Calibrated Data Augmentation for Scalable Markov Chain Monte Carlo

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

October 30, 2016

Abstract: Data augmentation (DA) 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 mismatch between the rate of convergence of the posterior distribution conditionally on augmented data and with augmented data marginalized out. In this article, we propose a calibrated DA (CDA) algorithm, which adjusts for this mismatch by slowing down the convergence of the conditional distribution via parameter marginalization or scale increasing. The bias introduced by changing the scale can be eliminated through a Metropolis-Hastings step, or in some cases, analytically bounded with suitable working parameters. CDA is easily applied to a broad variety of existing DA 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: Albert and Chib; Approximate Markov chain Monte Carlo; Big n; Data Augmentation; Maximal Correlation; Polya-Gamma; Scaling limits.

1 Introduction

With the deluge of data in many modern application areas, there is a 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 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 broad settings.

The standard workhouse for Bayesian posterior computation is Markov chain Monte Carlo (MCMC), and related sampling algorithms. Non-sampling alternatives, such as variational Bayes, tend to lack accuracy guarantees. However, it is well known that conventional MCMC algorithms are often unscalable for computation with large and complex data. Due to the iterative nature, the computational cost 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. As an analogy to travel in the real world, the former is the speed and the latter is the path distance to reach and tour the target. Any scaling obstacle in either factor will make MCMC inapplicable.

In recent years, a rich literature has developed focusing on increasing speed (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. However, very little attention has been paid to the problem of finding an efficient path to reach and explore the target distribution in big data settings. This issue is of substantial practical performance, as many routinely used algorithms, such as data augmentation samplers for probit (Albert and Chib, 1993) and logistic models (Polson et al., 2013), can suffer from extreme slow mixing for large sample sizes and become unusable as sample size increases (see Rajaratnam and Sparks (2015) and Johndrow et al. (2016) for details).

Such issues provide motivation for the development of algorithms for accelerating mixing in large sample size n settings. In this article, we focus on improving data augmentation (DA) based MCMC algorithms. Letting $\pi(\theta|y)$ denote the posterior distribution of the parameters θ given data y, such algorithms focus on the modified stationary distribution $\pi(\theta, z|y)$, with latent data z introduced in a careful manner so that $\pi(\theta|y) = \int \pi(\theta, z|y)dz$ and posterior sampling based on the conditional posterior distributions $\pi(z|\theta, y)$ and $\pi(\theta|z, y)$ is easier than directly sampling based on $\pi(\theta|y)$. For example, after data augmentation a simple blocked Gibbs sampler may be possible, while sampling based on the non-augmented posterior $\pi(\theta|y)$ may require substantial tuning. DA algorithms are used routinely, with the algorithms of Albert and Chib (1993) for probit models and Polson et al. (2013) for logistic regression and Poisson log linear models particularly popular.

The broad use of such DA algorithms is well justified given their simplicity, lack of tuning and good performance in a broad variety of settings. However, as the sample size n increases, substantial mixing problems can arise due to a mis-calibration issue. In particular, it is often the case that as n increases, the conditional posterior $\pi(\theta|z, y)$ concentrates at a faster rate than the marginal posterior $\pi(\theta|y)$. With $\pi(z|\theta y)$ strongly dependent on the previous state of θ, this can cause the step size of MCMC algorithms for updating θ to be much too small. Special cases of this pitfall of DA are well known in the literature. For example, the Albert and Chib (1993) DA algorithm for probit models for ordered categorical data updates latent threshold parameters from their uniform conditional posterior given the augmented data. As n increases, these uniform distributions have small support heavily correlated with the previous threshold value, leading to critical mixing problems even in moderate samples. In this specialized setting, it is common to marginalize out z in updating the latent thresholds using a Metropolis-Hastings (MH) step (Cowles, 1996). In this article,

our focus is on obtaining a broad and more fundamental solution to the mis-calibration problem.

There is a previous literature attempting to accelerate mixing in DA algorithms through the use of parameter expansion (PX) (Liu and Wu, 1999). The basic idea of PX-DA is that poor mixing is often due to high posterior dependence in parameters, which are updated separately within an MCMC algorithm. By introducing extra or redundant parameters into the MCMC algorithm, it is possible to dilute correlations in the parameters of interest and obtain improved mixing. PX-DA versions of the popular Albert and Chib (1993) algorithm have been developed (Liu and Wu, 1999), leading to some gains.

However, PX-DA does not address the mis-calibration issue, which is the focus of this article. In particular, we propose a calibrated DA (CDA) class of algorithms, which removes the dependency of the latent variable on the parameter, or calibrates the transition variance in the posterior $\pi(\theta|z,y)$ by adjusting the conditional variance. These adjustments can very significantly reduce autocorrelation in the Markov chain, dramatically improving mixing relative to DA or PX-DA in many very large n settings. Calibrating the variance leads to a bias in the stationary distribution so that the target is $\pi_r(\theta|y)$, which may differ slightly from $\pi(\theta|y)$. Metropolis-Hastings adjustment is employed to eliminate bias. We also show in several cases, this approximation error can be analytically bounded with suitable choice of the working parameter, leading to fast-mixing approximate MCMC. The proposed CDA approach is widely applicable, and we demonstrate the utility through probit, logistic and Poisson log-linear examples.

Section 2 proposes the general calibrated data augmentation (CDA) algorithm. Section 3 presents theory showing acceleration of mixing, accuracy of approximation, and ergodicity for the approximate MCMC algorithm without use of a MH adjustment. Section 4 provides simulation experiments and comparisons with existing algorithms. Section 5 contains an application to a large computational advertising data set. All the proofs are provided in the appendix.

2 Calibrated Data Augmentation

The primary role of data augmentation (DA) in MCMC algorithms is to make Gibbs sampling possible, because after augmentation conditional posterior distributions follow parametric forms that are easy to sample from. DA Gibbs proceeds by sampling from the conditional posterior distributions of the latent data based on $\pi(z|\theta,y)$ and from conditional posterior distributions of the parameters based on $\pi(\theta|z,y)$; each of these sampling steps can be broken up into a series of conditional updates for successive subsets of the latent data and parameters when the joint conditionals $\pi(z|\theta,y)$ and $\pi(\theta|z,y)$ are not available in a simple form.

The current article focuses on problems that arise due to a mis-calibration problem that occurs when (i) z is strongly dependent on the last value of θ, and (ii) $\text{var}(\theta|z,y)$ is substantially less than $\text{var}(\theta|y)$. The first condition is very common in Gibbs sampling due to the use of full conditional distributions. The second

condition is related to the rate difference of the concentration. Consider that the conditional posterior given the latent data is obtained by updating the posterior given the observed data with information in the latent data likelihood via Bayes rule as: $\pi(\theta|z,y) \propto \pi(\theta|y)L(z|\theta,y)$. Hence, as the amount of information in the latent data likelihood $L(z|\theta,y)$ increases, which typically occurs as the observed data sample size n increases, $\text{var}(\theta|z,y) \ll \text{var}(\theta|y)$. The combination of these two conditions naturally leads to a slow mixing problem, because the MCMC updates based on $\pi(\theta|z,y)$ will only explore a small region of $\pi(\theta|y)$. In general, the problem gets worse as the information in the latent data likelihood increases; inefficiency can arise even in moderate sample sizes but becomes critical in large samples.

The key novel idea of this article is to *calibrate* DA by adjusting the step size of the conditional distribution. Addressing either one of the conditions mentioned above leads to substantial acceleration. To remove the dependency of the latent variable on the last state, the calibrated DA (CDA) first marginalizes out the parameter from the Markov chain and then samples the parameter as an extra step at the end of each iteration. To calibrate the rate mismatch in the conditional posterior covariance, CDA induces an increase in $var(\theta|z,y)$ by inflating the variance parameter with a working parameter r. This leads up to an altered likelihood $\pi_r(y|\theta) = \int \pi_r(y|\theta,z)\pi_r(z)dz$. To reduce the bias, a correction term b is introduced so that $\pi_{r,b}(y|\theta)$ is close to $\pi(y|\theta)$. The samples from $\pi_{r,b}(y|\theta)$ can be used as either as approximate posterior or good proposal in the Metropolis-Hastings algorithm. We demonstrate a general numeric algorithm for $\pi(\theta|y,z)$ in the location-scale family, as well as analytical method for some special cases.

We now illustrate the two calibration strategies. As the parameters are commonly in multiple dimensions, we now divide the parameter into two sets $\theta = \{\theta_1, \theta_2\}$. Let θ_1 be the ones that mixes slowly.

2.1 Removing Dependency by Marginalization

Intuitively, if θ_1 concentrates rapidly near a function of z, but z depends on the last value of θ_1, slow mixing naturally occurs. Breaking the dependency would lead to significant improvement. The first strategy involves marginalize out θ_1 in the conditional distribution of z and θ_2, then sample new θ_1 based on $\pi(\theta_1|z,\theta_2,y)$, shown in Algorithm 1:

Algorithm 1 Marginalization based CDA

Based on the augmented joint distribution $\pi(z,y,\theta_1,\theta_2)$, obtain the marginal $\pi(z,y,\theta_2) = \int \pi(z,y,\theta_1,\theta_2)d\theta_1$.
for $step = 1 \dots N_{Steps}$ **do**
  Sample z from the marginal $\pi(z|\theta_2,y)$;
  Sample θ_2 from the marginal $\pi(\theta_2|z,y)$;
  Sample θ_1 from the full conditional $\pi(\theta_1|z,\theta_2,y)$;
end for

Obviously, this does not change the stationary distribution, but the sampling of the latent variable z no

longer relies on the previous value of θ_1. Therefore the generation of the new θ_1 is free from its last state.

This integration strategy is similar to the collapsed sampler proposed by Liu (1994), who focused on accelerating the imputation of z as the missing data by marginalizing out the the unimportant variables. The key difference is that θ_1 is important but marginalized out at first to exclude the information of the past, then it is sampled again at the end of each MCMC iteration. Note the first two steps in the algorithm already form a complete Markov chain and neither $\pi(z|\theta_2, y)$ or $\pi(\theta_2|z, y)$ depends θ_1. Therefore the autocorrelation of $E(\theta_1|z, \theta)$ is significantly reduced. A more rigorous theory will be provided in the next section.

Remark 1. *Due to the closed-form of $\pi(\theta_1|\theta_2, z, y)$ is commonly available in Gibbs sampling, it is possible to marginalize θ_1 out in most cases. The easiness of applying the new algorithm depends on if the marginal $\pi(z|\theta_2, y)$ and $\pi(\theta_2|z, y)$ can be readily sampled. When this is not true, it is possible to make some compromises such as reverting θ_2 to $\pi(\theta_2|\theta_1, z, y)$, as long as θ_2 does not impact greatly on the concentration of θ_1.*

We use the next example to show if $\pi(\theta_1|\theta_2, z, y)$ follows uniform distribution, the marginalization does not alter the distribution forms of the other posteriors, but leads to significant gains in mixing. Therefore, under such cases marginalization is always recommended.

Example 1: Threshold Updating in Ordinal Probit Regression

Consider the Albert and Chib (1993) DA algorithm for the probit regression with ordered categorical data, as aforementioned in the introduction. The likelihood is $L(y_i = j|x_i, \beta, \gamma) = \{\Phi(\gamma_j - x_i^T\beta) - \Phi(\gamma_{j-1} - x_i^T\beta)\}$, where Φ is the cumulative distribution function of standard normal, $\gamma = \{\gamma_0, \gamma_1, ...\gamma_k\}$ are the threshold parameters that correspond to the boundaries of k categories, with $\gamma_0 = -\infty$, $\gamma_k = -\infty$. To ensure identifiability, γ_1 is fixed at 0. A latent variable was discovered as the truncated normal distribution conditioned on β and γ, $z_i \sim \mathcal{N}_{(\gamma_{j-1},\gamma_j)}(x_i^T\beta, 1)$. For simplicity, flat priors are assumed for both β and γ. Thanks to the augmentation, the posterior for the other parameters can be sampled from normal $\beta \sim \mathcal{N}\{(x^Tx)^{-1}(x^Tz), (x^Tx)^{-1}\}$ and uniform $\gamma_j \sim \mathcal{U}\{\max_{i:y_i=j}(z_i), \min_{i:y_i=j+1}(z_i)\}$.

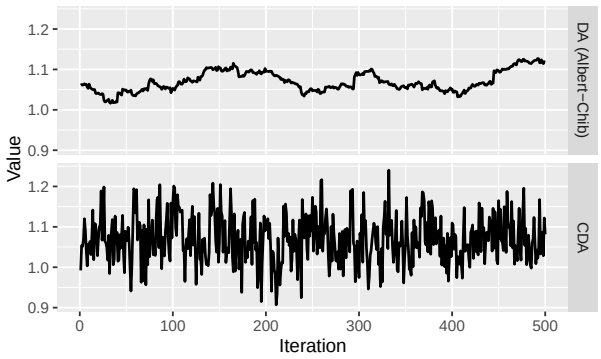
In the conditional distribution $\pi(z|\gamma, \beta, y)$, both the maximum of $z_i \sim \mathcal{N}_{(\gamma_{j-1},\gamma_j)}(x_i^T\beta, 1)$ and the minimum of $z_i \sim N_{(\gamma_j,\gamma_{j+1})}(x_i^T\beta, 1)$ approach rapidly to the boundary γ_j as the numbers of $y_i = j$ and $y_i = j + 1$ increase. Immediately at the next step, $\pi(\gamma_j|z, y)$ is confined by those two values. This causes critical mixing issue. To illustrate, we simulate only $1,000$ data points with $x_i^T\beta \sim \mathcal{N}(-1, 1)$ and threshold $\{-\infty, 0, 1, \infty\}$, obtaining 3 categories with count 504, 267 and 229. Even on this small sample with relatively balanced data, the estimated threshold γ_2 is critically slow (Figure 1).

Following Algorithm 1 with $\theta_1 = \{\gamma_2, \dots \gamma_{k-1}\}$ and $\theta_2 = \beta$, CDA first integrates the joint distribution $\pi(z, y, x, \beta, \gamma, y) = \prod_i \mathcal{N}(z_i|x_i^T\beta, 1) \prod_{j=1} \frac{1\{\max_{i:y_i=j}(z_i)<\gamma_j<\min_{i:y_i=j+1}(z_i)\}}{\min_{i:y_i=j+1}(z_i) - \max_{i:y_i=j}(z_i)}$ over $\gamma_2, \dots \gamma_{k-1}$, where each uniform

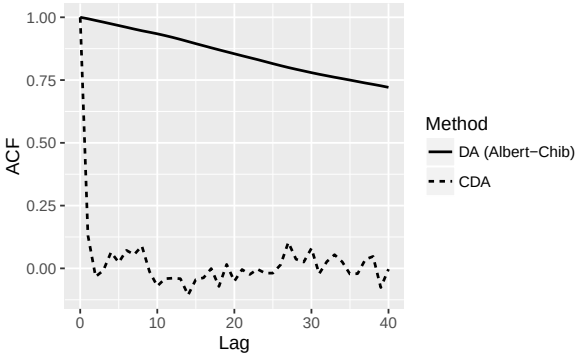
simply integrates to 1. The resulting algorithm is

$$
\begin{aligned}
&z_i \sim \mathcal{N}_{(-\infty,0)}(x_i^T\beta, 1) \text{ if } y_i = 1,\\
&z_i \sim \mathcal{N}_{(0,\infty)}(x_i^T\beta, 1) \text{ if } y_i > 1,\\
&\beta \sim \mathcal{N}\{(x^Tx)^{-1}(x^Tz), (x^Tx)^{-1}\},\\
&\gamma_j \sim U(z_{\left(\sum_i 1(y_i\le j)\right)}, z_{\left(\sum_i 1(y_i\le j)+1\right)}),
\end{aligned}
\tag{1}
$$

where $z_{\left(m\right)}$ denotes the mth order statistics. Due to the removal of correlation from the past, the mixing is substantially improved (Figure 1).



(a) Traceplot illustrating mixing performance of the original DA and CDA algorithms in the threshold updating in probit regression with ordered categorical data.



(b) Autocorrelation function (ACF) illustrating that the calibration improves the mixing from high correlation in 40 lags to an immediate drop to low correlation.

Figure 1: Panel (a) demonstrates the improvement in mixing and panel (b) shows autocorrelations change by calibrating Albert and Chib (1993) algorithm with threshold updating in ordered categorical data.

2.2 Fixing Variance Mismatch by Increasing Conditional Variance

CDA can directly correct the mismatch between the marginal and conditional variances. This relies on increasing the conditional variance via an approximate distribution. We assume the adjusting does not impact the posterior of $\pi(\theta_2|\theta_1, z, y)$, for notational simplicity, we use θ to represent of θ_1 and omit θ_2 in this section. CDA first induces an increase in $var(\theta|z, y)$ by modifying the parameters in each $\pi(z_i|\theta)$ with an working parameter r_i. For example, when $var(\theta|z, y)$ is a linear function of the $var(z_i|\theta)$ (e.g. $var(\theta|z, y) = [x^T diag\{1/var(z_i|\theta)\}x]^{-1}$), it is replaced with the $r_i var(z_i|\theta)$; when $var(\theta|z, y)$ is related to z_i (e.g. $var(\theta|z, y) = [x^T diag\{z\}x]^{-1}$), we change the parameters in $\pi(z_i|\theta)$ to generate smaller z_i. Let subscript r denote the modified ones. The modified marginal likelihood is obtained for each $L_r(y_i|\theta) = \int \pi(y_i, z_i|\theta)dz_i$. Without loss of generality, we assume that variance is monotonically increase in $r > 1$ and $r = 1$ corresponds to no change.

With the same distributional form, the integration will have the same distribution as $L(y_i|\theta)$ with difference parameters. Comparing the two, a bias correction term b_i is used to quantify the difference. For example, if $L(y_i|\theta)$ and $L_r(y_i|\theta)$ differs in the parameter as $x_i^T\theta$ and $\sqrt{r_i}x_i^T\theta$, then the second term is replaced as $\sqrt{r_i}x_i^T\theta + b_i$ with ideal bias correction near $(1-\sqrt{r_i})x_i^T\theta$. Then new the correction parameter is passed down in deriving the posterior $\pi_{r,b}(z_i|y_i,\theta)$ and $\pi_{r,b}(\theta|z,y)$. Note that although the exact bias $f_b(\theta, y_i, r_i)$ can be obtained as a function of θ, one would not use it to correct bias at every step as it would cause convergence issue.

With fixed r and b, the posterior sample from the new $\pi_{r,b}(\theta|y)$ is an approximate to the $\pi(\theta|y)$. To reduce the approximation error, the numeric adaptation can be made on r_i and b_i to increase $L_{r,b}(y_i\theta)/L(y_i|\theta)$. To obtain exact posterior, the sample from $\pi_{r,b}(\theta|y)$ can be used as proposal in Metropolis-Hastings step. Due to the similarity of the proposal to the true density, the proposed new parameter can be in multiple dimensions while enjoys high acceptance rate. In some special cases, analytical solution exists for good values of r_i and b_i with negligible approximation error, then the sample of $\pi_{r,b}(\theta|y)$ can be as the approximate posterior sample without the accept-reject step.

2.2.1 Numeric Exact Method

We first present a general method that works in most of the DA algorithms that requires calibration. This involves numeric adaptation of the working parameters r and b. The algorithm is listed in Algorithm 3. After the variance and bias adjustment terms are included, the procedures consists of two parts: in the tuning period, r and b are adapted using $L_{r,b}(y_i\theta)/L(y_i|\theta)$ at each step; in the sampling period, those parameters are fixed to ensure ergodicity.

Algorithm 2 Variance Adjusting CDA (Numeric)

Increase the variance of $\pi(\theta|z,y)$ by changing the parameter in $\pi(z_i|\theta,y_i)$ with r_i;
Integrate the modified density to obtain $L_r(y_i|\theta) = \int \pi_r(z_i|\theta)\pi(y_i|z_i,\theta)dz_i$;
Compare with $L(y_i|\theta)$, include another bias correction parameter b so that $L_{r,b}(y_i|\theta)$ can be equal to $L(y_i|\theta)$ with some b for all r.
Derive the analytical form of $f_b(\theta, y_i, r_i)$ so that $L_{r,b=f_b(\theta,y_i,r_i)}(y_i|\theta) = L(y_i|\theta)$;
Obtain $\pi_{r,b}(z_i|\theta,y_i)$ and $\pi_{r,b}(\theta|z,y)$;
Initialize r_i to a large value and $b_i = f_b(\theta, y_i, r_i)$ for $i = 1\dots n$;
for $step = 1\dots N_{Steps}$ **do**
  Generate individual z_i from $\pi_{r,b}(z_i|\theta,y_i)$;
  Generate θ' from $\pi_{r,b}(\theta'|z,y)$;
  Compute $\alpha_i = \frac{L(y_i|\theta',z_i)}{L(y_i|\theta,z_i)}$ for $i = 1\dots n$;
  Set $\theta = \theta'$ if $\mathcal{U}(0,1) < \frac{Q(\theta',\theta)}{Q(\theta,\theta')}\prod\alpha_i$, where $Q(\theta,\theta') = \int \pi_{r,b}(\theta|z,y)\pi_{r,b}(z|\theta',y)dz$
  if $step < N_{Tuning}$ **then**
    Set $b_i = f_b(\theta, y_i, r_i)$ for $i = 1\dots n$;
    If $\alpha_i < 1$, set r_i to $1 \vee (r_i\alpha_i)$ for $i = 1\dots n$;
  end if
end for

Remark 2. *The algorithm is generally applicable. For example, in the cases where $\pi(z_i|\theta, y_i)$ is in location-scale family, the variance adjustment can be made on the scale parameter while bias correction is applied on the location parameter. In most cases, $\frac{Q(\theta',\theta)}{Q(\theta,\theta')} = 1$ due to the symmetry.*

We now use the probit regression to illustrate the numeric method. Due to the conditional $\pi(z_i|\theta, y_i)$ is normal, CDA increases its variance and applies bias correction on the mean.

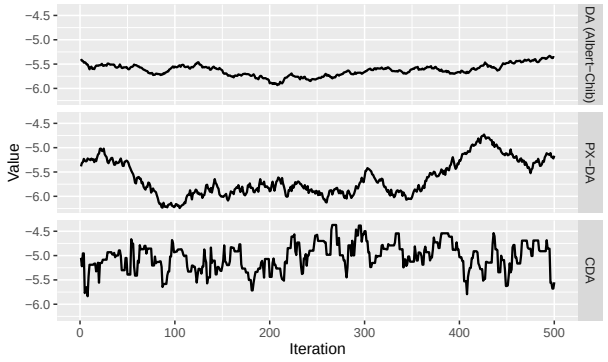
Example 2: Probit Regression with Rare Event

Consider a probit regression $\prod_{i=1}^n L(y_i|x_i^T\beta) = \prod_{i=1}^n \Phi(x_i^T\beta)^{y_i}\{1-\Phi(x_i^T\beta)\}^{(1-y_i)}$ with rare event so that the total occurrence of $y_i = 1$ is quite small compared to to n. We use $\beta = -5, 1$ corresponding to the intercept and $x_{i,1} \sim \mathcal{N}(0,1)$ to generate only 29 positive outcomes among $n = 10,000$. The posterior samples from $\beta \sim \mathcal{N}\{(x^Tx)^{-1}(x^Tz), (x^Tx)^{-1}\}$, conditional on $z_i \sim N_{(-\infty,0)}(x_i^T\beta, 1)$ if $y_i = 0$ and $z_i \sim N_{(0,\infty)}(x_i^T\beta, 1)$ if $y_i = 1$. For the intercept term, it concentrates at the rate of $O(1/n)$; while marginally, it concentrates much slower at $O(1/\log n)$. The slow mixing becomes particularly problematic when n is very large, which is common in rare event applications. In this testing case, the Albert and Chib (1993) DA algorithm suffers from extremely slow mixing. To compare, we test the parameter expansion algorithm (PX-DA) proposed by Liu and Wu (1999), where a redundant parameter is used in the probit link $\Phi(\alpha x_i^T\beta)$. PX-DA reduces the correlation to some extent, however, it does not solve the variance mismatch problem. As shown in Figure **??** (a) and (b), both DA and PX-DA result in extremely slow mixing.

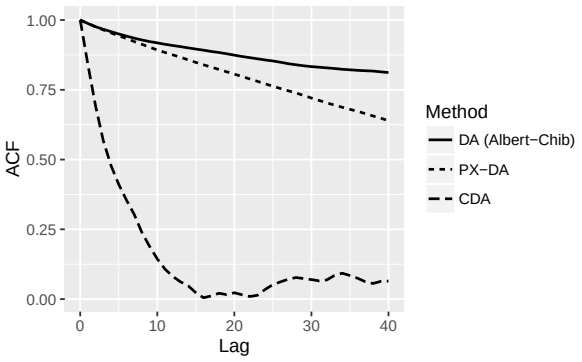
Noting $var(z_i|\beta) = 1$, CDA algorithm first replaces the it with $r_i > 1$ so that the conditional variance for β becomes $(x^T diag^{-1}\{r_i\}x)^{-1}$. The marginalization and adding bias correction leads to $L_{r,b}(y_i|\beta) = \Phi\{(x_i^T\beta + b_i)/\sqrt{r_i}\}^{y_i}[1-\Phi\{(x_i^T\beta + b_i)/\sqrt{r_i}\}]^{(1-y_i)}$. The resulting algorithm for the proposal is:

$$
\begin{aligned}
&z_i \sim N_{(-\infty,0)}(x_i^T\beta + b_i, r_i) \text{ if } y_i = 0 \\
&z_i \sim N_{(0,\infty)}(x_i^T\beta + b_i, r_i) \text{ if } y_i = 1 \\
&\beta \sim \mathcal{N}\{(x^T diag^{-1}\{r_i\}x)^{-1}, x^T diag^{-1}\{r_i\}(z - b_i), (x^T diag^{-1}\{r_i\}x)^{-1}\}
\end{aligned}
\tag{2}
$$

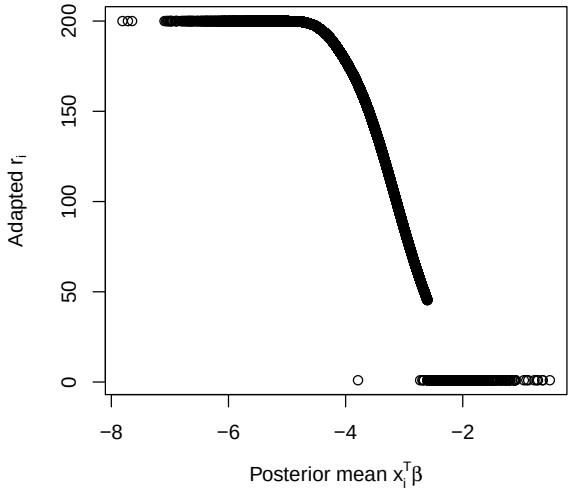
The calibrated algorithm leads to significant improvement of the mixing (Figure **??**a and b). In the numerically adapted proposal distribution (that is close to the true posterior distribution), it is interesting to note the tuned r_i is related to the posterior value of $x_i^T\beta$ (Figure **??**c). The very negative $x_i^T\beta$ (< -4) suffers the mis-match in the conditional and marginal variance, hence allows large r_i to adjust the step size. The bias correction is linear in $(\sqrt{r_i} - 1)x_i^T\beta$ as expected (Figure **??**d).



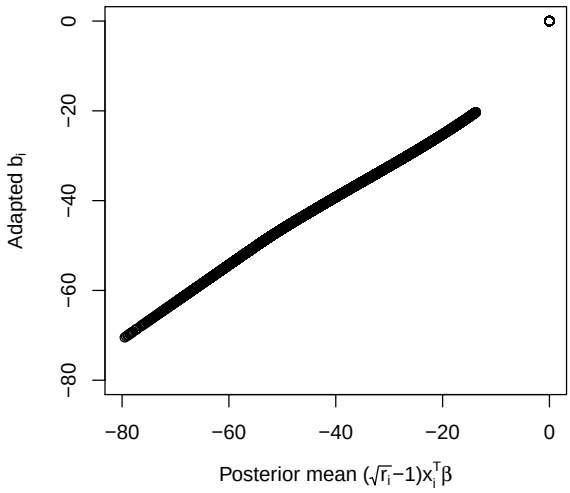
(a) Traceplot illustrating mixing performance of the original DA, parameter expanded DA and CDA algorithms in probit regression with rare event data.



(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 optimized r_i showing the room for variance increase is related to the value of $x_i^T\beta$.



(d) Numerically optimized b_i in the proposal appearing linear and close to the true bias in $(1-\sqrt{r_i})x_i^T\beta$.

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 degree of the variance increase in r_i (if $r_i = 1$: no increase) with respect to the value of $x_i^T\beta$. Panel (d) shows the bias correction term is close to the true bias if no correction were made.

2.2.2 Approximate Method

The numeric solution presented above is generally applicable. In some cases if the approximate error is strictly bounded, the algorithm can be further simplified to direct sampling without the Metropolis-Hastings step.

Let $d(\theta, \theta_{r,b})$ be the distance between the two posteriors based on the exact and approximate method. For example, it can be the Kullback-Leibler distance $E\sum_i\{\log L_{r,b}(y_i|\theta) - \log L(y_i|\theta)\}$. Alternatively, the more loose but useful bound is the norm distance between the first two moments $||\mathrm{E}(\theta|y) - \mathrm{E}_{r,b}(\theta|y)||$,

$||\text{var}(\theta|y) - \text{var}_{r,b}(\theta|y)||$. Both criteria would be carefully examined in the next section.

To check if there is a good analytical approximate, the method first proceeds in the same way as the numeric one by increasing the conditional variance and adjusting for correcting bias. Then it starts to differ by examining if the exact bias $f_b(\theta, y_i, r_i)$ can be approximated by a function without θ, at least in some region of θ.

If the approximation function exists, it is used for setting value of b in each step. This leaves one working parameter r to control the approximate error, quantified by the $d(\theta, \theta_{r,b})$. Therefore, r is used adaptively to bound over the posterior sample Θ^* so that $\sup_{\theta\in\Theta^*} d(\theta, \theta_{r,b}) \leq \epsilon$.

Algorithm 3 Variance Adjusting CDA (Approximate)

Increase the variance of $\pi(\theta|z, y)$ by changing the parameter in $\pi(z_i|\theta, y_i)$ with r_i;
Integrate the modified density to obtain $L_r(y_i|\theta) = \int \pi_r(z_i|\theta)\pi(y_i|z_i, \theta)dz_i$;
Compare with $L(y_i|\theta)$, include another bias correction parameter b so that $L_{r,b}(y_i|\theta)$ can be equal to $L(y_i|\theta)$ with some b for all r.
Derive the analytical form of $f_b(\theta, y_i, r_i)$ so that $L_{r,b=f_b(\theta,y_i,r_i)}(y_i|\theta) = L(y_i|\theta)$;
Check if $f_b(\theta, y_i, r_i)$ can be approximated by $b_i(y_i, r_i)$ in some region of θ.
Initialize r_i to a large value and b_i for $i = 1 \dots n$;

for $step = 1 \dots N_{Steps}$ **do**
  Generate individual z_i from $\pi_{r,b}(z_i|\theta, y_i)$;
  Generate θ from $\pi_{r,b}(\theta|z, y)$;
  If $\sup_{\theta\in\Theta^*} d(\theta, \theta_{r,b}) \leq \epsilon$ is not satisfied, update r_i and b_i;
end for

We found the analytical approximates exist for a wide range of popular distributions, such as logistic, Poisson log-linear and complementary log-log models. In the following example, we use a logistic regression to illustrate the analytical method.

Calibration Example 3: Mixed Effects Logistic Regression Consider a mixed effects logistic regression with $L(y_{ij}|\beta, \sigma^2) \propto \frac{\exp\{(x_{ij}^T\beta + s_{ij}^T\gamma_i)y_{ij}\}}{1+\exp(x_{ij}^T\beta + s_{ij}^T\gamma_i)}$, where $x_{ij}^T\beta$ is the fixed effect with β as the parameter of interest, $s_{ij}^T\gamma_i$ is the group random effect with $\gamma_i \overset{iid}{\sim} N(0, diag\{\sigma_l^2\})$ for all $i = 1 \dots n_i$. We use prior $\pi(\beta) \propto 1$ and $\pi(\sigma_l^2) = \sigma_l^{-2}$. Based on the Polya-Gamma DA proposed by Polson et al. (2013), the posterior is sampled from $z_{ij} \sim \mathcal{PG}(1, x_{ij}^T\beta + s_{ij}^T\gamma_i)$, $\gamma_i \sim \mathcal{N}\{(\sum_j s_{ij}z_{ij}s_{ij}^T + diag\{\sigma_l^{-2}\})^{-1}\sum_j s_{ij}(y_{ij} - \frac{1}{2} - z_{ij}x_{ij}^T\beta), (\sum_j s_{ij}z_{ij}s_{ij}^T + diag\{\sigma_l^{-2}\})^{-1}\}$, $\sigma_l^2 \sim \mathcal{IG}(n_i/2, \sum_i \gamma_{il}^2/2)$ and lastly $\beta \sim N\{(x^T diag\{z_{ij}\}x)^{-1}x^T(y - \frac{1}{2} - vec\{s_{ij}^T\gamma_i\}z), (x^T diag\{z_{ij}\}x)^{-1}\}$. Like the probit regression, slow mixing emerges when $\sum y_{ij}$ is small compared to the total sample size n.

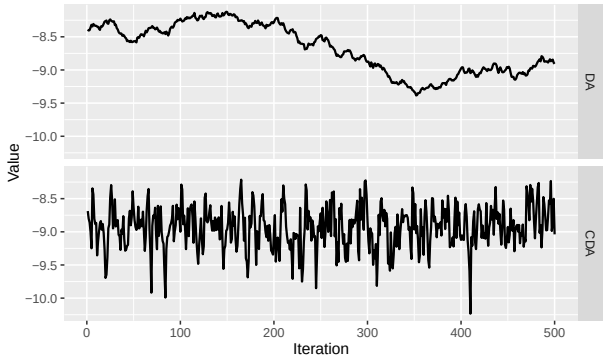
CDA adjusts the conditional variance $(x^T diag\{z_{ij}\}x)^{-1}$ by reducing the the value of z_{ij}. To achieve this, the original Polya-Gamma distribution is replaced by $z_{ij} \sim \mathcal{PG}(\frac{1}{r_{ij}}, x_{ij}^T\beta + s_{ij}^T\gamma_i + b_{ij})$ with $r_{ij} \geq 1$ and b_{ij} as the bias-correction term. The integration leads to $L_{r,b}(y_{ij}|\beta, \sigma^2) \propto \frac{\exp\{(x_{ij}^T\beta + s_{ij}^T\gamma_i)y_{ij}\}}{\{1+\exp(x_{ij}^T\beta + s_{ij}^T\gamma_i + b_{ij})\}^{1/r_{ij}}}$. Setting

$L_{r,b}(y_{ij}|\beta,\sigma^2) = L(y_{ij}|\beta,\sigma^2)$ yields the exact bias correction term $f_b(\theta, y_{ij}, r_{ij}) = \log[\{1+\exp(\eta_{ij})\}^{r_{ij}} - 1] - \eta_{ij} = \log\{r_{ij} + \frac{r_{ij}(r_{ij}-1)}{2!}\exp(\eta_{ij}) + \frac{r_{ij}(r_{ij}-1)(r_{ij}-2)}{3!}\exp(2\eta_{ij}) + \dots\}$ with $\eta_{ij} = x_{ij}^T\beta + s_{ij}^T\gamma_i$. The last equality suggests when $\exp(\eta_{ij})$ is close to 0, $f_b(\theta, y_{ij}, r_{ij})$ can be simply approximated by $b_{ij} = \log r_{ij}$. The choice for r_{ij} is discussed in the next section.

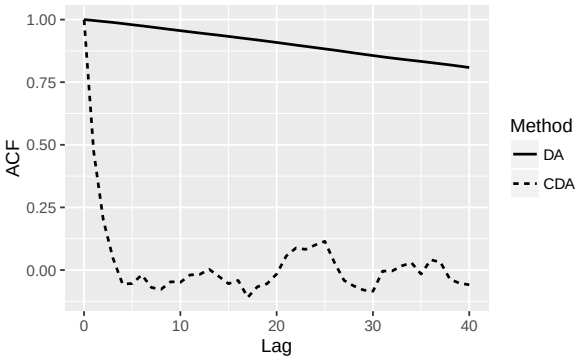
The resulting CDA algorithm is then:

$$
\begin{aligned}
z_{ij} &\sim \mathcal{PG}(\frac{1}{r_{ij}}, x_{ij}^T\beta + s_{ij}^T\gamma_i + \log r_{ij}) \\
\gamma_i &\sim \mathcal{N}\{(\sum_j s_{ij}z_{ij}s_{ij}^T + diag\{\sigma_j^{-2}\})^{-1}\sum_j s_{ij}\{y_{ij} - \frac{1}{2r_{ij}} - z_{ij}(\log r_{ij} + x_{ij}^T\beta)\}, (\sum_j s_{ij}z_{ij}s_{ij}^T + diag\{\sigma_j^{-2}\})^{-1}\} \\
\sigma_l^2 &\sim \mathcal{IG}(n/2, \sum_i \gamma_{il}^2/2) \\
\beta &\sim N\{(x^T diag\{z_{ij}\}x)^{-1}x^T\{y - \frac{1}{2} - z(\log r + vec\{s_{ij}^T\gamma_i\})\}, (x^T diag\{z_{ij}\}x)^{-1}\}
\end{aligned}
\tag{3}
$$

For illustration, we set $\beta = \{-9, 1\}$ as the intercept and the slope to $x_1 \sim \mathcal{N}(0,1)$, $n = 10^5$ and $n_i = 10^2$, and used $s_{ij} = 1$ as the random intercept with variance $\sigma^2 = 0.5$. To ensure identifiability, we fix $\gamma_1 = 0$. This setting leads to rare positive outcome $\sum y_{ij} = 50$. 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 mixed effects logistic regression.



(b) Autocorrelation function (ACF) illustrating the large difference between the mixing performance of the original 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 mixed logistic regression with rare event data, compared with the original (Polson et al., 2013).

3 Theory

3.1 Mixing Accelaration

The mixing of Markov chain is related to its geometric convergence rate. Let $\mathcal{P}(\theta,.)$ be the the Markov transition kernel and $\pi(.)$ be its stationary density and θ be the state in the state space Θ. The chain is geometrically ergodic if there exist $M:\mathcal{X}\to[0,\infty)$ 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 $||P_1-P_2||_{TV}=\sup_{\mathcal{A}\in\mathcal{F}}||P_1(\mathcal{A})-P_2(\mathcal{A})||$.

Consider a Hilbert space $L^2(\pi)=\{s(\theta):E\{s(\theta)\}=0,var\{s(\theta)\}<\infty\}$. The forward operator $\mathbf{F}$ can be defined as $\mathbf{F}s(\theta)=\int\mathcal{P}(\theta,\theta')s(\theta')d\theta'=E\{s(\theta')|\theta\}$, whose norm is equal to 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 directly related to the convergence rate ρ: when the chain is reversible with detailed balance (e.g. Metropolis-Hastings), $\lim_{k\to\infty}||\mathbf{F}^k||^{1/k}=\rho$; when the chain is non-reversible (e.g. Gibbs sampling), $||\mathbf{F}||^2$ is equal to the convergence rate of the reversibilized chain (Fill, 1991).

The two calibrating strategies proposed in the last section reduce the operator norm.

Theorem 1. *Let* $\mathbf{F}$ *and* $\mathbf{F}_{MS}$ *be the operators corresponding to the standard DA and the calibrated DA modified with marginalization and sampling (MS), then* $||\mathbf{F}_{MS}||\leq||\mathbf{F}||$.

Theorem 2. *Let* $\mathbf{F}$ *and* $\mathbf{F}_{IV}$ *be the operators corresponding to the standard DA and the calibrated DA modified with increased variance (IV). If the relative difference in marginal variance* $\frac{|var\{s(\theta)|y\}-var_{IV}\{s(\theta)|y\}|}{var\{s(\theta)|y\}}\leq\epsilon$ *and the variance increase* $\frac{var_{IV}\{\theta|z,y\}}{var\{\theta|z,y\}}\geq(1+\epsilon)$*, then* $||\mathbf{F}_{IV}||\leq||\mathbf{F}||$.

3.2 Error Control in Approximate CDA

For approximate CDA with increased variance, the error needs to be carefully controlled. One can bound the total variation distance $||P-P_{r,b}||_{TV}$, where P and $P_{r,b}$ are the measures for the stationary distributions corresponding to the exact and approximate algorithms. From this bound, one can control the errors of the posterior mean and variance.

Theorem 3. *Let* θ *be the p-element vector of* $\{\theta_j\}_{j=1\ldots p}$*. If the total variation distance between the two measures defined as above is small* $||P-P_{r,b}||_{TV}\leq\epsilon_1$*,* $\pi(\theta|y)$ *and* $\pi_{r,b}(\theta|y)$ *have the tail square integral negligible when* $|\theta|>M$*,* $E\theta_j^2 1_{|\theta_j|>M}\leq\epsilon_2$ *and* $E_r\theta_j^2 1_{|\theta_j|>M}\leq\epsilon_2$*. Let* M *be large enough so that* $\epsilon_2=o(\epsilon_1)$*, then the approximation errors between the first two central moments of* θ_j *have:*

$$|E(\theta_j|y)-E_{r,b}(\theta_j|y)|\leq 2M\epsilon_1+o(\epsilon_1),$$

$$|var(\theta_j|y) - var_{r,b}(\theta_j|y)| \leq 6M^2\epsilon_1 + o(\epsilon_1).$$

Remark 3. *The assumption on the tail is guaranteed by the existence of the second moment, which implies uniform integrablity. The bound on the variance difference is particularly useful in guaranteeing the average increased conditional variable could at most exceed the target variance by a small error* $E^z var_{r,b}(\theta_j|z,y) \leq var(\theta_j|y) + \epsilon$, *with* ϵ *as a polynomial function of* ϵ_1 *defined as above.*

The approximation error for the covariance can be similarly bounded:

Corollary 1. *If* $||P - P_{r,b}||_{TV} \leq \epsilon_1$ *and* $\pi(\theta|y)$ *and* $\pi_{r,b}(\theta|y)$, $E\theta_j^2 1_{|\theta_j|>M_j} \leq \epsilon_2$ *and* $E_{r,b}\theta_j^2 1_{|\theta_j|>M_j} \leq \epsilon_2$ *for all* j. *Let* M_j *be large enough so that* $\epsilon_2 = o(\epsilon_1)$, *then the approximation error between the covariances:*

$$|cov(\theta_{j_1}, \theta_{j_2}|y) - cov_{r,b}(\theta_{j_1}, \theta_{j_2}|y)| \leq 6M_{j_1}M_{j_2}\epsilon_1 + o(\epsilon_1).$$

As a special case, we consider in generalized linear model where the total likelihood can be broken into $L(y|B\theta) = \prod L(y_i|B_i^T\theta)$. Here the linear part $B_i^T\theta$ is set in general sense, which could be the combination of both fixed and random effects. Rather than controlling the total distance between $L(y|B\theta)$ and $L_{r,b}(y|B\theta)$, under some mild conditions of the predictor matrix B, one can focus on bounding the maximum error in individual $B_i^T\theta$, based on $L(y_i|B_i^T\theta)$ over $i = 1 \ldots n$.

Theorem 4. *If* B^TB *is full rank and let* $B^- := (B^TB)^{-1}B^T$, *the following inequalities hold:*

$$||E\theta - E\theta_{r,b}||_1 \leq ||B^-||_1||EB\theta - EB\theta_{r,b}||_\infty$$

$$||cov\theta - cov\theta_{r,b}||_1 \leq ||B^-||_1||B^-||_\infty||covB\theta - covB\theta_{r,b}||_\infty$$

Remark 4. *One reason for using these inequalities is that the norms* $||B^-||_1$ *and* $||B^-||_\infty$ *are usually quite small compared to* n. *For example, a well-conditioned matrix has* $||B^-||_\infty \approx p/n$ *and* $||B^-||_1 \approx p$.

We return to examine the approximate CDA for mixed effects logistic regression, and determined the suitable range for r_{ij}.

Calibration Example 3 (continued): Mixed Effects Logistic Regression

In the CDA for mixed effects logistic regression, the effective approximate likelihood is $L_{r,b}(y_{ij}|\eta_{ij}) = \frac{\Gamma(1/r_{ij}+1)}{\Gamma(y_{ij}+1)\Gamma(1/r_{ij}-y_{ij}+1)} \frac{\exp(\eta_{ij}+\log r_{ij})^{y_{ij}}}{\{1+\exp(\eta_{ij}+\log r_{ij})\}^{1/r_{ij}}}$. The distance between the approximate and the true likelihoods for each i has $||P_{r,b}(y_{ij}|\eta_{ij}) - P(y_{ij}|\eta_{ij}))||_{TV} \leq \{\frac{\sqrt{r_{ij}-1}\exp(\eta_{ij})}{2}\}1\{\eta_{ij} < -\log r_{ij} \leq 0\}$. The tail square integral $E\eta_i^2 1(|\eta_{ij}| > M) = O\{M^2\exp(-M)\}$. Given a maximally tolerable approximation error $\epsilon = 0.01$, the calculation leads to $r_{ij} \leq [\{\frac{10^{-3}}{\exp(\eta_{ij})}\}^2 \wedge \exp(-\eta_{ij})] \vee 1$. This suggests to apply calibration when $\eta_{ij} < -6.9$ approximately, which is consistent with the rare event scenario.

3.3 Ergodicity

The conditions for ergodicity should be met for Markov chains. In the CDA based on marginalization, obviously as long the original chain is ergodic, so is the marginalized chain is, due to its smaller norm of the forward operator. For the CDA based on variance increase, the adaption of the working parameters could influence the ergodicity. In the numeric method, we stop adaption after the tuning period; in the approximate solution, we rely on the the results from Roberts and Rosenthal (2007) that an adaptive Markov chain is uniformly ergodic, if (i) the transition kernel corresponding to each r is uniformly ergodicity (simultaneous convergence) and (ii) the total variation distance between kernels in two adjacent steps converges to 0 in probability (diminishing adaptation).

The first condition is simple to achieve since the approximation does not alter the distribution of the transition kernel but only with different parameterization, the condition can be met with only slight modification. For the diminishing adaptation, since r is adapted so that $\sup_{\theta\in\Theta^*} d(\theta, \theta_{r,b}) \leq \epsilon$ over the posterior sample Θ^*, the suitable r could be the extreme function of the posterior θ. When it is true, the convergence of the extreme function in probability leads to the convergence of the transition kernel.

Calibration Example 3 (continued): Mixed Effects Logistic Regression

In the CDA with mixed effects logistic regression, we set $r_{ij} = 1 \vee \inf_{\eta_{ij}:\theta\in\Theta^*}[\{\frac{10^{-3}}{\exp(\eta_{ij})}\}^2 \wedge \exp(-\eta_{ij})]$, which is related to the maximum of $\eta_{ij} = x_i^T\beta + s_{ij}\gamma_i$ if $\eta_{ij} < 0$. With finite posterior variance of β and γ_i, the value is convergent in probability via Chebyshev's inequality. The diminishing adaption for the numeric example is shown in Figure 4.

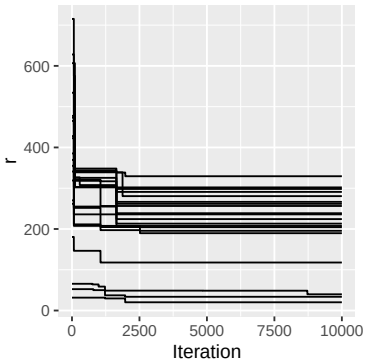


Figure 4: The diminishing adaptation of the working parameter r_{ij} in mixed effects logistic regression.

4 Real Data Application: Poisson Regression for Online Advertisement Tracking

We now apply CDA to a large data application in online advertisement tracking. The dataset contains the click-through count between pairs of website. There are $n = 59,792$ originating websites, where the advertisement is displayed, and 96 different target websites where the link points to. The training data is collected during a two-week period, the a non-overlapping set is collected during another two-week time for validation. For a given website, it is of commercial interests to predict the traffic from each originating one y_i, using the information from the other $p = 95$ websites $x_{i,1}, \ldots, x_{i,95}$. Therefore, this leads to a count data regression.

A significant proportion of the originating websites do not produce traffic to the target one. The outcome data are saturated with zeros and only 4.5% contain positive counts. For this scenario, Poisson log-linear model $y_i \sim Poisson\{\exp(x_i^T\beta)\}$ is known for slow mixing, due to the same variance mis-match caused by the rare positive outcomes. The predictor-based zero-inflated Poisson such as $y_i \sim \pi_i\{g(X_i^T\beta_1)\}1(y_i = 0) + \{1 - \pi_i(g(X_i^T\beta_1))\}Poisson\{\exp(X_i^T\beta_2)\}$ does not alleviate the mixing issue, on the contrary, mixing is even slower because of the high correlation of π_i and those with Poisson mean close to 0.

Rather, we consider calibrating the simpler model $y_i \sim Poisson\{\exp(\beta_0 + \sum_j x_{i,j}\beta_j)\}$ directly. Provided the mixing performance is satisfying, the intercept β_0 would be large and negative with significant uncertainty; at the same time, a small proportion can deviate from mean 0 based on their distinctively large predictors. This simple and tractable model would achieve the same goal as the predictor-based zero-inflated model.

Calibration Example 4: Poisson Log-Linear Regression with Zero-Inflated Data

We first derive an approximate CDA for Poisson log-linear regression. All the existing data augmentation methods for Poisson log-linear involve approximation. For Poisson with large mean, normal approximation works reasonably well; for small ones, Zhou et al. (2012) used negative binomial $\mathcal{NB}\{\alpha, \frac{\exp(x_{ij}^T\beta)}{\exp(x_{ij}^T\beta)+\alpha}\}$ with large α to connect to the Polya-Gamma augmentation. Here, we utilize a simpler limit form $L(y_i|x_i^T\beta) = \frac{\exp(y_i x_i^T\beta)}{\exp\{\exp(x_i^T\beta)\}y!} = \lim_{\lambda\to\infty} \frac{\exp(y_i x_i^T\beta)}{\{1+\exp(x_i^T\beta)/\lambda\}^\lambda y!}$. Using a flat prior on β, the Polya-Gamma augmentation leads to approximate posterior sampling $z_i \sim \mathcal{PG}\left\{\lambda, x_i^T\beta - \log\lambda\right\}$ and $\mathcal{N}\big[(x^T diag\{z_i\}x)^{-1}(x^T\big\{y - \lambda/2 + z\log\lambda\big\}), (x^T diag\{z_i\}x)^{-1}\big]$, with large λ.

Without calibration, the mixing is slow as the conditional variance for β is quite small, due to the large z_i caused in large λ approximation. To calibrate, we replace the above Polya-Gamma with $z_i \sim \mathcal{PG}(\frac{1}{r_i}, x_i^T\beta + b_i)$ and compare the integrated form with the Poisson density. This leads to the exact bias correction $f_b(\theta, y_i, r_i) = \log\frac{\exp\{\exp(x_i^T\beta)r_i\}-1}{\exp(x_i^T\beta)} = \log\{r_i + \frac{r_i^2}{2}\exp(x_i^T\beta) + \frac{r_i^3}{3!}\exp(2x_i^T\beta) + \ldots\}$, which can be

approximated b $b_i = \log r_i$ when $\exp(x_i^T\beta)$ is close to 0. When $r_i \to 0$, the bias is eliminated.

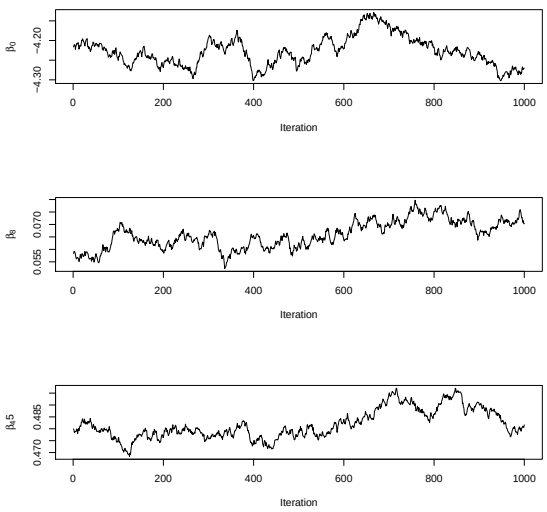
The resulting CDA is:

$$
\begin{aligned}
z_i &\sim \mathcal{PG}(\frac{1}{r_i}, x_{ij}^T\beta + \log r_i) \\
\beta &\sim N\{(x^T diag\{z_i\}x)^{-1}x^T(y - \frac{1}{2r} + z\log r), (x^T diag\{z_i\}x)^{-1}\}
\end{aligned}
\tag{4}
$$

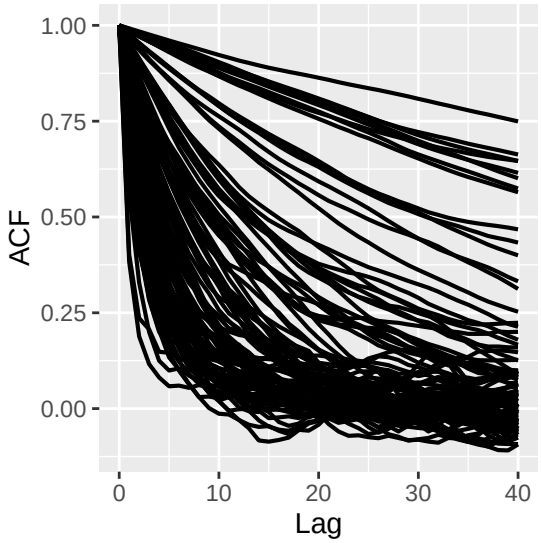
The effective approximate likelihood is $L_{r,b}(y_i|x_i^T\beta) = \frac{\Gamma(1/r_i+1)}{\Gamma(1/r_i-y_i+1)\Gamma(y_i+1)}\frac{\exp(x_i^T\beta+\log r_i)_i^y}{\{1+\exp(x_i^T\beta+\log r_i)\}^{1/r_i}}$ with $r_i < 1/(y_i-1)$. The total variation distance $||L(y_i|x_i^T\beta) - L_{r,b}(y_i|x_i^T\beta)||_{TV} \leq \frac{\sqrt{r}_i}{2}\exp(x_i^T\beta)$. The bound on the tail square integral is provided in the appendix. Given a maximally tolerable approximation error $\epsilon = 0.01$, the calculation leads to $r_i = \inf_{\theta\in\Theta^*}\{\frac{10^{-3.5}}{\exp(x_i^T\beta)}\}^2$.

We ran the uncalibrated DA with large $\lambda = 10,000$ and CDA for posterior computation. We started all three algorithms by assigning the same initial value $(x^Tx)^{-1}(x^T\log(y+1))$ to β. We ran each algorithm for 4000 steps and used the last $1,000$ as the posterior sample.

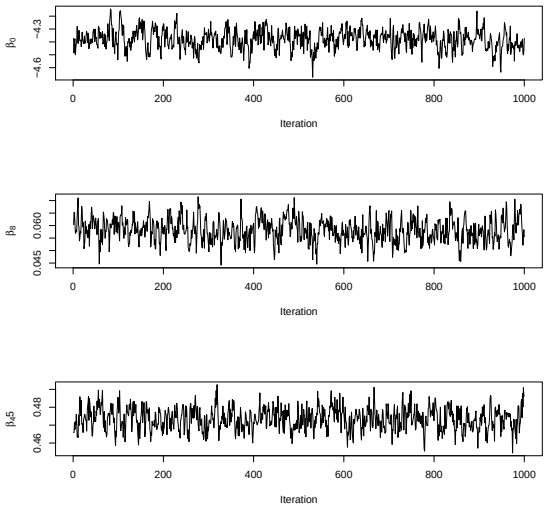
The mixing of DA and CDA is compared by traceplots and autocorrelation plots in Figure 5. DA shows slow mixing for several parameters (Figure5b), including the important intercept estimate β_0 (first plot in Figure5a). In contrast, CDA performs extremely very well in terms of mixing (Figure5d). This is shown by the low autocorrelation in all of the 96 parameters.



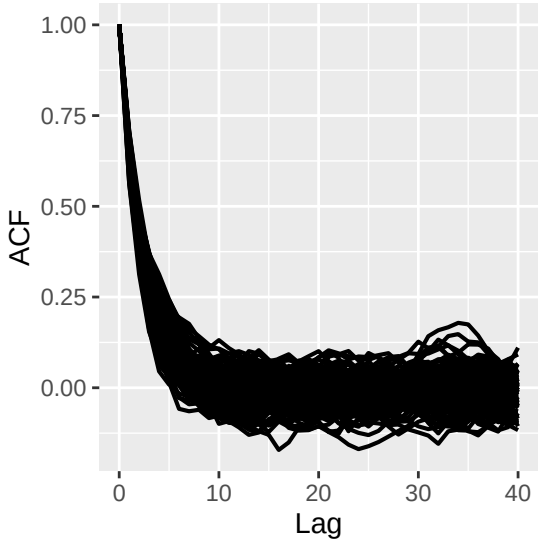
(a) Trace plots of three parameters from DA.



(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 5: Compared with DA, CDA produces much faster mixing posterior sample.

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 sum of slopes $\sum_{j=1}^{95}\beta_j$. To assess goodness-of-fit, we also evaluate 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 $\hat{\beta}$ the posterior mean. For prediction performance, we use the testing dataset collected $\{y_{new}, x_{new}\}$ on the same websites and $\tilde{y}_{i,new} = \exp(x_{i,new}\hat{\beta})$ as the estimator. We evaluate the cross-validation RMSE between $y_{i,new}$ and $\tilde{y}_{i,new}$.

As expected, the estimate for β_0 is quite negative, which is captured by the point estimates of both DA and CDA. However, DA severely underestimates the variance of the intercept. The estimates for the 95

covariates also differ greatly from CDA. Obviously, the poor mixing causes the Markov chain in DA to be trapped in a suboptimal state, whereas CDA performs exceptionally well in fit statistics and the validation error that is almost 4 times lower. More details in comparing the fitted and prediction are provided in the appendix.

To verify the result, we additionally ran Hamiltonian Monte Carlo (HMC) as the reference. HMC is known for its good mixing properties but very costly evaluation. The result of CDA agrees very well with HMC in both the posterior mean and standard deviation on all the 95 slope parameters (see appendix for comparison plots). For HMC, it requires significant tuning to reach ideal step size and length for proposal; while CDA only require one additional step in generating the latent variable. Therefore, CDA is significantly more efficient than HMC and took only 1/10 of the computing time.

	DA	CDA	HMC
β_0	-4.21 (0.042)	-4.38 (0.075)	-4.47 (0.071)
$\sum_{j=1}^{95} \beta_j$	-0.11 (0.063)	0.69 (0.053)	0.70 (0.055)
RMSE	32.86	5.06	4.88
D	182127.7	107076.9	106791.3
CV-RMSE	32.01	8.61	8.28
Steps to Converge	2000	50	500
Computing Time (per 2,000 steps)	50 mins	51 mins	600 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 sum-of-slopes are shown. The CDA shows much better fit statistics such as root-mean-squared error (RMSE) and deviance (D). In cross-validation (CV-RMSE), the CDA outperforms DA as well. The CDA converges much more rapidly than DA. Compared to the reference, CDA agrees with the HMC very well but takes significantly less time.

5 Discussion

The slow mixing is a severe problem that prevents data augmentation based MCMC from be applicable on large dataset. With data size increases and become complex, it is common for the parameters to deviate from the area that has reasonable mixing performance. As we show in the previous example, it does not only lead to an un-manageable increase in the computational time, but also could cause Markov chain to be trapped in the suboptimal state. Therefore, it is necessary to address this issue if we want to keep data augmentation useful in large data.

In this article, we propose a general class of solutions that calibrates this issue. Based on the data augmentation factorization, CDA either integrates out the parameter or increases its conditional variance, so that the step size is adjusted onto the same order of the marginal variance.

In the data application, we use Hamiltonian Monte Carlo as a good reference for parameter estimation. Here we draw a comparison between the variance increased CDA and HMC. The ideal Markov chain kernel would be to propose, based on the current state, a state that is as uncorrelated as possible; in the meantime,

with high acceptance probability to move to the new state. The HMC utilizes numerical simulation of Hamiltonian dynamics and a long walk to generate such a proposal. In contrast, the CDA directly utilizes the original density but with an increased conditional variance to reduce the correlation. The short distance between the proposal and current likelihood enables a good acceptance rate as well. The difference between the two is that the HMC is computationally costly since it relies on evaluation of gradients and multiple Hamiltonian steps for each proposal; whereas CDA does not use gradient but only one-step update, which is almost at the same low cost as the conventional DA method.

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*, (just-accepted):00–00, 2015.

Mary Kathryn Cowles. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6(2):101–111, 1996.

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.

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

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 Proof of Theorems

6.1.1 Proof of Theorem 1:

Let $\theta = \{\theta_1, \theta_2\}$ be the parameters that are divided into two parts. Let θ' and z' be the parameters and latent variables in the last iteration. Omitting y for the ease of notation, the square of maximal correlation can be represented as $||\mathbf{F}||^2 = \sup_{s(\theta,z),t(\theta,z)\in L^2(\pi)} \text{corr}\{s(\theta,z), t(\theta', z')\}^2 = \sup_{s(\theta,z)\in L^2(\pi)} \frac{\text{var}[E\{s(\theta,z)|\theta', z'\}]}{\text{var}\{s(\theta,z)\}}$.

The original DA samples in the order of $\{\theta_1', \theta_2'\} \to z' \to \{\theta_1, \theta_2\} \to z$, with $\text{var}[E\{s(\theta,z)|\theta', z'\}] = \text{var}[E\{s(\theta_1,\theta_2,z)|\theta_1', \theta_2', z'\}]$. The marginalization and sampling based CDA samples in the order of $\theta_2' \to z' \to \theta_2 \to z$, followed by $z' \to \theta_1'$ and $z \to \theta_1$ with $\text{var}[E\{s(\theta,z)|\theta', z'\}] = \text{var}[E\{s(\theta_1,\theta_2,z)|\theta_1', \theta_2', z'\}] = \text{var}[E\{s(\theta_1,\theta_2,z)|\theta_2', z'\}]$.

For better clarity, let E_X denote the integration over $P(dX)$.

$$
\begin{aligned}
\text{var}[E\{s(\theta_1,\theta_2,z)|\theta_2^{'},z'\}] &= E_{\theta_2^{'},z'}[E_{\theta_1,\theta_2,z}\{s(\theta_1,\theta_2,z)|\theta_2^{'},z'\}]^2 - (E_{\theta_2^{'},z'}[E_{\theta_1,\theta_2,z}\{s(\theta_1,\theta_2,z)|\theta_2^{'},z'\}])^2 \\
&= E_{\theta_2^{'},z'}[E_{\theta_1'}E_{\theta_1,\theta_2,z}\{s(\theta_1,\theta_2,z)|\theta_1',\theta_2^{'},z'\}]^2 - (E_{\theta_1',\theta_2^{'},z'}[E_{\theta_1,\theta_2,z}\{s(\theta_1,\theta_2,z)|\theta_1^{'},\theta_2^{'},z'\}])^2 \\
&\leq E_{\theta_2^{'},z'}E_{\theta_1'}[E_{\theta_1,\theta_2,z}\{s(\theta_1,\theta_2,z)|\theta_1',\theta_2^{'},z'\}]^2 - (E_{\theta_1',\theta_2^{'},z'}[E_{\theta_1,\theta_2,z}\{s(\theta_1,\theta_2,z)|\theta_1^{'},\theta_2^{'},z'\}])^2 \\
&= \text{var}[E\{s(\theta_1,\theta_2,z)|\theta_1^{'},\theta_2^{'},z'\}]
\end{aligned} \tag{5}
$$

This completes the proof.

6.1.2 Proof of Theorem 2:

Both DA and the variance increase CDA sample in the order of $\theta' \to z' \to \theta \to z$. Omitting y for the ease of notation, by Lemma 4 of Liu (1994), $\sup_{s(\theta)\in L^2(\pi)} \frac{\text{var}[E\{s(\theta,z)|\theta^{'},z'\}]}{\text{var}\{s(\theta,z)\}} = \sup_{s(\theta)\in L^2(\pi)} \frac{\text{var}[E\{s(\theta)|z'\}]}{\text{var}\{s(\theta)\}}$.

As $\frac{var_{IV}\{\theta|z\}}{var\{\theta|z\}} \geq (1+\epsilon)$ leads to $E[var_{IV}\{\theta|z\}] \geq (1+\epsilon)E[var\{\theta|z\}]$, $\frac{|\text{var}\{s(\theta)\}-\text{var}_{IV}\{s(\theta)\}|}{\text{var}\{s(\theta)\}} \leq \epsilon$ leads to $\text{var}_{IV}\{s(\theta)\} \leq (1+\epsilon)\text{var}\{s(\theta)\}$.

$$
1 - \frac{E[var_{IV}\{\theta|z\}]}{\text{var}_{IV}\{s(\theta)\}} \leq 1 - \frac{(1+\epsilon)E[var\{\theta|z\}]}{(1+\epsilon)\text{var}\{s(\theta)\}} \tag{6}
$$

Using $\text{var}\{s(\theta)\} = E[var\{s(\theta)|z\}] + var[E\{s(\theta)|z\}]$ and taking supremum on both sides complete the proof.

6.1.3 Proof of Theorem 3:

Without loss of generality, take $M \geq 1$, then $E|\theta_j|1_{|\theta_j|>M} \leq E\theta_j^2 1_{|\theta_j|>M} \leq \epsilon_2$.

$$
\begin{aligned}
|E(\theta_j|y) - E_{r,b}(\theta_j|y)| &= \int |\theta_j \pi(\theta_j|y) - \theta_j \pi_{r,b}(\theta_j|y)| d\theta_j \\
&\leq \int |\theta_j||\pi(\theta_j|y) - \pi_{r,b}(\theta_j|y)| d\theta_j \\
&= \int 1(|\theta_j| \leq M)|\theta_j| \cdot |\pi(\theta_j|y) - \pi_{r,b}(\theta_j|y)| d\theta_j + \int 1(|\theta_j| > M)|\theta_j||\pi(\theta_j|y) - \pi_{r,b}(\theta_j|y)| d\theta_j \\
&\leq M \int |\pi(\theta_j|y) - \pi_{r,b}(\theta_j|y)| d\theta_j + \int 1(|\theta_j| > M)|\theta_j|\pi(\theta_j|y) d\theta_j + \int 1(|\theta_j| > M)|\theta_j|\pi_{r,b}(\theta_j|y) d\theta_j \\
&\leq 2M\epsilon_1 + 2\epsilon_2 \\
&= 2M\epsilon_1 + o(\epsilon_1),
\end{aligned} \tag{7}
$$

where triangle inequality and the definition of total variation distance are used.

$$
\begin{aligned}
|\mathrm{var}(\theta_j|y) - \mathrm{var}_{r,b}(\theta_j|y)| &= |[E(\theta_j^2|y) - \{E(\theta_j|y)\}^2] - [E_{r,b}(\theta_j^2|y) - \{E_{r,b}(\theta_j|y)\}^2]| \\
&\leq |E(\theta_j^2|y) - [E_{r,b}(\theta_j^2|y)]| + |\{E(\theta_j|y)\}^2 - \{E_{r,b}(\theta_j|y)\}^2]| \\
&\leq 2M^2\epsilon_1 + 2\epsilon_2 + |\{E(\theta_j|y)\} - \{E_{r,b}(\theta_j|y)\}]| \cdot |\{E(\theta_j|y)\} + \{E_{r,b}(\theta_j|y)\}]| \\
&\leq 2M^2\epsilon_1 + 2\epsilon_2 + (2M\epsilon_1 + 2\epsilon_2)\{2E(\theta_j|y) + 2M\epsilon_1 + 2\epsilon_2\} \\
&\leq 2M^2\epsilon_1 + 2\epsilon_2 + (2M\epsilon_1 + 2\epsilon_2)\{2M + 2\epsilon_2 + 2M\epsilon_1 + 2\epsilon_2\} \\
&= 6M^2\epsilon_1 + o(\epsilon_1).
\end{aligned} \tag{8}
$$

To prove Corollary 1, using Cauchy-Schwarz inequality $\{E\theta_{j_1}\theta_{j_2}1(\theta_{j_1} > M_{j_1})1(\theta_{j_2} > M_{j_2})\}^2 \leq E\theta_{j_1}^2 1(\theta_{j_1} > M_{j_1})E\theta_{j_2}^2 1(\theta_{j_1} > M_{j_2}) = \epsilon_2^2$. Following the similar proof for variance, it can be derived that:

$$
|\mathrm{cov}(\theta_{j_1}, \theta_{j_2}|y) - \mathrm{cov}_{r,b}(\theta_{j_1}, \theta_{j_2}|y)| \leq 6M_{j_1}M_{j_2}\epsilon_1 + o(\epsilon_1).
$$

6.1.4 Proof of Theorem 4:

Since $\theta = B^- B\theta$, $\mathrm{cov}B\theta = B^- \mathrm{cov}B\theta B^{-T}$, applying Hölder's inequality:

$$
||E\theta - E\theta_{r,b}||_1 \leq ||B^-||_1 ||EB\theta - EB\theta_{r,b}||_\infty
$$

$$
||\mathrm{cov}\theta - \mathrm{cov}\theta_{r,b}||_1 \leq ||B^-||_\infty ||B^- \mathrm{cov}B\theta - B^- \mathrm{cov}B\theta_{r,b}||_1 \leq ||B^-||_1 ||B^-||_\infty ||\mathrm{cov}B\theta - \mathrm{cov}B\theta_{r,b}||_\infty
$$

6.2 Approximation Error in Logistic Regression

For better clarity, we renumber the double index ij using single index i.

6.2.1 Total Variation Distance

The individual Kullback-Leibler distance:

$$\begin{aligned} KL\{L_{r,b}(y_i|\eta_i)||L(y_i|\eta_i)\} &= \mathrm{E}\log\frac{\Gamma(1/r_i+1)r_i^{y_i}/\Gamma(1/r_i-y_i+1)}{\Gamma(2)/\Gamma(2-y_i)} + \log\frac{1+\exp(\eta_i)}{\{1+\exp(\eta_i)r_i\}^{1/r_i}} \\ &= \log\{1+\exp(\eta_i)\} - 1/r\log\{1+r\exp(\eta_i)\} \\ &\le \{(r_i-1)\frac{\exp(2\eta_i)}{2}\}1\{\exp(\eta_i)<1/r_i\} + \log\frac{1+\exp(\eta_i)}{\{1+\exp(\eta_i)r_i\}^{1/r_i}}1\{\exp(\eta_i)\ge 1/r_i\} \end{aligned} \tag{9}$$

With adaptive $r_i = 1$ if $\eta_i \ge -\log r_i$ and Pinsker's inequality,

$$||P_{r,b}(y_i|\eta_i) - P(y_i|\eta_i)||_{TV} \le \{\frac{\sqrt{r_i-1}\exp(\eta_i)}{2}\}1\{\eta_i < -\log r_i \le 0\}$$

6.2.2 Tail Integral

Consider each likelihood $L(y_i|p_i) = p_i^y(1-p)^{1-y_i}$ with $p_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$. Applying density transformation leads to $\pi(\eta_i|y_i) = \frac{\exp(\eta_i)\exp(y_i\eta_i)}{\{1+\exp(\eta_i)\}^3}$.

If $y_i = 1$,

$$\begin{aligned} \mathrm{E}\{\eta_i^2 1(|\eta_i|>M)\} &= E\{\eta_i^2 1(|\eta_i|>M, \eta_i\ge 0)\} + E\{\eta_i^2 1(|\eta_i|>M, \eta_i<0)\} \\ &\le \int_M^\infty \frac{\eta_i^2}{1+\exp(\eta_i)}d\eta_i + \int_{-\infty}^{-M}\eta_i^2\exp(2\eta_i)d\eta_i \\ &\le \int_M^\infty \eta_i^2\exp(-\eta_i)d\eta_i + \int_{-\infty}^{-M}\eta_i^2\exp(2\eta_i)d\eta_i \\ &= (M^2+2M+2)\exp(-M) + \frac{1}{4}(2M^2+2M+1)\exp(-2M). \end{aligned} \tag{10}$$

if $y_i = 0$,

$$\begin{aligned}
\mathrm{E}\{\eta_i^2 1(|\eta_i| > M)\} &= \mathrm{E}\{\eta_i^2 1(|\eta_i| > M, \eta_i \geq 0)\} + E\{\eta_i^2 1(|\eta_i| > M, \eta_i < 0)\} \\
&\leq \int_M^{\infty} \frac{\eta_i^2}{\{1+\exp(\eta_i)\}^2} d\eta_i + \int_{-\infty}^{-M} \eta_i^2 \exp(\eta_i) d\eta_i \\
&\leq \int_M^{\infty} \eta_i^2 \exp(-2\eta_i) d\eta_i + \int_{-\infty}^{-M} \eta_i^2 \exp(\eta_i) d\eta_i \\
&= \frac{1}{4}(2M^2 + 2M + 1)\exp(-2M) + (M^2 + 2M + 2)\exp(-M).
\end{aligned} \tag{11}$$

Therefore, the tail square integral is in $O(M^2 \exp(-M))$.

Consider the approximate density $L_{r,b}(y_i|\eta_i) = \frac{\Gamma(1/r_i+1)}{\Gamma(1/r_i-y_i+1)\Gamma(y_i+1)} p^{y_i}(1-p)^{(1/r_i-y_i)}$, where $p = \frac{\exp(\eta_i+\log r_i)}{\{1+\exp(\eta_i+\log r_i)\}}$ and $y_i < 1/r_i+1$. Applying density transformation leads to $\pi(\eta_i|y_i) = \frac{\Gamma(1/r_i+1)}{\Gamma(1/r_i-y_i+1)\Gamma(y_i+1)} \frac{\{r_i \exp(\eta_i)\}^{(y_i+1)}}{\{1+r_i \exp(\eta_i)\}^{(1/r_i+2)}}$.

$$\begin{aligned}
\mathrm{E}_{r,b}\{\eta_i^2 1(|\eta_i| > M)\} &\leq \int \eta_i^2 1(|\eta_i| > M) \frac{\Gamma(1/r_i+1) r_i^{y_i}}{\Gamma(1/r_i - y_i + 1)} \frac{r_i}{\Gamma(y_i+1)} \frac{\{\exp(\eta_i)\}^{(y_i+1)}}{\{1+r_i \exp(\eta_i)\}^{(1/r_i+2)}} d\eta_i \\
&\leq \int \eta_i^2 1(|\eta_i| > M) \frac{r_i}{y_i!} \frac{\{\exp(\eta_i)\}^{(y_i+1)}}{\{1+r_i \exp(\eta_i)\}^{(1/r_i+2)}} d\eta_i \\
&\leq \frac{1}{y_i! r_i^{y_i}} \int_M^{\infty} \frac{\eta_i^2}{1+r_i \exp(\eta_i)} d\eta_i + \frac{r_i}{y_i!} \int_{-\infty}^{-M} \eta_i^2 \exp\{\eta_i(y_i+1)\} d\eta_i \\
&\leq \frac{1}{y_i! r_i^{y_i+1}} \int_M^{\infty} \eta_i^2 \exp(-\eta_i) d\eta_i + \frac{r_i}{y_i!} \int_{-\infty}^{-M} \eta_i^2 \exp\{\eta_i(y_i+1)\} d\eta_i \\
&= \frac{1}{y_i! r_i^{y_i+1}} (M^2 + 2M + 2)\exp(-M) + \frac{r_i}{y_i!}(M^2 + 2M + 2)\exp(-M)
\end{aligned} \tag{12}$$

6.3 Approximation Error in Poisson Log-Linear Model

6.3.1 Total Variation Distance

With $\eta_i = x_i^T \beta$, the individual Kullback-Leibler distance:

$$KL\{L_{r,b}(y_i|\eta_i)||L(y_i|\eta_i)\} = \mathrm{E}\log\frac{\Gamma(1/r_i+1)r_i^{y_i}}{\Gamma(1/r_i-y_i+1)} + \log\frac{\exp\exp(\eta_i)}{\{1+\exp(\eta_i)r_i\}^{1/r_i}}$$

$$\leq \exp(\eta_i) - 1/r\log\{1+r\exp(\eta_i)\}$$

$$\leq \{r_i\frac{\exp(2\eta_i)}{2}\}1\{\exp(\eta_i) < 1/r_i\} + \log\frac{\exp\exp(\eta_i)}{\{1+\exp(\eta_i)r_i\}^{1/r_i}}1\{\exp(\eta_i) \geq 1/r_i\} \tag{13}$$

With adaptive $r_i = 0$ if $\eta_i \geq 1/r_i$ and Pinsker's inequality,

$$||P_{r,b}(y_i|\eta_i) - P(y_i|\eta_i)||_{TV} \leq \{\frac{\sqrt{r_i}\exp(\eta_i)}{2}\}1\{\eta_i < -\log r_i\}$$

6.3.2 Tail Integral

Consider each likelihood $L(y_i|p_i) = p_i^{y_i}\exp(-p_i)/y_i!$ with $p_i = \exp(\eta_i)$. Applying density transformation leads to $\pi(\eta_i|y_i) = \exp\{\eta_i(y_i+1)\}\exp\{-\exp(\eta_i)\}/y_i!$. Without loss of generality, assume $|M| \geq 1$.

$$E\{\eta_i^2 1(|\eta_i| > M)\} = E\{\eta_i^2 1(|\eta_i| > M, \eta_i \geq 0)\} + E\{\eta_i^2 1(|\eta_i| > M, \eta_i < 0)\}$$

$$\leq \int_M^\infty \frac{\exp\{\eta_i(y_i+3)\}\}}{\exp\{\exp(\eta_i)\}e^2 y_i!}d\eta_i + \int_{-\infty}^{-M}\frac{\eta_i^2\exp\{\eta_i(y_i+1)\}}{y_i!}d\eta_i \tag{14}$$

$$= \frac{IGamma(y_i+3, \exp(M)\}}{e^2 y_i!} + \frac{IGamma(3,(y_i+1)M)\}}{(y_i+1)^3 y_i!}.$$

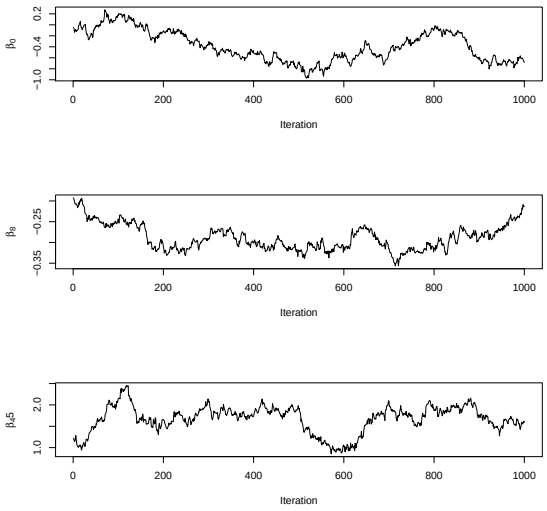
where $IGamma(a,b)$ is the incomplete Gamma function $\int_b^\infty t^{a-1}\exp(-t)dt$, equal to the $\{1-F(b)\}\Gamma(a)$, with $F(b)$ as the cumulative distribution function of gamma distribution $\mathcal{G}(a,1)$.

Similar to the logistic approximate, consider the approximate density $L_{r,b}(y_i|\eta_i) = \frac{\Gamma(1/r_i+1)}{\Gamma(1/r_i-y_i+1)\Gamma(y_i+1)}p^{y_i}(1-p)^{(1/r_i-y_i)}$, where $p = \frac{\exp(\eta_i+\log r_i)}{\{1+\exp(\eta_i+\log r_i)\}}$ and $y_i < 1/r_i+1$. Applying density transformation leads to $\pi(\eta_i|y_i) = \frac{\Gamma(1/r_i+1)}{\Gamma(1/r_i-y_i+1)\Gamma(y_i+1)}\frac{\{r_i\exp(\eta_i)\}^{(y_i+1)}}{\{1+r_i\exp(\eta_i)\}^{(1/r_i+2)}}$.

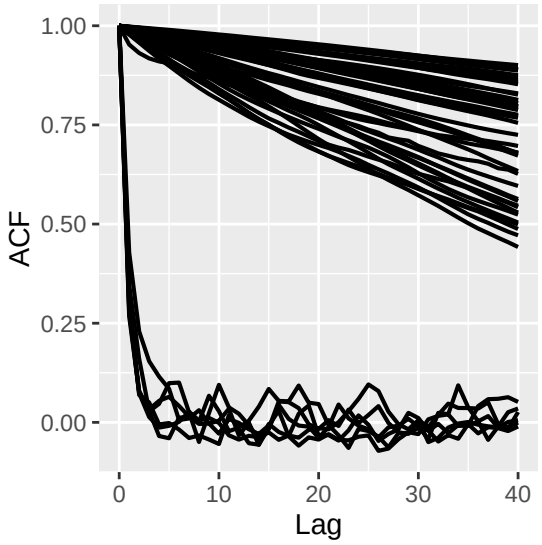
Note when $\eta_i > 0$ hence $r_i < 1/exp(0) = 1$, $\frac{1}{1+r_i\exp(\eta_i)\}^{(1/r_i)}} \leq \frac{1}{1+\exp(\eta_i)}$. Then,

$$
\begin{aligned}
\mathrm{E}_{r,b}\{\eta_i^2 1(|\eta_i| > M)\} &\leq \int \eta_i^2 1(|\eta_i| > M) \frac{\Gamma(1/r_i + 1) r_i^{y_i}}{\Gamma(1/r_i - y_i + 1)} \frac{r_i}{\Gamma(y_i + 1)} \frac{\{\exp(\eta_i)\}^{(y_i+1)}}{\{1 + r_i \exp(\eta_i)\}^{(1/r_i+2)}} d\eta_i \\
&\leq \int \eta_i^2 1(|\eta_i| > M) \frac{r_i}{y_i!} \frac{\{\exp(\eta_i)\}^{(y_i+1)}}{\{1 + r_i \exp(\eta_i)\}^{(1/r_i+2)}} d\eta_i \\
&\leq \int_M^\infty \frac{r_i}{y_i! r_i^{y_i+1}} \frac{\eta_i^2}{\{1 + r_i \exp(\eta_i)\}} d\eta_i + \frac{r_i}{y_i!} \int_{-\infty}^{-M} \eta_i^2 \exp\{\eta_i (y_i + 1)\} d\eta_i \\
&= \frac{1}{y_i! r_i^{(y_i+1)}} (M^2 + 2M + 2) \exp(-M) + \frac{r_i}{y_i!} \frac{IGamma(3, (y_i + 1)M)\}}{(y_i + 1)^3}
\end{aligned}
\tag{15}
$$

6.4 Mixing of Zero-inflated Poisson without Calibration



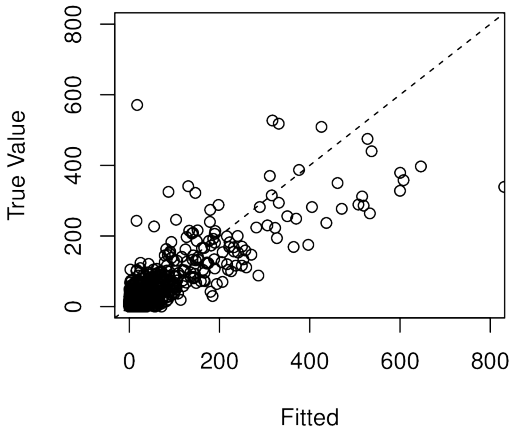
(a) Trace plots of three parameters from DA ZIP model



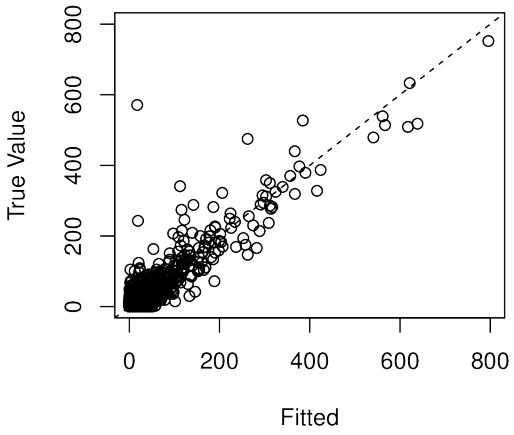
(b) Autocorrelation of all the 96 β's from DA ZIP model.

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

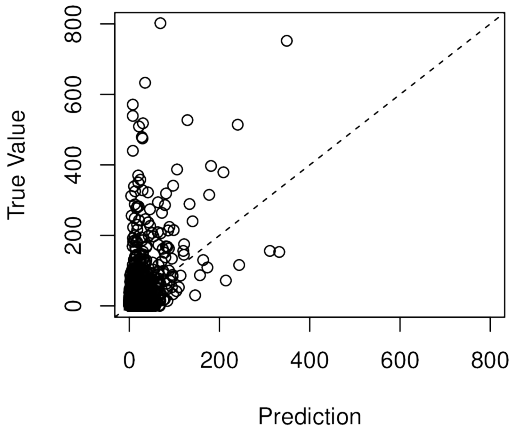
6.5 Goodness-of-Fit and Cross-Validation for Poisson Regression



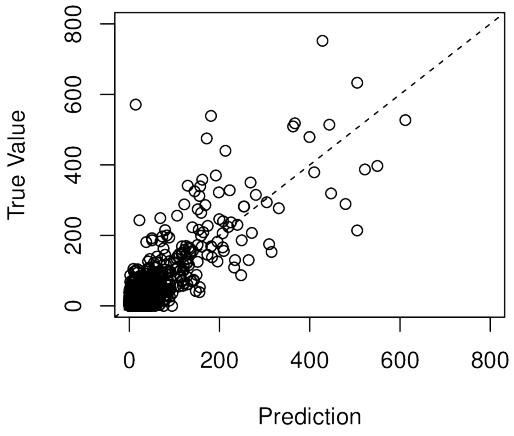
(a) Fitted vs true values using DA



(b) Fitted vs true values using CDA



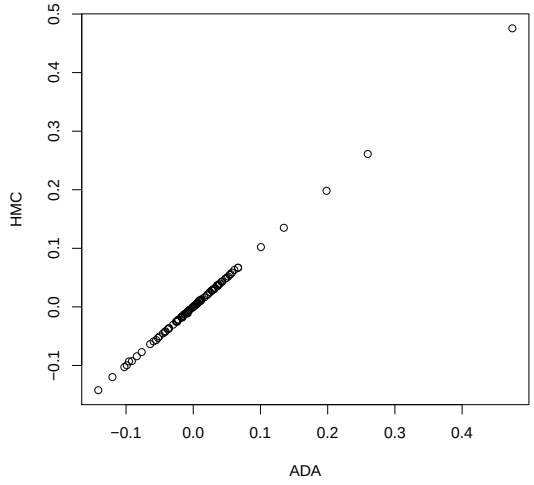
(c) Prediction vs true values using DA



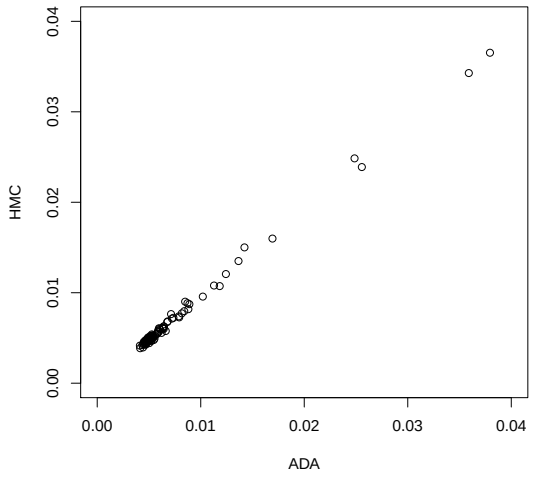
(d) Prediction vs true values using CDA

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

6.6 Comparing posterior samples of CDA with HMC



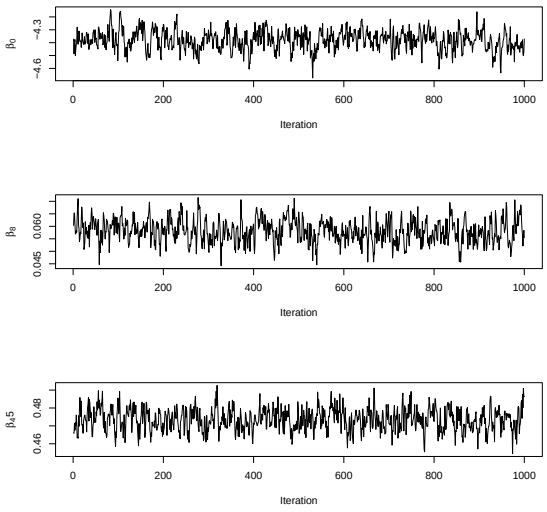
(a) Comparing posterior means for $\beta_1, \ldots \beta_{95}$ from the HMC and CDA.



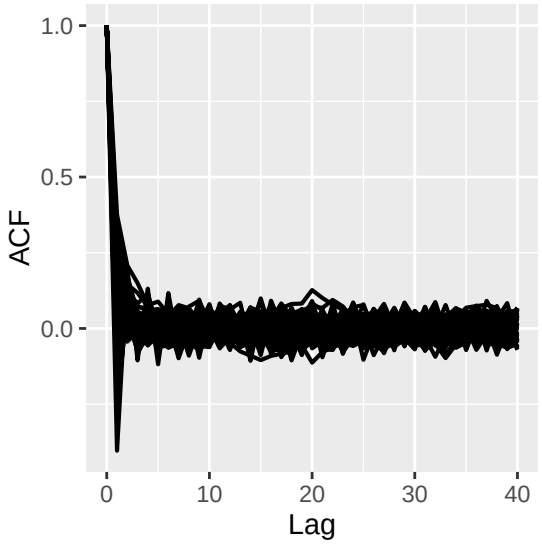
(b) Comparing posterior standard deviation for $\beta_1, \ldots \beta_{95}$ from the HMC and CDA.

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

6.7 Mixing of HMC



(a) Traceplots



(b) Autocorrelation

Figure 9: The posterior estimates produced by HMC.