

Approximating Bayesian inference by weighted likelihood

Xiaogang WANG

Key words and phrases: Empirical Bayes; entropy loss; hierarchical Bayes; James–Stein estimator; non-parametric regression; weighted likelihood.

MSC 2000: Primary 62H12; secondary 62B10.

Abstract: The author proposes to use weighted likelihood to approximate Bayesian inference when no external or prior information is available. He proposes a weighted likelihood estimator that minimizes the empirical Bayes risk under relative entropy loss. He discusses connections among the weighted likelihood, empirical Bayes and James–Stein estimators. Both simulated and real data sets are used for illustration purposes.

Approximer l'inférence bayésienne par la vraisemblance pondérée

Résumé : L'auteur propose l'emploi de la vraisemblance pondérée pour approximer l'inférence bayésienne en l'absence d'information externe ou a priori. Il propose un estimateur de vraisemblance pondérée qui minimise le risque de Bayes empirique sous l'entropie relative. Il établit des liens entre les estimateurs de vraisemblance pondérée, de Bayes empirique et de James–Stein. Des jeux de données réelles et simulées illustrent son propos.

1. INTRODUCTION

Many statistical applications involve multiple parameters and data sets that can be regarded as related or connected by the structure of the problem. For instance, data sets may be time series of ozone levels or they may be daily hospital admissions of different geographical sectors within the same region. The inferential interest, however, might focus on one parameter at a time. For example, one might want to examine the mean ozone level of a given year or the average weekly hospital admissions of a particular hospital. If the sample size is relatively large, then the classical likelihood approach would give reasonable results. When the sample size is small, however, the classical likelihood approach could face some serious challenges.

We first consider a motivating example. Suppose that a coin is tossed twice. Let $\theta_1 = P(\text{Head})$ for this coin and $\hat{\theta}_1$ denote the maximum likelihood estimator (MLE) of θ_1 . If the coin is indeed a fair one, it then follows that $P(\hat{\theta}_1 = 0 \text{ or } 1) = 1/2$ and $P(\hat{\theta}_1 = 1/2) = 1/2$. Thus the chance for the classical MLE to make a nonsensical decision about a fair coin is 50% if the sample size is 2. We suppose that the first coin is no longer available but one is allowed to flip a different coin twice. Let $\hat{\theta}_2$ denote the MLE of the second parameter θ_2 . The classical likelihood approach will ignore the second experiment since it comes from a different population. One interesting question would be whether one can still use the second experiment to derive a better estimate for θ_1 than the classical MLE. One might consider a convex combination of $\hat{\theta}_1$ and $\hat{\theta}_2$: $\hat{\theta}_1^e = \lambda_1 \hat{\theta}_1 + \lambda_2 \hat{\theta}_2$, where λ_1 and λ_2 are weights. The weight associated with the second experiment λ_2 should be chosen to reflect the degree of similarity between the two coins. For simplicity, we set each of the two weights to be 1/2. We denote the corresponding estimator by $\hat{\theta}_1^e$. Although $\hat{\theta}_1^e$ might not be the best choice of a convex combination, we examine the new estimator in order to demonstrate the advantage of combining information from different experiments. This new estimator has a much smaller mean squared error (MSE) than that of the MLE if the two parameters are close to each other. Suppose that for definiteness $\theta_1 = 0.5$ and $\theta_2 = 0.6$. It then follows that the probability for the new estimator $\hat{\theta}_1^e$ to make a nonsensical decision is 0.13, which is much smaller than that of the MLE. We also have that

$MSE(\tilde{\theta}_1^e)/MSE(\hat{\theta}_1) \approx 0.5$ if $\theta_1 = 0.5$ and $\theta_2 = 0.6$. Although this new estimator is slightly biased, it achieves a significantly smaller MSE when compared with the MLE. In fact, it can be verified that $MSE(\tilde{\theta}_1^e)/MSE(\hat{\theta}_1) < 0.75$ if $\theta_1, \theta_2 \in [0.35, 0.65]$. However, a complete pooling of different data sets might not always be desirable. In the motivating example, simply merging the two data sets in the likelihood inference is definitely not a sensible strategy if the second coin is entirely different from the first one. In fact, if $\theta_1 = 0.1$ and $\theta_2 = 0.9$, it then follows that $MSE(\tilde{\theta}_1^e)/MSE(\hat{\theta}_1) = 1.138$. Furthermore, the new estimator $\tilde{\theta}_1^e$ is grossly biased in this case. Therefore all related data sets should be evaluated to determine their relevance to the inference of the target parameter.

Cox (1981) gave an overview of some methods for the combination of data sets such as weighted means and pooling in the presence of over-dispersion. The report *Combining Information* by the U.S. National Research Council (1992) offered a survey on more recent techniques of combining information. Among all the methods for combining data sets, the Bayesian method is the most attractive and effective one. It is natural to model a problem involving several independent data sets hierarchically, with observable outcomes modelled conditionally on certain parameters. A hierarchical Bayesian framework nicely separates the information for the parameter of interest from that of other related experiments through the likelihood. Uncertainties associated with the parameters are expressed through the prior distribution of parameters. If the inferential interest is on one individual parameter, say θ_1 , then the posterior distribution is based on all data sets and the prior distribution. With the availability of Monte Carlo Markov Chain (MCMC) sampling, the hierarchical Bayes framework can be applied in very complex modelling scenarios such as spatial models; see Gelman, Carlin, Stein & Rubin (2003) and Gustafson, Hossain & McCandless (2005).

The hierarchical Bayes method could be very computationally intensive due to the employment of MCMC sampling. Wang, van Eeden & Zidek (2004) proposed using weighted likelihood to approximate the full Bayesian inference at the initial stage of a data analysis. Instead of using a prior distribution to connect all the parameters, the weighted likelihood method relies on likelihood weight functions which can be chosen adaptively to determine the degree of combination of related data sets. This method embraces the classical likelihood approach as it uses all available data sets for the inference of the parameter of interest. We show that the weighted likelihood method is also an approximate Bayes method under relative entropy loss. Since the weighted likelihood method does not use any prior information, it is most effective when there is no reliable external or prior information. The major advantage of the proposed method lies in the fact that it is usually much less computationally intensive than using MCMC for a hierarchical Bayesian analysis. For this reason, the weighted likelihood method can still serve as a useful exploratory tool even if prior information is available.

In order to introduce the weighted likelihood to be used in this article, we first review some related weighted likelihood methods and discuss their connections with our weighted likelihood.

1.1. Local likelihood methods.

The local likelihood introduced by Tibshirani & Hastie (1987) extended the idea of local fitting to likelihood-based regression models. Thus local regression may be viewed as a special case of the local likelihood procedure. Staniswalis (1989) defined her version of *local likelihood* in the context of non-parametric regression as follows:

$$LL_n(\theta) = \sum_{i=1}^n W\left(\frac{x_0 - x_i}{b}\right) \log f(y_i; \theta),$$

where the y_i denote realizations of random variables Y_i , the x_i are fixed and b is an unknown parameter. The general form of local likelihood was presented by Eguchi & Copas (1998). The

basic idea is to infuse local adaptation into likelihood by considering

$$L_n(\theta) = \sum_{i=1}^n K\left(\frac{y_i - t}{h}\right) \log f(y_i, \theta),$$

where $K = K\{(x_i - t)/h\}$ is a kernel function with center t and bandwidth h . The local maximum likelihood estimate $\hat{\theta}_t$ of a parameter in a statistical model $f(y; \theta)$ is defined by maximizing the weighted version of the likelihood function which gives a larger weight to a sample point near t . This does not give an unbiased estimating equation as it stands, and so the local likelihood proposed by Eguchi & Copas (1998) introduces a correction factor to ensure consistent estimation. The resulting *local maximum likelihood estimator* (LMLE), say $\hat{\theta}_{t,h}$, depends on the controllable variables t and h through the kernel function K . Intuitively, it is natural to think that $\hat{\theta}_{t,h}$ gains more information from the data points around t in the sample space.

1.2. Relevance weighted likelihood.

Hu (1994) and Hu & Zidek (2001) proposed a very general method for using all relevant sample information. They based their theory on what they called *relevance-weighted likelihood* (REWL). Let f_i be the unknown probability density function of the Y_i which is assumed to be independent of Y_j , $j \neq i$. The inferential interest is on a different probability density function f . At least in some qualitative sense, we assume that the f_i are thought to be “similar” to the target density f . Consequently the y_i are thought to be of some inferential value in our analysis though they are independently drawn from populations different from the target population. The relevance-weighted likelihood is defined as

$$\text{REWL}(\theta) = \prod_{i=1}^n f(y_i; \theta)^{\lambda_i},$$

where $\lambda_1, \dots, \lambda_n$ are likelihood weights.

It can be seen that this method generalizes the core of the local likelihood as the weight functions are no longer restricted to the kernel function. Hu & Rosenberger (2000) investigated two classes of weight functions, namely the exponential and the polynomial type in analyzing adaptive designs when time trends are present. The asymptotic properties of the relevance maximum weighted likelihood estimator, including asymptotic normality and consistency, were established in Hu (1997). Furthermore, Hu, Rosenberger & Zidek (2000) extended the relevance weighted likelihood framework for dependent sequences.

1.3. Weighted likelihood estimating equations.

Markatou, Basu & Lindsay (1997, 1998) proposed a method based on the weighted likelihood equation in the context of robust estimation. Their approach can be described as follows. Suppose that (Y_1, \dots, Y_n) is a random sample from the distribution $f(y; \theta)$. The *weighted likelihood equation* is defined as

$$\sum_{i=1}^n w(y_i, \hat{F}) \frac{\partial}{\partial \theta} \log f(y_i; \theta) = 0,$$

where \hat{F} is the empirical cumulative distribution function and $w(y_i, \hat{F})$ is a weight function. The weight function $w(y_i, \hat{F})$ is selected such that it has a value close to 1 if there is no evidence of model violation at y_i from the empirical distribution function. The weight function will be very close to 0 or exactly 0 at y_i if the empirical cumulative distribution function indicates a lack of fit at or near y_i . Thus, the role of the weight function is to down-weight points in the same sample that are inconsistent with the assumed model. In the framework of Markatou, Basu & Lindsay (1998), the parameter estimates are derived as the solution of a set of estimating equations. This formulation allows a different definition of robustness by allowing one to derive all solutions of

the estimating equations and thus identify to what extent the adopted model describes the data. Agostinelli & Markatou (2001) used the formulation similar to the relevance weighted likelihood in the context of testing.

1.4. Weighted likelihood.

The weighted likelihood employed in this article is closely related to but different from the relevance weighted likelihood. In the framework of the relevance weighted likelihood, the information about θ_1 builds up because the number of populations grows increasingly in close proximity to θ_1 . This is the paradigm commonly invoked in the context of non-parametric regression but it is not always the most natural one. In contrast, Wang, van Eeden & Zidek (2004) postulated a fixed number of populations with an increasingly large number of samples from these populations. This paradigm is more natural in situations in which the James–Stein estimator is derived.

Suppose that we are interested in a single population as in our motivating example. A single parameter or parameter vector θ_1 is of inferential interest. Information from other related populations, population 2, population 3, and so on, is available together with the direct information from population 1. Let $m - 1$ denote the total number of populations whose distributions are thought to “resemble” that of population 1. Let n_1, \dots, n_m denote the number of observations obtained from each individual population respectively. Let $\mathbf{X}_1, \dots, \mathbf{X}_m$ be random variables or vectors with marginal probability density functions $f_1(\cdot; \theta_1), \dots, f_m(\cdot; \theta_m)$, where $\mathbf{X}_i = (X_{i1}, \dots, X_{in_i})^\top$, $i = 1, \dots, m$. The parametric form of the joint distribution of $(\mathbf{X}_1, \dots, \mathbf{X}_m)$ is not assumed to be known. We are interested in the probability density function $f_1(\cdot; \theta_1) : \theta_1 \in \Theta$ of a study variable or vector of variables \mathbf{X}_1 , θ_1 being an unknown parameter or vector of parameters. At least in some qualitative sense, we assume that $f_2(\cdot; \theta_2), \dots, f_m(\cdot; \theta_m)$ are “similar” to $f_1(\cdot; \theta_1)$.

Weighted likelihood (WL) is then defined as:

$$\text{WL}(\theta_1) = \prod_{i=1}^m f_1(\mathbf{x}_i; \theta_1)^{\lambda_i}, \quad (1)$$

where the λ_i are likelihood weights and $\lambda_1 + \dots + \lambda_m = 1$.

We say that $\tilde{\theta}_1$ is a maximum weighted likelihood estimator (WLE) for θ_1 if

$$\tilde{\theta}_1 = \arg \sup_{\theta_1 \in \Theta} \text{WL}(\theta_1).$$

Note that the uniqueness of the maximizer is not assumed.

Throughout this article, we assume that for $i = 1, \dots, m$, X_{i1}, \dots, X_{in_i} are independent and identically distributed random variables. The weighted likelihood (WL) defined by (1) then becomes

$$\text{WL}(\theta_1) = \prod_{i=1}^m \prod_{j=1}^{n_i} f_1(x_{ij}; \theta_1)^{\lambda_i}.$$

The asymptotic properties of the weighted likelihood estimator (WLE) for fixed and adaptive weights are established in Wang, van Eeden & Zidek (2004). The weights for the related data sets are in place to down-weight possible bias of the WLE of θ_1 . The weights are crucial in weighted likelihood estimation. The kernel functions used in other weighted likelihood functions are natural candidates for the likelihood weights. We will discuss other types of adaptive weights in this article.

We now apply the weighted likelihood to the motivating example. To simplify the notation, let $Y_1 = X_{11} + X_{12}$ and $Y_2 = X_{21} + X_{22}$ denote the sample total from the first and second experiment respectively, where the X_{ij} are all Bernoulli random variables from the two experiments. The classical likelihoods for θ_1 and θ_2 are

$$L_1(\theta_1; Y_1) \propto \theta_1^{Y_1} (1 - \theta_1)^{2 - Y_1} \quad \text{and} \quad L_2(\theta_2; Y_2) \propto \theta_2^{Y_2} (1 - \theta_2)^{2 - Y_2}.$$

Thus, the classical MLEs of $\hat{\theta}_1$ and $\hat{\theta}_2$ are equal to $Y_1/2$ and $Y_2/2$, respectively.

Since one is actually only interested in the first coin, the weighted likelihood of θ_1 for some given weights is then defined as

$$WL(\theta_1; Y_1, Y_2) \propto L_1(\theta_1; Y_1)^{\lambda_1} L_2(\theta_1; Y_2)^{\lambda_2} = \theta_1^{\lambda_1 Y_1 + \lambda_2 Y_2} (1 - \theta_1)^{2 - (\lambda_1 Y_1 + \lambda_2 Y_2)}.$$

Note that θ_2 in the second classical likelihood L_2 is replaced by θ_1 in the construction of the weighted likelihood for θ_1 . In the weighted likelihood framework, the observation Y_2 might contain potentially very important information with respect to the inference on θ_1 . The weight λ_2 is intended to down-weight the potential bias and control the degree of combination of information.

It then follows that the WLE takes the form

$$\tilde{\theta}_1 = \lambda_1 \hat{\theta}_1 + \lambda_2 \hat{\theta}_2,$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are MLEs of θ_1 and θ_2 respectively. We emphasize that the WLE of θ_1 is indeed a linear combination that one might consider in the motivating example. We also remark that the MLE can be obtained from the WLE by setting λ_2 to zero.

1.5. Outline of the article.

In Section 2, we will derive the weighted likelihood as an approximate Bayes rule by minimizing empirical entropy loss. This result provides an information-theoretic justification for using weighted likelihood as an inference function. Various choices of weights are also discussed. In Section 3, we discuss the connections among the WL, the empirical Bayes and the James–Stein estimators when random variables are all normally distributed. We further generalize the result to the exponential family. Results of simulation studies are provided in Section 4. A case study applying all relevant methods to an educational experiment is presented in Section 5. Discussions are provided in Section 6.

2. APPROXIMATE BAYES DECISION BY THE WEIGHTED LIKELIHOOD ESTIMATOR

2.1. Approximate Bayes rule by non-parametric regression.

Let \mathcal{A} be a measurable space of decisions and let $\mathcal{L}: \mathbb{R} \times \mathcal{A} \rightarrow \mathbb{R}$ be a jointly measurable non-negative loss function. We assume that $E\{\mathcal{L}(\theta, a)\} < \infty$ for all $a \in \mathcal{A}$. Let $\delta: \mathbb{R}^d \rightarrow \mathcal{A}$ be a measurable decision rule for choosing $a \in \mathcal{A}$ after having observed X while θ is unknown. The quality of a decision rule is quantified in the risk function. To be more specific, for a decision rule $\delta(X)$, the risk function with respect to θ is defined as

$$R\{\theta, \delta(X)\} = E\{\mathcal{L}\{\theta, \delta(X)\}\}.$$

Prior information is used to further summarize the information of a decision rule contained in the risk function. The Bayes risk of a decision rule $\delta(X)$ with respect to a specific prior distribution $\pi(\theta)$ is defined as

$$B(\pi, \delta) = E_\pi\{R(\theta, \delta)\}.$$

The Bayes rule $\delta(X)$ associated with the minimum Bayes risk function must satisfy the following condition:

$$E\{\mathcal{L}(\theta, \delta(x))\} = \inf_{a \in \mathcal{A}} E\{\mathcal{L}(\theta, a) | X = x\}.$$

Thus the Bayes risk can be defined in terms of the posterior loss $E\{\mathcal{L}(\theta, a) | X = x\}$ for $a \in \mathcal{A}$.

Assume that we have $(\theta_1, X_1), \dots, (\theta_n, X_n)$ that are independent and identically distributed random vectors from the distribution of (θ, X) . If the knowledge of the prior is not fully known, Stone (1977) proposed to approximate the Bayes inference by using the idea of non-parametric regression. It then follows that

$$\hat{E}_n\{\mathcal{L}(\theta, a) | X\} = \sum_{k=1}^n W_{n,k}(X) \mathcal{L}(\theta_k, a),$$

for any given a and the weights $W_{n,k}$ are consistent weights defined in Stone (1977). A sequence of weights is said to be consistent if whenever $(Y_1, \mathbf{X}_1), \dots, (Y_n, \mathbf{X}_n)$ are independent and identically distributed, $r > 1$, and $E|Y|^r \leq \infty$, then $\widehat{E}_n(Y | \mathbf{X}) = E(Y | \mathbf{X})$ in L^r as $n \rightarrow \infty$. The necessary and sufficient conditions for obtaining consistent weights are established in Stone (1977).

2.2. Entropy loss and weighted likelihood.

Instead of using the popular mean squared error (MSE) as the loss function, we consider the relative entropy loss. Let $f(x | \theta)$ be the conditional density function for any given θ . For any other value of the parameter, say a , the relative entropy or Kullback–Leibler information is defined as

$$\mathcal{L}(\theta, a) = I(\theta; a) = \int \log \left\{ \frac{f(x | \theta)}{f(x | a)} \right\} f(x | \theta) dx.$$

Kullback (1959) described the relative entropy as the mean information per observation from $f(x | \theta)$ for discrimination of $f(x | a)$ for any given a in favour of $f(x | \theta)$ against $f(x | a)$. The entropy loss was also used in James & Stein (1961) in the estimation of the multinomial variance-covariance matrix. Ghosh & Yang (1988) considered entropy loss for simultaneous estimation of p independent means. Akaike (1974) derived the widely-used Akaike information criterion for model selection by using this as a predictive criterion. The relative entropy is also related to the Fisher information. Suppose that θ and $\theta + \delta$ are neighbouring points in a k -dimensional parameter space. Under certain regularity condition, as Kullback (1959) showed, $I(\theta; \theta + \delta)$ can be expressed as

$$I(\theta; \theta + \delta) \approx \frac{1}{2} \delta^\top F_\theta \delta,$$

where F_θ is the Fisher information matrix evaluated at θ .

We now derive the weighted likelihood function by minimizing the empirical relative entropy loss. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{in_i})$, where $x_{ij} \sim f(\cdot | \theta_i)$, $j = 1, \dots, n_i$, $i = 1, \dots, m$, and $n = n_1 + \dots + n_m$. By assigning equal weight to each observation in the same sample, we have

$$\widehat{E}_n \{ \mathcal{L}(\theta_1, a) | \mathbf{x}_1 \} = \sum_{k=1}^n W_{n,k}(\mathbf{x}_1) I(\theta_k, a) = \sum_{i=1}^m n_i w_i(\mathbf{x}_1) I(\theta_i, a).$$

For any given θ_i , a and \mathbf{x}_i , we can then estimate $I(\theta_i, a)$ by

$$\widehat{I}(\theta_i, a) = \frac{1}{n_i} \sum_{j=1}^{n_i} \log \frac{f(x_{ij} | \theta_i)}{f(x_{ij} | a)} \xrightarrow{a.s.} \mathcal{L}(\theta_i, a) \quad \text{as } n_i \rightarrow \infty.$$

It then follows that

$$\widehat{E}_n \{ \mathcal{L}(\theta_1, a) | \mathbf{x}_1 \} = \sum_{i=1}^m w_i(\mathbf{x}_1) \sum_{j=1}^{n_i} \log \frac{f(x_{ij} | \theta_i)}{f(x_{ij} | a)}.$$

We then have

$$\widehat{E} \{ \mathcal{L}(\theta_1, a) | \mathbf{x}_1 \} = \sum_{i=1}^m w_i(\mathbf{x}_1) \sum_{j=1}^{n_i} \log f(x_{ij} | \theta_i) - \sum_{i=1}^m w_i(\mathbf{x}_1) \sum_{j=1}^{n_i} \log f(x_{ij} | a). \quad (2)$$

The second term of equation (2) is independent of $\theta_1, \dots, \theta_m$. It then follows that the approximate Bayes rule must satisfy the following:

$$\widehat{\theta}_1^A(\mathbf{x}_1) = \arg \sup_{a \in \mathcal{A}} \sum_{i=1}^m w_i(\mathbf{x}_1) \sum_{j=1}^{n_i} \log f(x_{ij} | a).$$

Thus the approximate Bayes rule $\hat{\delta}(\mathbf{x}_1)$ of θ_1 under entropy loss is defined by a decision rule such that

$$\hat{\theta}_1^A(\mathbf{x}_1) = \arg \sup_{a \in \mathcal{A}} \sum_{i=1}^m w_i(\mathbf{x}_1) \sum_{j=1}^{n_i} \log f(x_{ij} | a).$$

It then follows that the approximate Bayes rule $\hat{\delta}(\mathbf{x}_1)$ of θ_1 under relative entropy loss must also maximize the following weighted likelihood:

$$\tilde{\theta}_1 = \hat{\theta}_1^A(\mathbf{x}_1) = \arg \sup_{a \in \mathcal{A}} \prod_{i=1}^m \prod_{j=1}^{n_i} f(x_{ij}; a)^{\lambda_i},$$

where $\lambda_i = w_i(\mathbf{x}_1)$ are consistent weights.

This shows that the WLE for θ_1 is indeed an approximate Bayes rule under entropy loss. Furthermore, it shows that weighted likelihood is an inference device for approximating Bayesian inference. The derivation of weighted likelihood generalizes the argument by Akaike (1974) that connects the likelihood principle with an information theoretic framework.

2.3. *Optimal adaptive weights for the weighted likelihood estimator.*

In order to approximate the Bayesian inference by non-parametric regression, Stone (1977) proposed the K -nearest neighbour weights and showed that they are consistent. Györfi, Kohler, Krzyzak & Walk (2002) showed that the kernel-type weights are also consistent. Theorem 2.1 in Stone (1977) provides necessary and sufficient conditions to obtain such consistent weights.

However, the kernel type of weights involves the specification of some kind of smoothing or tuning parameter. Cross-validation could be used to derive likelihood weights without any tuning parameter. In particular, Wang & Zidek (2005) used a delete-one approach in the cross-validation procedure. Let $\tilde{\theta}_1^{(-j)}$ be the WLE based on m samples without the j th data point from the first sample. This generalizes the two cases where either only the j th data point is deleted from the first sample or the j th data point from each sample is deleted. Note that $\tilde{\theta}_1^{(-j)}$ is a function of the weights λ_i . Let $1/n_1 D_{n_1}$ be the average discrepancy in the cross-validation given by

$$\frac{1}{n_1} D_{n_1}(\lambda_1, \dots, \lambda_m) = \frac{1}{n_1} \sum_{j=1}^{n_1} \{X_{1j} - \phi(\tilde{\theta}_1^{(-j)})\}^2,$$

with $\lambda_1 + \dots + \lambda_m = 1$. However, the cross-validated weights will not work if some or all of the parameters are identical. Also, when the summary statistics are provided without the actual data sets, then the cross-validated weights cannot be calculated.

When the cross-validation procedure is not applicable, we derive the optimum weights when the WLE takes a linear form and the pairwise distances among parameters are bounded. Assume that for $i = 1, \dots, m, j = 1, \dots, n_i$, the X_{ij} are random variables with $E(X_{ij}) = \theta_i$. Assume that the θ_i are all finite and $|\theta_i - \theta_1| \leq C_i$, where C_1, \dots, C_m are known constants. Let $B = (0, C_2, \dots, C_m)^\top$ and $\hat{\theta}_i = \hat{\theta}_i(x_{i1}, \dots, x_{in_i})$ be the MLE of $\theta_i, i = 1, \dots, m$.

THEOREM 1. *Assume that $V = \text{cov}(\hat{\theta})_{m \times m}$ is known and $(V + BB^\top)^{-1}$ is invertible. Suppose that $E(\hat{\theta}_i) = \theta_i$ for $i = 1, \dots, m$. If the weighted likelihood estimator takes the form of a linear combination of the $\hat{\theta}_i$, then the optimum weights which minimize the maximum of the mean squared error is given by*

$$\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)^\top = \frac{(V + BB^\top)^{-1} \mathbf{1}}{\mathbf{1}^\top (V + BB^\top)^{-1} \mathbf{1}}.$$

Detailed discussions including asymptotic properties of this set of weights can be found in Wang (2001). In practice, however, the C_i and the covariance matrix V are not known, thus the approximate optimal weights for the restricted mean problem are given by

$$\lambda^{(\text{ob})} = \frac{(\widehat{V} + \widehat{B}\widehat{B}^\top)^{-1}}{\mathbf{1}^\top(\widehat{V} + \widehat{B}\widehat{B}^\top)^{-1}\mathbf{1}}, \quad (3)$$

where \widehat{V} and \widehat{B} are estimates of V and B , respectively.

If the samples are independent of each other, it also follows that the optimal weights can be approximated by

$$\lambda_i^{(\text{oba})} = \frac{1}{D} \frac{1}{(\hat{\sigma}_i^2 + \widehat{C}_i^2)}, \quad i = 1, \dots, m, \quad (4)$$

where $C_1 = 0$ and $D = (\hat{\sigma}_1^2 + \widehat{C}_1^2)^{-1} + \dots + (\hat{\sigma}_m^2 + \widehat{C}_m^2)^{-1}$.

The formula for computing approximate optimal weights given by equation (4) offers some intuitive insights on how these optimal weights behave. The weight for a related population will take a relatively large value if and only if both the estimated variance and the estimated distance between θ_i and θ_1 are small. The corresponding weight assumes a small value if either σ_i^2 or \widehat{C}_i^2 is large.

3. BAYES METHODS AND THE WEIGHTED LIKELIHOOD ESTIMATOR

In this section, we discuss the connections between the approximate Bayes rules under MSE loss and the WLE. We also discuss the connection between the empirical Bayes and WL estimators.

3.1. Hierarchical and empirical Bayes methods.

Let $\phi_\eta(\theta)$ be the prior distribution where η is the vector of hyperparameters. Assume $\theta_1, \dots, \theta_m$ are independent and identically distributed with distribution $\pi_\eta(\theta)$ and $\mathbf{x}_i \sim f(\cdot | \theta_i)$ for $i = 1, \dots, m$. Suppose that θ_1 is of inferential interest. The marginal density by integrating out the observed vector of other “related” parameters $\boldsymbol{\theta}^r = (\theta_2, \dots, \theta_m)$ is given by

$$d_\eta(\mathbf{x}^r) = \int \pi_\eta(\boldsymbol{\theta}^r) f(\mathbf{x}^r | \boldsymbol{\theta}^r) d\boldsymbol{\theta}^r,$$

where

$$\mathbf{x}^r = (\mathbf{x}_2, \dots, \mathbf{x}_m), \quad \pi_\eta(\boldsymbol{\theta}^r) = \prod_{i=2}^m \pi_\eta(\theta_i) \quad \text{and} \quad f(\mathbf{x}^r | \boldsymbol{\theta}^r) = \prod_{i=2}^m f(\mathbf{x}_i | \theta_i).$$

Let $h(\eta)$ be the hyperprior distribution for η . We then have the induced prior density for θ_1 , viz.

$$\pi_{\mathbf{x}^r}(\theta_1) = \int h(\eta | \mathbf{x}^r) \pi_\eta(\theta_1) d\eta,$$

where $h(\eta | \mathbf{x}^r) = c h(\eta) d_\eta(\mathbf{x}^r)$ with c a normalizing constant. The posterior density of θ_1 based on all the data is then given by

$$p(\theta_1 | \mathbf{x}_1) = c \pi_{\mathbf{x}^r}(\theta_1) L_1(\mathbf{x}_1; \theta_1). \quad (5)$$

The hierarchical Bayes formula given by (5) neatly separates the “direct” information contained in $L_1(\mathbf{x}_1; \theta_1)$ and the information contained in other samples summarized by $\pi_{\mathbf{x}^r}(\theta_1)$.

If the hyperparameter vector η is unknown, then the maximum likelihood estimator of η based on other data \mathbf{x}^r can be derived from

$$\hat{\eta} = \arg \max_{\eta} d_\eta(\mathbf{x}^r).$$

The empirical Bayes alternative is then given by

$$p_{\hat{\eta}}(\theta_1 | \mathbf{x}_1) = c \pi_{\hat{\eta}}(\theta_1) L_1(\mathbf{x}_1; \theta_1).$$

The empirical Bayes estimate is derived by carrying out the following maximization:

$$\hat{\theta}_1^{\text{EB}} = \arg \max_{\theta_1 \in \Theta} \{ \log L_1(\mathbf{x}_1; \theta_1) + \log \pi_{\hat{\eta}}(\theta_1) \}.$$

In comparison, the WLE is derived as follows:

$$\tilde{\theta}_1^{\text{WLE}} = \arg \max_{\theta_1 \in \Theta} \left\{ \log L_1(\mathbf{x}_1; \theta_1) + \sum_{i=2}^m (\lambda_i / \lambda_1) \log L_i(\mathbf{x}_i; \theta_1) \right\},$$

where $\lambda_1 > 0$.

3.2. Approximate Bayes rule and the linear weighted likelihood estimator.

Denote the approximate Bayes rule under the squared error loss by $\hat{\theta}_1^S(\mathbf{X}_1)$. Stone (1977) showed that the approximate Bayes rule under squared error loss is given by $\hat{E}_n(\theta_1 | \mathbf{X}_1)$. It then follows that

$$\hat{E}_n(\theta_1 | \mathbf{X}_1) = \sum_{k=1}^n W_{kn}(\mathbf{X}_1) \theta_k,$$

where the W_{kn} are weights in non-parametric regression. This implies that the approximate Bayes rule under MSE is a linear combination of the θ_i . However, the random variables θ_i are, in fact, not observed. By replacing the θ_i by their MLEs, we then have that the approximate Bayes rule takes the form

$$\hat{\theta}_1^A(\mathbf{X}) = \sum_{i=1}^n \lambda_i(\mathbf{X}) \hat{\theta}_i(\mathbf{X}_i),$$

where $\hat{\theta}_i(\mathbf{X}_i)$ is the MLE of θ_i , $i = 1, \dots, m$. This demonstrates that the approximate Bayes rule under the MSE loss is indeed a linear combination of those individual MLEs derived from the samples.

For normal, Bernoulli and Poisson distributions, the WLE takes the form of linear combinations of the individual MLEs. For example, let $\mathbf{X}_i = (X_{i1}, \dots, X_{in_i})$, where $X_{ij} \sim N(\theta_i, 1)$, $j = 1, \dots, n_i$, $i = 1, \dots, m$. It then follows that the WLE takes the following form:

$$\tilde{\theta}_1(\mathbf{X}) = \sum_{i=1}^m \lambda_i \bar{X}_i,$$

where the λ_i are likelihood weights.

Thus, the WLE could be a linear combination of MLEs derived from each of the data sets. The general result is stated in the next two theorems.

THEOREM 2. Assume that $X_{ij} \stackrel{i.i.d.}{\sim} f(\cdot; \theta_i)$ for all $i = 1, \dots, m$ and $j = 1, \dots, n_i$. Let $\hat{\theta}_i(\mathbf{X}_i)$ denote the maximum likelihood estimator of θ_i derived from $\mathbf{X}_i = (X_{i1}, \dots, X_{in_i})$ for $i = 1, \dots, m$. If

$$\frac{\partial \log L_1(x_{i1}, \dots, x_{in_i}; \theta_1)}{\partial \theta_1} = n_i D_1(\theta_1) \{T(\mathbf{x}_i) - \theta_1\},$$

then

$$\tilde{\theta}_1 = \sum_{i=1}^m w_i \hat{\theta}_i(\mathbf{x}_i),$$

where $w_i = n_i \lambda_i / \sum_{i=1}^m n_i \lambda_i$ and the λ_i are likelihood weights for weighted likelihood.

THEOREM 3. For distributions of the exponential family, suppose that the maximum likelihood estimator of θ_1 takes the form of $g\{T(\mathbf{X}_1)\}$, where $T(\mathbf{X}_1)$ is the sufficient statistic. Then the weighted likelihood estimator of θ_1 takes the form $g\{\sum_{i=1}^m w_i T(\mathbf{X}_1)\}$, where $w_i = n_i \lambda_i / \sum_{i=1}^m n_i \lambda_i$.

The above two theorems imply that the WLE and the approximate Bayes rule under the mean squared error could coincide with each other for certain members of the exponential family if appropriate weights could be chosen.

3.3. Empirical Bayes, James–Stein estimator and the weighted likelihood estimator.

We now examine the relationships among the empirical Bayes, James–Stein and WL estimators when all variables are normally distributed. Consider the following sampling scheme of Efron (1996):

$$\theta_k \sim N(M, A), \quad \text{and} \quad Y_k | \theta_k \sim N(\theta_k, 1), \quad k = 1, \dots, m, \quad m > 3.$$

Suppose that the inferential interest is on θ_1 , the parameter of the first population. As shown in Efron (1996), the empirical Bayes estimates for M and A are then given by the following:

$$\widehat{M} = \frac{1}{m-1} \sum_{i=2}^m y_i; \quad \widehat{A} = \frac{1}{m-2} \sum_{i=2}^m (y_i - \widehat{M})^2.$$

The posterior density function of θ_1 with the estimated hyperparameters is then given by

$$\hat{\pi}(\theta_1 | y_1) \sim N[\widehat{M} + (1 - \widehat{B})(y_1 - \widehat{M}), 1 - \widehat{B}],$$

where $\widehat{B} = m\widehat{A}/(m-1)$.

It follows that the empirical Bayes estimator is

$$\hat{\theta}_1^{EB} = (1 - \widehat{B})y_1 + \sum_{i=1}^m \frac{\widehat{B}}{m-1} y_i. \quad (6)$$

It can be seen that the empirical Bayes estimator is a linear combination of the y_i .

The famous James–Stein estimator is given by

$$\hat{\theta}_1^{JS} = \bar{y} + (1 - \widehat{B}^{JS})(y_1 - \bar{y}),$$

where $\bar{y} = (y_1 + \dots + y_m)/m$ and $\widehat{B}^{JS} = (m-3)/\sum_{i=1}^m (y_i - \bar{y})^2$. It follows that

$$\hat{\theta}_1^{JS} = \left(1 - \frac{m-1}{m} \widehat{B}^{JS}\right) y_1 + \sum_{i=2}^m \frac{\widehat{B}^{JS}}{m-1} y_i. \quad (7)$$

It can be seen that the empirical Bayes and James–Stein estimator closely resemble each other. Efron & Morris (1973, Lemma 2) showed that the empirical Bayes estimator is close to but not as good as the James–Stein estimator.

For normal densities, the weighted likelihood estimator for θ_1 is also a linear combination of the x_i by Theorem 3.1. To be more specific, the WLE is given by

$$\hat{\theta}_1^{WLE} = \sum_{i=1}^m \lambda_i y_i, \quad (8)$$

where $\lambda_1 + \dots + \lambda_m = 1$.

By comparing equations (6), (7) and (8), we find that the empirical Bayes, the James–Stein and the weighted likelihood estimators all take the form of a linear combination of y_i for normal distributions. Consequently, the weights for the WLE can be chosen such that it coincides with either the empirical Bayes estimator or the James–Stein estimator. Observe that both empirical Bayes and James–Stein estimator assign equal weight to the y_i for $i > 2$. This implies that they do not utilize the differences among related data sets. The WLE, however, could assign unequal weights to different samples to reflect their possible varying relevance to θ_1 . This could be a very important advantage as shown in our simulation studies.

4. SIMULATION STUDIES

4.1. Simulation study based on binomial distributions.

We now go back to the motivating example. Recall that

$$Y_1 \sim \mathcal{BIN}(2, \theta_1) \quad \text{and} \quad Y_2 \sim \mathcal{BIN}(2, \theta_2).$$

Recall that the WLE of θ_1 takes a form of convex combination of the MLEs for θ_1 and θ_2 , i.e., $\hat{\theta}_1 = \lambda_1 \hat{\theta}_1 + \lambda_2 \hat{\theta}_2$. The optimal weights given by equation (3) actually require the knowledge of the bound between the two parameters. To be realistic, we could simply use the estimate of the bound between θ_1 and θ_2 and plug it into equation (3). Thus the adaptive weights take the following form:

$$\lambda_1^{\text{opt}} = \frac{d_2}{d_1 + d_2} \quad \text{and} \quad \lambda_2^{\text{opt}} = \frac{d_1}{d_1 + d_2},$$

where $d_1 = \hat{\theta}_1(1 - \hat{\theta}_1)$ and $d_2 = \hat{\theta}_2(1 - \hat{\theta}_2) + 2|\hat{\theta}_1 - \hat{\theta}_2|^2$. However, these adaptive weights are not very effective on combining the two samples in this simulation study. To be more specific, the weight λ_2 often assumes the value zero since d_1 equals zero quite often for small sample sizes. When d_1 equals zero, we then modify the adaptive weights as follows:

$$\lambda_1^{\text{opt}} = \frac{1}{1 + d_3} \quad \text{and} \quad \lambda_2^{\text{opt}} = \frac{d_3}{1 + d_3},$$

where $d_3 = 1/(1 + 2|\hat{\theta}_1 - \hat{\theta}_2|^2)$. Therefore, λ_1 is closer to 1 if $|\hat{\theta}_1 - \hat{\theta}_2|^2$ is relatively large. Without any external or prior information, we choose a uniform prior $\mathcal{U}(0, 1)$ and assume that θ_1 and θ_2 are independent and identically distributed. We remark that the Bayesian estimates were calculated using WinBugs with 1500 runs within each iteration.

The algorithm of this simulation study is described as follows:

ALGORITHM 4.1:

- Step 1: Generate $\theta_1^{[i]}$ and $\theta_2^{[i]}$ according to one of the four sampling schemes described below.
 - Step 2: Generate $Y_1^{[i]}$ and $Y_2^{[i]}$ independently from binomial distributions by using $\theta_1^{[i]}$ and $\theta_2^{[i]}$.
 - Step 3: Compute MLE, WLE with adaptive weights and Bayes estimator using MCMC sampling.
 - Step 4: Compute the squared error for the MLE, WLE and Bayes estimator.
- Repeat Steps 1 to 4 1000 times.

In order to evaluate and compare the performance of the WL and Bayes estimator, we consider four sampling schemes (SS) to generate θ_1 and θ_2 .

SS1: Fix $\theta_1 = 0.5$ and $\theta_2 = 0.6$.

SS2: Generate θ_1 and θ_2 independently from $\mathcal{U}(0, 1)$.

SS3: Generate θ_1 from $\mathcal{U}(0, 0.8)$ and $\theta_2 = \theta_1 + \mathcal{U}(0, 0.2)$.

SS4: Generate θ_1 from $\mathcal{U}(0.1, 0.3)$ and $\theta_2 = \theta_1 + \mathcal{U}(0, 0.2)$.

In sampling scheme 1, we fix the value of θ_1 and θ_2 for every iteration. In sampling schemes 2, 3 and 4, we generate random values for θ_1 and θ_2 for each iteration. The mean squared errors for the MLE, WL and Bayes estimators are listed in Table 1. It can be seen that WLE outperforms MLE across the four sampling schemes used in this simulation study. This shows clearly the advantage of combining information when the sample size from the target population is very small. In fact, WLE will have roughly the same MSE as the MLE, even when θ_1 equals 0.2 and θ_2 equals 0.8.

TABLE 1: Comparison of the mean squared errors for binomial distributions.

	MLE	WLE	Bayes	Bayes/WLE
SS1	0.124	0.069	0.031	45%
SS2	0.081	0.059	0.045	76%
SS3	0.094	0.054	0.044	81%
SS4	0.065	0.041	0.041	100%

The advantage of combining information by using the Bayes estimator is also evident in Table 1. The Bayes estimator offers further reduction when compared with the WLE in three of the four sampling schemes. In sampling scheme 1, the Bayes estimator offers 55% reduction in MSE when compared with that of WLE even though the parameter θ_1 takes a fixed value. This is not very surprising since the true value of θ_1 coincides with the mean of the assumed uniform prior, $\mathcal{U}(0, 1)$. This is clearly one of those cases in which the approximation by using WL to the full Bayesian inference is not satisfactory. Since the assumed prior coincides with the true prior $\mathcal{U}(0, 1)$ in sampling scheme 2, the Bayes estimator is guaranteed theoretically to be the optimal estimator. The WLE offers a MSE that is about 33% larger than that of the Bayes estimator. In sampling scheme 3, the approximation of WLE to the Bayes estimator is improved since the assumed prior differs from the true underlying prior distribution from which the parameters are generated. In sampling scheme 4, the assumed prior is quite different from the true prior. The approximation by using the WLE is actually perfect since both estimators have the same MSE.

4.2. Simulation study based on normal distributions.

We also carry out another simulation study to compare the performances of various estimators, namely, the MLE, James–Stein, empirical Bayes, WL and Bayes estimators.

Following the same sampling scheme considered in Efron (1996), we shall work with a simple two-level normal model of data Y_i , $i = 1, \dots, 5$, with group level effects θ_k , $i = 1, \dots, 5$, respectively, as follows:

$$Y_k | \theta_k \sim N(\theta_k, 1), \quad \theta_k \sim N(\mu_\theta, \sigma_\theta^2), \quad k = 1, \dots, m, \quad m > 3.$$

For simplicity, we set μ_θ equal to 0 and σ_θ equal to 1.

The algorithm for the second simulation study is described as follows:

ALGORITHM 4.2:

Step 1: Generate θ_i , $i = 1, \dots, 5$, from the standard normal distribution.

Step 2: Generate Y_i , $i = 1, \dots, 5$, from a normal distribution with mean θ_k and variance 1, respectively.

Step 3: Calculate the squared or absolute error of θ_1 for each of these five estimators.

Repeat Steps 1 to 3 1000 times.

We emphasize that θ_1 is of inferential interest. The MLE of θ_1 is simply X_1 . The Bayes estimator is actually 0.5*MLE if the values of the hyperparameters are known. The empirical Bayes and James–Stein estimators are given by equations (6) and (7) respectively. We use the optimal weights given by equation (3) with known variances to calculate the WLE of θ_1 . In fact, the optimal weights under the current sampling scheme can be approximated by

$$w_i = \frac{1}{D(1 + (y_i - y_1)^2)}, \quad i = 1, \dots, 5,$$

where

$$D = \sum_{i=1}^5 \frac{1}{1 + (y_i - y_1)^2}$$

is a normalizing constant. This set of approximate optimum adaptive weights actually generates quite similar performances when compared with those based on the optimal weights in this simulation study. In fact, the MSE for the WLE is about 0.679 using the approximate optimal weights and 0.669 using the optimal weights in this simulation study.

The MSEs for these five estimators are given in Table 2. The Bayes estimator is guaranteed theoretically to be the optimal estimator if the values of hyperparameters are known. We also note that the MSE of MLE is the largest. We then continue to compare the three remaining competitors: the empirical Bayes, James–Stein and the weighted likelihood estimators. The MSE values confirm that the empirical Bayes estimator is similar but not as good as the James–Stein estimator. Furthermore, we observe that the WLE outperforms the other estimators except the Bayes estimator with the correct knowledge of the prior. The mean absolute errors (MAE) for all five estimators are also provided in Table 2. The pattern is quite similar and the ordering of the MAEs is exactly the same as that of the MSEs. Since both the empirical Bayes and the James–Stein estimators assign equal weights to related populations, they simply ignore the fact that other populations might not be all equally influential with respect to the inference on the first population. We believe that WLE utilizes information contained in other samples in a more efficient way than the empirical Bayes and the James–Stein estimators. In summary, the WLE outperforms all its competitors except for the Bayes estimator with exact knowledge of the prior distribution.

TABLE 2: Comparison of mean squared errors and mean absolute errors.

	MLE	EB	JS	WLE	Bayes	Bayes/WLE
MSE	1.00	0.94	0.80	0.67	0.50	75%
MAE	0.79	0.74	0.71	0.65	0.56	86%

If reliable external or prior knowledge is available, the Bayes procedure should be used because it is theoretically guaranteed to be the optimal decision. If no prior knowledge is available, a realistic Bayesian analysis would employ a conditionally independent and identically distributed prior. For simplicity, we put a normal prior on μ_θ . The mean of the prior will be set to various values in this simulation study. One should assign a large value for the variance of μ_θ to reflect the fact that no external or prior knowledge is available. Therefore, any analysis will be

dominated by the data instead of the value of the prior mean. The parameter σ_θ^2 does not have any simple family of conjugate prior distributions because its marginal likelihood depends in a complex way on the data from all populations. However, if σ_θ^2 follows an inverse-gamma distribution, then the conditional distribution $p(\sigma_\theta^2 | \mu_\theta, x_1, \dots, x_m)$ also follows an inverse-gamma distribution. This conditional conjugacy allows σ_θ^2 to be updated easily using the Gibbs sampler (Gelfand & Smith 1990).

An inverse-gamma distribution denoted by $\mathcal{IG}(\varepsilon, \varepsilon)$ tries to assume some kind of non-informativeness within the conditionally conjugate family, with ε set to some low values such as 1 or 0.01. The results of using a combination of a normal prior distribution such as $N(0, 20^2)$ and an inverse-gamma distribution are given in Table 3. In general, we see that the MSEs of the Bayes estimator are universally smaller than those of the WLE. However, the fact that the ratios range from about 91% to 97% implies that the approximation using the WLE is quite satisfactory in this simulation study. Observe that the results do not change significantly for different choices of the prior mean. We note that the MSEs of the Bayes estimators by using other normal distributions such as $N(\mu_\theta, 10^2)$ and an inverse-gamma distribution with small values assigned to ε are quite similar to the results presented in Table 3. When no reliable external or prior information is available, a uniform non-informative prior might be a natural choice. The MSEs using some uniform prior distributions for the prior mean are provided in Table 4. We observe that the performances of the Bayes estimators using different uniform priors are comparable with that of the WLE. This seems to suggest that the WLE is very successful in approximating a full Bayesian inference in this simulation study when no reliable prior knowledge is available. We remark that all Bayesian estimates were obtained using *WinBugs* with 1500 runs within each iteration of the simulation study.

TABLE 3: Mean squared error based on normal and uniform prior distributions.

	$\mathcal{IG}(1, 1)$	$\mathcal{IG}(0.1, 0.1)$	$\mathcal{IG}(0.01, 0.01)$
$N(0, 20^2)$	0.61	0.61	0.62
$N(5, 20^2)$	0.63	0.63	0.63
$N(10, 20^2)$	0.66	0.65	0.65
	$\mathcal{IG}(1, 1)$	$\mathcal{IG}(0.1, 0.1)$	$\mathcal{IG}(0.01, 0.01)$
$\mathcal{U}(-10, 10)$	0.55	0.60	0.74
$\mathcal{U}(-20, 20)$	0.68	0.71	0.74

TABLE 4: Observed effects of special preparation on SAT-V scores in eight randomized experiments. Estimates are based on separate analyses for the eight experiments. They are from Rubin (1981).

School	Est. Treat. Effects	S.E.
A	28.39	14.9
B	7.94	10.2
C	-2.75	16.3
D	6.82	11.0
E	-0.64	9.4
F	0.63	11.4
G	18.01	10.4
H	12.16	17.6

5. CASE STUDY: COMBINING INFORMATION FROM EDUCATIONAL EXPERIMENTS

We illustrate weighted likelihood inference using a classical example of educational testing as described and analyzed in Gelman, Carlin, Stein & Rubin (2003) and compare the weighted likelihood estimator with other estimators.

5.1. Description of educational experiments and the maximum likelihood estimator.

A study was performed on an educational testing service to analyze the effects of special coaching programs on test scores. Separate randomized experiments were performed to estimate the effects of coaching programs for the SAT-V (Scholastic Aptitude Test-Verbal) in each of eight schools. The outcome variable was the score on a special administration of the SAT-V. There is no prior evidence to suggest that any of the eight schools was more effective than any other school. The estimated coaching effects (Est. Treat. Effects) and their standard errors (S.E.) were obtained by an analysis of covariance adjustment appropriate for a completely randomized experiment. Therefore, the sampling variances σ_j^2 are assumed to be known in this study. The summary statistics are provided in Table 5.

TABLE 5: Posterior means (PM) and standard deviations (SD) of the Bayes estimates for 8 schools with normal/uniform priors.

θ_k	$\tau^2 = 10^6$		$\tau^2 = 10^4$		$\tau^2 = 10^2$	
	PM	SD	PM	SD	PM	SD
θ_A	10.74	9.24	10.32	7.91	10.09	7.98
θ_B	7.19	6.40	7.19	6.23	7.18	6.01
θ_C	5.21	7.26	5.44	7.20	5.24	7.68
θ_D	6.54	6.73	6.87	6.41	6.91	6.42
θ_E	4.12	6.08	4.70	5.78	4.41	5.92
θ_F	5.23	7.00	5.55	6.21	5.32	6.43
θ_G	9.65	7.26	9.61	6.93	9.63	6.60
θ_H	7.42	8.37	7.58	7.52	7.19	7.34

Our goal is to estimate the treatment effect of each school, namely θ_i , $i = 1, \dots, 8$. The MLEs of these θ_i are simply the estimated treatment effects given by Table 5. The MLEs all ignore the fact that the eight schools are connected or related. Although the estimated treatment effects seem to be quite different from each other, it is very difficult to distinguish them statistically. For example, we can construct 95% confidence intervals by applying a simple normal analysis on each of the schools. All confidence intervals overlap with each other. Thus, there might be a loss of information by simply using the estimated treatment effects listed in Table 5. On the other hand, a pooled estimate of the treatment effect (7.9) does not seem to be reasonable, as discussed in Gelman, Carlin, Stein & Rubin (2003). The pooled estimate is calculated under the assumption that all parameters of these eight schools are identical. This assumption does not seem to hold because the estimated treatment effect of school A stands outside of the 95% confidence interval constructed by using the pooled estimate. Therefore a simple pooling does not seem to be a reasonable strategy. In this case study, we apply both hierarchical Bayes and the weighted likelihood methods to combine data across these eight schools in order to estimate the treatment effect of each school.

5.2. Bayesian estimates with different priors.

To fit a hierarchical Bayesian model for the SAT coaching data, we work with a simple two-level model of data X_i with group level effects θ_k as

$$\bar{X}_i | \theta_k \sim N(\theta_k, \sigma_k^2), \quad \theta_k \sim N(\mu_\theta, \sigma_\theta^2), \quad k = 1, \dots, m, \quad m > 3.$$

TABLE 6: Posterior means (PM) and standard deviations (SD) of the Bayes estimates for 8 schools with normal/ $\mathcal{IG}(0.01, 0.01)$ priors.

θ_k	$\tau^2 = 10^6$		$\tau^2 = 10^4$		$\tau^2 = 10^2$	
	PM	SD	PM	SD	PM	SD
θ_A	8.28	5.16	8.14	5.13	7.07	4.91
θ_B	7.31	4.63	7.17	4.60	6.18	4.41
θ_C	6.94	5.02	6.81	4.99	5.78	4.78
θ_D	7.24	4.53	7.11	4.51	6.10	4.31
θ_E	6.67	4.65	6.54	4.62	5.58	4.42
θ_F	7.05	4.69	6.92	4.66	5.92	4.47
θ_G	7.92	4.73	7.78	4.70	6.76	4.51
θ_H	7.58	4.68	7.43	4.65	6.39	4.43

TABLE 7: Estimates of the treatment effects for Schools A–H by using different methods.

θ_k	Bayes(G)	Bayes(U)	E-Bayes	J–S	WLE	MLE
θ_A	8.28	10.74	16.22	20.94	24.70	28.39
θ_B	7.31	7.19	9.39	5.85	7.92	7.94
θ_C	6.94	5.21	5.82	−2.02	−0.84	−2.75
θ_D	7.24	6.54	9.02	5.03	7.00	6.82
θ_E	6.67	4.12	6.52	−0.47	0.89	−0.64
θ_F	7.05	5.23	6.95	0.46	1.93	0.63
θ_G	7.92	9.65	12.76	13.28	16.18	18.01
θ_H	7.58	7.42	10.80	8.97	11.38	12.16

We first choose the prior as specified in Gelman, Carlin, Stein & Rubin (2003):

$$\mu_\theta \sim N(0, \tau^2) \quad \text{and} \quad \sigma_\theta^2 \sim \mathcal{U}(0, 10^3).$$

Gelman, Carlin, Stein & Rubin (2003) used 10^6 as the value for τ^2 and we set additional values of 10^2 and 10^4 , respectively, to derive Bayesian estimates. The posterior means derived from the Bayesian analysis are given in Table 6. We observe that the Bayesian estimates are quite different from the estimated treatment effects listed in Table 5 except for schools B and D . Furthermore, changing the value of τ^2 does not have much effect on the Bayesian estimates for the eight schools.

In order to make further comparisons, we also employ an inverse-gamma distribution for the variance. The common choices for the value of the parameter ε of an inverse-gamma distribution are 1, 0.1 and 0.01. We find that changing the value of ε does not change the results significantly. For brevity, we only provide the results when ε equals 0.01 in Table 7. We remark that all Bayesian estimates presented in Tables 6 and 7 are obtained using *WinBugs* with 1500 runs for the MCMC sampling.

5.3. Comparison of the weighted likelihood estimator and others.

The WLE using the optimum adaptive weights is also calculated for each school together with the James–Stein and the empirical Bayes estimates. They are provided in Table 7. Since there

are many sets of Bayesian estimates due to different choices of the hyperparameters and the prior distributions, we only select the first group of Bayesian estimates using either a uniform or an inverse-gamma prior for the prior variance and denote them by $\text{Bayes}(U)$ and $\text{Bayes}(G)$, respectively. In general, these estimates share some common features. For example, all the estimates for the schools B and D are quite similar to one another. However, for schools A, C and E, the estimates are quite different from one another.

To quantify the relative connections among all these estimates, we apply a hierarchical clustering algorithm to these groups of estimates. There are 3 ways to apply a hierarchical clustering algorithm, namely the simple, average and complete linkage. The three corresponding partitions all give a unified ordering: James–Stein estimator, WLE, MLE, empirical Bayes estimator, Bayesian estimators with uniform prior, and Bayesian estimators with inverse-gamma prior in this case. The grouping pattern is represented in Figure 1. It can be seen that the group of Bayesian estimates with a uniform prior closely resembles the group of Bayesian estimates with an inverse-gamma prior. Together with the empirical Bayes estimates, they form one cluster. This is not surprising due to the close relationship between the empirical Bayes estimator and traditional Bayesian estimator. The group of James–Stein estimates belongs to another cluster together with the groups of MLEs and WLEs. Furthermore, the groups of WLEs and MLEs are assigned into one sub-cluster. The overall grouping seems to be consistent with the known relationships among these estimators. However, it is a bit surprising to see that the group of James–Stein estimates is at the far end of the partition. This might be due to the fact that the James–Stein estimates all shrink the MLEs towards 0 by the same percentage. Therefore, they behave somewhat differently in this study. In general, the group of WLEs does share some common features with the groups of empirical Bayes, James–Stein and Bayes estimates using different priors.

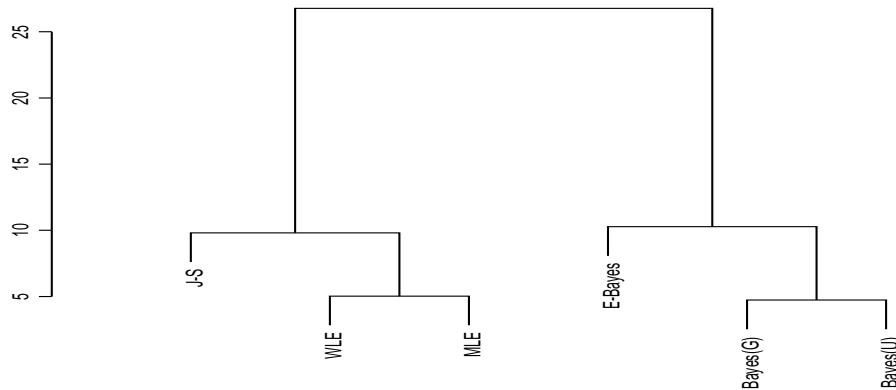


FIGURE 1: Grouping results of all the estimators by using a hierarchical clustering.

6. DISCUSSION

We have shown that the WLE is an approximate Bayes rule under empirical entropy loss. In addition, results from simulation studies seem to suggest that this approximation could be satisfactory when no reliable external or prior information is available. Since WLE usually takes the form of a convex combination of individual MLEs for many commonly used distributions of the exponential family, it is much less computationally intensive than the full Bayesian analysis with MCMC sampling. Although both are programmed in the same software package *R*, the full Bayesian inference with MCMC sampling usually takes about 1.5 hour to complete in our simulation studies, while the weighted likelihood method takes only a few seconds. Therefore

the proposed weighted likelihood approach could be a useful tool for an exploratory data analysis. Future work includes further investigation on different adaptive weights and comparing their strengths.

APPENDIX

Proof of Theorem 1. We are seeking the best linear combination of MLEs derived from the marginal distributions. As before, let us consider the MSE of the WLE. Writing

$$\tilde{\theta}_1 = \boldsymbol{\lambda}^\top \hat{\boldsymbol{\theta}} = \sum_{i=1}^m \lambda_i \hat{\theta}_i,$$

we can calculate

$$\begin{aligned} \text{MSE}(\tilde{\theta}_1) &= \text{E} \left\{ \sum_{i=1}^m \lambda_i (\hat{\theta}_i - \theta_1) \right\}^2 \quad \left(\text{since } \sum_{i=1}^m \lambda_i = 1 \right) \\ &= \sum_{i=1}^m \sum_{j=1}^m \lambda_i \lambda_j \{ \text{cov}(\hat{\theta}_i, \hat{\theta}_j) + (\theta_i - \theta_1)(\theta_j - \theta_1) \} \\ &\leq \boldsymbol{\lambda}^\top \text{cov}(\hat{\boldsymbol{\theta}}) \boldsymbol{\lambda} + \boldsymbol{\lambda}^\top \mathbf{B} \mathbf{B}^\top \boldsymbol{\lambda} \quad (\text{by the assumptions}) \\ &= \boldsymbol{\lambda}^\top (\mathbf{V} + \mathbf{B} \mathbf{B}^\top) \boldsymbol{\lambda}. \end{aligned}$$

Minimizing $\boldsymbol{\lambda}^\top (\mathbf{V} + \mathbf{B} \mathbf{B}^\top) \boldsymbol{\lambda}$ with the constraint $\lambda_1 + \cdots + \lambda_m = 1$ yields

$$\tilde{\theta}_1^* = \sum_{i=1}^m \lambda_i^* \theta_i,$$

where

$$\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)^\top = \frac{(\mathbf{V} + \mathbf{B} \mathbf{B}^\top)^{-1} \mathbf{1}}{\mathbf{1}^\top (\mathbf{V} + \mathbf{B} \mathbf{B}^\top)^{-1} \mathbf{1}}.$$

Proof of Theorem 2. The log-likelihood function for n_1 observations which are independent and identically distributed from the above distribution family can be written as

$$\ln L_1(X_{11}, \dots, X_{1n_1}; \theta_1) = A_1(\theta_1) \sum_{j=1}^{n_1} S(x_{1j}) + n_1 B_1(\theta_1) + \text{constant},$$

with

$$\frac{\partial \ln L_1(X_{11}, \dots, X_{1n_1}; \theta_1)}{\partial \theta_1} = A_1'(\theta_1) \sum_{i=1}^{n_1} S(x_{1i}) + n_1 B_1'(\theta_1) = n_1 \{ A_1'(\theta_1) T(\mathbf{x}^1) + B_1'(\theta_1) \}, \quad (9)$$

where

$$T(\mathbf{x}^1) = \frac{1}{n_1} \sum_{j=1}^{n_1} S(x_{1j}).$$

It is known (Lehmann 1983, p. 123) that for the exponential family, the necessary and sufficient condition for an unbiased estimator to achieve the Cramér–Rao lower bound is that there exists $D(\theta_1)$ such that

$$\frac{\partial \ln L_1(X_{11}, \dots, X_{1n_1}; \theta_1)}{\partial \theta_1} = n_1 D_1(\theta_1) \{ T(\mathbf{x}^1) - \theta_1 \}, \quad \forall \theta_1.$$

Under the condition

$$\frac{\partial \ln L_1(x_{i1}, \dots, x_{in_i}; \theta_1)}{\partial \theta_1} = n_i D_1(\theta_1) \{T(\mathbf{x}^i) - \theta_1\},$$

it can be seen that $T(\mathbf{x}^1)$ is the traditional MLE for θ_1 which achieves the Cramér–Rao lower bound and is unbiased as well. Then we have

$$\frac{\partial \ln \text{WL}}{\partial \theta_1} = \sum_{i=1}^m \lambda_i n_i D_1(\theta_1) \{T(\mathbf{x}^i) - \theta_1\},$$

and so the WLE is given by $\tilde{\theta}_1 = \sum_{i=1}^m w_i T(\mathbf{x}^i)$, where $w_i = n_i \lambda_i / \sum_{i=1}^m \lambda_i n_i$.

Proof of Theorem 3. From (9), the WLE satisfies

$$\sum_{i=1}^m \lambda_i n_i \{A'_1(\theta_1) T(\mathbf{x}^i) + B'_1(\theta_1)\} = 0,$$

which implies

$$A'_1(\theta_1) \left\{ \sum_{i=1}^m t_i \lambda_i T(\mathbf{x}^i) \right\} + B'_1(\theta_1) = 0,$$

where $t_i = n_i / \sum_{i=1}^m \lambda_i n_i$. Therefore the WLE of θ_1 takes the form

$$\tilde{\theta}_1 = g \left\{ \sum_{i=1}^m w_i T(\mathbf{x}^i) \right\},$$

where $w_i = n_i \lambda_i / \sum_{i=1}^m \lambda_i n_i$.

ACKNOWLEDGEMENTS

The author would like to thank the Editor, the Associate Editor and two anonymous referees for their valuable suggestions and comments that greatly improved the quality of this paper. The author also would like to thank Professors Neal Madras and Peter Peskun for their careful readings of the revised manuscript and very helpful comments. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- C. Agostinelli & M. Markatou (2001). Test of hypothesis based on weighted likelihood methodology. *Statistica Sinica*, 11, 499–514.
- H. Akaike (1973). Information theory and extension of the maximum likelihood principle. In *Second International Symposium on Information Theory: Tsahkadsor, Armenia, USSR, September 2–8, 1971* (B. N. Petrov & F. Csaki, eds.), Akadémiai Kiadó, Budapest, pp. 267–281.
- D. R. Cox (1981). Combination of data. *Encyclopedia of Statistical Sciences*, 2, John Wiley, pp. 45–52.
- B. Efron (1996). Empirical methods for combining likelihoods. *Journal of the American Statistical Association*, 96, 538–550.
- B. Efron & C. Morris (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68, 117–130.
- S. Eguchi & J. Copas (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society Series B*, 60, 709–724.
- A. E. Gelfand & A. F. M. Smith (1990). Sampling based approach to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.
- A. Gelman, J. B. Carlin, H. S. Stein & D. B. Rubin (2003). *Bayesian Data Analysis*. Chapman & Hall, New York.
- M. Ghosh & M. C. Yang (1988). Simultaneous estimation of multivariate precision matrix. *The Annals of Statistics*, 16, 278–291.

- P. Gustafson, S. Hossain & L. McCandless (2005). Innovative Bayesian methods for biostatistics and epidemiology. Chapter 26 in *Handbook of Statistics, Volume 25 : Bayesian Thinking, Modeling and Computation* (D. K. Dey & C. R. Rao, eds.), Elsevier, New York.
- Handbook of Statistics, Volume 25 : Bayesian Thinking, Modeling and Computation
- L. Györfi, M. Kohler, A. Krzyzak & H. Walk (2002). *A Distribution-Free Theory of Non-Parametric Regression*. Springer, New York.
- F. Hu (1994). *Relevance Weighted Smoothing and a new Bootstrap Method*. Doctoral Dissertation, Department of Statistics, The University of British Columbia, Vancouver, Canada.
- F. Hu (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators. *The Canadian Journal of Statistics*, 30, 45–59.
- F. Hu & W. F. Rosenberger (2000). Analysis of time trends in adaptive designs with application to a neurophysiology experiment. *Statistics in Medicine*, 19, 2067–2075.
- F. Hu, W. F. Rosenberger & J. V. Zidek (2000). The relevance weighted likelihood for dependent data. *Metrika*, 51, 223–243.
- F. Hu & J. V. Zidek (2001). The relevance weighted likelihood with applications. In *Empirical Bayes and Likelihood Inference (Montréal, QC, 1997)* (S. E. Ahmed & N. M. Reid, eds.), Springer, New York. pp. 211–235.
- W. James & C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, vol. 1, pp. 361–379.
- S. Kullback (1959). *Information Theory and Statistics*. Wiley, New York.
- E. L. Lehmann (1983). *Theory of Point Estimation*. Wiley, New York.
- M. Markatou, A. Basu & B. G. Lindsay (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*, 92, 215–232.
- M. Markatou, A. Basu & B. G. Lindsay (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93, 740–750.
- D. B. Rubin (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6, 377–401.
- J. G. Staniswalis (1989). The kernel estimate of a regression function in likelihood-based methods. *Journal of the American Statistical Association*, 89, 276–283.
- C. J. Stone (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5, 595–645.
- R. J. Tibshirani & T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559–567.
- U.S. National Research Council (1992). *Combining Information: Statistical Issues and Opportunities for Research*. National Academy Press, Washington, DC.
- X. Wang (2001). *Weighted Likelihood Estimation*. Doctoral Dissertation, Department of Statistics, The University of British Columbia, Vancouver, Canada.
- X. Wang, C. van Eeden & J. V. Zidek (2004). Asymptotic properties of the weighted likelihood estimators. *Journal of Statistical Planning and Inference*, 119, 37–54.
- X. Wang & J. V. Zidek (2005). Choosing likelihood weights by cross-validation. *The Annals of Statistics*, 33, 463–500.

Received 30 June 2004

Accepted 26 August 2005

Xiaogang WANG: stevenw@mathstat.yorku.ca

Department of Mathematics and Statistics

York University, Toronto, Ontario

Canada M3J 1P3