Chapter 23
Data-Based Priors for Bayesian Model Averaging

M. Ai, Y. Huang, and J. Yu

Abstract The uncertainty of models is now becoming one of the most important issues in the process of dealing with practical applications. In order to improve reliability and accuracy of inference, one usually adopts the model averaging method instead of selecting a single final model through a model selection procedure. Under the Bayesian framework, two upper bounds of the risk are derived and the posteriors are obtained by minimizing the bounds with a fixed prior. Then we propose two data-based algorithms to get proper priors for Bayesian model averaging in this paper. Simulations show that by using these priors, smaller mean squared prediction errors can be gotten both in synthetic data and real data studies, especially for the data of poor quality.

23.1 Introduction

It is common in practice that the observed data can be described by different models. A standard procedure to make inference is to choose a best model according to some criteria, such as model predictive ability, model fitting ability or many different information criteria like AIC and BIC. After selection, all the inferences and conclusions are made based on the assumption that the selected model is correct.

However, the drawbacks of this approach exist obviously. The selection of one particular model may lead to riskier decisions since it ignores the model uncertainty.

M. Ai (✉) · Y. Huang
LMAM, School of Mathematical Sciences and Center for Statistical Science,
Peking University, Beijing 100871, China
e-mail: myai@math.pku.edu.cn

Y. Huang
e-mail: huangyimin@pku.edu.cn

J. Yu
School of Mathematics and Statistics,
Beijing Institute of Technology, Beijing 100081, China
e-mail: yujunbeta@bit.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_23

In other words, if we choose a wrong model, the consequence will be disastrous. Moral-Benito already pointed out the concern in [8], "From a pure empirical viewpoint, model uncertainty represents a concern because estimates may well depend on the particular model considered." Therefore, combining multiple models to reduce the model uncertainty is very desirable.

As an alternative strategy, model averaging enables researchers to draw conclusions based on the whole universe of candidate models. In particular, researchers estimate all the candidate models and then compute a weighted average of all the estimates for the coefficient on X. There are two different approaches to model averaging in the literature including Frequentist Model Averaging (FMA) and Bayesian Model Averaging (BMA).

Frequentist approaches focus on improving prediction and use weighted mean of estimates from different models while Bayesian approaches focus on the probability that a model is true and consider priors and posteriors for different models. Ref. [4] suggested to use Bayesian inference to reduce the model uncertainty and pointed out the importance of the fragility of regression analysis to arbitrary decisions about the choice of control variables. Bayesian Model Averaging considers model uncertainty through the posterior distribution. The model posteriors are obtained by Bayes' theorem, and therefore allowing for combined estimation and prediction. Compared with the FMA approaches, there are a huge literature on the use of BMA in statistics.

Influenced by [4], most works were concentrated only on the linear models. Ref. [10] extended to generalized linear models by providing a straightforward approximation. For more details, refer to some landmark reviews such as [2, 8, 15] on BMA. Moreover, Refs. [6, 19] gave good estimators of the risk in linear mixed-effects models. For getting the posterior distribution of the weights, Ref. [17] gave a method called SOIL which can well separate the variables in the true model from the rest under some assumptions. However, they used a default prior for the procedure.

The Bayesian approaches have the advantage of using arbitrary domain knowledge through a proper prior. However, they can't guarantee the upper bound of the decision risk without assuming the truth of the prior. The Probably Approximately Correct (PAC) framework, first formulated by [7], was proposed to deal with this problem. It has been widely developed in recent years. Refs. [5, 11] gave tighter bounds in some specific cases. Ref. [1] provided an extended PAC-Bayes bound for learning the proper priors. But, they used the same data for learning the prior and the posterior simultaneously. This issue will make the ability of generalization worse.

There have been many recent developments in model averaging. Refs. [14, 18] presented two criteria, Mallows criterion and jackknife criterion, to determine the weights of model averaging. Their meanings are not as directly as minimizing the upper bound of the risk. They didn't build the relation between the risk and the criteria theoretically. Refs. [6, 19] gave good estimators of the risk in a certain type of models while our work doesn't specify the model type. For getting the posterior distribution of the weights, Ref. [17] gave a method without choosing a proper prior. Ref. [1] provided an extended PAC-Bayes bound for learning the proper

priors. Nevertheless, it involved reusing of the data which increased the probability of overfitting.

In this paper, we propose a specific risk bound under our settings and two data-based methods for adjusting the priors in PAC-Bayes framework. And, two practical algorithms are given accordingly. The main contributions of this work are the following. First, sequential batch sampling method is proposed to deal with the situation that there isn't historical data while the data can be sampled with the rules made by researchers. Second, when the historical data existed, we use similar old tasks to extract the mutual knowledge with the current task for adjusting the priors. Third, two theoretical risk bounds are provided for these two situations respectively. Fourth, empirical demonstration shows that the proposed meta-methods have excellent performances in the numerical studies.

The reminder of this paper is organized as follows. In Sect. 23.2, a standard risk bound and a practical sequential batch sampling method are established for obtaining a better prior in no previous data situation. Section 23.3 proposes the method to deal with historical similar data for the same purpose. Illustrative simulations given in Sect. 23.4 show that our algorithms will lead to more effective prediction and support our theoretical results. For real-world dataset, we apply the proposed methods to two real datasets and confirm the higher prediction accuracy of minimizing risk bound method. Section 23.5 concludes this paper with some discussions. Some proofs of theorems are delegated to the supplementary materials.

23.2 Sequential Adjustment of Priors

In a traditional supervised learning task, the learner needs to find an optimal *model* (or hypothesis) to fit the data, and then uses the learned model to make predictions. In the Bayesian approach, various models are allowed to fit the data. In particular, the learner needs to learn an optimal model *distribution* over the candidate models, and then uses the learned model distribution to make predictions.

More specifically, in a supervised learning task, we are given a set $S = \{(x_i, y_i)\}_{i=1}^n$ of i.i.d. samples drawn from an unknown distribution D over $\mathscr{X} \times \mathscr{Y}$, i.e., $(x_i, y_i) \sim D$. The goal is to find a model h in the candidate model set $\mathscr{H}$, a set of functions mapping features (feature vector) to responses, that minimizes the expected loss function $\mathbb{E}_{(x,y)\sim D} L(h, x, y)$, where L is a bounded loss function. Without loss of generality, we assume L is bounded by $[0, 1]$. In the Bayesian framework, a distribution Q over $\mathscr{H}$ is the purpose instead of searching a specific optimal model $h \in \mathscr{H}$. Therefore, the goal turns to finding the optimal model distribution Q, which minimizes $\mathbb{E}_{h\sim Q}\mathbb{E}_{(x,y)\sim D} L(h, x, y)$. Then one could use the weighted average of the models over $\mathscr{H}$ to make predictions, namely, $\widehat{y} = \mathbb{E}_{h\sim Q} h(x)$. More generally, we further assume that the candidate model set $\mathscr{H}$ consists of K classes of models $\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_K$ with $\mathscr{H} = \bigcup_{k=1}^K \mathscr{M}_k$. Each model class $\mathscr{M}_k$ is associated with a probability w_k, and for each model class $\mathscr{M}_k$, there is a distribution Q_k over $\mathscr{M}_k$. For example, a model class $\mathscr{M}_k$ could be a group of models obtained from the Lasso

method, and the hyper-parameter λ in Lasso follows a distribution Q_k. Another common example is that $\mathscr{M}_k$ is a group of neural networks with a certain architecture, and the weights of neural networks follow a joint distribution Q_k. In this way, the total distribution over $\mathscr{H}$ can be written as $\xi = (w, Q_1, \ldots, Q_K)$, where w consists of $w_1, \ldots, w_K$ with $\|w\|_1 = 1$. The goal of the learning task is to find an optimal distribution ξ which minimizes the expected risk $R(\xi, D) := \mathbb{E}_{h\sim\xi}\mathbb{E}_{(x,y)\sim D}L(h, x, y)$, and then the prediction is made by $\widehat{y} = \mathbb{E}_{h\sim\xi}h(x) = \sum_{k=1}^{K}[w_k \cdot \mathbb{E}_{h\sim Q_k}h(x)]$.

Since sample distribution D is unknown, the expected risk $R(\xi, D)$ cannot be computed directly. Therefore, it is usually be approximated by the empirical risk $\widehat{R}(\xi, S) := \mathbb{E}_{h\sim\xi}\sum_{(x_i,y_i)\in S}L(h, x_i, y_i)/|S|$ in practice, and ξ is learned by minimizing the empirical risk $\widehat{R}(\xi, S)$. When the sample size is large enough, it would be a good approximation. However, in many situations, we don't have so much data, which may lead to large difference between them. Thus, using the empirical risk $\widehat{R}(\xi, S)$ to approximate the expected risk $R(\xi, D)$ is not appropriate any longer.

We first study the difference between the empirical risk $\widehat{R}(\xi, S)$ and the expected risk $R(\xi, D)$. Based on the literature [7], we can obtain an upper bound of their difference which is stated as the following theorem.

Theorem 23.1 *Let ξ^0 be a prior distribution over $\mathscr{H}$ that must be chosen before observing the samples, and let $\delta \in (0, 1)$. Then with probability at least $1 - \delta$, the following inequality holds for all posterior distributions ξ over $\mathscr{H}$,*

$$R(\xi, D) \leq \widehat{R}(\xi, S) + \sqrt{\frac{\mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}(Q_k||Q_k^0) + \ln\frac{n}{\delta}}{2(n-1)}}, \tag{23.1}$$

where n is the cardinality of sample set S, and $\mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence between two distributions.[1]

According to the above theorem, it is clear that only when the sample size n is large, the difference $R(\xi, D) - \widehat{R}(\xi, S)$ can be guaranteed to be small. Thus, minimizing $\widehat{R}(\xi, S)$ may not lead to the minimizer of $R(\xi, D)$, which matches our intuition. To avoid the risk of the approximation, one can minimizes the upper bound of the expected risk $R(\xi, D)$ in stead of using the empirical risk $\widehat{R}(\xi, S)$ as an approximation. In particular, we denote the right hand side of Eq.(23.1) by $\overline{R}(\xi, \xi^0, S)$. Then one can learn the model distribution ξ by minimizing $\overline{R}(\xi, \xi^0, S)$. Intuitively, such choice of ξ for the learning task makes the worst case best.

Theorem 23.1 also indicates that the prior ξ^0 plays an important role. Since the choice of ξ balances the tradeoff between the empirical risk $\widehat{R}(\xi, S)$ and the regularization term, if the prior ξ^0 is far away from the true optimal model distribution ξ^*, the posterior ξ will also be bad. The best situation for optimizing the posterior ξ is that the prior ξ^0 exactly equals to the true optimal model distribution ξ^*. Then, the regularization term disappears. In other words, if there is a good prior ξ^0 which is close to ξ^*, the upper bound $\overline{R}(\xi, \xi^0, S)$ will be small. However, without

[1] $\mathrm{KL}(P||P^0)$ is defined as $\mathbb{E}_{x\sim P}\ln\frac{P(x)}{P^0(x)}$.

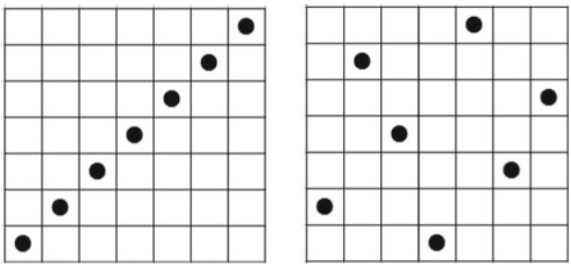
any prior knowledge, one can only use data to help obtain a better prior. The naive method is directly using the non-informative prior as ξ^0 for minimizing $\overline{R}(\xi, \xi^0, S)$ to get the posterior ξ. In this paper, we propose a more carefully designed method to get a better posterior than the naive method. In the following, we consider two different scenarioes for learning the prior. First, the data can be collected adaptively. The learner is allowed to do sampling in rounds and updates the prior distribution after each sampling. In each round, the learner can sample the data according to the prior distribution in the current round. Such iterative procedure updates the prior step by step. Ultimately, compared with dealing the whole data at once, this procedure of adjusting prior leads to a smaller upper bound. Moreover, it also gives an opportunity to choose some good sample sets for reducing the volatility of the estimators which is measured by $v(\xi, D) = \mathbb{E}_x \mathbb{E}_h (h(x) - \mathbb{E}_h h(x))^2$. The function $\widehat{v}(\xi, B) = \frac{1}{|B|} \sum_{x \in B} \mathbb{E}_h (h(x) - \mathbb{E}_h h(x))^2$ is defined to measure the volatility of the posterior ξ at the sample set B. The complete algorithm for sequential batch sampling is shown in Algorithm 6. Second, the data including the new task and other similar old tasks which have been already collected. The sequential sampling method can not be adopted in this scenario. Since the previous tasks are similar with the new task, we could use these old data to learn the prior for the new task. The details will be discussed in Sect. 23.3.

Algorithm 6 Sequential Batch Sampling Algorithm

1: Obtain a sample set B_1 from the sample space $\mathscr{X} \times \mathscr{Y}$ by a initial space-filling design.
2: Get the posterior ξ_1 based on the sample set B_1 by minimizing the risk bound with non-informative prior.
3: **for** $i = 2$ to b **do**
4: Search next sample set B_i ($|B_i| = n_b$) with the large volatility under the current posterior ξ_{i-1}, i.e., $\widehat{v}(\xi_{i-1}, B_i) > \gamma_i$ where γ is a given constant vector.
5: Get the posterior ξ_i based on the sample set B_i by minimizing the risk bound with the prior ξ_{i-1}.
6: **end for**
7: The final posterior is ξ_b.

For Algorithm 6, the data is processed in b steps. First, a space-filling design is used as initial experiment points to reduce the probability of overfitting caused by the unbalanced sampling. Traditional space-filling design aims to fill the input space with design points that are as "uniform" as possible in the input space. The uniformity of space-filling design is illustrated in Fig. 23.1. For next steps, uncertain points are needed to be explored. And, the uncertainty is measured by the volatility v. Hence, the batch with large volatility will be chosen. Note that if we set a huge γ, we will just explore a small region of the input space.

The setting of γ refers to [20]. However, in practice, it is found that this parameter γ does not matter much, since the results are similar with a wide range of γ. This procedure helps to reduce the variance of the estimator which is proved in [20] by sequential sampling. Furthermore, it also helps to adjust the prior in each step which is called learning the prior. The proposition is stated as below.



(a) Nonuniform design (b) Uniform design

Fig. 23.1 The illustration for uniform space-filling design

Proposition 23.1 *For $i = 1, 2, \ldots, b$, let $B_i = S$, ξ^* is the minimizer of the RHS of Eq. (23.1) with non-informative prior ξ^0 and ξ_i are obtained by Algorithm 6, then we have $\overline{R}(\xi_b, \xi_{b-1}, S) \leq \overline{R}(\xi^*, \xi^0, S)$.*

The above proposition can be understood straightforwardly. First, since we adjust the prior through the data step by step, the final prior ξ_{b-1} is better than the non-informative prior. Consequently, it receives the smaller expected risk. Second, we choose the sample sets sequentially with large volatility to do experiments in order to reduce uncertainty. The property is also confirmed in Sect. 23.4.

23.3 Priors Based on Historical Data

As mentioned in Sect. 23.2, when the data of historical tasks and the new tasks have already collected, sampling method can not be used any longer. Still, the learner needs a good prior for the reliable inferences. In order to get a good prior, it is helpful to extract the mutual knowledge from similar tasks. In particular, there are m sample tasks $T_1, \ldots, T_m$ i.i.d. generated from an unknown task distribution τ. For each sample task T_i, a sample set S_i with n_i samples is generated from an unknown distribution D_i. Without ambiguity, we use notation $\xi(\xi^0, S)$ to denote the posterior under the prior ξ^0 after observing the sample set S. The quality of a prior ξ^0 is measured by $\mathbb{E}_{D_i \sim \tau} \mathbb{E}_{S_i \sim D_i^{n_i}} R(\xi(\xi^0, S_i), D_i)$. Thus, the expected loss we want to minimize is

$$R(\xi^0, \tau) = \mathbb{E}_{D_i \sim \tau} \mathbb{E}_{S_i \sim D_i^{n_i}} R(\xi(\xi^0, S_i), D_i).$$

Similar to the single-task case, the above expected risk cannot be computed directly, thus the following empirical risk is used to estimate it:

$$\widehat{R}(\xi^0, S_1, \ldots, S_m) = \frac{1}{m} \sum_{i=1}^{m} \widehat{R}(\xi(\xi^0, S_i^{train}), S_i^{validation}),$$

where each sample set S_i is divided into a training set S_i^{train} and a validation set $S_i^{validation}$.

Consider the regression setting for task T. Suppose the true model is

$$y_T = f_T(x_T) + \sigma_T(x_T) \cdot \varepsilon_T,$$

where $f_T : \mathbb{R}^d \to \mathbb{R}$ is the function to be learned, the error term ε_T is assumed to be independent of X and has a known probability density $q(t), t \in \mathbb{R}$ with mean 0 and a finite variance. The unknown function $\sigma_T(x_T)$ controls the variance of the error at $X = x_T$. There are n_T i.i.d. samples $\{(x_{T,i}, y_{T,i})\}_{i=1}^{n_T}$ drawn from an unknown joint distribution of (x_T, y_T). Assume that there is a candidate model set $\mathscr{H}$. Each of them is a function mapping features (feature vector) to response, i.e., $h \in \mathscr{H} : \mathbb{R}^d \to \mathbb{R}$. To take the information of the old tasks, which can reflect the importance of each $h \in \mathscr{H}$, the following Algorithm 7 is proposed.

Algorithm 7 Historical Data Related Algorithm

1: **for** $i = 1$ to m **do**
2: Using T_i to obtain ξ_i by minimizing the risk bound with non-informative prior.
3: **end for**
4: **for** $i = 1$ to m **do**
5: Randomly split the data S_i into two parts $S_{i,n_i'}^{(1)} = (x_{i,\alpha}, y_{i,\alpha})_{\alpha=1}^{n_i'}$ for training and $S_{i,n_i'}^{(2)} = (x_{i,\alpha}, y_{i,\alpha})_{\alpha=n_i'+1}^{n_i}$ for validation.
6: **for** each $j \neq i$ **do**
7: Obtain estimates $\widehat{f}_{j,n_i'}(x, S_{i,n_i'}^{(1)}), \widehat{\sigma}_{j,n_i'}(x, S_{i,n_i'}^{(1)})$ with prior ξ_j.
8: Evaluate predictions on $S_{i,n_i'}^{(2)}$ and compute

$$E_j^i = \frac{\Pi_{\alpha=n_i'+1}^{n_i} q\left(\frac{y_\alpha - \widehat{f}_{j,n_i'}(x_{i,\alpha})}{\widehat{\sigma}_{j,n_i'}(x_{i,\alpha})}\right)}{\Pi_{\alpha=n_i'+1}^{n_i} \widehat{\sigma}_{j,n_i'}(x_{i,\alpha})}.$$

9: **end for**
10: **end for**
11: Repeat the random data segmentation more times and average the weights E_j^i after normalization to get $w_j^{(i)} (j \neq i)$.
12: Average all the $w_j^{(i)} (j \neq i)$ from $i = 1$ to m to obtain the final weights w_j.
13: The prior learned for a new task is $\xi^* = \sum_{i=1}^m w_i \xi_i$.

This algorithm is based on the cross-validation framework. First, using T_i to obtain the candidate priors ξ_i by minimizing the risk bound with non-informative prior. Cross-validation determines the importance of the priors. The j-th task is divided into two parts randomly. The first part is used to learn the posterior with the prior ξ_i. The second part is to evaluate the performance of the posterior by its likelihood function. This evaluation is inspired by [9]. To simplify the determination of the

weights, Ref. [9] proposed a frequentist approach to BMA. The Bayes' theorem was replaced by the Schwarz asymptotic approximation which could be viewed as using maximized likelihood function as the weights of the candidate models. The $\widehat{\sigma}$ on the denominator of E_j^i makes the weight larger if the model is accurate. This procedure repeats many times for each pair (i, j). Their averages reveal the importance of the priors. In the end, the ξ^* is obtained by weighted averaging them all. the property of this algorithm can be guaranteed by the following theorem.

The following regularization conditions are assumed for the results. First, q is assumed to be a known distribution with 0 and variance 1.

(C1) The functions f and σ are uniformly bounded, i.e., $\sup_x |f(x)| \leq A < \infty$ and $0 < c_m \leq \sigma(x) \leq c_M < \infty$ for constants A, c_m and c_M.

(C2) The error distribution q satisfies that for each $0 < s_0 < 1$ and $c_T > 0$, there exists a constant B such that

$$\int q(x) \ln \frac{q(x)}{\frac{1}{s} q(\frac{x-t}{s})} \mu(dx) \leq B((1-s)^2 + t^2)$$

for all $s_0 \leq s \leq s_0^{-1}$ and $-c_T \leq t \leq c_T$.

(C3) The risks of the estimators for approximating f and σ^2 decrease as the sample size increases.

For the condition (C1), note that, when we deal with k-way classification tasks, the responses belong to $\{1, 2, \ldots, k\}$ which is bounded obviously. Moreover, if the input space is a finite region which often happens in real datasets, most common functions are bounded uniformly. The constants A, c_m, c_M are involved in the derivation of the risk bounds, but they can be unknown in practice when we implement the Algorithm 7. The condition (C2) is satisfied by Gaussian, t (with degrees of freedom larger than two), double-exponential, and so on. The condition (C3) usually holds for a good estimating procedure, like consistent estimators. A model has consistency if the expected risk tends to zero when experimental size tends to infinity. Note that the conditions are satisfied in most situations.

Theorem 23.2 *Assume (C1)–(C3) are satisfied and σ_{T_i} is known. Then, the combined posterior ξ^* as given above satisfies*

$$R(\xi^*, \tau) \leq \inf_j \left(\frac{C_1}{\sum_{i \neq j}(n_i - n_i')} + \frac{C_2}{\sum_{i \neq j}(n_i - n_i')} \sum_{i \neq j} (n_i - n_i') \left[\widehat{R}(\xi_j^*, S_{i,n_i'}^{(2)}) + \sqrt{\frac{\mathrm{KL}(w_j{}^* || w_j) + \sum_{k=1}^K w_{j,k} \mathrm{KL}(Q_{j,k}^* || Q_{j,k}) + \ln \frac{n_i}{\delta}}{2(n_i - 1)}} \right] \right)$$

with probability at least $1 - \delta$, where the constant C_1, C_2 depend on the regularization conditions, π is the initial prior which should be non-informative prior and ξ_j^* is the minimizer of Eq. (23.1) with $\xi^0 = \xi_j$ and $S = S_{i,n_i'}^{(1)}$.

For simplify, we assume that the condition that σ_{T_i} is known in Theorem 23.2. In fact it is not a necessary condition, a more general case and corresponding proof can be found in Appendix.

In this general proof, it can be seen that variance estimation is also important for the Algorithm 7. Even if a procedure estimates f_T very well, a bad estimator of σ_T can substantially reduce its weight in the final estimator. Under the condition (C3), the risks of a good procedure for estimating f_T and σ_T usually decrease as the sample size increases. The influence of the number of testing points n_i' is quite clear. Smaller n_i' decreases the first penalty term but increases the main terms that involve the risks of each j. Moreover, Theorem 23.2 reveals the vital property that if one alternative model is consistent, the combined model will also have the consistency.

23.4 Simulations

In this section, some examples are shown to illustrate the procedure of Algorithms 6 and 7 and confirm Proposition 23.1. The method of minimizing the upper bound in Theorem 23.1 with non-informative prior is denoted by RBM (Risk Bound Method). Also, the SOIL method in [17] is under the comparison. The optimization for RHS of Eq. (23.1) in our algorithms is dealt by gradient descend method. R package "SOIL" is used to obtain the results of the SOIL method. First, we begin with linear models.

23.4.1 Synthetic Data Analysis

Example 23.1 The simulation data $\{(x_i, y_i)\}_{i=1}^n$ is generated for the RBM from the linear model $y_i = 1 + {x_i}^T\beta + \sigma\varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$, $\sigma \in \{1, 5\}$ and $x_i \sim N_d(0, \Sigma)$. For each element Σ_{ij} of Σ, $\Sigma_{ij} = \rho^{|i-j|}$ $(i \neq j)$ or 1 $(i = j)$ with $\rho \in \{0, 0.9\}$. The sequential batch sampling has b steps, and each step uses n/b samples followed Algorithm 6.

All the specific settings for parameters are summarized in Table 23.1, and the confidence level δ in Theorem 23.1 is set to 0.01. The Mean Squared Prediction Error(MSPE) $\mathbb{E}_x|f(x) - \widehat{f}(x)|^2$ and volatility defined in Sect. 23.2 are compared. They are obtained by sampling 1000 samples from the same distribution and computing their empirical MSPE $\sum_x |f(x) - \widehat{f}(x)|^2/10^3$ and volatility. For each model setting with a specific choice of the parameters (ρ, σ), we repeat 100 times and compute the average empirical value. The comparison among RBM, SOIL and SBS(Sequential Batch Sampling) are shown in Table 23.2.

The volatility of SOIL method is the smallest and very close to zero. This phenomenon shows that SOIL is focused on a few models, even just one model when the volatility equals to zero. Consequently, its MSPE is larger than other two

Table 23.1 Simulation settings of Example 23.1

Model	n	d	b	β
1	50	8	5	$(3, 1.5, 0, 0, 2, 0, 0, 0)^T$
2	150	50	5	$(1, 2, 3, 2, 0.75, 0, \ldots, 0)^T$
3	50	50	5	$(1, 1/2, 1/3, 1/4, 1/5, 1/6, 0, \ldots, 0)^T$

Table 23.2 Comparison among RBM, SOIL and SBS of Example 23.1

Model 1	(ρ, σ)	(0, 1)	(0, 5)	(0.9, 1)	(0.9, 5)
MSPE	RBM	2.03	48.23	3.71	53.83
	SOIL	2.13	53.21	2.17	53.21
	SBS	**1.71**	**14.08**	**3.25**	**26.40**
Volatility	RBM	1.64	3.47	1.31	0.49
	SOIL	0	0	0.002	0
	SBS	1.61	7.41	1.03	0.42
Model 2					
MSPE	RBM	1.97	46.26	1.46	35.97
	SOIL	2.01	50.23	1.96	49.78
	SBS	**1.93**	**38.69**	**1.38**	**12.92**
Volatility	RBM	1.60	2.72	3.38	7.48
	SOIL	0	0	0.001	0.01
	SBS	1.46	8.67	3.35	6.74
Model 3					
MSPE	RBM	1.67	42.06	1.24	38.51
	SOIL	1.99	49.80	1.93	47.99
	SBS	**1.65**	**27.32**	**1.23**	**29.44**
Volatility	RBM	0.27	1.54	0.74	3.39
	SOIL	0	0	0.02	0.36
	SBS	0.29	0.47	0.77	4.06

methods. SBS as a modification of RBM has similar results with RBM when σ is small. However, when σ is large, SBS performs much better than RBM. In this situation, the information of data is easily covered by big noises. Hence, a good prior which can provide more information is vital for this procedure.

Next example considers the same comparison but in non-linear models. In last example, the alternative models include the true model, but now the true non-linear model is approximated by many linear models.

Example 23.2 The simulation data $\{(x_i, y_i)\}_{i=1}^{50}$ is generated for the RBM from the non-linear models

1. $y_i = 1 + \sin(x_{i,1}) + \cos(x_{i,2}) + \varepsilon_i$,
2. $y_i = 1 + \sin(x_{i,1} + x_{i,2}) + \varepsilon_i$,

Table 23.3 Comparison among RBM, SOIL and SBS of Example 23.2

		Model 1	Model 2
MSPE	RBM	1.26	1.54
	SOIL	1.42	1.80
	SBS	**1.23**	**1.47**
Volatility	RBM	0.1	0.11
	SOIL	0.07	0.02
	SBS	0.11	0.14

where $\varepsilon_i \sim N(0, 1)$, and $x_i \sim N_8(0, I)$. The sequential batch sampling has 5 steps, and each step uses 10 samples followed Algorithm 6.

The results of Example 23.2 is listed in Table 23.3. Mostly, it is similar with the results of Example 23.1. The difference is that the volatility of SOIL becomes large when the model is completely non-linear. Using linear models to fit non-linear model obviously increases the model uncertainty, since none of the fitting models is correct.

The final example is under the situation that the data has been already collected. Hence, we can't use the SBS method to get the data. However, we have the extra data of many old similar tasks. In particular, we have the data of Example 23.1. Now, the new task is to fit a new model.

Example 23.3 The data of Example 23.1 with $(\rho, \sigma) = (0, 1)$ is given. The new task data $\{(x_i, y_i)\}_{i=1}^{20}$ is generated from the linear model $y_i = 1 + {x_i}^T\beta + \sigma\varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$, $\sigma \in \{1, 2, 3, 4, 5\}$, $\beta = \{1, -1, 0, 0, 0.5, 0, \ldots, 0\}$ and $x_i \sim N_{10}(0, I)$.

The method described in Algorithm 7 is denoted by HDR (Historical Data Related). The results in Fig. 23.2 show the high consistency with the last two examples. When σ is small, the different priors lead to similar result since the current data has key influence. However, when σ is large, the difference between RMB and HDR is huge. The reason is that the current data has been polluted by the strong noise. Hence, a good prior can provide the vital information about the model distribution.

23.4.2 Real Data Study

Here, we apply the proposed methods to two real datasets, BGS data and Bardet data, which are also used in [17].

First, the BGS data is with small d and from the Berkeley Guidance Study (BGS) by [13]. The dataset records 66 boys' physical growth measures from birth to eighteen

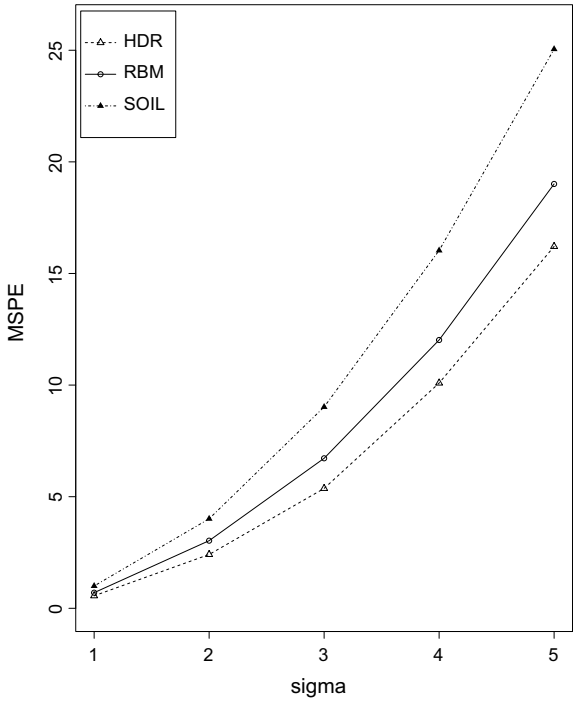


Fig. 23.2 Comparisons among the three methods in Example 23.3

years. Following [17], we consider the same regression model. The response is age 18 height and the factors include weights at ages two (WT2) and nine (WT9), heights at ages two (HT2) and nine (HT9), age nine leg circumference (LG9) and age 18 strength (ST18).

Second, for large d, the Bardet data collects tissue samples from the eyes of 120 twelve-week-old male rats. For each tissue, the RNAs of 31, 042 selected probes are measured by the normalized intensity valued. The gene intensity values are in log scale. Gene TRIM32, which causes the Bardet-Biedl syndrome, is the response in this study. The genes that are related to it are investigated. A screening method [3] is applied to the original probes. This screened dataset with 200 probes for each of 120 tissues is also used in [17].

Both cases are data-given cases that we can't use sequential batch sampling method. For the different setting of d, we assign corresponding similar historical data for two real datasets. The data of model 1 in Example 23.1 for the BGS data with small d. The data of model 3 in Example 23.1 for the Bardet data with large d.

We randomly sample 10 rows from the data as the test set to calculate empirical MSPE and volatility. The results are summarized in Table 23.4. From Table 23.4, we can see that both RBM and HDR have smaller MSPE than SOIL. However, HDR doesn't perform much better than RBM. This can be explained intuitively as follows. In theory, the historical tasks and the current task are assumed that they come from the

Table 23.4 Comparison among RBM, SOIL and HDR in real data

		BGS	Bardet
MSPE	RBM	13.54	0.0054
	SOIL	16.74	0.0065
	HDR	**13.06**	**0.0050**
Volatility	RBM	1.99	0.0013
	SOIL	0.43	0.0013
	HDR	1.84	0.0012

same task distribution. But in practice, how to measure the similarity between tasks is still a problem. Hence, an unrelated historical dataset may provide less information for the current prediction.

23.5 Concluding Remarks

This paper is based on the PAC-Bayes framework to study the model averaging problem. More concretely, the work is about how to assign the proper distribution on the candidate models. The work proposes specific upper bounds of the risks in different situations and aims to minimize them. In other words, it makes the worst situation best. For this purpose, two practical algorithms are provided to solve this optimization under two realistic situations respectively. One is that no previous data can be used, but the experimenters have the opportunity to design the sampling method before the collection of the data. The other one is that much historical data is given, the analysts should figure out a proper method to deal with these data. In the first case, the prior is adjusted step by step. Compared with dealing the whole data at once, this sequential method has the smaller upper bound of the risk. In the second case, using historical similar tasks to extract the information about the prior which is called meta-learning. The meta-learner is for the prior and the base-learner is for the posterior. Both methods are confirmed to be effective in our simulation and real data study.

However, some problems need to be investigated. First, in sequential batch sampling procedure, the volatility is used as a criterion to sample the data. This choice is based on our experience. There may exist other choices that have better results. Second, when a lot of historical data is available, many similar old tasks may be considered to extract more information for learning the new task better. How to define 'similar' is still an open problem. In practice, the similarity isn't measured by the data. Instead, it is judged by experts, which is not expected.

Acknowledgments The authors sincerely thank the editors and a referee for their valuable comments, which further improve this paper. The work is supported by NSFC grant 11671019, LMEQF and Beijing Institute of Technology Research Fund Program for Young Scholars.

Appendix

First, we review the classical PAC-Bayes bound [7, 12] with general notations.

Lemma 23.1 *Let $\mathscr{X}$ be a sample space and $\mathscr{F}$ be a function space over $\mathscr{X}$. Define a loss function $g(f, X) : \mathscr{F} \times \mathscr{X} \to [0, 1]$, and $S = \{X_1, \ldots, X_n\}$ be a sequence of n independent identical distributed random samples. Let π be some prior distribution over $\mathscr{F}$. For any $\delta \in (0, 1]$, the following bound holds for all posterior distributions ρ over $\mathscr{F}$,*

$$\mathbb{P}_S\left(\mathbb{E}_X\mathbb{E}_f g(f, X) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_f g(f, X_i) + \sqrt{\frac{\mathrm{KL}(\rho||\pi) + \ln\frac{n}{\delta}}{2(n-1)}}\right) \ge 1 - \delta. \tag{23.2}$$

Proof of Theorem 23.1 We use Lemma 23.1 to bound the expected risk with the following substitutions. The n samples are $X_i \triangleq z_i$. The function $f \triangleq h$ where $h \in \mathscr{H}$. The loss function $g(f, X) \triangleq L(h, z) \in [0, 1]$. The prior π is defined by $\pi \triangleq \xi^0$, in which we first sample k from $\{1, \ldots, K\}$ according to corresponding weights $\{w_1, \ldots, w_K\}$ and then sample h from Q_k. The posterior is defined similarly, $\rho \triangleq \xi$.

The KL-divergence term is

$$\begin{aligned}\mathrm{KL}(\rho||\pi) &= \mathbb{E}_f \ln\frac{\rho(f)}{\pi(f)} = \mathbb{E}_{k\in\{1,\ldots,K\}}(\mathbb{E}_h \frac{Q_k(h)}{Q_k^0(h)}|h \in \mathscr{M}_k)\\ &= \sum_{k=1}^{K} w_k \mathbb{E}_{h\in\mathscr{M}_k} \ln\frac{w_k Q_k(h)}{w_{0,k} Q_k^0(h)}\\ &= \mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}(Q_k||Q_k^0).\end{aligned} \tag{23.3}$$

Substituting the above into Eq. (23.2), it follows that

$$\begin{aligned}&\mathbb{P}_S\Bigg(\mathbb{E}_z\mathbb{E}_{k\in\{1,\ldots,K\}}\mathbb{E}_{h\in\mathscr{M}_k} L(h, z) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{k\in\{1,\ldots,K\}}\mathbb{E}_{h\in\mathscr{M}_k} L(h, z)\\ &+ \sqrt{\frac{\mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}((Q_k||Q_k^0) + \ln\frac{n}{\delta}}{2(n-1)}}\Bigg) \ge 1 - \delta.\end{aligned} \tag{23.4}$$

Using the notations in Sect. 23.2, we can rewrite the above as below,

$$\mathbb{P}_S\left(R(\xi, D) \le \widehat{R}(\xi, S) + \sqrt{(\mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}(Q_k||Q_k^0) + \ln\frac{n}{\delta})/(2n-2)}\right) \ge 1 - \delta. \tag{23.5}$$

Proof of Proposition 23.1 First, we proof that for $i = 2, \ldots, b$,

$$\overline{R}(\xi_i, \xi_{i-1}, B_i) \leq \overline{R}(\xi_{i-1}, \xi_{i-2}, B_{i-1}).$$

By definition of ξ_i,

$$\begin{aligned}
&\overline{R}(\xi_i, \xi_{i-1}, B_i) \leq \overline{R}(\xi_{i-1}, \xi_{i-1}, B_i) \\
&= \widehat{R}(\xi_{i-1}, B_i) + \sqrt{\ln \frac{n}{\delta}/(2n-2)} \\
&\leq \overline{R}(\xi_{i-1}, \xi_{i-2}, B_i) = \overline{R}(\xi_{i-1}, \xi_{i-2}, B_{i-1}).
\end{aligned}$$

Following these inequalities,

$$\overline{R}(\xi_b, \xi_{b-1}, S) = \overline{R}(\xi_b, \xi_{b-1}, B_b) \leq \overline{R}(\xi_1, \xi^0, B_1) = \overline{R}(\xi^*, \xi^0, S).$$

This finishes the proof.

Proof of Theorem 23.2 According to Theorem 1 in [16], we have

$$\begin{aligned}
R(\xi^*, \tau) \leq \inf_j \Bigg(&\frac{C_1}{\sum_{i \neq j}(n_i - n_i^{'})} \\
&+ \frac{C_2}{\sum_{i \neq j}(n_i - n_i^{'})} \sum_{i \neq j} \sum_{\alpha = n_i^{'}+1}^{n_i} \left[\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,\alpha}^2||^2 + R(\xi_j^*, D_i) \right] \Bigg),
\end{aligned} \tag{23.6}$$

where ξ_j^* is the minimizer of Eq. (23.1) with $\xi_0 = \xi_j$ and $S = S_{i,\alpha}^{(1)}$ denoted by $\xi_j^*(\xi_j, S_{i,\alpha}^{(1)})$.

For any $\alpha \geq n_i^{'}$ and an estimator satisfied the condition (C3), the inequalities $\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,n_i^{'}}^2||^2 \geq \mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,\alpha}^2||^2$ and $R(\xi_j^*(\xi_j, S_{i,n_i^{'}}^{(1)}), D_i) \geq R(\xi^*(\xi_j, S_{i,\alpha}^{(1)}), D_i)$ hold. Plugging into Eq. (23.6) for $\alpha = n_i^{'} + 1, \ldots, n_i$, it follows that

$$R(\xi^*, \tau) \leq \inf_j \left(\frac{C_1}{\sum_{i \neq j}(n_i - n_i^{'})} + \frac{C_2}{\sum_{i \neq j}(n_i - n_i^{'})} \sum_{i \neq j} \left[\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,n_i^{'}}^2||^2 + R(\xi_j^*, D_i) \right] \right),$$

where ξ_j^* is the minimizer of Eq. (23.1) with $\xi^0 = \xi_j$ and $S = S_{i,n_i^{'}}^{(1)}$.

Then, the result follows by the above inequality combined with Eq. (23.5). In order to obtain the form in Theorem 23.2, one only needs to note that if σ_{T_i} is known, the term $\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,n_i^{'}}^2||^2$ vanishes.

References

1. Amit, R., Meir, R.: Meta-learning by adjusting priors based on extended pac-bayes theory. In: International Conference on Machine Learning, pp. 205–214 (2018)
2. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T.: Bayesian model averaging: A tutorial. Stat. Sci. **14**(4), 382–401 (1999)
3. Huang, J., Ma, S., Zhang, C.H.: Adaptive lasso for sparse high-dimensional regression. Stat. Sin. **18**(4), 1603–1618 (2008)
4. Leamer, E.E.: Specification searches. Wiley, New York (1978)
5. Lever, G., Laviolette, F., Shawe-Taylor, J.: Tighter pac-bayes bounds through distribution-dependent priors. Theor. Comput. Sci. **473**(2), 4–28 (2013)
6. Liang, H., Zou, G., Wan, A.T.K., Zhang, X.: Optimal weight choice for frequentist model average estimators. J. Am. Stati. Assoc. **106**(495), 1053–1066 (2011)
7. Mcallester, D.A.: Pac-bayesian model averaging. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory, pp. 164–170 (1999)
8. Moral-Benito, E.: Model averaging in economics: An overview. J. Econ. Surv. **29**(1), 46–75 (2015)
9. Raftery, A.E.: Bayesian model selection in social research. Soc. Methodol. **25**(25), 111–163 (1995)
10. Raftery, A.E.: Approximate bayes factors and accounting for model uncertainty in generalised linear models. Biometrika **83**(2), 251–266 (1996)
11. Seeger, M.: Pac-bayesian generalisation error bounds for gaussian process classification. J. Mach. Learn. Res. **3**(2), 233–269 (2002)
12. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press (2014)
13. Tuddenham, R.D., Snyder, M.M.: Physical growth of california boys and girls from birth to eighteen years. Publ. Child Dev. **1**, 183–364 (1954)
14. Wan, A.T.K., Zhang, X., Zou, G.: Least squares model averaging by mallows criterion. J Econ. **156**(2), 277–283 (2010)
15. Wasserman, L.: Bayesian model selection and model averaging. J. Math. Psychol. **44**(1), 92–107 (2000)
16. Yang, Y.: Adaptive regression by mixing. J. Am. Stat. Assoc. **96**(454), 574–588 (2001)
17. Ye, C., Yang, Y., Yang, Y.: Sparsity oriented importance learning for high-dimensional linear regression. J. Am. Stat. Assoc. **2**, 1–16 (2016)
18. Zhang, X., Wan, A.T.K., Zou, G.: Model averaging by jackknife criterion in models with dependent data. J. Econ. **174**(2), 82–94 (2013)
19. Zhang, X., Zou, G., Liang, H.: Model averaging and weight choice in linear mixed-effects models. Biometrika **1**(1), 205–218 (2014)
20. Zhou, Q., Ernst, P.A., Morgan, K.L., Rubin, D.B., Zhang, A.: Sequential rerandomization. Biometrika **105**(3), 745–752 (2018)