Q-conjugate Message Passing for Efficient Bayesian Inference

Mykola Lukashchuk M.LUKASHCHUK@TUE.NL
İsmail Şenöz I.SENOZ@TUE.NL
Bert de Vries BERT.DE.VRIES@TUE.NL
Department of Electrical Engineering, Eindhoven University of Technology, the Netherlands

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

Bayesian inference in nonconjugate models such as Bayesian Poisson regression often relies on computationally expensive Monte Carlo methods. This paper introduces Q-conjugacy, a generalization of classical conjugacy that enables efficient closed-form variational inference in certain nonconjugate models. Q-conjugacy is a condition in which a closed-form update scheme expresses the solution minimizing the Kullback-Leibler divergence between a variational distribution and the product of two potentially unnormalized distributions. Leveraging Q-conjugacy within a local message passing framework allows deriving analytic inference update equations for nonconjugate models. The effectiveness of this approach is demonstrated on Bayesian Poisson regression and a model involving a hidden gamma-distributed latent variable with Gaussian-corrupted logarithmic observations. Results show that Q-conjugate triplets, such as (Gamma, LogNormal, Gamma), provide better speed-accuracy trade-offs than Markov Chain Monte Carlo.

Keywords: Bayesian inference, conjugacy, message passing, natural gradient, Poisson regression, variational inference

1. Introduction

Bayesian Poisson regression is a generalized linear model that links the logarithm of the mean of a Poisson distribution to a linear combination of predictors through regression coefficients. This model provides a way to describe the relationship between the expected count of an event and a set of explanatory variables. In the Bayesian framework, the goal is to obtain the posterior distribution of the regression coefficients, given the observed predictors and the count observations. Bayesian Poisson regression has been applied in various fields, including the analysis of kinematic driving events (Kim et al., 2013), crowd counting (Chan and Vasconcelos, 2009), mortality analysis (Stamey et al., 2008), and has been extensively discussed in the tutorial paper by Coxe et al. (2009).

Log-linear linking of the linear predictor and the regression coefficients with the Poisson mean is the most common choice, as the logarithm is the canonical link for the Poisson distribution family (Nelder and Wedderburn, 1972; D'Angelo and Canale, 2023). However, despite being a natural and plausible choice, a Bayesian Poisson regression model is nonconjugate, meaning that the posterior distribution of the regression coefficients given the observed data cannot be expressed in closed-form. As a result, researchers often employ Markov Chain Monte Carlo (MCMC) (FrüHwirth-Schnatter and Wagner, 2006) or Metropolis-Hastings (D'Angelo and Canale, 2023) sampling schemes to carry out inference by sampling from the posterior distribution of the coefficients. These methods may not

be efficient enough for hypothesis testing with large datasets or for real-time applications. This emphasizes the need for more efficient and scalable inference methods that can handle the lack of conjugacy in Bayesian Poisson regression while providing reasonably accurate posterior estimates.

Recent developments by Khan and Rue (2023) in the Bayesian learning rule (BLR) and by Akbayrak et al. (2022) on its local application to factor graph-based inference offer a promising approach to address this problem. The key BLR idea is to apply an iterative gradient update scheme with respect to the variational Free Energy. However, both works rely on approximation schemes or Monte Carlo-based estimators. We show that, with knowledge of the probabilistic model and a specific choice of the variational family, the gradient update can be expressed in closed-form.

We further extend the local BLR inference by introducing the concept of efficient variational inference triplets consisting of a prior, likelihood, and variational family, for which the local BLR yields closed-form updates. We call such triplets "Q-conjugate."

Our main contribution is that the BLR message passing scheme, detailed in Section 3, can be efficiently solved under the Q-conjugancy condition introduced in Section 4. Applying the BLR scheme to these Q-conjugate triplets enables us to derive efficient and closed-form update schemes for nonconjugate models. Specifically, in Section 5, we show that triplets such as (Gamma, LogNormal, Gamma) and (Normal, LogGamma, Normal) offer improved accuracy-power trade-offs compared to Stan MCMC (Carpenter et al., 2017) and the No-U-Turn Sampler (Hoffman and Gelman, 2014).

2. Background

2.1. Factor Graph Review

In this paper, we use a Forney-style factor graph (FFG) representation as the main computational framework for probabilistic models. An FFG represents a factorized model as a graph. Consider a factorized probabilistic model represented by a joint probability distribution (potentially not normalized) $f(x)$ that can be factorized into a product of positive functions (factors), $f_a(x_a)$, each defined over a subset of variables x_a

$$f(x) = \prod_{a \in \mathcal{V}} f_a(x_a), \tag{1}$$

where $\mathcal{V}$ is an indexed set of factors.

We can visualize this model as a Forney-style factor graph (FFG), which is a graphical representation of the factorization. In an FFG, each factor f_a is represented by a node $a \in \mathcal{V}$. The variables x_a that are arguments for factor f_a are represented by edges connected to node a. The edges connected to node a are denoted as $\mathcal{E}(a)$.

The resulting FFG is a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes representing factors, and $\mathcal{E}$ is the set of edges representing variables. By convention, nodes are typically indexed using letters such as a, b, c, while edges are denoted by i, j, k, unless otherwise specified.

The framework of an FFG also includes various subgraph definitions. For example, an edge-induced subgraph is defined as $\mathcal{G}(i) = (\mathcal{V}(i), i)$, which includes all nodes connected by edge i. In contrast, a node-induced subgraph is represented as $\mathcal{G}(a) = (a, \mathcal{E}(a))$, involving

all edges connected to node a. Moreover, we introduce a local subgraph termed $\mathcal{G}(a,i) = (\mathcal{V}(i), \mathcal{E}(a))$, which collects all the local nodes and edges around a and i, respectively. We follow the FFG framework presented in Şenöz et al. (2021), which requires that each edge is terminated by exactly two nodes. This can be accomplished by terminating half-edges with a "dummy" factor that is proportional to 1.

The FFG framework is particularly useful for inference in probabilistic models, as it is associated with an objective function called Bethe free energy (Yedidia et al., 2001). The Bethe Free Energy serves as a variational objective, supporting the derivation of local message-passing update rules that minimize this energy.

In particular, Şenöz et al. (2021) demonstrated that the marginal update rule for an FFG representation of model (1) can be obtained as stationary solutions of Bethe free energy (BFE) augmented with marginalization constraints. The messages are derived as the exponentiated Lagrange multipliers that enforce the marginalization constraints. Rigorously, (Şenöz et al., 2021, Theorem 1) shows that, given a subgraph $\mathcal{G}(b,i)$ as displayed in Fig. 1, the local stationary points of the minimization problem $\arg\min_q L[q,f]$, where the Lagrangian $L[q,f]$ is

$$\begin{aligned} L[q,f] = &\sum_{a\in\mathcal{V}} D_{\mathrm{KL}}[q_a||f_a] + \sum_{i\in\mathcal{E}} H[q_i] + \sum_{a\in\mathcal{V}} \psi_a \left[\int q_a(x_a)\mathrm{d}x_a - 1\right] + \sum_{i\in\mathcal{E}} \psi_i \left[\int q_i(x_i)\mathrm{d}x_i - 1\right] \\ &+ \sum_{a\in\mathcal{V}} \sum_{i\in\mathcal{E}(a)} \int \lambda_{ia}(x_i) \left[q_i(x_i) - \int q_a(x_a)\mathrm{d}x_{a\setminus i}\right] \mathrm{d}x_i \end{aligned} \tag{2}$$

are given by

$$\mu_{ic}(x_i) = \int f_b(x_b) \prod_{\substack{j\in\mathcal{E}(b)\\ j\neq i}} \mu_{jb}(x_j)\mathrm{d}x_j \tag{3a}$$

$$q_i(x_i) = \frac{\mu_{ib}(x_i)\mu_{ic}(x_i)}{\int \mu_{ib}(x_i)\mu_{ic}(x_i)\mathrm{d}x_i}\,. \tag{3b}$$

The auxiliary variables $\mu_{ic}(x_i)$ (and $\mu_{jb}(x_i)$) in (3a) serve as messages transmitted from node c (and b) to edge i (and j). For more information on factor graphs and message passing-based inference, we refer to Loeliger (2007) and Şenöz et al. (2021).

2.2. The LogGamma Distribution

In this paper, we use the LogGamma distribution, which is a probability distribution defined on the real line, with a probability density function given by

$$\mathcal{LG}(x \mid a,b) = \frac{e^{bx}e^{-e^x/a}}{a^b\Gamma(b)}, \quad -\infty < x < \infty, \tag{4}$$

where $a > 0$ is the scale parameter and $b > 0$ is the shape parameter.

We will also use an auxiliary unnormalized distribution on a multidimensional variable β. This function is derived from the LogGamma distribution and is defined as

$$\widetilde{\mathcal{LG}}(\beta \mid a,b,x) \triangleq \mathcal{LG}(\beta^\top x \mid a,b), \tag{5}$$

where x is a fixed vector of covariates, and α and β are the scale and shape parameters of the LogGamma distribution, respectively.

3. Problem Statement

To define a precise problem statement, we first follow the solution approach from Akbayrak et al. (2022) for an elementary Poisson log-linear model

$$p(y_i \mid \lambda_i) = \frac{{\lambda_i}^{y_i} e^{-\lambda_i}}{y_i!}, \quad \log(\lambda_i) = \beta^\top x_i. \tag{6}$$

To clarify the derivations, we will assume a Gaussian prior distribution for β. However, regardless of the prior choice, the model would be non-conjugate.

We factorize model (6) into a model of the form (1) by

$$p(y, x, \beta) = \mathcal{N}(\beta|\mu, \Sigma) \prod_{i=1}^{N} p(y_i \mid \lambda_i)\delta(\beta^\top x_i - \log \lambda_i). \tag{7}$$

The FFG visualization of (7) is shown in Fig. 2. Applying the marginal update rules (3) to our model, the posterior marginal of interest $q(\beta)$ is obtained in the following form (the derivation is provided in Appendix C)

$$\mu_l(\beta) \propto \prod_i \widetilde{\mathcal{LG}}(\beta|y_i + 1, 1, x_i) \tag{8a}$$

$$\mu_p(\beta) \propto \mathcal{N}(\beta|\mu, \Sigma) \tag{8b}$$

$$q(\beta) = \frac{\mu_p(\beta)\mu_l(\beta)}{\int \mu_p(\beta)\mu_l(\beta)\mathrm{d}\beta}. \tag{8c}$$

The main challenge lies in the fact that (8c) involves the multiplication of an unnormalized LogGamma distribution $\mu_l(\beta)$ and a normal distribution $\mu_p(\beta)$, which does not yield a closed-form expression due to the nonconjugacy of these distributions.

The key idea of Akbayrak et al. (2022) is to substitute the exact marginal computation with a parametric optimization. Consider a parametric family Q of distributions, defined as

$$Q = \left\{ q_\theta(\beta) \;\middle|\; \theta \in \Theta, \int q_\theta(\beta)\mathrm{d}\beta = 1 \right\}, \tag{9}$$

where Θ is a set of valid parameters for the distributions in Q. The parameters $\hat{\theta}$ of the approximate marginal $q_{\hat{\theta}}(\beta)$ are obtained by minimizing the free energy:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \Big(\mathbb{E}_{q_\theta}[\log q_\theta] - \mathbb{E}_{q_\theta}[\log \mu_l] - \mathbb{E}_{q_\theta}[\log \mu_p] \Big). \tag{10}$$

This approximation constitutes a local application of the BLR rule introduced by Khan and Rue (2023). Essentially, Akbayrak et al. (2022) substitutes the exact marginal computation (8) with the parametric optimization problem (10). If we have an effective way to

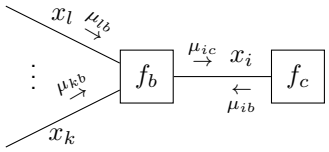

Figure 1: Visualization of a subgraph with indicated sum-product messages.

generate the solution $\hat{\theta}$ of the problem (10), we will obtain an efficient way to obtain the posterior marginal.

Although Khan and Rue (2023) proposed the BLR rule as a generic approach to inference in probabilistic models, the conditions to obtain a closed-form solution were not explicitly formulated. Similarly, Akbayrak et al. (2022) showcased how the BLR can be used as a generic tool to approximate inference algorithms on FFGs, but did not discuss the specific conditions under which the scheme becomes computationally efficient or accepts closed-form updates above the simple conjugacy. Identifying such conditions would enable the derivation of tractable update rules for the approximate posterior distribution, leading to more scalable and practical inference methods.

In view of the above, we define the following *problem statement:* Specify the conditions under which the inference scheme (10) becomes efficient and tractable, providing closed-form solutions or update rules for the approximate posterior distribution $q_{\hat{\theta}}(\beta)$.

4. Solution Proposal

4.1. Natural Gradient Optimization for Bayesian Poisson Regression

The problem (10) is essentially a parametric optimization problem, where the goal is to find the stationary points of the following optimization problem

$$\text{minimize} \quad \mathcal{F}(\theta)\,, \tag{11}$$

where θ is constrained to be in Θ. To solve this optimization problem, we propose employing natural gradient descent steps as introduced by the Bayesian Learning Rule (BLR) in Khan and Rue (2023). The BLR suggests updating the parameters θ of the approximate marginal using the following stationary point scheme with an update rule using a sequence of learning rates $\rho_t > 0$:

$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{q_\theta}[\log \mu_l] - \mathbb{E}_{q_\theta}[\log \mu_p] \tag{12a}$$

$$\theta_{t+1} = \arg\min_{\theta\in\Theta} \Big(\langle \nabla_\theta \mathcal{L}(\theta_t), \theta\rangle + \frac{1}{\rho_t} D_{\mathrm{KL}}[q_\theta(\beta)||q_{\theta_t}(\beta)] \Big), \tag{12b}$$

where θ_t represents the current estimate of the approximate marginal parameters at iteration t.

To ensure the tractability of the scheme (12), the choice of the family Q is crucial. For example, if Q is an exponential family of distributions,

$$q_\lambda(\beta) = h(\beta)\exp(\lambda^\top T(\beta) - A(\lambda)), \tag{13}$$

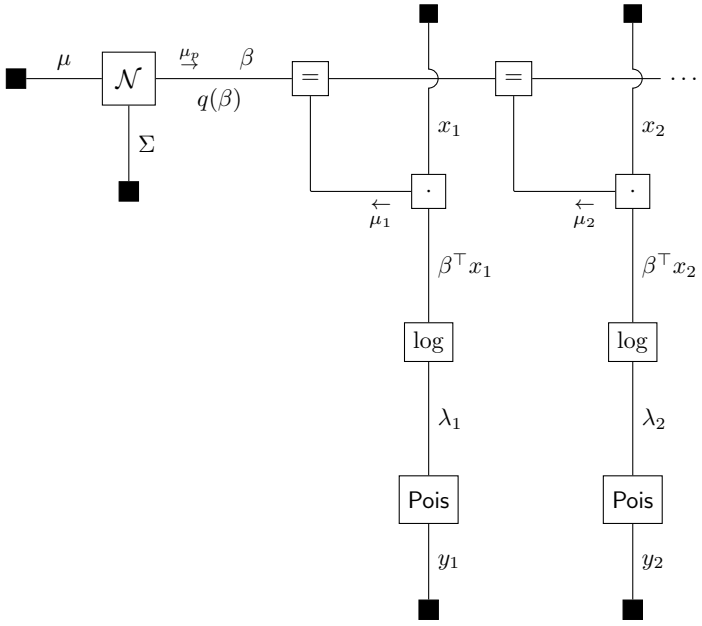

Figure 2: A Forney-style Factor Graph (FFG) representation of the Bayesian Poisson regression model (7) visualizes the factorization of the joint distribution on the observed data $x_1, \ldots, x_N; y_1, \ldots, y_N$, and the model parameter β. The posterior marginal $q(\beta)$ (8c) is obtained as a result of the product of the prior message μ_p (prior) and the likelihood messages $\mu_1, \ldots, \mu_n$ originating from the likelihood factors. Each likelihood message can be expressed as an auxilary distribution on β with parameters $\widetilde{\mathcal{LG}}(\beta|y_n + 1, 1, x_i)$ defined in Equation (5). Note that these likelihood messages are not normalized distributions over β. The combined likelihood message μ_l, which is the product of all individual likelihood messages μ_i, represents the overall contribution of the likelihood factors to the posterior. The main point of interest is the marginalization that occurs at the edge β (top left corner). At this point, the product of the prior message μ_p and the likelihood message μ_l is projected onto the normal family through a free energy minimization scheme (10). This projection enables the computation of an approximate posterior marginal $q(\beta)$ in a tractable form, although the exact posterior does not belong to the normal family due to the nonconjugacy of the likelihood with the normal prior.

where $h(\beta)$ is the base measure, λ is the natural parameter vector, $T(\beta)$ is the sufficient statistic vector, and $A(\lambda)$ is the log-partition function, then the BLR update rule (12b) simplifies to (Akbayrak et al., 2022, Appendix A)

$$\lambda_{t+1} = (1 - \rho_t)\lambda_t + \rho_t \widetilde{\nabla}_\lambda \mathcal{L}(\lambda_t), \tag{14}$$

where $\widetilde{\nabla}_\lambda$ denotes the natural gradient (Amari, 1998), defined as

$$\widetilde{\nabla}_\lambda \mathcal{L}(\lambda_t) = F(\lambda_t)^{-1} \nabla_\lambda \mathcal{L}(\lambda_t), \tag{15}$$

with $F(\lambda_t)$ being the Fisher information matrix at λ_t.

For this reason, in the next discussion, we will implicitly consider that we are dealing with a subfamily of exponential distributions whenever we denote the parameters by λ. When we are talking about any family, we will use θ.

In the following subsections, we introduce the concept of Q-conjugacy, which provides a principled way of selecting the family Q so that the expectations of the logarithms of μ_l and μ_p have closed-form expressions. Using the Q-conjugacy, we can derive an efficient gradient-based optimization scheme to find the optimal parameters $\hat{\theta}$.

4.2. Definition of Q-conjugacy

The choice of family Q is crucial to ensure the tractability of the optimization scheme derived from the BLR update rule, as shown in Equations (12) and (14). To make the scheme in closed=form, it is natural to consider scenarios in which we can compute the expectations of the logarithms of μ_l and μ_p under the variational distribution q_θ. When these expectations have closed-form expressions, the free energy gradient can be easily calculated, allowing the direct optimization of the objective with respect to the variational parameters θ. This condition can be summarized in the following definition 1.

Definition 1 (Q-conjugacy) *Let f and g be two positive functions such that*

$$\int f(x)g(x)\,\mathrm{d}x$$

is finite, and let

$$Q = \left\{ q_\theta(\beta) \;\middle|\; \theta \in \Theta, \int q_\theta(\beta)\mathrm{d}\beta = 1 \right\}$$

be a parametric family of distributions with a set of valid parameters Θ. Define the free energy objective function as

$$\mathcal{F}(\theta) = \mathbb{E}_{q_\theta}[\log q_\theta] - \mathbb{E}_{q_\theta}[\log f] - \mathbb{E}_{q_\theta}[\log g].$$

*f and g are **Q-conjugate** if the following conditions are met:*

1. *$\mathcal{F}(\theta)$ is a closed-form expression in terms of analytic functions of θ;*
2. *There exists a minimizer $\hat{\theta}$ of $\mathcal{F}(\theta)$:*

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{F}(\theta).$$

When a pair of functions are Q-conjugate, where Q is a subfamily of the exponential family, we can obtain a cheap scheme to find the stationary point (14), mitigating the main challenges faced by the BLR update scheme (12). The first condition of Q-conjugacy requires that expectations $\mathbb{E}_{q_\theta}[\log f]$ and $\mathbb{E}_{q_\theta}[\log g]$ can be expressed in closed-form as analytic functions of θ. This condition has significant implications for the optimization of free energy $\mathcal{F}(\theta)$. When expectations are available in closed-form, we can directly compute the gradient of $\mathcal{L}(\theta)$ with respect to θ, allowing efficient optimization of the free energy.

To illustrate the richness of the Q-conjugacy definition, we provide an example showing that classical conjugacy is a special case of Q-conjugacy.

Example 1 (Classical conjugacy as a special case of Q-conjugacy) *Let Q be a parametric family consisting of distributions from the exponential family. If $f(x)$ and $g(x)$ are two specific distributions within this family, then they are Q-conjugate.*

The detailed explanation of Example 1 is provided in Appendix A.

To further demonstrate that Q-conjugacy is a richer property than classical conjugacy, we present the example 2, which shows that the exponential and LogNormal distributions are Q-conjugate factors. The detailed derivation of this example is provided in Appendix B.

Example 2 (Q-conjugacy of Exponential and LogNormal Distributions) *Let $f(x;\kappa)$ be an exponential distribution with rate parameter $\kappa > 0$, and $g(x;\mu,\sigma)$ be a LogNormal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Define the exponential family Q as*

$$Q = \left\{ \frac{e^{-\lambda x}}{\lambda} \;\middle|\; \lambda > 0 \right\}.$$

Then f and g are Q-conjugate.

4.3. Q-conjugacy in Bayesian Poisson Regression

In the context of Bayesian Poisson regression (6), we can leverage the concept of Q-conjugacy to derive an efficient gradient-based optimization scheme to find the approximate posterior distribution. Recall that the exact posterior marginal $q(\beta)$ in (8c) involves the product of the likelihood message μ_l and the prior message μ_p (8).

The following lemma establishes the Q-conjugacy of the messages μ_l and μ_p when Q is the family of normal distributions.

Example 3 (Q-conjugacy in Bayesian Poisson Regression) *Consider the likelihood message μ_l and the prior message μ_p defined in (8). Let Q be the family of normal distributions. Then μ_l and μ_p are Q-conjugate.*

Proof First, we compute the expectation of the log-density of the LogGamma distribution over a univariate normal distribution with mean μ and variance σ^2:

$$\mathbb{E}_{\mathcal{N}(x|\mu,\sigma)}\left[\log \mathcal{LG}(x|\alpha,\beta)\right] = \mu\beta - \frac{\exp(\mu + \frac{\sigma^2}{2})}{\alpha} - \beta\log\alpha - \log\Gamma(\beta) \tag{16}$$

Next, we extend this result to the expectation over a multivariate normal distribution with mean μ and covariance Σ

$$\mathbb{E}_{\mathcal{N}(\beta|\mu,\Sigma)}[\log \widetilde{\mathcal{LG}}(\beta|a,1,x_i)] = \mu^\top x_i - \frac{\exp\left(\mu^\top x_i + \frac{x_i^\top \Sigma x_i}{2}\right)}{a} - \log a - \log\Gamma(1) \tag{17}$$

Using this result, we can show that the first condition of Q-conjugacy is satisfied:

1. The expectation of the log-likelihood message μ_l can be expressed as a closed-form analytic expression of the natural parameters $\lambda = (\lambda_1, \lambda_2)$ of the normal distribution:

$$\mathbb{E}_{q_\lambda}[\log \mu_l(\beta)] = \sum_i \frac{-\lambda_2^{-1}\lambda_1^\top x_i}{2} - \frac{\exp\left(-\frac{\lambda_2^{-1}\lambda_1^\top x_i}{2} - \frac{x_i^\top \lambda_2^{-1} x_i}{4}\right)}{a_i} - \log a_i - \log\Gamma(1), \tag{18}$$

   where $a_i = y_i + 1$.

   The expectation of $\log \mu_p$ is trivially a closed-form analytic expression of λ, since they are both normal distributions.

2. The existence of the minimizer $\hat{\lambda}$ of the free energy $\mathcal{F}(\lambda)$ can be established by noting that $\mathcal{F}(\lambda)$ is a bounded, continuous function that diverges at the boundary.

■

4.4. Summary

The main observation of Q-conjugacy is that the free-energy objective is more frequently known in closed-form than the marginal distribution. This enables efficient gradient-based optimization schemes for a wider range of models, even when the marginal distribution is not available in closed-form. Q-conjugacy allows deriving approximate posterior distributions through gradient descent, as in Bayesian Poisson regression, or obtain approximate posterior marginals in closed form, as for exponential and LogNormal distributions.

5. Numerical Illustrations

5.1. Experimental Scope

In this section, we evaluate the effectiveness and efficiency of our proposed closed-form gradient update scheme. We compare our approach with state-of-the-art MCMC sampling methods implemented in both Stan (Carpenter et al., 2017) and NumPyro (Phan et al., 2019) probabilistic programming languages. The No-U-Turn-Sampler (NUTS) (Hoffman and Gelman, 2014) implementation in Stan is selected for comparison due to its popularity. NumPyro's NUTS implementation is selected for its efficiency and widespread use. As demonstrated in Phan et al. (2019), NumPyro's adaptive variant of NUTS outperforms Stan's implementation on a wide range of models. Please keep in mind the following text:

The first experiment in Section 5.2 uses a Bayesian Poisson regression model (see Equation (6)) to examine the trade-offs between computational cost and accuracy when the posterior distribution can be easily sampled using MCMC methods. The second experiment in Section 5.3 illustrates the benefits of incorporating structural information through the variational family. This is because sampling-based inference methods may converge to an incorrect distribution form. This not only shows an improvement in effectiveness, but also an enhancement in accuracy compared to the sampling methods mentioned earlier.

All experiments are conducted on a machine equipped with an Intel Core i7-10700K CPU running at 3.80GHz and 64GB of RAM, ensuring consistent hardware performance across the compared methods. Our proposed approach is implemented in Julia 1.10.3 (Bezanson et al., 2017). For the MCMC sampling methods, we use Stan 2.32.2 through R language 4.4.0 (R Core Team, 2024) and NumPyro 0.15.0 with Python 3.10.3 (Van Rossum and Drake, 2009).

5.2. Poisson Regression

We generated synthetic data for a Poisson log-linear model (6) with sample sizes of 250, 2500, 5000, and 10000, and 5 covariates. The common parameter β is generated by independently sampling its components from a standard normal distribution, that is, $\beta_j \sim \mathcal{N}(0, 1)$ for $j = 1, \ldots, 5$. For each sample size n, we generate n independent samples y_i $(i = 1, \ldots, n)$

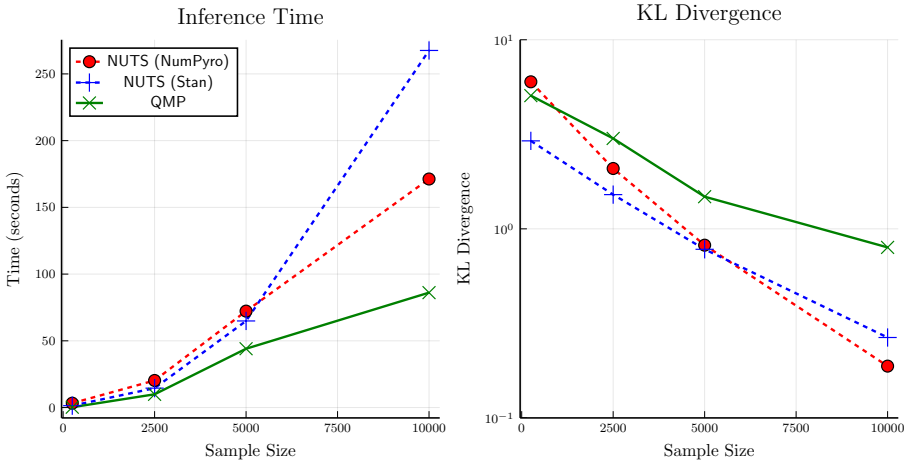

Figure 3: Comparison of inference time and KL divergence for QMP vs NUTS in NumPyro and Stan on synthetic data generated from a Poisson log-linear model with varying sample sizes (250, 2500, 5000, and 10000) and 5 covariates. The left subplot shows the inference time, while the right subplot displays the Kullback-Leibler (KL) divergence between the inferred posterior and the generative distribution of β, which we want to recover. QMP is significantly faster than NUTS in both NumPyro and Stan, requiring only 80 seconds for inference with 10,000 samples, compared to 175 and 275 seconds, respectively. This demonstrates QMP's scalability and efficiency for larger datasets. In terms of posterior quality, NUTS in NumPyro and Stan converge rapidly toward the generative distribution of β as the sample size increases, with KL decreasing from around 6 and 3 (for 250 samples) to 0.175 and 0.275 (for 10,000 samples). QMP shows a more gradual improvement, with the KL divergence reducing from 5 (for 250 samples) to 0.8 (for 10,000 samples).

from a Poisson distribution with mean $\lambda_i = \exp(\beta^\top x_i)$, where x_i represents the predictors for the i-th sample, with only one constraint: $\lambda_i \leqslant 100$ to ensure numerical stability for all algorithms.

The goal of the inference task is to estimate the posterior distribution of the regression coefficients $q(\beta)$, given the observed predictors x_i and counts y_i. We compare the performance of our proposed Q-conjugate Message Passing (QMP) algorithm with the NUTS implementations in NumPyro (N-NUTS) and Stan (S-NUTS).

The results are presented in Figure 3. Although N-NUTS and S-NUTS provide more accurate posterior estimates, especially for larger sample sizes, QMP offers a significant advantage in computational efficiency, making it suitable for large-scale or time-sensitive applications. The choice of inference framework depends on the specific requirements of the problem, such as computational resources, desired accuracy, and data scale.

5.3. Inference over a Positive Variable

We next consider a generative model with a Gamma-distributed latent variable and log-normal observations to evaluate QMP's inference performance in terms of accuracy and efficiency. The key difference from the previous experiment is that we infer the posterior over positive real numbers and have access to the true Bayesian posterior due to the one-dimensional nature of the problem, allowing assessment of the methods' accuracy. This model, using the (Gamma, Log-normal, Gamma) triplet, demonstrates the flexibility of the

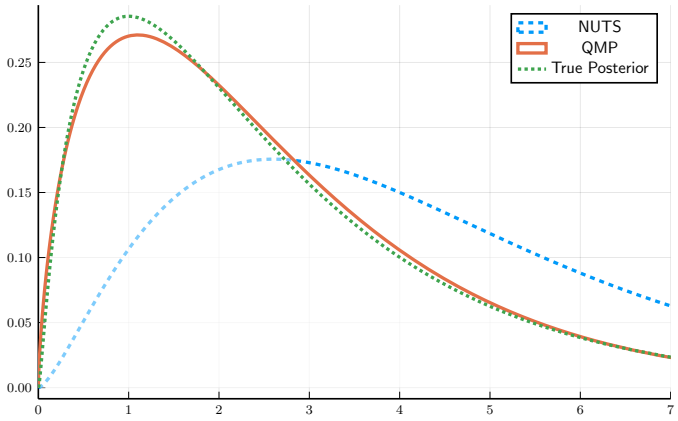

Figure 4: Inference results for the latent variable γ (the model (19)) inferred by NUTS, and the proposed QMP method. QMP more accurately captures the true posterior compared to NUTS, measured by a substantially lower KL divergence (see Table 1) to the true distribution, demonstrating its ability to better estimate uncertainty.

Table 1: We perform inference on the latent variable γ using the model specified in (19) with two different methods: NUTS (NumPyro) and the proposed Q-conjugate message passing. We compare the results of NUTS and Q-conjugate message passing in terms of average KL over 1000 generated points divergence from ground truth and inference execution time.

Method	KL	Time
QMP	0.03	87.242 μs
NUTS	0.35	5 s

QMP-based inference framework. In this triplet, the last Gamma specifies that we use a Gamma distribution as our variational family to approximate the posterior.

The model is specified as

$$p(x, \gamma) = \mathcal{N}(x \mid \log \gamma, 2)\Gamma(\gamma \mid 2, 2)\,. \tag{19}$$

Our goal is to infer a posterior $q(\gamma)$ given an observation x. We generate $1,000$ samples of the latent variable γ from $\Gamma(3, 1)$. We selected this distribution to be highly skewed and challenging to approximate using a normal distribution. This allows us to evaluate how well inference methods can handle cases where using the correct distributional form is essential for accurately estimating uncertainty. Next, for each γ sample, we generate an observation x from the normal distribution $\mathcal{N}(\log \gamma, 2)$. The large variance introduces significant noise, making an exact inversion for accurate estimation impossible.

We then performed inference for γ using both NumPyro's NUTS and QMP. The true posterior distribution for each data point was computed in quadratures, taking advantage of the one-dimensional nature of the problem. This allows us to assess the accuracy of the inference methods by comparing their results with the ground truth.

To compare the performance of NUTS and QMP, we calculated the Kullback-Leibler (KL) divergence between the inferred posterior and the true Bayesian posterior for each data point. Table 1 reports the average KL divergence, demonstrating the superior accuracy of QMP in capturing the true posterior distribution. Figure 4 shows the inferred posteriors for one of the $1,000$ runs, demonstrating the ability of QMP to approximate the true posterior more closely than NUTS. This example demonstrates how adding additional structure through the variational family can lead to improved accuracy compared to the MCMC methods.

6. Discussion

The work of Ranganath et al. (2014) has shown that stochastic gradient methods can efficiently solve the ELBO maximization problem, motivating the use of gradient-based approaches to solve (10). Following Khan and Rue (2023), we propose using natural gradient descent methods when the family is a parametric subfamily of the exponential distribution family. However, the success of these methods depends on the existence and computability of the free energy objective's gradient with respect to the variational parameters.

The BLR-based message passing scheme, as described in (10), suffers from high variance and convergence difficulties when applied with noisy natural gradients Akbayrak et al. (2022). To address this issue, we introduced the Q-conjugacy condition, under which the BLR message passing transforms into a closed-form update scheme over the parameters of the posterior marginal. As demonstrated by the numerical illustrations in Section 5, when the Q-conjugacy condition is satisfied, the BLR-based inference outperforms sampling methods, not only in terms of efficiency but also in terms of precision, particularly when the ability to select the structure of the marginal is leveraged.

It should be noted that D'Angelo and Canale (2023) proposed an effective Metropolis-Hastings method that relies on specific priors, such as horseshoe priors (Carvalho et al., 2010), to sample from posterior $q(\beta)$ (the model (6)). Although a direct comparison is not possible because the current Q-conjugate approach is not formulated for these priors, incorporating them into the message passing framework and comparing performance with specific inference schemes utilizing the posterior marginal form could be an interesting future research direction.

Moreover, recent work Kiral et al. (2023), which modifies the BLR rule to different forms of update steps beyond Euclidean geometry, could potentially yield significant performance improvements under the Q-conjugacy condition, as it transforms the Euclidean gradient into a Riemannian one for specific types of exponential family distributions, because Kiral et al. (2023) also for their approach to stay generic employ a noisy gradient estimator.

7. Conclusion

Our numerical experiments demonstrate that Q-conjugate triplets (Normal, LogGamma, Normal) and (Gamma, LogNormal, Gamma) offer improved accuracy-power trade-offs over MCMC methods, as shown in Sections 5.2 and 5.3, respectively. These findings enable efficient variational inference for real-world applications. Although based on carefully selected triplets, this work motivates research to discover additional triplets and develop generic, computationally efficient natural gradient-based inference methods.

Acknowledgments

This publication is part of the projects AUTO-AR and ROBUST (NWO: KICH3.LTP.20.006), which are partly financed by GN Hearing, the Eindhoven AI Systems Institute (EAISI), the Netherlands Enterprise Agency (RVO) and the Dutch Research Council (NWO).

References

S. Akbayrak, İsmail. Şenöz, A. Sarı, and B. de Vries. Probabilistic programming with stochastic variational message passing. *International Journal of Approximate Reasoning*, 148:235–252, Sept. 2022. ISSN 0888613X. doi: 10.1016/j.ijar.2022.06.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0888613X22000950.

S.-i. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2): 251–276, Jan. 1998. ISSN 0899-7667. doi: 10.1162/089976698300017746. URL https://doi.org/10.1162/089976698300017746.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL https://doi.org/10.1137/141000671.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76, 2017.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, June 2010. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asq017. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asq017.

A. B. Chan and N. Vasconcelos. Bayesian Poisson regression for crowd counting. In *2009 IEEE 12th International Conference on Computer Vision*, pages 545–551, Kyoto, Sept. 2009. IEEE. ISBN 978-1-4244-4420-5. doi: 10.1109/ICCV.2009.5459191. URL http://ieeexplore.ieee.org/document/5459191/.

S. Coxe, S. G. West, and L. S. Aiken. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment*, 91(2): 121–136, Feb. 2009. ISSN 0022-3891, 1532-7752. doi: 10.1080/00223890802634175. URL http://www.tandfonline.com/doi/abs/10.1080/00223890802634175.

L. D'Angelo and A. Canale. Efficient Posterior Sampling for Bayesian Poisson Regression. *Journal of Computational and Graphical Statistics*, 32(3):917–926, July 2023. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.2022.2123337. URL https://www.tandfonline.com/doi/full/10.1080/10618600.2022.2123337.

S. FrüHwirth-Schnatter and H. Wagner. Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika*, 93(4):827–841, Dec. 2006. ISSN 1464-3510, 0006-3444. doi: 10.1093/biomet/93.4.827. URL http://academic.oup.com/biomet/article/93/4/827/221862/Auxiliary-mixture-sampling-for-parameterdriven.

M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of machine learning research*, 15(1):1593–1623, 2014.

M. E. Khan and H. Rue. The Bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023.

S. Kim, Z. Chen, Z. Zhang, B. G. Simons-Morton, and P. S. Albert. Bayesian Hierarchical Poisson Regression Models: An Application to a Driving Study With Kinematic Events. *Journal of the American Statistical Association*, 108(502):494–503, June 2013. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2013.770702. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2013.770702.

E. M. Kiral, T. Möllenhoff, and M. E. Khan. The lie-group bayesian learning rule. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3352. PMLR, 2023.

S. Korl. A factor graph approach to signal modelling, system identification and filtering. *Series in Signal and Information Processing*, 15, 2005. ISSN 1616-671X. doi: 10.3929/ethz-a-005064226. URL https://www.research-collection.ethz.ch/handle/20.500.11850/82737.

J. H. Lambert. Observationes variae in mathesin puram. *Acta Helvetica Physico-Mathematico-Anatomico-Botanico-Medica*, 3:128–168, 1758.

H.-A. Loeliger. Factor Graphs and Message Passing Algorithms – Part 1: Introduction, 2007. URL http://www.crm.sns.it/media/course/1524/Loeliger_A.pdf. http://www.crm.sns.it/media/course/1524/Loeliger_A.pdf, last accessed on 3-4-2019.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Royal Statistical Society. Journal. Series A: General*, 135(3):370–384, May 1972. ISSN 0035-9238. doi: 10.2307/2344614. URL https://doi.org/10.2307/2344614.

D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

R. Ranganath, S. Gerrish, and D. Blei. Black Box Variational Inference. In S. Kaski and J. Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, Apr. 2014. PMLR. URL https://proceedings.mlr.press/v33/ranganath14.html.

İsmail. Şenöz, T. van de Laar, D. Bagaev, and B. de Vries. Variational Message Passing and Local Constraint Manipulation in Factor Graphs. *Entropy*, 23(7):807, July 2021. ISSN 1099-4300. doi: 10.3390/e23070807. URL https://www.mdpi.com/1099-4300/23/7/807. Publisher: Multidisciplinary Digital Publishing Institute.

J. D. Stamey, D. M. Young, and J. W. Seaman. A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression. *Statistics in Medicine*, 27(13):2440–2452, June 2008. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.3134. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.3134.

G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

E. W. Weisstein. Euler-mascheroni constant. *https://mathworld.wolfram.com/*, 2002.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, 13:24, 2001.

Appendix A. Proof of Classical Conjugacy as a Special Case of Q-conjugacy

Here we provide a proof that classical conjugacy is a special case of Q-conjugacy (Example 1)

Proof Consider a parametric family Q consisting of distributions from the exponential family, where each member can be expressed in the following form

$$q(x;\theta) = \exp(\eta(\theta)^\top T(x) - A(\theta)),$$

with $\eta(\theta)$ as the vector of natural parameters, $T(x)$ as the vector of sufficient statistics, and $A(\theta)$ as the logarithmic partition function that ensures normalization.

Let $f(x)$ and $g(x)$ be two specific distributions within this family, defined by

$$f(x) = q(x;\theta_f) = \exp(\eta(\theta_f)^\top T(x) - A(\theta_f)),$$
$$g(x) = q(x;\theta_g) = \exp(\eta(\theta_g)^\top T(x) - A(\theta_g)).$$

These distributions, by their parameterization and the inherent properties of the exponential family, meet the conditions for Q-Conjugacy:

1. The expectations of the logarithms of $f(x)$ and $g(x)$, under any distribution $q(x;\theta)$ in Q, are given by

$$\mathbb{E}_{q_\theta}[\log f(x)] = \eta_f^\top \nabla_\theta A(\theta) - A(\theta_f) \tag{20}$$
$$\mathbb{E}_{q_\theta}[\log g(x)] = \eta_g^\top \nabla_\theta A(\theta) - A(\theta_g)\,. \tag{21}$$

   These expressions demonstrate the analytic tractability required by Q-Conjugacy.

2. The unique stationary point of the free energy objective is given by

$$\eta(\theta^*) = \eta(\theta_f) + \eta(\theta_g).$$

   To demonstrate that θ^* is the global minimizer of the free energy $\mathcal{F}(\theta)$, we analyze the change in free energy from any point θ to θ^*. This change is given by:

$$\mathcal{F}(\theta) - \mathcal{F}(\theta^*) = A(\theta_f + \theta_g) - A(\theta) - (\eta_f + \eta_g - \eta(\theta))^\top \nabla_\theta A(\theta) \geqslant 0. \tag{22}$$

   The fact that this change is always positive establishes that θ^* is a global minimizer.

■

Appendix B. Q-conjugate triplets

Here we provide a proof that exponential distribution and LogNormal distribution form a Q-conjugate triplet with exponential family (Example 2)

Proof Consider the following factors:

$$f(x;\kappa) = \frac{e^{-\kappa x}}{\kappa}, \quad \kappa > 0,$$
$$g(x;\mu,\sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right), \quad \mu \in \mathbb{R}, \sigma > 0,$$

where f is the exponential distribution and g is the LogNormal distribution. Let Q be the exponential family defined by:

$$Q = \left\{ \frac{e^{-\lambda x}}{\lambda} \;\middle|\; \lambda > 0 \right\}.$$

1. For any $q(x;\lambda) \in Q$, the expectations of $\log f(x)$ and $\log g(x)$ are

$$\mathbb{E}q_\lambda[\log f(x)] = \log\lambda - \log\kappa,$$
$$\mathbb{E}q_\lambda[\log g(x)] = -\frac{-2(\gamma+\mu)\log(\lambda) + (\gamma+\mu)^2 + \log^2(\lambda) + \frac{\pi^2}{6}}{2\sigma^2} + \gamma - \log(\lambda) - \frac{1}{2}\log(2\pi) - \log(\sigma),$$

   where γ is the Euler-Mascheroni constant (Weisstein, 2002).These expressions are analytic functions of λ.

2. The free energy function $\mathcal{F}(\lambda)$ has a unique global minimum at λ^*, given by

$$\lambda^* = \kappa\sigma^2 \mathrm{W}\left(\frac{-\gamma - \mu - \frac{1}{\sigma^2}}{\kappa\sigma^2}\right), \tag{23}$$

   where W is the Lambert W function (Lambert, 1758).

Therefore, f and g are Q-conjugate factors, where $Q = \left\{ \frac{e^{-\lambda x}}{\lambda} \;\middle|\; \lambda > 0 \right\}$. ■

Appendix C. Derivation of the Posterior Marginal in Bayesian Poisson Regression

This appendix provides a detailed derivation of the posterior marginal (8)

$$q(\beta) \propto \mathcal{N}(\beta|\mu,\Sigma) \prod_i \widetilde{\mathcal{LG}}(\beta|y_i+1,1,x_i)$$

in the Bayesian Poisson regression model defined in the factorized form in the equation (7). The derivation process involves applying the marginal update rules (3) to the model

(7), which essentially reduces to the application of the Sum-Product algorithm (Korl, 2005, Equations 2.6 and 2.7).

First, applying the Sum-Product rule (Korl, 2005, Equations 2.6 and 2.7) to the factorized representation (7), we obtain the following expression (note that the observed values for $x_{1:N}$ and $y_{1:N}$ are denoted as $\hat{x}_{1:N}$ and $\hat{y}_{1:N}$, respectively)

$$q(\beta) = \mathcal{N}(\beta|\mu,\Sigma) \int \prod_{i=1}^{N} p(y_i, x_i, \beta)\delta(x_i - \hat{x}_i)\delta(y_i - \hat{y}_i)\mathrm{d}x_{1:N}\mathrm{d}\lambda_{1:N}\mathrm{d}y_{1:N}, \tag{24a}$$

$$p(y_i, x_i, \beta) = p(y_i \mid \lambda_i)\delta(\beta^\top x_i - \log\lambda_i). \tag{24b}$$

We can rewrite equation (24a) as a product of integrals, as each term depends only on variables with index i

$$q(\beta) = \mathcal{N}(\beta|\mu,\Sigma) \prod_{i=1}^{N} \int \delta(x - \hat{x}_i)\delta(y - \hat{y}_i)p(y_i \mid \lambda_i)\delta(\beta^\top x_i - \log\lambda_i)\mathrm{d}x_i\mathrm{d}\lambda_i\mathrm{d}y_i \tag{24c}$$

Then, we note that the following equality holds

$$\int p(y_i|\lambda_i)\delta(y_i - \hat{y}_i)\mathrm{d}y_i = \frac{e^{-\lambda_i}\lambda_i^{\hat{y}_i}}{\hat{y}_i!} \propto \Gamma(\lambda_i \mid \hat{y}_i + 1, 1) \tag{25}$$

Next, using the definition of the LogGamma distribution (equation (4)) and applying a change of variable $\upsilon_i = \log\lambda_i$, we obtain

$$\int \Gamma(\lambda_i \mid \hat{y}_i + 1, 1)\delta(\beta^\top x_i - \log\lambda_i)\mathrm{d}\lambda_i = \int \mathcal{LG}(\upsilon_i \mid \hat{y}_i + 1, 1)\delta(\beta^\top x_i - \upsilon_i)\mathrm{d}\upsilon_i \tag{26}$$

Finally, applying equations (25) and (26) to (24c) and the definition of the auxiliary distribution (5), we obtain the desired form of $q(\beta)$

$$q(\beta) \propto \mathcal{N}(\beta|\mu,\Sigma) \prod_{i} \widetilde{\mathcal{LG}}(\beta|\hat{y}_i + 1, 1, \hat{x}_i), \tag{27}$$

note that in equation (8), we are using a slight abuse of notation by referring to $\hat{x}_i$ as x_i and $\hat{y}_i$ as y_i.