

# Predictive Inference: A Review and New Developments

Donato M. Cifarelli, P. Muliere  
Bocconi University  
Viale Isonzo 25, 20135 Milano, Italy  
*michele.cifarelli@uni-bocconi.it, pietro.muliere@uni-bocconi.it*

Petrone, S.  
University of Insubria  
Via Ravasi 2, 21100 Varese, Italy  
*spetrone@eco.uninsubria.it*

## 1. Introduction

The general predictive problem related to a sequence  $\{X_n\}$  of random variables involves the evaluation of the probability of an event, dependent on the future realisations of some of the variables of the sequence, when the outcomes of a finite number of variables of the same sequence are assumed to be known. The treatment of this problem which takes into account only strictly observable events has been called *completely predictive*. The recourse to a parametric model, even if it is not necessary, generally simplifies the mathematical aspects of a predictive problem; this approach, which is referred to as *hypothetical*, is the one prevalent in the traditional Bayesian literature. However caution must be adopted when following such an approach; for instance, consistently with de Finetti for whom only observable facts are subject to probabilistic evaluation, it is possible to question the adoption of the hypothetical approach whenever one is not in the position to elicit a prior distribution for the parameter appearing in the model.

## 2. Parametric versus Nonparametric

The basic predictive assumption for a sequence of random variables is that of exchangeability. De Finetti style theorems characterise models in terms of invariance. The idea is that the statistician begins the model building phase by postulating reasonable symmetries for the distribution of the observable facts.

Let  $X_1, X_2, \dots$  be an exchangeable sequence of random variables defined on  $X \subseteq R$ . From de Finetti's representation theorem (de Finetti, 1937) there exists a random distribution function  $F$  conditional on which  $X_1, X_2, \dots$  are i.i.d. from  $F$ . That is, there exists a probability measure, defined on the space of probability measures on  $X$ , such that the joint distribution of  $X_1, X_2, \dots, X_n$ , for any  $n$ , can be written as

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \int \left\{ \prod_{i=1}^n F(A_i) \right\} \mu(dF)$$

where  $\mu$  is the de Finetti (or prior) measure. Therefore, if we assume only the exchangeability, the representation theorem involves an infinite dimensional parameter. This parameter is the weak limit (with probability P one) of the sequence of the empirical distribution functions. In order to justify the dependence of this limit to a finite-

dimensional parameter (hence a parametric approach) further assumptions must be introduced on the observables. For example, as we shall see, the existence of a predictive sufficient statistics.

### 3. Characterisation of Priors using Predictive Assumptions

The general predictive problem reduces to the computation of the conditional probability

$$P(X_{n+1} \in A | X_1, X_2, \dots, X_n)$$

for measurable sets  $A$ . The assumption of exchangeability and the representation theorem imply that

$$P(X_{n+1} \in A | X_1, X_2, \dots, X_n) = E(F(A) | X_1, X_2, \dots, X_n).$$

Without further assumptions, we need a prior on the infinite-dimensional parameter  $F$ . Unfortunately, when de Finetti (1935) suggested the general predictive approach, non-parametric priors were not known yet. Today, many proposals can be found in the literature, but there remains the problem of how to select the prior. One approach is to select  $\mu$  by appealing to prior information about  $F$  and attempting to incorporate this information into  $\mu$ . This is often a difficult task for non-parametric priors. Alternatively, we may try to describe our state of knowledge in terms of probabilistic assumptions on  $X_{n+1}$  given  $X_1, X_2, \dots, X_n$ , for  $n=1, 2, \dots$ , and consequently characterise the prior  $\mu$ .

Indeed, when  $\{X_n\}$  is an infinite sequence of random variables, the completely predictive approach to the construction of the law of the sequence is based on the specification of the distribution  $F_1$  of  $X_1$  and of the predictive distribution  $F_{n+1}$  of  $X_{n+1}$  given  $X_1, X_2, \dots, X_n$  for all  $n \geq 1$ . Whereas the Ionescu-Tulcea extension theorem states consistency conditions which guarantee the existence of a unique law for  $\{X_n\}$  determined by the sequence  $\{F_n\}$ , Fortini, Ladelli and Regazzini (2000) give necessary and sufficient conditions on the sequence  $\{F_n\}$  for the exchangeability of the law of  $\{X_n\}$ . This result characterises exchangeability in purely predictive terms; the de Finetti measure of the sequence  $\{X_n\}$  is then obtained by means of de Finetti's Representation Theorem.

Many priors used in Bayesian nonparametrics can easily be constructed following this approach; for example, the Dirichlet process (Regazzini, 1978; Lo, 1991), the Polya trees (Walker and Muliere, 1997a), the beta-Stacy (Walker and Muliere, 1997b), the Neutral to the right processes (Walker and Muliere (1999).

### 4. Predictive Sufficiency

For justifying a parametric approach, we need further assumptions on the observables, in addition to exchangeability. From a predictive point of view, an example of such assumptions is predictive sufficiency. Predictive sufficiency and its properties have been investigated in a number of papers among which: Campanino and Spizzichino (1981), Cifarelli and Regazzini (1980, 1981, 1982), Dawid (1982), Secchi (1987), Muliere

and Secchi (1992), Fortini, Ladelli and Regazzini (2000). Related notions of sufficiency have been studied by Lauritzen (1984,1988) and Diaconis and Freedman (1984).

In many practical situations, the researcher can assume, in addition to exchangeability, that a statistic  $T_n$  summarises all the information provided by  $X_1, X_2, \dots, X_n$  for predicting  $X_{n+1}$ . Then  $T_n$  is called *predictive sufficient statistic*.

When  $T_n$  is a linear function, Cifarelli and Regazzini (1982) have shown that, under some hypotheses, the probability law of  $X_1, X_2, \dots, X_n$  can be represented by means of a parametric model, where the model  $F$  is the limit of the sequence of predictive distributions of  $X_{n+1}$  and the prior on the parameter  $\Theta$  is the limit law of the sequence  $\{T_n\}$ . Fortini, Ladelli and Regazzini (2000) relax the hypotheses required.

This result does not say how to select the prior on  $\Theta$ . Anyway, Muliere and Secchi (1992) show that it is often reasonable to approximate the posterior distribution of  $\Theta$  given  $X_1, X_2, \dots, X_n$  by means of the distribution of  $\{T_n\}$  obtained by using a bootstrap procedure. This procedure results equivalent, from the completely predictive point of view, to those obtained by a Bayesian who decides to adopt a suitable improper prior distribution for  $\Theta$ .

## 5. Urn Schemes for Constructing Priors

In the context of Bayesian non-parametric inference, the importance of Blackwell and Mac-Queen's result (Blackwell and Mac-Queen, 1973) is that it gives a simple and concrete procedure for constructing an infinite sequence of random variables with Dirichlet process as de Finetti measure. The procedure has the additional advantage of making intuitively clear some of the mathematical properties of the Dirichlet process, like its conjugate property or the form of the predictive distribution of the  $(n+1)$ -th random variables generated by a Dirichlet process conditionally on the values of the first  $n$  variables.

In the spirit of Blackwell and MacQueen we present in this section a class of stochastic processes defined on a countable space of Polya urns which will be convenient for constructing more general classes of priors commonly used in Bayesian non-parametric inference, such as Polya trees and beta-Stacy processes. There are situations where the assumption of exchangeability for the sequence of observations is too restrictive or does not incorporate all the relevant information about the data. A weaker assumption is that of partial exchangeability, introduced by de Finetti (1938) and considered also by Diaconis and Freedman (1980). For the connections between the two ideas of partial exchangeability see Fortini, Ladelli, Petris and Regazzini (1999). When  $\{X_n\}$  is an infinite sequence of random variables with values in a discrete space, partial exchangeability (in the sense of Diaconis and Freedman) and recurrence imply that the law of the sequence is that of a mixture of Markov chains (Diaconis and Freedman, 1980); that is, conditionally on a random transition matrix  $\Pi$ ,  $\{X_n\}$  is a Markov chain with transition matrix  $\Pi$ . The prior distribution for  $\Pi$  may often be characterised in purely predictive terms; for example, Muliere, Secchi and Walker (2000) introduce an urn scheme called *reinforced urn process* which generates mixtures of Markov chains such that the law of  $\Pi$  is the product of Dirichlet processes. Reinforced urn processes have applications to survival analysis whenever individual specific data is modelled by a Markov chain and individuals from the population are assumed to be exchangeable.

## 6. Consistency

For subjective Bayesians like de Finetti and Savage probabilities represent degree of belief and there are no objective probability models. Bayesians learn from experience, so opinions based on very different priors will *merge* as data accumulate. A general result of this type was provided by Blackwell and Dubins (1962). A result relating *merging of opinions* and *posterior consistency* is discussed in Diaconis and Freedman (1986).

Roughly speaking, the posterior is consistent for  $F_0$  if it cumulates around  $F_0$  as the sample size increases, almost surely with respect to the product measure  $F_0^\infty$ . Bayesian nonparametric methods have only recently started to undergo asymptotic studies. Much of the papers are influenced by the paper of Diaconis and Freedman (1986). For a comprehensive review of this area see Wasserman (1998) and Ghoshal, Gosh and Ramamoorthi (1997). Recent results are in Walker and Hjort (2001) and Petrone and Wasserman (2001).

Our aim is to discuss consistency from a predictive point of view. In particular, we shall focus on the asymptotic behaviour of the predictive distribution. As shown in the previous sections, the representation theorem for exchangeable sequences and the results about predictive sufficiency ensure that the sequence of predictive distribution functions converges to the random distribution function  $F$  conditionally on which the observables are i.i.d. Our aim is to discuss how these results, which involve a *random* limit distribution  $F$ , are related to the notion of consistency of Doob (1948) or of Diaconis and Freedman (1986), in which  $F$  is the (fixed) true distribution.

On the other hand, starting from a paper by Diaconis and Freedman (1990), we might replace the true distribution function with the empirical distribution function. In particular it is of interest to study the asymptotic distance (in some sense) between the predictive distribution function and the empirical distribution function; results in this direction are proved in Berti and Rigo (1997).

## References

- Berti, P. and Rigo, P., (1997), A Glivenko-Cantelli theorem for exchangeable random variables, *Statistics and Probability Letters*, **32**, pp.385-391.
- Cifarelli, D.M. and Regazzini, E., (1982), Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability. *Exchangeability in Probability and Statistics* (G. Koch and F. Spizzichino, eds.), pp. 217-232, North-Holland, Amsterdam.
- Cifarelli, D.M., Muliere, P. and Secchi, P. (1999), Prior processes for Bayesian nonparametrics, *Technical Report n.377/P*, Politecnico di Milano.
- Cifarelli, D.M., Muliere, P. and Secchi, P. (2000), Urn schemes for constructing priors, *Technical Report n.429/P*, Politecnico di Milano.
- Fortini, S., Ladelli, L. and Regazzini, E., (2000), Exchangeability, Predictive distributions and Parametric Models, *Sankhya*, vol. **62**, Ser. A., pp. 86-109.
- Muliere, P. and Secchi, P., (1992), Exchangeability, Predictive Sufficiency and Bayesian Bootstrap, *Journal of the Italian Statistical Society*, **3**, pp.377-404.
- Muliere, P., Secchi, P. and Walker, S.G., (2000), Urn schemes and reinforced random walks, *Stochastic Processes and their Applications*, **88**, pp.59-78.
- Petrone, S. and Wasserman, L. (2001), Consistency of Bernstein posterior (under revision for *Journal of the Royal Statistical Society, Ser. B*).
- Walker, S. and Hjort, N.L. (2001), On Bayesian consistency. *Journal of the Royal Statistical Society, Ser. B*, to appear.
- Walker, S. and Muliere, P., (1999), A characterization of a Neutral to the right prior via an extension of Jhonson's sufficientness postulate, *The Annals of Statistics*, **27**, pp.589-599.