Calibrated Data Augmentation for Scalable Markov Chain Monte Carlo

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

November 30, 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; Approximate Markov chain Monte Carlo; 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 treatment of the behavior of MCMC for large n and/or p in the theory literature is arguably somewhat limited. Historically, many authors have focused on studying the mixing properties of MCMC by showing a general ergodicity condition, such as the geometric ergodicity condition of (cite Meyn and Tweedie, Roberts and Rosenthal). This generally yields bounds on the convergence rate and spectral gap of the Markov chain under consideration, 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 (cite Hairer, Mattingly and friends papers). Recent work verifies the existence of MCMC algorithms for computation in differential equation models with dimension-independent spectral gap (Hairer, Stuart, Vollmer). In this example, the algorithm under consideration is a 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 the popular data augmentation algorithms for posterior computation in probit, (Albert and Chib, 1993) and logistic models (Polson et al., 2013), showing that the algorithm fails 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

[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.

directly control the step sizes of the Gibbs sampler could be devised, it would be possible to make the mixing properties of the sampler insensitive to sample size by scaling the step sizes appropriately. This is similar to the conclusion of (HSV), 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 alters the invariant measure, one can use the Gibbs step as an efficient M-H proposal, thereby recovering the correct invariant, or view the resulting algorithm as a perturbation of the original Markov chain. We use bias correction through a second set of working parameters.

2 Calibrated Data Augmentation

The method is developed primarily in the context of Data augmentation Gibbs samples, which are based on the integral $\pi(\theta|y) = \int \pi(\theta|z,y)\pi(z|y)dz$ and take the general form

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

where f belongs to the location-scale family such as Gaussian. This allows conditional updates for the frequently high-dimensional parameter θ. As such, data augmentation is an important computational strategy when the difficulty of tuning Metropolis-Hastings algorithms is prohibitive due to dimension or complexity of the target distribution. An important class of data augmentation algorithms are those for generalized linear models (GLMs) with $\mathbb{E}[y_i \mid x_i, \theta] = g^{-1}(x_i\theta)$ and conditionally Gaussian prior on θ. We focus in particular on poisson log-linear, binomial logistic, and binomial probit as motivating examples.

The calibration of step sizes is best illustrated by an example. Consider the binomial probit, with sampling model

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

and improper prior $\pi(\beta) = 1$. The basic data augmentation algorithm (tanner and wong, albert and chib, etc) has the update rule

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

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

where $\text{No}_{[a,b]}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 truncated to the interval $[a, b]$. To make the Gibbs step sizes tunable, we introduce an auxiliary parameter r_i as the conditional variance of z_i. To reduce the bias caused by r_i, we adjust the mean by another auxiliary parameter b_i, giving

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi r_i}} \exp\left(-\frac{(z_i - x_i\beta - b_i)^2}{2r_i^2}\right) dz_i = \Phi(\frac{x_i\beta + b_i}{\sqrt{r_i}}), \tag{2}$$

yielding the update rule

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

$$\beta^* \mid z, X \sim \text{No}((X'R^{-1}X)^{-1}X'R^{-1}(z-b), (X'R^{-1}X)^{-1}),$$

where $R = \text{diag}(r_1, \ldots, r_n)$, $b = (b_1, \ldots, b_n)'$. Note that the parameter expansion algorithms of Liu and Wu (1999) and Meng and Van Dyk (1999) rescale β by $1/\sqrt{r}$, which does not change in the conditional variance for β. Our formulation allows a direct control of the step sizes through the value of r.

The update in (3) alters the invariant measure; thus, we use it as a M-H proposal. Letting $Q(\beta^*; \beta)$ be the proposal defined by (3), the proposal is accepted with probability

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

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

When the proposal is accepted $\beta_t = \beta^*$, the covariance:

$$\text{cov}(\beta_t \mid \beta_{t-1}, r, X, z) = (X'R^{-1}X)^{-1} + (X'R^{-1}X)^{-1}X'R^{-1}\,\text{cov}(z - b|R)R^{-1}X(X'R^{-1}X)^{-1}$$

so that the step size:

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

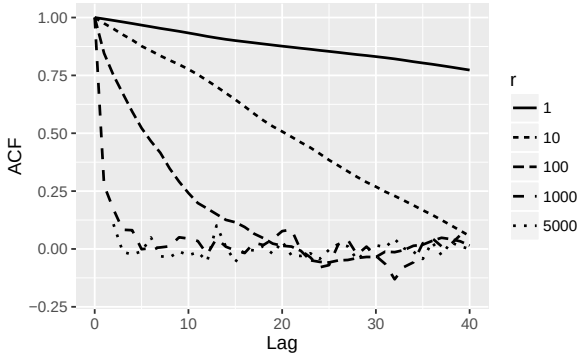
is lower bounded by linear combination of the r_i. In an intercept-only model, the variance is bounded by $\left(\sum_i r_i^{-1}\right)^{-1}$ via (5), $1/n$ times the harmonic mean of the r_i. Johndrow et al. (2016) show that when $\sum_i y_i = 1$ and $r_i = 1$, $\text{var}(\beta_t \mid \beta_{t-1})$ 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. Thus, to achieve step sizes consistent with the width of the space, we need

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

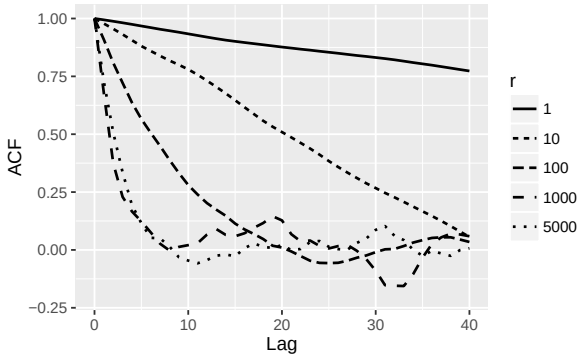
so if $r_i = r$ for all i

$$\frac{n}{r} \approx \log n \Rightarrow r \approx \frac{n}{\log n}.$$

To illustrate the effect of this calibration, consider an intercept only probit model, with $\sum_i y_i = 1$ and $n = 10^4$. We set $r_i = r$ for all i. To increase the acceptance rate of the M-H step, we use a rough estimate of $b_i = -3.7(\sqrt{r} - 1)$ to have the effective probability near $\Phi(-3.7) = 10^{-4}$ in the proposal distribution (later we will show the adaptation for b_i without using prior knowledge). Setting $r = 1$ in the proposal corresponds to the original Albert and Chib (1993) Gibbs sampler, which suffers from extremely slow mixing. As shown in Figure 1a, increasing the conditional variance by r leads to substantial decrease of the autocorrelation of β^* in the proposal distribution; after the M-H step at each iteration, the proposal is accepted as β into the Markov chain (Figure 1b) according to (9). This M-H step guarantees that the increase in the step size in the conditional variance does not exceed the real gap in the marginal. Consistent with the previous calculation, $r = 1000 \approx \frac{n}{\log n}$ leads to the near-optimal adjustment of the step size.



(a) Autocorrelation function (ACF) illustrating the effects of the variance increase of r on improving the mixing of the proposal distribution β^*.



(b) ACF illustrating the mixing of the posterior sample β after using Metropolis-Hastings step to accept the proposal from the calibrated distribution.

Figure 1: Panel (a) demonstrates the adjustment of the step size via increasing the conditional variance in the proposal β^* and panel (b) shows its effect after the proposal from the calibrated distribution is accepted as β into the Markov chain by the M-H criterion.

2.1 Adaptation of r and b

The chosen value for r and b are important to ensure good calibration and sufficiently large acceptance rate. Intercept-only probit is a special case where it is possible to estimate the high posterior density region, hence select b and r before sampling. To apply CDA more broadly, we need to be able to handle cases where this is not possible. In this section, we develop a strategy for adaptation of r and b. After the acceptance rate reaches a pre-set threshold, we stop adapting and collect the posterior sample.

In the adapting period, we start from an moderately large value for all r_i. And r_i is bounded in the region of $[1, \kappa]$, where the lower bound 1 corresponds to no adjustment and κ is a maximal value set to prevent numeric error. As the lower bound of the acceptance probability in (9) is the product of $\alpha_i = \frac{L_r(x_i\beta; y_i)L(x_i\beta^*; y_i)}{L_r(x_i\beta^*; y_i)L(x_i\beta; y_i)}$ over all i. When $\alpha_i < 1$, we decrease r_i to reduce the step size; when $\alpha_i > 1$, we increase

r_i to raise the step size. As an empirical value, we multiply r_i by $\sqrt{\alpha}_i$ at the end of each iteration. In the meantime, we set $b_i = x_i\beta(\sqrt{r_i} - 1)$ as it minimizes the difference between $\Phi(\frac{x_i\beta + b_i}{\sqrt{r_i}})$ and $\Phi(x_i\beta)$.

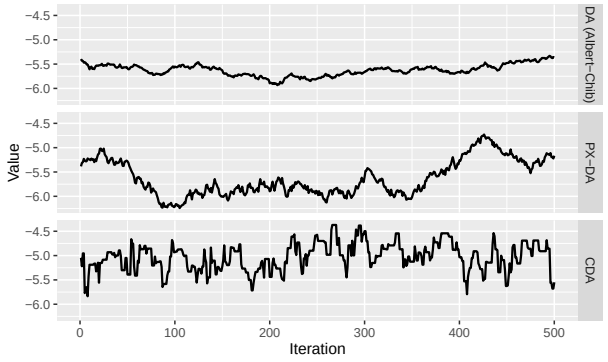
The acceptance rate is calculated as the total count of acceptance divided by the adapting iterations taken thus far. After it reaches satisfactory level (e.g. 0.3), we stop the adaption and keep r and b fixed. The posterior is collected as ordinary Metropolis-Hastings sampling.

To illustrate the case where r and b are difficult to pre-set, we consider a probit regression with an intercept and two predictors $x_{i,1}, x_{i,2} \sim \text{No}(1, 1)$, with $\beta = \{-5, 1, -1\}'$, generating 20 positive outcomes among $n = 10,000$.

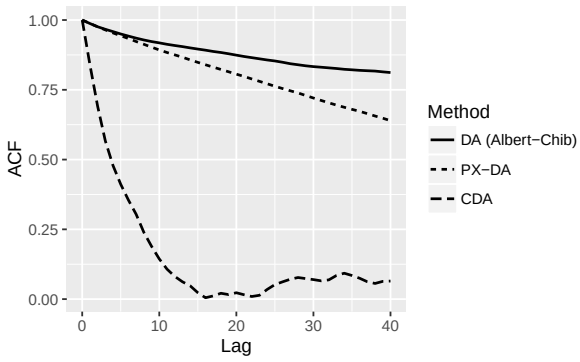
In this testing case, the Albert and Chib (1993) DA algorithm suffers from extremely slow mixing (Figure 2a and 2b). To compare with an early method, we tested the parameter expansion algorithm (PX-DA) proposed by Liu and Wu (1999). PX-DA only mildly reduces the correlation, as it does not solve the variance mismatch problem.

To calibrate, we apply CDA with initial value of 200 for all r_i and use the adaptive algorithm for 100 iterations, obtaining an acceptance rate of 0.36. We use the fixed r and b to collect the posterior sample. The calibrated algorithm leads to significant improvement of the mixing.

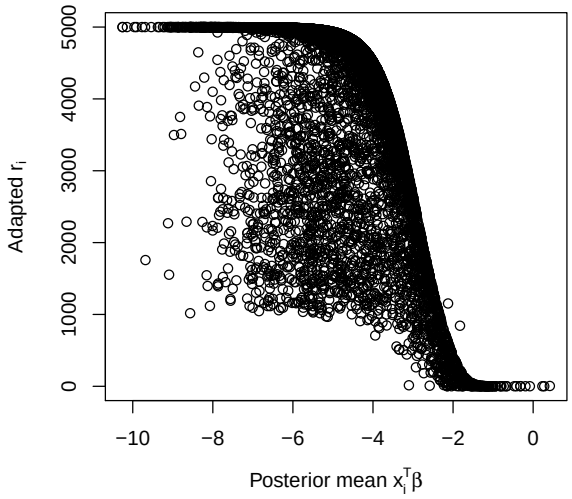
It is interesting to note relation between the adapted r_i and the posterior value of $x_i\beta$ (Figure 2c). The very negative $x_i\beta$ (< -2) suffers the mis-match in the conditional and marginal variance, hence allows large r_i to adjust the step size. The bias reducing term b_i is linear in $(\sqrt{r_i} - 1)x_i\beta$ as expected (Figure 2d).



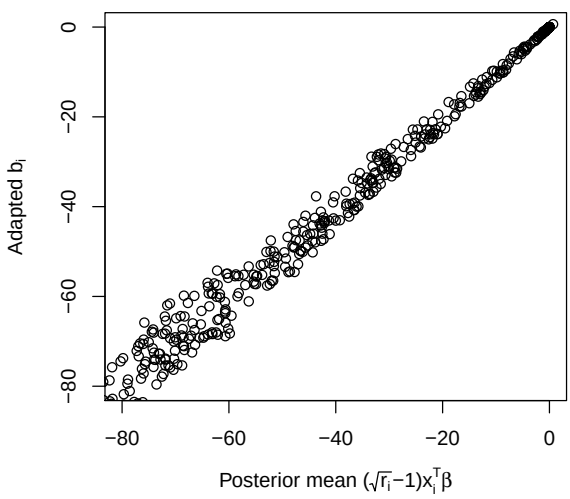
(a) Traceplot illustrating mixing performance of the original DA, parameter expanded DA and CDA algorithms in probit regression with rare event data.



(b) Autocorrelation function (ACF) illustrating the slow mixing of the DA and parameter expanded DA in rare event data, and CDA correcting this problem.



(c) Numerically adapted r_i showing the room for variance increase is related to the value of $|x_i\beta|$.



(d) Numerically adapted b_i learned during the tuning period is close to $(1-\sqrt{r_i})x_i\beta$ based on the posterior mean.

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). Panel (c) shows the room for the variance increase in r_i ($r_i = 1$: no increase) with respect to the value of $x_i\beta$. Panel (d) shows the adapted bias reducing term is close to the true bias based on the posterior mean.

2.2 Controlling Latent Variable in Conditional Variance

The success of CDA relies on the fact that the calibrated conditional with increased variance can still yield the similar marginal form as the original. Using $\pi(\theta|y) = \int \pi(\theta|z,y)\pi(z|y)dz$, when $\mathrm{cov}(\theta|z,y)$ does not involve the latent variable z (like in the probit example), the multiplication with r does not affect the integrability. Otherwise, when z is part of the variance, multiply r directly on the conditional variance might make the integration difficult. To solve this issue, one can influence the value of z by modifying $\pi(z|y)$ instead, while still maintain the closed-form marginal.

To illustrate, consider the logistic regression:

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

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

$$\begin{aligned} z_i &\sim \text{PG}(1, |x_i\beta|), \\ \beta &\sim \text{No}\left((X'ZX)^{-1}X'(y-\frac{1}{2}), (X'ZX)^{-1}\right), \end{aligned}$$

where $Z = diag(z_1, \ldots, z_n)$. This is due to the integration $L(y_i \mid x_i\beta) = \int \exp\{x_i\beta(y_i - 1/2)\} \exp(-\frac{z_i(x_i\beta)^2}{2}) \,\text{PG}(z_i \mid 1, 0)dz_i$. As described above, the direct multiplying r_i to z_i would make it difficult to integrate over z.

Observing the value of z is influenced by the first parameter in Polya-Gamma, we modify $\text{PG}(z_i \mid 1, 0)$ by replacing 1 with the auxiliary parameter r_i. This calibration does not affect the integrability, applying bias correction term b_i on $x_i\beta$, leading to:

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

yielding the update rule for the proposal:

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

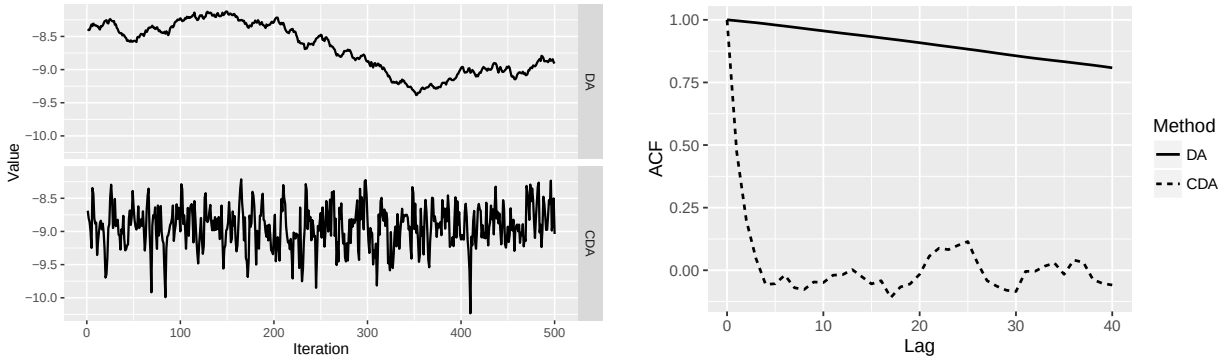
Unlike the probit example where the variance increase is deterministic, the step size tuning here is stochastic: utilizing $\mathbb{E}z_i = \frac{r_i}{2(|x_i\beta+b_i|)} \tanh(\frac{|x_i\beta+b_i|}{2})$, when $r_i < 1$, z_i will have small value with high probability, leading to large variance in $(X'ZX)^{-1}$.

Similar to probit example, M-H step can be derived to accept the proposal β^*:

$$1 \wedge \frac{\prod_i L_r(x_i\beta; y_i)L(x_i\beta^*; y_i)}{\prod_i L_r(x_i\beta^*; y_i)L(x_i\beta; y_i)},$$

where $L(\theta; y_i) = \frac{\exp(\theta y_i)}{1+\exp(\theta)}$. It then can derived the adaptive form for b_i should be $\log[\{1+\exp(x_i\beta)\}^{1/r_i} - 1] - x_i\beta$.

For illustration, we use a two parameter intercept-slope model with $x_1 \sim \text{No}(0, 1)$ and $\beta = \{-9, 1\}$. With $n = 10^5$, it leads to rare positive outcome $\sum y_i = 50$. We adapt the r and b for 100 steps, reaching acceptance rate of 0.9; then stopp adaption and collect the posterior sample. The different mixing performances in the original DA and the calibrated one are shown in Figure 3.



(a) Traceplot illustrating mixing performance of the original DA and CDA algorithms in logistic regression.

(b) Autocorrelation function (ACF) illustrating the different performances of the DA and CDA.

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 (Polson et al., 2013).

2.3 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 $\{\theta, \eta\}$, where only θ are the ones suffering from slow mixing. With Gibbs sampling, we alternatively sample $\eta \mid \theta, y$ and $\theta \mid \eta, y$. To calibrate, we now focus on $\theta \mid \eta, y$ and omit η for notational ease. Breaking down the likelihood into the product of the augmented likelihoods, each has:

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

where C is the constant free from θ, $m_i(\theta)$ is the function $m_i : \mathbb{R}^p \mapsto \mathbb{R}$. For example, $m_i(\beta) = x_i\beta$ in regression. Let the the conditional distribution for θ be:

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

We adjust Σ with the auxiliary parameter r_i: if Σ is free from z, we multiply r_i into $\pi\left(m_i(\theta)|z_i, y_i\right)$ like the probit example; otherwise, we change the parameters in $\pi(z_i|y_i)$ to influence the value of z_i like the logit example. With an additional parameter b_i to reduce the bias, we obtain the calibrated data augmentation:

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

Then the proposal update rule can be formed:

$$z_i \sim \pi(z_i|m_i(\theta) + b_i)$$

$$\theta^* \sim f(\mu(m_i(\theta) + b_i, z), \Sigma(z)).$$

and accepting new θ^* with probability:

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

To increase the acceptance rate, we adaptively tune r and b until the the acceptance rate reaches the preset threshold (e.g. 0.3). We start from an r_i that corresponds to moderate increase in the step size. Then at the end of each iteration, increase or decrease r_i based on $\alpha_i > 1$ or $\alpha_i < 1$ (assuming smaller r_i has larger chance of $\alpha_i > 1$). After the adaptation, we fix r and b and collect the posterior samples.

3 Theory: Mixing Acceleration

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. Therefore, we study how the calibration can solve this problem.

We first utilize another related quantity, the norm of the forward operator $||\mathbf{F}||$, 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).

The original DA samples in the sequence of $\theta' \to z' \to \theta \to z$. Omitting y for simpler notation, by Lemma 4 in Liu (1994):

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

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

The slow mixing is obviously due to $\mathbb{E}_{DA}[\mathrm{var}_{DA}\{s(\theta)|z'\}] \ll \mathrm{var}_{DA}\{s(\theta)\}$.

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

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

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 $\mathbb{E}[\mathrm{var}\{s(\theta)|z'\}]$ by γ times, 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 $p_0\gamma \geq 1$,*

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

As the result, with an increase in γ by raising the conditional variance, and a p_0 away from 0, one can significantly accelerate the mixing. The large p_0 is attributed to the similarity between L_r and L in (9), as the result of adaptation. For example, in the logit CDA:

$$\alpha_i = \frac{\{1 + \exp(x_i\beta)\}\{1 + \exp(x_i\beta^* + b_i)\}^{r_i}}{\{1 + \exp(x_i\beta^*)\}\{1 + \exp(x_i\beta + b_i)\}^{r_i}}$$

when $x_i\beta$ is negative (corresponding to large variance gap that causes slow mixing), the adapted bias reduction term $b_i = \log\{1/r_i + O(\frac{\exp(x_i\beta)}{r_i})\} \approx -\log r_i$. Then $\{1 + \exp(x_i\beta - \log r_i)\}^{r_i} = 1 + \exp(x_i\beta) + O(\frac{\exp(2x_i\beta)}{r_i})$.

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. $y_i \sim p\left(g(x_i\eta)\right)\delta_0 + \{1 - p\left(g(x_i\eta)\right)\}Poisson\{\exp(x_i\beta)\}$. Instead, it is rather useful to consider a simpler model $y_i \sim Poisson\{\exp(\beta_0 + \sum_j x_{i,j}\beta_j)\}$ with a quite negative intercept β_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:

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

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

$$\begin{aligned} z_i &\sim \mathrm{PG}\left(\lambda, x_i\beta - \log\lambda\right) \\ \beta &\sim \mathrm{No}\left(\left[(X'ZX)^{-1}X'\big(y - \lambda/2 + z\log\lambda\big), (X'ZX)^{-1}\right)\right. \end{aligned}$$

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

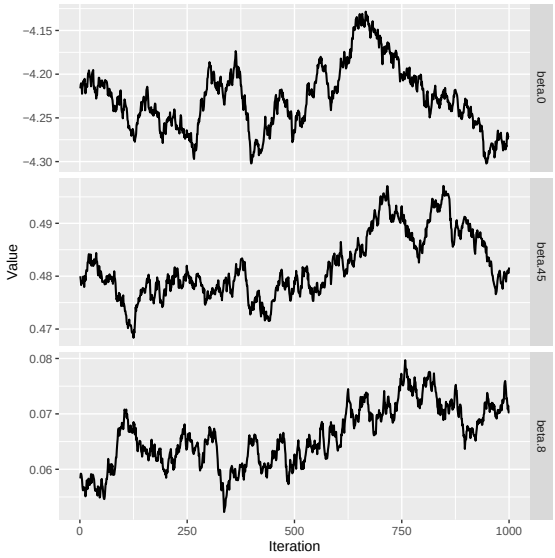
As λ control the magnitude of the latent z, the calibration is straightforward by replacing λ with small r_i and $-\log\lambda$ with b_i, giving the calibrated likelihood:

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

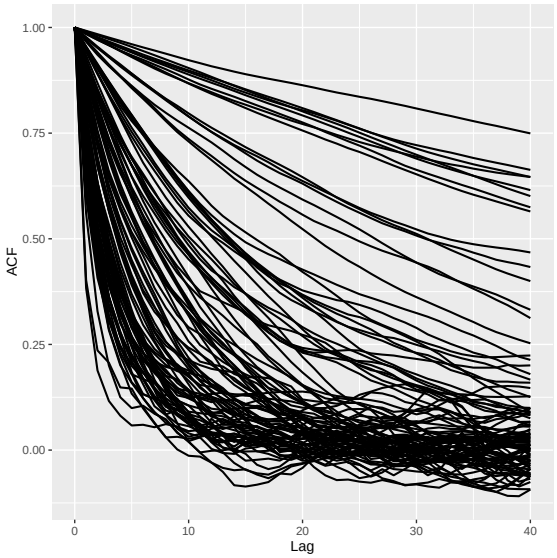
which generates the same sampling algorithm as the CDA in logit regression, except that $r_i > y_i$ to ensure the posterior properiety.

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.

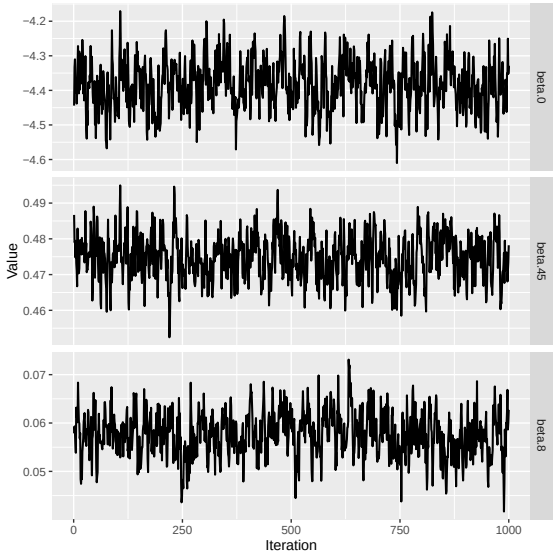
The mixing of DA and CDA is compared in traceplots and autocorrelation plots in Figure 4. DA shows slow mixing for several parameters (Figure4b), including the important intercept estimate β_0 (first plot in Figure4a). After calibration, the slow mixing problem is solved: *all* of 96 parameters show very low autocorrelation.



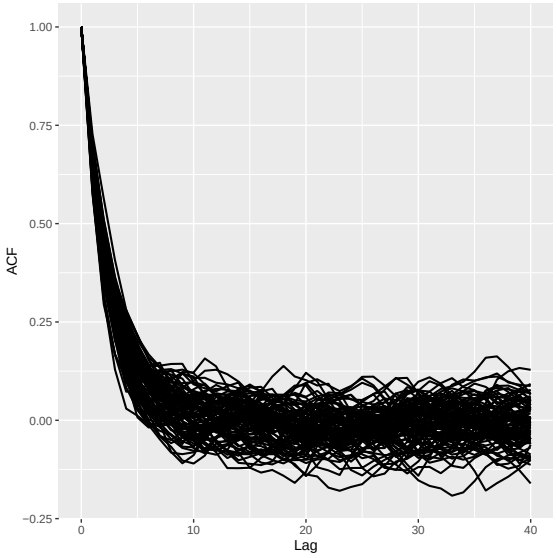
(a) Trace plots of three parameters from DA.



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



(c) Trace plots of three parameters from CDA.



(d) Autocorrelation of all the 96 β'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 β_0 and the norm of the coefficients $\sum_{j=0}^{95} |\beta_j|$. 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 $\mu_i = \exp(x_i\hat{\beta})$ and $\hat{\beta}$ as the posterior mean. For prediction performance, we use the testing dataset and $\hat{y}_{i,new} = \exp(x_{i,new}\hat{\beta})$ as the estimator. We evaluate the cross-validation RMSE between $y_{i,new}$ and $\hat{y}_{i,new}$.

As expected, the estimates for β_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
β_0	-4.21 (0.042)	-4.38 (0.075)	-4.47 (0.071)
$\sum_{j=0}^{95} \lvert\beta_j\rvert$	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.

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.

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.

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.

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.

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:

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

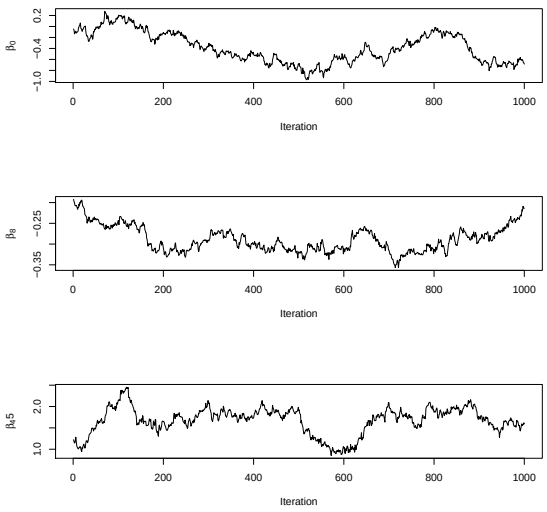
6.1.2 Theorem 1

With Lemma 1,

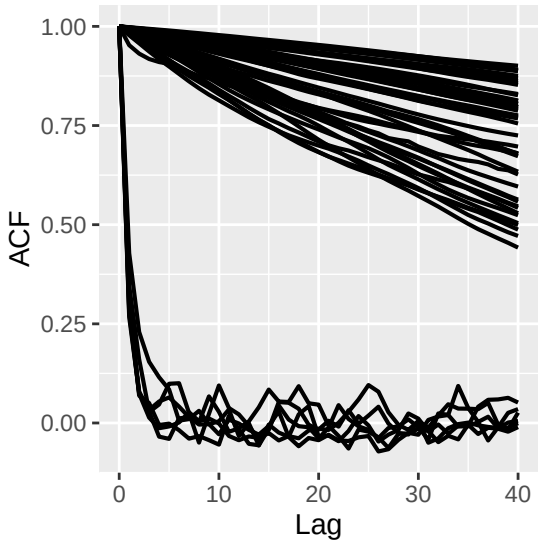
$$\begin{aligned}
\mathbb{E}[\text{var}_{CDA}\{s(\theta)|z', \theta'\}] \geq& p_0 \cdot \mathbb{E}[\text{var}_{CDA}(s(\theta^*)|z')] \\
\geq& p_0\gamma \cdot \mathbb{E}[\text{var}_{DA}(s(\theta^*)|z')].
\end{aligned}$$

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



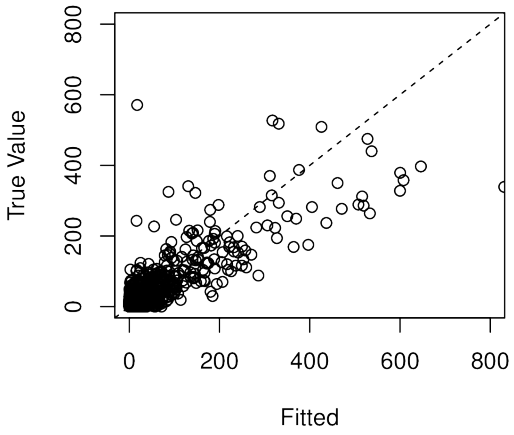
(a) Trace plots of three parameters from DA ZIP model



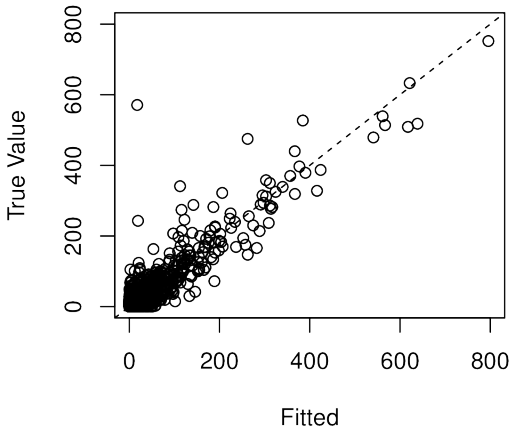
(b) Autocorrelation of all the 96 β's from DA ZIP model.

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

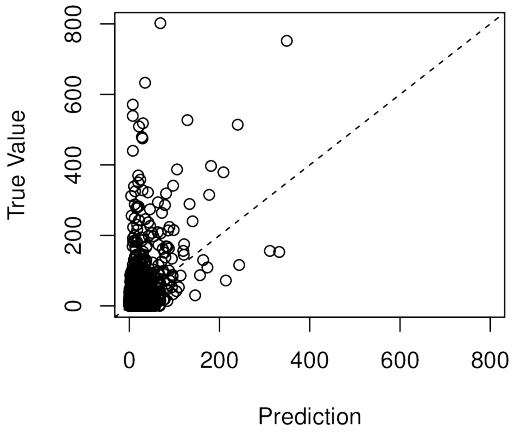
6.3 Goodness-of-Fit and Cross-Validation for Poisson Regression



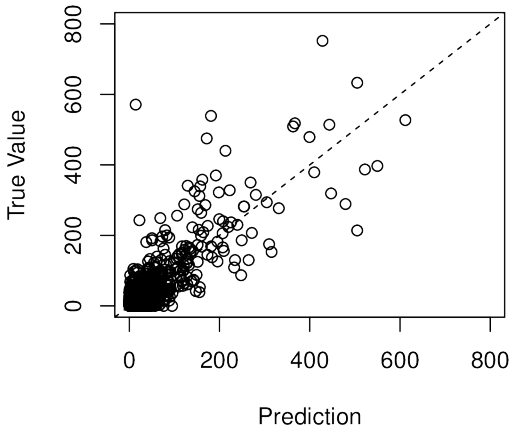
(a) Fitted vs true values using DA



(b) Fitted vs true values using CDA



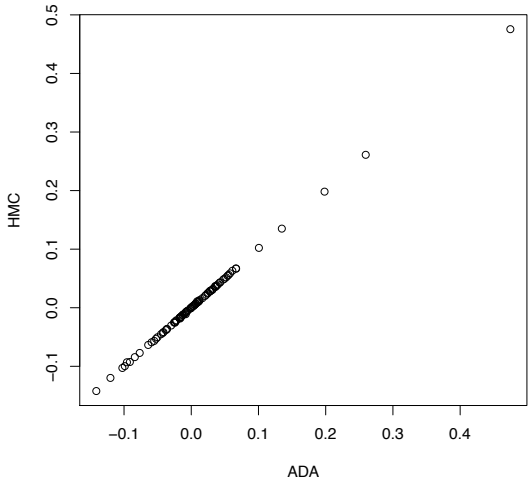
(c) Prediction vs true values using DA



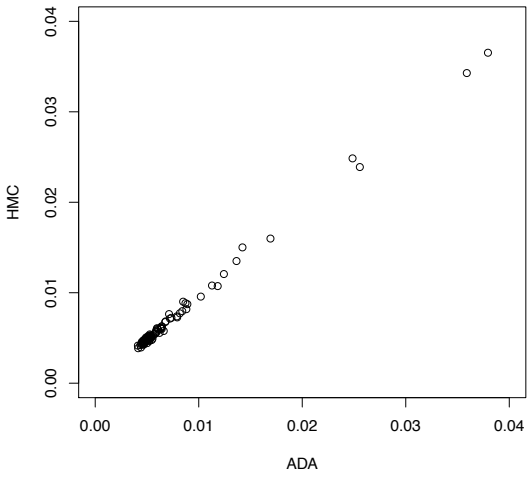
(d) Prediction vs true values using CDA

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



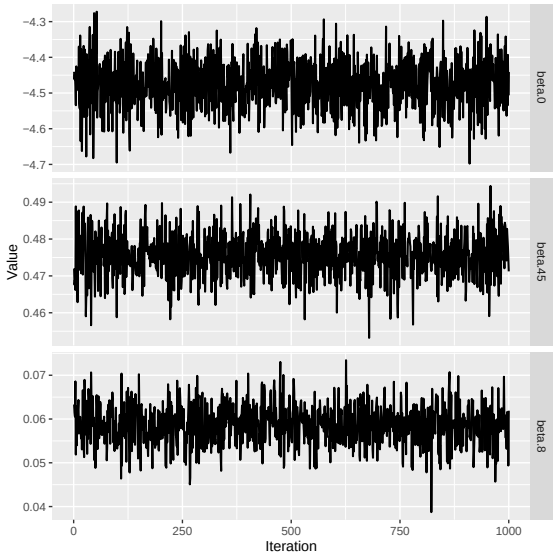
(a) Comparing posterior means for $\beta_1, \ldots \beta_{95}$ from the HMC and CDA. The RMSE between the two is 0.0007.



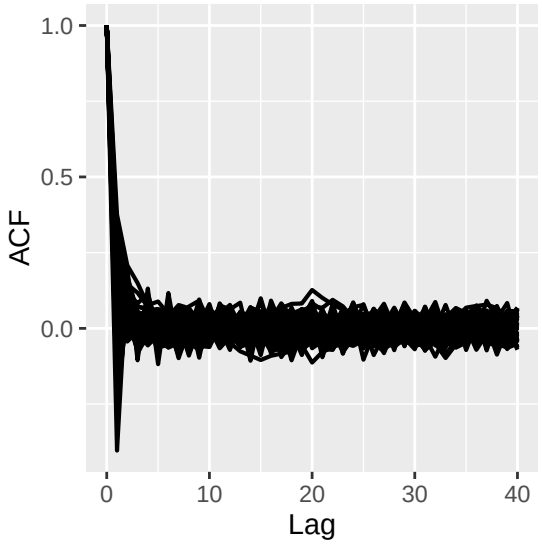
(b) Comparing posterior standard deviation for $\beta_1, \ldots \beta_{95}$ from the HMC and CDA. The RMSE between the two is 0.0004.

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.