Bayesian Analysis (2012) **7**, Number 1, pp. 37–44

Comment on Article by Sancetta

Bertrand Clarke[*]

1 Overview

This paper makes a landmark contribution in three senses.

First, it provides many results that are fundamentally important in their own right. I refer specifically to Theorems 3 and 8. Theorem 3 treats arbitrary loss functions by breaking the integral into two terms, one, I_t, where a difference of losses is bounded and another, Π_t, where a bound on the moments of a difference of losses must be used. (All notation here is the same as the author's unless noted otherwise.) The treatment of these two terms reveals the role of the relative entropy and how the tails of the loss affect the risk, respectively. This is a proof that makes us wiser.

Another fundamentally important result is Theorem 8. It can be regarded as an observation pursuant to Theorem 3, however, Theorem 8 is conceptually different because the true model need not be in the model class. This admission often goes under the genteel name of model mis-specification or model uncertainty. Traditionally (say in the 80's and 90's if not before), many statisticians would argue 'If I knew the true model were not in the model class I was using, I would choose a different model class.' However, as most of us who have attempted applied problems have found, we routinely use model classes on the grounds that we can get 'reasonable' answers. These days more careful analysts will often say things like 'I know the model is not true, but I think the error in approximation is no more damaging than the other sources of variability, at least for what I'm using the model to do.' By including Theorem 8, the author is recognizing that we can't assume the true model is in the class we are using so we have to know how this affects our techniques. I hope that in the future it will be routine to do this sort of analysis and commend the present author for his prescience.

Second, the paper provides a comprehensive overview of the cumulative risk in sequential prediction even though the main property of universality comes out of the Bayesian consistency literature. Thus, this paper is a timely synthesis of many ideas – some from the author and some scattered in various places throughout the literature. The comprehensiveness draws a line from Condition 1 – a variant on information denseness which goes back to Barron (1986) – to the basic universality of Theorem 4. This leads naturally to the predictive result of Theorem 3, the unique role of Bayes model averaging (BMA), the issue of switching reference classes (a variation of the use of multiple parametric models in BMA), Theorem 8 for wrong-model analysis, and finally to a general discussion of averaging for prediction versus prediction with expert advice. By drawing together so many threads, I hope this paper will become a key reference for researchers in the general area as well as for new researchers wanting to use or expand

[*]Dept. of Medicine, University of Miami, Miami, FL, bclarke2@med.miami.edu

DOI:10.1214/12-BA701A

the results.

Third, the paper is unabashedly predictive. That is, the paper regards the goal of prediction as the central task for a statistical method to accomplish well, implicitly regarding concerns such as model identification and decision making as derived from good prediction. This is important because it seems that the predictive perspective, advocated in statistics at least as early as Dawid (1982), has been becoming more widespread in recent years. This view is distinct from Frequentism or Bayesianism, but often reflects a sort of merger of the two. The Bayesian angle is seen in the present paper because the focus is on the marginal for the data and the Frequentist angle is seen in taking the expectations over the whole sample space under a fixed parameter value (when conditioning on the data at each stage of the prediction process would be the standard Bayesian approach). Even prediction with expert advice – which is intended to be distribution independent and hence neither Bayesian nor Frequentist – is included in the predictive view as in Sec. 6.3. Overall, the predictive view is an emerging perspective that may well continue to become more important in the coming years and the present paper is helping to elucidate these very current developments. For the interested reader, recent important publications are Ebrahimi et al. (2010) and van Erven et al. (to appear, 2012); a related approach is in Clarke (2010).

Separate from these encomia, I want to present an example and a conjecture that may permit a general asymptotic treatment of individual terms in $E_\theta D_{1,T}(P_\theta||P_w)$, the central quantity the author examines. After all, it is not enough to look at the cumulative risk. Prequentially, one is concerned with the risk at each stage of data accumulation because one must issue a prediction at the end of each stage. Then, I want to propose that the asymptotics for the individual terms can be used to obtain stagewise versions of the author's results.

2 Individual Risks

The author observes that $(1/T)E_\theta D_{1,T}(P_\theta||P_w) \to 0$ and that in fact

$$E_\theta D_{1,T}(P_\theta||P_w) = \sum_{t=0}^{T-1} E_\theta D(P_\theta||P_w(\cdot|X^t)) \sim \frac{d}{2}\ln T + C, \tag{1}$$

asymptotically, where C is a constant that can be identified, d is the dimension of θ, and P_w^n is the marginal for the data $X^t = (X_1, \ldots, X_t)$. (Conditioning on X_0 means the conditioning drops out.) Now, $\log T \sim \sum_{t=1}^T 1/t$ suggests the individual terms $E_\theta D(P_\theta||P_w(\cdot|X^t))$ behave like $d/(2t)$ asymptotically. In this section, we verify this in the normal example and sketch a proof more generally.

2.1 The Normal Example

To find $E_\theta D(P_\theta||P_w(\cdot|X^t))$, we have to find the predictive density for $P_w(\cdot|X^t)$, find its relative entropy distance to the true density, and then take the expectation over X^t in the true density. For the normal example, suppose we have t IID copies of $X \sim N(\theta, \nu^2)$ so that $\bar{X} = \bar{X}_t \sim N(\theta, \sigma^2)$ with $\sigma^2 = \nu^2/T$. Assume that $\theta \sim N(\mu, \tau^2)$ in which ν^2, μ, and τ^2 are known. Let $\rho = (1/\tau^2) + (1/\sigma^2)$ and write

$$\mu(\bar{x}) = \mu(\bar{x}_t) = \left(\frac{\sigma^2}{\sigma^2 + \tau^2}\right)\mu + \left(\frac{\tau^2}{\sigma^2 + \tau^2}\right)\bar{x}.$$

Now, it is well known that $(\Theta|\bar{x}) \sim N(\mu(\bar{x}), 1/\rho)$ and the predictive density for a new outcome x is

$$m(x|\bar{x}) = \int \left(\frac{1}{2\pi} e^{-(x-\theta)^2/s\nu^2}\right)\left(\sqrt{\frac{\rho}{2\pi}} e^{-(\theta-\mu(\bar{x}))^2/(2/\rho)}\right) d\theta. \qquad (2)$$

Completing the square in θ gives the identity

$$\begin{aligned} -\frac{\rho}{2}(\theta - \mu(\bar{x}))^2 - \frac{1}{2\nu^2}(x-\theta)^2 &= -\frac{1+\rho\nu^2}{2\nu^2}(\theta - \alpha)^2 \\ &- \frac{1}{2\nu^2}\left(x^2 + \rho\nu^2\mu(\bar{x})^2 - \frac{[x + \rho\nu^2\mu(\bar{x})]^2}{1+\rho\nu^2}\right) \qquad (3) \end{aligned}$$

where

$$\alpha = \frac{x + \rho\nu^2\mu(\bar{x})}{1 + \rho\nu^2}.$$

Using (3) in (2) we get

$$\begin{aligned} m(x|\bar{x}) &= \sqrt{\frac{\rho}{2\pi(1+\rho\nu^2)}} \times \int \frac{\sqrt{1+\rho\nu^2}}{\nu\sqrt{2\pi}} e^{-\frac{1+\rho\nu^2}{2\nu^2}(\theta-\alpha)^2} d\theta \\ &\times \ e^{-\frac{1}{2\nu^2}\left(x^2+\rho\nu^2\mu(\bar{x})^2 - \frac{[x+\rho\nu^2\mu(\bar{x})]^2}{1+\rho\nu^2}\right)} \\ &= \sqrt{\frac{\rho}{2\pi(1+\rho\nu^2)}} \times e^{-\frac{1}{2\nu^2}\left(x^2+\rho\nu^2\mu(\bar{x})^2 - \frac{[x+\rho\nu^2\mu(\bar{x})]^2}{1+\rho\nu^2}\right)}. \qquad (4) \end{aligned}$$

Completing the square in x gives the identity

$$\begin{aligned} &-\frac{1}{2\nu^2}\left(x^2 + \rho\nu^2\mu(\bar{x})^2 - \frac{[x+\rho\nu^2\mu(\bar{x})]^2}{1+\rho\nu^2}\right) \\ &= -\frac{\rho}{2(1+\rho\nu^2}(x - \mu(\bar{x}))^2 - \frac{1}{2\nu^2}\left(\rho\nu^2\mu(\bar{x})^2 - \frac{\rho\nu^2\mu(\bar{x})^2}{1+\rho\nu^2} - \frac{\rho^2\nu^4\mu(\bar{x})^2}{1+\rho\nu^2}\right), \qquad (5) \end{aligned}$$

where we also used $1 - 1/(1+\rho\nu^2) = \rho\nu^2/(1+\rho\nu^2)$. Now substitute (5) into (4). Since the integral of (4) over x must be one, we see that (5) must be one. So,

$$m(x|\bar{x}) = \sqrt{\frac{\rho}{2\imath(1+\rho\nu^2)}} e^{-\frac{\rho}{2(1+\rho\nu^2)}(x-\mu(\bar{x}))^2},$$

i.e., $m(x|\bar{x})$ is the density of a $N(\mu(\bar{x}), (1+\rho\nu^2)/\rho)$ random variable.

With mild abuse of notation, we can find the relative entropy between $p(x_{t+1}|\theta)$ and $m(x_{t+1}|\bar{x})$. It is

$$D(p(x_{t+1}|\theta)||m(x_{t+1}|\bar{x})) = \int \phi_{\theta,\mu}(x) \ln \frac{\phi_{\theta,\mu}(x)}{\phi_{\mu(\bar{x}),(1+\rho\nu^2)/\rho}(x)} dx, \tag{6}$$

where $\phi_{a,b^2}(\cdot)$ is the normal density with mean a and variance b^2. Writing in the form of the normal densities in (6) and simplifying (using $E(X-\theta)^2/\nu^2 = 1$ and adding and subtracting θ in the exponent of $\phi_{\mu(\bar{x}),(1+\rho\nu^2)/\rho}(x)$) gives that

$$D(p(x_{t+1}|\theta)||m(x_{t+1}|\bar{x})) = \frac{1}{2} \ln \frac{1+\rho\nu^2}{\rho\nu^2} + \left(\frac{1+\rho\nu^2}{\rho}\right) E_\theta(\theta - \mu(\bar{x}))^2. \tag{7}$$

Standard manipulations give that the expectation in the last term of (7) is

$$E_\theta(\theta - \mu(\bar{x}))^2 = \left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)^2 (\theta-\mu)^2 + \left(\frac{\tau^2}{\sigma^2+\tau^2}\right)^2 \frac{\nu^2}{n}. \tag{8}$$

Using (8) in (7) gives

$$\begin{aligned} &D(p(x_{t+1}|\theta)||m(x_{t+1}|\bar{x})) \\ &= \frac{1}{2} \ln \frac{1+\rho\nu^2}{\rho\nu^2} + \frac{1+\rho\nu^2}{\rho} \left[\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)^2 (\theta-\mu)^2 + \left(\frac{\tau^2}{\sigma^2+\tau^2}\right)^2 \frac{\nu^2}{n} \right]. \end{aligned} \tag{9}$$

Since $\sigma^2 = \nu^2/t$, is easy to see that, as $t \to \infty$,

$$\frac{1+\rho\nu^2}{\rho\nu^2} \to 1; \quad \frac{\rho}{2(1+\rho\nu^2)} \to \frac{1}{2\nu^2}; \quad \frac{\sigma^2}{\sigma^2+\tau^2} \to 0; \quad \text{and} \quad \frac{\tau^2}{\sigma^2+\tau^2} \to 1.$$

Using these convergences in (9) we see that $D(p(x_{t+1}|\theta)||m(x_{t+1}|\bar{x}))$ in (6) is approximately $1/(2t)$ and in fact, careful examination of the derivation gives that

$$D(p(x_{t+1}|\theta)||m(x_{t+1}|\bar{x})) = \frac{1}{2t} + o\left(\frac{1}{t}\right) \tag{10}$$

as $t \to \infty$. That is, the outer expectation in the sum in (1) does not need to be taken; the dependence on X^t drops out for the normal example. We suggest that analogous results can be derived for other exponential families with conjugate priors.

2.2 A General Case

Having seen the normal case we are ready to conjecture that in general

$$E_\theta D(P_\theta||P_w(\cdot|X^t)) = \frac{d}{2t} + o\left(\frac{1}{t}\right), \tag{11}$$

where θ is a d-dimensional real parameter. It's not hard to give a heuristic argument. Using asymptotic normality of the posterior, we get

$$\ln\left(\frac{p(x|\theta)}{\int p(x|\theta')w(\theta'|X^t)d\theta}d\theta\right) \approx \ln\left(\frac{p(x|\theta)}{\int p(x|\theta')\phi_{\hat{\theta},(t\hat{I}(\hat{\theta}))^{-1}}(\theta')}d\theta\right) \approx \ln\left(\frac{p(x|\theta)}{p(x|\hat{\theta})}\right), \quad (12)$$

where $\hat{I}(\cdot)$ is the empirical Fisher information and $\hat{\theta}$ is the maximum likelihood estimator (MLE). It may be possible to control the error in the first approximation. For instance, Bickel and Yahav (1969) use L^1, and Clarke and Barron (1988)(Appendix) and Clarke (1999) use relative entropy. (Many other contributors use other modes of convergence.) The second approximation might be formalized by using the rate of concentration of the normal density since it's set up for a Laplace approximation.

Given (12), we can write

$$E_\theta D(P_\theta||P_w(\cdot|X^t)) \approx E_{\theta,t}E_\theta \ln\left(\frac{p(x|\theta)}{p(x|\hat{\theta})}\right) = E_{\theta,t}D(P_\theta||P_{\hat{\theta}}) \quad (13)$$

where the expectation over the $t+1$ random variable is denoted E_θ and the expectation over the first t random variables is denoted $E_{\theta,t}$. Cencov (1981) proved an expansion for the risk of the MLE, $E_{\theta,t}D(P_\theta\|P_{\hat{\theta}})$, with leading term $d/(2t)$ and error $\mathcal{O}(n^{-3/2})$. However, it is enough to note here that for fixed $\hat{\theta}=\hat{\theta}(x^t)$, we have the Taylor expansion

$$D(P_\theta||P_{\hat{\theta}}) = \frac{1}{2}(\theta-\hat{\theta})'I(\theta)(\theta-\hat{\theta}) + o\left(\|\theta-\hat{\theta}\|^2\right), \quad (14)$$

where $I(\theta)$ is the Fisher information at θ. Expression (14) holds for $\hat{\theta}$ close to θ, a property ensured by the consistency of the MLE. Now, using (14) in (13) we get

$$\begin{aligned} E_{\theta,t}D(P_\theta||P_w(\cdot|X^t)) &= \frac{1}{2t}E_{\theta,t}t(\theta-\hat{\theta})'I(\theta)(\theta-\hat{\theta})\chi_{\|\theta-\hat{\theta}\|\le\delta} \\ &+ E_{\theta,t}o(\|\theta-\hat{\theta}\|^2)\chi_{\|\theta-\hat{\theta}\|\le\delta} \\ &+ E_{\theta,t}\int p(x|\theta)\ln\frac{p(x|\theta)}{p(x|\hat{\theta})}\chi_{\|\theta-\hat{\theta}\|>\delta}dx, \end{aligned}$$

for $\delta>0$, in which the last two terms are errors that must go to zero at rate $o(1/t)$.

For the first term in (15) note that $t(\theta-\hat{\theta})^2I(\theta)$ converges (in distribution) to a χ^2_d, the expectation converges to d (if uniform integrability is assumed, for instance). Essentially, this is Wilks' theorem in L^1 and versions of it are known cf. Clarke and Barron (1990). In this way the first time might be controlled as $d/(2t)+o(1/t)$. If one assumes uniformity of the Taylor expansion in (14) over a small neighborhood of θ's and $\hat{\theta}$'s then similar arguments might give that the second term in (15) is

$$\frac{1}{t}E_{\theta,t}t\|\theta-\hat{\theta}\|^2o(1)\chi_{\|\theta-\hat{\theta}\|\le\delta}$$

and hence also $o(1/t)$.

The third term in (15) seems more difficult and will depend on letting θ and $\hat{\theta}$ be close enough to each other that $p(x|\theta)/p(x|\hat{\theta})$ will be near one on a set of high probabiity. This will give that $|\ln p(x|\theta)/p(x|\hat{\theta})|$ is near zero on a set of high probability. However, we also want $P_\theta(\|\theta-\hat{\theta}\|>\delta)\to 0$ at a suitable rate as $t\to\infty$. For instance, Chebyshev gives that

$$P_\theta(t\|\theta-\hat{\theta}\|^2>t\delta^t)\leq\frac{1}{t\delta^2}E\chi_{\|\theta-\hat{\theta}\|>\delta}=o\left(\frac{1}{t}\right).$$

So, as long as the set on which $p(x|\theta)/p(x|\hat{\theta})$ is not small can be controlled (as it can trivially when X and θ are bounded), the conjecture should be true.

In principle, formalizing and justifying this sequence of approximations would lead to a proof of (11), at least when the MLE exists and is asymptotically normal as it is for instance with smooth, regular, full rank, exponential families equipped with smooth priors on open parameter spaces.

3 Relation to Sancetta's Paper

Asuming a result like (11) can be proved for a large class of parametric families, what does this have to do with the present paper? The possible answer is 'quite a lot'. Careful examination of the proof of Theorem 1, for instance, suggests we should be able to get a version of the bound on $\sup_\theta E^\theta D(P_\theta||P_w)$ given in the theorem (but not the universality). Moreover, examining the proof of Theorem 3 suggests we should be able to get a bound on

$$\sup_\theta E_\theta^{t-1}|\mathcal{L}(Z_t,f_t(w))-\mathcal{L}(Z_t,f_t(\theta))|,$$

asymptotically in t, merely by using the same hypotheses and technique of proof but treating an I_t and Π_t individually rather than taking their average over t. Likewise for the main bounds in Theorems 4, 5, 6, and 8, again apart from the universality clauses. Thus, a development for the t-th stage risk can be developed analogously to the present paper which focuses on the cumulative risk.

The key benefit of a development for the t-th stage risk is to reflect the fact that prediction is stagewise and at each stage we have the opportunity to change our predictor. So, if the bound in Theorem 8 is large and the inequality is believed to be tight we would be led to rechoose our predictive scheme. This would be a setting in which varying classes of predictors, as envisaged in Theorem 6, would be natural. We would want to use this to choose refinements of our predictors at each stage of a prequential strategy to make sure our methods were homing in on good predictors.

As a penultimate point, note that the asymptotics demonstrated in this paper show that universality is a link between Bayes consistency and prediction. However, the concept of universality strikes me as weak. The author admits that in most regular cases $E^\theta D_{1,T}(P_\theta||P_w)=\mathcal{O}(\ln T)$ so that the universality in Definition 1 is handily satisfied with room to spare: $\ln T$ is quite small compared to T. Are there important cases where the rate of universality is much larger than $(\ln T)/T$ so that the expected risk at stage t is also much higher than $d/(2t)$? It would be nice to derive t-th stage results analogous to (11) for families with other rates of convergence.

Finally, as I see it, if one focusses on an individual stage of prediction, say the t-th, then one is outside of the problem class for which universality is relevant. So, the goal of Bayes consistency while germane to predictive performance remains distinct from how one evaluates a predictor for actual usage. That is, Bayes consistency via cumulative risk is a helpful perspective for good stagewise prediction but Bayes consistency per se is conceptually disjoint from the goal of good stagewise prediction as encapsulated by controlling individual risks such as (11) or by extensions to the present paper as suggested at the beginning of this section.

References

Barron, A. R. (1986). "Discussion of Diaconis and Freedman: On the consistency of Bayes estimates." *Annals of Statistics*, 14: 26–30. 37

Bickel, P. and Yahav, J. (1969). "Some contributions to the asymptotic theory of Bayes solutions." *Z. Wahrscheinlichkeitstheorie und verw. Gebiete*, 11: 257–276. 41

Cencov, N. (1981). *Statistical Decision Rules and Optimal Inference*. Providence: American Mathematical Society. 41

Clarke, B. (1999). "Asymptotic normality of the posterior in relative entropy." *IEEE Transactions on Information Theory*, 45: 165–176. 41

— (2010). "Desiderata for a predictive theory of statistics." *Bayesian Analysis*, 5: 283–318. 38

Clarke, B. and Barron, A. R. (1988). "Information-theoretic asymptotics of Bayes methods." Technical Report 26, Statistics Department, University of Illinois. 41

— (1990). "Information-theoretic asymptotics of Bayes methods." *IEEE Transactions on Information Theory*, 38: 453–471. 41

Dawid, A. P. (1982). "The well-calibrated Bayesian." *Journal of the American Satistical Association*, 77: 605–610. 38

Ebrahimi, N., Soofi, E., and Soyer, R. (2010). "On the sample information about the parameter and prediction." *Statistical Science*, 25: 348–367. 38

van Erven, T., Grunewald, P., and de Rooij, S. (to appear, 2012). "Catching up faster by switching sooner: A prequential solution to the AIC-BIC dilemma." *Journal of the Royal Statistical Society Series B*. 38