Bayesian Inference for Gamma Models

Jingyu He
Department of Management Sciences
City University of Hong Kong

Nicholas Polson
Booth School of Business
The University of Chicago

Jianeng Xu
Booth School of Business
The University of Chicago

June 3, 2021

Abstract

We use the theory of normal variance-mean mixtures to derive a data augmentation scheme for models that include gamma functions. Our methodology applies to many situations in statistics and machine learning, including Multinomial-Dirichlet distributions, Negative binomial regression, Poisson-Gamma hierarchical models, Extreme value models, to name but a few. All of those models include a gamma function which does not admit a natural conjugate prior distribution providing a significant challenge to inference and prediction. To provide a data augmentation strategy, we construct and develop the theory of the class of Exponential Reciprocal Gamma distributions. This allows scalable EM and MCMC algorithms to be developed. We illustrate our methodology on a number of examples, including gamma shape inference, negative binomial regression and Dirichlet allocation. Finally, we conclude with directions for future research.

Key Words: Data Augmentation, Exponential Reciprocal Gamma, Pólya Gamma, Latent Dirichlet Allocation, Gamma Shape, Markov Chain Monte Carlo, Expectation-Maximization;

1 Introduction

Statistical models involving gamma functions are prevalent in statistics and machine learning. For example, topic models, negative binomial regression, time series count models, Poisson-Gamma hierarchical models, non-parametric Bayes, to name but a few (Rossell, 2009; Aktekin, Polson, and Soyer, 2018; Lijoi, Muliere, Prünster, and Taddei, 2016). Gamma distribution also serves as the conjugate prior for many model parameters, such as a normal precision and Poisson intensity. The normalizing constant depends on gamma function whose argument is the shape parameter. Bayesian Inference in gamma models is a long standing problem that presents significant technical and computational difficulties (Damsleth, 1975; Rossell, 2009; Miller, 2018). Similar issue also occurs to the learning of other widely-used gamma models. Table 1 gives a list of distributions, where the shape/concentration/dispersion parameters are nested in gamma functions.

By exploiting normal variance-mean mixture identities related to gamma function, we derive a general data augmentation strategy. Our main result is given in Proposition 1, according to which a MCMC algorithm is built in Proposition 2 and an expectation-maximization algorithm is developed in Section 3.1.

The main difficulty is to address the reciprocal gamma function $1/\Gamma(\alpha)$. Following Barndorff-Nielsen, Kent, and Sørensen (1982), Hartman (1976) and Roynette and Yor (2005), we represent $1/\Gamma(\alpha)$ as a mean-variance mixture of normals, then we define our new class of distributions named the exponential reciprocal gamma (ERG) class. This novel distribution places gamma models in the same footing as other commonly used Bayesian models, such as sparsity (lasso, horseshoe) and logit (Pólya-Gamma). As a by-product, we show how to nest the latter within our framework. Thus we unify the inference procedure for many models.

Our data augmentation strategy with ERG auxiliary variables may be utilized to design efficient Markov chain Monte Carlo (MCMC) algorithms in latent Dirichlet allocation (Blei, Ng, and Jordan, 2003), Beta-negative binomial models (Zhou, Hannah, Dunson, and Carin, 2012), and Gamma-Gamma (GaGa) hierarchical models (Rossell, 2009). It adds to the literature on Bayesian computation with auxiliary variables, which have proven useful in computing posterior distributions in logistic regression (Polson, Scott, and Windle, 2013), negative binomial regression (Pillow and Scott, 2012), multinomial factor models (Holmes and Held, 2006), support vector machines (Mallick, Ghosh, and Ghosh, 2005; Polson and Scott, 2011), and dependent multinomial models (Linderman, Johnson, and Adams, 2015).

Table 1: List of Gamma Models

Distribution	Likelihood	Applications
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	Gamma process, Poisson regression
Inverse Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	survival analysis, conjugate prior
Beta	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	order statistics, wavelet analysis
Dirichlet	$\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K x_k^{\alpha_k-1}$	topic model, Bayesian networks
Negative Binomial	$\frac{\Gamma(x+\alpha)}{\Gamma(\alpha)\Gamma(\alpha+1)} p^r (1-p)^x$	stock control problems, negative binomial regression

To illustrate our methodology, we show examples including gamma shape inference, negative binomial regression and multinomial-Dirichlet model. The first example has a posterior density which exactly matches with the above form. Hence our algorithms can be straightforwardly applied. The posterior samples are efficiently drawn and the posterior mode is easily found by EM algorithm. The second example extends our results to incorporate Pólya-Gamma mixture representation described in (Polson, Scott, and Windle, 2013). The last example demonstrates how our methodology generalizes to high-dimensional case. Conditioned on the auxiliary variables, all elements of the multivariate concentration parameter become mutually independent, which reduces the sampling difficulty significantly. We also suggest the use of normal approximation to speed up the sampling procedure in this example.

The ERG family of distributions is defined as an infinite convolution of Generalized inverse Gaussian (GIG) distributions and is related to the class of Pólya-Gamma (PG) distributions (Polson, Scott, and Windle, 2013) for logistic regression. Other mixture representations related to GIG are introduced in Zhang, Wang, Liu, Jordan, and Lawrence (2012) and Barndorff-Nielsen and Shephard (2012), with applications in sparse regression and stochastic volatility modelling. The ERG(0) distribution is also a special case of H_a^Γ distribution family studied in Roynette and Yor (2005), who provide a representation of the ratio gamma functions as a scale mixture of normals. This adds to scale mixtures results in Bayesian inference, see Andrews and Mallows (1974), Barndorff-Nielsen, Kent, and Sørensen (1982), West (1987), and Polson, Scott, and Windle (2013). Scale mixtures of normals are increasingly used in modeling complex high-dimensional distributions, and Bhattacharya, Chakraborty, and Mallick (2016) provide fast sampling strategies, adding to the practical use of scale mixture distributions in scalable stochastic simulations. Equivalently constructed scalable PG

sampling schemes are provided in Windle, Polson, and Scott (2014) and Glynn, Tokdar, Howard, and Banks (2019).

1.1 Connections with Existing Work

Obtaining random draws or finding the mode from a posterior distribution involving gamma functions is computationally challenging it requires accurate gamma function evaluations. Approximation methods have been proposed to handle the computational burden. Minka (2000) describes an efficient iterative schemes for maximum likelihood estimate of Dirichlet distribution. The likelihood of Dirichlet precision is approximated by a simpler function (gamma density) by matching the first two derivatives, while the multivariate Dirichlet mean is estimated separately by fixed-point iteration. Miller (2018) applies the same idea of derivative-matching and approximate the full conditional distribution of gamma shape parameter by a gamma density function. Rossell (2009) defines a gamma shape distribution in differential expression analysis. By approximating gamma function with Stirling's formula and evaluating the limit of the expression, the proposed distribution is roughly proportional to a gamma density. However, These ad-hoc methods are all essentially derived in univariate case. It's not straightforward to generalize them in multivariate cases as we need to deal with correlations, and the computation of approximating parameters itself will get cumbersome. Our framework instead provides an easy and unified way to derive both MCMC and EM algorithms for multivariate models, without the need to approximate density functions.

The rest of our paper proceeds as follows: Section 2 defines the class of ERG distributions; Section 3 illustrates our data augmentation strategy, developing a parameter expanded Gibbs sampler as well as EM algorithm; Section 4 presents examples of gamma shape inference, negative binomial regression and multinomial-Dirichlet model; and Section 5 concludes with directions for future research.

2 Exponential Reciprocal Gamma (ERG) Distribution

In this section, we present the theoretical development of the ERG distribution class, defining the ERG distribution by the form of its integral transform. In Section 2.1, we define a baseline case of the ERG distribution and prove that it is an infinite convolution of independent GIG distributions; in Section 2.2, the general class of exponential reciprocal distributions is constructed with an exponential tilting of the baseline case defined in Section 2.1. Other convolutions of GIG distribution are

discussed in Section 2.3.

2.1 ERG(0) Distribution

Let ERG(0) denote the exponential reciprocal gamma distribution. The parameter, which is a tilting parameter fixed at zero in this case, will be discussed in greater detail in Section 2.2.

Definition 2.1. Random variable X_0 has an exponential reciprocal gamma distribution, ERG(0), if and only if its density function $p_0(x)$ satisfies the following identity

$$E\left(e^{-s^2 X_0}\right) = \int_0^\infty e^{-s^2 x} p_0(x) dx = \frac{e^{-\gamma s}}{\Gamma(1+s)}, \qquad s > 0. \tag{1}$$

where $\gamma = -\psi(1) \approx 0.57721$ is the Euler-Mascheroni constant and $\psi(s) = \frac{d}{ds} \log \Gamma(s)$ is the digamma function.

Remark 1. *The product representation for the reciprocal gamma function due to Weierstrass is,*

$$\frac{e^{-\gamma s}}{\Gamma(1+s)} = \prod_{k=1}^{\infty} \left(1 + \frac{s}{k}\right) e^{-\frac{s}{k}}, \; s \in \mathbb{C}/\{0, -1, -2, ...\}$$

Remark 2. *Roynette and Yor (2005) prove the existence of an infinitely divisible distribution H_a^Γ, with density function $p_a^\Gamma(x)$, such that for $a > 0$*

$$E\left(e^{-\frac{1}{2}s^2 H_a^\Gamma}\right) = \int_0^\infty e^{-\frac{1}{2}s^2 x} p_a^\Gamma(x) dx = \frac{\Gamma(a)}{\Gamma(a+s)} e^{\psi(a)s}$$

which follows from the general product representation for the reciprocal gamma function,

$$\frac{\Gamma(a)}{\Gamma(a+s)} e^{\psi(a)s} = \prod_{k=0}^{\infty} \left(1 + \frac{s}{a+k}\right) e^{-\frac{s}{a+k}}. \tag{2}$$

Note that the representation in Remark 1 is a special case with $a = 1$. Hence

$$ERG(0) \overset{D}{=} \frac{1}{2} H_1^\Gamma.$$

Remark 3. *The ERG$(n, 0)$ is defines as follows. If $X_{n,0} \sim ERG(n, 0)$,*

$$E\left(e^{-s^2 X_{n,0}}\right) = \int_0^\infty e^{-s^2 x} p_{n,0}(x) dx = \frac{e^{-\gamma n s}}{\Gamma(1+s)^n}, \qquad n > 0, s > 0.$$

Note that ERG$(n, 0)$ is the equivalent in distribution to the sum of n independent ERG$(1, 0)$ when n is a positive integer.

2.2 General ERG(c) Distribution

We now construct the general class of ERG distributions, ERG(c), by exponentially tilting the ERG(0) distribution. The exponential tilting strategy – similar to the one used by Polson, Scott, and Windle (2013) – allows a second parameter $c \in \mathbb{R}^+$ to inform a priori the precision of the ERG random variable.

Definition 2.2. The ERG(c) distribution is constructed as an exponential tilting of the ERG(0) density. Its density function is

$$p_c(x) = Z_c \cdot \exp\left(-c^2 x\right) p_0(x), \quad x, c > 0.$$

The normalizing constant, namely $Z_c = 1/E\left(\exp\left(-c^2 X_0\right)\right)$ where X_0 is an ERG(0) random variable, can be calculated using the identity in Remark 1. Similarly, the integral identity of ERG(c) is given by

$$E\left(e^{-s^2 X_c}\right) = \prod_{k=1}^{\infty} \left(\frac{k + \sqrt{s^2 + c^2}}{k + c}\right) e^{-\frac{\sqrt{s^2+c^2}-c}{k}} \tag{3}$$

$$= \frac{\Gamma(1 + c)}{\Gamma(1 + \sqrt{s^2 + c^2})} e^{-\gamma(\sqrt{s^2+c^2}-c)}. \tag{4}$$

Our main result, presented in Theorem 1, is that a random variable $X_c \sim$ ERG(c) may be constructed from an infinite sum of independent generalized inverse Gaussian (GIG) random variables. The power of the result lies in the ability to identify previously unknown conditional posterior distributions.

Theorem 1. *The ERG(c) class of distributions can be constructed as an infinite sum of independent generalized inverse Gaussian (GIG) distributions as follows*

$$ERG(c) \overset{D}{=} \sum_{k=1}^{\infty} GIG\left(-\frac{3}{2}, 2c^2, \frac{1}{2k^2}\right).$$

In particular, when $c = 0$, the GIG distribution reduces to inverse gamma distribution. Hence,

$$ERG(0) \overset{D}{=} \sum_{k=1}^{\infty} \frac{1}{4k^2} \Gamma_k^{-1}.$$

where Γ_k^{-1} are i.i.d. inverse gamma random variables with shape $\frac{3}{2}$ and scale 1.

Proof. See Appendix A. □

The following theorem concerns the first two moments of ERG(c) which will be used to construct our EM algorithm in Section 3.1 and the approximate Gibbs sampler in Appendix D.

Theorem 2. *If G_k is a GIG random variable with $p = -\frac{3}{2}, a = 2c^2, b = \frac{1}{2k^2}$, then the mean and variance of the tail infinite sum $\sum_{k=N}^{\infty} G_k$ are*

$$\begin{aligned} E\left(\sum_{k=N}^{\infty} G_k\right) &= \frac{1}{2c}\left(\psi(N+c) - \psi(N)\right) \\ Var\left(\sum_{k=N}^{\infty} G_k\right) &= \frac{1}{4c^3}\left(\psi(N+c) - \psi(N) - c\psi'(N+c)\right), \end{aligned}$$

where $\psi(s) = \frac{d}{ds}\log\Gamma(s)$ is the digamma function. Setting $N = 1$ gives us the first two moments of ERG(c).

Proof. If g_k is a GIG random variable with $p = -\frac{3}{2}, a = 2c^2, b = \frac{1}{2k^2}$, then its mean and variance are given by

$$E(g_k) = \frac{1}{2}\left(\frac{1}{k^2 + ck}\right), \quad \text{Var}(g_k) = \frac{1}{4c}\left(\frac{1}{k^3 + 2ck^2 + c^2k}\right).$$

Theorem 2 is thus a direct application of Theorem 1. □

Remark 4. *The ERG(c) distribution class belongs to the family of generalized gamma convolutions (GGC), Bondesson (1992). Its Laplace transform in Equation (4) also satisfies*

$$E(e^{-sX_c}) = \exp\left\{\int_0^{\infty} (e^{-sx} - 1)\nu(x)dx\right\}, \quad s \geq 0.$$

Its Lévy density $\nu(x)$ and Thorin density $\mu(t)$ are

$$\begin{aligned} \nu(x) &= \frac{1}{x}\int_0^{\infty} e^{-tx}\mu(t)dt, \\ \mu(t) &= \mathbb{1}_{t \geq c^2} \frac{\psi(1 - \sqrt{c^2 - t}) + \psi(1 + \sqrt{c^2 - t}) + 2\gamma}{4\pi\sqrt{t - c^2}}. \end{aligned}$$

This result can be used to generate ERG random variables as it shows that it falls into the class of Generalized Gamma Convolution, see Bondesson (1982) and Rosiński (2001).

2.3 GIG Mixtures

The ERG distribution allows us to represent the unnormalized density $\frac{e^{ax}}{\Gamma(1+x)}$ as a normal variance-mean mixture. That is,

$$\frac{e^{ax}}{\Gamma(1+x)} = \int_0^\infty \phi(x \mid \mu(\omega), \sigma^2(\omega)) \cdot \tau(\omega) p_0(\omega) d\omega$$

where $\phi(\cdot \mid \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2 and $P_0(\cdot)$ is the distribution function of ERG(0). $\mu(\omega) = (a+\gamma)/(2\omega), \sigma^2(\omega) = 1/(2\omega)$ and $\tau(\omega) = \sqrt{\pi/\omega}\exp\{(a+\gamma)/4\omega\}$. In statistics and machine learning, probability distributions whose density function $p(x)$ is of the following form are used explicitly and implicitly:

$$p(x) = \int_0^\infty f(x \mid \theta(\omega)) p(\omega) d\omega. \tag{5}$$

Here $f(x \mid \theta(\omega))$ is some well-known density function, e.g. normal, and the mixing $p(\omega)$ is the distribution of a single GIG or an infinite convolution of GIG's. Combined with a data-augmentation scheme, the above mixture representation provides a powerful framework to solve many non-Gaussian models.

1. When $f = \phi$ and the mixing distribution is GIG, Polson and Scott (2013) give the variance-mean mixture representations for many common loss functions in regression and binary classification problems, which corresponds to different choices of the function $\theta(\omega) = (\mu(\omega), \sigma^2(\omega))$ and parameters of GIG. For example, absolute loss $L(y) = |y|$, hinge loss for support vector machine $L(y) = \max(1-y, 0)$ and check loss for quantile regression $L(y) = |y| + (2q-1)y$. The representations then help reduce those non-Gaussian models to Gaussian linear models with heteroscedastic errors. Note that GIG is a very general family with many common distributions as its special cases, such as gamma, inverse gamma and inverse Gaussian distribution.

2. For the logistic loss in binary classification $L(y) = \log(1+e^y)$, Polson and Scott (2013) show that it is also a normal variance-mean mixture. The mixing distribution is Pólya distribution, which is constructed as an infinite sum of exponentials. Note that exponential distribution is again a special case of GIG.

3. By choosing f to be the exponential power density, $f(x \mid \eta(\omega), q) \propto \exp\left\{-\frac{1}{2\eta(\omega)}|x|^q\right\}$, and the mixing distribution $p(\omega)$ to be GIG, Zhang, Wang, Liu, Jordan, and Lawrence (2012) introduce a sparsity-inducing prior called EP-GIG and develop EM algorithms for sparse learning. The density function of EP-GIG is given explicitly and special cases (generalized t distribution and exponential power-gamma distribution) are discussed when the mixing GIG reduces to inverse gamma and gamma respectively.

4. The Pólya-Gamma distribution class is proposed by Polson, Scott, and Windle (2013) to solve the inference problem in models with binomial likelihoods, including logistic regression and negative binomial regression. Pólya-Gamma distribution $\text{PG}(b, c)$ can be written as an infinite convolution of gamma distributions whose shapes are all equal to b and scales depend on c, or equivalently $\text{GIG}(b, 2, 0)$.

$$\text{PG}(b, c) \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{\text{GIG}(b, 2, 0)}{(k - 1/2)^2 + c^2/(4\pi^2)}.$$

Choose $f(z \mid \omega) = 2^{-b} \exp\left\{-\omega z^2/2 + \kappa z\right\}$ with $\kappa = a - b/2$ and $z = x'\beta$, the likelihood of logistic regression is a mixture

$$\frac{(e^z)^a}{(1 + e^z)^b} = \int_0^{\infty} f(z \mid \omega) p_{\text{PG}}(\omega \mid b, 0) d\omega. \tag{6}$$

Here $p_{\text{PG}}(\omega \mid b, 0)$ is the density of $\text{PG}(b, 0)$ and $f(z \mid \omega)$ is proportional to the normal density where the mean and variance are functions of ω.

5. Barndorff-Nielsen and Shephard (2012) use a normal variance-mean mixture as a general approach of building densities on the real line. Here $f(x \mid \theta(\omega)) = \phi(x \mid \mu + \beta\omega, \omega)$. When $p(\omega)$ is GIG density, the resulted mixture is generalized hyperbolic distribution, which includes many special cases such as normal inverse Gaussian, normal gamma, Laplace, skewed Student's t distribution. Furthermore, normal distribution can also be written as a limiting case of generalized hyperbolic distribution.

Table 2: Integral Representation for Gamma Functions

	Integral Representation	Auxiliary Variables
$\Gamma(\alpha)$	$\int_0^\infty x^{\alpha-1}e^{-x}dx$	Gamma
$\frac{1}{\Gamma(\alpha)}$	$\int_0^\infty \alpha e^{-\alpha^2 x+\gamma\alpha}p_0(x)dx$	ERG
$\frac{\Gamma(\alpha)}{\Gamma(\alpha+\beta)}$	$\int_0^1 \frac{1}{\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx$	Beta

3 MCMC and Data Augmentation

This section illustrates the data augmentation strategy and sampling scheme for Gamma inference using the ERG class distributions. First, notice that many Bayesian gamma models involve a posterior density of the form

$$p(x) = C \cdot \left(\prod_{\ell=1}^{L}\Gamma(g_\ell(x))\right)\left(\prod_{m=1}^{M}\frac{1}{\Gamma(h_m(x))}\right)\left(\prod_{n=1}^{N}\frac{\Gamma(j_n(x))}{\Gamma(j_n(x)+\beta_n)}\right)\cdot x^{p-1}e^{-ax^2+bx},\ x>0, \quad (7)$$

where $x \in \mathbb{R}^+$ is on the positive real line, and C is the normalizing constant. The arguments in gamma functions, $\{g_\ell(\cdot)\}_{\ell=1}^L, \{h_m(\cdot)\}_{m=1}^M$ and $\{j_n(\cdot)\}_{n=1}^N$ are nonnegative increasing linear functions of x on $(0,\infty)$. We assume $M \geq 1$, otherwise the form might not be integrable. Parameters $p, a, b, \beta_1, ..., \beta_N$ are scalars. β_n's are positive. p and a are nonnegative. In order to perform full posterior inference on the variable x, a sampling procedure for $p(x)$ is needed, while a Maximum A Posteriori (MAP) estimate requires finding the maximizer of it.

The idea of the data augmentation strategy is to introduce a group of auxiliary random variables $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)^T$ such that for $\boldsymbol{\omega} \in \Omega$,

$$p(x) = \int_\Omega p(x, \boldsymbol{\omega})d\boldsymbol{\omega}$$

and the joint density $p(x, \boldsymbol{\omega})$ after augmentation is easier to deal with, as it doesn't consist of gamma functions any longer.

Returning to Equation (7), the posterior density involves gamma functions $\Gamma(\cdot)$, reciprocal gamma functions $1/\Gamma(\cdot)$ and gamma ratios $\Gamma(\cdot)/\Gamma(\cdot+\beta)$. We will then express each of them using the corresponding integral representation in Table 2. This is equivalent to introducing an auxiliary random variable for each of them. The total number of the auxiliary random variables is then

$L + M + N$.

Proposition 1. The posterior density admits an integral representation as follows

$$p(x) = C \cdot \int_{(0,\infty)^{L+M}\times(0,1)^N} G(x,\boldsymbol{\tau}) \cdot H(x,\boldsymbol{\omega}) \cdot J(x,\boldsymbol{\eta}) \cdot Q(x) \cdot e^{-ax^2+bx} d\boldsymbol{\tau} d\boldsymbol{\omega} d\boldsymbol{\eta} \tag{8}$$

where

$$\begin{aligned}
G(x,\boldsymbol{\tau}) &= \prod_{\ell=1}^{L} \tau_\ell^{g_\ell(x)-1} e^{-\tau_\ell} \\
H(x,\boldsymbol{\omega}) &= \exp\left\{ -\sum_{m=1}^{M} h_m(x)^2 \omega_m + \gamma \sum_{m=1}^{M} h_m(x) \right\} \prod_{m=1}^{M} p_0(\omega_m) \\
J(x,\boldsymbol{\eta}) &= \prod_{n=1}^{N} \eta_n^{j_n(x)-1} (1-\eta_n)^{\beta_n - 1} \\
Q(x) &= x^{p-1} \prod_{m=1}^{M} h_m(x).
\end{aligned}$$

$p_0(\cdot)$ is the probability density of ERG(0).

Remark 5. *When $N \geq 1$ in the form (7), the third term $\left(\prod_{n=1}^{N} \frac{\Gamma(j_n(x))}{\Gamma(j_n(x)+\beta_n)}\right)$ can be absorbed into the first two terms. However, we still recommend using the Beta representation if possible. Otherwise the total number of auxiliary variables needed is increased by N.*

Remark 6. *We may generalize the form (7) in a few ways:*

1. *For multivariate x of dimension d, if g, h, j are all linear functions mapping from $\mathbb{R}^{+,d}$ to $\mathbb{R}^+$, then the data augmentation is still valid.*

2. *The x^p term can be replaced with a polynomial function of x, as long as it's always positive for $x > 0$.*

3. *If the posterior density $p(x)$ has extra factors which are not included in the form (7), the strategy still works as long as we can find integral representations for those extra factors.*

Although the posterior joint distribution in Equation (8) looks forbidding at the first glance, in many applications the linear functions g, h, j are simple enough, e.g. $h_m(x) = x$ for all m, and $N = 0$, which simplifies the expression a lot. More importantly, the conditional posteriors can be

derived easily. To derive that of τ_ℓ, for example,

$$p(\tau_\ell \mid x, \boldsymbol{\tau}_{-\ell}, \boldsymbol{\omega}, \boldsymbol{\eta}) = \tau_\ell^{g_\ell(x)-1} e^{-\tau_\ell} \cdot \frac{\left(\prod_{\ell' \neq \ell} \tau_{\ell'}^{g_{\ell'}(x)-1} e^{-\tau'_\ell} \cdot H(x, \boldsymbol{\omega}) \cdot J(x, \boldsymbol{\eta}) \cdot Q(x)\right)}{\int G(x, \boldsymbol{\tau}) \cdot H(x, \boldsymbol{\omega}) \cdot J(x, \boldsymbol{\eta}) \cdot Q(x) d\tau_\ell},$$

notice that it's proportional to $\tau_\ell^{g_\ell(x)-1} e^{-\tau_\ell}$. Therefore the conditional posterior distribution of τ_ℓ is $\Gamma(g_\ell(x), 1)$. Similarly for ω_m's and η_n's. For the conditional posterior of x, since g, h, j are all linear, it's proportional to $Q_1(x)e^{-Q_2(x)}$ where $Q_1(x)$ is a polynomial of degree $p + M - 1$ and $Q_2(x)$ is a quadratic function.

Before we summarize the conditional posteriors with respect to the augmented vector $(x, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta})$, the following definition of power truncated normal (PTN) distribution is useful.

Definition 3.1. The power truncated normal distribution, PTN(p, a, b), has density function

$$p(x) = C \cdot x^{p-1} e^{-ax^2+bx}, x > 0 \tag{9}$$

where $p, a > 0$ and $b \neq 0$.

Finally, the following proposition is helpful in developing the Gibbs sampler.

Proposition 2 (Gibbs Sampler)**.** If the joint probability density of the augmented vector $(x, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta})$ is proportional to $G(x, \boldsymbol{\tau}) \cdot H(x, \boldsymbol{\omega}) \cdot J(x, \boldsymbol{\eta}) \cdot Q(x)$ given in Proposition 1, then the conditional distributions are

$$\begin{aligned}
\tau_\ell \mid x, \boldsymbol{\tau}_{-\ell}, \boldsymbol{\omega}, \boldsymbol{\eta} &\sim \Gamma(g_\ell(x), 1), \quad \ell = 1, 2, ..., L \\
\omega_m \mid x, \boldsymbol{\tau}, \boldsymbol{\omega}_{-m}, \boldsymbol{\eta} &\sim \text{ERG}(h_m(x)), \quad m = 1, 2, ..., M \\
\eta_n \mid x, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta}_{-n} &\sim \text{Beta}(j_n(x), \beta_n), \quad n = 1, 2, ..., N \\
x \mid \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta} &\sim \sum_{m=0}^{M} \pi_k \cdot \text{PTN}(p + m, \tilde{a}, \tilde{b})
\end{aligned}$$

Here the conditional distribution of x is a finite discrete mixture of PTN distributions. $\{\pi_m\}_{m=0}^M$ are

proportional to the coefficients in the polynomial $Q(x)$.

$$\tilde{a} = a + \sum_{m=1}^{M} \omega_m \cdot (h'_m(0))^2$$
$$\tilde{b} = b + \left(\sum_{\ell=1}^{L} g'_\ell(0) \cdot \log \tau_\ell\right) + \left(\sum_{m=1}^{M} h'_m(0)(\gamma - 2\omega_m h_m(0))\right) + \left(\sum_{n=1}^{N} j'_n(0) \cdot \log \eta_n\right).$$

Furthermore, given x, all auxiliary random variables are conditionally independent. The sampling methods for ERG and PTN are given in Appendix B and C.

In many statistical applications, the number of auxiliary random variables grows linearly with the sample size and problem dimension, but the forms of g, h, j are relatively simple. Observing that $\tilde{a}$ and $\tilde{b}$ are both the sum of numerous terms, we may use normal variables to approximate them, by matching the moments, so that the sampling procedure for $(\boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta})$ can be skipped in Gibbs sampling. Appendix D gives the approximate Gibbs sampler when $g_\ell(x) = g(x)$ for all ℓ (similarly for h and j), and L, M, N are all large enough.

3.1 Expectation-Maximization Algorithm

By exploiting Proposition 1, MAP of the model can be found using an expectation-maximization algorithm. The complete-data log posterior is

$$\begin{aligned}
&\log p(x, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta}) \\
&= c_0 + \log G(x, \boldsymbol{\tau}) + \log H(x, \boldsymbol{\omega}) + \log J(x, \boldsymbol{\eta}) + \log Q(x) - ax^2 + bx \\
&= c_1 + \left(\sum_{\ell=1}^{L} g_\ell(x) \log \tau_\ell\right) + \left(\sum_{m=1}^{M} \gamma h_m(x) - h_m(x)^2 \omega_m\right) \\
&\quad + \left(\sum_{n=1}^{N} j_n(x) \log \eta_n\right) + \log Q(x) - ax^2 + bx
\end{aligned} \tag{10}$$

for some constants c_0, c_1 (with respect to x). In the t-th expectation step, we compute the expected value of $\log p(x, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta})$ under the current conditional posterior $p(\boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta} \mid x_{(t)})$, denoted as $C(x \mid x_{(t)})$. Then in the maximization step, $C(x \mid x^{(t)})$ is maximized as a function of x. We now derive the expectation and maximization steps.

- **The Expectation Step**

  From equation (10), notice that $\log p(x, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta})$ is linear in terms of $\log \tau_\ell, \omega_m$ and $\log \eta_n$. There-

fore, we replace them with their conditional expectations in the expectation step. Applying Proposition 2 yields

$$
\begin{aligned}
E\left(\log \tau_\ell \mid x_{(t)}\right) &= \psi(g_\ell(x_{(t)})), \quad \ell = 1, 2, ..., L \\
E\left(\omega_m \mid x_{(t)}\right) &= \frac{\psi(1 + h_m(x_{(t)})) + \gamma}{2h_m(x_{(t)})}, \quad m = 1, 2, ..., M \\
E\left(\log \eta_n \mid x_{(t)}\right) &= \psi(j_n(x_{(t)})) - \psi(j_n(x_{(t)}) + \beta_n), \quad n = 1, 2, ..., N.
\end{aligned}
$$

The derivation above uses the properties of gamma and beta distribution, as well as Theorem 2 which calculates the expectation of ERG(c) distribution. Finally, one can represent the function $C(x \mid x_{(t)})$ (up to a constant) as

$$
\begin{aligned}
C(x \mid x_{(t)}) = &\left(\sum_{\ell=1}^{L} g_\ell(x)\psi(g_\ell(x_{(t)}))\right) + \left(\sum_{m=1}^{M} \gamma h_m(x) - h_m(x)^2 \frac{\psi(1 + h_m(x_{(t)})) + \gamma}{2h_m(x_{(t)})}\right) \\
&+ \left(\sum_{n=1}^{N} j_n(x)\psi(j_n(x_{(t)})) - \psi(j_n(x_{(t)}) + \beta_n)\right) + \log Q(x) - ax^2 + bx
\end{aligned}
$$

Given the linearity of g, h, j functions, it can be further simplified as

$$
C(x \mid x_{(t)}) = \log Q(x) - \kappa_1 x^2 + \kappa_2 x \tag{11}
$$

for some constant $\kappa_1 > 0$ and $\kappa_2 \in \mathbb{R}$, depending on $x_{(t)}$.

- **The Maximization Step**

  Since $C(x \mid x_{(t)}) \to -\infty$ as $x \to \infty$, we conclude that $C(x \mid x_{(t)})$ as a function of x has a maximizer on $(0, \infty)$, which can be found numerically. Furthermore, when $h_m(0) = 0$ for all m, the unique maximizer has a closed form

$$
x^* := \arg\max_{x>0} C(x \mid x_{(t)}) = \frac{\kappa_2 + \sqrt{\kappa_2^2 + 8\kappa_1(p + M - 1)}}{4\kappa_1}. \tag{12}
$$

4 Examples

4.1 Inference for Gamma Shape

The gamma distribution, parameterized by shape α and rate β, is a component of many probability models. For instance, a gamma prior distribution for the precision parameter in Gaussian linear models is quite common. In fact, Normal-gamma distributions are workhorse models for shrinkage estimation in regression problems (Griffin and Brown, 2010). Gamma distribution is also widely used in modelling of extreme values, where it serves as the prior for the shape parameter of Pareto distribution (Arnold and Press (1989)) and helps to construct a quasi-conjugate prior for generalized Pareto distribution (Diebolt, El-Aroui, Garrido, and Girard (2005)). While a gamma prior distribution for a parameter is common, it is less common to model hyperparameters of the gamma distribution itself as random variables – particularly the shape parameter, α. Although posterior inference of the rate parameter is straightforward – since the gamma distribution itself is a conjugate prior for the rate parameter – posterior inference of the gamma shape parameter is a long-standing problem (Damsleth, 1975; Damien, Laud, and Smith, 1995; Rossell, 2009; Miller, 2018) and efficient posterior computation remains an open problem.

Damsleth (1975) discussed two conjugate priors for α, the gamma shape parameter. They are called GamCon distributions of type I and type II. Type I assumes the rate β is known, while the type II doesn't. In this section, we focus on the latter one and replicate Damsleth's example, showing how to utilize the data augmentation scheme and build algorithms for posterior inference.

Let's consider the following hierarchical model

$$\begin{aligned} x \mid \alpha, \beta &\sim \Gamma(\alpha, \beta), \\ \beta \mid \alpha &\sim \Gamma(\delta\alpha + 1, \delta\eta), \\ \alpha &\sim \xi_2(\eta/\mu, \delta). \end{aligned}$$

Here ξ_2 is GamCon distribution of type II. $\eta > \mu > 0, \delta > 0$. The probability density of $\xi_2(\mu, \delta)$ is

$$p(\alpha \mid \mu, \delta) = C_{\mu,\delta} \cdot \frac{\Gamma(\delta\alpha + 1)}{\Gamma(x)^\delta}(\delta\mu)^{-\delta\alpha},\ \alpha > 0, \mu > 1, \delta > 0.$$

Suppose observations $\boldsymbol{x} = (x_1, x_2, ..., x_n)'$ are independently and identically distributed gamma

random variables drawn from the above model. The likelihood is

$$f(\boldsymbol{x} \mid \alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i} \propto \frac{(x_g \beta)^{n\alpha}}{\Gamma(\alpha)^n} e^{-n x_g \beta}$$

where x_g is the geometrical mean, $x_g = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$.

The marginal posterior distribution of α is then calculated as

$$p(\alpha \mid \boldsymbol{x}) = \int_0^{\infty} f(\boldsymbol{x} \mid \alpha, \beta) p(\beta \mid \alpha) p(\alpha) d\beta$$

$$\text{where} \qquad p(\beta \mid \alpha) = \frac{(\delta\eta)^{\delta\alpha+1}}{\Gamma(\alpha\delta+1)} \beta^{\alpha\delta} e^{-\delta\eta\beta}, \text{ and} \qquad p(\alpha) = C \cdot \frac{\Gamma(\delta\alpha+1)}{\Gamma(\alpha)^{\delta}} (\delta\eta/\mu)^{-\delta\alpha}.$$

By construction, the marginal posterior of α given $\boldsymbol{x}$ also follows ξ_2, with updated parameters (η', μ', δ'). That is, $\alpha \mid \boldsymbol{x} \sim \xi_2(\eta'/\mu', \delta')$ and

$$\begin{aligned} \delta' &= \delta + n, \\ \eta' &= \frac{\delta}{\delta+n}\eta + \frac{n}{\delta+n} x_a, \\ \mu' &= \mu^{\frac{\delta}{\delta+n}} \cdot x_g^{\frac{n}{\delta+n}}. \end{aligned}$$

where x_a is the arithmetical mean, $x_a = \frac{1}{n}\sum_{i=1}^{n} x_i$. One can observe that η' is a weighted arithmetical mean of η and x_a with weights δ and n respectively and μ' is a weighted geometrical mean of μ and x_g, also with weights δ and n. Here x_a and x_g are two sufficient statistics. δ is viewed as the prior sample size while η and μ are the prior means. The problem of gamma shape inference is thus translated to that of ξ_2 distribution.

4.1.1 MCMC and EM for ξ_2 when $\delta \in \mathbb{N}^+$

We first develop the Gibbs sampler for a general GamCon distribution of type II, using data augmentation strategy. Suppose $x \sim \xi_2(x; \mu, \delta)$ and δ is a positive integer. Then the probability density can be rewritten in the form (7)

$$p(x \mid \mu, \delta) = C_{\mu,\delta} \cdot \Gamma(\delta x + 1) \left(\prod_{m=1}^{M} \frac{1}{\Gamma(x)} \right) e^{-\delta \log(\delta\mu) x} \tag{13}$$

Hence $L = 1$ with $g_1(x) = \delta x + 1$, $M = \delta$ with $h_m(x) = x$ and $N = 0$. $p = 1, a = 0$ and $b = -\delta \log(\delta\mu)$.

After introducing the auxiliary variables τ and $\omega_1, \omega_2, ..., \omega_\delta$, Proposition 2 then immediately gives the conditional posteriors:

$$\begin{aligned} \tau \mid x, \boldsymbol{\omega} &\sim \Gamma(\delta x + 1, 1) \\ \omega_1, \omega_2, ..., \omega_\delta \mid x, \tau &\overset{i.i.d.}{\sim} \text{ERG}(x) \\ x \mid \tau, \boldsymbol{\omega} &\sim \text{PTN}\left(\delta + 1, \sum_{i=1}^{\delta} \omega_i, \delta\left(\gamma - \log(\delta\mu/\tau)\right)\right) \end{aligned}$$

4.1.2 Damsleth Examples with Non-Informative Prior

Here we replicate Damsleth's example of no prior information with sample size $n = 5, 10, 30$. Putting $\delta = 0$, the posterior parameters are

$$\delta' = n, \quad \eta' = x_a, \quad \mu' = x_g.$$

The resulted posterior of gamma shape is thus $\alpha \mid \boldsymbol{x} \sim \xi_2(x_a/x_g, n)$. Instead of actually generating Gamma random variables, we directly use the sufficient statistics (x_a, x_g), given in Table 2 of Damsleth (1975). The true value of α is 5. The histogram of 5000 posterior samples for each case are shown in Figure 1. The colored solid lines denote the true posterior density and the black dashed line denotes the true α. We see that the posterior samples are indeed sampled from the target distributions. Figure 2 shows the sampling trace plots. In Table 3, the sample moments are compared with their theoretical counterparts calculated by numerical integration. When $n = 30$, the deviations from the theoretical values are $-1.0\%, -4.6\%, -8.2\%$ and 0.1% for mean, variance, skewness and kurtosis respectively.

Table 3: Posterior Moments

n	x_a	x_g	Method	Mean	Variance	Skewness	Kurtosis
5	7.19	6.05	Numerical Integration	4.768	5.399	0.997	4.494
			Posterior Sampling	4.779	5.630	1.305	6.148
10	5.57	5.01	Numerical Integration	6.271	5.780	0.783	3.921
			Posterior Sampling	6.101	4.857	0.634	3.427
30	5.09	4.26	Numerical Integration	3.252	0.585	0.490	3.361
			Posterior Sampling	3.250	0.605	0.537	3.155

To find the posterior mode of α, we exploit the EM algorithm developed in Section 3.1. Spe-

Figure 1: Histogram of Posterior Samples

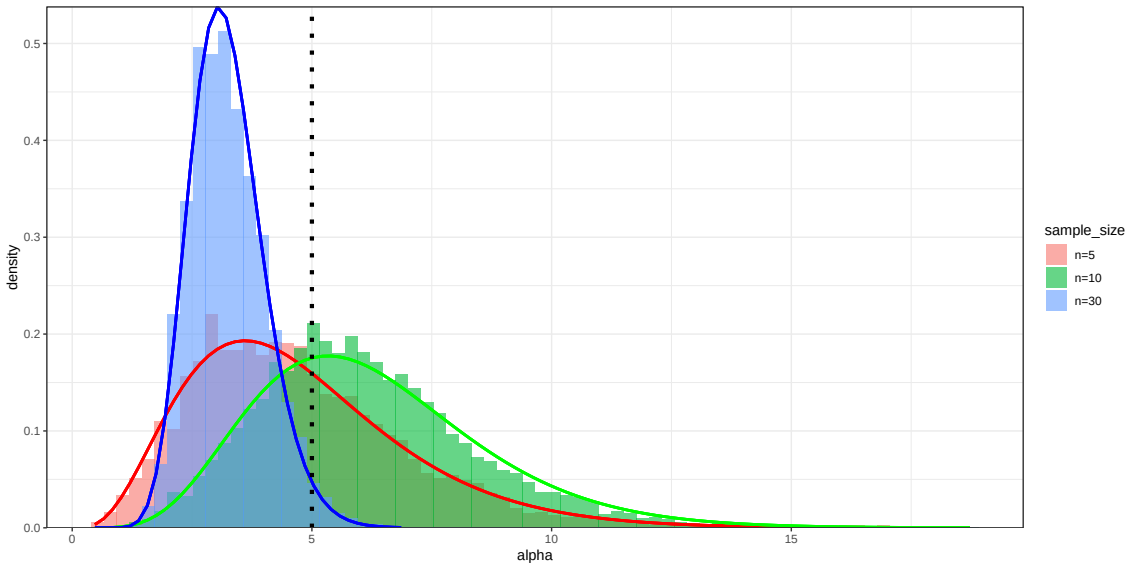
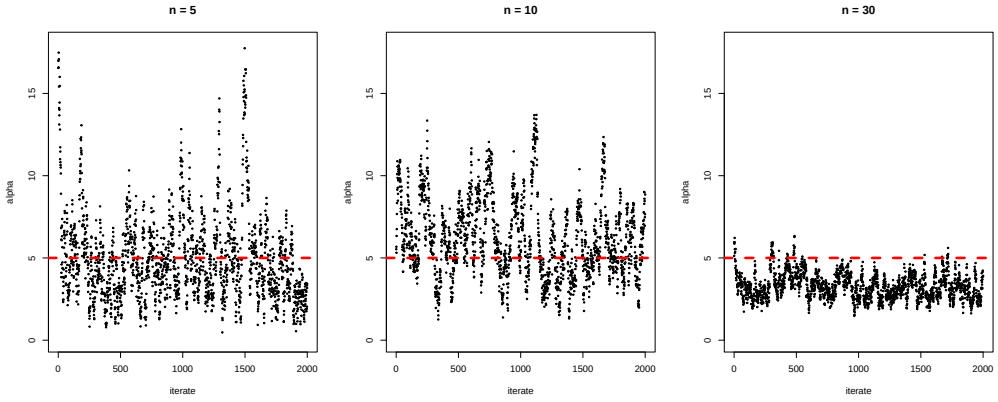


Figure 2: Trace Plot of Posterior Samples

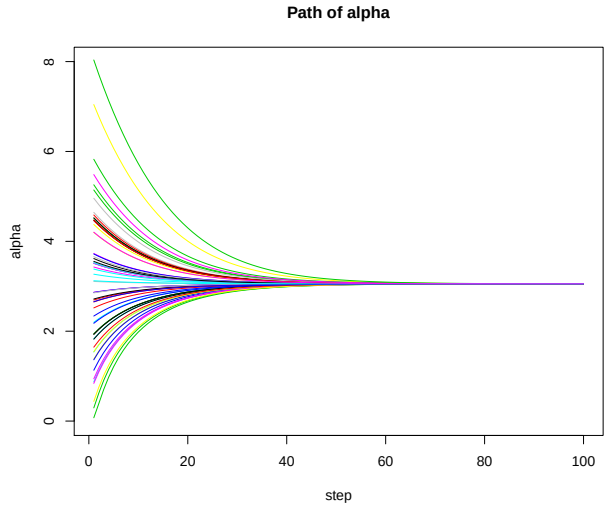


cially, in the expectation step, the conditional expected value of complete-data log posterior (at t-th iteration) given by Equation (11) now has the parameters

$$\begin{aligned}\log Q(x) &:= \delta' \log \alpha \\ \kappa_1 &:= \frac{\delta' \left(\gamma + \psi(\alpha^{(t)} + 1)\right)}{2\alpha^{(t)}} > 0 \\ \kappa_2 &:= \delta' \left(\gamma - \log \delta' \eta' / \mu' + \psi(\delta' \alpha^{(t)} + 1)\right)\end{aligned}$$

Then in the maximization step, α is updated by Equation (12), with $p - M + 1 = \delta'$. In Figure 3, we start with 30 different initial values in EM algorithm and show the optimizing paths for the case

Figure 3: Optimizing Paths of EM Algorithm



$n = 30$. In this particular case, the algorithm converges after around 50 steps. The 30 numerical solutions given by EM has mean 3.054, which matches with the maximizer found by Mathematica. And the standard deviation is as small as 0.0003.

4.2 Negative Binomial Regression

Next, we proceed to how our data augmentation strategy can be used to fit negative binomial regression models. The count data $\{y_i\}_{i=1}^n$ are assumed to follow the negative binomial distribution

$$y_i \mid r, p_i \sim \text{NB}(r, p_i), \quad p_i = \frac{1}{1 + e^{-\boldsymbol{x_i'}\boldsymbol{\beta}}}$$

where $\{\boldsymbol{x}_i\}_{i=1}^n$ are observed covariates and $\boldsymbol{\beta}$ is the regression coefficients. r is interpreted as the number of failures until the experiment is stopped, while the success probability p_i is related to $\boldsymbol{x_i'}\boldsymbol{\beta}$ via the logistic transformation. This model specification is equivalent to the following Gamma-Poisson mixture,

$$\begin{aligned} y_i \mid \lambda_i &\sim \text{Poisson}(\lambda_i), \\ \lambda_i \mid r, \boldsymbol{x_i'}\boldsymbol{\beta} &\sim \Gamma\left(r, e^{-\boldsymbol{x_i'}\boldsymbol{\beta}}\right). \end{aligned}$$

The likelihood is

$$f(y_i \mid \boldsymbol{x_i}, \boldsymbol{\beta}, r) = \frac{\Gamma(y_i + r)}{y_i! \cdot \Gamma(r)} \left(\frac{e^{\boldsymbol{x_i'\beta}}}{e^{\boldsymbol{x_i'\beta}} + 1}\right)^{y_i} \left(\frac{1}{e^{\boldsymbol{x_i'\beta}} + 1}\right)^{r}.$$

Pillow and Scott (2012) adopt exactly the same model with known r and use a data augmentation strategy for the inference on β. Our example here extends their method and allows for inference on both r and β. The parameter r controls the dispersion of observations, as the expectation of y_i is $re^{\boldsymbol{x_i'\beta}}$ and variance is $re^{\boldsymbol{x_i'\beta}}\left(1 + e^{\boldsymbol{x_i'\beta}}\right)$.

Let the prior for $\boldsymbol{\beta}$ and r be $N(\mathbf{0}, \Sigma)$ and $p(r)$. We then calculate the joint posterior of $(\boldsymbol{\beta}, r)$ as

$$\begin{aligned} p(\boldsymbol{\beta}, r \mid \boldsymbol{X}, y) &= C \cdot p(r) \exp\left(-\frac{1}{2}\boldsymbol{\beta'}\Sigma^{-1}\boldsymbol{\beta}\right) \prod_{i=1}^{n} \frac{\Gamma(y_i + r)}{\Gamma(r)} \left(\frac{e^{\boldsymbol{x_i'\beta}}}{e^{\boldsymbol{x_i'\beta}} + 1}\right)^{y_i} \left(\frac{1}{e^{\boldsymbol{x_i'\beta}} + 1}\right)^{r} \\ &= C \cdot p(r) \exp\left(-\frac{1}{2}\boldsymbol{\beta'}\Sigma^{-1}\boldsymbol{\beta}\right) \left(\frac{1}{\Gamma(r)}\right)^{n} \left(\prod_{i=1}^{n} \Gamma(y_i + r)\right) \left(\prod_{i=1}^{n} \frac{\left(e^{\boldsymbol{x_i'\beta}}\right)^{y_i}}{\left(e^{\boldsymbol{x_i'\beta}} + 1\right)^{r+y_i}}\right) \end{aligned}$$

Notice that if we assign gamma prior for r, then its conditional posterior density is close to the form (7), except for the extra factor $\prod_{i=1}^{n} \left(e^{\boldsymbol{x_i'\beta}}\right)^{y_i} / \left(e^{\boldsymbol{x_i'\beta}} + 1\right)^{r+y_i}$. Similarly for β, the posterior is close to normal density, except for the same factor. This is however not an issue as we can write this factor as a scale mixture of normals, where the mixing distribution is Pólya-Gamma from Polson, Scott, and Windle (2013). Setting $a = y_i$ and $b = r + y_i$, the key mixture representation in Equation (6) related to it now becomes

$$\begin{aligned} \frac{\left(e^{\boldsymbol{x_i'\beta}}\right)^{y_i}}{\left(e^{\boldsymbol{x_i'\beta}} + 1\right)^{r+y_i}} &\propto \int_0^{\infty} f(r, \boldsymbol{\beta} \mid \xi) \cdot p_{\text{PG}}(\xi_i \mid r + y_i, 0) d\xi_i \\ f(r, \boldsymbol{\beta} \mid \xi) &= \exp\left[\frac{1}{2}\left(-r(2\log 2 + \boldsymbol{x_i'\beta}) + y_i(\boldsymbol{x_i'\beta}) - \xi_i(\boldsymbol{x_i'\beta})^2\right)\right]. \end{aligned}$$

The integrand $f(r, \boldsymbol{\beta} \mid \xi)$ is an exponential density when viewed as a function of r, and a normal density when viewed as a function of $\boldsymbol{\beta}$.

Finally, we introduce the auxiliary variables $\boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\xi}$ which follow gamma, ERG and PG distribution respectively. Let $\Xi := \text{diag}(\xi_1, ..., \xi_n)$ and $\boldsymbol{z} := \left(\frac{y_1 - r}{2\xi_1}, ..., \frac{y_n - r}{2\xi_n}\right)'$ and $p(r) \sim \Gamma(a_0, b_0)$. The

conditional posteriors are derived as follows:

$$
\begin{aligned}
\tau_i \mid r, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{X}, \boldsymbol{y} &\sim \Gamma(y_i + r, 1) \\
\omega_i \mid r, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{X}, \boldsymbol{y} &\overset{i.i.d.}{\sim} \text{ERG}(r) \\
\xi_i \mid r, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{X}, \boldsymbol{y} &\sim \text{PG}(r + y_i, \boldsymbol{x_i'}\boldsymbol{\beta}) \\
\boldsymbol{\beta} \mid r, , \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{X}, \boldsymbol{y} &\sim N(\boldsymbol{m}, V) \\
r \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{X}, \boldsymbol{y} &\sim \text{PTN}(a_0 + n, a, b + b_0)
\end{aligned}
$$

where

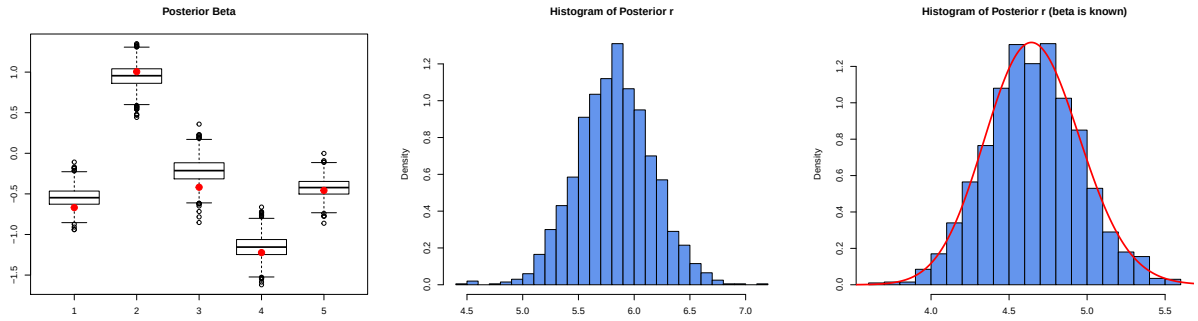
$$
\begin{aligned}
a &= \sum_{i=1}^{n} \omega_i \\
b &= n(\gamma - \log 2) + \sum_{i=1}^{n} \left(\log \tau_i - \boldsymbol{x_i'}\boldsymbol{\beta}/2 \right) \\
V &= \left(\Sigma^{-1} + X' \Xi X \right)^{-1} \\
m &= V X' \Omega z
\end{aligned}
$$

Figure 4 shows an illustrating simulation example where we set true $r = 5$ and generate $n = 100$ count observations. We draw the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^5$ and covariates $x_{ij} \overset{i.i.d.}{\sim} N(0, 0.5^2)$ for $1 \leq i \leq n$ and $1 \leq j \leq 5$. The prior for r is $p(r) \sim 1/r$ which is the limit case of gamma prior with $a_0 = 0, b_0 = 0$. For β, we choose $\Sigma = 10^6 I_5$ so that the prior information is relatively weak. The left panel shows the boxplots of posterior β samples, with those red dots denote the true β's. As the figure shows, all 5 β's fall in the 95% credible interval. The middle panel is the histogram of the posterior samples for r. Since the actual marginal posterior of r is hard to compute given the complicated form of the joint $p(\boldsymbol{\beta}, r \mid \boldsymbol{X}, y)$, in the right panel we use the same dataset, plug in the true β's, and run the Gibbs sampler again, so that we are able to compare the resulted histogram with the true density (red line). They are quite close to each other, indicating that the sampling procedure works well for this example.

4.3 Multinomial-Dirichlet Model

In this section, we develop the Markov chain Monte Carlo (MCMC) algorithm for fully posterior inference of the concentration parameter vector in the Dirichlet distribution. Such inference

Figure 4: Negative Binomial Regression Results



problems commonly arise in applied analyses of categorical data. Section 4.3.1 presents the general hierarchical multinomial-Dirichlet model class for which the ERG data augmentation scheme may be utilized. Section 4.3.2 develops a Gibbs sampler for inferring the concentration parameter in the Dirichlet distribution and conducts a simulation study comparing the performance of our data augmentation strategy with Metropolis-Hasting algorithm.

4.3.1 A Hierarchical Multinomial-Dirichlet Model

The multinomial-Dirichlet framework presented herein is closely related to the latent Dirichlet allocation model of Blei, Ng, and Jordan (2003) for topic modeling of text data, and we use text analysis as a motivating context. Suppose that for document $s \in \{1, ..., S\}$, each of N_s words in the document is independently allocated to K topics conditional on probability vector $\boldsymbol{p}_s = (p_{s1}, p_{s2}, \ldots, p_{sK})$. For each document s, the number of words allocated to each topic is denoted by $\boldsymbol{n}_s = (n_{s1}, \cdots, n_{sK})$, which follows a multinomial distribution. The sampling model for the count vector $\boldsymbol{n}_s$ is then a multinomial distribution given probability vector $\boldsymbol{p}_s$,

$$\boldsymbol{n}_s \mid N_s, \boldsymbol{p}_s \sim \text{Multinomial}\left(N_s, \boldsymbol{p}_s\right).$$

The probability vector $\boldsymbol{p}_s$ is the proportional allocation of each document to the K topics. In a Bayesian analysis, the probability vector for each document $\boldsymbol{p}_s$ is typically assigned a Dirichlet distribution with concentration parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$,

$$\boldsymbol{p}_s \mid \boldsymbol{\alpha} \sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right).$$

Rather than fixing $\boldsymbol{\alpha} = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$, as is common, we complete the model with a prior distribution $p(\boldsymbol{\alpha})$. This hierarchical prior distribution for $\boldsymbol{\alpha}$ facilitates more efficient information sharing across documents (observational units), and it yields practical advantages for out-of-sample prediction, which we discuss below. The model framework and ERG augmentation admit independent uniform, truncated normal, and exponential prior distributions for the elements α_k. Section 4.3.2 presents analyses based on independent gamma priors $p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} p(\alpha_k)$.

In application, model inferences are often summarized by the posterior predictive distribution for the topic proportion vector $\boldsymbol{p}^*$ in a new document. Computing the posterior predictive distribution $p\left(\boldsymbol{p}^* \mid \boldsymbol{n}_1, ..., \boldsymbol{n}_S\right) = \int_{\boldsymbol{\alpha}} p\left(\boldsymbol{p}^* \mid \boldsymbol{\alpha}\right) p(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, ..., \boldsymbol{n}_S) d\boldsymbol{\alpha}$ requires posterior computation of $p\left(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_S\right) \propto p\left(\boldsymbol{\alpha}\right) \prod_{s=1}^{S} p\left(\boldsymbol{n}_s \mid \boldsymbol{\alpha}\right)$; however, when the probability vectors $\boldsymbol{p}_s$ are integrated out of the multinomial likelihood, the marginal likelihood $p\left(\boldsymbol{n}_s \mid \boldsymbol{\alpha}\right)$ includes elements of $\boldsymbol{\alpha}$ inside the gamma function,

$$\prod_{s=1}^{S} p\left(\boldsymbol{n}_s \mid \boldsymbol{\alpha}\right) = \prod_{s=1}^{S} \left(\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} \left(n_{sk} + \alpha_k\right)\right)} \prod_{k=1}^{K} \frac{\Gamma\left(n_{sk} + \alpha_k\right)}{\Gamma\left(\alpha_k\right)} \right).$$

Because $\boldsymbol{\alpha}$ is nested inside the gamma function, computing $p(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_S)$ is a challenge. Previous inference strategies relied on approximations, but in Section 4.3.2 we introduce a new data augmentation scheme for computing the full posterior $p\left(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_S\right)$.

4.3.2 Data Augmentation and Simulation Study

Assume independent gamma prior distributions for each element of vector $\boldsymbol{\alpha}$ so that $p\left(\boldsymbol{\alpha}\right) \propto \prod_{k=1}^{K} \alpha_k^{a_0 - 1} e^{-b_0 \alpha_k}$, with hyperparamter a_0 and b_0. Note that gamma priors on each α_k give closed-form full conditional distributions in a Gibbs sampler, which is shown below. When $\alpha_k \sim \Gamma(a_0, b_0)$, where a_0 denotes the shape parameter and b_0 the rate, the expectation $E[\alpha_k] = a_0/b_0$. We can set $E[\alpha_k] = 1/K$, a standard choice for the Dirichlet concentration parameter, by choosing $a_0 = b_0/K$. The prior variance then depends on both the dimension of the Dirichlet distribution, K, and the rate parameter, b_0.

We now reorganize the multivariate posterior density of $\boldsymbol{\alpha}$ as

$$p(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_S) = \left(\prod_{k=1}^{K} f(\alpha_k) \right) \left(\prod_{s=1}^{S} \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k + N_s)} \right) \tag{14}$$

where

$$f(\alpha_k) = \left(\prod_{s=1}^{S} \Gamma(\alpha_k + n_{sk})\right)\left(\prod_{s=1}^{S} \frac{1}{\Gamma(\alpha_k)}\right) \alpha_k^{a_0-1} e^{-b_0\alpha_k}. \tag{15}$$

Note that for each α_k, Equation (15) is exactly of the form (7). And the extra factor, $\prod_{s=1}^{S} \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k + N_s)}$, can be replaced with a beta integral representation. A multivariate version of Proposition 2 produces the conditional posteriors as follows

$$\begin{aligned}
\tau_{sk} \mid \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\eta} &\sim \Gamma(\alpha_k + n_{sk}), \\
\omega_{sk} \mid \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\eta} &\sim \text{ERG}(\alpha_k), \\
\eta_s \mid \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\omega} &\sim \text{Beta}\left(\sum_{k=1}^{K} \alpha_k, N_s\right), \\
\alpha_k \mid \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta} &\sim \text{PTN}(S + a_0, a_k, b_k)
\end{aligned}$$

where $a_k = \sum_{s=1}^{S} \omega_{sk}, b_k = S\gamma - b_0 + \sum_{s=1}^{S} \log(\tau_{sk} \cdot \eta_s)$. Conditioned on the introduced auxiliary variables $(\boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta})$, all elements of the vector $\boldsymbol{\alpha}$ are now mutually independent, which significantly reduces the difficulty of sampling procedure as we can now sample separately from K univariate distributions.

The total number of auxiliary variables to be sampled at each iterate is $S(1 + 2K)$, which may greatly slow down the MCMC algorithm for large values of S and K. However, we observe that these variables affect the distribution of α_k only through the parameters (a_k, b_k). The summation form of (a_k, b_k) suggests that we may apply central limit theorem and approximate them by normal variables. Therefore, as mentioned in Section 3, the Gibbs sampling algorithm can be approximately simplified to

1. Initialize $a_k^{(0)}, b_k^{(0)}$ for $1 \leq k \leq K$.

2. At step t, sample $a_k^{(t)}$ and $b_k^{(t)}$ from $N(\mu_{a,k}, \sigma_{a,k}^2)$ and $N(\mu_{b,k}, \sigma_{b,k}^2)$ respectively, where

$$\begin{aligned}
\mu_{a,k} &= \frac{S}{2\alpha_k}\left(\psi(1+\alpha_k)+\gamma\right) \\
\sigma_{a,k}^2 &= \frac{\mu_{a,k}}{2\alpha_k^2} - \frac{S}{4\alpha_k^2}\psi'(1+\alpha_k) \\
\mu_{b,k} &= S\gamma - b_0 + S\psi(\alpha_0) + \sum_{s=1}^{S}\psi(\alpha_k+n_{sk}) - \psi(\alpha_0+N_s) \\
\sigma_{b,k}^2 &= S\psi'(\alpha_0) + \sum_{s=1}^{S}\psi'(\alpha_k+n_{sk}) - \psi'(\alpha_0+N_s)
\end{aligned}$$

   and $\alpha_0 = \sum_{k=1}^{K}\alpha_k$.

3. Sample $\alpha_k^{(t)}$ from PTN$(S+a_0, a_k^{(t)}, b_k^{(t)})$ for $1 \le k \le K$. Increase t by 1 and return to (2).

Fixing the dataset dimensions S and K, we consider two settings of the true $\boldsymbol{\alpha}$ for our simulation experiment: (A) heterogeneous $\alpha_k = k/K$ and (B) homogeneous $\alpha_k = 1/K$ for $k = 1, 2, ..., K$. Probability vector $\boldsymbol{p}_s$ are drawn independently from Dirichlet$(\boldsymbol{\alpha})$ and vector of counts $\boldsymbol{n}_s$ are drawn from Mutinomial $(N_s, \boldsymbol{p}_s)$ with $N_s = 500$ for $s = 1, 2, ..., S$. We set the hyperparamters $a_0 = b_0/K$ in independent gamma priors of α_k's, so that the prior expectations are all equal to $1/K$. The setting of our simulation experiment is summarized below:

- $S \in \{100, 1000\}, K \in \{10, 50\}$ and $b_0 \in \{0.1, 1, 5\}$.
- 4 MCMC algorithms in comparison:
  - DA: Gibbs sampler which iteratively samples $\boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta}$ and $\boldsymbol{\alpha}$.
  - DA-N: replace the parameters in the conditional posterior of $\boldsymbol{\alpha}$ with approximating normal variables and skip the sampling of $\boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta}$ in DA.
  - DA-E: replace the parameters in the conditional posterior of $\boldsymbol{\alpha}$ with corresponding expectations and skip the sampling of $\boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\eta}$ in DA.
  - MH: random-walk Metropolis-Hasting sampler.
- Metrics:

– Root Mean Square Error:

$$\text{RMSE} = \sqrt{\frac{1}{TK}\sum_{t=1}^{T}\|\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}\|^2}.$$

$T = 500$ is the posterior sample size.

– Effective sample size ratio (ESSR):

$$\text{ESSR} = \frac{1}{K}\sum_{k=1}^{K}\frac{\lambda_k^2}{\sigma_k^2},$$

which measures the serial correlation between posterior samples. λ^2 is the sample variance and σ^2 is the estimate of the spectral density at frequency zero.

– Algorithm running time in seconds

In simulation experiments, we initialize each α_k with a random draw from Lognormal$(0, 1/K)$ and multiply by $1/K$, so that the prior median is $1/K$. For each combination of (b_0, S, K), we run the simulation for 50 times (the first 200 samples are dropped each time). Table 4 and 5 show the results, averaged over 50 runs, for the homogeneous and heterogeneous setting respectively. Despite being slow, our data augmentation strategy with ERG auxiliary variables produces more accurate estimates of $\boldsymbol{\alpha}$ than Metropolis-Hasting does, in heterogeneous setting. While in homogeneous setting, the two methods have similar RMSE. Metropolis-Hasting gets significantly worse when K grows to 50. This is not surprising at all. As we notice that, the effective sample size ratios for Metropolis-Hasting are as low as 0.01, indicating strong serial correlations in posterior samples drawn by Metropolis-Hasting, of which the acceptance rate is around 0.02. Therefore, the RMSE for Metropolis-Hasting is entirely up to the distance between the initial $\boldsymbol{\alpha}_{(0)}$ and the true $\boldsymbol{\alpha}$. Our Gibbs sampler instead enjoys the advantage of being less sensitive to initialization and not requiring further tuning. The different choices of b_0 don't seem to affect the posteriors too much. Once replacing auxiliary variables with normal approximations or expectations, the Gibbs sampler gets much faster. Meanwhile, root mean squared errors and effective sample sizes get slightly improved. With DA-N and DA-E, we can expect the resulted posterior samples to have slightly less variation as well as less correlation.

Table 4: Homogeneous Setting

		$b_0 = 0.1$				$b_0 = 1$				$b_0 = 5$			
		Root Mean Square Error ($\times 10^3$)											
S	K	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH
100	10	20.93	20.95	18.95	20.56	20.25	20.36	18.31	19.99	20.22	20.25	18.38	20.19
100	50	7.78	8.03	6.93	5.88	7.71	7.95	6.86	6.03	7.66	7.95	6.84	5.95
1000	10	6.54	6.53	5.95	7.13	6.58	6.54	6.01	7.26	6.38	6.36	5.80	6.87
1000	50	2.49	2.51	2.19	2.48	2.49	2.51	2.20	2.50	2.48	2.50	2.18	2.46
		Effective Sample Size Ratio											
S	K	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH
100	10	0.32	0.33	0.32	0.04	0.32	0.32	0.32	0.04	0.32	0.32	0.32	0.04
100	50	0.08	0.08	0.07	0.01	0.08	0.08	0.07	0.01	0.08	0.08	0.07	0.01
1000	10	0.32	0.32	0.32	0.01	0.33	0.33	0.32	0.01	0.32	0.32	0.32	0.01
1000	50	0.08	0.08	0.07	0.01	0.08	0.07	0.07	0.01	0.08	0.08	0.07	0.01
		Running Time											
S	K	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH
100	10	0.62	0.03	0.02	0.04	0.60	0.03	0.02	0.04	0.63	0.03	0.02	0.04
100	50	3.16	0.07	0.03	0.05	3.01	0.07	0.03	0.05	2.99	0.07	0.03	0.05
1000	10	4.01	0.10	0.03	0.17	3.76	0.09	0.03	0.17	3.74	0.09	0.03	0.17
1000	50	19.85	0.53	0.11	0.46	18.47	0.52	0.11	0.44	18.50	0.52	0.11	0.44

5 Discussion

The class of Exponential Reciprocal Gamma (ERG) distributions are developed as mixing distributions for models with Gamma functions. This adds to the literature on normal variance-mean mixtures by showing that they extend to a wide class of applications. Our ensuing data augmentation strategy facilitates full posterior inference for parameters in models which were hitherto hard to provide inference. The algorithms are scalable and are a fast efficient simulation method for drawing from posterior distributions with applications to many area, such as non-parametric Bayes, latent Dirichlet allocation, Gamma-Gamma hierarchical models, extreme value models, and many other Bayesian mixture models.

The focus of our paper is on theoretical and algorithmic development of ERG auxiliary variables. Our work contributes to the literature on scale mixtures of normals (see, e.g., (Andrews and Mallows, 1974; West, 1987; Polson, Scott, and Windle, 2013)). We believe that the computational strategies developed here will provide the foundation for new and richly structured hierarchical gamma models. Applied Bayesian analyses of categorical data will benefit from increased model flexibility and information borrowing strategies.

Table 5: Heterogeneous Setting

		$b_0 = 0.1$				$b_0 = 1$				$b_0 = 5$			
		Root Mean Square Error ($\times 10^3$)											
S	K	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH
100	10	82.76	82.25	76.54	82.34	82.56	81.90	77.33	83.49	81.95	82.11	78.46	86.96
100	50	81.50	81.41	75.83	553.05	78.53	78.39	73.47	553.83	81.99	81.98	79.15	553.38
1000	10	26.11	26.00	24.50	41.00	25.76	25.61	24.18	41.08	25.88	25.71	24.25	41.39
1000	50	24.80	24.79	23.19	553.35	25.22	25.11	23.66	553.30	25.26	25.21	23.71	552.86
		Effective Sample Size Ratio											
S	K	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH
100	10	0.47	0.49	0.50	0.03	0.48	0.51	0.49	0.03	0.49	0.51	0.51	0.03
100	50	0.44	0.45	0.44	0.01	0.44	0.45	0.44	0.01	0.45	0.46	0.46	0.01
1000	10	0.48	0.50	0.50	0.02	0.47	0.50	0.50	0.01	0.47	0.50	0.50	0.01
1000	50	0.44	0.44	0.44	0.01	0.44	0.45	0.45	0.01	0.44	0.44	0.45	0.01
		Running Time											
S	K	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH	DA	DA-N	DA-E	MH
100	10	0.74	0.09	0.08	0.10	0.68	0.09	0.08	0.10	0.68	0.09	0.08	0.10
100	50	3.23	0.13	0.10	0.58	3.07	0.13	0.09	0.58	3.08	0.13	0.10	0.58
1000	10	3.92	0.09	0.05	0.20	3.77	0.09	0.05	0.20	3.76	0.09	0.05	0.19
1000	50	19.44	0.42	0.13	0.97	18.53	0.41	0.13	0.95	18.54	0.41	0.13	0.95

There are a number of avenues for future research. In particular, regularized scale allocation models can be implemented using data augmentation methods of Polson and Scott (2013) with ERG distribution. Barndorff-Nielsen, Blaesild, and Seshadri (1992) provide multivariate GIG distribution theory and relationships with Poisson processes.

References

Aktekin, T., N. Polson, and R. Soyer (2018). Sequential bayesian analysis of multivariate count data. *Bayesian Analysis 13*(2), 385–409.

Andrews, D. F. and C. L. Mallows (1974). Scale mixtures of Normal distributions. *Journal of the Royal Statistical Society, B*, 99–102.

Arnold, B. C. and S. J. Press (1989). Bayesian estimation and prediction for pareto data. *Journal of the American Statistical Association 84*(408), 1079–1084.

Barndorff-Nielsen, O., P. Blaesild, and V. Seshadri (1992). Multivariate distributions with general-

ized inverse gaussian marginals, and associated poisson mixtures. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 109–120.

Barndorff-Nielsen, O., J. Kent, and M. Sørensen (1982). Normal variance-mean mixtures and Z distributions. *International Statistical Review/Revue Internationale de Statistique*, 145–159.

Barndorff-Nielsen, O. E. and N. Shephard (2012). *Basics of Lévy processes*. Citeseer.

Bhattacharya, A., A. Chakraborty, and B. K. Mallick (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bondesson, L. (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability 14*(4), 855–869.

Bondesson, L. (1992). Generalized Gamma Convolutions and related classes of distributions and densities. *Lecture Notes in Statistics 76*.

Damien, P., P. W. Laud, and A. F. Smith (1995). Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *Journal of the Royal Statistical Society B*, 547–563.

Damsleth, E. (1975). Conjugate classes for Gamma distributions. *Scandinavian Journal of Statistics*, 80–84.

Diebolt, J., M.-A. El-Aroui, M. Garrido, and S. Girard (2005). Quasi-conjugate bayes estimates for gpd parameters and application to heavy tails modelling. *Extremes 8*(1-2), 57–78.

Glynn, C., S. T. Tokdar, B. Howard, and D. L. Banks (2019, 03). Bayesian analysis of dynamic linear topic models. *Bayesian Anal. 14*(1), 53–80.

Griffin, J. E. and P. J. Brown (2010, 03). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal. 5*(1), 171–188.

Hartman, P. (1976). Completely monotone families of solutions of n-th order linear differential equations and infinitely divisible distributions. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze 3*(2), 267–287.

Holmes, C. C. and L. Held (2006, 03). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*, 145–168.

Lijoi, A., P. Muliere, I. Prünster, and F. Taddei (2016). Innovation, growth and aggregate volatility from a bayesian nonparametric perspective. *Electronic Journal of Statistics 10*(2), 2179–2203.

Linderman, S., M. Johnson, and R. P. Adams (2015). Dependent multinomial models made easy: Stick-breaking with the pólya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pp. 3456–3464.

Mallick, B. K., D. Ghosh, and M. Ghosh (2005). Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 219–234.

McLeish, D. (2014). Simulating random variables using moment-generating functions and the saddlepoint approximation. *Journal of Statistical Computation and Simulation 84*(2), 324–334.

Miller, J. W. (2018). Fast and accurate approximation of the full conditional for Gamma shape parameters. *arXiv:1802.01610*.

Minka, T. (2000). Estimating a Dirichlet distribution. *Technical report, MIT*.

Pillow, J. and J. Scott (2012). Fully bayesian inference for neural models with negative-binomial spiking. *Advances in neural information processing systems 25*, 1898–1906.

Polson, N. G. and J. G. Scott (2013). Data augmentation for Non-Gaussian regression models using variance-mean mixtures. *Biometrika 100*(2), 459–471.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

Polson, N. G. and S. L. Scott (2011). Data augmentation for Support Vector Machines. *Bayesian Analysis 6*(1), 1–23.

Rosiński, J. (2001). Series representations of lévy processes from the perspective of point processes. In *Lévy processes*, pp. 401–415. Springer.

Rossell, D. (2009). GaGa: a parsimonious and flexible model for differential expression analysis. *The Annals of Applied Statistics 3*(3), 1035–1051.

Roynette, B. and M. Yor (2005). Couples de Wald indéfiniment divisibles. Exemples liés à la fonction gamma d'Euler et à la fonction zeta de Riemann. *Annales de l'institut Fourier 55*(4), 1219–1284.

West, M. (1987). On scale mixtures of Normal distributions. *Biometrika 74*(3), 646–648.

Windle, J., N. G. Polson, and J. G. Scott (2014). Sampling Polya-Gamma random variates: alternate and approximate techniques. *arXiv:1405.0506*.

Zhang, Z., S. Wang, D. Liu, M. I. Jordan, and N. Lawrence (2012). Ep-gig priors and applications in bayesian sparse learning. *Journal of Machine Learning Research 13*(6).

Zhou, M., L. Hannah, D. Dunson, and L. Carin (2012). Beta-negative binomial process and poisson factor analysis. In *Artificial Intelligence and Statistics*.

A Proof of Theorem 1

The generalized inverse Gaussian distribution, GIG(p, a, b), has probability density function

$$p(x) \propto x^{p-1} \exp\left\{-\frac{1}{2}(ax + b/x)\right\}, \quad a, b, x > 0, p \in \mathbb{R}.$$

It suffices to show that if $G_k \sim GIG\left(-\frac{3}{2}, 2c^2, \frac{1}{2k^2}\right)$, then the following integral identity holds,

$$E(e^{-s^2 G_k}) = \left(\frac{k + \sqrt{s^2 + c^2}}{k + c}\right) e^{-\frac{\sqrt{s^2+c^2}-c}{k}}.$$

The density of G_k given by

$$p_{k,c}(x) = m(k, c)\, x^{-\frac{5}{2}} \exp\left(-\frac{1}{4k^2 x} - c^2 x\right).$$

with normalizing constant,

$$m(k, c) = \frac{1}{\Gamma\left(\frac{3}{2}\right)} \frac{(2k)^{-3}}{ck^{-1} + 1} e^{ck^{-1}}.$$

It follows by the algebraic calculation,

$$\begin{aligned}
\int_0^\infty e^{-t^2 x} p_{k,c}(x) dx &= m(k, c) \int_0^\infty x^{-\frac{5}{2}} \exp\left(-\frac{1}{4k^2} x^{-1} - (t^2 + c^2)x\right) dx \\
&= \frac{m(k, c)}{m\left(k, \sqrt{t^2 + c^2}\right)} \\
&= \frac{(\sqrt{t^2 + c^2} k^{-1} + 1) \exp\left(\sqrt{t^2 + c^2} k^{-1}\right)}{(ck^{-1} + 1) \exp(ck^{-1})} \\
&= \left(\frac{k + \sqrt{t^2 + c^2}}{k + c}\right) e^{-\frac{\sqrt{t^2+c^2}-c}{k}},
\end{aligned}$$

as required.

B Simulating ERG Random Variables

We consider below 3 different ways to generate independent random variables from ERG distribution.

(a) Since Theorem 1, we can approximate an ERG(c) with the finite sum

$$X_N = \sum_{k=1}^{N-1} \text{GIG}\left(-\frac{3}{2}, 2c^2, \frac{1}{2k^2}\right) + \Gamma(a_N, b_N) \tag{16}$$

where the gamma random variable $\Gamma(a_N, b_N)$ are to approximate the tail part by matching the first two moments given in Theorem 2. The shape and rate parameter are

$$b_N = \frac{2c^2\left(\psi(N+c) - \psi(N)\right)}{\psi(N+c) - \psi(N) - c\psi'(N+c)}$$
$$a_N = \frac{b_N}{2c}\left(\psi(N+c) - \psi(N)\right)$$

(b) Since the generation of i.i.d. inverse gamma variables is faster than that of GIG variables, for small values of c, we may consider rejection sampling with ERG(0) as the proposal density.

1. Generate a sample W from ERG(0) using the method in (a) and U from Unif[0,1].
2. If $U < e^{-c^2 W}$, accept W as a sample drawn from ERG(c). Otherwise, reject W and return to the sampling step.

(c) McLeish (2014) shows that one can simulate random variables using the saddlepoint approximation, given the cumulant generating function $k(t) = \log E(e^{tW})$ is known. The saddlepoint approximation is

$$P(W \le x) \approx \Phi(w) + \phi(w)\left(\frac{1}{w} - \frac{1}{u}\right)$$
$$w = w(t) = \text{sgn}(t)\sqrt{2(tk'(t) - k(t))}$$
$$u = u(t) = t\sqrt{k''(t)}$$

where t solves $k'(t) = x$. $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and density of standard normal distribution. We can first generate a random variable T using inverse transform method, such that its cdf $F(t) = \Phi(w(t)) + \phi(w(t))\left(\frac{1}{w(t)} - \frac{1}{u(t)}\right)$, then $W = k'(T)$ has cdf given by the saddlepoint approximation above.

1. Genearte a sample U from Unif[0,1].

2. Solve $U = F(t)$ using Newton-Raphson iteration

$$
\begin{aligned}
t_{n+1} &= t_n - \frac{F(t_n) - U}{\sqrt{k''(t_n)}\phi(w(t_n))} \\
k(t) &= -\gamma\sqrt{c^2 - t} - \ln\Gamma(1 + \sqrt{c^2 - t}) + (\gamma c + \ln\Gamma(1 + c)) \\
k'(t) &= \frac{\gamma + \psi(1 + \sqrt{c^2 - t})}{2\sqrt{c^2 - t}} \\
k''(t) &= \frac{\gamma + \psi(1 + \sqrt{c^2 - t}) - \sqrt{c^2 - t}\psi'(1 + \sqrt{c^2 - t})}{4(c^2 - t)^{3/2}}
\end{aligned}
$$

3. $W = k'(t)$.

C Simulating PTN Random Variables

The probability density of a random variable PTN(p, a, b) is given as

$$
p(x \mid p, a, b) = \frac{x^{p-1}e^{-ax^2+bx}}{\int_0^\infty x^{p-1}e^{-ax^2+bx}dx}, \quad (x, p, a > 0, b \neq 0).
$$

Note that we can write it as a multiplication

$$
\begin{aligned}
p(x \mid p, a, b) &= \frac{x^{p-1}e^{-ax^2+bx}}{\int_0^\infty x^{p-1}e^{-ax^2+bx}dx} \\
&= \frac{x^{p-1}e^{-(\tau|b|-b)x}e^{-a(x-\tau|b|/2a)^2}}{\int_0^\infty x^{p-1}e^{-(\tau|b|-b)x}e^{-a(x-\tau|b|/2a)^2}dx} \\
&= ce^{-a(x-\tau|b|/2a)^2}g(x \mid p, \tau|b| - b)
\end{aligned}
$$

where $g(x|p, b)$ is the density of a gamma random variable with shape p and rate $\tau|b| - b > 0$.

$$
0 \leq e^{-a(x-\tau|b|/2a)^2} \leq 1,\ c = \frac{\int_0^\infty x^{p-1}e^{-(\tau|b|-b)x}dx}{\int_0^\infty x^{p-1}e^{-(\tau|b|-b)x}e^{-a(x-\tau|b|/2a)^2}dx} \geq 1
$$

We sample from $p(x \mid p, a, b)$ by rejection method:

1. Generate $X \sim \Gamma(p, \tau|b| - b)$ and $U \sim Unif[0, 1]$
2. Return X until $U \leq e^{-a(X-\tau|b|/2a)^2}$.

If $b > 0$, we set $\tau = \sqrt{1/4 + 2ap/b^2} + 1/2$; if $b < 0$, we set $\tau = \sqrt{1/4 + 2ap/b^2} - 1/2$. Then $e^{-a(E(X)-\tau|b|/2a)^2} = 1$.

D Approximating Gibbs Sampler

If $g_\ell(x) = g(x)$ for all ℓ (similarly for h and j), and L, M, N are all very large, then an approximate Gibbs sampler based on the conditional posterior in Proposition 2 is

1. Initialize $\tilde{a}^{(0)}, \tilde{b}^{(0)}$.

2. At step t, sample $(\tilde{a}^{(t)}, \tilde{b}^{(t)})$ from $N(\mu_a(t), \sigma_a^2(t))$ and $N(\mu_b(t), \sigma_b^2(t))$ respectively, where

$$\mu_a(t) = a + \frac{M(h'(0))^2}{2h(x^{(t-1)})}\left(\psi(1 + h(x^{(t-1)})) + \gamma\right)$$
$$\mu_b(t) = b + Lg'(0)\psi(g(x^{(t-1)})) + M\gamma h'(0) - 2(\mu_a(t) - a) + Nj'(0)\left(\psi(j(x^{(t-1)})) - \psi(j(x^{(t-1)}) + \beta_n)\right)$$
$$\sigma_a^2(t) = \frac{M(h'(0))^4}{4(h(x^{(t-1)}))^3}\left(\psi(1 + h(x^{(t-1)})) + \gamma - h(x^{(t-1)})\psi'(1 + h(x^{(t-1)}))\right)$$
$$\sigma_b^2(t) = L(g'(0))^2\psi'(g(x^{(t-1)})) + 4(h'(0))^2\sigma_a^2(t) + N(j'(0))^2\left(\psi'(j(x^{(t-1)})) - \psi'(j(x^{(t-1)}) + \beta_n)\right)$$

3. Sample $x^{(t)}$ from $\sum_{m=0}^{M} \pi_k \cdot \text{PTN}(p + m, \tilde{a}^{(t)}, \tilde{b}^{(t)})$.