Bayesian Inference for Pólya Inverse Gamma Models

Christopher Glynn
Paul College of Business and Economics
University of New Hampshire

Jingyu He
Booth School of Business
University of Chicago

Nicholas Polson
Booth School of Business
University of Chicago

Jianeng Xu
Booth School of Business
University of Chicago

First Draft: June 2018

This Draft: May 2019

Abstract

Probability density functions that include the gamma function are widely used in statistics and machine learning. The normalizing constants of gamma, inverse gamma, beta, and Dirichlet distributions all include model parameters as arguments in the gamma function; however, the gamma function does not naturally admit a conjugate prior distribution in a Bayesian analysis, and statistical inference of these parameters is a significant challenge. In this paper, we construct the Pólya-inverse Gamma (P-IG) distribution as an infinite convolution of Generalized inverse Gaussian (GIG) distributions, and we represent the reciprocal gamma function as a scale mixture of normal distributions. As a result, the P-IG distribution yields an efficient data augmentation strategy for fully Bayesian inference on model parameters in gamma, inverse gamma, beta, and Dirichlet distributions. To illustrate the applied utility of our data augmentation strategy, we infer the proportion of overdose deaths in the United States attributed to different opioid and prescription drugs with a Dirichlet allocation model.

Key Words: Pólya inverse Gamma, Pólya Gamma, Exponential reciprocal Gamma, Latent Dirichlet Allocation, Topic models, Gamma shape, Generalized Gamma Convolutions.

1 Introduction

Gamma, inverse gamma, beta, and Dirichlet probability distributions are core components of many Bayesian statistical and machine learning models. The normalizing constants of these distributions depend on gamma functions whose arguments include shape (gamma, inverse gamma) and concentration (beta, Dirichlet) parameters. Bayesian learning of parameters nested inside the gamma function presents significant technical difficulties, since there is no known conjugate prior distribution. In fact, inferring the shape parameter in the gamma distribution is a long-studied problem in Bayesian inference (Damsleth, 1975; Rossell et al., 2009; Miller, 2018).

In this paper, we develop the theoretical and algorithmic foundation of a Pólya-inverse Gamma (P-IG) data augmentation scheme for fully Bayesian inference of shape and concentration parameters in gamma, inverse gamma, and Dirichlet models, respectively. P-IG data augmentation may be utilized to design efficient Markov chain Monte Carlo (MCMC) algorithms in latent Dirichlet allocation (Blei et al., 2003), Beta-negative binomial models (Zhou et al., 2012), and Gamma-Gamma (GaGa) hierarchical models (Rossell et al., 2009). It adds to the literature on Bayesian computation with auxiliary variables, which have proven useful in computing posterior distributions in logistic regression (Polson et al., 2013), multinomial factor models (Holmes and Held, 2006), support vector machines (Mallick et al., 2005; Polson and Scott, 2011), and dependent multinomial models (Linderman et al., 2015).

The P-IG distribution is defined as an infinite convolution of Generalized inverse Gaussian (GIG) distributions and is related to the class of Pólya-Gamma (PG) distributions (Polson et al., 2013) for logistic regression. The Exponential reciprocal Gamma (E-RG) distribution is a special case of the P-IG distribution that has direct application to gamma shape inference. Our data-augmentation scheme builds on distributional results of Hartman (1976) and Roynette and Yor (2005), who provide a representation of the reciprocal gamma function as a scale mixture of normals. This adds to scale mixtures results in Bayesian inference, see Andrews and Mallows (1974), Barndorff-Nielsen et al. (1982), West (1987), and Polson et al. (2013). Scale mixtures of normals are increasingly used in modeling complex high-dimensional distributions, and Bhattacharya et al. (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 et al. (2014) and Glynn et al. (2019).

To illustrate the applied utility of our data augmentation strategy, we use a multinomial -

Dirichlet model to estimate the proportion of overdose deaths in the United States attributed to different opioid and prescription drugs. Robust quantification of uncertainty in the number of deaths attributed to opioids is of great interest to the public health community, and our P-IG approach provides a full posterior distribution, avoiding approximate EM-style algorithms such as Minka (2000) or the simulation approach of Miller (2018) and that taken by Rossell et al. (2009) in the class of GaGa models. We also present an application of gamma shape inference (West, 1992; Miller, 2018).

The rest of our paper is outlined as follows: Section 2 defines the class of P-IG distributions, relates the P-IG and Pólya-Gamma distributions, and identifies the Exponential reciprocal Gamma (E-RG) distribution as a special case of the P-IG class; Section 3 constructs data augmentation strategies in hierarchical multinomial-Dirichlet models, developing a parameter expanded Gibbs sampler for fully Bayesian inference of Dirichlet concentration parameters; Section 4 presents an augmentation strategy for fully Bayesian inference of the shape parameter in the gamma distribution; Section 5 presents an analysis of the opioid and prescription drug overdose data; and Section 6 concludes with directions for future research.

2 The Pólya-Inverse Gamma (P-IG) Distribution Class

In this section, we present the theoretical development of the P-IG distribution class, defining the P-IG distribution by the form of its Laplace transform. In Section 2.1, we define a specific case of the P-IG distribution and prove that it is an infinite convolution of independent GIG distributions; in Section 2.2, the general class of Pólya-Inverse Gamma distributions is constructed with an exponential tilting of the special case defined in Section 2.1; and in Section 2.3 we prove that the Exponential reciprocal Gamma (E-RG) distribution is a member of the P-IG distribution class, a result that relates ratios of gamma functions to the P-IG distribution.

2.1 The P-IG$(d, 0)$ distribution

Let P-IG$(\boldsymbol{d}, 0)$ denote the Pólya-inverse Gamma distribution where the infinite-dimension parameter vector $\boldsymbol{d} = (d_1, d_2, \cdots) > 0$ is a sequence of given positive constants. The second 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 ω has a Pólya-inverse Gamma distribution, P-IG$(\boldsymbol{d}, 0)$, with den-

sity $p(\omega \mid \boldsymbol{d}, 0)$ if its Laplace transform takes the form

$$E\left[e^{-\omega t^2}\right] = \int_0^\infty e^{-\omega t^2} p\left(\omega \mid \boldsymbol{d}, 0\right) d\omega = \prod_{k=1}^{\infty} \left(1 + \frac{|t|}{d_k}\right) e^{-\frac{|t|}{d_k}}. \tag{1}$$

We write $\omega \overset{D}{=} \text{P-IG}(\mathbf{d}, 0)$.

Remark 1. *With $d_k = k$, we have*

$$\frac{e^{-\gamma t}}{\Gamma(t+1)} = \int_0^\infty e^{-\omega t^2} p\left(\omega \mid \boldsymbol{d}, 0\right) d\omega, \quad t > 0$$

using the Hadamard factorization of the reciprocal Gamma function,

$$\frac{e^{-\gamma t}}{\Gamma(t+1)} = \prod_{k=1}^{\infty} \left(1 + \frac{t}{k}\right) e^{-\frac{t}{k}},$$

where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant.

Lemma 1. A P-IG$(\boldsymbol{d}, 0)$ random variable can be represented as an infinite convolution of reciprocal gamma distributions, equivalently expressed as an infinite convolution of GIG distributions,

$$w|\boldsymbol{d} \overset{D}{=} \sum_{k=1}^{\infty} R\Gamma\left(\frac{3}{2}, \frac{1}{4d_k^2}\right) = \sum_{k=1}^{\infty} GIG\left(-\frac{3}{2}, \frac{1}{\sqrt{2}d_k}, 0\right). \tag{2}$$

Proof. Let $w_k \sim R\Gamma(\frac{3}{2}, \beta_k)$, where $R\Gamma$ denotes the reciprocal (or inverse) gamma distribution. It has density

$$f_{w_k}(y) = \frac{\beta_k^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} y^{-\frac{5}{2}} e^{-\beta_k y^{-1}}, \ (y > 0),$$

where $\beta_k = 1/4d_k^2$, so that

$$\begin{aligned} E(e^{-t^2 w_k}) &= \int_0^\infty e^{-t^2 y} \frac{\beta_k^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} y^{-\frac{5}{2}} e^{-\beta_k y^{-1}} dy \\ &= \left(1 + \frac{|t|}{d_k}\right) e^{-\frac{|t|}{d_k}}. \end{aligned}$$

Therefore, by construction,

$$w|\boldsymbol{d} \overset{d}{=} \sum_{k=1}^{\infty} R\Gamma\left(\frac{3}{2}, \frac{1}{4d_k^2}\right) \Longrightarrow E_{w|\boldsymbol{d}}(e^{-t^2 w}) = \prod_{k=1}^{\infty} \left(1 + \frac{|t|}{d_k}\right) e^{-\frac{|t|}{d_k}}.$$

Since the reciprocal gamma distribution is a special case of the GIG[1] distribution, it follows that

$$w|\boldsymbol{d} \stackrel{d}{=} \sum_{k=1}^{\infty} GIG\left(-\frac{3}{2}, \frac{1}{\sqrt{2}d_k}, 0\right). \tag{3}$$

□

2.2 The General P-IG(d,c) Class

We construct the general class of P-IG distributions, P-IG$(\boldsymbol{d}, c)$, by exponentially tilting the P-IG$(\boldsymbol{d}, 0)$ class. The exponential tilting strategy – similar to the one used by Polson et al. (2013) – allows a second parameter $c \in \mathcal{R}$ to inform a priori the precision of the P-IG random variable.

Definition 2.2. The P-IG$(\boldsymbol{d}, c)$ distribution is constructed as an exponential tilting of the P-IG$(\boldsymbol{d}, 0)$ density, defined by

$$p\left(\omega \mid \boldsymbol{d}, c\right) = \frac{\exp\left(-\frac{c^2}{2}\omega\right) p\left(\omega \mid \boldsymbol{d}, 0\right)}{E_\omega\left[\exp\left(-\frac{c^2}{2}\omega\right)\right]}. \tag{4}$$

The normalizing constant, namely $E_\omega\left[\exp\left(-\frac{c^2}{2}\omega\right)\right]$, can be calculated using the Laplace transform identity in (1) which defines the P-IG distribution. The Laplace transform is given by

$$E_w(e^{-t^2 w}) = \prod_{k=1}^{\infty} \left(\frac{d_k + \sqrt{t^2 + c^2/2}}{d_k + c/\sqrt{2}}\right) e^{-\frac{\sqrt{t^2+c^2/2}}{d_k}} e^{\frac{c/\sqrt{2}}{d_k}}. \tag{5}$$

Our main result, presented in Theorem 1, is that a a random variable $\omega \sim$P-IG$(\boldsymbol{d}, c)$ may be constructed from an infinite sum of independent GIG-distributed random variables. The power of the result lies in the ability to identify previously unknown conditional posterior distributions in Bayesian inference and provide simulation strategies based on Generalized Gamma Convolutions (GGC) (Bondesson, 1992).

Theorem 1. *The P-IG$(\boldsymbol{d}, c)$ class of distributions can be constructed as an infinite sum of generalized inverse*

[1]The reciprocal gamma (RΓ) is a special case of the three-parameter generalized inverse Gaussian distribution, GIG(ν, δ, γ), with density function

$$p\left(x\right) = \frac{(\gamma/\delta)^\nu}{2K_\nu(\delta\gamma)} x^{\nu-1} \exp\left\{-\frac{1}{2}\left(\delta^2 x^{-1} + \gamma^2 x\right)\right\}, \quad x > 0.$$

Here $K_\nu(\cdot)$ is a modified Bessel function of the second kind.

Gaussian (GIG) distributions as follows

$$\text{P-IG}\left(\boldsymbol{d}, c\right) \stackrel{D}{=} \sum_{k=1}^{\infty} GIG\left(-\frac{3}{2}, \frac{1}{\sqrt{2}d_k}, |c|\right).$$

Proof. It suffices to show that Laplace transform of a $Y_k \sim GIG\left(-\frac{3}{2}, \frac{1}{\sqrt{2}d_k}, |c|\right)$ random variable is given by

$$E(e^{-t^2 Y_k}) = \left(\frac{d_k + \sqrt{t^2 + c^2/2}}{d_k + c/\sqrt{2}}\right) e^{-\frac{\sqrt{t^2+c^2/2}}{d_k}} e^{\frac{c/\sqrt{2}}{d_k}}.$$

The density of Y_k given by

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

with normalizing constant,

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

The Laplace transform follows by the algebraic calculation,

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

as required. □

Remark 2. *The popular Pólya Gamma distribution (Polson et al., 2013) with parameter $b > 0$ and $c \in \mathcal{R}$, denoted as $X \sim PG(b, c)$, is defined as an infinite convolution of gamma distributions. Because the gamma distribution is a special case of the GIG distribution, the PG(b, c) distribution can be represented as an infinite*

convolution of GIG distributions,

$$\omega \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{Gamma\left(b,1\right)}{(k-1/2)^2 + c^2/(4\pi^2)} \overset{D}{=} \sum_{k=1}^{\infty} Gamma\left(b, \left(2\pi^2(k-1/2)^2 + c^2/2\right)^{-1}\right)$$

$$\omega \overset{D}{=} \sum_{k=1}^{\infty} GIG\left(b, 0, \sqrt{\frac{2}{2\pi^2(k-1/2)^2 + c^2/2}}\right).$$

Thus, the PG(b,c) distribution is closely related to the P-IG distribution class through the infinite convolution if GIG distributions.

Remark 3. *Gamma function ratios appear in full conditional distributions in Bayesian nonparametric mixture models (Ferguson, 1973; Antoniak, 1974). For example, the distribution for the number of clusters in the Dirichlet Process mixture model, denoted k, depends on concentration parameter α and the number of observations n,*

$$p\left(k \mid \alpha, n\right) = \binom{n}{k} n! \alpha^k \frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha + n\right)}. \tag{6}$$

The ratio of gamma functions in 6 may be represented with the Beta function,

$$\frac{\Gamma\left(\lambda\right)}{\Gamma\left(\lambda + \alpha\right)} = \frac{\left(\lambda + \alpha\right) Beta\left(\lambda + 1, \alpha\right)}{\lambda \Gamma\left(\alpha\right)}. \tag{7}$$

Combining the likelihood in 6 and the Beta representation in 7 with a gamma prior $p\left(\alpha\right)$ enables conditional posterior sampling of concentration parameter α as a mixture of two gamma distributions. See Section 6 of Escobar and West (1995).

2.3 Exponential Reciprocal Gamma (E-RG) Models

The Exponential reciprocal Gamma (E-RG) distribution is constructed to provide a data augmentation strategy for reciprocal gamma functions, since its Laplace transform is given by a ratio of Gamma functions,

$$E_{\omega|a}\left[e^{-\omega t^2}\right] = \frac{\Gamma\left(a\right)}{\Gamma\left(a + t\right)}, \quad t > 0. \tag{8}$$

We write $\omega \overset{D}{=} \text{E-RG}\left(a\right)$ for $a > 0$. To show that this falls into the P-IG class, use the Hadamard-

Weierstrass factorization of the reciprocal Gamma function, see Roynette and Yor (2005), p 1265.

$$\frac{\Gamma(a)}{\Gamma(a+t)} = e^{-\psi(a)t} \prod_{k=0}^{\infty} \left(1 + \frac{t}{a+k}\right) e^{-\frac{t}{a+k}}, \tag{9}$$

where $\psi(a)$ is the digamma function. Hence an equivalent definition of the E-RG Laplace transform is

$$E_{\omega|a}\left[e^{-\omega t^2}\right] = e^{-\psi(a)t} \prod_{k=0}^{\infty} \left(1 + \frac{t}{a+k}\right) e^{-\frac{t}{a+k}}. \tag{10}$$

Recall that the Hadamard factorization of P-IG$(\boldsymbol{d}, 0)$ with $d_k = k$ in Remark 1,

$$\frac{e^{-\gamma t}}{\Gamma(t+1)} = \prod_{k=1}^{\infty} \left(1 + \frac{t}{k}\right) e^{-\frac{t}{k}}, \tag{11}$$

coincides with (10) when $a = 1$ and $\psi(1) = -\gamma$. Hence, the exponential reciprocal gamma E-RG(1) distribution is a special case of the Pólya-inverse gamma distribution with the sequence $d_k = k$, P-IG$((1, 2, 3, \ldots), 0)$.

Hartman (1976) and Roynette and Yor (2005) discuss the scale mixture of normals representation of $\Gamma(a)/\Gamma(a+t) = E_{w|a}\left[e^{-\omega t^2}\right]$. This is related to the Laplace transform identity in (11).

3 Inferring concentration parameters in multinomial-Dirichlet models

In this section, we develop Markov chain Monte Carlo (MCMC) algorithms for fully Bayesian inference of the concentration parameter vector in the Dirichlet distribution. Such inference problems commonly arise in applied analyses of categorical data. Section 3.1 presents the general hierarchical multinomial-Dirichlet model class for which the P-IG data augmentation scheme may be utilized. Section 3.2 develops a parameter expanded Gibbs sampler for inferring the concentration parameter in the Dirichlet distribution.

3.1 A hierarchical multinomial-Dirichlet model class

The multinomial-Dirichlet framework presented herein is closely related to the latent Dirichlet allocation model of Blei et al. (2003) for topic modeling in text data, and we use text analysis as a motivating context. Suppose that for document $m \in \{1, \ldots, M\}$, each of N_m words in the document is independently allocated to K topics conditional on probability vector $\boldsymbol{p}_m = (p_{m1}, p_{m2}, \ldots, p_{mK})$.

For each document m, the number of words allocated to each topic, $\boldsymbol{n}_m = (n_{m1}, \cdots, n_{mK})$, is modeled with a multinomial distribution. The sampling model for the count vector $\boldsymbol{n}_m$ is then a multinomial distribution given probability vector $\boldsymbol{p}_m$,

$$\boldsymbol{n}_m \mid \boldsymbol{p}_m \sim \text{Multinomial}\left(\boldsymbol{p}_m\right). \tag{12}$$

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

$$\boldsymbol{p}_m \mid \boldsymbol{\alpha} \sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right). \tag{13}$$

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 P-IG augmentation admit independent uniform, truncated normal, and exponential prior distributions for the elements α_k. Although reference priors $p(\alpha_k) \propto 1$ and exponential priors yield tractable full conditional distributions in the Gibbs sampler, we find in numerical experiments that they do not provide sufficient regularization for posterior convergence and advise against using them. Section 3.2 presents analyses based on independent truncated normal 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 $p\left(\boldsymbol{p}^* | \boldsymbol{n}_1, \ldots, \boldsymbol{n}_M\right) = \int_{\boldsymbol{\alpha}} p\left(\boldsymbol{p}^* \mid \boldsymbol{\alpha}\right) p(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_M) d\boldsymbol{\alpha}$ requires posterior computation of $p\left(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_M\right) \propto \prod_{m=1}^{M} p\left(\boldsymbol{n}_m \mid \boldsymbol{\alpha}\right) p\left(\boldsymbol{\alpha}\right)$; however, when the probability vectors $\boldsymbol{p}_m$ are integrated out of the multinomial likelihood, the marginal likelihood $p\left(\boldsymbol{n}_m \mid \boldsymbol{\alpha}\right)$ includes elements of $\boldsymbol{\alpha}$ inside the gamma function,

$$p\left(\boldsymbol{n}_m \mid \boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma\left[\sum_{k=1}^{K} \left(n_k + \alpha_k\right)\right]} \prod_{k=1}^{K} \frac{\Gamma\left(n_k + \alpha_k\right)}{\Gamma\left(\alpha_k\right)}. \tag{14}$$

Because $\boldsymbol{\alpha}$ is nested inside the gamma function, computing $p(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_M)$ is a challenge. Previous inference strategies relied on approximations, but in Section 3.2 we introduce a new data

augmentation scheme for computing the full posterior $p(\boldsymbol{\alpha} \mid \boldsymbol{n}_1, \ldots, \boldsymbol{n}_M)$.

3.2 Data augmentation strategies with P-IG auxiliary variables

The parameter expanded Gibbs sampler presented below introduces two auxiliary random variables, w_{mk} and η_m, to represent the gamma function with scale mixtures of normal distributions. Auxiliary variables w_{mk} and η_m in the scale mixture representation are iteratively conditioned on and then updated as part of the inference strategy for $\boldsymbol{\alpha}$.

Assume independent prior distributions for each element of vector $\boldsymbol{\alpha}$ so that $p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} p(\alpha_k)$. Note that truncated normal priors on each α_k give closed-form full conditional distributions in a Gibbs sampler, which is proven below. When $\alpha_k \sim TN(0, \tau^2)\mathbf{I}(\alpha_k > 0)$, where TN denotes the truncated normal distribution, the expectation $E[\alpha_k] = \sqrt{\frac{2}{\pi}}\tau$. We can set $E[\alpha_k] = \frac{1}{K}$, a standard choice for the Dirichlet concentration parameter, by choosing $\tau = \frac{1}{K}\sqrt{\frac{\pi}{2}}$. The important takeaway is that when the prior variance for the truncated normal depends on the dimension of the Dirichlet distribution, K, the expectation of α_k becomes a function of K as well.

Theorem 2. *The Gibbs sampler for data augmented multinomial-Dirichlet models is given by*

$$
\begin{aligned}
\eta_m \mid \boldsymbol{\alpha}, \boldsymbol{p} &\sim \Gamma\left(\sum_{k=1}^{K} \alpha_k + n_{m\bullet}, 1\right), \quad \forall m = 1, \cdots, M \\
w_{mk} \mid \boldsymbol{\alpha}, \boldsymbol{p} &\sim \textit{P-IG}\left(\boldsymbol{d}, \sqrt{2(n_{mk} + \alpha_k - 1)^2}\right), \quad d_k = k, \quad \forall i = 1, \cdots, K \\
\boldsymbol{p}_m \mid \boldsymbol{\alpha} &\sim \textit{Dirichlet}\left(n_{m1} + \alpha_1, \cdots, n_{mK} + \alpha_K\right) \\
\alpha_k \mid \eta, \boldsymbol{w} &\sim TN\left(\frac{b}{2a}, \frac{1}{2a}\right)\mathbf{I}(\alpha_k > 0)
\end{aligned}
\tag{15}
$$

where the value of a and b depend on the prior on α_k. Denote the truncated normal which truncates at $\alpha_k > 0$ by $TN(\cdot, \cdot)\mathbf{I}(\alpha_k > 0)$. Under the truncated normal prior $p(\alpha_k) \sim N(0, \tau^2)\mathbf{I}(\alpha_k > 0)$,

$$
\begin{aligned}
a &= \sum_{m=1}^{M} w_{mk} + \frac{1}{2\tau^2} \\
b &= \left[-2\sum_{m=1}^{M}(n_{mk} - 1)w_{mk} + \sum_{m=1}^{M}\log\eta_m + M\gamma + \sum_{m=1}^{M}\log p_{mk}\right].
\end{aligned}
$$

An essential aspect of this augmentation strategy is that all of these distributions are straightforward to simulate from.

Proof. Suppose the data is $\{n_{mk}\}_{m=1\cdots M,k=1\cdots K}$. Let $n_{m\bullet} = \sum_{k=1}^{K} n_{mk}$ and $n_{\bullet k} = \sum_{m=1}^{M} n_{mk}$. The likelihood $\boldsymbol{n}_m \mid \boldsymbol{p}_m$ and posterior $\boldsymbol{p}_m \mid \boldsymbol{n}_m$ for the probability vector $\boldsymbol{p}_m$ underlying observation $\boldsymbol{n}_m$ are given by

$$
\begin{aligned}
p\left(\boldsymbol{n}_m \mid \boldsymbol{p}_m\right) &\propto p_{m1}^{n_{m1}} \cdots p_{mK}^{n_{mK}} \\
p\left(\boldsymbol{p}_m \mid \boldsymbol{\alpha}, \boldsymbol{n}_m\right) &\propto \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k + n_{m\bullet}\right)}{\prod_{k=1}^{K} \Gamma\left(n_{mk} + \alpha_k\right)} \prod_{k=1}^{K} p_{mk}^{n_{mk}+\alpha_k-1} \\
&= \Gamma\left(\sum_{k=1}^{K} \alpha_k + n_{m\bullet}\right) \prod_{k=1}^{K} \left[\frac{1}{\Gamma(n_{mk}+\alpha_k)} p_{mk}^{n_{mk}+\alpha_k-1}\right] \\
&= \Gamma\left(\sum_{k=1}^{K} \alpha_k + n_{m\bullet}\right) \prod_{k=1}^{K} \left[\frac{1}{\Gamma(n_{mk}+\alpha_k)} e^{-\gamma(n_{mk}+\alpha_k-1)} e^{(\gamma+\log p_{mk})(n_{mk}+\alpha_k-1)}\right] \\
&= \int_0^\infty \eta_m^{\sum_{k=1}^{K} \alpha_k + n_{m\bullet} - 1} e^{-\eta_m} d\eta_m \prod_{k=1}^{K} \int_0^\infty e^{-(n_{mk}+\alpha_k-1)^2 w_{mk}} p(w_{mk}) dw_{mk} \\
&\quad \times \prod_{k=1}^{K} e^{(\gamma+\log p_{mk})(n_{mk}+\alpha_k-1)}.
\end{aligned}
\tag{16}
$$

Observe two points in 16: (i) the integral identity $\Gamma\left(\sum_{k=1}^{K} \alpha_k + n_{m\bullet}\right) = \int_0^\infty \eta_m^{\sum_{k=1}^{K} \alpha_k + n_{m\bullet} - 1} e^{-\eta_m} d\eta_m$ introduces an auxiliary random variable $\eta_m \sim \text{Gamma}\left(\sum_{k=1}^{K} \alpha_k + n_{m\bullet}, 1\right)$; and (ii) the integral identity $\frac{e^{-\gamma(n_{mk}+\alpha_k-1)}}{\Gamma(n_{mk}+\alpha_k)} = \int_0^\infty e^{-(n_{mk}+\alpha_k-1)^2 w_{mk}} p(w_{mk}) dw_{mk}$ is the Laplace transform of the E-RG distribution in Section 2.3, which is related to the Hadamard factorization of the P-IG$(\boldsymbol{d}, 0)$ distribution. This second integral identity introduces another auxiliary random variable $w_{mk} \sim$ P-IG$(\boldsymbol{d}, 0)$ (see Remark 1 and Equations 10 and 11 for the Hadamard factorization of the E-RG distribution).

The joint posterior for $p_1, \ldots, p_M$ is then

$$
\begin{aligned}
&p\left(\boldsymbol{p}_1, \cdots, \boldsymbol{p}_M \mid \boldsymbol{\alpha}, \boldsymbol{n}\right) \propto p\left(\boldsymbol{p}_m \mid \boldsymbol{\alpha}, \boldsymbol{n}\right) \cdots p\left(\boldsymbol{p}_M \mid \boldsymbol{\alpha}, \boldsymbol{n}\right) \times p\left(\boldsymbol{\alpha}\right) \\
&\propto \prod_{m=1}^{M} \left\{ \Gamma\left(\sum_{k=1}^{K} \alpha_k + n_{m\bullet}\right) \prod_{k=1}^{K} \left[\frac{1}{\Gamma\left(\left(n_{mk}+\alpha_k-1\right)+1\right)} e^{-\gamma(n_{mk}+\alpha_k-1)} e^{(\gamma+\log p_{mk})(n_{mk}+\alpha_k-1)}\right] \right\} \\
&= \prod_{m=1}^{M} \left\{ \int_0^\infty \eta_m^{\sum_{k=1}^{K} \alpha_k + n_{m\bullet} - 1} e^{-\eta_m} d\eta_m \prod_{k=1}^{K} \int_0^\infty e^{-(n_{mk}+\alpha_k-1)^2 w_{mk}} p\left(w_{mk}\right) dw_{mk} \prod_{k=1}^{K} e^{(\gamma+\log p_{mk})(n_{mk}+\alpha_k-1)} \right\} p\left(\boldsymbol{\alpha}\right).
\end{aligned}
$$

This leads to the posterior augmented by $\boldsymbol{w}$ and $\boldsymbol{\eta}$

$$
p\left(\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{\eta} \mid \boldsymbol{p}_1, \cdots, \boldsymbol{p}_M\right) = \prod_{m=1}^{M} \left\{ \eta_m^{\sum_{k=1}^{K} \alpha_k + n_{m\bullet} - 1} e^{-\eta_m} \prod_{k=1}^{K} e^{-(n_{mk}+\alpha_k-1)^2 w_{mk}} p\left(w_{mk}\right) \prod_{k=1}^{K} e^{(\gamma+\log p_{mk})(n_{mk}+\alpha_k-1)} \right\} p\left(\boldsymbol{\alpha}\right).
$$

Conditional on $\boldsymbol{\eta}$ and $\boldsymbol{w}$, the distribution of $\boldsymbol{\alpha}$ is then

$$
\begin{aligned}
&p\left(\boldsymbol{\alpha} \mid \boldsymbol{\eta}, \boldsymbol{w}\right) \propto \exp\left\{\log p(\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{\eta} \mid \boldsymbol{p}_1, \cdots, \boldsymbol{p}_M)\right\} \\
&\propto \exp\left\{\sum_{m=1}^{M}\left(\sum_{k=1}^{K}\alpha_k + n_{m\bullet} - 1\right)\log\eta_m - \sum_{m=1}^{M}\sum_{k=1}^{K}\left(n_{mk}+\alpha_k-1\right)^2 w_{mk} + \sum_{m=1}^{M}\sum_{k=1}^{K}\left(\gamma+\log p_{mk}\right)\left(n_{mk}+\alpha_k-1\right)\right\} p\left(\boldsymbol{\alpha}\right) \\
&\propto \exp\left\{\sum_{m=1}^{M}\sum_{k=1}^{K}\alpha_k\log\eta_m - \sum_{m=1}^{M}\sum_{k=1}^{K}\left(n_{mk}+\alpha_k-1\right)^2 w_{mk} + \sum_{m=1}^{M}\sum_{k=1}^{K}\left(\gamma+\log p_{mk}\right)\left(n_{mk}+\alpha_k-1\right)\right\} p\left(\boldsymbol{\alpha}\right) \\
&\propto \exp\left\{\sum_{k=1}^{K}\alpha_k\left(\sum_{m=1}^{M}\log\eta_m\right) - \sum_{k=1}^{K}\left(\sum_{m=1}^{M}(n_{mk}+\alpha_k-1)^2 w_{mk}\right) + \sum_{k=1}^{K}\left(\sum_{m=1}^{M}(\gamma+\log p_{mk})(n_{mk}+\alpha_k-1)\right)\right\} p\left(\boldsymbol{\alpha}\right).
\end{aligned}
$$

Therefore the conditional posterior of each α_k is

$$
\begin{aligned}
&p\left(\alpha_k \mid \boldsymbol{\eta}, \boldsymbol{\omega}\right) \\
&\propto \exp\left\{\alpha_k\left(\sum_{m=1}^{M}\log\eta_m\right) - \left(\sum_{m=1}^{M}\left(n_{mk}+\alpha_k-1\right)^2 w_{mk}\right) + \left(\sum_{m=1}^{M}\left(\gamma+\log p_{mk}\right)\left(n_{mk}+\alpha_k-1\right)\right)\right\} p\left(\alpha_k\right) \\
&\propto \exp\left\{\alpha_k\left(\sum_{m=1}^{M}\log\eta_m\right) - \sum_{m=1}^{M}\left(\alpha_k^2 w_{mk} + 2\alpha_k\left(n_{mk}-1\right)w_{mk} + \alpha_k\left(\gamma+\log p_{mk}\right)\right)\right\} p\left(\alpha_k\right) \\
&\propto \exp\left\{-\left(\sum_{m=1}^{M}w_{mk}\right)\alpha_k^2 + \left[-2\sum_{m=1}^{M}(n_{mk}-1)w_{mk} + \sum_{m=1}^{M}\log\eta_m + M\gamma + \sum_{m=1}^{M}\log p_{mk}\right]\alpha_k\right\} p\left(\alpha_k\right).
\end{aligned}
$$

The form of posterior $p\left(\alpha_k \mid \boldsymbol{\eta}, \boldsymbol{\omega}\right)$ depends on prior $p\left(\alpha_k\right)$. Under the normal prior $p\left(\alpha_k\right) \sim N\left(0, \tau^2\right)\mathbf{I}(\alpha_k > 0)$,

$$
\begin{aligned}
p\left(\alpha_k \mid \boldsymbol{\eta}, \boldsymbol{\omega}\right) &\propto \exp\left\{-\left(\sum_{m=1}^{M}w_{mk}\right)\alpha_k^2 + \left[-2\sum_{m=1}^{M}(n_{mk}-1)w_{mk} + \sum_{m=1}^{M}\log\eta_m + M\gamma + \sum_{m=1}^{M}\log p_{mk}\right]\alpha_k\right\} \\
&\quad \times \exp\left(-\frac{\alpha_k^2}{2\tau^2}\right)\mathbf{I}(\alpha_k > 0) \\
&:= \exp\left(-a\alpha_k^2 + b\alpha_k\right)\mathbf{I}(\alpha_k > 0) \\
&\propto \exp\left(-\frac{\left(\alpha_k - \frac{b}{2a}\right)^2}{1/a}\right)\mathbf{I}(\alpha_k > 0) = TN\left(\frac{b}{2a}, \frac{1}{2a}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
a &= \sum_{m=1}^{M}w_{mk} + \frac{1}{2\tau^2} \\
b &= \left[-2\sum_{m=1}^{M}(n_{mk}-1)w_{mk} + \sum_{m=1}^{M}\log\eta_m + M\gamma + \sum_{m=1}^{M}\log p_{mk}\right].
\end{aligned}
$$

The full conditional for $p\left(w_{mk} \mid \boldsymbol{\alpha}, \boldsymbol{p}\right)$ follows from the exponential tilting construction of P-

IG($\boldsymbol{d}, c$) in Definition 2.2. Starting with the joint posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\eta} \mid \boldsymbol{p}_1, \ldots, p_M)$, we focus on the proportionality including the single element $\boldsymbol{w}_{mk}$,

$$\begin{aligned} p(w_{mk} \mid \boldsymbol{\alpha}, \boldsymbol{p}) &\propto p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\eta} \mid \boldsymbol{p}_1, \ldots, p_M) \\ &\propto \exp\left\{-(n_{mk} + \alpha_k - 1)^2 w_{mk}\right\} p(w_{mk}). \end{aligned} \tag{17}$$

Since the prior distribution for the auxiliary variable w_{mk} is P-IG($\boldsymbol{d}, 0$), Equation 17 and Definition 2.2 imply that $w_{mk} \mid \boldsymbol{\alpha}, \boldsymbol{p} \sim$ P-IG $\left(\boldsymbol{d}, \sqrt{2(n_{mk} + \alpha_k - 1)^2}\right)$. □

An MCMC algorithm for the special case when $\boldsymbol{\alpha}$ is homogeneous (e.g., $\alpha_1 = \alpha_2 = \ldots = \alpha_K$) is presented in Appendix A.

4 Shape Inference of Gamma

The gamma distribution, parameterized by shape α and rate β, is a component of many probability models in Bayesian analysis. 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). 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. Posterior inference of the gamma shape parameter is a long-standing problem in Bayesian analysis (Damsleth, 1975; Damien et al., 1995; Rossell et al., 2009; Miller, 2018). Although posterior inference of the rate parameter is straightforward – since the gamma distribution itself is a conjugate prior for the rate parameter – there is no conjugate prior for the gamma shape parameter, and efficient posterior computation remains an open problem. In this section, we represent of the reciprocal gamma function in the $Ga(\alpha, \beta)$ density as a scale mixture of normals and utilize the P-IG data augmentation scheme to build an efficient MCMC algorithm.

Suppose $y_1, \ldots, y_n$ are independent and identically distributed observations modeled by a $Ga(\alpha, \beta)$ distribution. For observation y_i, the likelihood is

$$p(y_i \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta y_i}. \tag{18}$$

A natural prior for α is given by

$$p(\alpha \mid a, b, c) \propto \frac{a^{\alpha-1}\beta^{c\alpha}}{\Gamma(\alpha)^b}, \text{where } \alpha > 0$$

and a, b, c are given hyperparameters. Therefore, given data $(y_1, y_2, ..., y_n)$, the posterior distribution of α is

$$p(\alpha|a, b, c, \beta, y) \propto \frac{a^{\alpha-1}\beta^{c\alpha}}{\Gamma(\alpha)^b} \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta y_i} \propto \frac{1}{\Gamma(\alpha)^{b'}} \left(\beta'_y\right)^\alpha, \tag{19}$$

with updated hyperparameters $a' = a\prod_{i=1}^n y_i, b' = b + n, c' = c + n$, and $\beta'_y = a'\beta^{c'}$. We define $\tilde{\alpha} = \alpha - 1$ and reparameterize the right side of (19) with the goal of matching the E-RG structure in (10) and (11), which will facilitate posterior computation via P-IG data augmentation.

$$\begin{aligned}\frac{1}{\Gamma(\alpha)^{b'}} \left(\beta'_y\right)^\alpha &= \frac{e^{-\gamma b'\tilde{\alpha}}}{(\Gamma(\tilde{\alpha}+1))^{b'}} e^{\gamma b'\tilde{\alpha}} \left(\beta'_y\right)^{\tilde{\alpha}+1} \\ &= \left(\frac{e^{-\gamma\tilde{\alpha}}}{\Gamma(\tilde{\alpha}+1)}\right)^{b'} e^{\gamma b'\tilde{\alpha}} \left(\beta'_y\right)^{\tilde{\alpha}+1}.\end{aligned}$$

When b' is a nonnegative integer, we are able to introduce b' auxiliary i.i.d P-IG$(\boldsymbol{d}, 0)$ random variables, $\boldsymbol{w} = (w_1, \ldots, w_{b'})$, to represent $\left(\frac{e^{-\gamma\tilde{\alpha}}}{\Gamma(\tilde{\alpha}+1)}\right)^{b'}$ as a scale mixture of normals. The scale mixture representation appears in the product of the Laplace transforms of each auxiliary w_j, as in (20 - 21).

$$\frac{1}{\Gamma(\alpha)^{b'}} \left(\beta'_y\right)^\alpha \propto E_{\boldsymbol{w}}\left[e^{-\left(\sum_{j=1}^{b'} w_j\right)\tilde{\alpha}^2}\right] e^{\left(\gamma b' + \log\beta'_y\right)\tilde{\alpha}} \tag{20}$$

$$= \int_0^\infty e^{\left(\gamma b' + \log\beta'_y\right)\tilde{\alpha}} e^{-\left(\sum_{j=1}^{b'} w_j\right)\tilde{\alpha}^2} p(w) dw_1 \cdots dw_{b'} \tag{21}$$

This leads to a parameter expanded Gibbs sampling strategy with the full conditionals

$$\begin{aligned}p(\tilde{\alpha} \mid \boldsymbol{w}) &\propto e^{\left(\gamma b' + \log\beta'_y\right)\tilde{\alpha}} e^{-\left(\sum_{j=1}^{b'} w_j\right)\tilde{\alpha}^2} \\ &\sim N\left(\mu, \sigma^2\right)\Big|_{\{\tilde{\alpha} > -1\}} \\ w_j \mid \tilde{\alpha} &\sim \text{P-IG}\left(\mathbf{d}, \sqrt{2}\tilde{\alpha}\right),\ j = 1, 2, ..., b',\end{aligned}$$

where $d_k = k$, $\mu = \frac{\gamma b' + \log\beta'_y}{2\sum_{j=1}^{b'} w_j}$ and $\sigma^2 = \frac{1}{2\sum_{j=1}^{b'} w_j}$. The truncated normal full conditional $\tilde{\alpha} \mid \boldsymbol{w}$ ensures

that posterior samples of α are strictly positive. This straightforward Gibbs sampler provides a pathway for fully Bayesian inference in richly structured models of the gamma shape parameter.

Next we show a simple simulation study of gamma shape parameter inference. Suppose the data are generated from Gamma $(3, 2)$, with 200 observations. Figure 1 presents a histogram of 500 posterior samples, where the solid red line is the theoretical posterior density of α and the dashed black line is estimated density from posterior samples.

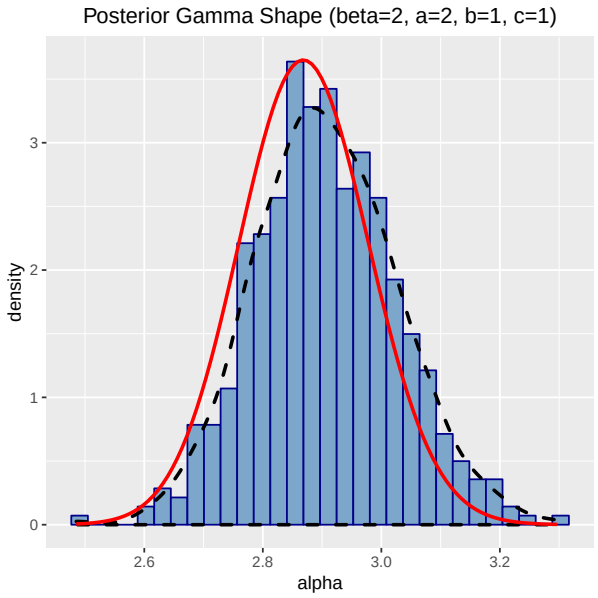

Figure 1: Posterior of Gamma shape parameter.

5 Application: Opioid and Prescription Drug Crisis

We present an analysis of opioid and prescription drug abuse data to illustrate the applied utility of the data augmentation scheme devised in Section 3. The opioid and prescription drug crisis continues to destroy lives in many parts of the United States. While the public discussion of the crisis focuses on opioid abuse, there are multiple drugs contributing to a larger pattern of substance abuse: cocaine, heroin, methadone, natural & semi-synthetic opioids, psychostimulants, and synthetic opioids. Estimating shared patterns of variation in state-level mortality rates is particularly important to public health officials. For example, identifying state-level characteristics associated with higher heroin overdose rates may inform public policy interventions. To estimate the underlying pattern in mortality rates by drug type, we model death counts with the multinomial-Dirichlet

framework presented in Section 3. The data in our analysis comes from the VSRR Provisional Drug Overdose Death Counts, a nationwide data set on mortality statistics from 2015 - 2018. For 19 of 50 states, a break down of deaths by drug type is provided. The underlying overdose rates are not directly observed, and our inference goal is to learn shared patterns of variation in state-level death rates by drug type.

State	Year	Cocaine	Synthetic	Heroin	Methadone	Nat.	Psych.
CT	2015	118	96	298	58	170	18
CT	2016	171	240	403	72	180	24
CT	2017	250	527	470	67	209	23
CT	2018	280	682	405	97	180	39
MD	2015	109	237	327	150	400	17
MD	2016	154	386	418	179	394	21

Table 1: Deaths by drug type: cocaine, synthetic opioids, heroin, methadone, natural opioids, and psychostimulants. The data is provided at the state level from 2015 - 2018, and the snapshot provided above is the first six rows.

Table 1 provides a snapshot of the VSRR data, which reports a count vector for deaths across six different drug types at the state-year level. Observe in Figure 2 that states exhibit distinct patterns of variation in empirical death rates. In some states/years, the largest proportion of overdose deaths is from synthetic opioids, while in others it is heroin. Significant state-year variation in normalized death counts motivates a hierarchical model for the proportion of deaths due to each drug type.

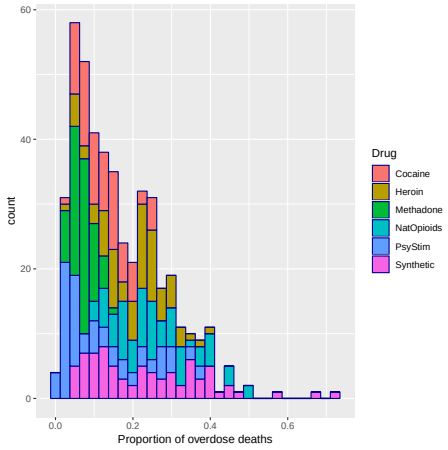

Figure 2: Empirical Rates of overdose by drug type.

In this analysis suppose that states and years are exchangeable, so that $m \in 1, \ldots, M$ indexes each individual state-year combination. Let N_m denote the total number of overdose deaths in

state-year m due to the six drugs under consideration. Let count vector $\boldsymbol{n}_m$ denote the number of deaths associated with each drug type. We model the count vector with a multinomial distribution conditional on the underlying state-year death proportions,

$$\boldsymbol{n}_m \mid \boldsymbol{p}_m \sim \text{Multinomial}\left(\boldsymbol{p}_m\right). \tag{22}$$

Probability vector $\boldsymbol{p}_m$ is the latent proportion of overdose deaths in each state-year associated with each drug type. As observed in Figure 2, variation in $\boldsymbol{p}_m$ at the state-year level is substantial, motivating a statistical model for $\boldsymbol{p}_m$ itself. We elicit conditionally independent Dirichlet prior distributions for each $\boldsymbol{p}_m$,

$$\boldsymbol{p}_m \mid \boldsymbol{\alpha} \sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right). \tag{23}$$

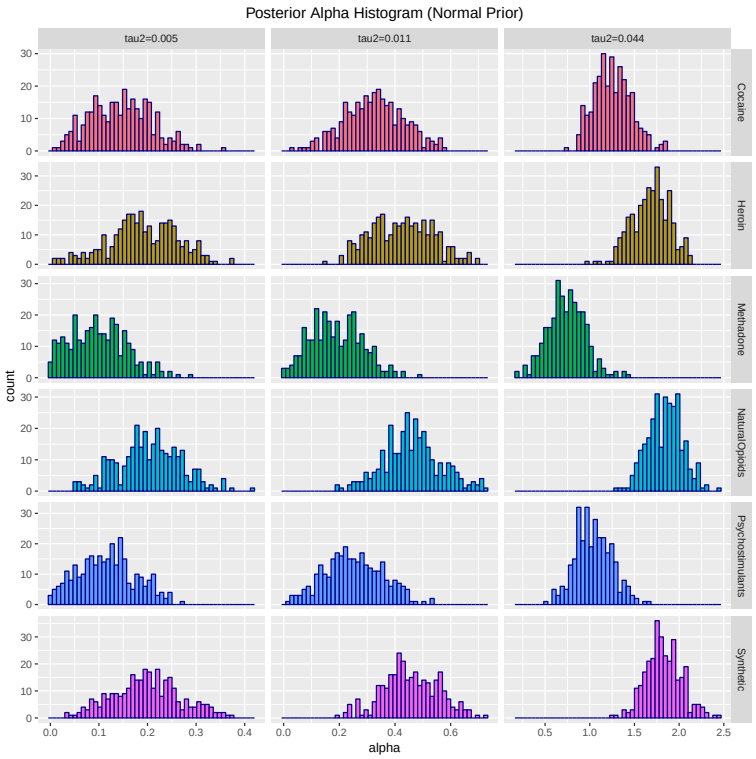

Figure 3: Posterior distributions for the concentration parameters $\alpha_k|n_1, \ldots, n_{76}$. Columns are posteriors under different prior choices of τ^2. The left column corresponds to the choice of τ such that $E[\alpha_k] = \frac{1}{3K}$; the middle column $E[\alpha_k] = \frac{1}{2K}$; and the right column $E[\alpha_k] = \frac{1}{K}$. Each row corresponds to the posterior α_k for one type of drug.

This hierarchical formulation enables the proportion of deaths associated with drug types to

vary significantly from one state-year to the next. The objective is to predict the proportion of deaths in a new state, $\boldsymbol{p}^*$, associated with each drug type. The full posterior predictive distribution for $\boldsymbol{p}^* \mid n_1, \ldots, n_M$ quantifies (with uncertainty) the relative proportion of abuse-related deaths by drug type at the aggregate level.

Following the algorithmic development in Theorem 2, we elicit truncated normal priors for $\boldsymbol{\alpha}$ with different expectations $E[\alpha_k]$ and compare the inferences (see Figure 3). Recall that when $\alpha_k \sim TN(0, \tau^2)\mathbf{I}(\alpha_k > 0)$, the prior expectation is $E[\alpha_k] = \sqrt{\frac{2}{\pi}}\tau$. Observe in Figure 3 that as prior variance τ^2 increases (with the smallest τ^2 in the left column), posterior estimates of α increase in magnitude, reflecting the larger prior mean. The posteriors become more diffuse as well, which reflects the increased prior variance. We also see in Figure 3 that the posterior distributions for the concentration parameters associated with heroin, natural opiods, and synthetic opiods are relatively larger than the concentration parameters for cocaine, methadone, and psychostimulants. As the expected value of the concentration parameter increases from $\frac{1}{3K}$ in the left column of Figure 3 to $\frac{1}{K}$ in the right column, the separation of the posteriors becomes more pronounced.

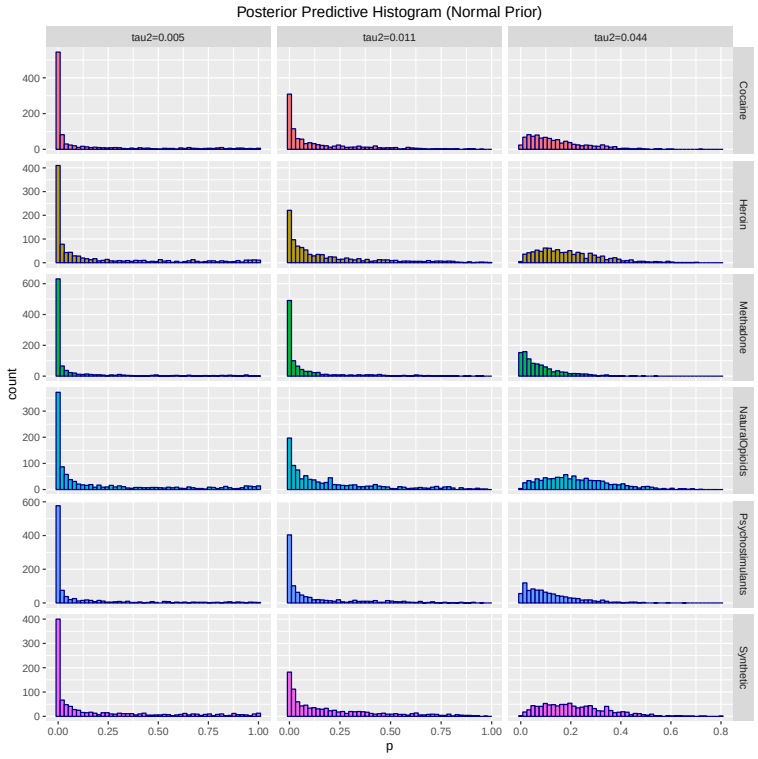

Figure 4: Posterior predictive distribution of $\boldsymbol{p}$, with normal prior on α. $\tau = 0.209, 0.104$, and 0.0696, same as Figure 3.

Computing the posterior for each α_k facilitates computation of a posterior predictive distribution for the proportion of deaths in a new state-year associated with each drug type. Rather than fixing α_k, we estimate the α_k and propagate uncertainty in the concentration parameters to predictions about the proportion of deaths attributed to each drug. Figure 4 illustrates that the posterior predicted proportion of deaths associated with heroin, natural opioids, and synthetic opioids is again relatively larger than the predictive values for cocaine, methadone, and psychostimulants. This is particularly true in the right hand column of Figure 4, where the prior mean of each concentration parameter is $E[\alpha_k] = \frac{1}{K}$, the prior specification which places the most predictive mass in the middle of the unit interval. By contrast, the $E[\alpha_k] = \frac{1}{3K}$ and $E[\alpha_k] = \frac{1}{2K}$ prior choices place significant predicted mass at the ends of the unit interval. Observe in the left and middle columns of Figure 4 that the predictive distributions are overly concentrated near zero, while the predictive distribution in the right column is more evenly spread along the unit interval.

6 Discussion

The Pólya-inverse Gamma distribution facilitates fully Bayesian posterior inference for concentration and shape parameters in Dirichlet and gamma statistical models, respectively. The P-IG distribution class is flexible and admits fast and efficient stochastic simulation methods in widely-used statistical models, such as latent Dirichlet allocation, Gamma-Gamma hierarchical models, and Bayesian nonparametric mixture models. The P-IG$(\boldsymbol{d}, c)$ distribution is constructed from an infinite convolution of GIG distributions and includes the E-RG distribution as a special case. It is the E-RG case that relates ratios of gamma functions to the Laplace transform of the P-IG distribution class, providing an efficient data augmentation strategy. Our parameter expanded Gibbs sampler leverages the scale mixture of normals representation of the E-RG distribution to estimate parameters nested in the gamma function.

The focus of the current paper is on theoretical and algorithmic development of the P-IG distribution class. Our work builds on distributional results of Hartman (1976) and Roynette and Yor (2005) and contributes to the literature on scale mixtures of normals (see, e.g., Andrews and Mallows (1974); West (1987); Polson et al. (2013)). We believe that the computational strategies developed here will provide the foundation for new and richly structured hierarchical models of Dirichlet concentration and gamma shape parameters. Applied Bayesian analyses of categorical data will benefit from increased model flexibility and information borrowing strategies.

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

References

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

Antoniak, C. E. (1974, 11). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Ann. Statist. 2*(6), 1152–1174.

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

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. (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.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Ferguson, T. S. (1973, 03). A bayesian analysis of some nonparametric problems. *Ann. Statist. 1*(2), 209–230.

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.

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.

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*.

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.

Rossell, D. et al. (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.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. *Working paper, Duke University*.

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

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

A MCMC for the case of homogeneous α

Under the special case $\alpha_1 = \alpha_2 = \cdots = \alpha_K = \alpha$. Need to learn α,

$$\frac{e^{-\gamma\alpha}}{\Gamma(\alpha+1)} = \prod_{k=1}^{\infty}\left(1+\frac{\alpha}{k}\right)e^{-\frac{\alpha}{k}} \tag{24}$$

From Roynette and Yor (2005) (IV 59)

$$\frac{\Gamma(\lambda)}{\Gamma(\lambda+\alpha)} = e^{-\alpha\psi(\lambda)}\prod_{k=1}^{\infty}\left(1+\frac{\alpha}{\lambda+k-1}\right)e^{-\frac{\alpha}{\lambda+k-1}} \tag{25}$$

By definition of PIG

$$E[e^{-\alpha^2 w}] = \frac{\Gamma(\lambda)}{\Gamma(\lambda+\alpha)}e^{\alpha\psi(\lambda)}, \tag{26}$$

where $w \sim PIG(\boldsymbol{d}, 0)$ and $d_k = \lambda + k - 1$. Let $n = \sum_{k=1}^{K} n_k$.

Theorem 3. *The Gibbs sampler for homogeneous α model is given by*

$$\begin{aligned}
\eta \mid \alpha, \boldsymbol{p} &\sim \Gamma(K\alpha + n, 1) \\
w_k \mid \alpha, \boldsymbol{p} &\sim PIG(\boldsymbol{d}, \sqrt{2(n_k+\alpha-1)^2}), \quad d_k = k \\
\boldsymbol{p} \mid \alpha &\sim Dir(n_1+\alpha, \cdots, n_K+\alpha) \\
\alpha \mid \eta, \boldsymbol{w} &\sim TN\left(\frac{b}{2a}, \frac{1}{2a}\right)\mathbf{I}(\alpha_k > 0)
\end{aligned}$$

where the value of a and b depend on form of prior $p(\alpha)$, TN for truncated normal with $\alpha_k > 0$. Truncated Normal prior $p(\alpha) \sim N(0, \tau^2)\mathbf{I}(\alpha_k > 0)$,

$$\begin{aligned}
a &= \sum_{k=1}^{K} w_k + \frac{1}{2\tau^2} \\
b &= -2\sum_i (n_k - 1)w_k + K\log\eta + \sum_i(\gamma + \log p_k).
\end{aligned}$$

Proof. Bayes rule requires the conditionals

$$
\begin{aligned}
f(\boldsymbol{n} \mid \boldsymbol{p}) &\propto p_1^{n_1} \cdots p_K^{n_K} \\
p(\boldsymbol{p}, \alpha \mid \boldsymbol{n}) &\propto \frac{\Gamma(K\alpha + n)}{\prod_{k=1}^K \Gamma(n_k + \alpha)} \prod_{k=1}^K p_k^{n_k+\alpha-1} \\
&= \Gamma(K\alpha + n) \prod_{k=1}^K \left[\frac{1}{\Gamma(n_k + \alpha)} p_k^{n_k+\alpha-1} \right] \\
&= \Gamma(K\alpha + n) \prod_{k=1}^K \left[\frac{1}{\Gamma((n_k + \alpha - 1) + 1)} e^{-\gamma(n_k+\alpha-1)} e^{(\gamma + \log p_k)(n_k+\alpha-1)} \right] \\
&= \int_0^\infty \eta^{K\alpha+n-1} e^{-\eta} d\eta \prod_{k=1}^K \int_0^\infty e^{-(n_k+\alpha-1)^2 w_k} p(w_k) dw_k \prod_{k=1}^K e^{(\gamma+\log p_k)(n_k+\alpha-1)} \\
&= \int p(\alpha, \boldsymbol{w}, \eta \mid \boldsymbol{p}) d\eta d\boldsymbol{w}
\end{aligned}
$$

Therefore, the augmented conditional posterior is

$$
p(\alpha, \boldsymbol{w}, \eta \mid \boldsymbol{p}) = \eta^{K\alpha+n-1} e^{-\eta} \prod_{k=1}^K e^{-(n_k+\alpha-1)^2 w_k} p(w_k) \prod_{k=1}^K e^{(\gamma+\log p_k)(n_k+\alpha-1)} p(\alpha)
$$

Conditional on η and w, distribution of α

$$
\begin{aligned}
p(\alpha \mid \eta, \boldsymbol{w}) &\sim \exp\left(\alpha K \log \eta - \sum_i (n_k + \alpha - 1)^2 w_k + \sum_i (\gamma + \log p_k)(n_k + \alpha - 1) \right) p(\alpha) \\
&\sim \exp\left(- \left(\sum_i w_k \right) \alpha^2 + \left(-2 \sum_i (n_k - 1) w_k + K \log \eta + \sum_i (\gamma + \log p_k) \right) \alpha \right) p(\alpha) \\
&:= \exp(-a\alpha^2 + b\alpha) \\
&\propto \exp\left(- \frac{\left(\alpha - \frac{b}{2a}\right)^2}{1/a} \right)
\end{aligned}
$$

Truncated Normal prior $p(\alpha) \sim N(0, \tau^2)\mathbf{I}(\alpha_k > 0)$,

$$
\begin{aligned}
a &= \sum_{k=1}^K w_k + \frac{1}{2\tau^2} \\
b &= -2 \sum_i (n_k - 1) w_k + K \log \eta + \sum_i (\gamma + \log p_k).
\end{aligned}
$$

which is a normal distribution with mean $b/2a$ and variance $1/2a$. □