mathematics

Article

On Consistency of the Bayes Estimator of the Density

Agustín G. Nogales

Departamento de Matemáticas, IMUEx, Universidad de Extremadura, 06006 Badajoz, Spain; nogales@unex.es

Abstract: Under mild conditions, strong consistency of the Bayes estimator of the density is proved. Moreover, the Bayes risk (for some common loss functions) of the Bayes estimator of the density (i.e., the posterior predictive density) goes to zero as the sample size goes to ∞. In passing, a similar result is obtained for the estimation of the sampling distribution.

Keywords: Bayesian density estimation; Bayesian estimation of the sampling distribution; posterior predictive distribution; consistency of the Bayes estimator

MSC: Primary: 62G07; 62G20; Secondary: 62F15

check for updates

Citation: Nogales, A.G. On Consistency of the Bayes Estimator of the Density. *Mathematics* **2022**, *10*, 636. https://doi.org/10.3390/math10040636

Academic Editors: Francisco German Badía and María D. Berrade

Received: 2 February 2022
Accepted: 17 February 2022
Published: 18 February 2022

1. Introduction

In a statistical context, since the expression *the probability of an event* A (usually denoted $P_\theta(A)$) depends on the unknown parameter, it is really a misuse of language. Before performing the experiment, this expression can be assigned a natural meaning from a Bayesian perspective as the prior predictive probability of A since it is the prior mean of the probabilities $P_\theta(A)$. However, in accordance with Bayesian philosophy, once the experiment has been carried out and the value ω has been observed, a more appropriate estimate of $P_\theta(A)$ is the posterior predictive probability given ω of A. The author has recently proved ([1]) that not only is this the Bayes estimator of $P_\theta(A)$ but that the posterior predictive distribution (resp. the posterior predictive density) is the Bayes estimator of the sampling distribution P_θ (resp. the density p_θ) for the squared variation total (resp. the squared L^1) loss function in the Bayesian experiment corresponding to an n-sized sample of the unknown distribution. It should be noted that the loss functions considered derive in a natural way from the commonly used squared error loss function when estimating a real function of the parameter.

The posterior predictive distribution is the cornerstone of Predictive Inference, which seeks to make inferences about a new unknown observation from a preceding random sample (see [2,3]). With that idea in mind, it has also been used in other areas such as model selection, testing for discordancy, goodness of fit, perturbation analysis, and classification (see additional fields of application in [1–5]). Furthermore, in [1], it has been presented as a solution for the Bayesian density estimation problem, giving several examples to illustrate the results and, in particular, to calculate a posterior predictive density. [3] provide many other examples of determining the posterior predictive distribution. But in practice, explicit evaluation of the posterior predictive distribution may be cumbersome, and its simulation may become preferable. The aforementioned work of [3] also constitutes a good reference for such simulation methods, and hence for the computation of the Bayes estimators of the density and the sampling distribution.

We would refer to the references cited in [1] for other statistical uses of the posterior predictive distribution and some useful ways to calculate it.

In this communication, we shall explore the asymptotic behaviour of the posterior predictive density as the Bayes estimator of the density, showing its strong consistency and that the Bayes risk goes to 0 as n goes to ∞.

2. The Framework

Let

$$(\Omega, \mathcal{A}, \{P_\theta\colon \theta \in (\Theta, \mathcal{T}, Q)\})$$

be a Bayesian experiment (where Q denotes de prior distribution on the parameter space $(\Theta, \mathcal{T})$), and consider the infinite product Bayesian experiment

$$(\Omega^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, \{P_\theta^{\mathbb{N}}\colon \theta \in (\Theta, \mathcal{T}, Q)\})$$

corresponding to an infinite sample of the unknown distribution P_θ. Let us write

$$I(\omega, \theta) := \omega, \quad J(\omega, \theta) := \theta, \quad I_n(\omega, \theta) := \omega_n \quad \text{and} \quad I_{(n)}(\omega) := \omega_{(n)} := (\omega_1, \ldots, \omega_n)$$

for integer n.

We suppose that $P^{\mathbb{N}}(\theta, A) := P_\theta^{\mathbb{N}}(A)$ is a Markov kernel. Let

$$\Pi_{\mathbb{N}} := P^{\mathbb{N}} \otimes Q$$

be the joint distribution of the parameter and the observations, i.e.,

$$\Pi_{\mathbb{N}}(A \times T) = \int_T P_\theta^{\mathbb{N}}(A) dQ(\theta), \quad A \in \mathcal{A}^{\mathbb{N}},\ T \in \mathcal{T}.$$

As $Q := \Pi_{\mathbb{N}}^J$ (i.e., the probability distribution of J with respect to $\Pi_{\mathbb{N}}$), $P_\theta^{\mathbb{N}}$ is a version of the conditional distribution (regular conditional probability) $\Pi_{\mathbb{N}}^{I|J=\theta}$. Analogously, P_θ^n is a version of the conditional distribution $\Pi_{\mathbb{N}}^{I_{(n)}|J=\theta}$.

Let $\beta_{Q,\mathbb{N}}^* := \Pi_{\mathbb{N}}^I$, the prior predictive distribution in $\Omega^{\mathbb{N}}$ (so that $\beta_{Q,\mathbb{N}}^*(A)$ is the prior mean of the probabilities $P_\theta^{\mathbb{N}}(A)$). Similarly, write $\beta_{Q,n}^* := \Pi_{\mathbb{N}}^{I_{(n)}}$ for the prior predictive distribution in Ω^n. So, the posterior distribution $P_{\omega,\mathbb{N}}^* := \Pi_{\mathbb{N}}^{J|I=\omega}$ given $\omega \in \Omega^{\mathbb{N}}$ satisfies

$$\Pi_{\mathbb{N}}(A \times T) = \int_T P_\theta^{\mathbb{N}}(A) dQ(\theta) = \int_A P_{\omega,\mathbb{N}}^*(T) d\beta_{Q,\mathbb{N}}^*(\omega), \quad A \in \mathcal{A}^{\mathbb{N}},\ T \in \mathcal{T}.$$

Denote by $P_{\omega_{(n)},n}^* := \Pi_{\mathbb{N}}^{J|I_{(n)}=\omega_{(n)}}$ for $\omega_{(n)} \in \Omega^n$ the posterior distribution given $\omega_{(n)} \in \Omega^n$.

Write $P_{\omega_{(n)},n}^{*\ P}$ for the posterior predictive distribution given $\omega_{(n)} \in \Omega^n$ defined for $A \in \mathcal{A}$ as

$$P_{\omega_{(n)},n}^{*\ P}(A) = \int_\Theta P_\theta(A) dP_{\omega_{(n)},n}^*(\theta).$$

So $P_{\omega_{(n)},n}^{*\ P}(A)$ is nothing but the posterior mean given $\omega_{(n)} \in \Omega^n$ of the probabilities $P_\theta(A)$.

In the dominated case, we can assume without loss of generality that the dominating measure μ is a probability measure (because of (1) below). We write $p_\theta = dP_\theta/d\mu$. The likelihood function $\mathcal{L}(\omega, \theta) := p_\theta(\omega)$ is assumed to be $\mathcal{A} \times \mathcal{T}$-measurable.

We have that, for all n and every event $A \in \mathcal{A}$,

$$\begin{aligned} P_{\omega_{(n)},n}^{*\ P}(A) &= \int_\Theta P_\theta(A) dP_{\omega_{(n)},n}^*(\theta) = \int_\Theta \int_A p_\theta(\omega') d\mu(\omega') dP_{\omega_{(n)},n}^*(\theta) \\ &= \int_A \int_\Theta p_\theta(\omega') dP_{\omega_{(n)},n}^*(\theta) d\mu(\omega'), \end{aligned}$$

which proves that

$$p_{\omega_{(n)},n}^{*\ P}(\omega') := \int_\Theta p_\theta(\omega') dP_{\omega_{(n)},n}^*(\theta)$$

is a μ-density of $P^{*\,P}_{\omega_{(n)},n}$ that we recognize as the posterior predictive density on Ω given $\omega_{(n)}$.

In the same way,

$$p^{*\,P}_{\omega,\mathbb{N}}(\omega') := \int_\Theta p_\theta(\omega')dP^*_{\omega,\mathbb{N}}(\theta)$$

is a μ-density of $P^{*\,P}_{\omega,\mathbb{N}}$, the posterior predictive density on Ω given $\omega \in \Omega^{\mathbb{N}}$.

In the following, we will assume the following additional regularity conditions:

(i) $(\Omega, \mathcal{A})$ is a standard Borel space;
(ii) Θ is a Borel subset of a Polish space and $\mathcal{T}$ is its Borel σ-field;
(iii) $\{P_\theta : \theta \in \Theta\}$ is identifiable.

According to [1], the posterior predictive distribution $P^{*\,P}_{\omega_{(n)},n}$ (resp. the posterior predictive density $p^{*\,P}_{\omega_{(n)},n}$) is the Bayes estimator of the sampling distribution P_θ (resp. the density p_θ) for the squared variation total (resp. the squared L^1) loss function in the product experiment $(\Omega^n, \mathcal{A}^n, \{P^n_\theta : \theta \in (\Theta, \mathcal{T}, Q)\})$. Analogously, the posterior predictive distribution $P^{*\,P}_{\omega,\mathbb{N}}$ (resp. the posterior predictive density $p^{*\,P}_{\omega,\mathbb{N}}$) is the Bayes estimator of the sampling distribution P_θ (resp. the density p_θ) for the squared variation total (resp. the squared L^1) loss function in the product experiment $(\Omega^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, \{P^{\mathbb{N}}_\theta : \theta \in (\Theta, \mathcal{T}, Q)\})$.

As a particular case of a well known result about the total variation distance between two probability measures and the L^1-distance between their densities, we have that

$$\sup_{A\in\mathcal{A}} \left| P^{*\,P}_{\omega_{(n)},n}(A) - P_\theta(A) \right| = \frac{1}{2} \int_\Omega \left| p^{*\,P}_{\omega_{(n)},n} - p_\theta \right| d\mu. \quad (1)$$

3. The Main Result

We ask whether the Bayes risk of the Bayes estimator $P^{*\,P}_{\omega_{(n)},n}$ of the sampling distribution P_θ goes to zero when $n \to \infty$, i.e., whether

$$\lim_n \int_{\Omega^{\mathbb{N}}\times\Theta} \sup_{A\in\mathcal{A}} \left| P^{*\,P}_{\omega_{(n)},n}(A) - P_\theta(A) \right|^2 d\Pi_{\mathbb{N}}(\omega,\theta) = 0.$$

In terms of densities, the question is whether the Bayes risk of the Bayes estimator $p^{*\,P}_{\omega_{(n)},n}$ of the density p_θ goes to zero when $n \to \infty$, i.e., whether

$$\lim_n \int_{\Omega^{\mathbb{N}}\times\Theta} \left(\int_\Omega \left| p^{*\,P}_{\omega_{(n)},n}(\omega') - p_\theta(\omega') \right| d\mu(\omega') \right)^2 d\Pi_{\mathbb{N}}(\omega,\theta) = 0.$$

Let us consider the auxiliary Bayesian experiment

$$(\Omega \times \Omega^{\mathbb{N}}, \mathcal{A} \times \mathcal{A}^{\mathbb{N}}, \{\mu \times P^{\mathbb{N}}_\theta : \theta \in (\Theta, \mathcal{T}, Q)\}).$$

For $\omega' \in \Omega$, $\omega \in \Omega^n$ and $\theta \in \Theta$, we will continue to write $I(\omega', \omega, \theta) = \omega$ and $J(\omega', \omega, \theta) = \theta$, and now we write $I'(\omega', \omega, \theta) = \omega'$.

The new prior predictive distribution is $\mu \times \beta^*_{Q,n}$ since

$$(\mu \times \Pi_{\mathbb{N}})^{(I', I_{(n)})}(A' \times A_{(n)}) = \mu(A') \cdot \beta^*_{Q,n}(A_{(n)}) = (\mu \times \beta^*_{Q,n})(A' \times A_{(n)}).$$

To compute the new posterior distributions, notice that

$$(\mu \times \Pi_{\mathbb{N}})(A' \times I^{-1}_{(n)}(A_{(n)}) \times T) = \int_{A' \times I^{-1}_{(n)}(A_{(n)})} (\mu \times \Pi_{\mathbb{N}})^{J|(I', I_{(n)}) = (\omega', \omega_{(n)})}(T) d(\mu \times \Pi_{\mathbb{N}})^{(I', I_{(n)})}(\omega', \omega_{(n)}).$$

On the other hand,

$$(\mu \times \Pi_{\mathbb{N}})(A' \times I_{(n)}^{-1}(A_{(n)}) \times T) = \mu(A') \cdot \Pi_{\mathbb{N}}(I_{(n)}^{-1}(A_{(n)}) \times T) =$$
$$\mu(A') \cdot \int_{A_{(n)}} P^*_{\omega_{(n)},n}(T) d\beta^*_{Q,n}(\omega_{(n)}) = \int_{A' \times A_{(n)}} P^*_{\omega_{(n)},n}(T) d(\mu \times \beta^*_{Q,n})(\omega', \omega_{(n)}).$$

So,

$$P^*_{\omega_{(n)},n} = (\mu \times \Pi_{\mathbb{N}})^{J|(I',I_{(n)})=(\omega',\omega_{(n)})}.$$

It follows that if $f \in L^1(Q)$ then

$$E_{P^*_{\omega_{(n)},n}}(f) = E_{\mu\times\Pi_{\mathbb{N}}}[f \mid (I', I_{(n)}) = (\omega', \omega_{(n)})].$$

when $\mathcal{A}'_{(n)} := (I', I_{(n)})^{-1}(\mathcal{A} \times \mathcal{A}^n)$, we have that $(\mathcal{A}'_{(n)})_n$ is an increasing sequence of sub-σ-fields of $\mathcal{A} \times \mathcal{A}^{\mathbb{N}}$ such that $\mathcal{A} \times \mathcal{A}^{\mathbb{N}} = \sigma(\cup_n \mathcal{A}'_{(n)})$. According to the martingale convergence theorem of Lévy, if Y is $(\mathcal{A} \times \mathcal{A}^{\mathbb{N}} \times \mathcal{T})$-measurable and $\mu \times \Pi_{\mathbb{N}}$-integrable then

$$E_{\mu\times\Pi_{\mathbb{N}}}(Y|\mathcal{A}'_{(n)})$$

converges $(\mu \times \Pi_{\mathbb{N}})$-a.e. and in $L^1(\mu \times \Pi_{\mathbb{N}})$ to $Y = E_{\mu\times\Pi_{\mathbb{N}}}(Y|\mathcal{A}' \times \mathcal{A}^{\mathbb{N}})$.

Let us consider the $\mu \times \Pi_{\mathbb{N}}$-integrable function

$$Y(\omega', \omega, \theta) := p_\theta(\omega').$$

We shall see that

$$p^{*\,P}_{\omega,\mathbb{N}}(\omega') = E_{\mu\times\Pi_{\mathbb{N}}}(Y \mid (I', I) = (\omega', \omega)). \tag{2}$$

Indeed, given $A' \in \mathcal{A}$ and $A \in \mathcal{A}^{\mathbb{N}}$, we have that

$$\begin{aligned}
\int_{(I',I)^{-1}(A'\times A)} p_\theta(\omega') d(\mu \times \Pi_{\mathbb{N}})(\omega', \omega, \theta) &= \int_A \int_\Theta \int_{A'} p_\theta(\omega') d\mu(\omega') dP^*_{\omega,\mathbb{N}}(\theta) d\beta^*_{Q,\mathbb{N}}(\omega) \\
&= \int_A \int_\Theta P_\theta(A') dP^*_{\omega,\mathbb{N}}(\theta) d\beta_{Q,\mathbb{N}}(\omega) = \int_A P^{*\,P}_{\omega,\mathbb{N}}(A') d\beta^*_{Q,\mathbb{N}}(\omega) \\
&= \int_{A'} \int_A p^{*\,P}_{\omega,\mathbb{N}}(\omega') d\mu(\omega') d\beta^*_{Q,\mathbb{N}}(\omega) = \int_{A'\times A} p^{*\,P}_{\omega,\mathbb{N}}(\omega') d(\mu \times \Pi_{\mathbb{N}})^{(I',I)}(\omega', \omega),
\end{aligned}$$

which proves (2).

Analogously, it can be shown that

$$p^{*\,P}_{\omega_{(n)},n}(\omega') = E_{\mu\times\Pi_{\mathbb{N}}}(Y \mid (I', I_{(n)}) = (\omega', \omega_{(n)})). \tag{3}$$

Hence, it follows from the aforementioned theorem of Lévy that

$$\lim_n p^{*\,P}_{\omega_{(n)},n}(\omega') = p^{*\,P}_{\omega,\mathbb{N}}(\omega'), \quad (\mu \times \Pi_{\mathbb{N}}) - \text{a.e.} \tag{4}$$

and

$$\lim_n \int_{\Omega\times\Omega^{\mathbb{N}}\times\Theta} \left| p^{*\,P}_{\omega_{(n)},n}(\omega') - p^{*\,P}_{\omega,\mathbb{N}}(\omega') \right| d(\mu \times \Pi_{\mathbb{N}})(\omega', \omega, \theta) = 0,$$

i.e.,

$$\lim_n \int_{\Omega^{\mathbb{N}}\times\Theta} \int_\Omega \left| p^{*\,P}_{\omega_{(n)},n}(\omega') - p^{*\,P}_{\omega,\mathbb{N}}(\omega') \right| d\mu(\omega') d\Pi_{\mathbb{N}}(\omega, \theta) = 0. \tag{5}$$

On the other hand, as a consequence of a known theorem of Doob (see Theorem 6.9 and Proposition 6.10 of [4], pp. 129, 130, we have that, for every $\omega' \in \Omega$,

$$\lim_n \int_\Theta p_{\theta'}(\omega') dP^*_{\omega_{(n)},n}(\theta') = p_\theta(\omega'), \quad P_\theta^{\mathbb{N}} - \text{a.e.}$$

for Q-almost every θ. Hence

$$\lim_n p^{*\ P}_{\omega_{(n)},n}(\omega') = p_\theta(\omega'), \quad P_\theta^{\mathbb{N}} - \text{a.e.}$$

for Q-almost every θ, i.e., given $\omega' \in \Omega$ there exists $T_{\omega'} \in \mathcal{T}$ such that $Q(T_{\omega'}) = 0$ and, $\forall \theta \notin T_{\omega'}$,

$$\lim_n p^{*\ P}_{\omega_{(n)},n}(\omega') = p_\theta(\omega'), \quad P_\theta^{\mathbb{N}} - \text{a.e.}$$

So, for $\theta \notin T_{\omega'}$, there exists $N_{\theta,\omega'} \in \mathcal{A}^{\mathbb{N}}$ such that $P_\theta^{\mathbb{N}}(N_{\theta,\omega'}) = 0$ and

$$\lim_n p^{*\ P}_{\omega_{(n)},n}(\omega') = p_\theta(\omega'), \quad \forall \omega \notin N_{\theta,\omega'}, \forall \theta \notin T_{\omega'}, \forall \omega' \in \Omega.$$

In particular,

$$\lim_n p^{*\ P}_{\omega_{(n)},n}(\omega') = p_\theta(\omega'), \quad \mu \times P_\theta^{\mathbb{N}} - \text{a.e.} \tag{6}$$

From (4) and (6), it follows that $p_\theta(\omega') = p^{*\ P}_{\omega,\mathbb{N}}(\omega')$, $\mu \times P_\theta^{\mathbb{N}} - \text{a.e.}$

From this and (5), it follows that

$$\lim_n \int_{\Omega^{\mathbb{N}} \times \Theta} \int_\Omega \left| p^{*\ P}_{\omega_{(n)},n}(\omega') - p_\theta(\omega') \right| d\mu(\omega') d\Pi_{\mathbb{N}}(\omega, \theta) = 0,$$

i.e., the risk of the Bayes estimator of the density for the L^1 loss function goes to 0 when $n \to \infty$.

It follows from this and (1) that

$$\lim_n \int_{\Omega^{\mathbb{N}} \times \Theta} \sup_{A \in \mathcal{A}} \left| P^{*\ P}_{\omega_{(n)},n}(A) - P_\theta(A) \right| d\Pi_{\mathbb{N}}(\omega, \theta) = 0,$$

i.e., the risk of the Bayes estimator of the sampling distribution P_θ for the variation total loss function goes to 0 when $n \to \infty$.

We ask whether these results remain true for the squared versions of the loss functions. The answer is affirmative because of the following general result: Let (X_n) be a sequence of r.r.v. on a probability space $(\Omega, \mathcal{A}, P)$ such that $\lim_n \int |X_n| dP = 0$. If there exists $a > 0$ such that $|X_n| \le a$, for all n, then $\lim_n \int |X_n|^2 dP = 0$ because

$$0 \le \int |X_n|^2 dP \le a \int |X_n| dP \to_n 0.$$

In our case $a = 2$, $P := \Pi_{\mathbb{N}}$ and

$$X_n := \int_\Omega \left| p^{*\ P}_{\omega_{(n)},n}(\omega') - p^{*\ P}_{\omega,\mathbb{N}}(\omega') \right| d\mu(\omega'), \quad \text{or} \quad X_n := \sup_{A \in \mathcal{A}} \left| P^{*\ P}_{\omega_{(n)},n}(A) - P_\theta(A) \right|.$$

So, we have proved the following result.

Theorem 1. *Let $(\Omega, \mathcal{A}, \{P_\theta \colon \theta \in (\Theta, \mathcal{T}, Q)\})$ be a Bayesian experiment dominated by a σ-finite measure μ. Let us assume that $(\Omega, \mathcal{A})$ is a standard Borel space, and that Θ is a Borel subset of a Polish space and $\mathcal{T}$ is its Borel σ-field. Assume also that the likelihood function $\mathcal{L}(\omega, \theta) := p_\theta(\omega) = \frac{dP_\theta}{d\mu}(\omega)$ is $\mathcal{A} \times \mathcal{T}$-measurable and the family $\{P_\theta \colon \theta \in \Theta\}$ is identifiable. Then:*

(a) *The posterior predictive density $p^*_{\omega_{(n)},n}{}^P$ is the Bayes estimator of the density p_θ in the product experiment $(\Omega^n, \mathcal{A}^n, \{P^n_\theta : \theta \in (\Theta, \mathcal{T}, Q)\})$ for the squared L^1 loss function. Moreover the risk function converges to 0 for both the L^1 loss function and the squared L^1 loss function.*
(b) *The posterior predictive distribution $P^*_{\omega_{(n)},n}{}^P$ is the Bayes estimator of the sampling distribution P_θ in the product experiment $(\Omega^n, \mathcal{A}^n, \{P^n_\theta : \theta \in (\Theta, \mathcal{T}, Q)\})$ for the squared variation total loss function. Moreover the risk function converges to 0 for both the variation total loss function and the squared variation total loss function.*
(c) *The posterior predictive density is a strongly consistent estimator of the density p_θ, i.e.,*

$$\lim_n p^*_{\omega_{(n)},n}{}^P(\omega') = p_\theta(\omega'), \quad \mu \times P^{\mathbb{N}}_\theta - a.e.$$

for Q-almost every $\theta \in \Theta$.

Funding: This research was funded by the Junta de Extremaura (SPAIN) grant number GR21044.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Nogales, A.G. On Bayesian estimation of densities and sampling distributions: The posterior predictive distribution as the Bayes estimator. *Stat. Neerl.* **2021**, accepted. [CrossRef]
2. Geisser, S. *Predictive Inference: An Introduction*; Chapman & Hall: New York, NY, USA, 1993.
3. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; CRC Press (Taylor & Francis Group): Boca Raton, FL, USA, 2014.
4. Ghosal, S.; Vaart, A.V.D. *Fundamentals of Nonparametric Bayesian Inference*; Cambridge University Press: Cambridge, UK, 2017.
5. Rubin, D.B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* **1984**, *12*, 1151–1172. [CrossRef]