Harvard University
Harvard University Biostatistics Working Paper Series

Year 2004 *Paper* 13

The Optimal Confidence Region for a Random Parameter

Hajime Uno* Lu Tian†
L.J. Wei‡

*Harvard University, huno@hsph.harvard.edu
†Northwestern University, lutian@northwestern.edu
‡Harvard University, wei@sdac.harvard.edu

The Optimal Confidence Region for a Random Parameter

Hajime Uno, Lu Tian, and L.J. Wei

Abstract

Under a two-level hierarchical model, suppose that the distribution of the random parameter is known or can be estimated well. Data are generated via a fixed, but unobservable realization of this parameter. In this paper, we derive the smallest confidence region of the random parameter under a joint Bayesian/frequentist paradigm. On average this optimal region can be much smaller than the corresponding Bayesian highest posterior density region. The new estimation procedure is appealing when one deals with data generated under a highly parallel structure, for example, data from a trial with a large number of clinical centers involved or genome-wide gene-expession data for estimating individual gene- or center-specific parameters simultaneously. The new proposal is illustrated with a typical microarray data set and its performance is examined via a small simulation study.

THE OPTIMAL CONFIDENCE REGION FOR A RANDOM PARAMETER

Hajime Uno
Department of Biostatistics, Harvard University, 677 Huntington Ave., Boston, MA 02115
huno@hsph.harvard.edu

Lu Tian
Department of Preventive Medicine, Northwestern University
680 N. Lake shore Drive, Chicago, IL
lutian@northwestern.edu

L.J. Wei
Department of Biostatistics, Harvard University, 677 Huntington Ave., Boston, MA 02115
wei@sdac.harvard.edu

Summary

Under a two-level hierarchical model, suppose that the distribution of the random parameter is known or can be estimated well. Data are generated via a fixed, but unobservable realization of this parameter. In this paper, we derive the smallest confidence region of the random parameter under a joint Bayesian/frequentist paradigm. On average this optimal region can be much smaller than the corresponding Bayesian highest posterior density region. The new estimation procedure is appealing when one deals with data generated under a highly parallel structure, for example, data from a trial with a large number of clinical centers involved or genome-wide gene-expression data for estimating individual gene- or center-specific parameters simultaneously. The new proposal is illustrated with a typical microarray data set and its performance is examined via a small simulation study.

Some Key words: Empirical Bayes; Gene-expression; Global clinical trials; Hierarchical model; Highest posterior density region.

1. INTRODUCTION

Let $\Pi' = (\Theta, \Gamma')$ be the row vector of *random* parameters with a known or well-estimated distribution function $G(\cdot)$. Here, Θ is the scalar parameter of interest and Γ is the vector of nuisance parameters. Conditional on $\Pi' = \pi' = (\theta, \gamma')$, let X be the observable random quantity with distribution function $F_\pi(x)$. Suppose that we are interested in making inferences about the unobservable θ based on the observed $X = x$.

Under the Bayesian paradigm, a $(1 - \alpha_0)$ credible region $B(x)$ is a set of θs such that

$$\text{pr}(\Theta \in B(X)|X = x) = 1 - \alpha_0, \tag{1.1}$$

where $0 < \alpha_0 < 1$, and the probability is generated by Θ conditional on $X = x$. Among these regions $B(x)$, the highest posterior density set $B_o(x)$ is the smallest one. Moreover, if the posterior density function $g_x(\theta)$ is continuous and non-uniform over every region in the space of Θ, there exists a constant c_x such that

$$B_o(x) = \{\theta : g_x(\theta) \geq c_x\}, \tag{1.2}$$

where c_x is determined via (1.1) (Box & Tiao, 1972, p.123). Note that c_x may vary substantially over the sample space of X. Under a frequentist paradigm, a $(1 - \alpha_0)$ confidence region $H(x)$ is a set of θs such that

$$\text{pr}(\Theta \in H(X)|\Theta = \theta) = 1 - \alpha_0 \tag{1.3}$$

without involving the "prior" distribution $G(\cdot)$.

For the present case, an inference procedure is expected to be utilized repeatedly for estimating different θs with different sets of data x. Hence, in evaluating the long-run performance of an

interval estimation procedure, one should be averaging over both the data X and the parameter Θ. Enlightened discussions of this joint frequentist/Bayesian principle can be found, for example, in Neyman (1977) and Bayarri & Berger (2004, Section 2.2). Under this paradigm, a $(1-\alpha_0)$ confidence interval $R(x)$ is a set of θs such that

$$\text{pr}(\Theta \in R(X)) = 1 - \alpha_0, \tag{1.4}$$

where the probability is generated by Θ and X jointly. Note that the Bayesian credible region $B(X)$ and the frequentist interval $H(X)$ are $R(X)$, but they have to satisfy rather stringent condition (1.1) for each observed x or (1.3) for each θ.

When the "prior" distribution $G(\cdot)$ is unknown, but can be estimated well without much error, $R(X)$ is approximately an empirical Bayes confidence region advocated by Morris (1983). Insightful discussions of such an estimation procedure can be found, for example, in the Comments on Morris (1983a) by Berger, Dempster, Hickely and Leonard. In their recent book, Carlin & Louis (2000) provided an excellent review on modern empirical Bayes inferences.

The class of intervals $R(X)$ is much larger than that of $B(X)$ or $H(X)$. An interesting question is how to identify the optimal region $R_o(X)$ in the sense that its expected size with respect to the measure generated by Θ and X is the smallest among the confidence sets $R(X)$. At first glance, this optimization problem seems prohibitively complex. In Section 2, we show that this optimal region is surprisingly easy to obtain. In fact, $R_o(x)$ is simply a set of θs such that

$$g_x(\theta) \geq c, \tag{1.5}$$

where c is a constant which satisfies (1.4). Therefore, if c_x in (1.2) is not constant, the corresponding Bayesian region $B_o(x)$ is not optimal. In Section 3, we show via a real example that

$R_o(x)$ can be quite different from $B_o(x)$, especially when the posterior density function is relatively flat. Moreover, we demonstrate via a simulation study that on average $R_o(X)$ can be quite smaller than $B_o(X)$.

The optimal interval $R_o(X)$ is attractive especially in the analysis of data generated from an experiment with a highly parallel structure, for example, data from a global trial with a large number of clinical centers involved or data generated from a typical genome-wide gene-expression experiment (Newton et al., 2004). One of the main goals for this type of studies is to make inferences simultaneously about individual center- or gene-level parameters (Efron, 1996, 2003). For the present case, x is the observed data from a specific center or gene, and one expects that approximately $100(1-\alpha_0)$ of all realized center- or gene-specific confidence regions cover their corresponding true θs. The long-run coverage probability (1.4) of $R_o(X)$ does not have to be interpreted with imaginary repetitions.

The burden of identifying the optimal $R_o(X)$ is to find the constant c in (1.5). In practice, this can be done via the standard Monte Carlo simulation method, which is illustrated in Section 3.

2. DERIVATION OF THE OPTIMAL CONFIDENCE REGION

Let $h(\cdot)$ be a non-negative, bounded function defined on the sample space of X. For any given function $h(\cdot)$, define the confidence region $J_X(h(X))$ for Θ, where

$$J_x(h(x)) = \{\theta : g_x(\theta) \geq h(x)\}. \tag{2.1}$$

Let

$$1-\alpha_x(h(x)) = \text{pr}(\Theta \in J_X(h(X))|X = x),$$

the coverage probability for θ given $X = x$. To identify the optimal $R_o(X)$ among all confidence regions $R(X)$ defined in (1.4), it is sufficient to consider the class $\mathcal{H}$ of sets $J_X(h(X))$, indexed by the function $h(\cdot)$ which satisfies the constraint

$$E\{\alpha_X(h(X))\} = \alpha_0, \tag{2.2}$$

where the expectation E is taken with respect to X. Note that if for any possible realization x, $\alpha_x(h(x)) = \alpha_0$, $J_X(h(X))$ is $B_o(X)$. Let $L_X(h(X))$ be the size of the region $J_X(h(X))$ and let its expected value be denoted by $S(h)$. Our task is to locate a $J_X(h(X))$ in $\mathcal{H}$, which minimizes $S(h)$.

First, consider the region $J_X(h(X))$ with $h(x)$ being constant d over x. The corresponding $\alpha_x(d)$ is a non-decreasing function of d, therefore, there exists c whose $\alpha_X(c)$ satisfies (2.2). We will show that the confidence region with $h(x) = c$ is optimal. To this end, for a non-constant function $h(\cdot)$, that is $\mathrm{pr}(h(X) = \text{constant}) \neq 1$, let $D_1 = \{x : h(x) < c\}$ and $D_2 = \{x : h(x) > c\}$. Then,

$$\mathrm{E}(I_2(X)\{\alpha_X(h(X)) - \alpha_X(c)\}) = \mathrm{E}(I_1(X)\{\alpha_X(c) - \alpha_X(h(X))\}) > 0, \tag{2.3}$$

where $I_k(X) = I(X \in D_k), k = 1, 2$, and $I(\cdot)$ is the indicator function. When $x \in D_1$, it is straightforward to show that for any realized x such that $\alpha_x(h(x)) < 1$,

$$\alpha_x(c) - \alpha_x(h(x)) < c\{L_x(h(x)) - L_x(c)\}. \tag{2.4}$$

Furthermore, $L_x(d)$ is monotone in d. This, coupled with (2.3) and (2.4), implies that

$$\frac{\mathrm{E}(I_1(X)\{\alpha_X(c) - \alpha_X(h(X))\})}{\mathrm{E}(I_1(X)\{L_X(h(X)) - L_X(c)\})} < c. \tag{2.5}$$

Similarly,

$$c < \frac{\mathrm{E}(I_2(X)\{\alpha_X(h(X)) - \alpha_X(c)\})}{\mathrm{E}(I_2(X)\{L_X(c) - L_X(h(X))\})}. \tag{2.6}$$

It follows from (2.3), (2.5) and (2.6) that

$$\frac{\mathrm{E}(I_2(X)\{L_X(c) - L_X(h(X))\})}{\mathrm{E}(I_1(X)\{L_X(h(X)) - L_X(c)\})} < \frac{\mathrm{E}(I_2(X)\{\alpha_X(h(X)) - \alpha_X(c)\})}{\mathrm{E}(I_1(X)\{\alpha_X(c) - \alpha_X(h(X))\})} = 1.$$

Therefore, for a non-constant $h(\cdot)$, the expected size $S(h)$ of the corresponding region $J_X(h(X))$ is strictly greater than that of $R_o(X)$.

3. A MICROARRAY EXAMPLE AND SIMULATION STUDY

We use a typical genome-wise microarray study to illustrate the new proposal. Suppose that under a specific cellular state, for each subject and each gene, the normalized expression level, which measures the abundance of the gene-specific RNA, is observed. Let n be the number of subjects and K be the number of genes in the study. Generally K is fairly large. For a given gene, suppose that the unobserved mean expression value is θ, a realization from the random parameter Θ, and the observed quantity x is the collection of n expression values. Here, $\Pi = \Theta$, that is, there are no nuisance random parameters. Assume that the prior distribution $G(\theta)$ of Θ is an inverse-Gamma, that is, the density function of Θ is proportional to

$$\theta^{-(a_2+1)} e^{-a_1/\theta}, \tag{3.1}$$

where a_1 and a_2 are known constants. Furthermore, assume that the distribution $F_\theta(x)$ of X given θ is a Gamma with density proportional to

$$x^{a_3-1} e^{-a_3 x/\theta}, \tag{3.2}$$

where a_3 is a known constant. It follows that the posterior density function of Θ is an inverse-Gamma with density proportional to

$$\theta^{-(a_2+na_3+1)}e^{-(a_1+na_3\bar{x})/\theta}, \tag{3.3}$$

where $\bar{x}$ is the observed sample mean of data x.

Note that the mean expression level is θ. In practice, a_1, a_2, and a_3 are unknown. Since the number of genes involved is quite large, one can use the maximum marginal likelihood estimates to approximate these fixed parameters. Similar hierarchical models have been proposed, for example, by Newton et al.(2001) and Kendziorski et al. (2003), for detecting differential gene expression between two cellular states. The data set we use for illustration is called "eset" from the computer package "Biobase" in Bioconductor *(www.bioconductor.org)*. This data set was generated using Affymetrix U95v2 chips at the Dana Farber Cancer Institute.

The data were normalized via the computer software dChip *(www.dchip.org)*. There are 500 genes involved in the study. For simplicity, we only consider the first n=13 subjects associated with Phenotype I. Furthermore, due to normalization, there are 111 genes whose expression values are negative. We deleted those observations in our analysis. This results in $K = 389$. Using the above hierarchical model, the marginal likelihood estimates for a_1, a_2 and a_3 are 61.01, 0.87 and 5.11, respectively. We assume that these estimates are the true values of those a's in (3.1) and (3.2).

To obtain the cutoff point c in (1.5), we approximate the probability of $\{g_X(\Theta) \geq \text{constant}\}$ using $100,000$ simulated pairs (θ, x), where $g_x(\theta)$ is proportional to (3.3). The average length of 389 optimal intervals $R_o(x)$ for θs is 135. On the other hand, the average length of 389 highest posterior density intervals $B_o(x)$ is 203. When the posterior density function for a specific gene

is flat, $R_o(x)$ can be quite different from $B_o(x)$. For instance, for gene "31444-s-at", the optimal interval is $(3930, 4354)$, which is much shorter than the corresponding highest posterior density region $(3281, 5325)$. On the other hand, if the posterior density is relatively narrow, these two types of intervals are quite similar. For instance, for gene "AFFX-MurFAS-at", the optimal interval $R_o(x) = (11.2, 25.0)$, which is slightly larger than $B_o(x) = (12.8, 21.0)$.

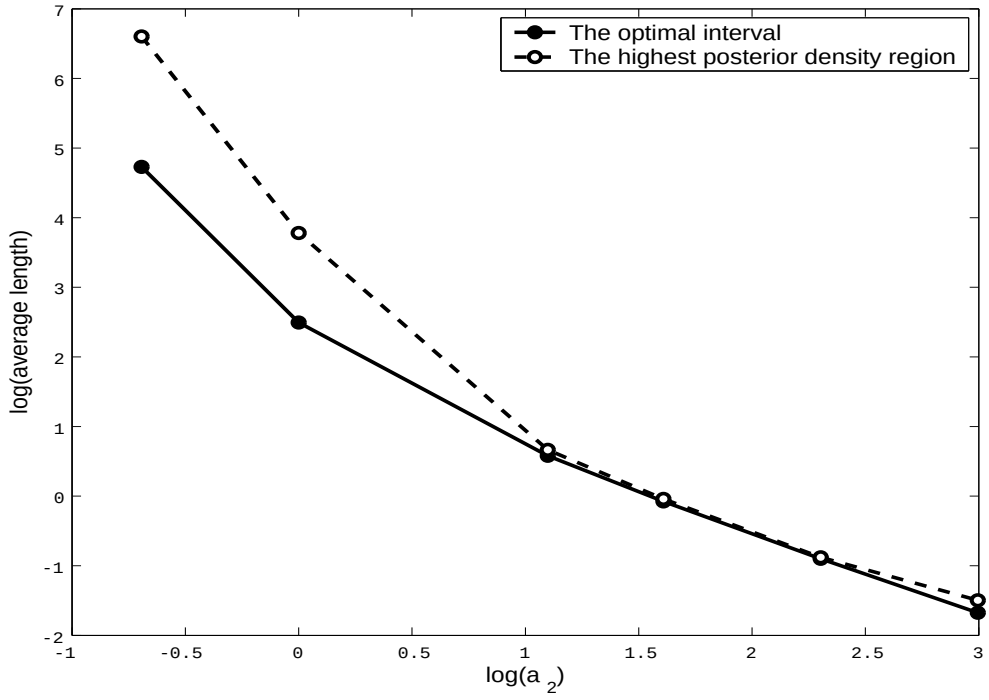
We also conducted a small simulation study to examine the performance of the optimal $R_o(X)$ under the above hierarchical model setting with $n = 20$ and with various sets of a_1, a_2 and a_3. For each simulation, we fixed the above parameters and generated 10,000 iterations. For each iteration, first we obtained a θ from the inverse-gamma distribution (3.1). We then generated 20 independent expression values x via (3.2). We used these 10,000 iterations to approximate the cutoff point c in (1.5) and also construct 10,000 95% interval estimates $R_0(x)$ and $B_o(x)$ for the corresponding θs. In general, we find that for a small a_2 in the prior distribution (3.1), the optimal region $R_o(X)$ significantly outperforms $B_o(X)$. On the other hand, when a_2 is relatively large, these two interval estimates are quite similar with respect to their average lengths. In Figure 1, we present a comparison between $R_o(X)$ and $B_o(X)$ with various values of a_2, but with $a_1 = 10$ and $a_3 = 5$. The improvement from $R_o(X)$ over $B_o(X)$ with respect to their lengths can be quite substantial. The empirical coverage probabilities of all interval estimates studied here are almost identical to their nominal levels.

4. REMARKS

Under the empirical Bayes paradigm, the hyper-parameters for the prior distribution $G(\cdot)$ have to be estimated. A naive empirical Bayes confidence interval for θ obtained via the Bayesian credible interval by replacing the hyper-parameters with the maximum marginal likelihood esti-

mates may not have the correct coverage probability. Various innovative calibrations of such an interval have been proposed, for example, by Morris (1983a,b), Laird & Louis (1987), Carlin & Gelfand (1990, 1991) and Datta et al. (2002). Similar calibrations may also be applicable to the optimal interval $R_o(X)$ proposed in this article.

Fig 1. Log(average length) of 0.95 confidence interval



REFERENCES

Bayarri, M. J. & Berger, J. O. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science* **19**, 58–80.

Box, G. & Tiao, G. (1972). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Carlin, B. & Gelfand, A. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association* **85**, 105–114.

Carlin, B. & Gelfand, A. (1991). A smaple reuse method for accurate parametric empirical Bayes confidence intervals. *Journal of the Royal Statisitcal Society, Ser. B* **53**, 189-200.

Carlin, B. & Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed.* London: Chapman & Hall.

Datta, G., Ghosh, M., Smith, D., & Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities on empirical Bayes confidence intervals. *Scandinavian Journal of Statistics* **29**, 139–152.

Efron, B. (1996). Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association* **91**, 538–550.

Efron, B. (2003). Robbins, empirical Bayes and microarrayes. *The Annals of Statistics* **31**, 366–378.

Kendziorski, C. M., Newton, M. A., Lan, H., & Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

Laird, N. & Louis, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* **82**, 739–750.

Morris, C. (1983a). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association* **78**, 47–55.

Morris, C. (1983b). Parametric empirical Bayes confidence intervals. In *Scientific inference, data analysis, and robustness,* 25–50. New York: Academic Press.

Newton, M., Kendziorski, C., Richmond, C. & Blattner, F. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Newton, M., Noueiry, A., Sparkar, D. & Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric heirarchical mixture method. *Biostatistics* **5**, 155–176.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthése* **36**, 97–131.