Bernoulli **27**(2), 2021, 1212–1238
https://doi.org/10.3150/20-BEJ1271

Minimax predictive density for sparse count data

KEISUKE YANO[1], RYOYA KANEKO[2] and FUMIYASU KOMAKI[3,4]

[1]*The Institute of Statistical Mathematics, 10-3 Midori cho, Tachikawa City, Tokyo, 190-8562, Japan. E-mail: yano@ism.ac.jp*
[2]*Tokyo Marine Holdings, Inc., 1-2-1 Marunouchi, Chiyoda-ku, Tokyo, 100-8050, Japan. E-mail: ryykaneko@gmail.com*
[3]*Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan. E-mail: komaki@g.ecc.u-tokyo.ac.jp*
[4]*RIKEN Center for Brain Science, 2-1 Hirosawa, Wako City, Saitama, 351-0198, Japan*

This paper discusses predictive densities under the Kullback–Leibler loss for high-dimensional Poisson sequence models under sparsity constraints. Sparsity in count data implies zero-inflation. We present a class of Bayes predictive densities that attain asymptotic minimaxity in sparse Poisson sequence models. We also show that our class with an estimator of unknown sparsity level plugged-in is adaptive in the asymptotically minimax sense. For application, we extend our results to settings with quasi-sparsity and with missing-completely-at-random observations. The simulation studies as well as application to real data illustrate the efficiency of the proposed Bayes predictive densities.

Keywords: Adaptation; high dimension; Kullback–Leibler divergence; missing at random; Poisson model; zero inflation

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 their uncertainty. It has a wide range of application in statistics, information theory, and machine learning. The simplest class of predictive densities is the 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 the class of *Bayes* predictive densities. A Bayes predictive density is the posterior mixture of densities of future observations. There is a vast literature on predictive density for statistical models in finite dimensions; see Section 1.2 for the literature review. Conversely, little is known about predictive density for statistical models in high dimensions. In prediction using sparse high-dimensional Gaussian models, [46,47] construct several predictive densities (including a Bayes predictive density) superior to all plug-in predictive densities.

The aim of this paper is to construct an efficient predictive density for high-dimensional sparse count data. The efficiency of a predictive density is measured by the supremum of the Kullback–Leibler risk under sparsity constraints. Sparsity in count data means that there exhibits an excess of zeros. See Section 1.1 for the formulation.

The motivation for analyzing sparse count data is well known. In analyzing high-dimensional count data, there often exhibits inflation of zeros. Data with an overabundance of zeros include samples from agriculture [22], environmental sciences [1], manufacturing [39], DNA sequencing [14], and terrorist attacks [14]. Another example (Japanese crime statistics) is presented in Section 4.

1.1. Problem setting and contributions

We summarize main results with the problem formulation ahead. Let X_i $(i = 1, 2, \dots, n)$ be a current observation independently distributed according to $\mathrm{Po}(r\theta_i)$, and let Y_i $(i = 1, 2, \dots, n)$ be a future observation independently distributed according to $\mathrm{Po}(\theta_i)$, where $\theta = (\theta_1, \dots, \theta_n)$ is an unknown parameter and r is a known constant. Constant r represents 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, this constant represents 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 $p(x \mid \theta)$ and $q(y \mid \theta)$, respectively:

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

Our target parameter space is the exact sparse parameter space which is defined as follows. Given $s \in (0, n)$, $\Theta[s] := \{\theta \in \mathbb{R}_+^n : \|\theta\|_0 \le s\}$, where $\|\cdot\|_0$ is the ℓ_0-norm given by $\|\theta\|_0 := \#\{i : \theta_i > 0\}$.

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

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

The corresponding risk (expected loss) is denoted by

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

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

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

To express high-dimensional settings under sparsity constraints, we employ the high dimensional asymptotics in which $n \to \infty$ and $\eta_n := s/n = s_n/n \to 0$. The value of s possibly depends on n and thus in what follows the dependence on n is often expressed, say, $s = s_n$.

Main theoretical contributions are summarized as follows:

(i) In Theorem 2.1, we identify the asymptotic minimax risk $\mathcal{R}(\Theta[s_n])$ and present a class of Bayes predictive densities attaining the asymptotic minimaxity;
(ii) In Theorem 2.2, we present an asymptotically minimax predictive density that is adaptive to an unknown sparsity.

In Theorem 2.1, we find that the sharp constant in the asymptotic minimax risk is controlled by the constant r. This constant highlights the interesting parallel between Gaussian and Poisson decision theories as discussed in Section 1.2. In Theorem 2.2, we show that a simple plug-in approach to choose the tuning parameter in the proposed class yields adaptive Bayes predictive densities. In addition, we obtain the corresponding results for quasi sparse Poisson models and for settings where current observations are missing completely at random in Section 3. These extensions are important in applications.

The 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 the sense of both predictive uncertainty quantification and point prediction.

The proposed class of predictive densities builds upon spike-and-slab prior distributions with improper slab priors. Interestingly, spike-and-slab prior distributions with slab priors having exponential tails do not yield asymptotically minimax predictive densities as Proposition 2.3 indicates. The proposed predictive densities are not only asymptotically minimax but also easily implemented by exact sampling.

1.2. Literature review

There is a rich literature on constructing predictive densities in fixed finite dimensions. Bayes predictive densities have been shown to dominate plug-in predictive densities in several instances. Studies of Bayes predictive densities date back to [2,3,48,49]. The first quantitative comparison of Bayes and plug-in predictive densities in a wide class of parametric models is [31]. [31] 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 [25] for asymptotic expansions of Bayes predictive densities. Minimax Bayes predictive densities for unconstrained parameter spaces are studied in [4,41]. Minimax predictive densities under parametric constraints are studied in [18,36,38]. Shrinkage priors for Bayes predictive densities under Gaussian models are investigated in [20,30,32,45]; see also [5,19,29] for the cases where the variances are unknown. Shrinkage priors for Bayes predictive densities under Poisson models are developed in [33,35]. The cases under α-divergence losses are covered by [13,37,43,52,57].

Relatively little is known about constructing predictive densities in high dimensions. [46,47] construct an asymptotically minimax predictive density for sparse Gaussian models. [53] obtained an asymptotically minimax predictive density for nonparametric Gaussian regression models under Sobolev constraints; thereafter, [56] obtained an adaptive minimax predictive density for these models. See also [54]. All above results employ Gaussian likelihood and the corresponding results for count data have been not known.

Poisson models deserve study in their own right as prototypical count data modeling ([6,10,27,33, 50]). Poisson models exhibit several correspondences to Gaussian models. [10,26,27,42] find the correspondence in estimation of means using the re-scaled squared loss defined as $\sum_{i=1}^{n} \theta_i^{-1}(\theta_i - \hat{\theta}_i(X))^2$. [21,33–35] find the correspondence in prediction using the Kullback–Leibler loss. In particular, [27, 42] find the correspondence in the asymptotic minimaxity under ellipsoidal and rectangle constraints in high-dimensional Poisson models using the re-scaled squared loss (the local Kullback–Leibler loss). In spite of the interesting correspondence in [27,42], 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 the results of asymptotic minimaxity in both estimation and prediction for sparse Poisson models, which are clearly parallel to the result for sparse Gaussian models by [47]; see Section 2.2 for detailed discussions. This paper also covers several new topics in predictive density under sparsity constraints: the adaptation to sparsity, quasi-sparsity, and missing completely at random.

Our strategy leverages spike-and-slab priors. In the literature, it is known that the choice of slab priors impacts on the statistical optimality [8,9,28,51]. But, the behavior has been studied only for (sub-)Gaussian models and the corresponding results for Poisson models have remained unavailable. In Proposition 2.3, we show that slab priors with tails as heavy as the exponential distribution suffer from the minimax sub-optimality. In Proposition 2.4, we also show that polynomially decaying slabs can attain the minimax optimality.

Relatively scarce are theoretical studies of zero-inflated or quasi zero-inflated Poisson models in high dimensions in spite of their importance. [14] constructs global-local shrinkage priors for high-

dimensional quasi zero-inflated Poisson models. The constructed priors have good theoretical properties of the shrinkage factors and of the multiple testing statistics. We confirm in Section 4 that our priors broadly outperform their priors in predictive density, which indicates our priors are more suitable for prediction. In contrast, we consider that their priors would be more suitable than our priors in multiple testing or in interpreting shrinkage factors. We shall also mention that in this direction, interesting and powerful extensions of [14] are now available in [23]. Appendix C in the supplementary material [55] provides the comparison of our predictive density to the Bayes predictive density based on the prior in [23].

1.3. Organization and notation

The rest of the paper is organized as follows. In Section 2, we present an asymptotically minimax predictive density and an adaptive minimax predictive density for sparse Poisson models, which is the main result in this paper. In Section 3, we present several extensions of the main result. In Section 4, we conduct simulation studies and present application to real data. In Section 5, we give proofs of main theorems (Theorems 2.1 and 2.2). In Section 6, we provide proofs of auxiliary lemmas used in Section 5. All proofs of propositions in Section 2 are given in Appendix A of the supplementary material. All proofs of propositions in Section 3 are given in Appendix B of the supplementary material.

Throughout the paper, we will use the following notations. The notation $a_n \sim b_n$ signifies that a_n/b_n converges to 1 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. The notation $o(a_n)$ indicates a term of which the absolute value divided by a_n goes to zero in n. For a function $f : \mathbb{N}^n \times \mathbb{N}^n \to \mathbb{R}$, the expectation $\mathbb{E}_\theta[f(X, Y)]$ indicates the expectation of $f(X, Y)$ with respect to $p(x \mid \theta)q(y \mid \theta)$. Likewise, for a function $g : \mathbb{N} \to \mathbb{R}$, the expectation $\mathbb{E}_\lambda[g(X_1)]$ indicates the expectation of $g(X_1)$ with respect to $\mathrm{Po}(\lambda)$. Constants $c_1, c_2, \ldots$ and $C_1, C_2, \ldots$ do not depend on n. Their values may be different at each appearance.

2. Predictive density for sparse Poisson models

2.1. Main results

This section presents main results for prediction using sparse Poisson models: the precise description of the asymptotic minimax risk; the construction of the class of asymptotically minimax predictive densities; and that of adaptive minimax predictive densities. Detailed discussions are provided in the subsequent subsection. Proofs of the theorems are presented in Section 5.

The first theorem describes the asymptotic minimax risk as well as the Bayes predictive density attaining the asymptotic minimaxity. For $r \in (0, \infty)$, let

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

For $h > 0$ and $\kappa > 0$, let $\Pi[h, \kappa]$ be an improper prior of the form

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

where δ_0 is the Dirac measure centered at 0.

Theorem 2.1. *Fix $r \in (0, \infty)$ and fix a sequence $s_n \in (0, n)$ such that $\eta_n = s_n/n = o(1)$. Then, the following holds*:

$$\mathcal{R}\big(\Theta[s_n]\big) \sim \mathcal{C} s_n \log\big(\eta_n^{-1}\big) \quad \text{as } n \to \infty.$$

Further, the predictive density $q_{\Pi[\eta_n,\kappa]}$ based on $\Pi[\eta_n, \kappa]$ with $\kappa > 0$ is asymptotically minimax: i.e.,

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[\eta_n,\kappa]}) \sim \mathcal{R}\big(\Theta[s_n]\big) \quad \text{as } n \to \infty.$$

The derivation of this theorem consists of (i) establishing a lower bound of $\mathcal{R}(\Theta[s_n])$ based on the Bayes risk maximization, and (ii) establishing an upper bound of it based on the Bayes predictive density $q_{\Pi[\eta_n,\kappa]}$.

The first theorem provides asymptotically minimax strategies, but the optimal strategies therein require the true value of η_n. The second theorem presents the optimal strategies without requiring the true value of s_n, that is, adaptive minimax predictive densities for sparse Poisson models. Let $\hat{s}_n := \max\{1, \#\{i : X_i \geq 1, i = 1, \ldots, n\}\}$, and let $\hat{\eta}_n := \hat{s}_n/n$.

Theorem 2.2. *Fix $r \in (0, \infty)$ and $\kappa > 0$. Then, the predictive density $q_{\Pi[\hat{\eta}_n,\kappa]}$ is adaptive in the asymptotically minimax sense on the class of exact sparse parameter spaces: i.e., for any sequence $s_n \in [1, n)$ such that $\sup_n s_n/n < 1$ and $\eta_n = s_n/n = o(1)$,*

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

The derivation builds upon evaluating the difference between two Kullback–Leibler risks $R(\theta, q_{\Pi[\eta_n,\kappa]})$ and $R(\theta, q_{\Pi[\hat{\eta}_n,\kappa]})$. We will show this difference is negligible uniformly in θ compared to the minimax risk. To check this, we use three properties of $\hat{s}_n$:

- The estimate $\hat{s}_n$ is bounded below by an absolute constant;
- The first and the second moments of $|\hat{s}_n/s_n - 1|$ are bounded above by an absolute constant;
- The estimate $\hat{s}_n$ can capture nearly the correct growth rate of s_n in the relatively dense regime, whenever the true value of θ is outside a vicinity of 0. Specifically, we will see that

$$\sup_{\theta : \max_i \theta_i > 1/\sqrt{\log s_n}} \mathbb{E}_\theta[\log s_n/\hat{s}_n] = O(\sqrt{\log s_n}) \quad \text{as } s_n \to \infty.$$

The first and the second properties make the difference negligible in the relatively sparse regime. The third property makes the difference negligible in the relatively dense regime. See Section 5.3 for the detail.

2.2. Discussions

Several discussions are provided in order.

2.2.1. *Prediction and estimation*, *Poisson and Gaussian*

Prediction and estimation: For comparison, let us consider estimating θ under the Kullback–Leibler risk $R_e(\theta, \hat{\theta}) := R(\theta, q(\cdot \mid \hat{\theta}))$ as in [16]. The minimax risk $\mathcal{E}(\Theta[s_n])$ for estimation is defined in such

a way that $\mathcal{E}(\Theta[s]) := \inf_{\hat{\theta}} \sup_{\theta \in \Theta[s]} R_{\mathrm{e}}(\theta, \hat{\theta})$. Since the minimax risk $\mathcal{E}(\Theta[s_n])$ for estimation can be viewed as the minimax risk for prediction when predictive densities are restricted to plug-in predictive densities, we always have $\mathcal{E}(\Theta[s_n]) \geq \mathcal{R}(\Theta[s_n])$.

The first proposition describes the asymptotic minimax risk for estimation. This proposition highlights a gap between $\mathcal{E}(\Theta[s_n])$ and $\mathcal{R}(\Theta[s_n])$. The second proposition indicates that the same data-dependent prior as in Theorem 2.2 yields an adaptive minimax estimator.

Proposition 2.1. *Fix $r \in (0, \infty)$ and fix a sequence $s_n \in (0, n)$ such that $\eta_n = s_n/n = o(1)$. Then, the following holds*:

$$\mathcal{E}\big(\Theta[s_n]\big) \sim \mathrm{e}^{-1} r^{-1} s_n \log\big(\eta_n^{-1}\big) \quad \text{as } n \to \infty.$$

Proposition 2.2. *Fix $r \in (0, \infty)$ and $\kappa > 0$. Then, the Bayes estimator $\hat{\theta}_{\Pi[\hat{\eta}_n, \kappa]}$ is adaptive in the asymptotically minimax sense on the class of exact sparse parameter spaces: for any sequence $s_n \in [1, n)$ such that $\sup_n s_n/n < 1$ and $\eta_n = s_n/n = o(1)$, we have*

$$\sup_{\theta \in \Theta[s_n]} R_{\mathrm{e}}(\theta, \hat{\theta}_{\Pi[\hat{\eta}_n, \kappa]}) \sim \mathcal{E}\big(\Theta[s_n]\big) \quad \text{as } n \to \infty.$$

According to Theorem 2.1 and Proposition 2.1, the rates (with respect to n) of minimax risks for estimation and for prediction are identical. But, the sharp constants of these minimax risks are different with r. The sharp constant of $\mathcal{R}(\Theta[s_n])$ (i.e., $\mathcal{C}$,) increases as r decreases but remains bounded above by 1, while that of $\mathcal{E}(\Theta[s_n])$ (, i.e., $\mathrm{e}^{-1} r^{-1}$) grows to infinity as r decreases. Further, $\mathcal{C} \sim \mathrm{e}^{-1} r^{-1}$ as r increases.

Poisson and Gaussian: [47] shows the asymptotic minimax risk for prediction using sparse Gaussian models is equal to $\{1/(1+r)\} s_n \log \eta_n^{-1}$ with r the ratio of sample sizes of current observations to those of future observations. Comparing our results with [47], we find interesting similarities between sparse Gaussian and sparse Poisson models. First, the rates with respect to n of these two problems are identical to $s_n \log \eta_n^{-1}$. Second, Figures 1 and 2 show the comparisons of the exact constants of minimax risks for sparse Poisson and Gaussian models. The vertical line indicates values of the risks and the 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 observation 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.

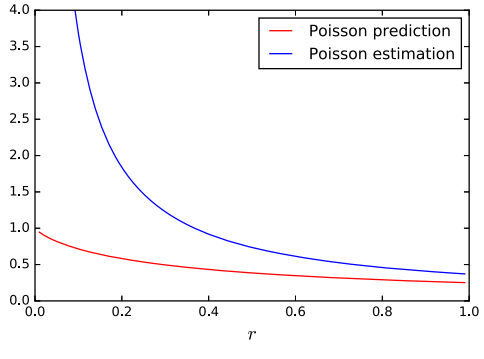

Figure 1. Predictive and estimative minimax risks for sparse Poisson models: the holizontal axis represents r.

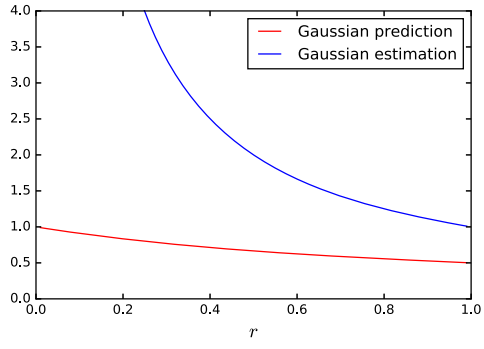

Figure 2. Predictive and estimative minimax risks for sparse Gaussian models: the holizontal axis represents r.

2.2.2. *Spike-and-slab priors*

Computation: Let us mention a computational advantage of using improper slab priors ahead. Bayes predictive densities often suffer from computational intractability because they may involve several numerical integrations. Using improper slab priors, we can avoid such a computational issue in our set-up. In fact, the Bayes predictive density based on $\Pi[h,\kappa]$ has the explicit form

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

where

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

The coordinate-wise marginal distribution of $q_{\Pi[h,\kappa]}$ is a zero-inflated negative binomial distribution and thus sampling from $q_{\Pi[h,\kappa]}$ is easy.

Condition on slab priors: Spike and slab priors has been recently well-investigated in sparse Gaussian models; see [8,9,28,51]. The existing results for sparse Gaussian models are summarized as follows:

- For point estimation, a slab prior with its tail at least as heavy as Laplace distribution yields a rate-optimal point estimator of the sparse mean (see [28]);
- For uncertainty quantification, Laplace slab prior yields a rate sub-optimal posterior ℓ^2-moment but Cauchy slab prior yields a rate-optimal one (see [8]).

Here we derive both necessary and sufficient conditions for the minimax prediction in sparse Poisson models.

Consider the following condition on a prior Π: For the posterior mean $\hat{\theta}_{\Pi}$, the asymptotic equation

$$\frac{|\hat{\theta}_{\Pi,i}(x) - x_i/r|}{x_i/r} \to 0 \quad \text{as } x_i \to \infty \tag{1}$$

holds for all i. This condition is a Poisson variant of tail-robustness of the posterior mean [7]. [23] studies this condition and a stronger condition in the context of the global-local shrinkage. The next proposition implies Condition (1) is necessary for the minimax prediction.

Proposition 2.3. *Fix a sequence $s_n \in (0, n)$ such that $\eta_n = s_n/n = o(1)$. If a spike-and-slab prior does not satisfy* (1)*, then we have*

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_\Pi)/\mathcal{R}\big(\Theta[s_n]\big) \to \infty \quad \text{as } n \to \infty.$$

Though Condition (1) is not sufficient, it becomes a useful criterion for checking the sub-optimality of a given predictive density. For example, a spike-and-slab prior with an exponentially decaying slab (e.g., Laplace slab) does not satisfy (1) and hence yields a sub-optimal predictive density.

Consider a spike-and-slab prior $\Pi_\gamma[\eta] := \prod_{i=1}^n \{(1-\eta)\delta_0(d\theta_i) + \eta\gamma(\theta_i)\, d\theta_i\}$ with a slab density γ. Make the following conditions on γ:

$$\sup_{\lambda > 0} \left| \lambda \frac{d}{d\lambda} \log \gamma(\lambda) \right| = \Lambda < \infty \tag{2}$$

and

$$\int_0^\infty \mathrm{e}^{-\lambda} \gamma(\lambda)\, d\lambda < \infty. \tag{3}$$

By the fundamental theorem of calculus, Condition (2) implies $c_1 \lambda^{-\Lambda} \le \gamma(\lambda) \le C_1 \lambda^{\Lambda}$ with some $c_1, C_1 > 0$. Condition (3) is a posterior integrability condition and is satisfied by all proper priors as well as ours. A class of slabs satisfying (2) and (3) includes half-Cauchy, Pareto, and regularly varying priors with index -1 discussed in [23,44]. We can easily check that spike-and-slab priors with slabs satisfying (2) and (3) meet Condition (1) and can show they yield the optimal predictive densities.

Proposition 2.4. *Fix $r \in (0, \infty)$ and fix a sequence $s_n \in (0, n)$ such that $\eta_n = s_n/n = o(1)$. Let $\Pi_\gamma[\eta_n]$ be a spike-and-slab prior with γ satisfying* (2) *and* (3) *Then, the predictive density $q_{\Pi_\gamma[\eta_n]}$ based on $\Pi_\gamma[\eta_n]$ is asymptotically minimax: that is,*

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi_\gamma[\eta_n]}) \sim \mathcal{R}\big(\Theta[s_n]\big) \quad \text{as } n \to \infty.$$

Optimal scaling: Our approach uses an improper prior within their mixture and therefore the scaling of the improper slab prior impacts on the resulting predictive density: $\Pi[L\eta_n, \kappa]$ for arbitrary $L > 0$ produces a different predictive density that is asymptotically minimax. This arbitrariness of the scale is well known in the objective Bayesian literature and worrisome in practice; see [15,24,40].

The next proposition provides a guideline for choosing L and removes this arbitrariness. Let

$$L^* := L^*_{r,\kappa} = \mathcal{C}/\mathcal{K} \quad \text{with } \mathcal{K} := \Gamma(\kappa+1)\frac{r^{-\kappa} - (r+1)^{-\kappa}}{\kappa}.$$

Proposition 2.5. *Fix $r \in (0, \infty)$ and $\kappa > 0$. Fix also a sequence $s_n \in (0, n)$ such that $\eta_n = s_n/n = o(1)$. Then, the predictive density $q_{\Pi[L\eta_n, \kappa]}$ with $L > 0$ and $\kappa > 0$ satisfies*

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[L\eta_n,\kappa]}) \le \mathcal{C} s_n \log\big(\eta_n^{-1}\big) - \mathcal{C} s_n \log L + \mathcal{K} s_n L + \Upsilon \tag{4}$$

with Υ terms that are independent of L or that are $O(s_n\eta_n)$, and L^* minimizes the right hand sides in (4) with respect to L.

This result shows that the scale of improper slab priors can be specified by the predictive setting (characterized by r). Our idea here is relevant to [24]: In [24], the scale of improper priors within their mixture is determined to yield log-posterior probabilities that coincide with log maximum likelihood plus an Akaike factor; see also the Appendix of [40]. In this light, [24] and this paper indicate that the specifications of the scale of improper priors within their mixture can be done from a predictive viewpoint.

3. Extensions

We present two extensions of our results: one is quasi-sparsity; the other is missing completely at random. These extensions are important in practice.

3.1. Quasi-sparsity

We introduce the notion of the quasi sparse parameter space. Given $s \in (0, n)$ and a threshold $\varepsilon > 0$, the quasi sparse parameter space is defined 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 value ε determines whether the parameter value of each coordinate is near-zero or not.

The next two propositions specifies the minimax risk over the quasi-sparse parameter space and presents an adaptive minimax predictive density.

Proposition 3.1. *Fix $r \in (0, \infty)$ and fix a sequence $s_n \in (0, \infty)$ such that $\eta_n = s_n/n = o(1)$. Fix also a shrinking sequence $\varepsilon_n > 0$ such that $\varepsilon_n = o(\eta_n)$. Then, for the quasi sparse parameter space $\Theta[s_n, \varepsilon_n]$, the following holds*:

$$\mathcal{R}\big(\Theta[s_n, \varepsilon_n]\big) := \inf_{\hat{q}} \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, \hat{q}) \sim \mathcal{C} s_n \log\big(\eta_n^{-1}\big) \quad \text{as } n \to \infty.$$

Further, the predictive density $q_{\Pi[\eta_n, \kappa]}$ based on $\Pi[\eta_n, \kappa]$ with $\kappa > 0$ is asymptotically minimax: *that is*,

$$\sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[\eta_n, \kappa]}) \sim \mathcal{R}\big(\Theta[s_n, \varepsilon_n]\big) \quad \text{as } n \to \infty.$$

Proposition 3.2. *Fix $r \in (0, \infty)$ and $\kappa > 0$. Then, the predictive density $q_{\Pi[\hat{\eta}_n, \kappa]}$ is adaptive in the asymptotically minimax sense on the class of quasi sparse parameter spaces*: *that is, for any sequence $s_n \in [1, n)$ such that $\sup_n s_n/n < 1$ and $\eta_n = s_n/n = o(1)$ and for any sequence $\varepsilon_n > 0$ such that $\varepsilon_n = o(\eta_n)$, we have*

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

3.2. Missing completely at random

We describe prediction using sparse Poisson models when the current observation is missing completely at random (MCAR). Let r_i's $(i = 1, 2, \ldots)$ be positive random variables. Given r_i $(i = 1, \ldots, n)$, let X_i $(i = 1, 2, \ldots, n)$ be a current observation independently distributed according to $\mathrm{Po}(r_i\theta_i)$, and let Y_i $(i = 1, 2, \ldots, n)$ be a future observation independently distributed according to $\mathrm{Po}(\theta_i)$, where θ_i $(i = 1, \ldots, n)$ is an unknown parameter. Suppose that $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ are independent. We denote by $R(\theta, \hat{q} \mid \{r_i\})$ the Kullback–Leibler risk conditioned on r_is. We also denote by $\overline{\mathcal{R}}(\Theta[s_n] \mid \{r_i\})$ the minimax Kullback–Leibler risk over $\Theta[s_n]$ conditioned on r_is.

To present mathematically unblemished results, we assume that r_is 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 sampling distribution G satisfies the following*: *(i)* $\mathbb{E}_G[r_1^2] < \infty$; *(ii)* $\mathbb{E}_G[r_1^{-2}] < \infty$.

Condition 3.1 (i) is usual. Condition 3.1 (ii) excludes any distribution G highly concentrated around 0 and is not stringent. Consider a longitudinal situation in which X_i $(i = 1, \ldots, n)$ is obtained as the sum of $\{X_{i,j} : j = 1, \ldots, r_i\}$, where r_i $(i = 1, \ldots, n)$ represents the sample size in the i-th coordinate, and for each i, $X_{i,j}$ $(j = 1, \ldots, r_i)$ follows $\mathrm{Po}(\theta_i)$. Condition 3.1 implies that for each coordinate there exists at least one observation: $r_i \geq 1$. Note that in our real data applications (Section 4.2), Condition 3.1 (ii) is satisfied.

The following propositions describe the asymptotic minimax risk and present an adaptive minimax predictive density. 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 $\overline{\mathcal{C}} := \overline{\mathcal{C}_n} = \sum_{i=1}^n \mathcal{C}_i / n$.

Proposition 3.3. *Fix a sequence $s_n \in (0, n)$ such that $\eta_n = s_n/n = o(1)$. Under Condition 3.1, we have*

$$\operatorname*{plim}_{n\to\infty} \overline{\mathcal{R}}\big(\Theta[s_n] \mid \{r_i\}\big) / \big\{\mathbb{E}_G[\overline{\mathcal{C}}] s_n \log\big(\eta_n^{-1}\big)\big\} = 1.$$

Further, the predictive density $q_{\Pi[\eta_n,\kappa]}$ based on $\Pi[\eta_n, \kappa]$ with $0 < \kappa \leq 1$ is asymptotically minimax:

$$\operatorname*{plim}_{n\to\infty} \overline{\mathcal{R}}\big(\Theta[s_n] \mid \{r_i\}\big) / R\big(\theta, q_{\Pi[\eta_n,\kappa]} \mid \{r_i\}\big) = 1.$$

Proposition 3.4. *Fix $\kappa \in (0, 1]$ and assume that Condition 3.1 holds. Then, the predictive density $q_{\Pi[\hat{\eta}_n,\kappa]}$ is adaptive in the asymptotically minimax sense on the class of exact sparse parameter spaces: for any sequence $s_n \in [1, n)$ such that $\sup_n s_n/n < 1$ and $\eta_n = s_n/n = o(1)$,*

$$\operatorname*{plim}_{n\to\infty} \overline{\mathcal{R}}\big(\Theta[s_n] \mid \{r_i\}\big) / R\big(\theta, q_{\Pi[\eta_n,\kappa]} \mid \{r_i\}\big) = 1.$$

Detailed discussions for the results herein are given in Appendix D.

Remark 3.1. We remark that the optimal scaling (in the sense of Proposition 2.5) for this set-up is $\overline{L}$, where

$$\overline{L} = \overline{\mathcal{C}}/\overline{\mathcal{K}} \quad \text{with } \overline{\mathcal{K}} = \Gamma(\kappa+1)\sum_{i=1}^{n}\left\{r_i^{-\kappa} - (r_i+1)^{-\kappa}\right\}/(n\kappa).$$

The derivation follows almost the same line as in the proof of Proposition 2.5 and is omitted.

4. Simulation studies and application to real data

4.1. Simulation studies

This subsection presents simulation studies to compare the performance of various predictive densities. The codes for implementing the proposed method are available at https://github.com/kyanostat/sparsepoisson.

Consider a sparse Poisson model described as follows. Parameter θ and observations X and Y are drawn from

$$\theta_i \sim \nu_i e_{S,i} \quad (i = 1, \dots, n),$$
$$X \mid \theta \sim \bigotimes_{i=1}^{n} \mathrm{Po}(r\theta_i), \qquad Y \mid \theta \sim \bigotimes_{i=1}^{n} \mathrm{Po}(\theta_i), \quad \text{and} \quad X \perp\!\!\!\perp Y \mid \theta,$$

respectively. Here,

- $\nu_1, \dots, \nu_n$ 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;
- $\nu_1, \dots, \nu_n$ and S are independent.

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

We compare the following four predictive densities:

- The Bayes predictive density based on $\Pi[L^*\hat{\eta}_n, \kappa]$ with L^* in Proposition 2.5;
- The Bayes predictive density based on the shrinkage prior in [33];
- The Bayes predictive density based on the Gauss hypergeometric prior in [14];
- The plug-in predictive density based on an ℓ_1-penalized estimator.

The second predictive density is shown in [33] to dominate the Bayes predictive density based on the Jeffreys prior. This predictive density has a hyper-parameter β and in simulation studies it is fixed to be 1. The third predictive density employs the global-local prior proposed in [14] and the specification of the hyper-parameters follows the online support pages the authors provide.

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,

Table 1. Comparison of predictive densities with $(n, s, r) = (200, 5, 1)$: the ℓ_1 distance, PLL, and 90%CP represent the mean ℓ_1 distance, the predictive log likelihood, and the empirical coverage probability based on a 90%-prediction set, respectively. For each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance

	$\Pi[L^*\hat{\eta}_n, 0.1]$	$\Pi[L^*\hat{\eta}_n, 1.0]$	GH	K04	ℓ_1 $(\lambda = 0.1)$
ℓ_1 distance	<u>18.8</u> (5.8)	21.9 (6.8)	104 (4.9)	96.5 (8.1)	22.1 (7.8)
PLL	<u>−15.4</u> (1.8)	−16.1 (1.6)	−66.3 (3.3)	−86.2 (8.8)	−Inf
90%CP (%)	92.6 (0.1)	95.8 (0.1)	<u>92.0</u> (1.5)	40.5 (24.4)	49.4 (21.6)

- the predictive log likelihood, that is, the log of the value of a predictive density at sampled Y and X, and
- the (empirical) coverage probability of Y on the basis of the joint 90%-prediction set constructed by a predictive density.

Tables 1 and 2 show the results of the comparison. The following abbreviations are used in the tables. The Bayes predictive density proposed in [14] is abbreviated to GH. The Bayes predictive density proposed in [33] 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 represents a mean ℓ_1 distance. The abbreviation PLL represents a predictive log likelihood. The abbreviation 90%CP represents the empirical coverage probability based on a 90%-prediction set.

The results have been summarized as follows. In regard to the ℓ_1 distances, samples from the predictive density based on $\Pi[L^*\hat{\eta}_n, 0.1]$ are closer to future observations than those of three other classes of predictive densities. In regard to the empirical coverage probabilities, the predictive densities based on $\Pi[L^*\hat{\eta}_n, 0.1]$ and the Gauss hypergeometric prior give the empirical coverage probabilities of Y that are relatively close to the nominal level. 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, an ℓ_1-penalized estimator returns zero for a coordinate at which the current observation is zero and most of the marginal predictive intervals degenerate into zero. This degeneracy also induces the divergence of a predictive log likelihood value of the plug-in predictive density based on an ℓ_1-penalized estimator.

Supplemental material provides additional numerical experiments including quasi-sparse, MCAR, large s_n settings, and the comparison with the recently developed prior distribution [23]. These show that the proposed predictive density with $\kappa = 0.1$ has stable predictive performance and this value of κ is suggested as a good default choice.

Table 2. Comparison of predictive densities with $(n, s, r) = (200, 5, 20)$: the ℓ_1 distance, PLL, and 90%CP represent the mean ℓ_1 distance, the predictive log likelihood, and the empirical coverage probability based on a 90%-prediction set, respectively. For each result, the averaged value is followed by the corresponding standard deviation. Underlines indicate the best performance

	$\Pi[L^*\hat{\eta}_n, 0.1]$	$\Pi[L^*\hat{\eta}_n, 1.0]$	GH	K04	ℓ_1 $(\lambda = 0.1)$
ℓ_1 distance	<u>14.0</u> (4.9)	14.5 (4.5)	15.7 (1.7)	22.5 (5.2)	14.1 (4.5)
PLL	<u>−13.3</u> (1.6)	−13.5 (1.5)	−15.6 (1.5)	−21.6 (2.2)	−Inf
90%CP (%)	<u>90.0</u> (0.0)	89.4 (0.0)	97.6 (0.7)	97.5 (1.4)	86.3 (3.9)

4.2. Application to real data

We apply our methods to Japanese crime data from an official database called *the number of crimes in Tokyo by type and town* [17]. This database reports the total numbers of crimes in Tokyo Prefecture. They are classified by town and also by the type of crimes. A motivation for this analysis comes from the importance of taking measures against future crimes by utilizing past crime data.

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 3 shows total counts of pickpockets from 2012 to 2017 for all towns in the wards. The scale of the pickpocket occurrences in each town is expressed by a gradation of colors: there have occurred more pickpockets in a deeper-colored town over 6 years. There have not occurred any pickpocket in white-colored towns. As seen from Figure 3, the data have zero or near-zero counts at a vast majority of locations, while having relatively large counts at certain locations.

The experimental settings are as follows. The data at the 978 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, in general, the ratio r of sample sizes is set as $r = 12$. However, some observations are missing because several towns, though in rare cases, did not report the counts.

As in Section 4.1, we compare the proposed predictive density $q_{\Pi[\overline{L}\hat{\eta}_n,\kappa]}$ (with $\overline{L}$ in Remark 3.1) to 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 $q_{\Pi[\overline{L}\hat{\eta}_n,\kappa]}$ is set as the simple estimator described before Theorem 2.2 with a slight modification: we use the mean of the numbers of values greater than 1 in each year as $\hat{s}_n$. The value of κ is fixed to be 0.1 as the numerical simulations suggest. We evaluate these predictive densities on the basis of the following two measures:

- The weighted ℓ_1 distance with the weight proportional to r between the mean vector of a predictive density and the data obtained in the first half of 2018, and
- the predictive log likelihood at the data obtained in the first half of 2018.

Figure 3. Pickpocket data: (Left) 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). There have occurred more pickpockets in a deeper-colored town over 6 years. There have occurred no pickpockets in white-colored towns. (Right) Coordinate-wise medians of the proposed predictive density $q_{\Pi[\overline{L}\hat{\eta}_n,0.1]}$ for the pickpockets in the first half of 2018.

Table 3. Comparison of predictive densities in pickpocket data by the weighted ℓ_1 distance (W-ℓ_1 distance) and the predictive log likelihood (PLL): underlines indicate the best performances

	$\Pi[\overline{L}\hat{\eta}_n, 0.1]$	GH	K04	ℓ_1 ($\lambda = 0.1$)
W-ℓ_1 distance	<u>273</u>	293	<u>273</u>	297
PLL	<u>−399</u>	<u>−399</u>	−429	−Inf

Table 3 shows a summary of comparisons. In all measures, the proposed predictive density $q_{\Pi[\overline{L}\hat{\eta}_n, 0.1]}$ has the best scores. The predictive density provides not only the mean but also the other statistics. The figure to the right of Figure 3 displays the coordinate-wise medians of the proposed predictive density. It highlights crime spots at which many pickpockets have occurred over the past 6 year, and at the same time shows potential crime spots that the spike-and-slab prior structure suggests.

5. Proofs of main theorems

This section presents proofs for the main theorems in Section 2.

5.1. Supporting lemmas

We begin with stating the supporting lemmas. For an estimator $\hat{\theta}$, let

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

For a prior Π of θ, let

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

and let $\hat{\theta}_\Pi(x; t) := (\hat{\theta}_\Pi(x; t), \ldots, \hat{\theta}_{\Pi,n}(x; t))$.

The first lemma reduces bounding $R(\theta, q_\Pi)$ to bounding $R_{\mathrm{e}}(\theta, \hat{\theta}_\Pi)$. The proof is given in Section 6.

Lemma 5.1. *Fix 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, we have*

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

The second and third lemma display useful formulae for Poisson random variables. The proofs are easy and we omit them.

Lemma 5.2. *Let X_1 be a random variable from the Poisson distribution with mean λ. Then, we have*

$$\mathbb{E}_\lambda \left[\frac{1}{X_1 + 1} \right] = \frac{1 - \mathrm{e}^{-\lambda}}{\lambda}.$$

Lemma 5.3. *Let X_1 be a random variable from the Poisson distribution with mean λ. Then, we have*

$$\mathbb{P}(X_1 - \lambda \le -x) \le \exp\left(-\frac{x^2}{2\lambda}\right), \quad 0 \le x \le \lambda.$$

5.2. Proof of Theorem 2.1

Step 1: *Lower bound on* $\mathcal{R}(\Theta[s_n])$

The Bayes risk minimization with respect to *block-independent* priors will give a lower bound on $\mathcal{R}(\Theta[s_n])$. Let $\Pi_{\mathrm{B},\nu}(d\theta)$ with $\nu > 0$ be a *block-independent* prior built as follows: divide $\{1, 2, \ldots, n\}$ into contiguous blocks $\{B_j : j = 1, 2, \ldots, s_n\}$ with each length $m_n := \lfloor \eta_n^{-1} \rfloor$. In each block B_j, draw $(\theta_{1+m_n(j-1)}, \ldots, \theta_{m_n j})$ 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 $I \in \{1, \ldots, m_n\}$ 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.

Start with deriving the explicit form of $\hat{\theta}_{\Pi_{\mathrm{B},\nu}}$. Let $\mathcal{X}_j := \{x^{(j)} = (x_1, \ldots, x_{m_n}) : \|x^{(j)}\|_0 \le 1\}$ $(j = 1, 2, \ldots, s_n)$. Observe that the Bayes formula yields, for $j = 1, \ldots, s_n - 1$ and for $i = 1 + m_n(j - 1), \ldots, m_n j$,

$$\hat{\theta}_{\Pi_{\mathrm{B},\nu},i}\big(x^{(j)}\big) = \frac{\sum_{k=1}^{m_n} \int \prod_{l \ne i} \{\theta_l^{x_l}\} \theta_i^{x_i+1} \, d\delta_0(\theta_1) \cdots d\delta_\nu(\theta_k) \cdots d\delta_0(\theta_{m_n})}{\sum_{k=1}^{m_n} \int \prod_{l=1}^{m_n} \theta_l^{x_l} \, d\delta_0(\theta_1) \cdots d\delta_\nu(\theta_k) \cdots d\delta_0(\theta_{m_n})}. \tag{5}$$

This implies that for each $j = 1, \ldots, s_n$ and for each $x^{(j)} \in \mathcal{X}_j$,

$$\hat{\theta}_{\Pi_{\mathrm{B},\nu},i}\big(x^{(j)}\big) = \begin{cases} \nu/m_n & \text{if } \big\|x^{(j)}\big\|_0 = 0, \\ \nu & \text{if } x_i^{(j)} \ne 0 \text{ and } x_k^{(j)} = 0 \text{ for } k \ne i, \\ 0 & \text{if otherwise,} \end{cases} \tag{6}$$

as well as that for $j = 1, \ldots, s_n$,

$$\sum_{i=1+m_n(j-1)}^{m_n j} \hat{\theta}_{\Pi_{\mathrm{B},\nu},i}\big(x^{(j)}\big) = \nu \quad \text{for } x^{(j)} \text{ such that } \big\|x^{(j)}\big\|_0 \le 1. \tag{7}$$

Fix θ in the support of $\Pi_{\mathrm{B},\nu}$. Then, we have

$$\begin{aligned} R_{\mathrm{e}}(t\theta, t\hat{\theta}_{\Pi_{\mathrm{B},\nu}}) &= \mathbb{E}_{t\theta} \sum_{i=1}^{n} t\theta_i \log \frac{\theta_i}{\hat{\theta}_{\Pi_{\mathrm{B},\nu},i}(X)} - t \sum_{i=1}^{n} \big(\theta_i - \mathbb{E}_{t\theta}\big[\hat{\theta}_{\Pi_{\mathrm{B},\nu},i}(X)\big]\big) \\ &= \sum_{j=1}^{s_n} \sum_{i=1+m_n(j-1)}^{m_n j} \mathbb{E}_{t\theta} t\theta_i \log \frac{\theta_i}{\hat{\theta}_{\Pi_{\mathrm{B},\nu},i}(X)}, \end{aligned}$$

where the second equality holds from (7). Letting $i(j) \in \{1 + m_n(j - 1), \ldots, m_n j\}$ denote the index with $\theta_{i(j)} = \nu$, we further have

$$R_{\mathrm{e}}(t\theta, t\hat{\theta}_{\Pi_{\mathrm{B},\nu}}) \ge \sum_{j=1}^{s_n} \mathbb{P}(X_{i(j)} = 0) t\nu \log \frac{\nu}{\nu/m_n} = s_n \mathrm{e}^{-t\nu} t\nu \log m_n.$$

This, together with Lemma 5.1, gives

$$R(\theta, q_{\Pi_{\mathrm{B},\nu}}) = \int_r^{r+1} \frac{R_{\mathrm{e}}(t\theta, t\hat{\theta}_{\Pi_{\mathrm{B},\nu}})}{t}\, dt \geq \bigl\{\mathrm{e}^{-r\nu} - \mathrm{e}^{-(r+1)\nu}\bigr\} s_n \log m_n.$$

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

$$\begin{aligned}\mathcal{R}\bigl(\Theta[s_n]\bigr) &\geq \inf_{\hat{q}} \int R(\theta, \hat{q})\, d\Pi_{\mathrm{B},\nu}(\theta) = \int R(\theta, q_{\Pi_{\mathrm{B},\nu}})\, d\Pi_{\mathrm{B},\nu}(\theta) \\ &\geq \bigl\{\mathrm{e}^{-r\nu} - \mathrm{e}^{-(r+1)\nu}\bigr\} s_n \log m_n.\end{aligned}$$

Maximizing the right-hand side in the above inequality with respect to ν presents the desired lower bound $\mathcal{R}(\Theta[s_n]) \geq \mathcal{C} s_n \log m_n$, which completes Step 1.

Step 2: *Upper bound on* $\mathcal{R}(\Theta[s_n])$

Let Π be an i.i.d. prior Π and consider the coordinate-wise Kullback–Leibler risk of the Bayes predictive density q_Π:

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

where $q_\Pi(y_i \mid x_i)$ is the marginal distribution of q_Π. Consider the following high-level condition on Π:

Condition 5.1. *There exist constants* $K \geq 1, C_1, C_2, C_3, C_4 > 0, C_5 < K, C_6 > 0$ *not depending on* n *for which we have*

(P1) $C_1\eta_n \leq \hat{\theta}_{\Pi,i}(X_i; t) \leq C_2\eta_n$ *for* $X_i = 0$;
(P2) $C_3 \leq \hat{\theta}_{\Pi,i}(X_i; t) \leq C_4$ *for* $1 \leq X_i \leq K$;
(P3) $0 < (X_i - C_5)/t \leq \hat{\theta}_{\Pi,i}(X_i; t) \leq (X_i + C_6)/t$ *for* $K < X_i$.

Under this condition, 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 conclude

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

Note that for $\Pi[\eta_n, \kappa]$ with $\kappa > 0$, the Bayes formula gives

$$\hat{\theta}_{\Pi[\eta_n,\kappa],1}(x_1; t) = \frac{0^{x_1+1} + \eta_n\Gamma(x_1 + \kappa + 1)/t^{x_1+\kappa+1}}{0^{x_1} + \eta_n\Gamma(x_1 + \kappa)/t^{x_1+\kappa}} = \begin{cases} \dfrac{\eta_n\Gamma(\kappa+1)/t^{\kappa+1}}{1 + \eta_n\Gamma(\kappa)/t^\kappa}, & x_1 = 0, \\ \dfrac{x_1 + \kappa}{t}, & x_1 \geq 1, \end{cases}$$

thereby implying $\Pi[\eta_n, \kappa]$ satisfies Condition 5.1.

For $\lambda > 0$ and $t \in (r, r+1)$, let $\hat{\rho}(\lambda, x_1; t) := t\lambda \log\{\lambda/\hat{\theta}_{\Pi,1}(x_1; t)\} - t\lambda + t\hat{\theta}_{\Pi,1}(x_1; t)$. Lemma 5.1 gives

$$\begin{aligned}\rho(\lambda) \le &\underbrace{\int_r^{r+1} \mathrm{E}_{t\lambda} 1_{X_1=0} \left\{ \lambda \log \frac{\lambda}{\hat{\theta}_{\Pi,1}(X_1; t)} - \lambda + \hat{\theta}_{\Pi,1}(X_1; t) \right\} dt}_{=:A_1} \\ &+ \underbrace{\int_r^{r+1} \mathrm{E}_{t\lambda} 1_{1 \le X_1 \le K} \left\{ \lambda \log \frac{\lambda}{\hat{\theta}_{\Pi,1}(X_1; t)} - \lambda + \hat{\theta}_{\Pi,1}(X_1; t) \right\} dt}_{=:A_2} \\ &+ \underbrace{\int_r^{r+1} \mathrm{E}_{t\lambda} 1_{K < X_1} \left\{ \lambda \log \frac{\lambda}{\hat{\theta}_{\Pi,1}(X_1; t)} - \lambda + \hat{\theta}_{\Pi,1}(X_1; t) \right\} dt}_{=:A_3}. \end{aligned} \tag{8}$$

From (P1) in Condition 5.1, we have

$$A_1 \le \left\{ \mathrm{e}^{-r\lambda} - \mathrm{e}^{-(r+1)} \right\} \left\{ \log \eta_n^{-1} + \log C_1^{-1} + \log \lambda \right\} + C_2 \eta_n. \tag{9}$$

From (P2) in Condition 5.1, we get

$$A_2 \le \sum_{k=1}^{K} \frac{(r+1)^k \lambda^k}{k!} \mathrm{e}^{-r\lambda} \left\{ \lambda \log \lambda + \lambda \log C_3^{-1} + C_4 \right\}. \tag{10}$$

To bound A_3, for $r \in (r, r+1)$, we write the integrand in A_3 as

$$\begin{aligned}&\mathrm{E}_{t\lambda} 1_{K < X_1} \left\{ \lambda \log \frac{\lambda}{\hat{\theta}_{\Pi,1}(X_1; t)} - \lambda + \hat{\theta}_{\Pi,1}(X_1; t) \right\} \\ &\quad = \underbrace{\mathrm{E}_{t\lambda} 1_{K < X_1} \lambda \log \left\{ \lambda / \hat{\theta}_{\Pi,1}(X_1; t) \right\}}_{=:A_{3,1}} + \underbrace{\mathrm{E}_{t\lambda} 1_{K < X_1} \left[-\lambda + \hat{\theta}_{\Pi,1}(X_1; t) \right]}_{=:A_{3,2}}. \end{aligned}$$

Lemma 5.2, together with Jensen's inequality, yields

$$\begin{aligned} A_{3,1} &= \mathrm{E}_{t\lambda} 1_{K < X_1} \left[\lambda \log \frac{t\lambda}{X_1 + 1} \right] + \mathrm{E}_{t\lambda} 1_{K < X_1} \left[\lambda \log \frac{X_1 + 1}{X_1 - C_5} \right] \\ &\le \lambda \left(1 - \mathrm{e}^{-t\lambda} \right) + \sum_{k=0}^{K} \frac{(r+1)^k \lambda^k}{k!} \mathrm{e}^{-r\lambda} \lambda |\log t\lambda| + \mathrm{E}_{t\lambda} 1_{K < X_1} \left[\lambda \log \frac{X_1 + 1}{X_1 - C_5} \right] \\ &\le \underbrace{\sum_{k=0}^{K} \frac{(r+1)^k \lambda^{k+1}}{k!} \mathrm{e}^{-r\lambda} \left\{ \max \left\{ |\log r\lambda|, \left| \log(r+1)\lambda \right| \right\} \right\}}_{=:F_1(\lambda, r)} + \underbrace{\mathrm{E}_{t\lambda} 1_{K < X_1} \left[\lambda \log \frac{X_1 + 1}{X_1 - C_5} \right]}_{=:A_{3,1,1}}. \end{aligned}$$

If $C_5 \le -1$, then $A_{3,1,1}$ is bounded above by 0. So, we can assume $C_5 > -1$. Observe that there exists some $c_1 > 1$ depending only on K and C_5 such that we have $(X_1 + 1)/(X_1 - C_5) < c_1$ for $X_1 > K$.

This gives

$$A_{3,1,1} \le \lambda \log c_1.$$

This bound is crude for large λ, and we consider another bound for large λ. Take λ° in such a way that $(r\lambda^\circ) - (r\lambda^\circ)^{3/4} - C_5 > 0$ and $r\lambda^\circ > 1$. Then Lemma 5.3 yields, for $\lambda > \lambda^\circ$,

$$A_{3,1,1} \le \mathrm{e}^{-(r\lambda)^{1/2}/2}\lambda \log c_1 + \lambda \log\biggl\{1 + \frac{1 + C_5}{(r\lambda) - (r\lambda)^{3/4} - C_5}\biggr\}.$$

Thus, we obtain $A_{3,1,1} \le F_2(\lambda, r)$ with

$$F_2(\lambda, r) := \begin{cases} \lambda \log c_1, & \lambda \le \lambda^\circ, \\ \mathrm{e}^{-(r\lambda)^{1/2}/2}\lambda \log c_1 + \lambda \log\biggl\{1 + \dfrac{1 + C_5}{(r\lambda) - (r\lambda)^{3/4} - C_5}\biggr\}, & \lambda > \lambda^\circ, \end{cases}$$

which, together with simple bounds on $A_{3,1,1}$ and $A_{3,2}$, gives

$$A_3 \le F_1(\lambda, r) + F_2(\lambda, r) + \sum_{k>K} \frac{(t\lambda)^k \mathrm{e}^{-t\lambda}(k + C_6)}{r(k!)}. \tag{11}$$

Taking the limit as $\lambda \to 0$ in the right-hand sides of (9), (10), and (11) gives $\rho(0) = O(\eta_n)$. Maximizing upper bounds in (9), (10), and (11) with respect to λ yields $\sup_{\lambda>0} \rho(\lambda) \le (C + o(1)) \log \eta_n^{-1}$. Thus, we obtain the desired upper bound

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

which completes the proof.

5.3. Proof of Theorem 2.2

Let $\bar{p} := \sum_{j=1}^{s_n} (1 - \mathrm{e}^{-r\theta_{[j]}})/s_n$, where $\theta_{[j]}$ $(j = 1, \dots, s_n)$ denotes the j-th largest component of $\{\theta_i : i = 1, \dots, n\}$.

5.3.1. *Supporting lemma*: *Properties of* $\hat{s}_n$

We start with summarizing the behaviour of $\hat{s}_n$ which is an important ingredient of the proof. The proof is given in Section 6.

Lemma 5.4. *The following hold for* $\theta \in \Theta[s_n]$:

(a) $\hat{s}_n \ge 1$;
(b) $\max\{\mathbb{E}_\theta |\hat{s}_n/s_n - 1|, \mathbb{E}_\theta |\hat{s}_n/s_n - 1|^2\} \le 3$;
(c) *If* $s_n \ge 4$ *and* $\theta_{[1]} \ge 1/\sqrt{\log s_n}$, *then, for sufficiently large* n *depending only on* r,

$$\mathbb{E}_\theta \log(s_n/\hat{s}_n) \le c_1 \max\bigl\{\sqrt{\log s_n}, s_n \exp(-c_2 s_n/\log s_n)\bigr\}$$

with positive constants c_1 *and* c_2 *depending only on* r.

This lemma indicates that $\hat{s}_n$ is not so far from s_n. In the sparse region (i.e., $s_n = o(n^{1/2})$), properties (a) and (b) are sufficient to prove Theorem 2.2. In the dense region (i.e., $s_n > cn^{1/2}$ for any $c > 0$), property (c) is additionally required.

5.3.2. *Proving Theorem 2.2*

Decompose the difference between the Kullback–Leibler divergences in such a way that

$$R(\theta, q_{\Pi[\hat{\eta}_n,\kappa]}) - R(\theta, q_{\Pi[\eta_n,\kappa]}) = \sum_{i=1}^{n} \underbrace{\mathbb{E}_\theta \log\left\{\frac{q_{\Pi[\eta_n,\kappa],i}(Y_i \mid X_i)}{q_{\Pi[\hat{\eta}_n,\kappa],i}(Y_i \mid X_i)}\right\}}_{:=D_i}$$
$$= \sum_{i\in\mathcal{A}} D_i + \sum_{i\notin\mathcal{A}} D_i, \tag{12}$$

where for $\theta \in \Theta[s_n]$, let $\mathcal{A} := \mathcal{A}(\theta) = \{i : \theta_i \neq 0\}$. The following three steps give an upper bound on the right-hand side of (12). In all steps, we use the following expression of $q_{\Pi[h,\kappa]}$:

$$q_{\Pi[h,\kappa]}(y \mid x) = \prod_{i=1}^{n} \left\{\omega_i \delta_0(y_i) + (1-\omega_i)\binom{x_i + y_i + \kappa - 1}{y_i}\left(\frac{r}{r+1}\right)^{x_i+\kappa}\left(1 - \frac{r}{r+1}\right)^{y_i}\right\}, \tag{13}$$

where

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

Step 1: Bounding D_i for $i \notin \mathcal{A}$. From (13), we have

$$D_i = \mathbb{E}_\theta \log\left\{\frac{1 + \hat{\eta}_n\Gamma(\kappa)/r^\kappa}{1 + \eta_n\Gamma(\kappa)/r^\kappa}\right\} + \mathbb{E}_\theta \log\left[\frac{1 + \{\eta_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}{1 + \{\hat{\eta}_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}\right]. \tag{14}$$

Since $\log(1+x) \leq x$ for $x > 0$, we have

$$\begin{aligned} \mathbb{E}_\theta \log\left\{\frac{1 + \hat{\eta}_n\Gamma(\kappa)/r^\kappa}{1 + \eta_n\Gamma(\kappa)/r^\kappa}\right\} &\leq \mathbb{E}_\theta \log\left\{1 + \frac{(\hat{\eta}_n - \eta_n)\Gamma(\kappa)/r^\kappa}{1 + \eta_n\Gamma(\kappa)/r^\kappa}\right\} \\ &\leq \mathbb{E}_\theta \log\left\{1 + \eta_n|\hat{s}_n/s_n - 1|\frac{\Gamma(\kappa)}{r^\kappa}\right\} \\ &\leq \eta_n \frac{\Gamma(\kappa)}{r^\kappa}\mathbb{E}_\theta|\hat{s}_n/s_n - 1| \\ &\leq c_1\eta_n \quad \text{with } c_1 := 3\frac{\Gamma(\kappa)}{r^\kappa} \end{aligned} \tag{15}$$

where the last inequality follows from Lemma 5.4 (b). Observe that

$$\frac{\{\eta_n - \hat{\eta}_n\}\{\Gamma(\kappa)/(r+1)^\kappa\}}{1 + \eta_n\{\Gamma(\kappa)/(r+1)^\kappa\}} \leq \frac{\eta_n\Gamma(\kappa)/(r+1)^\kappa}{1 + \eta_n\Gamma(\kappa)/(r+1)^\kappa} \leq \frac{\Gamma(\kappa)}{1 + \Gamma(\kappa)}.$$

Then, together with the inequality

$$-\log(1-x) \leq \{1/(1-U)^2\}x^2 + x \quad \text{for } 0 < x \leq U \text{ with some } 0 < U < 1,$$

this observation gives

$$\begin{aligned}
&\mathbb{E}_\theta \log\left[\frac{1+\{\eta_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}{1+\{\hat{\eta}_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}\right] \\
&\quad\le -\mathbb{E}_\theta \log\left[1-\frac{\{\eta_n-\hat{\eta}_n\}\{\Gamma(\kappa)/(r+1)^\kappa\}}{1+\eta_n\{\Gamma(\kappa)/(r+1)^\kappa\}}\right] \\
&\quad\le \eta_n\frac{\Gamma(\kappa)}{(r+1)^\kappa}\mathbb{E}_\theta|\hat{\eta}_n/\eta_n-1|+\left[\left\{1+\Gamma(\kappa)\right\}\eta_n\frac{\Gamma(\kappa)}{(r+1)^\kappa}\right]^2\mathbb{E}_\theta|\hat{\eta}_n/\eta_n-1|^2 \\
&\quad\le c_2\eta_n \quad \text{with } c_2 := 3\frac{\Gamma(\kappa)}{(r+1)^\kappa}+3\left[\left\{1+\Gamma(\kappa)\right\}\frac{\Gamma(\kappa)}{(r+1)^\kappa}\right]^2, \qquad (16)
\end{aligned}$$

where the last inequality follows from Lemma 5.4 (b). Combining (15) and (16) with (14) yields

$$D_i \le c_3\eta_n \quad \text{with } c_3 := c_1+c_2 \text{ for } i\notin\mathcal{A}. \qquad (17)$$

Step 2: Bounding D_i for $i\in\mathcal{A}$. Consider the following four cases: (i) $X_i=0$, $Y_i=0$; (ii) $X_i\ge 1$, $Y_i=0$; (iii) $X_i=0$, $Y_i\ge 1$; (iv) $X_i\ge 1$, $Y_i\ge 1$. In Case (i), we have

$$\begin{aligned}
&\log\left\{\frac{q_{\Pi[\eta_n,\kappa],i}(Y_i\mid X_i)}{q_{\Pi[\hat{\eta}_n,\kappa],i}(Y_i\mid X_i)}\right\} \\
&\quad=\log\left\{\frac{1+\hat{\eta}_n\Gamma(\kappa)/r^\kappa}{1+\eta_n\Gamma(\kappa)/r^\kappa}\right\}+\log\left[\frac{1+\{\eta_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}{1+\{\hat{\eta}_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}\right] \\
&\quad\le|\hat{s}_n/s_n-1|+\log\left[1+\Gamma(\kappa)/(r+1)^\kappa\right], \qquad (18)
\end{aligned}$$

where we use the inequalities

$$\log\left\{\frac{1+\hat{\eta}_n\Gamma(\kappa)/r^\kappa}{1+\eta_n\Gamma(\kappa)/r^\kappa}\right\}\le\log\left\{1+\frac{|\hat{\eta}_n-\eta_n|\Gamma(\kappa)/r^\kappa}{1+\eta_n\Gamma(\kappa)/r^\kappa}\right\}\le|\hat{s}_n/s_n-1|$$

and

$$\begin{aligned}
\log\left[\frac{1+\{\eta_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}{1+\{\hat{\eta}_n\Gamma(\kappa)/r^\kappa\}\{r/(r+1)\}^\kappa}\right]&\le\log\left[1+\left\{\eta_n\Gamma(\kappa)/r^\kappa\right\}\left\{r/(r+1)\right\}^\kappa\right] \\
&\le\log\left[1+\Gamma(\kappa)/(r+1)^\kappa\right].
\end{aligned}$$

Similarly, we get the following evaluations: In Case (ii),

$$\log\left\{\frac{q_{\Pi[\eta_n,\kappa],i}(Y_i\mid X_i)}{q_{\Pi[\hat{\eta}_n,\kappa],i}(Y_i\mid X_i)}\right\}=0. \qquad (19)$$

In Case (iii),

$$\log\left\{\frac{q_{\Pi[\eta_n,\kappa]}(Y_i\mid X_i)}{q_{\Pi[\hat{\eta}_n,\kappa]}(Y_i\mid X_i)}\right\}\le\log(s_n/\hat{s}_n)+|\hat{s}_n/s_n-1|. \qquad (20)$$

In Case (iv),

$$\log\left\{\frac{q_{\Pi[\eta_n,\kappa]}(Y_i\mid X_i)}{q_{\Pi[\hat{\eta}_n,\kappa]}(Y_i\mid X_i)}\right\}=0. \qquad (21)$$

From (18)–(21), we have, for $i \in \mathcal{A}$,

$$\begin{aligned} D_i &\le 2\mathbb{E}_\theta |\hat{s}_n/s_n - 1| + \log\bigl[1 + \Gamma(\kappa)/(r+1)^\kappa\bigr] + \mathbb{E}_\theta\bigl[1_{X_i=0, Y_i \ge 1} \log(s_n/\hat{s}_n)\bigr] \\ &\le 6 + \log\bigl[1 + \Gamma(\kappa)/(r+1)^\kappa\bigr] + \mathbb{E}_\theta\bigl[1_{X_i=0, Y_i \ge 1} \log(s_n/\hat{s}_n)\bigr], \end{aligned} \tag{22}$$

where the last inequality follows from Lemma 5.4 (b).

Step 3. Combining (17) and (22) with (12) gives

$$\begin{aligned} R(\theta, q_{\Pi[\hat{\eta}_n,\kappa]}) &\le R(\theta, q_{\Pi[\eta_n,\kappa]}) + c_3(n - s_n)\eta_n + s_n\bigl\{6 + \log\bigl[1 + \Gamma(\kappa)/(r+1)^\kappa\bigr]\bigr\} \\ &\quad + \sum_{i \in \mathcal{A}} \mathbb{E}_\theta\bigl[1_{X_i=0, Y_i \ge 1} \log(s_n/\hat{s}_n)\bigr] \\ &\le R(\theta, q_{\Pi[\eta_n,\kappa]}) + c_4 s_n + \sum_{i \in \mathcal{A}} \underbrace{\mathbb{E}_\theta\bigl[1_{X_i=0, Y_i \ge 1} \log(s_n/\hat{s}_n)\bigr]}_{:=T_i}, \end{aligned} \tag{23}$$

where $c_4 := c_3 + 6 + \log[1 + \Gamma(\kappa)/(r+1)^\kappa]$. We will show

$$\sum_{i \in \mathcal{A}} T_i = o\bigl(s_n \log(n/s_n)\bigr). \tag{24}$$

Since the number of indices in $\mathcal{A}$ is bounded above from s_n, it suffices to show $T_i = o(\log(n/s_n))$ uniformly in $\theta \in \Theta[s_n]$ and $i \in \mathcal{A}$. First, consider the case with $s_n = o(n^{1/2})$. Since $\log(s_n/\hat{s}_n) \le \log s_n$ from Lemma 5.4 (a), we get

$$T_i \le \log s_n = o\bigl(\log(n/s_n)\bigr). \tag{25}$$

Next consider the case with $s_n > cn^{1/2}$ for any $c > 0$. Since $\log(s_n/\hat{s}_n) \le \log s_n$ from Lemma 5.4 (a), we get, for $\theta \in \Theta[s_n]$ such that $\theta_{[1]} \le 1/\sqrt{\log s_n}$,

$$T_i \le \mathbb{E}_\theta\bigl[1_{Y_i \ge 1} \log(s_n/\hat{s}_n)\bigr] \le \bigl(1 - \mathrm{e}^{-\theta_{[1]}}\bigr) \log s_n \le \theta_{[1]} \log s_n \le \sqrt{\log s_n}. \tag{26}$$

Using Lemma 5.4 (c), we have, for $\theta \in \Theta[s_n]$ such that $\theta_{[1]} \ge 1/\sqrt{\log s_n}$,

$$T_i \le \mathbb{E}_\theta\bigl[\log(s_n/\hat{s}_n)\bigr] \le c_5 \sqrt{\log s_n}, \tag{27}$$

where c_5 is the constant depending only on r appearing in Lemma 5.4 (c). From (25)–(27), we obtain (24) and thus complete the proof.

6. Proofs of auxiliary lemmas

This section provides proofs of Lemmas 5.1 and 5.4.

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

Observe that the Kullback–Leibler risk is decomposed as

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

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

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

Introducing the random variable Z_t from $\bigotimes_{i=1}^n \mathrm{Po}(t\theta_i)$ $(t \in (r, r+1))$, we get

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

Therefore, it suffices to show

$$\frac{d}{dt}\mathbb{E}\left[\log\left\{\frac{p(Z_t \mid \theta; t)}{p_\Pi(Z_t; t)}\right\}\right] = \frac{R_{\mathrm{e}}(t\theta, t\hat{\theta}_\Pi(\cdot; t))}{t}, \quad t \in (r, r+1). \tag{28}$$

Differentiating $\mathbb{E}[\log\{p(Z_t \mid \theta; t)/p_\Pi(Z_t; t)\}]$ with respect to t yields

$$\begin{aligned}\mathbb{E}\big[\log\big\{p(Z_t \mid \theta; t)/p_\Pi(Z_t; t)\big\}\big] &= \mathbb{E}\left[\left\{\frac{d\log p(Z_t \mid \theta; t)}{dt}\right\}\log\left\{\frac{p(Z_t \mid \theta; t)}{p_\Pi(Z_t; t)}\right\}\right] \\ &\quad + \mathbb{E}\left[\frac{d\log p(Z_t \mid \theta; t)}{dt}\right] - \mathbb{E}\left[\frac{d\log p_\Pi(Z_t; t)}{dt}\right].\end{aligned} \tag{29}$$

Let e_i be the unit length vector in the i-th coordinate direction $(i = 1, \ldots, n)$. Together with the simple fact that

$$p_\Pi(Z_t + e_i; t)/p_\Pi(Z_t; t) = \hat{\theta}_{\Pi,i}(Z_t; t),$$

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 $f : \mathrm{N}^n \to \mathbb{R}$) yields

$$\begin{aligned}&\mathbb{E}\big[\big\{d\log p(Z_t \mid \theta; t)/dt\big\}\log\big\{p(Z_t \mid \theta; t)/p_\Pi(Z_t; t)\big\}\big] \\ &= \mathbb{E}\left[\left\{\sum_{i=1}^n \frac{Z_{t,i} - 1 - t\theta_i)}{t}\right\}\log\big\{p(Z_t \mid \theta; t)/p_\Pi(Z_t; t)\big\}\right] \\ &= \mathbb{E}\sum_{i=1}^n \theta_i\big[\log\big\{p(Z_t + e_i \mid \theta; t)/p(Z_t \mid \theta; t)\big\} - \log\big\{p_\Pi(Z_t + e_i; t)/p_\Pi(Z_t)\big\}\big] \\ &= \mathbb{E}\sum_{i=1}^n \theta_i \log\big\{\theta_i/\hat{\theta}_{\Pi,i}(Z_t; t)\big\}.\end{aligned} \tag{30}$$

Similarly, the identity $(d/dt)\log p_\Pi(x; t) = -\sum_{i=1}^n \{\hat{\theta}_{\Pi,i}(x; t) - x_i + 1\}$ gives

$$\begin{aligned}\mathbb{E}\left[\frac{d}{dt}\big\{\log p(Z_t \mid \theta; t)\big\}\right] &= \mathbb{E}\left[\sum_{i=1}^n \frac{Z_{t,i} - 1 - t\theta_i}{t}\right] \quad \text{and} \\ \mathbb{E}\left[\frac{d}{dt}\big\{\log p_\Pi(Z_t; t)\big\}\right] &= \mathbb{E}\left[-\sum_{i=1}^n \frac{\hat{\theta}_{\Pi,i}(Z_t; t) - Z_{t,i} + 1}{t}\right].\end{aligned} \tag{31}$$

Combining identities (30) and (31) with (29) gives (28), which completes the proof. □

Proof of Lemma 5.4. Property (a) is obvious by definition. Consider the bias of $\hat{s}_n - s_n$. Decompose $\#\{i : X_i \geq 1\}$ in such a way that $\#\{i : X_i \geq 1\} = \sum_{j=1}^{s_n} Z_j$ with independent Bernoulli random variable Z_j $(j = 1, \ldots, s_n)$ having the success probability $1 - \exp(-r\theta_{[j]})$. This decomposition gives

$$-\sum_{j=1}^{s_n} \mathrm{e}^{-r\theta_{[j]}} \leq \mathbb{E}_\theta(\hat{s}_n - s_n) \leq 0. \tag{32}$$

Consider the variance of $\hat{s}_n - s_n$. Since

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

we have

$$\mathbb{E}_\theta \left| \frac{\hat{s}_n - \mathbb{E}_\theta \hat{s}_n}{s_n} \right|^2 \leq \frac{1}{s_n^2} + \sum_{j=1}^{s_n} \frac{\mathrm{e}^{-r\theta_{[j]}}(1 - \mathrm{e}^{-r\theta_{[j]}})}{s_n^2}. \tag{33}$$

Given that $s_n \geq 1$, we get $\max\{\mathbb{E}|\hat{s}_n/s_n - 1|, \mathbb{E}|\hat{s}_n/s_n - 1|^2\} \leq 3$ from (32) and (33), which shows (b).

By the layer-cake representation and since $\sum_{j=1}^{s_n} Z_j \leq \hat{s}_n$, we have

$$\begin{aligned}
\mathbb{E}_\theta \log \frac{\eta_n}{\hat{\eta}_n} &= \int_0^{\log s_n} \mathbb{P}\left(\log \frac{\eta_n}{\hat{\eta}_n} > x \right) dx \\
&= \int_{1/s_n}^1 \mathbb{P}(\hat{s}_n < \beta s_n) \frac{d\beta}{\beta} \leq \int_{1/s_n}^1 \mathbb{P}\left(\sum_{j=1}^{s_n} Z_j < \beta s_n \right) \frac{d\beta}{\beta}.
\end{aligned}$$

Together with the Hoeffding inequality, this yields

$$\begin{aligned}
\mathbb{E}_\theta \log \frac{\eta_n}{\hat{\eta}_n} &\leq \int_{1/s_n}^1 \mathbb{P}\left(\sum_{j=1}^{s_n} Z_j - \mathbb{E}\left[\sum_{j=1}^{s_n} Z_j \right] < (\beta - \bar{p}) s_n \right) \frac{d\beta}{\beta} \\
&\leq \int_{1/s_n}^1 \frac{1}{\beta} \exp\left\{ -2s_n(\beta - \bar{p})^2 \right\} d\beta = \int_{1/s_n}^1 \exp\big(f(\beta) \big)\, d\beta,
\end{aligned} \tag{34}$$

where $f(\beta) := -2s_n(\beta - \bar{p})^2 - \log \beta$.

We employ the following bound on $f(\beta)$ to obtain an upper bound on the right-hand side in (34).

Lemma 6.1. *If $\bar{p}^2 > 1/s_n$ and $s_n \geq 4$, then we have*

$$f(\beta) \leq \max\big[f(1/s_n),\, f\big\{ (1/2)\big(\bar{p} + \sqrt{\bar{p}^2 - 1/s_n} \big) \big\} \big] \quad \textit{for } 1/s_n \leq \beta \leq 1.$$

Proof of Lemma 6.1. Observe that we have

$$\begin{aligned}
f''(\beta) > 0 \quad &\text{for } 1/s_n \leq \beta < 1/\sqrt{2s_n}, \\
f''(\beta) \leq 0 \quad &\text{for } 1/\sqrt{2s_n} \leq \beta \leq 1,
\end{aligned}$$

and

$$f'(\beta) \geq 0 \quad \text{for } \max\bigl\{1/s_n, (1/2)\bigl(\bar{p} - \sqrt{\bar{p}^2 - 1/s_n}\bigr)\bigr\} \leq \beta < (1/2)\bigl(\bar{p} + \sqrt{\bar{p}^2 - 1/s_n}\bigr),$$

$$f'(\beta) \leq 0 \quad \text{for } (1/2)\bigl(\bar{p} + \sqrt{\bar{p}^2 - 1/s_n}\bigr) \leq \beta \leq 1,$$

where the first two inequalities follow from $f''(\beta) = -4s_n + 1/\beta^2$, and the last two inequalities follow from $f'(\beta) = -4s_n(\beta - \bar{p}) - 1/\beta$ and $(1/2)(\bar{p} - \sqrt{\bar{p}^2 - 1/s_n}) < 1/s_n$. This observation gives the desired inequality. □

Go back to proving (c). Observe that $\bar{p}^2 > 1/s_n$ for $n \geq N$ with sufficiently large N depending only on r, because the assumption $\theta_{[1]} \geq 1/\sqrt{\log s_n}$ implies that $\bar{p} \geq 1 - \exp\{-r\theta_{[1]}\} \geq \widetilde{c}_1/\sqrt{\log s_n}$ with $\widetilde{c}_1$ depending only on r for sufficiently large n depending only on r. Thus, Lemma 6.1, together with (34), gives

$$\mathbb{E}_\theta \log \frac{\eta_n}{\hat{\eta}_n} \leq \exp\left[\max\left\{f(1/s_n), f\left(\frac{\bar{p} + \sqrt{\bar{p}^2 - 1/s_n}}{2}\right)\right\}\right] \quad \text{for } n \geq N. \tag{35}$$

Simple calculations yield

$$\begin{aligned} f(1/s_n) &= -2\widetilde{c}_1^2 s_n/\log s_n + 2\widetilde{c}_1/\sqrt{\log s_n} - 2/s_n + \log s_n \\ &\leq -\widetilde{c}_2 s_n/\log s_n + \log s_n + \widetilde{c}_3 \end{aligned} \tag{36}$$

with constants $\widetilde{c}_2$ and $\widetilde{c}_3$ depending only on $\widetilde{c}_1$, as well as

$$\begin{aligned} f\bigl\{\bigl(\bar{p} + \sqrt{\bar{p}^2 - 1/s_n}\bigr)/2\bigr\} &= -\frac{s_n}{2}\bigl(\bar{p} - \sqrt{\bar{p}^2 - 1/s_n}\bigr)^2 - \log \frac{\bar{p} + \sqrt{\bar{p}^2 - 1/s_n}}{2} \\ &= -\frac{1}{2s_n(\bar{p} + \sqrt{\bar{p}^2 - 1/s_n})^2} - \log \frac{\bar{p} + \sqrt{\bar{p}^2 - 1/s_n}}{2} \\ &\leq -1/\bigl(2s_n\bar{p}^2\bigr) - \log(\bar{p}/2) \\ &\leq -\widetilde{c}_4(\log s_n)/s_n + \log(\sqrt{\log s_n}) + \widetilde{c}_5 \end{aligned} \tag{37}$$

with constants $\widetilde{c}_4$ and $\widetilde{c}_5$ depending only on $\widetilde{c}_1$. Combining (36) and (37) with (35) completes the proof. □

7. Discussion and conclusions

We have studied asymptotic minimaxity in sparse Poisson sequence models. We have presented Bayes predictive densities that are adaptive in the asymptotically minimax sense.

Proposition 2.4 shows that spike-and-slab priors based on polynomially decaying slabs give asymptotically minimax predictive densities. This implies that the spike-and-slab approach is useful for sparse count data analysis. At the same time, asymptotic minimaxity does not tell which slab prior is the best, and selecting a single slab prior seems to require the other metrics, say, the computational complexity or the robustness. From the viewpoint of computational complexity, predictive densities based on our slab priors are easily implemented by exact sampling and are a good starting choice. [23] investigate

Bayesian tail robustness and derive yet another interesting heavy-tailed prior in the global-local shrinkage literature. It would also be helpful to understand more about similarities and differences between our work and [23] from the predictive perspective.

We can consider more general sparsity in count data. Quasi sparsity with a non-shrinking spike component is one such example. If the value of the spike component is known, a slight modification of our method would work. If the value is unknown, further consideration would be necessary.

Acknowledgements

We are thankful to the Editor, the Associate Editor, and anonymous referees for their helpful comments. We used the dataset of *the number of crimes in Tokyo prefecture by town and type* [17] by Tokyo metropolitan government and Tokyo Metropolitan Police Department. We used OpenStreetMap [12] and CartoDB [11] to creat Figure 3. This research was supported by JST CREST (JPMJCR1763), MEXT KAKENHI (16H06533), and JST KAKENHI (17H06570, 19K20222).

Supplementary Material

Supplement to "Minimax predictive density for sparse count data" (DOI: 10.3150/20-BEJ1271SUPP; .pdf). We provide proofs for all propositions and additional discussions.

References

[1] Agarwal, D.K., Gelfand, A.E. and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environ. Ecol. Stat.* **9** 341–355. MR1951713 https://doi.org/10.1023/A:1020910605990

[2] Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353 https://doi.org/10.1093/biomet/62.3.547

[3] Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika* **65** 53–59. MR0501450 https://doi.org/10.1093/biomet/65.1.53

[4] Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34** 2921–2938. MR2329473 https://doi.org/10.1214/009053606000000885

[5] Boisbunon, A. and Maruyama, Y. (2014). Inadmissibility of the best equivariant predictive density in the unknown variance case. *Biometrika* **101** 733–740. MR3254914 https://doi.org/10.1093/biomet/asu024

[6] Brown, L.D., Greenshtein, E. and Ritov, Y. (2013). The Poisson compound decision problem revisited. *J. Amer. Statist. Assoc.* **108** 741–749. MR3174656 https://doi.org/10.1080/01621459.2013.771582

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

[8] Castillo, I. and Mismer, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Stat.* **12** 3953–4001. MR3885271 https://doi.org/10.1214/18-EJS1494

[9] Castillo, I. and Szabó, B. (2020). Spike and slab empirical Bayes sparse credible sets. *Bernoulli* **26** 127–158. MR4036030 https://doi.org/10.3150/19-BEJ1119

[10] Clevenson, M.L. and Zidek, J.V. (1975). Simultaneous estimation of the means of independent Poisson laws. *J. Amer. Statist. Assoc.* **70** 698–705. MR0394962

[11] CartoDB contributors (2011). https://carto.com.

[12] OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org. Available at https://www.openstreetmap.org.

[13] Corcuera, J.M. and Giummolè, F. (1999). A generalized Bayes rule for prediction. *Scand. J. Stat.* **26** 265–279. MR1707658 https://doi.org/10.1111/1467-9469.00149

[14] Datta, J. and Dunson, D.B. (2016). Bayesian inference on quasi-sparse count data. *Biometrika* **103** 971–983. MR3620451 https://doi.org/10.1093/biomet/asw053

[15] Dawid, A.P. and Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Anal.* **10** 479–499. MR3420890 https://doi.org/10.1214/15-BA942

[16] Deledalle, C.-A. (2017). Estimation of Kullback–Leibler losses for noisy recovery problems within the exponential family. *Electron. J. Stat.* **11** 3141–3164. MR3694579 https://doi.org/10.1214/17-EJS1321

[17] Tokyo Metropolitan Police Department The number of crimes in tokyo prefecture by town and type. Available at http://www.keishicho.metro.tokyo.jp/about_mpd/jokyo_tokei/jokyo/ninchikensu.html.

[18] Fourdrinier, D., Marchand, É., Righi, A. and Strawderman, W.E. (2011). On improved predictive density estimation with parametric constraints. *Electron. J. Stat.* **5** 172–191. MR2792550 https://doi.org/10.1214/11-EJS603

[19] Fourdrinier, D., Marchand, É. and Strawderman, W.E. (2019). On efficient prediction and predictive density estimation for normal and spherically symmetric models. *J. Multivariate Anal.* **173** 18–25. MR3913045 https://doi.org/10.1016/j.jmva.2019.02.002

[20] George, E.I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Ann. Statist.* **34** 78–91. MR2275235 https://doi.org/10.1214/009053606000000155

[21] Ghosh, M. and Yang, M.C. (1988). Simultaneous estimation of Poisson means under entropy loss. *Ann. Statist.* **16** 278–291. MR0924871 https://doi.org/10.1214/aos/1176350705

[22] Hall, D.B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **56** 1030–1039. MR1815581 https://doi.org/10.1111/j.0006-341X.2000.01030.x

[23] Hamura, Y., Irie, K. and Sugasawa, S. On global-local shrinkage priors for count data. Available at arXiv:1907.01333.

[24] Hartigan, J. (2002). Bayesian regression using Akaike priors. Preprint, Yale Univ., New Haven, CT.

[25] Hartigan, J.A. (1998). The maximum likelihood prior. *Ann. Statist.* **26** 2083–2103. MR1700222 https://doi.org/10.1214/aos/1024691462

[26] Johnstone, I. (1984). Admissibility, difference equations and recurrence in estimating a Poisson mean. *Ann. Statist.* **12** 1173–1198. MR0760682 https://doi.org/10.1214/aos/1176346786

[27] Johnstone, I.M. and MacGibbon, K.B. (1992). Minimax estimation of a constrained Poisson vector. *Ann. Statist.* **20** 807–831. MR1165594 https://doi.org/10.1214/aos/1176348658

[28] Johnstone, I.M. and Silverman, B.W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135 https://doi.org/10.1214/009053604000000030

[29] Kato, K. (2009). Improved prediction for a multivariate normal distribution with unknown mean and variance. *Ann. Inst. Statist. Math.* **61** 531–542. MR2529965 https://doi.org/10.1007/s10463-007-0163-z

[30] Kobayashi, K. and Komaki, F. (2008). Bayesian shrinkage prediction for the regression problem. *J. Multivariate Anal.* **99** 1888–1905. MR2466542 https://doi.org/10.1016/j.jmva.2008.01.014

[31] Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 299–313. MR1439785 https://doi.org/10.1093/biomet/83.2.299

[32] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88** 859–864. MR1859415 https://doi.org/10.1093/biomet/88.3.859

[33] Komaki, F. (2004). Simultaneous prediction of independent Poisson observables. *Ann. Statist.* **32** 1744–1769. MR2089141 https://doi.org/10.1214/009053604000000445

[34] Komaki, F. (2006). A class of proper priors for Bayesian simultaneous prediction of independent Poisson observables. *J. Multivariate Anal.* **97** 1815–1828. MR2298891 https://doi.org/10.1016/j.jmva.2005.12.002

[35] Komaki, F. (2015). Simultaneous prediction for independent Poisson processes with different durations. *J. Multivariate Anal.* **141** 35–48. MR3390057 https://doi.org/10.1016/j.jmva.2015.06.008

[36] Kubokawa, T., Marchand, É., Strawderman, W.E. and Turcotte, J.-P. (2013). Minimaxity in predictive density estimation with parametric constraints. *J. Multivariate Anal.* **116** 382–397. MR3049911 https://doi.org/10.1016/j.jmva.2013.01.001

[37] L'Moudden, A. and Marchand, É. (2019). On predictive density estimation under α-divergence loss. *Math. Methods Statist.* **28** 127–143. MR3989319 https://doi.org/10.3103/S1066530719020030

[38] L'Moudden, A., Marchand, É., Kortbi, O. and Strawderman, W.E. (2017). On predictive density estimation for gamma models with parametric constraints. *J. Statist. Plann. Inference* **185** 56–68. MR3612671 https://doi.org/10.1016/j.jspi.2017.01.003

[39] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to random defects in manufacturing. *Technometrics* **34** 1–14.

[40] Leung, G. and Barron, A.R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **52** 3396–3410. MR2242356 https://doi.org/10.1109/TIT.2006.878172

[41] Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. Inf. Theory* **50** 2708–2726. MR2096988 https://doi.org/10.1109/TIT.2004.836922

[42] MacGibbon, B. (2010). Minimax estimation over hyperrectangles with implications in the Poisson case. In *Borrowing Strength: Theory Powering Applications – a Festschrift for Lawrence D. Brown*. *Inst. Math. Stat. (IMS) Collect.* **6** 32–42. Beachwood, OH: IMS. MR2798509

[43] Maruyama, Y. and Strawderman, W.E. (2012). Bayesian predictive densities for linear regression models under α-divergence loss: Some results and open problems. In *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*. *Inst. Math. Stat. (IMS) Collect.* **8** 42–56. Beachwood, OH: IMS. MR3202501 https://doi.org/10.1214/11-IMSCOLL803

[44] Maruyama, Y. and Takemura, A. (2008). Admissibility and minimaxity of generalized Bayes estimators for spherically symmetric family. *J. Multivariate Anal.* **99** 50–73. MR2408443 https://doi.org/10.1016/j.jmva.2007.01.002

[45] Matsuda, T. and Komaki, F. (2015). Singular value shrinkage priors for Bayesian prediction. *Biometrika* **102** 843–854. MR3431557 https://doi.org/10.1093/biomet/asv036

[46] Mukherjee, G. and Johnstone, I. On minimax optimality of sparse Bayes predictive density estimates. Available at arXiv:1707.04380.

[47] Mukherjee, G. and Johnstone, I.M. (2015). Exact minimax estimation of the predictive density in sparse Gaussian models. *Ann. Statist.* **43** 937–961. MR3346693 https://doi.org/10.1214/14-AOS1251

[48] Murray, G.D. (1977). A note on the estimation of probability density functions. *Biometrika* **64** 150–152. MR0448690 https://doi.org/10.2307/2335788

[49] Ng, V.M. (1980). On the estimation of parametric density functions. *Biometrika* **67** 505–506. MR0581751 https://doi.org/10.1093/biomet/67.2.505

[50] Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 157–163. Berkeley: Univ. California Press. MR0084919

[51] Ročková, V. and George, E.I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476 https://doi.org/10.1080/01621459.2016.1260469

[52] Suzuki, T. and Komaki, F. (2010). On prior selection and covariate shift of β-Bayesian prediction under α-divergence risk. *Comm. Statist. Theory Methods* **39** 1655–1673.

[53] Xu, X. and Liang, F. (2010). Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli* **16** 543–560. MR2668914 https://doi.org/10.3150/09-BEJ222

[54] Xu, X. and Zhou, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *J. Multivariate Anal.* **102** 1417–1428. MR2819959 https://doi.org/10.1016/j.jmva.2011.05.008

[55] Yano, K., Kaneko, R. and Komaki, F. (2021). Supplement to "Minimax predictive density for sparse count data." https://doi.org/10.3150/20-BEJ1271SUPP

[56] Yano, K. and Komaki, F. (2017). Asymptotically minimax prediction in infinite sequence models. *Electron. J. Stat.* **11** 3165–3195. MR3697133 https://doi.org/10.1214/17-EJS1312

[57] Zhang, F., Shi, Y., Ng, H.K.T. and Wang, R. (2018). Information geometry of generalized Bayesian prediction using α-divergences as loss functions. *IEEE Trans. Inf. Theory* **64** 1812–1824. MR3766316 https://doi.org/10.1109/TIT.2017.2774820

Received November 2019 and revised June 2020