EMPIRICAL BAYES ESTIMATION AND UNBIASED PREDICTION THEORY

Yushan Xiao*

This paper considers empirical Bayes estimators from the unbiased prediction point of view. The consideration is devoted to the normal case. It is shown that some empirical Bayes estimators are the best or the best conditional unbiased.

Key Words and Phrases: empirical Bayes estimation, unbiased predictor, conditionally unbiased predictor, James-Stein estimator.

1. Introduction

The correctness of Bayes methods depends very heavily on the accuracy of the knowledge of the prior distribution. When such knowledge is incomplete, the data can often give information about it, which are used in their places. This technique is called an empirical Bayes method. See Berger (1985) and references therein.

Let $(\lambda_1, X_1), \dots, (\lambda_n, X_n)$ be independent and identically distributed (*i.i.d.*) random pairs where λ_i are *i.i.d.* with a prior distribution G and the conditional distribution of X_i given λ_i has a known probability density function $f(x|\lambda_i)$ with respect to a σ -finite measure. $X_1, \dots, X_n$ are observable, whereas $\lambda_1, \dots, \lambda_n$ are not. Based on $X=(X_1, \dots, X_n)$, we consider a problem of estimating λ_n on the basis of the squared loss function. If G were known, then the Bayes estimator would be

$$\delta_G(X) = E_G(\lambda_n|X) = E_G(\lambda_n|X_n).$$

Since G is assumed to be unknown, the estimator is not available. However, since $X_1, \dots, X_n$ are *i.i.d.* with density $\int f(x|\lambda)dG(\lambda)$, it is reasonable to expect that X contains some information about G , which enables us to approximate δ_G . The resulting estimator is called an empirical Bayes estimator. The purpose of this paper is to explore the properties of empirical Bayes estimators.

Since the problem is to estimate an unknown random variable λ_n , we can consider an empirical Bayes estimator as a predictor in a statistical prediction problem. See Takeuchi (1975) for the details of statistical prediction problems. Some empirical Bayes estimators become unbiased predictors in terms of a statistical prediction theory. Hence, we shall consider their performances among unbiased predictors.

In Section 2 we shall define a statistical prediction problem and some concepts used in it. We also state some useful results. In Section 3 we shall

Received April, 1999. Revised August, 1999. Accepted September, 1999.

* Graduate School of Science and Technology, Kumamoto University, Kumamoto 860-8555, Japan.

take up the most familiar case where both priors and conditional distributions are normal. Unfortunately, the best unbiased predictors do not exist except in special cases. Hence we shall consider conditional unbiased predictors which were introduced into a statistical prediction problem by Ishii (1978). The concept seems more appropriate in the empirical Bayes setting. See Section 3.2.

2. Unbiased prediction theorem

Let X be an observable random vector and Y a unknown random variable. Suppose that the joint distribution of X and Y depends on an unknown parameter θ . Then the problem of predicting the value of Y based on X is called a statistical prediction problem. See Takeuchi (1975). We adopt a squared error as a loss function in the subsequent discussion. A statistic δ is said to be unbiased if

$$(2.1) \quad E_{\theta}\delta(X) = E_{\theta}Y, \quad \text{for all } \theta.$$

An unbiased predictor is said to be locally best unbiased at $\theta = \theta_0$ if it minimizes the risk at θ_0 among all unbiased predictors. If a locally best unbiased predictor is determined irrespective of θ , then it is said to be the best unbiased predictor. If not, then there dose not exist the best unbiased predictor.

Let

$$g(\theta, X) = E_{\theta}(Y|X),$$

which is equal to the Bayes estimator in the empirical Bayes setting. Suppose $\theta = (\theta_1, \dots, \theta_p)$ and let $f(x, \theta)$ be the density function of X with respect to a σ -finite measure. We also assume that $g(\theta, x)$ and $\log f(x, \theta)$ are differentiable with respect to θ . Let

$$I_{ij}(\theta) = E_{\theta} \left\{ \frac{\partial}{\partial \theta_i} \log f(X, \theta) \frac{\partial}{\partial \theta_j} \log f(X, \theta) \right\}, \quad i, j = 1, \dots, p$$

and

$$I(\theta) = (I_{ij}(\theta)),$$

$${}^t\alpha(\theta) = \left(E_{\theta} \frac{\partial}{\partial \theta_1} g(\theta, X), \dots, E_{\theta} \frac{\partial}{\partial \theta_p} g(\theta, X) \right),$$

where ${}^t\alpha(\theta)$ denotes the transposition of the vector $\alpha(\theta)$. Then the following result gives a lower bound for a risk of an unbiased predictor. The inequality is a version of the Cramér-Rao inequality to the prediction problem. For the proof, see Takeuchi (1975, p.16). See also Ishii (1969) for its generalization.

THEOREM 1. *Under some mild conditions on the distribution of X ,*

$$(2.2) \quad E_{\theta_0}(\delta(X) - Y)^2 \geq E_{\theta_0}(Y - g(\theta_0, X))^2 + {}^t\alpha(\theta_0)I(\theta_0)^{-1}\alpha(\theta_0),$$

for any unbiased predictor δ . The equality holds if and only if

$$(2.3) \quad \delta(X) = g(X, \theta_0) + \sum_{i=1}^p t_i \frac{\partial}{\partial \theta_i} \log f(X, \theta_0),$$

for some constants $t_1, \dots, t_p$.

Example 1. Let $X_i = (X_{1i}, X_{2i})$, $i=1, \dots, n+1$, be *i.i.d.* two-dimensional normal random vector with unknown mean vector μ and unknown covariance matrix Σ . We consider the problem of predicting $Y = X_{2,n+1}$ after observing $X = (X_1, \dots, X_n, X_{1,n+1})$. It is shown that there does not exist the best unbiased predictor by the existence of the locally best unbiased predictor dependent on the unknown parameter. See Takeuchi (1995, p.18).

In the above example, it seems natural to consider the following predictor,

$$(2.4) \quad \delta(X) = \bar{X}_2 + V_{21} V_{11}^{-1} (X_{1,n+1} - \bar{X}_1),$$

where

$$\bar{X} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^n X_i, \quad V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'.$$

Ishii (1978) gave its justification from the conditional unbiasedness point of view. See Takada (1982) for another justification.

Let $W(X)$ be a statistic. Then a predictor δ is said to be conditionally unbiased if

$$E_\theta(\delta(X) | W) = E_\theta(Y | W), \quad a.e.$$

for every θ . The conditional unbiasedness means the unbiasedness. A conditionally unbiased predictor is said to be the best if it minimizes the risk among all conditionally unbiased predictors.

Example 2. (Continued). Let $W = X_{1,n+1}$. Then δ in (2.4) is conditionally unbiased. Since $\bar{X}$ and V are complete and sufficient for the family of distribution of $X_1, \dots, X_n$, δ becomes the best conditionally unbiased predictor. See Ishii (1978, p. 70).

In order to get the best unbiased predictor, the following result may be useful in some examples. See Takeuchi (1975, p. 13) for the proof.

THEOREM 2. *A statistic $T(X)$ is sufficient for the family of distributions of X and*

$$(2.5) \quad g(\theta, X) = h(X) + \phi(\theta, T),$$

for some functions $h(\cdot)$ and $\phi(\cdot, \cdot)$. For any unbiased predictor δ let

$$\delta^*(T) = E(\delta(X) - h(X) | T) + h(X).$$

Then δ^* is also unbiased and the risk of δ^* is less than or equal to that of δ . Furthermore, if T is complete, then δ^* is the best unbiased predictor.

3. Empirical Bayes estimation

We now consider a particular distributional setting, namely the conditional distribution of X_i given λ_i is normal with mean λ_i and variance σ^2 and the prior distribution of λ_i is also normal with mean μ and variance τ^2 . Then the conditional distribution of $Y=\lambda_n$ given X_n becomes normal with mean $B\mu + (1-B)X_n$ and variance $(1-B)\sigma^2$ where $B=\sigma^2/(\sigma^2+\tau^2)$. So the Bayes estimator of λ_n becomes

$$(3.1) \quad \delta_G(X) = B\mu + (1-B)X_n.$$

We consider three circumstances about the parameters μ , τ^2 and σ^2 . First $\theta=\mu$ is the only unknown parameter. Second $\theta=(\mu, \tau^2)$ is unknown, while σ^2 is known. Third all parameters are unknown. So $\theta=(\mu, \tau^2, \sigma^2)$.

Let δ be any empirical Bayes estimator. Then

$$(3.2) \quad \begin{aligned} E_\theta(Y - \delta(X))^2 &= E_\theta(Y - \delta_G(X))^2 + E_\theta(\delta_G(X) - \delta(X))^2 \\ &= (1-B)\sigma^2 + E_\theta(\delta_G(X) - \delta(X))^2. \end{aligned}$$

The first term in (3.2) denotes the Bayes risk, whereas the second term is the risk due to using δ instead of δ_G .

3.1. $\theta=\mu$

Since the marginal distribution of X_i is normal with mean μ and variance $\sigma^2 + \tau^2$, μ is estimated by the sample mean $\bar{X} = 1/n \sum_{i=1}^n X_i$. So from (3.1) we consider the following empirical Bayes estimator

$$(3.3) \quad \delta_1(X) = B\bar{X} + (1-B)X_n.$$

It is easy to see that δ_1 is unbiased in the sense of (2.1). Now we get the next result.

THEOREM 3. δ_1 is the best unbiased predictor and

$$(3.4) \quad E(Y - \delta_1(X))^2 = (1-B)\sigma^2 + \frac{B\sigma^2}{n}.$$

PROOF. From (3.1)

$$g(\theta, X) = B\mu + (1-B)X_n,$$

and $\bar{X}$ is complete and sufficient. Hence it follows from Theorem 2 that δ_1 is the best unbiased predictor. The risk of δ_1 is easily obtained from (3.2).

3.2. $\theta=(\mu, \tau^2)$

We assume that $n \geq 4$. Let $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$. Then it is easy to see that

$$E_\theta\left(\frac{1}{S^2}\right) = \frac{1}{(\sigma^2 + \tau^2)(n-3)},$$

which suggests $\hat{B} = (n-3)\sigma^2/S^2$ as an estimator of B . So from (3.1) we consider the following empirical Bayes estimator

$$(3.4) \quad \delta_2(X) = \hat{B}\bar{X} + (1 - \hat{B})X_n.$$

This estimator is the component of the James-Stein estimator (e.g. Lehmann, 1983, p. 305). Note that each $(X_i - \bar{X})/S^2$, $i=1, \dots, n$, has the same distribution and $\sum_{i=1}^n (X_i - \bar{X})/S^2 = 0$. So

$$E_\theta \hat{B}(X_n - \bar{X}) = 0,$$

which implies

$$E_\theta \delta_2(X) = E_\theta X_n - E_\theta \hat{B}(X_n - \bar{X}) = E_\theta Y.$$

Hence δ_2 is unbiased in the sense of (2.1). The argument of Lehmann (1983, p. 305) leads to

$$(3.5) \quad E_\theta (Y - \delta_2(X))^2 = (1 - B)\sigma^2 + \frac{3B\sigma^2}{n}.$$

Some optimality about δ_2 may be expected among all unbiased predictors, which is, however, denied by the following result.

THEOREM 4. *There does not exist the best unbiased predictor when $\theta = (\mu, \tau^2)$ is unknown.*

PROOF. The straightforward calculation shows that the inequality in (2.2) becomes

$$(3.6) \quad E_{\theta_0} (Y - \delta(X))^2 \geq (1 - B_0)\sigma^2 + \frac{B_0\sigma^2}{n},$$

for any unbiased predictor δ where $\theta_0 = (\mu_0, \tau_0^2)$ and $B_0 = \sigma^2/(\sigma^2 + \tau_0^2)$. It is easy to see that the lower bound is attained by an unbiased predictor

$$(3.7) \quad \delta_0(X) = B_0\bar{X} + (1 - B_0)X_n,$$

and that δ_0 satisfies (2.3). Hence δ_0 is the locally best unbiased predictor at $\theta = \theta_0$. However, δ_0 does depend on θ_0 , and hence the best unbiased predictor does not exist.

Since the best unbiased predictor does not exist, we consider conditional unbiasedness instead of unbiasedness. In the empirical Bayes settings it seems natural to set $W = X_n$ in the definition of the conditional unbiasedness. Then δ is conditionally unbiased if and only if for each x_n

$$(3.8) \quad E_\theta \delta(X_1, \dots, X_{n-1}, x_n) = B\mu + (1 - B)x_n.$$

In the classical empirical Bayes methods the empirical Bayes estimator of $Y = \lambda_n$ is usually a function of X_n whose form depends only on $X_1, \dots, X_{n-1}$ (e.g. Robbins, 1964). The conditional unbiasedness requires that the expectation of the form should be equal to that of the Bayes estimator. So the concept of conditional unbiasedness seems to be more suitable here.

We assume that $n \geq 7$. Let

$$\bar{X}' = \frac{1}{n-1} \sum_{i=1}^{n-1} X_i, \quad S'^2 = \sum_{i=1}^{n-1} (X_i - \bar{X}')^2,$$

and

$$(3.9) \quad \delta'_2(X) = \hat{B}' \bar{X}' + (1 - \hat{B}') X_n,$$

where $\hat{B}' = (n-4)\sigma^2/S'^2$. Then we get the following result.

THEOREM 5. *The empirical Bayes estimator δ'_2 in (3.9) is the best conditionally unbiased predictor and*

$$(3.10) \quad E_\theta(Y - \delta'_2(X))^2 = (1 - B)\sigma^2 + \frac{3B\sigma^2(n-2)}{(n-6)(n-1)}.$$

PROOF. The conditional unbiasedness easily follows from the independence of X_n , $\bar{X}'$ and $\hat{B}'$. Since $(\bar{X}', S'^2)$ is complete and sufficient for the family of distributions of $X_1, \dots, X_{n-1}$, δ'_2 becomes the best conditionally unbiased predictor. The proof of (3.10) is given in the appendix.

Comparing (3.5) and (3.10), the risk of δ'_2 is great than that of δ_2 . However, δ_2 is not conditionally unbiased.

3.3. $\theta = (\mu, \tau^2, \sigma^2)$

First we consider the case that $\rho = \sigma^2/\tau^2$ is known. Thus $B = \rho/(1 + \rho)$ becomes known, so that from (3.1) the empirical Bayes estimator is

$$(3.11) \quad \delta'_1(X) = B\bar{X} + (1 - B)X_n,$$

which is equal to that in (3.3) where μ is the only unknown parameter. Hence the risk is equal to (3.4).

THEOREM 6. *If $\rho = \sigma^2/\tau^2$ is known, then δ'_1 in (3.11) is the best unbiased predictor.*

PROOF. Since $g(\theta, X)$ satisfies (2.5) and $(\bar{X}, S^2)$ is complete and sufficient for the family of distributions of X , the result follows from Theorem 2.

Next we consider the case that all parameters are unknown, but there exist a random variable S_0^2 independent of X and S_0^2/σ^2 is distributed as a chi-squared distribution with f degrees of freedom. We assume that $n \geq 4$ and consider

$$(3.12) \quad \delta_3(X, S_0^2) = \tilde{B}\bar{X} + (1 - \tilde{B})X_n$$

as an empirical Bayes estimator, where

$$\tilde{B} = \left(\frac{n-3}{f+2} \right) \frac{S_0^2}{S^2}.$$

δ_3 is the component of the James-Stein estimator. See Efron and Morris (1972). It is easy to see that δ_3 is unbiased in the sense of (2.1). From Theorem 7.1 of Efron and Morris (1972)

$$(3.13) \quad E_{\theta}(Y - \delta_3(X, S_0^2)) = (1 - B)\sigma^2 + \frac{2n+3f}{n(f+2)}B\sigma^2.$$

It follows from Theorem 2 that the lower bound for a risk of an unbiased predictor is equal to the right hand side of the inequality of (3.6) and that the predictor in (3.7) is the locally best unbiased. Hence, we have the following result.

THEOREM 7. *There does not exist the best unbiased predictor when $\theta = (\mu, \tau^2, \sigma^2)$ is unknown.*

Since the best unbiased predictor does not exist, we consider conditional unbiasedness instead of unbiasedness as in Section 3.2. A predictor δ is conditionally unbiased if and only if

$$E_{\theta}\delta(X_1, \dots, X_{n-1}, x_n, S_0^2) = B\mu + (1 - B)x_n,$$

for each x_n . We assume that $n \geq 7$. Consider

$$(3.14) \quad \delta'_3(X, \delta_0^2) = \tilde{B}'\bar{X}' + (1 - \tilde{B}')X_n,$$

where

$$\tilde{B}' = \left(\frac{n-4}{f} \right) \frac{S_0^2}{S'^2}.$$

Like the proof of Theorem 5, we can get the following result by using the fact that $\bar{X}'$, S'^2 and S_0^2 are sufficient and complete for the family of distributions of $X_1, \dots, X_{n-1}$ and S_0^2 . The proof of (3.15) is given in the appendix.

THEOREM 8. *The empirical Bayes estimator δ'_3 in (3.14) is the best conditionally unbiased predictor and*

$$(3.15) \quad E_{\theta}(Y - \delta'_3(X, \delta_0^2))^2 = (1 - B)\sigma^2 + \frac{2n^2 + 3nf - 6f - 8n}{f(n-1)(n-6)}B\sigma^2.$$

From (3.13) and (3.15) it turns out that the risk of δ'_3 is greater than that of δ_3 . However, δ_3 is not conditionally unbiased.

4. Remarks

The present paper has only dealt with the bivariate case. This specialization has been partly for the sake of clarity and convenience, but mostly because the concept of the conditional unbiasedness works well. Although it may be possible to consider the following multivariate extension, some difficulties may arise.

Suppose that $(\lambda_1, X_1), \dots, (\lambda_n, X_n)$ be independent random vectors where λ_i and X_i are q - and p - dimensional. Assume that the conditional distribution of X_i given λ_i is normal with mean vector $A_i\lambda_i$ and covariance matrix Σ and the prior distribution of λ_i is also normal with mean vector μ and covariance matrix Ψ , where A_i is a $p \times q$ known matrix.

A straightforward calculation shows that the Bayes estimator of λ_n is

$$(4.1) \quad \delta_G = \mu + \Psi A_n' (\Sigma + A_n \Psi A_n')^{-1} (X_n - A_n \mu).$$

Let us assume that $\theta = (\mu, \Psi)$ is unknown. The marginal distribution of X_i is normal with mean vector $A_i \mu$ and covariance matrix $\Sigma + A_i \Psi A_i'$. So it seems difficult to construct unbiased estimators of $\Psi A_n' (\Sigma + A_n \Psi A_n')^{-1}$ based on $X_1, \dots, X_{n-1}$ in order to get conditionally unbiased predictors unless $A_i = A_n, i=1, \dots, n-1$, which seems less interesting in the setting.

Even though $A_i = A_n, i=1, \dots, n-1$, difficulties may not disappear in some cases. Let A denote the common matrix of A_i and assume $\text{rank} A = q$ and $p > q$. Then we can rewrite (4.1) as

$$\delta_G = B A \mu + ((A' A)^{-1} A' - B) X_n,$$

where $B = (A' A)^{-1} A' \Sigma (\Sigma + A \Psi A')^{-1}$. Since the marginal distribution of X_i is normal with mean vector $A \mu$ and covariance matrix $\Sigma + A \Psi A'$, it is possible to get a conditionally unbiased predictor of λ_n . However, it seems to be difficult to get such a predictor through complete and sufficient statistics. So no optimality of a conditionally unbiased predictor can be addressed by the present method. A different method seems necessary to treat the case.

Acknowledgments

This work was done under the supervision of Professor Y. Takada, whose generous guidance and suggestions are gratefully appreciated. The author thanks the referee for helpful comments.

Appendix

PROOF of (3.10). It follows from (3.2) that

$$\begin{aligned} E_\theta (Y - \delta'_2(X))^2 &= (1-B)\sigma^2 + E_\theta (B\mu + (1-B)X_n - \delta'_2(X))^2 \\ &= (1-B)\sigma^2 + E_\theta (\hat{B}'(\bar{X}' - \mu) + (X_n - \mu)(B - \hat{B}'))^2 \\ &= (1-B)\sigma^2 + \left(\frac{\sigma^2 + \tau^2}{n-1} \right) E_\theta \hat{B}'^2 + (\sigma^2 + \tau^2) E_\theta (B - \hat{B}')^2 \end{aligned}$$

by the independence of $\bar{X}', \hat{B}'$ and X_n . Hence

$$\begin{aligned} E_\theta (Y - \delta'_2(X))^2 &= (1-B)\sigma^2 + \frac{B\sigma^2(n-4)}{(n-1)(n-6)} + \frac{2B\sigma^2}{n-6} \\ &= (1-B)\sigma^2 + \frac{3B\sigma^2(n-2)}{(n-6)(n-1)}, \end{aligned}$$

which completes the proof of (3.10)

PROOF of (3.15). Like the proof of (3.10), we have

$$\begin{aligned}
E_{\theta}(Y - \delta'_3(X, S_0^2))^2 &= (1-B)\sigma^2 + \left(\frac{\sigma^2 + \tau^2}{n-1}\right)E_{\theta}\tilde{B}^2 + (\sigma^2 + \tau^2)E_{\theta}(B - \tilde{B})^2 \\
&= (1-B)\sigma^2 + \frac{B\sigma^2(n-4)}{(n-1)(n-6)}\left(\frac{f+2}{f}\right) + \frac{2B\sigma^2(f+n-4)}{(n-6)f} \\
&= (1-B)\sigma^2 + \frac{2n^2 + 3nf - 6f - 8n}{f(n-1)(n-6)}B\sigma^2.
\end{aligned}$$

So the proof is completed.

REFERENCES

- [1] Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. 2nd. ed., Springer-Verlag.
- [2] Efron, B. and Morris, C. (1972). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part II: The Empirical Bayes Case. *J. Amer. Statist. Assoc.*, **67**, 130-139.
- [3] Ishii, G. (1969). Optimality of unbiased predictor. *Ann. Inst. Statist. Math.*, **21**, 471-488.
- [4] Ishii, G. (1978). Statistical prediction (in Japanese), in Basic Sugaku, No. 8, Gendai Sugakusha, 69-73.
- [5] Lehmann, E. L. (1983). Theory of Point Estimation. Wiley.
- [6] Robbins, H. (1964). The empirical Bayes approach to statistical decision problem. *Ann. Math. Statist.*, **35**, 1-20.
- [7] Takada, Y. (1982). A comment on best invariant predictors. *Ann. Statist.*, **10**, 971-978.
- [8] Takeuchi, K. (1975). Statistical Prediction Theory (in Japanese). Baihukan.