Exact Minimax Predictive Density for Sparse Count Data

Keisuke Yano¹, Ryoya Kaneko¹ and Fumiyasu Komaki¹,²

¹Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

²RIKEN Center for Brain Science, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan

Abstract: This paper discusses predictive densities under the Kullback–Leibler loss in high-dimensional sparse count data models. 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 asymptotically minimax Kullback–Leibler risks in sparse and quasi-sparse Poisson sequence models. We also provide a class of Bayes predictive densities that attain exact asymptotic minimaxity. To this end, we employ spike-and-slab priors with improper slab priors and introduce the notion of predictive specification of the scale in the mixture of improper priors. For application, we provide exact asymptotically minimax predictive densities that are adaptive to an unknown sparsity, and discuss the performance of the proposed Bayes predictive densities in settings where current observations are missing completely at random. The simulation studies as well as applications to real data demonstrate the efficiency of the proposed Bayes predictive densities.

MSC 2010 subject classifications: Primary 62M20; secondary 62C20, 62G20, 62P99.

Keywords and phrases: 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 within
statistical models in finite dimensions; see Subsection 1.2 for the literature review. One of the 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; refer to [28]. Conversely, little is known about predictive density within statistical models in high dimensions. Important works in this direction are [42, 41]. In [42, 41], several predictive densities (including a Bayes predictive density) superior to all plug-in predictive densities even in sparse high-dimensional Gaussian models are constructed.

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. In this paper, two types of sparsity are discussed: exact and quasi sparsity. Exact sparsity in count data means that there exhibits an excess of zeros. Quasi sparsity means that there exhibits an excess of near-zero counts. Exact sparsity is captured by the $\ell_0$-norm, and quasi sparsity is captured by the number of near-zero counts. See Subsection 1.1 for the formulation.

The 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 an overabundance of zeros include agriculture [21], environmental sciences [1], and manufacturing [34]. Data with an overabundance of near-zero counts include DNA sequencing and terrorist attacks [12]. Another example (Japanese crime statistics) is presented in Section 4.

1.1. Problem setting and contributions

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

$$p(x \mid \theta) = \prod_{i=1}^{n} \left\{ \frac{1}{x_i!} e^{-r\theta_i}(r\theta_i)^{x_i} \right\}$$
and
$$q(y \mid \theta) = \prod_{i=1}^{n} \left\{ \frac{1}{y_i!} e^{-\theta_i} \theta_i^{y_i} \right\}.$$
Two types of parameter spaces are of interest: the exact and the quasi sparse parameter spaces. Given \( s \in (0, n) \), the exact sparse parameter space is described as \( \Theta[s] := \{ \theta \in \mathbb{R}_+^n : \| \theta \|_0 \leq s \} \), where \( \| \cdot \|_0 \) is the \( \ell_0 \)-norm given by \( \| \theta \|_0 := |\{ i : \theta_i > 0 \}| \); Given \( s \in (0, n) \) and a threshold \( \varepsilon > 0 \), the quasi sparse parameter space 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 \( \varepsilon \) 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 \( \theta \). Similarly, the 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 \( n \to \infty \) and \( \eta_n := s/n = s_n/n \to 0. \) The values of \( s \) and \( \varepsilon \) possibly depend on \( n \) and thus in what follows the dependence on \( n \) is often expressed as \( s = s_n, \varepsilon = \varepsilon_n, \) and \( \eta = \eta_n. \)

The main theoretical contributions of this paper are summarized as follows: (i) The 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 in describing the ratio of the predictive minimax risks to the estimative minimax risks as shown in Proposition 2.1. This constant highlights the interesting parallel between Gaussian and Poisson decision theories as discussed in Subsection 1.2; (ii) A class of Bayes predictive densities attaining the exact asymptotic minimaxity is presented in Theorem 2.2; (iii) A class of exact asymptotically minimax predictive densities that are adaptive to an unknown
sparsity is also presented in Theorem 2.3; (iv) The proposed predictive densities are shown to attain the exact asymptotic minimaxity in the settings where current observations are missing completely at random (MCAR) in Section 3. Consideration for adaptation and for missing observations are important in application. In the two real-data examples in Section 4, there is no prior knowledge of \( s_n \), and there exhibit missing values of observations; hence we 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 with improper slab priors. The proposed class involves two ideas. The first idea is to control 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 2.2 indicates. To obtain predictive densities that are not only asymptotically minimax but are also easily implemented by the exact sampling, spike-and-slab priors with improper slab priors are used. The second idea is to specify the scale of an improper slab prior so as to minimize an upper bound of the minimax risks.

In the objective Bayesian literature, it is well-known that the scale of improper priors having the different dimensions within their mixture affects the posterior distribution unlike a single improper prior; see [23, 35, 13]. The scale is uniquely chosen to minimize an upper bound of supremum risks and is dependent on the constant \( r \). See Subsection 2.2 for more details.

1.2. Literature review

There is a rich literature on constructing predictive densities 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, 43, 3, 44]. The first quantitative comparison of Bayes and plug-in predictive densities in a wide class of parametric models is [28]. [28] 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 [22] for asymptotic expansions of Bayes predictive densities. Minimax Bayes predictive densities for unconstrained parameter spaces are studied in [36, 4]. These predictive densities under parametric constraints are studied
Shrinkage priors for Bayes predictive densities under Gaussian models are investigated in [29, 19, 27, 40]; see also [26, 5] for the cases where the variances are unknown. Shrinkage priors for Bayes predictive densities under Poisson models are developed in [30, 32]. The cases under α-divergence losses are covered by [11, 46, 39, 50].

Relatively little is known about constructing predictive densities in high dimensions. [42, 41] elegantly construct an asymptotically minimax predictive density for sparse Gaussian models. [47] obtained an asymptotically minimax predictive density for nonparametric Gaussian regression models under Sobolev constraints; thereafter, [49] obtained an adaptive minimax predictive density for these models. See also [48]. 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. Poisson models exhibit several correspondences to Gaussian models. [9, 24, 25, 38] find the correspondence in estimation of means using the re-scaled squared loss defined as \( \sum_{i=1}^{n} \theta_i^{-1}(\theta_i - \hat{\theta}_i(X))^2 \). [20, 30, 31, 32] find the correspondence in prediction using the Kullback–Leibler loss. In particular, [25, 38] find it in the exact asymptotic minimaxity under ellipsoidal and rectangle constraints in high-dimensional Poisson models using the re-scaled squared loss. In spite of the elegant correspondence in [25, 38], 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 exact asymptotic minimaxity in both estimation and prediction within sparse Poisson models, which are cleanly parallel to the result for sparse Gaussian models by [42]; see Subsection 2.2 for detailed discussions. Here we present a new strategy in constructing a predictive density, which covers several new topics in predictive density under sparsity constraints: Quasi-sparsity; The adaptation to sparsity; Missing Completely At Random (MCAR). See real-data examples in Subsection 4.2 for the importance of these topics.

Relatively scarce are theoretical studies of zero-inflated or quasi zero-inflated Poisson models in high dimensions in spite of their importance. [12] constructs 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 our own opinion, a prior should be constructed according to the context in which the prior is used. Our priors adopt the information of a predictive setting where they are used, i.e., the constant \( r \). This indicates our priors are suitable in prediction. We confirm in Section 4 that our priors
broadly outperform their priors in predictive density. In contrast, we consider that their priors would be more suitable than our priors in multiple testing or in interpreting shrinkage factors.

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 $n$. 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 $f : \mathbb{N}^n \times \mathbb{N}^n \rightarrow \mathbb{R}$, 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 $g : \mathbb{N} \rightarrow \mathbb{R}$, 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 for sparse Poisson models without missing observations. Theorems 2.1-2.3 are the main theorems. Theorem 2.1 provides exact asymptotic minimax risks; Theorem 2.2 presents a class of Bayes predictive densities attaining the exact asymptotic minimaxity; Theorem 2.3 presents a class of adaptive predictive densities. Section 3 provides the description of problem settings and theoretical results in predictive density for sparse Poisson models with MCAR settings. Section 4 reports simulation studies and applications to real data. Section 5 presents the outlines of Theorems 2.1 and 3.1. All proofs of theorems and propositions for Section 2 are provided in Appendix A. All proofs of theorems and propositions for Section 3 are provided in Appendix B.

2. Predictive density for sparse Poisson models

2.1. Main results

Results for sparse Poisson models without missing observations 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 subsequent subsection.

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

\[
C := C_r = \left( \frac{r}{r+1} \right)^r \left( \frac{1}{r+1} \right).
\]

**Theorem 2.1.** Fix \( r \in (0, \infty) \) and fix a sequence \( s_n \in (0, n) \) such that \( s_n/n = o(1) \).

(a) For the exact sparse parameter space \( \Theta[s_n] \), the following holds:

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

(b) For the quasi sparse parameter space \( \Theta[s_n, \varepsilon_n] \) with any shrinking sequence \( \varepsilon_n > 0 \) such that \( \varepsilon_n = o(\eta_n) \), the following holds:

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

The implication of this theorem in comparison with the estimative minimax risk is discussed in Subsection 2.2. The proof outline of this theorem is provided in Section 5. The full proof is provided in Appendices A.1-A.2.

The second theorem provides a class of Bayes predictive densities that attain the exact asymptotic minimaxity with the knowledge of sparsity \( s_n \). 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 \( \delta_0 \) is the Dirac measure centered at 0. The corresponding Bayes predictive density is well-defined and can be written as

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

where

\[
\omega_i := \begin{cases} 
\frac{1}{1 + \{1 + h \Gamma(\kappa)/r^\kappa\}} & \text{if } x_i = 0, \\
0 & \text{if } x_i \geq 1.
\end{cases}
\]
Here, the coordinate-wise marginal distribution of \( q_{\Pi[h,\kappa]} \) is just a zero-inflated negative binomial distribution, and sampling from \( q_{\Pi[h,\kappa]} \) is easy. Let
\[
L^* := L_{r,\kappa}^* = \frac{\mathcal{C}}{K} \quad \text{with} \quad K := \frac{r^{-\kappa} - (r + 1)^{-\kappa}}{\kappa}.
\]

**Theorem 2.2.** Fix \( r \in (0, \infty) \) and \( \kappa > 0 \). Fix also a sequence \( s_n \in (0, n) \) such that \( \eta_n = s_n/n = o(1) \). The predictive density \( q_{\Pi[L\eta_n,\kappa]} \) with \( L > 0 \) and \( \kappa > 0 \) satisfies
\[
\begin{align*}
\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[L\eta_n,\kappa]}) &\leq Cs_n \log(\eta_n^{-1}) - Cs_n \log L + Ks_n L + \Upsilon_1 \\
\sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[L\eta_n,\kappa]}) &\leq Cs_n \log(\eta_n^{-1}) - Cs_n \log L + Ks_n L + \Upsilon_2
\end{align*}
\]
and thus is asymptotically minimax: two asymptotic equalities
\[
\begin{align*}
\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[L\eta_n,\kappa]}) &\sim \mathcal{R}(\Theta[s_n]) \\
\sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[L\eta_n,\kappa]}) &\sim \mathcal{R}(\Theta[s_n, \varepsilon_n])
\end{align*}
\]
hold as \( n \to \infty \), where \( \varepsilon_n > 0 \) is a shrinking sequence such that \( \varepsilon_n = o(\eta_n) \), and \( \Upsilon_1 \) and \( \Upsilon_2 \) represent terms that are independent of \( L \) or that are \( O(s_n \eta_n) \). Furthermore, \( L^* \) minimizes the right hand sides in (1) with respect to \( L \).

The idea behind the proposed class with the scale \( L^* \) is discussed in Subsection 2.2. The proof is provided in Appendix A.3.

The third theorem presents adaptive minimax predictive densities based on the class \( \{\Pi[h,\kappa] : h > 0, \kappa > 0\} \) for sparse Poisson models. Let \( \bar{s}_n \) be given by \( \max\{1, \#\{i : X_i \geq 1, i = 1, \ldots, n\}\} \), and let \( \bar{\eta}_n := \bar{s}_n/n \).

**Theorem 2.3.** Fix \( r \in (0, \infty) \) and \( \kappa > 0 \).

(a) The predictive density \( q_{\Pi[L^*, \bar{\eta}_n, \kappa]} \) is adaptive in the exact minimax sense on the class of exact sparse parameter spaces: for any sequence \( s_n \in (0, n) \) such that \( \inf_n s_n > 0 \) and \( s_n = o(n^{1/2}) \), we have
\[
\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[L^*, \bar{\eta}_n, \kappa]}) \sim \mathcal{R}(\Theta[s_n]) \quad \text{as} \quad n \to \infty.
\]

(b) The predictive density \( q_{\Pi[L^*, \bar{\eta}_n, \kappa]} \) is adaptive in the exact minimax sense on the class of quasi sparse parameter spaces: for any shrinking sequence \( \varepsilon_n > 0 \) such that \( \varepsilon_n = o(\eta_n) \) and for any sequence \( s_n \in (0, n) \) such that \( \inf_n s_n > 0 \) and \( s_n = o(n^{1/2}) \), we have
\[
\sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[L^*, \bar{\eta}_n, \kappa]}) \sim \mathcal{R}(\Theta[s_n, \varepsilon_n]) \quad \text{as} \quad n \to \infty.
\]
The main ingredient of the proof is discussed in Subsection 2.2. The proof is provided in Appendix A.4.

2.2. Discussions

Several discussions are provided in order. All proofs of propositions are provided in Appendix A.5.

2.2.1. Prediction and estimation

On the basis of Theorem 2.1, 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

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

hold as $n \to 0$, where $s_n \in (0, n)$ is any sequence such that $\eta_n = s_n/n = o(1)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(\eta_n)$.

![Fig 1. Predictive and estimative minimax risks for sparse Poisson models: the horizontal axis represents $r$.](image1.png)

![Fig 2. Predictive and estimative minimax risks for sparse Gaussian models: the horizontal axis represents $r$.](image2.png)

According to Theorem 2.1 and Proposition 2.1, the exact asymptotic constants of predictive and estimative minimax risks ($R(\Theta|s_n])$ 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 as $r$ decreases but is bounded by 1, while that of estimative minimax risk grows to infinity as $r$ decreases. A similar phenomenon occurs in sparse
Gaussian models as found in [42]. However, the exact constants of predictive minimax risk are quite different in Poisson and Gaussian models, and they are not intuitively derivable.

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.

2.2.2. Spike-and-slab priors with improper slab priors

For Theorem 2.2, an idea behind the construction of $\Pi(L^*\eta_n, \kappa)$ with $L^*$ defined before Theorem 2.2 is discussed. We employ “spike-and-slab” priors with improper slab priors. Spike-and-slab priors are widely used in the literature of Bayesian sparse modelings; see [18, 45, 7, 8] for the recent developments. As discussed in the literature, the tail behavior of the slab prior strongly affects the performance, and thus the choice of a slab prior is substantially important. We provide two propositions indicating the effects 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 $\eta_n = s_n/n = o(1)$. For $\mu \geq 0$, $\nu = (\nu_1, \ldots, \nu_n) \in \mathbb{R}^n_+$, and $k = (k_1, \ldots, k_n) \in \mathbb{R}^n$, let

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

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

**Proposition 2.2.** Fix $\nu \in \mathbb{R}^n_+$ and a shrinking sequence $\mu = O(\eta_n)$. Then, the following holds:

$$\inf_{k \in \mathbb{R}^n_+} \sup_{\theta \in \Theta[s_n]} \frac{R(\theta, q_{\Pi_{SS}})}{R(\Theta[s_n])} \to \infty \text{ as } n \to \infty.$$
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; 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 $\eta_n = s_n/n = o(1)$. Let

$$\Pi_P := \prod_{i=1}^{n} \left\{ (1 - \eta_n) \delta_0(d\theta_i) + \eta_n \theta_i^{-2}1_{(1, \infty)}(d\theta_i) \right\}. $$

Proposition 2.3. Fix $r \in (0, \infty)$. Then, the following holds:

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi_P}) \sim R(\Theta[s_n]) \text{ as } n \to \infty. $$

Here we shall address the predictive specification of the scale $L$ in $\Pi[L\eta_n, \kappa]$, i.e., $L^*$. As the former part of Theorem 2.2 indicates, there is a practical ambiguity in the theoretical result that $\Pi[h, \kappa]$ yields the exact asymptotically minimax predictive density as long as $h = L\eta_n$ with $L > 0$. Our choice of $L^*$ avoids this ambiguity. The upper bound (1) is obtained by applying just one inequality

$$\sup_{\theta} \{f(\theta) + g(\theta)\} \leq \sup_{\theta} f(\theta) + \sup_{\theta} g(\theta) $$

with $f, g$ to the supremum risk; hence the scale $L^*$ is close to the second-order optimal among the class $\Pi[L\eta_n, \kappa]$. This also shows that the scale of improper priors within their mixture can be specified by the predictive setting (characterized by $r$) in which priors are used. Our idea here is relevant to [23]: In [23], 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 [35]. In this light, [23] and this paper indicate that the specifications of the scale can be done from a predictive viewpoint.

2.2.3. Adaptation

Discussions for Theorem 2.3 are provided.

First, the key idea of Theorem 2.3 is to evaluate a point-wise performance of an empirical Bayes predictive density with an estimator of the sparsity plugged-in. When a convenient estimator $\hat{s}_n$ for $s_n$ is available, plugging $\hat{\eta}_n := \hat{s}_n/n$ into $\eta_n$ of $\Pi[L^*\eta_n, \kappa]$ yields the empirical Bayes predictive density $q_{\Pi[L^*\hat{\eta}_n, \kappa]}$. The next proposition (Proposition 2.4) provides a point-wise performance guarantee for $q_{\Pi[L^*\hat{\eta}_n, \kappa]}$ with an estimator $\hat{s}_n$ that satisfies the following condition.

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
\[
\max \{ E_{\theta} |\hat{s}_n| / s_n - 1, E_{\theta} |\hat{s}_n| / s_n - 1 |^2 \} < \delta.
\]

Proposition 2.4. Fix \( r \in (0, \infty) \) and \( \kappa > 0 \). Suppose that Condition 2.1 holds. There exists a positive constant \( c \) depending only on \( r, \kappa \), and \( \gamma \) such that
(a) for \( \theta \in \Theta[\hat{s}_n] \) with any sequence \( s_n \in (0, n) \), we have
\[
R(\theta, q_{\Pi[L^{\star}\hat{\eta}_n,\kappa]}) \leq R(\theta, q_{\Pi[L^{\star}\eta_n,\kappa]}) + cs_n(1 + \delta + \log s_n);
\]
(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
\[
R(\theta, q_{\Pi[L^{\star}\hat{\eta}_n,\kappa]}) \leq R(\theta, q_{\Pi[L^{\star}\eta_n,\kappa]}) + cs_n(1 + \delta) + c\xi_n,
\]
where \( \xi_n := n \{ 1 - \exp(-r\varepsilon_n) \} \{ 1 + \delta + \log s_n \} \).

The estimator \( \max \{ 1, \# \{ i : X_i \geq 1, i = 1, \ldots, n \} \} \) for \( s_n \) satisfies Condition 2.1 with a universal positive constant \( \delta \) as the proof of Theorem 2.3 indicates, which is the base of the proof of Theorem 2.3; see Appendix A.4. Practically, the choice of an estimation for \( s_n \) impacts on the performance through \( \delta \) in Condition 2.1. So, \( \hat{s}_n \) should be replaced if there exists another better estimator; see the real-data examples in Subsection 4.2.

Second, Condition 2.1 is discussed. Consider the \( O(s_n \log s_n) \)-term in Proposition 2.4. This term can be improved provided that \( \delta < 1 \). However, the assumption \( \delta < 1 \) is considered to be too stringent for \( \theta \) near 0, since for \( i = 1, \ldots, n \), it is intrinsically difficult to test whether \( \theta_i > 0 \) or \( \theta_i = 0 \) for small \( \theta_i \); Hence, paying the cost of adding the \( O(s_n \log s_n) \)-term, we obtain Proposition 2.4 for any \( \delta > 0 \). Conversely, Theorem 2.3 requires an assumption on \( s_n \). The condition on \( s_n \) in Theorem 2.3 is discussed in Section 6.

3. Predictive Density for Sparse Poisson models with Missing-Completely-At-Random settings

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

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

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

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 $\mathcal{R}(\Theta[s_n])$ and $\mathcal{R}(\Theta[s_n, \varepsilon_n])$, 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 and it 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 Po($\theta_i$). Condition 3.1 implies that for each coordinate there exists at least one observation: $r_i \geq 1$. Also, for verifying Condition 3.1 (ii) in application, refer to Section 4.

### 3.1. Main results

In this section, results for sparse Poisson models with MCAR are presented. Notations for the 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

$$C_i := C_{r_i} = \left( \frac{r_i}{r_i + 1} \right)^{r_i} \left( \frac{1}{r_i + 1} \right).$$

Let $\bar{C} := \bar{C}_n = \sum_{i=1}^{n} C_i / n$.

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

$$L = \bar{C}/\tilde{C} \quad \text{with} \quad \tilde{C} = \sum_{i=1}^{n} \left\{ r_i^{-\kappa} - (r_i + 1)^{-\kappa} \right\}/(nk).$$

The third theorem provides adaptive predictive densities. Let $\hat{s}_n := \max\{1, \#\{i : X_i \geq 1, i = 1, \ldots, n\}\}$ and let $\eta_n := \hat{s}_n/n$.

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

$$\lim_{n \to \infty} \frac{\text{plim} \ \bar{C}(\Theta[s_n])}{\{E_G[\bar{C}]s_n \log(\eta_n^{-1})\}} = 1$$

$$\lim_{n \to \infty} \frac{\text{plim} \ \bar{C}(\Theta[s_n, \varepsilon_n])}{\{E_G[\bar{C}]s_n \log(\eta_n^{-1})\}} = 1$$

hold, where $s_n \in (0, n)$ is a sequence such that $\eta_n = s_n/n = o(1)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(\eta_n)$.

**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

$$\lim_{n \to \infty} \sup_{\theta \in [s_n]} R\left( \theta, q_{\Pi[\eta_n, \kappa]} \right)/\{E_G[\bar{C}]s_n \log(\eta_n^{-1})\} = 1$$

$$\lim_{n \to \infty} \sup_{\theta \in [s_n, \varepsilon_n]} R\left( \theta, q_{\Pi[\eta_n, \kappa]} \right)/\{E_G[\bar{C}]s_n \log(\eta_n^{-1})\} = 1$$

hold, where $s_n \in (0, n)$ is a sequence such that $\eta_n = s_n/n = o(1)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(\eta_n)$.

**Theorem 3.3.** Fix $\kappa \in (0, 1]$. Assume that Condition 3.1 holds.

(a) For exact sparse parameter spaces, the following asymptotic equality holds:

for any sequence $s_n \in (0, n)$ such that $\inf_n s_n > 0$ and $s_n = o(n^{1/2})$, we have

$$\lim_{n \to \infty} \sup_{\theta \in \Theta[s_n]} R\left( \theta, q_{\Pi[\eta_n, \kappa]} \right)/\{E_G[\bar{C}]s_n \log(\eta_n^{-1})\} = 1.$$

(b) For quasi sparse parameter spaces, the following asymptotic equality holds:

for any shrinking sequence $\varepsilon_n > 0$ such that $\varepsilon_n = o(\eta_n)$ and for any sequence $s_n \in (0, n)$ such that $\inf_n s_n > 0$ and $s_n = o(n^{1/2})$, we have

$$\lim_{n \to \infty} \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R\left( \theta, q_{\Pi[\eta_n, \kappa]} \right)/\{E_G[\bar{C}]s_n \log(\eta_n^{-1})\} = 1.$$

The proof outline of Theorem 3.1 is provided in Section 5. The full proofs of all above theorems are provided in Appendices B.1-B.3.
3.2. Discussions

In this subsection, by considering the following two settings, we compare Theorem 3.1 with Theorem 2.1:

(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 $l \to 0$, $G$ is weakly convergent to the Dirac measure $\delta_r$ centered at $r$ corresponding to a non-MCAR case. In Setting (B), $G$ is weakly convergent to the Dirac measure $\delta_1$ centered at 1 as $p \to 0$, and $G$ is weakly convergent to the Dirac measure $\delta_{1+N}$ as $p \to 1$.

In Figure 3, the vertical line indicates exact constants and the 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, the vertical line indicates exact constants and the 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 the 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 $l \to 0$. 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 [15, 16] for the literature of nonparametric regression in the presence of missing observations. Theorems 1
and 2 in [15] 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 if predictors are MCAR, the minimax MISE gets smaller as the missing probability approaches zero.

Next, we provide a proposition useful for comparing the estimation and prediction in cases with MCAR. This proposition provides the same comparison as in Subsection 2.2.

**Proposition 3.1.** Under Condition 3.1, two asymptotic equalities
\[
\lim_{n \to \infty} \frac{\mathcal{E}(\Theta[s_n]) \{\mathbb{E}[r^{-1}_1] e^{-1}s_n \log(\eta_n^{-1})\}}{\mathbb{E}G[r^{-1}_1] e^{-1}s_n \log(\eta_n^{-1})} = 1
\]
\[
\lim_{n \to \infty} \frac{\mathcal{E}(\Theta[s_n, \varepsilon_n]) \{\mathbb{E}[r^{-1}_1] e^{-1}s_n \log(\eta_n^{-1})\}}{\mathbb{E}G[r^{-1}_1] e^{-1}s_n \log(\eta_n^{-1})} = 1
\]
hold, where \(s_n \in (0, n)\) is a sequence such that \(\eta_n = s_n / n = o(1)\), and \(\varepsilon_n > 0\) is any shrinking sequence such that \(\varepsilon_n = o(\eta_n)\).

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**

Consider a sparse Poisson model without MCAR settings. Parameter \(\theta\) and observations \(X\) and \(Y\) are drawn from
\[
\theta_i \sim \nu_i e_{S,i} (i = 1, \ldots, n),
\]
\[
X \mid \theta \sim \otimes_{i=1}^n \text{Po}(r\theta_i),
\]
\[
Y \mid \theta \sim \otimes_{i=1}^n \text{Po}(\theta_i),
\]
respectively. Here,

- \(\nu_1, \ldots, \nu_n\) 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;
- \(\nu_1, \ldots, \nu_n\) and \(S\) are independent.
Here for a subset $J \subset \{1, \ldots, n\}$, $e_j$ indicates the vector whose $i$-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 C for the results with different choices of $(n, s, r)$.

We compare the following four predictive densities:

- The proposed predictive density based on $\Pi[L^*\hat{\eta}_n, \kappa]$;
- The Bayes predictive density based on the shrinkage prior proposed in [30];
- The Bayes predictive density based on the Gauss hypergeometric prior proposed in [12];
- The plug-in predictive density based on an $\ell_1$-penalized estimator.

For the proposed predictive density with $\Pi[L^*\hat{\eta}_n, \kappa]$, an estimator $\hat{s}_n$ for sparsity is the simple estimator described before Theorem 2.3. The second predictive density is shown in [30] to dominate the Bayes predictive density based on the Jeffreys prior. This predictive density has a hyper-parameter $\beta$ and in simulation studies it is fixed to be 1. The third predictive density employs the global-local prior proposed in [12] and the specification of the hyper-parameters follows the online support pages the authors provide.

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

- the mean of the $\ell_1$ 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 $X$, 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. The Bayes predictive density proposed in [12] is abbreviated to GH. The Bayes predictive density proposed in [30] is abbreviated to K04. The plug-in density based on an $\ell_1$-penalized estimator with regularization parameter $r\lambda$ is abbreviated to $\ell_1 (\lambda)$. The abbreviation $\ell_1$ distance represents a mean $\ell_1$ 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 $\ell_1$ 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 re-
Table 1
Comparison of predictive densities without MCAR and with \((n, s, r) = (200, 5, 1)\): the \(\ell_1\) distance, PLL, and 90\%CP represent the mean \(\ell_1\) 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 | \(\ell_1 (\lambda = 0.1)\) |
|---|---|---|---|---|
| \(\ell_1\) distance | 18.8 (5.8) | 21.9 (6.8) | 104 (4.9) | 96.5 (8.1) | 22.1 (7.8) |
| PLL | -15.4 (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) | 92.0 (1.5) | 40.5 (24.4) | 49.4 (21.6) |

Table 2
Comparison of predictive densities without MCAR and with \((n, s, r) = (200, 5, 20)\): the \(\ell_1\) distance, PLL, and 90\%CP represent the mean \(\ell_1\) 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 | \(\ell_1 (\lambda = 0.1)\) |
|---|---|---|---|---|
| \(\ell_1\) distance | 14.0 (4.9) | 14.5 (4.5) | 15.7 (1.7) | 22.5 (5.2) | 14.1 (4.5) |
| PLL | -13.3 (1.6) | -13.5 (1.5) | -15.6 (1.5) | -21.6 (2.2) | -Inf |
| 90\%CP (%) | 90.0 (0.0) | 89.4 (0.0) | 97.6 (0.7) | 97.5 (1.4) | 86.3 (3.9) |

gard 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 \(\ell_1\)-penalized estimator is too narrow to cover future observations. This is mainly because for this plug-in predictive density, an \(\ell_1\)-penalized estimator returns zero for a coordinate at which the current observation is zero and most of the marginal predictiev 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 \(\ell_1\)-penalized estimator.

4.1.2. Sparse Poisson model with Missing-Completely-At-Random settings

Consider a sparse Poisson model with MCAR. Parameter \(\theta\) and observations \(X\) and \(Y\) are drawn in the following way:

\[
\theta_i \sim \nu e_{S,i} \quad (i = 1, \ldots, n), \quad X \mid \theta \sim \times_{i=1}^{n} \text{Po}(r_i \theta_i), \quad Y \mid \theta \sim \times_{i=1}^{n} \text{Po}(\theta_i), \quad \text{and} \quad X \perp Y \mid \theta,
\]

where \(\nu_1, \ldots, \nu_n\) and \(S\) follow the same distributions as those in the previous subsection. In addition, \(r_i - 1 \quad (i = 1, \ldots, n)\) are independently drawn from the binomial distribution \(\text{Bi}(m, p)\) with the parameters \((m, p)\) 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 four predictive densities:

- The proposed predictive density based on \(\Pi[\hat{\eta}_n, \kappa]\);
- The Bayes predictive density based on the shrinkage prior proposed in [32];
- The Bayes predictive density based on the Gauss hypergeometric prior proposed in [12];
- The plug-in predictive density based on an \(\ell_1\)-penalized estimator.

An estimator \(\hat{s}_n\) is determined in the same manner as in the previous subsection. In [32], the second 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 \(\beta_i\) of the second predictive density is fixed to be 1.

In comparing the performance, we use the weighted \(\ell_1\) distance between the mean of the predictive density and a future observation. Here the weighted \(\ell_1\) 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 \(\ell_1\) distance.

**Table 3**

Comparison of predictive densities with MCAR and with \((n, s, m, p) = (200, 5, 1, 0.9)\): the \(W-\ell_1\) distance, PLL, and 90\%CP represent the weighted mean \(\ell_1\) 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.

| \(\Pi[\hat{\eta}_n, 0.1]\) | \(\Pi[\hat{\eta}_n, 1.0]\) | GH | K15 | \(\ell_1 (\lambda = 0.1)\) |
|-----------------|-----------------|---|----|------------------|
| \(W-\ell_1\) distance | 17.1 (5.6) | 26.8 (7.3) | 48.3 (5.6) | 25.6 (8.7) | 17.2 (5.8) |
| PLL | -14.7 (2.2) | -16.8 (2.7) | -42.3 (2.0) | -18.2 (3.4) | -Inf |
| 90\%CP (%) | 91.3 (0.1) | 71.8 (0.2) | 100 (0.0) | 61.7 (20.9) | 68.6 (13.6) |

**Table 4**

Comparison of predictive densities with MCAR and with \((n, s, m, p) = (200, 5, 10, 0.9)\): the \(W-\ell_1\) distance, PLL, and 90\%CP represent the weighted mean \(\ell_1\) 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.

| \(\Pi[\hat{\eta}_n, 0.1]\) | \(\Pi[\hat{\eta}_n, 1.0]\) | GH | K15 | \(\ell_1 (\lambda = 0.1)\) |
|-----------------|-----------------|---|----|------------------|
| \(W-\ell_1\) distance | 12.3 (4.2) | 12.4 (4.2) | 17.3 (1.8) | 15.2 (4.2) | 13.0 (4.4) |
| PLL | -12.2 (1.7) | -12.2 (1.7) | -19.3 (1.7) | -14.0 (1.9) | -Inf |
| 90\%CP (%) | 88.3 (4.2) | 88.5 (3.9) | 51.7 (8.6) | 78.3 (7.3) | 84.8 (4.6) |

Tables 3 and 4 show the results. In addition to the abbreviations used in the previous subsection, the abbreviation \(W-\ell_1\) distance represents a mean weighted
We see that the weighted $\ell_1$ distances by the proposed predictive density based on $\Pi[\hat{L}_{\eta_n}, 0.1]$ and the plug-in predictive density based on the $\ell_1$-penalized estimator are in the smallest level of all predictive densities compared here. As is the case without MCAR, the proposed predictive density with $\Pi[\hat{L}_{\eta_n}, 0.1]$ broadly has the coverage probability that is relatively close to the nominal level, whereas the other predictive densities (with one exception) do not.

In conclusion, regardless of whether there exists a missing of observations or not, these simulation studies show that the proposed predictive density is highly effective not only in point prediction but also in uncertainty quantification for future observations. Also, they suggest that the proposed predictive density with $\kappa = 0.1$ is stable than that with $\kappa = 1.0$.

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 for this analysis comes from the importance of taking measures against future crimes by utilizing 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 [14]. This database reports the total numbers of crimes in Tokyo Prefecture. They are classified by a town and also by the 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 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 occurred no pickpockets in white-colored towns. As seen from Figure 5, 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
Fig 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): there have occurred more pickpockets in a deeper-colored town over 6 years. There have occurred no pickpockets in white-colored towns.

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 Subsection 4.1, we compare the proposed predictive density $q_{\Pi}\{\hat{\eta}_{n}, \kappa\}$ with the three existing predictive densities, that is, the Bayes predictive density $\text{GH}$ based on a Gauss hypergeometric prior, the Bayes predictive density $\text{K04}$ based on the shrinkage prior, and the plug-in predictive density based on an $\ell_1$-regularized estimator. An estimator $\hat{s}_n$ used in $q_{\Pi}\{\hat{\eta}_{n}, \kappa\}$ is set as the simple estimator described before Theorem 2.3 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 $\kappa$ is fixed to be 0.1 as the numerical simulations suggest. We evaluate these predictive densities on the basis of the following three measures:

- The weighted $\ell_1$ distance used in Subsection 4.1.2 between the mean vector of a predictive density and the data obtained in the first half of 2018,
- the predictive log likelihood at the data obtained in the first half of 2018, and
- the means of coordinate-wise coverage probabilities of the data obtained
in the first half of 2018 based on 90%-prediction sets.

Table 5
Comparison of predictive densities in pickpocket data by the weighted \(\ell_1\) distance (W-\(\ell_1\) distance), the predictive log likelihood (PLL), and the mean of coordinate-wise coverage probabilities (CCP): underlines indicate the best performances.

|                  | \(\Pi[L_{\hat{\eta}_n}, 0.1]\) | GH | K04 | \(\ell_1 (\lambda = 0.1)\) |
|------------------|-------------------------------|----|-----|----------------------------|
| W-\(\ell_1\) distance | 273                           | 293| 273 | 297                        |
| PLL              | -399                          | -399| -429| -Inf                       |
| 90\%CCP (%)      | 93.0                          | 27.0| 84.2| 93.0                       |

Table 5 shows a summary of comparisons. In all measures, the proposed predictive density \(q_{\Pi[L_{\hat{\eta}_n}, 0.1]}\) has the best scores.

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 [12]: see Section 5 in [12]. 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 [12]. We also follow the pre-process of the data described in [12].

We apply the sparse Poisson model to the numbers of rare mutant alleles as follows. We denote by \(X_i\) \((i = 1, \ldots, 551)\) 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 \(r_i\) \((i = 1, \ldots, 551)\) is the double number of individuals whose \(i\)-th location are sequenced, and \(\theta_i\) 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. Since numerous fragments of DNA sequences are sampled and read at random, \(r_i\)'s are different in general (see Appendix E); 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 \(\otimes_{i=1}^n \text{Po}(\pi\theta_i)\) with \(\pi = 121412 = 2 \times 60706.\) We compare the propose prior \(\Pi[L_{\hat{\eta}_n}, \kappa]\) with the two existing priors, that is, the Gauss hypergeometric prior in [12] and the shrinkage prior in [30]. For the proposed prior, \(\kappa\) is set to 0.1 and the estimate \(\hat{s}_n\) to be plugged in is determined by
Fig 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 black points show the medians of marginal densities.

the k-means clustering with \( k = 2 \) (the resulting value of \( \hat{s}_n \) is 17). In this study, we give the qualitative comparison using samples from predictive densities.

Figure 6 shows marginal prediction intervals at a nominal level of 50%. The prediction intervals of the proposed predictive density \( q_{\mathcal{H}[\hat{\theta}_n, 0.1]} \) and of the Bayes predictive density based on 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 degenerate at most of the locations. In lower regimes of counts, there exists a dissimilarity between \( q_{\mathcal{H}[\hat{\theta}_n, 0.1]} \) and GH. Most of the medians of the intervals constructed by the predictive density based on GH are away from 0; while those constructed by \( q_{\mathcal{H}[\hat{\theta}_n, 0.1]} \) sometimes reach zero. Owing to the coordinate-wise independence, this dissimilarity is considered to reflect on that the proposed predictive density has more flexibility than GH.

5. Proof outlines

This section provides proof outlines of Theorems 2.1 and 3.1. Appendix A provides details of the proofs for Section 2. Appendix B provides details of the proofs for Section 3.

5.1. Outline of the proof of Theorem 2.1

The outline of the proof of Theorem 2.1 (a) is provided. The proof is divided into two parts: providing upper and lower bounds on \( \mathcal{R}(\Theta|s_n]) \).
A lower bound on \( R(\Theta[s_n]) \) 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 \( \{1, 2, \ldots, n\} \) 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 \( \nu e_I \) 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.

Consider the Bayes risk based on the prior distribution \( \Pi_{B,\nu} : B(\nu) := \int R(\theta, q_{\Pi_n}) d\Pi_{B,\nu}(\theta) \).

Since \( \Pi_{B,\nu} \) is supported on \( \Theta[s_n] \), the minimax risk is bounded below by \( B(\nu) \) for any \( \nu > 0 \) and thus bounded below by \( \sup_{\nu > 0} B(\nu) : R(\Theta[s_n]) \geq \sup_{\nu > 0} B(\nu) \).

It will be shown that \( \sup_{\nu > 0} B(\nu) \geq C s_n \log \eta_n^{-1} \), from which we will have \( R(\Theta[s_n]) \geq C s_n \log \eta_n^{-1} \).

An upper bound on \( R(\Theta[s_n]) \) is derived by bounding the coordinate-wise Kullback–Leibler risk of the Bayes predictive density \( q_{\Pi^*} \) based on \( \Pi^* = \Pi[\eta_n, \kappa] \) with \( \kappa > 0 \):

\[
\rho(\lambda) := E \log \frac{\exp(-\lambda) \lambda^{Y_i} / Y_i!}{q_{\Pi^*}(Y_i \mid X_i)}, \ \lambda > 0,
\]

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

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

from which we will have the following:

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

Here, the coordinate-wise independence of the predictive density \( q_{\Pi^*} \) is used.

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

\[
R_c(\theta, \hat{\theta}) := R(\theta, q(\cdot \mid \hat{\theta})) = E \sum_{i=1}^{n} \theta_i \log \frac{\theta_i}{\hat{\theta}_i(X)} - \theta_i + \hat{\theta}_i(X).
\]
For a prior $\Pi$ of $\theta$, let
\[
\hat{\theta}_i(x; t) = \int \theta_i p(x \mid t\theta) d\Pi(\theta) / \int p(x \mid t\theta) d\Pi(\theta), \quad i = 1, 2, \ldots, n.
\]

Let $\hat{\theta}(x; t) := (\hat{\theta}_1(x; t), \ldots, \hat{\theta}_n(x; t))$.

Lemma 5.1. For a prior $\Pi$ of $\theta$, if $\hat{\theta}_i(x; t)$ based on $\Pi$ is strictly larger than 0 for any $x \in \mathbb{N}^n$ and any $t \in (r, 1 + r)$, then, the equality
\[
R(\theta, q) = \int_r^{r+1} R_e(t\theta, \hat{\theta}(x; t)) dt
\]
holds.

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

Lemma 5.1 reduces bounding the predictive risk of the Bayes predictive density to bounding the estimative risk of the corresponding Bayes estimator. Splitting the sample space for the control of the Bayes estimator together with Lemma 5.1, we will complete the proof; see Appendices A.1-A.2.

5.2. The main ingredient of proof of Theorem 3.1

The main ingredient of the proof of Theorem 3.1 is provided in this subsection. The next lemma is the main ingredient and describes the exact asymptotic minimax risks for a setting with a fixed sequence \{r_i : i = 1, \ldots, n\} satisfying the following condition:

Condition 5.1. The asymptotic equality $\bar{C} := \sum_{i=1}^{n} C_i / n \sim \sum_{i \in J} C_i / s_n$ holds for any subset $J \subset \{1, \ldots, n\}$ satisfying $|J| = s_n$.

Lemma 5.2. 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 5.1 holds. Then,

(a) for the sparse parameter space $\Theta[s_n]$, the asymptotic equality
\[
\mathcal{R}(\Theta[s_n]) \sim \mathcal{T}_{s_n} \log(\eta^{-1}_n)
\]
holds as $n \to \infty$ and $\eta_n = s_n / n \to 0$;

(b) for the quasi sparse parameter space $\Theta[s_n, \varepsilon_n]$ with any shrinking sequence $\varepsilon_n$ such that $\varepsilon_n = o(\eta_n)$, the asymptotic equality
\[
\mathcal{R}(\Theta[s_n, \varepsilon_n]) \sim \mathcal{C}_{s_n} \log(\eta^{-1}_n)
\]
holds as $n \to \infty$ and $\eta_n \to 0$. 

Using Lemma 5.2, we will complete the proof of Theorem 3.1 easily. The outline of the proof of Lemma 5.2 will be given in the subsequent section. The full proof is given in Subsection B.4.1.

5.2.1. Outline of the proof of Lemma 5.2

The outline of the proof of Lemma 5.2 is provided. The proofs of both (a) and (b) of Lemma 5.2 are divided into two parts, providing upper and lower bounds on $R(\Theta[s_n])$ and $R(\Theta[s_n, \varepsilon_n])$, as with Theorem 2.1. However, compared to the proof of Theorem 2.1, an additional analysis on the difference of $r_i$'s is required in each part of the proof.

In constructing each bound for $R(\Theta[s_n])$ (and $R(\Theta[s_n, \varepsilon_n])$), the following formula for the Kullback-Leibler risk is employed instead of Lemma 5.1. For $i = 1, 2, \ldots, n$, let $t_i = t_i(\tau)$ ($i = 1, \ldots, n$) 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 $Z_i(\tau)$ ($i = 1, \ldots, n$) be a random variable independently distributed to Po($t_i(\tau)$ $\theta_i$). The density of $Z(\tau) = (Z_1(\tau), \ldots, Z_n(\tau))$ is denoted by

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

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

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

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

$$R_{\Pi}(\theta, \hat{\theta}_{\Pi}; \tau) := \mathbb{E}_{\theta, \tau} \left[ \sum_{i=1}^{n} t_i(\tau) \left\{ \theta_i \log \frac{\theta_i}{\hat{\theta}_{\Pi,i}(Z; \tau)} - \theta_i + \hat{\theta}_{\Pi,i}(Z; \tau) \right\} \right], \quad \tau \in [0, 1],$$

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

Lemma 5.3. [32] For a prior $\Pi$ of $\theta$, if $\hat{\theta}_{\Pi}(z; \tau)$ based on $\Pi$ is strictly larger than 0 for any $z \in \mathbb{N}^n$ and any $\tau \in [0, 1]$, then, the equality

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

holds.

For completing the proof, we will split the sample space to control the behavior of the Bayes estimator as with the proof of Theorem 2.1; see Appendix B.4.1.
6. Discussion and conclusion

In this paper, we have investigated the exact asymptotically minimax Kullback–Leibler risks in sparse and quasi-sparse Poisson sequence models. We have constructed Bayes predictive densities attaining exact asymptotic minimaxity using spike-and-slab priors with improper priors. The scale of improper priors is specified from the predictive point of view. For application, we have provided exact asymptotically minimax predictive densities that are adaptive to an unknown sparsity, and have discussed the performance in MCAR settings. The simulation studies as well as applications to real data have demonstrated the efficiency of the proposed densities.

For the future work, we shall discuss the condition on $s_n$ in Theorem 2.3. Consider the condition on $s_n$ in Theorem 2.3. This condition might be unavoidable if an estimator for $s_n$ is used, which is relevant to [6, 10]. For sparse Gaussian models, [6] constructed an adaptive estimator for $s_n$ in the regime $s_n = o(n^{1/2})$; while [6] removed a polynomially small proportion of non-zero components to construct an efficient estimator for $s_n$ in the regime $s_n \sim n^{\gamma}$ with $\gamma > 1/2$. [10] also found this phase transition in the estimation of non-smooth functionals. In addition, this condition on $s_n$ is conjectured to be unnecessary if another route is taken for constructing adaptive predictive densities. Either conjecture requires further theoretical developments in predictive density under sparsity constraints.

7. Acknowledgements

This research was supported by JST CREST (JPMJCR1763), MEXT KAKENHI (16H06533), and the Grant-in-Aid for Research Activity start-up (17H06570) from the JSPS.

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. Boisbunon and Y. Maruyama. Inadmissibility of the best equivariant predictive density in the unknown variance case. *Biometrika*, 101:733–740, 2014.

[6] A. Carpentier and N. Verzelen. Adaptive estimation of the sparsity in the Gaussian vector model. *Ann. Statist.*, 47:93–126, 2019.

[7] I. Castillo and R. Mismer. Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Statist.*, 12:3953–4001, 2018.

[8] I. Castillo and B. Szabó. Spike and slab empirical Bayes sparse credible sets. *To appear in Bernoulli*, 2019.

[9] L. Clevenson and J. Zidek. Simultaneous estimation of the means of independent Poisson laws. *J. American Statist. Assoc.*, 70:698–705, 1975.

[10] O. Collier, L. Comminges, and A. Tsybakov. On estimation of nonsmooth functionals of sparse normal means. arXiv:1805.10791.

[11] J. Corcuera and F. Giummolè. A generalized Bayes rule for prediction. *Scand. J. Statist.*, 26:266–279, 1999.

[12] J. Datta and D. Dunson. Bayesian inference on quasi-sparse count data. *Biometrika*, 103:971–983, 2016.

[13] P. Dawid and M. Musio. Bayesian model selection based on proper scoring rules. *Bayesian Anal.*, 10:479–499, 2015.

[14] The Metropolitan Police Department. The number of crimes in Tokyo prefecture by town and type. [http://www.keishicho.metro.tokyo.jp/about_mpd/jokyo_tokel/jokyo/ninchikensu.html](http://www.keishicho.metro.tokyo.jp/about_mpd/jokyo_tokel/jokyo/ninchikensu.html).

[15] S. Efromovich. Nonparametric regression with predictors missing at random. *J. American Statist. Assoc.*, 106:306–319, 2011.

[16] S. Efromovich. Adaptive nonparametric density estimation with missing observations. *J. Statist. Plann. Infer.*, 143:637–650, 2013.

[17] D. Fourdrinier, É. Marchand, A. Righi, and W. Strawderman. On improved predictive density estimation with parametric constraints. *Electron. J. Statist.*, 5:172–191, 2011.

[18] E. George and D. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87:731–747, 2000.

[19] E. George, F. Liang, and X. Xu. Improved minimax predictive densities under Kullback–Leibler loss. *Ann. Statist.*, 34:78–91, 2006.

[20] M. Ghosh and M.-C. Yang. Simultaneous estimation of poisson means under entropy loss. *Ann. Statist.*, 16:278–291, 1988.

[21] D. Hall. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56:1030–1039, 2002.

[22] J. Hartigan. The maximum likelihood prior. *Ann. Statist.*, 26:2083–2103,
1998.

[23] J. Hartigan. Bayesian regression using Akaike priors. New Haven, CT, Yale University, Preprint, 2002.

[24] I. Johnstone. Admissibility, difference equations and recurrence in estimating a Poisson mean. *Ann. Statist.*, 12:1173–1198, 1984.

[25] I. Johnstone and B. MacGibbon. Minimax estimation of a constrained Poisson vector. *Ann. Statist.*, 20:807–831, 1992.

[26] K. Kato. Improved prediction for a multivariate normal distribution with unknown mean and variance. *Ann. Ins. Statist. Math.*, 61:531–542, 2009.

[27] K. Kobayashi and F. Komaki. Bayesian shrinkage prediction for the regression problem. *J. Multivariate Anal.*, 99:1888–1905, 2008.

[28] F. Komaki. On asymptotic properties of predictive distributions. *Biometrika*, 83:299–313, 1996.

[29] F. Komaki. A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, 88:859–864, 2001.

[30] F. Komaki. Simultaneous prediction of independent Poisson observables. *Ann. Statist.*, 32:1744–1769, 2004.

[31] F. Komaki. A class of proper priors for Bayesian simultaneous prediction of independent Poisson observables. *J. Multivariate Anal.*, 97:1815–1823, 2006.

[32] F. Komaki. Simultaneous prediction for independent Poisson processes with different durations. *J. Multivariate Anal.*, 141:35–48, 2015.

[33] 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.

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

[35] G. Leung and A. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. on Information theory*, 52:3396–3410, 2006.

[36] 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.

[37] 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.

[38] B. MacGibbon. Minimax estimation over hyperrectangles with implications in the Poisson case. *IMS Collections*, 6:32–42, 2010.

[39] Y. Maruyama and W. Strawderman. Bayesian predictive densities for linear
regression models under $\alpha$-divergence loss: Some results and open problems. *IMS Collections*, 8:42–56, 2012.

[40] T. Matsuda and F. Komaki. Singular value shrinkage priors for Bayesian prediction. *Biometrika*, 102:843–854, 2015.

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

[42] G. Mukherjee and I. Johnstone. Exact minimax estimation of the predictive density in sparse gaussian models. *Ann. Statist.*, 43:937–961, 2015.

[43] G. Murray. A note on the estimation of probability density functions. *Biometrika*, 64:150–152, 1977.

[44] V. Ng. On the estimation of parametric density function. *Biometrika*, 67:505–506, 1980.

[45] V. Ročková and E. George. The Spick-and-Slab LASSO. *J. Amer. Statist. Assoc.*, 113:431–444, 2018.

[46] T. Suzuki and F. Komaki. On prior selection and covariate shift of $\beta$-Bayesian prediction under $\alpha$-divergence risk. *Comm. Statist. Theory and Methods*, 39:1655–1673, 2010.

[47] X. Xu and F. Liang. Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli*, 16:543–560, 2010.

[48] X. Xu and D. Zhou. Empirical Bayes predictive densities for high-dimensional normal models. *J. Multivariate Anal.*, 102:1417–1428, 2011.

[49] K. Yano and F. Komaki. Asymptotically minimax prediction in infinite sequence models. *Electron. J. Statist.*, 11:3165–3195, 2017.

[50] F. Zhang, Y. Shi, H. Ng, and R. Wang. Information geometry of generalized Bayesian prediction using $\alpha$-divergence as loss functions. *IEEE Tran. on Information Theory*, 64:1812–1824, 2018.
Supplementary material of “Exact Minimax Predictive Density for Sparse Count Data” by K. Yano, R. Kaneko and F. Komaki

Appendix A: Proofs for Section 2

A.1. Proof of Theorem 2.1 (a)

The proof is divided into two steps: constructing a lower bound on $R(\Theta[s_n])$ and constructing an upper bound on $R(\Theta[s_n])$.

Step 1: Lower bound on $R(\Theta[s_n])$

We begin with deriving the explicit form of $\hat{\theta}_{B,\nu,i}$. The marginal distribution of $X$ based on $\Pi_{B,\nu}$ is supported on $X_1 \times \ldots \times X_{s_n} \times \{0\}^{n-m_ns_n}$, where $X_j := \{x^{(j)} := (x_1,\ldots,x_{m_n}) : \|x^{(j)}\|_0 \leq 1 \} \ (j = 1, 2, \ldots, s_n)$. For each $j = 1, \ldots, s_n$ and for each $x^{(j)} \in X_j$, the Bayes formula gives the explicit form of the Bayes estimator $\hat{\theta}_{B,\nu,i}$:

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

Since for each $j = 1, \ldots, s_n$, the equality $\sum_{i=1}^{m_n} \hat{\theta}_{B,\nu,i}(X^{(j)}) = \nu$ holds for almost all $X^{(j)}$, the above explicit form of $\hat{\theta}_{B,\nu,i}$ yields $R_e(t\theta, t\hat{\theta}_{B,\nu}) \geq s_ne^{-r\nu}t\nu \log \lfloor \eta_n^{-1} \rfloor$ for $\theta$ in the support of $\Pi_{B,\nu}$. Together with Lemma 5.1, this inequality gives

\[
R(\theta, q_{B,\nu}) = \int_{t}^{t+1} R_e(t\theta, t\hat{\theta}_{B,\nu}) dt \geq \{e^{-r\nu} - e^{-(r+1)\nu}\}s_n \log \lfloor \eta_n^{-1} \rfloor
\]

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

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

Maximizing the rightmost hand side in the above inequality with respect to $\nu$, we obtain the desired lower bound $\mathcal{R}(\Theta[s_n]) \geq C s_n \log \lfloor \eta_n^{-1} \rfloor$, which completes Step 1.
Step 2: Upper bound on $\mathcal{R}(\Theta[s_n])$

The explicit form of $\hat{\theta}_{\Pi^*}(\cdot; t)$ is derived ahead. It 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

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

$$\hat{\theta}_{\Pi^*}(x_1; t) = \frac{x_1 + \kappa}{t}, \quad x_1 \geq 1.$$

The derivations are as follows. From the Bayes formula, we have

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

where we use the convention that $0^0 = 1$. Substituting $x_1 = 0$ into equality (2) yields the first identity. Substituting $x_1 \geq 1$ into equality (2) yields the second identity.

The explicit form of $\hat{\theta}_{\Pi^*}(\cdot; t)$ will give an upper bound on $\rho(\cdot)$. For $\lambda > 0$ and $t \in (r, r+1)$, let

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

From the explicit form of $\hat{\theta}_{\Pi^*}(\cdot; t)$, we have, for $x_1 = 0$,

$$\hat{\rho}(\lambda, x_1; t) = t \left\{ \lambda \log \eta_n^{-1} + \lambda \log \lambda - \lambda + \lambda \log(t^{x_1+1} + \eta_n t) + \frac{\eta_n}{t^{x_1+1} + \eta_n t} \right\},$$

and we have, for $x_1 \geq 1$,

$$\hat{\rho}(\lambda, x_1; t) = t \left\{ \lambda \log \left( \frac{t\lambda}{x_1 + \kappa} \right) - \lambda + \frac{x_1 + \kappa}{t} \right\}.$$

Together with Lemma 5.1, these give

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

$$\leq \int_r^{r+1} e^{-t\lambda} \left\{ \lambda \log \eta_n^{-1} + \lambda \log \lambda - \lambda + \lambda \log(t^{x_1+1} + \eta_n t) + \frac{\eta_n}{t^{x_1+1} + \eta_n t} \right\} \, dt$$

$$+ \int_r^{r+1} \mathbb{E}_{t\lambda} 1_{X_1 \geq 1} \left[ \lambda \log \left( \frac{t\lambda}{X_1 + \kappa} \right) - \lambda + \frac{X_1 + \kappa}{t} \right] \, dt$$

for $\lambda > 0$, and thus yield

$$\rho(\lambda) \leq \left\{ e^{-\lambda} - e^{-(r+1)\lambda} \right\} \log \eta_n^{-1} + e^{-\lambda} \lambda \log \lambda + \left\{ e^{-\lambda} - e^{-(r+1)\lambda} \right\} \log C$$

$$+ K\eta_n + \frac{F(\lambda)}{r} + \kappa(1 - e^{-(r+1)\lambda}) \log \left\{ \frac{r + 1}{r} \right\},$$

(3)
where $K$ is defined right before Theorem 2.2, $C := \{(r + 1)^\kappa + \eta_n(r + 1)\}$, and $F(\lambda) = F(\lambda, \kappa)$ is defined as follows: for $\kappa \geq 1$,
\[
F(\lambda, \kappa) := \lambda \log(1 - e^{-\lambda}),
\]
for $\kappa < 1$ and $\lambda \leq 1$,
\[
F(\lambda, \kappa) := 0,
\]
and for $\kappa < 1$ and $\lambda > 1$,
\[
F(\lambda, \kappa) := \lambda \log(1 - e^{-\lambda}) + \lambda e^{-\lambda^{1/2}} \log \kappa^{-1} + \lambda \log \left\{ 1 + \frac{1 - \kappa}{\lambda - \lambda^{3/4} + \kappa} \right\}.
\]
Here for $\kappa \geq 1$, we use
\[
E_{\lambda} \left[ \lambda \log \frac{\lambda}{X_1 + \kappa} \right] \leq E_{\lambda} \left[ \lambda \log \frac{\lambda}{X_1 + 1} \right] \leq \lambda \log E_{\lambda} \left[ \frac{\lambda}{X_1 + 1} \right] \leq \lambda \log(1 - e^{-\lambda}),
\]
where the second inequality follows from Jensen’s inequality and the last inequality follows since $E_{\lambda}[1/(X_1 + 1)] = (1 - e^{-\lambda})/\lambda$. For $\kappa < 1$, we use the following inequality with $\lambda > 1$:
\[
E_{\lambda} \left[ \lambda \log \frac{\lambda}{X_1 + \kappa} \right] \leq \lambda \log(1 - e^{-\lambda}) + E_{\lambda} \left[ \lambda \log \frac{X_1 + 1}{X_1 + \kappa} \right] \\
\leq \lambda \log(1 - e^{-\lambda}) + E_{\lambda} \left[ \lambda \log \left\{ 1 + \frac{1 - \kappa}{X_1 + \kappa} \right\} \right] \\
\leq \lambda \log(1 - e^{-\lambda}) + \lambda e^{-\lambda^{1/2}} \log \left\{ 1 + \frac{1 - \kappa}{\kappa} \right\} + \lambda \log \left\{ 1 + \frac{1 - \kappa}{\lambda - \lambda^{3/4} + \kappa} \right\},
\]
where the last inequality follows from the Bennett inequality for the Poisson distribution
\[
P(X_1 - \lambda \leq -x) \leq \exp\{-x^2/(2\lambda)\}, \ 0 \leq x \leq \lambda.
\]
Similarly, we obtain
\[
\rho(0) \leq K\eta_n. \tag{4}
\]
Maximizing the right hand side of (3) with respect to $\lambda$ gives $\sup_{\lambda > 0} \rho(\lambda) \leq (C + o(1)) \log \eta_n^{-1}$. Combined with (4), this yields the desired upper bound
\[
R(\Theta[s_n]) \leq (n - s_n)O(\eta_n) + (C + o(1))s_n \log \eta_n^{-1},
\]
which completes the proof. $\square$
A.2. Proof of Theorem 2.1 (b)

The proof follows almost the same line as in the proof of Theorem 2.1 (a).

Step 1: Lower bound on $R(\Theta[s_n, \varepsilon_n])$

We start extracting an exact sparse subspace in $\Theta[s_n, \varepsilon_n]$. Taking a sufficiently large $n$ (depending only on $r$) and for $\nu \in (0, 1)$, we have

$$\{ \theta \in \mathbb{R}_+^n : \theta[1] = \cdots = \theta[s_n] = \nu, \theta[s_n+1] = \cdots = \theta[n] = 0 \} \subset \Theta[s_n, \varepsilon_n],$$

(5)

where $\theta[i]$ is the $i$-th largest component of $\{\theta_i : i = 1, \ldots, n\}$. This gives

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

which completes Step 1.

Step 2: Upper bound on $R(\Theta[s_n, \varepsilon_n])$

Recall that $N(\theta, \varepsilon) := \#\{i : \theta_i > \varepsilon\}$. By definition of $N(\theta, \varepsilon)$, we have, for an independent prior $\Pi$,

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

(6)

where $R_i$ ($i = 1, \ldots, n$) is the Kullback-Leibler risk for the $i$-th coordinate. Since $N(\theta, \varepsilon) \leq s_n$ for $\theta \in \Theta[s_n, \varepsilon_n]$, we have

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

(7)

We put a prior $\Pi^* = \Pi[\eta_n, \kappa]$ into $\Pi$ and then we have $R_i(\lambda, q_{\Pi^*, \varepsilon_n}) = \rho(\lambda)$. From (3), we have $\sup_{0 < \lambda} R_i(\lambda, q_{\Pi^*, \varepsilon_n}) \leq C \log(n/s_n)$. Observe that for sufficiently small $\lambda$ (depending only on $r$ and $\kappa$), the first, the third, and the sixth terms in the right hand side of (3) are monotonically increasing with respect to $\lambda$, while the second and the forth terms of (3) are monotonically decreasing with respect to $\lambda$; hence substituting $\lambda = \varepsilon_n$ gives each supremum of the first, the third, and the sixth terms, while letting $\lambda \to 0$ gives each supremum of the second and the forth terms. Thus using the inequality $e^{-t\lambda} \geq 1 - t\lambda$ and using the right-continuity of $\lambda \log \lambda$ at $\lambda = 0$, we obtain $\sup_{\lambda \leq \varepsilon_n} R_i(\lambda, q_{\Pi, \varepsilon_n}) = O(\varepsilon_n \log \{\eta_n^{-1}\})$, which completes the proof. \(\square\)
A.3. Proof of Theorem 2.2

We provide the proof only for \( \Theta[s_n] \) and the proof for \( \Theta[s_n, \varepsilon_n] \) is straightforward. We begin with multiplying \( \eta_n \) by \( L \) in \( \Pi^* = \Pi[\eta_n, \kappa] \) of the proof of Theorem 2.1. Then, inequalities (3) and (4) give

\[
\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[L\eta_n, \kappa]}(\cdot)) = s_n \sup_{\lambda > 0} R_1(\lambda, q_{\Pi[L\eta_n, \kappa]}(\cdot), 1) + (n - s_n) R_1(0, q_{\Pi[L\eta_n, \kappa]}(\cdot), 1) \\
\leq C s_n \log(\eta_n^{-1} L^{-1}) + \Upsilon_1 + K L \log \eta_n \\
= C s_n \log(\eta_n^{-1}) + s_n \{K L - C \log L \} + \Upsilon_1,
\]

where \( \Upsilon_1 \) is defined as the supremum of the terms except the first term in the right hand side of (3) with respect to \( \lambda \). Observe that \( \Upsilon_1 \) consists of terms independent of \( L \) or terms that are \( o(s_n) \). This completes the proof of the former part of Theorem 2.2. A simple calculus gives the latter part of Theorem 2.2 and thus we complete the proof.

A.4. Proof of Theorem 2.3

We begin with showing that \( \delta \) for \( \hat{s}_n \) is bounded from above by a constant independent of \( n \). To this end, we use the bias-variance decomposition of \( \hat{s}_n - s_n \).

**Proof for Case (a):** Consider the bias of \( \hat{s}_n - s_n \). For \( j = 1, \ldots, s_n \), \( \theta[j] \) denotes the \( j \)-th largest component of \( \{\theta_i : i = 1, \ldots, n\} \). Since the decomposition \( \#\{i : X_i \geq 1\} = \sum_{j=1}^{s_n} Z_j \) holds with \( Z_j \) \( (j = 1, \ldots, s_n) \) independent and distributed according to the Bernoulli distribution having the success probability \( 1 - \exp(-r \theta[j]) \), we have

\[
-\sum_{j=1}^{s_n} e^{-r \theta[j]} \leq \mathbb{E}_\theta \hat{s}_n - s_n \leq 1. \tag{8}
\]

Consider the variance of \( \hat{s}_n - s_n \). Using the inequalities

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

we have

\[
\mathbb{E}_\theta |\hat{s}_n - \mathbb{E}_\theta \hat{s}_n|^2 \leq 1 + \sum_{j=1}^{s_n} e^{-r \theta[j]} \left(1 - e^{-r \theta[j]} \right) \tag{9}.
\]

Combining (8) with (9) shows that \( \max\{\mathbb{E}|\hat{s}_n/s_n - 1|, \mathbb{E}|\hat{s}_n/s_n - 1|^2\} \) is bounded from above by a positive constant depending only on \( \inf_n s_n \). Together with Proposition 2.4 (a), this completes the proof of Theorem 2.3 (a). \( \square \)
Proof for Case (b): The proof is almost the same as of Case (a).
Consider the bias of \( \hat{s}_n - s_n \). Since the decomposition \( \#\{i : X_i \geq 1\} = \sum_{j=1}^{n} Z_j \) holds with \( Z_j \) (\( j = 1, \ldots, n \)) independent and distributed according to the Bernoulli distribution having the success probability \( 1 - \exp(-r \theta_{[j]}) \), and since the inequality \( \theta_{[j]} = \varepsilon_n \) holds for \( j = s_n + 1, \ldots, n \), we have
\[
- \sum_{j=1}^{s_n} e^{-r \theta_{[j]}} \leq E[\theta|\hat{s}_n] - s_n \leq \sum_{j=s_n+1}^{n} (1 - \exp(-r \varepsilon_n)) \leq 1 + rn \varepsilon_n. \tag{10}
\]
Consider the variance of \( \hat{s}_n - s_n \). Using the inequality
\[
-1 - \sum_{j=1}^{n} (Z_j - E[Z_j]) \leq \hat{s}_n - E[\theta] \hat{s}_n \leq 1 + \sum_{j=1}^{n} (Z_j - E[Z_j]),
\]
we have
\[
E[\theta|\hat{s}_n] - E[\theta] \hat{s}_n |^2 \leq 1 + \sum_{j=1}^{s_n} e^{-r \theta_{[j]}} (1 - e^{-r \theta_{[j]}}) + nr \varepsilon_n. \tag{11}
\]
Combining (10) with (11) shows that \( \max\{E[\hat{s}_n/s_n - 1], E[\hat{s}_n/s_n - 1]^2\} \) is bounded from above by a positive constant depending only on \( \inf_n s_n \). Together with Proposition 2.4 (b), this completes the proof of Theorem 2.3 (b).

A.5. Proofs of Propositions

A.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, we employ the Bayes risk maximization based on a block-independent prior \( \Pi_{\text{B},\nu} \) with \( \nu > 0 \). In bounding \( \mathcal{E}(\Theta[s_n]) \) above, we use a thresholding estimator to simplify the proof. The thresholding estimator used in this proof is defined as follows: for each \( i = 1, 2, \ldots, n \),
\[
\hat{\theta}_{T,i}(x_i) = \begin{cases} 
\eta_n & x_i = 0, \\
(x_i + 1)/r & x_i \geq 1.
\end{cases}
\]
Consider a lower bound on \( \mathcal{E}(\Theta[s_n]) \). The explicit expression of \( \hat{\theta}_{\text{B},\nu} \) with \( \nu > 0 \) gives
\[
\int R(\theta, q(\cdot | \hat{\theta}_{\text{B},\nu})) d\Pi_{\text{B},\nu}(\theta) = s_n [e^{-r \nu} \nu \log\{\nu/(\nu|\eta_n)|\} + (1 - e^{-r \nu}) \nu \log(\nu/\nu)]
= s_n \nu e^{-r \nu} \log|\eta_n^{-1}|.
\]
and thus maximizing $\int R(\theta, q(\cdot | \hat{\theta}_{\Pi, \nu}))d\Pi_{\Pi, \nu}(\theta)$ with respect to $\nu$ 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

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

we bound the minimax risk $\mathcal{E}(\Theta|s_n)$ as

$$\mathcal{E}(\Theta|s_n) \leq (n - s_n)\tilde{\rho}(0) + s_n \sup_{\lambda > 0} \tilde{\rho}(\lambda).$$

Thus, deriving two (asymptotic) inequalities

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

will complete the proof.

Consider $\sup_{\lambda > 0} \tilde{\rho}(\lambda)$. For $\lambda > 0$, $\tilde{\rho}(\lambda)$ is expressed as

$$\tilde{\rho}(\lambda) = \mathbb{E}_{r\lambda}[1_{X_1 \geq 1}(X_1)[\lambda \log(\lambda/(X_1 + 1)) - \lambda + (X_1 + 1)/r]$$

$$+ \mathbb{E}_{r\lambda}[1_{X_1 = 0}(X_1)[\lambda \log(\lambda/\eta_n)] - \lambda + \eta_n].$$

This expression gives

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

$$+ \sup_{\lambda > 0} \{\lambda e^{-r\lambda}\} \log \eta_n^{-1} + \sup_{\lambda > 0} \mathbb{E}_{r\lambda} \lambda \log r(\lambda/(X_1 + 1)).$$

From inequality (12) and from the identity

$$\mathbb{E}_{r\lambda}[r/(X_1 + 1)] = \lambda^{-1}(1 - \exp(-r\lambda))$$

for $\lambda > 0$, we have

$$\sup_{\lambda > 0} \tilde{\rho}(\lambda) \leq 1 + \eta_n + \sup_{\lambda > 0} \lambda \exp(-r\lambda) \log \eta_n^{-1}.$$ 

Thus, we obtain the asymptotic equality

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

Similarly, we get the asymptotic inequality $\tilde{\rho}(0) = O(\eta_n)$. Hence, we obtain the desired upper bound on $\mathcal{E}(\Theta|s_n)$ and complete the proof. \qed
A.5.2. Proof of Proposition 2.2

Consider bounds on $\hat{\theta}_{\Pi_{SS,1}}(X_1; t)$. The Bayes estimate $\hat{\theta}_{\Pi_{SS,1}}(X_1; t)$ has the explicit expression

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

Combining this expression of $\hat{\theta}_{\Pi_{SS,1}}(X_1; t)$ with the simple algebra that $(a + b)/(c + d) \leq b/d$ for any $(a, b, c, d) \in \mathbb{R}_+^4$ such that $ad \leq bc$ gives

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

for sufficiently large $n$ depending only on $\nu_1$ and for $X_1 \geq 1$.

Using inequality (13) and Lemma 5.1, we have

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

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

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

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

A.5.3. Proof of Proposition 2.3

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

We begin with getting the following inequalities: there exist positive constants $C_1, C_2, C_3$ depending only on $r$ such that for $i = 1, \ldots, n$ and for $t \in (r, r+1)$, we have

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

where $\hat{\theta}_{\Pi_{P,i}}(x_i; t)$ is the Bayes estimate for $i = 1, \ldots, n$. These inequalities are obtained by applying the bounds from the proof of Proposition 2.2 to the Bayes estimate for $\Pi_P$. The proof then proceeds similarly to the proof of Proposition 2.2, using these inequalities to show that $\hat{\theta}_{\Pi_P}(X_1; t)$ is a good Bayes estimator for $X_1$.
where see the description before Lemma 5.1 for the definition of \( \hat{\theta}_{\Pi_r}(\cdot) \). We derive these bounds from the explicit form of \( \hat{\theta}_r(\cdot) \)

\[
\hat{\theta}_{\Pi_r}(x_i; t) = \frac{\eta_n \int_1^\infty \lambda^x_i e^{-t\lambda} d\lambda}{\int_0^{x_i} + \eta_n \int_1^\infty \lambda x_i^{-1} e^{-t\lambda} d\lambda}
\]

and from the identity

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

For \( \lambda > 0, x_i \in \mathbb{N}, \) and \( t \in (r, r+1), \) let

\[
\hat{\rho}_r(\lambda, x_1; t) := t \lambda \log \{ \lambda / \hat{\theta}_{\Pi_r}(x_1; t) \} - t \lambda + t \hat{\theta}_{\Pi_r}(x_1; t).
\]

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

\[
\begin{align*}
\hat{\rho}_r(\lambda, x_1; t) &\leq t \{ \lambda \log \eta_n^{-1} + \lambda \log \lambda - \lambda + \lambda \log C_1 + \eta_n C_2 \}, \quad x_1 = 0, \\
\hat{\rho}_r(\lambda, x_1; t) &\leq t \{ \lambda \log \{ t/C_3 \} - \lambda + C_3 \}, \quad x_1 = 1, \\
\hat{\rho}_r(\lambda, x_1; t) &\leq t \{ \lambda \log \{ t/(x_1 - 1) \} - \lambda + (x_1 - 1 + t)/t \}, \quad x_2 \geq 2.
\end{align*}
\]

For \( \lambda > 0, \) let

\[
\rho_r(\lambda) = \mathbb{E}_{\lambda} \log \frac{\exp(-\lambda) Y_1 / Y_1!}{q_{\Pi_r}(Y_1 | X_1)}.
\]

We will show that \( \rho_r(0) = O(\eta_n) \) and \( \sup_{\lambda > 0} \rho_r(\lambda) \leq (C + o(1)) s_n \log \eta_n^{-1} \).

Lemma 5.1 gives

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

and thus it suffices to bound \( \mathbb{E}_{\lambda} [\hat{\rho}_r(\lambda, X_1; t)] \) uniformly in \( t \in (r, r+1) \). Consider the decomposition

\[
\begin{align*}
\mathbb{E}_{\lambda} [\hat{\rho}_r(\lambda, X_1; t)]
&= \mathbb{E}_{\lambda} [1_{X_1 = 0} \hat{\rho}_r(\lambda, X_1; t)] + \mathbb{E}_{\lambda} [1_{X_1 = 1} \hat{\rho}_r(\lambda, X_1; t)] + \mathbb{E}_{\lambda} [1_{X_1 \geq 2} \hat{\rho}_r(\lambda, X_1; t)].
\end{align*}
\]

Bounding the first and the second terms in the right hand side of the above decomposition is straightforward; we have

\[
\lim_{\lambda \to 0} \int_r^{r+1} \frac{\mathbb{E}_{\lambda} [1_{X_1 = 0} \hat{\rho}_r(\lambda, X_1; t)]}{t} dt = O(\eta_n),
\]

\[
\sup_{\lambda > 0} \int_r^{r+1} \frac{\mathbb{E}_{\lambda} [1_{X_1 = 0} \hat{\rho}_r(\lambda, X_1; t)]}{t} dt = C \log \eta_n^{-1} + O(1),
\]

\[
\sup_{\lambda > 0} \int_r^{r+1} \frac{\mathbb{E}_{\lambda} [1_{X_1 = 1} \hat{\rho}_r(\lambda, X_1; t)]}{t} dt = C \log \eta_n^{-1} + O(1),
\]

\[
\sup_{\lambda > 0} \int_r^{r+1} \frac{\mathbb{E}_{\lambda} [1_{X_1 \geq 2} \hat{\rho}_r(\lambda, X_1; t)]}{t} dt = C \log \eta_n^{-1} + O(1),
\]

\[
\int_r^{r+1} \frac{\mathbb{E}_{\lambda} [1_{X_1 = 0} \hat{\rho}_r(\lambda, X_1; t)]}{t} dt = O(\eta_n),
\]

\[
\int_r^{r+1} \frac{\mathbb{E}_{\lambda} [1_{X_1 = 1} \hat{\rho}_r(\lambda, X_1; t)]}{t} dt = C \log \eta_n^{-1} + O(1),
\]

\[
\int_r^{r+1} \frac{\mathbb{E}_{\lambda} [1_{X_1 \geq 2} \hat{\rho}_r(\lambda, X_1; t)]}{t} dt = C \log \eta_n^{-1} + O(1),
\]
and

\[
\lim_{\lambda \to 0} \int_r^{r+1} \frac{E_{t\lambda}[1_{X_1=1}\hat{\rho}(\lambda, X_1; t)]}{t} dt = 0,
\]

for \( \lambda > 0 \)

\[
\sup_{\lambda > 0} \int_r^{r+1} \frac{E_{t\lambda}[1_{X_1=1}\hat{\rho}(\lambda, X_1; t)]}{t} dt = O(1).
\]

Consider bounding the third term. Since \( \log\{1/(X_1 - 1)\} \leq 0 \) for \( X_1 \geq 2 \) and since

\[
E_{t\lambda}[1_{X_1 \geq 2}(X_1 - 1 + t)/t] \leq \lambda + (1 - e^{-t\lambda}), \ t \in (r, r+1),
\]

we have

\[
\lim_{\lambda \to 0} \int_r^{r+1} \frac{E_{t\lambda}[1_{X_1 \geq 2}\hat{\rho}(\lambda, X_1; t)]}{t} dt = 0.
\]

Since we have

\[
\lambda E_{\lambda} \left[ 1_{X_1 \geq 2} \log \left( \frac{\lambda}{X_1 - 1} \right) \right] = \lambda E_{\lambda} \left[ 1_{X_1 \geq 2} \log \left( \frac{\lambda}{X_1 + 1} \right) \right] + \lambda E_{\lambda} \left[ 1_{X_1 \geq 2} \log \left( \frac{X_1 + 1}{X_1 - 1} \right) \right] \leq \lambda \log(1 - e^{-\lambda}) + \lambda e^{-\lambda^{1/2}/2} \log(2) + \lambda \log \left( 1 + \frac{2}{\lambda - \lambda^{3/4} - 1} \right)
\]

for \( \lambda > 2 \) from the Bennett inequality for the Poisson distribution, and since

\[
\lambda E_{\lambda} \left[ 1_{X_1 \geq 2} \log \left( \frac{\lambda}{X_1 - 1} \right) \right] \leq 2 \log 2
\]

for \( \lambda \leq 2 \), we have

\[
\sup_{\lambda > 0} \int_r^{r+1} \frac{E_{t\lambda}[1_{X_1 \geq 2}\hat{\rho}(\lambda, X_1; t)]}{t} dt = O(1),
\]

which provides the bound on the third term and thus completes the proof.

A.5.4. Proof of Proposition 2.4 (a)

The decomposition of the Kullback–Leibler divergence gives

\[
R(\theta, q_{\Pi[L^*\hat{\eta}_n, \kappa]}) = R(\theta, q_{\Pi[L^*\eta_n, \kappa]}) + \sum_{i=1}^{n} E_{\theta} \log \left\{ \frac{q_{\Pi[L^*\eta_n, \kappa], i}(Y_i \mid X_i)}{q_{\Pi[L^*\eta_n, \kappa], i}(Y_i \mid X_i)} \right\}
\]

\[
= R(\theta, q_{\Pi[L^*\eta_n, \kappa]}) + \sum_{i \in A} E_{\theta} \log \left\{ \frac{q_{\Pi[L^*\eta_n, \kappa], i}(Y_i \mid X_i)}{q_{\Pi[L^*\eta_n, \kappa], i}(Y_i \mid X_i)} \right\}
\]

\[
+ \sum_{i \notin A} E_{\theta} \log \left\{ \frac{q_{\Pi[L^*\eta_n, \kappa], i}(Y_i \mid X_i)}{q_{\Pi[L^*\eta_n, \kappa], i}(Y_i \mid X_i)} \right\},
\]

(14)
where for $\theta \in \Theta(s_n)$, let $A := A(\theta) = \{ i : \theta_i \neq 0 \}$. In what follows, we will bound the second and the third terms in the rightmost side of the above equality.

For $i \notin A$, we have

$$
\mathbb{E}_\theta \log \left\{ \frac{q_{(L^*, \eta_n, \kappa)}}{q_{(L^*, \eta_n, \kappa, i)}}(Y_i | X_i) \right\} = \mathbb{E}_\theta \log \left\{ \frac{1 + L^* \eta_n \Gamma(\kappa)/r^\kappa}{1 + L^* \eta_n \Gamma(\kappa)/r^\kappa} \right\} + \mathbb{E}_\theta \log \left[ 1 + \{ L^* \eta_n \Gamma(\kappa)/r^\kappa \} \{ r/(r + 1) \}^\kappa \right] \\
\leq c_1 \eta_n \delta
$$

(15)

with a positive constant $c_1$ depending only on $\kappa$ and $r$, where the last inequality follows from the following two bounds:

$$
\mathbb{E}_\theta \log \left\{ \frac{1 + L^* \eta_n \Gamma(\kappa)/r^\kappa}{1 + L^* \eta_n \Gamma(\kappa)/r^\kappa} \right\} \leq L^* \eta_n \frac{\Gamma(\kappa)}{r^\kappa} \delta; \quad (16)
$$

$$
\mathbb{E}_\theta \log \left[ 1 + \{ L^* \eta_n \Gamma(\kappa)/r^\kappa \} \{ r/(r + 1) \}^\kappa \right] \leq L^* \eta_n \frac{\Gamma(\kappa)}{(r + 1)\kappa} \delta + c_2 L^* \eta_n \delta, \quad (17)
$$

where $c_2$ is a positive constant depending only on $\kappa$. Here inequality (16) follows since $\log(1 + x) \leq x$ for $x > 0$. Inequality (17) follows since

$$
\mathbb{E}_\theta \log \left[ 1 + \{ L^* \eta_n \Gamma(\kappa)/r^\kappa \} \{ r/(r + 1) \}^\kappa \right] \leq -\mathbb{E}_\theta 1_{s_n \geq \delta} \log \left[ 1 - L^* \eta_n \Gamma(\kappa)/r^\kappa \right] \leq \eta_n \left( \frac{L^* \Gamma(\kappa)}{r + 1} \right) \delta + \left( 1 + L^* \Gamma(\kappa) \right) \eta_n \left( \frac{L^* \Gamma(\kappa)}{r + 1} \right)^2 \delta,
$$

where the second inequality follows since if $s_n \geq \delta$, then we have

$$
\frac{\eta_n - \delta}{1 + \eta_n \{ L^* \Gamma(\kappa)/r^\kappa \} \{ r/(r + 1) \}^\kappa} \leq \frac{\eta_n L^* \Gamma(\kappa)/(r + 1)^\kappa}{1 + \eta_n \{ L^* \Gamma(\kappa)/r^\kappa \} \{ r/(r + 1) \}^\kappa} \leq \frac{L^* \Gamma(\kappa)}{1 + L^* \Gamma(\kappa)}
$$

and since for any $x$ such that $0 < x \leq U$ for some $0 < U < 1$,

$$
-\log(1 - x) \leq \frac{1}{2} \log(1 - x^2) + x.
$$

For $i \in A$, 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 $i \notin A$, we have

$$
\log \left\{ \frac{q_{(L^*, \eta_n, \kappa)}}{q_{(L^*, \eta_n, \kappa, i)}}(Y_i | X_i) \right\} = \log \left\{ \frac{1 + L^* \eta_n \Gamma(\kappa)/r^\kappa}{1 + L^* \eta_n \Gamma(\kappa)/r^\kappa} \right\} + \log \left[ 1 + \{ L^* \eta_n \Gamma(\kappa)/r^\kappa \} \{ r/(r + 1) \}^\kappa \right] \\
\leq \frac{\delta}{s_n} - 1 + \log[1 + L^* \Gamma(\kappa)/(r + 1)^\kappa].
$$

(18)
In Case (ii), we have
\[
\log \left\{ \frac{q_{\Pi[L^*\eta_n,\kappa]}(Y_i | X_i)}{q_{\Pi[L^*\hat{\eta}_n,\kappa]}(Y_i | X_i)} \right\} = 0. \tag{19}
\]

In Case (iii), we have
\[
\log \left\{ \frac{q_{\Pi[L^*\eta_n,\kappa]}(Y_i | X_i)}{q_{\Pi[L^*\hat{\eta}_n,\kappa]}(Y_i | X_i)} \right\} = \log(\eta_n/\hat{\eta}_n) + \log \left[ \frac{1 + L^*\hat{\eta}_n \Gamma(\kappa)/r^\kappa}{1 + L^*\eta_n \Gamma(\kappa)/r^\kappa} \right] \leq \log(s_n/\gamma) + |\hat{s}_n/s_n - 1|, \tag{20}
\]
where the last inequality follows since \(\log(s_n/\hat{s}_n) \leq \log(s_n/\gamma)\) and since \(\log(1 + x) \leq x\) for \(x > 0\). In Case (iv), we have
\[
\log \left\{ \frac{q_{\Pi[L^*\eta_n,\kappa]}(Y_i | X_i)}{q_{\Pi[L^*\hat{\eta}_n,\kappa]}(Y_i | X_i)} \right\} = 0. \tag{21}
\]

From (18)-(21), for \(i \in \mathcal{A}\), we have
\[
\mathbb{E}_{\theta} \log \left\{ \frac{q_{\Pi[L^*\eta_n,\kappa]}(Y_i | X_i)}{q_{\Pi[L^*\hat{\eta}_n,\kappa]}(Y_i | X_i)} \right\} \leq 2\mathbb{E}_{\theta}|\hat{s}_n/s_n - 1| + \log(1 + L^*\Gamma(\kappa)/(r + 1)^\kappa) + \log(s_n/\gamma). \tag{22}
\]
Combining (15) and (22) with (14) completes the proof.

A.5.5. Proof of Proposition 2.4 (b)

The proof follows almost the same line as in the proof for Proposition 2.4 (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 (22). For \(i \not\in \mathcal{A}\) and for \(x_i \geq 1\), from (19) and (21), we have
\[
\mathbb{E}_{Y_i|\theta_i} \log \left\{ \frac{q_{\Pi[L^*\eta_n,\kappa]}(Y_i | x_i)}{q_{\Pi[L^*\hat{\eta}_n,\kappa]}(Y_i | x_i)} \right\} = 0. \tag{23}
\]
For \(i \not\in \mathcal{A}\) and for \(x_i = 0\), from (18) and (20), we have
\[
\mathbb{E}_{Y_i|\theta_i} \log \left\{ \frac{q_{\Pi[L^*\eta_n,\kappa]}(Y_i | x_i = 0)}{q_{\Pi[L^*\hat{\eta}_n,\kappa]}(Y_i | x_i = 0)} \right\} \leq |\hat{s}_n/s_n - 1| + \log(1 + L^*\Gamma(\kappa)/(r + 1)^\kappa) + (1 - e^{-r\varepsilon_n})|\hat{s}_n/s_n - 1| + \log(s_n/\gamma). \tag{24}
\]
Combining (22), (23), and (24) with (14) completes the proof.
Appendix B: Proofs for Section 3

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

B.1. Proof of Theorem 3.1

Admitting that Lemma 5.2 holds, Theorem 3.1 will be proved using the following lemma:

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

$$\left| \frac{R(\Theta[s_n])}{(\mathcal{E}_G \log s_n)} - 1 \right| \leq b_n + \sqrt{\frac{1}{2n}} \log \frac{2}{\delta}$$

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

The proof is given in Subsection B.4.2.

By substituting $\delta = \delta_n = e^{-n/\log n}$ in Lemma B.1 and by using Lemma 5.2, Theorem 3.1 is derived directly.

B.2. Proof of Theorem 3.2

The main ingredient of the proof of Theorem 3.2 is Lemma B.2 for a fixed sequence $\{r_i \in (0, \infty) : i \in \mathbb{N}\}$ satisfying Condition 5.1. For the simplicity, we conduct the proof for the case when replacing $\mathcal{L}$ by 1.

Lemma B.2 states that an improper prior $\Pi[\eta_n, \kappa]$ 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 5.1.

Lemma B.2. 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 5.1 holds. The predictive density $q_{\Pi[\eta_n, \kappa]}$ attains exact asymptotic minimaxity:

$$\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi[\eta_n, \kappa]}) \sim \overline{R}(\Theta[s_n])$$
$$\sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, q_{\Pi[\eta_n, \kappa]}) \sim \overline{R}(\Theta[s_n, \varepsilon_n])$$

hold as $n \to \infty$, where $s_n \in (0, n)$ is a sequence such that $\eta_n = s_n/n = o(1)$, and $\varepsilon_n > 0$ is any shrinking sequence such that $\varepsilon_n = o(\eta_n)$.

The proof is given in Subsection B.4.3.

Admitting that Lemma B.2, Theorem 3.2 is proved by Lemmas B.1 and B.2.
B.3. Proof of Theorem 3.3

For the proof of Case (a), replacing $e^{-r\theta_j}$ in (8) and (9) by $e^{-r_j\theta_j}$ completes the proof.

For the proof of Case (b), replacing $e^{-r\theta_j}$ in (10) and (11) by $e^{-r_j\theta_j}$ and replacing $\sum_{j=s_n+1}^{n}(1-\exp(-r_j\varepsilon_n))$ in (10) by $\sum_{j=s_n+1}^{n}(1-\exp(-r_j\varepsilon_n))$ complete the proof. \hfill \Box

B.4. Proofs of Lemmas for Section 3

B.4.1. Proof of Lemma 5.2

We provide only the proof for (a) and omit the proof for (b). Following the same argument as in Subsection A.2 completes the proof for case (b).

Step 1: Lower bound on $R(\Theta[s_n])$

A lower bound on $R(\Theta[s_n])$ builds upon Bayes risk maximization based on a varied-spike block-independent prior. Let

\[ \Pi_{\text{VB},\nu} \text{ with } \nu = (\nu^{(1)}, \ldots, \nu^{(s_n)}) \in \mathbb{R}^{m_n \times m_n} \times \cdots \times \mathbb{R}^{m_n \times \cdots \times \mathbb{R}^{n-m_n s_n}} \]

be a varied-spike 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 $(\theta_1+m_n(j-1), \ldots, \theta_{m_n})$ independently according to a single spike prior with spike strength parameter $\nu^{(j)} \in \mathbb{R}^{m_n}_+$, where a single spike prior with spike strength parameter $\nu^{(j)} \in \mathbb{R}^{m_n}_+$ is the distribution of $\nu^{(j)}_I e_I$ 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.

It is worth noting that a varied-spike block-independent prior may be different from a block-independent prior since the spike strength may be varied in each coordinate.

We begin with getting the explicit form of $\hat{\theta}_{\text{VB},\nu} = (\hat{\theta}^{(1)}_{\text{VB},\nu}, \ldots, \hat{\theta}^{(s_n)}_{\text{VB},\nu})$ in order to consider the Bayes risk based on the prior distribution $\Pi_{\text{VB},\nu}$. From the Bayes formula, we have, for $j = 1, \ldots, s_n - 1$ and for $k = 1, \ldots, m_n$,

\[ \hat{\theta}^{(j)}_{\Pi_{\text{VB},\nu,k}}(x) = \hat{\theta}^{(j)}_{\Pi_{\text{VB},\nu,k}}(x^{(j)}) = \begin{cases} \nu^{(j)}_k & \text{if } \|x^{(j)}\|_0 = 0, \\ \nu^{(j)}_k & \text{if } x^{(j)}_k \neq 0 \text{ and } x^{(j)}_l = 0 \text{ for } l \neq k, \\ 0 & \text{otherwise}, \end{cases} \]
where \( w_k^{(j)} := \exp(-r_k^{(j)} \nu_k^{(j)}) / \sum_{l=1}^{m_n} \exp(-r_l^{(j)} \nu_l^{(j)}) \).

This expression of \( \theta_{\text{VB},\nu} \) enables us to evaluate the estimative risk, that is, \( R_e(\theta, \hat{\theta}_{\text{VB},\nu}; \tau) \). By the coordinate-wise additive property of the Kullback-Leibler divergence, we decompose \( R_e(\theta, \hat{\theta}_{\text{VB},\nu}; \tau) \) as

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

where for \( j = 1, \ldots, s_n \),

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

and for \( j = s_n + 1 \),

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

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

\[
R^{(j)}(\theta^{(j)}; \tau)
= \mathbb{E}_{\theta^{(j)}} \sum_{k=1}^{m_n} i_k^{(j)} \left[ \hat{\theta}_k^{(j)} \log \left( \frac{\theta_k^{(j)}}{\hat{\theta}_{\text{VB},\nu,k}^{(j)}(Z^{(j)}; \tau)} \right) - \theta_k^{(j)} + \hat{\theta}_{\text{VB},\nu,k}^{(j)}(Z^{(j)}; \tau) \right]
= e^{-\gamma^{(j)}} \gamma^{(j)} \left\{ t_k^{(j)} \nu_k^{(j)} \log \frac{1}{u_k^{(j)}} - t_k^{(j)} \nu_k^{(j)} + \sum_{k=1}^{m_n} i_k^{(j)} u_k^{(j)} \nu_k^{(j)} \right\},
\]

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

\[
\int R^{(j)}(\theta^{(j)}; \tau) d\Pi_{\text{VB},\nu}(\theta)
= \frac{1}{m_n} \sum_{k=1}^{m_n} e^{-i_k^{(j)} \nu_k^{(j)}(1/u_k^{(j)}) - i_k^{(j)} \nu_k^{(j)} + \sum_{l=1}^{m_n} i_l^{(j)} u_l^{(j)} \nu_l^{(j)}}
\geq \frac{1}{m_n} \sum_{k=1}^{m_n} e^{-i_k^{(j)} \nu_k^{(j)}(1/u_k^{(j)}) - i_k^{(j)} \nu_k^{(j)}}.\]
Integrating the both hand sides of the above equality with respect to \( \tau \) over \([0, 1]\), we have

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

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 \( \Pi_{VB, \nu} \):

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

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

We next show that the asymptotic inequality

\[
\overline{R}(\Theta[s_n]) \geq \tilde{f}(\nu) \{ s_n \log(\eta_n^{-1}) \}(1 + o(1)) \quad (26)
\]

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

**Condition B.1.** there exists a positive constant \( C \) such that \( \max_{i} \nu_i^{(j)} \nu_i^{(j)} \leq C \) for any \( j = 1, \ldots, s_n \).

For \( \nu \) satisfying Condition B.1, the first term of (25) is rewritten as

\[
\frac{1}{m_n} \sum_{j=1}^{s_n} \sum_{k=1}^{m_n} f_k^{(j)}(\nu_k^{(j)}) \log \frac{1}{w_k^{(j)}} = \frac{1}{m_n} \sum_{j=1}^{s_n} \sum_{k=1}^{m_n} f_k^{(j)}(\nu_k^{(j)}) \left( \log m_n + \log \frac{1}{m_n w_k^{(j)}} \right) \\
= \tilde{f}(\nu) s_n \{ \log(\eta_n^{-1}) \}(1 + o(1)),
\]

because by definition of \( w_k^{(j)} \) we have

\[
\exp \left\{ - \max_{i=1, \ldots, m_n} \nu_i^{(j)} \nu_i^{(j)} + \nu_k^{(j)} \nu_k^{(j)} \right\} \leq \frac{1}{m_n w_k^{(j)}} \\
\leq \exp \left\{ - \min_{i=1, \ldots, m_n} \nu_i^{(j)} \nu_i^{(j)} + \nu_k^{(j)} \nu_k^{(j)} \right\}
\]
for any \(k = 1, \ldots, m\) and \(j = 1, \ldots, s_n\). For \(\nu\) satisfying Condition B.1, the second term of (25) is negligible compared to the first term.

Since Condition B.1 holds for \(\nu^0\) that maximizes the right hand side of (26), that is, \(\nu^0 = \log(1 + 1/r_i) (i = 1, \ldots, n)\), we obtain the desired lower bound by substituting \(\nu = \nu^0\), which completes Step 1.

**Step 2: Upper bound on \(\overline{\mathcal{R}}(\Theta|s_n)\)**

We derive an upper bound on \(\overline{\mathcal{R}}(\Theta|s_n)\) following almost the same procedure as in Subsection A.1. Let \(\Pi^* = \Pi[\eta_n, 1]\). Fix \(i = 1, 2, \ldots, n\). For \(\lambda_i > 0\), let

\[
\rho_i(\lambda_i) := \mathbb{E}_{\lambda_i} \log(\{\exp(-\lambda_i)\lambda_i^{z_i}/Y_i!\}/\{\eta_n, i(Y_1 | X_1)\}).
\]

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

\[
\hat{\theta}_{\Pi^*, i}(z_i; \tau) = (\eta_n/t_i(\tau)^{\kappa+1})/(1 + \eta_n/t_i(\tau))^{\kappa}, \quad z_i = 0,
\]

\[
\hat{\theta}_{\Pi^*, i}(z_i; \tau) = (z_i + 1)/t_i(\tau), \quad z_i \geq 1.
\]

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

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

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

\[
\hat{\rho}_i(\lambda_i, z_i; \tau) \leq \hat{t}_i(\tau)\{\lambda_i \log \eta_n^{-1} + \lambda_i \log \lambda_i - \lambda_i + \lambda_i \log C_1 + \eta_n C_2\}, \quad z_i = 0,
\]

\[
\hat{\rho}_i(\lambda_i, z_i; \tau) \leq \hat{t}_i(\tau)\{\lambda_i \log \hat{t}_i(\tau)\lambda_i/(z_i + 1)\} - \lambda_i + (z_i + 1)/t_i(\tau)\} \quad z_i \geq 1.
\]

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 A.1, we have, for sufficiently large \(n \in \mathbb{N}\) and for \(\lambda_i > 0\),

\[
\rho_i(\lambda_i) = \int_0^1 \mathbb{E}_{t_i(\tau)\lambda_i}[\hat{\rho}_i(\lambda_i, Z_1; \tau)]d\tau
\]

\[
\leq \int_0^1 \left[e^{-t_i(\tau)\lambda_i}\hat{t}_i(\tau)\lambda_i \log \eta_n^{-1} + e^{-t_i(\tau)\lambda_i}\hat{t}_i(\tau)\lambda_i \log(\lambda_i C_1)
\right.
\]

\[
+ \hat{t}_i(\tau)\lambda_i \log(1 - e^{-t_i(\tau)\lambda_i}) + \hat{t}_i(\tau)/t_i(\tau)(1 - e^{-t_i(\tau)\lambda_i})\] \] \(d\tau + \eta_n C_2.
\]

(27)
Inequality (27) yields
\[
\sup_{\lambda > 0} \rho_i(\lambda) \leq (C_i + o(1)) \log \eta_n^{-1}
\]
and a similar procedure yields the inequality \(\rho_i(0) = O(\eta_n)\).

Finally, we derive an upper bound on \(\mathcal{R}(\Theta[s_n])\) by combining asymptotic inequalities \(\sup_{\lambda > 0} \rho_i(\lambda) \leq (C_i + o(1)) \log \eta_n^{-1}\) and \(\rho_i(0) = O(\eta_n)\) for \(i = 1, \ldots, n\).

Let subscripts \([1], \ldots, [n]\) be the permutation of \(1, \ldots, n\) that satisfies \(C_{[1]} \geq \ldots \geq C_{[n]}\). The minimax risk \(\mathcal{R}(\Theta[s_n])\) is bounded as
\[
\mathcal{R}(\Theta[s_n]) \leq \sup_{\theta \in \Theta[s_n]} R(\theta, q_{H+}) = \sum_{i=s_n+1}^{n} \rho_i(0) + \sum_{i=1}^{s_n} \rho_i(\lambda) + \sum_{i=1}^{n} \rho_i(0).
\]
Together with Condition 5.1, the asymptotic inequalities
\[
\sup_{\lambda, > 0} \rho_i(\lambda) \leq (C_i + o(1)) \log \eta_n^{-1} \quad \text{and} \quad \rho_i(0) = O(\eta_n) \quad (i = 1, \ldots, n)
\]
give
\[
\mathcal{R}(\Theta[s_n]) \leq \sum_{i=1}^{s_n} \{C_i + o(1)\} \log \eta_n^{-1} + (n - s_n)O(\eta_n)
\]
\[
\sim \frac{s_n}{n} \sum_{i=1}^{n} C_i \{\log \eta_n^{-1}\} (1 + o(1)) \sim \bar{C}s_n \log \eta_n^{-1},
\]
from which we obtain the desired upper bound on \(\mathcal{R}(\Theta[s_n])\).

\(\Box\)

### B.4.2. Proof of Lemma B.1

The proof is divided into two parts: deriving the minimax risk with \(r_i\)’s conditioned, and taking expectation of that with respect to \(G\). Recall that \(\bar{C} := \frac{1}{n} \sum_{i=1}^{n} C_i\). The former part of the proof essentially is completed by Lemma 5.2.

Consider the latter part of the proof. Take the expectation of \(\bar{C}\) with respect to \(G\). Since \(\bar{C}\) is in \((0, 1)\), Hoeffding’s inequality gives
\[
\Pr\left(|\bar{C} - E_G \bar{C}| \geq t\right) \leq 2 \exp\left(-2nt^2\right), \quad t > 0.
\]
This implies that for any \(\delta \in (0, 1)\), we have
\[
|\bar{C} - E_G \bar{C}| \leq \sqrt{\frac{1}{2n} \log(2/\delta)}
\]
with probability higher than \(1 - \delta\). In combining (29) with the result of the former part, we make an additional analysis on the behaviors of \(C_1\) and \(C_2\) in
Condition 3.1 assures that
\[
R(\Theta[s_n]) - \mathcal{C}s_n \log(\eta_n^{-1}) = O_P(1).
\]
This implies that combining (29) with the result of the former part yields the inequality
\[
\left( \mathbb{E}_C - \sqrt{\frac{1}{2n}} \log(1/\delta) + b_n \right) s_n \log(\eta_n^{-1}) \leq R(\Theta[s_n]) \leq \left( \mathbb{E}_C + \sqrt{\frac{1}{2n}} \log(2/\delta) + b_n \right) s_n \log(\eta_n^{-1})
\]
with probability higher than \(1 - \delta\), where \(b_n\) is an \(o_p(1)\) term that is independent of \(\delta\). This completes the proof.

B.4.3. Proof of Lemma B.2

The proof follows essentially the same steps as those in Subsection A.3. We provide the proof only for \(\Theta[s_n]\). For the simplicity, we conduct the proof with \(L\) replaced by 1. Following the same argument as that of Theorem 5.2, we have, for \(i = 1, 2, \ldots, n\),
\[
\hat{\theta}_{\Pi_{[\eta_n, \kappa]}}, i(z_i; \tau) = \frac{\eta_n/t_i(\tau)^{\kappa+1}/(1 + \eta_n/t_i(\tau)^{\kappa})}{t_i(\tau)} \quad z_i = 0;
\]
\[
\hat{\theta}_{\Pi_{[\eta_n, \kappa]}}, i(z_i; \tau) = \frac{z_i + \kappa}{t_i(\tau)} \quad z_i \geq 1.
\]
The above equalities for \(\hat{\theta}_{\Pi_{[\eta_n, \kappa]}}\) show that there exists a positive constant \(C_1\) depending only on \(\kappa\) and \(\sup_i r_i\) for which we have
\[
\sup_{\lambda > 0} \mathbb{E}_\theta \log\left\{ \frac{\exp(-\lambda)\lambda^{Y_1}/Y_1!}{\{q_{\Pi_{[\eta_n, \kappa], 1}}(Y_1 | X_1)\}} \right\} \leq \mathcal{C} \log(\eta_n^{-1}) + C_1,
\]
\[
\lim_{\lambda \to 0} \mathbb{E}_\theta \log\left\{ \frac{\exp(-\lambda)\lambda^{Y_1}/Y_1!}{\{q_{\Pi_{[\eta_n, \kappa], 1}}(Y_1 | X_1)\}} \right\} = O(\eta_n).
\]
Thus, we have
\[
\sup_{\theta \in \Theta[s_n]} R(\theta, q_{\Pi_{[\eta_n, \kappa], 1}}) \leq (\mathcal{C} + o(1)) s_n \log(\eta_n^{-1}) + (n - s_n) O(\eta_n),
\]
which completes the proof.

B.5. Proof of Proposition 3.1

The proof is a combination of the following lemma and the same argument as in Subsection B.1:
Lemma B.3. 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 \( J \subset \{1, \ldots, n\} \) with \( |J| = s_n \). Then the asymptotic equalities

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

hold as \( n \to \infty \) and \( \eta_n \to 0 \), where \( \varepsilon_n \) is any shrinking sequence such that \( \varepsilon_n = o(\eta_n) \).

Once we obtain Lemma B.3, the remaining part of the proof of Proposition 3.1 is easy and thus we complete the proof. \( \square \)

Proof of Lemma B.3: In what follows, we prove Lemma B.3. We provide the proof only for \( \Theta[s_n] \). The proof follows almost the same line of the proof of Lemma 5.2. In bounding \( \mathcal{E}(\Theta[s_n]) \) below, we employ the Bayes risk based on the varied-spike block-independent prior \( \Pi_{VB,\nu} \) with \( \nu = (1/r_1, \ldots, 1/r_n) \in \mathbb{R}_+^n \). In bounding \( \mathcal{E}(\Theta[s_n]) \) above, we use a threshold estimator. The threshold estimator used in this proof is the same as in Subsection A.5.1: for each \( i = 1, \ldots, n \), \( \hat{\theta}_{T,i}(x_i) = \eta_i \) 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 \( \mathcal{E}(\Theta[s_n]) \). Observe that the Bayes risk based on \( \Pi_{VB,\nu} \) is decomposed 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

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

The explicit expression of \( \hat{\theta}_{\Pi_{VB,\nu}} \) with \( \nu = (1/r_i)_{i=1}^{n} \) gives

\[
B^{(j)}(\nu) = \frac{1}{m_n} \sum_{\theta \in \text{Supp}(\Pi_{VB,\nu})} \sum_{k=1}^{m_n} \mathbb{E}_\theta \left\{ \theta_k^{(j)} \log(\theta_k^{(j)}/\hat{\theta}_{\Pi_{VB,\nu}}^{(j)}) - \theta_k^{(j)} + \hat{\theta}_{\Pi_{VB,\nu}}^{(j)} \right\}
\]

\[
\geq \frac{1}{m_n} \sum_{k=1}^{m_n} (1/r_k^{(j)})e^{-1} \log m_n - \sum_{k=1}^{m_n} (1/r_k^{(j)})e^{-1}
\]

\[
\sim \frac{s_n}{n} \sum_{k=1}^{m_n} (1/r_k^{(j)})e^{-1} \log \eta_n^{-1},
\]

from which we have

\[
\int R(\theta, q(\cdot | \hat{\theta}_{\Pi_{VB,\nu}}))d\Pi_{VB,\nu}(\theta) \geq \left( \frac{1}{n} \sum_{i=1}^{n} r_i^{-1}e^{-1} + o(1) \right) s_n \log \eta_n^{-1}.
\]
This yields the desired lower bound on $\mathcal{F}(\Theta[s_n])$.

Consider an upper bound on $\mathcal{F}(\Theta[s_n])$. Let 
\[
\tilde{\rho}_i(\lambda_i) := E_{r_i, \lambda_i} [\lambda_i \log(\lambda_i / \hat{\theta}_{T,i}(X_i)) - \lambda_i + \hat{\theta}_{T,i}(X_i)]
\]
be a coordinate-wise estimative risk of $\hat{\theta}_{T,i}$. Following the same steps as in Subsection A.5.1, we get the following asymptotic inequalities: for $i = 1, \ldots, n$,
\[
\sup_{\lambda > 0} \tilde{\rho}_i(\lambda) = (1 + o(1))e^{-1/r_i - 1/s_n \log \eta_n^{-1}},
\]
\[
\tilde{\rho}_i(0) = O(\eta_n).
\]

Let subscripts $[1], \ldots, [n]$ be the permutation of $1, \ldots, n$ that satisfies $1/r_{[1]} \geq \ldots \geq 1/r_{[n]}$. The condition $\sum_{i=1}^n 1/nr_i \sim \sum_{i \in J} 1/(snr_i)$ for any subset $J \subset \{1, \ldots, n\}$ with $|J| = s_n$ implies that the minimax risk $\mathcal{F}(\Theta[s_n])$ is bounded above as follows:
\[
\mathcal{F}(\Theta[s_n]) \leq \sum_{i=1}^{s_n} \{r_i^{-1}e^{-1} + o(1)\} \log \eta_n^{-1} + (n - s_n)O(\eta_n)
\]
\[
\sim \frac{s_n}{n} \sum_{i=1}^n r_i^{-1}e^{-1} \{\log \eta_n^{-1}\}(1 + o(1)),
\]
from which we obtain the desired upper bound on $\mathcal{F}(\Theta[s_n])$ and thus complete the proof.$\square$

Appendix C: Supplemental experiments

This appendix presents supplemental experiments for readers’ better understanding.

C.1. Quasi-sparsity

This subsection provides simulation studies for quasi-sparsity. Parameter $\theta$ and observations $X$ and $Y$ are drawn from
\[
\theta_i \sim \nu_i e_{S,i} + \xi_i e_{S^c,i} \ (i = 1, \ldots, n),
\]
\[
X \mid \theta \sim \otimes_{i=1}^n \text{Po}(r\theta_i), \ Y \mid \theta \sim \otimes_{i=1}^n \text{Po}(\theta_i), \ \text{and} \ X \perp Y \mid \theta,
\]
respectively, where
- $\nu_1, \ldots, \nu_n$ are independently drawn from the gamma distribution with a shape parameter 10 and a scale parameter 1;
• $\xi_1, \ldots, \xi_n$ are independently and uniformly drawn from $[0, 10^{-2}]$;
• $S$ is drawn from the uniform distribution on all subsets having exactly $s$ and $S^c$ is its complement;
• $\nu_1,\ldots,\nu_n$ and $S$ are independent.

Here for a subset $J \subset \{1, \ldots, n\}$, $e_J$ 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.

|          | $L^* \hat{\eta}_{n, 0.1}$ | $L^* \hat{\eta}_{n, 1.0}$ | GH   | K04   | $\ell_1$ ($\lambda = 0.1$) |
|----------|-----------------------------|-----------------------------|-------|-------|----------------------------|
| $\ell_1$ distance | 13.8 (4.0) | 13.6 (4.0) | 18.1 (1.7) | 63.3 (8.3) | 17.6 (4.9) |
| PLL      | -19.0 (5.3) | -18.8 (5.3) | -20.8 (4.3) | -43.8 (4.9) | -Inf         |
| 90%CP (%) | 90.7 (2.8)  | 90.1 (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.

|          | $L^* \hat{\eta}_{n, 0.1}$ | $L^* \hat{\eta}_{n, 1.0}$ | GH   | K04   | $\ell_1$ ($\lambda = 0.1$) |
|----------|-----------------------------|-----------------------------|-------|-------|----------------------------|
| $\ell_1$ distance | 49.0 (8.5) | 48.8 (8.6) | 50.3 (3.0) | 206 (15) | 57.7 (9.1) |
| PLL      | -52.3 (5.6) | -52.2 (5.7) | -56.7 (4.5) | -121 (8.8) | -Inf         |
| 90%CP (%) | 89.8 (2.9)  | 89.5 (3.1)   | 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 does not depend on whether a parameter is exact sparse or quasi-sparse.

### C.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.

|          | $L^* \hat{\eta}_{n, 0.1}$ | $L^* \hat{\eta}_{n, 1.0}$ | GH   | K04   | $\ell_1$ ($\lambda = 0.1$) |
|----------|-----------------------------|-----------------------------|-------|-------|----------------------------|
| $\ell_1$ distance | 106.0 (11.5) | 106.4 (11.6) | 136.3 (4.76) | 117.1 (12.4) | 118.0 (13.0) |
| PLL      | -110.2 (4.5) | 110.4 (4.5) | -115.2 (4.6) | -117.8 (4.3) | -Inf         |
| 90%CP (%) | 89.0 (2.8)  | 89.8 (2.7)   | 92.1 (1.4) | 95.4 (2.4)  | 79.4 (5.4)  |
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.

| \(\ell_1\) distance | \([-0.3, 0.1]\) | \([-0.1, 1.0]\) | \(\text{GH}\) | \(\text{K04}\) | \(\ell_1 (\lambda = 0.1)\) |
|---------------------|-----------------|-----------------|---------|---------|------------------|
| PLL                | 208.0 (18.6)    | 208.4 (18.7)    | -190.3 (6.0) | 209.6 (12.4) | 221.3 (17.8)     |
| 90%CP (%)          | 88.9 (2.8)      | 89.7 (2.8)      | 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 works better in the predictive likelihood sense as the sparsity \(s_n\) is relatively large.

Appendix D: Formula for the Kullback–Leibler risk

In this appendix, we provide Lemma 5.1 for the sake of completeness. Let \(\Pi\) be a prior of \(\theta\) and suppose that the Bayes estimate \(\hat{\theta}_{\Pi}(x; t)\) based on \(\Pi\) 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 follows:

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

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 \(p(z | \theta; t) := \prod_{i=1}^n e^{-t\theta_i}t^{z_i-1}/z_i!\) and let \(p_{\Pi}(z; t) := \int p(z | \theta; t)d\Pi(\theta)\). From the sufficiency reduction, we have

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

Introducing the random variable \(Z_t\) from \(\otimes_{i=1}^n \text{Po}(t \theta_i)\) \((t \in (r, r + 1))\), we get the following expression of the risk \(R(\theta, q_{\Pi})\):

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

If we obtain the equality

\[
\frac{d}{dt} \mathbb{E} \left[ \log \left( \frac{p(Z_t | \theta; t)}{p_{\Pi}(Z_t; t)} \right) \right] = \frac{R_e(t \theta, \hat{\theta}_{\Pi}(\cdot; t))}{t}, \quad t \in (r, r + 1), \quad (30)
\]
we complete the proof. Differentiating $E[\log \left\{ \frac{p(Z_t | \theta; t)}{p_{\Pi}(Z_t; t)} \right\}]$ with respect to $t$ yields

$$E[\log \{ p(Z_t | \theta; t)/p_{\Pi}(Z_t; t) \}] = E \left[ \frac{d}{dt} \log \left\{ \frac{p(Z_t | \theta; t)}{p_{\Pi}(Z_t; t)} \right\} \right] + E \left[ \frac{d}{dt} \log p(Z_t | \theta; t) \right] - E \left[ \frac{d}{dt} \log p_{\Pi}(Z_t; t) \right].$$

(31)

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 ($E[\sum_{i=1}^{n} (Z_{t,i} - 1)f(Z_t)] = E[\sum_{i=1}^{n} t\theta_i f(Z_t + e_i)]$ for any function $f : \mathbb{N}^n \rightarrow \mathbb{R}$) yields

$$E \left[ \frac{d}{dt} \log \left\{ \frac{p(Z_t | \theta; t)}{p_{\Pi}(Z_t; t)} \right\} \right] = E \sum_{i=1}^{n} \theta_i \log \left\{ \frac{\theta_i}{\hat{\theta}_{\Pi,i}(Z_t; t)} \right\}.$$  

(32)

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

$$E \left[ \frac{d}{dt} \log p(Z_t | \theta; t) \right] = E \left[ \frac{d}{dt} \log p_{\Pi}(Z_t; t) \right] = E \left[ \frac{d}{dt} \log p_{\Pi}(Z_t; t) \right].$$

(33)

Combining these identities (32) and (33) with (31) gives (30), which completes the proof. 

$\square$

**Appendix E: The numbers of individuals in Subsection 4.2.2**

This appendix provides the histogram of the double numbers $(r,s)$ 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. From Figure 7, we see that most of locations are read from 100,000 to 120,000 times while a few locations are read less than 100,000 times. These differences can be explained by a mechanism of a sequencer: a DNA sequencer randomly reads fragments of a DNA sequence numerous times, and then reconstructs the whole sequence by combining all read fragments; hence, the numbers of individuals read at genetic locations are regarded as random variables and there occurs a difference in each value of them.