ABSTRACT. This paper describes how to specify probability models for data analysis via a backward induction procedure. The new approach yields coherent, prior-free uncertainty assessment. The backward induction approach is first demonstrated on two familiar models — the Bernoulli distribution and the Gaussian distribution — to compare the resulting specifications to their standard counterparts arising as Bayesian posterior distributions. The new approach is then applied to a kernel density estimator, which leads to a novel method for computing point-wise credible intervals in nonparametric density estimation.

1. INTRODUCTION

This paper presents a new approach to specifying probability models for data analysis, providing coherent, prior-free uncertainty assessment. The approach begins by assuming a sample size sufficiency condition; for some large $N$,

i) all $y_j$, for $j > N$, are independent and identically distributed with density function $p_N(y) \equiv p(y \mid y_{1:N})$ depending only on the sample $y_{1:N}$.

From this assumption, a sequence of predictive distributions is derived which satisfy a temporal coherence condition [Goldstein, 1983, Zabell, 2002, Parmigiani and Inoue, 2009]:

ii) For $p_t(y) \equiv p(y \mid y_{1:t})$,

$$p_t(y) = \int p_{t+1}(y \mid y_{t+1})p_t(y_{t+1})dy_{t+1},$$

for $0 < t < N$. 

1
With this coherent sequence of predictive distributions in hand, uncertainty intervals can be calculated via sequential forward simulation, starting from $p_n(y)$, based on an observed sample $y_{1:n}$.

Section 2 describes how uncertainty assessment can be performed without reference to a prior distribution, by working directly from a sequence of predictive distributions. Section 3 then describes how to derive a coherent sequence of predictive distributions by working backward from a specified $p_N(y)$. The working details of this approach are illustrated via two small examples and compared to the usual Bayesian posterior. Section 4 applies the method to a kernel density estimator, leading to an efficient method for producing point-wise credible intervals of an unknown density function.

2. Uncertainty assessment via sequential forward simulation

Although contemporary Bayesian statistics works predominately with probability models specified in terms of priors and likelihoods, it is possible to conduct posterior inference working directly with joint distributions on observables, a la de Finetti [de Finetti, 1974, 1975]. Recall the compositional representation of a joint distribution

$$p(y_{1:n}) = p_0(y_1)p_1(y_2 \mid y_1)p_2(y_3 \mid y_1:y_2)\ldots p_{n-1}(y_n \mid y_{1:(n-1)})$$

Posterior distributions can be derived from this sequence of predictive distributions, via forward simulation, as follows. First, with past (observed) data $y_{1:n}$ in hand, simulate $y^*_{n+1}$ from $p_n(y \mid y_{1:n})$. Then simulate $y^*_{n+2}$ from $p_{n+1}(y \mid y_{1:n}, y^*_{n+1})$, and then $y^*_{n+3}$ from $p_{n+2}(y \mid y_{1:n}, y^*_{(n+1):(n+2)})$, etc. Continue this process, sequentially
simulating a total of $m$ hypothetical future observations, arriving finally at distribution

\[(3)\quad p_N(y \mid y_{1:n}, y^*_N),\]

where $N = n + m$. From this distant-future predictive distribution, extract any summary of interest from $p_N(y \mid y_{1:n}, y^*_N)$; call it $\theta \equiv g[p_N]$. Typical choices for $g[\cdot]$ might be a mean, a quantile, a high density region or even the entire density function. Repeating this process, one performs a Monte Carlo integration over hypothetical future data realizations; each $\theta^{(j)}$ denoting property $g[\cdot]$ of a different $m$-step ahead posterior predictive distribution, corresponding to the $j$th simulated realization of future data $y^*_N$. The distant-future quantity $\theta$ is uncertain precisely because many different future realizations are possible.

Taking $N \to \infty$ makes the connection with the usual approach. A model parameter $\theta$ can be thought of as a functional $g[\cdot]$ of the posterior predictive distribution $p_N(y) \equiv p(y \mid y_{1:N})$ as $N \to \infty$ so that

\[(4)\quad \theta \equiv g[p_\infty(y)].\]

That is, supposing that $p(y_1, \ldots, y_\infty)$ is stipulated, $\theta$ simply picks off some feature of the conditional distribution of one element, given an infinite amount of past data.
Figure 1. Gray lines depict 20 simulated data sequences from the prior predictive; they terminate 1000 steps in the future at points that are uniformly distributed in the interval. Solid lines show 20 simulated data sequences beginning from the point \( n = 10 \) with an observed sample average of 0.7; restricting to sequences that run through the point \((10, 0.7)\) yields sample paths that terminate in a more concentrated region.

*Example: Bernoulli likelihood.* Suppose \( Y_i \sim \text{Bernoulli}(\theta) \) with prior \( \theta \sim \text{Uniform}(\alpha, \beta) \).

Integrating over this prior yields the following predictive updates

\[
p_t(y_{t+1} \mid y_{1:t}) = \text{Bernoulli}\left(\frac{\alpha_t}{\alpha_t + \beta_t}\right),
\]

\[
\alpha_t = \alpha_{t-1} + y_t,
\]

\[
\beta_t = \beta_{t-1} + 1 - y_t.
\]

Now, suppose \( n = 10 \) observations are observed, and that seven of them are ones: \( \sum_{i=1}^{n} y_i = 7 \). Figure 1 shows simulated predictive sequences 1000 steps into the future from the prior and from the posterior. Figure 2 shows that repeating this exercise 5000 times recapitulates the known Beta(8, 4) posterior distribution nicely.
Figure 2. The histogram at $t = 1000$ for 5000 simulated posterior predictive data sequences for $n = 10$, $\bar{y}_n = 0.7$; it nicely recapitulates the known Beta$(8, 4)$ posterior distribution for $\theta$, which is shown overlaid in black.

**Example: Gaussian likelihood with known variance.** Suppose $Y_i \sim N(\theta, 1)$ with prior $\theta \sim N(\mu_0, \phi_0^{-1})$. Integrating over this prior yields the following predictive updates

$$p_t(y_{t+1} \mid y_{1:t}) = N(\mu_t, 1 + 1/\phi_t),$$

$$\mu_t = \frac{y_t + \mu_{t-1}\phi_{t-1}}{1 + \phi_{t-1}},$$

$$\phi_t = 1 + \phi_{t-1}. \quad (6)$$

Forward simulation yields (approximate) posterior distributions over $\theta \equiv \bar{y}_N$, as in the Bernoulli example above and similarly recapitulates, as expected, the usual Bayesian posterior.

3. **Prior-free model specification via backward induction**

It is possible to determine the sequence in (2) not by integrating a specified likelihood over a specified prior distribution, but by iteratively solving for each term in
the product by directly enforcing (1), starting from $p_N(y)$ and working backward. This section works through this approach on two small examples. The next section uses the backward induction approach to derive a new method for nonparametric density estimation.

**Example: Bernoulli likelihood.** Assume that for a sample of size $N$ and $\bar{y} = N^{-1} \sum_i y_i$, a sufficiently accurate predictive distribution for $Y_{N+1}$ is Bernoulli($\bar{y}$). Write $\pi_t = \Pr(Y = 1 \mid y_{1:t})$ and $\pi_t(z) = \Pr(Y = 1 \mid Y_t = z)$. Plugging these definitions directly into (1) gives

$$\pi_{N-1} = \pi_N(1)\pi_{N-1} + \pi_N(0)(1 - \pi_{N-1})$$

$$= \left( \frac{N-1}{N} \bar{y}_{N-1} + \frac{1}{N} \right) \pi_{N-1} + \frac{N-1}{N} \bar{y}_{N-1}(1 - \pi_{N-1}),$$

$$= \bar{y}_{N-1}.$$

Repeating the same argument shows that the coherent predictive sequences use the current sample average at time $t$ as the prediction probability for observation $t + 1$.

Simulation from this sequence, as described in Section 2, yields a posterior distribution over $\theta \equiv \bar{y}_N$. Note that to duplicate the Bayesian solution demonstrated in the previous section, one can “seed” the backward induction procedure with two pseudo-observations, one of which is a one and the other a zero.

**Example: Gaussian distribution with known variance.** Assume that $Y_{N+1} \sim N(\bar{y}_N, 1)$ for a large fixed $N$. Equivalently, $Y_{N+1} = \bar{y}_N + \epsilon_N$ for $\epsilon_N \sim N(0, 1)$, or in terms of the random variable $Y_N$, $Y_{N+1} = \frac{1}{N} Y_N + \frac{N-1}{N} \bar{y}_{N-1} + \epsilon_N$. Because the sum of two Gaussians is again Gaussian, it is only necessary to find a Gaussian distribution for $Y_N$ that satisfies the above. Therefore, solving for the mean and variance gives
Figure 3. At $N = 20$ the predictive variance decays at a faster rate than the standard Bayesian model. By $N = 100$, the difference in the decay rates is nearly imperceptible. The horizontal axis of the second panel runs only to 20, rather than 100, for better visual comparison.

\[
EY_N = \bar{y}_{N-1} \left( \frac{N-1}{N} \right) + \frac{1}{N} EY_N + E\epsilon_N \implies EY_N = \bar{y}_{N-1}
\]

(8)

\[
VY_N = \frac{VY_N}{N^2} + V\epsilon_N \implies VY_N = \frac{N^2}{N^2 - 1} V\epsilon_N.
\]

Noting that $VY_N = \frac{N^2}{N^2 - 1} V\epsilon_N$ defines a recursion, one can compute

(9) \[
VY_t = \prod_{t+1 \leq j \leq N} \frac{j^2}{j^2 - 1} = \prod_{t+1 \leq j \leq m} (1 - j^{-2})^{-1}.
\]

for any $t$. How different this is from the usual Bayesian approach depends on the value of $N$. With orthodox Bayes, $N \to \infty$. Figure 3 shows how the variance decays for $N = 20$ versus $N = 100$, compared to the standard Bayesian approach in the previous section, with $\phi_0 = 0$. 

Example: Bayes rule. It is instructive to see how Bayes rule plays out in this framework...

4. A backward induced model for nonparametric density estimation

4.1. Coherent kernel density predictive distributions. In this section, the backward induction approach is used to derive a novel method for nonparametric density estimation with associated point-wise credible intervals. The method will be based on \( p_N(y \mid y_1:n) \) defined in terms of a kernel density estimator [Rosenblatt et al., 1956, Parzen, 1962, Silverman, 1986] of the form

\[
K_n^\tau(y) = \sum_{i=1}^{n} \phi(y \mid y_i, \tau),
\]

where \( \phi(y \mid \mu, \tau) \) is a normal density function with center \( \mu \) and “bandwidth” (variance) \( \tau \).

Begin by considering the marginalization consistency criterion applied to a kernel density estimator at sample size \( N \):

\[
p_{N-1}(y) = \int K_n^\tau(y)p_{N-1}(x)dx.
\]

Now “peel off” the \( N \)th observation \( x \equiv y_N \), obtaining

(10) \[
p_{N-1}(y) = \frac{N-1}{N}K_{N-1}^\tau(y) + \frac{1}{N} \int \phi(y \mid x, \tau)p_{N-1}(x)dx.
\]

Next, substitute (10) into itself:

\[
\frac{N-1}{N}K_{N-1}^\tau(y) + \frac{1}{N} \int \phi(y \mid x, \tau) \left[ \frac{N-1}{N}K_{N-1}^\tau(x) + \frac{1}{N} \int \phi(x \mid x', \tau)p_{N-1}(x')dx' \right] dx
\]
which simplifies to
\[
\frac{N-1}{N}K_{N-1}^{\tau}(y) + \frac{N-1}{N^{2}}K_{N-1}^{2\tau}(y) + \frac{1}{N^{2}} \int \int \phi(y \mid x, \tau)\phi(y \mid x, \tau)p_{N-1}(x')dx'dx.
\]

Exchanging the order of integration (and switching the names of \(x\) and \(x'\) for notational consistency), yields
\[
(11) \quad \frac{N-1}{N}K_{N-1}^{\tau}(y) + \frac{N-1}{N^{2}}K_{N-1}^{2\tau}(y) + \frac{1}{N^{2}} \int \phi(y \mid x, 2\tau)p_{N-1}(x)dx.
\]

Note that the third term in this expression is like the second term in expression (10), with \(N^{2}\) in place of \(N\) and \(2\tau\) in place of \(\tau\). Therefore, repeated substitution of (10) into the recursion gives an expanded representation of \(p_{N-1}(y)\) as
\[
(12) \quad p_{N-1}(y) = \sum_{j=1}^{\infty} \frac{N-1}{N^{j}}K_{N-1}^{j\tau}(y),
\]
which can be expressed as an expectation
\[
(13) \quad p_{N-1}(y) = EK_{N-1}^{Z\tau}(y).
\]

where \(Z \sim \text{Geometric}(\rho)\) for \(\rho = \frac{N-1}{N}\).

Moreover, because each term in (12) is itself a kernel density estimator and this representation involves only summation and convolution, we can apply the same process to obtain a nested sum expression for each predictive distribution at any number of steps back (\(N-2, N-3, \text{etc.}\)) simply by applying the mappings \(N \rightarrow N-1\) and \(\tau \rightarrow 2\tau\). Substitution and iteration yields
\[
(14) \quad p_{N-t} = \sum_{q=1}^{\infty} \cdots \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{N-1}{N^{j}} \frac{N-2}{(N-1)^{k}} \cdots \frac{N-t}{(N-t)^{q}} K_{N-t}^{q \cdots kj\tau}(y).
\]
Again, this can be seen as a nested expectation of independent geometric random variables $Z_h$ with parameters $\rho_h = \frac{N-h}{N-h+1}$ for $h = 1 \ldots t$:

$$p_{N-t}(y) = E_1E_2E_3 \ldots E_tK_{N-t}^{\prod_h Z_h \tau}(y).$$  

(15)

Observe that $K_{N-t}^{\prod_h Z_h \tau}(y)$ depends on the $Z_h$ variables only via their product. Defining

$$\chi_t = Z_1 \times Z_2 \times \ldots Z_t,$$

(16)

gives

$$p_{N-t}(y) = E K_{N-t}^{\chi_t \tau}(y)$$

(17)

where the expectation is now over $\chi_t$ for $t$ between 1 and $N - n$.

As a product of independent (but not identically distributed) geometric random variables, $\chi_t$ has no readily available closed form. However, a central limit theorem (in the log domain) suggests a reasonable log-normal approximation.

First, note that because the $Z_h$ geometric variables are independent, the product of their expectations gives the expectation of their product. Accordingly, $E\chi_t = \prod_{h=1}^{t} \rho_h \rho_h^{-1}$ with $\rho_h = \frac{N-h}{N-h+1}$. Similarly, $VZ_t = \frac{(1-\rho_1)}{\rho_1^2}$, so $EZ_t^2 = \frac{(2-\rho_1)}{\rho_1^2}$ and $V\chi_t = \prod_{h=1}^{t} (\frac{2-\rho_h}{\rho_h^2}) - \prod_{h=1}^{t} \rho_h^{-2}$ by properties of variance. Denote $E\chi_t \equiv \eta$ and $V\chi_t \equiv \nu$.

The log-normal approximation is improved by respecting the fact that $\chi_t \geq 1$. To that end, consider a log-normal random variable $\xi_t$ with mean $\eta - 1$ and variance $\nu$, which has parameters
\[ \mu = 2 \log (\eta - 1) - \frac{1}{2} \log (\nu + (\eta - 1)^2), \]
\[ \sigma = \sqrt{\log (1 + \nu/((\eta - 1)^2)}, \]

(18)

and set \( \chi_t = \xi_t + 1. \)

Note that the number of factors in the product defining \( \chi_t \) becomes small as \( t \) approaches \( N - n \), making the log-normal approximation inaccurate. This has an easy practical solution, however, which is to define the backward induction starting at \( N + a \) for \( a \) large enough that the log-normal central limit approximation obtains. Then, simply define \( N \) as the termination point for the forward simulation. Intuitively, this works because if \( N \) is thought to be large enough, then \( N + a \) also suffices, and \( p_N(y) \) and \( p_{N+a}(y) \) will be indistinguishable (by assumption).

Figures 4 and 5 illustrate the impact on the implied kernel for various values of \( t \).

The marginal kernel densities shown in Figure 4 were computed by numerical integration. At present, no convenient form is known for a log-normal scale mixture of normals. Fortunately, to implement the coherent density estimation proposed here, no evaluation of the density is required. Rather, it is only necessary to simulate from a kernel density distribution with a log-normal mixture of normal kernels, which can be done trivially as follows. At step \( t, \)

1. Select a location parameter \( u \) at random among the previous \( n + t - 1 \) data points (of which \( t - 1 \) are simulated).
2. Next, draw a scale parameter \( s \) from the log-normal distribution with parameters as in (18).
3. Finally, draw (pseudo-)observation \( y_{n+t}^* \) from \( N(u, \tau(s + 1)) \).
Figure 4. For $N = 1000$, $n = 50$ and $\tau = 0.04$, the implied kernel, marginally over $\chi_t$, is shown for $t = 1$ (dashed), $t = 400$ (dotted) and $t = 950$. At $t = N - n = 950$, the kernel is visually indistinguishable from a Gaussian kernel with variance 0.04.

Note that this forward simulation process yields independent samples of the distant future predictive $p_N(y|y_{1:n},y^*_ {(n+1):N})$, which may be obtained in parallel. This computational benefit makes the backward induced kernel density model an attractive alternative to Gaussian mixture models for density estimation, which require Markov chain algorithms [Escobar and West, 1995, Neal, 2000].

It was shown in West [1991] that among location-scale kernel density estimators, only the double-exponential (Laplace) kernel can give predictive densities satisfying (1). This result is not in conflict with the model here, because the sequence of kernels derived here are log-normal scale mixture of normals, which cannot be represented as a simple location-scale family.

4.2. Demonstrations.
Figure 5. The shifted log-normal mixing distribution becomes sharper as \( t \) approaches \( N - n \), collapsing to a near point-mass at \( \tau = 0.04 \) (shown in solid black). The dashed line shows the \( t = 1 \) one-step-ahead predictive diffuse mixing density for \( n = 50, N = 1000 \). The gray lines represent values of \( t \) between 10 and 950 in increments of 50.

4.2.1. Synthetic data. For this demonstration, \( n = 50 \) and \( n = 500 \) observations are drawn from a mixture of two Gaussians with equal weights:

\[
p(y) = \frac{1}{2} \phi(y \mid 2, 4) + \frac{1}{2} \phi(y \mid 10, 1).
\]

Each data set is fit using a backward induced kernel density procedure with \( N = 1000 \) and \( \tau = 0.08 \). These values were elicited by inspection of simulated data from mixtures of normals and the corresponding kernel density fit at different sample sizes and bandwidths. The resulting point estimate and uncertainty bands are depicted in Figures 6 and 7. As expected, the uncertainty bands of the \( n = 500 \) sample are much tighter than those of the \( n = 50 \) sample. For comparison, the R kernel density
Figure 6. Data are drawn from a mixture of two Gaussians, with \( n = 50 \). Three density estimates overlay the data histogram. Solid is the backward induced KDE with \( N = 1000 \) and \( \tau = 0.04 \); dashed is the true density; dotted is the R KDE with bandwidth select method SJ. One-thousand draws from the posterior density are shown in gray.

4.2.2. The galaxy data. The “galaxy data” have been widely used to exemplify Bayesian and non-Bayesian density estimation techniques. The data are 82 velocity measurements (in km/second) of galaxies obtained from an astronomical survey of the Corona Borealis region [Roeder, 1990]. Notable Bayesian papers using this data include Carlin and Chib [1995], Escobar and West [1995] and Bernardo [1999].

Figure 8 depicts the posterior mean for the \( N = 1000, \tau = 0.04 \) model, along with one-thousand posterior draws to provide visual uncertainty bands. Also depicted are
Figure 7. Data are drawn from a mixture of two Gaussians, with \( n = 500 \). Three density estimates overlay the data histogram. Solid is the backward induced KDE with \( N = 1000 \) and \( \tau = 0.08 \); dashed is the true density; dotted is the \( R \) KDE with bandwidth selection method SJ. One thousand draws from the posterior density are shown in gray. The uncertainty bands are much narrower with \( n = 500 \) than with \( n = 50 \).

the default kernel density estimate from the \( R \) software language and a histogram. Although the point estimate is less smooth than the default kernel density estimate, the posterior draws reflect substantial uncertainty, covering both the default kernel density estimate and the histogram contours.

4.3. Incorporation of prior information via pseudo-data.

4.4. Uncertainty reduction as \( n \to \infty \). The marginalization consistency property (1) entails that the sequence of predictive densities forms a Martingale sequence. Specifically, writing \( X_t \equiv p_n(y \mid Y_{1:t}) \), marginalization consistency can be stated as
The galaxy data of Roeder [1990] consists of \( n = 82 \) astronomical measurements. The posterior mean density is shown for the \( N = 1000 \) and \( \tau = 0.08 \) backward induced model (solid line). The dashed line depicts the default KDE in R.

The condition that \( \mathbb{E}(X_{t+1} \mid X_{1:t}) = X_t \). By solving the backward induction procedure, one has that \( \mathbb{E}(X_N) = X_n \), by construction; the one-step-ahead predictive density is the posterior mean. Because it is well-known that kernel density estimation is consistent, it follows directly that the posterior mean is also consistent. To study the concentration of the posterior about this mean, one can apply the Azuma-Hoeffding inequality. In particular, for any \( y \),

\[
|p_t(y) - p_{t+1}(y)| \leq \frac{\phi(0 \mid 0, \tau)}{t + 1},
\]

which follows from the fact that the kernel density is most peaked when the bandwidth equals \( \tau \) and the kernel is Gaussian, and that density functions are always
Figure 9. An application of the Azuma-Hoeffding inequality to the Martingale sequence of predictive kernel densities implies shrinking uncertainty about the posterior mean of the $m$-step ahead functional, as sample size increases. This illustration depicts the concentration of posterior mass as a progressively narrowing “uncertainty cone,” fanning out from the one step ahead distribution, as the observed sample size is pushed forward from $n$ to $n'$. Here $N = n + m$ and $N' = n' + m$ for a fixed $m$.

greater than or equal to zero. Therefore, Azuma-Hoeffding gives

$$\Pr\{|p_N(y) - p_n(y)| \geq \epsilon\} \leq 2 \exp \left( \frac{-\epsilon^2}{2c^2 \sum_{j=n+1}^{N+1} (j+1)^{-2}} \right),$$

$$= 2 \exp \left( \frac{-\epsilon^2}{2c^2 (\psi^{(1)}(n + 2) - \psi^{(1)}(n + m + 2))} \right)$$

where $\psi^{(1)}(\cdot)$ denotes the first derivative of the polygamma function, $c = \phi(0 \mid 0, \tau)$ and $N = n + m$. Thus, the asymptotic point-wise concentration is dictated by the growth of the difference $\psi^{(1)}(n + 2) - \psi^{(1)}(n + m + 2)$ as $n \to \infty$. It is easy to check that indeed this difference approaches zero as $n$ grows.
The temporal coherence condition plays the same role in the backward induction approach as exchangeability plays in defining traditional Bayesian probability models. In fact, an exchangeable model is always temporally coherent. However, interesting and useful models that satisfy these conditions need not be exchangeable — such as the kernel density model in the previous section. A kernel density model based on the double exponential distribution, as detailed in West [1991], is another example. In the discussion section of that paper, it is remarked that the double-exponential kernel density model does not correspond to any exchangeable distribution, because the likelihood evaluation depends on the ordering of the observed data. Note, however, that temporally coherent kernel density models are nonetheless learning symmetric in the following sense.

If the ordering of the first $n$ observations is unknown, arriving in a batch, one must average over permutations in order to evaluate their joint likelihood:

$$p(y_{1:n}) = \frac{1}{n!} \sum_{\pi \in \Pi} p_0(y_{\pi_1}) p_1(y_{\pi_2} \mid y_{\pi_1}) \cdots p_{n-1}(y_{\pi_n} \mid y_{\pi_1:\pi_{n-1}}), \tag{22}$$

where $\pi \in \Pi$ denotes a permutation of the indices 1 through $n$. However, observe that this averaging does not impact the conditional distribution of the unobserved future data $Y_{(n+1):N}$, so long as the observed data $y_{1:n}$ appears in each subsequent conditional distribution symmetrically:

$$p(y_{(n+1):N} \mid y_{1:n}) = \frac{1}{n!} \sum_{\pi \in \Pi} p_{1:n}(y_{\pi_1:n}) p_{(n+1):N}(y_{(n+1):N} \mid y_{\pi_1:n}) p_{1:n}(y_{\pi_1:n}), \tag{23}$$

$$= p_{(n+1):N}(y_{(n+1):N} \mid y_{\pi_1:n}) = p_{(n+1):N}(y_{(n+1):N} \mid y_{1:n}).$$
This implies, remarkably, that for a backward-induced model with permutation-invariant conditional distributions, the ordering of the observed data matters for likelihood evaluation (which requires permutation averaging), but does not matter for posterior inference via forward simulation.

The choice of the large-sample predictive density $p_N(\cdot \mid y_{1:N})$ plays the same role in the backward induction approach as the choice of a sufficient statistic does in an exchangeable Bayesian model. Determining which estimators can be tractably converted into probability models satisfying temporal coherence is an interesting question for future research.

**References**


