Journal of Statistical Planning and Inference 141 (2011) 3705–3715

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Bayesian predictive densities based on latent information priors

Fumiyasu Komaki [a,b,*]

[a] Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

[b] RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan

ARTICLE INFO

Article history:
Received 26 September 2010
Received in revised form
30 May 2011
Accepted 5 June 2011
Available online 14 June 2011

Keywords:
Essentially complete class
Jeffreys prior
Kullback–Leibler divergence
Minimaxity
Multinomial model
Reference prior

ABSTRACT

Construction methods for prior densities are investigated from a predictive viewpoint. Predictive densities for future observables are constructed by using observed data. The simultaneous distribution of future observables and observed data is assumed to belong to a parametric submodel of a multinomial model. Future observables and data are possibly dependent. The discrepancy of a predictive density to the true conditional density of future observables given observed data is evaluated by the Kullback–Leibler divergence. It is proved that limits of Bayesian predictive densities form an essentially complete class. Latent information priors are defined as priors maximizing the conditional mutual information between the parameter and the future observables given the observed data. Minimax predictive densities are constructed as limits of Bayesian predictive densities based on prior sequences converging to the latent information priors.

1. Introduction

We construct predictive densities for future observables by using observed data. Future observables and data are possibly dependent and the simultaneous distribution of them is assumed to belong to a submodel of a multinomial model. Various practically important models such as categorical models and graphical models are included in this class.

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets composed of k and l elements, and let x and y be random variables that take values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathcal{M}=\{p(x,y|\theta)|\theta\in\Theta\}$ be a set of probability densities on $\mathcal{X}\times\mathcal{Y}$. The model $\mathcal{M}$ is regarded as a submodel of the kl-nomial model. The model $\mathcal{M}$ is naturally regarded as a subset of the hyperplane $\{p=(p_{ij})|\sum_{i=1}^{k}\sum_{j=1}^{l}p_{ij}=1\}$ in Euclidean space $\mathbb{R}^{kl}$. In the following, we identify Θ with $\mathcal{M}$. Then, the parameter space Θ is endowed with the induced topology as a subset of $\mathbb{R}^{kl-1}$.

A predictive density $q(y;x)$ is defined as a function from $\mathcal{X}\times\mathcal{Y}$ to $[0,1]$ satisfying $\sum_{y\in\mathcal{Y}}q(y;x)=1(x\in\mathcal{X})$. The closeness of $q(y;x)$ to the true conditional probability density $p(y|x,\theta)$ is evaluated by the average Kullback–Leibler divergence:

$$R(\theta,q)=\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}, \tag{1}$$

where we define $c\log 0=-\infty$ $(c>0)$, $0\log 0=0$, $0\log(c/0)=0$ $(c\geq 0)$. Although the conditional probability $p(y|x,\theta)$ is not uniquely defined when $p(x|\theta)=0$, the risk value $R(\theta,q)$ is uniquely determined because $p(x,y|\theta)\log p(y|x,\theta)=0$ if $p(x|\theta)=0$.

[*] Correspondence address: Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
E-mail address: komaki@mist.i.u-tokyo.ac.jp

0378-3758 © 2011 Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.jspi.2011.06.009

First, we show that, for every predictive density $q(y;x)$, there exists a limit $\lim_{n\to\infty} p_{\pi_n}(y;x)$ of Bayesian predictive densities

$$p_{\pi_n}(y|x) := \frac{\int p(x,y|\theta)\,\mathrm{d}\pi_n(\theta)}{\int p(x|\theta)\,\mathrm{d}\pi_n(\theta)},$$

where $\{\pi_n\}_{n=1}^{\infty}$ is a prior sequence, such that $R(\theta,\lim_{n\to\infty} p_{\pi_n}(y;x)) \le R(\theta,q(y;x))$ for every $\theta \in \Theta$. In the terminology of statistical decision theory, this means that the class of predictive densities that are limits of Bayesian predictive densities is an essentially complete class.

Next, we investigate latent information priors defined as priors maximizing the conditional mutual information discussed in Section 3 between y and θ given x. We obtain a constructing method for a prior sequence $\{\pi_n\}_{n=1}^{\infty}$ converging the latent information prior, based on which a minimax predictive density $\lim_{n\to\infty} p_{\pi_n}(y|x)$ is obtained. We consider limits of Bayesian predictive densities to deal with conditional probabilities.

There exist important previous studies on prior construction by using the unconditional mutual information. The reference prior by Bernardo (1979, 2005) is a prior maximizing the mutual information between θ and y in the limit of the amount of information of y goes to infinity. It corresponds to the Jeffreys prior if there are no nuisance parameters; see Ibragimov and Hasminskii (1973) and Clarke and Barron (1994) for rigorous treatments. In coding theory, the prior maximizing the mutual information between y and θ is used for Bayes coding. It was shown that the Bayes codes for finite alphabet models based on the priors are minimax by Gallager (1979) and Davisson and Leon-Garcia (1980). In our framework, these settings correspond to prediction of y without x. In statistical applications, x plays an important role because it corresponds to observed data, although $\mathcal{X}$ is an empty set in the reference analysis and the standard framework of information theory; see also Komaki (2004) for the relation between statistical prediction and Bayes coding.

Geisser (1979), in the discussion of Bernardo (1979), discussed minimax prediction based on the risk function (1) as an alternative to the reference prior approach.

The latent information priors introduced in the present paper bridge these two approaches. The theorems obtained below clarify the relation between the conditional mutual information and minimax prediction based on observed data.

For Bayesian prediction of future observables by using observed data, Akaike (1983) discussed priors maximizing the mutual information between x and y and called them minimum information priors. Kuboki (1998) also proposed priors for Bayesian prediction based on an information theoretic quantity. These priors are different from latent information priors investigated in the present paper.

In Section 2, we prove that, for every predictive density $q(y;x)$, there exists a predictive density that is a limit of Bayesian predictive densities whose performance is not worse than that of $q(y;x)$. In Section 3, we introduce a construction method for minimax predictive densities as limits of Bayesian predictive densities. The method is based on the conditional mutual information between y and θ given x. In Section 4, we give some numerical results and discussions.

2. Limits of Bayesian predictive densities

In this section, we prove that the class of predictive densities that are limits of Bayesian predictive densities is an essentially complete class.

Throughout this paper, we assume the following conditions:

Assumption 1. Θ is compact.

Assumption 2. For every $x \in \mathcal{X}$, there exists $\theta \in \Theta$ such that $p(x|\theta) > 0$.

These assumptions are not restrictive. For Assumption 1, if Θ is not compact, we can regard the closure $\overline{\Theta}$ as the parameter space instead of Θ because we consider a submodel of a multinomial model. We do not lose generality by Assumption 2 because we can adopt $\mathcal{X}\backslash\{x_0\}$ instead of $\mathcal{X}$ if there exists $x_0 \in \mathcal{X}$ such that $p(x_0|\theta) = 0$ for every $\theta \in \Theta$.

We prepare several preliminary results to prove Theorem 1 below.

Let $\mathcal{P}$ be the set of all probability measures on Θ endowed with the weak convergence topology and the corresponding Borel algebra. By the Prohorov theorem and Assumption 1, $\mathcal{P}$ is compact.

When x and y are fixed, the function $\theta \in \Theta \to p(x,y|\theta) \in [0,1]$ is bounded and continuous. Thus, for every fixed $(x,y) \in \mathcal{X} \times \mathcal{Y}$, the function

$$\pi \in \mathcal{P} \to p_\pi(x,y) := \int p(x,y|\theta)\mathrm{d}\pi(\theta)$$

is continuous, because of the definition of weak convergence. Therefore, for every predictive density $q(y;x)$, the function from $\mathcal{P}$ to $[0,\infty]$ defined by

$$D_q(\pi) := \sum_{x,y} p_\pi(x,y)\log\frac{p_\pi(x,y)}{q(y;x)p_\pi(x)} = \sum_{x,y} p_\pi(x,y)\log p_\pi(x,y) - \sum_{x} p_\pi(x)\log p_\pi(x) - \sum_{(x,y):q(y;x)>0} p_\pi(x,y)\log q(y;x) - \sum_{(x,y):q(y;x)=0} p_\pi(x,y)\log q(y;x) \tag{2}$$

is lower semicontinuous, because the last term in (2) is lower semicontinuous and the other terms are continuous.

Lemma 1. *Let μ be a probability measure on Θ. Then, $\mathcal{P}_{\varepsilon\mu}=\{\varepsilon\mu+(1-\varepsilon)\pi|\pi\in\mathcal{P}\}$ $(0\le\varepsilon\le 1)$ is a closed subset of $\mathcal{P}$.*

Proof. Suppose that $\pi_\infty\in\mathcal{P}$ is the limit of a convergent sequence $\{\pi_k\}_{k=1}^{\infty}$ in $\mathcal{P}_{\varepsilon\mu}$. Since $\pi_k\in\mathcal{P}_{\varepsilon\mu}$,

$$\int f(\theta)\mathrm{d}\pi_k(\theta)-\varepsilon\int f(\theta)\,\mathrm{d}\mu(\theta)\ge 0$$

for every nonnegative bounded continuous function $f(\theta)$ on Θ. Thus,

$$\int f(\theta)\,\mathrm{d}\pi_\infty(\theta)=\lim_{k\to\infty}\int f(\theta)\,\mathrm{d}\pi_k(\theta)\ge\varepsilon\int f(\theta)\,\mathrm{d}\mu(\theta).$$

Hence, $\pi_\infty-\varepsilon\mu$ is a nonnegative measure. Therefore, $\pi_\infty\in\mathcal{P}_{\varepsilon\mu}$, and $\mathcal{P}_{\varepsilon\mu}$ is a closed set in $\mathcal{P}$. □

Lemma 2. *Let $f(\cdot)$ be a continuous function from $\mathcal{P}$ to $[0,\infty]$, and let μ be a probability measure on Θ such that $p_\mu(x):=\int p(x|\theta)\,\mathrm{d}\mu(\theta)>0$ for every $x\in\mathcal{X}$. Then, there is a probability measure π_n in*

$$\mathcal{P}_{\mu/n}:=\left\{\frac{1}{n}\mu+\left(1-\frac{1}{n}\right)\pi\middle|\pi\in\mathcal{P}\right\}\quad(n=1,2,3,\ldots)$$

such that $f(\pi_n)=\inf_{\pi\in\mathcal{P}_{\mu/n}}f(\pi)$. Furthermore, there exists a convergent subsequence $\{\pi'_m\}_{m=1}^{\infty}$ of $\{\pi_n\}_{n=1}^{\infty}$ and the equality $f(\pi'_\infty)=\inf_{\pi\in\mathcal{P}}f(\pi)$ holds, where $\pi'_\infty=\lim_{m\to\infty}\pi'_m$.

Proof. Note that there exists $\mu\in\mathcal{P}$ such that $p_\mu(x):=\int p(x|\theta)\,\mathrm{d}\mu(\theta)>0$ for every $x\in\mathcal{X}$ by Assumption 2. By Lemma 1, the sets $\mathcal{P}_{\mu/n}$ $(n=1,2,3,\ldots)$ are compact because they are closed subsets of a compact set $\mathcal{P}$. Thus, there is a probability measure π_n in $\mathcal{P}_{\mu/n}$ such that $f(\pi_n)=\inf_{\pi\in\mathcal{P}_{\mu/n}}f(\pi)$. There exists a convergent subsequence $\{\pi'_m\}_{m=1}^{\infty}$ of $\{\pi_n\}_{n=1}^{\infty}$ because $\mathcal{P}$ is compact.

Since $\mathcal{P}$ is compact and $f(\pi)$ is a continuous function of $\pi\in\mathcal{P}$, there exists $\hat{\pi}\in\mathcal{P}$ such that $f(\hat{\pi})=\inf_{\pi\in\mathcal{P}}f(\pi)$. Thus, $f(\pi'_\infty)\ge f(\hat{\pi})$, where $\pi'_\infty:=\lim_{m\to\infty}\pi'_m$. For every $\varepsilon>0$, there exists $\delta>0$ such that $\sup_{d(\hat{\pi},\pi)<\delta}f(\pi)\le f(\hat{\pi})+\varepsilon$, where d is the Prohorov metric on $\mathcal{P}$. We put

$$\hat{\pi}_n=\frac{1}{n}\mu+\frac{n-1}{n}\hat{\pi}\quad(n=1,2,3,\ldots).$$

Then, $\hat{\pi}_n\in\mathcal{P}_{\mu/n}$ and $\lim_{n\to\infty}\hat{\pi}_n=\hat{\pi}$. Thus, for every $\delta>0$, there exists a positive integer N such that $d(\hat{\pi},\hat{\pi}_n)<\delta$ $(n\ge N)$. If $n\ge N$, then $f(\pi'_\infty)\le f(\pi_n)\le f(\hat{\pi}_n)\le f(\hat{\pi})+\varepsilon$. Since $\varepsilon>0$ is arbitrary, we have $f(\pi'_\infty)\le f(\hat{\pi})$. Therefore, $f(\pi'_\infty)=f(\hat{\pi})$. □

The conditional probability $p_\pi(y|x)$ is not uniquely specified if $p_\pi(x)=0$. To resolve the problem, we consider a sequence of priors $\{\pi_n\}_{n=1}^{\infty}$ that satisfies $p_{\pi_n}(x)>0$ for every n and $x\in\mathcal{X}$. In the following, $\lim_{n\to\infty}p_{\pi_n}(y|x)$ is defined to be a map from $(x,y)\in\mathcal{X}\times\mathcal{Y}$ to the limit of the real number sequence $\{p_{\pi_n}(y|x)\}_{n=1}^{\infty}$. If there exist limits of sequence of real numbers $\{p_{\pi_n}(y|x)\}_{n=1}^{\infty}$ for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, we say the limit $\lim_{n\to\infty}p_{\pi_n}(y|x)$ of Bayesian predictive densities exists. Obviously, if the limit $\lim_{n\to\infty}p_{\pi_n}(y|x)$ exists, it is a predictive density because $0\le\lim_{n\to\infty}p_{\pi_n}(y|x)\le 1$ for every $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and $\sum_{y\in\mathcal{Y}}\lim_{n\to\infty}p_{\pi_n}(y|x)=1$ for every $x\in\mathcal{X}$.

Theorem 1.

(1) *Let $q(y;x)$ be a predictive density. If there exists $\hat{\pi}^q\in\mathcal{P}$ such that $D_q(\hat{\pi}^q)=\inf_{\pi\in\mathcal{P}}D_q(\pi)$ and $p_{\hat{\pi}^q}(x)>0$ for every $x\in\mathcal{X}$, then $R(\theta,p_{\hat{\pi}^q}(y|x))\le R(\theta,q(y;x))$ for every $\theta\in\Theta$.*

(2) *For every predictive density $q(y;x)$, there exists a convergent prior sequence $\{\pi_n^q\}_{n=1}^{\infty}$ such that $D_q(\lim_{n\to\infty}\pi_n^q)=\inf_{\pi\in\mathcal{P}}D_q(\pi)$, $\lim_{n\to\infty}p_{\pi_n^q}(y|x)$ exists, and $R(\theta,\lim_{n\to\infty}p_{\pi_n^q}(y|x))\le R(\theta,q(y;x))$ for every $\theta\in\Theta$.*

Proof. (1) Let $\mathcal{N}^q:=\{(x,y)\in\mathcal{X}\times\mathcal{Y}|q(y;x)=0\}$ and $\Theta^q:=\{\theta\in\Theta|\sum_{(x,y)\in\mathcal{N}^q}p(x,y|\theta)=0\}$. Let $\mathcal{P}^q$ be the set of all probability measures on Θ^q.

If $\Theta^q=\emptyset$, the assertion is obvious, because $R(\theta,q(y;x))=\infty$ for $\theta\notin\Theta^q$. We assume that $\Theta^q\ne\emptyset$ in the following. From (2), $D_q(\pi)<\infty$ if and only if $\pi\in\mathcal{P}^q$. Thus, if $\Theta^q\ne\emptyset$, then $D_q(\hat{\pi}^q)<\infty$ and $\hat{\pi}\in\mathcal{P}^q$.

Define

$$\tilde{\pi}_{\theta,u}:=u\delta_\theta+(1-u)\hat{\pi}^q,$$

for $\theta\in\Theta^q$ and $0\le u\le 1$, where δ_θ is the probability measure satisfying $\delta_\theta(\{\theta\})=1$. Then $\tilde{\pi}_{\theta,u}\in\mathcal{P}^q$, and we have

$$\frac{\partial}{\partial u}D_q(\tilde{\pi}_{\theta,u})\bigg|_{u=0}=\frac{\partial}{\partial u}\sum_{x,y}p_{\tilde{\pi}_{\theta,u}}(x,y)\log\frac{p_{\tilde{\pi}_{\theta,u}}(x,y)}{q(y;x)p_{\tilde{\pi}_{\theta,u}}(x)}\bigg|_{u=0}=\sum_{x,y}\left\{\frac{\partial}{\partial u}p_{\tilde{\pi}_{\theta,u}}(x,y)\bigg|_{u=0}\right\}\log\frac{p_{\hat{\pi}^q}(x,y)}{q(y;x)p_{\hat{\pi}^q}(x)}$$
$$=\sum_{x,y}p(x,y|\theta)\log\frac{p_{\hat{\pi}^q}(x,y)}{q(y;x)p_{\hat{\pi}^q}(x)}-\sum_{x,y}p_{\hat{\pi}^q}(x,y)\log\frac{p_{\hat{\pi}^q}(x,y)}{q(y;x)p_{\hat{\pi}^q}(x)}\ge 0.$$

Thus, if $\theta \in \Theta^q$,

$$R(\theta, p_{\hat{\pi}^q}(y|x)) = \sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\hat{\pi}^q}(y|x)} \le \sum\nolimits_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)} = R(\theta, q(y;x)) < \infty.$$

If $\theta \notin \Theta^q$, $R(\theta, q(y;x)) = \infty$ because $-\sum_{(x,y)\in\mathcal{N}^q} p(x,y|\theta)\log q(y;x) = \infty$.

Therefore, for every $\theta \in \Theta$, the inequality $R(\theta, p_{\hat{\pi}^q}(y|x)) \le R(\theta, q(y;x))$ holds.

(2) Define $\mathcal{N}^q$, Θ^q, and $\mathcal{P}^q$ as in the proof of (1). Then, Θ^q and $\mathcal{P}^q$ are compact subsets of Θ and $\mathcal{P}$, respectively.

If $\Theta^q = \emptyset$, the assertion is obvious, because $R(\theta, q(y;x)) = \infty$ for $\theta \notin \Theta^q$. We assume that $\Theta^q \neq \emptyset$ in the following. Let $\mathcal{X}^q := \{x \in \mathcal{X} | \exists\theta \in \Theta^q \text{ such that } p(x|\theta) > 0\}$ and μ^q be a probability measure on Θ^q such that $p_{\mu^q}(x) := \int p(x|\theta)\, \mathrm{d}\mu^q(\theta) > 0$ for every $x \in \mathcal{X}^q$.

Because $D_q(\pi)$ defined by (2) as a function of $\pi \in \mathcal{P}^q$ is continuous, there exists $\pi_n \in \mathcal{P}^q_{\mu^q/n} := \{(1/n)\mu^q + (1-1/n)\pi | \pi \in \mathcal{P}^q\}$ and $D_q(\pi_n) = \inf_{\pi\in\mathcal{P}^q_{\mu/n}} D_q(\pi)$. From Lemma 2, there exists a convergent subsequence $\{\pi'_m\}_{m=1}^{\infty}$ of $\{\pi_n\}_{n=1}^{\infty}$ such that $D_q(\pi'_\infty) = \inf_{\pi\in\mathcal{P}^q} D_q(\pi)$, where $\pi'_\infty = \lim_{m\to\infty}\pi'_m$.

Let n_m be the integer satisfying $\pi'_m = \pi_{n_m}$. We can take a subsequence $\{\pi'_m\}_{m=1}^{\infty}$ such that $0 < n_m/(n_{m+1}-n_m) < c$ for some positive constant c.

Since

$$\frac{n_m}{n_{m+1}}\pi'_m + \left(1-\frac{n_m}{n_{m+1}}\right)\delta_\theta = \frac{n_m}{n_{m+1}}\pi_{n_m} + \left(1-\frac{n_m}{n_{m+1}}\right)\delta_\theta \in \mathcal{P}^q_{\mu^q/n_{m+1}}$$

for every $\theta \in \Theta^q$, we have

$$\tilde{\pi}_{m,\theta,u} := u\left\{\frac{n_m}{n_{m+1}}\pi'_m + \left(1-\frac{n_m}{n_{m+1}}\right)\delta_\theta\right\} + (1-u)\pi'_{m+1} \in \mathcal{P}^q_{\mu^q/n_{m+1}}$$

for every $\theta \in \Theta^q$ and $0 \le u \le 1$. Thus,

$$\begin{aligned}
\frac{\partial}{\partial u}D_q(\tilde{\pi}_{m,\theta,u})\Big|_{u=0} &= \frac{\partial}{\partial u}\sum_{(x,y)\notin\mathcal{N}^q} p_{\tilde{\pi}_{m,\theta,u}}(x,y)\log\frac{p_{\tilde{\pi}_{m,\theta,u}}(x,y)}{q(y;x)p_{\tilde{\pi}_{m,\theta,u}}(x)}\Bigg|_{u=0} = \sum_{(x,y)\notin\mathcal{N}^q}\left\{\frac{\partial}{\partial u}p_{\tilde{\pi}_{m,\theta,u}}(x,y)\Big|_{u=0}\right\}\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} \\
&= \frac{n_m}{n_{m+1}}\sum_{(x,y)\notin\mathcal{N}^q} p_{\pi'_m}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} - \sum_{(x,y)\notin\mathcal{N}^q} p_{\pi'_{m+1}}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} \\
&\quad + \frac{n_{m+1}-n_m}{n_{m+1}}\sum_{(x,y)\notin\mathcal{N}^q} p(x,y|\theta)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} \ge 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\sum_{(x,y)\notin\mathcal{N}^q} p(x,y|\theta)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} &\ge \frac{n_{m+1}}{n_{m+1}-n_m}\sum_{(x,y)\notin\mathcal{N}^q} p_{\pi'_{m+1}}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} \\
&\quad - \frac{n_m}{n_{m+1}-n_m}\sum_{(x,y)\notin\mathcal{N}^q} p_{\pi'_m}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} \\
&= \frac{n_{m+1}}{n_{m+1}-n_m}\sum_{(x,y)\notin\mathcal{N}^q} p_{\pi'_{m+1}}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} \\
&\quad + \frac{n_m}{n_{m+1}-n_m}\left\{-\sum_{(x,y)\notin\mathcal{N}^q\cup\mathcal{N}^{\pi'_\infty}} p_{\pi'_m}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)}\right. \\
&\quad \left. - \sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}\setminus\mathcal{N}^q} p_{\pi'_m}(x,y)\log p_{\pi'_{m+1}}(y|x) + \sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}\setminus\mathcal{N}^q} p_{\pi'_m}(x,y)\log q(y;x)\right\} \\
&\ge \frac{n_{m+1}}{n_{m+1}-n_m}\sum_{(x,y)\notin\mathcal{N}^q} p_{\pi'_{m+1}}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} \\
&\quad + \frac{n_m}{n_{m+1}-n_m}\left\{-\sum_{(x,y)\notin\mathcal{N}^q\cup\mathcal{N}^{\pi'_\infty}} p_{\pi'_m}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} + \sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}\setminus\mathcal{N}^q} p_{\pi'_m}(x,y)\log q(y;x)\right\},
\end{aligned} \tag{3}$$

where $\mathcal{N}^{\pi'_\infty} := \{(x,y) \in \mathcal{X}\times\mathcal{Y} | p_{\pi'_\infty}(x,y) = 0\}$. Here, we have

$$\lim_{m\to\infty}\sum_{(x,y)\notin\mathcal{N}^q\cup\mathcal{N}^{\pi'_\infty}} p_{\pi'_m}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{q(y;x)p_{\pi'_{m+1}}(x)} = \sum_{(x,y)\notin\mathcal{N}^q\cup\mathcal{N}^{\pi'_\infty}} p_{\pi'_\infty}(x,y)\log\frac{p_{\pi'_\infty}(x,y)}{q(y;x)p_{\pi'_\infty}(x)}, \tag{4}$$

because $p_{\pi'_\infty}(x,y) > 0$ for every $(x,y) \notin \mathcal{N}^{\pi'_\infty}$, and

$$\lim_{m\to\infty} \sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}\setminus\mathcal{N}^q} p_{\pi'_m}(x,y)\log q(y;x) = 0 = -\sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}\setminus\mathcal{N}^q} p_{\pi'_\infty}(x,y)\log\frac{p_{\pi'_\infty}(x,y)}{q(y;x)p_{\pi'_\infty}(x)}. \tag{5}$$

Therefore, from (3)–(5), and $0 < n_m/(n_{m+1}-n_m) < c$, for every $\theta \in \Theta^q$,

$$\liminf_{m\to\infty} \sum_{(x,y)\notin\mathcal{N}^q} p(x,y|\theta)\log\frac{p_{\pi'_m}(x,y)}{q(y;x)p_{\pi'_m}(x)} \geq \sum_{(x,y)\notin\mathcal{N}^q} p_{\pi'_\infty}(x,y)\log\frac{p_{\pi'_\infty}(x,y)}{q(y;x)p_{\pi'_\infty}(x)} \geq 0. \tag{6}$$

By taking an appropriate subsequence $\{\pi''_k\}_{k=1}^\infty$ of $\{\pi'_m\}_{m=1}^\infty$, we can make the sequences of real numbers $\{p_{\pi''_k}(y|x)\}_{k=1}^\infty$ converge for all $(x,y) \in \mathcal{X}^q \times \mathcal{Y}$ because $p_{\pi'_m}(x) > 0$ $(x \in \mathcal{X}^q)$ and $0 \leq p_{\pi'_m}(x,y)/p_{\pi'_m}(x) \leq 1$.

Then, from (6), if $\theta \in \Theta^q$,

$$\begin{aligned} R\left(\theta, \lim_{k\to\infty} p_{\pi''_k}(y|x)\right) &= \sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{\lim_{k\to\infty} p_{\pi''_k}(y|x)} = \sum_{(x,y)\notin\mathcal{N}^q} p(x,y|\theta)\log\frac{p(y|x,\theta)}{\lim_{k\to\infty} p_{\pi''_k}(y|x)} \\ &\leq \sum_{(x,y)\notin\mathcal{N}^q} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)} = \sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)} = R(\theta, q(y;x)) < \infty. \end{aligned}$$

Note that although $\lim_{k\to\infty} p_{\pi''_k}(y|x)$ is not uniquely determined for $x \notin \mathcal{X}^q$, the risk $R(\theta, \lim_{k\to\infty} p_{\pi''_k}(y|x))$ does not depend on the choice of $\lim_{k\to\infty} p_{\pi''_k}(y|x)$ for such x, because $p(x|\theta) = 0$ if $\theta \in \Theta^q$ and $x \notin \mathcal{X}^q$.

If $\theta \notin \Theta^q$, $R(\theta, q(y;x)) = \infty$ because $-\sum_{(x,y)\in\mathcal{N}^q} p(x,y|\theta)\log q(y;x) = \infty$.

Hence, the risk of the predictive density defined by

$$\begin{cases} \lim_{k\to\infty} p_{\pi''_k}(y|x), & x \in \mathcal{X}^q \\ r(y;x), & x \notin \mathcal{X}^q, \end{cases}$$

where $r(y;x)$ is an arbitrary predictive density, is not greater than that of $q(y;x)$ for every $\theta \in \Theta$.

Therefore, by taking a sequence $\{\varepsilon_n \in (0,1)\}_{n=1}^\infty$ that converges rapidly enough to 0, we can construct a predictive density

$$\lim_{k\to\infty} p_{\varepsilon_k\overline{\mu}+(1-\varepsilon_k)\pi''_k}(y|x) = \begin{cases} \lim_{k\to\infty} p_{\pi''_k}(y|x), & x \in \mathcal{X}^q \\ p_{\overline{\mu}}(y|x), & x \notin \mathcal{X}^q \end{cases} \tag{7}$$

as a limit of Bayesian predictive densities based on priors $\varepsilon_k\overline{\mu}+(1-\varepsilon_k)\pi''_k$, where $\overline{\mu}$ is a measure on Θ such that $p_{\overline{\mu}}(x) > 0$ for every $x \in \mathcal{X}$.

Hence, the risk of the predictive density (7) is not greater than that of $q(y;x)$ for every $\theta \in \Theta$. □

We give two simple examples to clarify the meaning of Theorem 1 and its proof.

Example 1. Suppose that $\mathcal{X} = \{0,1,2\}$, $\mathcal{Y} = \{0,1\}$, $p(x,y|\theta) = \binom{2}{x}\theta^x(1-\theta)^{2-x}\theta^y(1-\theta)^{1-y}$, and $\Theta = [0,1]$. Let $q(y;x) = (x/2)^y(1-x/2)^{(1-y)}$, which is the plug-in predictive density with the maximum likelihood estimate $\hat{\theta} = x/2$. Then, $\mathcal{N}^q = \{(0,1),(2,0)\}$, $\Theta^q = \{0,1\}$, and $\mathcal{X}^q = \{0,2\}$. The prior defined by $\pi^{(w)} := w\delta_0 + (1-w)\delta_1 \in \mathcal{P}^q$ $(0 < w < 1)$ satisfies

$$D_q(\pi^{(w)}) = \inf_{\pi\in\mathcal{P}^q} D_q(\pi) = 0.$$

We set $\mu^q = \pi^{(w)}$, which satisfies $p_{\mu^q}(x) > 0$ for $x \in \mathcal{X}^q$. Then, we can set $\pi_n = \pi^{(w)}$ $(n = 1,2,3,\ldots)$ because $\pi^{(w)} \in \mathcal{P}^q_{\mu^q/n}$ and $D_q(\pi^{(w)}) = 0$. Then, $\lim_{n\to\infty} p_{\pi_n}(y|x) = p_{\pi^{(w)}}(y|x)$. Thus, $\pi'_\infty = \pi^{(w)}$ and $\mathcal{N}^{\pi'_\infty} = \mathcal{N}^q$.

The prior $\pi^{(w)}$ does not specify the conditional density $p_{\pi^{(w)}}(y|x=1)$ because $p_{\pi^{(w)}}(x=1) = 0$. We set $\overline{\mu}(\mathrm{d}\theta) = \mathrm{d}\theta$ and

$$\pi''_k = \frac{1}{k}\overline{\mu} + \left(1-\frac{1}{k}\right)\pi^{(w)}.$$

Then, $\lim_{k\to\infty} p_{\pi''_k}(y=0|x=0) = \lim_{k\to\infty} p_{\pi''_k}(y=1|x=2) = 1$ and $\lim_{k\to\infty} p_{\pi''_k}(y=0|x=1) = \lim_{k\to\infty} p_{\pi_k}(y=1|x=1) = 1/2$. The risk function of the predictive density $\lim_{k\to\infty} p_{\pi''_k}(y|x)$, which is a limit of the Bayesian predictive densities, is given by

$$R\left(\theta, \lim_{k\to\infty} p_{\pi''_k}(y|x)\right) = \begin{cases} 0, & \theta = 0 \in \Theta^q, \\ \infty, & \theta \in (0,1) = \Theta\setminus\Theta^q, \\ 0, & \theta = 1 \in \Theta^q \end{cases}$$

and coincides with $R(\theta, q(y;x))$.

Example 2. Suppose that $\mathcal{X} = \{0,1,2\}$, $\mathcal{Y} = \{0,1\}$, $\Theta = \{\theta_1,\theta_2\}$, $p((2,0)|\theta_1) = p((2,1)|\theta_1) = 0$, $p((0,0)|\theta_1) = p((1,1)|\theta_1) = 1/3$, $p((0,1)|\theta_1) = p((1,0)|\theta_1) = 1/6$, $p((2,0)|\theta_2) = p((2,1)|\theta_2) = (1-\varepsilon)/2$, and $p((0,0)|\theta_2) = p((0,1)|\theta_2) = p((1,0)|\theta_2) = p((1,1)|\theta_2) = \varepsilon/4$, where $0 < \varepsilon < 1$.

Consider a predictive density defined by $q(y=0;x=0)=q(y=1;x=1)=2/3$, $q(y=1;x=0)=q(y=0;x=1)=1/3$, $q(y=0;x=2)=1/3$, and $q(y=1;x=2)=2/3$. Then, $\mathcal{N}^q=\emptyset$, $\Theta^q=\Theta$, $\mathcal{P}^q=\mathcal{P}$, and $\mathcal{X}^q=\mathcal{X}$.

Then, $\hat{\pi}=\delta_{\theta_1}$ satisfies $D_q(\hat{\pi})=\inf_{\pi\in\mathcal{P}} D_q(\pi)=0$ because $p(y|x,\theta_1)=q(y;x)$ except for the case x=2. Since $p(x=2|\theta_1)=0$, $p_{\hat{\pi}}(y|x=2)$ is not uniquely determined. Thus, we consider a limit of Bayesian predictive densities.

Put $\mu=\delta_{\theta_1}/2+\delta_{\theta_2}/2$. It can be easily verified that $\pi_n=(1/n)\mu+(1-1/n)\delta_{\theta_1}$ satisfies $D_q(\pi_n)=\inf_{\pi\in\mathcal{P}_{\mu/n}} D_q(\pi)$. Then, $\lim_{n\to\infty} p_{\pi_n}(y|x=0)=p(y|x=0,\theta_1)=q(y;x=0)$, $\lim_{n\to\infty} p_{\pi_n}(y|x=1)=p(y|x=1,\theta_1)=q(y;x=1)$, $p_{\pi_n}(y|x=2)=p(y|x=\ 2,\theta_2)$ $\neq q(y;x=2)$. By calculation, we have $R(\theta_1,\lim_{n\to\infty} p_{\pi_n}(y|x))=R(\theta_1,q(y;x))=0$ and $R(\theta_2,\lim_{n\to\infty} p_{\pi_n}(y|x))=(\varepsilon/2)\log(9/8)<$ $R(\theta_2,q(y;x))=(1/2)\log(9/8)$. Thus, the performance of $\lim_{n\to\infty} p_{\pi_n}(y|x)$ is better than that of $q(y;x)$.

3. Latent information priors and minimax prediction

In this section, we construct minimax predictive densities that are limits of Bayesian predictive densities based on prior sequences converging to latent information priors defined below.

A predictive density $q(y;x)$ is said to be minimax if it satisfies the equality

$$\sup_{\theta\in\Theta}\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}=\inf_{\overline{q}}\sup_{\theta\in\Theta}\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{\overline{q}(y;x)}.$$

The conditional mutual information between y and θ given x is defined by

$$I_{\theta,y|x}(\pi):=\int\sum_{x,y} p(x,y|\theta)\log p(x,y|\theta)\,\mathrm{d}\pi(\theta)-\sum_{x,y} p_\pi(x,y)\log p_\pi(x,y)-\int\sum_{x} p(x|\theta)\log p(x|\theta)\,\mathrm{d}\pi(\theta)+\sum_{x} p_\pi(x)\log p_\pi(x),$$

which is a function of $\pi\in\mathcal{P}$. If $p_\pi(x)\neq 0$ for all $x\in\mathcal{X}$, then

$$I_{\theta,y|x}(\pi)=\int\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_\pi(y|x)}\,\mathrm{d}\pi(\theta).$$

Here, $I_{\theta,y|x}(\pi)$ is a quantity averaged over x. This definition of conditional mutual information is widely adopted in information theory; see, for example, Cover and Thomas (2006, p. 23). Since $u\log u$ $(0\leq u\leq 1)$ a bounded continuous function, $I_{\theta,y|x}(\pi)$ is a bounded continuous function of $\pi\in\mathcal{P}$.

We define a latent information prior as a prior $\hat{\pi}$ that satisfies $I_{\theta,y|x}(\hat{\pi})=\sup_{\pi\in\mathcal{P}} I_{\theta,y|x}(\pi)$. Intuitively speaking, when the parameter θ is distributed according to the latent information prior, θ has the maximum information about the future observable y under the condition that x is observed. Therefore, θ has the maximum amount of "latent" information, which we cannot observe through the data x. Thus, the latent information prior corresponds to the "worst case" and is naturally related to minimaxity. On the other hand, the minimum information prior discussed by Akaike (1983) is a prior maximizing the mutual information between the future observable y and the data x. This prior corresponds to the "best case" and is far from minimaxity.

The priors $\hat{\pi}$ and π_∞ in Theorem 2 below are the latent information priors.

Theorem 2.

(1) *Let $\hat{\pi}\in\mathcal{P}$ be a prior maximizing $I_{\theta,y|x}(\pi)$. If $p_{\hat{\pi}}(x)>0$ for all $x\in\mathcal{X}$, then $p_{\hat{\pi}}(y|x)$ is a minimax predictive density.*

(2) *There exists a convergent prior sequence $\{\pi_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} p_{\pi_n}(y|x)$ is a minimax predictive density and the equality $I_{\theta,y|x}(\pi_\infty)=\sup_{\pi\in\mathcal{P}} I_{\theta,y|x}(\pi)$ holds, where $\pi_\infty=\lim_{n\to\infty}\pi_n$.*

Proof. (1) Define $\tilde{\pi}_{\overline{\theta},u}:=u\delta_{\overline{\theta}}+(1-u)\hat{\pi}$ for $\overline{\theta}\in\Theta$ and $0\leq u\leq 1$. Then,

$$\begin{aligned}
\frac{\partial}{\partial u}I_{\theta,y|x}(\tilde{\pi}_{\overline{\theta},u})\bigg|_{u=0}&=\frac{\partial}{\partial u}\Bigg(\int\sum_{x,y} p(x,y|\theta)\log p(x,y|\theta)\,\mathrm{d}\tilde{\pi}_{\overline{\theta},u}(\theta)-\sum_{x,y} p_{\tilde{\pi}_{\overline{\theta},u}}(x,y)\log p_{\tilde{\pi}_{\overline{\theta},u}}(x,y)\\
&\quad-\int\sum_{x} p(x|\theta)\log p(x|\theta)\,\mathrm{d}\tilde{\pi}_{\overline{\theta},u}(\theta)+\sum_{x} p_{\tilde{\pi}_{\overline{\theta},u}}(x)\log p_{\tilde{\pi}_{\overline{\theta},u}}(x)\Bigg)\Bigg|_{u=0}\\
&=\sum_{x,y} p(x,y|\overline{\theta})\log p(x,y|\overline{\theta})-\int\sum_{x,y} p(x,y|\theta)\log p(x,y|\theta)\,\mathrm{d}\hat{\pi}(\theta)\\
&\quad-\sum_{x,y}\frac{\partial}{\partial u}p_{\tilde{\pi}_{\overline{\theta},u}}(x,y)\bigg|_{u=0}\log p_{\hat{\pi}}(x,y)-\sum_{x} p(x|\overline{\theta})\log p(x|\overline{\theta})\\
&\quad+\int\sum_{x} p(x|\theta)\log p(x|\theta)\,\mathrm{d}\hat{\pi}(\theta)+\sum_{x}\frac{\partial}{\partial u}p_{\tilde{\pi}_{\overline{\theta},u}}(x)\bigg|_{u=0}\log p_{\hat{\pi}}(x)\\
&=\sum_{x,y} p(x,y|\overline{\theta})\log\frac{p(x,y|\overline{\theta})}{p(x|\overline{\theta})}-\sum_{x,y} p(x,y|\overline{\theta})\log\frac{p_{\hat{\pi}}(x,y)}{p_{\hat{\pi}}(x)}
\end{aligned}$$

$$-\int \sum_{x,y} p(x,y|\theta)\log\frac{p(x,y|\theta)}{p(x|\theta)}\,\mathrm{d}\hat{\pi}(\theta)+\sum_{x,y} p_{\hat{\pi}}(x,y)\log\frac{p_{\hat{\pi}}(x,y)}{p_{\hat{\pi}}(x)}\le 0.$$

Noting that $p_{\hat{\pi}}(x)>0$ for every $x\in\mathcal{X}$ and that $p(x,y|\theta)\log p(y|x,\theta)=0$ if $p(x|\theta)=0$, we have

$$\sum_{x,y} p(x,y|\overline{\theta})\log\frac{p(y|x,\overline{\theta})}{p_{\hat{\pi}}(y|x)}\le\int\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\hat{\pi}}(y|x)}\,\mathrm{d}\hat{\pi}(\theta) \tag{8}$$

for every $\overline{\theta}\in\Theta$.

On the other hand, we have

$$\begin{aligned}\int\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\hat{\pi}}(y|x)}\,\mathrm{d}\hat{\pi}(\theta)&=\inf_q\int\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\hat{\pi}(\theta)\\ &\le\sup_{\pi\in\mathcal{P}}\inf_q\int\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\pi(\theta)\le\inf_q\sup_{\pi\in\mathcal{P}}\int\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\pi(\theta)\\ &=\inf_q\sup_{\theta\in\Theta}\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\le\sup_{\theta\in\Theta}\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\hat{\pi}}(y|x)}.\end{aligned} \tag{9}$$

The first equality is because the Bayes risk

$$\int R(\theta;q(y;x))\,\mathrm{d}\hat{\pi}(\theta)=\int\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\hat{\pi}(\theta)$$

with respect to $\hat{\pi}\in\mathcal{P}$ is minimized when

$$q(y;x)=p_{\hat{\pi}}(y|x):=\frac{\int p(x,y|\theta)\,\mathrm{d}\hat{\pi}(\theta)}{\int p(x|\theta)\,\mathrm{d}\hat{\pi}(\theta)};$$

see Aitchison (1975).

From (8) and (9), we have

$$\inf_q\sup_{\theta\in\Theta}\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}=\sup_{\theta\in\Theta}\sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\hat{\pi}}(y|x)}.$$

Therefore, the predictive density $p_{\hat{\pi}}(y|x)$ is minimax.

(2) Let μ be a probability measure on Θ such that $p_\mu(x):=\int p(x|\theta)\,\mathrm{d}\mu(\theta)>0$ for every $x\in\mathcal{X}$, and let $\pi_n\in\mathcal{P}_{\mu/n}:=\{\mu/n+(1-1/n)\pi|\pi\in\mathcal{P}\}$ be a prior satisfying $I_{\theta,y|x}(\pi_n)=\sup_{\pi\in\mathcal{P}_{\mu/n}} I_{\theta,y|x}(\pi)$. From Lemma 2, there exists a convergent subsequence $\{\pi'_m\}_{m=1}^\infty$ of $\{\pi_n\}_{n=1}^\infty$ and $I_{\theta,y|x}(\pi'_\infty)=\sup_{\pi\in\mathcal{P}} I_{\theta,y|x}(\pi)$, where $\pi'_\infty=\lim_{m\to\infty}\pi'_m$. Let n_m be the integer satisfying $\pi'_m=\pi_{n_m}$. As in the proof of Theorem 1, we can take a subsequence $\{\pi'_m\}_{m=1}^\infty$ such that $0<n_m/(n_{m+1}-n_m)<c$ for some positive constant c.

Then, for every $\overline{\theta}\in\Theta$,

$$\tilde{\pi}_{m,\overline{\theta},u}:=u\left\{\frac{n_m}{n_{m+1}}\pi'_m+\left(1-\frac{n_m}{n_{m+1}}\right)\delta_{\overline{\theta}}\right\}+(1-u)\pi'_{m+1}$$

belongs to $\mathcal{P}_{\mu/n_{m+1}}$ for $0\le u\le 1$, because $(n_m/n_{m+1})\pi'_m+(1-n_m/n_{m+1})\delta_{\overline{\theta}}\in\mathcal{P}_{\mu/n_{m+1}}$ and $\pi'_{m+1}\in\mathcal{P}_{\mu/n_{m+1}}$.

Thus,

$$\begin{aligned}\frac{\partial}{\partial u}I_{\theta,y|x}(\tilde{\pi}_{m,\overline{\theta},u})\bigg|_{u=0}&=\frac{\partial}{\partial u}\left(\int\sum_{x,y} p(x,y|\theta)\log p(x,y|\theta)\,\mathrm{d}\tilde{\pi}_{m,\overline{\theta},u}(\theta)-\sum_{x,y} p_{\tilde{\pi}_{m,\overline{\theta},u}}(x,y)\log p_{\tilde{\pi}_{m,\overline{\theta},u}}(x,y)\right.\\ &\qquad\left.-\int\sum_{x} p(x|\theta)\log p(x|\theta)\,\mathrm{d}\tilde{\pi}_{m,\overline{\theta},u}(\theta)+\sum_{x} p_{\tilde{\pi}_{m,\overline{\theta},u}}(x)\log p_{\tilde{\pi}_{m,\overline{\theta},u}}(x)\right)\bigg|_{u=0}\\ &=\frac{n_m}{n_{m+1}}\int\sum_{x,y} p(x,y|\theta)\log p(x,y|\theta)\,\mathrm{d}\pi'_m(\theta)+\left(1-\frac{n_m}{n_{m+1}}\right)\sum_{x,y} p(x,y|\overline{\theta})\log p(x,y|\overline{\theta})\\ &\qquad-\int\sum_{x,y} p(x,y|\theta)\log p(x,y|\theta)\,\mathrm{d}\pi'_{m+1}(\theta)-\sum_{x,y}\frac{\partial}{\partial u}p_{\tilde{\pi}_{m,\overline{\theta},u}}(x,y)\bigg|_{u=0}\log p_{\pi'_{m+1}}(x,y)\\ &\qquad-\frac{n_m}{n_{m+1}}\int\sum_{x} p(x|\theta)\log p(x|\theta)\,\mathrm{d}\pi'_m(\theta)-\left(1-\frac{n_m}{n_{m+1}}\right)\sum_{x} p(x|\overline{\theta})\log p(x|\overline{\theta})\\ &\qquad+\int\sum_{x} p(x|\theta)\log p(x|\theta)\,\mathrm{d}\pi'_{m+1}(\theta)+\sum_{x}\frac{\partial}{\partial u}p_{\tilde{\pi}_{m,\overline{\theta},u}}(x)\bigg|_{u=0}\log p_{\pi'_{m+1}}(x)\\ &=\left(1-\frac{n_m}{n_{m+1}}\right)\sum_{x,y} p(x,y|\overline{\theta})\log\frac{p(x,y|\overline{\theta})}{p(x|\overline{\theta})}-\left(1-\frac{n_m}{n_{m+1}}\right)\sum_{x,y} p(x,y|\overline{\theta})\log\frac{p_{\pi'_{m+1}}(x,y)}{p_{\pi'_{m+1}}(x)}\end{aligned}$$

$$\begin{aligned}&+\frac{n_m}{n_{m+1}}\int\sum_{x,y}p(x,y|\theta)\log\frac{p(x,y|\theta)}{p(x|\theta)}\,\mathrm{d}\pi'_m(\theta)-\int\sum_{x,y}p(x,y|\theta)\log\frac{p(x,y|\theta)}{p(x|\theta)}\,\mathrm{d}\pi'_{m+1}(\theta)\\&-\frac{n_m}{n_{m+1}}\sum_{x,y}p_{\pi'_m}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{p_{\pi'_{m+1}}(x)}+\sum_{x,y}p_{\pi'_{m+1}}(x,y)\log\frac{p_{\pi'_{m+1}}(x,y)}{p_{\pi'_{m+1}}(x)}\le 0.\end{aligned}$$

Noting that $p_{\pi'_m}(x)>0$ for every m and $x\in\mathcal{X}$ and that $p(x,y|\theta)\log p(y|x,\theta)=0$ if $p(x|\theta)=0$, we have

$$\begin{aligned}&\left(1-\frac{n_m}{n_{m+1}}\right)\sum_{x,y}p(x,y|\overline{\theta})\log\frac{p(y|x,\overline{\theta})}{p_{\pi'_{m+1}}(y|x)}+\frac{n_m}{n_{m+1}}\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\pi'_{m+1}}(y|x)}\,\mathrm{d}\pi'_m(\theta)\\&\quad-\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\pi'_{m+1}}(y|x)}\,\mathrm{d}\pi'_{m+1}(\theta)\le 0.\end{aligned}$$

Hence,

$$\begin{aligned}\sum_{x,y}p(x,y|\overline{\theta})\log\frac{p(y|x,\overline{\theta})}{p_{\pi'_{m+1}}(y|x)}\le&-\frac{n_m}{n_{m+1}-n_m}\left\{\int\sum_{(x,y)\notin\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\pi'_{m+1}}(y|x)}\,\mathrm{d}\pi'_m(\theta)\right.\\&\left.+\int\sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log p(y|x,\theta)\,\mathrm{d}\pi'_m(\theta)-\int\sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log p_{\pi'_{m+1}}(y|x)\,\mathrm{d}\pi'_m(\theta)\right\}\\&+\frac{n_{m+1}}{n_{m+1}-n_m}\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\pi'_{m+1}}(y|x)}\,\mathrm{d}\pi'_{m+1}(\theta)\\\le&-\frac{n_m}{n_{m+1}-n_m}\left\{\int\sum_{(x,y)\notin\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\pi'_{m+1}}(y|x)}\,\mathrm{d}\pi'_m(\theta)+\int\sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log p(y|x,\theta)\,\mathrm{d}\pi'_m(\theta)\right\}\\&+\frac{n_{m+1}}{n_{m+1}-n_m}\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\pi'_{m+1}}(y|x)}\,\mathrm{d}\pi'_{m+1}(\theta),\end{aligned}\tag{10}$$

where $\mathcal{N}^{\pi'_\infty}:=\{(x,y)\in\mathcal{X}\times\mathcal{Y}|p_{\pi'_\infty}(x,y)=0\}$. Here, we have

$$\lim_{m\to\infty}\int\sum_{(x,y)\notin\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log\frac{p(y|x,\theta)}{p_{\pi'_{m+1}}(y|x)}\,\mathrm{d}\pi'_m(\theta)=\int\sum_{(x,y)\notin\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log\frac{p(y|x,\theta)p_{\pi'_\infty}(x)}{p_{\pi'_\infty}(x,y)}\,\mathrm{d}\pi'_\infty(\theta)\tag{11}$$

and

$$\begin{aligned}\lim_{m\to\infty}\int\sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log p(y|x,\theta)\,\mathrm{d}\pi'_m(\theta)&=\int\sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log p(y|x,\theta)\,\mathrm{d}\pi'_\infty(\theta)\\&=\int\sum_{(x,y)\in\mathcal{N}^{\pi'_\infty}}p(x,y|\theta)\log\frac{p(y|x,\theta)p_{\pi'_\infty}(x)}{p_{\pi'_\infty}(x,y)}\,\mathrm{d}\pi'_\infty(\theta)=0,\end{aligned}\tag{12}$$

because $p(x,y|\theta)\log p(x,y|\theta)$ and $p(x|\theta)\log p(x|\theta)$ are bounded continuous functions of θ for every fixed (x,y).

From (10)–(12), and $0<n_m/(n_{m+1}-n_m)<c$, we have, for every $\overline{\theta}\in\Theta$,

$$\limsup_{m\to\infty}\sum_{x,y}p(x,y|\overline{\theta})\log\frac{p(y|x,\overline{\theta})}{p_{\pi'_m}(y|x)}\le\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)p_{\pi'_\infty}(x)}{p_{\pi'_\infty}(x,y)}\mathrm{d}\pi'_\infty(\theta).$$

By taking an appropriate subsequence $\{\pi''_k\}_{k=1}^\infty$ of $\{\pi'_m\}_{m=1}^\infty$, we can make $\{p_{\pi''_k}(y|x)\}_{k=1}^\infty$ converges for every (x,y) as $k\to\infty$. Then, for every $\overline{\theta}\in\Theta$,

$$\sum_{x,y}p(x,y|\overline{\theta})\log\frac{p(y|x,\overline{\theta})}{\lim_{k\to\infty}p_{\pi''_k}(y|x)}\le\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{\lim_{k\to\infty}p_{\pi''_k}(y|x)}\,\mathrm{d}\pi''_\infty(\theta),\tag{13}$$

where $\pi''_\infty=\pi'_\infty=\lim_{k\to\infty}\pi''_k$, because $\lim_{k\to\infty}p_{\pi''_k}(y|x)=p_{\pi''_\infty}(y|x)$ for x with $p_{\pi''_\infty}(x)>0$.

On the other hand, we have

$$\begin{aligned}\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{\lim_{k\to\infty}p_{\pi''_k}(y|x)}\,\mathrm{d}\pi''_\infty(\theta)&=\inf_q\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\pi''_\infty(\theta)\\&\le\sup_{\pi\in\mathcal{P}}\inf_q\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\pi(\theta)\le\inf_q\sup_{\pi\in\mathcal{P}}\int\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\pi(\theta)\\&=\inf_q\sup_{\theta\in\Theta}\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\le\sup_{\theta\in\Theta}\sum_{x,y}p(x,y|\theta)\log\frac{p(y|x,\theta)}{\lim_{k\to\infty}p_{\pi''_k}(y|x)}.\end{aligned}\tag{14}$$

The first equality is because the Bayes risk

$$\int R(\theta; q(y;x))\,\mathrm{d}\pi''_{\infty}(\theta) = \int \sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)}\,\mathrm{d}\pi''_{\infty}(\theta)$$

is minimized when $q(y;x) = p_{\pi''_{\infty}}(y|x)$; see Aitchison (1975). Although $p_{\pi''_{\infty}}(y|x)$ is not uniquely determined for x with $p_{\pi''_{\infty}}(x) = 0$, the Bayes risk does not depend on the choice of $p_{\pi''_{\infty}}(y|x)$ for such x.

From (13) and (14), we have

$$\inf_q \sup_{\theta\in\Theta} \sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{q(y;x)} = \sup_{\theta\in\Theta} \sum_{x,y} p(x,y|\theta)\log\frac{p(y|x,\theta)}{\lim_{k\to\infty} p_{\pi''_k}(y|x)}.$$

Therefore, the predictive density $\lim_{k\to\infty} p_{\pi''_k}(y|x)$ is minimax. □

4. Numerical results and discussions

Let $p(x|\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x}$ $(x = 0,1,\ldots,N)$, $p(y|\theta) = \binom{M}{y}\theta^y(1-\theta)^{M-y}$ $(y = 0,1,\ldots,M)$, and $\Theta = \{0.1k|k = 0,1,2,\ldots,10\}$ in which θ takes a value. Although this example is relatively simple in the sense that x and y are independent given θ, the behavior of priors is not trivial.

The latent information priors, which maximize $I_{\theta,y|x}(\pi)$, for 16 sets of values of (N,M) are obtained numerically; see Fig. 1.

The prior for $(N,M) = (0,1000)$ is almost uniform and is similar to the reference prior because the reference prior is the latent information prior with N=0 and $M\to\infty$. It is widely known the reference prior is uniform when the parameter space

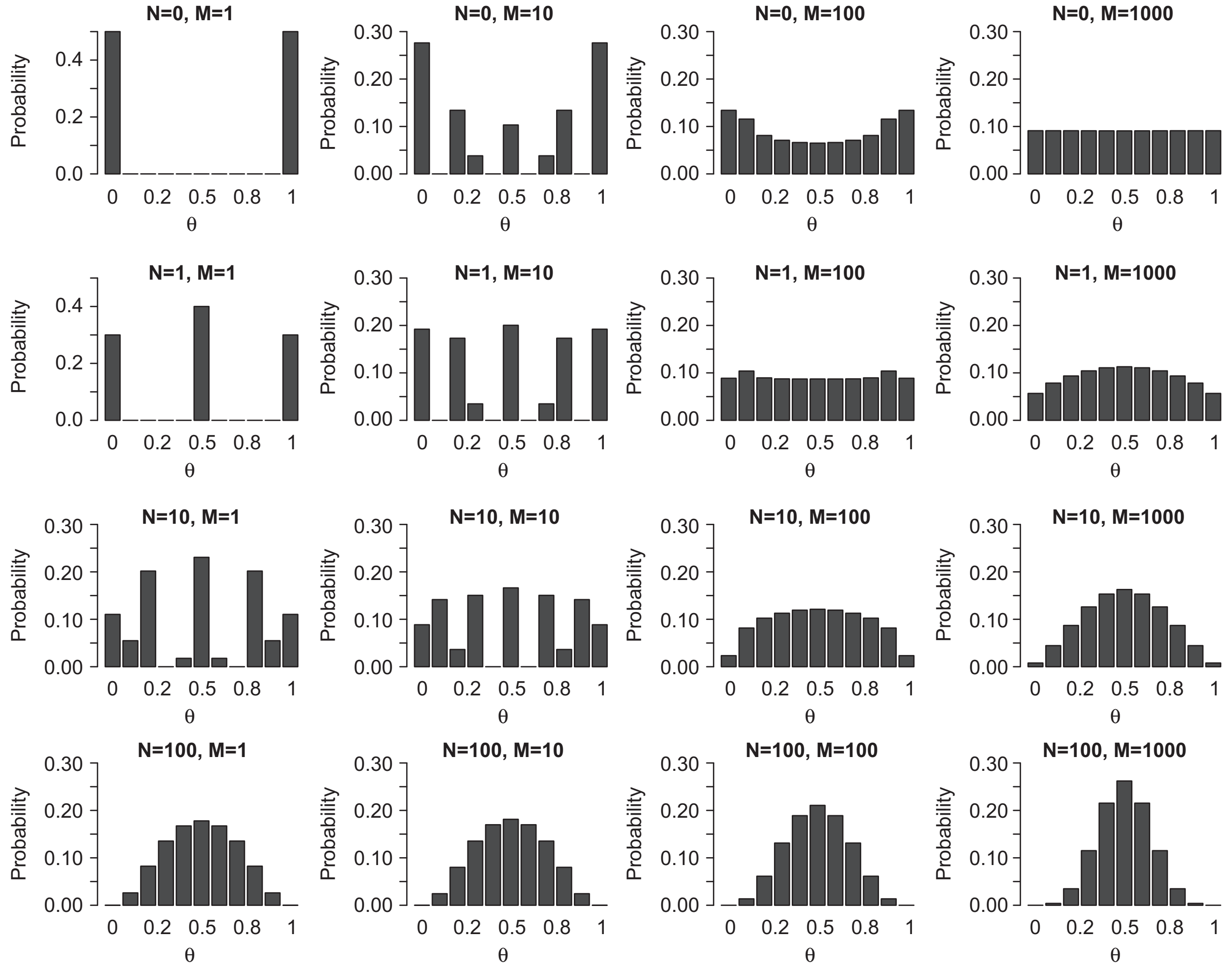

Fig. 1. Latent information priors for various (N,M) values.

is a finite set. The latent information prior for $(N,M)=(0,100)$ is similar to the histogram of the Jeffreys prior density $\theta^{-1/2}(1-\theta)^{-1/2}/B(1/2,1/2)$ for the binomial model with the ordinary parameter space $\Theta'=[0,1]$, which is different from the parameter space $\Theta=\{0.1k|k=0,1,2,\dots,10\}$, a finite subset of $\Theta'=[0,1]$, adopted here.

When $N=0$ and M is moderately large ($M=100$), the latent information prior is similar to the histogram of the Jeffreys prior density $\theta^{-1/2}(1-\theta)^{-1/2}/B(1/2,1/2)$, which is the reference prior on $\Theta'=[0,1]$. When $N=0$ and M is extremely large ($M=1000$), the latent information prior on Θ becomes almost uniform and is dissimilar from the histogram of the Jeffreys prior density on $\Theta'=[0,1]$. This is because we can distinguish almost completely all the points in the discrete parameter space Θ by using the information of y.

When both of N and M are small the priors assign weights only on a limited number of points in Θ. This corresponds to the phenomenon concerning the k-reference prior studied by Berger et al. (1989). The k-reference prior is the latent information prior with $N=0$ and $M=k$.

When N is large, the priors assign more weights to parameter values close to 0.5. The shapes of priors are quite different from the uniform density or the histogram of the Jeffreys prior for the binomial model with the ordinary parameter space $\Theta'=[0,1]$.

These observations show that the latent information priors strongly depend on (N,M). This indicates that we need to abandon the context invariance (see Dawid, 1983) of priors.

The relation between the conditional mutual information and predictive densities parallels to that between the unconditional mutual information and Bayes codes in information theory except for the care for the case $p_\pi(x)=0$. Many studies on the unconditional mutual information and minimax prediction and coding have been carried out; see, for example, Ibragimov and Hasminskii (1973), Gallager (1979), Davisson and Leon-Garcia (1980), Clarke and Barron (1994), and Haussler (1997). See also Grünwald and Dawid (2004) for discussions in a very general setting. Conditional mutual information is a fundamental quantity in information theory and naturally appeared in several previous studies in statistics such as Clarke and Yuan (2004), and Ebrahimi et al. (2010). The conditional mutual information $I_{\theta,y|x}(\pi)$ coincides with the Bayes risk of the Bayesian predictive density based on π. Therefore, it is natural that the prior maximizing $I_{\theta,y|x}(\pi)$ corresponds to minimax prediction based on data.

In general, the priors based on the unconditional mutual information and that based on the conditional mutual information are quite different. Latent information priors maximizing the conditional mutual information could play important roles in statistical applications. Although we have discussed submodels of multinomial models, essential part of our discussion seem to hold for more general models such as x and y are continuous random variables under suitable regularity conditions including compactness of the model as in the theory based on the unconditional mutual information studied by Haussler (1997).

The explicit forms of latent information priors are usually complex and difficult to obtain unless the parameter space is finite. For actual applications, it is important to develop approximation methods and asymptotic theory in various settings other than the situation $N=0, M\to\infty$ studied in the reference analysis. When $I_{\theta,y|x}(\pi)$ is close to $I_{\theta,y|x}(\hat{\pi})$, a prior π is considered to be close to $\hat{\pi}$ because $I_{\theta,y|x}(\pi)$ is a concave function of π. These topics require further research and will be discussed in other places.

Acknowledgments

The author would like to thank the associate editor and the referees for their constructive comments. This research was partially supported by Grant-in-Aid for Scientific Research (23300104, 23650144) and by the Aihara Project, the FIRST program from JSPS, initiated by CSTP.

References

Aitchison, J., 1975. Goodness of prediction fit. Biometrika 62, 547–554.

Akaike, H., 1983. On minimum information prior distributions. Annals of the Institute of Statistical Mathematics 35 (Part A), 139–149.

Berger, J.O., Bernardo, J.M., Mendoza, M., 1989. On priors that maximize expected information. In: Klein, J., Lee, J. (Eds.), Recent Developments of Statistics and its Applications. Freedom Academy, Seoul, pp. 1–20.

Bernardo, J.M., 1979. Reference posterior distributions for Bayesian inference (with discussion). Journal of Royal Statistical Society B 41, 113–147.

Bernardo, J.M., 2005. Reference analysis. In: Dey, K.K., Rao, C.R. (Eds.), Handbook of Statistics, vol. 25. Elsevier, Amsterdam, pp. 17–90.

Clarke, B.S., Barron, A.R., 1994. Jeffreys' prior is asymptotically least favorable under entropy risk. Journal of Statistical Planning and Inference 41, 36–60.

Clarke, B.S., Yuan, A., 2004. Partial information reference priors: derivation and interpretations. Journal of Statistical Planning and Inference 123, 313–345.

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory, second ed. Wiley-Interscience.

Davisson, L., Leon-Garcia, A., 1980. A source matching approach to finding minimax codes. IEEE Transactions on Information Theory 26, 166–174.

Dawid, A.P., 1983. Invariant Prior Distributions. In: Kotz, S., Johnson, N.L., Read, C.B. (Eds.), Encyclopedia of Statistical Sciences, vol. 4. Wiley-Interscience, New York, pp. 228–236.

Ebrahimi, N., Soofi, E.S., Soyer, R., 2010. On the sample information about parameter and prediction. Statistical Science 25, 348–367.

Gallager, R., 1979. Source coding with side information and universal coding. Technical Report LIDSP-937, M.I.T. Laboratory for Information and Decision Systems.

Geisser, S., 1979. Discussion on Reference posterior distributions for Bayesian inference by J.M. Bernardo. Journal of Royal Statistical Society B 41, 136–137.

Grünwald, P.D., Dawid, A.P., 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Annals of Statistics 32, 1367–1433.

Haussler, D., 1997. A general minimax result for relative entropy. IEEE Transactions on Information Theory 43, 1276–1280.

Ibragimov, I.A., Hasminskii, R.Z., 1973. On the information contained in a sample about a parameter. In: Second International Symposium on Information Theory. Akademiai, Kiado, Budapest, pp. 295–309.

Komaki, F., 2004. Simultaneous prediction of independent Poisson observables. Annals of Statistics 32, 1744–1769.

Kuboki, H., 1998. Reference priors for prediction. Journal of Statistical Planning and Inference 69, 295–317.