ESTIMATING A PRODUCT OF MEANS:
BAYESIAN ANALYSIS WITH REFERENCE PRIORS

by

James O. Berger and José M. Bernardo
Purdue University | University of Valencia

Author's Footnote

James O. Berger is the Richard M. Brumfield Distinguished Professor of Statistics, Statistics Department, Purdue University, West Lafayette, IN 47907; and José M. Bernardo is Professor of Statistics, Departamento de Estadística e I.O., Universidad de Valencia, 46071–Valencia, Spain. This work was supported by the U.S.-Spain Joint Committee for Scientific and Technological Co-operation, Grant CCB8409-025. The authors thank Kun-Liang Lu for performing the numerical computations, and the referees for very helpful comments.

◇

ABSTRACT

Suppose we observe $X \sim \mathcal{N}(\alpha, 1)$ and, independently, $Y \sim \mathcal{N}(\beta, 1)$ and are concerned with inference, mainly estimation and confidence statements, about the product of means $\theta = \alpha\beta$. This problem arises, most obviously, in situations of determining area based on measurements of length and width. It also arises in other practical contexts, however. For instance, in gypsy moth studies, the hatching rate of larvae per unit area can be estimated as the product of the mean of egg masses per unit area times the mean number of larvae hatching per egg mass. Approximately independent samples can be obtained for each mean (cf. Southwood (1978)).

A noninformative prior Bayesian approach to the problem is considered, in particular the *reference* prior approach of Bernardo (1979). An appropriate reference prior for the problem is developed, and relatively easily implementable formulas for posterior moments (e.g. the posterior mean and variance) and credible sets are derived. Comparisons with alternative noninformative priors and with classical procedures are also given.

The motivation for this work was in part the statistical importance of the problem and the difficulty in producing reasonable classical analyses, and in part to provide an interestingly complex example of a recently developed method of deriving reference priors for problems with nuisance parameters. This new method is briefly described in the paper. The problem is also of interest because of its mention in Efron (1986) as a situation for which "standard" noninformative prior Bayesian theories encounter difficulties.

KEY WORDS AND PHRASES: Noninformative Prior; Posterior Distribution; Nuisance Parameters.

1. INTRODUCTION

Suppose we observe $X \sim \mathcal{N}(\alpha, 1)$ and, independently, $Y \sim \mathcal{N}(\beta, 1)$ and are concerned with inference, mainly estimation and confidence statements, about the product of means $\theta = \alpha\beta$. We will, for the most part assume that $\alpha > 0$ and $\beta > 0$; Section 4 gives the modifications necessary for the unrestricted case. Also note that if X and Y have known variances σ_1^2 and σ_2^2, the problem can be reduced to that above by considering $X^* = x/\sigma_1, \alpha^* = \alpha/\sigma_1, Y^* = Y/\sigma_2$, and $\beta^* = \beta/\sigma_2$; since $\theta = \alpha\beta = \alpha^*\beta^*\sigma_1\sigma_2$, inferences about $\alpha^*\beta^*$ can be easily translated into inferences about θ.

In Section 2, we briefly review classical approaches to the problem, and indicate their inadequacies. Section 3 outlines a noninformative prior Bayesian approach to the problem, namely the reference prior approach of Bernardo (1979). To apply the reference prior approach here, an extension of the theory to deal with nuisance parameters was required. Berger and Bernardo (1987) present this extension, an outline of which, and application to the product of means problem, is given in Section 5. The product of means problem will be seen to be a particularly interesting application of the general theory from a foundational perspective.

Efron (1986) mentioned this problem as an example in which "standard" noninformative prior Bayesian theory encounters difficulties. Our motivation for considering this problem was, in part, to determine whether or not the reference prior approach would overcome the difficulties mentioned by Efron.

Section 6 compares several possible noninformative priors. Following Efron (1986), some of the comparisons are in frequentist terms.

2. CLASSICAL METHODS

The unbiased estimator of $\theta = \alpha\beta$ is, of course, $\hat{\theta}_U = xy$. This could be negative, an annoying possibility since θ is positive. Perhaps even worse is that, if (say) $x = y = -3$, then $\hat{\theta}_U = 9$, even though α and β must both be very close to zero.

The maximum likelihood estimator is $\hat{\theta}_M = x^+ y^+$, where "+" denotes the positive part. There are no obvious absurdities with this estimator, but reporting zero when it is known that both $\alpha > 0$ and $\beta > 0$ is awkward.

The variance of $\hat{\theta}_U$ is

$$V(\alpha,\beta) = E_{\alpha,\beta}(\hat{\theta}_U - \theta)^2 = \alpha^2 + \beta^2 + 1.$$

The variance of $\hat{\theta}_M$ (and its mean squared error) are fairly complicated, but approximately equal $V(\alpha,\beta)$ for moderate to large α and β. The difficulty with V is that

$$\sup_{(\alpha,\beta)} V(\alpha,\beta) = \infty;$$

this makes it difficult to report a "standard error". One might consider the estimated frequentist approach (cf. Kiefer (1977) and Berger (1987)), and report

$$\hat{V}(x,y) = x^2 + y^2 - 1.$$

Since $E_{\alpha,\beta}[\hat{V}] = \alpha^2 + \beta^2 + 1 = V(\alpha,\beta)$, this report can be claimed to be a valid frequentist estimated variance. Even this has problems, however: $\hat{V}$ can be negative, and even $\hat{V}^+$ (which is a conservatively valid frequentist report) has the annoying property of often being zero, a reported error which will be met with deserved skepticism.

Finding confidence sets for θ is also a very difficult problem. In fact, we do not know of any non-Bayesian approaches that would be likely to be successful.

The point of the above comments is to indicate the difficulties that a classical approach to the problem faces. Undoubtedly progress could be made in a classical direction, but it is very hard.

3. NONINFORMATIVE PRIOR BAYESIAN ANALYSIS

The "standard" noninformative prior for the problem is $\pi_u(\alpha,\beta) = 1$, since (α,β) is a location vector. Use of this prior leads easily to an estimate (the posterior mean) and standard error (the square root of the posterior variance). Indeed, analogously to Example 7 in Berger (1985, pp. 135 and 138), one obtains

$$\begin{aligned}\hat{\theta}_{\pi_u} &= \text{ posterior mean}\\ &= [x+\psi(x)][y+\psi(y)],\\ V_{\pi_u} &= \text{ posterior variance}\\ &= [1+x^2+x\psi(x)][1+y^2+y\psi(y)] - (\hat{\theta}_{\pi_u})^2,\end{aligned}$$

where

$$\psi(z) = \phi(z)/[1-\Phi(z)],$$

ϕ and Φ being the standard normal density and c.d.f., respectively.

As an estimate, $\hat{\theta}_{\pi_u}$ is perfectly sensible, being always positive, yet close to zero if x or y is quite negative, and being approximately equal to xy if x and y are moderate to large. Likewise, V_{π_u} is sensible; it is always positive, and increases to a maximum of x^2+y^2-1 for large x and y. For simplicity and sensibility, one could not hope to do much better.

Unfortunately, this is a problem where there can be considerable skewness in the "information" about θ (see Figure 1), so that an estimate and standard error can be quite inadequate in describing the location of θ. Credible sets (the Bayesian version of confidence sets) are thus needed, but the calculations for such can no longer be done in closed form. A pleasant surprise is that, although such Bayesian calculations involving (α,β) might be thought to require two-dimensional numerical integration, one-dimensional numerical integration actually suffices.

Before proceeding with this development, it is time to introduce two other noninformative priors that will be considered:

$$\pi_s = \sqrt{\alpha^2 + \beta^2} \quad \text{and} \quad \pi_r = (\alpha^2 + \beta^2)^{1/2}/(\alpha\beta).$$

The prior π_s is that which Efron (1986) reports as being the "best" noninformative prior, according to some numerical work, and was proposed by C. Stein based on reasoning in Stein (1982). This prior will also be seen to be the natural "reference" prior for the problem. The prior π_r is used in Section 5 for certain illustrative purposes. Overall, we follow Efron and Stein in recommending π_s.

The formulas for calculating with π_u, π_s, and π_r follow. For notational convenience, we rewrite the priors as

$$\pi_{ij}(\alpha, \beta) = (\alpha^2 + \beta^2)^{i/2}(\alpha\beta)^{-j}; \tag{3.1}$$

thus π_u, π_s, and π_r are π_{00}, π_{10}, and π_{11}, respectively. Also, define the functions

$$A(\omega) = \frac{x}{\omega} + y\omega,\ B(\omega) = x\omega - \frac{y}{\omega},\ C(\omega) = \omega^2 + \omega^{-2}$$

(which we will write A, B, and C for convenience),

$$\begin{aligned} \psi_0(t,\omega) &= 0,\ \psi_1(t,\omega) = \frac{1}{C},\ \psi_2(t,\omega) = \frac{\sqrt{t}}{C} + \frac{A}{C^2}, \\ \varphi_0(\omega) &= \frac{1}{\sqrt{C}},\ \varphi_1(\omega) = \frac{A}{C^{3/2}},\ \varphi_2(\omega) = \frac{A^2}{C^{5/2}} + \frac{1}{C^{3/2}}, \\ H_{ij}(t|x,y) &= \int_0^\infty \omega^{-1}C^{i/2}\phi(\frac{B}{\sqrt{C}})\{\psi_{i-2j+1}(t,\omega)\phi(\sqrt{tC} - \frac{A}{\sqrt{C}}) \\ &\qquad + \varphi_{i-2j+1}[1 - \Phi(\sqrt{tC} - \frac{A}{\sqrt{C}})]\}d\omega. \end{aligned}$$

<u>LEMMA 1</u>. *For the prior $\pi_{ij}(\alpha,\beta)$, the posterior density of $\theta = \alpha\beta$, given data (x,y), is*

$$\pi^*_{ij}(\theta) \equiv \pi_{ij}(\theta|x,y) = \frac{\theta^{(\frac{1}{2}i-j)}\int_0^\infty \omega^{-1}C^{i/2}\phi(\frac{B}{\sqrt{C}})\phi(\sqrt{\theta C} - \frac{A}{\sqrt{C}})d\omega}{2H_{ij}(0|x,y)}. \tag{3.2}$$

The posterior c.d.f. of θ is

$$F_{ij}(t|x,y) = \int_0^t \pi_{ij}(\theta|x,y)d\theta = 1 - \frac{H_{ij}(t|x,y)}{H_{ij}(0|x,y)}. \tag{3.3}$$

Posterior moments are given by

$$E^{\pi_{ij}(\theta|x,y)}[\theta^n] = \frac{H_{i(j-n)}(0|x,y)}{H_{ij}(0|x,y)}. \tag{3.4}$$

(See (A1) *in Appendix I for the definition of arbitrary* ψ_k *and* φ_k*.)*

Proof. Given in Appendix I. □

The important thing to note about Lemma 1 is that the c.d.f.s, F_{ij}, are expressible as one-dimensional integrals. Thus (3.3) can be used to easily determine quantiles of the posterior distribution. For instance, Table 1 presents, for a variety of x and y, quantiles of the posterior for θ corresponding to the prior π_s. An equal-tailed 90% credible region for θ when $x = 1, y = 1$ would thus be (0.17, 5.80); note that the posterior median for this data is 1.76, highlighting the skewness of the posterior. (The calculations were done by quadrature, using CADRE, on a Vax 11/780 computer; all numbers are accurate through the given digits.)

It is interesting to examine typical posterior densities for θ that result from use of π_u, π_s, and π_r. Figure 1 displays such posteriors for several possible (x, y) pairs. Note that all posteriors have mass piled up near zero when x and y are small, but for larger x and y the posteriors look more normal (except that π_r^* will always have a spike at zero). The skewness of the posteriors is also clearly revealed by these figures. Note finally that π_r^* is to the left of π_u^* which is to the left of π_s^*. It is interesting to simply look at these posteriors and judge which seems most reasonable intuitively; our choice on this basis is π_s^*.

4. BAYESIAN ANALYSIS FOR UNRESTRICTED α, β

When α and β are unrestricted, i.e. assumed only to lie in R^1, only slight modifications of the formulas in Section 3 are needed. The only change needed in the definition of the priors, is that π_r

Table 1. The .05, .25, .50, .75, and .95 Quantiles of the Posterior Corresponding to π_s

y	x								
	-1.5	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	8.0
	0.01	0.01	0.02	0.04	0.06	0.08	0.11	0.13	0.21
	0.07	0.08	0.13	0.21	0.32	0.45	0.59	0.73	1.16
-1.5	0.19	0.23	0.34	0.52	0.77	1.07	1.38	1.70	2.69
	0.43	0.51	0.74	1.09	1.55	2.09	2.66	3.26	5.09
	1.10	1.29	1.78	2.47	3.34	4.33	5.40	6.51	9.99
	0.01	0.02	0.03	0.04	0.07	0.10	0.14	0.17	0.27
	0.08	0.10	0.16	0.26	0.40	0.57	0.74	0.91	1.45
-1.0	0.23	0.27	0.41	0.64	0.96	1.32	1.70	2.10	3.31
	0.51	0.61	0.89	1.31	1.88	2.52	3.21	3.93	6.13
	1.29	1.50	2.08	2.90	3.92	5.08	6.32	7.61	11.65
	0.02	0.03	0.04	0.08	0.14	0.20	0.26	0.32	0.51
	0.13	0.16	0.26	0.45	0.71	1.01	1.32	1.63	2.57
0.0	0.34	0.41	0.65	1.04	1.58	2.19	2.83	3.47	5.46
	0.74	0.89	1.32	2.00	2.89	3.89	4.94	6.02	9.37
	1.78	2.08	2.91	4.10	5.55	7.17	8.90	10.68	16.26
	0.04	0.04	0.08	0.17	0.34	0.52	0.68	0.84	1.31
	0.21	0.26	0.45	0.83	1.44	2.11	2.77	3.42	5.39
1.0	0.52	0.64	1.04	1.76	2.77	3.90	5.05	6.21	12.85
	1.09	1.31	2.00	3.10	4.55	6.16	7.82	9.53	14.76
	2.47	2.90	4.10	5.80	7.90	10.20	12.60	15.08	22.79
	0.06	0.07	0.14	0.34	0.89	1.63	2.25	2.79	4.37
	0.32	0.40	0.71	1.44	2.71	4.19	5.63	7.02	11.13
2.0	0.77	0.96	1.58	2.77	4.53	6.50	8.49	10.48	16.46
	1.55	1.88	2.89	4.55	6.78	9.24	11.76	14.31	22.12
	3.34	3.92	5.55	7.90	10.80	13.95	17.20	20.53	30.83
	0.08	0.10	0.20	0.52	1.63	3.47	5.22	6.76	10.97
	0.45	0.57	1.01	2.11	4.19	6.70	9.20	11.62	18.69
3.0	1.07	1.32	2.19	3.90	6.50	9.47	12.46	15.43	24.34
	2.09	2.52	3.89	6.16	9.24	12.66	16.14	19.64	30.30
	4.33	5.08	7.17	10.20	13.95	18.02	22.20	26.45	39.51
	0.11	0.14	0.26	0.68	2.25	5.22	8.24	11.01	18.54
	0.59	0.74	1.32	2.77	5.63	9.20	12.79	16.28	26.49
4.0	1.38	1.70	2.83	5.05	8.49	12.46	16.47	20.47	32.39
	2.66	3.21	4.94	7.82	11.76	16.14	20.60	25.08	38.66
	5.44	6.32	8.90	12.60	17.20	22.20	27.31	32.50	48.38
	0.13	0.17	0.32	0.84	2.79	6.76	11.01	14.99	25.96
	0.73	0.91	1.63	3.42	7.02	11.62	16.28	20.86	34.23
5.0	1.70	2.10	3.47	6.21	10.48	15.43	20.47	25.48	40.44
	3.26	3.93	6.02	9.53	14.31	19.64	25.08	30.55	47.06
	6.51	7.61	10.68	15.08	20.53	26.45	32.50	38.62	57.31
	0.21	0.27	0.51	1.31	4.37	10.97	18.54	25.96	47.11
	1.16	1.45	2.57	5.39	11.13	18.69	26.49	34.23	57.04
8.0	2.69	3.31	5.46	9.74	16.46	24.34	32.39	40.44	64.49
	5.09	6.13	9.37	14.76	22.12	30.30	38.66	47.06	72.39
	9.99	11.65	16.26	22.79	30.83	39.51	48.38	57.31	84.57

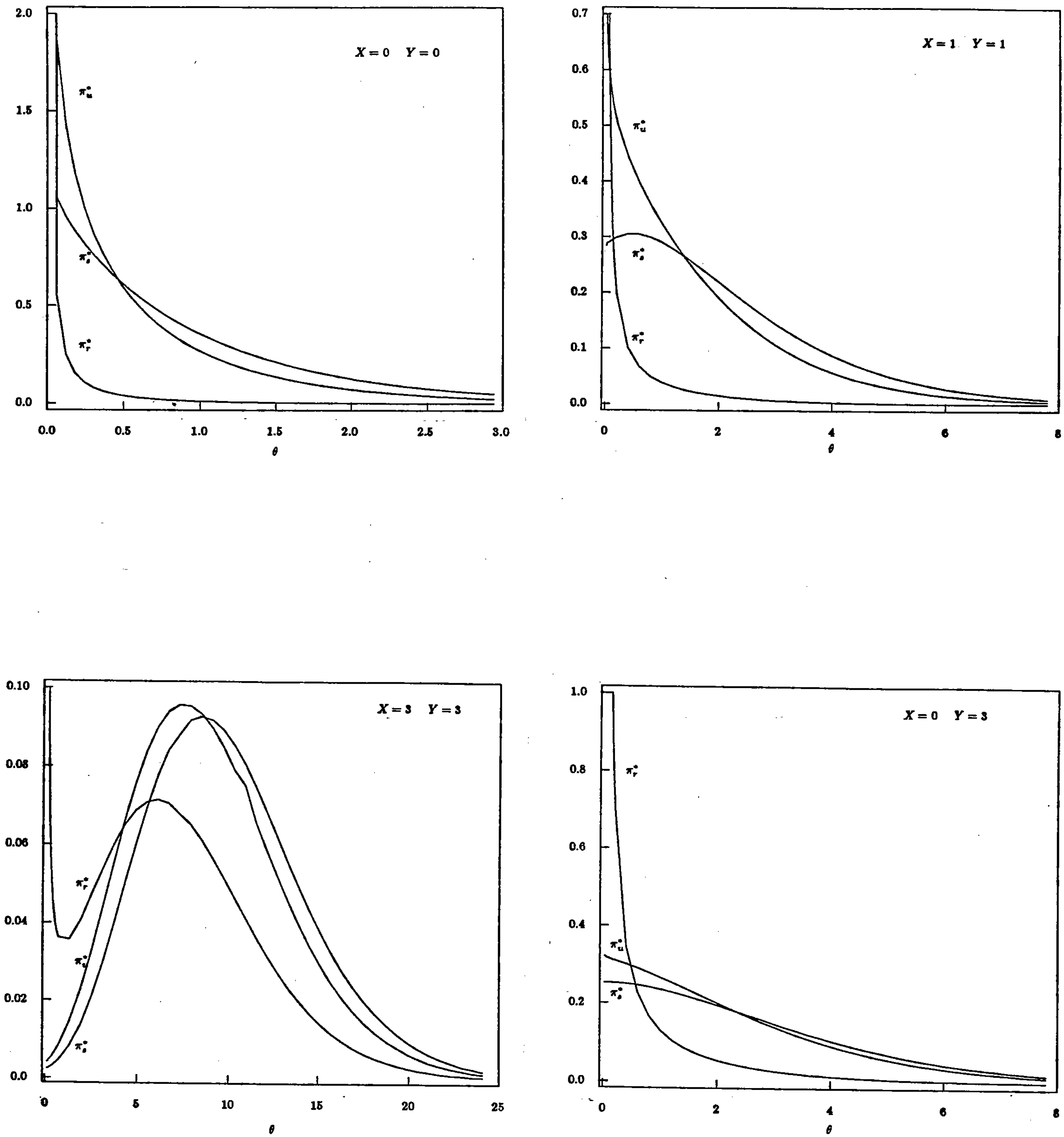

Figure 1. Posterior Distributions of θ. The posterior distributions corresponding to the three noninformative priors, π_u, π_s, and π_r, are graphed for four possible pairs of data.

be changed to $\pi_r(\alpha,\beta) = (\alpha^2+\beta^2)^{1/2}/|\alpha\beta|$. Lemma 1 becomes, defining

$$N_{ij}(\theta|x,y) = H_{ij}(0|x,y)\pi_{ij}(\theta|x,y),$$

$$D_{ij}(0) = H_{ij}(0|x,y) + H_{ij}(0|-x,y) + H_{ij}(0|x,-y) + H_{ij}(0|-x,-y)$$

and using the notation of Section 3:

<u>LEMMA 2</u>. *For the prior $\pi_{ij}(\alpha,\beta)$, the posterior density of $\theta = \alpha\beta$, given data (x,y), is*

$$\tilde{\pi}_{ij}(\theta|x,y) = \begin{cases} [N_{ij}(-\theta|-x,y) + N_{ij}(-\theta|x,-y)]/D_{ij}(0) & \text{if } \theta < 0 \\ [N_{ij}(\theta|x,y) + N_{ij}(\theta|-x,-y)]/D_{ij}(0) & \text{if } \theta > 0. \end{cases}$$

The posterior c.d.f. of θ is

$$F_{ij}(t|x,y) = \begin{cases} [H_{ij}(-t|-x,y) + H_{ij}(-t|x,-y)]/D_{ij}(0) & \text{if } t \le 0, \\ 1 - [H_{ij}(t|x,y) + H_{ij}(t|-x,-y)]/D_{ij}(0) & \text{if } t > 0. \end{cases}$$

<u>Proof</u>: Each integral can be divided into a sum of integrals over the four quadrants in (α,β) space. Separately transform each quadrant into the positive quadrant by sign changes on α and/or β. Note that the priors are unaffected by sign changes. As for the likelihood, defining (say) $\eta = -\alpha$ yields

$$exp\{-\frac{1}{2}(x-\alpha)^2\} = exp\{-\frac{1}{2}(x+\eta)^2\} = exp\{-\frac{1}{2}(-x-\eta)^2\};$$

any integral over positive η will now be exactly like those in Section 3, pretending that $-x$ is the data. The verification of Lemma 2 is then just bookkeeping. □

5. REFERENCE PRIORS

<u>5.1 Development</u>

Bernardo (1979) introduced the notion of a reference prior. The idea, for an experiment with density $f(x|\theta)$ and prior density $\pi(\theta)$, is to consider the amount of information about θ that the experiment can be expected to provide. Bernardo argues for using, as a measure of this information,

$$I^\theta\{f,\pi\} = \int\int f(x|\theta)\pi(\theta)\log\frac{\pi(\theta|x)}{\pi(\theta)}\,d\theta dx.$$

The reference prior is that π which maximizes this quantity, the rationale being that the larger this "information" is, the less informative is the prior.

For a variety of technical reasons, the reference prior is actually defined, not for the experiment $f(x|\theta)$ but via an asymptotic limit of i.i.d. repetitions of the experiment. In situations where asymptotic normality of the posterior holds, Bernardo (1979) showed that the reference prior for θ, providing there are no nuisance parameters, is Jeffreys's prior $\pi(\theta) = \sqrt{|I(\theta)|}$; here $I(\theta)$ is the expected Fisher information matrix, and $|A|$ denotes the determinant of A.

Suppose, now, that θ is the parameter of interest, but that ω is a nuisance parameter. Write the expected Fisher information matrix as

$$I(\theta,\omega) = \begin{pmatrix} I_{11}(\theta,\omega) & I_{12}(\theta,\omega) \\ I_{12}'(\theta,\omega) & I_{22}(\theta,\omega) \end{pmatrix},$$

the blocks corresponding to θ and ω in the usual way, and assume that asymptotic normality of the posterior holds. Bernardo (1979) suggests choosing conditional distributions $\pi(\omega|\theta)$, and then forming the "marginal" experiment for θ by integrating out over ω with respect to $\pi(\omega|\theta)$ and finding the reference prior $\pi(\theta)$ in this marginal experiment.

The hitch in this plan is the difficulty of choosing $\pi(\omega|\theta)$. Subjective choices are desirable, but somewhat defeat the motive of trying to be "noninformative" (though intuition would suggest that the influence of $\pi(\omega|\theta)$ might be substantially less than $\pi(\theta)$). A natural choice for $\pi(\omega|\theta)$ is the reference prior for ω in the experiment with θ assumed to be known. This works well when it turns out to be a proper distribution, but runs into normalization difficulties otherwise. Berger and Bernardo (1987) propose a scheme to circumvent these difficulties, a scheme which leads to the following program for determining the reference prior.

Step 1. Let $\pi(\omega|\theta)$ be the usual reference prior for ω with θ given, defined by

$$\pi(\omega|\theta) = |I_{2,2}(\theta,\omega)|^{1/2}. \tag{4.1}$$

Step 2. Choose a sequence $\Lambda_1 \subset \Lambda_2 \subset \ldots$ of subsets of the parameter space Λ for (θ, ω), such that $\bigcup_i \Lambda_i = \Lambda$, and such that $\pi(\omega|\theta)$ has finite mass on

$$\Omega_{i,\theta} = \{\omega : (\theta, \omega) \epsilon \Lambda_i\}$$

for all θ. Then normalize $\pi(\omega|\theta)$ on each $\Omega_{i,\theta}$, obtaining

$$p_i(\omega|\theta) = K_i(\theta)\pi(\omega|\theta)1_{\Omega_{i,\theta}}(\omega), \tag{4.2}$$

where 1_Ω denotes the indicator function on Ω and

$$K_i(\theta) = 1 \Big/ \int_{\Omega_{i,\theta}} \pi(\omega|\theta) d\omega. \tag{4.3}$$

Step 3. Find the marginal reference prior for θ, with respect to $p_i(\omega|\theta)$. This is (see Berger and Bernardo, 1987).

$$\pi_i(\theta) = exp\{\frac{1}{2} \int_{\Omega_{i,\theta}} p_i(\omega|\theta) \log(|I(\theta,\omega)|/|I_{2,2}(\theta,\omega)|) d\omega\}, \tag{4.4}$$

assuming the integral exists.

Step 4. Define the reference prior for (θ, ω), when ω is a nuisance parameter, by

$$\pi(\theta, \omega) = \lim_{i \to \infty} \left[\frac{K_i(\theta)\pi_i(\theta)}{K_i(\theta_0)\pi_i(\theta_0)} \right] \pi(\omega|\theta), \tag{4.5}$$

assuming the limit exists; here θ_0 is any fixed point.

The above program is not easy, supporting the statement in Efron (1986) that "the theory of Bayesian objectivity cannot be a simple one." The most serious difficulty is the need to choose the sequence $\{\Lambda_i\}$. We return to this issue after applying the above theory to the product of means example.

5.2 Reference Priors for the Product of Means

To apply the theory in Section 4.1, the nuisance parameter must first be selected. (It is shown in Berger and Bernardo (1987) that the reference prior does not depend on the particular parameterization chosen for the nuisance parameter.) For consistency with Appendix I, we choose

$$\omega = (\beta/\alpha)^{1/2}$$

as the nuisance parameter. Note that the transformation from $(\alpha,\beta) \to (\theta,\omega)$ is one-to-one with Hessian

$$H = \frac{\partial(\alpha,\beta)}{\partial(\theta,\omega)} = \begin{pmatrix} \frac{1}{2\omega\sqrt{\theta}} & -\frac{\sqrt{\theta}}{\omega^2} \\ \frac{\omega}{2\sqrt{\theta}} & \sqrt{\theta} \end{pmatrix}.$$

Since the information matrix for (α,β) in the original problem is the identity, it follows that the information matrix for (θ,ω) is

$$I(\theta,\omega) = H'\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} H = \begin{pmatrix} \frac{1}{4\theta}(\omega^2+\omega^{-2}) & \frac{\omega}{2}(1-\omega^{-4}) \\ \frac{\omega}{2}(1-\omega^{-4}) & \theta(1+\omega^{-4}) \end{pmatrix}.$$

Thus

$$|I_{22}(\theta,\omega)| = \theta(1+\omega^{-4}), \quad \frac{|I(\theta,\omega)|}{|I_{22}(\theta,\omega)|} = \frac{1}{\theta(\omega^2+\omega^{-2})}. \tag{4.6}$$

Step 1. Here

$$\pi(\omega|\theta) = |I_{22}(\theta,\omega)|^{1/2} = \sqrt{\theta}(1+\omega^{-4})^{1/2}.$$

Step 2. A natural sequence of sets $\{\Lambda_i^*\}$ to choose in (α,β) space is a collection of rectangles

$$\Lambda_i^* = (0,l_i) \times (0,k_i), \tag{4.7}$$

where $l_i \to \infty$ and $k_i \to \infty$. Transforming to (θ,ω) space yields, as the required θ-sections,

$$\Omega_{i,\theta} = (\sqrt{\theta}/l_i, k_i/\sqrt{\theta}).$$

Thus (4.3) and (4.2) become

$$K_i^{-1}(\theta) = \int\limits_{\sqrt{\theta}/l_i}^{k_i/\sqrt{\theta}} \sqrt{\theta}(1+\omega^{-4})^{1/2} d\omega, \tag{4.8}$$

$$p_i(\omega|\theta) = K_i(\theta)\sqrt{\theta}(1+\omega^{-4})^{1/2} 1_{\Omega_{i,\theta}}(\omega). \tag{4.9}$$

Step 3. Using (4.6) and (4.9), (4.4) becomes

$$\pi_i(\theta) = exp\{-\frac{1}{2} \int\limits_{\sqrt{\theta}/l_i}^{k_i/\sqrt{\theta}} K_i(\theta)\sqrt{\theta}(1+\omega^{-4})^{1/2} \log(\theta[\omega^2+\omega^{-2}])d\omega\}. \tag{4.10}$$

Step 4. In Appendix II, we show that the limit in (4.5) yields

$$\pi(\theta,\omega) = \sqrt{\theta}(1+\omega^{-4})^{1/2}. \tag{4.11}$$

Transforming back to (α,β) and noting that the Jacobian of the transformation is $(\beta/\alpha)^{1/2}$, one obtains the reference prior

$$(\alpha^2+\beta^2)^{1/2} = \pi_s(\alpha,\beta).$$

It is of considerable interest to investigate how the choice of $\{\Lambda_i\}$ affects the final result. The following theorem shows that quite general choices of $\{\Lambda_i\}$ yield the same reference prior here.

THEOREM 1. *Let Λ be any compact convex set in the positive quadrant which contains the origin and a line segment of each axis. Define $\Lambda_i = k_i\,\Lambda$ (i.e., the set formed by multiplying each point in Λ by k_i), where $k_i \to \infty$ as $i \to \infty$. Then π_s is the reference prior corresponding to $\{\Lambda_i\}$.*

Proof: Given in Appendix II. □

The convexity condition in the above theorem could most likely be eliminated, but the theorem does establish that the reference prior is quite insensitive to the choice of $\{\Lambda_i\}$. Part of the theoretical interest in this example, on the other hand, is that the choice of $\{\Lambda_i\}$ is not completely irrelevant. To demonstrate that this is so, consider choosing the Λ_i to be rectangles in (θ, ω) space; for instance, choose them so that

$$\Omega_{i,\theta} = (1/i, i).$$

Then, since $\pi(\omega|\theta)$ and $|I(\theta,\omega)|/|I_{22}(\theta,\omega)|$ factor into terms involving only θ and only ω, it is easy to check that the reference prior is

$$\pi(\theta, \omega) = \theta^{-1/2}(1 + \omega^{-4})^{1/2}.$$

Transforming back to (α, β) yields

$$(\alpha^2 + \beta^2)^{1/2}/(\alpha\beta) = \pi_r(\alpha, \beta).$$

This reference prior is not as arbitrary as it first appears to be, precisely because the given (θ, ω) transformation is that in which $|I_{22}|$ and $|I|/|I_{22}|$ factor into terms involving only θ and only ω. This is thus the transformation in which one might argue for "independence" of (θ, ω), leading rather naturally to consideration of rectangles for Λ_i in this space.

Although $\pi(\theta, \omega)$ can depend on the choice of the Λ_i, it will not depend on the choice of the nuisance parameter ω. Thus, if the Λ_i are defined as the appropriate transforms of (4.7), then $\pi(\theta, \omega)$ will turn out to be (4.11) no matter how ω is chosen. This is established in Berger and Bernardo (1987).

How is $\{\Lambda_i\}$ to be chosen when the choice matters (Note that this is the first studied example where it does matter.) We have no clearcut answer to this question, precisely because a dependence of the solution on the Λ_i is essentially an indication that some subjective input is needed; one cannot unambiguously define a reference prior.

One possible interpretation of the Λ_i is that they should reflect one's intuition concerning what is meant by "noninformative". For a normal mean μ, one will rarely be truly "noninformative" between $\mu = 5$ and $\mu = 10^{100}$, but one might be willing to model "noninformative" by saying that for some unknown large interval one wants to be noninformative. This notion would lead to choosing the Λ_i to be a series of nested intervals for μ which converge to R^1. Similar reasoning can be applied to the (α, β) situation, leading to the Λ_i^* in (4.7). Indeed, most frequently it would probably be natural to choose the Λ_i to be simple sets (rectangles, spheres, etc.) in the original parameterization of the problem; initially chosen parameterizations are often ones in which the analyst is roughly noninformative over natural sets. And note that Theorem 1 shows π_s to be the reference prior for any choice of such natural sets in the original parameterization. In this light, it is interesting to note that rectangles in (θ, ω) space (which lead to π_r) transform into wedges in (α, β) space; wedges are rather unnatural, implying that the boundaries ($\alpha = 0$ or $\beta = 0$) are somehow at "infinity".

6. COMPARISON OF NONINFORMATIVE PRIORS

The three noninformative priors, π_u, π_s, and π_r, all have some type of justification. The case for π_u is simply that the constant prior for (α, β) is standard. Many counterexamples have by now been created, however, which indicate that a "good" noninformative prior for the full parameter need not be good for lower dimensional functions of it. Thus, Efron (1986) observes: "The correct objective prior seems to depend on which parameter we want to estimate." Note that reference priors explicitly depend on which function of (α, β) one desires to estimate.

The case for π_s, argued here, is that it is the natural reference prior when θ is the parameter of interest (natural, in the sense that it corresponds to natural sequences $\{\Lambda_i\}$). The alternative reference prior π_r corresponded to a rather strange sequence $\{\Lambda_i\}$ (when considered in (α, β) space), and was mainly included to indicate the dependence of the reference prior on $\{\Lambda_i\}$. All in all, we would expect π_s to perform best.

What does "best" mean here One interpretation is simply that it should yield the most intuitively appealing results. To judge whether this is so, one might look at typical posteriors for each prior, as given in Figure 1. Examination of these figures reveals that π_r^* is highly counter-intuitive; the spike as $\theta \to 0$ ($\pi_r^*(\theta)$ grows at least as fast as a multiple of $\theta^{-1/2}$ as $\theta \to 0$) makes little intuitive sense. This spike exists because π_r itself blows up as $\theta = \alpha\beta \to 0$, so that one cannot justify the spike as being somehow indicated by the data. A referee has observed, however, that if $\log\theta$ were the parameter of interest, and if $\log\alpha$ and $\log\beta$ were the natural parameters in which to be noninformative (which might arise, for instance, if it were natural to think in terms of the orders of magnitude of α, β, and θ), then π_r might be quite reasonable. In particular, a transformation to $\log\theta$ removes the spike in π_r, and natural regions $\{\Lambda_i\}$ in $(\log\alpha,\ \log\beta)$–space result in π_r as the reference prior. While it would be hard to settle this issue outside of a clear practical context, we certainly are sympathetic to the underlying idea: in different contexts, different $\{\Lambda_i\}$ (and hence possibly different reference priors) might indeed be reasonable.

Comparison of π_u^* and π_s^* is more difficult. Looking at the data and the posteriors, one might judge that the π_u^* are shifted too far to the left, but this is not unarguably the case. Thus other criteria are needed to help distinguish between the two.

Various frequentist criteria have proved helpful in evaluating noninformative priors. The basic idea is to use the prior to generate a statistical procedure, and investigate the frequentist properties of the procedure. If the procedure resulting from one prior has substantially better properties than that resulting from another prior, then the latter prior is suspect. There is, of course, no guarantee that this approach to comparing priors will work. Note also that one cannot typically expect the procedure developed from the noninformative prior to have uniformly good frequentist properties; sensible conditional behavior and uniform frequentist properties are often simply not compatible.

The most common frequentist comparison of noninformative priors is via admissibility or risk dominance of resulting estimators (cf. Berger, 1985). Another method is to compare confidence

Table 2. Frequentist Coverage Probabilities of 0.05 and 0.95 Posterior Quantiles

(α,β)	$P_{.05}(\alpha,\beta)$		$1 - P_{.95}(\alpha,\beta)$	
	π_u^*	π_s^*	π_u^*	π_s^*
(0,0)	1	1	0	0
(1,1)	0.024	0.047	0.002	0.000
(2,2)	0.023	0.035	0.037	0.014
(3,3)	0.028	0.037	0.069	0.043
(4,4)	0.033	0.048	0.068	0.049
(5,5)	0.037	0.046	0.068	0.052

properties of sets arising from the posteriors. Indeed, Stein (1982) actually uses this approach to suggest "good" noninformative priors. Efron (1986) reports that π_s is the result for the product of means problem.

To compare π_u and π_s in this fashion, we follow the lead of Efron (1986) and investigate the γ^{th} posterior quantile, θ_γ, defined by $F(\theta_\gamma) = \gamma$, where F is the posterior c.d.f. as given in (3.3). In particular, we will calculate

$$P_\gamma(\alpha,\beta) = Pr_{\alpha,\beta}(\theta \le \theta_\gamma),$$

the frequentist probability that θ_γ (which depends on X and Y) is larger than the actual θ. Table 2 presents $P_{.05}(\alpha,\beta)$ and $1 - P_{.95}(\alpha,\beta)$ for various values of (α,β). A frequentist would want $P_\gamma(\alpha,\beta)$ to be close to γ, indicating that θ_γ exceeds θ the "correct" proportion of the time. Note that this table differs from that in Efron (1986) in that we are considering the problem with $\alpha > 0$ and $\beta > 0$, whereas he considered the unrestricted case; his numbers were also an approximation arising from a slightly different problem. The calculations in Table 2 were done by simulation, generating 4000 (X, Y) pairs for each (α,β), calculating the indicated posterior quantiles for each pair, and determining the proportion that exceeded $\theta = \alpha\beta$. This was done on a Vax 11/780 computer; the standard error of the entries in Table 2 is about .0035.

Note first that, when $\alpha = \beta = 0$ so $\theta = 0$, the posterior quantiles perform poorly in frequentist terms. But this is clearly an unavoidable conflict, and is of no help in choosing between π_u and π_s. By continuity of the coverage probability, this difficulty will persist for θ near zero; one simply cannot expect posterior quantiles to have "proper" frequentist behavior near the finite boundaries of parameter spaces (which can be interpreted as a criticism of demanding uniform frequentist behavior over the entire parameter space).

Thus, focus on the larger (α, β) in Table 2. Clearly the posterior quantiles for π_s^* yield frequentist error rates which are closer to the "ideal" 0.05 than are those for π_u^*. The posterior quantiles for π_u^* simply seem to be too small, being to the left of θ more often than one would desire. Another way of saying this is that the 90% credible interval $(\theta_{.05}, \theta_{.95})$ for π_u^* would miss to the right too often and to the left not often enough. This seems to be reasonably compelling evidence in support of our earlier intuition that π_u^* was indeed shifted a bit too far to the left. The posterior quantiles for π_s, on the other hand, seem much more balanced, yielding $P_\gamma(\alpha, \beta)$ closer to γ, and not having such a pronounced shift to the left. Thus all the evidence points to use of π_s as the noninformative prior for the problem.

REFERENCES

BERGER, J., 1985, *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

BERGER, J., 1987, "An Alternative: The Estimated Confidence Approach," Discussion of G. Casella's "Conditionally Acceptable Frequentist Solutions" in *Proceedings of the Fourth Purdue Symposium on Statistics*, S. S. Gupta and J. Berger (Eds.), New York: Springer-Verlag.

BERGER, J. and BERNARDO, J. M., 1987, "The Development of Reference Priors," Technical Report, Department of Statistics, Purdue University, West Lafayette, Indiana.

BERNARDO, J. M., 1979, "Reference Posterior Distributions for Bayesian Inference (with Discussion)," *Journal of the Royal Statistical Society B*, 41, 113–147.

EFRON, B., 1986, "Why Isn't Everyone a Bayesian ", *The American Statistician*, 40, 1–4.

JEFFREYS, H., 1961, *Theory of Probability (3rd edn.)*, London: Oxford University Press.

KIEFER, J., 1977, "Conditional Confidence Statements and Confidence Estimators (Theory and Methods)," *Journal of the American Statistical Association*, 72, 789–827.

SOUTHWOOD, T. 1978, *Ecological Methods With Particular Reference to the Study of Insect Populations*, London: Chapman and Hall.

STEIN, C., 1982, "On the Coverage Probability of Confidence Sets Based on a Prior Distribution", Technical Report No. 180, Department of Statistics, Stanford Univ., Stanford.

APPENDIX

I. Proof of Lemma 1.

Define $\omega = (\beta/\alpha)^{1/2}$, and change variables from (α, β) to (θ, ω). Note that $\alpha = \sqrt{\theta}/\omega$ and $\beta = \sqrt{\theta}\omega$. The Jacobian of the transformation is ω^{-1}, so that $\pi_{ij}(\alpha, \beta)$ transforms into

$$\begin{aligned}\tilde{\pi}_{ij}(\theta, \omega) &= \omega^{-1}[\left(\sqrt{\theta}\omega^{-1}\right)^2 + (\sqrt{\theta}\omega)^2]^{i/2}(\frac{\sqrt{\theta}}{\omega}\sqrt{\theta}\omega)^{-j} \\ &= \omega^{-1}\theta^{(\frac{1}{2}i-j)}[C(\omega)]^{i/2}.\end{aligned}$$

The likelihood becomes

$$\phi(x - \sqrt{\theta}\omega^{-1})\phi(y - \sqrt{\theta}\omega) = \phi\left(\frac{B}{\sqrt{C}}\right)\phi\left(\sqrt{\theta C} - \frac{A}{\sqrt{C}}\right),$$

so that the joint posterior density of (θ, ω) given (x, y) is proportional to

$$\theta^{(\frac{1}{2}i-j)}\omega^{-1}[C]^{i/2}\phi\left(\frac{B}{\sqrt{C}}\right)\phi\left(\sqrt{\theta C} - \frac{A}{\sqrt{C}}\right).$$

Integrating out over ω to find the marginal posterior for θ yields (3.2), subject to verification that $2H_{ij}(0|x, y)$ is the appropriate normalizing constant.

This last fact, together with (3.3) and (3.4), follow from the observation that

$$\int_t^\infty \theta^n \pi_{ij}(\theta|x,y)d\theta = \int_0^\infty \omega^{-1}C^{i/2}\phi\left(\frac{B}{\sqrt{C}}\right)\frac{g(\omega)}{2H_{ij}(0|x,y)}d\omega,$$

where

$$\begin{aligned} g(\omega) &= \int_t^\infty \theta^{(n+\frac{1}{2}i-j)}\phi\left(\sqrt{\theta C}-\frac{A}{\sqrt{C}}\right)d\theta \\ &= 2\int_{\sqrt{t}}^\infty \xi^{2n+i-2j+1}\phi\left(\xi\sqrt{C}-\frac{A}{\sqrt{C}}\right)d\xi \\ &= 2\{\psi_{2n+i-2j+1}(t,\omega)\phi(\sqrt{tC}-\frac{A}{\sqrt{C}})+\varphi_{2n+i-2j+1}(\omega)[1-\Phi(\sqrt{tC}-\frac{A}{\sqrt{C}})]\}; \end{aligned}$$

this last step can be verified by integration by parts, defining the ψ_m and φ_m by the recurrence relations

$$\begin{aligned} \psi_{m+1}(t,\omega) &= \frac{1}{C(\omega)}\left[t^{m/2}+A(\omega)\psi_m(t,\omega)+m\psi_{m-1}(t,\omega)\right] \\ \varphi_{m+1}(\omega) &= \frac{1}{C(\omega)}\left[A(\omega)\varphi_m(\omega)+m\varphi_{m-1}(\omega)\right], \end{aligned} \tag{A1}$$

the initializing functions being given before Lemma 1.

II. Proof of (4.11)

Separating the log term in (4.10) into [log θ + log $(\omega^2+\omega^{-2})$], it is clear that

$$\pi_i(\theta) = \frac{1}{\sqrt{\theta}}exp\{-\frac{1}{2}\sqrt{\theta}K_i(\theta)I_1^i(\theta)\},$$

where

$$I_1^i(\theta) = \int_{\sqrt{\theta}/l_i}^{k_i/\sqrt{\theta}} (1+\omega^{-4})^{1/2}\log(\omega^2+\omega^{-2})d\omega.$$

To evaluate this integral, break up the region of integration into the intervals $(\sqrt{\theta}/l_i,\varepsilon)$, $(\varepsilon,\varepsilon^{-1})$, and $(\varepsilon^{-1},\infty)$, where ε is small; call these integrals $I_{11}^i(\theta), R_1(\theta)$, and $I_{12}^i(\theta)$, respectively, observing that R_1 is a constant independent of k_i and l_i. Similarly, for

$$I_2^i(\theta) = \int_{\sqrt{\theta}/l_i}^{k_i/\sqrt{\theta}} (1+\omega^{-4})^{1/2}d\omega,$$

define $I_{21}^i(\theta), R_2(\theta)$, and $I_{22}^i(\theta)$.

Observe next that

$$(1+\omega^{-4})^{1/2}\log(\omega^2+\omega^{-2}) = \begin{cases} -2\omega^{-2}\log\omega + O(\varepsilon) & \text{on } (\frac{\sqrt{\theta}}{l_i},\varepsilon) \\ 2\log\omega + \omega^{-2}O(\varepsilon) & \text{on } (\varepsilon^{-1},\frac{k_i}{\sqrt{\theta}}), \end{cases}$$

$$(1+\omega^{-4})^{1/2} = \begin{cases} \omega^{-2}+O(\varepsilon) & \text{on } (\frac{\sqrt{\theta}}{l_i},\varepsilon) \\ 1+\omega^{-2}O(\varepsilon) & \text{on } (\varepsilon^{-1},\frac{k_i}{\sqrt{\theta}}). \end{cases}$$

Here we use $O(\varepsilon)$ to mean the absolute difference is no more than ε. Hence, letting R_j denote bounded functions of l_i or k_i,

$$\begin{aligned} I_{11}^i(\theta) &= \int_{\sqrt{\theta}/l_i}^{\varepsilon} -2\omega^{-2}\log\omega d\omega + O(\varepsilon) \\ &= R_3 - 2\frac{l_i}{\sqrt{\theta}}\left(\log\frac{\sqrt{\theta}}{l_i}+1\right) \\ I_{12}^i(\theta) &= \int_{\varepsilon^{-1}}^{k_i/\sqrt{\theta}} [2\log\omega + \omega^{-2}O(\varepsilon)]d\omega \\ &= R_4 + \frac{2k_i}{\sqrt{\theta}}(\log\frac{k_i}{\sqrt{\theta}}-1), \\ I_{21}^i(\theta) &= \int_{\sqrt{\theta}/l_i}^{\varepsilon} \omega^{-2}d\omega + O(\varepsilon) \\ &= R_5 + l_i/\sqrt{\theta}, \\ I_{22}^i(\theta) &= \int_{\varepsilon^{-1}}^{k_i/\sqrt{\theta}} [1+\omega^{-2}O(\varepsilon)]d\omega \\ &= R_6 + k_i/\sqrt{\theta}. \end{aligned}$$

Thus,

$$I_1^i(\theta) = R_7 + \frac{2}{\sqrt{\theta}}\{[l_i\ \log\ l_i + k_i\ \log\ k_i] - [(l_i+k_i)\ \log\ \sqrt{\theta} - k_i - l_i]\},$$

$$I_2^i(\theta) = R_8 + (l_i+k_i)/\sqrt{\theta}.$$

Hence

$$\begin{aligned} K_i(\theta) &= \frac{1}{\sqrt{\theta}}\ [R_8 + \frac{(l_i+k_i)}{\sqrt{\theta}}]^{-1} \qquad \text{(A2)} \\ &= (l_i+k_i)^{-1} + \sqrt{\theta}R_9(l_i+k_i)^{-2}, \end{aligned}$$

$$K_i(\theta)I_1^i(\theta) = \frac{2}{\sqrt{\theta}}\left\{\frac{[l_i\ (\log\ l_i+1)+k_i(\log\ k_i+1)]}{(l_i+k_i)} - \log\ \sqrt{\theta}\right\} + o(l_i+k_i). \tag{A3}$$

Thus

$$\frac{K_i(\theta)\pi_i(\theta)}{K_i(\theta_0)\pi_i(\theta_0)} = (1+o(l_i+k_i)), \tag{A4}$$

and the result follows.

III. Proof of Theorem 1.

Consider the curve defined by $\alpha\beta = \theta$ (θ consider fixed). It intersects the boundary of Λ_i in two points, that we refer to as $v_i^\theta = (\alpha_i^\theta, \theta/\alpha_i^\theta)$ and $w_i^\theta = (\theta/\beta_i^\theta, \beta_i^\theta)$. It is easy to see that

$$\Omega_{i,\theta} = \left(\left[\frac{\theta/\alpha_i^\theta}{\alpha_i^\theta}\right]^{1/2}, \left[\frac{\beta_i^\theta}{\theta/\beta_i^\theta}\right]^{1/2}\right) = \left(\frac{\sqrt{\theta}}{\alpha_i^\theta}, \frac{\beta_i^\theta}{\sqrt{\theta}}\right).$$

Observe, next, that

$$\lim_{i\to\infty}\frac{\alpha_i^\theta}{k_i} = \alpha_0 \quad \text{and} \quad \lim_{i\to\infty}\frac{\beta_i^\theta}{k_i} = \beta_0, \tag{A5}$$

where $(0,\alpha_0)$ and $(0,\beta_0)$ are the endpoints of the line segments formed by intersecting Λ with the respective axes. (This follows from continuity of the curve forming the boundary of Λ; $k_i^{-1}v_i^\theta$ and $k_i^{-1}w_i^\theta$ converge to $(0,\alpha_0)$ and $(0,\beta_0)$ along the curve.)

One now proceeds, as in the proof of (4.11), with l_i replaced by α_i^θ and k_i replaced by β_i^θ.

Then (A2), (A3), and (A4) become, using (A5),

$$\begin{aligned}
K_i(\theta) &= (\alpha_i^\theta + \beta_i^\theta)^{-1} + \sqrt{\theta} R_9 (\alpha_i^\theta + \beta_i^\theta)^{-2} \\
&= K_i^{-1}(\alpha_0 + \beta_0)^{-1}(1 + o(1)); \\
K_i(\theta) I_1^i(\theta) &= \frac{2}{\sqrt{\theta}} \left\{ \frac{[\alpha_i^\theta(\log \alpha_i^\theta + 1) + \beta_i^\theta(\log \beta_i^\theta + 1)]}{(\alpha_i^\theta + \beta_i^\theta)} - \log \sqrt{\theta} \right\} + o(\alpha_i^\theta + \beta_i^\theta) \\
&= \frac{2}{\sqrt{\theta}} \left\{ \frac{[\alpha_i^\theta(\log k_i + \log \alpha_0 + 1 + o(1)) + \beta_i^\theta(\log k_i + \log \beta_0 + 1 + o(1))]}{(\alpha_i^\theta + \beta_i^\theta)} \right. \\
&\qquad \left. - \log \sqrt{\theta} \right\} + o(k_i) \\
&= \frac{2}{\sqrt{\theta}} \left\{ (\log k_i + 1) + \frac{[k_i \alpha_0 (1 + o(1)) \log \alpha_0 + k_i \beta_0 (1 + o(1)) \log \beta_0]}{k_i(\alpha_0 + \beta_0)(1 + o(1))} \right. \\
&\qquad \left. - \log \sqrt{\theta} \right\} + o(1) \\
&= \frac{2}{\sqrt{\theta}} \left\{ \log k_i + 1 + \frac{(\alpha_0 \log \alpha_0 + \beta_0 \log \beta_0)}{(\alpha_0 + \beta_0)} - \log \sqrt{\theta} \right\} + o(1); \\
\frac{K_i(\theta)\pi_i(\theta)}{K_i(\theta_0)\pi_i(\theta_0)} &= (1 + o(1)).
\end{aligned}$$

This proves the result.