Electronic Journal of Statistics
Vol. 19 (2025) 1609–1632
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2369

Bayesian analysis with conditionally identically distributed sequences

Pier Giovanni Bissiri[1] **and Stephen G. Walker**[2]

[1] *Department of Economics, Management and Statistics, University of Milano-Biccoca, e-mail:* pier.bissiri@unimib.it

[2] *Department of Mathematics, University of Texas at Austin, e-mail:* s.g.walker@math.utexas.edu

Abstract: The paper undertakes Bayesian style inference using posterior distributions. The key difference is that we use an assumption of a conditionally identically distributed (c.i.d.) sequence rather than the more common exchangeable sequence. We show that there remains the existence of a prior and posterior while the updating mechanism is achieved through the predictive distributions. This is sufficient given a fundamental result of Doob which explained how posteriors can be constructed in the exchangeable case via predictive distributions. We model the predictive distributions using copulas ensuring the c.i.d. structure.

Keywords and phrases: copula model, MLE, predictive density.

Received June 2024.

1. Introduction

Exchangeability, [7], is one of the foundational concepts for Bayesian analysis, see [2]. The implication of assuming a sequence of variables, say $(X_1, X_2, \ldots)$ are exchangeable, is that a prior probability on a space of suitable distribution functions is guaranteed. In short, the sequence is conditionally i.i.d., where the conditioning is on a random distribution function. See also [12].

The assumption of exchangeability is what allows the Bayesian to set up a learning and updating process for the distribution of X_{n+1} given $X_{1:n}$. Indeed, the data generating mechanism is revised as more data become observed. The update works so that the order in which the observations arise does not matter. Hence, exchangeability can be seen as more of a tidying exercise, since there are multiple ways that it is possible to express a predictive model $p(x_{n+1} \mid x_{1:n})$ without the constraint that joint density functions become symmetric.

As models become more complex, so the exchangeable constraint can lead to intractable updates. One such well known example is the Dirichlet process mixture model, requiring MCMC methods to derive density estimators and full posterior inference. See, for example, [13].

A good reason for relaxing the exchangeable constraint is if it is believed the appropriate predictive density estimator at any sample size n is the density, from the chosen family of density functions modeling the data, evaluated at the maximum likelihood estimator. However, the one constraint we will impose

is that the marginal predictive density functions for each X_m, $m > n$, given $X_{1:n}$, for all n, are the same. This would appear to be a minimal necessary coherent structure for modeling in an i.i.d. setting, assumed for the observed data. This makes sense in that there can be no reason to predict X_{n+2} differently from X_{n+1} once $X_{1:n}$ have been seen. The sequence in this case is known as conditionally identically distributed (c.i.d.); see [6]. For more recent literature on c.i.d. sequences, see [1], [10] and [4].

The aim in the paper is to start with the assumption of c.i.d. for the sequence of variables, corresponding to and generalizing the assumption of exchangeability of the variables. The papers [5] and [3] provide examples of c.i.d. sequences which are not exchangeable. In the present paper, the c.i.d. sequences can be constructed in general using copula models and in particular we show how to construct the prior and the posterior using c.i.d. sequences. Here, we outline the key idea. A sequence $X_{1:\infty}$ is c.i.d. if $E(F_m \mid X_{1:n}) = F_n$ for all $m > n$, where F_m is the conditional distribution of X_{m+1} given $X_{1:m}$. A result from [6] says that F_m converges weakly almost surely to a random distribution which acts as a sample from the prior. So, in this paper, the general prior is the distribution of the sequence $X_{1:\infty}$. A specific prior for a statistic of interest S, for example the mean, is the distribution of $S_\infty = \theta(X_{1:\infty})$, which for the mean would be $S_\infty = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} X_i$. Similarly, the general posterior is the distribution of the sequence $X_{n+1:\infty}$ given $X_{1:n}$. Likewise this defines the posterior for a statistic of interest. In the parametric case we define the prior and posterior via the distribution of a well chosen estimator of the parameter. Hence, we can see immediately that the expected posterior distribution is F_n, the distribution for X_{n+1} given $X_{1:n}$, which is now fixed. It is this procedure which we wish to detail for a parametric model. It is necessary to construct the martingales again and the general way to do this is by using copulas. Indeed, copulas are necessary.

The layout of the paper is as follows: In Section 2 we set out the necessary background for the paper, specifically c.i.d. sequences and their properties. We also provide relevant information on copula models which are the main tool for the construction of c.i.d. sequences. Section 3 presents the main results, specifically how the c.i.d. sequences can characterise prior and posterior distributions. The idea is that the sequence $X_{1:\infty}$ characterize the prior while $X_{n+1:\infty}$ given $X_{1:n}$ characterize the posterior. For a prior and posterior on a parameter space we use the random MLE estimator associated with the relevant observations, demonstrating convergence from the infinite sequences. Briefly here, a c.i.d. sequence $X_{1:\infty}$ can induce a prior distribution on a parameter space Θ associated with model $f(x;\theta)$, $\theta \in \Theta$, by defining it to be the distribution of $\widehat{\theta}_\infty$, which is the MLE based on the sample $X_{1:\infty}$. Section 3 shows such a distribution exists. This set up is exactly as the Bayesian prior works, which follows from the work of Doob (1949). In Section 4 we provide some illustrations and Section 5 contains the proofs.

Before proceeding, it is helpful to detail the connection between the c.i.d. sequence and the model $f(x;\theta)$. The c.i.d. sequence, whether it starts at X_1 or X_{n+1} having seen $X_{1:n}$, defines a θ_∞ under certain conditions: see Theorem 3.2. By starting the sequence at $n+1$, so the c.i.d. sequence is $X_{n+1:\infty}$, leads to a

θ_∞ which represents a sample from the posterior given $X_{1:n}$. Whereas, if the sequence starts at 1, the θ_∞ would represent a sample from the prior. It is our work to show that in both cases such a θ_∞ exists.

The role played the model is that the c.i.d. sequence could start at $f(\cdot;\theta_n)$, where θ_n is a parameter estimator from $X_{1:n}$, so marginally all the $x_{n+1:\infty}$ are coming from this density function. See for example the illustration at the start of section 4. On the other hand, the c.i.d. sequence could be based on a nonparametric estimator using a copula model.

2. Preliminaries

Let $X_{1:\infty} = (X_1, X_2, \dots)$ be a sequence of real valued random observations. Precisely, for each $n = 1, 2, \dots$, $X_n = X_n(\omega)$ is a random variable defined on a probability space $(\Omega, \mathscr{F}, \mathrm{P})$ and valued into $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ where $\mathscr{B}(\mathbb{R})$ is the Borel sigma-field of $\mathbb{R}$. All random variables considered in this paper are defined on the probability space $(\Omega, \mathscr{F}, \mathrm{P})$.

In Bayesian statistics, the most common choice is to assign to the sequence $X_{1:\infty}$ an *exchangeable* distribution, namely a distribution that is invariant under finite permutations. In other words, the sequence $X_{1:\infty}$ has an exchangeable distribution if and only if for every $n = 1, 2, \dots$, the distribution of $(X_1, \dots, X_n)$ is the same of $(X_{\pi_1}, \dots, X_{\pi_n})$, for every permutation π of $\{1, \dots, n\}$. In such case, the random sequence $X_{1:\infty}$ itself is said to be exchangeable.

Exchangeability is known to be connected to the notion of random probability measure. A random probability measure on $\mathbb{R}$ is a measurable function defined on $(\Omega, \mathscr{F})$ and valued into $(\mathbb{P}, \mathscr{P})$, where $\mathbb{P}$ is the set of all probability measures on $\mathbb{R}$ and $\mathscr{P}$ is the sigma–field on $\mathbb{P}$ generated by the evaluation maps $p \to p(B)$, for B varying in $\mathscr{B}$.

The celebrated *de Finetti's representation theorem* tells us that the exchangeability of $X_{1:\infty}$ is tantamount to the existence of a random probability measure $\mu = \mu(\cdot)(\omega)$ such that the observations $X_1, X_2, \dots$ are conditionally i.i.d. given μ and the conditional distribution of X_1 given μ is μ. It is known that $\mu(B) = \mathrm{P}(X_1 \in B \mid \mathscr{T})$, for every $B \in \mathscr{B}(\mathbb{R})$, where $\mathscr{T}$ is the tail sigma-field, namely $\mathscr{T} = \cap_{n=1}^{+\infty} \sigma(X_{n:\infty})$, being $\sigma(X_{n:\infty})$ the sigma–field generated by $X_{n:\infty} = (X_n, X_{n+1}, \dots)$. Moreover, μ is known to be the the almost sure weak limit of both the empirical measure $\mu_n = \sum_{i=1}^n \delta_{X_i}/n$ and the predictive distribution $\mathrm{P}(X_{n+1} \in \cdot \mid X_{1:n})$, where we mean conditioning w.r.t. the sigma–field generated by the random vector $X_{1:n} = (X_1, \dots, X_n)$. Furthermore,

$$\mu(B) = \lim_{n \to +\infty} \mu_n(B), \tag{1}$$

$$\mu(B) = \lim_{n \to +\infty} P(X_{n+1} \in B \mid X_{1:n}), \tag{2}$$

almost surely and in L^1, for every $B \in \mathscr{B}(\mathbb{R})$. The random measure μ is called *the directing measure* of the sequence $X_{1:\infty}$. For more information about exchangeable sequences, see for instance [16].

A referee has pointed out that (2) implies (1). The proof, also supplied by the referee, uses the martingale sequence $M_0 = 0$ and

$$M_n = \sum_{i=1}^{n} \frac{1(X_i \in A) - P(X_i \in A \mid X_{1:i-1})}{i}.$$

Then $\sup_n E(M_n^2) < \infty$ so M_n converges almost surely and the Kronecker lemma yields

$$\mu_n(A) - n^{-1} \sum_{i=1}^{n} P(X_i \in A \mid X_{1:i-1}) \to 0 \quad \text{a.s.}$$

If (2) holds then $P(X_{n+1} \in A \mid X_{1:n}) \to \mu(A)$ and so $\mu_n(A) \to \mu(A)$.

In this paper, we consider a predictive approach rather than the traditional prior-posterior approach since we shall assess directly the predictive distributions following [9]. Moreover, as in [9], we relax the exchangeability condition considering instead the weaker condition of *conditional identical distribution* introduced and studied by [6].

As we will see, if we relax the assumption of exchangeability to c.i.d., the key elements of Bayesian inference, i.e. a prior and posterior distribution on a parameter space, remain. This has been defined appropriately. Crucially, it is the missing data $X_{n+1:\infty}$ given the observed data which are taken as c.i.d. The data are seen and so do not need a model. What is needed is a model for the missing data. There is no concern if many $p(X_{n+1:\infty} \mid X_{1:n})$ lead to a common posterior on the parameter space.

Definition 2.1. The random sequence $X_{1:\infty}$ is said to be *conditionally identically distributed* (c.i.d.) if and only if

$$\mathbb{E}(g(X_k)) = \mathbb{E}(g(X_1)) \tag{3}$$

$$\mathbb{E}(g(X_k) \mid X_{1:n}) = \mathbb{E}(g(X_{n+1}) \mid X_{1:n}), \quad \text{a.s.} \tag{4}$$

for all $k > n \geq 1$ and all bounded measurable functions $g : \mathbb{R} \to \mathbb{R}$.

The following Proposition provides some characterizations of c.i.d. sequences [6].

Proposition 1. The following statements are equivalent:

1. The random sequence $X_{1:\infty}$ is c.i.d.
2. The elements of the sequence $X_{1:\infty}$ are identically distributed and for every $n \geq 1$, the conditional distribution of X_{n+2} given $X_{1:n}$ is the same as the the conditional distribution of X_{n+1} given $X_{1:n}$.
3. $X_2 \sim X_1$, and for every $n \geq 1$,

$$(X_1, \ldots, X_n, X_{n+2}) \sim (X_1, \ldots, X_n, X_{n+1}), \tag{5}$$

where "$\sim$" means "distributed as".

4. If $g : \mathbb{R} \to \mathbb{R}$ is bounded and measurable, $W_0 = g(X_1)$ and

$$W_n = \mathbb{E}(g(X_{n+1}) \mid X_{1:n})$$

for every $n \geq 1$, then $W_{1:\infty}$ is a martingale with respect to the filtration generated by $X_{1:\infty}$, namely for every $n \geq 1$,

$$\mathbb{E}(W_{n+1} \mid X_{1:n}) = W_n, \quad \text{a.s.}$$

A relevant fact is that a c.i.d. sequence obeys a strong law of large numbers (SLLN).

Here we present a main result of [6] who proved that if $X_{1:\infty}$ is c.i.d. then there exists an random probability measure μ satisfying (1) and (2) which the authors of [6] call *representing measure*.

Proposition 2. If $X'_{1:\infty} = (X'_1, X'_2, \dots)$ is c.i.d. then there exists a random sequence $(Y, X_1, X_2, \dots)$ such that $X_{1:\infty} \sim X'_{1:\infty}$, $Y \sim X_1$ and

$$(X_1, \dots, X_n, Y) \sim (X_1, \dots, X_n, X_{n+1}), \tag{6}$$

for every $n \geq 1$.

Moreover, for every measurable $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}(|g(X_1)|) < \infty$, there exists a random variable V_g such that:

$$V_g = \lim_{n \to +\infty} \mathbb{E}(g(X_{n+1}) \mid X_{1:n}) = \mathbb{E}(g(Y) \mid X_{1:\infty}) = \int_{\mathbb{R}} g(x)\, \mu(\mathrm{dx}) \tag{7}$$

where the limit holds almost surely and in L^1, and μ is a random probability measure such that for every $B \in \mathscr{B}(\mathbb{R})$,

$$\mu(B) = \mathrm{P}(Y \in B \mid X_{1:\infty}). \tag{8}$$

Therefore, in particular, (1) and (2) hold true a.s. and in L^1, for every $B \in \mathscr{B}(\mathbb{R})$.

Proof. It follows from (5) and Kolmogorov's consistency theorem that there exists a random sequence $(Y, X_1, X_2, \dots)$ such that $X_{1:\infty} \sim X'_{1:\infty}$, $Y \sim X_1$ and (6) holds.

In order to prove the second equality in (7), note that the sequence $\mathbb{E}(g(Y) \mid X_{1:n})$ is a martingale converging a.s. and in L^1 to $\mathbb{E}(g(Y) \mid X_{1:\infty})$ by Lévy's Upwards Theorem. By (6),

$$\mathbb{E}(g(Y) \mid X_{1:n}) = \mathbb{E}(g(X_{n+1}) \mid X_{1:n})$$

and therefore the second equality in (7) holds a.s. and in L^1.

Note that third equality in (7) holds since Y is real valued and this ensures the existence of a regular conditional distribution of Y given $X_{1:\infty}$ [16, Theorem B.32, page 618]. □

In the following Theorem we provide a key limit result, the proof can be found in [6].

Theorem 2.2. *If $X_{1:\infty}$ is a c.i.d. random sequence and $g : \mathbb{R} \to \mathbb{R}$ is a measurable function such that $\mathbb{E}(|g(X_1)|) < \infty$, then the random variable V_g satisfying (7) is such that:*

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = V_g \tag{9}$$

almost surely and in L^1.

All exchangeable sequences are c.i.d. but the vice versa is not true. A very relevant fact about c.i.d. sequences is that there still exists a random probability measure μ satisfying (1) and (2) for every $B \in \mathscr{B}(\mathbb{R})$. This means that the object of inference μ is well defined. The existence of μ is proved by [6] resorting to the notion of stable convergence.

Before concluding this section, we recall the notion of copula which will be heavily used in the rest of the paper.

Definition 2.3. A *bivariate copula* is a bivariate cumulative distribution function on $[0,1]^2$ with uniform marginal distributions, namely a function $C : [0,1]^2 \to [0,1]$ such that:

1. for every $u, v \in [0,1]$, $C(u,0) = C(0,v) = 0$
2. for every $0 \le u_1 \le u_2 \le 1$ and every $0 \le v_1 \le v_2 \le 1$,

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$$

3. $C(u,1) = u$ for every $u \in [0,1]$ and $C(1,v) = v$ for every $v \in [0,1]$.

A very useful result about copulas is the well-known Sklar's Theorem:

Theorem 2.4 (Sklar's Theorem [17])**.** *If F is a bivariate cumulative distribution function with marginals F_1 and F_2, then there exists a copula C such that for all $y_1, y_2 \in [-\infty, +\infty]$,*

$$F(y_1,y_2) = C(F_1(y_1), F_2(y_2)). \tag{10}$$

If F_1 and F_2 are continuous then C is unique. Conversely, if C is a copula and F_1 and F_2 are distribution functions, then the function F defined by (10) is a joint cumulative distribution function with margins F_1 and F_2.

For more information about copulas see for instance [14].

3. Main results

We start by providing a characterization of c.i.d. sequences using copulas, as these will be used as the main tool for constructing such sequences.

Theorem 3.1. *The sequence $X_{1:\infty}$ is c.i.d. if and only if there exists a set of bivariate copulas $\mathscr{C} = \{C_{x_{1:n}} : n \ge 0,\ x_{1:n} \in \mathbb{R}^n \text{for } n \ge 1\}$*

such that

$$C_{x_{1:n}}(F(x \mid x_{1:n}), F(y \mid x_{1:n})) = \mathrm{P}(X_{n+1} \le x, X_{n+2} \le y \mid X_{1:n} = x_{1:n}), \tag{11}$$

for every $x_{1:n} = (x_1, \dots, x_n) \in \mathbb{R}^n$*, every* $x, y \in \mathbb{R}$*, and every* $n \ge 1$*, where*

$$F(x \mid x_{1:n}) = \mathrm{P}(X_{n+1} \le x \mid X_{1:n} = x_{1:n}),$$

for every $x \in \mathbb{R}$*.*

Moreover any set $\mathscr{C}$ *of copulas ensures the existence of a c.i.d. sequence* $X_{1:\infty}$ *satisfying* (11) *and together with the (marginal) distribution of* X_1 *uniquely characterizes the probability distribution of the sequence* $X_{1:\infty}$*.*

Proof. Let us prove the "if" part. It is known that for any copula C the function $u \to C(u, v)$ is continuous for every $v \in [0, 1]$ [see 14, Corollary 2.2.6, page 12]. Therefore, by 3 in Definition 2.3, (11) yields that for every $k \ge 1$ and every $y \in \mathbb{R}$

$$\begin{aligned}
&\mathrm{P}(X_{k+2} \le y \mid X_{1:k} = x_{1:k}) \\
&\quad = \lim_{x \to +\infty} \mathrm{P}(X_{k+1} \le x, X_{k+2} \le y \mid X_{1:k} = x_{1:k}) \\
&\quad = \lim_{x \to +\infty} C_{x_{1:k}}(F(x \mid x_{1:k}), F(y \mid x_{1:k})) \\
&\quad = C_{x_{1:k}}(1, F(y \mid x_{1:k})) \\
&\quad = F(y \mid x_{1:k}) = P(X_{k+1} \le y \mid X_{1:k} = x_{1:k}).
\end{aligned}$$

This implies that for every $n < k$,

$$\begin{aligned}
\mathrm{P}(X_{k+1} \le y \mid X_{1:n}) &= \mathbb{E}(\mathrm{P}(X_{k+1} \le y \mid X_{1:k-1}) \mid X_{1:n}) \\
&= \mathbb{E}(\mathrm{P}(X_k \le y \mid X_{1:k-1}) \mid X_{1:n}) \\
&= \mathrm{P}(X_k \le y \mid X_{1:n}),
\end{aligned}$$

which in turn by induction implies that

$$\mathrm{P}(X_k \le y \mid X_{1:n}) = \mathrm{P}(X_{n+1} \le y \mid X_{1:n})$$

for all $k > n$. and therefore the random sequence $X_{1:\infty}$ is c.i.d.

The "only if" part can be proved just applying Sklar's Theorem 2.4 to the conditional cumulative distribution function of the pair (X_{n+1}, X_{n+2}) given $X_{1:n} = x_{1:n}$. Indeed, such theorem ensures the existence of the copula $C_{x_{1:n}}$ satisfying (11) provided that $X_{1:\infty}$ is c.i.d., namely provided that the conditional cumulative distribution function of X_{n+1} given $X_{1:n}$ coincides with the conditional cumulative distribution function of X_{n+2} given $X_{1:n}$, which is $F(\cdot \mid X_{1:n})$, for every $n \ge 1$.

The last statement can be proved applying Ionescu-Tulcea Theorem noticing that the copula function $C_{x_{1:n}}$ allows to assess the conditional distribution function $F(\cdot \mid x_{1:n+1})$ through (11) on the basis of $F(\cdot \mid x_{1:n})$. □

The following Theorem establishes almost sure convergence of the maximum likelihood estimator (MLE) in presence of c.i.d. observations. The limiting random variable can be therefore considered to be the object of inference.

Before stating the next Theorem, let us introduce the likelihood function. Denote by Θ the parameter space and consider the function $f : \mathbb{R}\times\Theta \to [0, +\infty)$ such that $\int_{\mathbb{R}} f(x;\theta)\,\nu(\mathrm{dx}) = 1$, for every $\theta \in \Theta$ and some sigma-finite dominating measure ν on $\mathbb{R}$. Typically, ν is the Lebesgue measure.

Before stating the next Theorem, recall that μ denotes the representing measure of the c.i.d. sequence, namely the conditional distribution of Y given $X_{1:\infty}$.

Theorem 3.2. *If the following hold:*

1. *The random sequence $X_{1:\infty}$ is c.i.d.*
2. *The space Θ is a convex subset of $\mathbb{R}^d$,*
3. *The function $\theta \to \log f(x,\theta)$ is strictly concave and continuous for every $x \in \mathbb{R}$,*
4. *Either Θ is closed or for every $x \in \mathbb{R}$ the function $\theta \to f(x \mid \theta)$ can be extended with continuity to the closure $\overline{\Theta}$ of Θ (in the latter case, we let $f(x \mid \theta)$ be the extended function).*
5. *Either $\Theta \subseteq \mathbb{R}^d$ is bounded or*

$$\lim_{\|\theta\|\to+\infty} f(x,\theta) = 0 \tag{12}$$

   for every $x \in B$ and for some B such that $P(X_1 \in B) = 1$
6. *For every $\theta, \theta' \in \Theta^*$,*

$$\mathbb{E}(|\log f(X_1,\theta) - \log f(X_1,\theta'|) < +\infty \tag{13}$$

   where

$$\Theta^* = \{\theta \in \overline{\Theta} : P(X_1 \in \{x \in \mathbb{R} : f(x,\theta) > 0\}) = 1\}$$

7. *For every $\theta \in \overline{\Theta}$, there exists a $\rho > 0$, such that for every $\theta'' \in \Theta^*$ with $||\theta - \theta''|| < \rho$, we have:*

$$\mathbb{E}\left(\sup_{\theta'\in\Theta^*:||\theta'-\theta||<\rho} \log f(X_1,\theta') - \log f(X_1,\theta'')\right) < +\infty \tag{14}$$

8. *There exists a $\rho > 0$, such that for every $\theta \in \Theta$ with $||\theta|| > \rho$, we have:*

$$\mathbb{E}\left(\sup_{\theta'\in\Theta:||\theta'||>\rho} \log f(X_1,\theta') - \log f(X_1,\theta)\right) < +\infty \tag{15}$$

Then:

1. *For every $n \geq 1$, there uniquely exists a random variable*

$$\hat{\theta}_n = \hat{\theta}_n(X_1,\ldots,X_n)$$

such that

$$\frac{1}{n}\sum_{i=1}^{n}\{\log f(X_i,\hat{\theta}_n)-\log f(X_i,\theta_0)\} = \sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\{\log f(X_i,\theta)-\log f(X_i,\theta_0)\},$$

for every $\theta_0 \in \Theta$

2. *There uniquely exists a random variable* T *such that:*

$$\int_{\mathbb{R}}\{\log f(x,T)-\log f(x,\theta_0)\}\,\mu(\mathrm{dx}) = \sup_{\theta\in\Theta}\int_{\mathbb{R}}\{\log f(x,\theta)-\log f(x,\theta_0)\}\,\mu(\mathrm{dx}),$$

for every $\theta_0 \in \Theta$,

3. $\hat{\theta}_n \to T$ *a.s.*

The proof of the theorem is provided in Section 5. In this case T represents a sample from the prior distribution.

Our original condition, which covered the present conditions (6), (7) and (8), was the existence of a function $k(x)$ for which $\log f(x;\theta) \le k(x)$ for all $\theta \in \Theta$ and $x \in \mathbb{R}$, and $E|k(X_1)| < \infty$. This is precisely the condition appearing in [19] who considered an MLE with a misspecified model, for which there are similarities to our setting in that we have a random limit of the log likelihood function. However, the condition is quite strong. The new conditions are less strict and as always in such problems are seeking to provide a uniform law of large numbers type result for the convergence of the log likelihood. A general result in this direction using stochastic equicontinuity is provided by [15]; though his result is restricted to non-random limits. We need to extend slightly as we have a random limit.

Strict concavity of the log likelihood; i.e. condition 3. above, implies identifiability. Indeed, if the log likelihood is strictly concave then

$$\theta \to \int_{-\infty}^{\infty}\log f(x,\theta)\,f(x,\theta_0)\,\nu(\mathrm{d}x)$$

is also strictly concave and therefore it has a unique maximum which is attained at $\theta = \theta_0$ and therefore the Kullback-Leibler divergence between $f(x,\theta)$ and $f(x,\theta_0)$ is positive whenever $\theta \neq \theta_0$.

The following corollary is an application of the Theorem 3.2 to the exponential family.

Corollary 3.1. *Let*

$$f(x,\theta) = h(x)\,\exp\{\tau(\theta)\,T(x) - A(\theta)\}, \tag{16}$$

for some measurable function $T = (T_1, \ldots, T_d) : \mathbb{R} \to \mathbb{R}^d$*, for some continuous function* A *and some continuous invertible function* $\tau : \Theta \to H$*, where* H *is a convex subset of* $\mathbb{R}^d$*, and for every* $x \in \mathbb{R}$ *and every* $\theta \in \Theta$*.*

Assume that:

1. *The exponential family* (16) *is minimal, i.e.*

$$(\tau(\theta_1) - \tau(\theta_2))(T(x_1) - T(x_2)) = 0$$

   for some $\theta_1, \theta_2 \in \Theta$ *and some* $x_1, x_2 \in \mathbb{R}$ *if and only if* $\theta_1 = \theta_2$ *or* $x_1 = x_2$*.*
2. $\mathbb{E}(|T_j(X_1)|) < \infty$*,* $j = 1, \ldots, d$*,*
3. *Either the image of* τ *is bounded or there exists* B *such that* $P(X_1 \in B) > 0$ *and* $\lim_{||\eta|| \to \infty} A(\tau^{-1}(\eta)) - \eta T(x) = +\infty$ *for every* $x \in B$

Then the MLE $\hat{\theta}_n$ *converges almost surely.*

Proof. One can apply Theorem 3.2 considering the exponential family in canonical form, namely

$$\tilde{f}(x, \eta) = f(x, \tau^{-1}(\eta)) = h(x) \exp\{\eta\, T(x) - \tilde{A}(\eta)\},$$

where $\tilde{A} = A \circ \tau$. Indeed, if the exponential family is minimal then $\tilde{A}$ is strictly convex and therefore $\log \tilde{f}(x, \cdot)$ is strictly concave for every $x \in \mathbb{R}$. Condition 5 of Theorem 3.2 is ensured by condition 3 of the present theorem. The other conditions of Theorem 3.2 follow from continuity of τ and condition 2 of the present theorem. By Theorem 3.2, the MLE of the parameter η converges almost surely and then apply the invariance principle of the MLE and the continuous mapping theorem to obtain the result. □

4. Illustrations: Priors and posteriors

We start by showing how the c.i.d. sequence defines both a prior and posterior. To construct the c.i.d. sequence we use a copula model, which will be elaborated on later in this section. The prior construction is based on the theory that if the $X_{1:\infty}$ are c.i.d. and $f(x, \theta)$ is a parametric model then $\widehat{\theta}_\infty$ is a sample from the prior, and hence the characterization of the prior. This is indeed exactly how it is possible to characterize the prior under the assumption of an exchangeable sequence, see for example [8].

Here we demonstrate some key properties of the framework using an illustration. We sample X_1 from a standard normal density function and then construct $X_{2:M}$ for a suitably large M, using the copula model given by

$$F_{m+1}(x) = (1 - \alpha_m)\, F_m(x) + \alpha_m\, C_\rho(F_m(x), F_m(X_{m+1})), \tag{17}$$

where $X_{m+1} \sim F_m$, $\alpha_m = 1/m$ and

$$C_\rho(u, v) = \Phi\left(\frac{\Phi^{-1}(u) - \rho\, \Phi^{-1}(v)}{\sqrt{1 - \rho^2}}\right),$$

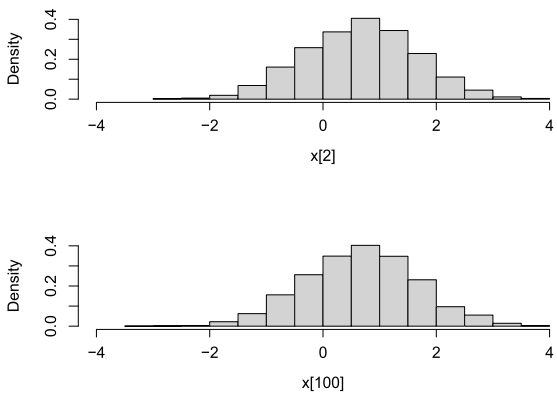

FIG 1. *Comparison of samples* $X[2]$ *and* $X[100]$.

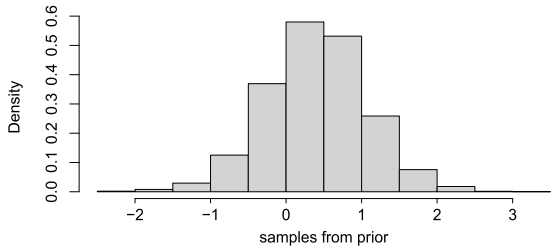

FIG 2. *Histogram of prior samples of sample means from c.i.d. sequences.*

taking $\rho = 0.95$.

Fig. 1 demonstrates the key feature of the c.i.d. sequence, namely, due to the martingale property of the sequence of $(F_m)_{m>1}$, the marginal distributions of X_l given X_1 is identical for all $l \geq 2$. In particular the figure shows this property for $l = 2$ and $l = 100$.

Samples from the prior are shown in Fig. 2. Each sample is obtained by running the c.i.d. sequence to $M = 100$ and then taking the MLE for the sequence which in this case is the sample mean.

In order to obtain the posterior samples, the same idea follows though now the c.i.d. sequence is started from X_{n+1} being taken from a normal model with mean the MLE from the data. The X_{n+1} for $l > 1$ are then taken according to the copula model, and the posterior samples are the means from each sequence, just as with the prior samples, the difference being that now the first n samples are the observations.

This paragraph has been pointed out by a referee: Suppose $P(X_{n+1} \in \cdot \big| X_{1:n})$ is the copula based predictive distribution. Then, this converges weakly a.s. to μ, since $X_{1:\infty}$ is c.i.d., but it may fail to converge in total variation a.s. In the second case, μ does not have a density even if the predictives are absolutely continuous with respect to the Lebesgue measure. In order that the predictives

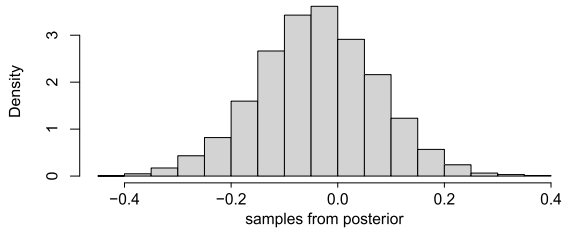

FIG 3. *Histogram of posterior samples of sample means from c.i.d. sequences.*

converges in total variation a.s., it is sufficient that $\alpha_n \to 0$ fast enough, which appears in (17). For instance, it is sufficient that $\sum_n \alpha_n < \infty$. See [11].

This represents the basic framework, the key components being the parametric model and the model for the c.i.d. sequence. The starting point for the c.i.d. sequence to characterize the posterior is started at the parametric model with the plug-in MLE.

In subsection 4.1 we consider a copula model for constructing the c.i.d. sequence, whereas in subsection 4.2 we use a normal based model, which includes a mixture of normal model.

4.1. Copula models

We characterize the c.i.d. sequence via non-data dependent copula; i.e. there exists a sequence of multivariate copula c_n such that

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)\, c_n(F(x_1), \ldots, F(x_n)),$$

where f is the marginal density, for all n. The difference with an exchangeable sequence is that the c_n are not symmetric in their arguments. So really we should write $c_n(u_1, \ldots, u_n) = c_{1,\ldots,n}(u_1, \ldots, u_n)$.

To get the update, we see that $f(x_i \mid x_1) = f(x_i)\, c_{1,i}(F(x_1), F(x_i))$ and so

$$f(x_2, \ldots, x_n \mid x_1) = \prod_{i=1}^{n} [f(x_i)\, c_{1,i}(F(x_1), F(x_i)]\, \frac{c_n(F(x_1), \ldots, F(x_n))}{\prod_{i=2}^{n} c_{1,i}(F(x_1), F(x_i))}.$$

Hence, the new conditional on x_1 copula are

$$\frac{c_n(F(x_1), \ldots, F(x_n))}{\prod_{i=2}^{n} c_{1,i}(F(x_1), F(x_i))}$$

and the new marginal densities are $f(x_i)c_{1,i}(F(x_1), F(x_i))$. Note then to characterize a c.i.d. sequence it must be that $c_{1,i}$ is the same for all i, so we can write it as $c_{1,\cdot}$.

This can be extended to the more general case $f(x_{n+1},\dots,x_m \mid x_1,\dots,x_n)$. First we find the new marginal density via

$$f(x_{n+1} \mid x_1,\dots,x_n) = f(x_{n+1})\frac{c_{1:n,\cdot}(F(x_1),\dots,F(x_n),F(x_{n+1}))}{c_{1:n}(F(x_1),\dots,F(x_n))}.$$

Therefore, the new marginal densities given $x_{1:n}$ are

$$f(x_i \mid x_1,\dots,x_n) = f(x_i)\frac{c_{1:n,\cdot}(F(x_1),\dots,F(x_n),F(x_i))}{c_{1:n}(F(x_1),\dots,F(x_n))},$$

for $i > n$, and the new copula for $x_{n+1},\dots,x_m$ is

$$\frac{c_{1:m}(F(x_1,\dots,F(x_n),F(x_{n+1}),\dots,F(x_m))}{\prod_{i=n+1}^{m} c_{1:n,\cdot}(F(x_1),\dots,F(x_n),F(x_i))}.$$

Hence, the prior is defined by the T from $f(x_1,\dots)$ and the posterior is defined by T from $f(x_{n+1},\dots \mid x_{1:n})$ and the implementation is done using the sequence of copula (c_i) with the constraints that $c_{1:m,i} = c_{1:m,j}$ for all $i,j > m$.

Here we demonstrate a straightforward illustration involving a normal mean model with known variance of 1. We generated $n = 50$ samples from a standard normal distribution and estimated the function with $\widehat{F}(x) = \Phi(x-\bar{x})$. The c.i.d. sequence $(x_m)_{m>n}$ is generated using the [9] scheme, i.e. $x_{m+1} \sim F_m$ and

$$F_{m+1}(x) = (1-\alpha_m)\,F_m(x) + H_m(F_m(x), F_m(x_{m+1}))$$

where

$$H_m(u,v) = \Phi\left(\frac{\Phi^{-1}(u) - \rho\Phi^{-1}(v)}{\sqrt{1-\rho^2}}\right)$$

for some ρ close to 1, which we took to be 0.95 and took the sequence $\alpha_m = 1/m$. This is of the form

$$f(x_i \mid x_{1:m}) = f_m(x_i) = f_{m-1}(x_i)\,c_m(F_{m-1}(x_i), F_{m-1}(x_m)),$$

for all $i > m$, which is a special case of the above, where we take the $c_m(u,v)$ to be a mixture of the independence copula and the Gaussian copula.

To get a sample from the posterior, we ran the martingale sequence (F_m) for 1000 iterations and then took the random parameter θ as the mean of the final random distribution. This was repeated 500 times to gather up a posterior sample and the histogram is presented in Fig. 4. The observed mean was 0.173 and the posterior sample mean was 0.177 with a sample variance of 0.016, which is close to $0.02 = 1/n$.

We repeat the experiment but now treating the variance also as unknown. The differences are now starting with the estimator $\widehat{F}(x) = \Phi((x-\bar{x})/S)$, where S is the sample variance. In Fig. 5 we plot the σ^2 posterior samples. The sample mean is 0.981, very close to 1, and the sample variance is 0.029, which is approximately the data sample variance divided by n.

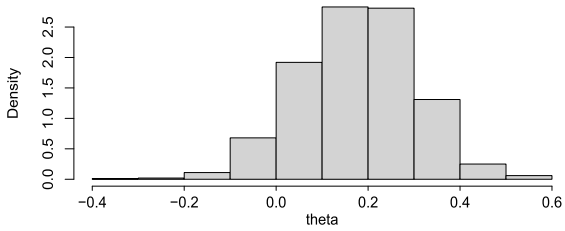

FIG 4. *Posterior samples of the means.*

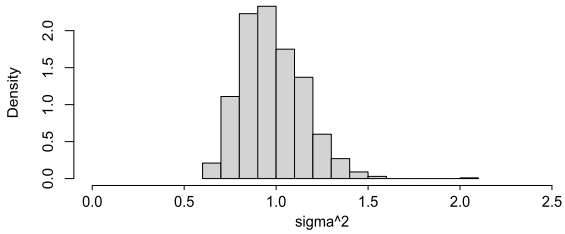

FIG 5. *Posterior samples of the variances.*

4.2. Normal models

In this section we construct c.i.d. sequences using the normal distribution. We start with a single normal component and then extend to a normal mixture model. The sequence (x_n) is constructed as follows. Take $x_1 \sim N(\theta_0, \sigma_0^2)$ for some constants θ_0 and $\sigma_0^2 > 0$ and then in general take

$$\theta_n = (1 - a_n)\,\theta_{n-1} + a_n\,x_n \quad \text{and} \quad \sigma_n^2 = \sigma_{n-1}^2(1 - a_n^2) \tag{18}$$

with

$$x_n = \theta_{n-1} + \sigma_{n-1}\,z_n, \tag{19}$$

where the (a_n) are a deterministic sequence in $(0, 1)$ and the (z_n) are an independent sequence of standard normal variables. It is easy to show that the sequence is c.i.d. To see this, we can write

$$x_{n+1} = (1 - a_n)\theta_{n-1} + a_n\,[\theta_{n-1} + \sigma_{n-1}\,z_n] + \sigma_n\,z_{n+1}$$

which becomes

$$x_{n+1} \sim \theta_{n-1} + \sqrt{\sigma_n^2 + a_n^2\,\sigma_{n-1}^2}\,z'$$

where z' is an independent standard normal variable.

Combining (18) and (19), we have that $\theta_n = \theta_{n-1} + a_n\sigma_{n-1}x_{n-1}$, so that:

$$\theta_n = \theta_0 + \sum_{j=1}^{n} a_j\sigma_{j-1}z_j$$

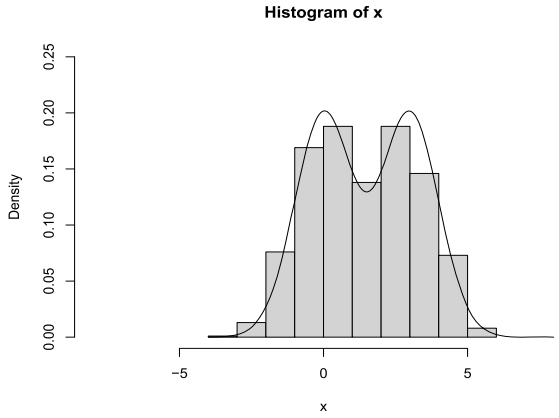

FIG 6. *c.i.d. sequence from normal mixture model.*

$$x_n = \theta_0 + \sum_{j=1}^{n-1} a_j \sigma_{j-1} z_j + \sigma_{n-1} z_n \tag{20}$$

Let θ_∞ be the almost sure limit of θ_n as n diverges to infinity, namely:

$$\theta_\infty = \theta_0 + \sum_{j=1}^{\infty} a_j \sigma_{j-1} z_j$$

The construction (20) is a special case of Example 1.2 of [6]. As proved by [6], x_n converges in probability if and only if $\sum_{n=1}^{\infty} a_n^2 = \infty$ and in this case, $\theta_\infty = \lim_{n\to\infty} x_n$. Moreover for x_n to converge almost surely it is sufficient that $\sum_{n=1}^{\infty} \prod_{j=1}^{n} (1 - a_j^2)^r < \infty$ for some $r > 0$.

The prior is the distribution of θ_∞, which is normal, and the mean is θ_0 with variance

$$\mathrm{Var}\,(\theta_\infty) = \sum_{n=1}^{\infty} a_n^2\, \sigma_{n-1}^2.$$

The posterior, given an observed sample $(x_{1:n})$, is also easy to derive, the mean for θ_∞ is now θ_n and the variance is $\sum_{m=n}^{\infty} a_{m+1}^2 \sigma_m^2$.

A special case arises by taking

$$b_{m+1}(1 - b_m) \cdots (1 - b_2) = \frac{1}{m(m+1)},$$

where $b_m = a_m^2$, for $m \geq 2$, since then $\mathrm{Var}\,(\theta_\infty \mid x_{1:n}) = \sigma_0^2/n$. We can ensure this by taking $b_m = 1/m$.

The normal model can be extended to cover a normal mixture model, of the form

$$f(x) = \sum_{j=1}^{k} w_j\, N(x \mid \theta_j, \sigma^2),$$

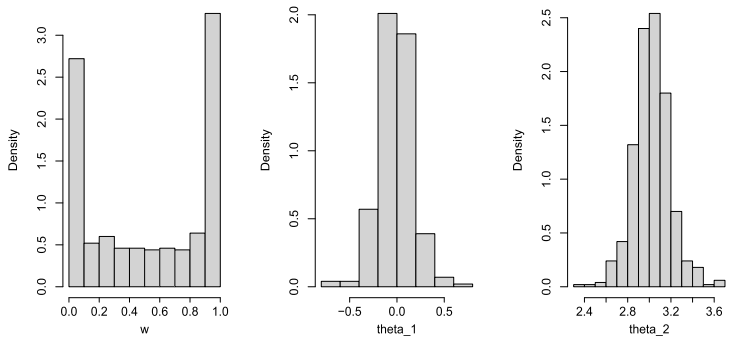

FIG 7. *Posterior density functions for mixture model parameters.*

where k is fixed and finite. The x_{m+1} is taken from this mixture model with parameters $(w_m, \theta_m, \sigma_m^2)$, and the updates of the parameters are

$$w_{j,m+1} = \frac{w_{j,m} N(x_{m+1}|\theta_{j,m}, \sigma_m^2)}{\sum_{j=1}^{k} w_{j,m} N(x_{m+1}|\theta_{j,m}, \sigma_m^2)}$$

with

$$\theta_{j,m+1} = (1 - a_m)\theta_{j,m} + a_m\, x_{m+1} \quad \text{and} \quad \sigma_{m+1}^2 = (1 - a_m^2)\sigma_m^2.$$

This will also generate a c.i.d. sequence if we only update the θ component from which the x was taken. That is, if $P(d_m = j \mid w_m) = w_{j,m}$ then

$$x_{m+1} = \theta_{d_m,m} + \sigma_m\, z \quad \text{and} \quad \theta_{d_m,m} = (1 - a_m)\theta_{d_m,m-1} + a_m\, x_m$$

with all the other $\theta_{\cdot,m}$ remaining unchanged.

For example, the starting density function for x_1 is a two component normal mixture, one with mean 0 and variance 1 and the other with mean 3 and variance 1, with weight 1/2. Fig. 6 shows a histogram of the samples (x_m) for $m = 1, \ldots, 1000$. Fig. 7 shows samples, 500 in each case, from the posterior density functions for the parameters w, θ_1 and θ_2. These are obtained by running 500 independent c.i.d. sequences. The idea here is that the data provides estimators $\widehat{\theta}_1 = 0$ and $\widehat{\theta}_2 = 3$ and $\widehat{w} = 1/2$ and this illustration merely shows how the procedure works in practice.

5. Proof of Theorem 3.2

A few preliminary results are needed in order to prove Theorem 3.2. The following Lemma shows that a maximizer for a strictly concave function is the unique global maximizer and that the maximum is well separated.

Lemma 5.1. *Let A be a convex closed subset of $\mathbb{R}^d$ and let $g : A \to \mathbb{R}$ be a strictly concave and continuous function that has a maximum at $x_0 \in A$. Then, x_0 is the unique global maximizer for g and, moreover, for every $\epsilon > 0$ such that*

$$B(x_0, \epsilon) = \{x \in A : ||x - x_0|| < \epsilon\},$$

we have that:

$$\sup_{x \in A \setminus B(x_0, \epsilon)} g(x) < g(x_0), \tag{21}$$

where we convene that $\sup_{x \in \emptyset} g(x) = -\infty$. Note that $B(x_0, \epsilon) \neq \emptyset$ for every $\epsilon > 0$ since $x_0 \in B(x_0, \epsilon)$ for all $\epsilon > 0$.

Proof. In order to show that x_0 is the unique global maximizer, assume that $x_1 \neq x_0$ is such that $g(x_1) \geq g(x_0)$. Then for every $0 < t < 1$, by strict concavity,

$$g(x_0 + t(x_1 - x_0)) = g((1-t)x_0 + t\,x_1) > (1-t)g(x_0) + tg(x_1) \geq g(x_0)$$

contradicting local maximality.

For any $x_1 \in A \setminus B(x_0, \epsilon)$, we have $g(x_1) < g(x_0)$. Let

$$x_2 = x_0 + \frac{\epsilon}{||x_1 - x_0||}(x_1 - x_0) = \frac{\epsilon}{||x_1 - x_0||}x_1 + \left(1 - \frac{\epsilon}{||x_1 - x_0||}\right)x_0.$$

Therefore, being $0 < \epsilon / \, ||x - x_0|| \leq 1$, by concavity,

$$g(x_2) \geq \frac{\epsilon}{||x_1 - x_0||}g(x_1) + \left(1 - \frac{\epsilon}{||x_1 - x_0||}\right)g(x_0) \geq g(x_1).$$

Clearly, $||x_2 - x_0|| = \epsilon$, namely

$$x_2 \in \{x \in A : ||x - x_0|| = \epsilon\} \subset A \setminus B(x_0, \epsilon) = \{x \in A : ||x - x_0|| \geq \epsilon\},$$

and therefore

$$\sup_{x \in A \setminus B(x_0, \epsilon)} g(x) = \sup_{x \in A:\, ||x - x_0|| = \epsilon} g(x).$$

Since the set $\{x \in A : ||x - x_0|| = \epsilon\}$ is compact, the continuous function g achieves its supremum on such set and therefore:

$$\sup_{x \in A:\, ||x - x_0|| = \epsilon} g(x) < g(x_0). \qquad \square$$

Lemma 5.2. *For all $\theta \in \Theta$, and for a fixed but arbitrary $\theta_0 \in \Theta$, and for every ω in Ω, let*

$$G(\theta)(\omega) = \int_{\mathbb{R}} \{\log f(x, \theta) - \log f(x, \theta_0)\}\, \gamma(\mathrm{dx})(\omega),$$

where $\gamma = \gamma(\cdot)(\omega)$ is a random probability measure such that $\mathbb{E}(\gamma(B)) = P(X_1 \in B)$ for every Borelian subset B of $\mathbb{R}$.

If the hypotheses 1–8 of Theorem 3.2 hold true, then there uniquely exists a random variable U such that

$$G(U) = \sup_{\theta \in \Theta} G(\theta).$$

Proof. For every $\omega \in \Omega$, the function $\theta \to G(\theta)(\omega)$ is strictly concave since the function $\theta \to \log f(x,\theta)$ is such (for every $x \in \mathbb{R}$). Being the function $\theta \to \log f(x,\theta)$ continuous for every $x \in \mathbb{R}$, by condition 7, one can verify by Lebesgue's Dominated Convergence Theorem that the function $\theta \to G(\theta)(\omega)$ is also continuous, for every $\omega \in \Omega$. Therefore, if Θ is compact, then by the Extreme Value Theorem, for every $\omega \in \Omega$, there exists $U(\omega)$ belonging to Θ that maximizes $\theta \to G(\theta)(\omega)$.

If Θ is not compact, but closed and unbounded, then one can proceed as follows. Let

$$K(x) = \sup_{\theta' \in \Theta^* : ||\theta'|| > \rho} \log f(X_1, \theta') - \log f(X_1, \theta_0)$$

observing that $K(\cdot)$, by (13) and (15), is integrable with respect to γ almost surely. By (12), and Fatou's Lemma, we have that

$$\begin{aligned}
&\liminf_{||\theta|| \to +\infty} \int_{\mathbb{R}} K(x)\gamma(\mathrm{dx}) - \mathrm{G}(\theta) \\
&\qquad = \liminf_{||\theta|| \to +\infty} \int_{\mathbb{R}} \{K(x) - \log f(x,\theta) + \log f(x,\theta_0)\}\, \gamma(\mathrm{dx}) \\
&\qquad \geq \int_{\mathbb{R}} \liminf_{||\theta|| \to +\infty} \{K(x) - \log f(x,\theta) + \log f(x,\theta_0)\} \gamma(\mathrm{dx}) = +\infty,
\end{aligned}$$

which implies that

$$\limsup_{||\theta|| \to +\infty} G(\theta)(\omega) = -\infty, \tag{22}$$

for every $\omega \in \Omega$. At this stage, just pick any $\theta_0 \in \Theta$ and let $A = \{\theta \in \Theta : G(\theta) \geq G(\theta_0)\}$. A is closed being G continuous and is bounded by (22). So, A is compact. By continuity, G attains a maximum on A. By definition of A, a maximizer of G on A is also a global maximizer. Since $G(\cdot)$ is strictly concave and defined on the convex set Θ, $U(\omega)$ must be the unique maximizer by Lemma 5.1.

In order to prove that we are dealing with a random variable, we need to show that $\omega \to U(\omega)$ is measurable. To this aim, recall that Θ is separable being a subset of $\mathbb{R}^d$ and therefore there exists a countable dense subset $\{q_1, q_2, \dots\}$ of Θ.

Let $F_k = \{q_1, \dots, q_k\}$ and let $U_k(\omega)$ be the maximizer of $G(\cdot)(\omega)$ on F_k with largest subscript. In other words, $U_k(\omega) = q_l$ if $l \in \{1, \dots, k\}$, $G(q_l)(\omega) \geq G(q_j)(\omega)$ for every $j \in \{1, \dots, k\}$ and $j < l$ whenever $G(q_j)(\omega) = G(q_l)(\omega)$. Being F_k a finite set, U_k is measurable and so is $\lim_{k\to\infty} U_k$. We shall now prove that $U = \lim_{k\to\infty} U_k$ so that U turns out to be measurable. To this aim, we show first that $G(U)(\omega) = \lim_{k\to\infty} G(U_k)(\omega) = \sup_{k\geq 1} G(U_k)$. Clearly, $G(U_k) \leq G(U)$, for every $k \geq 1$, so that $\sup_{k\geq 1} G(U_k) \leq G(U)$, so that we need to prove that $\sup_{k\geq 1} G(U_k) \geq G(U)$. To this aim, fix $\omega \in \Omega$, and let $(x_1, x_2, \dots)$ be a sequence of elements of $\{q_1, q_2, \dots\}$ such that x_k converges to $U = U(\omega)$. By continuity of $G(\cdot)$, we have that $G(x_k)$ converges to $G(U)$. Hence,

for every $\epsilon > 0$, there exists k such that $G(x_k, \theta_0) > G(U, \theta_0) - \epsilon$. At this stage, let j_0 be such that $x_k \in F_{j_0}$ and therefore $x_k \in F_j$ for every $j \geq j_0$ so that $G(U_j) \geq G(x_k) > G(U) - \epsilon$ for every $j \geq j_0$. So, $G(U)(\omega) = \lim_{k\to\infty} G(U_k)(\omega)$. This implies that $\lim_{k\to\infty} U_k = U$. Indeed, if that was not true, then there would exists $\epsilon > 0$ such that $||U_k - U|| > \epsilon$ eventually. By Lemma 5.1, this would imply that

$$\sup_{k\geq 1} G(U_k) \leq \sup_{x:||x-U||>\epsilon} G(x) < G(U)$$

and letting

$$\delta = G(U) - \sup_{x:||x-U||>\epsilon} G(x)$$

we would have that $G(U) - G(U_k) \geq \delta$ for every $k \geq 1$, namely that $G(U_k)$ does not converge to $G(U)$. □

Proposition 3. Assume the hypotheses 1–8 of Theorem 3.2 hold true, and set for every $\theta \in \Theta$,

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{\log f(X_i, \theta) - \log f(X_i, \theta_0)\}, \tag{23}$$

$$M(\theta) = \int_{\mathbb{R}} \{\log f(x, \theta) - \log f(x, \theta_0)\}\, \mu(\mathrm{dx}). \tag{24}$$

If S is a closed subset of Θ and $\theta_0 \in \Theta$, then:

$$\limsup_{n\to+\infty} \sup_{\theta\in S} M_n(\theta) \leq \sup_{\theta\in S} M(\theta),$$

almost surely.

Proof. Consider the function

$$\varphi(x, \theta, \theta_0, \rho) = \sup_{\theta'\in S:||\theta'-\theta||<\rho} \{\log f(x, \theta') - \log f(x, \theta_0)\},$$

where $\theta, \theta_0 \in S$, $x \in \mathbb{R}$ and $\rho > 0$. Exploiting again separability of S and continuity of $x \to \log f(x, \theta)$, one can show that $x \to \varphi(x, \theta, \theta_0, \rho)$ is measurable. Indeed, one can consider again a dense countable subset $\{q_1, q_2, \dots\}$ of S, let $F_k = \{q_1, \dots, q_k\}$ and verify that

$$\varphi(x, \theta, \theta_0, \rho) = \sup_{k\geq 1} \sup_{\theta'\in F_k:||\theta'-\theta||<\rho} \{\log f(x, \theta') - \log f(x, \theta_0)\}.$$

So φ is measurable in x and by condition 7 is integrable for some $\rho > 0$ with respect to the distribution of X_1 (or Y). Moreover, it is almost surely integrable with respect to the random measure μ, which is the conditional distribution of Y given $X_{1:\infty}$. Therefore, by Beppo Levi's Monotone Convergence Theorem, for

every θ there is an event B_θ such that $\mathrm{P}(B_\theta) = 1$ and for every $\omega \in B_\theta$,

$$\begin{aligned}\lim_{\rho \downarrow 0} &\int_{\mathbb{R}} \varphi(x, \theta, \theta_0, \rho)\, \mu(\mathrm{dx})(\omega) \\ &= \int_{\mathbb{R}} \{\log f(x, \theta) - \log f(x, \theta_0)\}\, \mu(\mathrm{dx})(\omega) \\ &= M(\theta)(\omega).\end{aligned}$$

Let $\epsilon > 0$. For each θ, find ρ_θ so that for every $\omega \in B_\theta$,

$$\int_{\mathbb{R}} \varphi(x, \theta, \theta_0, \rho_\theta)\, \mu(\mathrm{dx})(\omega) < \mathrm{M}(\theta)(\omega) + \epsilon. \tag{25}$$

At this stage, if S is unbounded, consider the function

$$\tilde{\varphi}(x, \theta', \delta) = \sup_{\theta \in S : ||\theta|| > \delta} \{\log f(x, \theta) - \log f(x, \theta')\},$$

where $\theta' \in S$, $x \in \mathbb{R}$ and $\delta > 0$. The function $\tilde{\varphi}$ can be shown to be measurable similarly to φ. Moreover, for some $\delta > 0$, it is integrable with respect to μ almost surely thanks to conditions 6 and 8. By the Strong Law of Large Numbers Theorem 2.2, for every $\theta, \theta' \in S$ such that $||\theta|| > \delta$ and $||\theta'|| > \delta$,

$$\begin{aligned}\limsup_{n \to \infty} &\sup_{\theta \in S : ||\theta|| > \delta} M_n(\theta) - M_n(\theta') \\ &= \limsup_{n \to \infty} \sup_{\theta \in S : ||\theta|| > \delta} \frac{1}{n} \sum_{i=1}^{n} \{\log f(X_i, \theta) - \log f(X_i, \theta')\} \\ &\leq \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}(X_i, \theta', \delta) \\ &= \int_{\mathbb{R}} \tilde{\varphi}(x, \theta', \delta)\, \mu(\mathrm{dx}).\end{aligned} \tag{26}$$

By (12), we have that $\lim_{\delta \to +\infty} \tilde{\varphi}(x, \theta', \delta) = -\infty$ for every x belonging to a set with positive μ–measure and therefore by Beppo Levi's Monotone Convergence Theorem,

$$\lim_{\delta \to +\infty} \int_{\mathbb{R}} \tilde{\varphi}(x, \theta', \delta)\, \mu(\mathrm{dx}) = -\infty.$$

Hence, the quantity (26) is negative for some large δ and therefore, using again the Strong Law of Large Numbers Theorem 2.2 so that $\lim_{n \to \infty} M_n(\theta') = M(\theta')$ almost surely, we have that:

$$\limsup_{n \to \infty} \sup_{\theta \in S : ||\theta|| > \delta_0} M_n(\theta) < \sup_{\theta' \in S : ||\theta'|| > \delta_0} M(\theta') \tag{27}$$

for some large δ_0.

At this stage, let $S_0 = \{\theta \in S : ||\theta|| \leq \delta_0\}$. Being S closed, S_0 is compact. The spheres $B(\theta, \rho_\theta) = \{\theta' : ||\theta' - \theta|| < \rho_\theta\}$ cover S_0 so that by compactness

of S_0 there exists a finite subcover, say, $S_0 \subset \cup_{j=1}^m B(\theta_j, \rho_{\theta_j})$. For each $\theta \in S_0$ there exists an index j such that $\theta \in B(\theta_j, \rho_{\theta_j})$. From the definition of φ, $\log f(x,\theta) - \log f(x,\theta_0) \leq \varphi(x, \theta_j, \rho_{\theta_j})$, for all $x \in \mathbb{R}$. Hence,

$$\begin{aligned} M_n(\theta) &= \frac{1}{n}\sum_{i=1}^n \{\log f(X_i,\theta) - \log f(X_i,\theta_0)\} \\ &\leq \frac{1}{n}\sum_{i=1}^n \varphi(X_i, \theta_j, \theta_0, \rho_{\theta_j}). \end{aligned}$$

So that

$$\sup_{\theta\in S_0} M_n(\theta) \leq \sup_{j=1,\ldots,m} \frac{1}{n}\sum_{i=1}^n \varphi(X_i, \theta_j, \theta_0, \rho_{\theta_j}). \tag{28}$$

By the Strong Law of Large Numbers Theorem 2.2, the following holds almost surely for $j = 1, \ldots, m$:

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \varphi(X_i, \theta_j, \theta_0, \rho_{\theta_j}) = \int_{\mathbb{R}} \varphi(x, \theta_j, \theta_0, \rho_{\theta_j})\, \mu(\mathrm{dx}),$$

which by (25) yields:

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \varphi(X_i, \theta_j, \theta_0, \rho_{\theta_j}) \leq M(\theta_j) + \epsilon,$$

almost surely for $j = 1, \ldots, m$, which implies:

$$\limsup_{n\to+\infty} \sup_{j=1,\ldots,m} \frac{1}{n}\sum_{i=1}^n \varphi(X_i, \theta_j, \theta_0, \rho_{\theta_j}) \leq \sup_{j=1,\ldots,m} M(\theta_j) + \epsilon,$$

almost surely, which in turn by (28) yields that:

$$\limsup_{n\to+\infty} \sup_{\theta\in S_0} M_n(\theta) \leq \sup_{\theta\in S_0} M(\theta) + \epsilon,$$

almost surely. Because it is true for all $\epsilon > 0$, it is true for $\epsilon = 0$, namely:

$$\limsup_{n\to+\infty} \sup_{\theta\in S_0} M_n(\theta) \leq \sup_{\theta\in S_0} M(\theta). \tag{29}$$

Combining (29) with (27), the thesis is obtained. □

At this stage, we are in position to prove Theorem 3.2, whose proof is based on Wald's classical proof of strong consistency of the maximum likelihood estimator in the i.i.d. setting [18].

Proof of Theorem 3.2. Consider M and M_n as defined by (23) and (24) in the statement of Proposition 3. At this stage, consider G appearing in Lemma 5.2. If γ is the empirical measure μ_n then $G = M_n$ by (23), and if instead $\gamma = \mu$

then $G = M$ by (24). Therefore, both 1 and 2 are just a consequence of Lemma 5.2.

At this stage, our goal is to prove 3. To this aim, for every $\omega \in \Omega$, and every $k = 1, 2, \ldots$ let $R_k = R_k(\omega)$ be such that

$$\sup_{\theta : ||\theta - T|| \geq R_k} M(\theta) \leq \inf_{\theta : ||\theta - T|| \leq 1/k} M(\theta) - 1/k, \tag{30}$$

and $R_k \downarrow 0$ as $k \to \infty$. Such sequence R_k exists due to (strict) concavity of M.

At this stage, fix $\rho > 0$ and let $k \geq 1$ be an integer such that $R_k + 1/k < \rho$. Moreover, consider a countable or finite covering for Θ made of balls of radius $1/k$ and let $\tilde{\Theta}$ be the set of centers of such balls. One can consider an at most countable collection of disjoint sets $D_{\theta'} \subset B(\theta', 1/k)$, being $\theta' \in \tilde{\Theta}$, which is a partition of Θ, i.e.

$$\Theta = \bigcup_{\theta' \in \tilde{\Theta}} D_{\theta'}.$$

At this stage, let $S_{\theta'} = \{\theta \in \Theta : ||\theta - \theta'|| \geq \rho\}$. By Proposition 3, we have that $\mathrm{P}(E_{\theta'}) = 1$ if

$$E_{\theta'} = \left\{ \limsup_{n \to +\infty} \sup_{\theta \in S_{\theta'}} M_n(\theta) - M_n(\theta') \leq \sup_{\theta \in S_{\theta'}} M(\theta) - M(\theta') \right\},$$

and therefore:

$$\begin{aligned} 1 &= \sum_{\theta' \in \tilde{\Theta}} P(T \in D_{\theta'}) \\ &= \sum_{\theta' \in \tilde{\Theta}} \mathrm{P}(\{T \in D_{\theta'}\} \cap E_{\theta'}). \end{aligned} \tag{31}$$

If θ belongs to $S_{\theta'}$, then

$$||\theta - T|| \geq ||\theta - \theta'|| - ||T - \theta'|| \geq \rho - 1/k > R_k.$$

and by (30) if $T \in D_{\theta'}$ then $M(\theta) < M(\theta')$. Reasoning as in the proof of Lemma 5.2, the function $M(\cdot)(\omega)$ achieves its maximum value on $S_{\theta'}$. Let $W(\omega) = \sup_{\theta \in S_{\theta'}} M(\theta) - M(\theta')$, then $W(\omega) < 0$ and (31) yields:

$$\begin{aligned} 1 &= \sum_{\theta' \in \tilde{\Theta}} \mathrm{P}\left(T \in D_{\theta'},\ \limsup_{n \to +\infty} \sup_{\theta \in S_{\theta'}} M_n(\theta) - M_n(\theta') \leq W \right) \\ &= \sum_{\theta' \in \tilde{\Theta}} \mathrm{P}\left(T \in D_{\theta'},\ \exists n_0 : \forall n > n_0,\ \sup_{\theta \in S_{\theta'}} M_n(\theta) - M_n(\theta') < 0 \right). \end{aligned} \tag{32}$$

Being $M_n(\hat{\theta}_n) \geq M_n(\theta')$, (32) yields:

$$\begin{aligned}
1 &= \sum_{\theta' \in \tilde{\Theta}} \mathrm{P}\left(T \in D_{\theta'},\ \exists n_0 : \forall n > n_0, \hat{\theta}_n \notin S_{\theta'}\right) \\
&= \sum_{\theta' \in \tilde{\Theta}} \mathrm{P}\left(T \in D_{\theta'},\ \exists n_0 : \forall n > n_0, ||\hat{\theta}_n - \theta'|| < \rho\right) \\
&\leq \sum_{\theta' \in \tilde{\Theta}} \mathrm{P}\left(T \in D_{\theta'},\ \exists n_0 : \forall n > n_0, ||\hat{\theta}_n - T|| < \rho + 1/k\right) \\
&= \mathrm{P}\left(\bigcup_{\theta' \in \tilde{\Theta}} \{T \in D_{\theta'}\} \cap \{\exists n_0 : \forall n > n_0, ||\hat{\theta}_n - T|| < \rho + 1/k\}\right) \\
&= \mathrm{P}\left(\exists n_0 : \forall n > n_0, ||\hat{\theta}_n - T|| < \rho + 1/k\right).
\end{aligned}$$

Since ρ and $1/k$ can be arbitrarily small the proof is complete. □

Acknowledgments

The authors are grateful for the comments and suggestions of two referees on an earlier version of the paper.

References

[1] Bassetti, F., Crimaldi, I. and Leisen, F. (2010). Conditionally identically distributed species sampling sequences. *Advances in Applied Probability* **42** 433–459. MR2675111

[2] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Analysis*. Wiley.

[3] Berti, P., Dreassi, E., Leisen, F., Pratelli, L. and Rigo, P. (2023). Bayesian predictive inference without a prior. *Statistica Sinica* **33** 2405–2429. MR4647040

[4] Berti, P., Dreassi, E., Pratelli, L. and Rigo, P. (2021). Asymptotics of certain conditionally identically distributed sequences. *Statistics and Probability Letters* **168** 108923. MR4149322

[5] Berti, P., Dreassi, E., Pratelli, L. and Rigo, P. (2021). A class of models for Bayesian predictive inference. *Bernoulli* **27** 702–726. MR4177386

[6] Berti, P., Pratelli, L. and Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *The Annals of Probability* **32** 2029–2052. MR2073184

[7] de Finetti, B. (1930). Funzione caratteristica di un fenomeno aleatorio. *Mem. Acad. Naz. Lincei* **4** 86–133.

[8] Doob, J. L. (1949). Application of the theory of martingales. *Actes du Colloque International des probabilites et ses applications* 23–27. MR0033460

[9] Fong, E., Holmes, C. and Walker, S. G. (2023). Martingale posterior distributions. *Journal of the Royal Statistical Society Series B*. in press. MR4726953
[10] Fortini, S., Petrone, S. and Sporysheva, P. (2018). On a notion of partially conditionally identically distributed sequences. *Stochastic Processes and their Applications* **128** 819–846. MR3758339
[11] Garelli, S., Leisen, F., Pratelli, L. and Rigo, P. (2024). Asymptotics of predictive distributions driven by sample means and variances. *arxiv.org/abs/2403.16828*.
[12] Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80** 470–501. MR0076206
[13] Hjort, N., Holmes, C., Mueller, P. and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press. MR2722987
[14] Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science & Business Media, New York. MR2197664
[15] Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica* **59** 1161–1167. MR1113551
[16] Schervish, M. J. (1995). *Theory of Statistics*. Springer: New York, USA. MR1354146
[17] Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris* **8** 229–231. MR0125600
[18] Wald, A. (1949). Note on the consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics* **20** 595–601. MR0032169
[19] White, H. (1982). Maximum likelihood estimation with misspecified models. *Econometrica* **50** 1–25. MR0640163