Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Ultimate Pólya Gamma Samplers–Efficient MCMC for Possibly Imbalanced Binary and Categorical Data

Gregor Zens, Sylvia Frühwirth-Schnatter & Helga Wagner

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2023.2259030

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Ultimate Pólya Gamma Samplers – Efficient MCMC for Possibly Imbalanced Binary and Categorical Data

Gregor Zens[a,b], Sylvia Frühwirth-Schnatter[b], and Helga Wagner[c]

[a]International Institute for Applied Systems Analysis, Laxenburg, Austria; [b]WU Wien, Vienna, Austria; [c]JKU Linz, Linz, Austria

ABSTRACT
Modeling binary and categorical data is one of the most commonly encountered tasks of applied statisticians and econometricians. While Bayesian methods in this context have been available for decades now, they often require a high level of familiarity with Bayesian statistics or suffer from issues such as low sampling efficiency. To contribute to the accessibility of Bayesian models for binary and categorical data, we introduce novel latent variable representations based on Pólya-Gamma random variables for a range of commonly encountered logistic regression models. From these latent variable representations, new Gibbs sampling algorithms for binary, binomial, and multinomial logit models are derived. All models allow for a conditionally Gaussian likelihood representation, rendering extensions to more complex modeling frameworks such as state space models straightforward. However, sampling efficiency may still be an issue in these data augmentation based estimation frameworks. To counteract this, novel marginal data augmentation strategies are developed and discussed in detail. The merits of our approach are illustrated through extensive simulations and real data applications. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2020
Accepted July 2023

KEYWORDS
Bayesian; Data augmentation; Gibbs sampling; MCMC boosting; Parameter expansion

1. Introduction

Applied statisticians and econometricians commonly have to deal with modeling binary or categorical outcome variables. Widely used tools for analyzing such data include probit as well as binary, multinomial, and binomial logit regression models. Bayesian approaches toward inference are very useful in this context, as they allow to easily extend the standard regression framework to more complex settings such as random effects or state space models. However, as opposed to regression models with Gaussian outcomes, their implementation can be demanding from a computational viewpoint (Chopin and Ridgway 2017).

One strategy to implement sampling-based inference relies on importance sampling (Zellner and Rossi 1984) or various types of Metropolis-Hastings (MH) algorithms (Rossi, Allenby, and McCulloch 2005), exploiting directly the non-Gaussian likelihood. However, these algorithms often require careful tuning and substantial experience with Bayesian computation, especially in more complex frameworks like state space models.

Routine Bayesian computation for these type of data more often relies on Markov chain Monte Carlo (MCMC) algorithms based on data augmentation (DA, Tanner and Wong 1987). As shown by the seminal paper of Albert and Chib (1993), the binary probit model admits a latent variable representation where the latent variable equation is linear in the unknown parameters, with an error term following a standard normal distribution. As simulating the latent variables is easy when the parameters are known, the latent variable representation admits a straightforward Gibbs sampler using one level of DA, where the unknown parameters are sampled from a conditionally Gaussian model. This strategy works also for more complex models, such as probit state space or random effects models.[1]

However, MCMC estimation based on DA is less straightforward for a logit model which still admits a latent variable representation that is linear in the unknown parameters, but exhibits an error term that follows a logistic distribution. Related latent variable representations with non-Gaussian errors exist for multinomial logit (MNL) models (Frühwirth-Schnatter and Frühwirth 2010) and logistic regression models for binomial outcomes (Fussl, Frühwirth-Schnatter, and Frühwirth 2013). While the latent variables usually can be easily sampled, sampling the unknown parameters is more involved due to the non-Gaussian error terms.

A common solution relies on a scale-mixture representation of the non-Gaussian error distribution and introduces the corresponding scale parameters as a second level of DA. Conveniently, the unknown model parameters can then be sampled from a conditionally Gaussian regression model. Examples include a representation of the logistic distribution involving the Kolmogoroff-Smirnov distribution (Holmes and Held 2006) and

[1] There is also an active literature on posterior simulation tools for probit and logit regression models that does not rely on DA. For instance, Durante (2019) introduces a framework for conjugate analysis of the probit model that has been generalized subsequently, see Anceschi et al. (2023) for a review. Sen et al. (2020) use a sampling framework for logistic regression based on piecewise deterministic Monte Carlo processes. We provide a discussion of these and other alternative methods in Appendix A.1.

CONTACT Gregor Zens zens@iiasa.ac.at International Institute for Applied Systems Analysis, Laxenburg, Austria.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

highly accurate finite scale-mixture approximations (Frühwirth-Schnatter and Frühwirth 2007, 2010; Frühwirth-Schnatter et al. 2009). A seminal paper in this context is Polson, Scott, and Windle (2013) which avoids any explicit latent variable representation. They derive the *Pólya-Gamma sampler* that exploits a mixture representation of the non-Gaussian likelihood of the marginal model based on the Pólya-Gamma distribution and works with a single level of DA.

In this article, we propose a new sampling scheme involving the Pólya-Gamma distribution. Instead of working with the marginal model, we introduce a new mixture representation of the logistic distribution based on the Pólya-Gamma distribution in the latent variable representation of the logit model. Similar to Holmes and Held (2006) and Frühwirth-Schnatter and Frühwirth (2010), we use DA and introduce the Pólya-Gamma mixing variables as a second set of latent variables. Our new Pólya-Gamma mixture representation has the advantage that the joint posterior distribution of all augmented variables is easy to sample from, as the Pólya-Gamma mixing variable follows a tilted Pólya-Gamma distribution conditional on the latent utilities. This allows to sample the unknown model parameters from a conditionally Gaussian model, facilitating posterior simulation in complex frameworks such as state space or random effects models.

A commonly encountered challenge when working with MCMC methods based on DA is poor mixing. For binary and categorical regressions, this issue is especially pronounced for imbalanced data, where the success probability is either close to zero or one for the majority of the observations, see the excellent work of Johndrow et al. (2019). Neither the original Pólya-Gamma sampler of Polson, Scott, and Windle (2013) with a single level of DA, nor our new Pólya-Gamma sampler with two levels of DA, are an exception to this rule.

To resolve this issue, we introduce imbalanced marginal data augmentation (iMDA) as a *boosting* strategy to make our new sampler as well as the original probit sampler of Albert and Chib (1993) robust to possibly imbalanced data. This strategy is inspired by earlier work on marginal data augmentation (MDA) for binary and categorical data (Liu and Wu 1999; McCulloch, Polson, and Rossi 2000; van Dyk and Meng 2001; Imai and van Dyk 2005). Starting from a latent variable representation of the binary model, we expand the latent variable representation with the help of two unidentified "working parameters." One parameter is a global scale parameter for the latent variable, which has been shown to improve mixing considerably by Liu and Wu (1999), among others. However, this strategy alone does not resolve slow mixing when dealing with highly imbalanced data. To address this, we introduce an additional, unknown location parameter, which improves mixing considerably in the case of imbalanced data. As iMDA only works in the context of a latent variable representation, this strategy cannot be applied to the original Pólya-Gamma sampler of Polson, Scott, and Windle (2013) due to the lack of such a representation. In comparison, our new Pólya-Gamma representation of the logit model is very generic and is easily combined with iMDA, not only for binary regression models, but also for more flexible models such as binary state space models. We refer to a sampling strategy combining a Pólya-Gamma mixture representation with iMDA as an *ultimate* Pólya-Gamma (UPG) sampler due to its efficiency.

A further contribution of this article is to show that such an UPG sampler can be derived for other non-Gaussian regression problems, including models for categorical and binomial data. For the MNL model, commonly a logit model based on a (partial) differenced random utility model (dRUM) representation is applied to sample the category specific parameters, see for example Holmes and Held (2006), Frühwirth-Schnatter and Frühwirth (2010), or Polson, Scott, and Windle (2013). Using this partial dRUM representation, we derive a new sampler for the MNL model in this article. Since the latent variable equation is linear in the unknown parameters and involves a logistic error distribution, we use once more the Pólya-Gamma mixture representation of the logistic distribution and introduce the mixing variables as additional latent variables. For binomial models, a latent variable representation which did not involve a choice equation was introduced by Fussl, Frühwirth-Schnatter, and Frühwirth (2013). Since an explicit choice equation is needed to apply iMDA, we derive a new latent variable representation for binomial data which involves error terms that follow generalized logistic distributions. We introduce Pólya-Gamma mixture representations of these distributions and use the resulting auxiliary variables as an additional latent layer. Both for MNL models and for binomial models, this DA scheme leads to a conditionally Gaussian posterior and allows to sample all unknowns through efficient block moves. Again, we apply iMDA to derive UPG samplers which mix well, also in the context of imbalanced data.

Overall, we find that the various algorithms show highly competitive performance when compared to alternative DA frameworks, which we demonstrate via extensive simulation studies. In addition, we present real world data examples that further illustrate the merits of our approach. The underlying algorithms for probit regression and logistic regression models for binary, categorical and binomial outcomes have been made available in the R package `UPG`, which is available on *CRAN* (Zens, Frühwirth-Schnatter, and Wagner 2021).

The remainder of the article is structured as follows. Section 2 introduces the UPG sampler. This sampling strategy is extended to categorical data in Section 3 and to binomial data in Section 4. In Section 5, the UPG sampler is compared to alternative DA algorithms. Section 6 applies the framework to binary state space models and discusses the utility of the approach in the context of mixture-of-experts models. Section 7 concludes.

2. Ultimate Pólya-Gamma Samplers for Binary Data

2.1. Latent Variable Representations for Binary Data

Models for a vector of N binary observations $\mathbf{y} = (y_1, \ldots, y_N)$ are defined by

$$\Pr(y_i = 1|\lambda_i) = F_\varepsilon(\log \lambda_i), \tag{1}$$

where λ_i depends on exogenous variables and unknown parameters $\boldsymbol{\beta}$, for example, $\log \lambda_i = \mathbf{x}_i \boldsymbol{\beta}$ in a standard binary regression model. Choosing the cdf $F_\varepsilon(\varepsilon) = \Phi(\varepsilon)$ of the standard normal distribution leads to the probit model $\Pr(y_i = 1|\lambda_i) = \Phi(\log \lambda_i)$, whereas the cdf $F_\varepsilon(\varepsilon) = e^\varepsilon/(1 + e^\varepsilon)$ of the logistic distribution leads to the logit model

$$\Pr(y_i = 1|\lambda_i) = \lambda_i/(1 + \lambda_i).$$

A latent variable representation of model (1) involving a latent utility z_i is given by:

$$y_i = I\{z_i > 0\}, \;\; z_i = \log \lambda_i + \varepsilon_i, \quad \varepsilon_i \sim f_\varepsilon(\varepsilon_i), \tag{2}$$

where $f_\varepsilon(\varepsilon) = F'_\varepsilon(\varepsilon) = \phi(\varepsilon)$ is equal to the standard normal pdf for a probit model and equal to $f_\varepsilon(\varepsilon) = e^\varepsilon/(1+e^\varepsilon)^2$ for a logit model.

In Bayesian inference, the set of observed data $\mathbf{y} = (y_1, \dots, y_N)$ can be augmented with the latent variables $\boldsymbol{z} = (z_1, \dots, z_N)$ in (2) to obtain the set of complete data $(\boldsymbol{z}, \mathbf{y})$, facilitating the implementation of MCMC algorithms. As shown by Albert and Chib (1993), this single level of DA involving $\boldsymbol{z}$ leads to a straightforward Gibbs sampler for the probit model. With $\log \lambda_i = \mathbf{x}_i \boldsymbol{\beta}$, the following two-step sampling *Scheme 1* can be set up under a Gaussian prior $p(\boldsymbol{\beta})$:

(Z) Given $\boldsymbol{\beta}$, sample the latent variables z_i for each $i = 1, \dots, N$ independently from $p(z_i|\boldsymbol{\beta}, \mathbf{y})$ (see Appendix A.4.1);
(P) sample the unknown parameters $\boldsymbol{\beta}$ conditional on $\boldsymbol{z}$ from the Gaussian posterior $p(\boldsymbol{\beta}|\boldsymbol{z}, \mathbf{y})$ derived from regression model (2).

Two main challenges are associated with such MCMC schemes, namely slow convergence and a lack of closed form posteriors for the unknown parameters, such as $p(\boldsymbol{\beta}|\boldsymbol{z}, \mathbf{y})$, outside of probit models. We address both issues in this article.

First, to boost MCMC convergence, we rely on MDA in the spirit of Liu and Wu (1999). In that paper, the scale-based transformation $\tilde{z}_i = \sqrt{\delta} z_i$, depending on a "working parameter" δ, is used to define the expanded probit regression model

$$y_i = I\{\tilde{z}_i > 0\}, \;\; \tilde{z}_i = \sqrt{\delta}\mathbf{x}_i \boldsymbol{\beta} + \tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \sim \mathcal{N}(0, \delta). \tag{3}$$

In model (3), the likelihood $p(\tilde{\boldsymbol{z}}|\delta)$ of $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \dots, \tilde{z}_N)$, marginalized w.r.t. $\boldsymbol{\beta}$, is available in closed form and yields an inverse Gamma posterior $p(\delta|\tilde{\boldsymbol{z}})$ under a conjugate prior $p(\delta)$. Assuming prior independence of δ and $\boldsymbol{\beta}$, this allows to rescale the latent variables $\boldsymbol{z}$ without involving $\boldsymbol{\beta}$. Specifically, a draw $\tilde{\delta}$ from the working prior $p(\delta)$ is used to "propose" a scale-move $\tilde{z}_i = \sqrt{\tilde{\delta}} z_i$ in system (3), based solely on prior information. Then, an updated value δ^{new} is sampled from the posterior $p(\delta|\tilde{\boldsymbol{z}})$ and the proposed scale-move is immediately "corrected" (using a posteriori information) via the inverse transformation $z_i^{\text{new}} = \tilde{z}_i/\sqrt{\delta^{\text{new}}}$, before $\boldsymbol{\beta}$ is updated conditional on $\boldsymbol{z}^{\text{new}}$. This extends *Scheme 1* to *Scheme 2*:

(Z) Sample from $p(\boldsymbol{z}|\boldsymbol{\beta}, \mathbf{y})$ as in *Scheme 1*;
(B-S) move from $\boldsymbol{z}$ to $\boldsymbol{z}^{\text{new}}$ using a scale-based expansion move under prior $p(\delta)$;
(P) sample from $p(\boldsymbol{\beta}|\boldsymbol{z}^{\text{new}}, \mathbf{y})$ as in *Scheme 1*.

The boosted *Scheme 2* always provides better convergence results than *Scheme 1*, see van Dyk and Meng (2001) and Hobert and Marchev (2008) for further theoretical results. Indeed, as an example in Liu and Wu (1999) illustrates, Step (B-S) improves efficiency considerably in cases where the coefficient of determination in the latent regression model is large, as long as the data are balanced. However, DA schemes are in general known to be slowly mixing for imbalanced datasets where only a few cases with $y_i = 1$ or $y_i = 0$ among the N data points are observed (Johndrow et al. 2019). Indeed, sampling under *Scheme 2* is still highly inefficient in such cases, as will be illustrated in Section 2.2.

A first major contribution of this article is to protect DA algorithms for binary and categorical data against imbalanced data by using, in addition to a scale-based transformation, a location-based expansion $\tilde{z}_i = z_i + \gamma$, depending on a "working parameter" γ, to define the expanded version

$$y_i = I\{\tilde{z}_i > \gamma\}, \;\; \tilde{z}_i = \gamma + \log \lambda_i + \varepsilon_i, \tag{4}$$

of the binary regression model (2).

As opposed to (3), the choice equation in (4) depends on γ and defines a likelihood $p(\mathbf{y}|\gamma, \tilde{\boldsymbol{z}})$. In a probit regression model, the likelihood $p(\tilde{\boldsymbol{z}}|\gamma)$ of the latent data, marginalized w.r.t. $\boldsymbol{\beta}$, is available in closed form. In combination with the likelihood $p(\mathbf{y}|\gamma, \tilde{\boldsymbol{z}})$ and a Gaussian working prior $p(\gamma)$, a Gaussian posterior $p(\gamma|\tilde{\boldsymbol{z}}, \mathbf{y})$, truncated to the interval $[L, U)$ defined by, respectively, the maximum utility L of the outcomes where $y_i = 0$ and the minimum utility U of the outcomes where $y_i = 1$, is obtained. Assuming prior independence of γ and $\boldsymbol{\beta}$ then allows to shift the latent variables $\boldsymbol{z}$ without involving $\boldsymbol{\beta}$. Similar to the scale-based expansion, a location-move $\tilde{z}_i = z_i + \tilde{\gamma}$ is proposed using a draw $\tilde{\gamma}$ from the working prior $p(\gamma)$, before being immediately "corrected" via the inverse transformation $z_i^{\text{new}} = \tilde{z}_i - \gamma^{\text{new}} = z_i + \tilde{\gamma} - \gamma^{\text{new}}$ using a draw γ^{new} from the posterior distribution $p(\gamma|\tilde{\boldsymbol{z}}, \mathbf{y})$, see Section 2.3 for further details. Subsequently, the regression coefficients $\boldsymbol{\beta}$ are sampled conditional on $\boldsymbol{z}^{\text{new}}$. We find that performing such a location-based expansion step before a scale-based transformation yields dramatic improvement compared to *Scheme 1* and *Scheme 2*, also in cases where the data are imbalanced, see Sections 2.2 and 5 for further illustration.

A second main contribution of this article is to take location-based and scale-based parameter expansion beyond the probit regression model by introducing new latent variable representations for binary, binomial and multinomial logit models. For binary logit models, a second level of DA is introduced to deal with the logistic error term. For this, we apply a new mixture representation of the logistic distribution,

$$f_\varepsilon(\varepsilon_i) = e^{\varepsilon_i}/(1+e^{\varepsilon_i})^2 = \frac{1}{4}\int e^{-\omega_i \varepsilon_i^2/2} p(\omega_i)\, d\omega_i, \tag{5}$$

where $\omega_i \sim \mathcal{PG}(2, 0)$ follows a Pólya-Gamma distribution (Polson, Scott, and Windle 2013), see Appendix A.2.1 and A.2.2 for details. This representation is very convenient, as the conditional posterior $\omega_i \mid \varepsilon_i \sim \mathcal{PG}(2, |\varepsilon_i|)$ of ω_i given ε_i is a tilted Pólya-Gamma distribution which is easy to sample from, see Polson, Scott, and Windle (2013). For a binary logit model with $\log \lambda_i = \mathbf{x}_i \boldsymbol{\beta}$, this new representation allows constructing a Pólya-Gamma sampler that extends *Scheme 1* in the following way:

(Z) sample the latent variable z_i from $p(z_i|\boldsymbol{\beta}, y_i)$ independently for each i in the latent variable model (2) (see Algorithm 1 and Appendix A.4.1) and sample the scale parameter ω_i conditional on z_i and $\boldsymbol{\beta}$ from $\omega_i|z_i, \boldsymbol{\beta} \sim \mathcal{PG}(2, |z_i - \mathbf{x}_i\boldsymbol{\beta}|)$;
(P) sample the unknown parameters $\boldsymbol{\beta}$ conditional on the latent variables $\mathbf{z} = (z_1, \dots, z_N)$ and $\boldsymbol{\omega} = (\omega_1, \dots, \omega_N)$ from the conditionally Gaussian posterior $p(\boldsymbol{\beta}|\boldsymbol{\omega}, \mathbf{z}, \mathbf{y})$.

While this scheme is easy to implement, it can be slowly mixing, like any such sampler. To deal with this issue, we additionally include the two parameter expansion steps introduced above, performing first a location-based and then a scale-based transformation. We refer to the resulting sampling scheme as *Scheme 3* and provide full theoretical and computational details in Section 2.3. In later sections, we extend this strategy to logistic regression models for categorical and binomial outcomes.

While our boosting strategy is inspired by Liu and Wu (1999) and related to earlier work on MDA for binary and categorical data (McCulloch, Polson, and Rossi 2000; van Dyk and Meng 2001; Imai and van Dyk 2005), it generalizes this literature in several aspects. Importantly, it works for any binary data model with a latent variable representation. In addition, freeing the location of the threshold γ in model (4) leads to an MCMC scheme that is well mixing, even in cases of extremely imbalanced data, see much of the remainder of this article for further illustration. A related strategy to improve mixing behavior in the context of data augmentation algorithms is outlined in Duan, Johndrow, and Dunson (2018). In their contribution, the authors use location and scale parameters to reparametrize augmented likelihood functions in binary and count data regression models. These calibration parameters have to be set manually, and the authors propose an involved optimization procedure based on large sample arguments and approximations to determine suitable values. The resulting algorithms show efficiency gains that are comparable to the marginal data augmentation proposed in this article when analyzing datasets with many observations and rare outcomes. A potential downside of the approach of Duan, Johndrow, and Dunson (2018) is that the optimization procedure relies on the inverse of the observed Fisher information and the sampler uses Metropolis-Hastings updates. Both may result in scaling issues when many covariates are present. In such settings, a pure data augmentation approach as proposed in this article may prove more effective. Importantly, our approach is also fully automatic and does not rely on any approximations in the sense that tuning-free and exact Gibbs updates for the location and scale parameters are derived. We give details on the resulting posterior simulation scheme in Section 2.3. Before presenting these details, we illustrate the specific roles of the location and scale parameters γ and δ using heuristic arguments in the next section.

2.2. Illustration and Intuition

As a first illustration of the potential merits of the proposed iMDA scheme in imbalanced logistic regression settings, we compare estimation efficiency of the popular Pólya-Gamma sampler from Polson, Scott, and Windle (2013) with a plain DA sampler as in *Scheme 1*, a scale-based parameter expansion scheme (as in *Scheme 2*) and the proposed approach based on location-based and scale-based expansion (as in *Scheme 3*) in Figure 1. A more systematic comparison will be given in Section 5. It is clearly visible that the UPG sampler outperforms all other samplers in terms of efficiency. Notably, these efficiency gains are realized despite introducing two layers of latent auxiliary variables, which usually increases autocorrelation in the posterior draws significantly. This is counteracted by our novel iMDA strategy based on the working parameters γ and δ.

We start with the role of δ, the working parameter used for scale-based expansion of the latent utility equation. Broadly speaking, this scale-based expansion will be highly effective in scenarios where the coefficient of determination in the latent utility model is high. In such settings, the current parameter draw almost perfectly determines the location of the latent utilities and vice versa. As a result, the MCMC chain is only able to move very slowly. To resolve this issue, δ artificially decreases the coefficient of determination via increasing the error variance in the latent utility equation. In turn, this decreases the dependency of the latent utilities and the regression coefficients, directly enabling larger steps of the Markov chain. In other words, δ is used to make the posterior of the latent utilities in the expanded model more diffuse than the posterior of the utilities in the original model. Similar as well as more formal arguments and further illustration of such scale-based expansion steps have been discussed for instance in Liu and Wu (1999) or Imai and van Dyk (2005).

However, a scale-based expansion alone is usually not enough to fully resolve the issue that step sizes become small relative to the range of the high posterior density region in imbalanced data settings (Johndrow et al. 2019). This can be seen from the unsatisfactory performance of the PX-DA sampler in Figure 1 and has also been discussed in Duan, Johndrow, and Dunson (2018). In our approach, this issue is effectively offset through the location-based expansion of the latent utility model. In this section, we aim to illustrate the mechanism behind this strategy

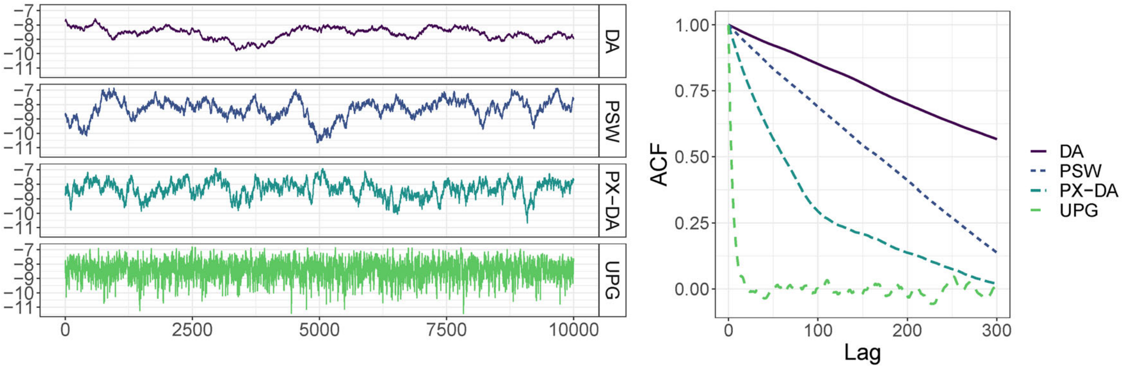


Figure 1. MCMC draws and corresponding autocorrelation functions of an intercept-only logistic regression model fitted using plain data augmentation (DA, *Scheme 1*), the original Pólya-Gamma sampler (PSW), a MDA sampler with scale-based expansion (PX-DA, *Scheme 2*) and a DA sampler with scale- and location-based expansion (UPG, *Scheme 3*). Two out of $N = 10{,}000$ binary observations are nonzero.

through a small numerical exercise and defer full details to Section 2.3.

To investigate how the location-based expansion influences step sizes of the Markov chain, we consider three datasets with $N = 100$ observations each. One dataset is balanced, while the others are imbalanced, with success probabilities 99% and 1%, respectively. We simulate 25,000 replications of a single MCMC iteration for a grid of starting positions of the intercept β_0, using $\mathcal{N}(0, 100)$ prior distributions for both β_0 and γ. For each starting position and for each replication, we save the absolute step size of a plain DA sampler (*Scheme 1*) and the step size of a sampler with an additional location-based expansion step, as well as the realized shift $\tilde{\gamma} - \gamma^{\text{new}}$ in the sampler including the location-based expansion step.

The results are summarized in Figure 2. The left panel shows the log average step size of the plain DA scheme. It is evident that step sizes decrease significantly when exploring posterior regions that reach far into the positive (negative) part of the real line in imbalanced scenarios with high (low) success probabilities. The purpose of the location-based expansion is to counteract this issue via shifting the utilities by $\tilde{\gamma} - \gamma^{\text{new}}$, directly leading to larger step sizes of the Markov chain. The average shift for each dataset and value of β_0 is depicted in the middle panel of Figure 2. The magnitude of the shift, $|\tilde{\gamma} - \gamma^{\text{new}}|$, is equivalent to the increase in step size in the location-expanded sampler. While step sizes increase everywhere, the improvement is particularly large in the tails of the posterior density in imbalanced datasets, where standard DA algorithms are usually highly inefficient. In addition, the shift-move evidently acts as a "push into the right direction" that systematically leads the Markov chain back toward the highest posterior density region, effectively avoiding staying in the tails of the posterior distribution for too long. The log average step sizes of the location-expanded sampler are shown in the right panel of Figure 2. As expected from the preceding discussion, the most significant step size improvements are observed in the tail regions of the posterior distribution in the imbalanced cases.

2.3. MCMC Details for Binary Logit Regression Models

The latent utility representation of the binary logit model is

$$y_i = I\{z_i > 0\}, \;\; z_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{LO}, \tag{6}$$

where $\mathcal{LO}$ is the logistic distribution. We assume $\boldsymbol{\beta} \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{A}_0)$ follows a multivariate Gaussian distribution a priori, where $\boldsymbol{A}_0$ is either fixed or equipped with a hierarchical structure, for example, to define a shrinkage prior (see e.g., Piironen and Vehtari 2017). The first block of the MCMC scheme consists of two steps that simulate the two sets of latent variables, $\mathbf{z}$ and $\boldsymbol{\omega}$. Given $\boldsymbol{\beta}$ and the outcome y_i, we sample z_i for each i from $p(z_i|\lambda_i, y_i)$ in the logistic model (6) where $\log \lambda_i = \mathbf{x}_i \boldsymbol{\beta}$. Then, the Pólya-Gamma scale parameters are simulated from $\omega_i | z_i, \boldsymbol{\beta} \sim \mathcal{PG}(2, |z_i - \mathbf{x}_i \boldsymbol{\beta}|)$.

For given latent variables, a location-based parameter expansion step, based on a working prior $p(\gamma) = \mathcal{N}(0, G_0)$, is then applied. For this, a prior draw $\tilde{\gamma} \sim \mathcal{N}(0, G_0)$ is used to "propose", for each $i = 1, \ldots, N$, a location move $\tilde{z}_i = z_i + \tilde{\gamma}$ in the expanded model

$$y_i = I\{\tilde{z}_i > \gamma\}, \qquad \tilde{z}_i = \gamma + \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \tag{7}$$

while ω_i is unaffected. Conditional on the latent variables $\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_N)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$, but marginally w.r.t. $\boldsymbol{\beta}$, the conditional distribution $\gamma | \boldsymbol{\omega}, \tilde{\mathbf{z}} \sim \mathcal{N}(g_N, G_N)$ is Gaussian where:

$$\begin{aligned} G_N &= (G_0^{-1} + \sum_{i=1}^{N} \omega_i - m_b^\top \mathbf{B}_N m_b)^{-1}, \\ g_N &= G_N(m_\gamma - m_b^\top \mathbf{B}_N \mathbf{m}_N(\tilde{\mathbf{z}})), \\ \mathbf{B}_N &= (\boldsymbol{A}_0^{-1} + \sum_{i=1}^{N} \omega_i \mathbf{x}_i^\top \mathbf{x}_i)^{-1}, \end{aligned} \tag{8}$$

$$\mathbf{m}_N(\tilde{\mathbf{z}}) = \sum_{i=1}^{N} \omega_i \mathbf{x}_i^\top \tilde{z}_i, \quad m_b = \sum_{i=1}^{N} \omega_i \mathbf{x}_i^\top, \quad m_\gamma = \sum_{i=1}^{N} \omega_i \tilde{z}_i,$$

as is easily shown, see Appendix A.4.2. Since the choice equation in (7) depends on γ, $p(\gamma|\boldsymbol{\omega}, \tilde{\mathbf{z}})$ has to be combined with the likelihood $p(\mathbf{y}|\gamma, \tilde{\mathbf{z}})$ of the observed outcomes $\mathbf{y} = (y_1, \ldots, y_N)$ to define the posterior $p(\gamma|\boldsymbol{\omega}, \tilde{\mathbf{z}}, \mathbf{y})$. The derivation of the likelihood $p(\mathbf{y}|\gamma, \tilde{\mathbf{z}})$ is a generic step in our sampler which does not involve the specification of λ_i:

$$\begin{aligned} p(\mathbf{y}|\gamma, \tilde{\mathbf{z}}) &\propto \prod_{i:y_i=0} I\{\gamma \geq \tilde{z}_i\} \prod_{i:y_i=1} I\{\gamma < \tilde{z}_i\} \\ &\propto I\{L(\tilde{\gamma}) \leq \gamma < U(\tilde{\gamma})\}, \end{aligned} \tag{9}$$

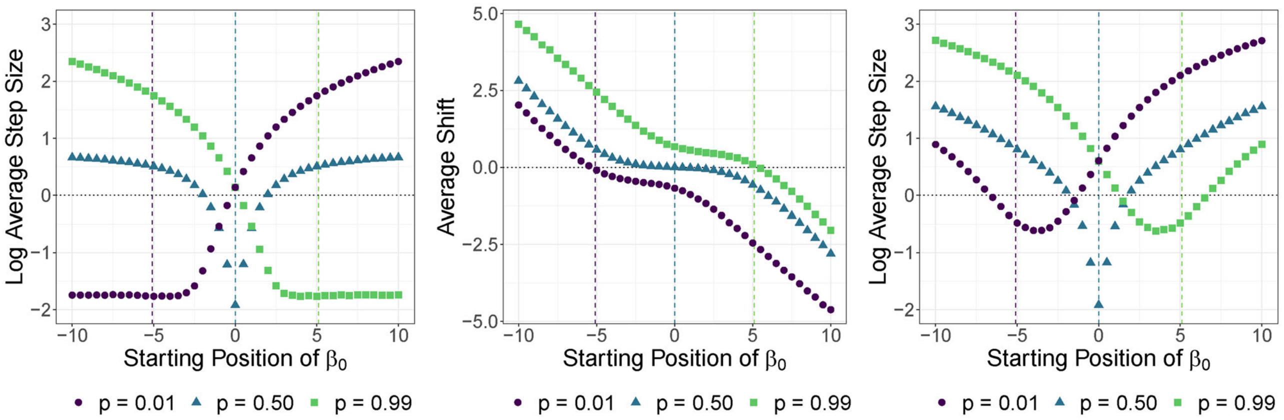


Figure 2. Illustration of the mechanism behind the location-based expansion. (left) Log average step size of a plain DA sampler. (middle) Realized shift of utilities. (right) Log average step size of a sampler with location-based expansion. Dotted lines are the means of the posterior distribution of β_0 under a $N(0, 100)$ prior.

where $I\{\cdot\}$ is the indicator function and $L(\tilde{\gamma}) = \max_{i:y_i=0} \tilde{z}_i = \max_{i:y_i=0} z_i + \tilde{\gamma}$ and $U(\tilde{\gamma}) = \min_{i:y_i=1} \tilde{z}_i = \min_{i:y_i=1} z_i + \tilde{\gamma}$. If no outcome $y_i = 0$ is observed, then $L(\tilde{\gamma}) = -\infty$; if no outcome $y_i = 1$ is observed, then $U(\tilde{\gamma}) = +\infty$. Hence, $p(\gamma|\boldsymbol{\omega}, \tilde{\mathbf{z}}, \mathbf{y}) \propto p(\mathbf{y}|\gamma, \tilde{\mathbf{z}}) p(\gamma|\boldsymbol{\omega}, \tilde{\mathbf{z}})$ is equal to a truncated version of the Gaussian posterior (8):

$$\gamma|\boldsymbol{\omega}, \tilde{\mathbf{z}}, \mathbf{y} \sim \mathcal{N}\left(g_N, G_N\right) I\{L(\tilde{\gamma}) \leq \gamma < U(\tilde{\gamma})\}. \tag{10}$$

An updated working parameter γ^{new} is sampled from (10) and the proposed location-based move is 'corrected' based on a posteriori information by defining the shifted utilities $z_i^L = \tilde{z}_i - \gamma^{\text{new}} = z_i + \tilde{\gamma} - \gamma^{\text{new}}$.

This location-based move is followed by a scale-based expansion, using an inverse Gamma $\mathcal{G}^{-1}(d_0, D_0)$ working prior $p(\delta)$. Similar to before, $\tilde{\delta}$ is sampled from $p(\delta)$ and used to propose, for each $i = 1, \ldots, N$, a scale-based move $\tilde{z}_i = \sqrt{\tilde{\delta}} z_i^L$ in the expanded model

$$y_i = I\{\tilde{z}_i > 0\}, \qquad \tilde{z}_i = \sqrt{\delta}\mathbf{x}_i\boldsymbol{\beta} + \sqrt{\delta}\varepsilon_i. \tag{11}$$

Conditional on the Pólya-Gamma scale parameters ω_i, it follows that

$$\begin{aligned} p(\tilde{z}_i|\omega_i, \delta, \boldsymbol{\beta}) &\propto \frac{1}{\sqrt{\delta}} \exp\left\{-\frac{\omega_i}{2}\left(\frac{\tilde{z}_i}{\sqrt{\delta}} - \mathbf{x}_i\boldsymbol{\beta}\right)^2\right\} \\ &= \frac{1}{\sqrt{\delta}} \exp\left\{-\frac{\omega_i}{2}\left(\sqrt{\frac{\tilde{\delta}}{\delta}} z_i^L - \mathbf{x}_i\boldsymbol{\beta}\right)^2\right\}. \end{aligned}$$

Hence, conditional on δ, $\tilde{\delta}$ and the shifted utilities $\mathbf{z}^L = (z_1^L, \ldots, z_N^L)$, the posterior $\boldsymbol{\beta}|\delta, \tilde{\delta}, \mathbf{z}^L, \boldsymbol{\omega} \sim \mathcal{N}\left(\sqrt{\tilde{\delta}/\delta}\mathbf{b}_N, \mathbf{B}_N\right)$ is Gaussian with $\mathbf{b}_N = \mathbf{B}_N \mathbf{m}_N(\mathbf{z}^L)$ and $\mathbf{m}_N(\mathbf{z}^L)$ and $\mathbf{B}_N$ as in (8). Furthermore, conditional on $\mathbf{z}^L$, but marginally w.r.t. $\boldsymbol{\beta}$, the posterior $\delta|\tilde{\delta}, \mathbf{z}^L, \boldsymbol{\omega} \sim \mathcal{G}^{-1}\left(d_N, D_N(\tilde{\delta})\right)$ is inverse Gamma with following moments:

$$\begin{aligned} d_N &= d_0 + \frac{N}{2}, \\ D_N(\tilde{\delta}) &= D_0 + \frac{\tilde{\delta}}{2}\left(\sum_{i=1}^{N} \omega_i (z_i^L - \mathbf{x}_i\mathbf{b}_N)^2 + \mathbf{b}_N^\top \boldsymbol{A}_0^{-1}\mathbf{b}_N\right). \end{aligned} \tag{12}$$

An updated working parameter δ^{new} is sampled from $\mathcal{G}^{-1}\left(d_N, D_N(\tilde{\delta})\right)$ and the proposed scale-based move is corrected by defining the rescaled utilities $z_i^{LS} = \sqrt{\tilde{\delta}/\delta^{\text{new}}} z_i^L$. This concludes the scale-based expansion and $\boldsymbol{\beta}|\mathbf{z}^{LS}, \boldsymbol{\omega}$ is sampled conditional on $\mathbf{z}^{LS}$ or, equivalently, from the Gaussian posterior $\boldsymbol{\beta}|\delta^{\text{new}}, \tilde{\delta}, \mathbf{z}^L, \boldsymbol{\omega} \sim \mathcal{N}\left(\sqrt{\tilde{\delta}/\delta^{\text{new}}}\mathbf{b}_N, \mathbf{B}_N\right)$. As Algorithm 1 illustrates, many steps in this ultimate Pólya-Gamma (UPG) sampler are generic and easily extended to more complex models for binary data, as will be illustrated in Section 6.

Algorithm 1 The ultimate Pólya-Gamma sampler for binary data.

```
Choose starting values for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ and repeat the following steps:

(Z)   For each $i = 1, \ldots, N$, sample $z_i = \log \lambda_i + F_\varepsilon^{-1}(y_i + U_i(1 - y_i - \pi_i))$ in model (2), where $U_i \sim \mathcal{U}[0, 1]$, $\pi_i = F_\varepsilon(\log \lambda_i)$, and $F_\varepsilon^{-1}(p) = \Phi^{-1}(p)$ for the probit and $F_\varepsilon^{-1}(p) = \log p - \log(1 - p)$ for the logit model. For a logit model, sample $\omega_i|z_i, \log \lambda_i \sim \mathcal{PG}\left(2, |z_i - \log \lambda_i|\right)$.
(B-L) Location-based parameter expansion: sample $\tilde{\gamma} \sim \mathcal{N}(0, G_0)$ and propose utilities $\tilde{z}_i = z_i + \tilde{\gamma}$ for $i = 1, \ldots, N$. Sample $\gamma^{\text{new}}$ from $\gamma|\boldsymbol{\omega}, \tilde{\mathbf{z}}, \mathbf{y}$ and define shifted utilities $z_i^L = \tilde{z}_i - \gamma^{\text{new}}$. For a binary regression model, $p(\gamma|\boldsymbol{\omega}, \tilde{\mathbf{z}}, \mathbf{y})$ is given by the truncated Gaussian-posterior in (10).
(B-S) Scale-based parameter expansion: sample $\tilde{\delta} \sim \mathcal{G}^{-1}(d_0, D_0)$ and sample $\delta^{\text{new}}$ from $\delta|\tilde{\delta}, \mathbf{z}^L, \boldsymbol{\omega}$. Define rescaled utilities $z_i^{LS} = \sqrt{\tilde{\delta}/\delta^{\text{new}}} z_i^L$. For a binary regression model, $\delta|\tilde{\delta}, \mathbf{z}^L, \boldsymbol{\omega} \sim \mathcal{G}^{-1}\left(d_N, D_N(\tilde{\delta})\right)$ is an inverse Gamma distribution, with $d_N$ and $D_N(\tilde{\delta})$ given by (12).
(P)   Sample the unknown parameter in $\log \lambda_i$ conditional on $\mathbf{z}^{LS}$. For a binary regression model, $\boldsymbol{\beta}|\delta^{\text{new}}, \tilde{\delta}, \mathbf{z}^L, \boldsymbol{\omega} \sim \mathcal{N}\left(\sqrt{\tilde{\delta}/\delta^{\text{new}}}\mathbf{B}_N\mathbf{m}_N(\mathbf{z}^L), \mathbf{B}_N\right)$ where $\mathbf{m}_N(\mathbf{z}^L)$ and $\mathbf{B}_N$ are given by (8).
```

3. Ultimate Pólya-Gamma Samplers for Categorical Data

Let $\{y_i\}$, $i = 1, \ldots, N$, be a sequence of categorical data, where y_i is equal to one of at least three unordered categories. The categories are labeled by $L = \{0, \ldots, m\}$, and for any k the set of all categories but k is denoted by $L_{-k} = L \setminus \{k\}$. We assume that the observations are mutually independent and that for each $k \in L$ the probability of y_i taking the value k depends on covariates $\mathbf{x}_i$ in the following way:

$$\Pr(y_i = k|\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_m) = \pi_{ki}(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_m) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_k)}{\sum_{l=0}^{m} \exp(\mathbf{x}_i\boldsymbol{\beta}_l)}, \tag{13}$$

where $\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_m$ are category specific unknown parameters of dimension d. To make the model identifiable, the parameter $\boldsymbol{\beta}_{k_0}$ of a baseline category k_0 is set equal to $\mathbf{0}$: $\boldsymbol{\beta}_{k_0} = \mathbf{0}$. Thus, the parameter $\boldsymbol{\beta}_k$ is relative to the baseline category k_0 in terms of the change in log-odds. In the following, we assume without loss of generality that $k_0 = 0$. A more general version of the multinomial logit (MNL) model (again with baseline $k_0 = 0$) reads:

$$\Pr(y_i = k|\boldsymbol{\beta}) = \lambda_{ki} / \left(1 + \sum_{l=1}^{m} \lambda_{li}\right), \tag{14}$$

where $\lambda_{1i}, \ldots, \lambda_{mi}$ depend on unknown parameters $\boldsymbol{\beta}$, while $\lambda_{0i} = 1$. For the standard MNL regression model (13), for instance, $\log \lambda_{ki} = \mathbf{x}_i\boldsymbol{\beta}_k$ for $k = 1, \ldots, m$.

Our starting point is writing the MNL model as a random utility model (RUM), see McFadden (1974):

$$u_{ki} = \log \lambda_{ki} + \epsilon_{ki}, \quad k = 0, \ldots, m, \tag{15}$$

$$y_i = k \Leftrightarrow u_{ki} = \max_{l \in L} u_{li}. \tag{16}$$

Thus, the observed category is equal to the category with maximal utility. If the errors $\epsilon_{0i}, \ldots, \epsilon_{mi}$ in (15) are iid random variables from an extreme value (EV) distribution, then the MNL model (14) results as marginal distribution of the categorical variable y_i.

Conditional on y_i, the posterior distribution $p(\mathbf{u}_i|\lambda_{ki}, y_i)$ of the latent utilities $\mathbf{u}_i = (u_{0i}, \ldots, u_{mi})$ is of closed form and easy to sample from, see Proposition 1 which is proven in Appendix A.3.

Proposition 1. Given y_i, realizations from the distribution $p(u_{0i}, \dots, u_{mi}|\lambda_{1i} \dots, \lambda_{mi}, y_i)$ can be represented as

$$e^{-u_{ki}} = -\frac{\log U_i}{1+\sum_{l=1}^{m} \lambda_{li}} - \frac{\log V_{ki}}{\lambda_{ki}} I\{y_i \neq k\}, \quad k = 0, \dots, m, \tag{17}$$

where U_i and $V_{0i}, \dots, V_{mi}$ are $m+1$ iid uniform random numbers.

Utilizing Proposition 1 together with a mixture approximation of the extreme value distribution to sample all unknown parameters jointly via two levels of data augmentation (Frühwirth-Schnatter and Frühwirth 2007) turned out to be inefficient. Noting that the choice equation (16) can be rewritten as a choice between any category k and all its alternatives in L_{-k}, Frühwirth-Schnatter and Frühwirth (2010) derive the partial dRUM representation of a RUM model and show that it has an explicit form if the errors $\epsilon_{0i}, \dots, \epsilon_{mi}$ in (15) are iid random variables from an extreme value distribution.

For a multinomial regression model this yields following well-known representation (see e.g., Holmes and Held 2006):

$$\begin{aligned} z_{ki} &= \mathbf{x}_i\boldsymbol{\beta}_k - \xi_{ki}(\boldsymbol{\beta}_{-k}) + \varepsilon_{ki}, \quad \varepsilon_{ki} \sim \mathcal{LO}, & (18)\\ y_i &= \begin{cases} k, & z_{ki} > 0, \\ \neq k, & z_{ki} \leq 0. \end{cases} & (19) \end{aligned}$$

where the error term ε_{ki} follows a logistic distribution, $z_{ki} = u_{ki} - \max_{\ell \in L_{-k}} u_{\ell i}$ is the utility gap between category k and all its alternatives and the offset $\xi_{ki}(\boldsymbol{\beta}_{-k})$ is defined as

$$\xi_{ki}(\boldsymbol{\beta}_{-k}) = \log\left(1 + \sum_{\ell \neq \{k,0\}} \exp(\mathbf{x}_i\boldsymbol{\beta}_\ell)\right).$$

While Frühwirth-Schnatter and Frühwirth (2010) use a very accurate finite mixture approximation for the logistic distribution for MCMC estimation, in this article we derive an ultimate Pólya-Gamma sampler based on the partial dRUM representation and proceed similarly as in Section 2. We use Proposition 1 to sample the utilities $u_{0i}, \dots, u_{mi}$ in the RUM model (15) and to define the utility gap z_{ki} between category k and all its alternatives. Given the utility gap z_{ki}, we exploit the Pólya-Gamma mixture representation of the logistic distribution in (18) with category specific latent variables ω_{ki} which are sampled from $\omega_{ki}|\boldsymbol{\beta}, z_{ki} \sim \mathcal{PG}(2, |\varepsilon_{ki}|)$, where $\varepsilon_{ki} = z_{ki} - \mathbf{x}_i\boldsymbol{\beta}_k + \xi_{ki}(\boldsymbol{\beta}_{-k})$.

To handle imbalanced data, we apply location- and scale-based boosting as in Section 2 with category-specific working parameters γ_k and δ_k. For instance, location-based boosting using $\tilde{z}_{ki} = z_{ki} + \tilde{\gamma}_k$ where $\tilde{\gamma}_k \sim \mathcal{N}(0, G_0)$, yields the following expanded model:

$$\begin{aligned} \tilde{z}_{ki} &= \gamma_k + \mathbf{x}_i\boldsymbol{\beta}_k - \xi_{ki}(\boldsymbol{\beta}_{-k}) + \varepsilon_{ki}, \quad \varepsilon_{ki} \sim \mathcal{LO}, & (20)\\ y_i &= \begin{cases} k, & \tilde{z}_{ki} > \gamma_k, \\ \neq k, & \tilde{z}_{ki} \leq \gamma_k. \end{cases} & (21) \end{aligned}$$

Conditional on the latent variables $\boldsymbol{\omega}_k = (\omega_{k1}, \dots, \omega_{kN})$ and $\tilde{\mathbf{z}}_k = (\tilde{z}_{k1}, \dots, \tilde{z}_{kN})$, (20) defines a Gaussian posterior distribution $p(\gamma_k|\boldsymbol{\beta}_{-k}, \boldsymbol{\omega}_k, \tilde{\mathbf{z}}_k)$, marginally w.r.t. $\boldsymbol{\beta}_k$. Similarly as in Section 2, the choice equation (21) defines a likelihood function $p(\mathbf{y}|\gamma_k, \tilde{\mathbf{z}}_k)$ which restricts γ_k to the interval $[L(\tilde{\gamma}_k), U(\tilde{\gamma}_k))$, where $L(\tilde{\gamma}_k) = \max_{y_i \neq k} z_{ki} + \tilde{\gamma}_k$ and $U(\tilde{\gamma}_k) = \min_{y_i = k} z_{ki} + \tilde{\gamma}_k$. Full details on the UPG sampler for multinomial logistic regression models are provided in Appendix A.4.3.

4. Ultimate Pólya-Gamma Samplers for Binomial Data

In this section, we consider models with binomial outcomes, that is, models of the form

$$y_i \sim \text{BiNom}(N_i, \pi_i), \qquad \text{logit}\, \pi_i = \log \lambda_i, \qquad i = 1, \dots, N, \tag{22}$$

with $\log \lambda_i = \mathbf{x}_i\boldsymbol{\beta}$ for a standard binomial regression model. As shown in Johndrow et al. (2019), Bayesian inference for binomial regression models based on the Pólya-Gamma sampler (Polson, Scott, and Windle 2013) is sensitive to imbalanced data. Similarly, the latent variable representation of binomial models of Fussl, Frühwirth-Schnatter, and Frühwirth (2013) is sensitive to imbalanced data, as we will show in Section 5. As for a logit model (which results for $N_i \equiv 1$), applying iMDA would be an option to improve mixing. However, Fussl, Frühwirth-Schnatter, and Frühwirth (2013) provide no explicit choice equation, which is needed for iMDA. The goal of this section is to define an UPG sampler which combines a new latent variable representation of binomial models, based on Pólya-Gamma mixture representations of generalized logistic distributions, with iMDA to protect the algorithm against imbalanced data.

4.1. A New Latent Variable Representation for Binomial Data

In Theorem 2, we introduce a new latent variable representation for binomial outcomes where two latent variable equations, both linear in $\log \lambda_i$, with error terms following generalized logistic distributions are utilized. An explicit choice equation is provided which relates latent variables w_i and v_i to the observed binomial outcome y_i. We show in Theorem 3 that, conditional on y_i, the posterior distribution of the latent variables is of closed form and easy to sample from, see Appendix A.3. for a proof of both theorems.

Theorem 2 (Latent variable representation of a binomial model). For $0 < y_i < N_i$, a binomial logistic model has the following random utility representation:

$$\begin{aligned} w_i &= \log \lambda_i + \varepsilon_{w,i}, \quad \varepsilon_{w,i} \sim \mathcal{GL}_{\text{II}}(k), \qquad (23)\\ v_i &= \log \lambda_i + \varepsilon_{v,i}, \quad \varepsilon_{v,i} \sim \mathcal{GL}_{\text{I}}(N_i - k), \\ y_i &= k \Leftrightarrow w_i > 0,\ v_i \leq 0, \end{aligned}$$

where $\mathcal{GL}_{\text{I}}(\nu)$ and $\mathcal{GL}_{\text{II}}(\nu)$ are, respectively, the generalized logistic distributions of type I and type II. For $y_i = 0$, the model reduces to

$$v_i = \log \lambda_i + \varepsilon_{v,i}, \quad \varepsilon_{v,i} \sim \mathcal{GL}_{\text{I}}(N_i), \quad y_i = 0 \Leftrightarrow v_i \leq 0.$$

For $y_i = N_i$, the model reduces to

$$w_i = \log \lambda_i + \varepsilon_{w,i}, \quad \varepsilon_{w,i} \sim \mathcal{GL}_{\text{II}}(N_i), \quad y_i = N_i \Leftrightarrow w_i > 0.$$

For $N_i = 1$, the logistic model results, as both $\mathcal{GL}_{\text{I}}(\nu)$ and $\mathcal{GL}_{\text{II}}(\nu)$ reduce to a logistic distribution for $\nu = 1$. For $y_i = 0$, $z_i = v_i$, whereas for $y_i = 1$, $z_i = w_i$, and the choice equation reduces to $y_i = I\{z_i > 0\}$.

Theorem 3 (Sampling the utilities in the binomial RUM). Given y_i and holding all model parameters in λ_i fixed, the latent variables $w_i|\lambda_i, (y_i > 0)$ and $v_i|\lambda_i, (y_i < N_i)$ are conditionally independent. The distributions of $w_i|\lambda_i, (y_i > 0)$ and $v_i|\lambda_i, (y_i < N_i)$ are equal in distribution to

$$w_i = \log\left((1+\lambda_i)\frac{1}{W_i^{1/y_i}} - \lambda_i\right), \quad y_i > 0, \tag{24}$$

$$v_i = -\log\left(\frac{1+\lambda_i}{\lambda_i}\frac{1}{V_i^{1/(N_i-y_i)}} - \frac{1}{\lambda_i}\right), \quad y_i < N_i, \tag{25}$$

where W_i and V_i are iid uniform random numbers.

4.2. Ultimate Pólya-Gamma Samplers for Binomial Data

The two main building blocks for the UPG sampler for binomial data are a Gaussian mixture representation of the involved generalized logistic distributions based on the Pólya-Gamma distribution and the application of iMDA to handle imbalanced data.

A random variable ε following the generalized logistic distribution of type I or II can be represented as a normal mixture,

$$\begin{aligned} f_\varepsilon(\varepsilon) &= c(a,b)\frac{(e^\varepsilon)^a}{(1+e^\varepsilon)^b} \\ &= \frac{c(a,b)}{2^b}\exp(\kappa\varepsilon)\int_0^\infty \exp(-\frac{\omega\varepsilon^2}{2})p(\omega)d\omega, \end{aligned} \tag{26}$$

with $\kappa = a - b/2$ and the Pólya-Gamma distribution $\omega \sim \mathcal{PG}(b, 0)$, introduced by Polson, Scott, and Windle (2013) serving as mixing measure, see Appendix A.2.1 to A.2.3. For $y_i > 0$, the type II generalized logistic distribution $\varepsilon_{w,i} \sim \mathcal{GL}_{\text{II}}(y_i)$ in (23) has such a representation with:

$$\kappa_{w,i} = \frac{1-y_i}{2}, \qquad \omega_{w,i} \sim \mathcal{PG}(y_i+1, 0),$$

see (A.11). Similarly, for $y_i < N_i$, the type I generalized logistic distribution $\varepsilon_{v,i} \sim \mathcal{GL}_{\text{I}}(N_i - y_i)$ in (23) has such a representation with

$$\kappa_{v,i} = \frac{N_i - y_i - 1}{2}, \qquad \omega_{v,i} \sim \mathcal{PG}(N_i - y_i + 1, 0),$$

see (A.7). Note that $\kappa_{w,i} = 0$ for $y_i = 1$ and $\kappa_{v,i} = 0$ for $y_i = N_i - 1$. Hence, for $N_i = 1$, the Pólya-Gamma mixture approximation (5) of a logistic model involving $\mathcal{PG}(2, 0)$ results. For $N_i > 1$, $\kappa_{v,i} > 0$ for $0 \le y_i \le N_i - 2$ and $\kappa_{w,i} < 0$ for $2 \le y_i \le N_i$. This leads to a slightly more challenging sampler than for binary and multinomial models.

For each $i = 1, \ldots, N$, we introduce the latent variables $\mathbf{z}_i = (w_i, \omega_{w,i}, v_i, \omega_{v,i})$, if $0 < y_i < N_i$, $\mathbf{z}_i = (w_i, \omega_{w,i})$, if $y_i = N_i$, and $\mathbf{z}_i = (v_i, \omega_{v,i})$, if $y_i = 0$. Conditional on λ_i, the latent variables $w_i|\lambda_i, (y_i > 0)$ and $v_i|\lambda_i, (y_i < N_i)$ are sampled from Theorem 3 without conditioning on $\omega_{w,i}$ and $\omega_{v,i}$. Given w_i and v_i, the parameters $\omega_{w,i}|w_i, (y_i > 0), \lambda_i$ and $\omega_{v,i}|v_i, (y_i < N_i), \lambda_i$ are independent and follow (tilted) Pólya-Gamma distributions:

$$\begin{aligned} \omega_{w,i}|w_i, y_i, \lambda_i &\sim \mathcal{PG}(y_i + 1, |w_i - \log\lambda_i|), \quad y_i > 0, \\ \omega_{v,i}|v_i, y_i, \lambda_i &\sim \mathcal{PG}(N_i - y_i + 1, |v_i - \log\lambda_i|), \quad y_i < N_i. \end{aligned} \tag{27}$$

To handle imbalanced data, we apply location- and scale-based boosting as in the previous sections, based on the working parameters γ and δ. Location-based boosting, for instance, uses $\tilde{\gamma} \sim \mathcal{N}(0, G_0)$ to define $\tilde{w}_i = w_i + \tilde{\gamma}$ and $\tilde{v}_i = v_i + \tilde{\gamma}$ in the following expanded version of model (23) with an explicit choice equation involving γ:

$$\begin{aligned} \tilde{w}_i &= \gamma + \log\lambda_i + \varepsilon_{w,i}, \quad y_i > 0, \\ \tilde{v}_i &= \gamma + \log\lambda_i + \varepsilon_{v,i}, \quad y_i < N_i, \end{aligned} \tag{28}$$

$$y_i = k \Leftrightarrow \begin{cases} \tilde{v}_i \le \gamma < \tilde{w}_i, & 0 < k < N_i, \\ \gamma \ge \tilde{v}_i, & k = 0, \\ \gamma < \tilde{w}_i, & k = N_i. \end{cases} \tag{29}$$

Full details on the UPG sampler for binomial data are provided in Appendix A.4.4.

5. Comparison with Other Sampling Strategies

This section compares the proposed sampling framework with other DA approaches for posterior simulation in binary and categorical regression models. Specifically, we conduct a large scale simulation study to establish the efficiency of our approach in imbalanced scenarios relative to other DA approaches. However, from a practical point of view, a number of alternative estimation algorithms that do not rely on DA are available for binary and categorical regression modeling. These algorithms can be highly efficient, and relying on them is often a reasonable choice. Hence, a thorough discussion of the unique advantages and disadvantages of the DA strategy outlined in this article—and DA schemes in general—is warranted, and we provide such a discussion in Appendix A.1.

A set of systematic simulations is carried out to compare the efficiency of our approach to other popular Bayesian sampling schemes that involve DA. The main results are based on simulations with varying levels of imbalancedness, where imbalancedness is either induced by fixing the number of successes at two and increasing the sample size, or fixing the sample size at $N = 1000$ and varying the intercept term in the data generating process. Each Markov chain was run for 10,000 iterations after an initial burn-in period of 2000 iterations. To gain robustness with respect to the computed inefficiency factors, each simulation is repeated 100 times and median results across these replications are reported. The computation of the inefficiency factors is based on an estimate of the spectral density of the posterior chain evaluated at zero.[2] In this section, we present results on various logistic regression models, while additional results for probit regression models and tabulated simulation results can be found in Appendix A.6.

For binary logistic regression, we compare the sampling scheme outlined in Section 2.3 (UPG), the Pólya-Gamma sampler of Polson, Scott, and Windle (2013) (PSW) and the auxiliary mixture DA scheme outlined in Frühwirth-Schnatter and Frühwirth (2010) (FSF). To assess sampling efficiency for the MNL model, we compare the MNL sampler proposed in Section 3 (UPG) with the sampling scheme of Polson, Scott, and Windle (2013) (PSW) and the partial dRUM sampler of Frühwirth-Schnatter and Frühwirth (2010) (FSF) in a setting with three categories. For the simulations with varying sample sizes, the first two categories are observed twice each and the

[2] Estimating the spectral density at zero is accomplished via R package `coda` (Plummer et al. 2006) and is based on fitting an autoregressive process to the posterior draws.

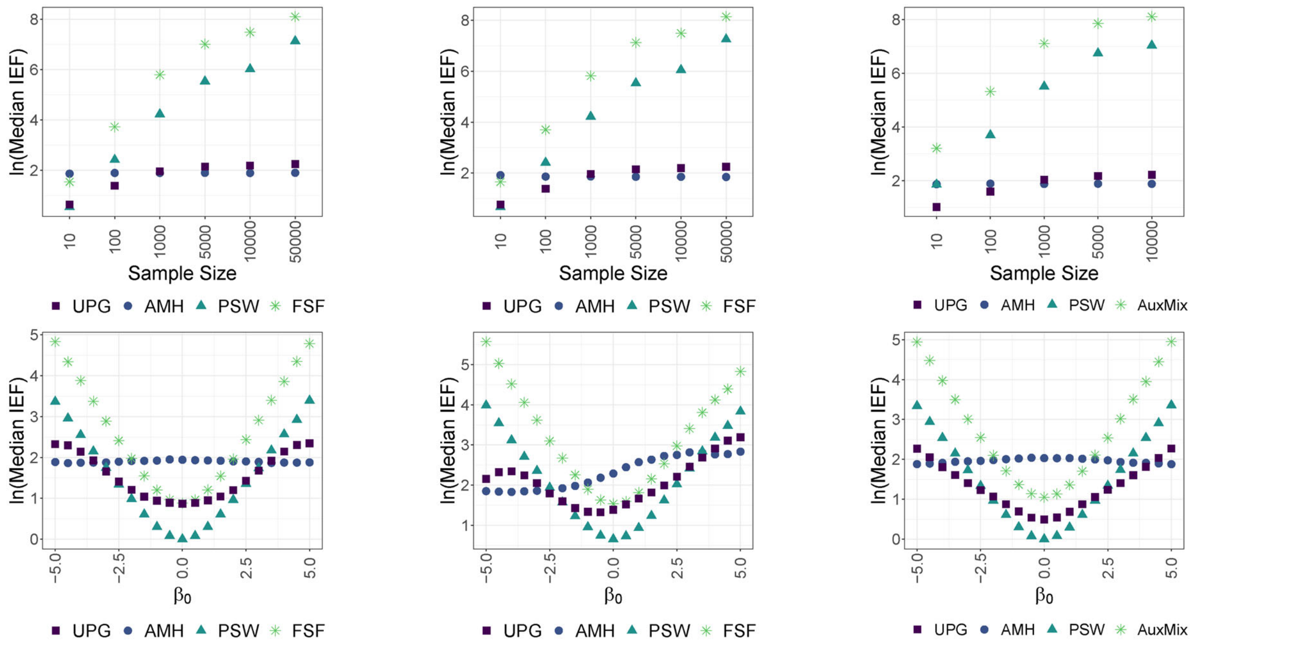


Figure 3. Sampling efficiency of intercept β_0 fitted to datasets with increasing sample size N and two successes (top) and varying intercepts β_0 with $N = 1000$ (bottom) for binary logistic regression (left), multinomial logistic regression (middle) and binomial logistic regression (right). Y-axis is on the log-scale and results are medians across 100 replications.

remaining $N - 4$ observations fall into the baseline category. For the varying intercept simulations, the intercept of the first category is varied while the other intercepts are fixed at zero. Finally, to illustrate the efficiency gains in the case of logistic regression analysis of binomial data, we compare the approach outlined in Section 4 (UPG) to the sampling scheme of Polson, Scott, and Windle (2013) (PSW) and to the auxiliary mixture sampler introduced in Fussl, Frühwirth-Schnatter, and Frühwirth (2013) (AuxMix). For all observations, we assume $N_i = 5$ trials. In all simulations, an adaptive Metropolis-Hastings sampler (AMH) is included as a benchmark as well. Throughout all simulation settings, independent $\mathcal{N}(0, 10)$ priors are specified on the regression parameters, and we choose $\gamma \sim \mathcal{N}(0, 100)$ and $\delta \sim \mathcal{IG}(2.5, 1.5)$ as working prior for the iMDA algorithms.

The results of the main simulation exercise are summarized in Figure 3. The empirical inefficiency factors confirm that standard DA techniques exhibit extremely inefficient sampling behavior when confronted with imbalanced data, as shown theoretically and empirically in Johndrow et al. (2019). The MDA strategy we propose alleviates this issue and allows for rather efficient estimation also in highly imbalanced data settings.

6. Applications to More Complex Models

6.1. Application to a Binary State Space Model

Let $\{y_t\}$ be a time series of binary observations, observed for $t = 1, \ldots, T$, taking one of two possible values labeled $\{0, 1\}$. The probability that y_t takes the value 1 depends on covariates $\mathbf{x}_t$, including a constant, through time-varying parameters $\boldsymbol{\beta}_t$ as follows:

$$\Pr(y_t = 1|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T) = \frac{\exp(\mathbf{x}_t\boldsymbol{\beta}_t)}{1 + \exp(\mathbf{x}_t\boldsymbol{\beta}_t)}. \tag{30}$$

We assume that conditional on knowing $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T$, the observations are mutually independent. A commonly used model for describing the time-variation of $\boldsymbol{\beta}_t$ reads:

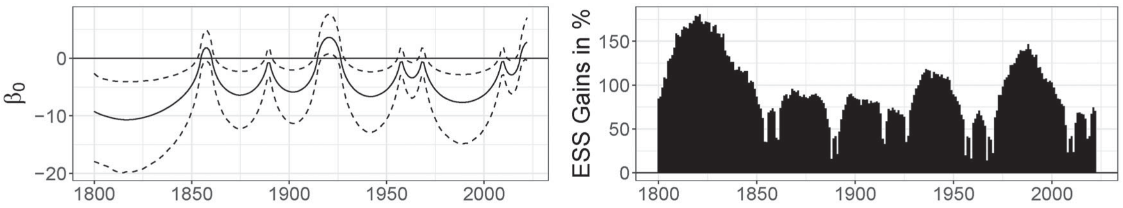
$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}\right), \tag{31}$$

with $\boldsymbol{\beta}_0 \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{P}_{0|0}\right)$ and $\mathbf{Q} = \text{Diag}\left(\theta_1, \ldots, \theta_d\right)$, where $\theta_1, \ldots, \theta_d$ are unknown variances. MCMC estimation of binary state space models (SSM) is challenging. Single-move sampling of $\boldsymbol{\beta}_t$ is potentially very inefficient (Shephard and Pitt 1997), while blocked MH updates require suitable proposal densities in a high-dimensional space (Gamerman 1998). Within the DA framework, a latent utility z_t of choosing category 1 is introduced for each y_t:

$$y_t = 1 \Leftrightarrow z_t > 0, \qquad z_t = \mathbf{x}_t\boldsymbol{\beta}_t + \varepsilon_t. \tag{32}$$

Given $\mathbf{z} = \{z_t\}$, this SSM is conditionally Gaussian for a probit link, but conditionally non-Gaussian for a logit link. Frühwirth-Schnatter and Frühwirth (2007) implemented an auxiliary mixture sampler for a binary logit SSM. Alternatively, using the Pólya-Gamma mixture representation of the logistic distribution of ε_t yields a conditionally Gaussian SSM which allows multi-move sampling of the entire state process $\boldsymbol{\beta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\}$ using FFBS (Frühwirth-Schnatter 1994; Carter and Kohn 1994) in a similar fashion as for a probit SSM. To achieve robustness against imbalance, we extend the iMDA scheme introduced in Section 2 to SSMs, see Appendix A.5 for details.

To illustrate the gains in sampling efficiency for binary SSMs, we apply the UPG framework to an example dataset on severe global pandemics. The data covers $T = 222$ years from 1800 to 2022 and documents disease episodes characterized by a worldwide spread and a death toll of more than 75,000. In addition, we focus on diseases that are characterized by relatively short periods of activity, hence, excluding pandemics such as



(a) Posterior of Time-Varying Constant.

(b) Gains in Effective Sample Size.

Figure 4. Panel (a) shows the posterior of a local level model fitted to the global pandemic data (solid line: posterior mean, dashed lines: 0.05 and 0.95 posterior quantiles). Panel (b) shows the percentage gains in effective sample size when iMDA is applied, averaged across 10 independent chains.

HIV/AIDS. This results in a total of eight pandemic events falling into the sample period, starting with a bubonic plague outbreak between 1855 and 1860 and ending with the global outbreak of COVID-19, starting in 2019.[3]

For years featuring a global pandemic, $y_t = 1$, and $y_t = 0$ otherwise. A pandemic is observed in roughly 1 out of 8 years with high state persistence, rendering the dataset relatively imbalanced. We fit a logistic local level model to the data, once with and once without iMDA, using $\theta \sim \mathcal{IG}(5, 2)$ and $P_{0|0} = 100$ as prior settings. The Gibbs sampler is iterated 100,000 times after an initial burn-in period of 10,000 iterations. This numerical study is repeated 10 times. One of the resulting posterior distributions (based on the UPG sampler) is shown in Panel (a) of Figure 4. The time-varying intercept evolves smoothly, as is typical for binary state space models. The estimated path is characterized by long periods without severe pandemics, interrupted by short pandemic episodes. In Panel (b), the percentage gains in effective sample size of the sampler with iMDA relative to the plain sampler are plotted for each year. The iMDA scheme described in Appendix A.5 is able to significantly improve sampling efficiency in all years. The most pronounced gains—up to 200% improvement in effective sample size—are observed during prolonged "imbalanced" periods where the outcome does not change. Averaging across all periods, the inefficiency factors are roughly halved, from about 96 in the plain sampler to around 45 in the UPG sampling scheme.

6.2. Application to Logistic Mixture-of-Experts Regression Models

Let y_i ($i = 1, \ldots, N$) be a grouped binary outcome with $C_i = j$ denoting that observation i belongs to group $j = 1, \ldots, J$. A logistic mixture-of-experts regression model with H ($h = 1, \ldots, H$) components takes the form

$$\begin{aligned} p(y_i \mid C_i = j, \boldsymbol{x}_i, \boldsymbol{w}_j) &= \sum_{h=1}^{H} \eta_{jh}(\boldsymbol{w}_j) Ber(\zeta_{ih}(\boldsymbol{x}_i)) \\ \zeta_{ih}(\boldsymbol{x}_i) &= \frac{\exp(\boldsymbol{x}_i \boldsymbol{\beta}_h)}{1 + \exp(\boldsymbol{x}_i \boldsymbol{\beta}_h)} \\ \eta_{jh}(\boldsymbol{w}_j) &= \frac{\exp(\boldsymbol{w}_j \boldsymbol{\psi}_h)}{\sum_{l=1}^{H} \exp(\boldsymbol{w}_j \boldsymbol{\psi}_l)} \end{aligned} \tag{33}$$

where H logistic regression "experts" are used to model cluster-specific success probabilities $\zeta_{ih}(\boldsymbol{x}_i)$ using individual-level covariates $\boldsymbol{x}_i$ and a multinomial logistic regression plays the role of a "gating function", modeling the mixture weights $\eta_{jh}(\boldsymbol{w}_j)$ based on group-level covariates $\boldsymbol{w}_j$. This model has good approximation properties (Jiang and Tanner 1999) and is popular in model-based clustering and ensemble learning. Furthermore, developing efficient inferential tools is an important research avenue (Sharma, Saxena, and Rai 2019). A thorough treatment of mixture-of-experts models is given in Gormley and Frühwirth-Schnatter (2019).

The model in (33) naturally involves multiple layers of hierarchy, multi-modal posteriors and discrete parameter spaces, potentially rendering inference with general purpose posterior simulation tools difficult.[4] As a result, DA algorithms are popular tools for the estimation of mixture-of-experts models (Gormley and Frühwirth-Schnatter 2019). However, imbalanced data and large samples may lead to convergence issues. In model (33), both the success probabilities ζ_{ih} and the mixture weights η_{jh} may be imbalanced.

The methodology proposed in the present article is a potential remedy in such scenarios, as both the logistic regression experts and the gating function can be estimated using DA with additional location-based and scale-based parameter expansion steps. We demonstrate in a numerical exercise in Appendix A.6.3 that our iMDA scheme indeed leads to sizeable efficiency gains with respect to all involved regression parameters in simulated data. In Appendix A.7, we further illustrate logistic mixture-of-experts regression models in a large-sample real world application on maternal education and child mortality. Again, effective sample sizes increase as soon as iMDA is introduced.

7. Concluding Remarks

Due to a wide range of applications in many areas of applied science, much attention has been dedicated to the development of estimation algorithms for generalized linear models. In the past decades, various DA algorithms have been brought forward that have steadily increased accessibility and popularity of Bayesian estimation techniques in the context of regression models for binary and categorical outcomes. In this article, we introduce new sampling algorithms based on Pólya-Gamma mixture representations for estimation of these models. The

[3] The data is sourced from *https://en.wikipedia.org/wiki/List_of_epidemics* and the sources therein.

[4] See Appendix A.1 for further discussion.

algorithms are easily implemented, intuitively appealing and allow for a conditionally Gaussian posterior distribution of the regression effects in binary, multinomial and binomial logistic regression frameworks. To counteract potentially inefficient sampling behavior, we develop a novel parameter expansion strategy and apply it to the introduced sampling algorithms as well as to probit frameworks. This results in a competitive level of sampling efficiency, even in scenarios where outcomes are heavily imbalanced, as is demonstrated via extensive simulation studies and real data applications.

A number of future research avenues worth exploring come readily to mind. First, the proposed family of DA and MCMC boosting schemes could be extended to accommodate other types of limited outcomes such as ordered or count data. Second, we approached the problem of efficiency comparisons mostly empirically and left theoretical aspects largely unexplored. Extending the theoretical results of Choi and Hobert (2013) and Johndrow et al. (2019), among others, might be fruitful and assessing convergence rates of the proposed sampling schemes more formally may reveal additional insights. Finally, it is well-known that scale-based parameter expansion leads to faster convergence of expectation-maximization algorithms (Liu, Rubin, and Wu 1998). It may be worth to investigate whether the proposed location-based expansion leads to additional efficiency gains in this context.

Supplementary Materials

The online supplement contains the technical appendix as well as replication materials.

Acknowledgments

The authors would like to thank Darjus Hosszejni, two anonymous referees and an associate editor for helpful comments and suggestions.

Disclosure Statement

No potential conflict of interest was reported by the authors.

ORCID

Gregor Zens https://orcid.org/0000-0003-2253-8736

References

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [1,2,3]

Anceschi, N., Fasano, A., Durante, D., and Zanella, G. (2023), "Bayesian Conjugacy in Probit, Tobit, Multinomial Probit and Extensions: A Review and New Results," *Journal of the American Statistical Association*, 118, 1451–1469. [1]

Carter, C. K., and Kohn, R. (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81, 541–553. [9]

Choi, H. M., and Hobert, J. P. (2013), "The Polya-Gamma Gibbs Sampler for Bayesian Logistic Regression is Uniformly Ergodic," *Electronic Journal of Statistics*, 7, 2054–2064. [11]

Chopin, N., and Ridgway, J. (2017), "Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation," *Statistical Science*, 32, 64–87. DOI:10.1214/16-STS581 [1]

Duan, L. L., Johndrow, J. E., and Dunson, D. B. (2018), "Scaling Up Data Augmentation MCMC via Calibration," *Journal of Machine Learning Research*, 19, 1–34. [4]

Durante, D. (2019), "Conjugate Bayes for Probit Regression via Unified Skew-Normal Distributions," *Biometrika*, 106, 765–779. [1]

Frühwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15, 183–202. [9]

Frühwirth-Schnatter, S., and Frühwirth, R. (2007), "Auxiliary Mixture Sampling with Applications to Logistic Models," *Computational Statistics & Data Analysis*, 51, 3509–3528. [2,7,9]

——— (2010), "Data Augmentation and MCMC for Binary and Multinomial Logit Models," in *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, eds. T. Kneib and G. Tutz, pp. 111–132, Heidelberg: Physica-Verlag. [1,2,7,8]

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009), "Improved Auxiliary Mixture Sampling for Hierarchical Models of non-Gaussian Data," *Statistics and Computing*, 19, 479–492. [2]

Fussl, A., Frühwirth-Schnatter, S., and Frühwirth, R. (2013), "Efficient MCMC for Binomial Logit Models," *ACM Transactions on Modeling and Computer Simulation*, 23, 3:1–3:21. [1,2,7,9]

Gamerman, D. (1998), "Markov Chain Monte Carlo for Dynamic Generalized Linear Models," *Biometrika*, 85, 215–227. [9]

Gormley, I. C., and Frühwirth-Schnatter, S. (2019), "Mixture of Experts Models," in *Handbook of Mixture Analysis*, eds. S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert, pp. 271–307, Boca Raton, FL: CRC Press. [10]

Hobert, J. P., and Marchev, D. (2008), "A Theoretical Comparison of the Data Augmentation, Marginal Augmentation and PX-DA Algorithms," *The Annals of Statistics*, 36, 532–554. [3]

Holmes, C. C., and Held, L. (2006), "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression," *Bayesian Analysis*, 1, 145–168. [1,2,7]

Imai, K., and van Dyk, D. A. (2005), "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation," *Journal of Econometrics*, 124, 311–334. [2,4]

Jiang, W., and Tanner, M. A. (1999), "Hierarchical Mixtures-of-Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation," *The Annals of Statistics*, 27, 987–1011. [10]

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019), "MCMC for Imbalanced Categorical Data," *Journal of the American Statistical Association*, 114, 1394–1403. [2,3,4,7,9,11]

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion to Accelerate EM: The PX-EM Algorithm," *Biometrika*, 85, 755–770. [11]

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274. [2,3,4]

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics*, 99, 173–193. [2,4]

McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behaviour," in *Frontiers of Econometrics*, ed. P. Zarembka, pp. 105–142, New York: Academic Press. [6]

Piironen, J., and Vehtari, A. (2017), "Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors," *Electronic Journal of Stasitistics*, 11, 5018–5051. [5]

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006), "CODA: Convergence Diagnosis and Output Analysis for MCMC," *R News*, 6, 7–11. [8]

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models Using Pólya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349. [2,3,4,7,8,9]

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005), *Bayesian Statistics and Marketing*, Chichester: Wiley. [1]

Sen, D., Sachs, M., Lu, J., and Dunson, D. B. (2020), "Efficient Posterior Sampling for High-Dimensional Imbalanced Logistic Regression," *Biometrika*, 107, 1005–1012. [1]

Sharma, A., Saxena, S., and Rai, P. (2019), "A Flexible Probabilistic Framework for Large-Margin Mixture of Experts," *Machine Learning*, 108, 1369–1393. [10]

Shephard, N., and Pitt, M. K. (1997), "Likelihood Analysis of non-Gaussian Measurement Time Series," *Biometrika*, 84, 653–667. [9]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [1]

van Dyk, D., and Meng, X.-L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50. [2,3,4]

Zellner, A., and Rossi, P. E. (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics*, 25, 365–393. [1]

Zens, G., Frühwirth-Schnatter, S., and Wagner, H. (2021), "Efficient Bayesian Modeling of Binary and Categorical Data in R: The UPG Package," arXiv preprint arXiv:2101.02506. [2]