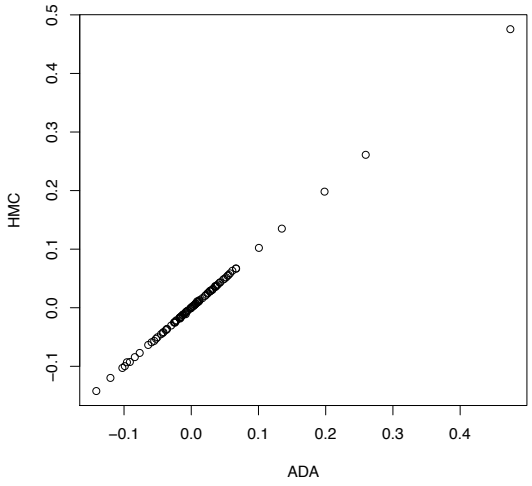
(c) Prediction vs true values using DA



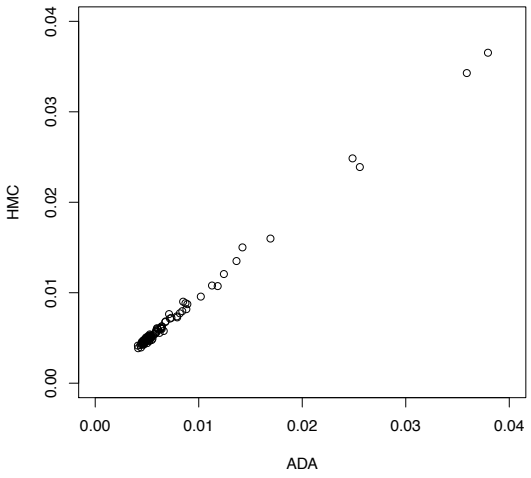
(d) Prediction vs true values using CDA

Figure 6: 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



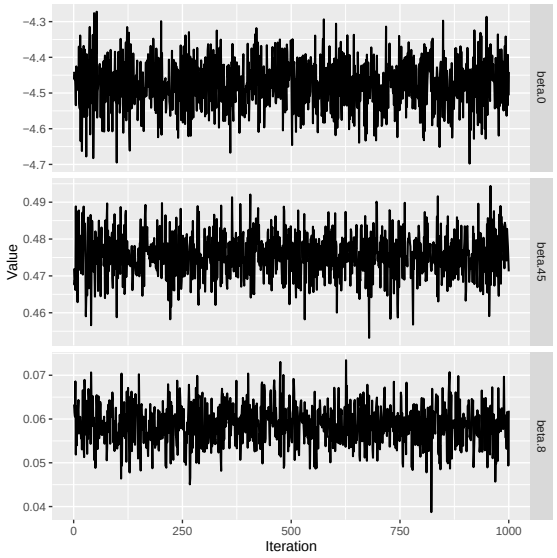
(a) Comparing posterior means for $\theta_1, \ldots \theta_{95}$ from the HMC and CDA. The RMSE between the two is 0.0007.



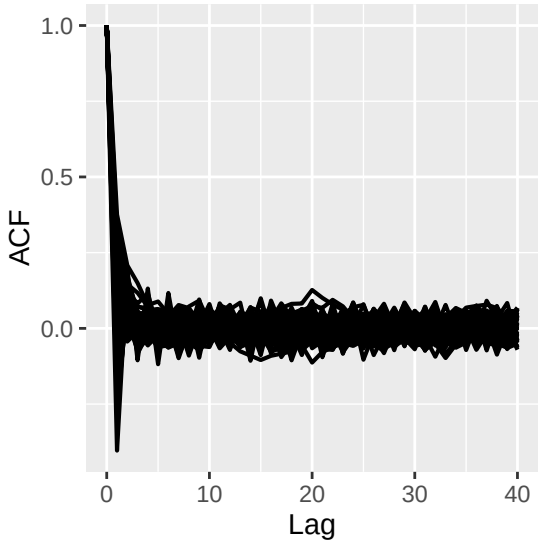
(b) Comparing posterior standard deviation for $\theta_1, \ldots \theta_{95}$ from the HMC and CDA. The RMSE between the two is 0.0004.

Figure 7: The results from CDA and HMC agree very well.

6.5 Mixing of HMC



(a) Traceplots



(b) Autocorrelation

Figure 8: The posterior estimates produced by HMC.