Bayesian Analysis (2022) **17**, Number 2, pp. 545–564

On Global-Local Shrinkage Priors for Count Data[*]

Yasuyuki Hamura[†], Kaoru Irie[‡], and Shonosuke Sugasawa[§]

Abstract. Global-local shrinkage priors have been recognized as a useful class of priors that can strongly shrink small signals toward prior means while keeping large signals unshrunk. Although such priors have been extensively discussed under Gaussian responses, in practice, we often encounter count responses. Previous contributions on global-local shrinkage priors cannot be readily applied to count data. In this paper, we discuss global-local shrinkage priors for analyzing a sequence of counts. We provide sufficient conditions under which the posterior mean is unshrunk for very large signals, known as the tail robustness property. Then, we propose tractable priors to satisfy those conditions approximately or exactly and develop a custom posterior computation algorithm for Bayesian inference without tuning parameters. We demonstrate the proposed methods through simulation studies and an application to a real dataset.

Keywords: heavy tailed distribution, Markov Chain Monte Carlo, Poisson distribution, tail robustness.

1 Introduction

High-dimensional count data appear in various scientific fields, including genetics, epidemiology, and social science. Frequently, in such data, many counts are moderate except for some outliers (very large counts). For example, in crime statistics, the number of occurrences of a specific crime is likely to be small or moderate in many regions. Yet, one observes several regions with unexplained high crime rate. Detecting such "hotspots" is undoubtedly of interest in crime statistics. In this context, using a Poisson-gamma model is inappropriate as the gamma prior shrinks all observations uniformly, including the large signals. Hence, meaningful regions with large signals may be overlooked. A desirable prior should therefore account for both small and large signals, and realize flexible shrinkage effects on Poisson rates.

This type of priors has been studied as global-local shrinkage priors for Gaussian observations. Most notably, the horseshoe prior (Carvalho et al., 2010) has been proposed to detect sparse signals in high-dimensional continuous observations. In hierarchical models, these priors have been adopted for random effect distributions in small area

arXiv: 1907.01333

[*] The authors are supported by the Japan Society for the Promotion of Science (JSPS KAKENHI) grant numbers: 20J10427, 17K17659, and 18K12757.

[†] Graduate School of Economics, The University of Tokyo. JSPS Research Fellow, yasu.stat@gmail.com

[‡] Faculty of Economics, The University of Tokyo, irie@e.u-tokyo.ac.jp

[§] Center for Spatial Information Science, The University of Tokyo, sugasawa@csis.u-tokyo.ac.jp

https://doi.org/10.1214/21-BA1263

estimation (Tang et al., 2018) or default Bayesian analysis (Bhadra et al., 2016). For recent developments, see Bhadra et al. (2019) and the references therein.

While extensively studied for Gaussian data, global-local shrinkage priors have not been fully developed for count data. This is despite the wide usage of hierarchical Poisson models in applications such as disease mapping (e.g., Wakefield, 2006; Lawson, 2013). The theory related to the Poisson likelihoods has been well developed (e.g., Brown et al., 2013; Yano et al., 2018), but not necessarily from the viewpoint of global-local shrinkage. The conjugate gamma priors for the Poisson rate parameters have been routinely adopted in practice: the global-local shrinkage is modeled via hyperpriors for the gamma scale parameters (e.g., Zhu et al., 2019). In this context, Datta and Dunson (2016) studied the use of the generalized hypergeometric distribution for the scale parameter and its shrinkage effect, focusing on the analysis of zero-inflated count data. In contrast, our research is concerned with the heavy-tail property of global-local shrinkage priors, which ensures large signals are exempt from being shrunk.

Our objective is to consider the effect of the hyperprior on the posterior means of Poisson rates in terms of their robustness. To do this, we first define the mathematical concept of tail-robustness for the Bayes estimators. A robust Bayes estimator should keep large signals unshrunk while strongly shrinking the small signals toward prior means. We formally define this concept as *tail-robustness* in equation (2.3). Sufficient conditions for tail-robustness are given in Theorem 2.1 and Corollary 1.

Our class of priors to be utilized is restricted by the tail-robustness requirement for the Bayes estimators. The conditions in Theorem 2.1 reveal the importance of local shrinkage induced by the individual scale parameter of the gamma distribution customized for each Poisson rate. The theorem supports the use of two classes of hyperpriors proposed in Section 3: the inverse-gamma prior and the newly introduced extremely heavy-tailed prior. The inverse-gamma prior is a well-known distribution and can be easily integrated into the model. The asymptotic bias for large signals is shown to be negligible. Hence, the inverse-gamma prior is "approximately" tail-robust. The extremely heavily tailed prior is a new class of probability distributions that, in contrast to the inverse-gamma prior, satisfies the conditions for tail-robustness and is exactly tail-robust. Both priors are conditionally conjugate for most parameters in the model. This allows for a fast and efficient posterior analysis using the Gibbs sampler.

In the numerical study of our paper, we demonstrate the theoretically guaranteed properties of tail-robustness for those priors in a setting where the standard Poisson-gamma model suffers from over-shrinkage of the Bayes estimators in the presence of outliers. We empirically show the differences between the two proposed priors: the inverse-gamma prior is better in the point estimations for small signals and has a greater shrinkage effect toward the prior mean; meanwhile, the extremely heavy-tailed prior is successful in quantifying the uncertainty for large counts, as shown in the coverage rates of posterior credible intervals. Despite this difference, both priors perform almost equally in the analysis of the actual crime data in Japan by detecting the crime hotspots that are overlooked by the Poisson-gamma models.

The rest of the paper is organized as follows. In Section 2, we define the model and tail-robustness, and derive sufficient conditions for local priors to satisfy the tail-robustness. In Section 3, we propose two local priors and provide efficient posterior computation algorithms using Gibbs sampling. We further discuss some properties of the implied marginal priors and posteriors of the Poisson rate. Section 4 describes the numerical experiments for the extensive comparison of the proposed priors and other commonly used priors/estimators under various settings. Section 5 presents an application using the real crime data of the Tokyo metropolitan area in Japan. The R code implementing the proposed method is available at our GitHub repository (https://github.com/sshonosuke/GLSP-count).

2 Tail-robustness under count responses

2.1 Hierarchical models for count data

Our model has the following hierarchical representation, where m observations $y_1, \dots, y_m$ are conditionally independent and modeled as,

$$y_i|\lambda_i \sim \mathrm{Po}(\eta_i \lambda_i), \quad \lambda_i|u_i \sim \mathrm{Ga}(\alpha, \beta/u_i), \quad u_i \sim \pi(u_i), \tag{2.1}$$

for $i = 1, \dots, m$, where $Po(\eta_i\lambda_i)$ is the Poisson distribution with rate $\eta_i\lambda_i$, and $Ga(\alpha, \beta/u_i)$ the gamma distribution with shape α and rate β/u_i whose (conditional) mean is $u_i\alpha/\beta$. In addition, $\eta_i \in (0, \infty)$ is a known offset, $(\alpha, \beta) \in (0, \infty)^2$ are the hyperparameters, and $u_i \in (0, \infty)$ is a local scale parameter. The offset term, η_i, can be a structured part characterizing dependence on covariates, as we will examine in Section 5. Next, we set $\eta_i = 1$ for simplicity. The two rate parameters of the gamma prior, β and u_i^{-1}, control the global and local shrinkage effects, respectively. Under this model, the Bayes estimator of Poisson rate λ_i is the posterior mean

$$\begin{aligned}\widetilde{\lambda}_i &= \mathrm{E}\Big[\frac{u_i}{\beta + u_i}(\alpha + y_i)\Big|y_i\Big] \\ &= y_i - \mathrm{E}\Big[\frac{\beta}{\beta + u_i}\Big(y_i - \frac{\alpha u_i}{\beta}\Big)\Big|y_i\Big],\end{aligned} \tag{2.2}$$

where the expectation is taken with respect to the marginal posterior of u_i so that the conditional posterior mean of λ_i shrinks y_i toward the prior mean $\alpha u_i/\beta$. Throughout the paper, we consider proper priors for u_i only. The use of improper priors for u_i results in an improper marginal of λ_i. Furthermore, the posterior distribution of λ_i will not successfully reflect the prior information and will fail to produce a satisfactory Bayes estimator.

2.2 Tail-robustness of the posterior mean

We consider the appropriate choice of the prior $\pi(u_i)$ in terms of the shrinkage effect realized in the Bayes estimator $\widetilde{\lambda}_i$. As stated in the introduction, the estimator should not shrink toward the prior mean when a large signal is observed. This property is named as tail-robustness (e.g., Carvalho et al., 2010). The tail-robustness is mathematically defined by the following condition:

$$\lim_{y_i \to \infty} |\widetilde{\lambda}_i - y_i| = 0. \tag{2.3}$$

This means that the (mean) absolute error loss tends to zero as $y_i \to \infty$. For fixed u_i, the Bayes estimator $(\alpha + y_i)/(1 + \beta/u_i)$ clearly loses the tail-robustness. This motivates the hierarchical model with the prior $\pi(u_i)$ on u_i. Throughout this paper, our primary interest is in the property defined in equation (2.3). Nevertheless, there are other definitions of tail-robustness related to various loss functions. We discuss this issue in detail in Section S3 of the Supplementary Materials (Hamura et al., 2021).

To consider the tail-robustness, the next theorem is useful in evaluating the asymptotic bias $\lim_{y_i \to \infty}(\widetilde{\lambda}_i - y_i)$ for a variety of priors. The proof is given in Sections S1 and S2 of the Supplementary Materials.

Theorem 2.1. *Assume that $\pi(\cdot)$ is strictly positive and continuously differentiable. Suppose that $\pi(\cdot)$ satisfies the following two conditions:*

$$\int_0^1 |u\pi'(u)| du < \infty, \tag{A1}$$

$$\xi \equiv \lim_{u \to \infty} \frac{u\pi'(u)}{\pi(u)} \quad \textit{exists in } [-\infty, \infty]. \tag{A2}$$

Then the asymptotic bias of $\widetilde{\lambda}_i$ is $1 + \xi$, that is,

$$\lim_{y_i \to \infty}(\widetilde{\lambda}_i - y_i) = 1 + \xi.$$

The asymptotic bias of $\widetilde{\lambda}_i$ under $y_i \to \infty$ can be characterized by the tail behavior of the mixing distribution $\pi(\cdot)$. This condition is similar to but significantly different from that of Gaussian response (e.g., Tang et al., 2018). From Theorem 2.1, it follows that $\xi = -1$ is the sufficient condition for the estimator to be tail-robust, and this is summarized in the following corollary.

Corollary 1. *Under the conditions (A1) and*

$$\lim_{u \to \infty} \frac{u\pi'(u)}{\pi(u)} = -1, \tag{A3}$$

the Bayes estimator $\widetilde{\lambda}_i$ is tail-robust and satisfies $|\widetilde{\lambda}_i - y_i| \to 0$ as $y_i \to \infty$.

The crucial assumption in the above corollary is (A3). This describes the desirable tail behavior of the prior distribution of u_i. In fact, (A3) is sufficient for $\psi(u) = u\pi(u)$ to be slowly varying as $u \to \infty$, that is, $\lim_{u \to \infty} \psi(\kappa u)/\psi(u) = 1$ for all $\kappa > 0$ (e.g., see Seneta, 1976, equation (1.11)). This implies that, for the marginal prior $p(\lambda_i) = \int_0^\infty Ga(\lambda_i|\alpha, \beta/u_i)\pi(u_i)du_i$, we have $\lambda_i p(\lambda_i) \sim \lambda_i \pi(\lambda_i)$ as $\lambda_i \to \infty$ under the regularity condition that justifies the interchange of the limit and integral. In other words, under this assumption, the marginal densities of λ_i and u_i are asymptotically equivalent in their tails.

An example of priors that satisfies assumption (A3) is $\pi(u) \propto 1/u$. In many cases, (A3) requires priors to be of this form; see Section S4 of the Supplementary Materials for more details. However, this prior is improper. In other words, $\pi(\cdot)$ must be extremely heavy-tailed for $\widetilde{\lambda}_i$ to be tail-robust. In contrast, (A1) is merely a technical requirement for the proof.

One notable feature imposed by Corollary 1 is that the sufficient conditions for the tail-robustness, (A1) and (A3), are independent of the values of hyperparameters α and β. This setting about the hyperparameters greatly differentiates our proposed approach from those in other studies, for example, Proposition 1 of Datta and Dunson (2016), where the tail-robustness is discussed for the limiting values of hyperparameters, that is, $\beta \to \infty$ or $\beta \to 0$.

3 Global-local shrinkage priors for count data

3.1 Proposed priors

Under the hierarchical model (2.1), we propose two families of priors for u_i. Each is indexed by a hyperparameter $\gamma \in (0, \infty)$, which can be estimated in practice.

The first prior is the inverse gamma (IG) prior given by following density:

$$\pi_{\text{IG}}(u_i; \gamma) = \frac{\gamma^\gamma}{\Gamma(\gamma)} \frac{1}{{u_i}^{1+\gamma}} e^{-\gamma/u_i}, \tag{3.1}$$

where $\gamma > 0$. This prior is denoted by $\text{IG}(\gamma, \gamma)$. It is proper and conditionally conjugate, which simplifies the posterior computation by Markov Chain Monte Carlo (MCMC) methods. From Theorem 2.1, it follows that $\lim_{y_i \to \infty}(\widetilde{\lambda}_i - y_i) = -\gamma$. This indicates that the IG prior approximately satisfies the tail-robustness when γ is small. Both shape and rate parameters of the proposed IG prior are γ, so that we have $E[1/u_i] = 1$ and the parameter β in equation (2.1) is identified as the global shrinkage factor.

Next, we introduce the extremely heavy-tailed (EH) prior, defined by the following density

$$\pi_{\text{EH}}(u_i; \gamma) = \frac{\gamma}{1 + u_i} \frac{1}{\{1 + \log(1 + u_i)\}^{1+\gamma}}, \tag{3.2}$$

for $\gamma > 0$. The EH prior can be seen as a modification of the scaled-beta prior (Armagan et al., 2011); the details on the connection to the scaled-beta prior are discussed in Section S4 of the Supplementary Materials. The additional logarithm function in equation (3.2) contributes to the integrability of the density function. The use of log-terms is often seen in the literature on decision-theoretic statistical theory (for example, see Maruyama and Strawderman 2020, Remark 4.1). This prior is proper because

$$\int_0^\infty \pi_{\text{EH}}(u; \gamma) du = \Big[- \{1 + \log(1 + u)\}^{-\gamma} \Big]_0^\infty = 1.$$

The notable property of the EH prior is that this distribution is exactly tail-robust. This is because it satisfies the condition of Corollary 1 since

$$\frac{u {\pi_{\text{EH}}}'(u; \gamma)}{\pi_{\text{EH}}(u; \gamma)} = u \Big\{ - \frac{1}{1+u} - \frac{1+\gamma}{1 + \log(1+u)} \frac{1}{1+u} \Big\} \to -1,$$

as $u \to \infty$.

The densities and tail-behaviors of the proposed priors are summarized in Table 1 together with those of the Gauss hypergeometric (GH) prior considered in Datta and Dunson (2016). The GH prior is dependent on the global rate parameter β, but its density tail (the asymptotic functional form of the density as $u_i \to \infty$) is independent of β and identical to that of the half-Cauchy prior (Carvalho et al., 2010). The density tail of the EH prior is heavier than those of the GH and IG priors regardless of γ. This difference originates from the log-term in the EH density and is essential for the exact tail-robustness of the EH prior.

	Density kernel of u_i	Density tail as $u_i \to \infty$
$GH(1/2, 1/2, \gamma, 1/\beta)$	$u_i^{-1/2}(1+u_i)^{-\gamma}(\beta+u_i)^{\gamma-1}$	$u_i^{-3/2}$
$IG(\gamma, \gamma)$	$u_i^{-(\gamma+1)}e^{-\gamma/u_i}$	$u_i^{-(\gamma+1)}$
$EH(\gamma)$	$(1+u_i)^{-1}\{1+\log(1+u_i)\}^{-(1+\gamma)}$	$u_i^{-1}(\log u_i)^{-(1+\gamma)}$

Table 1: Densities of GH, IG, and EH priors.

Finally, we note the parametrization by $\kappa = 1/(1+u) \in (0,1)$, which also clarifies the difference between the proposed priors and others. The implied density of the EH prior in the scale of κ is $\pi_{EH}(\kappa) = \gamma\kappa^{-1}/\{1+\log(1/\kappa)\}^{1+\gamma}$. This expression shows that the EH prior can be viewed as an extension of the improper beta prior, $Be(0,1)$. The resulting EH prior is proper; the additional log-term in the density of the EH prior ensures the propriety. The other priors, including the half-Cauchy prior, remain in the class of beta distributions in κ-scale and do not have log-terms in their densities.

3.2 Posterior computation

The computation of the Bayes estimator is based on MCMC methods. Because the proposed priors are mostly conditionally conjugate, sampling from most conditional posterior distributions is straightforward. Here, we outline the posterior sampling procedure with the proposed priors. The detailed step-by-step Gibbs sampler algorithm (partially collapsed Gibbs sampler, van Dyk and Park, 2008) is described in Section S7 of the Supplementary Materials.

We first discuss the parameters $(\lambda_{1:m}, \alpha, \beta)$, which are included in all models regardless of the choice of priors for u_i. In practice, we assign prior distributions for α and β in practice. Here, we consider the gamma priors; $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$ and $\beta \sim \mathrm{Ga}(a_\beta, b_\beta)$. We set $a_\alpha = b_\alpha = a_\beta = b_\beta = 1$ as default and will use this choice of the hyperparameters in the numerical studies in Sections 4 and 5. While the gamma prior for β is conditionally conjugate, the gamma prior for α is not. However, using the augmentation technique by Zhou and Carin (2013), we can derive an efficient Gibbs sampling method as described in Section S7 of the Supplementary Materials.

For the model with the IG prior, the scale parameter u_i has a known conditional posterior, while the conditional posterior of the hyperparameter γ is difficult to directly sample from. Several computationally sophisticated options are available for the sampling of γ. However, we simply use the random-walk Metropolis-Hastings method with the uniform prior $\gamma \sim U(\varepsilon_1, \varepsilon_2)$ for fixed small $\varepsilon_1 > 0$ and large $\varepsilon_2 > 0$. We set $\varepsilon_1 = 0.001$ and $\varepsilon_2 = 150$ as default.

The new EH prior is not conditionally conjugate for u_i, despite its simple closed form density function in equation (3.2). To develop an efficient sampling algorithm, we introduce a novel augmentation approach using two positive valued latent variables v_i and w_i, given by the following integral formula:

$$\pi_{\mathrm{EH}}(u_i;\gamma) = \iint_{(0,\infty)^2} \pi_{\mathrm{EH}}(u_i, v_i, w_i;\gamma) dv_i dw_i,$$

where

$$\begin{aligned}\pi_{\mathrm{EH}}(u_i, v_i, w_i;\gamma) &= Ga(u_i|1, v_i)Ga(v_i|w_i, 1)Ga(w_i|\gamma, 1)\\ &= \frac{w_i^{\gamma-1} v_i^{w_i}}{\Gamma(\gamma)\Gamma(w_i)} \exp\left\{-w_i - v_i(1+u_i)\right\}.\end{aligned}$$

Using the above expression, we see that the full conditional distribution of u_i is the generalized inverse Gaussian (GIG) distribution. We can also obtain familiar forms of the conditional posterior distributions of the other parameters, (v_i, w_i). Further details are in Section S7 of the Supplementary Materials. For the shape parameter γ in the EH prior, we use the gamma prior $\gamma \sim \mathrm{Ga}(a_\gamma, b_\gamma)$ which is conditionally conjugate. We set $a_\gamma = b_\gamma = 1$ for simplicity.

3.3 Marginal prior distributions for λ_i

Here, we consider the behavior of the marginal density of λ_i in the limit of $\lambda_i \to \infty$ and $\lambda_i \to 0$. Note that information on the behavior of the marginal density of λ_i around zero is also important to understand the amount of shrinkage effect toward zero. We discuss this here.

Under a general prior $\pi(u_i;\gamma)$, the marginal prior distribution for λ_i is given by

$$\begin{aligned}p(\lambda_i;\alpha,\beta,\gamma) &= \int_0^\infty \frac{\beta^\alpha/{u_i}^\alpha}{\Gamma(\alpha)} {\lambda_i}^{\alpha-1} e^{-(\beta/u_i)\lambda_i} \pi(u_i;\gamma) du_i\\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{1}{x^\alpha} e^{-\beta/x} \pi(\lambda_i x;\gamma) dx.\end{aligned}$$

We continue the computation of this density for the two classes of priors: π_{IG} and π_{EH}.

For the IG prior $\pi(u_i;\gamma) = \pi_{\mathrm{IG}}(u_i;\gamma)$, we have

$$p(\lambda_i;\alpha,\beta,\gamma) = \frac{(\beta/\gamma)^\alpha}{B(\alpha,\gamma)} \frac{{\lambda_i}^{\alpha-1}}{\{1+(\beta/\gamma)\lambda_i\}^{\alpha+\gamma}}.$$

This implies the beta distribution, namely, $(\beta/\gamma)\lambda_i/\{1+(\beta/\gamma)\lambda_i\} \sim \mathrm{Beta}(\alpha,\gamma)$. Regarding the tail property of the marginal density, we have $p(\lambda_i;\alpha,\beta,\gamma) = O({\lambda_i}^{-1-\gamma})$ as $\lambda_i \to \infty$. For a sufficiently small γ, the marginal prior of λ_i can be heavily tailed and almost equivalent to λ_i^{-1} in the tail. This observation is coherent with the γ-dependent asymptotic bias of the Bayes estimator, $\lim_{y_i\to\infty}(\widetilde{\lambda}_i - y_i) = -\gamma$.

Note that due to the heavy tail of this density, the prior mean of λ_i does not exist if $\gamma \leq 1$, as is easily verified by the direct computation. In this situation, it is difficult to interpret the prior from the viewpoint of shrinkage. This is because the prior mean (to which the estimator is shrunk) does not exist. For those who prefer priors with finite means, we recommend the modification of the IG prior to $\mathrm{IG}(\gamma+1,\gamma)$, $\gamma > 0$. This instead increases the asymptotic bias slightly to $-\gamma-1$. In contrast, the density at the origin depends on the value of α. In particular, $\lim_{\lambda_i \to 0} p(\lambda_i;\alpha,\beta,\gamma) = \infty$ for $\alpha < 1$, while the limit becomes a positive constant for $\alpha = 1$ and zero for $\alpha > 1$. This helps in interpreting the choice of, or the posterior inference for, hyperparameter α.

For the EH prior, the marginal density is evaluated around zero as follows. For $\pi(u_i;\gamma) = \pi_{\mathrm{EH}}(u_i;\gamma)$, we have

$$\begin{aligned} p(\lambda_i;\alpha,\beta,\gamma) &= \frac{\beta^\alpha \gamma}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^\alpha} \frac{1}{1+\lambda_i x} \frac{1}{\{1+\log(1+\lambda_i x)\}^{1+\gamma}} dx \\ &\to \frac{\beta^\alpha \gamma}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^\alpha} dx \\ &= \begin{cases} (\alpha-1)^{-1}\beta\gamma & \text{if } \alpha > 1 \\ \infty & \text{if } \alpha \leq 1 \end{cases} \end{aligned}$$

as $\lambda_i \to 0$, by the monotone convergence theorem. Thus, $\lim_{\lambda_i \to 0} p(\lambda_i;\alpha,\beta,\gamma) > 0$ and non-decreasing as $\alpha \to 0$, implying the stronger shrinkage of small signals toward the global prior mean for small α. For the tail property, we have

$$\begin{aligned} \lim_{\lambda_i \to \infty} \frac{p(\lambda_i;\alpha,\beta,\gamma)}{\pi_{\mathrm{EH}}(\lambda_i;\gamma)} &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^\alpha} \left[\lim_{\lambda_i \to \infty} \frac{1+\lambda_i}{1+\lambda_i x} \left\{ \frac{1+\log(1+\lambda_i)}{1+\log(1+\lambda_i x)} \right\}^{1+\gamma} \right] dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^{\alpha+1}} dx = 1. \end{aligned}$$

Therefore, $p(\lambda_i;\alpha,\beta,\gamma) \sim \pi_{\mathrm{EH}}(\lambda_i;\gamma) \sim \gamma {\lambda_i}^{-1} (\log \lambda_i)^{-1-\gamma}$ as $\lambda_i \to \infty$. This means that the marginal prior $p(\lambda_i;\alpha,\beta,\gamma)$ is proper but has a sufficiently heavy tail so that the model can accommodate large signals. For the computations verifying these results, see Section S5 of the Supplementary Materials.

The marginal distributions of λ_i with $\alpha = \beta = 2$ under the proposed IG and EH priors with $\gamma = 1$ and $\gamma = 0.5$, and the GH prior with $\gamma = 1$ are visually illustrated in Figure 1. The IG prior with $\gamma = 0.5$ has the almost same tail-behavior as the GH prior. This is because the tail-behavior of the density of u_i under the IG prior with $\gamma = 1$ is equivalent to that of GH as shown in Table 1. Moreover, the density tail under the EH prior is heavier than those under the IG and GH priors, which is also consistent with Table 1.

3.4 Marginal posterior distributions for λ_i

We briefly describe the flexibility of the proposed prior distributions compared with the conventional gamma prior for λ_i. The conditional posterior distribution of λ_i given u_i is

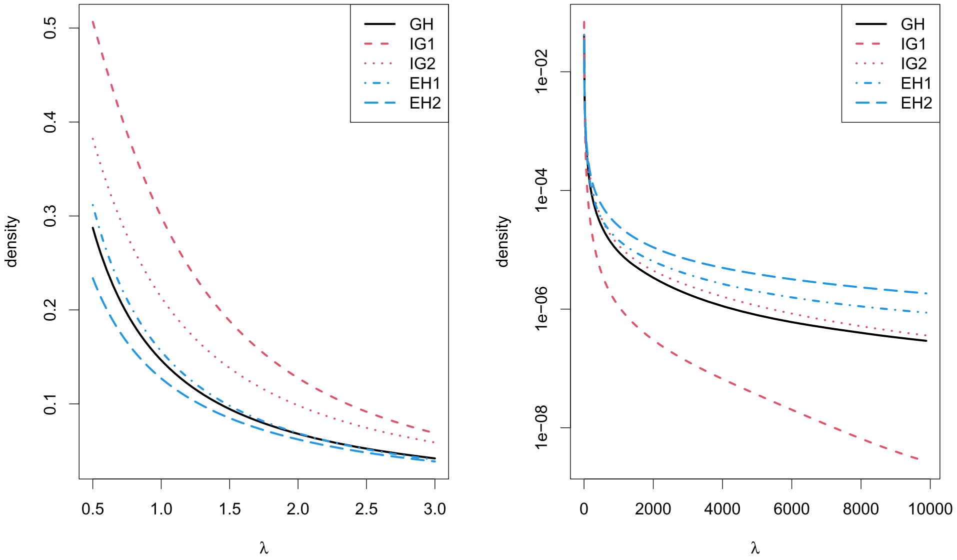

Figure 1: Left: Marginal densities of λ_i with $\alpha = \beta = 2$ under the Gauss hypergeometric prior (GH) with $\gamma = 1$, inverse-gamma priors with $\gamma = 1$ (IG1) and $\gamma = 0.5$ (IG2), and extremely heavily-tailed priors with $\gamma = 1$ (EH1) and $\gamma = 0.5$ (EH2). The GH and EH densities are evaluated by the Monte Carlo integration. Right: The marginal densities of the five prior distributions in the tail. The vertical axis is logarithmic.

$\mathrm{Ga}(y_i + \alpha, 1 + \beta/u_i)$ under the model (2.1). Therefore, the marginal posterior distribution of λ_i is obtained as the mixture of the gamma distribution with respect to the marginal posterior distribution of u_i. Note that the use of the gamma prior distribution for λ_i with no hierarchical prior (and $u_i = 1$) leads to the posterior distribution $\mathrm{Ga}(y_i + \alpha, 1 + \beta)$. We set $\alpha = \beta = 2$ and show the marginal posterior density of λ_i with several values of y_i in Figure 2. Notably, under a moderate signal, such as $y_i = 1$, the posterior distributions of λ_i are almost the same among the conventional gamma prior and the proposed global-local shrinkage priors. In contrast, under large values of y_i, the posterior densities of the proposed methods are significantly different from the one based on the gamma prior. The proposed priors flexibly shift the posterior location toward large signals, while the gamma prior over-shrinks the posterior density toward zero. As noted in the previous section, the hyperparameter γ in the inverse gamma (IG) distribution is directly related to the asymptotic bias. Furthermore, Figure 2 shows that the IG prior with the smaller γ produces heavier-tailed posterior density functions than that with the larger γ.

4 Simulation study

We here investigate the finite sample performance of the proposed method together with some existing methods. We generated the independent sequence of counts from

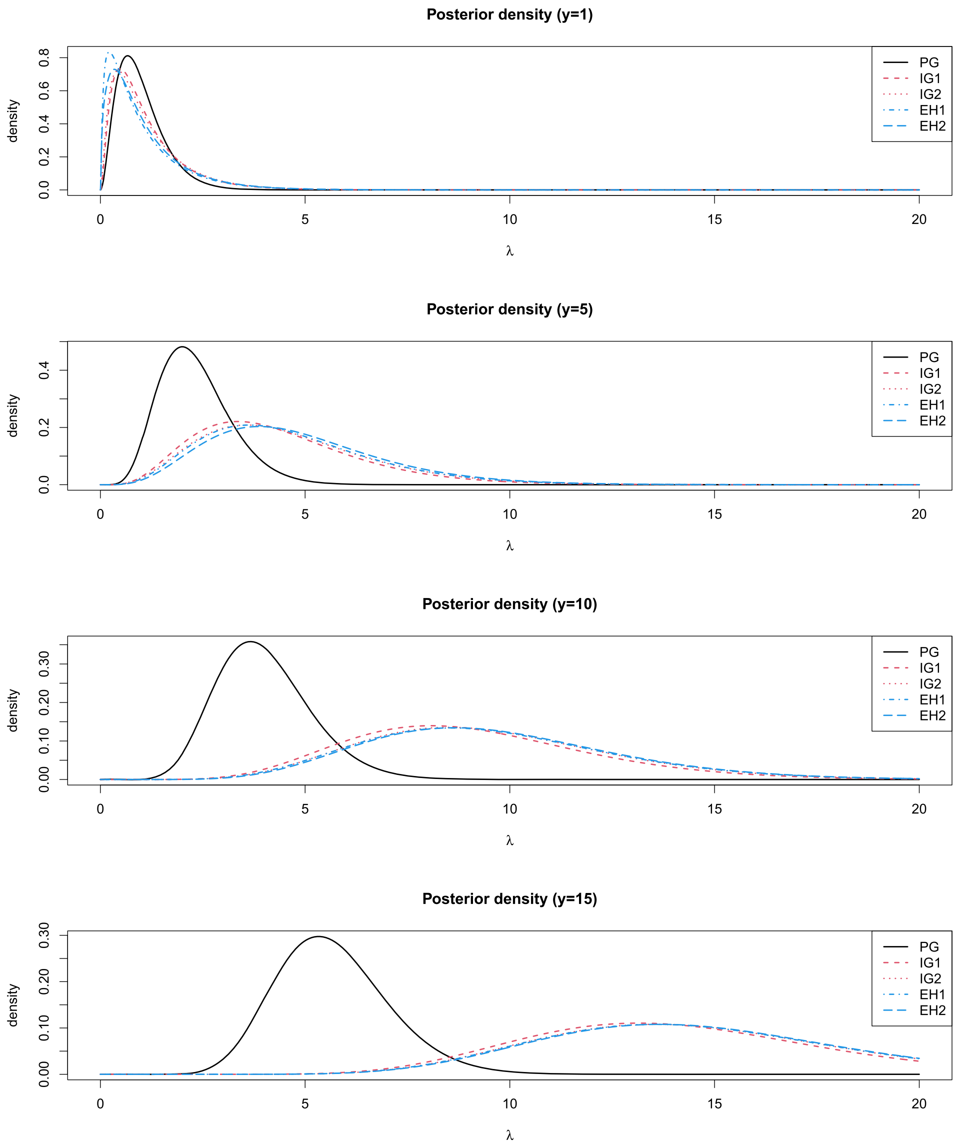

Figure 2: Marginal posterior distributions for λ_i with $\alpha = \beta = 2$ based on the conventional gamma prior (PG), the proposed inverse gamma prior with $\gamma = 1$ (IG1) and $\gamma = 0.5$ (IG2), and the proposed extremely heavy-tailed prior with $\gamma = 1$ (EH1) and $\gamma = 0.5$ (EH2). Each row corresponds to a difference value of $y_i \in \{1, 5, 10, 15\}$.

$y_i \sim \mathrm{Po}(\lambda_i \eta_i)$ for $i = 1, \ldots, m$ with $m = 200$. The adjustment term η_i was generated from $U(1,5)$, and assumed to be known. For the generating process for λ_i, we considered this mixture: $\lambda_i \sim (1-\omega) f_0 + \omega f_1$, where f_0 and f_1 denote distributions of moderate and large signals, respectively. Note that ω denotes the proportion of large signals (outliers). For the settings of f_0 and f_1, we adopted the following four scenarios:

$$\begin{aligned}
&\text{(I)} \quad f_0 = \mathrm{Ga}(2,2), \quad f_1 = \mathrm{Ga}(10,2)\\
&\text{(II)} \quad f_0 = 0.75\mathrm{Ga}(2,2) + 0.25\delta(1), \quad f_1 = \mathrm{Ga}(10,2)\\
&\text{(III)} \quad f_0 = 0.5\mathrm{Ga}(2,2) + 0.5\delta(1), \quad f_1 = \mathrm{Ga}(10,2)\\
&\text{(IV)} \quad f_0 = U(0,2), \quad f_1 = 4 + |t_3|,
\end{aligned}$$

where $\delta(1)$ is the point mass on 1, $U(0,2)$ is the uniform distribution on $[0,2]$, and t_3 is the t-distribution with 3 degrees of freedom. In scenarios (II) and (III), the moderate signals are more concentrated around 1 and have less variation compared to the continuous prior $\mathrm{Ga}(2,2)$ in scenario (I). We define the outlying and non-outlying values of λ_i as those generated from f_1 and f_0, respectively. In each scenario, we considered two values of ω, namely, 0.05 and 0.1.

We considered the estimation of λ_i using the following six priors/methods:

- IG: The proposed method with the inverse gamma prior for u_i.
- EH: The proposed method with the extremely heavy-tailed prior for u_i.
- GH: The Gauss hypergeometric prior proposed by Datta and Dunson (2016).
- PG: The gamma distribution for λ_i with $u_i = 1$, or the Poisson-gamma model.
- KW: Nonparametric empirical Bayes method (Kiefer and Wolfowitz, 1956; Koenker and Mizera 2014).
- ML: Maximum likelihood (non-shrinkage) estimator, that is, y_i.

We assigned prior distributions for the hyperparameters in the two proposed methods, as illustrated in Section 3.2. In the GH method, the hyperparameters were estimated by the empirical Bayes method recommended in Datta and Dunson (2016). Then, 3,000 posterior samples were generated directly from the posterior distribution of λ_i. We assigned the gamma priors for the hyperparameters in the PG method and used the prior distributions given in Section 3.2 for the hyperparameter in the IG and EH methods. The three methods require computations by the MCMC method; for each dataset, we generated 3,000 posterior samples after discarding 500 samples as a burn-in period. We computed point estimates of λ_i, where we used the posterior mean as point estimation in the first four methods. The performance of these point estimators is evaluated by the mean squared error (MSE) and mean absolute percentage error (MAPE) defined as the averaged values of $(\widehat{\lambda}_i - \lambda_i)^2$ and $|\widehat{\lambda}_i - \lambda_i|/\lambda_i$, respectively. Since MAPE for extremely small λ_i can take extremely high values, MAPE is evaluated without samples that satisfy $\lambda_i < 0.001$. These measures were calculated separately for outlying and non-outlying values of the true λ_i's. We also computed 95% credible intervals of λ_i based on the first four Bayesian methods and evaluated the performance using the coverage probability (CP) and average length (AL). We repeated the experiment 1,000 times and report the averages of MSE, MAPE, CP, and AL below.

In Table 2, we present the averaged values of the MSEs and MAPEs in all scenarios. For non-outlying values, the IG, PG, and KW methods are quite comparable and better than the other methods in MSE. Meanwhile, the EH and GH methods perform better in MAPE. The GH method suffers from the worse MSE for non-outlying values in all scenarios; its strong shrinkage effect simply mismatches the design of our simulation study where the observations are not zero-inflated.

For outlying values, the point estimates of the PG method tend to be worse than ML in MSE. This is clearly due to the over-shrinkage problem of the PG method discussed earlier. However, other than this, there is no clear structure in the comparison of the EH, IG, and GH methods. The EH method, for which we verified the exact tail-robustness in Theorem 1, should perform better than the other models in MSE for outliers. Unexpectedly, in our simulation study, the GH method outperforms the EH method in several scenarios. The difference between the two methods may be emphasized if we consider larger λ_i (or y_i) and/or increase the MCMC iterations.

In terms of MAPE for outliers, the IG, EH, and GH methods are almost indistinguishable. In fact, another concept of tail-robustness can be considered for the MAPE-type loss function to explain this result. The Poisson-gamma model with any proper prior for scale u_i, including the IG, EH, and GH models, can achieve such tail-robustness. We name this property *weak tail-robustness*. Further details are in the Supplementary Materials (S3).

In Table 3, we report averaged values of the CPs and ALs of 95% credible intervals of the four Bayesian methods. For outliers in all experiments, the PG method has the narrowest intervals with the lowest empirical coverage rates, while the GH method obtains the widest intervals with the highest coverage rates. In comparison, the results of the IG and EH methods are moderate; they improve the coverage with narrower credible intervals. We also find that the coverage performance of the EH method is better than that of the IG method.

We checked the performance of the MCMC algorithm for the IG, EH, and IG methods under scenario (I) with $\omega = 0.1$. The averaged values of the inefficiency factors of $\lambda_1, \ldots, \lambda_m$ under the IG, EH, and PG methods were 1.17, 4.39, and 1.01, respectively. This shows that the resulting inefficiency factors seem acceptable, but that of the EH method is slightly higher than those of the other methods. This is partly because the number of latent parameters used in the Gibbs sampling of the EH method is large compared with the other methods. In Section S6 of the Supplementary Materials, we report the additional simulation studies with large sample size, namely, $m = 400$, and computation time of the four Bayesian methods.

Scenario	ω		IG	EH	GH	PG	KW	ML
(I)	0.05	MSE-n	**0.24**	0.28	0.42	0.25	0.26	0.40
		MSE-o	3.30	2.86	**2.80**	3.86	3.08	2.84
		MAPE-n	0.64	**0.57**	0.65	0.63	0.67	0.62
		MAPE-o	0.21	**0.19**	**0.19**	0.23	0.21	0.19
(I)	0.1	MSE-n	**0.26**	0.29	0.42	0.28	0.28	0.40
		MSE-o	2.99	2.76	2.69	3.01	**2.58**	2.73
		MAPE-n	0.64	**0.58**	0.65	0.63	0.67	0.61
		MAPE-o	0.20	**0.19**	**0.19**	0.20	**0.19**	0.19
(II)	0.05	MSE-n	**0.22**	0.27	0.43	0.23	0.23	0.40
		MSE-o	3.46	2.90	**2.80**	4.31	3.06	2.84
		MAPE-n	0.58	**0.52**	0.61	0.57	0.60	0.58
		MAPE-o	0.22	0.20	**0.19**	0.24	0.21	0.19
(II)	0.1	MSE-n	**0.24**	0.28	0.43	0.27	**0.24**	0.40
		MSE-o	3.05	2.79	2.78	3.13	**2.60**	2.81
		MAPE-n	0.59	**0.54**	0.62	0.59	0.62	0.58
		MAPE-o	0.20	**0.19**	**0.19**	0.20	**0.19**	0.19
(III)	0.05	MSE-n	0.19	0.26	0.43	0.21	**0.18**	0.40
		MSE-o	3.79	3.03	**2.90**	5.02	3.17	2.94
		MAPE-n	0.50	**0.47**	0.57	0.50	0.48	0.55
		MAPE-o	0.23	0.20	**0.19**	0.26	0.21	0.20
(III)	0.1	MSE-n	0.22	0.28	0.44	0.26	**0.20**	0.41
		MSE-o	3.09	2.78	2.80	3.25	**2.54**	2.82
		MAPE-n	0.53	**0.50**	0.58	0.53	0.51	0.55
		MAPE-o	0.20	**0.19**	**0.19**	0.21	**0.19**	0.19
(IV)	0.05	MSE-n	0.21	0.27	0.40	0.21	**0.20**	0.40
		MSE-o	2.38	**1.97**	2.01	2.71	2.52	2.07
		MAPE-n	1.43	1.03	**0.95**	1.35	1.33	0.63
		MAPE-o	0.25	**0.22**	**0.22**	0.27	0.25	0.22
(IV)	0.1	MSE-n	**0.23**	0.28	0.42	0.24	**0.23**	0.40
		MSE-o	2.12	**1.95**	2.02	2.14	2.03	2.07
		MAPE-n	1.40	1.06	**0.98**	1.30	1.35	0.63
		MAPE-o	0.23	0.22	0.22	0.23	**0.21**	0.22

Table 2: Averaged values of mean squared errors (MSE) and mean absolute percentage errors (MAPE) in non-outlying (-n) and outlying (-o) areas under four scenarios with $m = 200$ and $\omega \in \{0.05, 0.1\}$. The best results among the model-based methods (other than ML) are highlighted in bold.

5 Data analysis

We apply the proposed methods to the analysis of crime data using the generalized linear model with Poisson likelihood and random effects. This model has been adopted for various datasets in applied statistics; for example, in areal count data in disease mapping (Lawson, 2013). In such an application, the Poisson rate λ_i (defined below) is not merely an adjustment of areal effects. Rather, it is an important parameter

			CP				AL			
Scenario	ω		IG	EH	GH	PG	IG	EH	GH	PG
(I)	0.05	n	96.0	96.2	95.6	96.6	1.93	2.01	2.32	1.99
		o	88.1	91.7	94.3	80.8	5.57	5.81	6.27	4.83
	0.1	n	96.3	96.4	95.7	96.6	2.01	2.05	2.33	2.10
		o	90.7	92.4	94.8	88.7	5.71	5.83	6.25	5.20
(II)	0.05	n	96.2	96.3	95.5	96.9	1.90	2.02	2.36	1.98
		o	87.0	91.7	94.6	77.0	5.49	5.75	6.23	4.65
	0.1	n	96.4	96.4	95.5	96.8	2.00	2.07	2.37	2.12
		o	90.2	92.3	94.8	87.3	5.71	5.83	6.28	5.12
(III)	0.05	n	96.7	96.4	95.4	97.3	1.88	2.04	2.40	1.97
		o	84.8	90.9	94.1	69.9	5.42	5.73	6.23	4.47
	0.1	n	96.9	96.5	95.3	97.1	1.98	2.09	2.40	2.12
		o	89.8	92.2	94.8	86.0	5.69	5.82	6.27	5.03
(IV)	0.05	n	93.9	95.6	95.4	95.2	1.89	2.01	2.29	1.91
		o	84.5	91.4	94.3	77.5	4.35	4.83	5.33	3.80
	0.1	n	94.7	95.8	95.5	95.7	1.99	2.05	2.33	2.04
		o	88.0	91.6	94.6	85.8	4.51	4.85	5.32	4.13

Table 3: Coverage probabilities (CP) and average lengths (AL) of 95% credible intervals in non-outlying (n) and outlying (o) areas under four scenarios with $m = 200$ and $\omega \in \{0.05, 0.1\}$.

interpreted as the intrinsic relative risk of the region i (e.g., Li et al., 2010). Here, we incorporate the idea of covariate adjustment into crime risk modeling.

The dataset consists of the numbers of police-recorded crimes in the Tokyo metropolitan area (provided by the University of Tsukuba and publicly available online; "GIS database of the number of police-recorded crime at O-aza, chome in Tokyo, 2009–2017," available at `https://commons.sk.tsukuba.ac.jp/data_en`). We focused on the number of violent crimes in $m = 2855$ local towns in the Tokyo metropolitan area in 2015. For auxiliary information about each town, we used area (km^2), population densities at noon and night, the density of foreign people, the percentage of single-person households, and the average duration of residence. These help adjust the crime risk. Let y_i be the observed count of violent crimes, a_i be the area, and x_i be the vector of the standardized auxiliary information of the i^{th} local town. We are interested in the crime rates after adjusting the risk using the auxiliary information. We employed the following Poisson regression model:

$$y_i|\lambda_i \sim \text{Po}(\lambda_i \eta_i), \qquad \eta_i = \exp(\log a_i + x_i^t \delta), \tag{5.1}$$

independently for $i = 1, \ldots, m$, where δ is a vector of unknown regression coefficients. Under the model (5.1), the random effect for local town i, denoted by λ_i, can be interpreted as an adjustment risk factor that is not explained by the auxiliary information. For most local towns, the offset term explains the variation of crime rates. Hence, the adjustment risk factor is expected to be small. Yet, the adjustment risk might be extremely high in some local towns that we want to detect. This is precisely where global-local shrinkage priors fit, for which we employed the proposed IG and EH priors for λ_i. We adopted $N(0, 100)$ as a prior distribution of each component of δ; we found that the following result was robust to the choice of prior variance. For posterior inference, we simply used Gibbs sampling in which posterior samples of $(\lambda_1, \ldots, \lambda_m)$ and δ are iteratively drawn from their full conditional distributions. Conditional on δ, we can use the posterior computation algorithm for λ_i provided in Section 3.2. Meanwhile, given λ_i's, the full conditional distribution of δ is not of a familiar form. The detailed algorithm customized for the sampling of δ is based on the independent Metropolis–Hasting method (given in Section S7 of the Supplementary Materials). For comparison, we also applied the gamma distribution for λ_i as considered in Section 4 (again denoted by PG hereafter). We did not consider the GH prior in this example since the empirical Bayes method in Datta and Dunson (2016) is not available under the generalized linear model in equation (5.1). Similarly, we did not apply the KW method. In each Gibbs sampler, we generated 20,000 posterior samples after discarding 3,000 posterior samples as burn-in.

We first computed posterior means of risk factor λ_i by using the three methods. The spatial pattern of the estimates is shown in Figure 3. Notably, the proposed two methods, IG and EH, provide similar estimates of λ_i in most areas and successfully detect several local towns whose risk factors are extremely high. In contrast, such extreme towns are less emphasized, or not detected at all, by the PG method because it significantly underestimates the true risk factors. In Figure 4, we plotted the estimates of λ_i based on the proposed methods against that of the PG method. This clearly highlights the underestimation of outlying signals caused by the PG method.

Next, we detected ten local towns with the largest posterior means of λ_i. For these towns, we computed 95% credible intervals of λ_i, as shown in the left panel of Figure 5. This panel again shows the over-shrinkage problem of the PG method in both point and interval estimations (posterior means and credible intervals). The posterior credible intervals computed by the PG method are narrower, suggesting the underestimation of posterior uncertainty. We also randomly selected another ten local towns with moderate estimates of λ_i and 95% credible intervals, as shown in the right panel of the same figure. The difference between the three methods is almost negligible for these towns. These observations exemplify that the proposed methods can avoid the over-shrinkage problem for large signals while their performance in the other towns is almost the same as the standard PG method.

6 Discussion

The Poisson-gamma model is a convenient tool, but can be restrictive since the observational mean and variance given λ_i must always be equal. Furthermore, both conditional prior mean and variance of λ_i are controlled by the common local parameter u_i under the gamma prior $\mathrm{Ga}(\alpha, \beta/u_i)$ for λ_i. This affects both the baseline and amount of shrinkage. This property is not seen in the Gaussian case, where the local parameter appears in the prior variance and controls only the amount of local shrinkage. This makes the role of local parameters clear. In this sense, the local parameter in equation

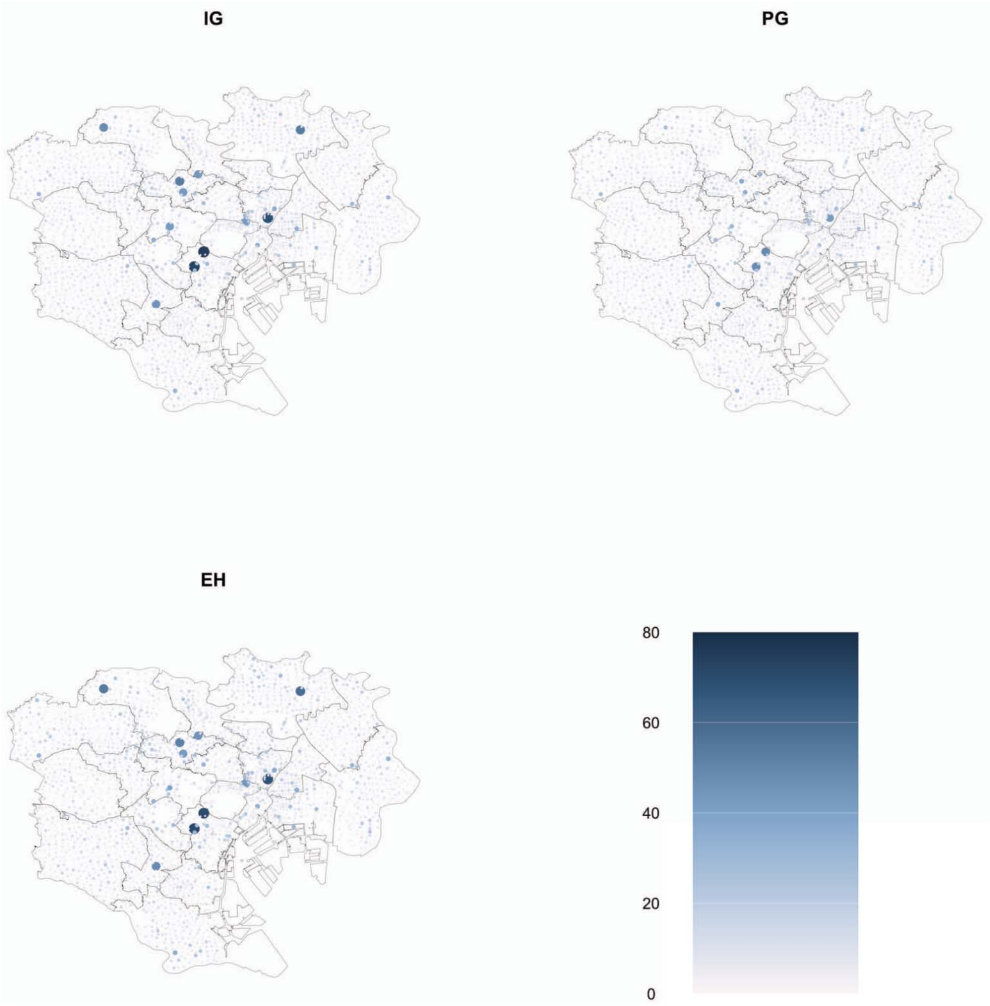

Figure 3: Posterior means of risk factors λ_i based on IG, EH, and PG methods.

(2.1) might be less interpretable. Meanwhile, this setting also enables us to carry out posterior computation easily and has been studied intensively in the literature (e.g., Datta and Dunson, 2016). Future research could pursue an alternative setting for hierarchical modeling of a sequence of counts under which the role of the local parameters is properly restricted and interpreted.

From the viewpoint of methodological research, this paper is primarily focused on the point and interval estimation of the Poisson rate. The estimation of high-dimensional counts can be cast as other statistical problems such as multiple testing. Detailed investigation in such directions would have extended the scope of this paper; thus, we leave it to a valuable future study.

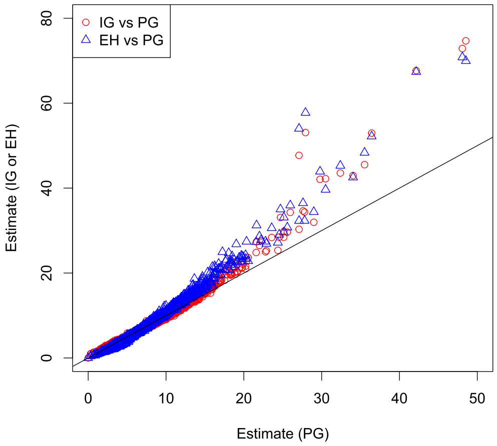

Figure 4: Scatter plot of posterior means of risk factors λ_i based on IG, EH, and PG methods.

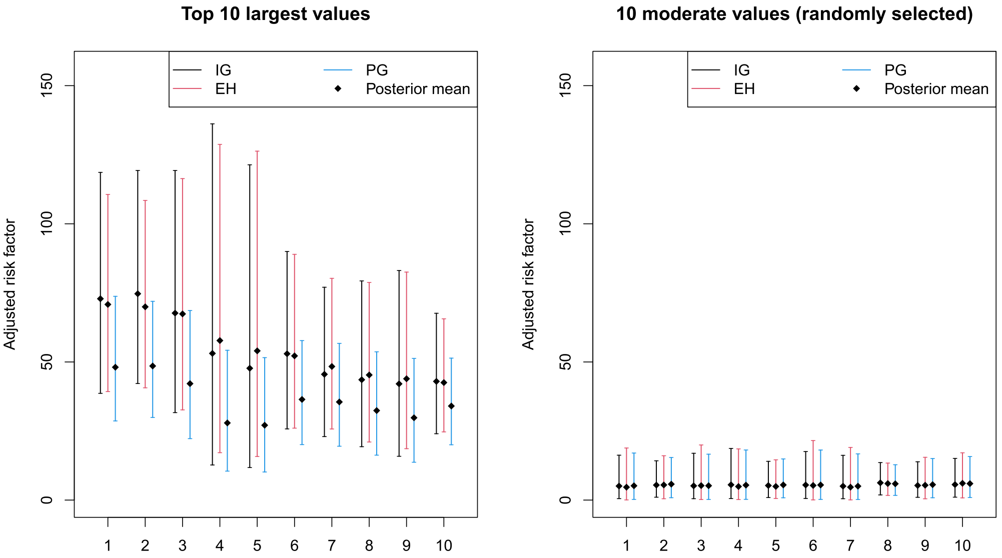

Figure 5: 95% credible intervals for areas with highest 10 posterior means (left) and for 10 randomly selected areas with moderate posterior means (right) of adjusted risk factors.

The newly introduced EH prior represents the probability distributions that satisfy the conditions for tail-robustness given in Theorem 2.1. However, the class of priors that satisfy those conditions is not limited to the EH prior. In theory, we can consider a general class of priors with the following density which is also proper and tail-robust:

$$\pi(u_i) \propto \frac{{u_i}^{\gamma_1 - 1}}{(1+\gamma_2 u_i)^{\gamma_1}} \frac{1}{\{1+\gamma_3 \log(1+\gamma_4+u_i)\}^{1+\gamma_5}}.$$

The hyperparameters $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ increase the flexibility of the model and could improve the EH prior equipped with a single parameter γ. However, the posterior inference under this prior is challenging due to the intractable normalizing constant that involves those hyperparameters. The full-Bayes inference for the hyperparameters is not as straightforward as that of the EH prior. The inference with fixed hyperparameters is feasible by utilizing the same parameter augmentation in Section 3.2, but raises the problem of hyperparameter tuning. We leave the development of this extension to future work, which could be useful in more structured models for count data.

Supplementary Material

Supplementary Materials for "On Global-local Shrinkage Priors for Count Data" (DOI: 10.1214/21-BA1263SUPP; .pdf). Supplementary Materials for "On Global-local Shrinkage Priors for Count Data" include technical details regarding the proofs of Theorem 2.1 (S1 and S2), the related tail-robustness properties (S3), the derivation of the EH prior (S4), and the evaluation of the marginal prior distribution of λ_i with the EH prior (S5). It also provides additional simulation results (S6) and computational details of the MCMC algorithm (S7)

References

Armagan, A., Clyde, M. & Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In *Advances in neural information processing systems*, 523–531. 549

Bhadra, A., Datta, J., Polson, N. G. & Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* **103**, 955–969. MR3620450. doi: https://doi.org/10.1093/biomet/asw041. 546

Bhadra, A., Datta, J., Polson, N. G. & Willard, B. T. (2019). Lasso meets horseshoe: A survey. *Statistical Science* **35**, 405–427. MR4017521. doi: https://doi.org/10.1214/19-STS700. 546

Brown, L. D., Greenshtein, E., & Ritov, Y. (2013). The Poisson compound decision problem revisited. *Journal of the American Statistical Association* **108**, 741–749. MR3174656. doi: https://doi.org/10.1080/01621459.2013.771582. 546

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 545, 547, 550

Datta, J. & Dunson, D. V. (2016). Bayesian inference on quasi-sparse count data. *Biometrika* **103**, 971–983. MR3620451. doi: https://doi.org/10.1093/biomet/asw053. 546, 549, 550, 555, 559, 560

Hamura, Y., Irie, K., & Sugasawa, S. (2021). "Supplementary Material of "On Global-local Shrinkage Priors for Count Data"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1263SUPP. 548

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**, 887–906. MR0086464. doi: https://doi.org/10.1214/aoms/1177728066. 555

Koenker, R. & Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rule. *Journal of the American Statistical Association* **109**, 674–685. MR3223742. doi: https://doi.org/10.1080/01621459.2013.869224. 555

Lawson, A. B. (2013). Bayesian disease mapping: hierarchical modeling in spatial epidemiology. Chapman and Hall/CRC. MR2484272. doi: https://doi.org/10.1201/b14073. 546, 557

Li, H., Graubardn, B. I., & Gail, M. H. (2010). Covariate adjustment and ranking methods to identify regions with high and low mortality rates. *Biometrics* **66**, 613–620. MR2758842. doi: https://doi.org/10.1111/j.1541-0420.2009.01284.x. 558

Maruyama, Y. & Strawderman, W. E. (2020). Admissible Bayes equivariant estimation of location vectors for spherically symmetric distributions with unknown scale. *Annals of Statistics* **48**, 1052–1071. MR4102687. doi: https://doi.org/10.1214/19-AOS1837. 549

Seneta, E. (1976). Regularly varying functions. Springer-Verlag Berlin Heidelberg. MR0453936. doi: https://doi.org/10.1007/BFb0079658. 548

Tang, X., Ghosh, M., Ha, N., & Sedransk, J. (2018). Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association* **113**, 1476–1489. MR3902223. doi: https://doi.org/10.1080/01621459.2017.1419135. 546, 548

van Dyk, D. A. & Park, T. (2019). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association* **103**, 790–796. MR2524010. doi: https://doi.org/10.1198/016214508000000409. 550

Wakefield, J. (2006). Disease mapping and spatial regression with count data. Oxford University Press. 546

Yano, K., Kaneko, R., & Komaki, F. (2018). Exact Minimax Predictive Density for Sparse Count Data. arXiv:1812.06037. 546

Zhou, M. & Carin, L. (2013). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 307–320. doi: https://doi.org/10.1109/TPAMI.2013.211. 550

Zhu, A., Ibrahim, J. G., & Love, M. I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092. doi: https://doi.org/10.1093/bioinformatics/bty895. 546

Acknowledgments

The authors thank the editor, associate editor, and anonymous referees for their constructive suggestions.