Computational Statistics and Data Analysis 104 (2016) 183–196

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Using the Bayesian Shtarkov solution for predictions

Tri Le *, Bertrand Clarke

Department of Statistics, University of Nebraska-Lincoln, 340 Hardin Hall North, Lincoln, NE, USA

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form 11 May 2016
Accepted 30 June 2016
Available online 15 July 2016

Keywords:
Bayes
Prequential
Model average
Stacking
Shtarkov predictor
Bagging

ABSTRACT

The Bayes Shtarkov predictor can be defined and used for a variety of data sets that are exceedingly hard if not impossible to model in any detailed fashion. Indeed, this is the setting in which the derivation of the Shtarkov solution is most compelling. The computations show that anytime the numerical approximation to the Shtarkov solution is 'reasonable', it is better in terms of predictive error than a variety of other general predictive procedures. These include two forms of additive model as well as bagging or stacking with support vector machines, Nadaraya–Watson estimators, or draws from a Gaussian Process Prior.

1. Introduction

What kind of inference can we do when we do not believe the data were generated by a model? The most obvious answer to this question is prediction[1]: As long as there is something to measure we can make a guess as to its next value. The act of making the guess does not by itself even require there be anything stable enough about the data generator (DG) to make good prediction feasible. Moreover, predicting is more general than modeling because every model corresponds to a predictor but not every predictor corresponds to a model. Hence, if there is no model for a given DG we are essentially forced to predict using a larger class of predictors than models represent, i.e., it is not reasonable to limit ourselves to models for prediction. One effect of this in a Bayes context is to change the meaning of the prior.

This situation is far from unusual. Indeed, one can argue that many of the most important data that are gathered were not generated by a model, or, more precisely, they were not generated by any mechanism for which modeling per se is likely to be helpful. We use the term $\mathcal{M}$-open to label this class of problems, see Bernardo and Smith (2000) and Clyde and Iversen (2013). Specifically, we say a problem is $\mathcal{M}$-open when there is no model that accurately describes the mechanism by which the DG generated the data. Operationally, when we say this we mean that on intuitive and pragmatic grounds it is more reasonable to abandon rather than continue the search for a true model.

Let us review three techniques that have been proposed for $\mathcal{M}$-open data.

One of the earliest techniques intended for $\mathcal{M}$-open data is due to Shtarkov (1987). He recognized that if there is no model it may make sense to imagine a collection of 'experts' regarded as density functions who issue predictions. Then, at each time step the best expert can be identified in the sense of regret under log-loss. This approach has been extended in Vovk (2001) and Cesa-Bianchi and Lugosi (2006). Techniques for computing the Shtarkov solution were first presented in Kontkanen and Myllymaki (2007). Although Shtarkov's formulation was not Bayesian, the frequentist Shtarkov predictor

* Corresponding author.
E-mail addresses: tle20@unl.edu (T. Le), bclarke3@unl.edu (B. Clarke).

[1] R code for generating Bayes Shtarkov predictions presented here is in given in an annex to the electronic version of this paper.

http://dx.doi.org/10.1016/j.csda.2016.06.018

is asymptotically Bayes (see Clarke, 2007) and it is easy to write down the Bayes version. (Here and below, we abbreviate 'Bayesian' to 'Bayes' wherever possible, for brevity.)

A second approach for prediction is stacking, due to Wolpert (1992). Given a list of candidate models, weights for their predictions can be derived by minimizing a criterion similar to cross-validation. There are several versions of this minimization problem depending on the constraints imposed on the weights. Stacking has been studied by Breiman (1996b), Ting and Witten (1999) amongst others and explicitly extended to $\mathcal{M}$-open problems by Clyde and Iversen (2013). Le and Clarke (2015) showed that stacking can be asymptotically regarded as the Bayes action under several loss functions.

A third technique is bagging, see Breiman (1996a). Despite its origins in classification, bagging has also been used, usually without comment, for regression problems that are $\mathcal{M}$-open. For instance, Strobl et al. (2009) provide several examples as well as a good discussion of the key features of bagging in practice. It can be shown that bagging is asymptotically a specific form of Bayes model averaging (BMA), see Le and Clarke (unpublished).

A separate issue from how predictions from components are combined is the selection of the components themselves. While it is desirable to choose components that will yield good predictors, this cannot be known in advance. So, for $\mathcal{M}$-open DG's we want components that are flexible.

For bagging and stacking we have used three classes of components. First is the Nadaraya–Watson (NW) estimator. Given a data set, we can draw, say, ten bootstrap samples and therefore generate ten NW estimators. In the $\mathcal{M}$-open case it does not make sense to call them estimators since there is nothing to estimate. However, we do so for convenience. Now we can 'bag' the ten NW estimators by taking the average of the predictions they make at a new value of the explanatory variables. Alternatively, we can stack the ten NW estimators or more precisely the predictors they generate. NW estimators can be regarded as Bayes by using a prior on the smoothing parameter. This is computationally infeasible for the scale of our work here.

Second, given a kernel function, we can obtain the posterior from a Gaussian process prior (GPP) and use it to generate predicted values similar to the way NW estimators were used.

A third class of components that we use is support vector machines (SVM's). These are also based on kernel functions. Although it does not seem to have been formally proved, SVM regression and Gaussian process regression are not equivalent, see Rasmussen and Williams (2006) Sec. 6.4.1. Moreover, while more familiar from classification than regression, SVM's do give a regression function under an ϵ-insensitive loss. Again, taking ten bootstrap samples leads to ten SVM regression functions that can be bagged or stacked. The form of solution is based on the Representer Theorem, see Kimeldorf and Wahba (1973). More recently, Chakraborty et al. (2012) developed Bayes estimation in this context.

A fourth class of components that we use only with the Shtarkov predictor is the multinomial. Even though the experts combined in a Shtarkov predictor may be discrete or continuous, here we must discretize any explanatory variables so that predictions can be computed. For independent data, the 'experts' then naturally assume a multinomial distribution. The dependent variable must also be discretized to minimize computational difficulties and avoid having to choose specific parametric forms for the experts. While the computational procedure was proposed by Kontkanen and Myllymaki (2007) in the frequentist case, we are the first to evaluate how well it performs in contrast to other techniques.

Given independent data of the form $(y_i, x_i)|_{i=1}^n$ where x_i is a vector of explanatory variables, we form various predictors $\hat{y}_{i+1}(\cdot)$ for $Y_{i+1}(x_{i+1})$ using the first i data points. We evaluate these predictors by their cumulative squared prediction error, namely

$$CPE = \sum_{i=1}^{n} (\hat{y}_i(x_i) - y_i)^2. \tag{1}$$

Thus our evaluation is prequential, see Dawid (1984).

For a sequence of $\mathcal{M}$-open data sets we compare the CPE's of 10 different predictors using up to two explanatory variables (chosen by highest correlation with the y_i's). Six predictors come from the combination of bagging or stacking with NW, GPP, and SVM predictors. The other four are forms of the Shtarkov predictor: no side information, one of two explanatory variables as side information, and two explanatory variables as side information. For comparison purposes, we also generate predictions using additive models via the Bayes LASSO and the horseshoe prior, even though these are intended for use outside the class of $\mathcal{M}$-open problems. To get a better assessment of CPE, often (1) is averaged over several permutations of the data. Also, if several permutations of the data are used the standard deviation of the errors can be found at each time step, $i = 1, \ldots, n$. We have not done this here because it was too computationally demanding and probably not necessary given the sample sizes we have used in our examples.

Our main finding here is that when Bayes Shtarkov solutions are feasible to compute reliably at least one of them outperforms the other six methods. We attribute this to the fact that the optimality property satisfied by Shtarkov predictors is the most desirable one for $\mathcal{M}$-open problems. This is not to say that the naive use of Bayes Shtarkov solutions will always be the best. Indeed, we found many cases where side information made for a worse predictor than the absence of side information. Moreover, computing Shtarkov solutions reliably is very difficult with existing data storage. Amongst the other predictors, we also noted some regularities but they were not as strong. First, for $\mathcal{M}$-open problems, stacking with SVM's tended to do well in the sense that stacked SVM's were always one of the top three methods in terms of CPE. Second, in some cases (not shown here) where the problems were $\mathcal{M}$-open but not as hard the best results were typically obtained by stacking NW estimators.

From a high level, these findings are not a surprise. We expect that for more complex problems a more flexible method such as stacking SVM's should perform well and we expect that for a slightly less complex problem a slightly less flexible method such as stacking NW estimators should perform well. However, our results suggest that if enough data and computing power were available then the Shtarkov predictors would always be the best in the hardest $\mathcal{M}$-open problems.

The structure of this paper is as follows. In Section 2 we formally present the predictors we will compare computationally. In particular, we provide a Bayes version of the Shtarkov predictor and indicate how to find it computationally. In Section 3 we explain exactly what we have computed and present our results for five data sets. In a brief conclusion section we discuss the broader implications of our work. Technical details not germane to the main flow of the paper are relegated to the two Appendices.

2. Model averages

Here we list three model averages—the Bayes Shtarkov predictor, stacking, and bagging. Even though the Bayes Shtarkov we use is immediately derivable from published results on the frequentist case, we present the details for the sake of completeness. We only discuss stacking and bagging to provide their definitions and comment on their Bayesian interpretation.

2.1. Bayes Shtarkov predictors

2.1.1. Bayes Shtarkov predictors without side information

Start by considering online prediction of arbitrary sequences $y_1, y_2, \ldots$, drawn from a finite discrete set $\mathcal{Y}$. Interest focusses on the case that no probability distribution can be assumed for a sequence of length n, say $y^n = (y_1, y_2, \ldots, y_n)$. This is the paradigm $\mathcal{M}$-open statistical prediction problem for random variables taking values in a finite set.

This problem can be regarded as a sequential game between Nature, N, and a Forecaster, F, permitting F to access a collection of experts indexed by $\theta \in \Theta \subset \mathbb{R}^k$ for some k. In the special case of log-loss, each round of the game proceeds as follows. Each expert announces a density say p_θ. Given this, F announces a density $q(\cdot)$ that will be used to predict the value N issues. Finally, N issues y and pays F $\log q(y)$. If this number is negative, it is the amount of money F pays N and this concludes the round. See Shtarkov (1987) and Cesa-Bianchi and Lugosi (2006) for details of this game and its properties.

Now suppose n independent rounds of this game are to be played. At the nth round each expert θ announces a density $p(y^n \mid \theta)$ for y^n. F receives these p_θ's and chooses the density $q(y^n)$ by trying to match the performance of the best expert θ for predicting y^n. Then, N reveals y^n and incurs the loss (or gain) $\log q(y^n)$. How should F use the p_θ's to choose q? Obviously, the best expert will incur the loss $\min_\theta \log 1/p(y^n \mid \theta)$.

In the Bayes version of the game between N and F, F has access to experts that are weighted by a prior $w(\theta)$. So, we want to choose q to minimize the maximum regret

$$\sup_{y^n}\left[\log\frac{1}{q(y^n)} - \inf_\theta \log\frac{1}{w(\theta)p(y^n \mid \theta)}\right] = \sup_{y^n}\left[\sup_\theta \log\frac{w(\theta)p(y^n \mid \theta)}{q(y^n)}\right]. \tag{2}$$

The solution q_{opt} to (2) that we henceforth call the Bayes Shtarkov predictor (for the discrete case) is given in the following theorem.

Theorem 2.1. *The optimum of (2) is*

$$\begin{aligned} q_{\text{opt}}(y^n) &= \arg_q\left[\inf_{q\in\mathcal{P}}\left(\sup_{y^n}\sup_\theta \log\frac{w(\theta)p(y^n \mid \theta)}{q(y^n)}\right)\right] \\ &= \frac{w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))}{\sum_{y^n} w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))}, \end{aligned} \tag{3}$$

where $\tilde\theta$ is the posterior mode.

Proof. This is a straightforward modification of Shtarkov (1987). □

In the continuous case, the sum is replaced by the corresponding integral and (3) becomes

$$q_{\text{opt}}(y^n) = \frac{w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))}{\int w(\tilde\theta(y^n))p(y^n \mid \tilde\theta(y^n))dy^n}. \tag{4}$$

Our next result gives sufficient conditions for (4) to exist and is adapted from Rissanen (1996).

Theorem 2.2. *Assume the following.*

(i) *Let $I_n(\theta)$ be the nth stage Fisher information and suppose there is an $I(\theta)$ so that*

$$I_n(\theta) = -\frac{1}{n} E\left[\frac{\partial^2 \log p(Y^n \mid \theta)}{\partial\theta_i \partial\theta_j}\right] \to I(\theta) \quad \text{as } n \to \infty,$$

and $\exists c_1, c_2$ so that $0 < c_1 \le |I(\theta)| \le c_2 < \infty$ for all $\theta \in \Theta$.

(ii) *The elements of $I(\theta)$ are continuous in Θ.*

(iii)

$$\int_\Theta \sqrt{|I(\theta)|}\, d\theta < \infty.$$

(iv) *The posterior mode, $\tilde{\theta}$, satisfies the central limit theorem,*

$$\xi = \sqrt{n}(\tilde{\theta}(y^n) - \theta) \xrightarrow{L} N(0, I^{-1}(\theta)),$$

uniformly for $\theta \in \Theta$.

(v)

$$I(y^n, \tilde{\theta}) = \left(-\frac{1}{n}\left\{\frac{\partial^2 \log p(y^n \mid \theta)}{\partial\theta_i \partial\theta_j}\right\}_{\theta=\tilde{\theta}}\right)_{i,j=1,\ldots,k} < C_0 < \infty,$$

where C_0 is a positive-definite matrix. In addition, the family

$$I_{ij}(y^n, \theta(\xi)) = -\frac{1}{n}\frac{\partial^2 \log p(y^n \mid \theta(\xi))}{\partial\xi_i \partial\xi_j},$$

as a function of the standardized variable ξ is equicontinuous at $\xi = 0$ for $n \ge 1$, $1 \le i, j \le k$.

Then, the integral in (4) is finite.

Remark. For IID processes, Ferguson (2002) gives conditions under which assumption (iv) holds.

Proof. See Appendix A. □

Regardless of how $q_{\text{opt}}(y^n)$ is computed, Bayes Shtarkov predictors are ratios

$$q_{\text{opt}}(y_{n+1} \mid y^n) = \frac{q_{\text{opt}}(y^{n+1})}{q_{\text{opt}}(y^n)}, \tag{5}$$

and can be used prequentially.

In general, using the Bayes Shtarkov predictor requires that we compute the normalizing constant, i.e., the denominator, in (3) or (4). Here, however, we will only use the mode of (5) (see Section 3 for a justification). So, it will be enough to ignore computing the denominator in (3) or (4). For the sake of completeness, we indicate in Appendix B how the denominator in (3) or (4) would be computed if it were desirable to use, say, the mean or median of (5) rather than its mode.

2.1.2. Bayes Shtarkov predictors with side information

So far we have not included any explanatory variables when predicting Y_{n+1}. However, it is common in practice to have explanatory variables; these are usually called side information when Shtarkov predictors are used since the functional form of the dependence of Y_{n+1} on x_{n+1} is unspecified.

When the Y's and x's are discrete, q_{opt} can be found as follows. Let $(x_i, y_i)|_{i=1}^n$ where $x_j \in \mathcal{X} = \{1, 2, \ldots, M\}$ and $y_j \in \mathcal{Y} = \{1, 2, \ldots, K\}$, $j = 1, \ldots, n$. Now divide y^n into M subsequences y^{n_m} corresponding to each value $x = m$, $m = 1, \ldots, M$, i.e., y^{n_m} is the subsequence of y^n for which the corresponding value of the explanatory variable is m. The form of q_{opt} derived in Xie and Barron (2000) Sec. IX is stated in the next result; see also Cesa-Bianchi and Lugosi (2006) Chap. 9.

Theorem 2.3. *Let $q_{\text{opt}}(y^{n_m})$ be of the form given in Theorem 2.1. Then,*

$$q_{\text{opt}}(y^n \mid x^n) = \prod_{m=1}^{M} q_{\text{opt}}(y^{n_m}).$$

General Bayes Shtarkov predictors for continuous Y's or x's do not seem to have been derived except in the sense that they can be regarded as limits of discrete cases. However, Cesa-Bianchi and Lugosi (2006) Chap. 11 review prediction with side information using restricted families of predictors.

2.2. Stacking

Stacking was first introduced by Wolpert (1992) and studied primarily as a predictor in numerous contexts such as regression (Breiman, 1996b; Clarke, 2003; Sill et al., 2009), classification and distance learning (Ting and Witten, 1999; Ozay and Vural, 2012), and density estimation (Smyth and Wolpert, 1999). Stacking has also been used to estimate error rates (Rokach, 2010).

The basic idea is that if J candidate signal plus noise models of the form $Y = f_j(x) + \epsilon$ for $j = 1, \ldots, J$ are available then they can be usefully combined to give the predictor

$$\hat{Y}_{stack}(x) = \sum_{j=1}^{J} \hat{w}_j \hat{f}_j(x),$$

where $\hat{f}_j$ is an estimate of f_j. The $\hat{w}_j$'s are obtained by invoking an optimality property similar to cross-validation (CV). More formally, let $\hat{f}_{j,-i}$ be the estimate of f_j using $n-1$ of the n data points and dropping the ith one. Then the estimated weight vector $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_J)$ is

$$\hat{w} = \arg\min_{w \in \mathbb{R}^J} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} w_j \hat{f}_{j,-i}(x_i) \right)^2. \tag{6}$$

Expression (6) corresponds to leave-one-out CV but can be readily modified to correspond to leave-K-out CV. A Bayesian interpretation of stacking can be found in Le and Clarke (2015).

2.3. Bagging

Bagging ('bootstrap aggregating'), introduced by Breiman (1996a), is a general strategy to improve the accuracy of model-based predictors. Usually, the model is thought to be good in the sense of being unbiased but gives predictors that are highly variable so that bagging will help stabilize it. The basic strategy is as follows. Given a sample, fit a model $\hat{f}(x)$ and consider predicting the response for a new value of the explanatory variable x. A bagged predictor for Y at x is found by drawing B bootstrap samples from the training data, using each sample to produce an $\hat{f}_b(x)$, and taking an average

$$\hat{Y}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x).$$

Bagging has received much attention and is frequently used, especially in classification. However, there remains relatively little understanding how bagging works apart from the results in Breiman (1996b) and Buhlmann and Yu (2002). It has also been argued that bagging is asymptotically a form of BMA, see Le and Clarke (unpublished).

3. Presentation of computational results

In this section we apply the techniques described in Section 2 to five data sets. All of these data sets have explanatory variables that we have only used when benchmarking our ten nonlinear predictors to additive models. For simplicity, we chose the two explanatory variables that had the largest correlations to the response variable. The reason we reduced to two variables, generically denoted x_1 and x_2, was that the techniques based on Bayes Shtarkov predictor had to be discretized to be computed and this required that no cells be empty to ensure that the right hand side of Theorem 2.3 would always be well-defined as a density. When we tried more variables, or a finer discretization, Bayes Shtarkov predictors were impossible to compute; in the results below there are several cases where we had to make other adjustments so the Bayes Shtarkov predictors could be found. Unless remarked otherwise, our Bayes Shtarkov predictors use a discretization of the response into 20 cells and a discretization of single variable side information into 20 cells but discretization of two variable side information into 16 cells (four for each variable). In all cases, the appropriate percentiles were used as break points for the discretizations. In addition, we always used the Dirichlet distribution with $\alpha = 1$ as a prior for choosing the hyper-parameters in the multinomial resulting from the discretization of Y. This is just the uniform distribution over the parameters in the multinomial; computational details are given in Appendix B.

The ten nonlinear methods that we compared are as follows. Six came from using stacking and bagging with GPP's, NW's, and SVM's. All kernels used were the default radial basis function kernels with the analog of the variance/bandwidth estimated internally to the R programs, in this case `kernlab`, `np` and `e1071`. The other four were Bayes Shtarkov predictors with (i) no side information, (ii) one of two variables as side information (two cases), and (iii) two variables as side information. The two variables chosen as side information for the Bayes Shtarkov were used in all cases of GPP's, NW's, and, SVM's to improve comparability of results. For the stacking point predictor we used the weights from the optimization in Section 2.2 and for the bagging predictors we used the formula in Section 2.3. We obtained point predictors from the Shtarkov optimization by choosing the mode of $q_{\text{opt}}(\cdot \mid y^n)$ in (5). This is reasonable because the mode typically summarized the location of the $q_{\text{opt}}(\cdot \mid y^n)$'s better than the median or mean did.

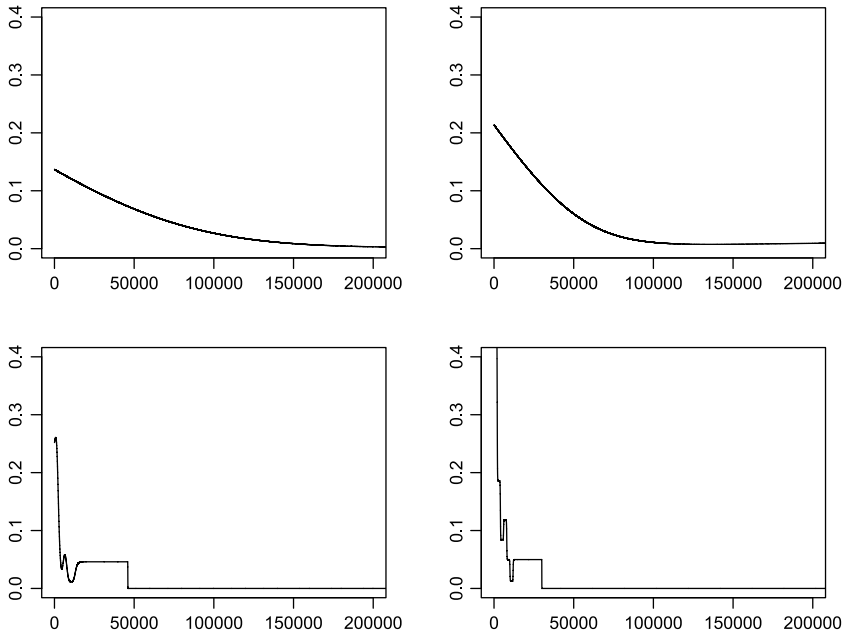

Fig. 1. Plots of the conditional density (5) for the 517th data point given the previous 516 data points for the data set Online News Popularity.

An example of this is seen in Fig. 1 in which graphs of the final predictive densities from our first example, i.e., the conditional densities (5) for our first data set in which the four panels correspond to the four types of side information that we considered. Since the densities in Fig. 1 are strongly skewed to the right, the mean of (5) is not reasonable. Also, the median often occurs where the density is very low. Moreover, even though the median may naively seem the most representative of the location of $q_{\text{opt}}(\cdot \mid y^n)$ among the common location measures, using the discretization usually makes the midpoint of the modal cell well-defined and a better predictor than the midpoint of the median cell. Since graphs similar to Fig. 1 can be generated for the other data sets we used, and strongly skewed graphs such as those in the bottom row of Fig. 1 predominated, the mode seemed best for general usage even though it is not ideal. A more delicate analysis would adaptively choose the best predictor for a given stage in the predictive sequence. However, this is difficult to automate and confounds the comparison of methods with the choice of predictor.

The final two methods we included in our comparisons were additive models using the Bayes LASSO and the horseshoe prior. We used these in two ways. First, we restricted them to the same two explanatory variables as we used in each example. Then we recomputed the CPE's using all the explanatory variables thereby allowing the Bayes LASSO or horseshoe prior to do the variable selection automatically.

As noted earlier, our results suggest that when the Shtarkov predictors can be found effectively, they are best for genuinely $\mathcal{M}$-open data sets. However, the impediments to computing increase with the number of iterations i.e., the number of predictions to be made is high, and the discretization required to compute good approximations may be too fine. Consequently, to obtain some of our comparative statements about Bayes Shtarkov predictors we have had to limit the number of iterations. One of the unusual features of the Shtarkov predictors is that often side information i.e., explanatory variables, was harmful. This may be an artifact of the computing or it may be that the side information was misleading. We return to this issue in Section 4.

In the next subsections we present our CPE results for five $\mathcal{M}$-open data sets, namely, Online News Popularity, Abalone Female, CompActiv, Soil Moisture, and Abalone Male data sets. We separated the entire Abalone data set into Abalone Male and Abalone Female because sex had a large effect on size and we only permitted two explanatory variables. Wherever possible we used a sample size of 517 with a burn-in of 267 and hence 250 predictions could be used to calculate the CPE. (This was chosen as reasonable given the wide variety of data sets we explored.)

3.1. *Online News Popularity*, $n = 517$

As our first example, consider the Online News Popularity data set publicly available from the UC Irvine Machine Learning Repository. There are 58 non-trivial explanatory variables related to the number of shares in social networks (popularity). We took x_1 to be the 'maximum of the average keyword shares' and x_2 to be the 'average of the average keyword shares'; details and references can be found at http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity. The actual sample size is 39 797 but for computational convenience, we randomly selected $n = 517$ data points.

The CPE's for the 14 predictors are in Tables 1, 2, and 3. The lowest three CPE's are in bold and the lowest CPE is starred. The numbers in the tables are truncated to the four most significant digits; fewer digits are given only when this is not possible.

Table 1
CPE's for six averaging methods for Online News Popularity, scaled in millions.

	GPP using x_1 and x_2	NW using x_1 and x_2	SVM using x_1 and x_2
Stacking	8752	9764	5187
Bagging	9478	9487	5597

Table 2
CPE's for four Shtarkov predictors for Online News Popularity, scaled in millions.

Sht, no side info.	Sht, side info. x_1 and x_2	Sht, side info. x_1	Sht, side info. x_2
5062	3432×10	**4904**	**4806***

Table 3
CPE's for Bayes LASSO and Bayes horseshoe predictors for Online News Popularity, scaled in millions.

Bayes LASSO using x_1 and x_2	Bayes horseshoe using x_1 and x_2
1067×10	9988
Bayes LASSO using all variables	Bayes horseshoe using all variables
2905×10	3028×10

Table 4
CPE's for six averaging methods for Abalone Female.

	GPP using x_1 and x_2	NW using x_1 and x_2	SVM using x_1 and x_2
Stacking	2041	2050	**1467**
Bagging	2073	2144	**1684**

Table 5
CPE's for four Shtarkov predictors for Abalone Female. In this case, we used 50 iterations, adding five data points per iteration. We then multiplied the result by five.

Sht, no side info.	Sht, side info. x_1 and x_2	Sht, side info. x_1	Sht, side info. x_2
650*	9613	8441	1534×10

Table 1 shows that among the model average predictors, stacking with SVM's does best and bagging with SVM's is second best. The best results, however, are seen in Table 2. The Shtarkov predictor with no side information or with one of x_1 and x_2 did best. The Shtarkov predictor with both x_1 and x_2 as side information was the worst of the ten methods. We suggest that this occurs because when one variable is used as side information we discretized into 20 cells but when two variables were used as side information we discretized each of them into four cells giving 16 total. Had we been able to use a substantially finer discretization for the two variable side information predictor, and been able to compute it effectively, it might have done best. For instance, if we used 20 cells for x_1 and for x_2 we would have 400 cells total and many would be empty, for our sample size, making computation impossible.

Table 3 shows that conventional additive models do very poorly for complex data. Moreover, permitting automatic variable selection in Bayes LASSO or horseshoe gives the worst results. This pattern suggests that combining variable selection with prediction in one additive procedure is generally going to give inadequate predictive performance. In the one example where this ordering was reversed (Abalone Male), the performance of all four additive methods was so poor none were worth considering.

3.2. *Abalone Female*, $n = 517$

As our second example, we consider the Abalone Female data set publicly available from the UC Irvine Machine Learning Repository. There are 7 non-trivial explanatory variables related to the age of a female abalone. We took x_1 to be 'wholeweight' and x_2 to be 'shellweight'; details and references can be found at http://archive.ics.uci.edu/ml/datasets/Abalone. The sample size is 1307 but for computational convenience, we randomly selected $n = 517$ data points. The CPE's for the ten predictors are in Tables 4 and 5. The lowest three CPE's are in bold and the lowest CPE is starred.

Table 4 shows that, as expected, stacking with SVM's and bagging with SVM's did best among the six averaging methods. When we computed the Shtarkov predictors, we encountered problems with data storage because some cells, while not void, had few points in them resulting is very small values that were below the accuracy threshold of our computing. As a result, the computations for the Shtarkov predictors for Abalone Female could not be done as for the first data set.

Table 6
CPE's for Bayes LASSO and Bayes horseshoe predictors for Abalone Female.

Bayes LASSO using x_1 and x_2	Bayes horseshoe using x_1 and x_2
1672×10	1671×10
Bayes LASSO using all variables	Bayes horseshoe using all variables
2242×10	2281×10

Table 7
CPE's for four Shtarkov predictors for Abalone Female. In this case, we used ten iterations, adding five data points per iteration.

Sht, no side info.	Sht, side info. x_1 and x_2	Sht, side info. x_1	Sht, side info. x_2
556	601	686	707

Table 8
CPE's for six averaging methods for CompActiv.

	GPP using x_1 and x_2	NW using x_1 and x_2	SVM using x_1 and x_2
Stacking	4034×10	$\mathbf{1759 \times 10}$	$\mathbf{1382 \times 10}^*$
Bagging	4159×10	$\mathbf{1983 \times 10}$	2062×10

Table 9
Row II: CPE's for three Shtarkov predictors for CompActiv. In this case, we used 50 iterations, adding five data points per iteration. We then multiplied the resulting CPE by five. Row III: CPE's for four Shtarkov predictors for CompActiv. In this case, we used ten iterations, adding five data points per iteration.

Sht, no side info.	Sht, side info. x_1 and x_2	Sht, side info. x_1	Sht, side info. x_2
	1823×10^2	1098×10^2	1346×10^2
1780	1226	1263	1408

Instead, at each iteration we added five data points rather than one so that the number of predictions was one-fifth of before, i.e., 50 rather than 250. Thus, after the burn-in of 267, we predicted y_{268} and then added five more data points to predict y_{273}, and so on. To compensate for this, we multiplied the CPE's by five; these values are in Table 5 and can be regarded as roughly comparable to those in Table 4. If this reasoning is valid, the Shtarkov predictor with no side information is best and the poor performance of the other three Shtarkov predictor can be explained as before: The discretization is so crude that the extra information is misleading.

Table 6 is qualitatively identical to Table 3. Hence, additive model predictors may be ignored in this example.

3.2.1. *Abalone Female computations for later comparisons*

For the three data sets CompActiv, Soil Moisture, and Abalone Male we were unable to compute the Shtarkov predictors for 250 iterations and in some cases not even for 50 iterations. So, to permit comparisons for those three data sets with Abalone Female we record Table 7. We chose a new set of 517 data points at random from the original 1307. We used a burn-in of 267 and performed the same computations as went into Table 5 but stopped at ten iterations (five extra data points per iteration) and recorded the CPE. We get similar results in that side information is not helpful at this level of discretization. However, the variability of the CPE is seen in that without side information the 10-iteration Shtarkov CPE is not much smaller than the 50-iteration Shtarkov CPE so it likely grows slowly with the number of iterations. However, when side information is used, it is damaging at a much faster rate.

3.3. *CompActiv*, $n = 517$

As our third example, we consider the CompActiv data set. The response variable is the portion of time that CPU's run in user mode. There are twenty one explanatory variables related to the response variables; we took x_1 to be the number of page faults caused by address translation and x_2 the number of disk blocks available for page swapping; the data, details, and references can be found at www.cs.toronto.edu/~delve/data/comp-activ. The sample size is 8192 but for computational convenience, we randomly selected $n = 517$ data points. CPE's for six of the ten predictors are in Table 8. CPE's for the other four (Shtarkov) predictors are in Table 9 for two settings. The lowest three CPE's are in bold and the lowest CPE is starred. Since the predictors from the additive models again performed so poorly (see Table 10) we ignore them for the purposes of the present discussion.

Table 8 shows that, as expected, stacking SVM's has the lowest CPE among the model average predictors. In this example, however, the second and third lowest CPE's are from stacking or bagging with NW. Since the data are complex, it is no

Table 10
CPE's for Bayes LASSO and Bayes horseshoe predictors for CompActiv.

Bayes LASSO using x_1 and x_2	Bayes horseshoe using x_1 and x_2
7103×10^2	7060×10^2
Bayes LASSO using all variables	Bayes horseshoe using all variables
1702×10^3	1707×10^3

Table 11
CPE's for six averaging methods for Soil Moisture.

	GPP using x_1 and x_2	NW using x_1 and x_2	SVM using x_1 and x_2
Stacking	359	356	**339***
Bagging	377	**356**	**349**

surprise that sometimes a method that does not involve stacking or SVM's does reasonably well. However, note that the CPE for stacking SVM's is much lower than the second or third best methods.

In this example, it was impossible to compute any of the Shtarkov predictors for all 250 single data point iterations so we had no choice but to consider 50 iterations (five data points per iteration) and scale up the results by multiplying by five as in row II of Table 9. Indeed, even for this restricted case, we were unable to compute the CPE in the absence of side information. However, the CPE's in this case were much worse than for the predictors in Table 8. This shows that naively scaling is inappropriate. Hence we regard the performance of Shtarkov in row III of Table 9 are most reliable and representative.

To get a complete series of CPE's that we could compare we had to reduce to ten iterations as in Table 7. In this setting, we argued Bayes Shtarkov is the best. Note that as in the other examples, naively scaling errors give results that are not representative. What seems to be more reasonable i.e., representative of the CPE performance of Bayes Shtarkov is that when the CPE is small, it is much more accurate and stable than when it is extremely large. Otherwise put, it is unreasonable to compare these CPE's with those in Table 8 or with row II in Table 9 because the degree of extrapolation is so high.

In the present case, using side information x_1 and x_2 is best. This corresponds to our usual intuition that more information is better but this is the only example where two side variables actually produce the best Shtarkov predictor given the fineness of discretization we were able to implement computationally. As before, the main impediment to getting good performance from the Shtarkov predictors remains computational – too many numbers that are too small arise – but the discrepancy between the CPE's for Shtarkov and, say, stacking SVM's is so large that it is difficult to regard Shtarkov predictors as likely to be best outside the exceptional cases where they can be readily computed. In the Soil Moisture and Abalone Male we also computed Shtarkov predictors using ten iterations and we discuss those comparisons later.

3.4. *Soil Moisture*, $n = 517$

As our fourth example, consider the Soil Moisture data set. The response variable is an interpolated measure of topsoil moisture. There are six explanatory variables; three are for location (two for location on a grid, one for elevation), two for soil electrical resistivity, and one for a standard 'wetness index' that is a function of elevation; see Franz et al. (2015) for a detailed description. We took x_1 as the north–south location of a point where a measurement was taken and x_2 as the wetness index. The sample size is 18 973 but for computational convenience, we randomly selected $n = 517$ data points. CPE's for six of the ten predictors are in Table 11. CPE's for the other four (Shtarkov) predictors are in Table 12 for two settings. The lowest three CPE's are in bold and the lowest CPE is starred. From this point on we omit further discussion of the additive predictors because their performances in terms of CPE were worse than the worst of the other 10 methods by factors of five or higher.

The results for this data set are qualitatively nearly the same as for the CompActiv data set even though the origins of the data are very different. That is, Table 11 shows that stacking with SVM's is best, bagging with SVM's is second best and bagging with NW is third best among the model averages. The actual CPE values are much closer to each other than in Table 8 but the best methods are the same apart from stacking with NW being in third place for the CompActiv data.

As before, the most reliable and representative CPE's from the Bayes Shatarkov method show are the ones that are lowest, even when we had to reduce the number of iterations or increase the number of data points per iterations. Hence, we surmise from Table 12 that Bayes Shtarkov with either sources of side information are predictively the best.

Likewise, row II in Table 12 shows that using x_1 as side information does best (apart from the Shtarkov predictor in the absence of side information that could not be effectively computed). On the other hand, row III shows that taking x_1 as side information is best (but only by a small margin) and for CompActiv using both x_1 and x_2 was best. As before, the impediment to comparing the Shtarkov predictors to the other six predictors directly was computational. One other difference between our analyses of the CompActiv and the Soil Moisture data is that the best of the rescaled Shtarkov predictors for the Soil Moisture data had a CPE of 518 which is higher than the CPE's of the other six predictors, but much closer in performance than for the CompActiv data.

Table 12
Row II: CPE's for four Shtarkov predictors for Soil Moisture. In this case, we used 50 iterations, adding five data points per iteration. We then multiplied the resulting CPE by five. Row III: CPE's for four Shtarkov predictors for Soil Moisture. In this case, we used ten iterations, adding five data points per iteration.

Sht, no side info.	Sht, side info. x_1 and x_2	Sht, side info. x_1	Sht, side info. x_2
	3338	518	701
33	94	21	23

Table 13
CPE's for six averaging methods for Abalone Male.

	GPP using x_1 and x_2	NW using x_1 and x_2	SVM using x_1 and x_2
Stacking	**1776***	2619	**1854**
Bagging	**1842**	2809	2162

Table 14
Row II: CPE's for four Shtarkov predictors for Abalone Male. In this case, we used ten iterations, adding five data points per iteration. Row III: The corresponding results for Abalone Female.

Sht, no side info.	Sht, side info. x_1 and x_2	Sht, side info. x_1	Sht, side info. x_2
321	311	130	298
556	601	686	707

3.5. *Abalone Male*, $n = 517$

As a final $\mathcal{M}$-open example, we consider the Abalone Male data set publicly available from the UC Irvine Machine Learning Repository. There are 7 non-trivial explanatory variables related to the age of male abalone. We took x_1 to be 'height' and x_2 to be 'shellweight'; details and references can be found at http://archive.ics.uci.edu/ml/datasets/Abalone. The sample size is 1528 but for computational convenience, we randomly selected $n = 517$ data points. The CPE's for the ten predictors are in Tables 13 and 14. The lowest three CPE's are in bold and the lowest CPE is starred.

Table 13 shows that the six averaging methods separate into two classes of three based on CPE's. Specifically, bagging and stacking with GPP's had the lowest CPE's while stacking with SVM's was third best and their CPE's were relatively similar. The other three methods did noticeably worse even though they had CPE's that were not too different from each other.

This data set was the most refractory of the data sets we analyzed here. So, it is not necessarily surprising that the stacking GPP's was unexpectedly effective. That is, given sufficiently difficult data, it is no surprise that sometimes a method that does not involve stacking or SVM's does best. It may be that in this example GPP's put more mass around good predictors than NW does because NW is more flexible than GPP's.

One way in which Abalone Male was refractory is that we were unable to do even 50 iterations for the Shtarkov predictors. We were only able to do ten reliably. So, we only compare our results here with row III from Tables 9 and 12 (for CompActiv and Soil Moisture) and computed the corresponding results for Abalone Female. For Abalone Male using x_1 only is best (lowest CPE) and at this number of iterations it seems decisive. For Abalone Female using no side information is best although by a smaller margin. For CompActiv, using x_1 and x_2 is best, but using only x_1 is not much worse. For Soil Moisture, using x_1 only is best, using x_2 only is nearly as good, and the other two cases are decisively worse. From these, the only summary statement that seems possible is that using x_2 alone in a Shtarkov predictor is generally a poor choice, whereas there are cases in which the other three possibilities outperform. This is not a surprise because x_2 has a lower correlation with the response than x_1 does. However, the facts that (i) no side information can be best and (ii) using both x_1 and x_2 can be worst are counterintuitive although we have suggested that part of the problem arises from the discretization.

Note that the Shtarkov solutions for Abalone Female were best and the Shtarkov solutions for Abalone Male are actually better, suggesting again the Bayes Shtarkov predictors are best for $\mathcal{M}$-open data. That is, the principle remains that when the Bayes Shtarkov results are convincing, they outperform the model averaging methods on difficult problems.

4. Conclusions

Overall, we found that in a series of examples, Bayes Shtarkov solutions essentially always performed best for $\mathcal{M}$-open problems when they could be effectively computed. The problems with finding good approximations to exact, formal Bayes Shtarkov solutions are exactly what one would expect: Any discretization has to have enough data points in each cell that the approximation is valid. It is unclear how to assure this in general although many of our examples succeed in doing so and hence give Bayes Shtarkov solutions that outperform model average solutions that should work well for complex data. Unsurprisingly, the finer the discretization, the longer the running time.

We found the surprising result that Bayes Shtarkov predictors with no side information often performed better than the Bayes Shtarkov predictor with side information. We explain this by the increased problems due to discretization that occur

when more explanatory variable are used although other explanations may be possible, too. Likely these points apply to the non-Bayes Shtarkov predictors as well. Indeed, examples not shown here suggest that (i) the more cells that are used in the discretization, the better, provided they are populated, and (ii) the effect on the CPE can be very large—certainly large enough to change the performance ordering of the techniques.

As a secondary point, we found that among the six non-Shtarkov predictors stacking with SVM's was always one of the best choices. For less difficult data sets (not examined here) we also found that among these six predictors that stacking with NW was the best choice. Consequently, we recommend that if the Bayes Shtarkov (or non-Bayes Shtarkov) predictors are infeasible to compute, these other predictors may be next best.

Our comparisons were made in squared error distance and hence our conclusions would likely be valid for many other L^p-type distances as well. Would they be the same for, say, a general invariant measure of prediction such as log-loss? The answer is maybe. The reason is that the performance of predictive techniques can vary widely depending on which invariant measure is used. For instance, Gneiting et al. (2007) and Gneiting (2011) recognize different scoring functions can lead to different predictors being optimal and that bringing in scoring functions without a specific motivation is merely introducing more variability that will be similar to model uncertainty.

Here, log-loss is the sense of distance used to derive the Shtarkov-type predictors so it would be circular to argue 'optimizing with respect to log-loss gives something optimal with respect to log-loss'. Hence, there is good reason to use a different distance. Moreover, log-loss does not make sense for point predictors that do not arise from a density such as our use of SVM's or GPP's. So, while we have defaulted to squared error, we recognize that its main justification stems from the Prequential Principle that we think is a fundamental link between observables and predictions.

To conclude, let us clarify the concept of 'model' as it has been used in an atypical sense throughout this paper. One point of view is that a model is any computable probability distribution. We do not disagree with this, but we do prefer a more general setting in which a model is essentially an action, possibly a conditional action e.g., given side information, in an action space, in a predictive decision theory problem. The problem does not even have to be fully specified. At root, for us, a model is something that emits predictions that can be compared with outcomes of a data generator. Thus, $q_{\text{opt}}(y^n)$ is a probability density and hence a model—even though it does not correspond to a stochastic process. (Marginalizing out Y_n from $q_{\text{opt}}(y^n)$ does not yield $q_{\text{opt}}(y^{n-1})$.) Also, SVM's are merely the solution to an optimization problem. Here we regard them as models even though (i) they do not correspond to a computable distribution and (ii) there is no obvious sense in which they converges to a limit as the sample size increases (due to the presence of x_i in the argument of the kernel function). Indeed, the dependence of an SVM on the data make it incompatible with the traditional statistical notion of a parametric or non-parametric family.

Our general notion of a model seems warranted by the complexity and lack of systematic properties data sets such as those we have analyzed here and is consistent with regarding the data as coming from a data generator that is not a priori describable in any particular terms.

Acknowledgments

Le gratefully acknowledges the extensive support provided by Holland Computing Center and Clarke gratefully acknowledges support from NSF grant # DMS-1419754.

Appendix A. Proof of Theorem 2.2

The proof in Rissanen (1996) still holds up to the inequality labeled (26) with $\hat{\theta}$ replaced by $\tilde{\theta}$. That is, we suppose the k-dimensional parameter space is discretized into k-dimensional rectangles $R(\theta^d, d_n)$ with axes parallel to the axes of the parameter space, sidelength r_n, and centered at values denoted θ^d. Write $d = d_n = r_n/\sqrt{n}$ so that $d = o(1/\sqrt{n})$ if $r = r_n \to 0$. To be specific, let $\tilde{\theta}^d = \tilde{\theta}^d(y^n)$ i.e., $\tilde{\theta}$ quantized to precision d. Now, for $\tilde{\theta} \in R(\theta^d, d_n)$, Taylor expanding gives

$$\log \frac{w(\tilde{\theta})p(y^n \mid \tilde{\theta})}{w(\theta_d)p(y^n \mid \theta_d)} = (\theta_d - \tilde{\theta})^T \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log w(\theta^*) \right)_{i,j} (\theta_d - \tilde{\theta}) - \frac{n}{2}(\theta_d - \tilde{\theta})^T \hat{I}(\theta^*)(\theta_d - \tilde{\theta})$$

for some θ^* on the line joining $\tilde{\theta}$ and θ_d where $\hat{I}(\cdot)$ denotes the empirical Fisher information matrix. Re-arranging and upper bounding give

$$\log \frac{p(y^n \mid \tilde{\theta})}{p(y^n \mid \theta^d)} \leq Cr^2,$$

for some positive constant C using both clauses of Assumption (v). Let

$$P_d(\theta_d) = \int_{\{\tilde{\theta}(y^n) \in R(\theta^d, d_n)\}} p(y^n \mid \theta^d) \mathrm{d}y^n.$$

Since $p(\cdot \mid \theta)$ is a probability density for every θ,

$$\hat{p}_d(y^n) = \frac{p(y^n \mid \tilde{\theta}^d)}{\sum_{\theta_d} P_d(\theta_d)}$$

is well defined, and we can write the decomposition

$$\log \frac{w(\tilde{\theta}^d)p(y^n \mid \tilde{\theta}^d)}{\hat{p}_d(y^n)} = \log \frac{w(\tilde{\theta})p(y^n \mid \tilde{\theta})}{\hat{p}_d(y^n)} + \log \frac{p(y^n \mid \tilde{\theta}^d)}{p(y^n \mid \tilde{\theta})} + \log \frac{w(\tilde{\theta}^d)}{w(\tilde{\theta})}. \tag{7}$$

Since $w(\theta)$ is continuous, the last term in (7) satisfies $|\log w(\tilde{\theta}^d)/w(\tilde{\theta})| \leq \eta$ for any $\eta > 0$ provided the discretization is fine enough.

Given (7), the proof in Rissanen (1996) establishes there is a $c > 0$ and a $K > 0$ so that as n increases

$$\left| \log \frac{p(y^n \mid \tilde{\theta})w(\tilde{\theta})}{\hat{p}_d(y^n)} - \frac{k}{2}\log n2\pi - \log \int_\Theta \sqrt{|I(\theta)|}d\theta \right| \leq Kr + (c + C)r^2 + \eta,$$

cf. inequality (27) in Rissanen (1996). The rest of the proof in Rissanen (1996) holds to give that for $\epsilon(r) = Kr + (c + C)r^2$,

$$e^{-\epsilon(r)-\eta} \leq \frac{\int w(\tilde{\theta}(y^n))p(y^n \mid \tilde{\theta}(y^n))dy^n}{(n/2\pi)^{k/2}\int_\Theta \sqrt{|I(\theta)|}d\theta} \leq e^{\epsilon(r)+\eta},$$

as n increases, thereby establishing the theorem.

Appendix B. Computing the constant in a Shtarkov solution

As noted in Section 2.1.1, and argued at the beginning of Section 3, we used the mode of the Shtarkov solution. Therefore we did not need to compute the constant in the denominators of (3) or (4). However, for predictors such as the mean or median, it is necessary to compute the constant. We start with the case of discrete Y since it is easier than the case of continuous Y. In our computing, we used the discrete case only because our algorithms only permit discrete side information.

For the discrete case, the sum in (3) is over all $|\mathcal{Y}|^n$ terms which is intractable when n is large. Kontkanen and Myllymaki (2007) and Barron et al. (2014) have proposed algorithms for computing the denominators for frequentist Shtarkov predictors. Here, we extend Kontkanen and Myllymaki (2007) to the Bayes Shtarkov case, (3) and (4), since it is more general. We also describe the Roos (2008) method for the continuous case for the sake of completeness.

Case I (*Discrete Y*). Let $\mathcal{Y} = \{1, 2, \ldots, k, \ldots, K\}$, h_k be the number of occurrences of k in y^n, and the prior $w(\theta)$ be the Dirichlet distribution $Dir(\alpha, \ldots, \alpha)$. Then $\tilde{\theta}$ is

$$\tilde{\theta}_k = \frac{\alpha + h_k - 1}{\alpha K + n - K}.$$

We want to compute the denominator in (3),

$$S(K, n) = \sum_{y^n} w(\tilde{\theta}(y^n))p(y^n \mid \tilde{\theta}(y^n)).$$

Arguments similar to those used (Kontkanen and Myllymaki, 2007) give

$$S(K, n) = \sum_{n_1+n_2=n} \frac{n!}{n_1!n_2!}\left(\frac{\alpha K_1 + n_1 - 1}{\alpha K + n - 2}\right)^{n_1}\left(\frac{\alpha K_2 + n_2 - 1}{\alpha K + n - 2}\right)^{n_2} S(K_1, n_1)S(K_2, n_2),$$

where $K_1 + K_2 = K$, and

$$S(K + 2, n) = S(K + 1, n) + \frac{n}{K}S(K, n).$$

Therefore, we have the following algorithm for computing $q_{\text{opt}}(y^n)$; the running time of this algorithm is $O(n + K) = O(n)$ i.e., linear, provided K is fixed.

Step 1. Count the occurrence $h_1, \ldots, h_K$ form the sequence y^n.
Step 2. Compute the numerator in (3) from

$$w(\tilde{\theta}(y^n))p(y^n \mid \tilde{\theta}(y^n)) = \prod_{k=1}^{K}\left(\frac{\alpha + h_k - 1}{\alpha K + n - K}\right)^{\alpha+h_k-1}.$$

Step 3. Set $S(1, n) = 1$.
Step 4. Compute

$$S(2, n) = \sum_{n_1+n_2=n} \frac{n!}{n_1!n_2!}\left(\frac{\alpha + n_1 - 1}{2\alpha + n - 2}\right)^{n_1}\left(\frac{\alpha + n_2 - 1}{2\alpha + n - 2}\right)^{n_2}.$$

Step 5. For $k = 1$ to $K - 2$, compute

$$S(k+2, n) = S(k+1, n) + \frac{n}{k} S(k, n).$$

Step 6. Set

$$q_{\text{opt}}(y^n) = \frac{\text{result in Step 2}}{\text{result in Step 5}}.$$

Case II (*Continuous Y*). As noted in Roos (2008), the integral in (4) can only be solved in closed form for some specific models so it is important to have a more general computational procedure. If we do not discretize Y, Roos (2008) provides a Monte Carlo style approximation. Implementing this would require choosing a specific form for the 'experts' $p(y^n \mid \theta)$. Choosing such a form is akin to model selection and hence an open question although it is tempting to suggest that the Pearson distributions, see Giuard (1984) for a summary, can be regarded as a sort of 'universal' parametric family.

Let $f(y^n) = w(\tilde{\theta}(y^n))p(y^n \mid \tilde{\theta}(y^n))$ assuming p is known. As a default, the natural choice for w is an objective prior such as Jeffreys' truncated to a compact set of θ's and normalized. Now, the denominator $S = \int_{\mathcal{Y}^n} f(z)dz$ can be found as follows.

Step 1. Compute the numerator $w(\tilde{\theta}(y^n))p(y^n \mid \tilde{\theta}(y^n))$ in (4).
Step 2. Draw a sample $z_1, \ldots, z_m$ from the distribution f/S by using MCMC (see Roos, 2008) without knowing S.
Step 3. Compute

$$\left(\frac{1}{m|\mathcal{Y}^n|} \sum_{i=1}^{m} \frac{1}{f(z_i)} \right)^{-1}. \tag{8}$$

This converges to S almost everywhere as $m \to \infty$.
Step 4. Set

$$q_{\text{opt}}(y^n) \approx \frac{\text{result in Step 1}}{\text{result in Step 3}}.$$

The estimator in (8) is sometimes called the harmonic mean estimator; see Neal (2008) who argued that in many cases (8) would not perform well due to lack of convergence. In fact, Roos (2008) proves that (8) does converge as $m \to \infty$. The resolution between these two positions may be that the convergence rate is very slow. Effectively, this is another reason why we used discretization. We comment that if there are cases where (8) converges quickly, then, for the continuous case without side information we would be able to bypass the problems with the fineness of the discretization and in principle get better results. However, the continuous case with side information would remain unresolved.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.06.018.

References

Barron, A., Roos, T., Watanabe, K., 2014. Bayesian properties of normalized maximum likelihood and its fast computation. arXiv:1401.7116.
Bernardo, J., Smith, A., 2000. Bayesian Theory. John Wiley & Sons, Chichester.
Breiman, L., 1996a. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 1996b. Stacked regressions. Mach. Learn. 24, 49–64.
Buhlmann, P., Yu, B., 2002. Analyzing bagging. Ann. Statist. 30, 927–961.
Cesa-Bianchi, N., Lugosi, G., 2006. Prediction, Learning, and Games. Cambridge University Press, New York.
Chakraborty, S., Ghosh, M., Mallick, B., 2012. Bayesian nonlinear regression for large p small n problems. J. Multivariate Anal. 108, 28–40.
Clarke, B., 2003. Bayes model averaging and stacking when model approximation error cannot be ignored. J. Mach. Learn. Res. 683–712.
Clarke, B., 2007. Information optimality and Bayesian modelling. J. Econometrics 138, 405–429.
Clyde, M., Iversen, E., 2013. Bayesian model averaging in the M-open framework. In: Damien, P., Dellaportas, P., Polson, N., Stephens, D. (Eds.), Bayesian Theory and Applications. Oxford University Press, Oxford, pp. 484–498.
Dawid, P., 1984. The prequential approach. J. Roy. Statist. Soc. 147, 287–292.
Ferguson, T., 2002. A Course in Large Sample Theory. CRC Press, Florida.
Franz, T., Wang, T., Avery, W., Finkenbiner, C., Brocca, L., 2015. Combined analysis of soil moisture measurements from roving and fixed cosmic ray neutron probes for multiscale real-time monitoring. Geophys. Res. Lett. 42 (9), 3389–3396.
Giuard, V., 1984. Systems of one-dimensional continuous distributions and their application in simulation studies. In: Robustness of Statistical Methods and Nonparametric Statistics. Reidel Publishing Company, pp. 43–52.
Gneiting, T., 2011. Making and evaluating point forecasts. J. Amer. Statist. Assoc. 106, 746–762.
Gneiting, T., Balabdaoui, F., Raftery, A., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 243–268.
Kimeldorf, G., Wahba, G., 1973. Some results on tchebycheffian spline functions. J. Math. Anal. Appl. 33, 82–95.
Kontkanen, P., Myllymaki, P., 2007. A linear-time algorithm for computing the multinomial stochastic complexity. Inform. Process. Lett. 103, 227–233.
Le, T., Clarke, B., 2015. A bayes interpretation of stacking for m-complete and m-open settings in revision for Bayesian Analysis.
Le, T., Clarke, B., 2015. Prediction in m-open problems. (unpublished manuscript).

Neal, R., 2008. The harmonic mean of the likelihood: Worst Monte Carlo method ever. https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/ (Last accessed 05.05.16).
Ozay, M., Vural, F.T.Y., 2012. A new fuzzy stacked generalization technique and analysis of its performance. arXiv:1204.0171.
Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. The MIT Press, Massachusetts.
Rissanen, J., 1996. Fisher information and stochastic complexity. IEEE Trans. Inform. Theory 42, 40–47.
Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33, 1–39.
Roos, T., 2008. Monte Carlo estimation of minimax regret with an application to mdl model selection. In: Information Theory Workshop. pp. 284–288.
Shtarkov, Y., 1987. Universal sequential coding of single messages. Probl. Inf. Transm. 23, 3–17.
Sill, J., Takacs, G., Mackey, L., Lin, D., 2009. Feature-weighted linear stacking. arxiv.org/pdf/0911.0460.
Smyth, P., Wolpert, D., 1999. Linearly combining density estimators via stacking. Mach. Learn. J. 36, 59–83.
Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol. Methods 14, 323–348.
Ting, K.M., Witten, I., 1999. Issues in stacked generalization. J. Artificial Intelligence Res. 10, 271–289.
Vovk, V., 2001. Competitive on-line statistics. Internat. Statist. Rev. 69, 213–248.
Wolpert, D., 1992. Stacked generalization. Neural Netw. 5, 241–259.
Xie, Q., Barron, A., 2000. Asymptotic minimax regret for data compression, gambling, and prediction. IEEE Trans. Inform. Theory 46, 431–445.