Bayesian Analysis (2024) **19**, Number 1, pp. 77–97

Sparse Bayesian Inference on Gamma-Distributed Observations Using Shape-Scale Inverse-Gamma Mixtures*

Yasuyuki Hamura†, Takahiro Onizuka‡, Shintaro Hashimoto§ and Shonosuke Sugasawa¶

Abstract. In various applications, we deal with high-dimensional positive-valued data that often exhibits sparsity. This paper develops a new class of continuous global-local shrinkage priors tailored to analyzing gamma-distributed observations where most of the underlying means are concentrated around a certain value. Unlike existing shrinkage priors, our new prior is a shape-scale mixture of inverse-gamma distributions, which has a desirable interpretation of the form of posterior mean and admits flexible shrinkage. We show that the proposed prior has two desirable theoretical properties; Kullback-Leibler super-efficiency under sparsity and robust shrinkage rules for large observations. We propose an efficient sampling algorithm for posterior inference. The performance of the proposed method is illustrated through simulation and two real data examples, the average length of hospital stay for COVID-19 in South Korea and adaptive variance estimation of gene expression data.

Keywords: gamma distribution, Kullback-Leibler super-efficiency, Markov chain Monte Carlo, tail-robustness.

1 Introduction

In various statistical applications, we often face a sequence of positive-valued observations such as machine failure time, store waiting time, survival time under a certain disease, an income of a certain group, and so on. A common feature of the data is "sparsity" in the sense that most of the underlying means of observations are concentrated around a certain value (grand mean) while a small part of the means is significantly away from the grand mean. To reflect the sparsity structure, a useful Bayesian technique is an idea of "global-local shrinkage" (e.g. Polson and Scott, 2012) that provides adaptive and flexible shrinkage estimation of underlying means; when the observations are around the grand mean, the posterior mean strongly shrinks the observation toward the grand mean, but the observations that are away from the grand mean remain unshrunk.

This paper proposes a new framework for sparse Bayesian inference on a sequence

arXiv: 2203.08440

*The authors' research was supported in part by JSPS KAKENHI Grant Numbers 22K20132, 20J10427, 19K11852, 21K13835, and 21H00699 from the Japan Society for the Promotion of Science.

†Graduate School of Economics, Kyoto University, yasu.stat@gmail.com

‡Department of Mathematics, Hiroshima University, t-onizuka@hiroshima-u.ac.jp

§Department of Mathematics, Hiroshima University, s-hashimoto@hiroshima-u.ac.jp

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

 https://doi.org/10.1214/22-BA1348

of positive-valued observations by using gamma sampling distributions for observations and develops a novel class of global-local shrinkage priors for positive-valued heterogeneous mean parameters based on shape-scale mixtures of inverse-gamma distributions. Specifically, we introduce a scaled beta (SB) distribution and its extension called inverse rescaled beta (IRB) distribution as mixing distributions in the shape-scale mixture. We discuss distributional properties, tail decay rate, and concentration around the origin of the proposed priors and develop an efficient sampling scheme from the posterior distribution. Moreover, we reveal two theoretical properties of the proposed prior, tail-robustness for large means and Kullback-Leibler supper-efficiency under sparsity.

There are several works on shrinkage inference of a sequence of positive-valued data. Under the gamma sampling model (as in our proposal), simultaneous estimation for rate/scale parameters was considered by several authors decades ago (e.g., Berger, 1980; Ghosh and Parsian, 1980; DasGupta, 1986; Dey et al., 1987). However, the classical framework does not take into account sparsity and provides only universal shrinkage regardless of the observed values. To address the sparsity in positive-valued data, Donoho and Jin (2006) proposed a threshold-type estimator with the false discovery rate control, but the sampling model is an exponential distribution (a special case of gamma distribution). Therefore, its applicability is quite limited. Recently, Lu and Stephens (2016) proposed an empirical Bayes shrinkage method customized for variance estimation using a χ^2-distribution (a special case of gamma distribution) for the observed sampling variance and a finite mixture of inverse-gamma priors for the true variance. More recently, Kwon and Zhao (2022) also proposed a novel formulation called F-modeling for variance shrinkage estimator in terms of nonparametric empirical Bayes. However, this approach does not address sparsity, and no theoretical results are discussed. More importantly, the existing methods only produce point estimates of the underlying means. In contrast, the proposed method can obtain full information on posterior distributions, enabling us to carry out uncertainty quantification.

In Bayesian analysis, the methodology and application of "global-local shrinkage priors" have been developed last decades. Under Gaussian sequence or normal linear regression models, there have been a variety of shrinkage priors including the most famous horseshoe (Carvalho et al., 2010) prior and its related priors (e.g. Armagan et al., 2013; Bhadra et al., 2017; Bhattacharya et al., 2015; Hamura et al., 2020; Zhang et al., 2020). Such prior is known to have an attractive shrinkage property, making it possible to strongly shrink small observations toward zero while keeping large observations unshrunk. Recently, techniques of global-local shrinkage priors for Gaussian data are extended to the (quasi-)sparse count data (e.g. Datta and Dunson, 2016; Hamura et al., 2022b). Although several theoretical properties (e.g., Kullback-Leibler supper-efficiency and tail-robustness) have been revealed under the Gaussian and Poisson sampling distributions, theoretical properties of global-local shrinkage under the gamma sampling model are not fully discussed. Furthermore, the theoretical development of the proposed prior requires substantial work due to the form of shape-scale mixtures that are rather different from the existing global-local shrinkage priors. To fill the gap, we contribute to the theoretical development of global-local shrinkage by showing Kullback-Leibler supper-efficiency and tail-robustness under the gamma sampling model.

The remainder of the paper is structured as follows. In Section 2, we introduce settings and our hierarchical model, and we propose a global-local shrinkage prior based on a kind of beta distribution. Furthermore, we illustrate the properties of the marginal prior and posterior distributions for λ_i, and also discuss the selection of hyperparameters of the proposed priors. An efficient posterior computation algorithm is constructed via the Markov chain Monte Carlo method. In Section 3, we show two theoretical properties of the proposed priors. The performance of the proposed method is demonstrated through numerical studies in Section 4, and we apply the method to two real datasets related to the average length of hospital stay for COVID-19 in South Korea and variance estimation of gene expression data in Section 5. Proofs and technical details are given in the Supplementary Material (Hamura et al., 2022). R code implementing the proposed methods is available at Github repository (https://github.com/sshonosuke/GLSP-gamma/).

2 Sparse Bayesian inference on gamma-distributed observations

2.1 Settings and models

Suppose we observe a sequence of gamma-distributed observations, denoted by $y_1, \ldots, y_n$. For each $i = 1, \ldots, n$, we assume the following gamma model y_i:

$$y_i \mid \lambda_i \sim \mathrm{Ga}\left(\delta_i, \frac{\delta_i}{\lambda_i \eta_i}\right), \tag{2.1}$$

where $\mathrm{Ga}(\alpha, \beta)$ denotes a gamma distribution with shape parameter α and rate parameter β, δ_i is a fixed constant, and λ_i is a parameter of interest. Under the model, $E(y_i) = \lambda_i \eta_i$ and η_i is a structural component that may be modeled to incorporate covariates and other external information (e.g., spatial information). In what follows, we assume $\eta_i = 1$ for simplicity, under which λ_i is interpreted as the mean of y_i, but all the computation algorithms and analytical results are valid for the general form of η_i as long as η_i is conditioned on. As considered in Lu and Stephens (2016), if y_i and λ_i are sampling and true variances, respectively, the choice is $\delta_i = n_i/2$, where n_i is a sample size used to compute y_i. Moreover, if y_i is a sample mean based on n_i samples generated from an exponential distribution $\mathrm{Exp}(1/\lambda_i)$, it holds that $\delta_i = n_i$, and it reduces the framework of a sequence of exponential data when $n_i = 1$, considered in Donoho and Jin (2006). In the present framework, our interest lies in the simultaneous estimation of the sequence of positive-valued means $\lambda = (\lambda_1, \ldots, \lambda_n)$ by combining information of a given set of data $y = (y_1, \ldots, y_n)$. In particular, we focus on the structure that most observations are located around the grand mean while some observations are very large. To carry out flexible Bayesian inference even under this situation, we employ an idea of global-local shrinkage that can provide customized shrinkage estimation of λ_i depending on the location of the observed value y_i.

Specifically, we consider the following prior distribution for λ_i:

$$\lambda_i \mid u_i \sim \mathrm{IG}(1 + \tau u_i, \beta \tau u_i), \quad i = 1, \ldots, n, \tag{2.2}$$

where β and τ are unknown global parameters and u_i is a local parameter related to the customized shrinkage rule. The prior mean of λ_i is $E(\lambda_i) = \beta$ so that β is interpreted as a grand mean of underlying heterogeneous means. On the other hand, since $\mathrm{Var}(\lambda_i) = \beta^2/(\tau u_i - 1)$ as long as $\tau u_i > 1$, τ and u_i control the scale of the prior. Unusual parametrization of (2.2) is the dependence of both shape and scale parameters on the local parameter u_i so that setting a mixing distribution for u_i leads to a class of shape-scale mixtures of inverse-gamma distributions. However, this parametrization is essential to interpret the form of posterior means of λ_i.

Under the inverse-gamma prior (2.2), the conditional posterior distributing of λ_i given u_i is $\mathrm{IG}(1 + \delta_i + \tau u_i, \delta_i y_i + \beta\tau u_i)$, so that the posterior mean of λ_i is given by

$$E(\lambda_i \mid y_i) = E\left(\frac{\delta_i y_i + \beta\tau u_i}{\delta_i + \tau u_i} \mid y_i\right) = \beta + \{1 - E(\kappa_i \mid y_i)\}\,(y_i - \beta),$$

where $\kappa_i = \tau u_i/(\delta_i + \tau u_i) \in (0, 1)$ is known as *shrinkage factor* that determines the amount of shrinkage of y_i toward the grand mean β. As desirable properties of κ_i, $E(\kappa_i \mid y_i)$ should be close to 1 when y_i is close to the grand mean, leading to strong shrinkage toward β, while $E(\kappa_i \mid y_i)$ should be sufficiently small for y_i having large $y_i - \beta$ to prevent bias caused by over-shrinkage. We also note that the global parameter τ determines the overall shrinkage effect, whereas the local parameter u_i allows κ_i to vary over different observations.

2.2 Global-local shrinkage priors

Our hierarchical model can be expressed as

$$y_i \mid \lambda_i \sim \mathrm{Ga}\left(\delta_i, \frac{\delta_i}{\lambda_i}\right), \qquad \lambda_i \mid u_i \sim \mathrm{IG}(1 + \tau u_i, \beta\tau u_i), \qquad u_i \sim \pi(\cdot),$$

where priors for β and τ are discussed at the end of this subsection. For the local parameter u_i, we suggest two prior distributions. The first one is the scaled beta (SB) prior

$$\pi_{\mathrm{SB}}(u_i) = \frac{1}{B(a, b)} \frac{{u_i}^{a-1}}{(1 + u_i)^{a+b}},$$

where $a, b > 0$ are hyperparameters and $B(a, b)$ is the beta function. The SB distribution is also known as the beta prime distribution (e.g. Johnson et al., 1995), and the related family of distributions has been often used in Bayesian statistics (e.g. Pérez et al., 2017; Hamura et al., 2021), especially in the context of shrinkage priors. As an alternative prior, we newly propose the inverse rescaled beta (IRB) prior

$$\pi_{\mathrm{IRB}}(u_i) = \frac{1}{B(b, a)} \frac{1}{u_i(1 + u_i)} \frac{\{\log(1 + 1/u_i)\}^{b-1}}{\{1 + \log(1 + 1/u_i)\}^{b+a}}.$$

Note that the IRB prior for u_i is equivalent to using the rescaled beta prior (Hamura et al., 2021) for $1/u_i$.

Here, we summarize basic properties of $\pi_{\mathrm{SB}}(u_i)$ and $\pi_{\mathrm{IRB}}(u_i)$ under $u_i \to 0$ and $u_i \to \infty$. As is well known, the SB prior has the following properties.

- *Concentration at the origin.* As $u_i \to 0$, we have $\pi_{\mathrm{SB}}(u_i) \propto {u_i}^{a-1}$. In particular, $\pi_{\mathrm{SB}}(\kappa_i) \to \infty$ as $\kappa_i \to 0$ if and only if $a < 1$.

- *Tail decay.* As $u_i \to \infty$, we have $\pi_{\mathrm{SB}}(u_i) \propto {u_i}^{-1-b}$. In particular, $\pi_{\mathrm{SB}}(\kappa_i) \to \infty$ as $\kappa_i \to 1$ if and only if $b < 1$.

Meanwhile, ignoring log factors, we see that $\pi_{\mathrm{IRB}}(u_i)$ has the following properties:

- *Concentration at the origin.* As $u_i \to 0$, we have $\pi_{\mathrm{IRB}}(u_i) \approx {u_i}^{-1}$. In particular, $\pi_{\mathrm{IRB}}(\kappa_i) \approx {\kappa_i}^{-1} \to \infty$ as $\kappa_i \to 0$ whatever the value of $a > 0$ is. This is in contrast to the case of the SB prior.

- *Tail decay.* As $u_i \to \infty$, we have $\pi_{\mathrm{IRB}}(u_i) \propto {u_i}^{-1-b}$. In particular, $\pi_{\mathrm{IRB}}(\kappa_i) \to \infty$ as $\kappa_i \to 1$ if and only if $b < 1$. This is exactly as in the case of the SB prior.

In the context of existing global-local shrinkage priors, the concentration at both $\kappa_i = 0$ and $\kappa_i = 1$ is closely related to the properties of shrinkage and tail robustness of the marginal prior of the parameter of interest (e.g. Carvalho et al., 2010; Datta and Dunson, 2016). However, as shown in the subsequent section, the concentration at $\kappa_i = 0$ leads to unnecessary shrinkage toward the origin in our framework, possibly because the local parameter depends not only on scale but also on shape unlike the existing formulation of global-local shrinkage. Hence, we should not pursue the concentration at $\kappa_i = 0$ in the proposed model. In fact, as shown in Proposition 2.1, the choice of a (controlling the concentration at $\kappa_i = 0$) is not related to the performance of shrinkage and tail robustness as the marginal prior of λ_i.

We discuss the priors for β and τ. Remember that β is a grand mean (i.e. shrinkage target of the posterior mean) of λ_i and τ controls the overall shrinkage. It would be possible to fix β or assign an informative prior for β if the user has much information about β. On the other hand, when there is not much prior information on β and τ, we recommend using proper but slightly diffuse priors. In our numerical studies, we use priors, $\beta \sim \mathrm{Ga}(0.1, 0.1)$ and $\tau \sim \mathrm{Ga}(0.1, 0.1)$ as default priors, which are conditionally conjugate. Although improper priors can be assigned for β and τ, checking the posterior propriety given a certain form of improper prior is not straightforward due to the complicated hierarchical forms of the model. In the Supplementary Material, we discuss the conditions of posterior propriety under some forms of improper priors. For example, using $\pi(\beta) \propto 1/\beta$ combined with a proper gamma prior for τ leads to posterior propriety under some conditions. Furthermore, we also note that the standard improper priors for scale parameters such as $\pi(\beta) \propto 1$ or $\pi(\beta) \propto 1/\beta$ may not be necessarily reasonable under the hierarchical gamma model, that is, it is not clear whether these priors can be justified as objective ones such as reference priors. Since we assume subjective priors for λ_i and u_i, we may be able to consider reference priors for β and τ using an idea of partial information prior (Sun and Berger, 1998), but we do not pursue the detailed argument here.

2.3 Marginal prior for λ_i

In this section, we consider the behavior of the marginal prior of λ_i. We assume that the grand mean β and global shrinkage parameter τ are fixed at 1 for simplicity so that the grand mean is 1 in the following discussion. We first discuss the roles of the hyperparameters, a and b, of the proposed priors, and then we propose particular choices of the hyperparameters.

The goal is to select a and b so that the marginal prior of λ_i should ideally (G1) not be thick at the origin and have (G2) a fat right-tail and (G3) a spike at 1. We provide the following analytical results concerning the behavior of the marginal prior for λ_i.

Proposition 2.1. *Suppose that either* $\pi(u_i) = \pi_{\mathrm{SB}}(u_i) \propto {u_i}^{a-1}/(1+u_i)^{a+b}$ *or* $\pi(u_i) = \pi_{\mathrm{IRB}}(u_i) \propto [1/\{u_i(1+u_i)\}]\{\log(1+1/u_i)\}^{b-1}/\{1+\log(1+1/u_i)\}^{b+a}$. *Then the marginal prior* $p(\lambda_i)$ *of* λ_i *has the following properties:*

(i) *As* $\lambda_i \to 0$,

$$p(\lambda_i) \approx \begin{cases} {\lambda_i}^{a-1}, & \text{if } \pi(u_i) = \pi_{\mathrm{SB}}(u_i), \\ {\lambda_i}^{-1}, & \text{if } \pi(u_i) = \pi_{\mathrm{IRB}}(u_i). \end{cases}$$

(ii) *As* $\lambda_i \to \infty$,

$$p(\lambda_i) \approx {\lambda_i}^{-2}.$$

(iii) *As* $\lambda_i \to 1$,

$$p(\lambda_i) \to \begin{cases} \infty, & \text{if } b \le 1/2, \\ C_1 < \infty, & \text{if } b > 1/2 \end{cases}$$

for some finite positive constant $0 < C_1 < \infty$.

Proposition 2.1 shows that our three goals are achieved whenever $b \le 1/2$ and $a > 1$ for the SB prior, and that goal (G1) is impossible to achieve under the IRB prior but (G2) and (G3) are achieved for $b \le 1/2$. This result is obtained from a more general theorem (Theorem S1), given in the Supplementary Material. Theorem S1 provides equivalents for the tail densities and density at 1 of λ under different priors for u_i and relies on convergence theorems and approximations to prove them. We note that log factors are ignored in the above statement.

In more detail, Part (i) corresponds to shrinkage and non-shrinkage for small λ_i under the SB prior with $a > 1$ and the IRB prior, respectively. Part (ii) corresponds to robustness for large λ_i (i.e., the posterior mean of λ_i does not shrink large y_i) under the proposed priors; if we fix u_i, then we necessarily have $p(\lambda_i) \propto {\lambda_i}^{-2-u_i} < {\lambda_i}^{-2}$ as $\lambda_i \to \infty$. Part (iii) corresponds to shrinkage for moderate λ_i under the proposed priors with $b \le 1/2$; if we fix u_i, then $p(\lambda_i)$ never diverges at $\lambda_i = 1$.

In other words, the left tail of $\pi(u_i)$ can affect the left tail of $p(\lambda_i)$ if we use the SB prior with $a \leq 1$ or the IRB prior; the right tail of $p(\lambda_i)$ is guaranteed to be sufficiently heavy for any values of the hyperparameters; we can expect that a sufficient amount of prior probability mass is put around $\lambda_i = 1$ if we choose $b \leq 1/2$ for the SB and IRB priors. Based on these findings, we propose to use $a > 1$ for the SB prior and $b \leq 1/2$ for both the SB and IRB priors. In particular, our default choices are $a = 2$ and $b = 1/2$ for both the priors.

The marginal prior densities of λ_i under the SB and IRB priors are illustrated in Figure 1. As expected, it can be seen from the right panel that the right tail of $p(\lambda_i)$ is heavier under the proposed priors than the global shrinkage prior (denoted by GL in Figure 1) when u_i is fixed, that is, $\lambda_i \sim \mathrm{IG}(2, 1)$. Also, it is confirmed that the IRB prior makes the right tail heavier than the SB prior. The left panel shows that the hyperparameter a of the SB and IRB priors causes a trade-off between undesirable tail thickness at the origin and desirable tail thickness at infinity. However, for the case of the SB prior, we at least have that $p(\lambda_i) \to 0$ as $\lambda_i \to 0$ for $a = 2$ and for $a = 3$. The most remarkable point we want to stress here is that under each of the proposed priors, $p(\lambda_i)$ has a spike at $\lambda_i = 1$. This means that a large shrinkage effect is expected when we use one of the proposed priors, and this is quite in contrast to the case of fixing $u_i = 1$, where the mode of $p(\lambda_i)$ is significantly shifted to the left.

Finally, the choice $a = 2$ may seem slightly strange in the literature on global-local shrinkage priors. Under $u_i \sim \mathrm{SB}(a, b)$, the shrinkage factor $\kappa_i = u_i/(1 + u_i)$ follows the beta distribution $\mathrm{Beta}(a, b)$. The well-known horseshoe prior (Carvalho et al., 2010) corresponds to the case $(a, b) = (1/2, 1/2)$, and the resulting prior distribution of κ_i is $\mathrm{Beta}(1/2, 1/2)$, which has the popular U-shaped density. For our model, we do not adopt the choice $(a, b) = (1/2, 1/2)$, since setting $a = 1/2$ causes unexpected tail-robustness (or lack of desirable shrinkage toward the grand mean) around the origin and since using $a > 1$ does not affect tail-robustness around infinity much (see also Section 3.1).

2.4 Marginal posterior of λ_i

Here, we discuss the flexibility of the proposed prior distributions. As an artificial example, we suppose that $m = 50$ observations (the first 46 observations are 5 and the others are $7, 15, 30, 50$) are observed. Furthermore, we set $\delta_i = 5$. We show marginal posterior distributions of the shrinkage factor κ_i given $y \in \{7, 15, 30, 50\}$ in Figure 2. The marginal posterior under the global shrinkage prior ($u_i = 1$) does not depend on y and over-shrinks the posterior density under a large signal such as $y = 30$ and $y = 50$. Also, the global shrinkage method does not have strong shrinkage near the grand mean when $y = 7$. On the other hand, Figure 2 shows that the posterior of κ_i under the SB and IRB priors change flexibly according to the observed values, as expected from the design of the priors. Comparing the two priors, it can be seen that the IRB posterior is more concentrated around $\kappa_i = 0$ than the SB prior when y_i is large. Therefore, we recommend using IRB prior to situations where tail-robustness is required.

We further investigate the behavior of the posterior distribution through posterior means and variances of λ_i as a function of y_i. To see the properties of the local shrinkage

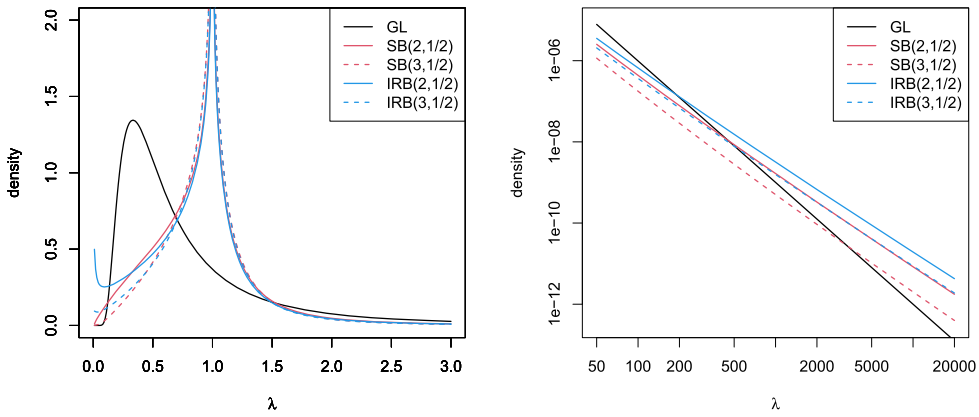

Figure 1: Marginal prior densities for λ_i. The right panel is an enlarged version of the left panel in the log-scaled tail region.

property, the hyperparameters in the three priors are fixed to their posterior means obtained to make Figure 2. We set $(\log y_1, \ldots, \log y_{100})$ to equally-spaced 100 points from -4 to 4, and computed posterior means and variances of λ_i $(i = 1, \ldots, 100)$ based on the three priors. The results are shown in Figure 3. It is observed that both the posterior mean and variance of the GL prior are simple functions of y_i. On the other hand, the proposed two priors, SB and IRB, strongly shrink y_i around the grand mean while do not shrink large or small y_i. Moreover, the posterior variances of the proposed two priors are small around the grand mean due to the strong shrinkage property and those are large when the observed value is large.

2.5 Posterior computation

We provide an efficient Metropolis within the Gibbs algorithm for our model by using the approximation method of Miller (2019). Here, we consider the case of the SB prior. The details of posterior computation under the IRB prior are given in the Supplementary Material. In order to simplify sampling of τ, we make the change of variables $\nu_i = \tau u_i$ for $i = 1, \ldots, n$. Then the overall posterior distribution of $(\lambda, \beta, \tau, \nu)$ given y is expressed by

$$p(\lambda, \beta, \tau, \nu \mid y) \propto \pi(\beta)\pi(\tau)\frac{1}{\tau^n}\prod_{i=1}^{n}\left\{\pi(\nu_i/\tau)\frac{\beta^{\nu_i+1}{\nu_i}^{\nu_i}}{\Gamma(\nu_i)}\frac{1}{{\lambda_i}^{\nu_i+2}}e^{-\beta\nu_i/\lambda_i}\frac{1}{{\lambda_i}^{\delta_i}}\exp\left(-\frac{\delta_i y_i}{\lambda_i}\right)\right\},$$

where $\nu = (\nu_1, \ldots, \nu_n)$. Since the SB prior density is expressed as

$$\pi_{\mathrm{SB}}(u_i) = \frac{1}{B(a,b)}\frac{{u_i}^{a-1}}{(1+u_i)^{a+b}} = \frac{1}{\Gamma(a)\Gamma(b)}\int_0^\infty {t_i}^{a+b-1}e^{-t_i}{u_i}^{a-1}e^{-t_i u_i}\mathrm{d}t_i$$

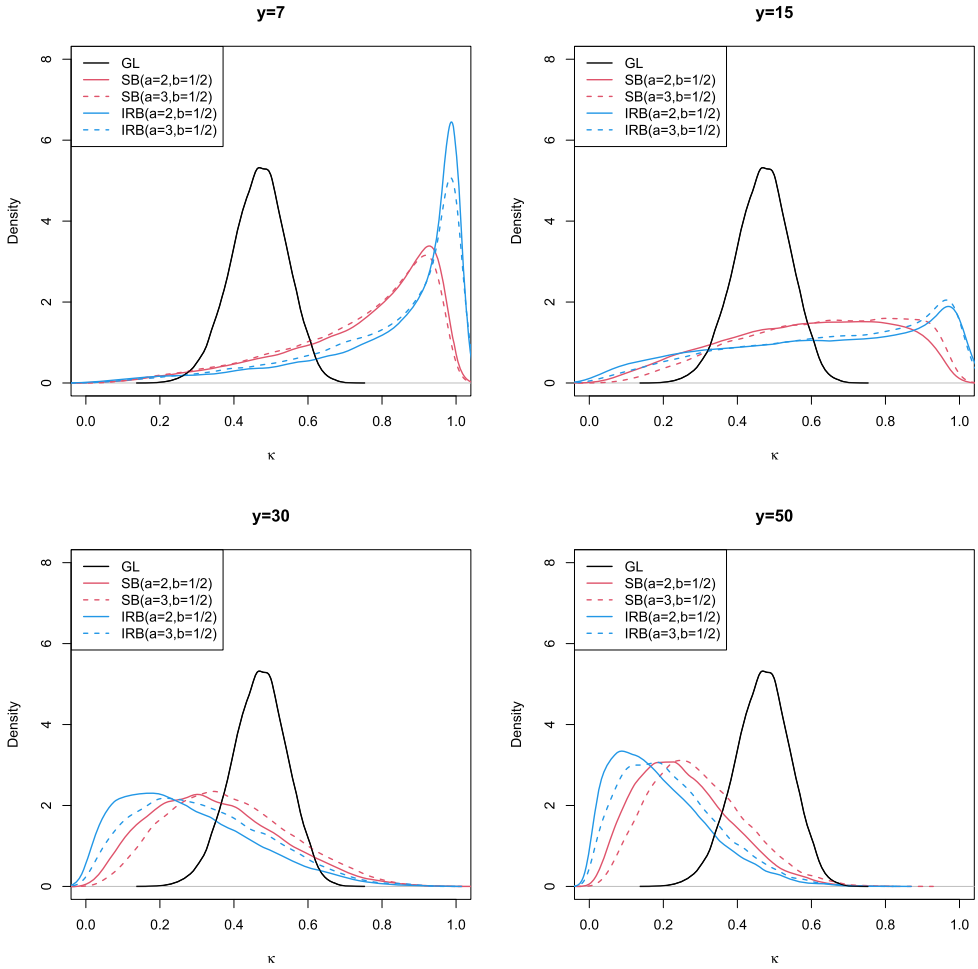

Figure 2: Marginal posterior densities for the shrinkage factor κ_i under four types of observed values.

for all $i = 1, \ldots, n$, it follows that

$$p(\lambda, \beta, \tau, \nu \mid y) \propto \int_{(0,\infty)^n} \left[\pi(\beta)\pi(\tau) \frac{1}{\tau^{na}} \right.$$
$$\times \prod_{i=1}^{n} \left\{ {t_i}^{a+b-1} e^{-t_i} {\nu_i}^{a-1} e^{-t_i \nu_i / \tau} \frac{\beta^{\nu_i+1} {\nu_i}^{\nu_i}}{\Gamma(\nu_i)} \frac{1}{{\lambda_i}^{\nu_i+2}} \right.$$
$$\left. \left. \times e^{-\beta\nu_i/\lambda_i} \frac{1}{{\lambda_i}^{\delta_i}} e^{-(\delta_i y_i)/\lambda_i} \right\} \right] \mathrm{d}t.$$

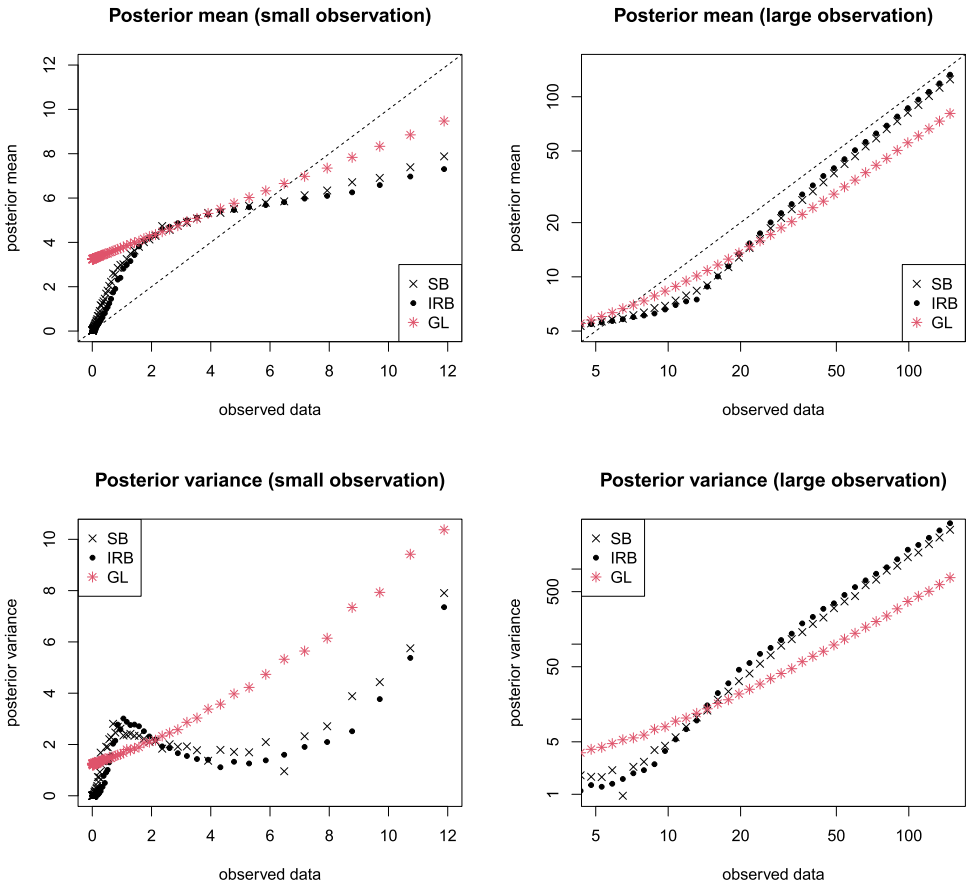

Figure 3: Posterior means and variances of λ_i for various observed values.

We consider $t = (t_1, \ldots, t_n) \in (0, \infty)^n$ as a set of additional latent variables. For the global parameters, we consider the conjugate gamma priors $\pi(\beta) = \mathrm{Ga}(\beta \mid a_\beta, b_\beta)$ and $\pi(\tau) = \mathrm{Ga}(\tau \mid a_\tau, b_\tau)$.

The variables λ, β, τ, t, and ν are updated in the following way.

- Sample $\lambda_i \sim \mathrm{IG}(\delta_i + \nu_i + 1, \delta_i y_i + \beta\nu_i)$ independently for $i = 1, \ldots, n$.
- Sample $\beta \sim \mathrm{Ga}(\sum_{i=1}^n \nu_i + n + a_\beta, \sum_{i=1}^n \nu_i/\lambda_i + b_\beta)$.
- Sample $\tau \sim \mathrm{GIG}(-na + a_\tau, 2b_\tau, 2\sum_{i=1}^n t_i\nu_i)$, where $\mathrm{GIG}(a, b, \gamma)$ has density proportional to $x^{a-1}\exp(-bx/2 - \gamma/2x)$.
- Sample $t_i \sim \mathrm{Ga}(a + b, 1 + \nu_i/\tau)$ independently for $i = 1, \ldots, n$.

- The full conditional distribution of ν_i is proportional to

$$\prod_{i=1}^{n}\{\mathrm{Ga}(\nu_i \mid a, t_i/\tau)\mathrm{Ga}(1/\lambda_i \mid \nu_i, \beta\nu_i)\},$$

  which can be accurately approximated by using the method of Miller (2019) for each $i = 1, \ldots, n$. The method is based on the gamma approximation of intractable probability density function by matching the first- and second-derivatives of log densities. We use the approximate full conditional distributions as proposal distributions in independent Metropolis-Hastings (MH) steps.

The full conditional distributions of parameters and latent variables other than ν_i are of familiar forms. Even for the full conditional of v_i, we can efficiently sample from the distribution. Note that the number of latent variables in the proposed priors is larger than that of GL prior to exhibit global-local shrinkage properties. Hence, the computation time of the MCMC (Markov chain Monte Carlo) algorithm with the proposed priors can be longer than that of the GL prior. Specifically, in the example given in Section 2.4, the computation times of SB and IRB to generate 5000 posterior samples are around 5 seconds while that of GL is less than 1 second. Such an increase in computational costs would be a reasonable price for the desirable shrinkage properties.

3 Theoretical properties

In this section, we analytically compare properties of different priors for u_i and, in particular, show two properties of the proposed priors, namely, tail-robustness for large observations (Section 3.1) and desirable Kullback-Leibler risk bound under sparsity (Section 3.2). For simplicity, we fix $\beta = \tau = 1$ in what follows so that all the theoretical results are conditional on the hyperparameters.

3.1 Tail-robustness for large observations

For a prior $\pi(u_i)$ of local parameter u_i, we consider the class given by

$$\sup_{u \geqslant 1}\{u\pi(u)\} < \infty, \tag{3.1}$$

$$\pi(u) \sim C\frac{u^{\alpha-1}}{\{1+\log(1+1/u)\}^{1+\gamma}} \quad \text{as } u \to 0 \text{ for some } \alpha \geqslant 0 \text{ and } \gamma \geqslant -1, \tag{3.2}$$

where C is a positive constant. The notation $f(x) \sim g(x)$ means $\lim_{x\to 0} f(x)/g(x) = 1$. Condition (3.1) is a technical condition satisfied by most priors. Condition (3.2) is a condition on the tail of $\pi(u_i)$ at the origin and is satisfied by both the SB and the IRB priors. Because we consider proper distributions only in this paper, the case of $\alpha = 0$ and $\gamma \leq 0$ is excluded.

We consider the tail robustness of the Bayes estimator of λ_i given by

$$\hat{\lambda}_i = \hat{\lambda}_i^{\mathrm{ML}} - E(\kappa_i \mid y_i)(\hat{\lambda}_i^{\mathrm{ML}} - 1),$$

where $\hat{\lambda}_i^{\mathrm{ML}} = y_i$ and $\kappa_i = u_i/(\delta_i + u_i)$. Specifically, we show that the expected shrinkage factor, $E(\kappa_i \mid y_i)$, converges to zero as $y_i \to \infty$.

Theorem 3.1. *There exists a function* $\kappa^* : (0, \infty) \to (0, \infty)$ *such that*

$$E(\kappa_i \mid y_i) \sim \frac{1}{\delta_i}(1+\alpha)\kappa^*(\delta_i y_i) \to 0$$

as $y_i \to \infty$.

Since the local parameter depends on not only the scale parameter but also the shape parameter, the evaluation of the posterior mean requires a detailed investigation of integrals involving gamma functions, where the details of the proof are given in the Supplementary Material. The constant $\alpha \geqslant 0$ is related to the tail of $\pi(u_i)$ at the origin. The heavier the tail is, the faster the expected shrinkage factor converges to zero. Finally, we note that if we fix $u_i = 1$, then $E(\kappa_i \mid y_i) = 1/(\delta_i + 1)$ does not converge to 0 as $y_i \to \infty$.

In the Supplementary Material, we further investigate the rate of $\kappa^*(y)$, which shows that $\kappa^*(y) = 1/\log y$ as $y \to \infty$. This means that $E(\kappa_i \mid y_i)$ converges to 0 very slowly as $y_i \to \infty$, while it remains a positive constant when $u_i = 1$ even under $y_i \to \infty$. This property of $E(\kappa_i \mid y_i)$ indicates that $(E[\lambda_i \mid y_i] - y_i)/y_i \to 0$ as $y_i \to \infty$, which is known as weakly tail-robust (Hamura et al., 2022b). Such property is also adopted to show the robustness of shrinkage under count response (Datta and Dunson, 2016) and correlated normal response (Okano et al., 2022).

In the Supplementary Material, we also investigate the behavior of $E(\kappa_i \mid y_i)$ as $y_i \to 0$, where it is shown that $E(\kappa_i \mid y_i) \to 0$ as $y_i \to 0$ if either $\pi(u_i) = \pi_{\mathrm{SB}}(u_i)$ with $\delta_i \geq a$ or $\pi(u_i) = \pi_{\mathrm{IRB}}(u_i)$. This indicates that tail-robustness for a small observation is also established.

3.2 Kullback-Leibler super-efficiency under sparsity

We now consider the predictive efficiency for the proposed method (e.g. Polson and Scott, 2010; Carvalho et al., 2010; Datta and Dunson, 2016). In particular, we discuss the Kullback-Leibler divergence between the true sampling density and the Bayes predictive density under the proposed global-local shrinkage prior. We consider the following one-dimensional model

$$y \sim \mathrm{Ga}\left(\delta, \frac{\delta}{\lambda}\right), \quad \lambda \sim \mathrm{IG}(1+u, u), \quad u \sim \pi(u).$$

In the above model, let $f(y \mid \lambda) = \mathrm{Ga}(y \mid \delta, \delta/\lambda)$ and let λ_0 be the true value of λ. We define the Kullback-Leibler (KL) divergence between $f(y \mid \lambda)$ and $f(y \mid \lambda_0)$ by $D^{\mathrm{KL}}(\lambda_0, \lambda) = D^{\mathrm{KL}}(f(y \mid \lambda_0), f(y \mid \lambda))$. Then we have

$$D^{\mathrm{KL}}(\lambda_0, \lambda) = \delta\left(\frac{1/\lambda}{1/\lambda_0} - 1 - \log\frac{1/\lambda}{1/\lambda_0}\right) = \delta\left(\frac{\lambda_0}{\lambda} - 1 - \log\frac{\lambda_0}{\lambda}\right).$$

Furthermore, the KL neighborhood around λ_0 is defined by

$$A_\varepsilon(\lambda_0) = \left\{\lambda \in (0,\infty) \mid D^{\mathrm{KL}}(\lambda_0, \lambda) < \varepsilon\right\}.$$

We assume that the prior $p(\lambda)$ is information dense in the sense of $\mathrm{pr}(\lambda \in A_\varepsilon(\lambda_0)) > 0$ for all $\varepsilon > 0$. From the Proposition 4 in Barron (1987), we have the Cesáro-mean risk R_n is expressed by

$$R_n \leq \varepsilon - n^{-1} \log \mathrm{pr}(\lambda \in A_\varepsilon(\lambda_0)), \tag{3.3}$$

where $R_n = n^{-1} \sum_{k=1}^n D^{\mathrm{KL}}(f(y \mid \lambda_0) \mid \hat{f}_k(\lambda))$ and $\hat{f}_k(\lambda)$ is the Bayes predictive density under KL divergence using the posterior density based on $k \leq n$ observations $y_1, \ldots, y_k$. We now evaluate the prior probability $\mathrm{pr}(\lambda \in A_\varepsilon(\lambda_0))$ in the right-hand side of (3.3) when $\lambda_0 = 1$.

Although we proved the theorem for the univariate case, the convergence in the multivariate case is derived from a component-wise application.

Theorem 3.2. *Assume that the true sampling model is* $\mathrm{Ga}(\delta, \delta/\lambda_0)$. *For* $\lambda_0 \neq 1$, *the Cesáro-mean risk for Bayes predictive density* $\hat{f}_n$, *which is the posterior mean of the density function* $f(\cdot \mid \lambda)$, *satisfies*

$$R_n = O\left(n^{-1} \log n\right).$$

If $\lambda_0 = 1$ *and if* $\pi(u) \propto u^{-1-b}$ *as* $u \to \infty$ *for some* $0 < b \leq 1/2$, *then*

$$R_n = O\left\{n^{-1} (\log n - \log\log n)\right\}.$$

The proof of the theorem is given in the Supplementary Material. The results indicate that the Cesáro-mean risk achieves the optimal rate of convergence for the finite-dimensional parametric family when $\lambda_0 \neq 1$, while the risk has the super-efficient rate of Kullback-Leibler convergence for $\lambda_0 = 1$. The latter phenomenon is called *Kullback-Leibler super-efficiency*, which is a kind of higher-order optimality, and such results are commonly adopted to show theoretical superiority in handling sparsity in the context of global-local shrinkage priors (e.g. Polson and Scott, 2010; Carvalho et al., 2010; Datta and Dunson, 2016). Theorem 3.2 relates the right tail of $\pi(u_i)$ to the risk given in (3.3). To achieve Kullback-Leibler super-efficiency, it is sufficient to use $\pi(u_i)$ with a sufficiently heavy tail ($b \leq 1/2$). Thus, b plays a role in controlling sparsity at the grand mean. We remark that fixing $u_i = 1$ corresponds to using a point mass prior for u_i and hence to violation of the sufficient condition that $\pi(u) \propto u^{-1-b}$ as $u \to \infty$.

4 Simulation studies

We evaluate the performance of Bayesian and frequentist shrinkage methods under gamma response. Let $y_i \sim \mathrm{Ga}(\delta_i, \delta_i/\lambda_i)$ for $i = 1, \ldots, n(= 200)$ and $\delta_i = 5$. We consider the following six scenarios of the true mean λ_i:

(Scenario 1) $\lambda_i \sim 0.95\delta_\mu + 0.05\mathrm{Ga}(20\mu, 2)$, (Scenario 2) $\lambda_i \sim 0.9\delta_\mu + 0.1\mathrm{Ga}(20\mu, 2)$,

(Scenario 3) $\lambda_i \sim 0.95\delta_\mu + 0.05\mu|t_3|$, (Scenario 4) $\lambda_i \sim 0.9\text{Ga}(5\mu, 5) + 0.1\mu|t_1|$,
(Scenario 5) $\lambda_i \sim 0.9\delta_\mu + 0.1\text{Ga}(10\mu, 2)$, (Scenario 6) $\lambda_i \sim 0.85\delta_\mu + 0.15\text{Ga}(10\mu, 2)$,

where $\mu = 5$, δ_a denotes a point mass at a and t_c denotes a t-distribution with c degrees of freedom. In the first three scenarios, most of the true means λ_i is exactly equal to $\mu = 5$, and a small part of true means are very large compared with μ. In scenario 4, most true means are concentrated around μ (not exactly equal to 0).

For the simulated data, we apply six methods, the proposed scaled beta (SB) and inverse rescaled beta (IRB) priors, the global shrinkage (GL) prior (setting $u_i = 1$ in the proposed model), shrinkage estimators given by DasGupta (1986) (DG), adaptive variance shrinkage estimators by Lu and Stephens (2016) (VS), and maximum likelihood (ML) estimator y_i. Note that the DG method is to provide decision-theoretic point estimates of λ_i by minimizing a weighted quadratic loss function, and the VS method uses a finite mixture of inverse-gamma distributions as a prior distribution for λ_i. The tuning parameters in the SB and IRB priors are set to $a = 2$ and $b = 1/2$. We used non-informative gamma priors, $\beta \sim \text{Ga}(0.1, 0.1)$ and $\tau \sim \text{Ga}(0.1, 0.1)$ for SB, IRB and GL priors. For the Bayesian methods, 3000 posterior samples are generated after discarding the first 2000 samples as burn-in. Note that we used the R package "vashr" (`https://github.com/mengyin/vashr`) to apply the VS method, where the degrees of freedom of χ^2-distribution is set to $2\delta_i (= 10)$.

We first investigate the shrinkage property of the proposed global-local shrinkage priors compared with the other methods. In Figure 4, we show scatter plots of observed values and point estimates (posterior means for the Bayesian methods) produced by five shrinkage methods under scenario 1. It is observed that the standard shrinkage methods, GL, DG, and VS, linearly shrink the observed value y_i, that is, the shrinkage factor is constant regardless of y_i. On the other hand, the proposed SB and IRB priors more strongly shrink the observed values around $\lambda_i = 5$, showing the adaptive shrinkage property of the global-local shrinkage prior.

We next evaluate mean absolute percentage error (MAPE), defined as $n^{-1}\sum_{i=1}^n \lambda_i^{-1}|\lambda_i - \widehat{\lambda}_i|$ with a point estimate $\widehat{\lambda}_i$. We present boxplots of MAPE for 1000 replications in Figure 5. The results indicate that the proposed SB and IRB provide more accurate point estimates than the other methods in all the scenarios, except for IRB under Scenario 3. The amount of improvement of the proposed methods is remarkable when the null and non-null signals are well-separated, as in Scenarios 1 and 2. Comparing SB and IRB, SB tends to provide a smaller overall MAPE than IRB. To compare the two methods more precisely, we also computed MAPE only for non-null signals. The averaged values of MAPE for non-null signals are given in Table 1, which shows that IRB performs slightly better than SB for the estimation of non-null signals, and this is consistent with the stronger tail-robustness property of IRB than that of SB.

Furthermore, we computed the coverage probability (CP) and average length (AL) of 95% credible/confidence intervals. We only consider the ML method for the frequentist methods since DG and VS do not provide interval estimation. The 95% confidence interval of ML can be obtained as $(y_i/\text{P}_G(0.975; \delta_i, \delta_i), y_i/\text{P}_G(0.025; \delta_i, \delta_i))$, where $\text{P}_G(\cdot; \alpha, \beta)$ denotes the probability function of $\text{Ga}(\alpha, \beta)$. The CP and AL averaged over 1000 Monte

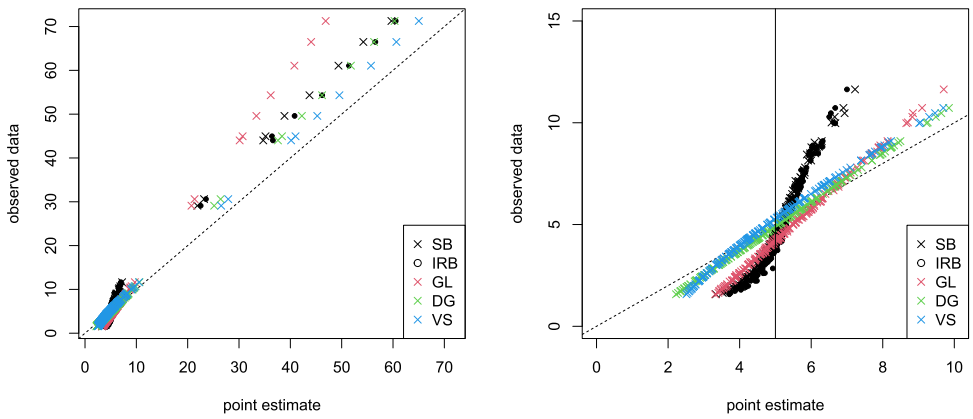

Figure 4: Scatter plots of observed values (ML) and point estimates obtained from five shrinkage methods. The right panel is an enlarged version of the left panel. The vertical line in the right panel indicates the location of null signals.

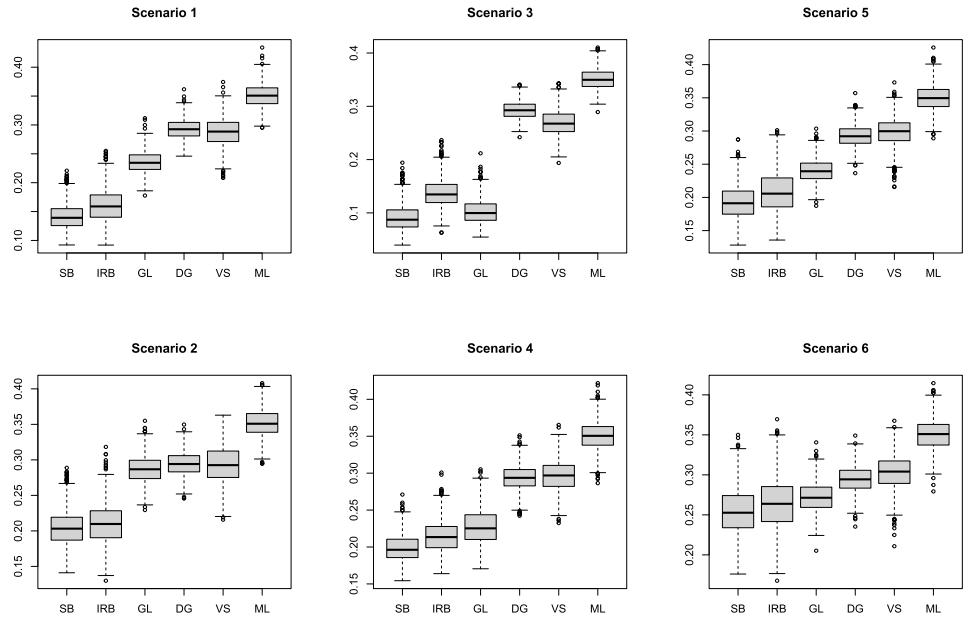

Figure 5: Boxplots of mean absolute percentage errors (MAPE) for 1000 Monte Carlo replications.

Carlo replications are given in Table 2. It can be seen that all three Bayesian methods have empirical CP values larger than the nominal level of 0.95 except in Scenario 6, whereas the interval lengths of SB and IRB tend to be smaller than GL and ML in all the scenarios.

Scenario	1	2	3	4	5	6
SB	0.365	0.359	0.304	0.292	0.388	0.384
IRB	0.360	0.360	0.284	0.286	0.385	0.386

Table 1: Mean absolute percentage errors (MAPE) for non-null signals averaged over 1000 Monte Carlo replications.

	Coverage probability				Average length			
Scenario	SB	IRB	GL	ML	SB	IRB	GL	ML
1	98.7	98.4	97.6	95.0	1.13	1.19	1.58	2.59
2	97.3	97.1	97.1	95.0	1.32	1.33	1.79	2.59
3	98.1	98.4	97.9	95.0	0.81	1.09	0.91	2.59
4	95.6	96.0	97.2	95.0	1.12	1.23	1.43	2.59
5	96.6	96.3	96.7	95.0	1.23	1.27	1.56	2.59
6	94.4	94.0	96.2	95.0	1.37	1.39	1.69	2.59

Table 2: Coverage probabilities and average lengths of 95% credible/confidence intervals averaged over 1000 Monte Carlo replications.

5 Real data example

5.1 Average admission period of COVID-19 in Korea

We first apply the global-local shrinkage techniques to estimate the average length of hospital stay of COVID-19-infected persons. We use the data set available at Kaggle (https://www.kaggle.com/kimjihoo/coronavirusdataset), where the date of admission and discharge is observed for 1587 individuals in Korea. We then group these individuals regarding 98 cities and three classes of age, young (39 or less), middle (from 40 to 69), and old (70 or more), resulting in $n = 185$ groups after omitting empty groups. Assuming exponential distributions with group-specific mean for admission period (days) of each individual, the group-wise sample mean is distributed as $\mathrm{Ga}(n_i, n_i/\lambda_i)$ for $i = 1, \ldots, n$, where n_i is the number of individuals within the ith group and λ_i is the true mean of admission period specific to the ith group. Note that n_i ranges from 1 to 258, and the scatter plot of n_i and y_i are given in Figure 7.

We apply the proposed SB prior as well as GL and DG methods. Regarding the prior distributions for grand mean β in the SB and GL models, we assign a non-informative prior, $\mathrm{Ga}(0.1, 0.1)$. Furthermore, we set $a = 2$ and $b = 1/2$ in the proposed prior distributions. The posterior means for SB and GL are computed based on 10000 posterior samples (after discarding 3000 samples), whose histograms are shown in Figure 6. The posterior mean of the grand mean β in the SB model was 22.0 (95% credible interval was $(21.3, 22.8)$), which is consistent with the evidence that the average admission period is around 21 (e.g. Jang et al., 2021). On the other hand, the posterior mean of the grand mean β in the GL model was 24.4, where 95% credible interval was $(22.5, 26.6)$. We also present the histogram of shrinkage estimates made by DG. It is observed that SB strongly shrinks the observed values toward the grand mean β so that most of the posterior means of the average admission period are concentrated around the grand mean.

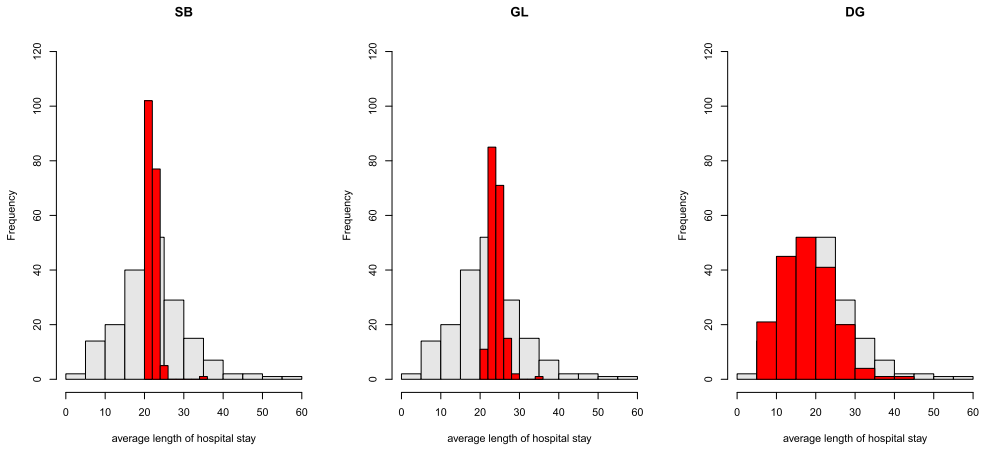

Figure 6: Histograms of shrinkage estimates (red) and observed values (gray) of the average admission period.

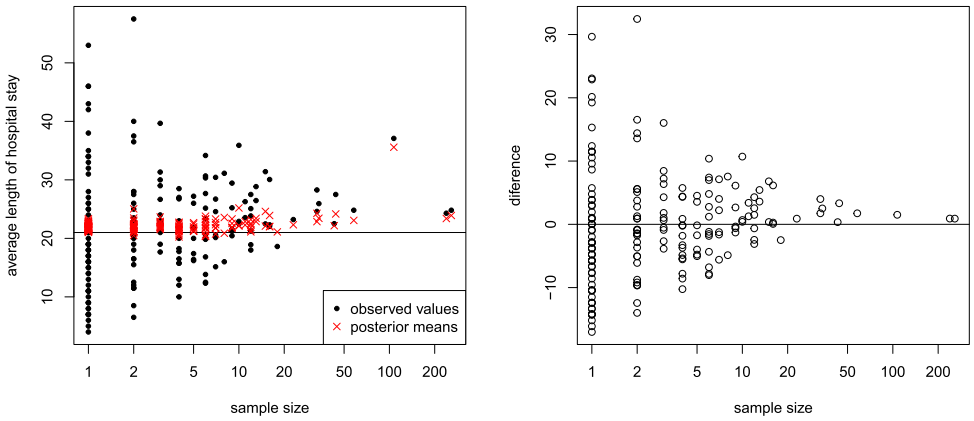

Figure 7: Left: Scatter plot of sample size n_i and average admission. Right: Scatter plot of sample size n_i and difference of y_i and posterior mean of λ_i.

This is because most groups having large sample means have small sample sizes, and such unreliable information is strongly shrunk. We found that only a single group (old age class of Gyeongsan-si) has a much larger average admission period, about 35 days. Since the sample size of this group is 107, and the sample mean is about 37, the posterior result seems reasonable. To see more detailed results, we present scatter plots of observed values and posterior means against sample size n_i, in Figure 7. It is observed that the amount of shrinkage (i.e., the difference between observed values and posterior means) decreases as n_i increases, and observations having small sample sizes strongly shrunk toward the grand mean. From Figure 6, it can also be seen that GL also provides reasonably shrunk estimates of λ_i and DG does not, but the proposed SB prior can provide

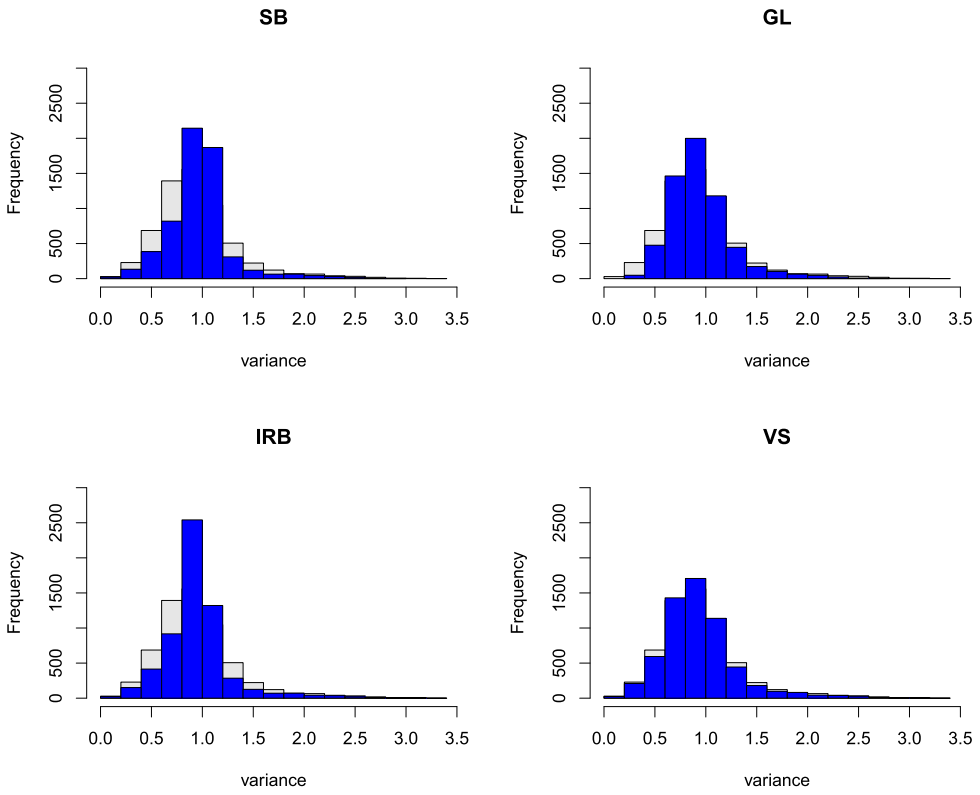

Figure 8: Histograms of shrinkage estimates (blue) and observed values (gray) of variances of gene expression data.

strongly shrunk point estimates. Moreover, the average length of 95% credible intervals made by SB was 19.0, which was considerably smaller than the 22.9 produced by GL.

5.2 Variance estimation of gene expression data

We next apply the shrinkage methods to variance estimation of gene expression data. As noted in Lu and Stephens (2016), in gene expression analysis aimed at identifying deferentially expressed genes, accurate estimation of the unknown variance is an essential step since it directly relates to the degree of statistical significance. We use a popular prostate cancer dataset from Singh et al. (2002). In this dataset, there are gene expression values for $n = 6033$ genes for 50 subjects in control subjects. We compute sampling variances of n gene expressions, distributed as $\mathrm{Ga}(n_i/2, n_i/2\lambda_i)$ for $i = 1, \ldots, n$, where λ_i is the true variance of the ith gene expression. By assigning non-informative priors, $\mathrm{Ga}(0.1, 0.1)$ for β in the SB, IRB, and GL models as well as the VS method. In the Bayesian methods, we computed posterior means using 2000 posterior samples after discarding the first 1000 samples. The histograms of posterior means are shown in Figure 8. As confirmed

in the previous example, we can see that the proposed SB and IRB priors can provide more shrunk estimates than the other methods. Furthermore, average lengths of 95% credible intervals made by SB and IRB were 0.640 and 0.639, respectively, which are smaller than 0.663 by GL. This shows the efficiency of the proposed priors.

6 Discussion

We proposed a new class of continuous global-local shrinkage priors for high-dimensional positive-valued parameters based on shape-scale mixtures of inverse-gamma distributions. Although this paper focuses on a sequence of gamma-distributed observations, it can also be useful in other models. One notable application would be using a flexible error distribution in a regression model for gamma-distributed observations, such as gamma regression or accelerated failure time models. For the latter model, the proposed distribution may cast an alternative to the Bayesian nonparametric approach (e.g. Hanson, 2006; Kuo and Mallick, 1997), and some comparisons would be an interesting future study. Although the approximate sampling method of Miller (2019) is adopted in sampling from the local parameter in the proposed priors, it might be worth implementing the more recent data augmentation technique of Hamura et al. (2022a).

Supplementary Material

Supplementary Materials for "Sparse Bayesian inference on gamma-distributed observations using shape-scale inverse-gamma mixtures" (DOI: 10.1214/22-BA1348SUPP; .pdf). Supplementary material available online includes proof of the theoretical results.

References

Armagan, A., Dunson, D., and Lee, J. (2013). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23: 119–143. MR3076161. 78

Barron, A. R. (1987). "Are Bayes rules consistent in information?" In *Open problems in communication and computation*, 85–91. 89

Berger, J. (1980). "Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters." *The Annals of Statistics*, 8(3): 545–571. MR0568720. 78

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). "The horseshoe+ estimator of ultra-sparse signals." *Bayesian Analysis*, 12(4): 1105–1131. MR3724980. doi: https://doi.org/10.1214/16-BA1028. 78

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace priors for optimal shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 78

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97: 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 78, 81, 83, 88, 89

DasGupta, A. (1986). "Simultaneous estimation in the multiparameter gamma distribution under weighted quadratic losses." *The Annals of Statistics*, 14(1): 206–219. MR0829563. doi: https://doi.org/10.1214/aos/1176349850. 78, 90

Datta, J. and Dunson, D. (2016). "Bayesian inference on quasi-sparse count data." *Biometrika*, 103(4): 971–983. MR3620451. doi: https://doi.org/10.1093/biomet/asw053. 78, 81, 88, 89

Dey, D., Ghosh, M., and Srinivasan, C. (1987). "Simultaneous estimation of parameters under entropy loss." *Journal of Statistical Planning and Inference*, 15: 347–363. MR0879215. doi: https://doi.org/10.1016/0378-3758(86)90108-4. 78

Donoho, D. and Jin, J. (2006). "Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data." *The Annals of Statistics*, 34(6): 2980–3018. MR2329475. doi: https://doi.org/10.1214/009053606000000920. 78, 79

Ghosh, M. and Parsian, A. (1980). "Admissible and minimax multiparameter estimation in exponential families." *Journal of Multivariate Analysis*, 10: 551–564. MR0599689. doi: https://doi.org/10.1016/0047-259X(80)90069-X. 78

Hamura, Y., Irie, K., and Sugasawa, S. (2020). "Shrinkage with Robustness: Log-Adjusted Priors for Sparse Signals." *arXiv preprint arXiv:2001.08465*. 78

Hamura, Y., Irie, K., and Sugasawa, S. (2021). "Robust Hierarchical Modeling of Counts under Zero-inflation and Outliers." *arXiv preprint arXiv:2106.10503*. 80

Hamura, Y., Irie, K., and Sugasawa, S. (2022a). "On Data Augmentation for Models Involving Reciprocal Gamma Functions." *Journal of Computational and Graphical Statistics*. 95

Hamura, Y., Irie, K., and Sugasawa, S. (2022b). "On global-local shrinkage priors for count data." *Bayesian Analysis*, 17(2): 545–564. MR4483230. doi: https://doi.org/10.1214/21-ba1263. 78, 88

Hamura, Y., Onizuka, T., Hashimoto S., and Sugasawa, S. (2022). "Supplementary Materials for "Sparse Bayesian inference on gamma-distributed observations using shape-scale inverse-gamma mixtures"" *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1348SUPP. 79

Hanson, T. E. (2006). "Modeling censored lifetime data using a mixture of gammas baseline." *Bayesian Analysis*, 1(3): 575–594. MR2221289. doi: https://doi.org/10.1214/06-BA119. 95

Jang, S. Y., Seon, J.-Y., Yoon, S.-J., Park, S.-Y., Lee, S. H., and Oh, I.-H. (2021). "Comorbidities and factors determining medical expenses and length of stay for admitted COVID-19 patients in Korea." *Risk Management and Healthcare Policy*, 14. 92

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons. MR1326603. 80

Kuo, L. and Mallick, B. (1997). "Bayesian semiparametric inference for the accelerated failure-time model." *Canadian Journal of Statistics*, 25(4): 457–472. 95

Kwon, Y. and Zhao, Z. (2022). "On F-modelling-based empirical Bayes estimation of variances." *Biometrika*. doi: https://doi.org/10.1093/biomet/asac019. 78

Lu, M. and Stephens, M. (2016). "Variance adaptive shrinkage (vash): flexible empirical Bayes estimation of variances." *Bioinformatics*, 32(22): 3428–3434. 78, 79, 90, 94

Miller, J. W. (2019). "Fast and accurate approximation of the full conditional for gamma shape parameters." *Journal of Computational and Graphical Statistics*, 28(2): 476–480. MR3974896. doi: https://doi.org/10.1080/10618600.2018.1537929. 84, 87, 95

Okano, R., Hamura, Y., Irie, K., and Sugasawa, S. (2022). "Locally Adaptive Bayesian Isotonic Regression using Half Shrinkage Priors." *arXiv preprint arXiv:2208.05121*. 88

Pérez, M.-E., Pericchi, L. R., and Ramírez, I. C. (2017). "The scaled beta2 distribution as a robust prior for scales." *Bayesian Analysis*, 12(3): 615–637. MR3655869. doi: https://doi.org/10.1214/16-BA1015. 80

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. MR3204017. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. 88, 89

Polson, N. G. and Scott, J. G. (2012). "Local shrinkage rules, Lévy processes and regularized regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2): 287–311. MR2899864. doi: https://doi.org/10.1111/j.1467-9868.2011.01015.x. 77

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., et al. (2002). "Gene expression correlates of clinical prostate cancer behavior." *Cancer cell*, 1(2): 203–209. 94

Sun, D. and Berger, J. O. (1998). "Reference priors with partial information." *Biometrika*, 85(1): 55–71. MR1627242. doi: https://doi.org/10.1093/biomet/85.1.55. 81

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior." *Journal of the American Statistical Association*, 1–13. MR4436318. doi: https://doi.org/10.1080/01621459.2020.1825449. 78

Acknowledgments

We thank two anonymous reviewers for useful suggestions, which improved the quality of this work.