Scientific Research Publishing

Advances in Pure Mathematics, 2025, 15(5), 319-338
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Generalized Bayesian Inference for Regression-Type Models with an Intractable Normalizing Constant

Qinqin Gan, Wanzhou Ye

Department of Mathematics, College of Science, Shanghai University, Shanghai, China
Email: gqq20001205@163.com, wzhy@shu.edu.cn

How to cite this paper: Gan, Q.Q. and Ye, W.Z. (2025) Generalized Bayesian Inference for Regression-Type Models with an Intractable Normalizing Constant. *Advances in Pure Mathematics*, **15**, 319-338.
https://doi.org/10.4236/apm.2025.155016

Received: May 9, 2025
Accepted: May 20, 2025
Published: May 23, 2025

Abstract

Regression models with intractable normalizing constants are valuable tools for analyzing complex data structures, yet parameter inference for such models remains highly challenging—particularly when observations are discrete. In statistical inference, discrete state spaces introduce significant computational difficulties, as the normalizing constant often requires summation over extremely large or even infinite sets, which is typically infeasible in practice. These challenges are further compounded when observations are independent but not identically distributed. This paper addresses these issues by developing a novel generalized Bayesian inference approach tailored for regression models with intractable likelihoods. The key idea is to employ a specific form of generalized Fisher divergence to update beliefs about the model parameters, thereby circumventing the need to compute the normalizing constant. The resulting generalized posterior distribution can be sampled using standard computational tools, such as Markov Chain Monte Carlo (MCMC), effectively avoiding the intractability of the normalizing constant.

Keywords

Intractable Normalizing Constant, Fisher Divergence, Conway-Maxwell-Poisson Regression

1. Introduction

In statistics, many models possess an intractable normalizing constant [1]. Inference for these models is complicated, because the normalizing functions of their probability distributions include the parameters of interest. In Bayesian analysis, they result in so-called doubly intractable posterior distributions which pose sig-

nificant computational challenges [2].

In Bayesian inference, the posterior distribution of parameters θ given observed data y is given by:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}, \tag{1}$$

where $p(y) = \int p(y \mid \theta)\, p(\theta) \mathrm{d}\theta$ is the marginal likelihood. The posterior is **intractable** when $p(y)$ cannot be computed analytically. The problem becomes **doubly intractable** when the likelihood itself contains an unknown normalizing constant:

$$p(y \mid \theta) = \frac{f(y,\theta)}{Z(\theta)},\ Z(\theta) = \int f(y,\theta) \mathrm{d}y. \tag{2}$$

The doubly intractable property introduces two fundamental challenges: the failure of standard Markov Chain Monte Carlo (MCMC) methods and the curse of dimensionality. To address these challenges, researchers have proposed a variety of strategies, including exact methods based on auxiliary variables (e.g., Exchange Algorithm [3] [4]), approximate methods (e.g., Approximate Bayesian Computation [5]), and likelihood-free alternatives (e.g., Noise Contrastive Estimation [6]). However, these approaches often suffer from limitations, such as high computational cost or restrictive model assumptions.

KSD-Bayes [7] is the first generalized Bayesian inference method proposed for handling models with intractable likelihoods. It constructs a generalized posterior distribution by minimising a loss function known as the Kernel Stein Discrepancy (KSD). This approach avoids the need to explicitly compute the normalising constant, offers robustness to model misspecification, and, under certain conditions, allows for either standard MCMC sampling or closed-form solutions. In short, KSD-Bayes serves as a more flexible and robust alternative to traditional Bayesian inference. However, the generalised posterior in KSD-Bayes is dependent on a user-specified kernel function, which poses challenges in discrete domains. In such domains, natural choices of kernel are often lacking, and even when available, they can be computationally expensive. To address this issue, DFD-Bayes [8] was developed. DFD-Bayes is a generalised Bayesian inference approach based on discrete Fisher divergence, which eliminates the need for user-specified kernels and is specifically designed for intractable likelihood problems in discrete settings. However, DFD-Bayes is currently applicable only to independent and identically distributed (i.i.d.) data and does not extend to regression models. Consequently, there is a pressing need to relax the i.i.d. assumption and develop generalised Bayesian inference methods suitable for regression models. Motivated by this gap, the aim of this paper is to propose a novel generalised Bayesian inference framework for regression models with intractable normalising constants, applicable to independent but not necessarily identically distributed (INID) discrete data.

The rest of this paper is organized as following: in Section 2, we provide background of this work. Section 3 introduces the specific model structure and pre-

sents the DSFD-Bayes methodology. Section 4 presents Monte Carlo studies and case studies on real datasets, demonstrating the effectiveness of our approach. Finally, Section 5 concludes the paper with discussion.

2. Background

The aim of this section is to briefly review Generalized Bayesian Inference.

2.1. Limitations of Standard Bayesian Inference

Standard Bayesian inference, as a core paradigm of probabilistic modeling, relies critically on three fundamental assumptions: 1) correct specification of the parametric likelihood function; 2) appropriate choice of prior distribution; and 3) rigorous application of Bayes' theorem. However, modern statistical modeling practice demonstrates that these assumptions face significant challenges in complex data analysis scenarios.

First, in the case of model misspecification (M-open), when the true data-generating mechanism $f_0(x)$ is not within the assumed model family $\{p(x|\theta)\}_{\theta\in\Theta}$, posterior inferences may become unreliable [9]. This limitation is particularly pronounced in low-dimensional parameter inference.

Second, for models containing latent variables or complex dependency structures, the likelihood function $p(x|\theta)$ often involves high-dimensional integration, making the normalization constant $Z(\theta)$ analytically intractable. Even when using proxy models for approximation, parameter estimation through the KL divergence minimization remains constrained by model specification accuracy.

2.2. Generalized Bayesian Inference

Building upon these motivations, Bissiri *et al.* [10] proposed focusing inference exclusively on the target quantities of interest, rather than requiring full specification of the data-generating process. They produce what they refer to as a general Bayesian update—a coherent method to produce a posterior distribution over some quantity without relying on a full model for the observations, when considering the non-inferential Bayesian decision problem [11]. Given a prior distribution $\pi(\theta)$, an observed dataset $\{y_i\}_{i=1}^n$, and a positive scaling parameter $w>0$, the generalized Bayesian framework replaces the likelihood with a loss function $D_n(\theta)$ to define the posterior update, resulting in the generalized posterior:

$$\pi_n^D(\theta) \propto \pi(\theta)\exp\left(-wD_n(\theta)\right). \tag{3}$$

In this formulation, $D_n(\theta)$ represents a loss function quantifying the discrepancy between the parameter θ and the observed data, while w serves as a calibration parameter governing the influence of the loss. Notably, when the loss function is chosen as the negative log-likelihood, *i.e.*, $D_n(\theta) = -\sum_{i=1}^n \log p(y_i|\theta)$, the generalized posterior reduces to the standard Bayesian posterior, thereby recovering classical Bayesian inference as a special case.

The generalized Bayesian framework introduces the concept of a loss function

into the updating process, offering a fundamentally new perspective for statistical modeling and parameter inference. Compared to the traditional Bayesian paradigm, its primary advantage lies in its remarkable flexibility. Rather than requiring full specification of a probabilistic model, the generalized Bayesian approach establishes a connection between parameters and observed data through a carefully designed loss function. This feature makes it particularly well-suited for applications where the true data-generating mechanism is unknown or excessively complex. In practice, researchers can tailor the choice of loss function to the inferential objective—for example, employing the squared loss for mean estimation or the absolute error loss for median estimation—thus allowing more precise targeting of specific quantities of interest.

Importantly, the generalized Bayesian framework exhibits strong robustness properties. Traditional Bayesian inference can suffer from substantial bias under model misspecification; in contrast, the generalized approach mitigates this issue by directly targeting the loss associated with the parameter of interest, thereby reducing reliance on the correctness of the global model specification. This robustness naturally accommodates a wide range of robust statistical methods, such as the use of the Huber loss for outlier-resistant estimation. Furthermore, the framework demonstrates powerful information integration capabilities: it is capable of incorporating not only traditional random samples but also non-random information sources, such as expert knowledge, and can effectively handle special settings like partial likelihoods commonly encountered in survival analysis.

Nevertheless, the application of the generalized Bayesian framework demands careful attention to several critical aspects. First, the choice and calibration of the loss function are of central importance, as they directly influence the reliability of the resulting inference. Researchers must thoughtfully select an appropriate loss form tailored to the specific context and cautiously determine any associated calibration parameters. Second, computational challenges can arise, particularly because the generalized posterior often lacks a closed-form expression. In such cases, inference typically relies on numerical methods such as Markov chain Monte Carlo (MCMC), and the computational burden may become substantial in high-dimensional parameter spaces.

3. Methodology

In this section, we first present a general framework for probabilistic regression-type models with an intractable normalizing constant and then present Discrete Slope Fisher Divergence. Finally, we will propose Discrete Slope Fisher Divergence. All random variables in this paper are defined on the foundational probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For instance, the expectation of a random variable $\boldsymbol{y}$ is expressed as $\mathbb{E}(\boldsymbol{y}) = \int_{\Omega} \boldsymbol{y} \mathrm{d}\mathbb{P}$.

3.1. Regression-Type Models with an Intractable Normalizing Constant on a Discrete Domain

We present a specialized framework for probabilistic regression-type models with

an intractable normalizing constant defined exclusively on discrete domains. Suppose we have independent observations $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n \in \mathcal{D}$ from unknown distributions $q(\boldsymbol{y}\mid \boldsymbol{x}_1),\cdots,q(\boldsymbol{y}\mid \boldsymbol{x}_n)$, where $\mathcal{D}$ is a discrete domain (e.g., $\mathcal{D}=\{0,1\}^d$ for binary data or $\mathcal{D}=\{0,1,2,\cdots\}^d$ for count data), with associated covariates $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$. We consider parametric models $p(\cdot\mid \boldsymbol{x}_i,\boldsymbol{\theta})$ for $i=1,\cdots,n$, where $\boldsymbol{\theta}\in\mathbb{R}^p$ is a vector of unknown parameters. Under the fixed design setting, the observations $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n$ are independent but not necessarily identically distributed.

Our discrete-domain framework for regression-type models with an intractable normalizing constant is given by:

$$p(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta})=\frac{1}{Z_i(\boldsymbol{\theta})}p_0(\boldsymbol{y})\tilde{p}_1(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta}), \tag{4}$$

where:

- $p_0(\boldsymbol{y})$ is a known function of $\boldsymbol{y}$ that encodes domain-specific information (e.g., boundary constraints or structural zeros);
- $\tilde{p}_1(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta})$ models the relationship between the discrete response and covariates;
- $Z_i(\boldsymbol{\theta})$ is the normalizing constant (partition function) defined as:

$$Z_i(\boldsymbol{\theta})=\sum_{\boldsymbol{y}\in\mathcal{D}}p_0(\boldsymbol{y})\tilde{p}_1(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta}). \tag{5}$$

If the model (4) is correctly specified, the true data-generating distribution satisfies $q(\boldsymbol{y}\mid \boldsymbol{x}_i)\propto p_0(\boldsymbol{y})\tilde{p}_1(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta}_0)$ for $i=1,\cdots,n$, where $\boldsymbol{\theta}_0$ represents the true parameter vector.

This framework provides the foundation for developing and analyzing regression models for discrete data with intractable normalizing constants, where the computational challenges of normalization are inherent to the discrete nature of the domain.

Example (Conway-Maxwell-Poisson regression model [12]). Let $y\in\mathbb{N}_0$, where $\mathbb{N}_0$ is the set of non-negative integers, with $p_0(y)=1$ and

$$\tilde{p}_1(y\mid \boldsymbol{x}_i,\boldsymbol{\theta})=\frac{\lambda_i^y}{(y!)^\nu},$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta}^\top,\nu)^\top$, $\nu\geq 0$ represents the dispersion parameter, and $\lambda_i=\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$ generalizes the Poisson mean parameter. Then model (4) yields the CMP regression with normalizing constant:

$$Z_i(\boldsymbol{\theta})=\sum_{s=0}^{\infty}\frac{\lambda_i^s}{(s!)^\nu}.$$

The CMP distribution encompasses three important special cases:

- **Geometric**: When $\nu=0$ (with $\lambda_i<1$)
- **Poisson**: When $\nu=1$
- **Bernoulli**: In the limit $\nu\to\infty$

For $\nu\in[0,1)$, the CMP describes data that are overdispersed relative to a Poisson distribution with the same mean; For $\nu>1$, the CMP model is appropriate

for underdispersed data.

3.2. A Discrete Slope Fisher Divergence

Fisher divergence, originally proposed for score matching [13], measures the discrepancy between two probability distributions by comparing their score functions (gradients of log-densities). However, its reliance on gradient operators limits its applications to continuous and differentiable models. To overcome this limitation, Lyus [14] introduced the generalized Fisher divergence, which extends the original framework by replacing gradients with general linear operators, enabling broader applicability—including discrete data and non-gradient-based models.

Based on the Fisher divergence for data-generating distribution $q(\boldsymbol{y}\mid \boldsymbol{x}_i)$ and parametric model distribution $p(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta})$ on $\mathbb{R}^d$, the divergence $D_{\mathcal{L}}(q_*\,\|\,p_*)$ is defined as:

$$D_{\mathcal{L}}(q_*\,\|\,p_*)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla\log q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)-\nabla\log p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})\right\|^2\right], \tag{6}$$

where ∇ is the gradient operator with respect to $\boldsymbol{y}$, $\mathbb{E}$ represents the expectation under the true distribution $q(\cdot\mid \boldsymbol{x}_i)$.

Based on the main idea from Lyu (2012), the generalized Fisher divergence $D_{\mathcal{L}}(q_*\,\|\,p_*)$ for $q_*=\prod_{i=1}^{n}q(\boldsymbol{y}\mid \boldsymbol{x}_i)$ and $p_*=\prod_{i=1}^{n}p(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta})$ is defined as:

$$D_{\mathcal{L}}(q_*\,\|\,p_*)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left\|\frac{\mathcal{L}q(\boldsymbol{y}\mid \boldsymbol{x}_i)}{q(\boldsymbol{y}\mid \boldsymbol{x}_i)}-\frac{\mathcal{L}p(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta})}{p(\boldsymbol{y}\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right\|^2\right) \tag{7}$$

where $\|\cdot\|$ denotes the Euclidean norm.

To handle discrete data, Lyu investigated a special case of the linear operator $\mathcal{L}$ in Equation (7), called the marginalization operator $\mathcal{M}$. This operator is defined as:

$$\mathcal{M}q(\boldsymbol{y}\mid \boldsymbol{x}_i):=\left(\mathcal{M}_1q(\boldsymbol{y}\mid \boldsymbol{x}_i),\cdots,\mathcal{M}_dq(\boldsymbol{y}\mid \boldsymbol{x}_i)\right)^{\top}$$

where each component operator is given by:

$$\mathcal{M}_jq(\boldsymbol{y}\mid \boldsymbol{x}_i):=\sum_{y_j}q(\boldsymbol{y}\mid \boldsymbol{x}_i)\ \text{ for } i=1,\cdots,n \text{ and } j=1,\cdots,d.$$

However, in the one-dimensional discrete case, this specific linear operator (namely, the marginalization operator $\mathcal{M}$) fails to eliminate the influence of the normalization constant. To address this limitation, [15] proposed a new linear operator called the **forward difference operator**.

The traditional Fisher divergence compares the slopes of log-density functions. However, since discrete data density functions lack continuous slopes at certain points, directly comparing the log-gradients of density functions may not be appropriate for discrete data. Therefore, [15] considered a new linear operator $\mathcal{L}$ defined as:

$$\mathcal{L}p(y\mid \boldsymbol{x}_i,\boldsymbol{\theta}):=p(y^{+}\mid \boldsymbol{x}_i,\boldsymbol{\theta})-p(y\mid \boldsymbol{x}_i,\boldsymbol{\theta}),\ i=1,\cdots,n,$$

where $y^{+}=y+1$. This linear operator represents the "discrete slope" of

$p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ at point y, which is the discrete counterpart of the log-density gradient in continuous distributions. When y takes values from a finite set, special treatment is required at boundary points. Specifically, when y is at the upper boundary, define $p(y^+ \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = 0$. Similarly, define $p(y^- \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = 0$ for $y^- = y - 1$ at the lower boundary.

Notably, this linear operator provides the "slope" of $p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ at point y in discrete settings. We can derive (ignoring constant terms):

$$\frac{\mathcal{L}p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})} = \frac{p(y+1 \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})} - 1$$

This shows that the fundamental principle of this method is to make the ratio $\frac{p(y^+ \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})}$ as close as possible to the corresponding ratio from the data distribution, *i.e.*, $\frac{q(y^+ \mid \boldsymbol{x}_i)}{q(y \mid \boldsymbol{x}_i)}$.

To avoid division-by-zero problems when computing slopes, introducing the slope transformation proposed by [16]:

$$t(u) = \frac{1}{1+u}$$

Under this transformation, if the probability density approaches zero causing the ratio to diverge to infinity, the transformed value becomes $t(\infty) = 0$.

Therefore, a new generalized Fisher divergence between q_* and p_* is defined as:

$$\begin{aligned} D_{\mathrm{DSF}}(q_* \,\|\, p_*) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Bigg\{ & \left[t\left(\frac{p(y_i^+ \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) - t\left(\frac{q(y_i^+ \mid \boldsymbol{x}_i)}{q(y_i \mid \boldsymbol{x}_i)} \right) \right]^2 \\ & + \left[t\left(\frac{p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p(y_i^- \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) - t\left(\frac{q(y_i \mid \boldsymbol{x}_i)}{q(y_i^- \mid \boldsymbol{x}_i)} \right) \right]^2 \Bigg\}, \end{aligned} \tag{8}$$

where, q_i and p_i denote $q(y \mid \boldsymbol{x}_i)$ and $q(y \mid \boldsymbol{x}_i, \boldsymbol{\theta})$, respectively. In this paper, we refer to this divergence as the Discrete Slope Fisher Divergence.

To extend to the multivariate case, the linear operator $\mathcal{L}$ is defined as:

$$\mathcal{L}p(\boldsymbol{y} \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \begin{pmatrix} \vdots \\ \mathcal{L}_j p(\boldsymbol{y} \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ p(\boldsymbol{y}^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}) - p(\boldsymbol{y} \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \\ \vdots \end{pmatrix},$$

where for $j = 1, \cdots, d$, we define:

$$\boldsymbol{y}^{(j+)} = (y_1, \cdots, y_j + 1, \cdots, y_d)^\top.$$

If $\boldsymbol{y}$ has bounded support, when $\boldsymbol{y}$ is at the boundary, we set:

$$p(\boldsymbol{y}^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = 0.$$

Similarly, we define:

$$\boldsymbol{y}^{(j-)} = \left(y_1, \cdots, y_j - 1, \cdots, y_d\right)^{\top},$$

and at the boundary of the response domain, we set:

$$p\left(\boldsymbol{y}^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right) = 0.$$

Analogous to the univariate case, the Discrete Slope Fisher Divergence between q_* and p_* is defined as:

$$D_{\mathrm{DSF}}\left(q_* \,\|\, p_*\right) = \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{DSF}}\left(q_i \,\|\, p_i\right) \tag{9}$$

where

$$\begin{aligned} D_{\mathrm{DSF}}\left(q_i \,\|\, p_i\right) = \mathbb{E}\Bigg[\sum_{j=1}^{d}\Bigg\{\Bigg[t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right) - t\left(\frac{q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right)}{q\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i\right)}\right)\Bigg]^2 \\ + \Bigg[t\left(\frac{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right) - t\left(\frac{q\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i\right)}{q\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i\right)}\right)\Bigg]^2\Bigg\}\Bigg] \end{aligned} \tag{10}$$

Theorem 1 Equation (10) can be decomposed as:

$$\begin{aligned} D_{\mathrm{DSF}}\left(q_i \,\|\, p_i\right) = \mathbb{E}\Bigg[\sum_{j=1}^{d}\Bigg\{t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right)^2 + t\left(\frac{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right)^2 \\ - 2t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right)\Bigg\}\Bigg] + C \end{aligned} \tag{11}$$

where C is a constant independent of the parameters $\boldsymbol{\theta}$.

This theorem shows that (10) is tractable. Its proof is given in the Appendix. The discrete Slope Fisher divergence between a model p_θ and an empirical distribution $q_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$, is computed as

$$\begin{aligned} D_{\mathrm{DSF}}\left(q_n \,\|\, p_\theta\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\Bigg\{t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right)^2 + t\left(\frac{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right)^2 \\ - 2t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right)\Bigg\} + C\left(q_n\right), \end{aligned} \tag{12}$$

where $C\left(q_n\right)$ is a constant depending on the empirical density p_n but not on $\boldsymbol{\theta}$.

3.3. A Generalised Posterior

We are now in a position to present DSFD-Bayes.

Given a prior distribution π over the parameter space Θ and a statistical model p_θ as defined in Equation (4) on a discrete domain—a regression-type model with an intractable normalization constant parameterized by $\theta \in \Theta$ —

where the observed response variables are independent but not necessarily identically distributed, the **DSFD-Bayes posterior distribution** is defined as:

$$\pi_n^D(\theta) \propto \pi(\theta)\exp\left(-wnD_{\mathrm{DSF}}(q_n \,\|\, p_\theta)\right), \tag{13}$$

where $w>0$ is a tuning constant to be specified (see Section 3.3 in [8] for guidance on selecting an appropriate weighting factor w). The θ-independent constant term $C(q_n)$ in $D_{\mathrm{DSF}}(q_n \,\|\, p_\theta)$ cancels out during the normalization of the DSFD-Bayes posterior. Hence, $C(q_n)$ can be safely ignored in practical computations.

The DSFD-Bayes posterior is directly amenable to standard posterior sampling techniques, such as Markov chain Monte Carlo (MCMC) and the No-U-Turn Sampler (NUTS), as the actual computation does not involve any intractable normalizing constants. The resulting DSFD-Bayes method can be computed at cost $O(nd)$ linear in the size of the dataset.

4. Experiments

This paper validates the proposed method through two simulation studies—one overdispersed and one underdispersed—and presents a data analysis case involving underdispersion. All analyses were conducted on a HUAWEI computer equipped with an Intel Core i9-12900H processor (2.5 GHz) and 16 GB of RAM, running Windows 11 and Python 3.10.16.

We investigate a relatively simple model—the Conway—Maxwell—Poisson (CMP) regression model, and the CMP regression model was chosen as our experimental framework based on the following considerations: With its single dispersion parameter ν, the CMP model generalizes both Poisson and Bernoulli regression characteristics. This "parsimonious yet expressive" property makes it an ideal testbed that captures the essential complexity of real-world statistical models while preventing over-parameterization from confounding computational efficiency evaluations.

4.1. Simulation Studies

We simulated two datasets from the Conway-Maxwell-Poisson (CMP) regression model: 1) **underdispersion** with dispersion parameter $\nu=1.5$; 2) **overdispersion** with $\nu=0.8$. In both simulation settings, the true regression coefficients were set to $\boldsymbol{\beta}=(0.5,-1.2,0.8)$, where 0.5 is the intercept term.

This study employs the Conway-Maxwell-Poisson (CMP) regression model to generate simulated data, evaluating model performance through carefully controlled generation processes for both covariates and response variables. The covariate generation is meticulously designed to simulate real-world data scenarios, where the explanatory variable X_1 is sampled from the standard normal distribution $\mathcal{N}(0,1)$ with 1000 samples, and X_2 is drawn from the normal distribution $\mathcal{N}(1,0.5)$ with the same sample size. To ensure experimental reproducibility, we set the random seed to 1 and include an intercept column vector $\mathbf{1}$,

ultimately constructing the design matrix $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2]$. This design maintains the stochastic characteristics of continuous variables while guaranteeing result reproducibility through fixed random seeds.

The response variable generation achieves non-i.i.d. properties through a parameter heterogeneity mechanism. Each response variable Y_i corresponds to a unique rate parameter $\lambda_i = \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})$, where the regression coefficients are set as $\boldsymbol{\beta} = (0.3, 0.5, -0.2)^\top$. Since the CMP distribution lacks an analytical sampling method, we employ Metropolis-Hastings algorithm for MCMC sampling: initial values $y^{(0)}$ are sampled from Poisson (2) distribution, followed by 2000 burn-in iterations. During this process, we set shared dispersion parameters to make all response variables exhibit either under-dispersion or over-dispersion characteristics.

We compared two methods: the standard Bayesian inference and our proposed DSFD-Bayes method, and the method for calibrating the weight w in DSFD-Bayes is described in Section 3.3.

For the underdispersed case, the prior distribution for the regression coefficients was specified as a multivariate normal distribution $\mathcal{N}(\mathbf{0}, I)$, and the prior for the dispersion parameter ν was set as a log-normal distribution with log-mean 0 and log-standard deviation 0.5. For the overdispersed case, the prior for $\boldsymbol{\beta}$ was also specified as $\mathcal{N}(\mathbf{0}, I)$, while the prior for ν was chosen as a log-normal distribution with log-mean −0.5 and log-standard deviation 0.5.

We used the No-U-Turn Sampler (NUTS) to sample from all posterior distributions. The specific settings were as follows: 5000 post-warmup samples were retained after discarding 5000 warm-up iterations, and 10 independent chains were run. In all cases, the $\hat{R}$ statistics for the four parameters were below 1.01, indicating good convergence.

Table 1 and Table 2 present comparative evaluations of DSFD-Bayes versus standard Bayes under under-dispersed ($\nu = 1.5$) and over-dispersed ($\nu = 0.8$) conditions, respectively. The evaluation results suggest that the DSFD-Bayes generally outperforms the standard Bayes. DSFD-Bayes consistently yields smaller absolute biases and exhibits a reasonable decreasing trend in standard deviations as the sample size increases. Its 95% credible intervals tend to cover the true parameter values in most cases, with more appropriate interval widths. In contrast, the credible intervals produced by the standard Bayesian method are often too narrow and deviate from the true values.

The evaluation results suggest that the DSFD-Bayes method generally outperforms the standard Bayesian approach. DSFD-Bayes consistently yields smaller absolute biases and exhibits a reasonable decreasing trend in standard deviations as the sample size increases. Its 95% credible intervals tend to cover the true parameter values in most cases, with more appropriate interval widths. In contrast, the credible intervals produced by the standard Bayesian method are often too narrow and deviate from the true values.

In terms of applicability, DSFD-Bayes demonstrates greater robustness, per-

forming well across a wide range of sample sizes. The standard Bayesian method, on the other hand, tends to fail in small-sample settings due to the overwhelming influence of the prior, and is only potentially reliable in large-sample cases. These findings indicate that DSFD-Bayes offers a better balance between prior information and data, making it a more reliable choice for Bayesian inference.

Table 1. Comparison of performance metrics for parameter estimation using DSFD-Bayes and Bayes methods for different sample sizes ($\nu = 1.5$).

	DSFD-Bayes			Bayes		
Parameter	Bias	SD	95% CI	Bias	SD	95% CI
			$n = 10$			
β_0	−0.1656	0.2016	(−0.0619, 0.7429)	14.8625	0.7317	(14.0575, 16.5038)
β_1	0.8602	0.0887	(−0.5151, −0.1643)	0.5365	0.0002	(−0.6639, −0.6633)
β_2	−0.2978	0.1911	(0.1290, 0.8749)	3.6892	0.0001	(4.4889, 4.4893)
ν	−0.7817	0.1216	(0.4900, 0.9620)	6.7955	0.4314	(7.5355, 8.9768)
			$n = 50$			
β_0	−0.8410	0.0267	(−0.3934, −0.2891)	1.5275	0.0024	(2.0236, 2.0317)
β_1	0.8565	0.0123	(−0.3677, −0.3196)	0.8541	0.0044	(−0.3540, −0.3394)
β_2	−0.5653	0.0210	(0.1935, 0.2765)	−2.4191	0.0266	(−1.6665, −1.5775)
ν	−1.3282	0.0119	(0.1487, 0.1953)	−1.1706	0.0016	(0.3269, 0.3323)
			$n = 100$			
β_0	−2.8805	0.7161	(−3.7940, −0.9872)	1.3351	0.0004	(1.8343, 1.8358)
β_1	1.1286	0.8506	(−1.9555, 1.4455)	0.8484	0.0005	(−0.3524, −0.3508)
β_2	−2.2036	0.7836	(−2.9597, 0.1096)	−2.6799	0.0027	(−1.8849, −1.8754)
ν	0.2217	0.9582	(0.5191, 4.2530)	−0.9369	0.0003	(0.5626, 0.5637)
			$n = 150$			
β_0	−0.2400	0.2973	(−0.3199, 0.8321)	1.3845	0.0003	(1.8840, 1.8850)
β_1	0.8605	0.1507	(−0.6516, −0.0523)	0.8918	0.0007	(−0.3093, −0.3072)
β_2	−0.5587	0.2507	(−0.2422, 0.7509)	−2.6532	0.0008	(−1.8547, −1.8519)
ν	−0.9300	0.1550	(0.3002, 0.9041)	−1.0164	0.0002	(0.4833, 0.4840)
			$n = 200$			
β_0	−0.2508	0.5806	(−0.8783, 1.3975)	1.4505	0.0007	(1.9495, 1.9515)
β_1	0.7173	0.4130	(−1.3580, 0.2573)	0.8979	0.0003	(−0.3028, −0.3017)
β_2	−0.4258	0.5417	(−0.6746, 1.4571)	−2.5694	0.0020	(−1.7727, −1.7660)
ν	−0.7506	0.2855	(0.3182, 1.4075)	−0.9134	0.0003	(0.5862, 0.5871)

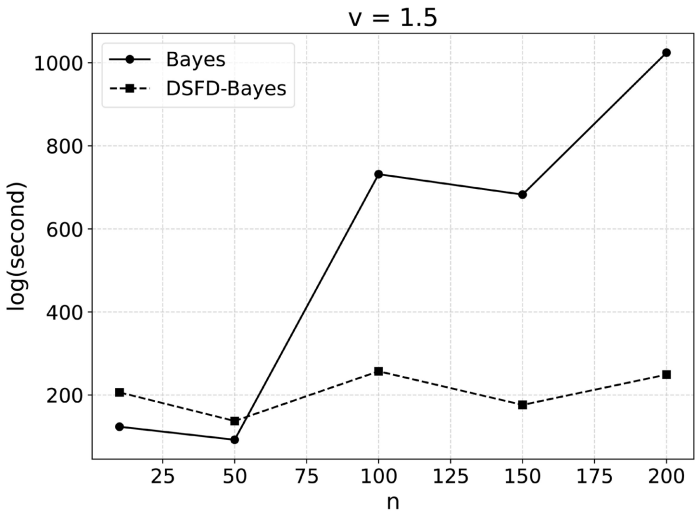
Note: SD = Standard Deviation of the Estimate; CI = Credible Interval.

Table 2. Comparison of performance metrics for parameter estimation using DSFD-Bayes and Bayes methods for different sample sizes ($\nu = 0.8$).

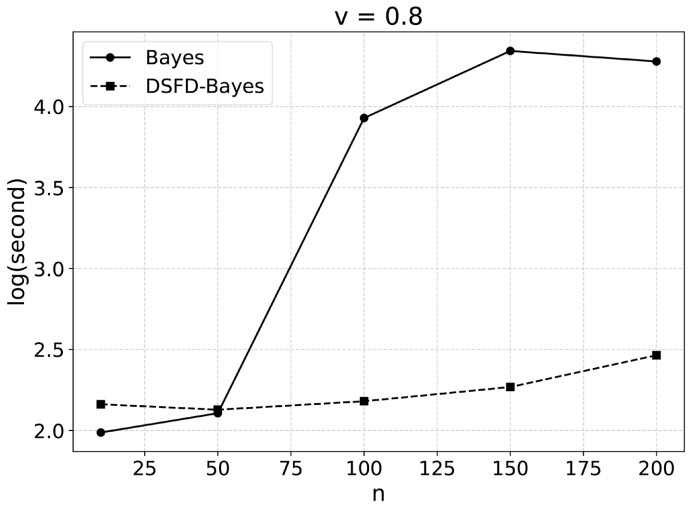
	DSFD-Bayes			Bayes		
Parameter	Bias	SD	95% CI	Bias	SD	95% CI
			$n = 10$			
β_0	−0.0396	0.4317	(−0.3697, 1.3089)	1.4959	0.0463	(1.9087, 2.0707)
β_1	−0.0690	0.2873	(−1.8351, −0.7181)	1.3508	0.0236	(0.1087, 0.1900)
β_2	−0.0300	0.4924	(−0.1492, 1.7416)	−1.6168	0.1284	(−1.0286, −0.5840)
ν	−0.1300	0.1154	(0.4758, 0.9304)	−0.5851	0.0017	(0.2120, 0.2169)
			$n = 50$			
β_0	0.0460	0.0218	(0.5027, 0.5884)	1.2085	0.0001	(1.7083, 1.7087)
β_1	−0.0869	0.0093	(−1.3055, −1.2687)	0.8769	0.0002	(−0.3234, −0.3228)
β_2	−0.0411	0.0178	(0.7242, 0.7937)	−2.8052	0.0007	(−2.0063, −2.0041)
ν	0.1360	0.0025	(0.9314, 0.9412)	−0.2629	0.0000	(0.2371, 0.2372)
			$n = 100$			
β_0	0.0412	0.0399	(0.4631, 0.6202)	1.1342	0.0071	(1.6201, 1.6483)
β_1	−0.0934	0.0195	(−1.3320, −1.2551)	0.8874	0.0021	(−0.3167, −0.3084)
β_2	−0.0139	0.0315	(0.7254, 0.8470)	−2.8845	0.0028	(−2.0897, −2.0803)
ν	0.0055	0.0088	(0.7896, 0.8244)	−0.2736	0.0016	(0.5232, 0.5295)
			$n = 150$			
β_0	−0.2278	0.0028	(0.2669, 0.2779)	1.1623	0.0035	(1.6560, 1.6684)
β_1	0.0009	0.0012	(−1.2014, −1.1968)	0.8980	0.0016	(−0.3052, −0.2989)
β_2	0.0861	0.0022	(0.8817, 0.8904)	−2.8611	0.0017	(−2.0641, −2.0581)
ν	−0.1277	0.0008	(0.6707, 0.6739)	−0.3707	0.0012	(0.4268, 0.4319)
			$n = 200$			
β_0	−0.0375	0.0246	(0.4144, 0.5090)	1.1607	0.0029	(1.6589, 1.6660)
β_1	−0.0353	0.0110	(−1.2565, −1.2134)	0.8976	0.0013	(−0.3043, −0.2994)
β_2	−0.0179	0.0216	(0.7406, 0.8249)	−2.8601	0.0013	(−2.0613, −2.0579)
ν	0.0048	0.0047	(0.7962, 0.8146)	−0.3705	0.0010	(0.4275, 0.4311)

Note: SD = Standard Deviation of the Estimate; CI = Credible Interval.

As shown in Figure 1, although DSFD-Bayes incurs slightly higher computational cost than Standard Bayes in small-sample scenarios, its computational time becomes significantly lower as the sample size increases, with a marked difference observed between the two methods. While DSFD-Bayes demonstrates superior computational efficiency in larger datasets, its relatively higher costs in small-sample settings may constrain its practical adoption in resource-constrained environments.



(a) Underdispersed ($v = 1.5$)



(b) Overdispersed ($v = 0.8$)

Figure 1. Comparison of Computational Time between DSFD-Bayes and Standard Bayes.

4.2. Case Studies

We use the airfreight breakage dataset from [17], which contains 10 batches of air cargo transportation records. The explanatory variable (X_i) is the number of transfers each batch experienced during transportation (a discrete count), and the response variable (Y_i) is the number of damaged ampules out of 1000 in each batch (also a discrete count). This dataset exhibits underdispersion.

We model the real dataset using a Conway-Maxwell-Poisson regression model. The regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)$ are assigned a multivariate normal prior $N(\mathbf{0}, 4\boldsymbol{I})$, and the dispersion parameter v follows a $\text{Gamma}(7,1)$ prior.

Figure 2 shows the estimated dispersion parameter $v > 1$, consistent with the under-dispersion property of the data. A comparison of the forest plots reveals that DSFD-Bayes yields more concentrated parameter estimates with narrower intervals, indicating greater stability and reliability.

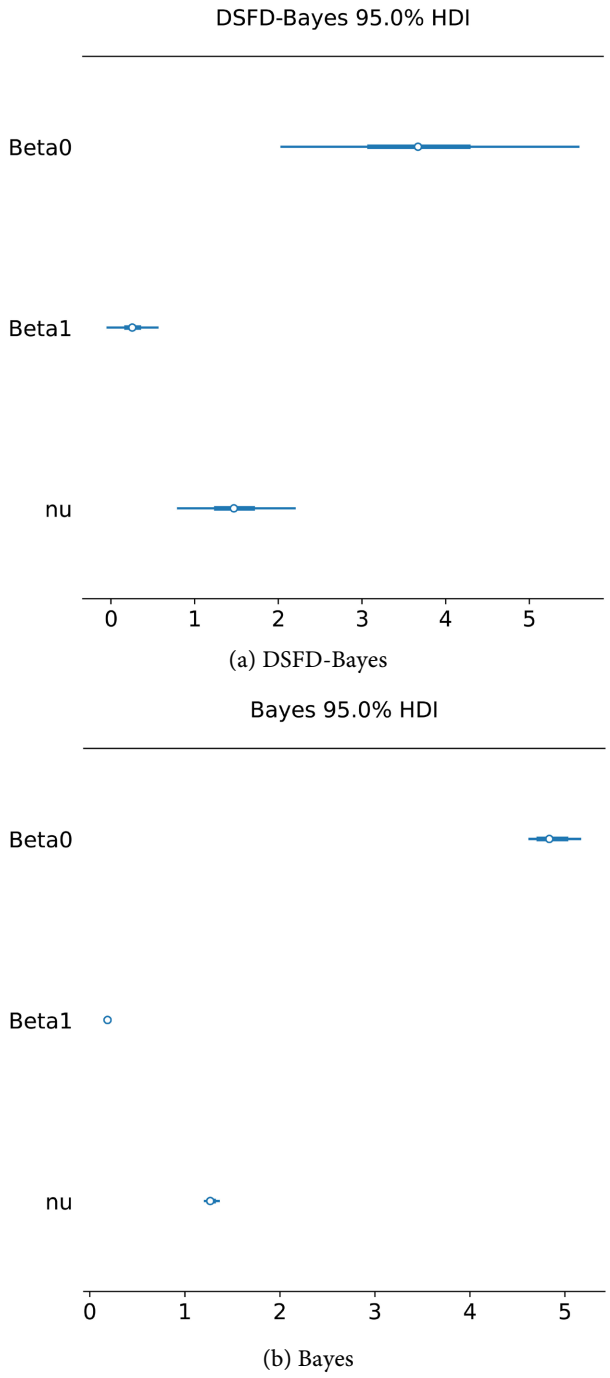


Figure 2. Comparison of posterior parameter estimates between DSFD-Bayes and Standard Bayes (95% Credible Intervals).

To compare the predictive performance of the standard Bayes and the DSFD-Bayes, we conduct a posterior predictive check. **Figure 3** and **Figure 4** present the posterior predictive distributions of the two methods for the number of defects. As shown in **Figure 4** (standard Bayes), while the posterior mean generally follows the trend of the observed data, there is considerable variability, especially in regions with higher observed values. The widened credible intervals in those areas

suggest that the standard Bayesian model struggles to capture heteroskedasticity and local structure, resulting in low confidence in predicting extreme values. In contrast, Figure 3 demonstrates that DSFD-Bayes significantly improves predictive performance. Its posterior mean closely aligns with the observed data, and the prediction intervals are narrower and more adaptive across different regions. Particularly in the low-to-moderate response range, the credible intervals remain compact while still covering the true observations, indicating better calibration.

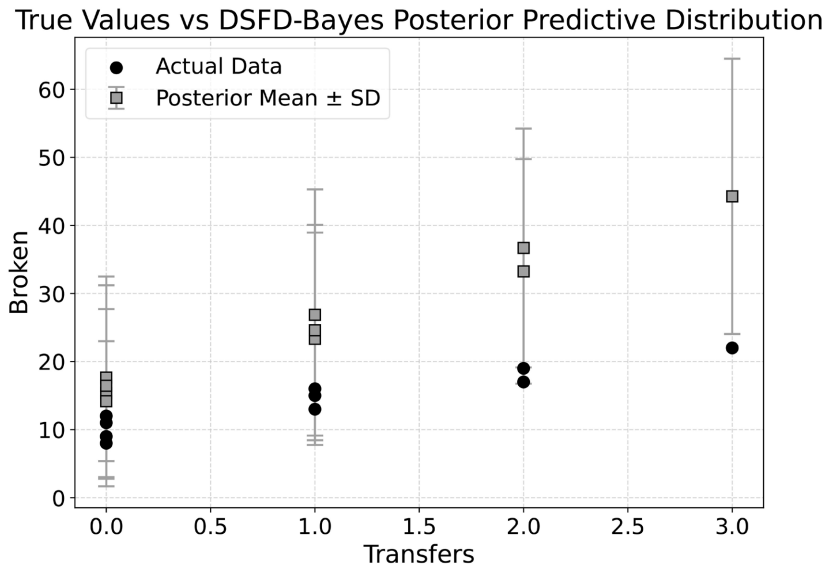


Figure 3. Comparison between observed data and dsfd-bayes posterior predictive estimates.

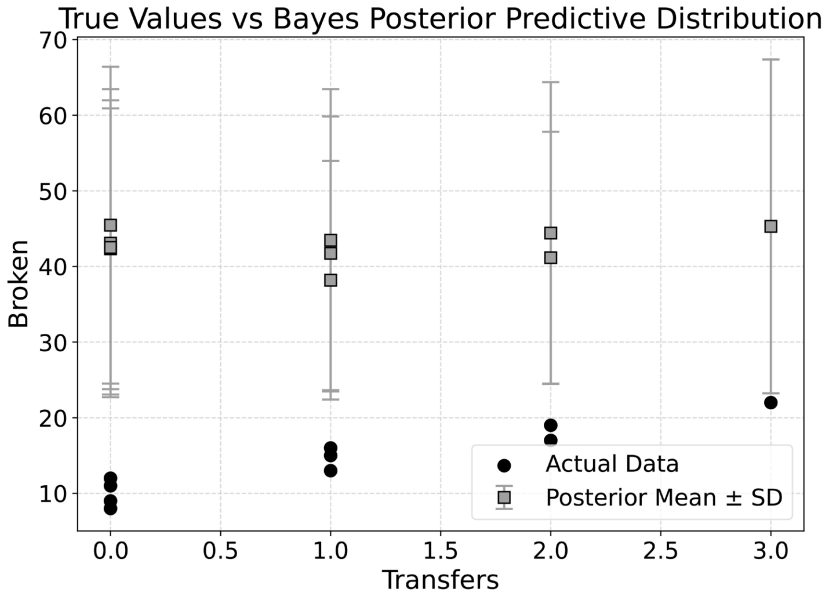


Figure 4. Comparison between observed data and bayes posterior predictive estimates.

5. Conclusions and Suggestions

This paper proposes a novel generalized Bayesian inference method called DSFD-

Bayes for discrete regression models with intractable normalizing constants. The method overcomes the limitations of existing KSD-Bayes and DFD-Bayes approaches, and is applicable to independent but not necessarily identically distributed (INID) discrete data, filling a gap in generalized Bayesian inference for regression modeling. Both simulation studies and empirical analysis demonstrate that the DSFD-Bayes performs better in model adaptability and robustness. However, a limitation of this study is that the proposed method is currently restricted to regression-type models, and the lack of exploration into more diverse model classes—such as time series or hierarchical models—restricts the general applicability of the proposed method. This limitation suggests a promising direction for future research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Jin, I.H. and Liang, F. (2014) Use of SAMC for Bayesian Analysis of Statistical Models with Intractable Normalizing Constants. *Computational Statistics & Data Analysis*, **71**, 402-416. https://doi.org/10.1016/j.csda.2012.07.005

[2] Park, J. and Haran, M. (2018) Bayesian Inference in the Presence of Intractable Normalizing Functions. *Journal of the American Statistical Association*, **113**, 1372-1390. https://doi.org/10.1080/01621459.2018.1448824

[3] Murray, I., Ghahramani, Z. and Mackay, D. (2006) MCMC for Doubly-Intractable Distributions. *Proceedings of the* 22*nd Conference on Uncertainty in Artificial Intelligence*, Cambridge, 13-16 July 2006, 359-366.

[4] Møller, J., Pettitt, A.N., Reeves, R. and Berthelsen, K.K. (2006) An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants. *Biometrika*, **93**, 451-458. https://doi.org/10.1093/biomet/93.2.451

[5] Marin, J., Pudlo, P., Robert, C.P. and Ryder, R.J. (2011) Approximate Bayesian Computational Methods. *Statistics and Computing*, **22**, 1167-1180. https://doi.org/10.1007/s11222-011-9288-2

[6] Gutmann, M. and Hyvärinen, A. (2010) Noise-Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models. *Proceedings of the* 13*th International Conference on Artificial Intelligence and Statistics*, Sardinia, 13-15 May 2010, 297-304.

[7] Matsubara, T., Knoblauch, J., Briol, F. and Oates, C.J. (2022) Robust Generalised Bayesian Inference for Intractable Likelihoods. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **84**, 997-1022. https://doi.org/10.1111/rssb.12500

[8] Matsubara, T., Knoblauch, J., Briol, F. and Oates, C.J. (2023) Generalized Bayesian Inference for Discrete Intractable Likelihood. *Journal of the American Statistical Association*, **119**, 2345-2355. https://doi.org/10.1080/01621459.2023.2257891

[9] Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1999) Bayesian Model Choice: What and Why? In: Bernardo, J.M., *et al.*, Eds., *Bayesian Statistics*, Vol. 6, Oxford University Press, 343-370. https://doi.org/10.1093/oso/9780198504856.003.0015

[10] Bissiri, P.G., Holmes, C.C. and Walker, S.G. (2016) A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society Series B*: *Statistical*

Methodology, **78**, 1103-1130. https://doi.org/10.1111/rssb.12158

[11] Jewson, J., Smith, J.Q. and Holmes, C. (2018) Principles of Bayesian Inference Using General Divergence Criteria. *Entropy*, **20**, Article No. 442. https://doi.org/10.3390/e20060442

[12] Sellers, K.F. and Shmueli, G. (2010) A Flexible Regression Model for Count Data. *The Annals of Applied Statistics*, **4**, 943-961.

[13] Hyvrinen, A. and Dayan, P. (2005) Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, **6**, 695-709.

[14] Lyu, S. (2012) Interpretation and Generalization of Score Matching. *Proceedings of the* 25*th Conference on Uncertainty in Artificial Intelligence*, Montreal, 18-21 June 2009, 359-366.

[15] Xu, J., Scealy, J.L., Wood, A.T. and Zou, T. (2022) Generalized Score Matching for Regression.

[16] Hyvärinen, A. (2007) Some Extensions of Score Matching. *Computational Statistics & Data Analysis*, **51**, 2499-2512. https://doi.org/10.1016/j.csda.2006.09.003

[17] Kutner, M.H., Nachtsheim, C.J. and Neter, J. (1984) Applied Linear Regression Models. *Technometrics*, **26**, 415-416. https://doi.org/10.1080/00401706.1984.10487998

Appendix

Proof of Theorem 1. Recall that the transformation function $t\left(f_1(y)/f_2(y)\right)$ returns zero when $f_2(y)=0$, and returns one when $f_1(y)=0$. In other words, the transformation $t\left(f_1(y)/f_2(y)\right)$ yields a constant value whenever either $f_1(y)$ or $f_2(y)$ is zero. Therefore, for simplicity, we assume that both $q(y\mid \boldsymbol{x}_i)$ and $p(y\mid \boldsymbol{x}_i,\boldsymbol{\theta})$ are non-zero for all $y\in\mathcal{D}$. Note that we have

$$
\begin{aligned}
D_{\mathrm{DSF}}(q_i,p_i) &= \sum_{y\in D} q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)\sum_{j=1}^{d}\left\{\left[t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right)-t\left(\frac{q\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i\right)}{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}\right)\right]^2\right.\\
&\quad\left.+\left[t\left(\frac{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}\right)-t\left(\frac{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}{q(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i)}\right)\right]^2\right\}\\
&= \sum_{y\in D} q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)\sum_{j=1}^{d}\left\{t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right)^2+t\left(\frac{q\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i\right)}{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}\right)^2\right.\\
&\quad-2t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right)t\left(\frac{q\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i\right)}{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}\right)\\
&\quad+t\left(\frac{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}\right)^2+t\left(\frac{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}{q\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i\right)}\right)^2\\
&\quad\left.-2t\left(\frac{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}\right)t\left(\frac{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}{q\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i\right)}\right)\right\}\\
&= \sum_{y\in D} q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)\sum_{j=1}^{d}\left[t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right)^2-2t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right)+C\right.\\
&\quad\left.+t\left(\frac{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}\right)^2+C\right]
\end{aligned}
$$

After expanding the squares, reorganizing the summation terms, and separating out the constant term C, we obtain:

$$
\begin{aligned}
D_{\mathrm{DSF}}(q_i,p_i) &= \sum_{y\in D} q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)\sum_{j=1}^{d}\left\{t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right)^2+t\left(\frac{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}\right)^2\right\}\\
&\quad-2\sum_{y\in D} q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)\sum_{j=1}^{d}\left\{t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}\right)t\left(\frac{q\left(\boldsymbol{y}_i^{(j+)}\mid \boldsymbol{x}_i\right)}{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}\right)\right.\\
&\quad\left.+t\left(\frac{p(\boldsymbol{y}_i\mid \boldsymbol{x}_i,\boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)}\right)t\left(\frac{q(\boldsymbol{y}_i\mid \boldsymbol{x}_i)}{q\left(\boldsymbol{y}_i^{(j-)}\mid \boldsymbol{x}_i\right)}\right)\right\}+C
\end{aligned}
$$

where C does not depend on $\boldsymbol{\theta}$. Firstly examining the second term:

$$\sum_{\boldsymbol{y}\in D} q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \sum_{j=1}^{d} \left\{ t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) t\left(\frac{q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right)}{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)} \right) + t\left(\frac{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)} \right) t\left(\frac{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}{q\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i\right)} \right) \right\}$$

According to the definition of the transformation function $t(u) = \dfrac{1}{1+u}$, we have:

$$t\left(\frac{q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right)}{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)} \right) = \frac{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) + q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right)}$$

$$t\left(\frac{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}{q\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i\right)} \right) = \frac{q\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i\right)}{q\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i\right) + q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}$$

Substituting into the second term's summation:

$$\sum_{\boldsymbol{y}\in D} q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \sum_{j=1}^{d} \left\{ t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) \frac{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) + q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right)} + t\left(\frac{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)} \right) \frac{q\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i\right)}{q\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i\right) + q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)} \right\}$$

This part can be split into two independent summation terms:

$$\begin{aligned} &\sum_{\boldsymbol{y}\in D} q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \sum_{j=1}^{d} t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) \frac{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) + q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right)} \\ &+ \sum_{\boldsymbol{y}\in D} q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right) \sum_{j=1}^{d} t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) \frac{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}{q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) + q\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i\right)} \\ &= \sum_{\boldsymbol{y}\in D} q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \sum_{j=1}^{d} t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) \end{aligned}$$

Therefore, returning to the original expression:

$$\begin{aligned} D_{\mathrm{DSF}}(q_i, p_i) &= \sum_{\boldsymbol{y}\in D} q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \sum_{j=1}^{d} \left\{ t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right)^2 + t\left(\frac{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)} \right)^2 \right\} \\ &\quad - 2 \sum_{\boldsymbol{y}\in D} q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \sum_{j=1}^{d} t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right) + C \\ &= \sum_{\boldsymbol{y}\in D} q(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \sum_{j=1}^{d} \left\{ t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right)^2 + t\left(\frac{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{p\left(\boldsymbol{y}_i^{(j-)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)} \right)^2 \right\} \end{aligned}$$

$$-2t\left(\frac{p\left(\boldsymbol{y}_i^{(j+)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}\right)\Bigg\}+C$$

This completes the proof.