International Journal of Approximate Reasoning 78 (2016) 138–152

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Frequentistic approximations to Bayesian prevision of exchangeable random elements [☆]

Donato Michele Cifarelli [a], Emanuele Dolera [b,*], Eugenio Regazzini [c,d]

[a] Dipartimento di Scienze delle Decisioni, Università commerciale L. Bocconi, 20136 Milano, Italy
[b] Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, 41125 Modena, Italy
[c] Dipartimento di Matematica, Università di Pavia, 27100 Pavia, Italy
[d] CNR-IMATI, Milano, Italy

ARTICLE INFO

Article history:
Received 31 January 2016
Received in revised form 23 May 2016
Accepted 16 June 2016
Available online 11 July 2016

Keywords:
Asymptotics of prior and posterior distributions
Exchangeability
Inverse probabilities
Probability distances
Random measures

ABSTRACT

Posterior and predictive distributions for m future trials, given the first n elements of an infinite exchangeable sequence $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$, are considered in a nonparametric Bayesian setting. The former distribution is compared to the unit mass at the empirical distribution $\tilde{\mathfrak{e}}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\tilde{\xi}_i}$ of the n past observations, while the latter is compared to the m-fold product $\tilde{\mathfrak{e}}_n^m$. Comparisons are made by means of distinguished probability distances inducing topologies that are equivalent to (or finer than) the topology of weak convergence of probability measures. After stating almost sure convergence to zero of these distances as n goes to infinity, the paper focuses on the analysis of the rate of approach to zero, so providing a quantitative evaluation of the approximation of posterior and predictive distributions through their frequentistic counterparts $\delta_{\tilde{\mathfrak{e}}_n}$ and $\tilde{\mathfrak{e}}_n^m$, respectively. Characteristic features of the present work, with respect to more common literature on Bayesian consistency, are: first, comparisons are made between entities which depend on the n past observation only; second, the approximations are studied under the actual (exchangeable) law of the $\tilde{\xi}_n$'s, and not under hypothetical product laws $\mathfrak{p}_0^\infty$, as $\mathfrak{p}_0$ varies among the admissible determinations of a random probability measure.

1. Introduction

In the present paper the term *prevision*[1] will be used to designate both any activity directed to evaluation of probabilities of future (or, at least, till not known) events on the basis of an observed frequency, and the result of such an activity. Thus, prevision mingles with probabilistic inductive reasoning, and an important field of application of prevision is that of statistical problems, classically characterized by the circumstance that the events considered therein are generally thought of as *analogous* events.[2] Under this very same circumstance, frequentistic approaches to statistics look at observable single events—or more general random elements $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$ taking values in some space $\mathbb{X}$, like in the rest of the present work—as independent and identically distributed (i.i.d.) with a common probability distribution (p.d.) that can be approximated by

☆ This paper is part of the virtual special issue on Bayesian nonparametrics, edited by A. Lijoi, A. Mira and A. Benavoli.

* Corresponding author.
E-mail addresses: emanuele.dolera@unimore.it, emanuele.dolera@unipv.it (E. Dolera), eugenio.regazzini@unipv.it (E. Regazzini).

[1] The term *prevision* is a translation of the Italian *previsione*, adopted by de Finetti, which is used in the English translation of his treatise [24]. See Translator's note on page 151.

[2] It is well-known that modern statistics deals with a variety of different situations in which this kind of analogy is not supposed. Therefore, the relative statistical modeling departs from the classical one considered here.

http://dx.doi.org/10.1016/j.ijar.2016.06.007

observed (empirical) frequency. Laws of large numbers and allied results are then invoked to assert that such an approximation improves as the number of observations goes to infinity. Bayesian statisticians translate the aforesaid analogy into a less restrictive property, that is the *exchangeability* of the $\tilde{\xi}_n$'s. As a consequence, any correct expression of Bayesian prevision must rely on a conditional p.d. for till now unknown observable random elements, given the frequency distribution of observed random elements. The expectation, due to the analogy of the observable elements here realized in the form of exchangeability, is that we are willing to be influenced more and more by the observed frequency as the size of experience increases. The present paper hinges upon the ground of this intuitive expectation. In fact, its possible truth and, even more, any suitable quantification of its validity would provide us with invaluable information about the approximation of Bayesian previsions by frequentistic ones which—as already explained—although cruder, are of easier evaluation. This circumstance comes to the fore, for example, within the so-called *empirical Bayes* approach, which tries to justify partial replacement of orthodox Bayesian reasoning with frequentistic elements. See, e.g., Robbins [40,41], Efron [18] and Remark 2 in Section 3 of the present paper. Moreover, an interpretation of the aforesaid approximation in terms of loss functions will be given in the course of this Introduction.

The present work, which is part of a wide-ranging research, focuses on the discrepancy between posterior (predictive of m future observations, respectively) distribution, given n past observations, and the point mass at (the m-fold product of, respectively) the empirical distribution of the same past observations, when n goes to infinity. The idea to compare a Bayesian inference to any of its frequentistic counterparts goes back, for different motives, to classical authors, such as Laplace [30], Poincaré [37], Bernstein [6], von Mises [34,35], de Finetti [19,20,22], Romanovsky [42], and has had remarkable developments also in recent years, at least in two directions: the *consistency* of Bayesian procedures from a frequentistic point of view, and the *Bernstein–von Mises* phenomenon concerning a version of the central limit theorem for Bayesian estimators, in order to provide confidence regions connected with the aforesaid consistency issue. By way of example, see Schwartz [43], Diaconis and Freedman [13], Barron, Schervish and Wasserman [4], Ghosal, Ghosh and van der Vaart [26], Ghosh and Ramamoorthy [27] for the consistency, and this last book together with Freedman [25] for the Bernstein–von Mises phenomenon. To explain the connection with the present work, one must say that, especially in recent times, these fields of research have aligned themselves more and more with the interpretation of Bayesian inferences as procedures aimed at producing suitable estimators of unknown quantities, whose efficiency is checked from a frequentistic viewpoint. This body of literature on consistency ends up interpreting Bayesian rules as mere functions of the observations and it has more an operational aim, by offering a motivation for frequentists to use Bayesian procedures. To appreciate the peculiarity of our work with respect to the aforesaid lines of research, one should thoroughly retrieve the Bayesian approach to statistical inference, in the spirit of the solution to the problem of *inverse probabilities* provided by de Finetti [19,20] in his earliest papers on exchangeability. Nowadays, Doob [15] is commonly credited as the author of the solution to a generalized form of the same problem, recalled in Theorem 1 of this paper. Indeed, if one reckons that the Bayesian way of thinking indicates, *lato sensu*, the correct way of making statistical inference, it is fair to pursue the above-mentioned goals of approximating posterior and predictive distributions by more tractable laws—typically obtained by frequentistic procedures—depending only on past observations. Doob's theorem is then replaced by a statement concerning the almost sure (a.s.) convergence to zero of any weak probability distance (see Subsection 2.3 below for more information) between the posterior distribution $q(\tilde{\boldsymbol{\xi}}^{(n)}, \cdot)$ and $\delta_{\tilde{\mathfrak{e}}_n}$, the point mass at the empirical measure $\tilde{\mathfrak{e}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{\xi}_i}$, as $n \to +\infty$. Successively, one can deduce the a.s. convergence to zero of any weak probability distance between the predictive distribution of m future observations and the m-fold product $\tilde{\mathfrak{e}}_n^m := \underbrace{\tilde{\mathfrak{e}}_n \otimes \cdots \otimes \tilde{\mathfrak{e}}_n}_{m\text{-times}}$, as $n \to +\infty$, for every $m \in \mathbb{N}$. Moreover, the main results in the present paper involve only *finitary*—hence, empirically ascertainable—entities. In this respect, see Bassetti [5] for the connection with finite exchangeable sequences. At this stage, one can appreciate the further step, made in Theorems 3, 4 and 5, to provide quantitative estimations of the error in the aforesaid approximations. More precisely, considering by way of example the comparison of $q(\tilde{\boldsymbol{\xi}}^{(n)}, \cdot)$ with $\delta_{\tilde{\mathfrak{e}}_n}$, there are a positive (non-random) sequence b_n, going to infinity with n, and a suitable constant $L > 0$ such that, for every $\varepsilon, \eta > 0$, there exists some index $n_0 = n_0(\varepsilon, \eta) \in \mathbb{N}$ satisfying

$$\rho\left(\left\{\max_{\nu \le n \le \nu+m} b_n \mathrm{d}_{[[\mathbb{X}]]}(q(\tilde{\boldsymbol{\xi}}^{(n)}), \delta_{\tilde{\mathfrak{e}}_n}) \le L + \varepsilon\right\}\right) \ge 1 - \eta$$

for every $\nu \ge n_0$ and $m \in \mathbb{N}$, where ρ denotes the p.d. that makes the $\tilde{\xi}_n$'s exchangeable, and $\mathrm{d}_{[[\mathbb{X}]]}$ is a suitable probability distance to be specified in Subsection 2.3. See Remark 2 in Section 3 for more explanation. Allied results, formulated in similar frameworks, can be found in Diaconis and Freedman [14] and in Berti et al. [7]. To capture the spirit of statements of this kind, one can think of $\mathrm{d}_{[[\mathbb{X}]]}(q(\tilde{\boldsymbol{\xi}}^{(n)}), \delta_{\tilde{\mathfrak{e}}_n})$ as loss consequent upon the approximation of the posterior distribution with its "natural" frequentistic counterpart, based on the sample $\tilde{\boldsymbol{\xi}}^{(n)}$. In fact, one has a sequence of losses, converging to zero as $n \to +\infty$ with ρ-probability 1, and an interesting problem is that of estimating the rapidity of such a convergence. For this purpose, one can think of $1/b_n$ as a sort of certainty equivalent loss, for a sample size n, and seek for sequences $\{1/b_n\}_{n \ge 1}$ decreasing to zero, suited to guarantee an overall coverage of the loss. Then, statements under discussion say how such a coverage has to be understood: starting from a suitable sample size, the ratio of the random loss to its certainty equivalent is definitively not allowed to cross a constant barrier. Obviously, there are general reasons, for example of an economic