The Korean Communications
in Statistics Vol. 6, No. 1, 1999
pp. 193-205

A Bayesian Test Criterion for the Behrens-Fisher Problem

Hea-Jung Kim[1)]

Abstract

An approximate Bayes criterion for Behrens-Fisher problem (testing equality of means of two normal populations with unequal variances) is proposed and examined. Development of the criterion involves derivation of approximate Bayes factor using the imaginary training sample approach introduced by Spiegelhalter and Smith (1982). The proposed criterion is designed to develop a Bayesian test criterion having a closed form, so that it provides an alternative test to those based upon asymptotic sampling theory (such as Welch's t test). For the suggested Bayes criterion, numerical study gives comparisons with a couple of asymptotic classical test criteria.

1. Introduction

Suppose two independent samples $X_1(1),\ldots,X_{n_1}(1)$ and $X_1(2),\ldots,X_{n_2}(2)$ are obtained from two univariate normal populations $N(\mu_1,\sigma_1^2)$ and $N(\mu_2,\sigma_2^2)$, respectively, where the population variances are unknown and not the same. For the two samples case, many statisticians have drawn attention to the problem of testing the equality of the two means, and this problem is called the Behrens-Fisher problem. See for example Rencher(1995), Johnson and Wichern(1992), Hogg and Tanis (1993), Scheffé(1943) for testing the problem with sampling theory approach. Hattmansperger(1973) and Ghosh(1975) respectively suggested a nonparametric test and a Stein-type two-stage procedure for the problem. For Bayesian approach, Johnson and Weerahandi(1988) gave a Bayesian solution to the Behrens-Fisher problem. They provided exact and approximate methods for calculating posterior probabilities for the difference between two normal means. Then they constructed $1-\alpha$ Bayesian credible interval to test $H_0:\mu_1=\mu_2$, $\sigma_1^2\neq\sigma_2^2$(using Lindley's method). However, the Bayesian method can be criticized in two reasons. First, it needs massive computations for a numerical integration involved in calculating the posterior probability. Second, as pointed by Lee(1988), Lindley's method is not the best way of summarizing posterior beliefs of H_0(comparing to formal Bayes test criterion, i.e. Bayes factor).

The aim of this paper is to develop an alternative Bayes criterion for testing the

1) Professor, Department of Statistics, Dongguk University, Seoul, 100-715, Korea.

Behrens-Fisher problem which eliminates the problems of the Bayes test mentioned above. The development involves calculation of posterior probability of each possible model($M_1:\mu_1=\mu_2$, $\sigma_1^2\neq\sigma_2^2$ and $M_2:\mu_1\neq\mu_2$, $\sigma_1^2\neq\sigma_2^2$) via a default Bayes factor, and suggests a test criterion in terms of the posterior probabilities so that the Behrens-Fisher problem can be resolved in a simple and formally justifiable way. In Section 2, we review the default Bayes factor obtained from applying the device of imaginary training sample method introduced by Smith and Spiegelhalter(1980) and Spiegelhalter and Smith(1982). In Section 3 we suggest a Bayesian criterion for comparing(or testing) the hypothesized models in the Behrens-Fisher problem. Section 4 examines the performance of the suggested criterion and notes some comparisons with traditional tests. Finally, Section 5 includes some concluding remarks.

2. Default Bayes Factor

Suppose we have data D, assumed to have arisen under one of two alternative models M_1 and M_2 having probability densities $p(D\mid\theta_i, M_i)$, where parameter vectors are unknown. Given an improper prior $\pi(\theta_i\mid M_i)$ for the parameter of each model, together with prior p_i of each model being true, the data produce the posterior probability of M_i being true as

$$p(M_i\mid D)=\frac{p(D\mid M_i)p_i}{\sum_{j=1}^{2}p(D\mid M_j)p_j},\quad i=1,2, \tag{2.1}$$

where $p_1+p_2=1$ and the density $p(D\mid M_i)$ is called the marginal likelihood or predictive density of D and is defined as

$$p(D\mid M_i)=C_i\int p(D\mid\theta_i, M_i)\pi(\theta_i\mid M_i)d\theta_i, \tag{2.2}$$

where C_i denotes the normalizing constant so that $C_i\,\pi(\theta_i\mid M_i)$ may be proper over the parameter space Θ_i, $\theta_i\in\Theta_i$.

The Bayes factor(cf. Jeffreys, 1961) for M_1 against M_2 is defined by

$$B_{12}=\frac{p(D\mid M_1)}{p(D\mid M_2)}=\frac{C_1\int p(D\mid\theta_1, M_1)\pi(\theta_1\mid M_1)d\theta_1}{C_2\int p(D\mid\theta_2, M_2)\pi(\theta_2\mid M_2)d\theta_2}. \tag{2.3}$$

As seen above, the Bayes factor denotes the ratio of the posterior odds of M_1 to its prior odds, regardless of the value of the prior odds. Thus B_{12} can be viewed as the weighted likelihood ratio of M_1 to M_2 and hence can be a criterion for measuring comparative support of the data for the two models(cf. Kass and Raftery 1995). The posterior probability (2.1) that M_1 is true is then expressed in terms of B_{12}.

In the model comparison, most Bayesians today prefer to use noninformative priors that are typically improper. Such priors are defined only up to a constant multiple, and hence the Bayes factor is itself a multiple of this arbitrary constant(cf. Kass and Raftery 1995). As is the case, computing B_{12} in equation (2.3) requires specification of a ratio of undefined constants, C_1/C_2. Thus, it is not possible to utilize B_{12} for the comparison between M_1 and M_2. One effort to resolve this problem(specification of C_1/C_2) is the "training sample method" that puts aside a training sample which is combined with the improper prior to produce a proper prior distribution. This idea was introduced by Lampers(1971), and other implementations have been suggested under the name of partial Bayes factors(O'Hagan 1991), intrinsic Bayes factors(Berger and Perrichi 1996), and fractional Bayes factors(O'Hagan 1995). However, the training sample approach is clearly impractical if complete data set itself is rather small, or if the data derive from a highly structured situation, such as complete randomized block experiment. In such a case, setting a training sample aside would destroy the overall structure(say, symmetry and orthogonality) of the remaining observations.

Another solution to resolve the specification problem of C_1/C_2, which does not involves the above problems attached to the training sample approach, is the "imaginary training sample method" of Spiegelhalter and Smith(1982) and Smith and Spiegelhalter (1980) which uses imaginary observations to determine ratio of undefined constants(C_1/C_2) in the Bayes factor. In this paper, we will invoke the imaginary training sample method in order to specify C_1/C_2 in the comparison of M_1 and M_2.

3. Test Criterion

Suppose we have two independent univariate normal populations Π_1 and Π_2 each specified by a model M_i, $i=1,2$, where M_i defines the distribution of each population distribution $\Pi_k \sim N(\mu_k, \sigma_k^2)$, $k=1,2$, where parameters are unknown. Let our interest of model comparison be homogeneity/heterogeneity of the means between two populations. Then the model specification becomes

$$M_1 : \mu_1 = \mu_2 = \mu \ , \ \ \sigma_1^2 \neq \sigma_2^2 \qquad \text{versus} \qquad M_2 : \mu_1 \neq \mu_2 \ , \ \ \sigma_1^2 \neq \sigma_2^2. \tag{3.1}$$

Let $X_1(k), X_2(k), \dots, X_{n_k}(k)$ be independent sample of size n_k from Π_k with distribution $N(\mu_k, \sigma_k^2)$, $k=1,2$, and let denote the two independent samples from Π_1 and Π_2 as D. If we define $\overline{X}(k)=\sum_{j=1}^{n_k} X_j(k)/n_k$ and $V_k=\sum_{j=1}^{n_k}(X_j(k)-\overline{X}(k))^2$. Then the data D is to have arisen under M_1 and M_2 according to respective probability densities given by

$$p(D \mid \mu, \sigma_1^2, \sigma_2^2, M_1)=\prod_{k=1}^{2}(2\pi)^{-\frac{n_k}{2}}\sigma_k^{-n_k}\exp\left\{-\frac{1}{2\sigma_k^2}\Omega_k\right\}, \tag{3.2}$$

$$p(D \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, M_2)=\prod_{k=1}^{2}(2\pi)^{-\frac{n_k}{2}}\sigma_k^{-n_k}\exp\left\{-\frac{1}{2\sigma_k^2}\Omega_k^*\right\}, \tag{3.3}$$

where $\Omega_k=V_k+n_k(\mu-\overline{X}(k))^2$ and $\Omega_k^*=V_k+n_k(\mu_k-\overline{X}(k))^2$.

Since our interest focuses primary on a statement concerning to relative probability that D comes from one or the other of the model, and not about making probability statement about where a parameter lies, we shall use a particular convenient prior density to reflect a noninformative information about the unknown parameters. In this paper, we shall be concerned with the case here both priors $\pi(\mu, \sigma_1^2, \sigma_2^2 \mid M_1)$ and $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid M_2)$ have improper limiting form of the univariate normal-inverse chi-square conjugate priors(cf. DeGroot 1970 Smith and Spiegelhaiter 1980) that can be written by

$$\pi(\mu, \sigma_1, \sigma_2 \mid M_1) \;=\; C_1\prod_{k=1}^{2}(2\pi\sigma_k^2)^{-1/4}\sigma_k^{-2}, \tag{3.4}$$

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid M_2) \;=\; C_2\prod_{k=1}^{2}(2\pi\sigma_k^2)^{-1/2}\sigma_k^{-2}, \tag{3.5}$$

where C_i, $i=1,2$, are undefined constants.

Lemma 3.1. Under the improper priors (3.4) and (3.5), respective marginal likelihoods conditional on M_1 and M_2 are

$$p(D \mid M_1)=C_1\exp\left\{-\frac{Q}{2}\right\}\prod_{k=1}^{2}\{\Delta_k n_k(n_k-1)^{-1/4}(2\pi)^{-n_k/2}V_k^{-n_k/2}\}\left[1+O_p\left(\frac{1}{n_k}\right)\right],$$

$$p(D \mid M_2)=C_2\prod_{k=1}^{2}\{\Delta_k^* n_k^{-1/2}(2\pi)^{-n_k/2}V_k^{-n_k/2}\},$$

where $\Delta_k=2^{(n_k+1/2)/2}\Gamma\left(\frac{n_k+1/2}{2}\right)$, $\Delta_k^*=2^{n_k/2}\Gamma\left(\frac{n_k}{2}\right)$,

$Q=(\overline{X}(1)-\overline{X}(2))^2/(s_1^2/n_1+s_2^2/n_2)$, $s_k^2=V_k/(n_k-1)$.

Proof. See appendix for the derivation of $p(D \mid M_1)$, $p(D \mid M_2)$ is obtained from the following procedure. According to the definition of the marginal likelihood (2.2), (3.3) and (3.5) yield

$$p(D \mid M_2) = C_2 \int (2\pi)^{-(n_1+n_2+2)/2} \prod_{k=1}^{2} \sigma_k^{-(n_k+3)} \times \exp\left\{-\frac{1}{2}\sum_{k=1}^{2}(V_k + n_k(\overline{X}(k)-\mu_k)^2/\sigma_k^2\right\} \prod_{k=1}^{2} d\mu_k \prod_{k=1}^{2} d\sigma_k^2. \quad (3.6)$$

Integrate the integrand in (3.6) with respect to μ_k's using univariate normal normalizing constant. This gives

$$C_2 \int (2\pi)^{-(n_1+n_2)/2} \prod_{k=1}^{2}\{n_k^{-1/2}\sigma_k^{-(n_k+2)}\} \exp\left\{-\frac{1}{2}\sum_{k=1}^{2} V_k/\sigma_k^2\right\} \prod_{k=1}^{2} d\sigma_k^2. \quad (3.7)$$

Then the desired marginal likelihood $p(D \mid M_2)$ is found by integrating with respect σ_k^2's, using the inverse chi-squared normalizing constants (cf. Lee 1988).

The marginal likelihoods, $p(D \mid M_1)$ and $p(D \mid M_2)$, are clearly indeterminate owing to the presence of the undefined constants, C_1 and C_2. Therefore, as $n \to \infty$, Bayes factor for comparing M_1 with M_2 is itself a multiple of these arbitrary constants such that

$$B_{12} = \frac{P(D \mid M_1)}{P(D \mid M_2)} = \frac{C_1}{C_2} 2^{1/2} \exp\left\{-\frac{Q}{2}\right\} \prod_{k=1}^{2}\left(\frac{n_k}{n_k - 1}\right)^{1/4} \prod_{k=1}^{2} \frac{\Gamma\left\{\frac{n_k + 1/2}{2}\right\}}{\Gamma\left\{\frac{n_k}{2}\right\}}, \quad (3.8)$$

where $Q = (\overline{X}(1) - \overline{X}(2))^2 / \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)$, $s_k^2 = V_k/\nu_k$, $\nu_k = (n_k - 1)$.

A solution for this problem is to construct a default Bayes factor by means of imaginary training sample. This method, so called the imaginary training sample method, for assigning the constants C_1/C_2 has been proposed by Spiegelhalter and Smith(1982). The basic idea, a variation on a theme of Good(1947), is to image that a imaginary training sample data set is available which is defined as follows.

Definition 3.1 (Spiegelhalter and Smith 1982). A data set is called the imaginary training data set if it is available, which

(i) involves the smallest possible sample size permitting a comparison of M_1 and M_2 ;

(ii) provides maximum possible support for M_1, so that it may yield the Bayes factor $B_{12} \approx 1$.

Using the definition, we can obtain the imaginary training sample to eliminate the indeterminacy of the Bayes factor in (3.8). Lemma 3.1 and the condition (i) of Definition 3.1 require that a minimal size of the imaginary training sample is 2 for each population $\Pi_k \sim N(\mu_k, \sigma_k^2)$ defined by M_2, $k = 1, 2$, because we need at least 2 observations in order to be able to estimate σ_k^2 and μ_k. Since $M_1 \subset M_2$, we see that the imaginary training sample of size 2 for each population is minimal sample size for the comparison of M_1 and M_2. Under the imaginary training sample of total sample size 2×2, the approximate Bayes factor (3.8) will be evaluated as

$$\frac{P(D \mid M_1)}{P(D \mid M_2)} = 2\frac{C_1}{C_2}\exp\left\{-\frac{Q^*}{2}\right\}\Gamma\{5/4\}^2, \tag{3.9}$$

where $Q^* = (\overline{X}^*(1) - \overline{X}^*(2))^2 / \left(\frac{s_{1*}^2}{2} + \frac{s_{2*}^2}{2}\right)$ that is the square of Fisher's Z test statistic (value $Q^*=0$ giving maximum support to M_1), and $\overline{X}^*(k)$ and s_{k*}^2 are respective unbiased estimates of μ_k and σ_k^2 obtained from the imaginary training sample. Suppose we denote Λ by

$$\exp\left\{-\frac{Q^*}{2}\right\},$$

then Λ is a function of Z test statistic for testing the null hypothesis that M_1 is true(cf. Hogg and Tanis 1993). The Bayes factor (3.9) can be expressed in terms of Λ, so that

$$\frac{P(D \mid M_1)}{P(D \mid M_2)} = 2\frac{C_1}{C_2}\Gamma\{5/4\}^2\Lambda. \tag{3.10}$$

Since the statistic Λ takes a value $0 \le \Lambda \le 1$, the condition (ii) of Definition 3.1 leads to the value of Λ to be one (i.e. $Q^* \to \infty$) achieving maximum support to M_1. Furthermore, if we set the Bayes factor (3.9) (arising from this imaginary experiment) equal to one, we can immediately deduce, from Definition 3.1, that

$$\frac{C_1}{C_2} = 2^{-1}\Gamma\{5/4\}^{-2}. \tag{3.11}$$

Theorem 3.1. If n_1 and n_2 are large, the approximate Bayes factor B_{12}^I for comparing M_1 with M_2, obtained from the imaginary training sample method under the priors (3.4) and (3.5), is reduces to a function of Fisher's Z-test statistic (cf. Hogg and Tanis 1993)

$$B_{12}^I = 2^{-1/2}\Gamma\{5/4\}^{-2}\exp\left\{-\frac{Z^2}{2}\right\}\prod_{k=1}^{2}\frac{\Gamma\left\{\frac{n_k+1/2}{2}\right\}}{\Gamma\left\{\frac{n_k}{2}\right\}}\left(\frac{n_k}{n_k-1}\right)^{1/4}, \tag{3.12}$$

where $Z=\frac{\overline{X}(1)-\overline{X}(2)}{(s_1^2/n_1+s_2^2/n_2)^{1/2}}$, s_1^2 and s_2^2 are the usual unbiased estimators of σ_1^2 and σ_2^2 (the two population variances).

Proof. Substituting C_1/C_2 of (3.11) into (3.8), we have the result.

This yields the following Bayes criterion for testing M_1 versus M_2.

Corollary 3.1. The posterior probability of M_1 is given by

$$P(M_1 \mid D)=\frac{\pi_1 B_{12}^I}{\pi_2+\pi_1 B_{12}^I}, \tag{3.13}$$

where π_i denotes the prior probability of M_i, $i=1,2$.

Proof. Under the prior probabilities, Bayes theorem gives

$$P(M_1 \mid D) = P(D \mid M_1)\pi_1 \ / \ \left(\sum_{i=1}^{2} P(D \mid M_i)\pi_i\right).$$

Since $B_{12}^I=P(D \mid M_1)/P(D \mid M_2)$, expressing $P(M_1 \mid D)$ in terms of B_{12}^I, we have the result.

If $P(M_1 \mid D)$, is larger than $1/2$, then we choose M_1 as a model best supported by the data, D. Otherwise, we choose M_2, and in case $P(M_1 \mid D)=1/2$, we may randomly choose one of the two model. Note that when $\pi_1=\pi_2=1/2$, $P(M_1 \mid D) > 1/2$ is equivalent to $B_{12}^I > 1$.

4. Numerical Study

The Bayes factor is a summary of the evidence provided by the data in favor of M_1 as opposed to M_2. Jeffreys(1961) has proposed, for nested models, $M_1 \subset M_2$, the following order of magnitude interpretations of B_{12}(see Kass and Raftery(1995) for the other order of magnitude).

Range	Evidence
$B_{12} > 1$	evidence supports M_1
$1 > B_{12} > 10^{-1/2}$	very slight evidence against M_1
$10^{-1/2} > B_{12} > 10^{-1}$	moderate evidence against M_1
$10^{-1} > B_{12} > 10^{-2}$	strong to very strong evidence against M_1
$10^{-2} > B_{12}$	decisive evidence against M_1

We shall summarize the p-values of classical two-tailed tests for H_0 that M_1 is true corresponding to critical values of the Bayes factor (B_{12}^I) in (3.12) for some particular combinations of n_1, n_2 and $c = s_1^2/(s_1^2+s_2^2)$. The standard tests considered here are those using Fisher's Z test statistic

$$Z = (\overline{X}(1) - \overline{X}(2)) \,/\, (s_1^2/n_1 + s_2^2/n_2)^{1/2} \;\sim\; N(0,1),$$

and Welch's t test statistic(cf. Hogg and Tanis 1993)

$$Z = (\overline{X}(1) - \overline{X}(2)) \,/\, (s_1^2/n_1 + s_2^2/n_2)^{1/2} \;\sim\; t_{([r])},$$

where $[r]$ is the greatest integer part of

$$r = \{c^2/(n_1-1) + (1-c)^2/(n_2-1)\}^{-1} \text{ and } c = s_1^2 \,/\, (s_1^2+s_2^2).$$

TABLE 1. p-values of Fisher's Z Test(F) and Welch's t Test(W) with Given c.

		B_{12}^I						B_{12}^I			
n_1	Test	1	$10^{-1/2}$	0.1	0.01	n_1	Test	1	$10^{-1/2}$	0.1	0.01
					$(n_2 = n_1)$						
5	F	.408	.084	.021	.001	10	F	.247	.056	.015	.001
	$W(c=.1)$	.446	.146	.071	.026		$W(c=.1)$	.272	.083	.033	.008
	$W(c=.7)$	.435	.128	.055	.016		$W(c=.7)$	.264	.075	.027	.005
15	F	.188	.045	.012	.001	20	F	.155	.037	.010	.001
	$W(c=.1)$	.205	.060	.022	.004		$W(c=.1)$	.169	.049	.017	.003
	$W(c=.7)$	.200	.056	.019	.003		$W(c=.7)$	.165	.045	.015	.002
50	F	.087	.022	.006	.000	100	F	.057	.015	.004	.000
	$W(c=.1)$	.092	.026	.008	.001		$W(c=.1)$	.060	.016	.005	.000
	$W(c=.7)$	.090	.025	.007	.001		$W(c=.7)$	.059	.016	.004	.000

		B^I_{12}						B^I_{12}			
n_1	Test	1	$10^{-1/2}$	0.1	0.01	n_1	Test	1	$10^{-1/2}$	0.1	0.01
					$(n_2=2\times n_1)$						
5	F	.314	.068	.017	.001	10	F	.195	.046	.012	.001
	$W(c=.1)$	.336	.096	.037	.009		$W(c=.1)$	.208	.058	.020	.003
	$W(c=.7)$	.345	.108	.047	.013		$W(c=.7)$	.212	.062	.023	.004
15	F	.150	.036	.010	.001	20	F	.124	.031	.008	.000
	$W(c=.1)$	.158	.044	.014	.001		$W(c=.1)$	.131	.036	.011	.001
	$W(c=.7)$	.161	.046	.015	.002		$W(c=.7)$	.133	.037	.012	.001
50	F	.071	.018	.005	.000	100	F	.046	.012	.003	.000
	$W(c=.1)$	.073	.020	.006	.000		$W(c=.1)$	.047	.013	.003	.000
	$W(c=.7)$	.074	.020	.006	.000		$W(c=.7)$	.048	.013	.004	.000

From Table 1, the correspondence between Bayes factor and p values of the classical test statistics may be roughly summarized as follows. For moderate values of total sample size ($n_1=15,20$), the critical values $10^{-1/2}$, 10^{-1}, and 10^{-2} of the Bayes factor correspond to $6.0-3.1$, $4.9-1.0$ and $0.4-0.0$ percent values of the classical test statistics. In all cases, for large experiments ($n_1=50,100$), evidence at a very high significance level is required for the Bayes factor to favor strongly the more complex hypothesis. This phenomenon is related to the "Lindley paradox" so that the result of classical and Bayesian analyses may differ more and more as $n_1\rightarrow\infty$. For fixed Z, the phenomenon is easily seen from the fact that Stirling's formula asymptotically leads to

$$B^I_{12} \propto \exp\left\{-\frac{Z^2}{2}\right\}\prod_{k=1}^{2}\left(\frac{n_k+1/2}{2}\right)^{1/4},$$

and hence $B^I_{12}\rightarrow\infty$. Consequently $P(M_2\mid D)$ is of order $1/n^{1/2}$, $n=\text{Min}\{N_1,N_2\}$ and thus $P(M_1\mid D)\rightarrow 1$. For small experiments ($n_1=5,10$), the critical values $10^{-1/2}$, 10^{-1}, and 10^{-2} of the Bayes factor correspond to high percent values of the classical test statistics, and hence the approximate Bayes factor in (3.12) seems to require a more delicate investigation.

5. Concluding Remarks

We have suggested an approximate Bayes test criterion for the univariate Behrens- Fisher problem via a development of the imaginary training sample method introduced by Spiegelhalter and Smith(1982). The development is pertaining to the comparison of homo/heteroscedasticity of two normal means with heteroscedastic variances. The appeal of the method is that it provides a simple method for evaluating a value for the arbitrary constant attached to the Bayes factor (using with improper priors) to coming at a Bayes test criterion. It is seen that the Bayes criterion so obtained is expressed as a function of classical test criterion. So that it is possible to compare the suggested test criterion with classical test criteria in terms of p-value. The numerical study notes that (i) the criterion generally gives more conservative critical value than the classical tests do for moderate and large sample sizes and (ii) the criterion provides an automatic assessment of "significance", taking into the size and structure of the experiment. Thus this study can be taken as an another illustration of the result by Berger and Sellke (1987).

Suggested Bayes factor for the comparison of homo/heteroscedasticity of two normal means can be extended to the case of the multivariate Behrens-Fisher problem. A study pertaining to this extension is not unimportant, and hence it is left as a future study of interest.

Appendix : Proof of Lemma 3.1

Let respective joint density and prior density be (3.2) and (3.4). Then it follows from the definition of the marginal likelihood that

$$p(D \mid M_1) = C_1 \int (2\pi)^{-(n_1+n_2+1)/2} \prod_{k=1}^{2} \{\sigma_k^{-(n_k+5/2)}\} \cdot \times \exp\left\{-\frac{1}{2}\sum_{k=1}^{2}\{V_k + n_k(\overline{X}(k)-\mu)^2\}/\sigma_k^2\right\} \prod_{k=1}^{2} d\sigma_k^2 \, d\mu$$

Integrate the integrand with respect to σ_k^2's using the inverse chi-squared normalizing constants (cf. Lee 1988). This gives

$$C_1 \int (2\pi)^{-(n_1+n_2+1)/2} \prod_{k=1}^{2} \{\Delta_k V_k^{-(n_k+1/2)/2} (1+W_k)^{-(n_k+1/2)/2}\} d\mu, \tag{6.1}$$

where $W_k = n_k(\overline{X}(k)-\mu)^2 / V_k$.

Note that, under M_1, $p \lim_{n_k\to\infty} \overline{X}(k) = \mu$, $p \lim_{n_k\to\infty} V_k/n_k = \sigma_k^2$ and $W_k = O_p\left(\frac{1}{n_k}\right)$; that is $p \lim_{n_k\to\infty} n_k W_k =$ a constant. Hence, as $n_k \to \infty$, W_k approaches zero, in probability, so that

$$p \lim_{n_k\to\infty} (1+W_k)^{5/4} = 1. \tag{6.2}$$

Now consider $(1+W_k)^{\nu_k/2}$, where $\nu_k = n_k - 1$.

$$
\begin{aligned}
(1+W_k)^{\nu_k/2} &= \exp\left\{\frac{\nu_k}{2}\log(1+W_k)\right\}, \\
&= \exp\left\{\frac{\nu_k}{2}\left[W_k-\frac{W_k^2}{2}+\frac{W_k^3}{3}-\cdots\right]\right\}, \\
&= \exp\left\{\frac{\nu_k}{2}W_k\right\}\cdot\exp\left\{-\frac{\nu_k}{4}W_k^2+\frac{\nu_k}{6}W_k^3+\cdots\right\}.
\end{aligned}
$$

It is easy to see that since $W_k=O_p(n_k^{-1})$, $W_k{}^{\ell}=O_p(n_k^{-\ell})$. Therefore,

$$(1+W_k)^{\nu_k/2}=\exp\left\{\frac{\nu_k}{2}W_k\right\}\cdot\left[1+O_p\left(\frac{1}{n_k}\right)\right]. \tag{6.3}$$

Substituting (6.2) and (6.3) into (6.1), assuming n_k's are large, and neglecting terms of order n_k^{-1} in probability, gives approximately

$$
\begin{aligned}
&C_1(2\pi)^{-(n_1+n_2+1)/2}\prod_{k=1}^{2}\{\Delta_k V_k^{-(n_k+1/2)/2}\}\int\prod_{k=1}^{2}\exp\left\{-\sum_{k=1}^{2}\frac{\nu_k}{2}W_k\right\}d\mu \\
&\quad= C_1(2\pi)^{-(n_1+n_2+1)/2}\exp\left\{-\frac{Q}{2}\right\}\prod_{k=1}^{2}\{\Delta_k V_k^{-(n_k+1/2)/2}\}\int\prod_{k=1}^{2}\exp\left\{-\frac{1}{2}(\mu-\mu_0)^2/J\right\}d\mu, \qquad (6.4)
\end{aligned}
$$

where $J^{-1}=\sum_{k=1}^{2}n_k s_k^{-2}$, $s_k^2=V_k/\nu_k$, $\mu_0=K\overline{X}(1)+(I-K)\overline{X}(2)$, $K=n_1/(s_1^2 J)$,
$Q=(\overline{X}(1)-\overline{X}(2))^2/(s_1^2/n_1+s_2^2/n_2)$.

Thus the desired marginal likelihood $p(D\mid M_1)$ can be found by integrating with respect to μ using the univariate normal normalizing constant

$$p(D\mid M_1)=C_1(2\pi)^{-(n_1+n_2)/2}\prod_{k=1}^{2}\{\Delta_k V_k^{-(n_k+1/2)/2}\}\times J^{-1/2}\exp\left\{-\frac{Q}{2}\right\}\left[1+O_p\left(\frac{1}{n_k}\right)\right], \tag{6.5}$$

where $n=\text{Min}\{n_1,n_2\}$. Now

$$J^{-1/2}=\prod_{k=1}^{2}\{(n_k(n_k-1))^{-1/4}V_k^{1/4}\}\prod_{k=1}^{2}(1+U_k)^{-1/4}, \tag{6.6}$$

where $U_1=(s_1^2/n_1)^{1/2}(s_2^2/n_2)^{-1}(s_1^2/n_1)^{1/2}$ and $U_2=(s_2^2/n_2)^{1/2}(s_1^2/n_1)^{-1}(s_2^2/n_2)^{1/2}$.

Note that the values of U_1 and U_2 approach zero in probability, as n_1 and n_2 approach infinity, i.e.

$$p\lim_{n\to\infty}(1+U_k)^{-1/4}=1 \quad , \quad k=1,2, \tag{6.7}$$

where $n=\mathrm{Min}\{n_1, n_2\}$. Substituting the results (6.6) and (6.7) (neglecting terms of order n^{-1} in probability) into (6.5) yields $p(D \mid M_1)$ of Lemma 3.1.

References

[1] Berger, J. O. and Perrichi, L. R. (1996). "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, Vol. 91, 109-122.

[2] Berger, J. O. and Sellke, T. (1987). "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence," *Journal of the American Statistical Association*, Vol. 82, 112-122.

[3] DeGroot, M. H. (1970). *Optimal Statistical Decisions*, McGraw Hill, New York.

[4] Ghosh, B. K. (1975). "A two-stage procedure for the Behrens-Fisher problem," *Journal of the American Statistical Association*, Vol. 70, 457-462.

[5] Good, I. J. (1947). *Probability and the Weighing of Evidence*, Haffner, New York.

[6] Hettmansperger, T. P. (1973). "A large sample conservative test for location with unknown scale parameters," *Journal of the American Statistical Association*, Vol. 68, 466-468.

[7] Hogg, R. V. and Tanis, E. A. (1993). *Probability and Statistical Inference*, Prentice Hall, New Jersey.

[8] Jeffreys, H. (1961). *Theory of Probability*, Oxford: University Press.

[9] Johnson, R. A. and Weerahandi, S. (1988). "A Bayesian solution to the multivariate Behrens-Fisher problem," *Journal of the American Statistical Association*, Vol. 83, 145-149.

[10] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

[11] Kass, R. E. and Raftery, A. E. (1995). "Bayes factors," *Journal of the American Statistical Association*, Vol. 90, 773-795.

[12] Lampers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*, University Press, Rotterdam.

[13] Lee, P. M. (1988). *Bayesian Statistics : An Introduction*, Wiley, New York.

[14] O'Hagan, A. (1991). Contribution to the discussion of "Posterior Bayes Factors," *Journal of the Royal Statistical Society, Ser. B*, Vol. 53, 137.

[15] O'Hagan, A. (1995). "Fractional Bayes factors for model comparison," *Journal of the Royal Statistical Society, Ser. B*, Vol. 57, 99-138.

[16] Rencher, A. C. (1995). *Methods of Multivariate Analysis*, Wiley, New York.

[17] Smith, A. F. M. and Spiegelhalter, D. J. (1980). "Bayes factors and choice criteria for

linear models," *Journal of the Royal Statistical Society, Ser. B*, Vol. 42, 213-220.

[18] Spiegelhalter, D. J. and Smith, A. F. M. (1982). "Bayes Factors for Linear and Log-linear Models with Vague Prior Information," *Journal of the Royal Statistical Society, Ser. B*, Vol. 44, 377-387.