

Article

Efficient EM Estimation for the Pogit Model via Polya-Gamma Augmentation

Iván Gutiérrez ^{1,*} , Sandra Ramírez ^{2,*}  and Leonardo Jofré ³ 

¹ Departamento de Economía y Administración, Facultad de Economía y Negocios, Universidad Andrés Bello, Santiago 8370134, Chile

² Departamento de Ciencias Naturales y Matemáticas, Facultad de Ingeniería y Ciencias, Pontificia Universidad Javeriana, Cali 760031, Colombia

³ Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile; lnjofre@mat.uc.cl

* Correspondence: ivan.gutierrez@unab.cl (I.G.); smramirez@javerianacali.edu.co (S.R.)

Abstract

The Poisson-logistic (pogit) model is widely used for count data with latent intensities, with applications including under-reporting correction and share-of-wallet estimation, yet existing estimation methods do not scale well to large datasets. We propose a new expectation-maximization (EM) algorithm for the standard pogit model based on Polya-Gamma data augmentation, which yields a conditionally Gaussian complete-data likelihood with closed-form EM-updates. The resulting EM algorithm has low per-iteration cost and naturally accommodates computational enhancements, including quasi-Newton acceleration and mini-batch implementations. These features enable efficient inference on datasets with millions of observations. Simulation studies and real-data applications demonstrate substantial computational improvements without loss of statistical accuracy, and comparisons with direct maximum-likelihood optimization routines show that the proposed method provides a scalable and competitive alternative for large-scale pogit estimation.

Keywords: EM algorithm; pogit; under-reporting; share-of-wallet

1. Introduction

Count data models (see, e.g., [1]) with latent exposure or reporting mechanisms are common in many empirical settings, including marketing analytics [2], epidemiology [3], official statistics [4], and gender-violence research [5,6]. Among these, the Poisson-logistic (pogit) model [7] has emerged as a flexible and interpretable framework for modeling observed counts subject to under-reporting or partial observability. By combining a binomial observation equation with a Poisson exposure model, the pogit specification allows researchers to disentangle reporting behavior from the underlying intensity process, while retaining a regression structure that facilitates inference and interpretation.

The pogit model has found applications in diverse areas. In studies of under-reporting, it provides a principled way to account for missing or censored events by explicitly modeling the probability that an event is observed (see, e.g., [7]). In marketing and consumer analytics, it has been used for share-of-wallet (SoW) estimation, where observed purchases represent a noisy subset of latent purchase opportunities (see, e.g., [2]). These applications have motivated a growing methodological literature, including extensions to negative-binomial exposures [4], as well as parametric and nonparametric Bayesian formulations that incorporate prior information and/or variable selection (see, e.g., [5,8,9]).



Academic Editor: Michel Broniatowski

Received: 12 January 2026

Revised: 30 January 2026

Accepted: 5 February 2026

Published: 11 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Despite these advances, a major practical limitation of existing pogit methodologies is their lack of scalability. While the latent count can be analytically marginalized and maximum-likelihood estimation can in principle be based on the observed data likelihood, the resulting objective function is highly nonlinear and tightly couples the reporting and intensity components through a multiplicative structure. This makes direct likelihood maximization numerically delicate in practice, particularly as sample size and covariate dimension grow, or when parameters are weakly identified. Bayesian approaches face related difficulties, as Markov chain Monte Carlo methods must explore high-dimensional posteriors with strong dependence across model components. As a consequence, pogit models remain difficult to deploy in large-scale applications, despite their conceptual suitability for precisely such settings.

In this article, we address this literature gap by introducing a new expectation-maximization (EM) algorithm [10] for efficient estimation of the standard pogit model. Our approach builds on Polya-Gamma data augmentation [11], which yields conditionally Gaussian complete-data likelihoods and enables closed-form updates in the M-step. Specifically, we adapt the Polya-Gamma-based EM framework developed by Scott et al. [12] for logistic regression to the pogit setting, combining it with the approximate augmentation strategy for Poisson models introduced by D'Angelo et al. [13]. This synthesis results in an EM algorithm with a simple structure and closed-form updates throughout. Crucially, the resulting procedure is naturally amenable to online and mini-batch variants, making it possible to fit pogit models to datasets of unprecedented size using streaming data, without sacrificing statistical efficiency.

The remainder of the article is organized as follows. Section 2 introduces the pogit model and reviews its likelihood structure. Section 3 presents the proposed Polya-Gamma-based EM algorithm and discusses its computational properties, including quasi-Newton acceleration and mini-batch extensions for large-scale data. Section 4 describes additional computational enhancements. Section 5 reports results from simulation studies assessing convergence, finite-sample performance, and computational efficiency, with comparisons to direct numerical maximum-likelihood estimation. Section 6 applies the method to real datasets. Section 7 concludes.

Contributions

This article makes the following contributions:

- We introduce a new expectation-maximization algorithm for the standard pogit model based on Polya-Gamma data augmentation. By combining an exact augmentation for the binomial component with a controlled approximation for the Poisson component, the complete-data log-likelihood becomes quadratic in the regression parameters. This yields closed-form expressions for all E-step expectations and reduces the M-step to simple weighted least-squares updates, resulting in a fully analytic EM procedure with low per-iteration computational cost.
- We show that the resulting EM algorithm admits scalable online and mini-batch variants. In particular, the method can be applied to datasets with millions of observations using mini-batch updates, making pogit models feasible in large-scale applications where existing methods break down.
- We evaluate the statistical and computational performance of the proposed estimator using both simulated and real datasets. The results demonstrate fast convergence, stable behavior across sample sizes, and competitive estimation accuracy.
- We provide a systematic comparison with direct numerical maximization of the observed-data likelihood using generic maximum-likelihood routines, showing that

the proposed method delivers substantial runtime improvements while exhibiting stable finite-sample behavior, robust parameter recovery, and numerical stability.

2. Model

2.1. Hierarchical Specification

We consider the standard Poisson-logistic (pogit) model for count data with latent intensity and partial observability. For each observational unit $i = 1, \dots, N$, let y_i denote the observed count and n_i an unobserved latent count representing the total number of underlying events or opportunities. The model is defined hierarchically as

$$\begin{aligned}(y_i | n_i, \theta_i) &\stackrel{\text{ind}}{\sim} \text{Bin}(n_i, \theta_i), \\ (n_i | \lambda_i) &\stackrel{\text{ind}}{\sim} \text{Poisson}(E_i \lambda_i),\end{aligned} \quad i = 1, \dots, N,$$

where $\stackrel{\text{ind}}{\sim}$ denotes conditional independence across observational units, conditional on the model parameters and covariates, E_i is a known offset, $\theta_i \in (0, 1)$ is the probability that a latent event is observed, and $\lambda_i > 0$ is the latent intensity. The first-level binomial equation captures partial observability: conditional on n_i , only a fraction θ_i of events is recorded. The second-level Poisson equation models heterogeneity in the total number of latent events across observational units.

A key property of the pogit model is that the latent count admits a simple predictive distribution, as stated in the following property:

Property 1. For each observational unit $i = 1, \dots, N$ in a pogit model, $(n_i - y_i | y_i, \theta_i, \lambda_i) \sim \text{Poisson}(E_i \lambda_i (1 - \theta_i))$.

This result follows from the Poisson thinning property (see, e.g., [14]) and plays an important role in the EM algorithm developed below.

2.2. Regression Structure

Both the reporting probability and the latent intensity are linked to covariates through separate regression specifications:

$$\begin{aligned}\theta_i &= \text{sigmoid}(\eta_{i1}), & i &= 1, \dots, N, \\ \lambda_i &= \exp(\eta_{i2}), & j &= 1, 2,\end{aligned}$$

where $\eta_{ij} = x'_{ij} \beta_j$ denotes the linear predictor associated with component j for observational unit i , $x_{ij} \in \mathbb{R}^{K_j}$ is a column vector of observed covariates (with x'_{ij} denoting its transpose), $\beta_j \in \mathbb{R}^{K_j}$ is the corresponding vector of regression coefficients, K_j denotes the number of covariates entering component j , and $\text{sigmoid}(x) \equiv 1/(1 + e^{-x})$ is the sigmoid (or expit) function. This specification allows the reporting mechanism and the latent intensity to depend on distinct, possibly overlapping, covariate sets. Such flexibility is essential in applications such as under-reporting correction and share-of-wallet estimation. Moreover, the chosen link functions yield interpretable parameters and facilitate likelihood augmentation.

Throughout the paper, the model is defined in terms of the natural parameters (θ_i, λ_i) , while many derivations and algorithmic steps are expressed in terms of the associated linear predictors (η_{i1}, η_{i2}) , where $\theta_i = \text{sigmoid}(\eta_{i1})$ and $\lambda_i = \exp(\eta_{i2})$. This reparameterization is purely deterministic and one-to-one. For notational simplicity and computational convenience, we work with whichever representation is more appropriate in a given context, and the mapping between the two is made explicit whenever it is used.

Figure 1 provides a graphical representation of the model using plate notation [15]. Regression coefficients $\beta = (\beta'_1, \beta'_2)'$ are shared across observational units, while n_i and y_i are unit-specific latent and observed variables.

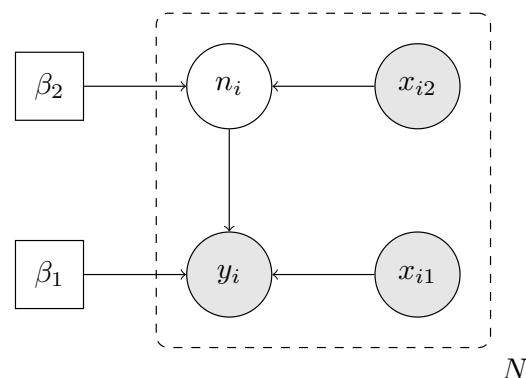


Figure 1. Graphical representation of the pogit model. Circles denote random variables and arrows indicate conditional dependence. Shaded nodes represent observed quantities.

2.3. Observed Likelihood and Identification

Marginalizing over the latent count n_i yields

$$(y_i \mid \theta_i, \lambda_i) \stackrel{\text{ind}}{\sim} \text{Poisson}(E_i \theta_i \lambda_i), \quad i = 1, \dots, N,$$

see, e.g., Kingman [14]. The observed-data likelihood, therefore, depends on the parameters only through the product $\theta_i \lambda_i$. As a consequence, without additional structure, the reporting probability and the latent intensity are not separately identified. In particular, for any $c > 0$, the transformation $\theta_i \mapsto c\theta_i$ and $\lambda_i \mapsto \lambda_i/c$ leaves the likelihood unchanged.

Identification is restored by the regression structure, which links θ_i and λ_i to covariates through distinct linear predictors. Under mild regularity conditions, this structure ensures local identification of the parameter vector β , that is, that there is a neighborhood of the true β without observationally equivalent parameter values (in the sense of producing the exact same likelihood). The identification result is not merely technical. Local identification is a necessary condition for the observed-data likelihood to admit a locally unique maximizer in a neighborhood of the true parameter value, thereby ruling out flat ridges generated by observationally equivalent parameterizations. This property underpins meaningful likelihood-based inference by ensuring that distinct parameter values correspond to distinct data-generating processes in a local neighborhood. While local identification does not preclude the presence of nearly flat regions of the likelihood, it guarantees local uniqueness of the true parameter, which is a necessary condition for well-defined EM-based estimation. A formal definition of local identification and the associated regularity concepts are provided in Appendix A.

Theorem 1. Let $i = 1, \dots, N$, and let $x'_i = (x'_{i1}, x'_{i2})$ and $z'_i = ((1 - \theta_i)x'_{i1}, x'_{i2})$. Suppose $\{(y_i, x_i)\}_{i=1}^N$ is an i.i.d. sequence such that $E[z_i z'_i]$ is positive definite. Then the pogit parameters are locally identified.

Proof. See Appendix A. $\square$

Remark 1 (Overlap and weak identification). Local identification may hold even when x_{i1} and x_{i2} overlap. However, strong overlap or collinearity renders the Fisher information matrix nearly singular, leading to weak identification in finite samples. Applied work, therefore, often relies on

distinct covariates across hierarchical levels (see, e.g., [2]), or incorporates auxiliary data in which the counts n_i are observed (see, e.g., [8]).

Remark 2 (Local versus global identification). *Theorem 1 establishes local, but not global, identification. As shown by Brennan et al. [16], when covariates entering the intensity equation are a subset of those entering the reporting equation, the model may be locally identified while remaining globally unidentified. Common remedies include reducing covariate overlap, incorporating auxiliary observations on latent counts, or imposing sign restrictions informed by expert judgment.*

From a computational perspective, the observed-data likelihood is highly nonlinear in β . Direct maximization requires numerical optimization and becomes increasingly costly as the sample size and covariate dimension grow. Moreover, the multiplicative interaction between θ_i and λ_i often leads to instability in large samples. These features motivate the search for an augmented representation with simpler structure.

2.4. A Naive EM Algorithm

The hierarchical formulation naturally suggests an EM algorithm based on the augmented likelihood $p(y, n | \beta) = \prod_i p(y_i | n_i, \theta_i) p(n_i | \lambda_i)$, that is

$$p(y, n | \beta) = \prod_i \text{Bin}(y_i | n_i, \theta_i) \text{Poisson}(n_i | E_i \lambda_i).$$

Given a current iterate $\beta^{(t)}$, such an algorithm replaces n_i by $E[n_i | y_i, \beta^{(t)}]$ in the E-step and updates $\beta = (\beta'_1, \beta'_2)'$ via binomial and Poisson regressions in the M-step. While formally valid, this approach is computationally unattractive: both regressions require iterative solvers, so each EM iteration contains nested optimization loops. As a result, the procedure scales poorly and offers little computational advantage over direct likelihood maximization.

The key insight of this article is that augmenting only with the latent counts n_i is *insufficient* to obtain a scalable EM algorithm, because the resulting complete likelihood remains non-quadratic in the regression parameters.

To overcome this limitation, we exploit the fact that both components of the pogit model admit conditionally Gaussian representations under suitable augmentations. The binomial component admits an exact Polya-Gamma augmentation, while the Poisson component can be accurately approximated by a negative-binomial pmf that also yields a Polya-Gamma augmentation.

Introducing Polya-Gamma variables in addition to the latent counts renders the complete log-likelihood quadratic in (β_1, β_2) . The resulting EM algorithm features closed-form E-step expectations and M-steps that reduce to weighted least-squares problems with explicit solutions. This structure yields low per-iteration cost and naturally accommodates large-scale extensions. The construction of this augmented likelihood and the corresponding EM updates are developed in the next section.

3. An Improved EM Algorithm

In this section, we present a scalable expectation–maximization (EM) algorithm for the pogit model that exploits Polya-Gamma augmentation to obtain a quadratic complete-data log-likelihood and closed-form updates in both the E- and M-steps. As mentioned in Section 2, the algorithm is based on the Polya-Gamma distribution, so we start by explaining this distribution and its key properties.

3.1. The Polya-Gamma Distribution

The Polya-Gamma distribution was introduced by Polson et al. [11] as part of a new data augmentation for logistic regression models. A random variable $w \geq 0$ is said to

follow a Polya-Gamma distribution with parameters (b, c) , denoted $w \sim \text{PG}(b, c)$, where $b > 0$ and $c \in \mathbb{R}$, if it admits the representation

$$w \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)},$$

where $\{g_k\}_{k \geq 1}$ are independent $\text{Gamma}(b, 1)$ random variables.

The Polya-Gamma distribution has two key properties.

Property 2. For any $a, \psi \in \mathbb{R}$ and $b > 0$,

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b} \int_0^\infty \exp(\kappa\psi - w\psi^2/2) p(w | b, 0) dw,$$

where $\kappa = a - b/2$ and $w \sim \text{PG}(b, 0)$.

This property will be useful for our EM algorithm, as it transforms binomial and negative-binomial likelihoods into Gaussian kernels conditional on the latent variable w .

Property 3. Let $w \sim \text{PG}(b, c)$. Then, $E[w] = (b/4) \text{tanhc}(c/2)$ for any $c > 0$ and $E[w] = b/4$ otherwise, where $\text{tanhc}(x) := \tanh(x)/x$.

This property will be particularly convenient for our EM algorithm, as it will allow the E-step to be computed analytically without numerical integration.

3.2. The Augmented Likelihood

We now derive our likelihood augmentation. The central idea is to augment the pogit model so that the complete-data log-likelihood becomes quadratic in the regression parameters (β_1, β_2) , yielding closed-form updates in the M-step and eliminating inner optimization loops. For ease of exposition, we present the main steps of the likelihood augmentation here; the full derivation is deferred to Appendix B.

Consider the i th contribution to the naive complete likelihood,

$$p(n_i, y_i | \beta) = \text{Bin}(y_i | n_i, \theta_i) \text{Poisson}(n_i | E_i \lambda_i).$$

We treat the two components separately.

Before deriving the augmented likelihood contributions, recall that $\eta_{i1} = x'_{i1} \beta_1$ and $\eta_{i2} = x'_{i2} \beta_2$ denote the linear predictors associated with the binomial (reporting) and latent intensity components, respectively.

3.2.1. Binomial Component

Using the Polya-Gamma identity in Property 2 and rearranging terms, the binomial likelihood can be written as

$$\text{Bin}(y_i | n_i, \theta_i) \propto E_{w_{i1}} \exp(\kappa_{i1} \eta_{i1} - w_{i1} \eta_{i1}^2 / 2),$$

where $E_{w_{i1}}$ denotes expectation with respect to $w_{i1} \sim \text{PG}(n_i, 0)$, $\kappa_{i1} = y_i - n_i/2$, and the symbol $\propto$ denotes equality up to a multiplicative constant independent of the parameters of interest. Therefore, conditional on w_{i1} , the binomial likelihood contribution can be expressed as an exponential quadratic form in the linear predictor η_{i1} , that is, as a Gaussian kernel (up to a normalizing constant).

3.2.2. Poisson Component

Unlike the binomial likelihood, the Poisson likelihood does not admit an exact Polya-Gamma representation. However, it is well-known that the $\text{NegBin}(r_i, E_i \lambda_i / (r_i + E_i \lambda_i))$ distribution converges in distribution to the $\text{Poisson}(E_i \lambda_i)$ distribution as $r_i \rightarrow \infty$ (see, e.g., [14]). Unlike the binomial likelihood, the Poisson likelihood does not admit an exact Polya-Gamma representation. We, therefore, approximate it using a negative-binomial distribution with parameter $r_i \gg 0$, where r_i controls the accuracy of the approximation.

Hence, we can approximate

$$\text{Poisson}(n_i \mid E_i \lambda_i) \approx \text{NegBin}(n_i \mid r_i, E_i \lambda_i / (r_i + E_i \lambda_i)),$$

for some $r_i \gg 0$. Using Property 2 and simplifying, we obtain

$$\text{Poisson}(n_i \mid E_i \lambda_i) \propto E_{w_{i2}} \exp(\kappa_{i2} \eta_{i2} - w_{i2} \eta_{i2}^2 / 2 - \frac{w_{i2}}{2} (\log(r_i / E_i))^2),$$

where $E_{w_{i2}}$ denotes expectation with respect to $w_{i2} \sim \text{PG}(n_i + r_i, 0)$, $\kappa_{i2} = (n_i - r_i) / 2 + w_{i2} \log(r_i / E_i)$. Therefore, conditional on w_{i2} , the approximated Poisson likelihood can be expressed as an exponential quadratic form in the linear predictor η_{i2} , that is, as a Gaussian kernel (up to a normalizing constant). This representation is key to obtaining closed-form updates in the M-step.

Remark 3. The use of a negative-binomial approximation to enable Polya-Gamma augmentation for Poisson models was first proposed by D'Angelo et al. [13] in the context of Bayesian Poisson regression, and shown to improve computational efficiency relative to earlier approaches.

Remark 4. Although the proposed augmentation is based on an approximation rather than an exact identity, its quality is fully controllable: accuracy can be made arbitrarily high by increasing r_i , with the only practical limitation being numerical stability. In practice, even moderate values of r_i already yield excellent approximations; in our experiments, $r_i = 100$ produced indistinguishable parameter estimates. For fixed r_i , the proposed procedure is an exact EM algorithm for a well-defined approximating likelihood based on a negative-binomial representation of the Poisson component, and this likelihood converges pointwise to the plogit likelihood as $r_i \rightarrow \infty$.

3.2.3. Augmented Log-Likelihood

Introducing the aforementioned w_{ij} 's to the model, we obtain the following augmented log-likelihood:

$$\log p(y, n, w \mid \beta) \doteq \sum_{ij} (\kappa_{ij} \eta_{ij} - w_{ij} \eta_{ij}^2 / 2) - \sum_i w_{i2} (\log(r_i / E_i))^2 / 2,$$

where $\doteq$ denotes equality up to additive constants that do not depend on the parameters of interest (in this case, β).

3.3. EM Steps

Let $\beta^{(t)}$ denote the current iterate of β . As with any EM algorithm, our procedure updates this value following two general steps:

- E-step: compute the Q-function, $Q(\beta) := E[\log p(y, n, w \mid \beta) \mid y, \beta^{(t)}]$.
- M-step: update β by maximizing the Q-function.

As we shall see, both steps admit closed-form expressions, leading to a simple and efficient EM algorithm.

3.3.1. E-Step

First, note that $Q(\beta)$ depends on β only through η . In particular,

$$Q(\beta) \doteq \sum_{ij} (\hat{\kappa}_{ij} \eta_{ij} - \hat{w}_{ij} \eta_{ij}^2 / 2) - \sum_i \hat{w}_{i2} (\log(r_i / E_i))^2 / 2,$$

where $\hat{a} = E[a \mid y, \beta^{(t)}]$ denotes the conditional expectation of any random variable a (for instance, κ_{ij}). We now explain how to compute $\hat{\kappa}_{ij}$ and $\hat{w}_{ij}$.

Computing $\hat{\kappa}_{ij}$. This is straightforward as it is linear in (n, w) :

$$\begin{aligned} \hat{\kappa}_{i1} &= y_i - \hat{n}_i / 2, \\ \hat{\kappa}_{i2} &= (\hat{n}_i - r_i) / 2 + \hat{w}_{i2} \log(r_i / E_i). \end{aligned}$$

Computing $\hat{n}_i$. This is also straightforward. Property 1 implies that $E[n_i \mid y, \beta]$ is equal to $y_i + E_i(1 - \theta_i)\lambda_i = y_i + E_i \exp(\eta_{i2}) / (1 + \exp(\eta_{i1}))$, so evaluating at $\beta = \beta^{(t)}$ gives

$$\hat{n}_i = y_i + E_i \exp(\hat{\eta}_{i2}) / (1 + \exp(\hat{\eta}_{i1})).$$

Computing $\hat{w}_{ij}$. This is more challenging, but it is well-known that

- $(w_{i1} \mid y, n, \beta) \sim \text{PG}(n_i, \eta_{i1})$ [11].
- $(w_{i2} \mid y, n, \beta) \sim \text{PG}(n_i + r_i, \eta_{i2} - \log(r_i / E_i))$ [13].

Hence, using Property 3 and the law of iterated expectations, we obtain

$$\begin{aligned} \hat{w}_{i1} &= 0.25 \hat{n}_i \tanhc(0.5 \hat{\eta}_{i1}), \\ \hat{w}_{i2} &= 0.25 (\hat{n}_i + r_i) \tanhc(0.5 (\hat{\eta}_{i2} - \log(r_i / E_i))). \end{aligned}$$

In summary, $Q(\beta)$ has a closed-form expression.

Remark 5. In order to apply the E-step successfully, the functions $\text{sigmoid}(\cdot)$ and $\tanhc(\cdot)$ must be evaluated in a numerically stable way. In particular, we use

$$\text{sigmoid}(x) = \begin{cases} \exp(x) / (1 + \exp(x)), & \text{if } x < 0 \\ 1 / (1 + \exp(-x)), & \text{otherwise} \end{cases}$$

and

$$\tanhc(x) = \begin{cases} 1 - x^2/3 + 2x^4/15 - 17x^6/315, & \text{if } |x| < \epsilon \\ \tanh(x)/x, & \text{otherwise} \end{cases}$$

In our experiments, $\epsilon = 10^{-4}$ worked fine.

Remark 6. Exploiting the closed-form mean of the Polya-Gamma distribution within an expectation-maximization framework was first proposed by Scott et al. [12] in the context of logistic regression, where it was shown to yield substantial computational gains relative to naive optimization strategies.

3.3.2. M-Step

As $Q(\beta)$ is quadratic in β , maximization yields closed-form solutions. For $j = 1, 2$, let X_j denote the design matrix collecting the covariates x'_{ij} associated with component j across all observational units. By standard least-squares theory, the solution is given by

$$\beta_j^{(t+1)} = (X_j' \hat{W}_j X_j)^{-1} X_j' \hat{\kappa}_j, \quad j = 1, 2$$

where $\hat{W}_j$ is a diagonal matrix with entries $\hat{w}_{ij}$.

In this way, each EM iteration reduces to two independent weighted least-squares problems that require no any inner iterative procedure. In addition, each update has a structure particularly well suited for parallel and mini-batch extensions.

Remark 7. The Polya-Gamma augmentation used here is closely related to constructions commonly exploited in variational Bayes (VB) methods (see, e.g., [17]) for logistic models [18]. The present approach, however, differs in both objective and implementation. Our algorithm is an expectation-maximization procedure that directly targets the observed-data likelihood of the plogit model, rather than a variational lower bound, and imposes no factorization assumptions on latent variables. All conditional expectations in the E-step are computed exactly under the augmented model, and stochastic approximation is used solely to scale the computation of sufficient statistics in large samples. As a result, the method retains the likelihood-based interpretation and asymptotic properties of maximum-likelihood estimation, while achieving computational efficiency comparable to VB approaches.

4. Computational Enhancements

The EM algorithm introduced in Section 3 has two key computational advantages: each iteration has low cost due to closed-form updates, and the algorithm admits a simple representation in terms of sufficient statistics. Nevertheless, two practical challenges remain in large-scale applications. First, when the sample size N is very large, even inexpensive full-batch iterations can become costly. Second, like most fixed-point algorithms, EM may converge slowly when the likelihood surface is flat or parameters are weakly identified.

In this section, we address these challenges using two complementary strategies. To scale the algorithm to massive datasets, we develop a mini-batch EM scheme based on Robbins–Monro stochastic approximation. To accelerate convergence (even in samples of moderate size), we combine the EM updates with a quasi-Newton extrapolation technique known as SQUAREM. These enhancements target distinct computational bottlenecks and can be used independently or in combination.

4.1. Mini-Batch EM via Robbins–Monro

To reduce computational cost when N is large, we adopt an online EM approach based on stochastic approximation [19], following the general framework of Cappé and Moulines [20]. The key observation is that the EM algorithm derived in Section 3 depends on the data only through additive sufficient statistics, making it particularly well suited for mini-batching.

Recall that the M-step for component $j \in \{1, 2\}$ depends on the statistics

$$S_j^{(t)} = X_j' \hat{W}_j X_j, \quad s_j^{(t)} = X_j' \hat{\kappa}_j.$$

Given a random mini-batch $B_t \subseteq \{1, \dots, N\}$, unbiased estimators are

$$\hat{S}_j^{(t)} = \frac{N}{|B_t|} \sum_{i \in B_t} \hat{w}_{ij} x_{ij} x_{ij}', \quad \hat{s}_j^{(t)} = \frac{N}{|B_t|} \sum_{i \in B_t} \hat{\kappa}_{ij} x_{ij},$$

where randomness arises solely from subsampling.

A naive approach would replace $(S_j^{(t)}, s_j^{(t)})$ directly with their mini-batch counterparts. However, although unbiased, these estimators exhibit persistent sampling noise that prevents convergence of the resulting EM iterations. To stabilize the procedure, Cappé and Moulines [20] propose updating the sufficient statistics using diminishing step sizes, that is

$$\begin{aligned} S_j^{(t+1)} &= (1 - \gamma_t) S_j^{(t)} + \gamma_t \hat{S}_j^{(t)}, \\ s_j^{(t+1)} &= (1 - \gamma_t) s_j^{(t)} + \gamma_t \hat{s}_j^{(t)}, \end{aligned}$$

where $S_j^{(0)} = 0$, $s_j^{(0)} = 0$, and the step size sequence $\{\gamma_t\}_{t \geq 1}$ shrinks in such a way that $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$.

In practice, we use

$$\gamma_t = \begin{cases} c, & t \leq t_{\text{burn}}, \\ c/(t - t_{\text{burn}} + t_0)^a, & t > t_{\text{burn}}, \end{cases}$$

with $c > 0$, $a \in (1/2, 1)$, and $t_{\text{burn}}, t_0 > 0$. The initial constant phase stabilizes early iterations, while the polynomial decay ensures asymptotic convergence.

To further reduce variance and improve finite-sample performance, we apply *Polyak averaging* [21,22] to the parameter iterates after burn-in. Specifically, the reported estimator is

$$\tilde{\beta} = (T - t_{\text{burn}})^{-1} \sum_{t > t_{\text{burn}}} \beta^{(t)},$$

which achieves optimal asymptotic variance for stochastic approximation schemes [20].

This mini-batch EM procedure preserves the structure of the exact Polya-Gamma EM updates while reducing per-iteration complexity to $O(|B_t|K_j^2)$, independent of the total sample size. As a result, the method scales naturally to datasets with millions of observations.

Remark 8. Mini-batch EM algorithms based on stochastic approximation may require a large number of iterations and can exhibit substantial variability in the raw parameter iterates. For this reason, we recommend using a sufficiently large iteration budget (at least 10,000 iterations in our experiments) and assessing convergence in terms of the Polyak-averaged estimator $\tilde{\beta}$ rather than the instantaneous iterate $\beta^{(t)}$. This practice stabilizes inference and aligns with standard recommendations in stochastic approximation.

Remark 9. Unlike full-batch EM, mini-batch EM does not guarantee a monotone increase in the observed-data likelihood. While monotonicity can in principle be enforced via safeguarding steps, our simulation experiments indicate that the proposed algorithm exhibits stable behavior without such modifications. For this reason, and to preserve computational simplicity, we adopt the unsafeguarded version in our implementation.

4.2. Quasi-Newton Acceleration via SQUAREM

Even in full-batch settings, EM algorithms may converge slowly when the likelihood surface is flat or parameters are weakly identified. To accelerate convergence, we complement our method with a quasi-Newton extrapolation technique known as SQUAREM [23].

Let $M(\beta)$ denote the EM update mapping induced by one full E- and M-step, so that the standard EM iteration is $\beta^{(t+1)} = M(\beta^{(t)})$. SQUAREM treats EM as a fixed-point iteration and constructs an accelerated update by extrapolating along the EM trajectory. Starting from $\beta^{(t,0)} := \beta^{(t)}$, define

$$\begin{aligned} \beta^{(t,h)} &= M(\beta^{(t,h-1)}), \quad h = 1, 2, \\ r &= \beta^{(t,1)} - \beta^{(t,0)}, \\ v &= \beta^{(t,2)} - \beta^{(t,1)} - r. \end{aligned}$$

Here, r represents the first-order displacement induced by EM, while v captures curvature by measuring deviations from linearity along the EM path.

The accelerated update is then given by

$$\beta_{\text{SQ}}^{(t+1)} = \beta^{(t)} - 2\alpha r + \alpha^2 v,$$

where the step size α is chosen to approximately minimize the norm of the residual. Following Varadhan and Roland [23], we set

$$\alpha = -\|r\|_2 / \|v\|_2.$$

A safeguarding step ensures monotonicity of the observed log-likelihood, reverting to the standard EM update whenever the extrapolated iterate decreases the likelihood.

SQUAREM is particularly well-suited to our Polya-Gamma-based EM algorithm. First, each EM iteration is deterministic and inexpensive, making the cost of additional EM evaluations negligible relative to the gains in convergence speed. Second, the M-step consists of weighted least-squares updates, which vary smoothly with the parameters and favor quasi-Newton extrapolation. Third, the method is entirely generic and requires no modification of the underlying EM structure.

In our empirical experiments, SQUAREM substantially reduces the number of EM iterations required to reach convergence, often by an order of magnitude, while preserving numerical stability. For this reason, we recommend SQUAREM as the default acceleration strategy when fitting the plogit model in moderate to large samples.

5. Simulation Study

This section evaluates the proposed EM algorithm along three complementary dimensions. First, we study its finite-sample estimation behavior under moderate sample sizes, focusing on signal recovery and dispersion across Monte Carlo replications. Second, we assess numerical robustness with respect to the negative-binomial approximation parameter underlying the Polya-Gamma augmentation. Third, we examine computational scalability in large samples and quantify the gains achieved by standard acceleration techniques. Together, these experiments are designed to validate the statistical accuracy, numerical stability, and practical scalability of the proposed estimation framework.

Remark 10. All simulations were conducted on a desktop computer equipped with an Intel® Core™ i7-9750H CPU (6 cores, 2.60GHz) and 16 GB of RAM, running Windows 11 (64-bit).

5.1. Finite-Sample Estimation Behavior

We begin by examining the finite-sample behavior of the EM estimator in moderate sample sizes. In all scenarios, each model component includes nine covariates (i.e., $K_1 = K_2 = 9$). For each observational unit i , the covariate vectors are generated independently according to $x_{ij} \sim N(0, I_9/3)$. The scaling factor $1/3$ is chosen so that, if all regression coefficients are set equal to one, the resulting linear predictors $\eta_{ij} = x'_{ij}\beta_j$ have variance of moderate magnitude. This normalization keeps the linear predictors within a numerically stable range for the logistic and exponential link functions throughout the simulations.

The true parameter vectors are fixed at

$$\beta_1 = (1, 2, 0, \dots, 0), \quad \beta_2 = (2, 1, 0, \dots, 0),$$

so that only the first two coefficients in each vector are nonzero. This design induces a sparse signal structure and allows us to assess signal recovery.

Throughout the simulation study, the exposure term is fixed at $E_i \equiv 1$ for all i . We consider two sample sizes: $N = 500$ (Scenario 1) and $N = 1000$ (Scenario 2). For each scenario, $R = 100$ independent datasets are generated from the data-generating process described above. In each replication, the EM algorithm is initialized randomly and iterated until convergence, with a maximum of 1000 iterations allowed. Convergence is declared when the maximum relative change across all parameter components between successive iterations falls below 10^{-4} .

Figure 2 summarizes the resulting Monte Carlo distributions of the EM estimates. The nonzero coefficients in β_1 and β_2 are accurately recovered in both scenarios, while coefficients that are truly zero remain tightly concentrated around zero. Increasing the sample size from 500 to 1000 leads to a visible reduction in dispersion across replications, indicating improved estimator concentration in larger samples.

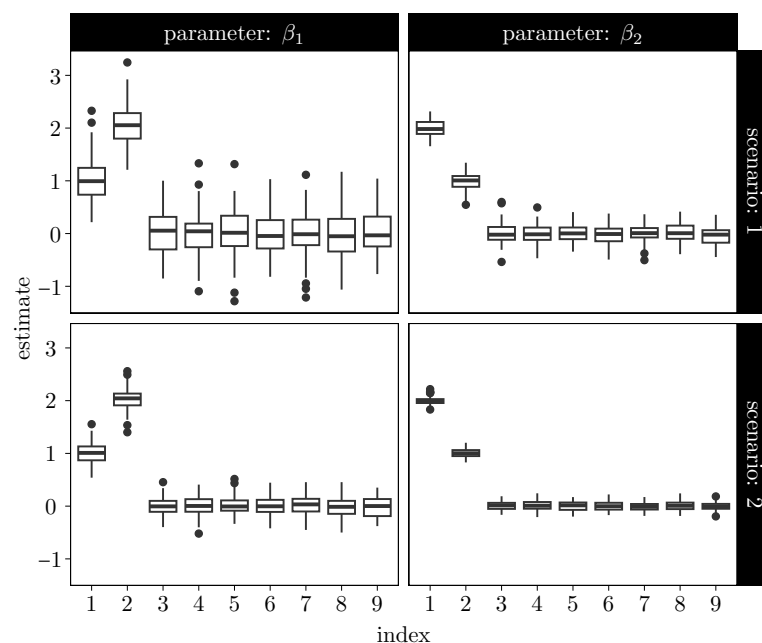


Figure 2. Boxplots of EM estimates across $R = 100$ Monte Carlo replications for the coefficients $\beta_1 = (1, 2, 0, \dots, 0)$ (left) and $\beta_2 = (2, 1, 0, \dots, 0)$ (right) for two sample sizes: $N = 500$ (top) and $N = 1000$ (bottom). The horizontal axis corresponds to the index of the regression coefficient within each parameter vector (i.e., the k th component of β_j). Coefficients that are truly zero concentrate near zero, while nonzero coefficients are recovered with reduced dispersion as sample size increases.

5.2. Robustness to the Approximation Parameter

Having established stable finite-sample behavior, we next examine robustness with respect to the tuning parameter r controlling the accuracy of the negative-binomial approximation used to obtain a Polya-Gamma augmentation of the Poisson component. While larger values of r yield a closer approximation to the Poisson likelihood, excessively large choices may introduce unnecessary numerical overhead and potential numerical issues. Evaluating sensitivity to this parameter is, therefore, important for practical implementation.

Figure 3 reports Monte Carlo boxplots for selected coefficients (indices 1 and 2) of β_1 and β_2 across values $r \in \{30, 60, \dots, 210\}$, with $r_i \equiv r$ for all i . Across all panels, the empirical distributions remain stable as r varies. While Monte Carlo variability is present, no systematic shifts or trends are observed as r increases from relatively small to moderately large values.

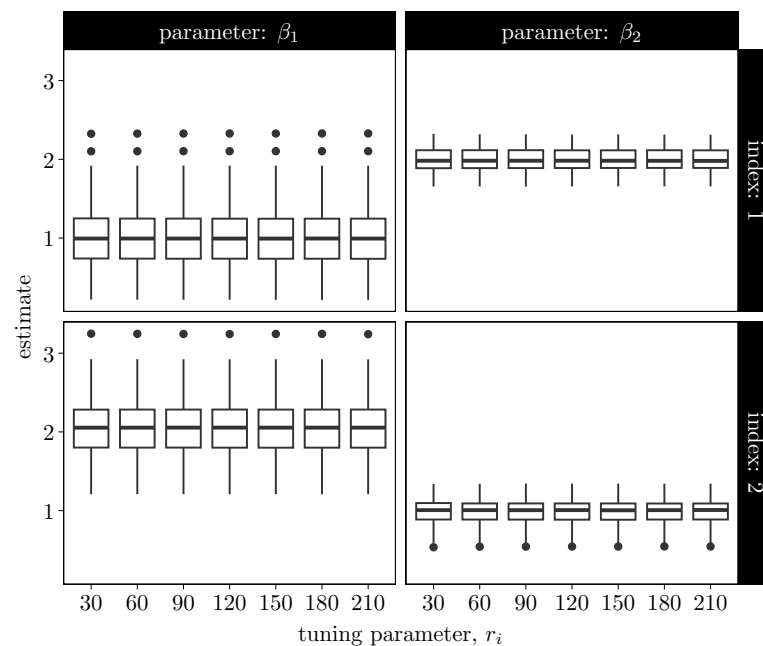


Figure 3. Sensitivity of EM estimates to the negative-binomial tuning parameter r_i . Boxplots are shown for selected coefficients (indices 1 and 2, corresponding to the first two components of each coefficient vector) of β_1 and β_2 across values $r_i \in \{30, 60, \dots, 210\}$. No systematic changes in the estimates are observed as r_i varies.

These results indicate that the proposed EM algorithm is not unduly sensitive to the precise choice of the tuning parameter, provided that r is chosen sufficiently large to yield an accurate quadratic approximation of the Poisson component. In particular, moderate values such as $r = 100$, which we adopt as a default in subsequent experiments, already yield estimation behavior comparable to that observed for larger values.

Remark 11. To complement the results of this particular experiment, we included three additional figures in Appendix C, showing the sensitivities of the estimates' standard errors, the observed likelihood and the effective number of iterations to r . Together, these figures show that the effective number of iterations grows with r , but standard errors and the log-likelihood tend to stabilize at $r \sim 100$. These supplementary results strengthen our practical recommendation of $r = 100$.

5.3. Computational Scalability and Acceleration

Having established finite-sample accuracy and numerical robustness, we now turn to computational performance. These experiments focus on scalability in large samples and are designed to assess whether the proposed EM algorithm and its accelerated variants remain practical in regimes where existing pogit implementations become computationally prohibitive.

We consider three complementary comparisons. First, we benchmark the proposed deterministic EM algorithm against direct maximum-likelihood (ML) optimization of the observed-data likelihood. Second, we evaluate the impact of mini-batch estimation via Robbins–Monro updates on runtime scalability. Third, we assess the gains from quasi-Newton acceleration using SQUAREM. In all cases, methods target the same likelihood and are run under identical initialization/stopping rules. For the Robbins–Monro mini-batch EM variant, stepsize sequences were chosen according to standard diminishing-stepsizes conditions; all tuning parameters are reported in Supplementary File S1.

Figures 4 and 5 report total runtime (measured in seconds) as a function of sample size (measured in hundreds of thousands of observations). Figure 4 compares the proposed

EM algorithm with direct ML optimization (implemented in Julia using BFGS via the R function `stats::optim()`); Figure 5 compares standard EM with its SQUAREM-accelerated version and Figure 6 compares full-batch EM with a Robbins–Monro mini-batch variant.

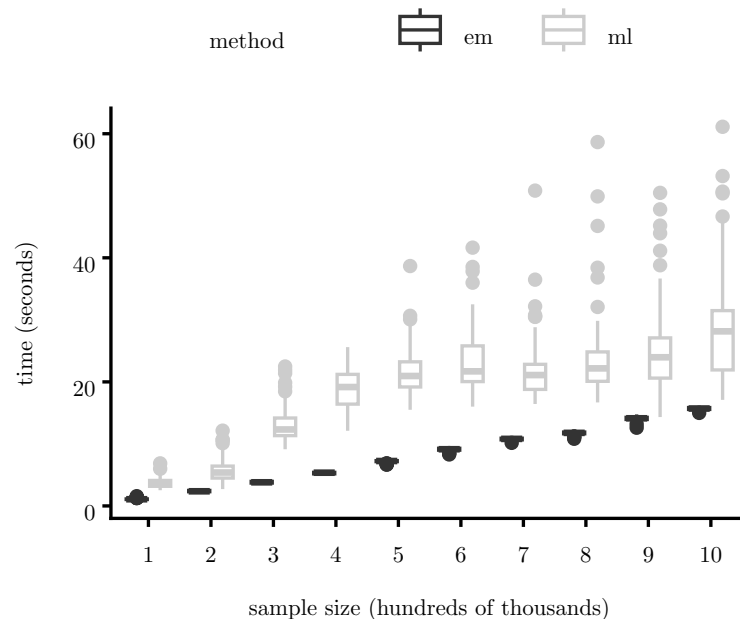


Figure 4. Runtime scaling of the proposed EM algorithm as a function of sample size. Across all sample sizes, EM is consistently faster than direct ML optimization, with the difference becoming more pronounced as N increases. For both algorithms, we allowed a maximum of 1000 iterations, and convergence was declared if the maximum relative change across all parameter components between successive iterations falls below 10^{-4} .

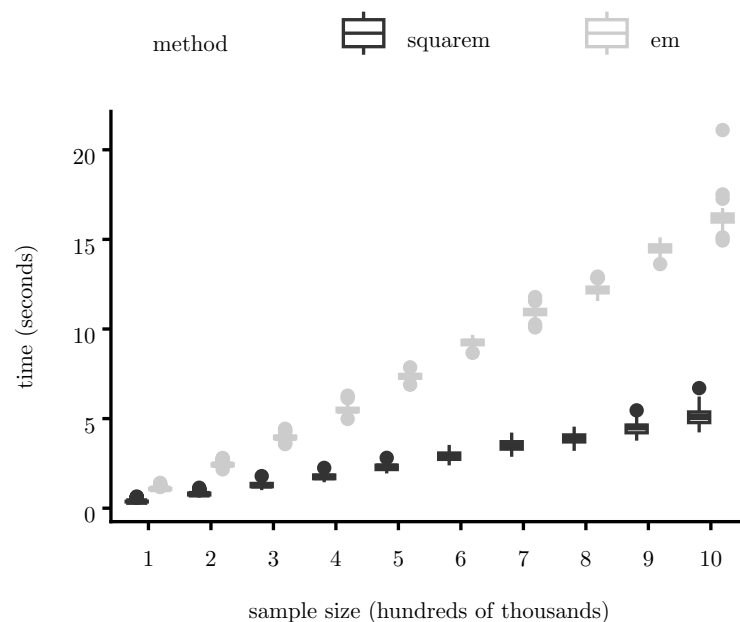


Figure 5. Runtime comparison between standard EM and SQUAREM-accelerated EM. SQUAREM applies safeguarded squared extrapolation and yields substantial runtime reductions for large samples. For both algorithms, we allowed a maximum of 1000 iterations, and convergence was declared if the maximum relative change across all parameter components between successive iterations falls below 10^{-4} .

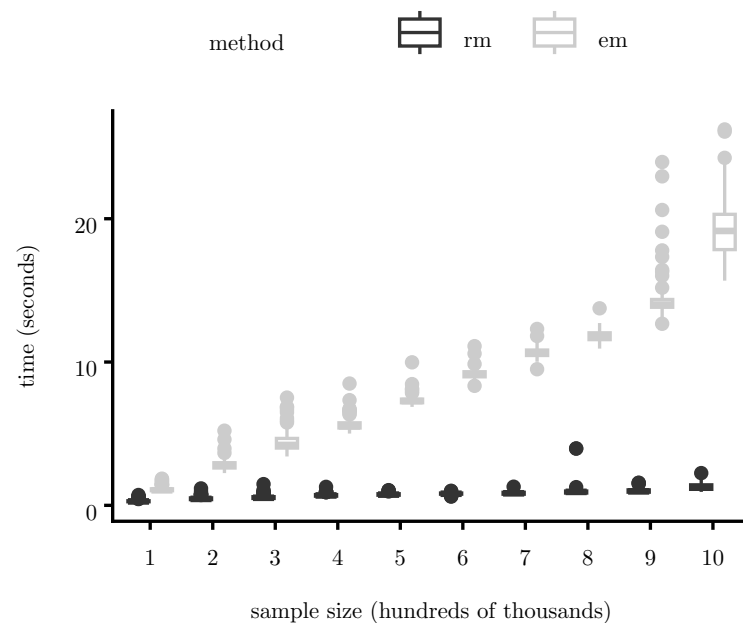


Figure 6. Runtime comparison between full-batch EM and Robbins–Monro mini-batch EM as a function of sample size. All methods use identical initialization and stopping rules. For the first algorithm, we allowed a maximum of 1000 iterations, and convergence was declared if the maximum relative change across all parameter components between successive iterations falls below 10^{-4} . For the second estimator, we allowed a maximum of 10,000 iterations, and convergence was declared if the maximum relative change across all parameter components between successive Polyak-averages falls below 10^{-4} .

Across all methods, runtime grows approximately linearly with sample size over the ranges considered. The proposed EM algorithm consistently outperforms direct ML optimization, with a consistent runtime advantage across all sample sizes considered. The Robbins–Monro variant further reduces runtime in very large samples by lowering per-iteration cost, while SQUAREM substantially decreases the number of iterations required for convergence. These gains are achieved without compromising numerical stability or likelihood monotonicity.

Overall, the simulation results demonstrate that the proposed Poly-Gamma-based EM algorithm combines stable finite-sample behavior, robustness to approximation choices, and scalability to large datasets. These properties address key practical limitations of existing pogit implementations and support the use of the method in modern large-scale applications.

Finally, to facilitate transparency and reproducibility of the computational evidence reported above, we provide the Julia version 1.12.3 and R version 4.4.2 code used to generate all simulation results in Supplementary File S1.

6. Real-World Application

We apply the proposed pogit model, estimated via a Pólya-Gamma-based EM framework, to an openly available dataset derived from Amazon purchase histories [24], with a specific focus on purchases of Apple products within the technology and electronics category. The purpose of this empirical application is to illustrate the use of the model and the practical implementation of the proposed estimation procedure in a realistic, large-scale empirical setting.

This application is not intended to demonstrate superior empirical fit or to provide a comprehensive benchmarking exercise. Rather, it is designed to illustrate the feasibility, interpretability, and numerical stability of the proposed EM framework when applied to

a challenging publicly available dataset characterized by severe partial observability and heterogeneous consumer behavior.

The dataset combines detailed longitudinal Amazon purchase histories with respondent-level demographic and household information for $N = 5027$ users in the United States. Data were collected through a consent-centric crowdsourcing protocol, under which participants voluntarily provided exports of their personal Amazon purchase histories spanning 1 January 2018 to 19 March 2023, together with survey-based sociodemographic information. This design yields a rich observational dataset linking transactional behavior over time to individual and household characteristics, while ensuring informed consent and preserving user privacy.

A comprehensive descriptive and exploratory analysis of this dataset has been previously reported in Berke et al. [25]. The dataset and its construction are described in detail in Berke et al. [24] and have since been used in several independent studies examining distinct dimensions of consumer behavior (e.g., [26–28]). Building on this established empirical foundation, the present study does not revisit descriptive statistics or exploratory analyses previously reported in the literature, and instead proceeds directly to the model-based analysis under partial observability.

In the present study, the dataset is used under a deliberately constructed information scenario that mirrors the decision environment faced by a focal firm. Although the underlying data contain transactions for multiple brands within the technology and electronics category, the estimation strategy intentionally restricts the information available for model fitting to Apple-specific transactional histories observed on Amazon, together with respondent-level sociodemographic and household characteristics. This restriction is methodological rather than data-driven and reflects the realistic constraint that firms typically lack access to competitors' sales and to consumers' total category expenditure. Within this setting of partial observability, the *pogit* model illustrates how latent wallet allocation and underlying demand intensity can be inferred using focal-firm transaction data alone.

The model specification captures two distinct latent behavioral mechanisms. An intensity component governs the customer's overall demand for technology and electronics, defined on a normalized latent scale and corresponding to the Size-of-Wallet (SioW). An allocation component governs the fraction of this latent category demand allocated to the focal firm, corresponding to the share-of-wallet (SoW). Identification under partial observability is achieved through the joint hierarchical structure of the model and a deliberate separation of covariates across these latent components. Apple-specific transactional and behavioral features enter the allocation mechanism, whereas broader sociodemographic and household characteristics enter the intensity mechanism.

In this empirical application, all customers are observed over the same calendar window and are, therefore, assumed to face identical exposure. Accordingly, the offset term is fixed at a constant value, $E_i \equiv 1$ for all i , and cross-sectional heterogeneity in overall purchasing activity is captured entirely through the latent intensity $\lambda_i = \exp(x'_{2i}\beta_2)$.

To operationalize the model in a count-data setting, we transform Apple expenditure observed on Amazon during the evaluation window T_2 (1 November 2021 to 19 March 2023) using a variance-to-mean scaling motivated by the moment structure of the Poisson distribution; a model-based and data-driven justification of this transformation is provided in Appendix D. Specifically, total monetary spending is rescaled by a constant c , estimated exclusively from the training sample, and subsequently rounded to obtain a count-like response variable compatible with the *pogit* specification. This transformation induces a normalized discrete scale shared by the observed response y_i and the latent exposure n_i , thereby allowing monetary purchase volumes to be modeled coherently within a Poisson–

binomial structure. Importantly, neither y_i nor n_i should be interpreted as literal counts of physical transactions, orders, or items; both quantities represent discretized proxies for relative spending intensity defined on a common latent scale.

Predictors are computed exclusively from transactions in T_1 (1 January 2018 to 31 October 2021), whereas the response is computed exclusively from transactions in T_2 (1 November 2021 to 19 March 2023); the train/test partition is defined at the respondent level and applied consistently across both windows. Categorical predictors encoded as single-selection factors are included using an omitted reference level, which defines the reference group, whereas for multiple-selection categorical variables, all indicator categories are retained. All quantitative covariates are standardized prior to estimation. For quantitative transactional predictors, correlation-based screening is performed using Spearman rank correlations to accommodate potential nonlinearity and heavy-tailed distributions. All data-adaptive preprocessing steps—including winsorization thresholds, standardization parameters, and correlation-screening rules—are learned exclusively on the training sample ($n_{train} = 1336$) and subsequently applied unchanged to the held-out test sample ($n_{test} = 334$), thereby ensuring strict prevention of information leakage.

Although the full Amazon dataset permits observation of technology and electronics spending across multiple brands within the Amazon channel during the evaluation window T_2 , information on non-Apple spending is used exclusively ex post as a channel-restricted benchmark for validation and contextualization of the model-implied SoW.

To ensure transparency and reproducibility of the empirical application, we provide a complete technical report in Supplementary File S2, including data pre-processing procedures, design matrices, coefficient estimates, and detailed catalogs of predictor and response variables. Reproducible implementations of the real-data application in Julia and R are provided in Supplementary File S3.

Against this methodological background, Tables 1 and 2 summarize the main empirical patterns recovered by the plogit model under the full specification and after covariate selection, respectively. The full specification incorporates the complete set of covariates considered in the empirical analysis and reveals that most of the model's explanatory content is concentrated in a relatively small subset of predictors related to Apple-specific transactional behavior and household characteristics—most notably RFM variables, income categories, and platform usage intensity—while the remaining covariates contribute limited additional explanatory power, motivating a more parsimonious specification to facilitate interpretation.

In the full specification (Table 1), variation in SoW is primarily driven by Apple-specific transactional behavior. Both Frequency and Monetary enter with positive and statistically significant coefficients (estimates 0.622 and 0.347, respectively), indicating that customers who purchase Apple products more frequently and spend more on the focal brand allocate a larger share of their latent category demand to Apple, whereas Recency does not reach statistical significance once the full set of transactional and demographic controls is included. Heterogeneity in SioW is mainly associated with household income and platform usage intensity: the Income 100–149k category exhibits a positive and statistically significant association with latent purchase intensity (estimate 0.357), while the highest income group does not reach conventional significance levels. Platform engagement plays a central role, with accounts shared by three individuals (Use = 3) displaying substantially higher latent category demand (estimate 0.589). In addition, the life-event indicator Life: Became pregnant enters with a negative and statistically significant coefficient, suggesting a temporary reduction in latent purchasing intensity, whereas other demographic and life-event controls do not exhibit statistically significant effects.

Table 1. Estimated coefficients for the plogit model (full specification). Confidence intervals are at the 95% level. Statistically significant *p*-values (5%) are shown in bold.

Covariate	Estimates	<i>p</i> -Value ^{g,h}	Lower CI	Upper CI
SoW component				
Intercept	−0.736	0.044	−1.452	−0.021
Recency	−0.155	0.207	−0.396	0.086
Frequency	0.622	<0.001	0.289	0.956
Monetary	0.347	0.004	0.111	0.582
A.M.Spend ^a	0.158	0.834	−1.320	1.637
M.M.Spend ^b	−0.073	0.919	−1.483	1.337
R.A.M.Units ^c	−0.026	0.866	−0.333	0.280
SioW/intensity component				
Intercept	−0.437	0.207	−1.115	0.241
Race: Black ^d	0.345	0.166	−0.144	0.833
Race: White ^d	0.262	0.262	−0.196	0.721
Race: Asian ^d	0.330	0.168	−0.139	0.800
Race: Other ^d	0.234	0.479	−0.414	0.882
Race: Native American ^d	−0.424	0.344	−1.303	0.454
Race: Pacific Islander ^d	−0.606	0.515	−2.431	1.220
Life: Lost job ^d	−0.283	0.093	−0.613	0.047
Life: Moved residence ^d	−0.104	0.379	−0.335	0.128
Life: Divorce ^d	0.132	0.792	−0.849	1.113
Life: Had a child ^d	0.003	0.991	−0.523	0.529
Life: Became pregnant ^d	−0.863	0.027	−1.630	−0.096
Age 25–34 ^e	0.027	0.851	−0.255	0.309
Age 35–44 ^e	−0.032	0.843	−0.345	0.281
Age 45–54 ^e	0.169	0.309	−0.156	0.493
Age 55–64 ^e	0.263	0.183	−0.124	0.649
Age ≥ 65 ^e	−0.499	0.153	−1.183	0.186
Income 50–74 k ^e	0.326	0.035	0.022	0.630
Income 75–99 k ^e	0.097	0.572	−0.239	0.433
Income 100–149 k ^e	0.357	0.021	0.053	0.661
Income ≥ 150 k ^e	0.301	0.070	−0.025	0.626
Income < 25 k ^e	−0.016	0.941	−0.438	0.406
Income: Prefer not to say ^e	−0.021	0.963	−0.909	0.867
Use = 2 ^e	−0.023	0.840	−0.247	0.201
Use = 3 ^e	0.589	<0.001	0.323	0.855
Use = 4+ ^e	0.281	0.131	−0.084	0.646
States ^f	0.014	0.748	−0.069	0.096

^a A.M.Spend: average monthly spend (USD) over the last 12 months. ^b M.M.Spend: maximum monthly spend (USD) over the last 12 months. ^c R.A.M.Units: ratio of average monthly units to peak monthly units (12-month window). ^d Race, Life: multiple-selection indicator variables. ^e Age, Income, Use: single-selection factor indicators. ^f States: number of distinct shipping states. ^g The *p*-values were computed assuming normality. ^h The underlying variances were computed with the open product gradient method.

In the parsimonious specification (Table 2), SoW is sharply characterized by Apple-specific behavioral variables in the allocation component. Frequency and Monetary remain the dominant determinants, with positive and statistically significant coefficients (estimates 0.305 and 0.210, respectively), confirming the central role of repeated purchasing and cumulative spending in shaping wallet allocation toward the focal firm. In contrast to the full model, Recency now enters with a negative and statistically significant coefficient (estimate −0.143), indicating that more recent Apple purchases are associated with a higher share of wallet once irrelevant covariates are removed. Under the same parsimonious specification, heterogeneity in SioW is primarily driven by household income and platform usage intensity: the Income 50–74 k and Income 100–149 k categories are positively and

significantly associated with latent category demand (estimates 0.396 and 0.432, respectively), while platform engagement remains a dominant factor, with accounts shared by three individuals (Use = 3) exhibiting a strong and highly significant positive association with intensity (estimate 0.615). Consistent with the full specification, the life-event indicator Life: Became pregnant retains a negative and statistically significant effect, whereas other retained income and usage categories do not reach conventional significance levels.

Table 2. Estimated coefficients for the plogit model after covariate selection. Confidence intervals are at the 95% level. Statistically significant *p*-values (5%) are shown in bold.

Covariate	Estimates	<i>p</i> -Value ^{c,d}	Lower CI	Upper CI
SoW component				
Intercept	−2.390	0.123	−5.429	0.650
Recency	−0.143	0.022	−0.265	−0.020
Frequency	0.305	0.002	0.114	0.497
Monetary	0.210	0.002	0.078	0.341
SioW/intensity component				
Intercept	1.078	0.445	−1.691	3.848
Life: Lost job ^a	−0.122	0.438	−0.431	0.187
Life: Moved residence ^a	−0.136	0.238	−0.362	0.090
Life: Divorce ^a	0.088	0.860	−0.885	1.060
Life: Had a child ^a	0.024	0.925	−0.483	0.531
Life: Became pregnant ^a	−0.858	0.026	−1.610	−0.105
Income 50–74 k ^b	0.396	0.010	0.093	0.700
Income 75–99 k ^b	0.236	0.160	−0.093	0.565
Income 100–149 k ^b	0.432	0.005	0.132	0.733
Income ≥ 150 k ^b	0.318	0.059	−0.012	0.648
Income < 25 k ^b	0.068	0.750	−0.350	0.486
Income: Prefer not to say ^b	0.170	0.693	−0.674	1.014
Use = 2 ^b	0.032	0.776	−0.186	0.250
Use = 3 ^b	0.615	<0.001	0.354	0.875
Use = 4+ ^b	0.277	0.126	−0.078	0.633

^a Life: multiple-selection indicator variables. ^b Income, Use: single-selection factor indicators. ^c The *p*-values were computed assuming normality. ^d The underlying variances were computed with the open product gradient method.

7. Conclusions

This paper develops a scalable expectation-maximization algorithm for the Poisson-logistic (plogit) model, a classical framework for count data subject to partial observability. By exploiting a Polya-Gamma data augmentation, the proposed approach yields a quadratic complete-data log-likelihood and closed-form updates in both the E- and M-steps. As a result, each EM iteration is computationally inexpensive, numerically stable, and well-suited to large datasets.

The primary contribution is methodological. Unlike existing frequentist and Bayesian approaches, which typically rely on generic numerical optimization or Markov chain Monte Carlo methods, the proposed EM formulation scales naturally with sample size and covariate dimension. The algorithm admits deterministic full-batch updates, mini-batch variants based on Robbins–Monro stochastic approximation, and quasi-Newton acceleration via SQUAREM, all targeting the same observed-data likelihood. Simulation experiments confirm stable finite-sample behavior, robustness to the negative-binomial approximation underlying the Polya-Gamma construction, and substantial computational gains relative to direct maximum-likelihood optimization.

The proposed framework also suggests several directions for future research. First, the quadratic structure of the M-step makes the algorithm particularly amenable to regularization in high-dimensional settings. Penalized extensions based on ℓ_1 penalties or global–local shrinkage priors, such as the horseshoe, could be incorporated either through penalized M-steps or additional latent-variable augmentations. Second, the Polya-Gamma construction naturally links the present EM approach to variational Bayes methods. In a related direction, exploring variational Bayes approximations built on similar augmentation schemes may yield fast approximate Bayesian procedures that complement the EM algorithm developed here.

More broadly, the results highlight the role of Polya-Gamma augmentation as a unifying computational device for efficient inference in models combining nonlinear link functions and discrete outcomes. By making frequentist estimation of plogit models feasible at scale, the proposed EM algorithm expands the practical applicability of these models in modern empirical settings.

Finally, it is worth noting that the proposed EM estimator targets the maximum–likelihood solution of the plogit model, and therefore, inherits the usual asymptotic properties of likelihood–based inference under standard regularity conditions. While Bayesian Polya-Gamma formulations naturally provide finite-sample posterior uncertainty, the present approach relies on asymptotic inference based on the observed Fisher information, yielding a simple and scalable frequentist alternative suited to large datasets.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/e28020207/s>. Supplementary File S1: Replication package for the simulation experiments (scripts to reproduce the simulation study) (ZIP). Supplementary File S2: Report of the real application (variable catalog, design-matrix construction, and preprocessing details) (PDF). Supplementary File S3: Replication package for the real application (datasets and Julia/R scripts to construct the design matrices and reproduce the empirical analyses) (ZIP). Reference [29] is cited in the Supplementary Materials.

Author Contributions: Conceptualization, I.G., L.J., and S.R.; methodology, I.G., L.J., and S.R.; software, I.G. and S.R.; writing, original draft preparation, I.G. and S.R.; writing—review and editing, I.G. and S.R.; funding acquisition and resources, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: S.R. received sponsorship from Pontificia Universidad Javeriana, Cali, Colombia (ROR: <https://ror.org/03etyjw28>, accessed on 6 February 2026), through the project under grant agreement “PUJC-CTD01 Nuevos Doctores”. I.G. was supported by CONICYT PFCHA/Doctorado Becas Chile/2020-21201742.

Data Availability Statement: The empirical application in our manuscript is based on an open-access dataset, originally introduced and documented in Berke et al. [24]. No proprietary or restricted data are used in this study. While the raw data are openly available from the original source, the variables analyzed in our paper are derived and constructed by the authors through preprocessing and feature-engineering steps required for the implementation of the proposed plogit model. The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors would like to thank the referees for their valuable suggestions and recommendations, which significantly contributed to improving the quality of this manuscript. S.R. gratefully acknowledges the Departamento de Ciencias Naturales y Matemáticas, Facultad de Ingeniería y Ciencias at Pontificia Universidad Javeriana, Cali (Colombia), and I.G. gratefully acknowledges the Departamento de Economía y Administración, Facultad de Economía y Negocios at Universidad Andrés Bello, Santiago (Chile), for the research time and institutional support made available within their academic duties.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Local Identification Conditions

In this section, we study the local identification of the plogit parameters. Local identification is the relevant notion in our context, as it characterizes whether parameters can be uniquely recovered—up to arbitrarily small perturbations—from the distribution of the observed data, and it is the standard concept underlying asymptotic inference in nonlinear parametric models.

We write $x'_i = (x'_{i1}, x'_{i2})$, where x_{i1} and x_{i2} denote the covariate vectors associated with the reporting and intensity components, respectively.

Throughout this section, x_i denotes the vector of observed covariates (regressors), while y_i denotes the observed count response.

Definition A1. Let $\theta \in \Theta \subset \mathbb{R}^M$ ($M < \infty$) denote the vector of model parameters, and let P_θ denote the probability distribution of the observed data $x = (x_1, \dots, x_n)$. Two parameter points θ and θ' are said to be observationally equivalent if $P_\theta = P_{\theta'}$. A parameter point θ_0 is locally identified if there exists an open neighborhood $\mathcal{U}(\theta_0)$ such that no $\theta \in \mathcal{U}(\theta_0) \setminus \{\theta_0\}$ is observationally equivalent to θ_0 .

Local identification ensures that the likelihood function has a unique maximizer in a neighborhood of the true parameter value, ruling out local ridges or flat directions that would prevent reliable inference.

Rothenberg [30] provides sufficient conditions for local identification. In order to express them succinctly, we need the notions of regular models and regular points.

Definition A2. Consider a parametric statistical model $(S, \mathcal{F})$, where S is the sample space and $\mathcal{F} = \{f(\cdot | \theta) : \theta \in \Theta\}$ is a family of probability densities or mass functions. The model is said to be regular if

1. Θ is an open subset of $\mathbb{R}^M$, for some $M < \infty$;
2. the support of $f(\cdot | \theta)$ does not depend on θ ;
3. there exists an open set $\Psi \supseteq \Theta$ such that $f(y | \theta)$ and $\log f(y | \theta)$ are continuously differentiable with respect to θ for all $\theta \in \Psi$ and all y in the support;
4. the Fisher information matrix

$$I(\theta) = \mathbb{E}[\nabla_\theta \log f(y | \theta) \nabla_\theta \log f(y | \theta)']$$

exists and is continuous on Θ .

Definition A3. Let $M(\theta)$ be a matrix whose entries are continuous functions of θ on Θ . A point $\theta_0 \in \Theta$ is called a regular point of $M(\cdot)$ if there exists an open neighborhood of θ_0 over which $M(\theta)$ has constant rank.

Remark A1. These definitions extend immediately to regression models by replacing $\log f(y | \theta)$ with $\log f(y | x, \theta)$ and taking expectations over the joint distribution of (x, y) .

Under these conditions, local identification can be characterized entirely in terms of the Fisher information matrix.

Theorem A1 (Rothenberg, 1971). Consider a regular parametric model with parameter $\theta \in \Theta$ and Fisher information matrix $I(\theta)$. If θ_0 is a regular point of $I(\theta)$, then θ_0 is locally identified if and only if $I(\theta_0)$ is non-singular.

With these definitions and theorems, we are now ready to prove Theorem 1. To this end, we start with the following lemma:

Lemma A1. *Given a pogit model, define $z'_i = ((1 - \theta_i)x'_{i1}, x'_{i2})$. Suppose $\{(y_i, x_i)\}$ is an i.i.d. sequence. Then, the Fisher information is given by $I(\beta) = NI_1(\beta)$, with $I_1(\beta) = E[\theta_1 \lambda_1 z_1 z'_1]$.*

Proof. For i.i.d. data, the Fisher information is $I(\beta) = N E[E[s_1 s'_1 | x]]$, where s_1 is the 1st contribution to the score function. To compute such a contribution, observe that the first contribution to the likelihood is

$$\ell_1 = \log \text{Poisson}(y_1 | \lambda_1 \theta_1) = -\theta_1 \lambda_1 + y_1 \log(\theta_1 \lambda_1) - \log y_1!,$$

Deriving this expression with respect to β_1 and β_2 , the following is obtained

$$s_1(\beta) = \begin{bmatrix} \nabla_{\beta_1} \ell_1 \\ \nabla_{\beta_2} \ell_1 \end{bmatrix} = \begin{bmatrix} (y_1 - \theta_1 \lambda_1)(1 - \theta_1)x_{11} \\ (y_1 - \theta_1 \lambda_1)x_{12} \end{bmatrix} = (y_1 - \theta_1 \lambda_1)z_1.$$

Thus, $s_1 s'_1 = (y_1 - \theta_1 \lambda_1)^2 z_1 z'_1$. Now, under the model, $(y_1 | x_1) \sim \text{Poisson}(\lambda_1 \theta_1)$. Hence, exploiting some well-known properties of the Poisson distribution, the following expression is obtained

$$E[(y_1 - \theta_1 \lambda_1)^2 | x_1] = V[y_1 | x_1] = \theta_1 \lambda_1,$$

so $E[s_1 s'_1 | x_1] = \theta_1 \lambda_1 z_1 z'_1$ and the result follows. $\square$

With this result, we can now prove Theorem 1.

Proof of Theorem 1. Clearly, the pogit model is regular, so it only remains to prove that $I(\beta_0)$ is non-singular for any regular point β_0 . On the other hand, by Lemma A1, $I(\beta_0) = NE[\theta_1 \lambda_1 z_1 z'_1]$ (with β evaluated at β_0). Hence, it suffices to prove that $E[\theta_1 \lambda_1 z_1 z'_1]$ is a positive definite matrix. To do so, suppose the opposite. Then,

$$\zeta' E[\theta_1 \lambda_1 z_1 z'_1] \zeta = E[\theta_1 \lambda_1 (z'_1 \zeta)^2] = 0$$

for some $\zeta \neq 0_{K_1+K_2}$, where K_1 and K_2 denote the dimensions of x_{i1} and x_{i2} , respectively, which implies that $z'_1 \zeta = 0$ (a.s.) because $\theta_1 \lambda_1 > 0$, but $z'_1 \zeta$ cannot be zero (a.s.) because then $E[z_1 z'_1] = 0$, so $P(z'_1 \zeta = 0) < 1$; yielding a contradiction. $\square$

Appendix B. Full Derivation of the Complete Likelihood

Throughout this section, for each observational unit $i = 1, \dots, N$, we work with linear predictors $\eta_{i1} = x'_{i1} \beta_1$ and $\eta_{i2} = x'_{i2} \beta_2$, which correspond to the binomial (reporting) and latent intensity components of the pogit model, respectively. For notational convenience, likelihood expressions are often written in terms of these linear predictors rather than the natural parameters θ_i and λ_i , where $\theta_i = \text{sigmoid}(\eta_{i1})$ and $\lambda_i = \exp(\eta_{i2})$.

Before proceeding, note that the auxiliary quantities a , b , and ψ introduced below are temporary variables used solely to apply the Polya-Gamma identity in Property 2. They are explicitly defined within each likelihood component and should not be interpreted as structural parameters of the pogit model.

Throughout the augmented likelihood derivations, $E_{w_{ij}}$ denotes expectation with respect to the Polya-Gamma latent variable w_{ij} , with $j \in \{1, 2\}$, introduced by Property 2.

Here $r_i \gg 0$ is a fixed approximation parameter controlling the negative-binomial approximation to the Poisson distribution.

Appendix B.1. Binomial Component

We know $\theta_i = \text{sigmoid}(\eta_{i1})$. Hence,

$$\begin{aligned}\text{Bin}(y_i | n_i, \theta_i) &\propto \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i} \\ &\propto \exp(\eta_{i1})^{y_i} / (1 + \exp(\eta_{i1}))^{n_i} \\ &\propto (e^\psi)^a / (1 + e^\psi)^b,\end{aligned}$$

where $\psi = \eta_{i1}$, $a = y_i$ and $b = n_i$. Applying Property 2 gives the desired result.

Appendix B.2. Poisson Component

We know $\lambda_i = e^{\eta_{i2}}$ so $E_i \lambda_i = r_i e^{\eta_{i2} - \log(r_i/E_i)}$ and

$$\begin{aligned}\text{Poisson}(n_i | E_i \lambda_i) &\approx \text{NegBin}(n_i | r_i, E_i \lambda_i / (r_i + E_i \lambda_i)) \\ &\propto \binom{r_i + n_i - 1}{r_i - 1} \left(\frac{r_i}{r_i + E_i \lambda_i} \right)^{r_i} \left(\frac{E_i \lambda_i}{r_i + E_i \lambda_i} \right)^{n_i} \\ &\propto \frac{(r_i e^{\eta_{i2} - \log(r_i/E_i)})^{n_i}}{(r_i + r_i e^{\eta_{i2} - \log(r_i/E_i)})^{n_i + r_i}} \\ &\propto \frac{(e^{\eta_{i2} - \log(r_i/E_i)})^{n_i}}{(1 + e^{\eta_{i2} - \log(r_i/E_i)})^{n_i + r_i}} \\ &\propto (e^\psi)^a / (1 + e^\psi)^b\end{aligned}$$

where $\psi = \eta_{i2} - \log(r_i/E_i)$, $a = n_i$ and $b = r_i + n_i$. Applying Property 2 gives

$$\text{Poisson}(n_i | E_i \lambda_i) \propto E_{w_{i2}} \exp\{(a - b/2)\psi - w_{i2}\psi^2/2\}$$

with $w_{i2} \sim \text{PG}(b, 0)$. However,

$$\begin{aligned}(a - b/2)\psi &\doteq (a - b/2)\eta_{i2}, \\ w_{i2}\psi^2/2 &\doteq w_{i2}\eta_{i2}^2/2 - w_{i2}\log(r_i/E_i)\eta_{i2} + \frac{w_{i2}}{2}(\log(r_i/E_i))^2.\end{aligned}$$

Noting that the symbol $\doteq$ denotes equality up to additive terms independent of the linear predictor and absorbable into the normalizing constant, we have

$$\begin{aligned}(a - b/2)\psi - w_{i2}\psi^2/2 &\doteq (a - b/2)\eta_{i2} - w_{i2}\eta_{i2}^2/2 + w_{i2}\log(r_i/E_i)\eta_{i2} - \frac{w_{i2}}{2}(\log(r_i/E_i))^2 \\ &\doteq ((r_i - n_i)/2 + w_{i2}\log(r_i/E_i))\eta_{i2} - w_{i2}\eta_{i2}^2/2 - \frac{w_{i2}}{2}(\log(r_i/E_i))^2 \\ &\doteq \kappa_{i2}\eta_{i2} - w_{i2}\eta_{i2}^2/2 - \frac{w_{i2}}{2}(\log(r_i/E_i))^2,\end{aligned}$$

with κ_{i2} defined in the main document.

Exponentiating and taking expectations, we obtain

$$\text{Poisson}(n_i | E_i \lambda_i) \propto E_{w_{i2}} \exp\{\kappa_{i2}\eta_{i2} - w_{i2}\eta_{i2}^2/2 - \frac{w_{i2}}{2}(\log(r_i/E_i))^2\},$$

as stated in the main document.

Appendix C. Additional Simulation Experiments

This appendix reports additional simulation experiments designed to assess the impact of the Poisson-to-negative binomial approximation underlying the Polya–Gamma

augmentation on likelihood-based inference. In particular, we examine the sensitivity of the standard errors, effective iterations and observed-data likelihood to the choice of the approximation parameter. These supplementary results complement the main simulation study by providing direct diagnostics for inferential stability and support the practical recommendations adopted in the main text.

Figure A1 illustrates how likelihood-based standard errors vary with the negative-binomial approximation parameter r , where the standard errors were calculated using OPG (outer product gradient) method. To this end, we simulated 100 datasets of size $N = 500$ using the same data-generating process used in Section 5. Across the range of values considered, the estimated standard errors exhibit only minor fluctuations and no discernible trend as r increases. This indicates that, in the simulation settings examined, inferential uncertainty is largely insensitive to the precise choice of r once it is sufficiently large.

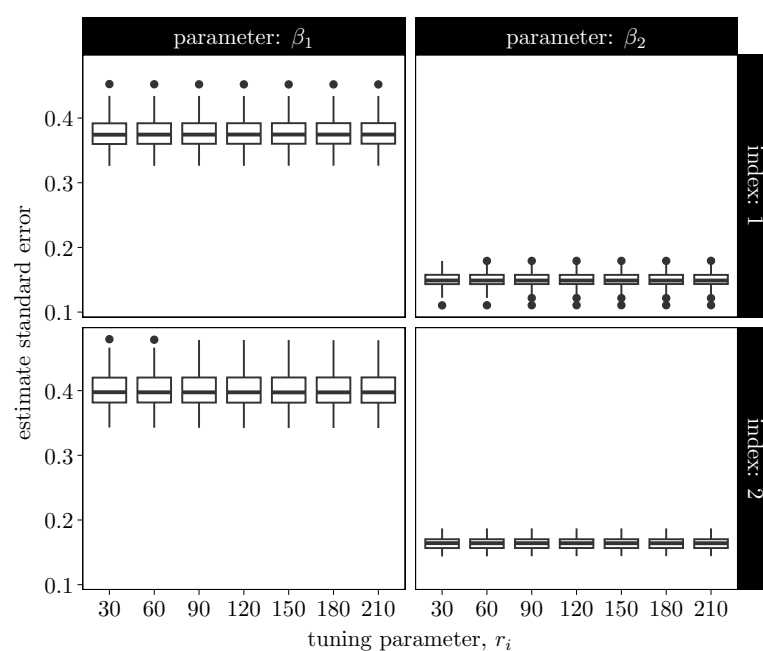


Figure A1. Estimate standard errors as a function of the tuning parameter r_i . The standard errors are relatively stable across values of r_i .

Figure A2 shows how the observed-data log-likelihood varies with the negative-binomial approximation parameter r . To improve readability and avoid an overcrowded spaghetti plot, we track the likelihood trajectories for five simulated datasets of size $N = 500$, generated as described in Section 5. As shown in the figure, the likelihood increases rapidly for small values of r and then stabilizes as r grows. In all trajectories, changes become negligible once r reaches values around $r \approx 100$, indicating that the approximation is sufficiently accurate beyond this point. This behavior supports our practical recommendation to use moderate values of r , as larger choices yield no meaningful improvement in likelihood.

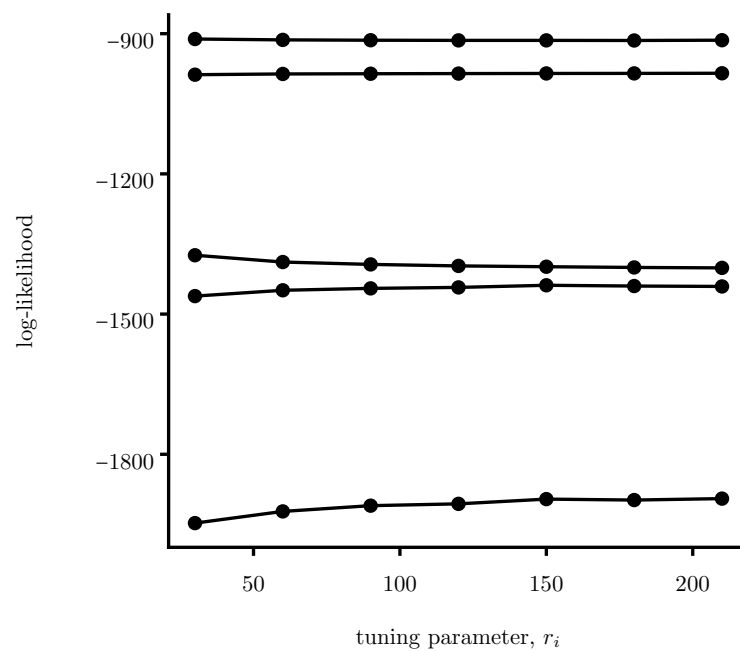


Figure A2. Observed log-likelihood as a function of the tuning parameter r_i , for 5 simulated datasets. The log-likelihood are relatively stable for $r_i \geq 100$.

Figure A3 documents how the effective number of EM iterations varies with the approximation parameter r . The number of iterations increases monotonically with r , indicating higher computational cost for more refined approximations. At the same time, earlier results show that estimates, standard errors, and likelihood values stabilize for moderate values of r . Taken together, these findings suggest that choosing r around 100 is sufficient in practice, as larger values mainly increase runtime without yielding material inferential gains.

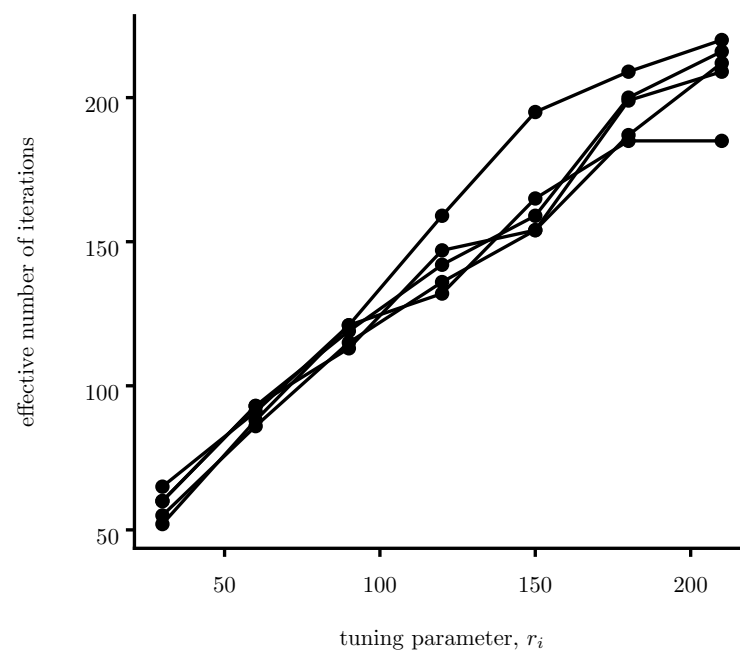


Figure A3. Effective number of iterations as a function of the tuning parameter r_i , for 5 simulated datasets. The log-likelihood are relatively stable for $r_i \geq 100$.

Appendix D. Discretization of Monetary Spending

Following Glady and Croux [2], when monetary volumes are observed, spending can be discretized into count-like units prior to estimation in Poisson-based models. In the real data application, we consider Apple spending on Amazon within the technology/electronics universe $\mathcal{U}$ over the evaluation window T_2 , denoted by $\text{spend}_i^{\text{Apple}|\text{Amz},\mathcal{U}}$.

This appendix provides a methodological justification of the discretization procedure used in the real-world application, grounded in the statistical properties of the data and the Poisson modeling framework. To avoid information leakage, all quantities used to define the discretization scale are computed exclusively on the training sample. Let $\mathcal{I}_{\text{tr}}$ denote the training index set. We define the discretization constant as

$$c = \frac{\hat{\sigma}_{\text{tr}}^2 \text{Apple}|\text{Amz},\mathcal{U}}{\overline{\text{spend}}_{\text{tr}}^{\text{Apple}|\text{Amz},\mathcal{U}}},$$

that is, the ratio between the cross-sectional variance and the mean of Apple spending in the training set. This choice is motivated by the Poisson assumption underlying the model, under which the mean and variance coincide on the latent model scale.

Prior to computing c , extreme values are mitigated using a Tukey rule-based winsorization procedure, with thresholds estimated on the training sample and subsequently applied unchanged to the test sample. The discretized response is then defined as

$$y_i = \text{round}\left(\frac{\text{spend}_i^{\text{Apple}|\text{Amz},\mathcal{U}}}{c}\right).$$

The operator $\text{round}(\cdot)$ denotes rounding to the nearest integer. The resulting variable y_i is a normalized, count-like proxy for monetary spending that is compatible with the Poisson likelihood employed in the plogit model. Importantly, neither y_i nor the latent category size n_i should be interpreted as literal counts of transactions or items; both quantities are defined on a common discretized scale designed to capture relative spending intensity.

References

1. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2013.
2. Glady, N.; Croux, C. Predicting customer wallet without survey data. *J. Serv. Res.* **2009**, *11*, 219–231. [\[CrossRef\]](#)
3. Stoner, O.; Economou, T.; Drummond Marques da Silva, G. A hierarchical framework for correcting under-reporting in count data. *J. Am. Stat. Assoc.* **2019**, *114*, 1481–1492. [\[CrossRef\]](#)
4. Papadopoulos, G. Immigration status and property crime: An application of estimators for underreported outcomes. *IZA J. Migr.* **2014**, *3*, 12. [\[CrossRef\]](#)
5. Polettini, S.; Arima, S.; Martino, S. An investigation of models for under-reporting in the analysis of violence against women in Italy. *Soc. Indic. Res.* **2024**, *175*, 1007–1026. [\[CrossRef\]](#)
6. Bradshaw, C.; Blei, D.M. A Bayesian model of underreporting for sexual assault on college campuses *Ann. Appl. Stat.* **2024**, *18*, 3146–3164. [\[CrossRef\]](#)
7. Winkelmann, R.; Zimmermann, K.F. *Poisson-Logistic Regression*; Working Paper 93–18; Department of Economics, University of Munich: Munich, Germany, 1993.
8. Dvořák, J. M.; Wagner, H. Sparse Bayesian modelling of underreported count data. *Stat. Model.* **2015**, *16*, 24–46. [\[CrossRef\]](#)
9. Arima, S.; Polettini, S.; Pasculli, G.; Gesualdo, L.; Pesce, F.; Procaccini, D.-A. A Bayesian nonparametric approach to correct for underreporting in count data. *Biostatistics* **2023**, *25*, 904–918. [\[CrossRef\]](#)
10. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38. [\[CrossRef\]](#)
11. Polson, N.G.; Scott, J.G.; Windle, J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Stat. Assoc.* **2013**, *108*, 1339–1349. [\[CrossRef\]](#)
12. Scott, J.G.; Sun, S. Expectation-maximization for logistic regression. *arXiv* **2013**, arXiv:1306.0040. [\[CrossRef\]](#)

13. D'Angelo, S.; Canale, A. Efficient Posterior sampling for Bayesian Poisson regression. *J. Comput. Graph. Stat.* **2023**, *32*, 917–926. [CrossRef]
14. Kingman, J.F.C. *Poisson Processes*; Oxford University Press: Oxford, UK, 1992.
15. Buntine, W. Operations for learning with graphical models. *J. Artif. Intell. Res.* **1994**, *2*, 159–225. [CrossRef]
16. Brennan, J.; Bannick, M.; Kassebaum, N.; Wilner, L.; Thomson, A.; Aravkin, A.; Zheng, P. Analysis and Methods to Mitigate Effects of Under-Reporting in Count Data. *arXiv* **2021**, arXiv:2109.12247. [CrossRef]
17. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
18. Durante, D.; Rigon, T. Conditionally conjugate mean-field variational Bayes for logistic models. *Statist. Sci.* **2019**, *34*, 472–485. [CrossRef]
19. Robbins, H.; Monro, S. A Stochastic Approximation Method. *Ann. Math. Statist.* **1951**, *22*, 400–407. [CrossRef]
20. Cappé, O.; Moulines, E. On-line expectation–maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 593–613. [CrossRef]
21. Polyak, B. A new method of stochastic approximation type. *Autom. Remote Control* **1990**, *51*, 937–946.
22. Polyak, B.; Juditsky, A. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **1992**, *30*, 838–855. [CrossRef]
23. Varadhan, R.; Roland, C. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.* **2008**, *35*, 335–353. [CrossRef]
24. Berke, A.; Calacci, D.; Mahari, R.; Yabe, T.; Larson, K.; Pentland, S. Open e-commerce 1.0, five years of crowdsourced U.S. Amazon purchase histories with user demographics. *Sci. Data* **2024**, *11*, 491. [CrossRef]
25. Berke, A.; Mahari, R.; Pentland, S.; Larson, K.; Calacci, D. Insights from an experiment crowdsourcing data from thousands of U.S. Amazon users: The importance of transparency, money, and data use. *Proc. ACM Hum.-Comput. Interact.* **2024**, *8*, 466. [CrossRef]
26. Ifrim, A.-M.; Oncioiu, I. A hybrid numerical–semantic clustering algorithm based on scalarized optimization. *Algorithms* **2025**, *18*, 607. [CrossRef]
27. Hang, Z.; Zeng, L. The Cost of Inaccessibility: Retail Discrimination and Mobility-Constrained Consumers. Kilts Center at Chicago Booth Marketing Data Center Paper. 2025. Available online: <https://ssrn.com/abstract=5496018> (accessed on 1 February 2026).
28. Berke, A.; Calacci, D.; Pentland, S.; Larson, K. Evaluating Amazon effects and the limited impact of COVID-19 with purchases crowdsourced from U.S. consumers. *PLoS ONE* **2025**, *20*, e0336571. [CrossRef]
29. Li, X.; Shemshadi, A.; Olech, Ł.; Michalewicz, Z. Customer wallet share estimation for manufacturers based on transaction data. In *Data Mining*; Springer: Singapore, 2019; pp. 171–182. [CrossRef]
30. Rothenberg, T.J. Identification in parametric models. *Econometrica* **1971**, *39*, 577–591. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.