

ASYMPTOTICALLY MINIMAX PREDICTIVE DENSITY FOR SPARSE COUNT DATA

KEISUKE YANO, RYOYA KANEKO, AND FUMIYASU KOMAKI

ABSTRACT. Predictive density estimation under the Kullback–Leibler loss in high-dimensional sparse count data models is discussed. In particular, Poisson sequence models under sparsity constraints are discussed. Sparsity in count data implies zero-inflation or quasi zero-inflation, that is, situations where there exists an excess of zeros or near-zero counts. We investigate the exact asymptotic minimax Kullback–Leibler risks in sparse and quasi-sparse Poisson sequence models, providing a class of Bayes predictive densities that attain exact asymptotic minimaxity. For application, we discuss adaptation to an unknown sparsity, and also discuss the performance of the proposed Bayes predictive densities in settings where current observations are missing completely at random. Both simulation studies and applications to real data show the efficiency of the proposed Bayes predictive densities.

1. INTRODUCTION

Predictive density is a probability density of future observations on the basis of current observations. It is used not only to estimate future observations but also to quantify uncertainty of them. It has a wide range of application in statistics, information theory, and machine learning. The simplest class of predictive densities is a class of *plug-in* predictive densities. A plug-in predictive density is constructed by substituting an estimator into an unknown parameter of a statistical model. Another class of predictive densities is a class of *Bayes* predictive densities. A Bayes predictive density is the posterior mixture of densities of future observations. There is a vast literature in predictive density estimation within statistical models in finite dimensions; see Subsection 1.2 for the literature review. One of important findings in the literature is that one can construct the Bayes predictive density better than a plug-in predictive density under the Kullback–Leibler loss; see [25]. In contrast, little is known about predictive density estimation within statistical models in high dimensions. Important works in this direction are [41, 40]. [41, 40] construct several predictive densities (including a Bayes predictive density) superior to all plug-in predictive densities even in sparse high-dimensional Gaussian models.

The aim of this paper is to construct an efficient predictive density for high-dimensional sparse count data. Efficiency of a predictive density is measured by the supremum of the Kullback–Leibler risk under sparsity constraints. Two types of sparsity are discussed in this paper: one is *exact* sparsity. Exact sparsity in count data means that there exhibits an excess of zeros. The other

Date: December 17, 2018.

Key words and phrases. Adaptation; High dimension; Kullback–Leibler divergence; Missing Completely At Random; Poisson model; Zero inflation.

K. Yano is supported by the Grant-in-Aid for Research Activity start-up (17H06570) from the JSPS and by JST CREST (JPMJCR1763).

is *quasi* sparsity. Quasi sparsity means that there exhibits an excess of near-zero counts. Exact sparsity is captured by ℓ_0 -norm, and quasi sparsity is captured by the number of near-zero counts. See Subsection 1.1 for the formulation.

Motivation for analyzing sparse count data is well-known. In analyzing high-dimensional count data, there often exhibits exact sparsity and quasi sparsity corresponding to inflation of zeros and near-zero counts, respectively. Data with overabundance of zeros include agriculture [18], environmental sciences [1], and manufacturing [33]. Data with overabundance of near-zero counts include DNA sequencing and terrorist attacks [8]. Another example (Japanese crime statistics) is presented in Section 4.

1.1. Problem setting and contributions. Main results are summarized with the problem formulation ahead. Let X_i ($i = 1, 2, \dots, n$) be a current observation independently distributed according to $\text{Po}(r\theta_i)$, and let Y_i ($i = 1, 2, \dots, n$) be a future observation independently distributed according to $\text{Po}(\theta_i)$, where $\theta = (\theta_1, \dots, \theta_n)$ is an unknown parameter and r is a known constant. Constant r signifies the ratio of the mean of the i -th ($i = 1, \dots, n$) current observation to that of the i -th future observation. By sufficiency reduction, this constant means the ratio of sample sizes of current observations to those of future observations. Suppose that $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ are independent. The densities of X and Y with parameter θ are denoted by $p(x | \theta)$ and $q(y | \theta)$, respectively:

$$p(x | \theta) = \prod_{i=1}^n \left\{ \frac{1}{x_i!} e^{-r\theta_i} (r\theta_i)^{x_i} \right\} \text{ and } q(y | \theta) = \prod_{i=1}^n \left\{ \frac{1}{y_i!} e^{-\theta_i} \theta_i^{y_i} \right\}.$$

Two types of parameter spaces are of interest: one is the exact sparse parameter space. Given $s \in (0, n)$, it is described as $\Theta[s] := \{\theta \in \mathbb{R}_+^n : \|\theta\|_0 \leq s\}$, where $\|\cdot\|_0$ is ℓ_0 -norm given by $\|\theta\|_0 := \#\{i : \theta_i > 0\}$. The other is the quasi sparse parameter space. Given $s \in (0, n)$ and a threshold $\varepsilon > 0$, it is described as $\Theta[s, \varepsilon] := \{\theta \in \mathbb{R}_+^n : N(\theta, \varepsilon) \leq s\}$, where $N(\theta, \varepsilon) := \#\{i : \theta_i > \varepsilon\}$, $\varepsilon > 0$. A threshold ε determines whether the parameter value of each coordinate is near-zero or not.

The performance of a predictive density \hat{q} is evaluated by the Kullback–Leibler loss

$$L(\theta, \hat{q}(\cdot; x)) = \sum_{y \in \mathbb{N}^n} q(y | \theta) \log \frac{q(y | \theta)}{\hat{q}(y; x)}.$$

The corresponding risk (expected loss) is denoted by

$$R(\theta, \hat{q}) = \sum_{x \in \mathbb{N}^n} \sum_{y \in \mathbb{N}^n} p(x | \theta) q(y | \theta) \log \frac{q(y | \theta)}{\hat{q}(y; x)}.$$

The minimax Kullback–Leibler risk over $\Theta[s]$ is defined as

$$\mathcal{R}(\Theta[s]) := \mathcal{R}_n(\Theta[s]) = \inf_{\hat{q}} \sup_{\theta \in \Theta[s]} R(\theta, \hat{q}).$$

For comparison, the minimax Kullback–Leibler risk among plug-in predictive densities is denoted by $\mathcal{E}(\Theta[s]) := \inf_{\hat{q}} \sup_{\theta \in \Theta[s]} R(\theta, q(\cdot | \hat{\theta}))$. The minimax risk $\mathcal{R}(\Theta[s])$ ($\mathcal{E}(\Theta[s])$) is called predictive (estimative) minimax risk, (respectively,) because $\mathcal{R}(\Theta[s])$ depends on the whole class of predictive

densities, while $\mathcal{E}(\Theta[s])$ depends on the whole class of estimators of θ . Likewise, two minimax Kullback–Leibler risks $\mathcal{R}(\Theta[s, \varepsilon])$ and $\mathcal{E}(\Theta[s, \varepsilon])$ over $\Theta[s, \varepsilon]$ are defined.

To express high-dimensional settings under sparsity constraints, we employ the high dimensional asymptotics in which $n \rightarrow \infty$ and $s/n = s_n/n \rightarrow 0$. The values of s and ε possibly depend on n and thus in what follows the dependence on n is often expressed as $s = s_n$ and $\varepsilon = \varepsilon_n$.

Main theoretical contributions of this paper are as follows. (i) exact asymptotic minimax risks $\mathcal{R}(\Theta[s_n])$ and $\mathcal{R}(\Theta[s_n, \varepsilon_n])$ over $\Theta[s_n]$ and $\Theta[s_n, \varepsilon_n]$ are identified in Theorem 2.1. Constant r is shown to be the only key parameter to describe the ratio of the predictive minimax risks to the estimative minimax risks as shown in Proposition 2.1: the ratio decreases as r increases; (ii) a class of Bayes predictive densities attaining the exact asymptotic minimaxity is presented in Theorem 2.2. Two types of adaptation to unknown sparsity are also discussed; (iii) the proposed predictive densities are also shown to attain the exact asymptotic minimaxity in the settings where current observations are missing completely at random (MCAR) as discussed in Section 3. Consideration to adaptation and to missing observations are important in application. In two real-data examples in Section 4, we do not have a priori knowledge of s_n and there exhibit missing values of observations; hence we have to take care of adaptation and missing observations.

Practical effectiveness of the proposed Bayes predictive densities is examined by both simulation studies and applications to real data in Section 4. These studies show that the proposed Bayes predictive densities are effective in senses of not only point prediction but also predictive uncertainty quantification.

The proposed class of predictive densities builds upon spike-and-slab prior distributions. What is important here is the tail behavior of a slab prior. Interestingly, spike-and-slab prior distributions with slab priors having exponential tails do not yield asymptotically exact minimax predictive densities as Proposition 2.5 indicates. To obtain predictive densities that are not only asymptotically minimax but also easily implemented by the exact sampling, we use spike-and-slab priors with improper slab priors. It is known that the posterior distributions based on spike-and-slab priors with improper priors are affected by the scaling of improper priors; see [20, 34, 9]. Here we calibrate the scale as posterior distributions yield asymptotically minimax optimal predictive densities. Details are presented in Subsection 2.2.

Theoretically, one of important properties of spike-and-slab priors in Poisson sequence models is that they yield easy-to-analyze predictive densities. Owing to the easy-to-analyze form, we establish user-friendly lower and upper bounds on the posterior means, and we take adaptation into consideration in our analysis. The bounds on the posterior means are used to bound the Kullback–Leibler risk of the Bayes predictive densities in conjunction with an integral representation of the Kullback–Leibler risk in Poisson sequence models; see Subsection 5.2. The adaptation result is obtained by the easy-to-analyze form in conjunction with the decomposition of the Kullback–Leibler divergence; see Subsection 2.2.2. Computationally, sampling from the proposed predictive densities is easy because a predictive density in the proposed class is the product of zero-inflated negative-binomial distributions.

1.2. Literature review. There is a rich literature on predictive density estimation in fixed finite dimensions. In the literature, Bayes predictive densities are showed to dominate plug-in predictive densities. Studies of Bayes predictive densities date back to [2, 42, 3, 43]. The first quantitative comparison of Bayes and plug-in predictive densities in a wide class of parametric models is [25]. [25] showed that there exists a Bayes predictive density that dominates a plug-in predictive density under the Kullback–Leibler loss, employing asymptotic expansions of Bayes predictive densities; see also [12, 45, 19] for asymptotic expansions of Bayes predictive densities. Minimax Bayes predictive densities for unconstrained parameter spaces are studied in [35, 4, 30]. Minimax Bayes predictive densities under parametric constraints are studied in [15, 32, 36]. Shrinkage priors for Bayes predictive densities under Gaussian models are investigated in [26, 16, 24, 39]; see also [23, 5] for the cases where the variances are unknown. Shrinkage priors for Bayes predictive densities under Poisson models are developed in [27, 31]. The cases under α -divergence losses are covered by [7, 44, 38, 49].

Relatively little is known about predictive density estimation in high dimensions. [41, 40] elegantly construct asymptotically minimax predictive density for sparse Gaussian models. [46] obtained an asymptotically minimax predictive density for nonparametric Gaussian regression models under Sobolev constraints; later, [48] obtained an adaptive minimax predictive density for these models. See also [47]. All above results employ Gaussian likelihood and the corresponding results for count data have been not known. Our results are the first contribution to investigating efficient predictive density estimation for high-dimensional count data.

Poisson models are prototypical in analyzing count data as Gaussian models are in analyzing continuous data. Several authors have investigated correspondences between Gaussian models and Poisson models. [6, 21, 22, 37] discuss the correspondence in estimation of means using the re-scaled squared loss defined as $\sum_{i=1}^n \theta_i^{-1}(\theta_i - \hat{\theta}_i(X))^2$. [17, 27, 28, 31] discuss the correspondence in prediction using the Kullback–Leibler loss. In particular, [22, 37] study the exact asymptotic minimaxity under ellipsoidal and rectangle constraints in high-dimensional Poisson models using the re-scaled squared loss, providing results that are parallel with those in Gaussian models. In spite of the elegant correspondence in [22, 37], the re-scaled squared loss is not compatible with sparsity: the loss diverges if $\theta_i = 0$ and $\hat{\theta}_i(X) \neq 0$ for at least one index i .

Employing the Kullback–Leibler divergence, this paper presents results on exact asymptotic minimaxity in both estimation and prediction within sparse Poisson models, which are neatly parallel to the result for sparse Gaussian models by [41]; see Subsection 2.2 for detailed discussions. But, our results employ a different strategy in constructing a predictive density. This strategy makes the calculation of the Kullback–Leibler risk easy in Poisson models and enables us to consider the adaptation to sparsity. Furthermore, we cover results on several new topics in minimax predictive density estimation that come from the nature of count data analysis, that is, quasi-sparsity and Missing Completely At Random (MCAR); for the set-up of prediction with MCAR, see Section 3; for the importance of these topics, see real-data examples in Subsection 4.2.

Theoretical studies of zero-inflated or quasi zero-inflated Poisson models in high dimensions are relatively scarce in spite of their importance. [8] presents elegant local-global shrinkage priors for high-dimensional quasi zero-inflated Poisson models, providing theoretical properties of the shrinkage factors and of the multiple testing statistics. In general, a prior distribution should

be constructed according to the context in which the prior is used. We confirm that our priors outperform their priors in predictive density estimation in Section 4; while we consider that their priors would be more suitable than our priors in multiple testing.

1.3. Notations. Additional notations are summarized here. The notation $a_n \sim b_n$ signifies that a_n/b_n converges to 1 as n goes to infinity. The notation $a_n \asymp b_n$ signifies that a_n/b_n converges to a constant as n goes to infinity. The notation $O(a_n)$ indicates a term of which the absolute value divided by a_n is bounded for a large n .

For a function $f : \mathbb{N}^n \times \mathbb{N}^n \rightarrow \mathbb{R}$, the expectation $\mathbb{E}_\theta[f(X, Y)]$ indicates the expectation of $f(X, Y)$ with respect to $p(x | \theta)q(y | \theta)$. Likewise, for a function $g : \mathbb{N} \rightarrow \mathbb{R}$, the expectation $\mathbb{E}_\lambda[g(X_1)]$ indicates the expectation of $g(X_1)$ with respect to $\text{Po}(\lambda)$.

1.4. Organization. The rest of this paper is organized as follows. Section 2 presents theoretical results in predictive density estimation in sparse Poisson models without MCAR. Theorems 2.1 and 2.2 are the main theorems that provide exact asymptotic minimax risks and a class of Bayes predictive densities attaining the exact asymptotic minimaxity. Proposition 2.1 provides a Discussion on comparison between prediction and estimation. Propositions 2.2 and 2.4 provide adaptive predictive densities. Section 3 provides the description of problem settings and theoretical results in predictive density estimation in sparse Poisson models with MCAR. Section 4 reports simulation studies and applications to real data. All proofs of theorems and propositions for Section 2 are provided in Section 5. All proofs of theorems and propositions for Section 3 are provided in Section 6.

2. PREDICTIVE DENSITY ESTIMATION IN SPARSE POISSON MODELS

2.1. Main results. Results for sparse Poisson models without MCAR are provided in order; the precise description of the exact asymptotic minimax risk, and the construction of Bayes predictive densities that attain exact asymptotic minimaxity. Discussions on comparison between prediction and estimation and on adaptation are provided in the sequel subsection.

The first theorem presents a precise description of the exact asymptotic minimax risk. For $r \in (0, \infty)$, let

$$\mathcal{C} := \mathcal{C}_r = \left(\frac{r}{r+1} \right)^r \left(\frac{1}{r+1} \right).$$

Theorem 2.1. *Fix $r \in (0, \infty)$ and fix a sequence $s_n \in (0, n)$ such that $s_n = o(n)$.*

(a) *For the exact sparse parameter space $\Theta[s_n]$, the asymptotic equality*

$$\mathcal{R}(\Theta[s_n]) \sim \mathcal{C} s_n \log(n/s_n)$$

holds as $n \rightarrow \infty$.

(b) *For the quasi sparse parameter space $\Theta[s_n, \varepsilon_n]$ with any shrinking sequence $\varepsilon_n > 0$ such that $\varepsilon_n = o(s_n/n)$, the asymptotic equality*

$$\mathcal{R}(\Theta[s_n, \varepsilon_n]) \sim \mathcal{C} s_n \log(n/s_n)$$

holds as $n \rightarrow \infty$.

The implication of this theorem in comparison with the estimative minimax risk is discussed in Subsection 2.2. The proof of this theorem is provided in Subsections 5.2-5.3.

The second theorem provides a class of Bayes predictive densities that attain the exact asymptotic minimaxity with the knowledge of sparsity s_n . For cases where sparsity s_n is unknown, adaptation to s_n is discussed in Subsection 2.2.2. For $t > 0$ and $\kappa > 0$, let $\Pi[t, \kappa]$ be an improper prior of the form

$$\Pi[t, \kappa](d\theta) = \prod_{i=1}^n \left\{ \delta_0(d\theta_i) + \frac{t}{n} \theta^{\kappa-1} 1_{(0, \infty)}(d\theta_i) \right\},$$

where δ_0 is the Dirac measure centered at 0. The corresponding Bayes predictive density is well-defined and given by

$$q_{\Pi[t, \kappa]}(y \mid x) = \prod_{i=1}^n \left\{ \omega_i \delta_0(y_i) + (1 - \omega_i) \binom{x_i + y_i + \kappa - 1}{y_i} \left(\frac{r}{r+1} \right)^{x_i + \kappa} \left(1 - \frac{r}{r+1} \right)^{y_i} \right\},$$

where

$$\omega_i := \begin{cases} 1 / \{1 + (t/n) \Gamma(\kappa) / r^\kappa\} & \text{if } x_i = 0, \\ 0 & \text{if } x_i \geq 1. \end{cases}$$

Here, the coordinate-wise marginal distribution of $q_{\Pi[t, \kappa]}$ is just a zero-inflated negative binomial distribution, and sampling from $q_{\Pi[t, \kappa]}$ is easy.

Theorem 2.2. *Fix $r \in (0, \infty)$ and $\kappa > 0$. Fix also a sequence $s_n \in (0, n)$ such that $s_n = o(n)$. The predictive density $q_{\Pi[s_n, \kappa]}$ is asymptotically minimax: two asymptotic equalities*

$$\begin{aligned} \sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[s_n, \kappa]}) &\sim \mathcal{R}(\Theta[s_n]) \\ \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[s_n, \kappa]}) &\sim \mathcal{R}(\Theta[s_n, \varepsilon_n]) \end{aligned}$$

hold as $n \rightarrow \infty$, where $\varepsilon_n > 0$ is a shrinking sequence such that $\varepsilon_n = o(s_n/n)$.

2.2. Discussions. Several discussions are provided in order.

2.2.1. Prediction and estimation. Minimax risks for estimation and prediction are compared. The minimax risk $\mathcal{E}(\Theta[s_n])$ restricted to plug-in predictive densities is provided in the following proposition.

Proposition 2.1. *Fix $r \in (0, \infty)$. The asymptotic equalities*

$$\mathcal{E}(\Theta[s_n]) \sim \mathcal{E}(\Theta[s_n, \varepsilon_n]) \sim e^{-1} r^{-1} s_n \log(n/s_n)$$

hold as $n \rightarrow \infty$, where $s_n \in (0, n)$ is any sequence such that $s_n = o(n)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(s_n/n)$.

According to Theorem 2.1 and Proposition 2.1, the exact asymptotic constants of predictive and estimative minimax risks ($\mathcal{R}(\Theta[s_n])$ and $\mathcal{E}(\Theta[s_n])$) are different with r , and the rates of convergence of predictive and estimative minimax risks are identical with n . The constant r is the key parameter in describing the exact constants of the risks: the exact constant of predictive minimax risk increases

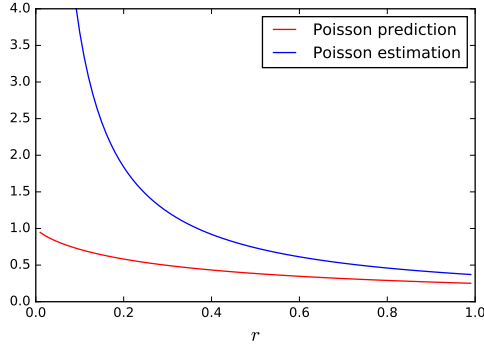


FIGURE 1. Predictive and estimative minimax risks for sparse Poisson models: the horizontal axis represents r .

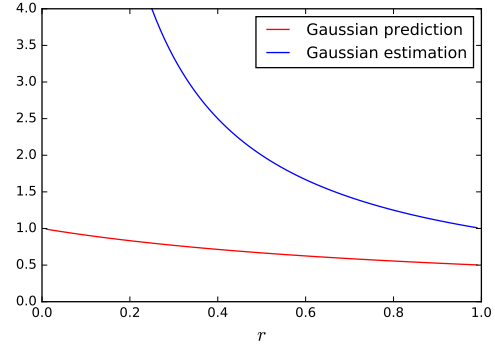


FIGURE 2. Predictive and estimative minimax risks for sparse Gaussian models: the horizontal axis represents r .

as r decreases but is bounded by 1, while that of estimative minimax risk grows to infinity as r decreases. The similar phenomenon occurs in sparse Gaussian models as found in [41]. However, the exact constants of predictive minimax risk are quite different in Poisson and Gaussian models, and they are not derivable intuitively.

Figures 1 and 2 show comparisons of the exact constants of minimax risks for sparse Poisson and Gaussian models. A vertical line indicates values of the risks and a horizontal line indicates values of r . They show the similarity of the behavior with respect to r of minimax risks in Poisson and Gaussian cases. An interesting finding in comparison of Poisson and Gaussian cases is that the exact constants of predictive minimax risks in both cases get closer to 1 as r approaches to 0.

2.2.2. Adaptation. Two types of adaptation to an unknown sparsity s are discussed. First, a class of adaptive Bayes predictive densities is presented for cases where the sparsity level is high: $1 \leq \inf_n s_n \leq \sup_n s_n < \infty$. Second, the performance of an empirical Bayes predictive density with an estimator \hat{s}_n plugged-in is discussed.

The first proposition provides a class of Bayes predictive densities that attain the exact asymptotic minimaxity without the knowledge of sparsity, provided that the sparsity level is high: $1 \leq \inf_n s_n \leq \sup_n s_n < \infty$.

Proposition 2.2. Fix $r \in (0, \infty)$ and $\kappa > 0$. Fix a sequence $t_n \in (0, n)$ such that $1 \leq \inf_n t_n \leq \sup_n t_n < \infty$. The predictive density $q_{\Pi[t_n, \kappa]}$ is adaptive in the exact minimax sense on the classes $\{\Theta[s_n] : 1 \leq \inf_n s_n \leq \sup_n s_n < \infty\}$ and $\{\Theta[s_n, \varepsilon_n] : 1 \leq \inf_n s_n \leq \sup_n s_n < \infty, \varepsilon_n = o(1/n), \varepsilon_n > 0\}$: two asymptotic equalities

$$\begin{aligned} \sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[t_n, \kappa]}) &\sim \mathcal{R}(\Theta[s_n]) \\ \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[t_n, \kappa]}) &\sim \mathcal{R}(\Theta[s_n, \varepsilon_n]) \end{aligned}$$

hold as $n \rightarrow \infty$, where $s_n \in (0, n)$ is any sequence such that $1 \leq \inf_n s_n \leq \sup_n s_n < \infty$, and $\varepsilon_n > 0$ is any sequence such that $\varepsilon_n = o(1/n)$.

Remark that a sequence t_n in Proposition 2.2 is chosen independently from the true sparsity s_n ; one can choose the all-one sequence as t_n .

Next, an empirical Bayes predictive density with an estimator of the sparsity plugged-in is discussed. In application, there exist situations in which a convenient estimator \hat{s}_n for s_n is available. Then, plugging \hat{s}_n into t of $\Pi[t, \kappa]$ yields the empirical Bayes predictive density $q_{\Pi[\hat{s}_n, \kappa]}$. We provide a point-wise performance guarantee for $q_{\Pi[\hat{s}_n, \kappa]}$ in Proposition 2.3. Using Proposition 2.3, we construct an asymptotically adaptive empirical Bayes predictive density for cases with s_n such that $s_n = o(n^{1/2})$ in Proposition 2.4. For simplicity, we make the following condition on \hat{s}_n :

Condition 2.1. *The following are satisfied:*

- (i) $\hat{s}_n \geq \gamma$ a.s. for some $\gamma > 0$;
- (ii) There exists $\delta > 0$ for which we have $\max\{\mathbb{E}_\theta|\hat{s}_n/s_n - 1|, \mathbb{E}_\theta|\hat{s}_n/s_n - 1|^2\} < \delta$.

An instance of an estimator \hat{s}_n satisfying Condition 2.1 is discussed after Proposition 2.3.

The following proposition shows that under Condition 2.1, the empirical Bayes predictive density $q_{\Pi[\hat{s}_n, \kappa]}$ works as the Bayes predictive density $q_{\Pi[s_n, \kappa]}$ does.

Proposition 2.3. *Fix $r \in (0, \infty)$ and $\kappa > 0$. Suppose that Condition 2.1 holds. There exists a positive constant c depending only on r , κ , and γ such that*

- (a) *for $\theta \in \Theta[s_n]$ with any sequence $s_n \in (0, n)$, we have*

$$R(\theta, q_{\Pi[\hat{s}_n, \kappa]}) \leq R(\theta, q_{\Pi[s_n, \kappa]}) + c(s_n\delta + s_n^2/n + s_n \log s_n);$$

- (b) *for $\theta \in \Theta[s_n, \varepsilon_n]$ with any sequence $s_n \in (0, n)$ and any sequence $\varepsilon_n > 0$, we have*

$$R(\theta, q_{\Pi[\hat{s}_n, \kappa]}) \leq R(\theta, q_{\Pi[s_n, \kappa]}) + c(s_n\delta + s_n^2/n + \xi_n),$$

where $\xi_n := n\{1 - \exp(-\varepsilon_n)\}\{1 + \delta + \log s_n\}$.

For a simple (crude) estimator \hat{s}_n , the following proposition provides an upper bound on δ and the maximum risk of $q_{\Pi[\hat{s}_n, \kappa]}$ over $\Theta[s_n]$ and $\Theta[s_n, \varepsilon_n]$. Consider that \hat{s}_n is given by $\max\{1, \#\{i : X_i \geq 1, i = 1, \dots, n\}\}$. Obviously, \hat{s}_n satisfies Condition 2.1 (i).

Proposition 2.4. *Fix $r \in (0, \infty)$. Let $s_n \in (0, n)$ be a sequence such that $\inf_n s_n > 0$ and $s_n = o(n^{1/2})$.*

- (a) *For the exact sparse parameter space $\Theta[s_n]$, $\max\{\mathbb{E}|\hat{s}_n/s_n - 1|, \mathbb{E}|\hat{s}_n/s_n - 1|^2\}$ is bounded above by a positive constant depending only on $\inf_n s_n$. Furthermore, for any $\kappa > 0$, the following asymptotic equality holds:*

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[\hat{s}_n, \kappa]}) \sim \mathcal{R}(\Theta[s_n]) \text{ as } n \rightarrow \infty.$$

- (b) *For the quasi sparse parameter space $\Theta[s_n, \varepsilon_n]$ with any shrinking sequence $\varepsilon_n > 0$ such that $\varepsilon_n = o(s_n/n)$, $\max\{\mathbb{E}|\hat{s}_n/s_n - 1|, \mathbb{E}|\hat{s}_n/s_n - 1|^2\}$ is bounded above by a positive constant depending only on $\inf_n s_n$. Furthermore, for any $\kappa > 0$, the following asymptotic equality holds:*

$$\sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[\hat{s}_n, \kappa]}) \sim \mathcal{R}(\Theta[s_n, \varepsilon_n]) \text{ as } n \rightarrow \infty.$$

There are comments on Propositions 2.3-2.4. The $O(s_n/n)$ and $O(s_n \log s_n)$ -terms in Proposition 2.3 can be improved provided that $\delta < 1$. However, the assumption that $\delta < 1$ is considered to be too stringent for θ near 0, since for $i = 1, \dots, n$, it is intrinsically difficult to test whether $\theta_i > 0$ or $\theta_i = 0$ for small θ_i . Paying the cost of adding the $O(s_n/n)$ and $O(s_n \log s_n)$ - terms, we obtain Proposition 2.3 for any $\delta > 0$. Instead, Proposition 2.4 requires an assumption on s_n .

2.2.3. Tail conditions for spike-and-slab priors. An idea behind the construction of $\Pi[t, \kappa]$ is discussed. We employ “spike-and-slab” priors, getting a hint from the satisfactory result of using a spike-and-slab prior to construct an asymptotically minimax Bayes predictive density for sparse Gaussian models by [40]. Spike-and-slab priors are widely used in the literature of Bayesian sparse modelings. However, the choice of a slab prior is substantially important since the tail behavior of the slab prior strongly affects the performance. We provide two propositions indicating the effect of the tail behavior on the performance.

The first proposition shows an example in which spike-and-slab priors with slab priors having exponential tails suffer from the sub-optimality. Let $s_n \in (0, n)$ be a sequence such that $s_n = o(n)$. For $\mu \geq 0$, $\nu = (\nu_1, \dots, \nu_n) \in \mathbb{R}_+^n$, and $k = (k_1, \dots, k_n) \in \mathbb{R}^n$, let

$$\Pi_{\text{SS}} := \prod_{i=1}^n \left\{ \left(1 - \frac{s_n}{n}\right) \delta_\mu(d\theta_i) + \frac{s_n}{n} \text{Ga}(d\theta_i; \nu_i, k_i) \right\},$$

where $\text{Ga}(\cdot; \nu_1, k_1)$ is Gamma distribution with inverse scale parameter ν_1 and shape parameter k_1 . The sub-optimality is discussed within the exact sparse parameter space.

Proposition 2.5. *Fix $\nu \in \mathbb{R}_+^n$ and fix a shrinking sequence $\mu = \mu_n = O(s_n/n)$. Then, the asymptotic inequality*

$$\inf_{k \in \mathbb{R}_+^n} \sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi_{\text{SS}}}) \Big/ \mathcal{R}(\Theta[s_n]) \rightarrow \infty$$

holds as $n \rightarrow \infty$.

Note that the sup-optimality of spike-and-slab priors with bounded uniform priors in sparse Gaussian models is reported in [40].

It is of theoretical interest to reveal a condition on the tail of a slab prior under which the Bayes predictive density based on a spike-and-slab prior is asymptotically minimax. The second proposition shows that a polynomial decay of the tail of a slab prior is allowed and thus there exists a proper prior distribution of which the posterior mixture attains exact asymptotic minimaxity. Let $s_n \in (0, n)$ be a sequence such that $s_n = o(n)$. Let

$$\Pi_{\text{P}} := \prod_{i=1}^n \left\{ \left(1 - \frac{s_n}{n}\right) \delta_0(d\theta_i) + \frac{s_n}{n} \theta_i^{-2} 1_{(1, \infty)}(d\theta_i) \right\}.$$

Proposition 2.6. *Fix $r \in (0, \infty)$. Then, the following asymptotic equality holds:*

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi_{\text{P}}}) \sim \mathcal{R}(\Theta[s_n]) \text{ as } n \rightarrow \infty.$$

3. PREDICTIVE DENSITY ESTIMATION IN SPARSE POISSON MODELS WITH MCAR

In this section, a series of results discussed in Section 2 is generalized into settings with MCAR.

Prediction in sparse Poisson models with MCAR is formulated as follows. Let r_i 's ($i = 1, 2, \dots$) be positive random variables. Given r_i ($i = 1, \dots, n$), let X_i ($i = 1, 2, \dots, n$) be a current observation independently distributed according to $\text{Po}(r_i\theta_i)$, and let Y_i ($i = 1, 2, \dots, n$) be a future observation independently distributed according to $\text{Po}(\theta_i)$, where θ_i ($i = 1, \dots, n$) is an unknown parameter. Suppose that $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ are independent. The true densities of X and Y with parameter θ given r_i s are denoted by $p(x | \theta)$ and $q(y | \theta)$, that is

$$p(x | \theta) = \prod_{i=1}^n \left\{ \frac{1}{x_i!} e^{-r_i\theta_i} (r_i\theta_i)^{x_i} \right\} \quad \text{and} \quad q(y | \theta) = \prod_{i=1}^n \left\{ \frac{1}{y_i!} e^{-\theta_i} \theta_i^{y_i} \right\}.$$

Recall that the parameter space is restricted to $\Theta[s_n] = \{\theta \in \mathbb{R}_+^n : \|\theta\|_0 \leq s_n\}$ and $\Theta[s_n, \varepsilon_n] = \{\theta \in \mathbb{R}_+^n : N(\theta, \varepsilon_n) \leq s_n\}$, where $N(\theta, \varepsilon) := \#\{i : \theta_i > \varepsilon\}$ for $\theta \in \mathbb{R}^n$ and for $\varepsilon > 0$. The minimax Kullback-Leibler risks over $\Theta[s_n]$ and $\Theta[s_n, \varepsilon_n]$ conditioned on r_i s are denoted by $\overline{\mathcal{R}}(\Theta[s_n])$ and $\overline{\mathcal{R}}(\Theta[s_n, \varepsilon_n])$, respectively.

For the sequel theoretical analyses, we assume that r_i s are independent and identically distributed according to a sampling distribution G , and make the following condition on G . Let \mathbb{E}_G be the expectation with respect to G .

Condition 3.1. *A distribution G satisfies the following: (i) $\mathbb{E}_G[r_1] < \infty$; (ii) $\mathbb{E}_G[1/r_1^2] < \infty$.*

Condition 3.1 (i) is usual. Condition 3.1 (ii) excludes any distribution G that is highly concentrated around 0. Condition 3.1 (ii) is not stringent; consider a longitudinal situation in which X_i ($i = 1, \dots, n$) is obtained as the sum of $\{X_{i,j} : j = 1, \dots, r_i\}$, where r_i ($i = 1, \dots, n$) signifies the sample size in the i -th coordinate, and for each i , $X_{i,j}$ ($j = 1, \dots, r_i$) follows $\text{Po}(\theta_i)$. Condition 3.1 implies that for each coordinate there exists at least one observation: $r_i \geq 1$. For verifying Condition 3.1 (ii) in application, see also Section 4.

3.1. Main results. Results for sparse Poisson models with MCAR are presented in order. Notations for theorems are introduced ahead. Fix an infinite sequence $\{r_i \in (0, \infty) : i \in \mathbb{N}\}$ such that $0 < \inf_i r_i \leq \sup_i r_i < \infty$. For any $i \in \mathbb{N}$, let

$$\mathcal{C}_i := \mathcal{C}_{r_i} = \left(\frac{r_i}{r_i + 1} \right)^{r_i} \left(\frac{1}{r_i + 1} \right).$$

Let $\overline{\mathcal{C}} := \overline{\mathcal{C}}_n = \sum_{i=1}^n \mathcal{C}_i / n$.

The following theorems provide theoretical results for settings with MCAR. The first theorem describes the exact minimax risk. The second theorem states that $q_{\Pi[s_n, \kappa]}$ with $0 < \kappa \leq 1$ attains exact asymptotic minimaxity regardless of a sampling distribution G of r_i s whenever G satisfies Condition 3.1.

Theorem 3.1. *Under Condition 3.1, two asymptotic equalities*

$$\begin{aligned} \lim_{n \rightarrow \infty} \overline{\mathcal{R}}(\Theta[s_n]) / \{\mathbb{E}_G[\mathcal{C}_1] s_n \log(n/s_n)\} &= 1 \\ \lim_{n \rightarrow \infty} \overline{\mathcal{R}}(\Theta[s_n, \varepsilon_n]) / \{\mathbb{E}_G[\mathcal{C}_1] s_n \log(n/s_n)\} &= 1 \end{aligned}$$

hold, where $s_n \in (0, n)$ is a sequence such that $s_n = o(n)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(s_n/n)$.

Theorem 3.2. Fix $\kappa \in (0, 1]$. Under Condition 3.1, the predictive density $q_{\Pi[s_n, \kappa]}$ attains exact asymptotic minimaxity regardless of a sampling distribution G of r_i s: two asymptotic equalities

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} \sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[s_n, \kappa]}) / \{\mathbb{E}_G[\mathcal{C}_1] s_n \log(n/s_n)\} &= 1 \\ \text{plim}_{n \rightarrow \infty} \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[s_n, \kappa]}) / \{\mathbb{E}_G[\mathcal{C}_1] s_n \log(n/s_n)\} &= 1 \end{aligned}$$

hold, where $s_n \in (0, n)$ is a sequence such that $s_n = o(n)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(s_n/n)$.

3.2. Discussions. We compare Theorem 3.1 with Theorem 2.1. We consider two settings:

- (A) G is the Gamma distribution with shape parameter r/l and scale parameter l for $0 < l \leq 1$ and $r \geq 2$;
- (B) G is the distribution of $1 + S$, where S follows the binomial distribution with trial number N and success probability p .

In Setting (A), the mean remains r for any l and the variance is rl , which means that as $l \rightarrow 0$, G is weakly convergent to the Dirac measure δ_r centered at r corresponding to a non-MCAR case. In Setting (B), G is weakly convergent to the Dirac measure δ_1 centered at 1 as $p \rightarrow 0$, and G is weakly convergent to the Dirac measure δ_{1+N} as $p \rightarrow 1$.

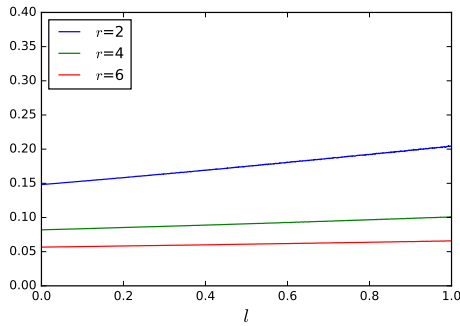


FIGURE 3. Comparison of $\mathbb{E}_G[\mathcal{C}_1]$ in Setting (A): the horizontal line indicates l of G .

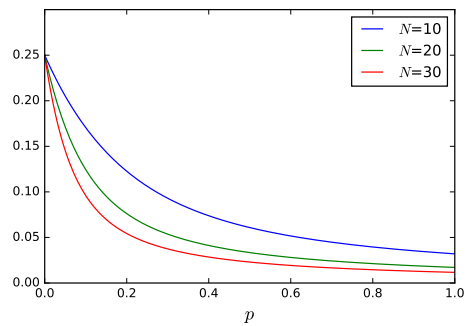


FIGURE 4. Comparison of $\mathbb{E}_G[\mathcal{C}_1]$ in Setting (B): the horizontal line indicates p of G .

In Figure 3, a vertical line indicates exact constants and a horizontal line indicates values of l . The blue line denotes the case with $r = 2$, the green line denotes the case with $r = 4$, and the red line denotes the case with $r = 6$. In Figure 4, a vertical line indicates exact constants and a horizontal line indicates values of p . The blue line denotes the case with $N = 10$, the green line denotes the case with $N = 20$, and the red line denotes the case with $N = 30$.

Figures 3 and 4 show exact constants of predictive minimax risks in Settings (A) and (B), respectively. According to Figure 3, the constant gets larger as the variance of G increases. The constant

in a MCAR case approaches to that in a non-MCAR case in the limit $l \rightarrow 0$. According to Figure 4, the constant gets smaller as the missing probability $1 - p$ gets smaller. Further, the numerical result in Setting (B) is consistent to the results in [13, 14] for the literature of nonparametric regression in the presence of missing observations. Theorems 1 and 2 in [13] provide tight lower and upper bounds on mean integrated squared errors (MISE) in nonparametric regression with predictors missing at random. Those theorems also provide an exact asymptotically minimax estimator for MCAR cases and show that the minimax MISE gets smaller as the missing probability approaches zero if predictors are MCAR.

Next, we provide a proposition useful to comparison between estimation and prediction in cases with MCAR.

Proposition 3.1. *Under Condition 3.1, two asymptotic equalities*

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} \mathcal{E}(\Theta[s_n]) / \{\mathbb{E}_G[r_1^{-1}] e^{-1} s_n \log(n/s_n)\} &= 1 \\ \text{plim}_{n \rightarrow \infty} \mathcal{E}(\Theta[s_n, \varepsilon_n]) / \{\mathbb{E}_G[r_1^{-1}] e^{-1} s_n \log(n/s_n)\} &= 1 \end{aligned}$$

hold, where $s_n \in (0, n)$ is a sequence such that $s_n = o(n)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(s_n/n)$.

4. SIMULATION STUDIES AND APPLICATIONS TO REAL DATA

4.1. Simulation studies. In this section, we present simulation studies to compare the performance of various predictive densities.

4.1.1. Sparse Poisson model without MCAR. First, consider a sparse Poisson model without MCAR. Parameter θ and observations X and Y are drawn from

$$\begin{aligned} \theta_i &\sim \nu_i e_{S,i} \quad (i = 1, \dots, n), \\ X \mid \theta &\sim \otimes_{i=1}^n \text{Po}(r\theta_i), \quad Y \mid \theta \sim \otimes_{i=1}^n \text{Po}(\theta_i), \quad \text{and } X \perp\!\!\!\perp Y \mid \theta, \end{aligned}$$

respectively, where

- ν_1, \dots, ν_n are independent samples from the gamma distribution with a shape parameter 10 and a scale parameter 1;
- S is drawn from the uniform distribution on all subsets having exactly s elements;
- ν_1, \dots, ν_n and S are independent;

Here for a subset $J \subset \{1, \dots, n\}$, e_J indicates the vector of which the i -th component is 1 if $i \in J$ and 0 otherwise. We examine two cases $(n, s, r) = (200, 5, 1)$ and $(200, 5, 20)$, and generate 500 current observations (X 's) and 500 future observations (Y 's). See Appendix A for the results with different choices of (n, s, r) .

We compare the following five classes of predictive densities:

- the proposed empirical Bayes predictive densities based on $\Pi[\hat{s}_n, \kappa]$.
- the proposed Bayes predictive densities based on $\Pi[1, \kappa]$;
- the Bayes predictive density based on the shrinkage prior proposed in [27];
- the Bayes predictive density based on the Gauss Hypergeometric prior proposed in [8];

- the plug-in predictive density based on an ℓ_1 -penalized estimator.

For the empirical Bayes predictive density with $\Pi[\hat{s}_n, \kappa]$, an estimate of sparsity \hat{s}_n to be plugged in is determined by the k-means clustering with $k = 2$. The proposed predictive densities $q_{\Pi[\hat{s}_n, \kappa]}$ and $q_{\Pi[1, \kappa]}$ are shown to be adaptive in Propositions 2.2 and 2.4. The third predictive density is shown in [27] to dominate the Bayes predictive density based on the Jeffreys prior. In simulation studies, a hyperparameter β of the third predictive density is fixed to 1.

The performance of predictive densities is evaluated by the following three measures:

- the mean of the ℓ_1 distance ($\sum_{i=1}^n |u_i - v_i|$ for $u, v \in \mathbb{R}^n$) between the mean of a predictive density and a future observation;
- the predictive log likelihood, that is, the log of the value of a predictive density at sampled Y and X ;
- the (empirical) coverage probabilities of Y on the basis of the joint 70%, 90%-prediction sets constructed by a predictive density.

Tables 1 and 2 show the results of comparison. The following abbreviations are used. The Bayes predictive density proposed in [8] is abbreviated to GH. The Bayes predictive density proposed in [27] is abbreviated to K04. The plug-in density based on an ℓ_1 -penalized estimator with regularization parameter $r\lambda$ is abbreviated to ℓ_1 (λ). The abbreviation ℓ_1 distance signifies a mean ℓ_1 distance. The abbreviation PLL signifies a predictive log likelihood. The abbreviation $\alpha\%$ CP signifies the empirical coverage probability based on an $\alpha\%$ -prediction set.

TABLE 1. Comparison of predictive densities without MCAR and with $(n, s, r) = (200, 5, 1)$: ℓ_1 distance, PLL and $\alpha\%$ CP signify the mean ℓ_1 distance, the predictive log likelihood, and the empirical coverage probability based on an $\alpha\%$ -prediction set, respectively. For each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance.

	$\Pi[\hat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K04	ℓ_1 ($\lambda = 0.1$)
ℓ_1 distance	21.6(6.5)	<u>19.3</u> (6.4)	21.3(7.1)	104(4.9)	96.5(8.1)	22.1(7.8)
PLL	-17.2(2.0)	<u>-15.3</u> (1.8)	-15.7(1.6)	-66.3(3.28)	-86.2(8.8)	-Inf
70%CP (%)	81.6(17.3)	<u>71.8</u> (18.7)	72.2(18.1)	77.0(4.86)	22.9(19.7)	26.7(16.8)
90%CP (%)	94.5(9.3)	<u>90.6</u> (11.1)	91.4(10.2)	92.0(1.50)	40.5(24.4)	49.4(21.6)

The results are summarized as follows. In regard to ℓ_1 distances, samples from Bayes predictive densities based on $\Pi[\hat{s}_n, \kappa]$ and $\Pi[1, \kappa]$ are closer to future observations than those of three other classes of predictive densities. In regard to empirical coverage probabilities, the Bayes predictive densities based on $\Pi[\hat{s}_n, \kappa]$ and $\Pi[1, \kappa]$ give the empirical coverage probabilities of Y that are relatively close to the nominal level. The prediction sets of the Bayes predictive densities K04 and GH are too broad to have the correct level for prediction sets. The prediction set of the plug-in predictive density based on the ℓ_1 -penalized estimator is too narrow to cover future observations. This is mainly because for this plug-in predictive density, most of the marginal predictive intervals degenerate into zero, since an ℓ_1 -penalized estimator returns zero for a coordinate at which the

TABLE 2. Comparison of predictive densities without MCAR and with $(n, s, r) = (200, 5, 20)$: ℓ_1 distance, PLL and $\alpha\%$ CP signify the mean ℓ_1 distance, the predictive log likelihood, and the empirical coverage probability based on an $\alpha\%$ -prediction set, respectively. For each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance.

	$\Pi[\hat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K04	ℓ_1 ($\lambda = 0.1$)
ℓ_1 distance	12.6(4.2)	<u>12.9</u> (4.2)	13.0(4.5)	15.7(1.67)	22.5(5.2)	14.1(4.5)
PLL	<u>-12.8</u> (1.6)	-12.9(0.7)	-12.9(1.5)	-15.6(1.5)	-21.6(2.2)	-Inf
70%CP (%)	<u>70.4</u> (4.6)	<u>69.6</u> (4.5)	69.5(4.4)	89.4(1.6)	91.1(3.3)	63.0(5.4)
90%CP (%)	<u>89.8</u> (3.0)	89.1(3.2)	89.5(2.9)	97.6(0.71)	97.5(1.4)	86.3(3.9)

current observation is zero. This degeneracy also induces diverging a predictive log likelihood value of the plug-in predictive density based on an ℓ_1 -penalized estimator.

4.1.2. *Sparse Poisson model with MCAR.* Next, consider a sparse Poisson model with MCAR. Parameter θ and observations X and Y are drawn in the following way:

$$\theta_i \sim \nu_i e_{S,i} \ (i = 1, \dots, n),$$

$$X \mid \theta \sim \otimes_{i=1}^n \text{Po}(r_i \theta_i), \ Y \mid \theta \sim \otimes_{i=1}^n \text{Po}(\theta_i), \text{ and } X \perp\!\!\!\perp Y \mid \theta,$$

where ν_1, \dots, ν_n and S follow the same distributions as those in the previous simulation study. In addition, $r_i - 1$ ($i = 1, \dots, n$) are independently drawn from the binomial distribution $\text{Bi}(m, p)$ of which the parameters (m, p) are either $(1, 0.9)$ or $(10, 0.9)$. We set $(n, s) = (200, 5)$, and generate 500 current observations (X 's) and 500 future observations (Y 's).

We compare the following five classes of predictive densities:

- the proposed empirical Bayes predictive densities based on $\Pi[\hat{s}_n, \kappa]$.
- the proposed Bayes predictive densities based on $\Pi[1, \kappa]$;
- the Bayes predictive density based on the shrinkage prior proposed in [31];
- the Bayes predictive density based on the Gauss Hypergeometric prior proposed in [8];
- the plug-in predictive density based on an ℓ_1 -penalized estimator.

An estimator \hat{s}_n is determined in the same manner as in the previous subsection. In [31], the third predictive density is shown to dominate the Bayes predictive density based on the Jeffreys prior in the case where the numbers of observations are coordinate-wise different. In simulation studies, each hyper parameter β_i of the third predictive density is fixed to 1.

In comparing the performance, we use the weighted ℓ_1 distance between the mean of the predictive density and a future observation. Here the the weighted ℓ_1 distance between $u, v \in \mathbb{R}^n$ is given by $\sum_{i=1}^n r_i |u_i - v_i| / (\sum_{i=1}^n r_i / n)$. For the construction of prediction sets, we also use this weighted ℓ_1 distance.

Tables 3 and 4 show the results. In addition to the abbreviations used in the previous subsection, the abbreviation W- ℓ_1 distance signifies a mean weighted ℓ_1 distance. We see that the weighted ℓ_1 distances by the Bayes predictive densities based on $\Pi[\hat{s}_n, \kappa]$ and $\Pi[1, \kappa]$ are in the smallest level of

TABLE 3. Comparison of predictive densities with MCAR and with $(n, s, m, p) = (200, 5, 1, 0.9)$: $W\text{-}\ell_1$ distance, PLL and $\alpha\%$ CP signify the weighted mean ℓ_1 distance, the predictive log likelihood, and the empirical coverage probability based on an $\alpha\%$ -prediction set, respectively. For each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance.

	$\Pi[\widehat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K15	$\ell_1 (\lambda = 0.1)$
$W\text{-}\ell_1$ distance	17.5(5.4)	<u>15.1</u> (5.0)	16.2(5.5)	48.3(5.6)	25.6(8.7)	17.1(5.8)
PLL	-15.1(2.2)	<u>-13.9</u> (1.6)	-14.2(1.6)	-42.3(2.0)	-18.2(3.4)	-Inf
70%CP (%)	74.6(1.5)	<u>71.2</u> (13.2)	<u>71.2</u> (11.8)	99(0.0)	34.4(18.1)	42(13.8)
90%CP (%)	91.9(7.8)	<u>90.3</u> (8.0)	90.8(6.8)	100(0.0)	61.7(20.9)	68.6(13.6)

TABLE 4. Comparison of predictive densities with MCAR and with $(n, s, m, p) = (200, 5, 10, 0.9)$: $W\text{-}\ell_1$ distance, PLL and $\alpha\%$ CP signify the weighted mean ℓ_1 distance, the predictive log likelihood, and the empirical coverage probability based on an $\alpha\%$ -prediction set, respectively. For each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance.

	$\Pi[\widehat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K15	$\ell_1 (\lambda = 0.1)$
$W\text{-}\ell_1$ distance	13.4(4.8)	<u>10.7</u> (3.9)	12.3(4.4)	17.3(1.8)	15.2(4.2)	13.0(4.4)
PLL	-13.1(1.8)	<u>-11.9</u> (1.7)	-12.7(1.7)	-19.3(1.7)	-14.0(1.9)	-Inf
70%CP (%)	<u>69.9</u> (6.2)	69.4(5.1)	66.9(5.4)	14.1(3.2)	52.3(9.9)	61.0(6.4)
90%CP (%)	<u>90.0</u> (3.4)	89.7(3.3)	87.8(3.6)	51.7(8.6)	78.3(7.3)	84.8(4.6)

all predictive densities compared here. As is the case without MCAR, the proposed Bayes predictive densities have coverage probabilities that are relatively close to the nominal level, whereas the other three predictive densities do not.

In conclusion, these simulation studies show that in sparse Poisson models, the proposed Bayes predictive densities are highly effective not only in point prediction but also in uncertainty quantification for future observations regardless of whether there exists a missing of observations or not.

4.2. Applications to Real Data. In this section, we apply our methods to two real datasets; Japanese crime data and exome sequencing data.

4.2.1. Pickpockets in Tokyo Prefecture. We discuss the validity of the proposed predictive densities empirically using Japanese crime data. A motivation of this analysis comes from the importance of taking measures against future crimes by utilizing the past crime data.

We apply our methods to crime data from an official database called *the number of crimes in Tokyo by type and town* [11]. This database reports the total numbers of crimes in Tokyo Prefecture. They are classified by town and also by type of crimes.

We use pickpocket data from 2012 to the first half of 2018 at 978 towns in eight wards (Bunkyo Ward, Chiyoda Ward, Chuo Ward, Edogawa Ward, Koto Ward, Minato Ward, Sumida Ward and

Taito Ward). Figure 5 shows total counts of pickpockets from 2012 to 2017 at all towns in the wards. A blue circle indicates the count of pickpockets at each town, with the radius corresponding to the amount of pickpockets at each town. Cream-colored circles correspond to towns at which there have not occurred any pickpockets. Figure 5 indicates that the data have zero or near-zero counts at a vast majority of locations, but have relatively large counts in certain locations.

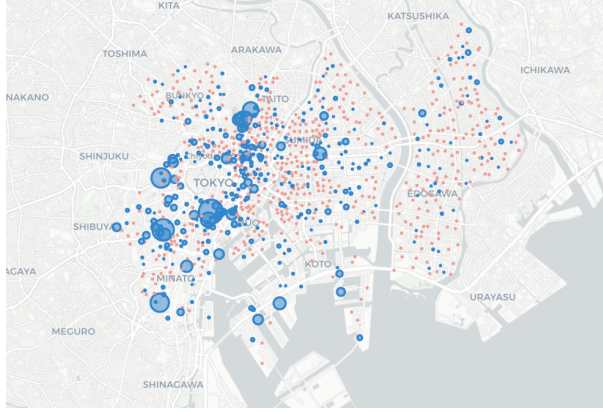


FIGURE 5. Total numbers of pickpockets from 2012 to 2017 in eight wards (Bunkyo Ward, Chiyoda Ward, Chuo Ward, Edogawa Ward, Koto Ward, Minato Ward, Sumida Ward and Taito Ward): blue circles indicate the counts of pickpockets and cream-colored circles indicate towns at which there have not occurred pickpockets.

The experimental settings are as follows. The data at the 361 towns from 2012 to 2017 are used as current observations. The data in the first half of 2018 are used as future observations. Since the counts in the first half of 2018 would be considered as the half of the total counts in 2018, the ratio r of sample sizes is, in general, set as $r = 12$. But, there occurs missing of observations because several towns, though in rare cases, did not report the counts.

As in Subsection 4.1, we compare the proposed predictive densities $q_{\Pi[\hat{s}_n, \kappa]}$ and $q_{\Pi[1, \kappa]}$ with the three existing predictive densities, that is, the Bayes predictive density GH based on a Gauss Hypergeometric prior, the Bayes predictive density K04 based on the shrinkage prior, and the plug-in predictive density based on an ℓ_1 -regularized estimator. An estimator \hat{s}_n used in $q_{\Pi[\hat{s}_n, \kappa]}$ is set as the simple estimator described before Proposition 2.4 with a slight modification: we use the mean of the numbers of values greater than 1 in each year as \hat{s}_n . We evaluate these predictive densities on the basis of the following three measures:

- the weighted ℓ_1 distance used in Subsection 4.1.2 between the mean vector of a predictive density and the obtained data of the first half in 2018;
- the predictive log likelihood at the obtained data of the first half in 2018;
- the means of coordinate-wise coverage probabilities of the obtained data of the first half in 2018 based on 70% and 90%-prediction sets.

Table 5 shows a summary of comparisons. In regard to weighted ℓ_1 distance, there are few differences; in regard to predictive log likelihood, $q_{\Pi[\hat{s}_n, \kappa]}$, $q_{\Pi[1, \kappa]}$ and the Bayes predictive density based on the Gauss Hypergeometric prior outperform others; in regard to 70% coordinate-wise

TABLE 5. Comparison of predictive densities in pickpocket data by the weighted ℓ_1 distance (W- ℓ_1 distance), the predictive log likelihood (PLL), and the mean of coordinate-wise coverage probabilities (CCP): underlines indicate the best performance.

	$\Pi[\widehat{s}_n, 0.1]$	$\Pi[1, 0.5]$	$\Pi[1, 1]$	GH	K04	ℓ_1 ($\lambda = 0.1$)
W- ℓ_1 distance	<u>273</u>	281	292	293	<u>273</u>	297
PLL	<u>-394</u>	-458	-479	-399	-429	-Inf
70%CCP (%)	91.5	91.5	91.5	7	<u>84.2</u>	90.1
90%CCP (%)	<u>93.0</u>	<u>93.0</u>	<u>93.0</u>	27	84.2	<u>93.0</u>

coverage probability, all of the methods work poorly; in regard to 90% coordinate-wise overage probability, $q_{\Pi[\widehat{s}_n, \kappa]}$ and $q_{\Pi[1, \kappa]}$ outperform others. Overall, the proposed empirical Bayes predictive density $q_{\Pi[\widehat{s}_n, \kappa]}$ is better than the other predictive densities.

4.2.2. *Rare mutation rates in an oncogene.* We consider an application of the proposed methods to exome sequencing data from a huge database called the Exome Aggregation Consortium (ExAC). ExAC reports the total numbers of mutant alleles in each genetic position along the whole exome, gathered from 60706 unrelated individuals. The analysis in this subsection is motivated by [8]: see Section 5 in [8]. We focus on rare allele mutations in a gene PIK3CA; For the importance of analysing rare allele mutations and the choice of the gene, see [8]. We also follow the pre-process of the data described in [8].

We apply the sparse Poisson model to the numbers of rare mutant alleles as follows. We denote by X_i ($i = 1 \dots, 551$) the number of rare mutant alleles in the i -th position. We assume that X_i is distributed according to $\text{Po}(r_i \theta_i)$. Here r_i ($i = 1, \dots, 551$) is the double number of individuals whose i -th location are sequenced, and θ_i is the frequency rate common to individuals in the i -th position. The doubling is necessary because each individual has two copies of each allele. r_i 's are different in general (see Appendix C), since numerous fragments of DNA sequences are sampled and read at random; hence the data is regarded as having an MCAR structure.

Our goal is to predict the behavior of rare allele mutations under the assumption that all individuals could be sequenced at all positions and the sequencing depth is uniform across the gene; the target mutation counts Y_i 's are assumed to be distributed according to $\otimes_{i=1}^n \text{Po}(\bar{r} \theta_i)$ with $\bar{r} = 121412 = 2 \times 60706$. We compare the propose prior $\Pi[\widehat{s}_n, \kappa]$ with the two existing priors, that is, the Gauss hypergeometric prior in [8] and the shrinkage prior in [27]. For the proposed prior, κ is set to 0.1 and the estimate \widehat{s}_n to be plugged in is determined by the k-means clustering with $k = 2$ (the resulting value of \widehat{s}_n is 17). In this study, we give both quantitative and qualitative comparisons of above three priors; for quantitative comparison, we use a risk unbiased estimator of the Kullback–Leibler risk; for qualitative comparison, we use samples from predictive densities.

First, we discuss the quantitative evaluation. The quantitative evaluation measure is the Poisson unbiased risk estimator (PUKLA), considered in [10]. PUKLA is defined as an unbiased risk estimator of the *estimative* Kullback–Leibler risk (up to additive terms independent on $\widehat{\theta}(X)$) for

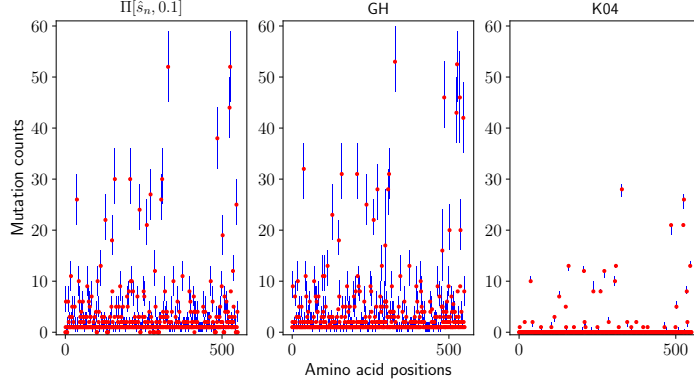


FIGURE 6. Comparison of Bayes predictive densities in terms of marginal 50%-prediction intervals. For each i -th position, the top and the bottom of the blue lines indicate 25% and 75% percentiles, respectively. The red points show the medians of marginal densities.

Poisson models. PUKLA for an estimator $\hat{\theta}$ is given as follows:

$$\hat{R}(X; q(\cdot; \hat{\theta})) = \sum_{i=1}^n \left[\hat{\theta}_i(X) - \frac{X_i \log \hat{\theta}_i(X_i - e_i)}{r_i} \right],$$

where e_i ($i = 1, \dots, n$) is defined as $(e_i)_i = 1$ and $(e_i)_j = 0$ for $j \neq i$. Although it is natural to use an unbiased estimator of the *predictive* risk in evaluating predictive densities, the predictive risk unbiased estimator of the Poisson sequence model is not known and the derivation itself is outside the scope of this study; hence we instead adopt PUKLA as the evaluation measure. The smaller value of PUKLA is preferable. Note that the estimative risk of a Bayes estimator is interpreted as the predictive risk of the corresponding Bayes predictive density in infinitesimal prediction; see Section 3 in [29]. The resulting values of PUKLA of $\Pi[\hat{s}_n, 0.1]$, GH, and K04 are 0.1812, 0.1904, and 0.1860, respectively. It is seen that the proposed prior outperforms others in the sense of PUKLA.

Next, we draw qualitative comparisons between the Bayes predictive densities based on the three priors. Figure 6 shows marginal prediction intervals at the nominal level 50%. The prediction intervals of the proposed predictive density $q_{\Pi[\hat{s}_n, 0.1]}$ and of the Bayes predictive density GH show apparently similar behaviors: they shrink at positions whose counts are low, and they remain the scales at positions whose counts are large. The prediction intervals by K04 are degenerated at the most of locations. In lower regimes of counts, there exists a dissimilarity between $q_{\Pi[\hat{s}_n, 0.1]}$ and GH. Most of the medians of the intervals constructed by GH are away from 0; while those constructed by $q_{\Pi[\hat{s}_n, 0.1]}$ sometimes reach zero. This dissimilarity is considered to reflect on that the proposed predictive density have more flexibility than GH owing to the coordinate-wise independence.

5. PROOFS FOR SECTION 2

5.1. Outline of the proof of Theorem 2.1 (a). The outline of the proof of Theorem 2.1 (a) is provided ahead. The proof is divided into two parts: providing upper and lower bounds on $\mathcal{R}(\Theta[s_n])$. Let $\eta_n := s_n/n$.

A lower bound on $\mathcal{R}(\Theta[s_n])$ builds upon Bayes risk maximization based on a *block-independent* prior. Let $\Pi_{B,\nu}(d\theta)$ with $\nu > 0$ be a *block-independent* prior built as follows: divide $\{1, 2, \dots, n\}$ into contiguous blocks B_j ($j = 1, 2, \dots, s_n$) with each length $m_n := \lfloor n/s_n \rfloor$. In each block B_j , draw $(\theta_{1+m_n(j-1)}, \dots, \theta_{m_n j})$ independently according to a single spike prior with spike strength $\nu > 0$, where a single spike prior with spike strength $\nu > 0$ is the distribution of νe_I with a uniformly random index $I \in \{1, \dots, m_n\}$ and a unit length vector e_i in the i -th coordinate direction. Finally, set $\theta_i = 0$ for the remaining $n - m_n s_n$ components.

Consider the Bayes risk based on the prior distribution $\Pi_{B,\nu}$: $\mathcal{B}(\nu) := \int R(\theta, q_{\Pi_B}) d\Pi_{B,\nu}(\theta)$. Since $\Pi_{B,\nu}$ is supported on $\Theta[s_n]$, the minimax risk is bounded below by $\mathcal{B}(\nu)$ for any $\nu > 0$ and thus bounded below by $\sup_{\nu>0} \mathcal{B}(\nu)$: $\mathcal{R}(\Theta[s_n]) \geq \sup_{\nu>0} \mathcal{B}(\nu)$. It will be shown that $\sup_{\nu>0} \mathcal{B}(\nu) \geq \mathcal{C} s_n \log \eta_n^{-1}$, from which we will have $\mathcal{R}(\Theta[s_n]) \geq \mathcal{C} s_n \log \eta_n^{-1}$.

An upper bound on $\mathcal{R}(\Theta[s_n])$ is derived through bounding the coordinate-wise Kullback–Leibler risk of the Bayes predictive density q_{Π^*} based on $\Pi^* = \Pi[s_n, 1]$:

$$\rho(\lambda) := \mathbb{E}_\lambda \log[\{\exp(-\lambda) \lambda^{Y_1} / Y_1!\} / \{q_{\Pi^*,i}(Y_1 | X_1)\}], \quad \lambda > 0,$$

where $q_{\Pi^*}(y_i | x_i)$ is the marginal distribution of q_{Π^*} . It will be shown that

- $\rho(0) = O(\eta_n)$;
- $\sup_{\lambda>0} \rho(\lambda) \leq (\mathcal{C} + o(1)) s_n \log \eta_n^{-1}$,

from which we will have

$$\begin{aligned} \mathcal{R}(\Theta[s_n]) &\leq \sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi^*}) \\ &= (n - s_n) \rho(0) + s_n \sup_{\lambda>0} \rho(\lambda) \leq (n - s_n) O(\eta_n) + (\mathcal{C} + o(1)) s_n \log \eta_n^{-1}. \end{aligned}$$

Here, the coordinate-wise independence of the predictive density q_{Π^*} is used.

In bounding $\mathcal{R}(\Theta[s_n])$ both below and above, the following formula for the Kullback–Leibler risk is employed. For an estimator $\hat{\theta}$, let

$$R_e(\theta, \hat{\theta}) := R(\theta, q(\cdot | \hat{\theta})) = \mathbb{E}_\theta \sum_{i=1}^n \left[\theta_i \log \frac{\theta_i}{\hat{\theta}_i(X)} - \theta_i + \hat{\theta}_i(X) \right].$$

For a prior Π of θ , let

$$\hat{\theta}_{\Pi,i}(x; t) = \int \theta_i p(x | t\theta) d\Pi(\theta) \Big/ \int p(x | t\theta) d\Pi(\theta), \quad i = 1, 2, \dots, n.$$

Let $\hat{\theta}_\Pi(x; t) := (\hat{\theta}_{\Pi,1}(x; t), \dots, \hat{\theta}_{\Pi,n}(x; t))$.

Lemma 5.1. For a prior Π of θ , if $\hat{\theta}_\Pi(x; t)$ based on Π is strictly larger than 0 for any $x \in \mathbb{N}^n$ and any $t \in (r, 1 + r)$, then, the equality

$$R(\theta, q_\Pi) = \int_r^{r+1} \frac{R_e(t\theta, t\hat{\theta}_\Pi(\cdot; t))}{t} dt$$

holds.

This formula is known in the literature, but the proof is given in Appendix B for the sake of completeness.

5.2. Proof of Theorem 2.1 (a).

Lower bound on $\mathcal{R}(\Theta[s_n])$. First, the explicit form of $\hat{\theta}_{\Pi_{B,\nu}}$ is derived. Note that the marginal distribution of X based on $\Pi_{B,\nu}$ is supported on $\mathcal{X}_1 \times \dots \times \mathcal{X}_{s_n} \times \{0\}^{n-m_n s_n}$, where $\mathcal{X}_j := \{x^{(j)} := (x_1, \dots, x_{m_n}) : \|x^{(j)}\|_0 \leq 1\}$ ($j = 1, 2, \dots, s_n$). It follows from the Bayes formula that for each $j = 1, \dots, s_n$ and for each $x^{(j)} \in \mathcal{X}_j$, the Bayes estimator $\hat{\theta}_{\Pi_{B,\nu},i}$ ($i = 1 + m_n(j-1), \dots, m_n j$) is of the form

$$\hat{\theta}_{\Pi_{B,\nu},i}(x^{(j)}) = \begin{cases} \nu/m_n & \text{if } \|x^{(j)}\|_0 = 0, \\ \nu & \text{if } x_i^{(j)} \neq 0 \text{ and } x_k^{(j)} = 0 \text{ for } k \neq i, \\ 0 & \text{if otherwise.} \end{cases}$$

Second, we have, for θ in the support of $\Pi_{B,\nu}$,

$$R_e(t\theta, t\hat{\theta}_{\Pi_{B,\nu}}) = s_n e^{-t\nu} t\nu \log \lfloor n/s_n \rfloor,$$

where we use the expression of $\hat{\theta}_{\Pi_{B,\nu}}$, and use the fact that for each $j = 1, \dots, s_n$, the equality $\sum_{i=1+m_n(j-1)}^{m_n j} \hat{\theta}_{\Pi_{B,\nu},i}(X^{(j)}) = \nu$ holds for almost all $X^{(j)}$. From Lemma 5.1 and from the above inequality, we have, for θ in the support of $\Pi_{B,\nu}$,

$$R(\theta, q_{\Pi_{B,\nu}}) = \int_r^{r+1} \frac{R_e(t\theta, t\hat{\theta}_{\Pi_{B,\nu}})}{t} dt \geq \{e^{-r\nu} - e^{-(r+1)\nu}\} s_n \log \lfloor \eta_n^{-1} \rfloor.$$

Taking expectation of $R(\theta, q_{\Pi_{B,\nu}})$ with respect to $\Pi_{B,\nu}$ yields

$$\mathcal{R}(\Theta[s_n]) \geq \inf_{\hat{q}} \int R(\theta, \hat{q}) d\Pi_{B,\nu}(\theta) = \int R(\theta, q_{\Pi_{B,\nu}}) d\Pi_{B,\nu}(\theta) \geq \{e^{-r\nu} - e^{-(r+1)\nu}\} s_n \log \lfloor \eta_n^{-1} \rfloor.$$

Maximizing the rightmost hand side in the above inequality with respect to ν , the desired lower bound $\mathcal{R}(\Theta[s_n]) \geq \mathcal{C} s_n \log \lfloor \eta_n^{-1} \rfloor$ is obtained.

Upper bound on $\mathcal{R}(\Theta[s_n])$. To bound $\rho(\lambda)$, further knowledge of $\hat{\theta}_{\Pi^*}(\cdot; t)$ is needed. The explicit form of $\hat{\theta}_{\Pi^*}(x; t)$ depends only on whether x is 0 or not. The following two identities for $\hat{\theta}_{\Pi^*}(\cdot; t)$ show the form of $\hat{\theta}_{\Pi^*}$: for $t \in (r, r+1)$, we have:

$$\begin{aligned} \hat{\theta}_{\Pi^*,1}(x_1; t) &= \frac{\eta_n/t^{\kappa+1}}{1 + \eta_n/t^\kappa}, & x_1 = 0; \\ \hat{\theta}_{\Pi^*,1}(x_1; t) &= (x_1 + 1)/t, & x_1 \geq 1. \end{aligned}$$

The derivations are as follows. It follows from the Bayes formula that for $x_1 \in \mathbb{N}$ we have

$$\widehat{\theta}_{\Pi^*,1}(x_1; t) = \frac{0^{x_1+1} + \eta_n \Gamma(x_1 + 2)/t^{x_1+2}}{0^{x_1} + \eta_n \Gamma(x_1 + 1)/t^{x_1+1}}, \quad (1)$$

where we use the convention that $0^0 = 1$. Substituting $x_1 = 0$ into equality (1) yields the first identity. Substituting $x_1 \geq 1$ into equality (1) yields the second identity.

Using the above identities for $\widehat{\theta}_{\Pi^*}$, an upper bound on $\rho(\cdot)$ (and thus on $\mathcal{R}(\Theta[s_n])$) will be derived. For $\lambda > 0$ and $t \in (r, r+1)$, let

$$\widehat{\rho}(\lambda, x_1; t) := t\lambda \log\{\lambda/\widehat{\theta}_{\Pi^*,1}(x_1; t)\} - t\lambda + t\widehat{\theta}_{\Pi^*,1}(x_1; t).$$

From the above two identities for $\widehat{\theta}_{\Pi^*}(\cdot; t)$, we obtain the following two inequalities for $\widehat{\rho}(\lambda, x_1; t)$: there exist positive constants C_1 and C_2 depending only on r for which for sufficiently large $n \in \mathbb{N}$ (depending only on r), we have

$$\begin{aligned} \widehat{\rho}(\lambda, x_1; t) &\leq t\{\lambda \log \eta_n^{-1} + \lambda \log \lambda - \lambda + \lambda \log C_1 + \eta_n C_2\}, & x_1 = 0, \\ \widehat{\rho}(\lambda, x_1; t) &\leq t[\lambda \log\{t\lambda/(x_1 + 1)\} - \lambda + (x_1 + 1)/t], & x_1 \geq 1. \end{aligned}$$

Constants C_1 and C_2 are given by $(r+2)^2$ and $(1/r^2)$, respectively. It follows that for $\lambda > 0$, we have

$$\begin{aligned} \rho(\lambda) &= \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[\widehat{\rho}(\lambda, X_1; t)]}{t} dt \\ &\leq \int_r^{r+1} \left\{ e^{-t\lambda} \lambda \log \eta_n^{-1} + e^{-t\lambda} \lambda \log(\lambda C_1) + \lambda \log(1 - e^{-t\lambda}) + (1 - e^{-t\lambda})/t \right\} dt + \eta_n C_2, \end{aligned} \quad (2)$$

where we use the inequalities for $\widehat{\rho}(\lambda; x_1)$, Jensen's inequality for the function $\log(1/\cdot)$, and the identity $\mathbb{E}_\lambda[1/(X_1 + 1)] = \{1 - \exp(-\lambda)\}/\lambda$ for $\lambda > 0$.

Finally, the inequality $\sup_{\lambda>0} \rho(\lambda) \leq (\mathcal{C} + o(1)) \log \eta_n^{-1}$ follows from inequality (2) and the inequality $\rho(0) = O(\eta_n)$ follows from the bound on $\widehat{\rho}(\lambda, x_1; t)$ in case with $x_1 = 0$. Thus, the desired upper bound

$$\mathcal{R}(\Theta[s_n]) \leq (n - s_n)O(\eta_n) + (\mathcal{C} + o(1))s_n \log \eta_n^{-1}$$

is obtained and the proof is completed. \square

5.3. Proof of Theorem 2.1 (b).

Lower bound on $\mathcal{R}(\Theta[s_n, \varepsilon_n])$. A lower bound is established by the fact that we have, for a sufficiently large n (depending only on r) and for $\nu \in (0, 1)$,

$$\{\theta \in \mathbb{R}_+^n : \theta_{[1]} = \cdots = \theta_{[n/\lfloor n/s_n \rfloor]} = \nu, \theta_{[n/\lfloor n/s_n \rfloor + 1]} = \cdots = \theta_{[n]} = 0\} \subset \Theta[s_n, \varepsilon_n], \quad (3)$$

where $\theta_{[i]}$ is the i -th largest component of $\{\theta_i : i = 1, \dots, n\}$. The relation (3) is immediately derived by definition of $\Theta[s_n, \varepsilon_n]$. Using (3), we have

$$\mathcal{R}(\Theta[s_n, \varepsilon_n]) \geq \sup_{\nu \in (0,1)} \int R(\theta, q_{\Pi_{B,\nu}}) d\Pi_{B,\nu}(\theta) \geq \{\mathcal{C}s_n \log(n/s_n)\}(1 + o(1)).$$

This establishes a lower bound on $\mathcal{R}(\Theta[s_n, \varepsilon_n])$.

Upper bound on $\mathcal{R}(\Theta[s_n, \varepsilon_n])$. Recall that $N(\theta, \varepsilon) := \#\{i : \theta_i > \varepsilon\}$. An upper bound is established by the fact that we have, for an independent prior Π ,

$$R(\theta, q_\Pi) \leq N(\theta, \varepsilon) \sup_{\varepsilon_n < \lambda} R_i(\lambda, q_{\Pi, i}) + (n - N(\theta, \varepsilon)) \sup_{\lambda \leq \varepsilon_n} R_i(\lambda, q_{\Pi, i}), \quad (4)$$

where R_i ($i = 1, \dots, n$) is the Kullback–Leibler risk for the i -th coordinate and the inequality (4) is shown by definition of $N(\theta, \varepsilon)$. Since $N(\theta, \varepsilon_n) \leq s_n$ for $\theta \in \Theta[s_n, \varepsilon_n]$, we have

$$R(\theta, q_\Pi) \leq s_n \sup_{0 < \lambda} R_i(\lambda, q_{\Pi, i}) + n \sup_{\lambda \leq \varepsilon_n} R_i(\lambda, q_{\Pi, i}). \quad (5)$$

Here we put a prior $\Pi^* = \Pi[s_n, 1]$ into Π . Following the same lines in upper-bounding $\mathcal{R}(\Theta[s_n])$, $R_i(\lambda, q_{\Pi^*, i})$ is bounded above as follows: for some positive constants C_1 and C_2 depending only on r such that for sufficiently large $n \in \mathbb{N}$ (depending only on r), for $\lambda > 0$,

$$\begin{aligned} R_i(\lambda, q_{\Pi^*, i}) &\leq \int_r^{r+1} [e^{-t\lambda} \lambda \log \eta_n^{-1} + e^{-t\lambda} \lambda \log(\lambda C_1) \\ &\quad + \lambda \log(1 - e^{-t\lambda}) + (1 - e^{-t\lambda})/t] dt + \eta_n C_2. \end{aligned} \quad (6)$$

The inequality $\sup_{0 < \lambda} R_i(\lambda, q_{\Pi^*, i}) \leq C \log(n/s_n)$ follows immediately from inequality (6). For the evaluation of $\sup_{\lambda \leq \varepsilon_n} R_i(\lambda, q_{\Pi^*, i})$, note that for sufficiently small λ (depending only on r), the first, the second and the fourth terms in the right hand side of (6) are monotonically increasing with respect to λ , while the third term of (6) is monotonically decreasing with respect to λ ; hence each supremum of the first, the second, and the fourth terms over $\lambda \in (0, \varepsilon_n)$ is obtained by substituting $\lambda = \varepsilon_n$, while the supremum of the third term is obtained by letting $\lambda \rightarrow 0$. Thus using the inequality $e^{-t\lambda} \geq 1 - t\lambda$ and using the right-continuity of $\lambda \log \lambda$ at $\lambda = 0$, it is shown that

$$\sup_{\lambda \leq \varepsilon_n} R_i(\lambda, q_{\Pi^*, i}) \leq \{e^{-r\varepsilon_n} - e^{-(r+1)\varepsilon_n}\} \log \eta_n^{-1} + e^{-(r+1)\varepsilon_n} \varepsilon_n \log(\varepsilon_n C_1) + \varepsilon_n + \eta_n C_2. \quad (7)$$

Since $e^{-r\lambda} - e^{-(r+1)\lambda} = O(\lambda)$, it follows that $\sup_{\lambda < \varepsilon_n} R_i(\lambda, q_{\Pi^*, i})$ is $O(\varepsilon_n \log\{\eta_n^{-1} \varepsilon_n\})$. Thus, we obtain

$$\mathcal{R}(\Theta[s_n, \varepsilon_n]) \leq \{C s_n \log \eta_n^{-1}\} (1 + o(1)),$$

which completes the proof. \square

5.4. Proof of Theorem 2.2. The proof will be completed by replacing $\Pi^* = \Pi[s_n, 1]$ by $\Pi[s_n, \kappa]$ with $\kappa > 0$ in the proof of bounding $\mathcal{R}(\Theta[s_n])$ and $\mathcal{R}(\Theta[s_n, \varepsilon_n])$ above. We provide the proof only for $\Theta[s_n]$. Following the same argument as in the proof of Theorem 2.1, it is shown that for $t \in (r, r+1)$, the following equalities hold:

$$\begin{aligned} \widehat{\theta}_{\Pi[s_n, \kappa], 1}(x_1; t) &= (\eta_n / t^{\kappa+1}) / (1 + \eta_n / t^\kappa) & x_1 &= 0; \\ \widehat{\theta}_{\Pi[s_n, \kappa], 1}(x_1; t) &= (x_1 + \kappa) / t, & x_1 &\geq 1; \end{aligned}$$

Using the above equalities for $\widehat{\theta}_{\Pi[s_n, \kappa]}$, there exists a positive constant C_1 depending only on κ and r for which we have

$$\begin{aligned} \sup_{\lambda > 0} \mathbb{E}_\theta \log[\{\exp(-\lambda)\lambda^{Y_1}/Y_1!\}/\{q_{\Pi[s_n, \kappa], 1}(Y_1 | X_1)\}] &\leq C \log(\eta_n^{-1}) + C_1, \\ \lim_{\lambda \rightarrow 0} \mathbb{E}_\theta \log[\{\exp(-\lambda)\lambda^{Y_1}/Y_1!\}/\{q_{\Pi[s_n, \kappa], 1}(Y_1 | X_1)\}] &= O(\eta_n), \end{aligned}$$

where let $q_{\Pi[s_n, \kappa], 1}(y_1 | x_1)$ be the marginal distribution of $q_{\Pi[s_n, \kappa]}$. Thus, we have

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[s_n, \kappa]}) \leq (C + o(1))s_n \log \eta_n^{-1} + (n - s_n)O(\eta_n),$$

which completes the proof. \square

5.5. Proofs of Propositions.

5.5.1. Proof of Proposition 2.1. The proof follows almost the same line as in the proof of Theorem 2.1. We provide the proof only for $\Theta[s_n]$. In bounding $\mathcal{E}(\Theta[s_n])$ below, the Bayes risk maximization based on a block-independent prior $\Pi_{B, \nu}$ with $\nu > 0$ is employed. In bounding $\mathcal{E}(\Theta[s_n])$ above, a thresholding estimator is used for simplification of the proof. The thresholding estimator used in this proof is defined as follows: for each $i = 1, 2, \dots, n$, if $x_i = 0$, $\widehat{\theta}_{T, i}(x_i) = \eta_n$ and if $x_i \geq 1$, $\widehat{\theta}_{T, i}(x_i) = (x_i + 1)/r$.

Consider a lower bound on $\mathcal{E}(\Theta[s_n])$. It follows from the explicit expression of $\widehat{\theta}_{\Pi_{B, \nu}}$ with $\nu > 0$ that we have

$$\begin{aligned} \int R(\theta, q(\cdot | \widehat{\theta}_{\Pi_{B, \nu}})) d\Pi_{B, \nu}(\theta) &= s_n [e^{-r\nu} \nu \log\{\nu/(\nu \lfloor \eta_n \rfloor)\} + (1 - e^{-r\nu}) \nu \log(\nu/\nu)] \\ &= s_n \nu e^{-r\nu} \log \lfloor \eta_n^{-1} \rfloor \end{aligned}$$

and thus maximizing $\int R(\theta, q(\cdot | \widehat{\theta}_{\Pi_{B, \nu}})) d\Pi_{B, \nu}(\theta)$ with respect to ν yields the inequality

$$\mathcal{E}(\Theta[s_n]) \geq e^{-1} r^{-1} s_n (1 + o(1)) \log \eta_n^{-1}.$$

Consider an upper bound on $\mathcal{E}(\Theta[s_n])$. Introducing the notation

$$\widetilde{\rho}(\lambda) := \mathbb{E}_{r\lambda} [\lambda \log(\lambda/\widehat{\theta}_{T, 1}(X_1)) - \lambda + \widehat{\theta}_{T, 1}(X_1)],$$

the minimax risk $\mathcal{E}(\Theta[s_n])$ is bounded as $\mathcal{E}(\Theta[s_n]) \leq (n - s_n) \widetilde{\rho}(0) + s_n \sup_{\lambda > 0} \widetilde{\rho}(\lambda)$. Thus, deriving two (asymptotic) inequalities

$$\widetilde{\rho}(0) = O(\eta_n) \text{ and } \sup_{\lambda > 0} \widetilde{\rho}(\lambda) = (1 + o(1)) e^{-1} r^{-1} s_n \log \eta_n^{-1}$$

will complete the proof.

Consider $\sup_{\lambda > 0} \widetilde{\rho}(\lambda)$. For $\lambda > 0$, $\widetilde{\rho}(\lambda)$ is given by

$$\begin{aligned} \widetilde{\rho}(\lambda) &= \mathbb{E}_{r\lambda} 1_{X_1 \geq 1}(X_1) [\lambda \log\{\lambda/(X_1 + 1)\} - \lambda + (X_1 + 1)/r] \\ &\quad + \mathbb{E}_{r\lambda} 1_{X_1 = 0}(X_1) [\lambda \log\{\lambda/(\eta_n)\} - \lambda + \eta_n]. \end{aligned}$$

From this expression, $\tilde{\rho}(\lambda)$ is bounded above as

$$\begin{aligned} & \sup_{\lambda>0} \tilde{\rho}(\lambda) \\ & \leq \sup_{\lambda>0} \{1 - e^{-r\lambda} + \eta_n e^{-r\lambda} + e^{-r\lambda} \lambda \log \lambda\} + \sup_{\lambda>0} \{\lambda e^{-r\lambda}\} \log \eta_n^{-1} + \sup_{\lambda>0} \mathbb{E}_{r\lambda} \lambda \log \{r\lambda/(X_1 + 1)\}. \end{aligned} \quad (8)$$

The inequality $\sup_{\lambda>0} \tilde{\rho}(\lambda) \leq 1 + \eta_n + \sup_{\lambda>0} \lambda \exp(-r\lambda) \log \eta_n^{-1}$ follows from inequality (8) and from the identity $\mathbb{E}_{r\lambda}[r/(X_1 + 1)] = \lambda^{-1}(1 - \exp(-r\lambda))$ for $\lambda > 0$. Thus, the asymptotic equality

$$\sup_{\lambda>0} \tilde{\rho}(\lambda) = (1 + o(1))e^{-1}r^{-1}s_n \log \eta_n^{-1}$$

holds. Similarly, the asymptotic inequality $\tilde{\rho}(0) = O(\eta_n)$ holds. Hence, the desired upper bound on $\mathcal{E}(\Theta[s_n])$ is obtained and the proof is completed. \square

5.5.2. *Proof of Proposition 2.2.* Noticing that $s_n \log(n/s_n) \sim s_n \log n$ when $1 \leq \inf_n s_n \leq \sup_n s_n < \infty$, the proof is completed replacing $\eta_n = s_n/n$ in Theorem 2.2 by $1/n$. \square

5.5.3. *Proof of Proposition 2.3 (a).* Using the decomposition of the Kullback–Leibler divergence, we have

$$\begin{aligned} R(\theta, q_{\Pi[\hat{s}_n, \kappa]}) &= R(\theta, q_{\Pi[s_n, \kappa]}) + \sum_{i=1}^n \mathbb{E}_\theta \log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | X_i)} \right\} \\ &= R(\theta, q_{\Pi[s_n, \kappa]}) + \sum_{i \in \mathcal{A}} \log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | X_i)} \right\} + \sum_{i \notin \mathcal{A}} \log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | X_i)} \right\}, \end{aligned} \quad (9)$$

where for $\theta \in \Theta[s_n]$, let $\mathcal{A} := \mathcal{A}(\theta) = \{i : \theta_i \neq 0\}$. In what follows, we will bound the second and the third terms in the rightmost side of the above equality.

For $i \notin \mathcal{A}$, we have

$$\begin{aligned} & \mathbb{E}_\theta \log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | X_i)} \right\} \\ &= \mathbb{E}_\theta \log \left\{ \frac{1 + \hat{s}_n \Gamma(\kappa)/(nr^\kappa)}{1 + s_n \Gamma(\kappa)/(nr^\kappa)} \right\} + \mathbb{E}_\theta \log \left[\frac{1 + \{s_n \Gamma(\kappa)/(nr^\kappa)\} \{r/(r+1)\}^\kappa}{1 + \{\hat{s}_n \Gamma(\kappa)/(nr^\kappa)\} \{r/(r+1)\}^\kappa} \right] \\ &\leq 2 \frac{s_n \Gamma(\kappa)}{nr^\kappa} \mathbb{E}_\theta |\hat{s}_n/s_n - 1| + \frac{1}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 + \frac{1}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 \mathbb{E}_\theta \{1 + |\hat{s}_n/s_n - 1|\}^2 \\ &\leq 2 \frac{s_n \Gamma(\kappa)}{nr^\kappa} \mathbb{E}_\theta |\hat{s}_n/s_n - 1| + \frac{3}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 + \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 \mathbb{E}_\theta |\hat{s}_n/s_n - 1|^2 \end{aligned} \quad (10)$$

where the last two inequalities follow since $x - x^2/2 \leq \log(x+1) \leq x$ for $x > 0$ and since $xy \leq 2x^2 + 2y^2$ for $x, y > 0$.

For $i \in \mathcal{A}$, we consider the following four cases: (i) $X_i = 0, Y_i = 0$; (ii) $X_i \geq 1, Y_i = 0$; (iii) $X_i = 0, Y_i \geq 1$; (iv) $X_i \geq 1, Y_i \geq 1$. In Case (i), by the same argument as in the case for $i \notin \mathcal{A}$, we have

$$\log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | X_i)} \right\} \leq 2 \frac{s_n \Gamma(\kappa)}{nr^\kappa} |\hat{s}_n/s_n - 1| + \frac{3}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 + \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 |\hat{s}_n/s_n - 1|^2. \quad (11)$$

In Case (ii), we have

$$\log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | X_i)} \right\} = 0. \quad (12)$$

In Case (iii), we have

$$\begin{aligned} \log \left\{ \frac{q_{\Pi[s_n, \kappa]}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa]}(Y_i | X_i)} \right\} &= \log(s_n/\hat{s}_n) + \log \left\{ \frac{1 + \hat{s}_n \Gamma(\kappa)/(nr^\kappa)}{1 + s_n \Gamma(\kappa)/(nr^\kappa)} \right\} \\ &\leq \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right) |\hat{s}_n/s_n - 1| + \frac{1}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 + |\hat{s}_n/s_n - 1| + \log s_n/\gamma, \end{aligned} \quad (13)$$

where the last inequality follows since $-\log(1 + |\hat{s}_n/s_n - 1|) \leq \log(s_n/\hat{s}_n) \leq \log s_n$, and since $-x \leq -\log(1 + x)$ for $x > 0$. In Case (iv), we have

$$\log \left\{ \frac{q_{\Pi[s_n, \kappa]}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa]}(Y_i | X_i)} \right\} = 0. \quad (14)$$

From (11)-(14), for $i \in \mathcal{A}$, we have

$$\begin{aligned} \mathbb{E}_\theta \log \left\{ \frac{q_{\Pi[s_n, \kappa]}(Y_i | X_i)}{q_{\Pi[\hat{s}_n, \kappa]}(Y_i | X_i)} \right\} \\ \leq \left\{ 3 \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right) + 1 \right\} \mathbb{E}_\theta |\hat{s}_n/s_n - 1| + 3 \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 + \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 \mathbb{E}_\theta |\hat{s}_n/s_n - 1|^2 + \log \frac{s_n}{\gamma}. \end{aligned} \quad (15)$$

Combining (10) and (15) with (9) completes the proof. \square

5.5.4. *Proof of Proposition 2.3 (b).* The proof follows almost the same line as in the proof for Proposition 2.3 (a). For $\theta \in \Theta[s_n, \varepsilon_n]$, let $\mathcal{A} := \mathcal{A}(\theta) = \{i : \theta_i > \varepsilon_n\}$.

For $i \in \mathcal{A}$, we use the bound (15). For $i \notin \mathcal{A}$ and for $x_i \geq 1$, from (12) and (14), we have

$$\mathbb{E}_{Y_i | \theta_i} \log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | x_i)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | x_i)} \right\} = 0. \quad (16)$$

For $i \notin \mathcal{A}$ and for $x_i = 0$, from (11) and (13), we have

$$\begin{aligned} \mathbb{E}_{Y_i | \theta_i} \log \left\{ \frac{q_{\Pi[s_n, \kappa], i}(Y_i | x_i = 0)}{q_{\Pi[\hat{s}_n, \kappa], i}(Y_i | x_i = 0)} \right\} \\ \leq 2 \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right) |\hat{s}_n/s_n - 1| + \frac{1}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 + \frac{1}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 \{1 + |\hat{s}_n/s_n - 1|\}^2 \\ + (1 - e^{-\varepsilon_n}) \left[\left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right) |\hat{s}_n/s_n - 1| + \frac{1}{2} \left(\frac{s_n \Gamma(\kappa)}{nr^\kappa} \right)^2 + |\hat{s}_n/s_n - 1| + \log s_n \right]. \end{aligned} \quad (17)$$

Combining (15), (16), and (17) with (9) completes the proof. \square

5.5.5. *Proof of Proposition 2.4.* We use the bias-variance decomposition of $\hat{s}_n - s_n$ to prove Proposition 2.4.

Proof for Case (a): Consider the bias of $\hat{s}_n - s_n$. For $j = 1, \dots, s_n$, $\theta_{[j]}$ denotes the j -th largest component of $\{\theta_i : i = 1, \dots, n\}$. It follows that

$$-\sum_{j=1}^{s_n} e^{-\theta_{[j]}} \leq \mathbb{E}_\theta \hat{s}_n - s_n \leq 1 \quad (18)$$

since the decomposition $\#\{i : X_i \geq 1\} = \sum_{j=1}^{s_n} Z_j$ holds, where Z_j ($j = 1, \dots, s_n$) is independent and distributed according to Bernoulli distribution with the success probability $1 - \exp(-r\theta_{[j]})$.

Consider the variance of $\widehat{s}_n - s_n$. It follows that

$$\mathbb{E}_\theta |\widehat{s}_n - \mathbb{E}_\theta \widehat{s}_n|^2 \leq 1 + \sum_{j=1}^{s_n} e^{-r\theta_{[j]}} (1 - e^{-r\theta_{[j]}}) \quad (19)$$

since the inequalities

$$-1 + \left\{ \sum_{j=1}^{s_n} (Z_j - \mathbb{E} Z_j) \right\} \leq \widehat{s}_n - \mathbb{E}_\theta \widehat{s}_n \leq 1 + \left\{ \sum_{j=1}^{s_n} (Z_j - \mathbb{E} Z_j) \right\}$$

hold.

Combining (18) with (19) completes the proof of the former part of Proposition 2.4 (a). Combining this with Proposition 2.3 (a) completes the proof of the latter part of Proposition 2.4 (a). \square

Proof for Case (b): The proof is almost the same as of Case (a).

Consider the bias of $\widehat{s}_n - s_n$. It follows that

$$-\sum_{j=1}^{s_n} e^{-\theta_{[j]}} \leq \mathbb{E}_\theta [\widehat{s}_n] - s_n \leq \sum_{j=s_n+1}^n (1 - \exp(-r\varepsilon_n)) \leq 1 + rn\varepsilon_n \quad (20)$$

since the decomposition $\#\{i : X_i \geq 1\} = \sum_{j=1}^n Z_j$ holds, where Z_j ($j = 1, \dots, n$) is independent and distributed according to Bernoulli distribution with the success probability $1 - \exp(-r\theta_{[j]})$, and the inequality $\theta_{[j]} \leq \varepsilon_n$ holds for $j = s_n + 1, \dots, n$.

Consider the variance of $\widehat{s}_n - s_n$. It follows that

$$\mathbb{E}_\theta |\widehat{s}_n - \mathbb{E}_\theta \widehat{s}_n|^2 \leq 1 + \sum_{j=1}^{s_n} e^{-r\theta_{[j]}} (1 - e^{-r\theta_{[j]}}) + nr\varepsilon_n \quad (21)$$

since the inequality

$$-1 - \sum_{j=1}^n (Z_j - \mathbb{E} Z_j) \leq \widehat{s}_n - \mathbb{E}_\theta \widehat{s}_n \leq 1 + \sum_{j=1}^n (Z_j - \mathbb{E} Z_j)$$

holds.

Combining (20) with (21) complete the proof of the former part of Proposition 2.4 (b). Combining this with Proposition 2.3 (b) completes the proof of the latter part of Proposition 2.4 (b). \square

5.5.6. *Proof of Proposition 2.5.* Consider bounds on $\widehat{\theta}_{\Pi_{SS},1}(X_1; t)$. The Bayes estimate $\widehat{\theta}_{\Pi_{SS},1}(X_1; t)$ is explicitly given as

$$\widehat{\theta}_{\Pi_{SS},1}(X_1; t) = \frac{(1 - \eta_n)\mu^{X_1+1} \exp(-t\mu) + \eta_n \nu_1^{k_1} (\nu_1 + t)^{-X_1-k_1-1} \Gamma(X_1 + k_1 + 1) / \Gamma(k_1)}{(1 - \eta_n)\mu^{X_1} \exp(-t\mu) + \eta_n \nu_1^{k_1} (\nu_1 + t)^{-X_1-k_1} \Gamma(X_1 + k_1) / \Gamma(k_1)}.$$

Combining this expression of $\widehat{\theta}_{\Pi_{SS},1}(X_1; t)$ with the simple algebra that $(a + b)/(c + d) \leq b/d$ for any $(a, b, c, d) \in \mathbb{R}_+^4$ such that $ad \leq bc$, we have, for sufficiently large n depending only on ν_1 and for $X_1 \geq 1$,

$$\widehat{\theta}_{\Pi_{SS},1}(X_1; t) \leq (X_1 + k_1) / (\nu_1 + t). \quad (22)$$

Using inequality (22) and Lemma 5.1, we have

$$R(\theta, q_{\Pi_{\text{SS}}}) \geq \int_r^{r+1} \mathbb{E}_{t\theta} \left[1_{X_1 \geq 1}(X_1) \left\{ \theta_1 \log \frac{\theta_1}{X_1 + k_1} + \theta_1 \log(\nu_1 + t) - \theta_1 + \hat{\theta}_{\Pi_{\text{SS}},1}(X_1; t) \right\} \right] dt.$$

Since $\hat{\theta}_{\Pi_{\text{SS}},1}(X_1; t)/\{\theta_1/(\nu_1 + t)\} \rightarrow 1$ almost surely as $\theta_1 \rightarrow \infty$, we have, for any $(\theta_2, \dots, \theta_n) \in \mathbb{R}_+^{n-1}$,

$$\liminf_{\theta_1 \rightarrow \infty} \frac{R(\theta, q_{\Pi_{\text{SS}}})}{\theta_1} \geq \int_1^{1+r} \{\log(\nu_1 + t) - 1 + 1/(\nu_1 + t)\} dt > 0$$

and thus $\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi_{\text{SS}}}) \rightarrow \infty$ for any $k \in \mathbb{R}_+^n$, which completes the proof. \square

5.5.7. *Proof of Proposition 2.6.* The proof steps are the same as in the proof of Theorem 2.2. For notational simplicity, we consider $\Pi_P = \prod_{i=1}^n \{\delta_0(d\theta_i) + (s_n/n)\theta_i^{-2}1_{(1,\infty)}(d\theta_i)\}$.

First, there exist positive constants C_1, C_2, C_3 depending only on r such that for $i = 1, \dots, n$ and for $t \in (r, r+1)$, we have

$$\begin{aligned} \eta_n C_1 &\leq \hat{\theta}_{\Pi_P, i}(x_i; t) \leq \eta_n C_2, & x_i &= 0, \\ \hat{\theta}_{\Pi_P, i}(x_i; t) &= C_3, & x_i &= 1, \\ (x_i - 1)/t &\leq \hat{\theta}_{\Pi_P, i}(x_i; t) \leq (x_i - 1 + t)/t, & x_i &\geq 2, \end{aligned}$$

where see the description before Lemma 5.1 for the definition of $\hat{\theta}_{\Pi_P}(\cdot; \cdot)$. These bounds are derived from the explicit form of $\hat{\theta}_P(\cdot; \cdot)$

$$\hat{\theta}_{\Pi_P}(x_i; t) = \frac{\eta_n \int_1^\infty \lambda^{x_i-1} e^{-t\lambda} d\lambda}{0^{x_i} + \eta_n \int_1^\infty \lambda^{x_i-2} e^{-t\lambda} d\lambda}$$

and from the identity

$$(x_i - 1) \int_1^\infty \lambda^{x_i-2} e^{-t\lambda} d\lambda - t \int_1^\infty \lambda^{x_i-1} e^{-t\lambda} d\lambda = -e^{-t}, \quad x_i \geq 0, \quad t > 0.$$

Second, for $\lambda > 0$, $x_1 \in \mathbb{N}$, and $t \in (r, r+1)$, let

$$\hat{\rho}(\lambda, x_1; t) := t\lambda \log\{\lambda/\hat{\theta}_{\Pi_P}(x_1; t)\} - t\lambda + t\hat{\theta}_{\Pi_P}(x_1; t).$$

From the above inequalities for $\hat{\theta}_{\Pi_P}(\cdot; \cdot)$, we have, for $\lambda > 0$ and $t \in (r, r+1)$,

$$\begin{aligned} \hat{\rho}(\lambda, x_1; t) &\leq t\{\lambda \log \eta_n^{-1} + \lambda \log \lambda - \lambda + \lambda \log C_1 + \eta_n C_2\}, & x_1 &= 0, \\ \hat{\rho}(\lambda, x_1; t) &\leq t[\lambda \log\{t\lambda/C_3\} - \lambda + C_3], & x_1 &= 1, \\ \hat{\rho}(\lambda, x_1; t) &\leq t[\lambda \log\{t\lambda/(x_1 - 1)\} - \lambda + (x_1 - 1 + t)/t], & x_1 &\geq 2. \end{aligned}$$

Finally, for $\lambda > 0$, let

$$\rho(\lambda) = \mathbb{E}_\lambda \log\{\{\exp(-\lambda)\lambda^{Y_1}/Y_1!\}/\{q_{\Pi_P}(Y_1 | X_1)\}\}.$$

It will be shown that $\rho(0) = O(\eta_n)$ and $\sup_{\lambda > 0} \rho(\lambda) \leq (C + o(1))s_n \log \eta_n^{-1}$. From Lemma 5.1, we have

$$\rho(\lambda) = \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[\hat{\rho}(\lambda, X_1; t)]}{t} dt$$

and thus all we have to do is to bound $\mathbb{E}_{t\lambda}[\widehat{\rho}(\lambda, X_1; t)]$ uniformly in $t \in (r, r+1)$. Consider the decomposition

$$\mathbb{E}_{t\lambda}[\widehat{\rho}(\lambda, X_1; t)] = \mathbb{E}_{t\lambda}[1_{X_1=0}\widehat{\rho}(\lambda, X_1; t)] + \mathbb{E}_{t\lambda}[1_{X_1=1}\widehat{\rho}(\lambda, X_1; t)] + \mathbb{E}_{t\lambda}[1_{X_1 \geq 2}\widehat{\rho}(\lambda, X_1; t)].$$

Bounding the first and the second terms in the right hand side of the above decomposition is straightforward. As a result, we have

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[1_{X_1=0}\widehat{\rho}(\lambda, X_1; t)]}{t} dt &= O(\eta_n), \\ \sup_{\lambda > 0} \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[1_{X_1=0}\widehat{\rho}(\lambda, X_1; t)]}{t} dt &= \mathcal{C} \log \eta_n^{-1} + O(1), \\ \lim_{\lambda \rightarrow 0} \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[1_{X_1=1}\widehat{\rho}(\lambda, X_1; t)]}{t} dt &= 0, \\ \sup_{\lambda > 0} \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[1_{X_1=1}\widehat{\rho}(\lambda, X_1; t)]}{t} dt &= O(1). \end{aligned}$$

The remaining part of the proof is to show that

$$\lim_{\lambda \rightarrow 0} \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[1_{X_1 \geq 2}\widehat{\rho}(\lambda, X_1; t)]}{t} dt = 0 \text{ and } \sup_{\lambda > 0} \int_r^{r+1} \frac{\mathbb{E}_{t\lambda}[1_{X_1 \geq 2}\widehat{\rho}(\lambda, X_1; t)]}{t} dt = O(1).$$

The former equality holds since $\log\{1/(X_1 - 1)\} \leq 0$ for $X_1 \geq 2$ and since

$$\mathbb{E}_{t\lambda}[1_{X_1 \geq 2}(X_1 - 1 + t)/t] \leq \lambda + (1 - e^{-t\lambda}), \quad t \in (r, r+1).$$

The latter inequality is obtained by the inequalities

$$\lambda \mathbb{E}_\lambda[1_{X_1 \geq 2} \log\{\lambda/(X_1 - 1)\}] \leq 4 \log 4 \quad \text{for } 0 < \lambda \leq 4, \quad (23)$$

$$\lambda \mathbb{E}_\lambda[1_{X_1 \geq 2} \log\{\lambda/(X_1 - 1)\}] \leq \lambda \log\{\lambda \exp(-\lambda^{1/2}/2) + \lambda/(\lambda - \lambda^{3/4} - 1)\} \quad \text{for } 4 < \lambda, \quad (24)$$

where inequality (23) follows since $\log(1/(x_1 - 1)) \leq 0$ for $x_1 \geq 2$ and inequality (24) follows by introducing $Z = (X_1 - \lambda)/\sqrt{\lambda}$ and by applying the Bennett inequality for the Poisson distribution to Z . Thus, we complete the proof. \square

6. PROOFS FOR SECTION 3

The proofs for the results in Section 3 are given in this section.

6.1. Proof of Theorem 3.1. The main ingredient of the proof of Theorem 2.1 is Lemma 6.1 describing the exact asymptotic minimax risks for a setting with a fixed sequence $\{r_i : i = 1, \dots, n\}$ satisfying the following condition:

Condition 6.1. *The asymptotic equality $\bar{\mathcal{C}} := \sum_{i=1}^n \mathcal{C}_i/n \sim \sum_{i \in J} \mathcal{C}_i/s_n$ holds for any subset $J \subset \{1, \dots, n\}$ satisfying $|J| = s_n$.*

Lemma 6.1. *Fix an infinite sequence $\{r_i \in (0, \infty) : i \in \mathbb{N}\}$ such that $0 < \inf_i r_i \leq \sup_i r_i < \infty$. Suppose Condition 6.1 holds.*

(a) For the sparse parameter space $\Theta[s_n]$, the asymptotic equality

$$\overline{\mathcal{R}}(\Theta[s_n]) \sim \overline{\mathcal{C}} s_n \log(n/s_n)$$

holds as $n \rightarrow \infty$ and $s_n/n \rightarrow 0$.

(b) For the quasi sparse parameter space $\Theta[s_n, \varepsilon_n]$ with any shrinking sequence ε_n such that $\varepsilon_n = o(s_n/n)$, the asymptotic equality

$$\overline{\mathcal{R}}(\Theta[s_n, \varepsilon_n]) \sim \overline{\mathcal{C}} s_n \log(n/s_n)$$

holds as $n \rightarrow \infty$ and $s_n/n \rightarrow 0$.

The proof is given in Subsection 6.3.2.

Admitting that Lemma 6.1 holds, Theorem 3.1 will be proved using the following lemma:

Lemma 6.2. Fix $\delta \in (0, 1)$. Under Condition 3.1, the asymptotic inequality

$$|\overline{\mathcal{R}}(\Theta[s_n]) / (\mathbb{E}_G \mathcal{C}_1 s_n \log \eta_n^{-1}) - 1| \leq b_n + \sqrt{\{1/(2n)\} \log(2/\delta)}$$

holds with probability higher than $1 - \delta$, where b_n is an $o_p(1)$ term independent of δ .

The proof is given in Subsection 6.3.3.

Go back to the proof of Theorem 3.1. By substituting $\delta = \delta_n = \exp(-n/\log n)$ in Lemma 6.2 and by using Lemma 6.1, Theorem 3.1 is derived directly. \square

6.2. Proof of Theorem 3.2. The main ingredient of the proof of Theorem 3.2 is Lemma 6.3 for a fixed sequence $\{r_i \in (0, \infty) : i \in \mathbb{N}\}$ satisfying Condition 6.1.

Lemma 6.3 states that an improper prior $\Pi[s_n, \kappa]$ yields Bayes predictive density attaining the exact asymptotic minimaxity even for a setting with a fixed sequence $\{r_i : i \in \mathbb{N}\}$ satisfying Condition 6.1.

Lemma 6.3. Fix $\kappa > 0$. Fix an infinite sequence $\{r_i \in (0, \infty) : i \in \mathbb{N}\}$ such that $0 < \inf_i r_i \leq \sup_i r_i < \infty$. Suppose Condition 6.1 holds. The predictive density $q_{\Pi[s_n, \kappa]}$ attains exact asymptotic minimaxity:

$$\begin{aligned} \sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[s_n, \kappa]}) &\sim \overline{\mathcal{R}}(\Theta[s_n]) \\ \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[s_n, \kappa]}) &\sim \overline{\mathcal{R}}(\Theta[s_n, \varepsilon_n]) \end{aligned}$$

hold as $n \rightarrow \infty$, where $s_n \in (0, n)$ is a sequence such that $s_n = o(n)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(s_n/n)$.

The proof is given in Subsection 6.3.4.

Admitting that Lemma 6.3, Theorem 3.2 is proved by Lemmas 6.2 and 6.3. \square

6.3. Proofs of Lemmas 6.1-6.3.

6.3.1. *Outline of the proof of Lemma 6.1.* The outline of the proof of Lemma 6.1 is provided ahead. The proofs of both (a) and (b) of Lemma 6.1 are divided into two parts, providing upper and lower bounds on $\overline{\mathcal{R}}(\Theta[s_n])$ and $\overline{\mathcal{R}}(\Theta[s_n, \varepsilon_n])$, as with Theorem 2.1. Compared to the proof of Theorem 2.1, additional care about the difference of r_i 's is required in each part of the proof.

In constructing each bound for $\overline{\mathcal{R}}(\Theta[s_n])$ (and $\overline{\mathcal{R}}(\Theta[s_n, \varepsilon_n])$), the following formula for the Kullback-Leibler risk is employed instead of Lemma 5.1. For $i = 1, 2, \dots, n$, let $t_i = t_i(\tau)$ ($i = 1, \dots, n$) be a smooth and monotonically increasing function of $\tau \in [0, 1]$ such that $t_i(0) = r_i$ and $t_i(1) = 1 + r_i$. Using $t_i(\tau)$, let $Z_i(\tau)$ ($i = 1, \dots, n$) be a random variable independently distributed to $\text{Po}(t_i(\tau)\theta_i)$. The density of $Z(\tau) = (Z_1(\tau), \dots, Z_n(\tau))$ is denoted by

$$p(z \mid \theta; \tau) = \prod_{i=1}^n \left[\frac{\exp\{-t_i(\tau)\theta_i\} \{t_i(\tau)\theta_i\}^{z_i}}{z_i!} \right].$$

By definitions, X and Y follow the same distributions as those of $Z(0)$ and $Z(1) - Z(0)$, respectively. For a prior Π of θ , let

$$\widehat{\theta}_{\Pi, i}(z; \tau) := \int \theta_i p(z \mid \theta; \tau) d\Pi(\theta) \Big/ \int p(z \mid \theta; \tau) d\Pi(\theta), \quad i = 1, \dots, n.$$

For a Bayes estimator $\widehat{\theta}_{\Pi}$ based on Π , let

$$R_e(\theta, \widehat{\theta}_{\Pi}; \tau) = \mathbb{E}_{\theta, \tau} \left[\sum_{i=1}^n t_i(\tau) \left\{ \theta_i \log \frac{\theta_i}{\widehat{\theta}_{\Pi, i}(Z; \tau)} - \theta_i + \widehat{\theta}_{\Pi, i}(Z; \tau) \right\} \right],$$

where $\mathbb{E}_{\theta, \tau}$ is the expectation with respect to $p(\cdot \mid \theta; \tau)$.

Lemma 6.4. [31] *For a prior Π of θ , if $\widehat{\theta}_{\Pi}(z; \tau)$ based on Π is strictly larger than 0 for any $z \in \mathbb{N}^n$ and any $\tau \in [0, 1]$, then, the equality*

$$R(\theta, q_{\Pi}) = \int_0^1 R_e(\theta, \widehat{\theta}_{\Pi}; \tau) d\tau$$

holds.

6.3.2. *Proof of Lemma 6.1.* We provide only the proof for (a) and omit the proof for (b). Following the same argument as in Subsection 5.3 completes the proof for case (b).

Lower bound on $\overline{\mathcal{R}}(\Theta[s_n])$. A lower bound on $\overline{\mathcal{R}}(\Theta[s_n])$ is constructed through Bayes risk maximization based on a *varied-spike block-independent* prior. Let $\Pi_{\text{VB}, \nu}$ with $\nu = (\nu^{(1)}, \dots, \nu^{(s_n)}) \in \mathbb{R}^{m_n} \times \mathbb{R}^{m_n} \times \dots \times \mathbb{R}^{m_n} \times \mathbb{R}^{n-m_n s_n}$ be a *varied-spike block-independent* prior built as follows: divide $\{1, 2, \dots, n\}$ into contiguous blocks B_j ($j = 1, 2, \dots, s_n$) with each length $m_n := \lfloor n/s_n \rfloor$. In each block B_j , draw $(\theta_{1+m_n(j-1)}, \dots, \theta_{m_n j})$ independently according to a single spike prior with spike strength parameter $\nu^{(j)} \in \mathbb{R}_+^{s_n}$, where a single spike prior with spike strength parameter $\nu^{(j)} \in \mathbb{R}_+^{s_n}$ is the distribution of $\nu_I^{(j)} e_I$ with a uniformly random index $I \in \{1, \dots, m_n\}$ and a unit length vector e_i in the i -th coordinate direction. Finally, set $\theta_i = 0$ for the remaining $n - m_n s_n$ components. In general, a varied-spike block-independent prior is different from a block-independent prior since the spike strength may be varied in each coordinate.

First, in order to consider the Bayes risk based on the prior distribution $\Pi_{\text{VB},\nu}$, the explicit form of $\hat{\theta}_{\Pi_{\text{VB},\nu}} = (\hat{\theta}_{\Pi_{\text{VB}}}^{(1)}, \dots, \hat{\theta}_{\Pi_{\text{VB}}}^{(s_n)})$ is derived as follows: for $j = 1, \dots, s_n - 1$ and for $k = 1, \dots, m_n$,

$$\hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(x) = \hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(x^{(j)}) = \begin{cases} w_k^{(j)} \nu_k^{(j)} & \text{if } \|x^{(j)}\|_0 = 0, \\ \nu_k^{(j)} & \text{if } x_k^{(j)} \neq 0 \text{ and } x_l^{(j)} = 0 \text{ for } l \neq k, \\ 0 & \text{otherwise,} \end{cases}$$

where $w_k^{(j)} := \exp(-r_k^{(j)} \nu_k^{(j)}) / \sum_{l=1}^{m_n} \exp(-r_l^{(j)} \nu_l^{(j)})$.

Using this expression of $\hat{\theta}_{\Pi_{\text{VB},\nu}}$, we evaluate the estimative risk $R_e(\theta, \hat{\theta}_{\Pi_{\text{VB},\nu}}; \tau)$. By the coordinate-wise additive property of the Kullback-Leibler divergence, $R_e(\theta, \hat{\theta}_{\Pi_{\text{VB},\nu}}; \tau)$ is decomposed as

$$R_e(\theta, \hat{\theta}_{\Pi_{\text{VB},\nu}}; \tau) = \sum_{j=1}^{s_n-1} R^{(j)}(\theta^{(j)}; \tau) + R^{(s_n)}(0; \tau),$$

where for $j = 1, \dots, s_n - 1$,

$$R^{(j)}(\theta^{(j)}; \tau) := \mathbb{E}_{\theta; \tau} \sum_{k=1}^{m_n} \dot{t}_k^{(j)}(\tau) \left[\theta_k^{(j)} \log \{ \theta_k^{(j)} / \hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(Z^{(j)}; \tau) \} - \theta_k^{(j)} + \hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(Z^{(j)}; \tau) \right]$$

and for $j = s_n$,

$$R^{(s_n)}(\theta^{(s_n)}; \tau) := \mathbb{E}_{\theta; \tau} \sum_{k=1}^{n-m_n s_n} \dot{t}_k^{(j)}(\tau) \left[\theta_k^{(j)} \log \{ \theta_k^{(j)} / \hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(Z^{(j)}; \tau) \} - \theta_k^{(j)} + \hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(Z^{(j)}; \tau) \right].$$

Second, we evaluate $R^{(j)}(\theta^{(j)}; \tau)$. Fix $j \in \{1, \dots, s_n - 1\}$. For notational brevity, we omit τ in $t_i(\tau)$'s. For $\theta^{(j)}$ in the j -th block of θ in the support of $\Pi_{\text{VB},\nu}$, we have

$$\begin{aligned} R^{(j)}(\theta^{(j)}; \tau) &= \mathbb{E}_{\theta; \tau} \sum_{k=1}^{m_n} \dot{t}_k^{(j)} \left[\theta_k^{(j)} \log \{ \theta_k^{(j)} / \hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(Z^{(j)}; \tau) \} - \theta_k^{(j)} + \hat{\theta}_{\Pi_{\text{VB},\nu,k}}^{(j)}(Z^{(j)}; \tau) \right] \\ &= e^{-t_\gamma^{(j)} \nu_\gamma^{(j)}} \left\{ \dot{t}_\gamma^{(j)} \nu_\gamma^{(j)} \log \frac{1}{w_\gamma^{(j)}} - \dot{t}_\gamma^{(j)} \nu_\gamma^{(j)} + \sum_{k=1}^{m_n} \dot{t}_k^{(j)} w_k^{(j)} \nu_k^{(j)} \right\}, \end{aligned}$$

where we denote by $\gamma = \gamma(j)$ the location in which the element is a spike. Taking the expectation with respect to $\Pi_{\text{VB},\nu}$ yields

$$\begin{aligned} \int R^{(j)}(\theta^{(j)}; \tau) d\Pi_{\text{VB},\nu}(\theta) &= \frac{1}{m_n} \sum_{k=1}^{m_n} e^{-t_k^{(j)} \nu_k^{(j)}} \left(\dot{t}_k^{(j)} \nu_k^{(j)} \log \frac{1}{w_k^{(j)}} - \dot{t}_k^{(j)} \nu_k^{(j)} + \sum_{l=1}^{m_n} \dot{t}_l^{(j)} w_l^{(j)} \nu_l^{(j)} \right) \\ &\geq \frac{1}{m_n} \sum_{k=1}^{m_n} e^{-t_k^{(j)} \nu_k^{(j)}} \left\{ \dot{t}_k^{(j)} \nu_k^{(j)} \log(1/w_k^{(j)}) - \dot{t}_k^{(j)} \nu_k^{(j)} \right\}. \end{aligned}$$

Integrating the both hand sides of the above equality with respect to τ over $[0, 1]$, we have

$$\int_0^1 \left[\int R^{(j)}(\theta^{(j)}; \tau) d\Pi_{\text{VB},\nu}(\theta) \right] d\tau \geq \frac{1}{m_n} \sum_{k=1}^{m_n} \left\{ f_k^{(j)}(\nu_k^{(j)}) \log(1/w_k^{(j)}) - f_k^{(j)}(\nu_k^{(j)}) \right\},$$

where $f_k^{(j)}(\lambda) := \exp\{-r_k^{(j)}\lambda\} - \exp\{-(1+r_k^{(j)})\lambda\}$, $\lambda > 0$. By summing up the block-wise risk evaluation, we have the following lower bound on the overall Bayes risk of $\Pi_{VB,\nu}$:

$$\begin{aligned}\overline{\mathcal{R}}(\Theta[s_n]) &\geq \sum_{j=1}^{s_n} \int_0^1 \left[\int R^{(j)}(\theta^{(j)}; \tau) d\Pi_{VB,\nu}(\theta) \right] d\tau \\ &\geq \frac{1}{m_n} \sum_{j=1}^{s_n} \sum_{k=1}^{m_n} f_k^{(j)}(\nu_k^{(j)}) \log \frac{1}{w_k^{(j)}} - \frac{1}{m_n} \sum_{j=1}^{s_n} \sum_{k=1}^{m_n} f_k^{(j)}(\nu_k^{(j)}) \\ &= \frac{1}{m_n} \sum_{j=1}^{s_n} \sum_{k=1}^{m_n} f_k^{(j)}(\nu_k^{(j)}) \log \frac{1}{w_k^{(j)}} - \frac{1}{m_n} \sum_{i=1}^n f_i(\nu_i),\end{aligned}\tag{25}$$

where $f_i(\lambda) := \exp\{-r_i\lambda\} - \exp\{-(1+r_i)\lambda\}$, $\lambda > 0$.

Third, we show that the asymptotic inequality

$$\overline{\mathcal{R}}(\Theta[s_n]) \geq \bar{f}(\nu) \{s_n \log(n/s_n)\} (1 + o(1))\tag{26}$$

holds, where $\bar{f}(\nu) := \sum_{i=1}^n f_i(\nu_i)/n$, provided that the following condition holds for ν :

Condition 6.2. *there exists a positive constant C such that $\max_l r_l^{(j)} \nu_l^{(j)} \leq C$ for any $j = 1, \dots, s_n$.*

For ν satisfying Condition 6.2, the first term of (25) is evaluated as

$$\begin{aligned}\frac{1}{m_n} \sum_{j=1}^{s_n} \sum_{k=1}^{m_n} f_i^{(j)}(\nu_i^{(j)}) \log \frac{1}{w_k^{(j)}} &= \frac{1}{m_n} \sum_{j=1}^{s_n} \sum_{k=1}^{m_n} f_i^{(j)}(\nu_i^{(j)}) \left(\log m_n + \log \frac{1}{m_n w_k^{(j)}} \right) \\ &= \bar{f}(\nu) s_n \{ \log(n/s_n) \} (1 + o(1)),\end{aligned}$$

because it follows by definition of $w_k^{(j)}$ that for any $k = 1, \dots, m_n$ and $j = 1, \dots, s_n$,

$$\exp \left\{ - \max_{l=1, \dots, m_n} r_l^{(j)} \nu_l^{(j)} + r_k^{(j)} \nu_k^{(j)} \right\} \leq \frac{1}{m_n w_k^{(j)}} \leq \exp \left\{ - \min_{l=1, \dots, m_n} r_l^{(j)} \nu_l^{(j)} + r_k^{(j)} \nu_k^{(j)} \right\}.$$

For ν satisfying Condition 6.2, the second term of (25) is negligible compared to the first term.

Finally, Condition 6.2 holds for ν° that maximizes the right hand side of (26), that is, $\nu_i^\circ = \log(1 + 1/r_i)$ ($i = 1, \dots, n$), and the desired lower bound is obtained by substituting $\nu = \nu^\circ$.

Upper bound on $\overline{\mathcal{R}}(\Theta[s_n])$. An upper bound on $\overline{\mathcal{R}}(\Theta[s_n])$ is derived through almost the same procedure as in Subsection 5.2. Let $\Pi^* = \Pi[s_n, 1]$. Fix $i = 1, 2, \dots, n$. For $\lambda_i > 0$, let

$$\rho_i(\lambda_i) := \mathbb{E}_{\lambda_i} \log[\{\exp(-\lambda_i) \lambda_i^{Y_1} / Y_1!\} / \{q_{\Pi^*,i}(Y_1 \mid X_1)\}].$$

To bound $\rho_i(\lambda_i)$, we employ the following equalities related to the behavior of the Bayes estimator $\hat{\theta}_{\Pi^*}$: for $\tau \in [0, 1]$, we have

$$\begin{aligned}\hat{\theta}_{\Pi^*,i}(z_i; \tau) &= (\eta_n/t_i(\tau)^{\kappa+1}) / (1 + \eta_n/t_i(\tau)^\kappa), & z_i &= 0, \\ \hat{\theta}_{\Pi^*,i}(z_i; \tau) &= (z_i + 1)/t_i(\tau), & z_i &\geq 1.\end{aligned}$$

For $\lambda_i > 0$ and $\tau \in [0, 1]$, let

$$\hat{\rho}_i(\lambda_i, z_1; \tau) := \dot{t}_i(\tau) \left[\lambda \log\{\lambda / \hat{\theta}_{\Pi^*,1}(z_1; \tau)\} - \lambda_i + \hat{\theta}_{\Pi^*,1}(z_1; \tau) \right].$$

Using the above equalities for $\hat{\theta}_{\Pi^*}$, the following bounds on $\hat{\rho}_i(\lambda_i, z_1; \tau)$ are derived: there exist positive constants C_1, C_2 such that we have

$$\begin{aligned}\hat{\rho}_i(\lambda_i, z_i; \tau) &\leq \dot{t}_i(\tau) \{ \lambda_i \log \eta_n^{-1} + \lambda_i \log \lambda_i - \lambda_i + \lambda_i \log C_1 + \eta_n C_2 \}, & z_1 = 0, \\ \hat{\rho}_i(\lambda_i, z_1; \tau) &\leq \dot{t}_i(\tau) \{ \lambda_i \log \{ \dot{t}_i(\tau) \lambda_i / (z_i + 1) \} - \lambda_i + (z_i + 1) / \dot{t}_i(\tau) \} & z_i \geq 1.\end{aligned}$$

Here $C_1 = \sup_i (r_i + 2)^2$ and $C_2 = 1 / (\inf_i r_i^2)$.

By the same way as in Subsection 5.2, it follows that for sufficiently large $n \in \mathbb{N}$ and for $\lambda_i > 0$, we have

$$\begin{aligned}\rho_i(\lambda_i) &= \int_0^1 \mathbb{E}_{t_i(\tau) \lambda_i} [\hat{\rho}_i(\lambda_i, Z_1; \tau)] d\tau \\ &\leq \int_0^1 \left[e^{-t_i(\tau) \lambda_i} \dot{t}_i(\tau) \lambda_i \log \eta_n^{-1} + e^{-t_i(\tau) \lambda_i} \dot{t}_i(\tau) \lambda_i \log (\lambda_i C_1) \right. \\ &\quad \left. + \dot{t}_i(\tau) \lambda_i \log (1 - e^{-t_i(\tau) \lambda_i}) + \dot{t}_i(\tau) / t_i(\tau) (1 - e^{-t_i(\tau) \lambda_i}) \right] d\tau + \eta_n C_2.\end{aligned}\quad (27)$$

Thus, the inequality

$$\sup_{\lambda > 0} \rho_i(\lambda) \leq (\mathcal{C}_i + o(1)) \log \eta_n^{-1}$$

follows from (27), and the inequality $\rho_i(0) = O(\eta_n)$ follows from the above bound on $\hat{\rho}_i(\lambda_i, z_1; \tau)$ for $z_1 = 0$.

Finally, we derive an upper bound on $\bar{\mathcal{R}}(\Theta[s_n])$ by using asymptotic inequalities $\sup_{\lambda > 0} \rho_i(\lambda) \leq (\mathcal{C}_i + o(1)) \log \eta_n^{-1}$ and $\rho_i(0) = O(\eta_n)$ for $i = 1, \dots, n$. Let subscripts $[1], \dots, [n]$ be the permutation of $1, \dots, n$ that satisfies $\mathcal{C}_{[1]} \geq \dots \geq \mathcal{C}_{[n]}$. The minimax risk $\bar{\mathcal{R}}(\Theta[s_n])$ is bounded as

$$\bar{\mathcal{R}}(\Theta[s_n]) \leq \sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi^*}) = \sum_{i=1}^{s_n} \sup_{\lambda > 0} \rho_{[i]}(\lambda) + \sum_{i=s_n+1}^n \rho_{[i]}(0). \quad (28)$$

By using the asymptotic inequalities

$$\sup_{\lambda_i > 0} \rho_i(\lambda_i) \leq (\mathcal{C}_i + o(1)) \log \eta_n^{-1} \text{ and } \rho_i(0) = O(\eta_n) \text{ (} i = 1, \dots, n \text{)}$$

and by using Condition 6.1, we have

$$\begin{aligned}\bar{\mathcal{R}}(\Theta[s_n]) &\leq \sum_{i=1}^{s_n} \{ \mathcal{C}_{[i]} + o(1) \} \log \eta_n^{-1} + (n - s_n) O(\eta_n) \\ &\sim \frac{s_n}{n} \sum_{i=1}^n \mathcal{C}_i \{ \log \eta_n^{-1} \} (1 + o(1)) \sim \bar{\mathcal{C}}_{s_n} \log \eta_n^{-1}.\end{aligned}$$

Therefore the desired upper bound on $\bar{\mathcal{R}}(\Theta[s_n])$ is obtained. \square

6.3.3. Proof of Lemma 6.2. The proof is two-fold: deriving the minimax risk with r_i s conditioned, and then taking expectation of that with respect to G . Recall that $\bar{\mathcal{C}} := \sum_{i=1}^n \mathcal{C}_i / n$. The former part of the proof essentially is completed by Lemma 6.1.

Consider the latter part of the proof. Take the expectation of \mathcal{C}_1 with respect to G . Since $\mathcal{C}_1 = \mathcal{C}_{r_1}$ is included in $(0, 1)$ for $r_1 \in (0, \infty)$, it follows from Hoeffding's inequality that we have

$$\Pr(|\bar{\mathcal{C}} - \mathbb{E}_G \mathcal{C}_1| \geq t) \leq 2 \exp(-2nt^2), \quad t > 0.$$

Therefore, for any $\delta \in (0, 1)$, the inequality

$$|\bar{\mathcal{C}} - \mathbb{E}_G \mathcal{C}_1| \leq \sqrt{\{1/(2n)\} \log(2/\delta)} \quad (29)$$

holds with probability higher than $1 - \delta$. In combining (29) with the result of the former part, we have to care about the behaviors of C_1 and C_2 in (27). For possible values of these constants, see the sentence just before (27). Condition 3.1 assures that the residual

$$\bar{\mathcal{R}}(\Theta[s_n]) - \bar{\mathcal{C}} s_n \log(\eta_n^{-1})$$

is $O_P(1)$. Thus, combining (29) with the result of the former part yields the inequality

$$\begin{aligned} \left(\mathbb{E}_G \mathcal{C}_1 - \sqrt{\{1/(2n)\} \log(1/\delta)} + b_n \right) s_n \log \eta_n^{-1} &\leq \bar{\mathcal{R}}(\Theta[s_n]) \\ &\leq \left(\mathbb{E}_G \mathcal{C}_1 + \sqrt{\{1/(2n)\} \log(2/\delta)} + b_n \right) s_n \log \eta_n^{-1} \end{aligned}$$

with probability higher than $1 - \delta$, where b_n is an $o_p(1)$ term that is independent of δ . This completes the proof. \square

6.3.4. Proof of Lemma 6.3. The proof follows essentially the same steps as those in Subsection 5.4. We provide the proof only for $\Theta[s_n]$. Following the same argument as that of Theorem 6.1, it is shown that for $i = 1, 2, \dots, n$, we have

$$\begin{aligned} \hat{\theta}_{\Pi[s_n, \kappa], i}(z_i; \tau) &= (\eta_n / t_i(\tau)^{\kappa+1}) / (1 + \eta_n / t_i(\tau)^{\kappa}) & z_i &= 0; \\ \hat{\theta}_{\Pi[s_n, \kappa], i}(z_i; \tau) &= (z_i + \kappa) / t_i(\tau) & z_i &\geq 1. \end{aligned}$$

Using the above equalities for $\hat{\theta}_{\Pi[s_n, \kappa]}$, there exists a positive constant C_1 depending only on κ and $\sup_i r_i$ for which we have

$$\begin{aligned} \sup_{\lambda > 0} \mathbb{E}_\theta \log[\{\exp(-\lambda) \lambda^{Y_1} / Y_1!\} / \{q_{\Pi[s_n, \kappa], 1}(Y_1 | X_1)\}] &\leq \bar{\mathcal{C}} \log \eta_n^{-1} + C_1, \\ \lim_{\lambda \rightarrow 0} \mathbb{E}_\theta \log[\{\exp(-\lambda) \lambda^{Y_1} / Y_1!\} / \{q_{\Pi[s_n, \kappa], 1}(Y_1 | X_1)\}] &= O(\eta_n). \end{aligned}$$

Thus, we have $\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[s_n, \kappa]}) \leq (\bar{\mathcal{C}} + o(1)) s_n \log \eta_n^{-1} + (n - s_n) O(\eta_n)$, which completes the proof. \square

6.4. Proof of Proposition 3.1. The proof is a combination of the following lemma and the same argument as in Subsection 6.1:

Lemma 6.5. *Fix an infinite sequence $\{r_i \in (0, \infty) : i \in \mathbb{N}\}$ such that $0 < \inf_i r_i \leq \sup_i r_i < \infty$. Suppose that $\sum_{i=1}^n 1/(nr_i) \sim \sum_{i \in J} 1/(s_n r_i)$ for any subset $J \subset \{1, \dots, n\}$ with $|J| = s_n$. Then the asymptotic equalities*

$$\bar{\mathcal{E}}(\Theta[s_n]) \sim \bar{\mathcal{E}}(\Theta[s_n, \varepsilon_n]) \sim e^{-1} \sum_{i=1}^n (r_i^{-1}/n) s_n \log(n/s_n)$$

hold as $n \rightarrow \infty$ and $s_n/n \rightarrow 0$, where ε_n is any shrinking sequence such that $\varepsilon_n = o(s_n/n)$.

Once we obtain Lemma 6.5, the remaining part of the proof of Proposition 3.1 is easy and thus we complete the proof. \square

Proof of Lemma 6.5: In the following, we prove Lemma 6.5. We provide the proof only for $\Theta[s_n]$. The proof follows almost the same line of the proof of Lemma 6.1. In bounding $\bar{\mathcal{E}}(\Theta[s_n])$ below, the Bayes risk based on the varied-spike block-independent prior $\Pi_{VB,\nu}$ with $\nu = (1/r_1, \dots, 1/r_n) \in \mathbb{R}_+^n$ is employed. In bounding $\bar{\mathcal{E}}(\Theta[s_n])$ above, a threshold estimator is used. The threshold estimator used in this proof is the same as in Subsection 5.5.1: for each $i = 1, \dots, n$, $\hat{\theta}_{T,i}(x_i) = \eta_n$ if $x_i = 0$ and $\hat{\theta}_{T,i}(x_i) = (x_i + 1)/t_i$ if $x_i \geq 1$.

Consider a lower bound on $\bar{\mathcal{E}}(\Theta[s_n])$. First, decompose the Bayes risk based on $\Pi_{VB,\nu}$ into a block-wise Bayes risk as $\int R(\theta, q(\cdot | \hat{\theta}_{\Pi_{VB,\nu}})) d\Pi_{VB,\nu}(\theta) = \sum_{j=1}^{s_n} B^{(j)}(\nu)$, where

$$B^{(j)}(\nu) = \int R(\theta^{(j)}, q(\cdot | \hat{\theta}_{\Pi_{VB,\nu}}^{(j)})) d\Pi_{VB,\nu}(\theta), \quad j = 1, \dots, s_n.$$

For the evaluation of $B^{(j)}(\nu)$, it follows from the explicit expression of $\hat{\theta}_{\Pi_{VB,\nu}}^{(j)}$ with $\nu = (1/r_i)_{i=1}^n$, that we have

$$\begin{aligned} B^{(j)}(\nu) &= \frac{1}{m_n} \sum_{\theta \in \text{Supp}(\Pi_{VB,\nu})} \sum_{k=1}^{m_n} \mathbb{E}_{\theta} \left\{ \theta_k^{(j)} \log(\theta_k^{(j)} / \hat{\theta}_{\Pi_{VB,\nu}}^{(j)}) - \theta_k^{(j)} + \hat{\theta}_{\Pi_{VB,\nu}}^{(j)} \right\} \\ &\geq \frac{1}{m_n} \sum_{k=1}^{m_n} (1/r_k^{(j)}) e^{-1} \log m_n - \sum_{k=1}^{m_n} (1/r_k^{(j)}) e^{-1} \\ &\sim \frac{s_n}{n} \sum_{k=1}^{m_n} (1/r_k^{(j)}) e^{-1} \log \eta_n^{-1}. \end{aligned}$$

Therefore, we have

$$\int R(\theta, q(\cdot | \hat{\theta}_{\Pi_{VB,\nu}})) d\Pi_{VB,\nu}(\theta) \geq \left(\frac{1}{n} \sum_{i=1}^n r_i^{-1} e^{-1} + o(1) \right) s_n \log \eta_n^{-1}.$$

Thus the desired lower bound on $\bar{\mathcal{E}}(\Theta[s_n])$ is obtained.

Consider an upper bound on $\bar{\mathcal{E}}(\Theta[s_n])$. Let

$$\tilde{\rho}_i(\lambda_i) := \mathbb{E}_{r_i \lambda_i} [\lambda_i \log(\lambda_i / \hat{\theta}_{T,i}(X_i)) - \lambda_i + \hat{\theta}_{T,i}(X_i)]$$

be a coordinate-wise estimative risk of $\hat{\theta}_{T,i}$. Following the same steps as in Subsection 5.5.1, the asymptotic inequalities $\sup_{\lambda > 0} \tilde{\rho}_i(\lambda) = (1 + o(1)) e^{-1} r_i^{-1} s_n \log \eta_n^{-1}$ and $\tilde{\rho}_i(0) = O(\eta_n)$ hold for $i = 1, \dots, n$. Let subscripts $[1], \dots, [n]$ be the permutation of $1, \dots, n$ that satisfies $1/r_{[1]} \geq \dots \geq 1/r_{[n]}$. Using the condition $\sum_{i=1}^n 1/(nr_i) \sim \sum_{i \in J} 1/(s_n r_i)$ for any subset $J \subset \{1, \dots, n\}$ with $|J| = s_n$, the minimax risk $\bar{\mathcal{E}}(\Theta[s_n])$ is bounded above as follows:

$$\bar{\mathcal{E}}(\Theta[s_n]) \leq \sum_{i=1}^{s_n} \{r_{[i]}^{-1} e^{-1} + o(1)\} \log \eta_n^{-1} + (n - s_n) O(\eta_n) \sim \frac{s_n}{n} \sum_{i=1}^n r_i^{-1} e^{-1} \{\log \eta_n^{-1}\} (1 + o(1)).$$

Therefore the desired upper bound on $\bar{\mathcal{E}}(\Theta[s_n])$ is obtained. \square

APPENDIX A. SUPPLEMENTAL EXPERIMENTS

This appendix presents supplemental experiments for readers' better understanding.

A.1. Quasi-sparsity. This subsection provides simulation studies for quasi-sparsity. Parameter θ and observations X and Y are drawn from

$$\begin{aligned} \theta_i &\sim \nu_i e_{S,i} + \xi_i e_{S^c,i} \quad (i = 1, \dots, n), \\ X \mid \theta &\sim \otimes_{i=1}^n \text{Po}(r\theta_i), Y \mid \theta \sim \otimes_{i=1}^n \text{Po}(\theta_i), \text{ and } X \perp\!\!\!\perp Y \mid \theta, \end{aligned}$$

respectively, where

- ν_1, \dots, ν_n are independent samples from the gamma distribution with a shape parameter 10 and a scale parameter 1;
- ξ_1, \dots, ξ_n are independent samples from the uniform distribution on $[0, 0.01]$;
- S is drawn from the uniform distribution on all subsets having exactly s and S^c is its complement;
- ν_1, \dots, ν_n and S are independent.

Here for a subset $J \subset \{1, \dots, n\}$, e_J indicates the vector of which the i -th component is 1 if $i \in J$ and 0 if otherwise. We examine two cases $(n, s, r) = (200, 5, 20)$ and $(n, s, r) = (200, 20, 20)$, and generate 500 current observations (X 's) and 500 future observations (Y 's).

TABLE 6. Comparison of predictive densities without MCAR, with $(n, s, r) = (200, 5, 20)$, and with quasi-sparsity: for each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance. The same abbreviations as in Table 1 are used.

	$\Pi[\hat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K04	$\ell_1 \ (\lambda = 0.1)$
ℓ_1 distance	<u>14.2</u> (4.0)	15.5(4.7)	16.6(4.5)	18.1(1.7)	63.3(8.3)	17.6(4.9)
PLL	-22.4(0.1)	-22.3(8.7)	-25.0(10.1)	<u>-20.8</u> (4.3)	-43.8(4.9)	-Inf
70%CP (%)	<u>69.1</u> (4.9)	<u>69.1</u> (4.6)	<u>70.9</u> (4.6)	1.8(0.8)	18.6(10.4)	61.9(6.0)
90%CP (%)	89.2(3.6)	89.4(2.9)	<u>90.2</u> (2.9)	45.6(18.9)	43.5(15.0)	85.5(4.2)

TABLE 7. Comparison of predictive densities without MCAR, with $(n, s, r) = (200, 20, 20)$, and with quasi-sparsity: for each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance.

	$\Pi[\hat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K04	$\ell_1 \ (\lambda = 0.1)$
ℓ_1 distance	<u>49.8</u> (8.4)	51.9(8.5)	54.0(8.8)	50.3(3.0)	206(15)	57.7(9.1)
PLL	<u>-55.0</u> (8.2)	-58.4(8.1)	-60.9(9.2)	-56.7(4.5)	-121(8.8)	-Inf
70%CP (%)	68.5(4.6)	68.5(4.6)	<u>70.0</u> (4.4)	0.0(0.0)	0.0(0.0)	57.5(5.8)
90%CP (%)	88.8(2.8)	88.7(3.2)	<u>89.6</u> (2.7)	50.3(3.1)	0.0(0.0)	82.8(4.3)

Tables 6 and 7 show that the performance of the proposed predictive densities do not depend on whether a parameter is exact sparse or quasi-sparse.

A.2. **Effect of s .** This subsection provides simulation studies highlighting the effect of s . The set-up except for s is the same as that in Subsection 4.1.1.

TABLE 8. Comparison of predictive densities without MCAR, with $(n, s, r) = (200, 50, 20)$, and with exact sparsity: for each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance.

	$\Pi[\widehat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K04	$\ell_1 (\lambda = 0.1)$
ℓ_1 distance	112.4(13.2)	<u>109.0</u> (12.6)	111.6(12.8)	136.3(4.76)	117.1(12.4)	118.0(13.0)
PLL	-112.9(4.8)	<u>-109.6</u> (4.8)	-110.6(4.7)	-115.2(4.6)	-117.8(4.3)	-Inf
70%CP (%)	<u>69.2</u> (4.3)	68.2(4.7)	68.5(4.5)	75.8(2.9)	82.8(4.2)	51.9(6.9)
90%CP (%)	<u>89.3</u> (2.8)	88.9(2.8)	88.8(3.0)	92.1(1.4)	95.4(2.4)	79.4(5.4)

TABLE 9. Comparison of predictive densities without MCAR, with $(n, s, r) = (200, 100, 20)$, and with exact sparsity: for each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance.

	$\Pi[\widehat{s}_n, 0.1]$	$\Pi[1, 0.1]$	$\Pi[1, 0.5]$	GH	K04	$\ell_1 (\lambda = 0.1)$
ℓ_1 distance	204.5(17.5)	204.1(17.7)	203.9(17.4)	<u>190.3</u> (6.0)	209.6(12.4)	221.3(17.8)
PLL	-201.9(6.1)	-201.5(6.1)	-201.5(6.0)	<u>-191.9</u> (1.9)	-207.4(6.1)	-Inf
70%CP (%)	<u>68.6</u> (4.3)	67.8(4.4)	67.9(4.3)	100.0(0.0)	76.5(3.9)	43.2(5.8)
90%CP (%)	<u>89.1</u> (3.1)	88.5(3.1)	88.7(2.9)	100.0(0.0)	92.9(2.2)	72.1(5.6)

Tables 8 and 9 find that the Bayes predictive density based on a Gauss Hypergeometric prior work better as the sparsity s_n is relatively large.

APPENDIX B. FORMULA FOR THE KULLBACK–LEIBLER RISK

In this appendix, Lemma 5.1 is proved for the sake of completeness. Let Π be a prior of θ and suppose that the Bayes estimate $\widehat{\theta}_\Pi(x; t)$ based on Π is strictly larger than 0 for any $x \in \mathbb{N}^n$ and any $t \in (r, r + 1)$.

Consider the following decomposition of the Kullback–Leibler risk:

$$R(\theta, q_\Pi) = \mathbb{E}_\theta[\log\{s(Y, X | \theta)/s_\Pi(Y, X)\}] - \mathbb{E}_\theta[\log\{p(X | \theta)/p_\Pi(X)\}],$$

where $s(y, x | \theta) = p(x | \theta)q(y | \theta)$, $s_\Pi(y, x) := \int s(y, x | \theta)d\Pi(\theta)$, and $p_\Pi(x) := \int p(x | \theta)d\Pi(\theta)$. For $z \in \mathbb{N}^n$ and $t \in (r, r + 1)$, let $p(z | \theta; t) := \prod_{i=1}^n e^{-t\theta_i} (t\theta_i)^{z_i-1}/z_i!$ and let $p_\Pi(z; t) := \int p(z | \theta; t)\Pi(d\theta)$. It follows from the sufficiency reduction that

$$\mathbb{E}_\theta[\log\{s(Y, X | \theta)/s_\Pi(Y, X)\}] = \mathbb{E}_\theta[\log\{p(X + Y | \theta; r + 1)/p_\Pi(X + Y; r + 1)\}].$$

Introducing the random variable Z_t from $\otimes_{i=1}^n \text{Po}(t\theta_i)$ ($t \in (r, r + 1)$), $R(\theta, q_\Pi)$ is expressed as

$$R(\theta, q_\Pi) = \int_r^{r+1} \frac{d}{dt} \mathbb{E}[\log\{p(Z_t | \theta; t)/p_\Pi(Z_t; t)\}] dt$$

If the equality

$$(d/dt)\mathbb{E}[\log\{p(Z_t | \theta; t)/p_\Pi(Z_t; t)\}] = R_e(t\theta, t\hat{\theta}_\Pi(\cdot; t))/t, \quad t \in (r, r+1) \quad (30)$$

is derived, the proof is completed. Differentiating $\mathbb{E}[\log\{p(Z_t | \theta; t)/p_\Pi(Z_t; t)\}]$ with respect to t yields

$$\begin{aligned} \mathbb{E}[\log\{p(Z_t | \theta; t)/p_\Pi(Z_t; t)\}] &= \mathbb{E}[\{d \log p(Z_t | \theta; t)/dt\} \log\{p(Z_t | \theta; t)/p_\Pi(Z_t; t)\}] \\ &\quad + \mathbb{E}[d \log p(Z_t | \theta; t)/dt] - \mathbb{E}[d \log p_\Pi(Z_t; t)/dt]. \end{aligned} \quad (31)$$

For the first term in the right hand side of equality (31), it follows that

$$\begin{aligned} &\mathbb{E}[\{d \log p(Z_t | \theta; t)/dt\} \log\{p(Z_t | \theta; t)/p_\Pi(Z_t; t)\}] \\ &= \mathbb{E} \left[\left\{ \sum_{i=1}^n \frac{Z_{t,i} - 1 - t\theta_i}{t} \right\} \log\{p(Z_t | \theta; t)/p_\Pi(Z_t; t)\} \right] \\ &= \mathbb{E} \sum_{i=1}^n \theta_i [\log\{p(Z_t + e_i | \theta; t)/p(Z_t | \theta; t)\} - \log\{p_\Pi(Z_t + e_i; t)/p_\Pi(Z_t; t)\}] \\ &= \mathbb{E} \sum_{i=1}^n \theta_i \log\{\theta_i/\hat{\theta}_{\Pi,i}(Z_t; t)\} \end{aligned}$$

from Hudson's lemma ($\mathbb{E}[\sum_{i=1}^n (Z_{t,i} - 1)f(Z_t)] = \mathbb{E}[\sum_{i=1}^n t\theta_i f(Z_t + e_i)]$ for any function $f: \mathbb{N}^n \rightarrow \mathbb{R}$) and from the simple fact that $p_\Pi(Z_t + e_i; t)/p_\Pi(Z_t; t) = \hat{\theta}_{\Pi,i}(Z_t; t)$, where e_i is a unit length vector in the i -th coordinate direction ($i = 1, \dots, n$). Similarly, for the second and third terms in the right hand side of equality (31), it follows that

$$\begin{aligned} \mathbb{E} \left[\frac{d}{dt} \{\log p(Z_t | \theta; t)\} \right] &= \mathbb{E} \left[\sum_{i=1}^n \frac{Z_{t,i} - 1 - t\theta_i}{t} \right] \text{ and} \\ \mathbb{E} \left[\frac{d}{dt} \{\log p_\Pi(Z_t; t)\} \right] &= \mathbb{E} \left[- \sum_{i=1}^n \frac{\hat{\theta}_{\Pi,i}(Z_t; t) - Z_{t,i} + 1}{t} \right] \end{aligned}$$

from the identity $(d/dt) \log p_\Pi(x; t) = -\sum_{i=1}^n \{\hat{\theta}_{\Pi,i}(x; t) - x_i + 1\}$. Thus, we obtain (30), which completes the proof. \square

APPENDIX C. THE NUMBERS OF INDIVIDUALS IN SUBSECTION 4.2.2

This appendix provides the histogram of the double numbers (r_i s) of individuals sequenced at each genetic position.

Figure 7 shows that r_i s are varied but concentrated around the certain value. The summary statistics are as follows: the mean is 113783.36; the median is 119236; the standard deviation is 15888.35; the skewness is -4.02; the kurtosis is 17.42; r_i s are bounded above by 121412. r_i s are bounded below by 12752.

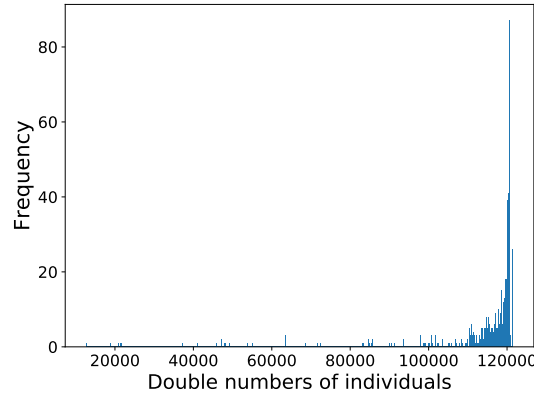


FIGURE 7. The histogram of the double numbers of individuals; the vertical line denotes the counts of genetic positions and the horizontal line denotes the double numbers of individuals sequenced at each genetic position.

REFERENCES

- [1] D. Agarwal, D. Gelfand, and S. Citron-Pousty. Zero-inflated model with application to spatial count data. *Environ. and Ecolog. Statist.*, 9:341–355, 2002.
- [2] J. Aitchison. Goodness of prediction fit. *Biometrika*, 62:547–554, 1975.
- [3] H. Akaike. A new look at Bayes procedure. *Biometrika*, 65:53–59, 1978.
- [4] M. Aslan. Asymptotically minimax Bayes predictive densities. *Ann. Statist.*, 34:2921–2938, 2006.
- [5] A. Boissunon and Y. Maruyama. Inadmissibility of the best equivariant predictive density in the unknown variance case. *Biometrika*, 101:733–740, 2014.
- [6] L. Clevenson and J. Zidek. Simultaneous estimation of the means of independent poisson laws. *J. American Statist. Assoc.*, 70:698–705, 1975.
- [7] J. Corcuera and F. Giummolè. A generalized Bayes rule for prediction. *Scand. J. Statist.*, 26:266–279, 1999.
- [8] J. Datta and D. Dunson. Bayesian inference on quasi-sparse count data. *Biometrika*, 103:971–983, 2016.
- [9] P. Dawid and M. Musio. Bayesian model selection based on proper scoring rules. *Bayesian Anal.*, 10:479–499, 2015.
- [10] C.-A. Deledalle. Estimation of Kullback–Leibler losses for noisy recovery problems within the exponential family. *Electron. J. Statist.*, 11:3141–3164, 2017.
- [11] The Metropolitan Police Department. The number of crimes in tokyo prefecture by town and type. http://www.keishicho.metro.tokyo.jp/about_mpd/jokyo_tokei/jokyo/ninchikensu.html.
- [12] I. Dunsmore. Asymptotic prediction analysis. *Biometrika*, 63:627–630, 1975.
- [13] S. Efromovich. Nonparametric regression with predictors missing at random. *J. American Statist. Assoc.*, 106:306–319, 2011.
- [14] S. Efromovich. Adaptive nonparametric density estimation with missing observations. *J. Statist. Plann. Infer.*, 143:637–650, 2013.
- [15] D. Fourdrinier, É. Marchand, A. Righi, and W. Strawderman. On improved predictive density estimation with parametric constraints. *Electron. J. Statist.*, 5:172–191, 2011.
- [16] E. George, F. Liang, and X. Xu. Improved minimax predictive densities under Kullback–Leibler loss. *Ann. Statist.*, 34:78–91, 2006.
- [17] M. Ghosh and M.-C. Yang. Simultaneous estimation of poisson means under entropy loss. *Ann. Statist.*, 16:278–291, 1988.

- [18] D. Hall. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56:1030–1039, 2002.
- [19] J. Hartigan. The maximum likelihood prior. *Ann. Statist.*, 26:2083–2103, 1998.
- [20] J. Hartigan. Bayesian regression using Akaike priors. New Haven, CT, Yale University, Preprint, 2002.
- [21] I. Johnstone. Admissibility, difference equations and recurrence in estimating a Poisson mean. *Ann. Statist.*, 12:1173–1198, 1984.
- [22] I. Johnstone and B. MacGibbon. Minimax estimation of a constrained Poisson vector. *Ann. Statist.*, 20:807–831, 1992.
- [23] K. Kato. Improved prediction for a multivariate normal distribution with unknown mean and variance. *Ann. Ins. Statist. Math.*, 61:531–542, 2009.
- [24] K. Kobayashi and F. Komaki. Bayesian shrinkage prediction for the regression problem. *J. Multivariate Anal.*, 99:1888–1905, 2008.
- [25] F. Komaki. On asymptotic properties of predictive distributions. *Biometrika*, 83:299–313, 1996.
- [26] F. Komaki. A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, 88:859–864, 2001.
- [27] F. Komaki. Simultaneous prediction of independent Poisson observables. *Ann. Statist.*, 32:1744–1769, 2004.
- [28] F. Komaki. A class of proper priors for Bayesian simultaneous prediction of independent Poisson observables. *J. Multivariate Anal.*, 97:1815–1823, 2006.
- [29] F. Komaki. Shrinkage priors for Bayesian prediction. *Ann. Statist.*, 34:808–819, 2006.
- [30] F. Komaki. Bayesian predictive densities based on latent information priors. *J. Statist. Plan. Infer.*, 12:3705–3715, 2011.
- [31] F. Komaki. Simultaneous prediction for independent Poisson processes with different durations. *J. Multivariate Anal.*, 141:35–48, 2015.
- [32] T. Kubokawa, É. Marchand, W. Strawderman, and J.-P. Turcotte. Minimaxity in predictive density estimation with parametric constraints. *J. Multivariate Anal.*, 116:382–397, 2013.
- [33] D. Lambert. Zero-inflated Poisson regression, with an application to random defects in manufacturing. *Technometrics*, 34:1–14, 1992.
- [34] G. Leung and A. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. on Information theory*, 52:3396–3410, 2006.
- [35] F. Liang and A. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. on Information theory*, 50:2708–2726, 2004.
- [36] A. L’Moudden, É. Marchand, O. Kortbi, and W. Strawderman. On predictive density estimation for Gamma models with parametric constraints. *J. Statist. Plan. Infer.*, 185:56–68, 2017.
- [37] B. MacGibbon. Minimax estimation over hyperrectangles with implications in the Poisson case. *IMS Collections*, 6:32–42, 2010.
- [38] Y. Maruyama and W. Strawderman. Bayesian predictive densities for linear regression models under α -divergence loss: Some results and open problems. *IMS Collections*, 8:42–56, 2012.
- [39] T. Matsuda and F. Komaki. Singular value shrinkage priors for Bayesian prediction. *Biometrika*, 102:843–854, 2015.
- [40] G. Mukherjee and I. Johnstone. On minimax optimality of sparse Bayes predictive density estimates. arXiv:1707.04380.
- [41] G. Mukherjee and I. Johnstone. Exact minimax estimation of the predictive density in sparse gaussian models. *Ann. Statist.*, 43:937–961, 2015.
- [42] G. Murray. A note on the estimation of probability density functions. *Biometrika*, 64:150–152, 1977.
- [43] V. Ng. On the estimation of parametric density function. *Biometrika*, 67:505–506, 1980.
- [44] T. Suzuki and F. Komaki. On prior selection and covariate shift of β -Bayesian prediction under α -divergence risk. *Comm. Statist. Theory and Methods*, 39:1655–1673, 2010.
- [45] P. Vidoni. A simple predictive density based on the p^* -formula. *Biometrika*, 82:855–863, 1995.

- [46] X. Xu and F. Liang. Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli*, 16:543–560, 2010.
- [47] X. Xu and D. Zhou. Empirical Bayes predictive densities for high-dimensional normal models. *J. Multivariate Anal.*, 102:1417–1428, 2011.
- [48] K. Yano and F. Komaki. Asymptotically minimax prediction in infinite sequence models. *Electron. J. Statist.*, 11:3165–3195, 2017.
- [49] F. Zhang, Y. Shi, H. Ng, and R. Wang. Information geometry of generalized Bayesian prediction using α -divergence as loss functions. *IEEE Tran. on Information Theory*, 64:1812–1824, 2018.

(K. Yano) DEPARTMENT OF MATHEMATICAL INFORMATICS, GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY, UNIVERSITY OF TOKYO, 7-3-1 HONGO, BUNKYO-KU, TOKYO 113-0033, JAPAN.

E-mail address: yano@mist.i.u-tokyo.ac.jp

(R. Kaneko) DEPARTMENT OF MATHEMATICAL INFORMATICS, GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY, UNIVERSITY OF TOKYO, 7-3-1 HONGO, BUNKYO-KU, TOKYO 113-0033, JAPAN.

E-mail address: ryoya_kaneko@mist.i.u-tokyo.ac.jp

(F. Komaki) DEPARTMENT OF MATHEMATICAL INFORMATICS, GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY, UNIVERSITY OF TOKYO, 7-3-1 HONGO, BUNKYO-KU, TOKYO 113-0033, JAPAN.

E-mail address: komaki@mist.i.u-tokyo.ac.jp