On Bayesian estimation of densities and sampling distributions: The posterior predictive distribution as the Bayes estimator

Agustín G. Nogales

Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain

Correspondence
Agustín G. Nogales, Departamento de Matemáticas, Universidad de Extremadura, Avda. de Elvas s/n, 06006 Badajoz, Spain.
Email: nogales@unex.es

Funding information
Consejería de Educación y Empleo, Junta de Extremadura, Grant/Award Number: GR18016

Optimality results for three interesting Bayesian estimation problems are presented in this paper: the estimation of the sampling distribution for the squared total variation function, the estimation of the density for the $L^1$-squared loss function and the estimation of a real distribution function for the $L^\infty$-squared loss function. The posterior predictive distribution provides the solution to these problems. Some examples are presented to illustrate it.

**KEYWORDS**
Bayesian density estimation, Bayesian estimation of the sampling distribution, posterior predictive distribution

**MOS SUBJECT CLASSIFICATION**
Primary 62F15; 62G07 Secondary 62C10

1 | INTRODUCTION AND BASIC DEFINITIONS

In the next pages, the problems of estimation of a density or a probability measure (or even of a distribution function in the real case) are considered under the Bayesian point of view. These problems are addressed in a number of previous references such as Ghosh and Ramamoorthy (2003, chapter 5), Lijoi and Prünster (2010, section 3.4), Lo (1984), Ferguson (1983) or, recently, Marchand and Sadeghkhan (2018), to mention just a few. Popular choices for Bayesian density estimation are Dirichlet-process mixture models, due to their large support and the ease of their implementation (see Bean et al., 2016). Ghosal and Van der Vaart (2017, p. 121) contain a brief historical review on Bayesian density estimation. Theorem 2 below is an optimality result on Bayesian density estimation—and Bayesian estimation of the sampling distribution—that must
be considered as a functional extension of corollary 1.2.(a) in Lehmann and Casella (1998, p. 228) which yields the Bayes estimator of a real function of the parameter for the squared error loss function.

The posterior predictive distribution has been presented as the keystone in Predictive Inference, which seeks to make inferences about a new unknown observation from the previous random sample, in contrast to the greater emphasis that statistical inference makes on the estimation and contrast of parameters since its mathematical foundations in the early 20th century (see Geisser, 1993 or Gelman et al., 2014). 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 or classification (see additional fields of application in Geisser (1993) and Rubin (1984), but never as a possible solution for the Bayesian density estimation problem.

Here, the posterior predictive density appears as the optimal estimator of the density for the $L^1$-squared loss function and this is true whatever be the prior distribution. Besides, the posterior predictive distribution is the optimal estimator of the probability measures $P_\theta$ for the squared total variation loss function, and, in the real case, the posterior predictive distribution function becomes the optimal estimator of the sampling distribution function for $L^\infty$-squared loss function. This way, Theorem 2 is not a simply existing result of an optimal estimator of the density but rather it shows that the optimal estimator is the posterior predictive density. Moreover, the proofs of Theorems 1 and 2 show that the square in the total variation, $L^1$ and $L^\infty$ loss functions come from the quadratic error loss function used in the estimation of a real function of the parameter. In this sense, these loss functions should be considered as natural for their respective estimation problems, as natural as euclidean distance can be. Finally, the results are general enough (in fact, Markov kernel theory has been used to obtain a general presentation of the posterior predictive distribution) to simultaneously cover continuous and discrete, univariate and multivariate, parametric and nonparametric cases.

Several examples are presented in Section 4 to illustrate the results. Gelman et al. (2014) contain many other examples of determination of the posterior predictive distribution. But in practice, the explicit evaluation of the posterior predictive distribution could be cumbersome and its simulation may become preferable. The study of Gelman et al. (2014) is also a good reference for such simulation methods and, hence, for the computation of the Bayes estimators of the density and the sampling distribution.

For ease of reading, proofs have been postponed to Section 5.

In what follows we will place ourselves in a general framework for the Bayesian inference, as is described in Barra (1971). An appendix is included to establish the notations and the basic statistical and probabilistic concepts that will be used in the paper.

**2 | BAYESIAN ESTIMATION OF PROBABILITIES, SAMPLING DISTRIBUTIONS AND DENSITIES**

According to Bayesian philosophy, given $A \in \mathcal{A}$, a natural estimator of $f_A(\theta) := P_\theta(A)$ is the posterior mean of $f_A$, which coincides with the posterior predictive probability of $A$, $T(\omega) := P^\omega P(A)$. In fact, this is the Bayes estimator of $f_A$ (see Theorem 1.(i)).

So, the posterior predictive distribution $P^\omega P$ appears as the natural Bayesian estimator of the probability distribution $P_\theta$. 
To estimate probability measures, the squared total variation loss function

$$W_1(Q, P) := \sup_{A \in \mathcal{A}} |Q(A) - P(A)|^2,$$

will be considered. An estimator of \( f(\theta) := P_\theta \) is a Markov kernel \( M : (\Omega, \mathcal{A}) \to (\Omega, \mathcal{A}) \) so that, being observed \( \omega \in \Omega \), \( M(\omega, \cdot) \) is a probability measure on \( \mathcal{A} \) which is considered as an estimation of \( f \). We wonder if the Bayes mean risk of the estimator \( M^* := P^{*P} \) is less than that of any other estimator \( M \) of \( f \), that is, we wonder if

$$\int_{\Omega \times \Theta} \sup_{A \in \mathcal{A}} |P^*(A) - P_\theta(A)|^2 d\Pi(\omega, \theta) \leq \int_{\Omega \times \Theta} \sup_{A \in \mathcal{A}} |M(\omega, A) - P_\theta(A)|^2 d\Pi(\omega, \theta).$$

Theorem 1.(ii) below gives the answer.

An estimator of the density \( p_\theta \) on \((\Omega, \mathcal{A}, \{P_\theta : \theta \in (\Theta, T, Q)\})\) is a measurable map \( m : (\Omega^2, \mathcal{A}^2) \to \mathbb{R} \) in such a way that, being observed \( \omega \in \Omega \), the map \( \omega' \mapsto m(\omega, \omega') \) is an estimation of \( p_\theta \).

It is well known (see Ghosal & Van der Vaart, 2017, p. 126) that, given two probability measures \( Q \) and \( P \) on \((\Omega, \mathcal{A})\) having densities \( q \) and \( p \) with respect to a \( \sigma \)-finite measure \( \mu \),

$$\sup_{A \in \mathcal{A}} |Q(A) - P(A)| = \frac{1}{2} \int |q - p| d\mu.$$

So the Bayesian estimation of the sampling distribution \( P_\theta \) for the squared total variation loss function corresponds to the Bayesian estimation of its density \( p_\theta \) for the \( L^1 \)-squared loss function

$$W'_1(q, p) := \left( \int |q - p| \, d\mu \right)^2.$$

The next theorem also solves the estimation problem of the density.

**Theorem 1.** Let \((\Omega, \mathcal{A}, \{P_\theta : \theta \in (\Theta, T, Q)\})\) be a Bayesian statistical experiment dominated by a \( \sigma \)-finite measure \( \mu \), where the \( \sigma \)-field \( \mathcal{A} \) is supposed to be separable. We suppose that the likelihood function \( L(\omega, \theta) := P_\theta(\omega) = dP_\theta(\omega)/d\mu \) is \( \mathcal{A} \otimes T \)-measurable.

(i) Given \( A \in \mathcal{A} \), the posterior predictive probability \( P^{P^*}_\omega(A) \) of \( A \) is the Bayes estimator of the probability \( P_\theta(A) \) of \( A \) for the squared error loss function

$$W(x, \theta) := (x - P_\theta(A))^2.$$

Moreover, if \( X \) is a real statistics with finite mean, its posterior predictive mean

$$E_{P^{P^*}_\omega}(X) = \int_{\Theta} \int_{\Omega} X(\omega') dP_\theta(\omega') dP^{*P}_\omega(\theta),$$

is the Bayes estimator of \( E_\theta(X) \).

(ii) The posterior predictive distribution \( P^{P^*}_\omega \) is the Bayes estimator of the sampling distribution \( P_\theta \) for the squared total variation loss function

$$W_1(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|^2.$$
(iii) The posterior predictive density

\[ b_{Q,\omega}^*(\omega') := \frac{dP_\omega^P}{d\mu}(\omega') = \int_\Theta p_\omega(\omega')p_\omega^*(\theta)dQ(\theta). \]

is the Bayes estimator of the density \( p_\theta \) for the \( L^1 \)-squared loss function

\[ W_1'(p, q) := \left( \int_\Omega |p - q|d\mu \right)^2. \]

### 3 Bayesian Estimation of Sampling Distributions and Densities from a Sample

More generally, an estimator of \( f(\theta) := P_\theta \) from a sample of size \( n \) of this distribution is a Markov kernel

\[ M_n : (\Omega^n, A^n) \rightarrow (\Omega, A). \]

Let us consider the Markov kernel

\[ P_n : (\Theta, T) \rightarrow (\Omega^n, A^n), \]

defined by \( P_n(\theta, A) = P^\theta(A), A \in A^n, \theta \in \Theta \). We write \( \Pi_n := P_n \otimes Q \), so that

\[ \Pi_n(A \times T) = \int_T P^\theta_n(A)dQ(\theta), \quad A \in A^n, T \in T. \]

The corresponding prior predictive distribution is

\[ \beta_{Q,n}^*(A) = \int_\Theta P^\theta_n(A)dQ(\theta) = \Pi^I_n(A), \]

where \( I(\omega, \theta) = \omega \) for \( \omega \in \Omega^n \). Let us write \( I_i(\omega) = \omega_i \) and \( \tilde{I}_i(\omega, \theta) = \omega_i \), for \( \omega \in \Omega^n \) and \( i = 1, \ldots, n \). Hence

\[ \left( \beta_{Q,n}^* \right)^I_i(A_i) = \int_\Theta P_\omega(A_i)dQ(\theta) = \beta_{Q,n}^*(A_i), \]

and

\[ \Pi^I_n(A_i \times T) = \int_T P_\omega(A_i)dQ(\theta). \]

so

\[ \left( \beta_{Q,n}^* \right)^I_i = \beta_{Q,n}^* \quad \text{and} \quad \Pi^I_n = \Pi. \]

Denoting \( J(\omega, \theta) = \theta \), the posterior distribution \( P_{\omega,n}^*: = \Pi^I_n|_{I=\omega} \), \( \omega \in \Omega^n \), is defined in such a way that

\[ \Pi_n(A \times T) = \int_A P_{\omega,n}^*(T)d\beta_{Q,n}^*(\omega). \]
The $\mu^n$-density of $P^n_\theta$ is

$$p_{\theta,n}(\omega) := \frac{dP^n_\theta}{d\mu^n}(\omega) = \prod_{i=1}^np_{\theta}(\omega_i) \quad \text{for} \quad \omega = (\omega_1, \ldots, \omega_n) \in \Omega^n.$$ 

The posterior density given $\omega \in \Omega^n$ is of the form

$$p_{\omega,n}^*(\theta) := \frac{dP_{\omega,n}^*}{dQ}(\theta) \propto p_{\theta,n}(\omega).$$

According to Theorem 1.(ii), the Markov kernel

$$(P_n^*)^* : (\Omega^n, \mathcal{A}) \Rightarrow (\Omega^n, \mathcal{A}^n)$$

defined by

$$(P_n^*)^*(\omega, A) := (P_{\omega,n}^*)^n(A) = \int_{\Theta} P^n_\theta(A)dP_{\omega,n}^*(\theta),$$

is the Bayes estimator of the product probability measure $f_n(\theta) := P^n_\theta$. That is to say

$$\int_{\Omega^n \times \Theta \times A \in A^e} \sup |(P_{\omega,n}^*)^n(A) - P^n_\theta(A)|^2d\Pi_n(\omega, \theta) \leq \int_{\Omega^n \times \Theta \times A \in A^e} \sup |M(\omega, A) - P^n_\theta(A)|^2d\Pi_n(\omega, \theta),$$

for every estimator $M : (\Omega^n, \mathcal{A}^n) \Rightarrow (\Omega^n, \mathcal{A}^n)$ of $P^n_\theta$.

The next theorem shows how marginalizing the posterior predictive distribution $(P_{\omega,n}^*)^n$ we can get the Bayes estimator of the sampling probability measure $P_\theta$ or its density.

**Theorem 2** (Bayesian density estimation from a sample of size $n$). Let $(\Omega, \mathcal{A}, \{P_\theta : \theta \in (\Theta, \mathcal{T}, Q)\})$ be a Bayesian statistical experiment dominated by a $\sigma$-finite measure $\mu$, where the $\sigma$-field $\mathcal{A}$ is supposed to be separable. We suppose that the likelihood function $L(\omega, \theta) := p_\theta(\omega) = dP_\theta(\omega)/d\mu$ is $\mathcal{A} \otimes \mathcal{T}$-measurable. Let $n \in \mathbb{N}$. All the estimation problems below are referred to the product Bayesian statistical experiment $(\Omega^n, \mathcal{A}^n, \{P^n_\theta : \theta \in (\Theta, \mathcal{T}, Q)\})$ corresponding to a $n$-sized sample of the observed unknown distribution. Let $I_1(\omega_1, \ldots, \omega_n) := \omega_1$.

(i) Given $A \in \mathcal{A},$

$$P_{\omega,n}^*P(A) = \left[\left(\frac{p_{\omega,n}^*}{P^n_\theta}\right)^n\right]_{I_1}(A)$$

is the Bayes estimator of the probability $P_\theta(A)$ of $A$ for the squared error loss function

$$W(x, \theta) := (x - P_\theta(A))^2.$$

(ii) The distribution

$$\left[\left(\frac{p_{\omega,n}^*}{P^n_\theta}\right)^n\right]_{I_1}$$

of the projection $I_1$ under the posterior predictive probability $(P_{\omega,n}^*)^n$, that coincides with the posterior predictive distribution $P_{\omega,n}^*P$ on $(\Omega, \mathcal{A})$, is the Bayes estimator of the sampling
distribution $P_\theta$ for the squared total variation loss function

$$W_1(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|^2.$$  

(iii) The posterior predictive density

$$b^*_{Q, o, n}(\omega') := \frac{dP^*_{\omega, n}}{d\mu}(\omega') = \int_\Theta p_\theta(\omega') p^*_{o, n}(\theta) dQ(\theta).$$

is the Bayes estimator of the density $p_\theta$ for the $L^1$-squared loss function

$$W_1'(p, q) := \left( \int_\Omega |p - q| d\mu \right)^2.$$

We end this section with a remark that address the problem of estimating a real distribution function.

**Remark.** (Bayesian estimation of a distribution function). When $P_\theta$ is a probability distribution on the line, we may be interested in the estimation of its distribution function $F_\theta(t) := P_\theta(\{ -\infty, t \})$. An estimator of such a distribution function is a map

$$F : (x, t) \in \mathbb{R}^n \times \mathbb{R} \mapsto F(x, t) := M(x, ] - \infty, t],$$

for a Markov kernel $M : (\mathbb{R}^n, \mathcal{R}) \mapsto (\mathbb{R}, \mathcal{R})$, where $\mathcal{R}$ denotes the Borel $\sigma$-field on $\mathbb{R}$.

According to the previous results, given $t \in \mathbb{R}$,

$$F^*_x(t) := \left[ \left( P^*_{x, n} \right)^{\text{int}} \right]_t (\{ -\infty, t]) = \int_{-\infty}^t \int_{\Theta} p_\theta(y) \cdot p^*_{x, n}(\theta) dQ(\theta) d\mu^n(y),$$

is the Bayes estimator of $F_\theta(t)$ for the squared error loss function. So

$$\int_{\mathbb{R}^n \times \Theta} |F^*_x(t) - F_\theta(t)|^2 d\Pi(x, \theta) \leq \int_{\mathbb{R}^n \times \Theta} |F(x, t) - F_\theta(t)|^2 d\Pi(x, \theta),$$

for any other estimator $F$ of $F_\theta$. Since

$$\sup_{t \in \mathbb{R}} |F(x, t) - F_\theta(t)| = \sup_{r \in \mathbb{Q}} |F(x, r) - F_\theta(r)|,$$

we have that, given $(x, \theta) \in \mathbb{R}^n \times \Theta$ and $k \in \mathbb{N}$, there exists $r_k \in \mathbb{Q}$ such that

$$C(x, \theta) - \frac{1}{k} \leq |F^*_x(r_k) - F_\theta(r_k)|,$$

where $C(x, \theta) := \sup_{t \in \mathbb{R}} |F^*_x(t) - F_\theta(t)|^2$, and hence (see Remark 3 at the end of Section 6)

$$\int_{\mathbb{R}^n \times \Theta} C(x, \theta) d\Pi(x, \theta) \leq \int_{\mathbb{R}^n \times \Theta} |F^*_x(r_k) - F_\theta(r_k)|^2 d\Pi(x, \theta) + \frac{1}{k}$$

$$\leq \int_{\mathbb{R}^n \times \Theta} \sup_{t \in \mathbb{R}} |F(x, t) - F_\theta(t)|^2 d\Pi(x, \theta) + \frac{1}{k}. $$
We have proved that the posterior predictive distribution function \( F^*_x \) is the Bayes estimator of the distribution function \( F_\theta \) for the \( L^2 \)-squared loss function

\[
W''(F, G) = \left( \sup_{t \in \mathbb{R}}|F(t) - G(t)| \right)^2.
\]

4 | EXAMPLES

Example 1. Let \( P_\theta \) the normal distribution \( N(\theta, \sigma^2_\theta) \) with unknown mean \( \theta \in \mathbb{R} \) and known variance \( \sigma^2_\theta \). Let \( Q := N(\mu, \tau^2) \) be the prior distribution where the mean \( \mu \) and variance \( \tau^2 \) are known constants. It is well known that the posterior distribution is

\[
P^*_x = N(m_n(x), s^2_n)
\]

where

\[
m_n(x) = \frac{n\tau^2 x + \sigma^2_\theta}{n\tau^2 + \sigma^2_\theta} \quad \text{and} \quad s^2_n = \frac{\tau^2 \sigma^2_\theta}{n\tau^2 + \sigma^2_\theta}.
\]

It can be shown that the distribution of \( I_1 \) with respect to the posterior predictive distribution is

\[
\left( (P^*_x)^{\text{post}} \right) = N(m_n(x), \sigma^2_\theta + s^2_n).
\]

For the details, the reader is addressed to Boldstat (2004, p. 185), where the distribution of \( I_1 \) with respect to the posterior predictive distribution is referred to as the predictive distribution for the next observation given the observation \( x \).

So \( M^*_x(x, \cdot) := N(m_n(x), \sigma^2_\theta + s^2_n) \) is the Bayes estimator of the sampling distribution \( N(\theta, \sigma^2_\theta) \) for the squared total variation loss function and the density of \( N(m_n(x), \sigma^2_\theta + s^2_n) \) is the Bayes estimator of the density of \( N(\theta, \sigma^2_\theta) \) for the \( L^1 \)-squared loss function.

Example 2. Let \( G(\alpha, \beta) \) be the distribution gamma with parameters \( \alpha, \beta > 0 \) and \( P_\theta := G(1, \theta^{-1}) \), whose density is \( p_\theta(x) = \theta \exp\{-\theta x\} \) for \( x > 0 \).

So \( P^*_0 \) is the joint distribution of a sample of size \( n \) of an exponential distribution of parameter \( 1/\theta \) and its density is \( p_{\theta,n}(x) = \theta^n \exp\{-\theta \sum_i x_i\} \) for \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n_+ \).

Consider the prior distribution \( Q := G(1, \lambda^{-1}) \) for some known \( \lambda > 0 \).

Since, for \( a > 0 \),

\[
\int_0^\infty \theta^n \exp\{-a\theta\} d\theta = \frac{n!}{a^{n+1}}.
\]

we have that the posterior density given \( x \in \mathbb{R}^n_+ \) is

\[
p^*_n(x) = \frac{\lambda + \sum_i x_i}{n!} \theta^n \exp\{-\theta(\lambda + \sum_i x_i)\}.
\]

So, denoting by \( \mu_n \) the Lebesgue measure on \( \mathbb{R}^n_+ \), the density of the posterior predictive probability given \( x \) is

\[
d(P^*_n)^{\text{post}}(x') = \int_\Theta p_{\theta,n}(x') \cdot p^*_n(\theta) d\theta = \frac{(2n)!}{n!} \left( \frac{\lambda + \sum_i x_i}{\lambda + \sum_i x_i'} \right)^{n+1}.
\]
According to the previous results, this is the Bayes estimator of the joint density $p_{\theta,n}$ for the loss function

$$W'_n(q, p) := \left( \int_{\mathbb{R}^n} |q - p| d\mu_n \right)^2,$$

while the posterior predictive distribution $(P^*_{x,n})^n$ is the Bayes estimator of the sampling distribution $P^n_{\theta}$ for the squared total variation loss function on $(\Omega^n, A^n)$.

Moreover, the image $M^n_{k}(x, \cdot) := \left( P^*_{x,n} \right)^n = \int_{\mathbb{R}^n} \frac{dM^n_{x}(x', \cdot)}{d\mu_1}(x') = I_1(P^*_{x,n})^n = P^n_{x,n}$ is the Bayes estimator of the probability distribution $P^n_{\theta}$ for the squared total variation on $(\Omega, A)$ and its density

$$x' > 0 \mapsto \frac{dM^n_{x}(x', \cdot)}{d\mu_1}(x') = \int_0^\infty p_\theta(x') \cdot p^n_{x,n}(\theta) d\theta = \frac{(n + 1)(\lambda + \sum_{i=1}^n x_i)^{n+1}}{(\lambda + x' + \sum_{i=1}^n x_i)^{n+2}},$$

is the Bayes estimator of the density $p_\theta$ for the $L^1$-squared loss function $W'_n$.

Example 3. Let $P^n_{\theta}$ be the Poisson distribution with parameter $\theta > 0$ whose probability function (or density with respect to the counter measure $\mu_1$ on $\mathbb{N}_0$) is $p_\theta(k) = \exp(-\theta) \frac{\theta^k}{k!}$ for $k \in \mathbb{N}_0$.

So $P^n_{\theta}$ is the joint distribution of a sample of size $n$ of a Poisson distribution of parameter $\theta$ and its probability function (or density with respect to the counter measure $\mu_n$ on $\mathbb{N}_0^n$) is $p_{\theta,n}(k) = \exp(-n\theta) \frac{\theta^k}{\prod_{i=1}^n (k_i)!}$ for $k = (k_1, \ldots, k_n) \in \mathbb{N}_0^n$, where $|k|_1 := \sum_{i=1}^n k_i$.

Consider the prior distribution $Q := G(1, \lambda^{-1})$ for some known $\lambda > 0$.

It is readily shown that the posterior distribution given $k \in \mathbb{N}_0^n$ is the gamma distribution $G\left(|k|_1 + 1, \frac{1}{\lambda + n}\right)$ whose density is

$$p_{k,n}^*(\theta) = \frac{(\lambda + n)^{|k|_1+1}}{|k|_1!} \cdot \theta^{|k|_1} \cdot \exp\{-\theta(\lambda + n)\}.$$ 

So the probability function of the posterior predictive probability given $k \in \mathbb{N}_0^n$ is

$$\frac{d(P^n_{k,n})^n}{d\mu_n}(k') = \int_{\Theta} p_{\theta,n}(k') \cdot p_{k,n}^*(\theta) d\theta = \frac{(|k'|_1 + |k|_1)!}{\prod_{i=1}^n (k_i)! \cdot (|k|_1)!} \cdot \frac{(\lambda + n)^{|k'|_1+1}}{(\lambda + 2n)^{|k'|_1+|k|_1+1}}.$$

According to the previous results, this is the Bayes estimator of the joint density $p_{\theta,n}$ for the loss function

$$W'_n(q, p) := \left( \int_{\mathbb{N}_0^n} |q - p| d\mu_n \right)^2,$$

while the posterior predictive distribution $(P^n_{k,n})^n$ is the Bayes estimator of the sampling distribution $P^n_{\theta}$ for the squared total variation loss function on $\mathbb{N}_0^n$.

Moreover, the image $M^n_{k}(k', \cdot) := \left( P^n_{k,n} \right)^n = \int_{\mathbb{R}^n} \frac{dM^n_{k}(k', \cdot)}{d\mu_1}(k') = \int_0^\infty p_\theta(k') \cdot p_{k,n}^*(\theta) d\theta = \frac{(k' + |k|_1)!}{k'! \cdot (|k|_1)!} \cdot \frac{(\lambda + n)^{|k|_1+1}}{(\lambda + n + 1)^{k' + |k|_1+1}}.$
is the Bayes estimator of the probability function \( p_\theta \) for the loss function \( W'_1 \).

**Example 4.** Let \( P_\theta \) be the Bernoulli distribution with parameter \( \theta \in (0, 1) \) whose probability function is \( p_\theta(k) := \theta^k(1-\theta)^{n-k}, k = 0, 1. \) So \( P^n_\theta \) is the joint distribution of a sample of size \( n \) of a Bernoulli distribution with parameter \( \theta \) and its probability function is

\[
p_{\theta,n}(k) = \theta^{||k||} (1-\theta)^{n-||k||}, \quad k \in \{0, 1\}^n
\]

where \( ||k|| := \sum_{i=1}^k k_i. \) Consider the uniform distribution on the unit interval as prior distribution. So, the posterior distribution given \( k \in \{0, 1\}^n \) is the Beta distribution

\[
P^n_{k,n} = B(||k|| + 1, n - ||k|| + 1)
\]

with parameters \( ||k|| + 1 \) and \( n - ||k|| + 1 \). Hence, denoting \( \mu_n \) for the counter measure on \( \{0, 1\}^n \) and \( \beta \) the Euler beta function, the probability function of the posterior predictive probability given \( k \in \{0, 1\}^n \) is

\[
\frac{d(P^n_{k,n})}{d\mu_n}(k') = \int_\Theta p_{\theta,n}(k') \cdot p^n_{k,n}(\theta) \, d\theta
\]

\[
= \frac{\beta(||k|| + ||k'|| + 1, 2n - ||k|| - ||k'|| + 1)}{\beta(||k|| + 1, n - ||k|| + 1)}
\]

\[
= \frac{\Gamma(n+2)}{\Gamma(2n+2)} \cdot (||k||+1)!(2n-||k'||+1)!
\]

This is the Bayes estimator of the joint probability function \( p_{\theta,n} \) for the loss function \( W'_n(q,p) := \left( \int_{\{0,1\}^n} |q-p|d\mu_n \right)^2 \), while the posterior predictive distribution \( P^n_{k,n} \) is the Bayes estimator of the sampling distribution \( P^n_\theta \) for the squared total variation loss function on \( \{0, 1\}^n \).

Moreover, the image \( M^n_k(k, \cdot) := \left[ (P^n_{k,n}) \right]_{k=0}^1 = I_1(P^n_{k,n}) \) is the Bayes estimator of the probability distribution \( P_\theta \) for the squared total variation on \( \{0, 1\} \) and its probability function

\[
k' \in \{0, 1\} \quad \mapsto \frac{dM^n_k(k, \cdot)}{d\mu_1}(k') = \int_0^1 p_\theta(k') \cdot \delta^n_{k,n}(\theta) \, d\theta
\]

\[
= \frac{\Gamma(n+2)}{\Gamma(2n+2)} \cdot (k'+||k||)!(2n-k'-||k||)!
\]

is the Bayes estimator of the probability function \( p_\theta \) for the \( L^1 \)-squared loss function \( W'_1 \).

\[ \square \]

### 5 Proofs

**Proof of Theorem 1.** (i) Notice that, writing \( f_A(\theta) := P_\theta(A) \),

\[
P^n_{\omega P}(A) = \int_\Theta P_\theta(A) dP^n_\theta(\theta) = E_{P^n_{\omega}}(f_A),
\]

that, as a consequence of Jensen’s inequality (see Lehmann & Casella, 1998, p. 228), is the Bayes estimator of \( f_A \) for the quadratic error loss function.
In the same way, if $X$ is a real integrable statistic on $(\Omega, \mathcal{A})$ and $f(\theta) := E_\theta(X)$, we have that

$$E_{P_\omega}^* (X) = \int_{\Theta} \int_{\Omega} X(\omega')dP_\theta(\omega')dP_\omega(\theta) = E_{P_\omega}^* (f),$$

is the Bayes estimator of $f$, the mean of $X$.

(ii) According to (i), given $A \in \mathcal{A}$,

$$\int_{\Omega \times \Theta} \left| P_{\omega}^* P(\omega) - P(\theta) \right|^2 d\Pi(\omega, \theta) \leq \int_{\Omega \times \Theta} \left| X(\omega) - P(\theta)(A) \right|^2 d\Pi(\omega, \theta),$$

for any real measurable function $X$ on $(\Omega, \mathcal{A})$. If $\mathcal{A}$ is a separable $\sigma$-field, there exists a countable algebra $\mathcal{A}_0$ such that $\mathcal{A} = \sigma(\mathcal{A}_0)$. In particular, it follows that

$$\sup_{A \in \mathcal{A}} |M(\omega, A) - P(\theta)(A)|^2 = \sup_{A \in \mathcal{A}_0} |M(\omega, A) - P(\theta)(A)|^2$$

is $(\mathcal{A} \otimes \mathcal{T})$-measurable. Given $(\omega, \theta) \in \Omega \times \Theta$, let

$$C(\omega, \theta) := \sup_{A \in \mathcal{A}} \left| P_{\omega}^* P(\omega) - P(\theta)(A) \right|^2,$$

and, given $n \in \mathbb{N}$, choose $A_n \in \mathcal{A}_0$ so that

$$C - \frac{1}{n} \leq \left| P_{\omega}^* P(A_n) - P(\theta)(A_n) \right|^2.$$

It follows from this that

$$\int_{\Omega \times \Theta} Cd\Pi \leq \int_{\Omega \times \Theta} \left| P_{\omega}^* P(A_n) - P(\theta)(A_n) \right|^2 d\Pi(\omega, \theta) + \frac{1}{n} \leq \int_{\Omega \times \Theta} \sup_{A \in \mathcal{A}} |M(\omega, A) - P(\theta)(A)|^2 d\Pi(\omega, \theta) + \frac{1}{n},$$

and this gives the proof as $n$ is arbitrary. To refine the proof from a measure-theoretical point of view, a judicious use of the Ryll–Nardzewski and Kuratowski measurable selection theorem would also be helpful. See the details in Remark at the end of the section.

(iii) It follows from (ii) that, to estimate the density $p_\theta$, the posterior predictive density

$$b_{Q,\omega}(\omega') := \frac{dP_{\omega}^* P}{d\mu}(\omega')$$

minimizes the Bayes mean risk for the loss function

$$W'_1(q, p) := \left( \int |q - p| \, d\mu \right)^2,$$

that is,

$$E_\Pi \left[ \left( \int |b_{Q,\omega}^* - p_\theta| \, d\mu \right)^2 \right] \leq E_\Pi \left[ \left( \int |m(\omega, \cdot) - p_\theta| \, d\mu \right)^2 \right]$$

for any measurable function $m : \Omega \times \Omega \to [0, \infty)$ such that $\int_{\Omega} m(\omega, \omega')d\mu(\omega') = 1$ for every $\omega$. □
Proof of Theorem 2. (i) Given $A \in \mathcal{A}^n$, Theorem 1.(i) shows that the posterior predictive probability $(P_{o,n}^*)^\pi_n(A)$ of $A$ is the Bayes estimator of $f_A(\theta) := P_{o}^n(A)$ in the product Bayesian statistical experiment, as

$$(P_{o,n}^*)^\pi_n(A) = \int_\Theta P_{\theta}^n(A)dP_{o,n}^*(\theta) = E_{P_{o,n}^*}(f_A),$$

that is,

$$\int_{\Omega^* \times \Theta} \left| (P_{o,n}^*)^\pi_n(A) - P_{\theta}^n(A) \right|^2 d\Pi_n(\omega, \theta) \leq \int_{\Omega^* \times \Theta} \left| X(\omega) - P_{\theta}^n(A) \right|^2 d\Pi_n(\omega, \theta),$$

for any other estimator $X : (\Omega^n, \mathcal{A}^n) \rightarrow \mathbb{R}$ of $f_A$. In particular, given $A \in \mathcal{A}$, applying this result to $I^{-1}_1(A) = A \times \Omega^{n-1} \in \mathcal{A}^n$, we obtain that

$$\int_{\Omega^* \times \Theta} \left| (P_{o,n}^*)^\pi_n(I^{-1}_1(A)) - P_{\theta}(A) \right|^2 d\Pi_n(\omega, \theta) \leq \int_{\Omega^* \times \Theta} \left| X(\omega) - P_{\theta}(A) \right|^2 d\Pi_n(\omega, \theta)$$

for any other estimator $X : (\Omega^n, \mathcal{A}^n) \rightarrow \mathbb{R}$ of $g_A := P_{\theta}(A)$.

(ii) Being $\mathcal{A}$ a separable $\sigma$-field, there exists a countable algebra $\mathcal{A}_0$ such that $\mathcal{A} = \sigma(\mathcal{A}_0)$. In particular, it follows that

$$\sup_{A \in \mathcal{A}} |M(\omega, A) - P_{\theta}(A)|^2 = \sup_{A \in \mathcal{A}_0} |M(\omega, A) - P_{\theta}(A)|^2,$$

is $(\mathcal{A} \otimes \mathcal{T})$-measurable. Given $(\omega, \theta) \in \Omega^* \times \Theta$, let

$$C_n(\omega, \theta) := \sup_{A \in \mathcal{A}} \left| (P_{o,n}^*)^\pi_n(I^{-1}_1(A)) - P_{\theta}(A) \right|^2,$$

and given $k \in \mathbb{N}$, choose $A_k \in \mathcal{A}_0$ so that

$$C_n - \frac{1}{k} \leq \left| (P_{o,n}^*)^\pi_n(I^{-1}_1(A_k)) - P_{\theta}(A_k) \right|^2.$$

It follows that

$$\int_{\Omega^* \times \Theta} C_n d\Pi_n \leq \int_{\Omega^* \times \Theta} \left| (P_{o,n}^*)^\pi_n(I^{-1}_1(A_k)) - P_{\theta}(A_k) \right|^2 d\Pi_n(\omega, \theta) + \frac{1}{k}\leq \int_{\Omega^* \times \Theta} \sup_{A \in \mathcal{A}} |M(\omega, A) - P_{\theta}(A)|^2 d\Pi_n(\omega, \theta) + \frac{1}{k},$$

for any Markov kernel $M : (\Omega^n, \mathcal{A}^n) \rightarrow (\Omega, \mathcal{A})$ and, being $k$ arbitrary, this proves that

$$M_n^*(\omega, A) := (P_{o,n}^*)^\pi_n(I^{-1}_1(A)),$$

is the Bayes estimator of $f(\theta) := P_{\theta}$ for the squared total variation loss function in the Bayesian statistical experiment

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

corresponding to a $n$-sized sample of the observed distribution. See Remark below.
(iii) Note that, given $A \in \mathcal{A}$, Fubini’s theorem yields

$$(P^*_{\omega,n})^p(I_1^{-1}(A)) = \int_{\Theta} P_\theta(A) dP^*_{\omega,n}(\theta) = \int_A \int_{\Theta} p_\theta(\omega') \cdot P^*_{\omega,n}(\theta) dQ(\theta) d\mu(\omega'),$$

where $p^*_{\omega,n}$ denotes the posterior density with respect to the prior distribution $Q$. Hence, for $\omega \in \Omega^n$, the $\mu$-density of $M^*_n(\omega, \cdot)$ is

$$\frac{dM^*_n(\omega, \cdot)}{d\mu}(\omega') = \int_{\Theta} p_\theta(\omega') \cdot P^*_{\omega,n}(\theta) dQ(\theta),$$

and this is the Bayes estimator of the sampling density $p_\theta$ for the loss function $W'$. □

Remark. (A precision on measure-theoretical technicalities in the proofs of the previous results). We detail the proof of Theorem 1.(ii), being that of Theorem 2.(ii) (and even that of the last remark of Section 3) similar. It follows from Theorem 1.(i) that, given $(\omega, \theta) \in \Omega \times \Theta$, and writing

$$C(\omega, \theta) := \sup_{A \in \mathcal{A}} \left| P^*_{\omega}(A) - P_\theta(A) \right|^2,$$

we have that, given $n \in \mathbb{N}$, there exists $A_n(\omega, \theta) \in \mathcal{A}_0$ so that

$$C(\omega, \theta) - \frac{1}{n} \leq \left| P^*_{\omega}(A_n(\omega, \theta)) - P_\theta(A_n(\omega, \theta)) \right|^2.$$

To continue the proof we will use the Ryll-Nardzewski and Kuratowski measurable selection theorem as appears in Bogachev (2007, p. 36). With the notations of this book, we make $(T, \mathcal{M}) = (\Omega \times \Theta, \mathcal{A} \otimes \mathcal{T})$ and $X = \mathcal{A}_0$ (the countable field generating $\mathcal{A}$). Given $n \in \mathbb{N}$, let us consider the map $S_n : \Omega \times \Theta \to \mathcal{P}(X)$ defined by

$$S_n(\omega, \theta) = \left\{ A \in \mathcal{A}_0 : C(\omega, \theta) - \frac{1}{n} \leq \left| P^*_{\omega}(A) - P_\theta(A) \right|^2 \right\}.$$

We have that $\emptyset \neq S_n(\omega, \theta) \subset X$ and $S_n(\omega, \theta)$ is closed for the discrete topology on $\mathcal{A}_0$. Moreover, given an open set $U \subset \mathcal{A}_0$,

$$\left\{ (\omega, \theta) : S_n(\omega, \theta) \cap U \neq \emptyset \right\} \in \mathcal{A} \otimes \mathcal{T},$$

because, given $A \in \mathcal{A}_0$,

$$\left\{ (\omega, \theta) : S_n(\omega, \theta) \ni A \right\} = \left\{ (\omega, \theta) : C(\omega, \theta) - \left| P^*_{\omega}(A) - P_\theta(A) \right|^2 \leq \frac{1}{n} \right\} \in \mathcal{A} \otimes \mathcal{T}.$$

So, according to the measurable selection theorem cited above, there exists a measurable map $s_n : (\Omega \times \Theta, \mathcal{A} \otimes \mathcal{T}) \to (\mathcal{A}_0, \mathcal{P}(\mathcal{A}_0))$ such that $s_n(\omega, \theta) \in S_n(\omega, \theta)$ for every $(\omega, \theta)$, or, which is the same,

$$C(\omega, \theta) - \frac{1}{n} \leq \left| P^*_{\omega}(s_n(\omega, \theta)) - P_\theta(s_n(\omega, \theta)) \right|^2.$$

It follows that
\[
\int_{\Omega \times \Theta} C(\omega, \theta) d\Pi(\omega, \theta) \leq \int_{\Omega \times \Theta} \left| P^P_\omega(\omega, s_n(s_n(\omega, \theta))) - P_\theta(\omega, \theta) \right|^2 d\Pi(\omega, \theta) + \frac{1}{n}
\]

which gives the proof as \(n\) is arbitrary.

\[\square\]

**ACKNOWLEDGEMENTS**

This paper has been supported by the Junta de Extremadura (Spain) under the grant Gr18016.

**ORCID**

Agustín G. Nogales ORCID: https://orcid.org/0000-0002-7201-7608

**REFERENCES**

Barra, J. R. (1971). *Notions fondamentales de statistique mathématique*. Paris: Dunod.

Bean, A., Xu, X., & MacEachern, S. (2016). Transformations and Bayesian density estimation. *Electronic Journal of Statistics, 10*, 3355–3373.

Bogachev, V. I. (2007). *Measure theory* (Vol. II). Berlin, Germany: Springer.

Boldstat, W. M. (2004). *Introduction to Bayesian statistics*. Hoboken, NJ: Wiley.

Ferguson, T. S. (1983). *Bayesian density estimation by mixtures of normal distributions*. In M. Haseeb Rizvi, J. S. Rustagi & D. Siegmund (Eds.), *Recent advances in statistics* (pp. 287–302). New York, NY: Academic Press.

Florens, J. P., Mouchart, M., & Rolin, J. M. (1990). *Elements of Bayesian statistics*. New York, NY: Marcel Dekker.

Geisser, S. (1993). *Predictive inference: An introduction*. Dordrecht, Netherlands: Springer Science+ Business Media.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Ghosal, S., & Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge, UK: Cambridge University Press.

Ghosh, J. K., & Ramamoorthy, R. V. (2003). *Bayesian nonparametrics*. New York, NY: Springer.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York, NY: Springer.

Lijoi, A., & Prünster, I. (2010). *Models beyond the Dirichlet process*. In N. L. Hjort, C. Holmes, P. Müller, & S. G. Walker (Eds.), *Bayesian nonparametrics*. Cambridge, UK: Cambridge Series in Statistical and Probabilistic Mathematics.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *The Annals of Statistics, 12*(1), 351–357.

Marchand, É., & Sadeghkhani, A. (2018). Predictive density estimation with additional information. *Electronic Journal of Statistics, 12*, 4209–4238.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statisticians. *The Annals of Statistics, 12*(4), 1151–1172.

---

**How to cite this article:** Nogales, A. G. (2022). On Bayesian estimation of densities and sampling distributions: The posterior predictive distribution as the Bayes estimator. *Statistica Neerlandica, 76*(2), 236–250. https://doi.org/10.1111/stan.12258

**APPENDIX**

Let us briefly recall some basic concepts about Markov kernels, mainly to fix the notations. In the next, \((\Omega, \mathcal{A}), (\Omega_1, \mathcal{A}_1)\) and so on will denote measurable spaces.
Definition 1. (1) (Markov kernel) A Markov kernel \( M_1 : (\Omega, A) \to (\Omega_1, A_1) \) is a map \( M_1 : \Omega \times A_1 \to [0, 1] \) such that: (i) \( \forall \omega \in \Omega, M_1(\omega, \cdot) \) is a probability measure on \( A_1 \); (ii) \( \forall A_1 \in A_1, M_1(\cdot, A_1) \) is \( A \)-measurable.

(2) (Image of a Markov kernel) The image (or probability distribution) of a Markov kernel \( M_1 : (\Omega, A, P) \to (\Omega_1, A_1) \) on a probability space is the probability measure \( P^{M_1} \) on \( A_1 \) defined by \( P^{M_1}(A_1) := \int_{\Omega} M_1(\omega, A_1) \, dP(\omega) \).

(3) (Composition of Markov kernels) Given two Markov kernels \( M_1 : (\Omega_1, A_1) \to (\Omega_2, A_2) \) and \( M_2 : (\Omega_2, A_2) \to (\Omega_3, A_3) \), its composition is defined as the Markov kernel \( M_2M_1 : (\Omega_1, A_1) \to (\Omega_3, A_3) \) given by

\[
M_2M_1(\omega_1, A_3) = \int_{\Omega_2} M_2(\omega_2, A_3)M_1(\omega_1, d\omega_2).
\]

Remarks (1) (Markov kernels as extensions of the concept of random variable) The concept of Markov kernel extends the concept of random variable (or measurable map). A random variable \( T_1 : (\Omega, A, P) \to (\Omega_1, A_1) \) will be identified with the Markov kernel \( M_{T_1} : (\Omega, A, P) \to (\Omega_1, A_1) \) defined by \( M_{T_1}(\omega, A_1) = \delta_{T_1(\omega)}(A_1) = I_{A_1}(T_1(\omega)) \), where \( \delta_{T_1(\omega)} \) denotes the Dirac measure—the degenerate distribution—at the point \( T_1(\omega) \), and \( I_{A_1} \) is the indicator function of the event \( A_1 \). In particular, the probability distribution \( P^{M_{T_1}} \) of \( M_{T_1} \) coincides with the probability distribution \( P^{T_1} \) of \( T_1 \) defined as \( P^{T_1}(A_1) := P(T_1 \in A_1) \).

(2) Given a Markov kernel \( M_1 : (\Omega_1, A_1) \to (\Omega_2, A_2) \) and a random variable \( X_2 : (\Omega_2, A_2) \to (\Omega_3, A_3) \), we have that \( M_{X_2}M_1(\omega_1, A_3) = M_1(\omega_1, X_2^{-1}(A_3)) = M_1(\omega_1, \cdot)^{X_2}(A_3) \). We write \( X_2M_1 := M_{X_2}M_1 \).

Let \( (\Omega, A, \{ P_\theta : \theta \in (\Theta, T, Q) \}) \) be a Bayesian statistical experiment where \( Q \) is the prior distribution, a probability measure on the measurable space \((\Theta, T)\). \( (\Omega, A) \) is the sample space and \((\Theta, T)\) is the parameter space.

When needed, we shall suppose that \( P_\theta \) has a density (or Radon-Nikodym derivative) \( p_\theta \) with respect to a \( \sigma \)-finite measure \( \mu \) on \( A \) and that the likelihood function \( L : (\omega, \theta) \in (\Omega \times \Theta, A \otimes T) \to L(\omega, \theta) := p_\theta(\omega) \) is measurable. So we have a Markov kernel \( P : (\Theta, T) \to (\Omega, A) \) defined by \( P(\theta, A) := P_\theta(A) \). Let \( P^* : (\Omega, A) \to (\Theta, T) \) the Markov kernel determined by the posterior distributions. In fact, if we denote by \( \Pi \) the only probability measure on \( A \otimes T \) such that

\[
\Pi(A \times T) = \int_T P_\theta(A) dQ(\theta), \quad A \in A, T \in T, \quad (A1)
\]

then \( P^* \) is defined in such a way that

\[
\Pi(A \times T) = \int_A P^*_\omega(T) d\beta^*_Q(\omega), \quad A \in A, T \in T, \quad (A2)
\]

where \( \beta^*_Q \) denotes the so called prior predictive probability, defined by

\[
\beta^*_Q(\omega, A) = \int_\Theta P_\theta(A) dQ(\theta), \quad A \in A.
\]

In other terms, \( \beta^*_Q = Q^P \), the probability distribution of the Markov kernel \( P \) with respect to the prior distribution \( Q \).

The probability measure \( \Pi \) integrates all the basic ingredients of the Bayesian model, and these ingredients can be essentially derived from \( \Pi \), something that would allow us to identify the
Bayesian model as the probability space \((\Omega \times \Theta, \mathcal{A} \otimes \mathcal{T}, \Pi)\) (so is done, for instance, in Florens et al. (1990)).

It is well known that, for \(\omega \in \Omega\), the posterior density with respect to the prior distribution is proportional to the likelihood. Namely

\[
p^\omega_\omega(\theta) := \frac{dP^\omega_\omega}{dQ}(\theta) = C(\omega)p_\theta(\omega),
\]

where \(C(\omega) = \left[\int_\Theta p_\theta(\omega)dQ(\theta)\right]^{-1}\).

This way we obtain a statistical experiment \((\Theta, \mathcal{I}, \{P^\omega_\omega : \omega \in \Omega\})\) on the parameter space \((\Theta, \mathcal{I})\). We can reconsider the Markov kernel \(P\) defined on this statistical experiment

\[
P : (\Theta, \mathcal{I}, \{P^\omega_\omega : \omega \in \Omega\}) \succ (\Omega, A).
\]

Since \(P^\omega_\omega P(A) = \int_\Theta p_\theta(A)dP^\omega_\omega(\theta)\), for \(A \in A\), it is called the posterior predictive distribution on \(A\) given \(\omega\), and the statistical experiment image of \(P\) is

\[
(\Omega, A, \{P^\omega_\omega : \omega \in \Omega\}).
\]

Note that, given \(\omega \in \Omega\), according to Fubini’s Theorem,

\[
P^\omega_\omega P(A) = \int_\Theta p_\theta(A)dP^\omega_\omega(\theta) = \int_\Theta \int_A p_\theta(\omega')d\mu(\omega')p^\omega_\omega(\theta)dQ(\theta)
= \int_A \int_\Theta p_\theta(\omega')p^\omega_\omega(\theta)dQ(\theta)d\mu(\omega').
\]

So, the posterior predictive density is

\[
\frac{dP^\omega_\omega P}{d\mu}(\omega') = \int_\Theta p_\theta(\omega')p^\omega_\omega(\theta)dQ(\theta).
\]

If we consider the composition of the Markov kernels \(P^\omega\) and \(P\):

\[
(\Omega, A) \succ P^\omega \succ (\Theta, \mathcal{T}) \succ P \succ (\Omega, A),
\]

defined by

\[
PP^\omega(\omega, A) := \int_\Theta P_\theta(A)dP^\omega_\omega(\theta) = \int_\Theta \int_A p_\theta(\omega')p^\omega_\omega(\theta)dQ(\theta)d\mu(\omega'),
\]

we have that

\[
\frac{dPP^\omega(\omega, \cdot)}{d\mu}(\omega') = \int_\Theta p_\theta(\omega')p^\omega_\omega(\theta)dQ(\theta).
\]

Notice that \(PP^\omega(\omega, \cdot) = (P^\omega_\omega)^P\).

Remark. Because of (1), we introduce the notation \(\Pi := P \otimes Q\). So, (2) reads as \(\Pi := \beta_Q^* \otimes P^\omega\). Hence, after observing \(\omega \in \Omega\), replacing the prior distribution \(Q\) by the posterior distribution \(P^\omega_\omega\), we get the probability distribution \(\Pi_\omega := P \otimes P^\omega_\omega\) on \(A \otimes \mathcal{T}\). According to (3), \(PP^\omega(\omega, A) = \Pi_\omega(A \times \Theta) = \Pi^\omega_\omega(A)\) where \(I(\omega, \theta) = \omega\). This way the posterior predictive distribution \((P^\omega_\omega)^P\) given \(\omega\) appears as the marginal \(\Pi_\omega\)-distribution on \(\Omega\).