Statistical Science
2005, Vol. 20, No. 2, 141–162
DOI 10.1214/088342304000000279

From Unit Root to Stein's Estimator to Fisher's k Statistics: If You Have a Moment, I Can Tell You More

Xiao-Li Meng

This article is dedicated to my mentor and friend George Tiao on the occasion of his 70th birthday

Abstract. Any general textbook that discusses moment generating functions (MGFs) shows how to obtain a moment of positive-integer order via differentiation, although usually the presented examples are only illustrative, because the corresponding moments can be calculated in more direct ways. It is thus somewhat unfortunate that very few textbooks discuss the use of MGFs when it becomes the simplest, and sometimes the only, approach for analytic calculation and manipulation of moments. Such situations arise when we need to evaluate the moments of ratios and logarithms, two of the most common transformations in statistics. Such moments can be obtained by differentiating and *integrating* a joint MGF of the underlying untransformed random variables in appropriate ways. These techniques are examples of multivariate Laplace transform methods and can also be derived from the fact that moments of negative orders can be obtained by integrating an MGF. This article reviews, extends and corrects various results scattered in the literature on this joint-MGF approach, and provides four applications of independent interest to demonstrate its power and beauty. The first application, which motivated this article, is for the exact calculation of the moments of a well-known limiting distribution under the unit-root AR(1) model. The second, which builds on Stigler's Galtonian perspective, reveals a straightforward, non-Bayesian constructive derivation of the Stein estimator, as well as convenient expressions for studying its risk and bias. The third finds an exceedingly simple bound for the bias of a sample correlation from a bivariate normal population, namely the magnitude of the relative bias is not just of order n^{-1}, but actually is bounded above by n^{-1} for all sample sizes $n \geq 2$. The fourth tackles the otherwise intractable problem of studying the finite-sample optimal bridge in the context of bridge sampling for computing normalizing constants. A by-product of the joint-MGF approach is that positive-order fractional moments can be easily obtained from an MGF without invoking the concept of fractional differentiation, a method used by R. A. Fisher in his study of k statistics 45 years before it reappeared in the probability literature.

Xiao-Li Meng is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA (e-mail: meng@stat.harvard.edu).

Key words and phrases: AR(1) model, bias, bridge sampling, Efron–Morris estimator, fractional derivative, history of statistics, James–Stein estimator, Laplace transform, normalizing constants, R. A. Fisher, unit root, Wiener process.

PROLOGUE

Any book (e.g., Johnson, Kotz and Kemp, 1992) that discusses moment generating functions (MGFs) shows how to obtain a moment with positive integer order by differentiating an MGF. Specifically, suppose $M_X(t) = E[e^{tX}]$ exists in a neighborhood of $t = 0$. Then

$$E(X^k) = M_X^{(k)}(0), \tag{0.1}$$

where $M_X^{(k)}(t)$ denotes the kth derivative of $M_X(t)$. It is much less well known, however, that by *integrating* $M_X(t)$ in appropriate ways, we can obtain the expectation of $g(X)$ for a variety of choices of g. For example, if $g(x)$ is the Laplace transform of $h(t)$, that is,

$$g(x) = \int_0^\infty h(t)e^{-tx}\,dt, \tag{0.2}$$

then, when Fubini's theorem is applicable, we have

$$\begin{aligned} E[g(X)] &= \int_0^\infty h(t)E[e^{-tX}]\,dt \\ &= \int_0^\infty M_X(-t)h(t)\,dt. \end{aligned} \tag{0.3}$$

At first glance, (0.3) may not seem to be very useful. It simply replaces the direct integration, with respect to the probability measure defined by X, needed to evaluate $E[g(X)]$ analytically with another integration. As we all know, integration is generally much harder than differentiation. Could this be the reason that formulas such as (0.3) are almost never mentioned in any textbook that presents MGFs (but see Cressie and Borkent, 1986)? Putting it differently, is there any value for identity (0.3) and ones similar to it to be a part of our general textbook knowledge of MGFs?

Many statisticians perhaps never wonder or care about such a question, nor did I until 1996.

1. A MOTIVATING STORY: TAKING A MOMENT WITH UNIT ROOT

For a newcomer to time series analysis, as I was around 1996 when I was working with Professor George Tiao, "unit root" is often both a fascinating and frustrating topic, regardless of whether one's main interest is in application or in theory. From a practical point of view, models that involve unit roots are fascinating and important because they can capture phenomena with "phase transition" type characteristics (e.g., from stationarity to nonstationarity), to borrow a common phrase from physics. The various inferential implications of such models or even the existence of such transitions in particular studies are ever-debatable. See, for example, the entire theme issue of *Journal of Applied Econometrics* [1991, **6**(4)], which was devoted to a lively and vigorous debate on "Classical and Bayesian methods of testing for unit roots" and their implications for analyzing the gross national product of the United States from 1909 to 1970 and other economic time series. The very different conclusions reported by different articles in that issue highlight the frustration a practitioner may face. Because phase transition pushes nonrobustness to its extreme, the inferential conclusions are often frustratingly sensitive to the assumptions posited for either frequentist or Bayesian approaches. Furthermore, the actual analysis is often more complicated because of the nonapplicability of off-the-shelf methods, such as standard normal asymptotics.

This complication also fascinates those who are more theoretically oriented. Because normal asymptotics are so prevalent in general statistics, nonnormal asymptotics are much more intriguing to theoreticians. Indeed, it is generally said and believed that articles with nonnormal asymptotics have a substantially higher publication acceptance rate than those with normal asymptotics. However, the theoretical thrill does not come without frustration, even for tasks as basic as verifying whether an intuitively standardized random quantity indeed has mean 0 and variance 1. A well-known statistic in the literature for testing the unit-root AR(1) model (e.g., Dickey and Fuller, 1979) provides a perfect illustration of such frustration. This statistic was one of the key stepping stones in building popular testing procedures for unit roots in current practice (e.g., Elliott, Rothenberg and Stock, 1996).

Specifically, suppose we adopt the standard AR(1) model (with known variance for simplicity)

$$\begin{aligned} y_t &= \phi y_{t-1} + \varepsilon_t, \\ \varepsilon_t &\overset{\text{i.i.d.}}{\sim} N(0,1), \quad t = 1, 2, \ldots, n, \end{aligned} \tag{1.1}$$

with the convention that $y_0 = 0$, and we are interested in estimating and/or testing the model parameter ϕ. A common method is to use the standard least-squares estimator, which yields

$$\hat{\phi}_n = \frac{\sum_{t=0}^{n-1} y_t y_{t+1}}{\sum_{t=0}^{n-1} y_t^2}.$$

The analogy to standard linear regression also suggests how we might standardize $\hat{\phi}_n - \phi$ to arrive at an approximate confidence interval or hypothesis test. If this were the standard linear regression in the form of $y_t = \phi x_t + \varepsilon_t$, then the variance would be $(\sum_{t=1}^{n} x_t^2)^{-1}$. Since for (1.1), $x_t = y_{t-1}$, we are led to consider

$$\tau_n \triangleq \left(\sum_{t=0}^{n-1} y_t^2\right)^{1/2} (\hat{\phi}_n - \phi)$$

as our standardization of $\hat{\phi}_n$. However, although the analogy is suggestive, for indeed τ_n has the usual $N(0, 1)$ asymptotic distribution when $|\phi| < 1$, it would be a most unforgivable mistake in basic statistics or probability to rely on the analogy to assert that τ_n is asymptotically standardized in general.

Indeed, the asymptotic distribution of τ_n is no longer $N(0, 1)$ as soon as $|\phi| \geq 1$. The case of $\phi = 1$ has received the most attention in the literature, partly because of its implications in practice, because it models random walk phenomenon (e.g., in economic time series), and partly because it signifies unit-root type problems in general. [Indeed, for the Studentized statistic studied by White (1959), the asymptotic normality holds also for $|\phi| > 1$, making the case of $|\phi| = 1$ even more fascinating and unique.] When $\phi = 1$, it is well known that (see Chan and Wei, 1987; Tanaka, 1996, Chapter 3)

$$\begin{aligned} \tau_n &= \left(\sum_{t=0}^{n-1} y_t^2\right)^{1/2} (\hat{\phi}_n - \phi) \\ &\xrightarrow{\mathcal{D}} \frac{(W^2(1) - 1)/2}{[\int_0^1 W^2(t)\, dt]^{1/2}} \triangleq \tau, \end{aligned} \tag{1.2}$$

where $W(t)$ is the standard Wiener process on $[0, 1]$.

Clearly τ is not $N(0, 1)$, but it is conceivable that it might still have mean 0 and variance 1. Putting it differently, it is not immediate what its mean or variance should be just by inspecting its stochastic expression as given in (1.2); the dependence between the numerator and the denominator of τ complicates such a determination. Of course one can always resort to simulation, which would easily indicate that the mean of τ_n is in the vicinity of -0.4 when n becomes large enough. However, one would find that it is much more difficult to rule out $V(\tau) = 1$ by simulation, as we discuss shortly. Furthermore, when analytical calculations can be done relatively easily, there is no rational argument for not performing them, especially when simulations are used as an investigation tool. There is really no more reliable way to validate any simulation other then by checking its output against known analytic results.

The question then is whether there exists a relatively simple method for analytically calculating the moments of τ. The answer turns out to be a pleasant yes, although apparently this is not a generally well recognized fact, judging from my initial failure (during 1996–1997) to find the answer after a relatively extensive literature search and consultation in both statistics and econometrics. The search and research were most rewarding, because what they revealed was not just a neat trick for analytic calculation of the moments of τ, but rather a class of powerful tools for analytical evaluation of moments of ratios and logarithms, two of the most common transformations in statistics. The problem of calculating the moments of τ nicely illustrates the power of this class of methods.

Specifically, if we let $X = \frac{1}{2}(W^2(1) - 1)$ and $Y = \int_0^1 W^2(t)\, dt$, then $\tau = X/\sqrt{Y}$. White (1958) established that (with a minor correction by Abadir, 1993; also see Rao, 1978) the joint MGF for X and Y, $M_{X,Y}(t_1, t_2) = E(\exp(t_1 X + t_2 Y))$, is given by

$$\begin{aligned} &M_{X,Y}(-t_1, -t_2) \\ &\quad = \exp\left(\frac{t_1}{2}\right) \\ &\qquad \cdot \left[\cosh(\sqrt{2t_2}) + \frac{t_1}{\sqrt{2t_2}} \sinh(\sqrt{2t_2})\right]^{-1/2}, \\ &\qquad\qquad t_1 \in R,\ t_2 \geq 0. \end{aligned} \tag{1.3}$$

Since the joint MGF uniquely determines the distribution of (X, Y) (under regularity conditions), it also determines any moment of any $g(X, Y)$, which gives us hope of directly linking a moment to the joint MGF without first inverting the MGF to obtain the density. This is indeed possible for $g(X, Y) = X^k/Y^b$, where k is a nonnegative integer and b is arbitrary, as Lemma 1 of Section 2 asserts that (under very mild regularity conditions)

$$E\left(\frac{X^k}{Y^b}\right) = \frac{1}{\Gamma(b)} \int_0^\infty M_{X,Y}^{(k,0)}(0, -t) t^{b-1}\, dt, \tag{1.4}$$

where $M^{(k_1,k_2)}(t_1,t_2)$ denotes $(\partial^{k_1+k_2}M(t_1,t_2))/\partial t_1^{k_1}\partial t_2^{k_2}$. Applying (1.4) with (1.3) then yields straightforwardly

$$E(\tau) = -\frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{1}{\sqrt{\cosh(s)}}\left[1-\frac{\tanh(s)}{s}\right]ds = -0.42309564\ldots \tag{1.5}$$

and

$$E(\tau^2) = \frac{1}{4}\int_0^\infty \frac{1}{\sqrt{\cosh(s)}}\left[s+3\frac{\tanh^2(s)}{s}\right]ds - 1 = 1.14159507\ldots, \tag{1.6}$$

which implies that $V(\tau) = E\tau^2 - (E\tau)^2 = 0.96258515\ldots$. Therefore, not only $E(\tau)\neq 0$, but also $V(\tau)\neq 1$, although, somewhat intriguingly, $V(\tau)$ is much closer to 1 than $E(\tau)$ is to 0. [Why is that? Lin (2003) speculated that the negative correlation between the numerator and the denominator of τ attenuates its intended mean much more than its intended variance.] The closeness of $V(\tau)$ to 1 makes it much more difficult to determine $V(\tau)\neq 1$ via simulating the distribution of τ_n, because one can (and should) always wonder whether the observed difference is due to finite n, however large, even if the Monte Carlo error is of no concern. All reported numerical values and/or digits in (1.5)–(1.6) are obtained and confirmed with a number of different numerical integration routines, such as MATLAB and Maple. [Note intriguingly that $E(\tau^2)$ has its first six digits the same as $\pi-2$!] These numerical values are also consistent with the values reported in Gonzalo and Pitarakis (1998), $E(\tau) = -0.4231$ and $E(\tau^2) = 1.1417$, except that (1.6) implies that the last digit in their $E(\tau^2)$ should be 6, a reflection of rounding discrepancy between their numerical series-expansion evaluation and our numerical integration evaluation. Using the same integration approach, Gonzalo and Pitarakis (1998) also obtained moments for several other related statistics.

This article is the result of encouragement from all those who shared my joy over the simplicity of (1.5)–(1.6) [in contrast to the expression of τ in (1.2)] and who assured me that I was not alone in feeling that something is missing in our general textbook knowledge of the joint-MGF approach. This article is thus intended to narrow the gap, so others can spend less time and experience less frustration than I did dealing with similar analytic problems. Section 2 contains a number of variations and extensions of (1.4), as well as related theoretical and historical development. Sections 3, 4 and 5 present, respectively, three applications in the contexts of the Stein estimator, bias of a sample correlation and optimal bridge sampling; Section 6 traces the relevant history back to a work by R. A. Fisher in 1930 on k statistics. All four of these sections are intended to be self-contained (except for their common references to Section 2) and therefore can be read in any order. Section 5 is most technically involved because it tackles an open problem in a theoretical study of bridge sampling, so it should not be a part of any relaxing bedtime reading unless you have trouble falling asleep.

Finally, a disclaimer is necessary for any article that attempts to overview a topic of such diversity. The 76 listed references (and the references therein) are the result of literally months of search and research, vertically and horizontally, of the literature. Undoubtedly they are only a (nonrandom) sample of the articles that could be cited, in view of the enormous literature on Laplace transforms—because identities like (1.4) are part of the general multivariate Laplace transform techniques—and related topics in and outside of statistics and econometrics. I thus offer my apology to those whose relevant contribution is not given appropriate credit in this article, and express my gratitude to anyone who can further enrich my—as well as others'—knowledge on this topic.

2. THE JOINT-MGF APPROACH

2.1 Integrating a Moment Generating Function

As a direct analogy to the well-known formula (0.1), but with the differentiation operator replaced by the integration operator, one can verify that (e.g., Cressie, Davis, Folks and Policello, 1981) if $P(Y>0)=1$, then for any positive integer k,

$$E(Y^{-k}) = \int_0^\infty\cdots\int_0^\infty M_Y\left(-\sum_{i=1}^k t_i\right)dt_1\cdots dt_k. \tag{2.1}$$

This, however, is not a very useful formula because it involves multidimensional integration when $k>1$. [One can also successively integrate a probability generating function to obtain negative integer moments; see Chao and Strawderman (1972) and the references therein.]

A more useful formula is obtained by noticing that for any $y>0$ and $b>0$,

$$y^{-b} = \frac{1}{\Gamma(b)}\int_0^\infty t^{b-1}e^{-ty}\,dt, \tag{2.2}$$

that is, $g(y) = y^{-b}$ is the Laplace transform of $h(t) = t^{b-1}/\Gamma(b)$. Consequently, by identity (0.3), where Fubini's theorem is obviously applicable because the integrand here is nonnegative, we obtain

$$E(Y^{-b}) = \frac{1}{\Gamma(b)} \int_0^\infty t^{b-1} M_Y(-t)\, dt, \tag{2.3}$$

which not only avoids multidimensional integration, but more importantly handles negative fractional moments [obviously the right-hand side of (2.1) can be simplified into that of (2.3) by a change of variables via $s_1 = \sum_{i=1}^k t_i$ and $s_j = t_j$ for $j \geq 2$]. For integer b, identity (2.3) was given in Cressie, Davis, Folks and Policello (1981).

Combining (0.1) and (2.3), we obtain the following lemma, which is a more complete and rigorous formulation of several previous results (e.g., Sawa, 1972; Mehta and Swamy, 1978; Cressie, Davis, Folks and Policello, 1981). To simplify notation, we use P as a generic notation for probability measure of any (joint or marginal) random variable and use λ_d^+ for the Lebesgue measure on the d-dimensional product space $(0, \infty)^d$. To ensure the lemma is applicable as generally as possible, we adopt the notion of *quasi-integrability* of a function f, which only requires either $(f)^+$ or $(f)^-$ to be integrable, where $(\cdot)^+$ and $(\cdot)^-$ are the standard positive-part and negative-part functions, respectively. This relaxation on integrability makes verification of the conditions needed for (1.4) much easier, as in our motivating example (see below).

LEMMA 1. *Suppose k is a nonnegative integer and $b > 0$, $P(Y > 0) = 1$, $M_{X,Y}(t_1, 0)$ exists in a neighborhood of $t_1 = 0$, and X^k/Y^b is quasi-integrable with respect to P. Then $M_{X,Y}^{(k,0)}(0, -t_2)t_2^{b-1}$ is quasi-integrable with respect to λ_1^+ and the identity*

$$E\left(\frac{X^k}{Y^b}\right) = \frac{1}{\Gamma(b)} \int_0^\infty M_{X,Y}^{(k,0)}(0, -t)t^{b-1}\, dt \tag{2.4}$$

holds, where the values $\pm\infty$ are allowed.

Although (2.4) is not hard to verify formally, we provide a rigorous proof of Lemma 1 in Appendix A in view of some oversights in the literature concerning special cases of Lemma 1. Cressie, Davis, Folks and Policello (1981) considered the case when b is an integer and stated that (2.4) holds "when either integral exists," without assuming the quasi-integrability of X^k/Y^b. A simple example indicates that this assumption cannot be relaxed. Take $X = Z$, $Y = Z^2$ and $k = b = 1$ in (2.4), where $Z \sim N(0, 1)$. Then the left-hand side of (2.4) does not exist but the right-hand side is zero because $M_{Z,Z^2}^{(1,0)}(0, t_2) = 0$ for all $t_2 \geq 0$. The same error was made earlier in Sawa (1972), who considered Lemma 1 in the case when b and k are the same integer. Sawa's (1972) result has been extended and applied frequently in econometrics for evaluating moments of estimators of coefficients for various models, such as simultaneous equation models (e.g., Mehta and Swamy, 1978), dynamic regression models (e.g., Hoque, 1985; Peters, 1989) and many autoregressive or autoregressive integrated moving average (ARIMA) type models (e.g., Sawa, 1978; De Gooijer, 1980; Evans and Savin, 1981, 1984; Nankervis and Savin, 1988; Abadir and Larsson, 1996, 2001; Pitarakis, 1998). The error in Sawa's (1972) result was spotted by Mehta and Swamy (1978). However, due to the way they constructed their proof, Mehta and Swamy (1978, page 8) seemed to imply that the source of the error occurred in the interchange of integration with differentiation that leads to (A.2) in our proof given in Appendix A. Our proof shows that (A.2) holds in general and the error was in a subsequent interchange of integrals when no condition of Fubini's theorem was satisfied, as our simple counterexample shows.

Although not every statistician cares about regularity conditions as such, they are important in applications such as our motivating example. Our goal there was to compute the moments of τ, but we did not even know whether these moments exist or, at least, it is not obvious why they do. Lemma 1 provides a very effective way to determine the existence of any moment of positive-integer order of τ as well as its value. Specifically, for any integer $k > 0$, it is easy to derive from (1.3), by using the differentiation chain rule, that

$$\begin{aligned} M_{X,Y}^{(k,0)}(0, -t_2) &= \frac{(-1)^k}{2^k\sqrt{\cosh(\sqrt{2t_2})}} \\ &\quad \cdot \sum_{i=0}^k (-1)^i (2i-1)!! \binom{k}{i} \left[\frac{\tanh(\sqrt{2t_2})}{\sqrt{2t_2}}\right]^i, \end{aligned} \tag{2.5}$$

where $(2i-1)!! = (2i-1)(2i-3)\cdots 1$. Clearly, for k even, the conditions of Lemma 1 are satisfied because any fixed-sign random variable (i.e., $\tau^k \geq 0$) is quasi-integrable and thus (2.4) is applicable. For k odd, we have $|\tau|^k < 1 + \tau^{k+1}$ and thus τ^k is integrable since τ^{k+1} is integrable, because the right-hand side of (2.5)

is integrable over $t_2 \in (0, \infty)$ for any $k \geq 0$. Consequently, for any integer $k > 0$, $E(\tau^k)$ is finite and its value can be found via (2.4) with $b = k/2$. Letting

$$a_{i,k} = \int_0^\infty \frac{[\tanh(s)]^i}{\sqrt{\cosh(s)}} s^{k-i-1}\, ds,$$

then a simple change of variable $s = \sqrt{2t_2}$ yields

$$\begin{aligned} E(\tau^k) &= E\left(\frac{X^k}{Y^{k/2}}\right) \\ &= \frac{(-1)^k}{2^{(3k-2)/2}\Gamma(k/2)} \\ &\quad \cdot \sum_{i=0}^{k} (-1)^i (2i-1)!! \binom{k}{i} a_{i,k}, \end{aligned} \tag{2.6}$$

which gives (1.5)–(1.6) when $k = 1, 2$. For higher moments (e.g., those needed for exact skewness and kurtosis of τ) the following recursive formula, obtained via integration by parts, is useful for reducing the computational burden:

$$\begin{aligned} a_{i,k} &= \frac{2(k-i-1)}{2i-1} a_{i-1,k-2} \\ &\quad + \frac{2(i-1)}{2i-1} a_{i-2,k-2}, \quad 0 < i < k,\ k > 2, \end{aligned} \tag{2.7}$$

where $a_{i,k}$ is zero when $i < 0$.

Although one might not find the general expressions (2.6)–(2.7) in the published literature, the joint MGF approach was used in the literature to deal with similar moment calculations involving ratios. In addition to the aforementioned work by Gonzalo and Pitarakis (1998), Tanaka (1996, Chapter 1) used this approach to compute the moments of $\tau^* = \frac{1}{2}(W^2(1) - 1)/[\int_0^1 W^2(t)\, dt]$, and Nielsen (1997) applied the same approach to find the expansions of the moments of τ_n^2. In particular, (1.6) was given by Nielsen (1997) as the mean of τ^2. This is also an example of using the joint MGF method to find moments of a ratio of quadratic forms of normal variables, a class of problems we discuss further in Section 2.2.

2.2 Fractional Moments and Fractional Derivatives

In most of the literature mentioned in Section 2.1, b was restricted to be an integer. The extension to noninteger b is immediate since (2.3) holds for nonintegers as presented by Stuart and Ord (1987, page 101). This trivial extension turns out to be important because (i) it facilitates exact calculation of moments of "Studentization," $T_n = (\hat{\theta}_n - \theta)/\sqrt{\hat{V}_n}$ (in an obvious notation) for finite or infinite n, as in the unit-root problem, (ii) it suggests further useful generalizations such as those provided in Section 2.3 and (iii) by letting $Y = X$, it provides a formula for evaluating positive fractional moments $E(X^a)$, where $a > 0$ and $P(X \geq 0) = 1$. Cressie, Davis, Folks and Policello (1981) mentioned the possibility of using fractional derivatives to evaluate fractional moments via the MGF [i.e., a generalization of (2.1)] and Laue's (1980) work to connect fractional moments with fractional derivatives of a characteristic function. Indeed, such results were presented by Wolfe (1975) and, in fact, were used by Fisher in 1930 (see Section 6); similar results were also presented by Cressie and Borkent (1986) and were further discussed by Jones (1987a, b). The derivation provided below, as a simple consequence of (2.4), is more straightforward and appealing to researchers who are unfamiliar with the concept of fractional derivatives. The results are obviously equivalent, because both methods are used to establish the same identity. Indeed, a fractional derivative is defined in terms of integration; see Ross (1975) for an overview of fractional calculus.

Specifically, suppose $P(X > 0) = 1$, $M_X(t)$ exists in a neighborhood of $t = 0$ and a is a positive noninteger [X can be negative for certain choices of a (e.g., $1/3$), but we avoid such a complication here]. Let $\lceil a \rceil$ be the smallest integer that exceeds a and let $\langle a \rangle = \lceil a \rceil - a$. Now let $k = \lceil a \rceil$, $b = \langle a \rangle$ and $Y = X$ in (2.4). Since $M_{X,X}(t_1, t_2) = M_X(t_1 + t_2)$, Lemma 1 implies

$$\begin{aligned} E(X^a) &= E\left(\frac{X^{\lceil a \rceil}}{X^{\langle a \rangle}}\right) \\ &= \frac{1}{\Gamma(\langle a \rangle)} \int_0^\infty M_X^{(\lceil a \rceil)}(-t) t^{\langle a \rangle - 1}\, dt, \end{aligned} \tag{2.8}$$

and one side of (2.8) is finite if and only if the other side is. Note that (2.8) reduces to (0.1) when a is an integer, by taking $\langle a \rangle \to 0$ (this can be verified directly or by using fractional derivatives). We also note in passing that by letting $Y = X^{-1}$ and $b = a$ in (2.3), we can obtain $E(X^a)$ by integrating the MGF of X^{-1}:

$$E(X^a) = \frac{1}{\Gamma(a)} \int_0^\infty M_{X^{-1}}(-t)\, t^{a-1}\, dt. \tag{2.9}$$

This identity was used by Shepp and Lloyd (1966) in their study of limiting distributions of cycle lengths in a random permutation. For another discussion of the relationship between positive and negative moments, see Piegorsch and Casella (1985) and the accompanying comments, as well as Khuri and Casella (2002).

Historically, the derivation of (2.8) was presented by Mathai (1991) when X is a positive quadratic form in a (possibly singular) normal variable. Mathai's (1991) argument, however, is really general because the normality was used there only to justify interchange of integrals, as in deriving (2.3), which was also presented by Mathai (1991) under the normality setting. There also have been a considerable number of articles on calculating (integer) moments of ratios of quadratic forms in normal variables, and the joint MGF approach seems to be the most popular one; see, for example, Jones (1986), Morin (1992) and Tsui and Ali (1994) and references therein. Closely related work includes finding the exact distribution of a ratio via the inverse Mellin transform of its moments; see Provost and Rudiuk (1994) and references therein, and the book by Springer (1979). As we see here, the technique is useful in general as long as the required MGF is available. This was emphasized by Jones (1987b), who discussed the use of fractional derivatives in general for computing multivariate fractional moments.

As a simple illustration of the use of (2.8) in a nonnormal case, consider the stable distribution on $(0, \infty)$, which has the MGF (see Feller, 1971, pages 448–449)

$$M_X(-t) = \exp(-ct^\alpha), \quad c > 0,\ t > 0,\ 0 < \alpha < 1.$$

For $0 < a < 1$, we have $\lceil a \rceil = 1$ and $\langle a \rangle = 1 - a$, and thus by (2.8), after letting $s = ct^\alpha$,

$$E(X^a) = \frac{c^{a/\alpha}}{\Gamma(1-a)} \int_0^\infty e^{-s} s^{-a/\alpha}\, ds. \tag{2.10}$$

Since the right-hand side of (2.10) is infinite when $\alpha \le a < 1$, we conclude that $E(X^a) = \infty$ whenever $a \ge \alpha$. When $0 < a < \alpha$, the right-hand side of (2.10) is $c^{a/\alpha}\Gamma(1 - a/\alpha)/\Gamma(1-a)$, which is also the value of $E(X^a)$ when $a \le 0$, as can be verified directly by using (2.3). The same result was obtained by Wolfe (1975) via the fractional differentiation approach. Although fractional moments are often more of theoretical interest, there has been some work on using fractional moments in constructing estimators; see, for example, From and Saxena (1989) and the references cited there.

The identity (2.8) also reminds us that we can extend Lemma 1 to cases where the k in X^k is noninteger (again, assuming X is nonnegative, although this assumption can be relaxed). The price one has to pay for this generality is the need for double integrations, as seen in the following lemma, the proof of which is again deferred to Appendix A.

LEMMA 2. *Suppose a is a positive noninteger, $b > 0$ and $P(X \ge 0, Y > 0) = 1$. Then*

$$E\left(\frac{X^a}{Y^b}\right) = \frac{1}{\Gamma(\langle a \rangle)\Gamma(b)} \int_0^\infty \int_0^\infty M_{X,Y}^{(\lceil a \rceil, 0)}(-t_1, -t_2) \cdot t_1^{\langle a \rangle - 1} t_2^{b-1}\, dt_1\, dt_2 \tag{2.11}$$

and one side is finite if and only if the other side is.

Note that as a special case of Lemma 2 (i.e., with $Y = 1$), the condition $P(X > 0) = 1$ for (2.8) can be relaxed to $P(X \ge 0) = 1$, and the condition that $M_X(t)$ exists in a neighborhood of $t = 0$ is also not needed [but recall $M_X(t)$ always exists for $t \le 0$ when X is nonnegative]. We also note that if we do not insist on directly integrating the joint MGF, then we can express (2.11) in an equivalent form described by Evans and Savin (1981):

$$E\left(\frac{X^a}{Y^b}\right) = \frac{1}{\Gamma(b)} \int_0^\infty E[X^a e^{-tY}] t^{b-1}\, dt.$$

This result eliminates the double integration in (2.11), but, as a trade-off, it requires the expression of $E[X^a e^{-tY}]$, which may not be directly available to the investigator even if the joint MGF is. Of course, if one recognizes and uses the fact that $E[X^a e^{-tY}]$ can be obtained via differentiating and integrating the joint MGF, then effectively one is implementing (2.11).

2.3 Further Extensions and Variations

Lemmas 1 and 2 can be easily extended to more general identities that may be useful for analytic calculations of more complicated moments, such as multivariate moments (see, e.g., Jones, 1987b and Mathai, 1991). The following Theorem 1 is a general result that includes both Lemmas 1 and 2 as special cases, although its proof is essentially the same as that for Lemmas 1 and 2 (combined) but only with more complicated notation (hence the proof of Theorem 1 is omitted from Appendix A). It also covers a formula used by Davies, Pate and Petruccelli (1985) to find exact moments of the sample cross correlations of multivariate autoregressive moving average models. For simplicity, an operation applied to a vector means that it is applied componentwise [e.g., $\lceil \mathbf{a} \rceil \triangleq (\lceil a_1 \rceil, \lceil a_2 \rceil, \ldots, \lceil a_M \rceil)^\top$].

THEOREM 1. *Let* $\mathbf{X} = \{X_1, \ldots, X_L\}$, $\mathbf{Y} = \{Y_1, \ldots, Y_M\}$ *and* $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, *where* $P(\mathbf{Y} \geq 0, \mathbf{Z} > 0) = 1$. *Let* $\mathbf{k} = \{k_1, \ldots, k_L\}$ *be* L *nonnegative integers, let* $\mathbf{a} = \{a_1, \ldots, a_M\}$ *be* M *positive nonintegers and let* $\mathbf{b} = \{b_1, \ldots, b_N\}$ *be* N *positive numbers. Suppose* $\prod_{l=1}^{L} X_l^{k_l} \prod_{m=1}^{M} Y_m^{a_m} / \prod_{n=1}^{N} Z_n^{b_n}$ *is quasi-integrable with respect to* P *and that* $M_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\mathbf{t}_L, \mathbf{0}_M, \mathbf{0}_N)$ *exists in a neighborhood of* $\mathbf{t}_L = \mathbf{0}_L$, *where* $\mathbf{0}_D$ *denotes a* D*-dimensional vector of zeros. Then* $M_{\mathbf{X},\mathbf{Y},\mathbf{Z}}^{(\mathbf{k},\lceil\mathbf{a}\rceil,\mathbf{0}_N)}(\mathbf{0}_L, -\mathbf{u}, -\mathbf{v}) \prod_{m=1}^{M} u_m^{\langle a_m\rangle - 1} \prod_{n=1}^{N} v_n^{b_n - 1}$ *is quasi-integrable with respect to* λ_{M+N}^{+} *and the identity*

$$
\begin{aligned}
&E\left(\frac{\prod_{l=1}^{L} X_l^{k_l} \prod_{m=1}^{M} Y_m^{a_m}}{\prod_{n=1}^{N} Z_n^{b_n}}\right) \\
&\quad = \left\{\int_{\mathbf{0}_M}^{\infty_M} \int_{\mathbf{0}_N}^{\infty_N} \left[M_{\mathbf{X},\mathbf{Y},\mathbf{Z}}^{(\mathbf{k},\lceil\mathbf{a}\rceil,\mathbf{0}_N)}(\mathbf{0}_L, -\mathbf{u}, -\mathbf{v}) \right.\right. \\
&\qquad \left.\left. \cdot \prod_{m=1}^{M} u_m^{\langle a_m\rangle - 1} \prod_{n=1}^{N} v_n^{b_n - 1} \right] d\mathbf{u}\, d\mathbf{v} \right\} \\
&\qquad \cdot \left\{\prod_{m=1}^{M} \Gamma(\langle a_m\rangle) \prod_{n=1}^{N} \Gamma(b_n)\right\}^{-1}
\end{aligned} \tag{2.12}
$$

holds, where the values $\pm\infty$ *are allowed.*

The above extension is more or less obvious once we see the common patterns given by Lemmas 1 and 2. That is, any positive-integer moments are taken care of by differentiation and any pure positive fractional moments (i.e., after taking out the largest positive-integer moments) and negative moments are dealt with by integration. However, the extension to $W = \log(X/Y)$ takes a much less obvious form. The extension is possible because $M_W(t) = E[(X/Y)^t]$ and thus we can use Lemma 2 to connect the moments of W to $M_{X,Y}(t_1, t_2)$. The following theorem is a result of such a connection.

THEOREM 2. *For any* $t_1 > 0, t_2 > 0$ *and nonnegative integer* k, *let*

$$
\begin{aligned}
g_k(t_1, t_2) &= \left\{\frac{\partial^k}{\partial t^k}\left[\left(\frac{t_2}{t_1}\right)^t \frac{\sin(\pi t)}{\pi t}\right]\right\}_{t=0} \\
&= \sum_{j=0}^{[k/2]} \alpha_{j,k} (\log t_2 - \log t_1)^{k-2j},
\end{aligned} \tag{2.13}
$$

where $[k/2]$ *is the integer part of* $k/2$ *and* $\alpha_{j,k} = (-1)^j \pi^{2j} k!/[(2j+1)!(k-2j)!]$. *Suppose* $P(X > 0, Y > 0) = 1$. *Then* $(\log X - \log Y)^k$ *is quasi-integrable with respect to* P *if and only if* $M_{X,Y}^{(1,1)}(-t_1, -t_2) \cdot g_k(t_1, t_2)$ *is quasi-integrable with respect to* λ_2^{+}. *Furthermore,*

$$
\begin{aligned}
&E(\log X - \log Y)^k \\
&\quad = \int_0^\infty \int_0^\infty M_{X,Y}^{(1,1)}(-t_1, -t_2) g_k(t_1, t_2)\, dt_1\, dt_2.
\end{aligned} \tag{2.14}
$$

In particular,

$$
\begin{aligned}
&E(\log X) \\
&\quad = -\int_0^\infty M_X^{(1)}(-t) \log t\, dt - \gamma,
\end{aligned} \tag{2.15}
$$

$$
\begin{aligned}
&\operatorname{Var}(\log X) \\
&\quad = \int_0^\infty M_X^{(1)}(-t)(\log t)^2\, dt \\
&\qquad - \left(\int_0^\infty M_X^{(1)}(-t) \log t\, dt\right)^2 - \frac{\pi^2}{6}
\end{aligned} \tag{2.16}
$$

and

$$
\begin{aligned}
&E(\log X - \log Y)^2 \\
&\quad = \int_0^\infty \int_0^\infty M_{X,Y}^{(1,1)}(-t_1, -t_2) \\
&\qquad \cdot (\log t_2 - \log t_1)^2\, dt_1\, dt_2 - \frac{\pi^2}{3},
\end{aligned} \tag{2.17}
$$

where $\gamma = 0.57721566490\ldots$ *is Euler's constant. In all of these identities the value* ∞ *is allowed.*

The proof for Theorem 2, as given in Appendix A, is considerably more involved than that for Lemmas 1 and 2 (and hence Theorem 1), partially because the conclusion of Theorem 2 is stronger—(2.14) holds in the strongest possible sense, as given by the "if and only if" statement. This is possible because $M_{X,Y}^{(1,1)}(t_1, t_2)$ is nonnegative even though $\log(X/Y)$ does not have a fixed sign in general. This new result provides a way to tackle the otherwise intractable problem for determining the finite-sample optimal bridge in bridge sampling (Meng and Wong, 1996), as reported in Section 5.

3. A GALTONIAN CONSTRUCTION OF THE JAMES–STEIN ESTIMATOR

3.1 Stigler's Galtonian Perspective Revisited

Consider the well-known setting for shrinkage estimators, $x_i \overset{\text{indep}}{\sim} N(\theta_i, 1)$, $i = 1, \ldots, k$, where $\theta = (\theta_1, \ldots, \theta_k)^\top$ are unknown parameters. Perhaps the most startling and well-known discovery in classical sta-

tistics is that the obvious estimator of θ, $\mathbf{x} = (x_1, \ldots, x_k)^\top$, is uniformly dominated in terms of the composite quadratic risk $R(\theta, \hat{\theta}) = \sum_{i=1}^k E_\theta(\theta_i - \hat{\theta}_i)^2$ by any estimator of the form

$$\hat{\theta}_c = \left(1 - \frac{c}{\sum_{i=1}^k x_i^2}\right)\mathbf{x}, \quad 0 < c < 2(k-2), \tag{3.1}$$

when $k \geq 3$. Within this class, the best choice of c is $k-2$, which corresponds to the well-known James–Stein estimator (James and Stein, 1961). Since the discovery of the James–Stein estimator, there have been many interesting extensions; see, for example, Brandwein and Strawderman (1990). Also see Maatta and Casella (1990) for a related development concerning variance estimation.

The Bayesian explanation for (3.1) is almost immediate [see, e.g., Efron and Morris (1973)]. Among various non-Bayesian explanations of this "paradoxical" phenomenon, Stigler's (1990) Galtonian perspective is particularly appealing and insightful. Stigler argued that the "obvious" estimator $\hat{\theta}_0 = \mathbf{x}$ is inferior because it corresponds to regressing $\mathbf{x}$ on θ [since $E(\mathbf{x}|\theta) = \theta$], which is the "wrong" regression line when we want to predict θ from $\mathbf{x}$, which should use $E(\theta|\mathbf{x})$. Since we do not make any distributional assumption about θ, we would approximate $E(\theta|\mathbf{x})$, say, by $\beta\mathbf{x}$, which includes $\hat{\theta}_0 = \mathbf{x}$ as a special case. Stigler then invoked the idea of "data augmentation," namely, if we had the values of θ, then the best choice of β under the loss $L(\theta, \hat{\theta}) = \sum_{i=1}^k (\theta_i - \hat{\theta}_i)^2$ would be

$$\hat{\beta}_\theta = \frac{\sum_{i=1}^k \theta_i x_i}{\sum_{i=1}^k x_i^2}. \tag{3.2}$$

Since θ is unknown, we would like to estimate $\hat{\beta}_\theta$. Stigler noticed that if the numerator in (3.2) is replaced by its *unbiased* estimator $\sum_{i=1}^k x_i^2 - k$, then (3.2) is in the form of (3.1) with $c = k$. The need to have at least three points of (θ_i, x_i) is also quite intuitive from the Galtonian perspective, because with only two points the two regression lines, θ on $\mathbf{x}$ and $\mathbf{x}$ on θ, would be the same. The extra constraint of zero intercept appears to be compensated on average by the fact that $E(\theta_1 X_2 - \theta_2 X_1 | \theta_1, \theta_2) = 0$ for any θ_1 and θ_2.

There is a small disappointment in the argument above, since it did not lead to the best choice $c = k-2$. Furthermore, the choice $c = k$ satisfies $0 < c < 2(k-2)$ only when $k > 4$, a condition stronger than necessary. It turns out that this problem can be easily fixed if we seek an unbiased estimator for $\hat{\beta}_\theta$ of (3.2) itself, instead of just for its numerator. This is because

$$E_\theta\left(1 - \frac{k-2}{\sum_{i=1}^k x_i^2}\right) = E_\theta\left(\frac{\sum_{i=1}^k \theta_i x_i}{\sum_{i=1}^k x_i^2}\right) \quad \text{for any } \theta \tag{3.3}$$

when $k \geq 3$. Although Stigler (1990) did not use (3.3) to interpret this unbiasedness property of the James–Stein choice of c, he proved (3.3) and used it in an elegant proof of the fact that (3.1) dominates $\hat{\theta}_0 = \mathbf{x}$. Stigler provided two proofs of (3.3): one relies on an invariance argument and the other uses Stein's (1981) integration-by-parts formula concerning the normal distribution. Lemma 1 leads to a rather simple proof, but more importantly its use with the Galtonian perspective leads to the following straightforward constructive derivation of the James–Stein estimator, a derivation that reveals how naturally we first choose (3.1) as our candidate class and then impose $E_\theta(\hat{\beta}) = E_\theta(\hat{\beta}_\theta)$ to arrive at the optimal choice $c = k-2$.

3.2 A Non-Bayesian Constructive Derivation

Inspired by the Galtonian perspective, we seek $\hat{\beta}(\mathbf{x})$ such that $\hat{\theta} = \hat{\beta}(\mathbf{x})\mathbf{x}$ (strictly) dominates $\hat{\theta}_0 = \mathbf{x}$ in terms of $R(\theta, \hat{\theta})$. By taking expectations of both sides of the well-known regression decomposition $\sum(\theta_i - \beta x_i)^2 = (\beta - \hat{\beta}_\theta)^2 \sum x_i^2 + \sum(\theta_i - \hat{\beta}_\theta x_i)^2$, where $\hat{\beta}_\theta$ is given by (3.2), we have

$$R(\theta, \hat{\theta}) = E_\theta\left[(\hat{\beta}(\mathbf{x}) - \hat{\beta}_\theta)^2 \left(\sum_{i=1}^k x_i^2\right)\right] + \sum_{i=1}^k E_\theta(\theta_i - \hat{\beta}_\theta x_i)^2. \tag{3.4}$$

We thus only need to deal with the first term on the right-hand side of (3.4), which we denote by $D(\hat{\beta}(\mathbf{x})|\theta)$. At this moment, we (as non-Bayesian) have little idea what the form of $\hat{\beta}(\mathbf{x})$ might be, but it is intuitively clear that it is impossible to minimize $D(\hat{\beta}(\mathbf{x})|\theta)$ over all possible $\hat{\beta}(\mathbf{x})$ simultaneously for all θ. We thus restrict the class of candidates for $\hat{\beta}(\mathbf{x})$, and the simplest general class for $\hat{\beta}$ appears to be $\hat{\beta}(\mathbf{x}, \alpha)$, that is, a class indexed by a scalar quantity α. That is, much like reducing all possible models by parameterizing, the simplest type is a parametric family indexed by a scalar parameter α. We still have no idea what this class/family looks like, but if $\hat{\beta}(\mathbf{x}, \alpha)$ is a differentiable function of α, then seeking the optimal

α amounts to solving

$$\begin{aligned}&\frac{\partial}{\partial\alpha}D(\hat\beta(\mathbf{x},\alpha)|\theta)\\ &\quad=2E_\theta\left[(\hat\beta(\mathbf{x},\alpha)-\hat\beta_\theta)\left(\frac{\partial\hat\beta(\mathbf{x},\alpha)}{\partial\alpha}\sum_{i=1}^k x_i^2\right)\right]\\ &\quad=0\end{aligned}\tag{3.5}$$

simultaneously for all θ. Evidently, the easiest way to solve (3.5) is to set

$$\frac{\partial\hat\beta(\mathbf{x},\alpha)}{\partial\alpha}\left(\sum_{i=1}^k x_i^2\right)=\text{ constant}\tag{3.6}$$

and

$$E_\theta[\hat\beta(\mathbf{x},\alpha)-\hat\beta_\theta]=0\quad\text{for all }\theta.\tag{3.7}$$

The differential equation (3.6) immediately suggests that

$$\hat\beta(\mathbf{x},\alpha)=c_0+\frac{c_1\alpha}{\sum_{i=1}^k x_i^2},\tag{3.8}$$

and the requirement that $\hat\beta=1$ belong to this class (to ensure that the optimal estimator we find dominates $\hat\theta_0=\mathbf{x}$) sets $c_0=1$. We can rewrite $c_1\alpha$ as α since α is arbitrary, and then the unbiasedness requirement (3.7) is equivalent to determining α_0 such that

$$E_\theta\left(\frac{\sum_{i=1}^k\theta_i x_i-\alpha_0}{\sum_{i=1}^k x_i^2}\right)=1\quad\text{for all }\theta.\tag{3.9}$$

Once such an α_0 is found, then (3.4) indeed is minimized by $\hat\theta=\hat\beta(\mathbf{x},\alpha_0)\mathbf{x}$ because $D(\hat\beta(\mathbf{x},\alpha)|\theta)$ is a convex quadratic function of α for $\hat\beta(\mathbf{x},\alpha)$ given by (3.8). Lemma 1 is a very handy tool for searching for (instead of proving) such an α_0. Note that by the Cauchy–Schwarz inequality, the random variable in (3.9) is bounded above by $\|\theta\|^2+|\alpha_0|(\sum x_i^2)^{-1}$ and thus the left-hand side of (3.9) exists when $k>2$. Therefore, Lemma 1 is applicable as long as $k\ge3$.

To apply Lemma 1, we let $X=\sum_{i=1}^k\xi_i x_i-\alpha_0$ and $Y=\sum_{i=1}^k x_i^2$, and we use general $\xi=(\xi_1,\ldots,\xi_k)^\top$ in X instead of θ because of the bias calculations we discuss later. Then the joint MGF of (X,Y) is easily obtained by forming an appropriate quadratic form for each x_i, $i=1,\ldots,k$, as

$$\begin{aligned}&M_{X,Y}(t_1,t_2)\\ &\quad=(1-2t_2)^{-k/2}\\ &\qquad\cdot\exp\left[-t_1\alpha_0+\frac{\|\xi\|^2t_1^2+2\xi^\top\theta t_1+2\|\theta\|^2t_2}{2(1-2t_2)}\right].\end{aligned}$$

Since $M_{X,Y}(t_1,0)$ exists for any t_1 and

$$\begin{aligned}&M_{X,Y}^{(1,0)}(0,-t_2)\\ &\quad=(1+2t_2)^{-k/2}\left[\frac{\xi^\top\theta}{1+2t_2}-\alpha_0\right]\exp\left(-\frac{\|\theta\|^2t_2}{1+2t_2}\right),\end{aligned}$$

by (2.4), after letting $s=(1+2t_2)^{-1}$ and performing an integration by parts, we obtain

$$\begin{aligned}&E\left(\frac{\sum_{i=1}^k\xi_i x_i-\alpha_0}{\sum_{i=1}^k x_i^2}\right)\\ &\quad=\frac{\xi^\top\theta}{\|\theta\|^2}-\frac{1}{2}\left[(k-2)\frac{\xi^\top\theta}{\|\theta\|^2}+\alpha_0\right]\gamma_{k-2}(\|\theta\|),\end{aligned}\tag{3.10}$$

where

$$\gamma_m(x)=\int_0^1 s^{(m-2)/2}\exp\left(\frac{x^2}{2}(s-1)\right)ds.\tag{3.11}$$

Therefore, when $\xi=\theta$,

$$\begin{aligned}&E\left(\frac{\sum_{i=1}^k\theta_i x_i-\alpha_0}{\sum_{i=1}^k x_i^2}\right)\\ &\quad=1-\frac{\gamma_{k-2}(\|\theta\|)}{2}[(k-2)+\alpha_0]\end{aligned}\tag{3.12}$$

and thus (3.9) holds if and only if $\alpha_0=-(k-2)$, the James–Stein choice.

Similar arguments can be used for the Efron–Morris (1973) estimator

$$\hat\theta_c^{\mathrm{EM}}=\bar x\mathbf{1}_{k\times1}+\left(1-\frac{c}{\sum_{i=1}^k(x_i-\bar x)^2}\right)(\mathbf{x}-\bar x\mathbf{1}_{k\times1}),\qquad 0<c<2(k-3),$$

where $\bar x$ is the average of $\{x_1,\ldots,x_n\}$ and $\mathbf{1}_{k\times1}=(1,\ldots,1)^\top$. In particular, the best choice of $c=k-3$ corresponds to the unbiased estimator of the regression slope based on the "augmented data," that is,

$$\begin{aligned}&E\left(1-\frac{k-3}{\sum_{i=1}^k(x_i-\bar x)^2}\right)\\ &\quad=E\left(\frac{\sum_{i=1}^k(x_i-\bar x)(\theta_i-\bar\theta)}{\sum_{i=1}^k(x_i-\bar x)^2}\right)\end{aligned}$$

when $k\ge4$. [It is easy to verify that, after a rotation transformation, the above expression is the same as (3.3) with k replaced by $k-1$.]

3.3 Convenient Expressions for Studying Risk and Bias of $\hat{\theta}_c$

The identity (3.12) also provides a trivial way to calculate the risk of $\hat{\theta}_c$ in (3.1). Because

$$\sum_{i=1}^{k}(\hat{\theta}_i - \theta_i)^2 = \sum_{i=1}^{k}(x_i - \theta_i)^2 - 2c + 2c\left(\frac{\sum_{i=1}^{k}\theta_i x_i + c/2}{\sum_{i=1}^{k} x_i^2}\right),$$

setting $\alpha_0 = -c/2$ in (3.12) yields

$$R(\theta, \hat{\theta}_c) = k - \frac{c}{2}[2(k-2) - c]\gamma_{k-2}(\|\theta\|). \tag{3.13}$$

Because $\gamma_{k-2}(\|\theta\|) > 0$ for any θ, if we only look for a nonconstructive proof, then (3.13) is all we need because it trivially implies that (1) $\hat{\theta}_c$ dominates $\hat{\theta}_0$ if and only if $0 < c < 2(k-2)$ and (2) $R(\theta, \hat{\theta}_c)$ is minimized when $c = k - 2$. Among rigorous proofs of Stein's paradox (e.g., as discussed in Stigler, 1990), this perhaps is the most elementary one given (3.13), which itself is a straightforward application of Lemma 1.

To calculate the bias of $\hat{\theta}_c$, we need the more general identity (3.10). Specifically, by setting $\alpha_0 = 0$, $\xi_j = 1$ and $\xi_i = 0$ for $i \neq j$ in (3.10) in turn for each j, we obtain

$$\text{Bias}(\hat{\theta}_c) \triangleq E(\hat{\theta}_c) - \theta = -\frac{c\gamma_k(\|\theta\|)}{2}\theta. \tag{3.14}$$

In deriving (3.14) we have used the simple recursive relationship

$$\gamma_m(x) = x^{-2}[2 - (m-2)\gamma_{m-2}(x)], \quad m \geq 3, \tag{3.15}$$

which is obtained via integration by parts.

Although there are many other ways to calculate risk or bias of $\hat{\theta}_c$ (e.g., using series expansions as in Bock, Judge and Yancey, 1984), expressions (3.13) and (3.14) are particularly convenient for certain theoretical derivations [as well as for numerical evaluation because of (3.15)] by taking advantage of the known properties of the γ function of (3.11). To better facilitate our discussion, we first "standardize" the index c in $\hat{\theta}_c$ via $b = [c - (k-2)]/(k-2)$ and, accordingly, with a slight abuse of notation, rewrite $\hat{\theta}_c$ as $\hat{\theta}_b$. This reindexing removes the dependence of c on k, that is, the region $0 < c < 2(k-2)$ is transformed into $-1 < b < 1$, where $b = 0$ indexes the James–Stein estimator and $b = -1$ is the maximum likelihood estimator (MLE). More importantly, it explicitly displays the symmetry of the risk as a function of b (for fixed $\|\theta\|$):

$$R(\theta, \hat{\theta}_b) = k - \tfrac{1}{2}(k-2)^2(1-b^2)\gamma_{k-2}(\|\theta\|). \tag{3.16}$$

In contrast, $\|\text{Bias}(\hat{\theta}_b)\|^2$ is a strictly increasing function of $b \in [-1, 1]$ for fixed $\|\theta\|$ because (3.14) implies

$$\|\text{Bias}(\hat{\theta}_b)\|^2 = \tfrac{1}{4}(k-2)^2(1+b)^2\|\theta\|^2\gamma_k^2(\|\theta\|). \tag{3.17}$$

Comparing (3.16) with (3.17) reveals some interesting features of $\hat{\theta}_b$ for $b \in [-1, 1]$. For example, we observe that for any $0 \leq b \leq 1$, $\hat{\theta}_{-b}$ and $\hat{\theta}_b$ have the identical maximal risk among $\hat{\Theta}_b \triangleq \{\hat{\theta}_\beta, |\beta| \leq b\}$, yet $\hat{\theta}_{-b}$ has the least bias (in terms of its magnitude), whereas $\hat{\theta}_b$ has the maximum bias in the same class $\hat{\Theta}_b$. In particular, $\hat{\theta}_1$ can be viewed as the "worst" estimator within the entire class $\{\hat{\theta}_b, |b| \leq 1\}$ because it has the largest risk as well as the largest bias. This is directly visible from Figures 1 and 2, which plot the risk surface and the bias-squared surface, respectively, for $k = 3$ and $k = 4$.

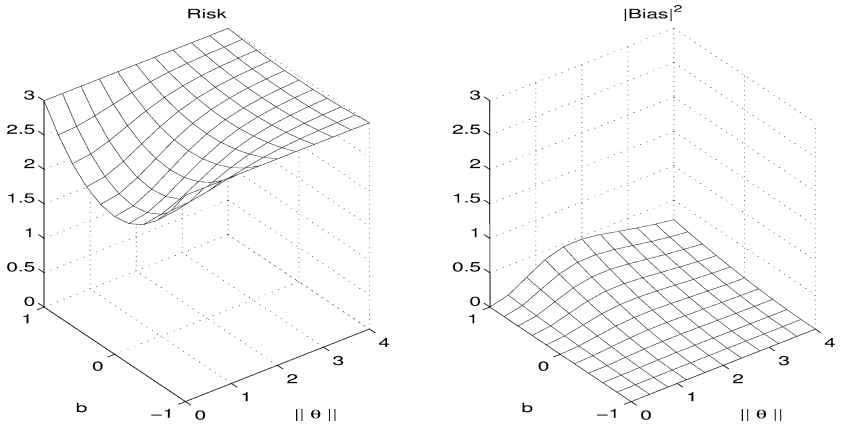

FIG. 1. *Risk and bias-squared surfaces of $\hat{\theta}_b$ when $k = 3$.*

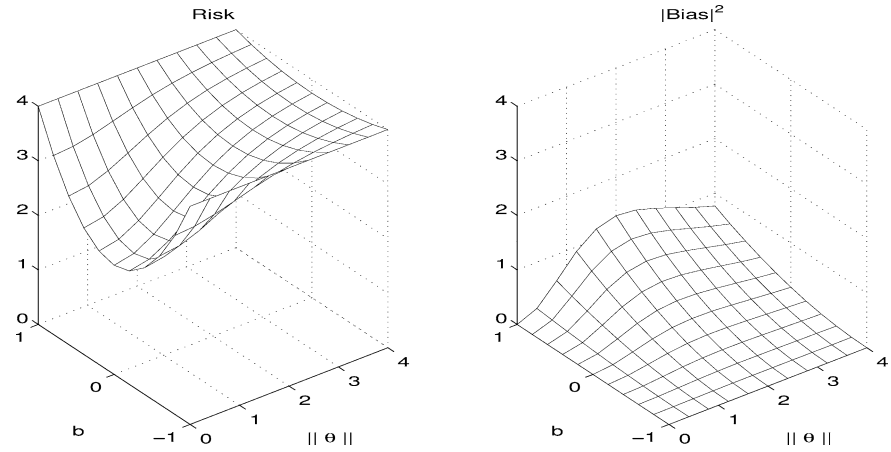

FIG. 2. *Risk and bias-squared surfaces of $\hat{\theta}_b$ when $k = 4$.*

The figures also reveal that for fixed b, the risk is a monotone increasing function as $\|\theta\|$ moves from zero to infinity. This can be easily proved via (3.16) because $\gamma_m(x)$ is a strictly decreasing function of x with $\gamma_m(0) = 2/m$ and $\gamma_m(\infty) = 0$. Consequently, for any $-1 \le b \le 1$, the risk of $\hat{\theta}_b$ increases strictly from $2 + b^2(k-2)$ to k as $\|\theta\|$ moves away from the origin to infinity. This, of course, confirms our intuition that as the true θ moves away from zero, the shrinkage factor in (3.1) becomes closer and closer (stochastically) to 1 and, therefore, $\hat{\theta}_b$ should have increased risk until it reaches the maximal value k, because $\hat{\theta}_b$ behaves more and more like the MLE $\mathbf{x}$.

The same intuition also suggests that the bias of $\hat{\theta}_b$ decreases as the underlying θ moves away from zero since the MLE $\mathbf{x}$ is unbiased for θ. It is evident from the figures, however, that $\|\operatorname{Bias}(\hat{\theta}_b)\|^2$ is not a monotone function of $\|\theta\|$ for fixed b. Indeed, it is easy to verify from (3.17) that $\|\operatorname{Bias}(\hat{\theta}_b)\|$ approaches zero when $\|\theta\|$ approaches either zero or infinity and reaches its maximum when $\|\theta\| = z_k$, the solution of $\gamma_k(z) = 2/(z^2 + k - 1)$. [Why the absolute bias $\|\operatorname{Bias}(\hat{\theta}_b)\|$ reaches the maximum at this particular value is a theoretical curiosity for which an intuitive explanation is yet to be found.] This does not imply that our intuition is wrong. It actually is accurate: What went wrong was the inappropriate mathematical formulation of our intuition. The correct formulation is to use *relative bias*, relative to the size of the estimand θ, which indeed in general is a more meaningful measure of bias. It is then trivial to see from (3.17) that the relative magnitude of the bias $\|\operatorname{Bias}(\hat{\theta}_c)\|/\|\theta\| = (k-2)(1+b)\gamma_k(\|\theta\|)/2$ monotonically decreases to zero with the increase of $\|\theta\|$. Furthermore, because

$$\sup_{\theta \in R^k} \frac{\|\operatorname{Bias}(\hat{\theta}_b)\|}{\|\theta\|} = \lim_{\|\theta\| \to 0} \frac{\|\operatorname{Bias}(\hat{\theta}_b)\|}{\|\theta\|} = \frac{c}{2}\gamma_k(0) = \left(1 - \frac{2}{k}\right)(1+b),$$

we learn that the maximal magnitude of the relative bias occurs at the origin. For the James–Stein estimator, this maximal relative bias is $1 - 2k^{-1}$, which monotonically increases with k but is bounded above by 1.

As a further illustration of the utilities of (3.16)–(3.17), because $\lim_{x\to\infty} x^2\gamma_m(x)/2 = 1$ for any $m \ge 1$ [for $m \ge 3$, this is a consequence of (3.15) and for $m = 1, 2$, it can be verified directly], we can easily derive the rate at which the reduction in risk (compared to the MLE) and $\|\operatorname{Bias}(\hat{\theta}_b)\|^2$ approach zero as $\|\theta\| \to \infty$:

$$k - R(\theta, \hat{\theta}_b) \asymp \frac{(k-2)^2(1-b^2)}{\|\theta\|^2}, \qquad \|\operatorname{Bias}(\hat{\theta}_b)\|^2 \asymp \frac{(k-2)^2(1+b)^2}{\|\theta\|^2}. \tag{3.18}$$

These "Cauchy density" type tails are quite visible in Figures 1 and 2 (for the risk surfaces, when viewed upper side down). Furthermore, because $\gamma_m(x) \le 2/x^2$ for all $m \ge 2$ [again a consequence of (3.15) when $m \ge 3$ and directly verifiable for $m = 2$, but not for $m = 1$], the right-hand side of each equivalent relationship in (3.18) also serves as a sharp upper bound of the corresponding left-hand side, except for the case of $k - R(\theta, \hat{\theta}_b)$ when $k = 3$ [because $\gamma_1(x) > 2/x^2$ when $x \ge x_0$, the solution of $\gamma_1(x) = 2/(x^2 - 1)$].

4. BOUNDING THE BIAS IN A SAMPLE CORRELATION

As a simple illustration of the usefulness of Theorem 1, going beyond bivariate MGFs as covered by Lemmas 1 and 2, consider the problem of estimating a population correlation ρ by a sample correlation r_n, where n indexes the sample size of a simple random sample. It is well known that, unlike a sample mean or a sample variance, r_n is generally a biased estimator of ρ, and the exact distribution of r_n is generally very complicated, even when the underlying joint distribution is bivariate normal. Indeed, assuming joint normality, Fisher (1915) via an elegant geometric argument found the density function of r_n to be

$$f(r) = \frac{(1-\rho^2)^{(n-1)/2}}{\pi\Gamma(n-2)}(1-r^2)^{(n-4)/2} \cdot \frac{d^{n-2}}{d(\rho r)^{n-2}}\left\{\frac{\arccos(-\rho r)}{\sqrt{1-\rho^2 r^2}}\right\}, \quad -1 \le r \le 1. \tag{4.1}$$

It is clear that $f(r)$ is not particularly easy to manipulate analytically, for example, for finding the mean of r_n. By expressing (4.1) in terms of a hypergeometric function, we can express $E(r_n)$ via a hypergeometric series (e.g., Hotelling, 1953; Stuart and Ord, 1987, pages 529–531),

$$E(r_n) = \frac{\rho\Gamma^2(\frac{1}{2}n)}{\Gamma\{\frac{1}{2}(n-1)\}\Gamma\{\frac{1}{2}(n+1)\}} \cdot F\left(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n+1), \rho^2\right), \tag{4.2}$$

where

$$F(\alpha,\beta,\gamma,z) = 1 + \frac{\alpha\cdot\beta}{\gamma\cdot 1}z + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)\cdot 1\cdot 2}z^2 + \cdots.$$

(Higher-order moments can also be expressed via hypergeometric series, as in Johnson, Kotz and Balakrishnan, 1995, Chapter 32.) We now show that with the help of Theorem 1, we can find an integral representation of $E(r_n)$ from which we can easily derive some useful bounds for the relative bias $(E(r_n)-\rho)/\rho$ without having to deal with (4.1) or (4.2), and the results do not rely on large-sample approximations.

BOUND I. *Suppose (x_i, y_i), $i = 1, \ldots, n$, are a simple random sample from a bivariate normal distribution with population correlation ρ. Then, for any $n \ge 2$,*

$$-\frac{1}{n} < \frac{E(r_n)-\rho}{\rho} < 0 \quad \textit{when } 0 < \rho^2 < 1, \qquad E(r_n) = \rho \quad \textit{when } p = 0, \pm 1. \tag{4.3}$$

PROOF. As detailed in Appendix B, applying Theorem 1 to the current setting yields

$$E(r_n) = \frac{2(n-1)\rho}{\pi} \cdot \int_0^\infty\int_0^{\pi/2}\left[(1-\rho^2)\sin^2(2\theta)r^2 + 2r + 1\right]^{-(n+1)/2} d\theta\, dr. \tag{4.4}$$

When $0 < \rho^2 < 1$, since for $r > 0$ and $0 < \theta < \pi/2$,

$$2r+1 < (1-\rho^2)\sin^2(2\theta)r^2 + 2r + 1 < (r+1)^2,$$

we obtain from (4.4) that

$$\left(1-\frac{1}{n}\right) = (n-1)\int_0^\infty \frac{dr}{(1+r)^{n+1}} < \frac{E(r_n)}{\rho} < (n-1)\int_0^\infty \frac{dr}{(1+2r)^{(n+1)/2}} = 1$$

and hence (4.3) [$E(r_n) = \rho$ when $\rho^2 = 0$ or 1 is obvious]. □

Bound I is handy to use in practice, since it says that for $1 > \rho > 0$, r_n underestimates ρ, and for $-1 < \rho < 0$, r_n overestimates ρ, but the absolute relative bias [i.e., $|E(r_n) - \rho|/|\rho|$] never exceeds $1/n$ for any n (≥ 2). Thus the finite-sample bias is usually of no practical concern as long as n is not too small (e.g., $n \ge 10$). [Nevertheless, an unbiased estimate of ρ, in the form of $r_n F(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}(n-1), 1-r_n^2)$, where $F(\alpha,\beta,\gamma,z)$ is the hypergeometric series used in (4.2), was suggested by Olkin and Pratt (1958).] Whether this exceedingly neat bound also holds for some other distributions is a question of both theoretical and practical interest.

Although bound (4.3) is sufficient and appealing for most practical purposes, it can be improved upon if one is willing to carry out a few more algebraic steps, as illustrated below.

BOUND II. *Assuming the same condition as in Bound* I, *for any $n \ge 4$,*

$$-\frac{1}{n-3}\frac{1-\rho^2}{1+\rho^2} < \frac{E(r_n)-\rho}{\rho} < 0 \quad \textit{when } 0 < \rho^2 < 1. \tag{4.5}$$

PROOF. Since $(a+bt)^{-(n+1)/2}$ is a convex function of t when a and b are positive, by the Jensen inequality, we obtain from (4.4) that

$$\begin{aligned}\frac{E(r_n)}{\rho} &> (n-1)\\ &\quad\cdot\int_0^\infty\Bigg[(1-\rho^2)r^2\\ &\qquad\cdot\left(\frac{2}{\pi}\int_0^{\pi/2}\sin^2(2\theta)\,d\theta\right)\\ &\qquad+2r+1\Bigg]^{-(n+1)/2}dr \qquad (4.6)\\ &= (n-1)\\ &\quad\cdot\int_0^\infty\left[\frac{(1-\rho^2)}{2}r^2+2r+1\right]^{-(n+1)/2}dr.\end{aligned}$$

Let $R(t)=a+bt+ct^2$ and $\Delta=4ac-b^2$. Then

$$\begin{aligned}&\int[R(t)]^{-(n+1)/2}\,dt\\ &\quad=\frac{2(2ct+b)[R(t)]^{-(n-1)/2}}{(n-1)\Delta} \qquad (4.7)\\ &\qquad+\frac{4(n-2)c}{(n-1)\Delta}\int[R(t)]^{-(n-1)/2}\,dt.\end{aligned}$$

Applying (4.7) to (4.6) yields, when $n\ge 4$,

$$\begin{aligned}\frac{E(r_n)}{\rho} &> \frac{2}{1+\rho^2}\\ &\quad-\frac{1-\rho^2}{1+\rho^2}(n-2)\\ &\qquad\cdot\int_0^\infty\left[\frac{(1-\rho^2)}{2}r^2+2r+1\right]^{-(n-1)/2}dr\\ &> \frac{2}{1+\rho^2}\\ &\quad-\frac{1-\rho^2}{1+\rho^2}(n-2)\int_0^\infty(1+2r)^{-(n-1)/2}\,dr\\ &= 1-\frac{1-\rho^2}{1+\rho^2}\frac{1}{n-3},\end{aligned}$$

which proves (4.5). □

Compared to Bound I, Bound II is closer (for large n) to the asymptotic expansion of (4.2), which is (see Stuart and Ord, 1987, page 531)

$$\frac{E(r_n)-\rho}{\rho}=-\frac{(1-\rho^2)}{2n}+O\left(\frac{1}{n^2}\right),\quad \rho\neq 0.$$

If desired, one can repeatedly use (4.7) to improve the bound in (4.5). Of course, the bounds become more and more complicated and thus lose their practical value. However, the derivation of (4.5) (as well as its further refinement) demonstrates that Theorem 1 can lead to rather accurate bounds without ever invoking large-sample arguments, illustrating the potential usefulness of Theorem 1 (and its various special cases and extensions) in finite-sample theoretical investigations.

5. SEARCHING FOR FINITE-SAMPLE OPTIMAL BRIDGE

Bridge sampling is a generalization of importance sampling for simulating (ratios of) normalizing constants of probability models. Computing normalizing constants is a common computational problem in statistics as well as in other fields such as physics and genetics. The basic setting for bridge sampling is easy to describe, yet the problems to which it is applicable can be exceedingly complex (e.g., computing exchange frequencies in quantum crystals; see Ceperley, 1995, pages 341–343). Indeed, the method originated in computational physics for computing free-energy differences, a problem that essentially defeats the standard importance sampling technique (see Bennett, 1976). For recent theoretical and empirical studies of bridge sampling and closely related methods, refer to Meng and Wong (1996), Meng and Schilling (1996, 2002), DiCiccio, Kass, Raftery and Wasserman (1997), Gelman and Meng (1998) and Kong, McCullagh, Meng, Nicolae and Tan (2003). We discuss only material that is directly related to our current topic and we ignore all regularity conditions.

Suppose we have two densities $p_i(\omega)$, $i=0,1$, with respect to a common measure $\mu(\omega)$. We can evaluate $p_i(\omega)$ up to a normalizing constant c_i: $p_i(\omega)=q_i(\omega)/c_i$, $i=0,1$. We also have draws $\{\omega_{ij}, j=1,\ldots,n_i\}$ from $p_i(\omega)$. The powerful Markov chain Monte Carlo (e.g., Metropolis algorithm) allows us to simulate from densities with unknown normalizing constants; here we assume draws from p_0 are independent of draws from p_1. Our goal here is to use these draws to estimate $r=c_1/c_0$. The bridge sampling relies on the following simple identity to construct estimators for r. For simplicity, suppose p_0 and p_1 share a common support Ω [but see Voter (1985) and Meng

and Schilling (2002) when this assumption fails], and suppose $\alpha(\omega)$ is a nonnegative function on Ω such that $0 < \int \alpha(\omega) p_0(\omega) p_1(\omega) \mu(d\omega) < \infty$. Then it is trivial to verify that

$$\frac{c_1}{c_0} = \frac{E_0[q_1(\omega)\alpha(\omega)]}{E_1[q_0(\omega)\alpha(\omega)]}, \tag{5.1}$$

where E_i denotes the expectation with respect to $p_i(\omega)$, $i = 0, 1$. Thus, given α and the draws $\{\omega_{ij}, j = 1, \ldots, n_i; i = 0, 1\}$, a simulation-consistent estimate of r is

$$\hat{r}_\alpha = \frac{(1/n_0)\sum_{j=1}^{n_0} q_1(\omega_{0j})\alpha(\omega_{0j})}{(1/n_1)\sum_{j=1}^{n_1} q_0(\omega_{1j})\alpha(\omega_{1j})}. \tag{5.2}$$

An obvious question then is the choice of α. Indeed, if we choose $\alpha = q_0^{-1}$, then we have the standard importance sampling estimator (e.g., Ott, 1979), which has large variability when the χ^2 distance between p_0 and p_1 is large. By sensibly choosing α, we can reduce this variability by orders of magnitude, since a good choice of α can "bridge" p_0 and p_1, and thus effectively shorten the distance between p_0 and p_1 [see Meng and Wong (1996) and Gelman and Meng (1998) for detailed discussions]. The specific variability we refer to here is the mean-squared error of $\log \hat{r}_\alpha$, which is asymptotically equivalent to the relative mean-squared error of $\hat{r}_\alpha$, $E(\hat{r}_\alpha - r)^2/r^2$. The log scale not only makes the error symmetric about p_0 and p_1, but it is also more relevant in many applications (e.g., log-likelihood ratios; free-energy differences).

To find the α that minimizes $R_{n_0,n_1}^{(\alpha)} \triangleq E[\log \hat{r}_\alpha - \log r]^2$ in general is a very difficult problem even asymptotically (see Romero, 2003). This is because $R_{n_0,n_1}^{(\alpha)}$ in general is a very complicated functional of α due to the fact that the draws from the same p_i, $i = 0, 1$, are not necessarily independent (recall we generally obtain draws using Markov chain Monte Carlo). However, when the draws are independent, it can be shown (e.g., Meng and Wong, 1996) that the α that minimizes the asymptotic $R_{n_0,n_1}^{(\alpha)}$ is given by

$$\begin{aligned} \alpha_{\text{opt}}(\omega) &\propto \frac{1}{s_0 p_0(\omega) + s_1 p_1(\omega)} \\ &\propto \frac{1}{s_0 r q_0(\omega) + s_1 q_1(\omega)}, \end{aligned} \tag{5.3}$$

where $s_i = n_i/(n_0 + n_1)$, $i = 0, 1$, are assumed to be strictly between 0 and 1 asymptotically. Since this α_{opt} depends on the unknown r, Meng and Wong (1996) constructed an iterative sequence that converges to an estimator which achieves the asymptotic minimum error. Empirical and theoretical evidence (e.g., Meng and Schilling, 1996; Servidea, 2002; Romero, 2003) suggests that (5.3) is typically a sensible choice of α even with dependent draws. It is therefore useful to explore the finite-sample cases under the independence assumption.

Given the intuitive nature (i.e., a mixture of p_0 and p_1) of α_{opt} in (5.3), it is of great theoretical interest to find out what else can make $\log \hat{r}_\alpha$ even more accurate with finite samples. Since the asymptotic mean-squared error ignores the bias in $\log \hat{r}_\alpha$, we expect the bias to have an important role to play in determining the finite-sample optimal bridge; the issue of reducing finite-sample bias is also of some practical interest (e.g., see Meng and Schilling, 1996). Theorem 2 makes it possible to investigate such questions because it allows us to derive an exact expression for $R_{n_0,n_1}^{(\alpha)}$.

To apply Theorem 2, let

$$X_0 = \frac{1}{n_0}\sum_{j=1}^{n_0} p_1(\omega_{0j})\alpha(\omega_{0j})$$

and

$$X_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} p_0(\omega_{1j})\alpha(\omega_{1j}).$$

Then, using the fact that $\log \hat{r}_\alpha - \log r = \log X_0 - \log X_1$ and $M_{X_0,X_1}(t_1, t_2; \alpha) = M_{X_0}(t_1; \alpha) M_{X_1}(t_2; \alpha)$, (2.17) can be simplified to

$$\begin{aligned} R_{n_0,n_1}^{(\alpha)} &= \int_0^\infty [M_{X_0}^{(1)}(-t;\alpha) + M_{X_1}^{(1)}(-t;\alpha)](\log t)^2\, dt \\ &\quad - 2\left[\int_0^\infty M_{X_0}^{(1)}(-t;\alpha)\log t\, dt\right] \\ &\quad \cdot \left[\int_0^\infty M_{X_1}^{(1)}(-t;\alpha)\log t\, dt\right] - \frac{\pi^2}{3}. \end{aligned} \tag{5.4}$$

In deriving (5.4) we have used the fact that $\int_0^\infty M_{X_i}^{(1)}(-t;\alpha)\, dt = 1$, $i = 0, 1$.

Under the further assumption that draws from each p_i are independent, we have

$$M_{X_i}(t;\alpha) = M_i^{n_i}\left(\frac{t}{n_i};\alpha\right), \quad i = 0, 1, \tag{5.5}$$

where $M_i(t;\alpha)$ is the MGF of $p_{1-i}(\omega)\alpha(\omega)$ with $\omega \sim p_i(\omega)$, $i = 0, 1$. Combining (5.5) and (5.4) gives us an exact expression of $R_{n_0,n_1}^{(\alpha)}$ as a functional of α, which can then be maximized by using the calculus of variations. The details of this exercise are rather involved and thus are deferred to Appendix C. The end result is

that the optimal α_O must satisfy the integral equation (we use the $\propto$ notation because α_O is determined up to a proportionality constant)

$$\alpha_O(\omega) \propto \frac{1}{s_0 p_0(\omega) W_0(\omega;\alpha_O) + s_1 p_1(\omega) W_1(\omega;\alpha_O)}, \tag{5.6}$$

where

$$W_i(\omega;\alpha_O) = \int_0^\infty \int_0^t \xi_{1-i}(t;\alpha_O) \cdot \exp\left(-\frac{s}{n_{1-i}}\,\alpha_O(\omega)p_i(\omega)\right) ds\,dt, \quad i=0,1, \tag{5.7}$$

$$\xi_i(t;\alpha) = \left(\log t - C_{1-i}^{(\alpha)}\right) \cdot M_i^{n_i-1}\left(-\frac{t}{n_i};\alpha\right), \quad i=0,1, \tag{5.8}$$

and $C_{1-i}^{(\alpha_O)}$ is a constant given in (C.2) of Appendix C.

Thus, heuristically speaking, the finite solution differs from the asymptotic solution (5.3) by incorporating additional "weights" (i.e., in addition to s_0 and s_1) $W_i(\omega;\alpha_O)$ when forming the mixture. Of course, the limit of $W_i(\omega;\alpha_O)$ must be free of ω and i in view of (5.3). This is indeed the case, because it is easy to verify that for any $\alpha(\omega) > 0$,

$$\lim_{n_0,n_1\to\infty} W_i(\omega;\alpha) = \int_0^\infty t[\log t + (\gamma + \log\beta_\alpha)]\exp(-t\beta_\alpha)\,dt = \beta_\alpha^{-2},$$

where γ is Euler's constant and $\beta_\alpha = M_0^{(1)}(0;\alpha) = M_1^{(1)}(0;\alpha) = \int \alpha(\omega)p_0(\omega)p_1(\omega)\mu(d\omega) > 0$.

Although it is impossible to solve (5.6) analytically for α_O in general, expressions (5.6)–(5.8) allow us to explore the structure of the finite-sample optimal solution. For example, we observe that the $\xi_i(t;\alpha)$ function defined in (5.8), which plays a key role in determining the finite-sample "weights" W_i, also plays a key role in determining the finite-sample bias of $\log\hat{r}_\alpha$, because

$$\begin{aligned}\int_0^\infty \xi_i(t;\alpha) M_{1-i}^{(1)}\left(-\frac{t}{n_{1-i}};\alpha\right) dt &= (-1)^i\left(C_0^{(\alpha)} - C_1^{(\alpha)}\right)\\ &= (-1)^i E[\log X_0 - \log X_1], \quad i=0,1,\end{aligned}$$

which is the (positive or negative) finite-sample bias of $\log\hat{r}_\alpha$. The extent to which the finite-sample solution reduces the mean-squared error and the bias relative to the asymptotic choice (5.3) can be investigated by asymptotic expansions of W_i, $i=0,1$. Clearly this investigation is beyond the scope of this paper; here our main purpose is to demonstrate how Theorem 2 has made such a previously intractable investigation possible.

6. ONCE AGAIN THERE IS A LINK TO FISHER

As discussed in Section 1, various cases of Theorem 1, particularly Lemmas 1 and 2, have appeared in the literature due to their usefulness in analytical moment calculations. Even Theorem 2 is a mathematical consequence, albeit not in an obvious form. Cressie, Davis, Folks and Policello (1981) mentioned that the earliest work of this sort they were able to trace was Williams (1941). In fact, the relevant literature can be traced back to Fisher (1930), who considered the problem of finding finite-sample moments of sample measures of "departure from normality," including the sample skewness and kurtosis. Fisher made explicit use of the moment generating function (which Fisher called characteristic function even though his definition did not involve $i=\sqrt{-1}$), and established a symbolic relationship between a joint MGF of a set of random variables and a joint MGF of functions of these random variables when the first MGF is evaluated at the origin. From this relationship and under the normality assumption, Fisher arrived at the identity

$$E\left(\frac{k_3^a k_4^b k_5^c\cdots}{k_2^r}\right) = \frac{E(k_3^a k_4^b k_5^c\cdots)}{(d^r/dt_2^r)M_{k_2}(t_2)|_{t_2=0}}, \tag{6.1}$$

where k_j is Fisher's k statistic of order j (i.e., k_j is the jth sample cumulant), $M_{k_2}(t_2)$ is the MGF of k_2 (= sample variance), $a,b,c,\ldots$ are integers and r is chosen such that the ratio inside the left-hand side is scale invariant. While (6.1) is a reexpression of Fisher's identity for the sake of modern readers, the notation d^r/dt_2^r was explicitly used by Fisher (1930, page 28) and it was clearly for a fractional derivative, since only $2r$ was guaranteed to be an integer. The use of fractional differentiation is more evident from Fisher's symbolic operations that led to (6.1), where he introduced the notation $D_3D_2^{3/2}$ and $D_5D_2^{5/2}$, "where D_p stands for d/dt_p" (Fisher, 1930, page 28).

However, not unusual in his writing, Fisher (1930, page 28) used these operators "without discussing what meaning should be attached to the fractional indices."

In particular, he gave

$$\left.\frac{d^r}{dt_2^r} M_{k_2}(t_2)\right|_{t_2=0} = \frac{(n+2r-3)\cdots(n+1)(n-1)}{(n-1)^r}\sigma^{2r}, \quad n \geq 2, \tag{6.2}$$

which, of course, is $E(k_2^r) = \sigma^{2r} E(\chi_{n-1}^{2r})/(n-1)^r$, as can be verified from (2.8) with $X = \chi_{n-1}^2$. It is possible that Fisher had first obtained (6.2) for integer r and then formally generalized it to noninteger r in the obvious way. For the normal problem, this generalization turns out to be irrelevant since "r is always an integer save for the odd moments which necessarily vanish" (Fisher, 1930, page 25), because the numerator on the right-hand side of (6.1) is zero when $2r$ is odd and thus these operators "find in fact only zero terms on which to operate" (Fisher, 1930, page 28). It would certainly be interesting to learn what meaning Fisher would have attached to these operators had he chosen to work with nonnormal distributions, because how to define a fractional differentiation operator was still a topic of discussion and research more than four decades later (e.g., Ross, 1975; Johnson, 1975).

Fisher's symbolic method is difficult to apply in general without the normality assumption, as Bowman and Shenton (1992) demonstrated. The simplicity under the normality comes from the fact that the ratio in the left-hand side of (6.1) is independent of its denominator (because the ratio is scale invariant) and thus the expectation of the ratio is the ratio of expectations, which leads to (6.1). In an earlier work where he introduced the celebrated k statistics, Fisher (1929) worked out how to obtain $E(k_3^a k_4^b k_5^c \cdots)$ via a joint MGF; thus what his symbolic approach effectively accomplished was to evaluate $E(k_2^r)$ through the rth derivative of $M_{k_2}(t)$ at $t = 0$, a predecessor of (2.8). Although Fisher's method is not very effective without the independence structure, it made it clear that the moments of a ratio, including fractional moments, can be obtained directly from the joint MGF of its numerator and denominator, which is the central theme for the later work discussed in this paper. Furthermore, it directly stimulated more workable approaches, such as (2.8).

For example, in an attempt to overcome the difficulty with Fisher's method, Bowman and Shenton (1992) arrived at (2.8) with $a = 1/2$, which they used to obtain a series expansion of the mean of the sample standard deviation from a modified normal distribution. They also applied a version of (2.4) to derive exact expressions of moments of ratios of central and noncentral sample moments when the data are generated from a uniform on the unit interval, and they emphasized the power of the MGF method in obtaining these finite-sample results. It is with the same emphasis that this paper attempts to unify and extend various results as given in Section 2, and the presented applications are intended to demonstrate the usefulness of these identities in finite-sample theoretical studies, such as the one that Fisher (1930) pursued.

EPILOGUE

Moment generating functions indeed generate all kinds of moments, many more than most of our textbooks have ever taught us. The possibility of using (0.3) is simply endless, because it works for infinitely many pairs of g and h that satisfy (0.2) (cf. Gradshteyn and Ryzhik, 1992), as emphasized by Cressie and Borkent (1986). For example,

$$E\left[\frac{X}{X^2-1}\right] = \int_0^\infty \cosh(t) M_X(-t)\, dt$$

when $P(X > 1) = 1$. Although the majority of these identities remain at most a mathematical curiosity, our knowledge about their existences can bring us some happy (research) moments, as the examples in this article intend to demonstrate.

In addition, our toolkit can be further expanded if we replace the Laplace transform by the Fourier transform or even by a hybrid transform such as $E[e^{it_1X+t_2Y}]$, where $i = \sqrt{-1}$, as suggested by Professor K. Lange in a personal exchange.

Of course, all these would be stories for another day, if you have a second moment.

APPENDIX A: PROOFS FOR SECTION 2

PROOF OF LEMMA 1. Choose an ε such that $M_{X,Y}(t_1, 0)$ exists when $|t_1| \leq \varepsilon$. For any x and positive y,

$$\sum_{j=0}^{\infty} \frac{|t_1 x|^j}{j!} \exp(-t_2 y) \leq e^{\varepsilon x} + e^{-\varepsilon x} \quad \text{if } |t_1| \leq \varepsilon \text{ and } t_2 \geq 0. \tag{A.1}$$

Since the right-hand side of (A.1) is integrable with respect to P under our assumption, we can interchange integration with summation (by Theorem 16.7

of Billingsley, 1995, page 211) to obtain

$$\sum_{j=0}^{\infty} \frac{t_1^j}{j!} E(X^j e^{-t_2 Y}) = E[\exp(t_1 X - t_2 Y)] = M_{X,Y}(t_1, -t_2) \quad \text{for all } |t_1| \le \varepsilon,\ t_2 > 0.$$

It follows then (see Billingsley, 1995, page 543) that, for any integer $k \ge 0$ and $t \ge 0$, $M_{X,Y}^{(k,0)}(0, -t^{1/b})$ exists and is given by

$$M_{X,Y}^{(k,0)}(0, -t^{1/b}) = E[X^k \exp(-t^{1/b} Y)]. \tag{A.2}$$

Let $B(X, Y; t) = X^k \exp(-t^{1/b} Y)$. Since X^k / Y^b is quasi-integrable, we have, say, $E[(X^k)^+ / Y^b] < \infty$. Then

$$E\left[\int_0^{\infty} (B(X, Y; t))^+ \, dt\right] = E\left[(X^k)^+ \int_0^{\infty} \exp(-t^{1/b} Y)\, dt\right] = \Gamma(b+1) E\left[\frac{(X^k)^+}{Y^b}\right] < \infty.$$

Consequently, by Fubini's theorem (for quasi-integrability; see Neveu, 1965, Chapter III.2) and by (A.2), $M_{X,Y}^{(k,0)}(0, -t^{1/b}) \triangleq E[B(X, Y; t)]$ is quasi-integrable with respect to λ_1^+, and (2.4) holds. □

PROOF OF LEMMA 2. For any $t_1 > 0$ and $t_2 \ge 0$, since

$$\sum_{j=0}^{\infty} (|tx|^j / j!) \exp(-t_1 x - t_2 y) \le \exp\bigl(-(t_1 - |t|)x\bigr),$$

which is integrable with respect to P when $|t| < t_1$, we have (again by Theorem 16.7 of Billingsley, 1995)

$$\sum_{j=0}^{\infty} \frac{t^j}{j!} E\bigl(X^j \exp(-t_1 X - t_2 Y)\bigr) = M_{X,Y}(t - t_1, -t_2) \quad \text{for any } |t| < t_1.$$

It follows then (again see Billingsley, 1995, page 543), for any positive integer k, that

$$M_{X,Y}^{(k,0)}(-t_1, -t_2) = E[X^k \exp(-t_1 X - t_2 Y)] \quad \text{if } t_1 > 0 \text{ and } t_2 \ge 0.$$

The rest of the proof follows trivially from Fubini's theorem, noting that the value of $M_{X,Y}^{(\lceil a \rceil, 0)}(-t_1, -t_2)$ at $t_1 = 0$ is not relevant for (2.11). □

PROOF OF THEOREM 2. For any positive x, y, t_1 and t_2, let

$$h_t(x, y; t_1, t_2) = xy \exp(-t_1 x - t_2 y) \left(\frac{t_2}{t_1}\right)^t \frac{\sin(\pi t)}{\pi t}, \tag{A.3}$$

where the index t is an arbitrary real number and the right-hand side of (A.3) is defined to be $xy \cdot \exp(-t_1 x - t_2 y)$ when $t = 0$. Since

$$\frac{\sin(\pi t)}{\pi t} = \sum_{j=0}^{\infty} (-1)^j \frac{(\pi t)^{2j}}{(2j+1)!} \quad \text{for all } |t| < \infty,$$

it is easy to see that $h_t(x, y; t_1, t_2)$ is differentiable with respect to t to any order. Furthermore, for a given positive integer k and $1 > \varepsilon > 0$, one can find a constant $c(\varepsilon, k)$ such that

$$\begin{aligned} &\sup_{|t| \le \varepsilon} \left| \frac{\partial^k h_t(x, y; t_1, t_2)}{\partial t^k} \right| \\ &\quad \le c(\varepsilon, k) xy \exp(-t_1 x - t_2 y) \\ &\qquad \cdot \left(\left(\frac{t_2}{t_1}\right)^{\varepsilon} + \left(\frac{t_2}{t_1}\right)^{-\varepsilon} \right) \\ &\qquad \cdot (1 + |\log t_2 - \log t_1|^k). \end{aligned} \tag{A.4}$$

Since the right-hand side of (A.4) is integrable in (t_1, t_2) with respect to λ_2^+ when $\varepsilon < 1$, the following interchange of integration with differentiation is justified by the dominated convergence theorem (DCT):

$$\begin{aligned} &\int_0^{\infty} \int_0^{\infty} xy \exp(-t_1 x - t_2 y) g_k(t_1, t_2)\, dt_1\, dt_2 \\ &\quad = \int_0^{\infty} \int_0^{\infty} \left\{ \frac{\partial^k h_t(x, y; t_1, t_2)}{\partial t^k} \right\}_{t=0} dt_1\, dt_2 \\ &\quad = \left\{ \frac{\partial^k}{\partial t^k} \int_0^{\infty} \int_0^{\infty} h_t(x, y; t_1, t_2)\, dt_1\, dt_2 \right\}_{t=0} \\ &\quad = \left\{ \frac{\partial^k}{\partial t^k} \left(\frac{x}{y}\right)^t \right\}_{t=0} = (\log x - \log y)^k. \end{aligned} \tag{A.5}$$

In deriving (A.5), we used the identity $\Gamma(1-t)\Gamma(1+t) = \pi t / \sin(\pi t)$ for $0 \le t < 1$.

Now suppose that $(\log X - \log Y)^k$ is quasi-integrable with respect to P, say, $E[((\log X - \log Y)^k)^+] < \infty$. Let $B(X, Y; t_1, t_2) = XY \exp(-t_1 X - t_2 Y) g_k(t_1, t_2)$ and $B_i(X, Y; t_1, t_2) = XY \exp(-t_1 X - t_2 Y)(\log t_2 - \log t_1)^i$, $i = 0, 1, 2, \ldots$. Then from (2.13) we have

$$(B(X, Y; t_1, t_2))^+ \le \sum_{j=0}^{[k/2]} |\alpha_{j,k}| (B_{k-2j}(X, Y; t_1, t_2))^+. \tag{A.6}$$

Using the fact that $((a+b)^i)^+ \le 2^{i-1}[((a)^i)^+ + ((b)^i)^+]$ for any positive integer i, we have $(B_i(X,Y;t_1,t_2))^+ \le 2^{i-1}XY\exp(-t_1X-t_2Y)[|\log(t_2Y) - \log(t_1X)|^i + ((\log Y - \log X)^i)^+]$. This implies that for any $1 \le i \le k$,

$$\begin{aligned} E\Big[\int_0^\infty\int_0^\infty (B_i(X,Y;t_1,t_2))^+\,dt_1\,dt_2\Big] &\le 2^{i-1}\Big\{\int_0^\infty\int_0^\infty \exp(-s_1-s_2) \\ &\qquad \cdot |\log(s_2/s_1)|^i\,ds_1\,ds_2 \\ &\qquad + E\big[\big((\log(X/Y))^k\big)^+\big]+1\Big\} \\ &< \infty. \end{aligned}$$

It follows then, by (A.6), DCT and Fubini's theorem, that $B(X,Y;t_1,t_2)$ is quasi-integrable with respect to $P\times\lambda_2^+$, which implies that $E[B(X,Y;t_1,t_2)]$ is quasi-integrable with respect to λ_2^+. Using a similar argument as for (A.2), one can easily show that $M_{X,Y}^{(1,1)}(-t_1,-t_2) = E(XY\exp(-t_1X-t_2Y))$ for any $t_1>0$ and $t_2>0$. Therefore, $M_{X,Y}^{(1,1)}(-t_1,-t_2)g_k(t_1,t_2) = E[B(X,Y;t_1,t_2)]$ is quasi-integrable with respect to λ_2^+ and (2.14) holds because of (A.5) and Fubini's theorem.

Next we suppose, without loss of generality, that

$$\begin{aligned} &\big(M_{X,Y}^{(1,1)}(-t_1,-t_2)g_k(t_1,t_2)\big)^+ \\ &\quad \triangleq M_{X,Y}^{(1,1)}(-t_1,-t_2)(g_k(t_1,t_2))^+ \end{aligned}$$

is integrable with respect to λ_2^+. It follows immediately by Fubini's theorem that $B(X,Y;t_1,t_2)$ is quasi-integrable with respect to $P\times\lambda_2^+$ since

$$\begin{aligned} &\int_0^\infty\int_0^\infty E[(B(X,Y;t_1,t_2))^+]\,dt_1\,dt_2 \\ &\quad = \int_0^\infty\int_0^\infty M_{X,Y}^{(1,1)}(-t_1,-t_2)(g_k(t_1,t_2))^+\,dt_1\,dt_2 \\ &\quad < \infty. \end{aligned}$$

By Fubini's theorem and (A.5),

$$(\log X - \log Y)^k \triangleq \int_0^\infty\int_0^\infty B(X,Y;t_1,t_2)\,dt_1\,dt_2$$

is quasi-integrable with respect to P and (2.14) holds.

Identities (2.15)–(2.17) follow from (2.14) because $g_1(t_1,t_2) = \log t_2 - \log t_1$ and $g_2(t_1,t_2) = (\log t_2 - \log t_1)^2 - \pi^2/3$. To simplify the expressions, we used the following facts: $\int_0^\infty e^{-t}\log t\,dt = \Gamma'(1) = -\gamma$, $\int_0^\infty e^{-t}(\log t)^2\,dt = \Gamma''(1) = \gamma^2 + \pi^2/6$ and $\int_0^\infty\int_0^\infty M_{X,Y}^{(1,1)}(-t_1,-t_2)\,dt_1\,dt_2 = 1$, which is (2.14) when $k=0$. □

APPENDIX B: DERIVATION OF $E(r_n)$ IN SECTION 4

Without loss of generality, we can assume $\binom{x_i}{y_i} \overset{\text{i.i.d.}}{\sim} N(0, \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right))$. With a simple rotation we have

$$E(r_n) = E\Big[\frac{\sum_{i=1}^{n-1} x_i y_i}{\sqrt{\sum_{i=1}^{n-1}x_i^2}\sqrt{\sum_{i=1}^{n-1}y_i^2}}\Big] \triangleq E\Big[\frac{X}{\sqrt{Y_1}\sqrt{Y_2}}\Big].$$

To apply Theorem 1, we first calculate

$$\begin{aligned} &M_{X,Y_1,Y_2}(t_1,t_2,t_3) \\ &\quad = \Big\{\frac{1}{2\pi|\Sigma|^{1/2}} \\ &\qquad \cdot \iint \exp\Big[-\frac{1}{2}(x,y)(\Sigma^{-1}-T)\begin{pmatrix}x\\y\end{pmatrix}\Big] \\ &\qquad\qquad \cdot dx\,dy\Big\}^{n-1}, \end{aligned}$$

where $T = \left(\begin{smallmatrix}2t_2 & t_1\\ t_1 & 2t_3\end{smallmatrix}\right)$ and $\Sigma = \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)$. Thus,

$$\begin{aligned} &M_{X,Y_1,Y_2}(t_1,t_2,t_3) \\ &\quad = \Big(\frac{1}{|\Sigma|^{1/2}|\Sigma^{-1}-T|^{1/2}}\Big)^{n-1} \\ &\quad = |I-\Sigma T|^{-(n-1)/2} \\ &\quad = \big[(1-2t_2)(1-2t_3) - 4\rho^2t_2t_3 \\ &\qquad\qquad - 2\rho t_1 - (1-\rho^2)t_1^2\big]^{-(n-1)/2} \end{aligned}$$

and

$$\begin{aligned} &M_{X,Y_1,Y_2}^{(1,0,0)}(0,-t_2,-t_3) \\ &\quad = \frac{(n-1)\rho}{[4t_2t_3(1-\rho^2)+2(t_2+t_3)+1]^{(n+1)/2}}. \end{aligned}$$

Since $M_{X,Y_1,Y_2}(t_1,0,0)$ exists for (at least) $|t_1| \le (1+\rho)^{-1}$ and $E(|r_n|)\le 1$, we can apply Theorem 1 (with $L=1$, $M=0$, $N=2$ and $k_1=1$, $b_1=b_2=1/2$) to obtain

$$\begin{aligned} E(r_n) &= \frac{(n-1)\rho}{\Gamma(\frac{3}{2})\Gamma(\frac{3}{2})} \\ &\quad \cdot \int_0^\infty\int_0^\infty [4t_2^2t_3^2(1-\rho^2) \\ &\qquad + 2(t_2^2+t_3^2)+1]^{-(n+1)/2}\,dt_2\,dt_3. \end{aligned} \tag{B.1}$$

With the transformation $t_2 = \sqrt{r}\sin\theta$ and $t_3 = \sqrt{r}\cos\theta$, (B.1) becomes (4.4).

APPENDIX C: DERIVATION OF OPTIMAL α IN SECTION 5

To derive the optimal α that minimizes $R^{(\alpha)}_{n_0,n_1}$, we follow the calculus of variations approach with a simple modification to accommodate the constraint that α is positive. We note that for any $h(\omega)$ that is bounded on Ω, we can find an $\varepsilon_h > 0$ such that when $|\varepsilon| < \varepsilon_h$, $[1 + \varepsilon h(\omega)]\alpha(\omega)$ is positive on Ω. Now if α_O minimizes $R^{(\alpha)}_{n_0,n_1}$, then $\xi(\varepsilon) \stackrel{\Delta}{=} R^{((1+\varepsilon h)\alpha_O)}_{n_0,n_1}$ must achieve its minimum at $\varepsilon = 0$ and thus $\xi'(0) = 0$. This implies, from (5.4) and (5.5), that

$$\int_0^\infty [G_0(t;\alpha_O,h) + G_1(t;\alpha_O,h)]\,dt = 0,$$

where, for $i = 0, 1$,

$$\begin{aligned} &G_i(t;\alpha_O,h) \\ &\quad = \frac{\partial^2 M_i^{n_i}(s/n_i;(1+\varepsilon h)\alpha_O)}{\partial\varepsilon\,\partial s}\bigg|_{(\varepsilon=0,s=-t)} \\ &\qquad \cdot \big[(\log t)^2 - 2C^{(\alpha_O)}_{n_{1-i}}\log t\big] \\ &\quad = A_i^{(\alpha_O)}(t)E_i\bigg[\exp\Big(-\frac{t}{n_i}\alpha_O p_{1-i}\Big)\alpha_O p_{1-i}h\bigg] \\ &\qquad - \frac{1}{n_i}B_i^{(\alpha_O)}(t) \\ &\qquad \cdot E_i\bigg[\exp\Big(-\frac{t}{n_i}\alpha_O p_{1-i}\Big)(\alpha_O p_{1-i})^2 h\bigg] \end{aligned} \tag{C.1}$$

with

$$\begin{aligned} C_i^{(\alpha_O)} &= \int_0^\infty M_i^{n_i-1}\Big(-\frac{t}{n_i};\alpha_O\Big) \\ &\qquad \cdot M_i^{(1)}\Big(-\frac{t}{n_i};\alpha_O\Big)\log t\,dt, \end{aligned} \tag{C.2}$$

$$\begin{aligned} A_i^{(\alpha_O)}(t) &= \frac{\partial[tM_i^{n_i-1}(-t/n_i;\alpha_O)]}{\partial t} \\ &\qquad \cdot \big[(\log t)^2 - 2C^{(\alpha_O)}_{1-i}\log t\big], \end{aligned} \tag{C.3}$$

$$\begin{aligned} B_i^{(\alpha_O)}(t) &= tM_i^{n_i-1}\Big(-\frac{t}{n_i};\alpha_0\Big) \\ &\qquad \cdot \big[(\log t)^2 - 2C^{(\alpha_O)}_{1-i}\log t\big]. \end{aligned} \tag{C.4}$$

It follows then, by interchanging the integrations, that

$$\begin{aligned} 0 = \int_\Omega \Bigg\{ &\int_0^\infty \bigg[A_0^{(\alpha_O)}(t)\exp\Big(-\frac{t}{n_0}\alpha_O(\omega)p_1(\omega)\Big) \\ &\qquad + A_1^{(\alpha_O)}(t)\exp\Big(-\frac{t}{n_1}\alpha_O(\omega)p_0(\omega)\Big)\bigg]dt \\ &- \bigg[\int_0^\infty B_0^{(\alpha_O)}(t)\exp\Big(-\frac{t}{n_0}\alpha_O(\omega)p_1(\omega)\Big)dt\bigg] \\ &\cdot \frac{\alpha_O(\omega)p_1(\omega)}{n_0} \\ &- \bigg[\int_0^\infty B_1^{(\alpha_O)}(t)\exp\Big(-\frac{t}{n_1}\alpha_O(\omega)p_0(\omega)\Big)dt\bigg] \\ &\qquad \cdot \frac{\alpha_O(\omega)p_0(\omega)}{n_1}\Bigg\} \\ &\cdot \alpha_O(\omega)p_0(\omega)p_1(\omega)h(\omega)\mu(d\omega). \end{aligned} \tag{C.5}$$

This implies, because $h(\omega)$ is arbitrary besides being bounded, that the expression inside the brace of (C.5), as a function of ω, must be zero on Ω almost surely with respect $\mu(\omega)$. This yields, after an integration by parts as implied by (C.3) and (C.4),

$$\begin{aligned} &\int_0^\infty \xi_0(t;\alpha_O)\exp\Big(-\frac{t}{n_0}\alpha_O(\omega)p_1(\omega)\Big)dt \\ &\quad + \int_0^\infty \xi_1(t;\alpha_O)\exp\Big(-\frac{t}{n_1}\alpha_O(\omega)p_0(\omega)\Big)dt \\ &\quad = 0, \end{aligned} \tag{C.6}$$

where $\xi_i(t;\alpha_O)$ is given by (5.8). Using the fact that $\int_0^t e^{-as}\,ds = (1 - e^{-at})/a$ and the fact that if $\alpha_O(\omega)$ satisfies (C.6), then $c\alpha_O(\omega)$ must satisfy (C.6) for any c ($\neq 0$) that is independent of ω because $\hat{r}_\alpha \stackrel{\Delta}{=} \hat{r}_{c\alpha}$ [see (5.2)], we can rewrite (C.6) as (5.6).

ACKNOWLEDGMENTS

This article owes its existence to encouragement, comments and help from many colleagues and friends. My foremost thanks go to George Tiao for introducing me to the world of unit roots, and for sharing its joy and frustration through a joint project that motivated this article. This article was presented at the NSF-NBER Time-Series Conference held in Chicago, September 19–20, 2003 in honor of George Tiao's 70th birthday. The comments and encouragement from the discussant, J.-L. Lin, and from many audience members are particularly acknowledged. My thanks also go to, among many others, P. Billingsley, D. Banks, M. Bock, N. Chan, N. Cressie, A. Dasgupta,

P. Donnelly, V. Dukic, S. Kotz, K. Lange, W. Rosenberger, S. Stigler, F. Vaida and D. Wallace. Special thanks go to M. Romero and M. Wichura for their independent verification of a number of theoretical and numerical results presented in this paper.

I also want to thank Editor George Casella for his appreciation and encouragement of this project over the years, and an Associate Editor and two referees for a set of simply the most congenial and helpful review reports I have ever received: I never have had a happier "referee-report-reading" moment than this one. This research was supported in part by NSA Grant MDA 904-96-1-0007 and NSF Grants DMS 95-05043, 96-26691, 00-72827 and 02-04552.

REFERENCES

ABADIR, K. M. (1993). The limiting distribution of the autocorrelation coefficient under a unit root. *Ann. Statist.* **21** 1058–1070.

ABADIR, K. M. and LARSSON, R. (1996). The joint moment generating function of quadratic forms in multivariate autoregressive series. *Econometric Theory* **12** 682–704.

ABADIR, K. M. and LARSSON, R. (2001). The joint moment generating function of quadratic forms in multivariate autoregressive series: The case with deterministic components. *Econometric Theory* **17** 222–246.

BENNETT, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Computational Phys.* **22** 245–268.

BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York.

BOCK, M. E., JUDGE, G. G. and YANCEY, T. A. (1984). A simple form for the inverse moments of non-central χ^2 and F random variables and certain confluent hypergeometric functions. *J. Econometrics* **25** 217–234.

BOWMAN, K. O. and SHENTON, L. R. (1992). Some exact expressions for the mean and higher moments of functions of sample moments. *Ann. Inst. Statist. Math.* **44** 781–798.

BRANDWEIN, A. C. and STRAWDERMAN, W. E. (1990). Stein estimation: The spherically symmetric case. *Statist. Sci.* **5** 356–369.

CEPERLEY, D. M. (1995). Path integrals in the theory of condensed helium. *Rev. Modern Phys.* **67** 279–355.

CHAN, N. H. and WEI, C. Z. (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Ann. Statist.* **15** 1050–1063.

CHAO, M. T. and STRAWDERMAN, W. E. (1972). Negative moments of positive random variables. *J. Amer. Statist. Assoc.* **67** 429–431.

CRESSIE, N. and BORKENT, M. (1986). The moment generating function has its moments. *J. Statist. Plann. Inference* **13** 337–344.

CRESSIE, N., DAVIS, A. S., FOLKS, J. L. and POLICELLO, G. E. (1981). The moment-generating function and negative integer moments. *Amer. Statist.* **35** 148–150.

DAVIES, N., PATE, M. B. and PETRUCCELLI, J. D. (1985). Exact moments of the sample cross correlations of multivariate autoregressive moving average time series. *Sankhyā Ser. B* **47** 325–337.

DE GOOIJER, J. G. (1980). Exact moments of the sample autocorrelations from series generated by general ARIMA processes of order (p, d, q), $d = 0$ or 1. *J. Econometrics* **14** 365–379.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92** 903–915.

DICKEY, D. A. and FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* **74** 427–431.

EFRON, B. and MORRIS, C. N. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

ELLIOTT, G., ROTHENBERG, T. and STOCK, J. H. (1996). Efficient tests for an autoregression unit root. *Econometrica* **64** 813–836.

EVANS, G. B. A. and SAVIN, N. E. (1981). Testing for unit roots. I. *Econometrica* **49** 753–779.

EVANS, G. B. A. and SAVIN, N. E. (1984). Testing for unit roots. II. *Econometrica* **52** 1241–1269.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521.

FISHER, R. A. (1929). Moments and product moments of sampling distributions. *Proc. London Math. Soc. (2)* **30** 199–238.

FISHER, R. A. (1930). The moments of the distribution for normal samples of measures of departure from normality. *Proc. Roy. Soc. London Ser. A* **130** 16–28.

FROM, S. G. and SAXENA, K. M. L. (1989). Estimating parameters from mixed samples using sample fractional moments. *J. Statist. Plann. Inference* **21** 231–244.

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185.

GONZALO, J. and PITARAKIS, J. (1998). On the exact moments of asymptotic distributions in an unstable AR(1) with dependent errors. *Internat. Econom. Rev.* **39** 71–88.

GRADSHTEYN, I. S. and RYZHIK, I. M. (1992). *Table of Integrals, Series, and Products*, corrected and enlarged ed. Academic Press, San Diego, CA.

HOQUE, A. (1985). The exact moments of forecast error in the general dynamic model. *Sankhyā Ser. B* **47** 128–143.

HOTELLING, H. (1953). New light on the correlation coefficient and its transforms (with discussion). *J. Roy. Statist. Soc. Ser. B* **15** 193–232.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. California Press, Berkeley.

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1995). *Continuous Univariate Distributions* **2**, 2nd ed. Wiley, New York.

JOHNSON, N. L., KOTZ, S. and KEMP, A. W. (1992). *Univariate Discrete Distributions*, 2nd ed. Wiley, New York.

JOHNSON, P. D., JR. (1975). An algebraic definition of fractional differentiation. *Fractional Calculus and Its Applications. Lecture Notes in Math.* **457** 226–231. Springer, Berlin.

JONES, M. C. (1986). Expressions for inverse moments of positive quadratic forms in normal variables. *Austral. J. Statist.* **28** 242–250.

JONES, M. C. (1987a). Inverse factorial moments. *Statist. Probab. Lett.* **6** 37–42. Correction **6** 369.

JONES, M. C. (1987b). On moments of ratios of quadratic forms in normal variables. *Statist. Probab. Lett.* **6** 129–136. Correction **6** 369.

KHURI, A. and CASELLA, G. (2002). The existence of the first negative moment revisited. *Amer. Statist.* **56** 44–47.

KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. and TAN, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 585–618.

LAUE, G. (1980). Remarks on the relation between fractional moments and fractional derivatives of characteristic functions. *J. Appl. Probab.* **17** 456–466.

LIN, J. L. (2003). Discussion of "From unit root to Stein's estimator to Fisher's k statistics: If you have a moment, I can tell you more," by X.-L. Meng. Presented at the NSF-NBER Time-Series Conference, Chicago, September 19–20, 2003.

MAATTA, J. M. and CASELLA, G. (1990). Developments in decision-theoretic variance estimation (with discussion). *Statist. Sci.* **5** 90–120.

MATHAI, A. M. (1991). On fractional moments of quadratic expressions in normal variables. *Comm. Statist. Theory Methods* **20** 3159–3174.

MEHTA, J. S. and SWAMY, P. A. V. B. (1978). The existence of moments of some simple Bayes estimators of coefficients in a simultaneous equation model. *J. Econometrics* **7** 1–13.

MENG, X.-L. and SCHILLING, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *J. Amer. Statist. Assoc.* **91** 1254–1267.

MENG, X.-L. and SCHILLING, S. (2002). Warp bridge sampling. *J. Comput. Graph. Statist.* **11** 552–586.

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860.

MORIN, D. (1992). Exact moments of ratios of quadratic forms. *Metron* **30** 59–78.

NANKERVIS, J. C. and SAVIN, N. E. (1988). The exact moments of the least-squares estimator for the autoregressive model: Corrections and extensions. *J. Econometrics* **37** 381–388.

NEVEU, J. (1965). *Mathematical Foundations of the Calculus of Probability*. Holden-Day, San Francisco.

NIELSEN, B. (1997). Bartlett correction of the unit root test in autoregressive models. *Biometrika* **84** 500–504.

OLKIN, I. and PRATT, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Ann. Math. Statist.* **29** 201–211.

OTT, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American J. Human Genetics* **31** 161–175.

PETERS, T. A. (1989). The exact moments of OLS in dynamic regression models with nonnormal errors. *J. Econometrics* **40** 279–305.

PIEGORSCH, W. W. and CASELLA, G. (1985). The existence of the first negative moment. *Amer. Statist.* **39** 60–62. Comments by N. L. Johnson, **39** 240 and J. Hannan, **39** 326.

PITARAKIS, J. (1998). Moment generating functions and further exact results for seasonal autoregressions. *Econometric Theory* **14** 770–782.

PROVOST, S. B. and RUDIUK, E. M. (1994). The exact density function of the ratio of two dependent linear combinations of chi-square variables. *Ann. Inst. Statist. Math.* **46** 557–571.

RAO, M. M. (1978). Asymptotic distribution of an estimator of the boundary parameter of an unstable process. *Ann. Statist.* **6** 185–190. Correction **8** 1403.

ROMERO, M. (2003). On two topics with no bridge: Bridge sampling with dependent draws and bias of the multiple imputation variance estimator. Ph.D dissertation, Dept. Statistics, Univ. Chicago.

ROSS, B. (1975). A brief history and exposition of the fundamental theory of fractional calculus. *Fractional Calculus and Its Applications*. *Lecture Notes in Math.* **457** 1–36. Springer, Berlin.

SAWA, T. (1972). Finite sample properties of the k-class estimators. *Econometrica* **40** 653–680.

SAWA, T. (1978). The exact moments of the least squares estimator for the autoregressive model. *J. Econometrics* **8** 159–172.

SERVIDEA, J. (2002). Bridge sampling with dependent random draws: Techniques and strategy. Ph.D. dissertation, Dept. Statistics, Univ. Chicago.

SHEPP, L. A. and LLOYD, S. P. (1966). Ordered cycle lengths in a random permutation. *Trans. Amer. Math. Soc.* **121** 340–357.

SPRINGER, M. D. (1979). *The Algebra of Random Variables.* Wiley, New York.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

STIGLER, S. M. (1990). The 1988 Neyman memorial lecture: A Galtonian perspective on shrinkage estimators. *Statist. Sci.* **5** 147–155.

STUART, A. and ORD, J. K. (1987). *Kendall's Advanced Theory of Statistics* **1**. *Distribution Theory*, 5th ed. Oxford Univ. Press, London.

TANAKA, K. (1996). *Time Series Analysis*: *Nonstationary and Noninvertible Distribution Theory.* Wiley, New York.

TSUI, A. K. and ALI, M. M. (1994). Exact distributions, density functions and moments of the least squares estimator in a first-order autoregressive model. *Comput. Statist. Data Anal.* **17** 433–454.

VOTER, A. F. (1985). A Monte Carlo method for determining free-energy differences and transition state theory rate constants. *J. Chemical Physics* **82** 1890–1899.

WHITE, J. S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *Ann. Math. Statist.* **29** 1188–1197.

WHITE, J. S. (1959). The limiting distribution of the serial correlation coefficient in the explosive case. II. *Ann. Math. Statist.* **30** 831–834.

WILLIAMS, J. D. (1941). Moments of the ratio of the mean square successive difference to the mean square difference in samples from a normal universe. *Ann. Math. Statist.* **12** 239–241.

WOLFE, S. J. (1975). On moments of probability distribution functions. *Fractional Calculus and Its Applications*. *Lecture Notes in Math.* **457** 306–316. Springer, Berlin.