Computing the asymptotic power of a Euclidean-distance test for goodness-of-fit*

William Perkins

School of Mathematics
Georgia Institute of Technology
686 Cherry St.
Atlanta, GA 30332-0160
e-mail: perkins@math.gatech.edu

Gary Simon

IOMS Department
Stern School of Business
NYU
44 West 4th St.
New York, NY 10012
e-mail: gsimon@stern.nyu.edu

and

Mark Tygert

Courant Institute of Mathematical Sciences
NYU
251 Mercer St.
New York, NY 10012
e-mail: tygert@aya.yale.edu

Abstract: A natural (yet unconventional) test for goodness-of-fit measures the discrepancy between the model and empirical distributions via their Euclidean distance (or, equivalently, via its square). The present paper characterizes the statistical power of such a test against a family of alternative distributions, in the limit that the number of observations is large, with every alternative departing from the model in the same direction. Specifically, the paper provides an efficient numerical method for evaluating the cumulative distribution function (cdf) of the square of the Euclidean distance between the model and empirical distributions under the alternatives, in the limit that the number of observations is large. The paper illustrates the scheme by plotting the asymptotic power (as a function of the significance level) for several examples.

AMS 2010 subject classifications: Primary 62G10, 62F03; secondary 65C60.
Keywords and phrases: rms, root-mean-square, significance, statistic.

*Supported in part by NSF Grant OISE-0730136, an NSF Postdoctoral Research Fellowship, a DARPA Young Faculty Award, and an Alfred P. Sloan Research Fellowship.

Contents

1. Introduction

Given n observations, each falling in one of m bins, we would like to test if these observations are consistent with having arisen as independent and identically distributed (i.i.d.) draws from a specified probability distribution p_0 over the m bins (p_0 is known as the "model"). A natural measure of the deviation between p_0 and the observations is the square x_a of the Euclidean distance between the actually observed distribution of the draws and the expected distribution p_0, that is,

$$x_a = \sum_{k=1}^{m} ((y_a)_k - (p_0)_k)^2, \tag{1}$$

where $(y_a)_1$, $(y_a)_2$, ..., $(y_a)_m$ are the proportions of the n observations falling in bins 1, 2, ..., m, respectively.

The "P-value" is then defined to be the probability that $X_0 \geq x_a$, where X_0 would be the same as x_a, but constructed from n draws that definitely are taken i.i.d. from p_0, that is,

$$X_0 = \sum_{k=1}^{m} ((Y_0)_k - (p_0)_k)^2, \tag{2}$$

where $(Y_0)_1$, $(Y_0)_2$, ..., $(Y_0)_m$ are the proportions of n i.i.d. draws from p_0 falling in bins 1, 2, ..., m, respectively. When calculating the P-value — the probability that $X_0 \geq x_a$ — we view X_0 as a random variable while viewing x_a as a fixed number. If the P-value is small, then we can be confident that the observed draws were not taken i.i.d. from the model p_0.

To characterize the statistical power of the P-value based on the Euclidean distance, we consider n i.i.d. draws from the alternative distribution

$$p_a = p_0 + a/\sqrt{n}, \tag{3}$$

where a is a vector whose m entries satisfy $\sum_{k=1}^{m} a_k = 0$. We thus need to calculate the distribution of the square X_a of the Euclidean distance,

$$X_a = \sum_{k=1}^{m} ((Y_a)_k - (p_0)_k)^2, \tag{4}$$

where $(Y_a)_1, (Y_a)_2, \ldots, (Y_a)_m$ are the proportions of n i.i.d. draws from p_a falling in bins $1, 2, \ldots, m$, respectively. Section 4 below provides an efficient method for calculating the cumulative distribution function (cdf) of $n \cdot X_a$ in the limit that the number n of draws is large. Section 5 below then describes how to use such a method to plot the cdf of the P-values; this cdf is the same as the statistical power function of the hypothesis test based on the Euclidean distance (as a function of the significance level). Presenting this method is the principal purpose of the present paper, complementing the earlier discussions of Perkins, Tygert, and Ward (2011b) and Perkins, Tygert, and Ward (2011a), which compare the Euclidean distance with classical statistics such as χ^2, the log–likelihood-ratio G^2, and other members of the Cressie-Read power-divergence family; Perkins, Tygert, and Ward (2011b) and Perkins, Tygert, and Ward (2011a) review the classical statistics and provide detailed comparisons.

As reviewed, for example, by Kendall et al. (2009) and Rao (2002), $m \cdot n \cdot X_a$ defined in (4) converges in distribution to a noncentral χ^2 in the limit that the number n of draws is large, when the model p_0 is a uniform distribution. When p_0 is nonuniform, $m \cdot n \cdot X_a$ converges in distribution to the sum of the squares of independent Gaussian random variables in the limit that the number n of draws is large, as shown by Moore and Spruill (1975) and reviewed in Section 2 below. Section 3 provides integral representations for the cdf of the sum of the squares of independent Gaussian random variables and applies suitable quadratures for their numerical evaluation. Section 4 summarizes the numerical method obtained by combining Sections 2 and 3. Section 5 summarizes a scheme for plotting the asymptotic power (as a function of the significance level) using the method of Section 4. Section 6 illustrates the methods via several numerical examples.

The extension to models with nuisance parameters is straightforward, following Perkins, Tygert, and Ward (2011c); the present paper focuses on the simpler case in which the model p_0 is a single, fully specified probability distribution.

2. Preliminaries

This section states Theorem 2.1, which is a special case of Theorem 4.2 of Moore and Spruill (1975). Before stating the theorem, we need to set up some notation. The set-up amounts to an algorithm for computing the real numbers $\sigma_1, \sigma_2, \ldots, \sigma_{m-1}$ and $\zeta_1, \zeta_2, \ldots, \zeta_{m-1}$ used in Theorem 2.1, where m is an integer greater than 1.

First, we aim to define the positive real numbers $\sigma_1, \sigma_2, \ldots, \sigma_{m-1}$, given any $m \times 1$ vector p_0 whose entries are all positive. We define D to be the diagonal

$m \times m$ matrix

$$D_{j,k} = \begin{cases} \frac{1}{(p_0)_j}, & j = k \\ 0, & j \neq k \end{cases} \tag{5}$$

for $j, k = 1, 2, \ldots, m$. We define H to be the $m \times m$ matrix

$$H_{j,k} = \begin{cases} 1 - \frac{1}{m}, & j = k \\ -\frac{1}{m}, & j \neq k \end{cases} \tag{6}$$

for $j, k = 1, 2, \ldots, m$. Note that H is an orthogonal projector. We define $B = HDH$, so that B is the self-adjoint $m \times m$ matrix

$$B_{j,k} = \begin{cases} \frac{1}{(p_0)_j} - \frac{1}{m}\left(\frac{1}{(p_0)_j} + \frac{1}{(p_0)_k}\right) + \frac{1}{m^2}\sum_{l=1}^{m} \frac{1}{(p_0)_l}, & j = k \\ -\frac{1}{m}\left(\frac{1}{(p_0)_j} + \frac{1}{(p_0)_k}\right) + \frac{1}{m^2}\sum_{l=1}^{m} \frac{1}{(p_0)_l}, & j \neq k \end{cases} \tag{7}$$

for $j, k = 1, 2, \ldots, m$. As a self-adjoint matrix whose rank is $m - 1$ (after all, $B = HDH$, H is an orthogonal projector whose rank is $m - 1$, and D is a full-rank diagonal matrix), B given in (7) has an eigendecomposition

$$B = Q \Lambda Q^{\top}, \tag{8}$$

where Q is a real unitary $m \times m$ matrix and Λ is a diagonal $m \times m$ matrix such that $\Lambda_{m,m} = 0$. Finally, we define the positive real numbers $\sigma_1, \sigma_2, \ldots, \sigma_{m-1}$ via the formula

$$\sigma_k^2 = 1 / \Lambda_{k,k} \tag{9}$$

for $k = 1, 2, \ldots, m - 1$, where $\Lambda_{1,1}, \Lambda_{2,2}, \ldots, \Lambda_{m,m}$ are the diagonal entries of Λ from the eigendecomposition (8).

Next, we define the real numbers $\zeta_1, \zeta_2, \ldots, \zeta_{m-1}$, given both p_0 and an $m \times 1$ vector a such that $\sum_{k=1}^{m} a_k = 0$. We define the $(m-1) \times 1$ vector

$$\eta = \tilde{Q}^{\top} a, \tag{10}$$

where $\tilde{Q}$ is the leftmost $m \times (m-1)$ block of Q from the eigendecomposition (8), that is, $\tilde{Q}$ is the same as Q after deleting the last column of Q. We can then define the real numbers $\zeta_1, \zeta_2, \ldots, \zeta_{m-1}$ via the formula

$$\zeta_k = \eta_k / \sigma_k \tag{11}$$

for $k = 1, 2, \ldots, m - 1$, where η is defined in (10) and σ is defined in (9).

With this notation, we can state the following special case of Theorem 4.2 of Moore and Spruill (1975).

Theorem 2.1. *Suppose that m is an integer greater than one, p_0 is a probability distribution over m bins (that is, p_0 is an $m \times 1$ vector whose entries are all positive and $\sum_{k=1}^{m} (p_0)_k = 1$), a is an $m \times 1$ vector such that $\sum_{k=1}^{m} a_k = 0$, and*

$(Y_n)_1$, $(Y_n)_2$, ..., $(Y_n)_m$ are the proportions of draws falling in bins 1, 2, ..., m, respectively, out of a total of n i.i.d. draws from the probability distribution

$$p_a = p_0 + a/\sqrt{n}. \tag{12}$$

Suppose further that X_n is the random variable

$$X_n = n \sum_{k=1}^{m} ((Y_n)_k - (p_0)_k)^2. \tag{13}$$

Then, X_n converges in distribution to the random variable

$$X_\infty = \sum_{k=1}^{m-1} \sigma_k^2 \, (Z_k + \zeta_k)^2 \tag{14}$$

as n becomes large, where Z_1, Z_2, ..., Z_{m-1} are i.i.d. Gaussian random variables of zero mean and unit variance, σ_1, σ_2, ..., σ_{m-1} are the positive real numbers defined in (9), and ζ_1, ζ_2, ..., ζ_{m-1} are the real numbers defined in (11). The values of σ_1, σ_2, ..., σ_{m-1} do not depend on the vector a; the values of ζ_1, ζ_2, ..., ζ_{m-1} do depend on a.

Remark 2.2. The $m \times m$ matrix B defined in (7) is the sum of a diagonal matrix and a low-rank matrix. The methods of Gu and Eisenstat (1994, 1995) for computing the eigenvalues of such a matrix B and computing the result of applying $Q^\top$ from (8) to an arbitrary vector require only either $\mathcal{O}(m^2)$ or $\mathcal{O}(m \log(m))$ floating-point operations. The $\mathcal{O}(m^2)$ methods of Gu and Eisenstat (1994, 1995) are usually more efficient than the $\mathcal{O}(m \log(m))$ method of Gu and Eisenstat (1995), unless m is impractically large.

3. Integral representations

This section describes efficient algorithms for evaluating the cdf of the sum (14) of the squares of independent Gaussian random variables. The bibliography of Duchesne and de Micheaux (2010) gives references to possible alternatives to the methods of the present section. Our principal tool is the following theorem, representing the cdf as an integral suitable for evaluation via quadratures (see, for example, Remark 3.2 below); the theorem expresses formula 7 of Rice (1980) in the same form as formula 8 of Perkins, Tygert, and Ward (2011b).

Theorem 3.1. *Suppose that ℓ is a positive integer, Z_1, Z_2, ..., Z_ℓ are i.i.d. Gaussian random variables of zero mean and unit variance, and σ_1, σ_2, ..., σ_ℓ and ζ_1, ζ_2, ..., ζ_ℓ are real numbers. Suppose in addition that X is the random variable*

$$X = \sum_{k=1}^{\ell} \sigma_k^2 \, (Z_k + \zeta_k)^2. \tag{15}$$

Then, the cdf F of X is

$$F(x) = \int_0^\infty \mathrm{Im} \left(\frac{e^{1-y} \, e^{iy\sqrt{\ell}} \, \prod_{k=1}^{\ell} e^{\zeta_k^2 (1 - w_k(y))/(2 w_k(y))}}{\pi \left(y - \frac{1}{1 - i\sqrt{\ell}} \right) \prod_{k=1}^{\ell} \sqrt{w_k(y)}} \right) dy \tag{16}$$

for any positive real number x, where

$$w_k(y) = 1 - 2(y-1)\sigma_k^2/x + 2iy\sigma_k^2\sqrt{\ell}/x, \tag{17}$$

and $F(x) = 0$ for any nonpositive real number x. The square roots in (16) denote the principal branch, and Im *takes the imaginary part.*

Remark 3.2. An efficient means of evaluating (16) numerically is to employ adaptive Gaussian quadratures; see, for example, Section 4.7 of Press et al. (2007). Good choices for the lowest orders of the quadratures used in the adaptive Gaussian quadratures are 10 and 21, for double-precision accuracy.

The remainder of the present section (particularly Remark 3.5) discusses the numerical stability of the method of Remark 3.2 and recalls an alternative integral representation suitable for use when the method of Remark 3.2 is not guaranteed to be numerically stable. The following lemma, proven in Remark 3.2 of Perkins, Tygert, and Ward (2011b), ensures that the denominator in (16) is not too small.

Lemma 3.3. *Suppose that ℓ is a positive integer, and r_1, r_2, ..., r_ℓ and y are positive real numbers. Suppose further that (in parallel with formula (17) above)*

$$w_k = 1 - r_k(y-1) + r_k iy\sqrt{\ell} \tag{18}$$

for $k = 1, 2, \ldots, \ell$.

Then,

$$\left| \prod_{k=1}^{\ell} \sqrt{w_k} \right| > e^{-1/4}. \tag{19}$$

The following lemma ensures that the numerator in (16) is not too large, provided that $e^{\zeta_k^2/2}$ is not large.

Lemma 3.4. *Suppose that r, y, and ℓ are positive real numbers and (in parallel with formulae (17) and (18) above)*

$$w = 1 - r(y-1) + riy\sqrt{\ell}. \tag{20}$$

Then,

$$\left| \frac{1-w}{w} \right| \le \sqrt{1 + \frac{1}{\ell}}. \tag{21}$$

Proof. Defining

$$z = \frac{1}{y} \tag{22}$$

and

$$c = 1 + \frac{1}{r}, \tag{23}$$

we obtain that

$$\frac{1-w}{w} = -\frac{1-z-i\sqrt{\ell}}{1-cz-i\sqrt{\ell}}. \tag{24}$$

It follows from (24) that

$$\left|\frac{1-w}{w}\right|^2 = \frac{(1-z)^2+\ell}{(1-cz)^2+\ell}. \tag{25}$$

It follows from (22) that $z \geq 0$ and from (23) that $c \geq 1$, and hence

$$cz - 1 \geq z - 1. \tag{26}$$

If $z \geq 1$, then (26) yields that

$$(cz-1)^2 \geq (z-1)^2, \tag{27}$$

which in turn yields that

$$\frac{(1-z)^2+\ell}{(1-cz)^2+\ell} \leq \frac{(1-z)^2+\ell}{(1-z)^2+\ell} = 1. \tag{28}$$

If $z \leq 1$, then (recalling that $z \geq 0$, too)

$$\frac{(1-z)^2+\ell}{(1-cz)^2+\ell} \leq \frac{(1-z)^2+\ell}{\ell} \leq \frac{1+\ell}{\ell}. \tag{29}$$

We see from (28) and (29) that, in all cases,

$$\frac{(1-z)^2+\ell}{(1-cz)^2+\ell} \leq 1 + \frac{1}{\ell}. \tag{30}$$

Combining (25) and (30) yields (21). □

Remark 3.5. The bound (19) shows that the integrand in (16) is not too large for any nonnegative y, provided that the numerator of (16) is not too large. An upper bound on the numerator follows immediately from (21):

$$\left|\prod_{k=1}^{\ell} e^{\zeta_k^2(1-w_k(y))/(2w_k(y))}\right| \leq \prod_{k=1}^{\ell} e^{\zeta_k^2\sqrt{1+1/\ell}/2}. \tag{31}$$

For any particular application, we can check that the right-hand side of (31) is not too many orders of magnitude in size, guaranteeing that applying quadratures to the integral in (16) cannot lead to catastrophic cancellation in floating-point arithmetic. Naturally, it is also possible to check on the magnitude of the integrand in (16) during its numerical evaluation, indicating even better numerical stability than guaranteed by our *a priori* estimates. See Theorem 3.7 and Remark 3.8 below for an alternative integral representation suitable for use when the right-hand side of (31) is large.

Remark 3.6. The bound in (31) is quite pessimistic. In fact, the real part of $(1 - w_k(y))/(2w_k(y))$ is often nonpositive, so that

$$\left| e^{\zeta_k^2 (1 - w_k(y))/(2w_k(y))} \right| \leq 1. \tag{32}$$

If the right-hand side of (31) is large, then we can use the method of Imhof (1961), Davies (1980), and others, applying numerical quadratures to the integral in the following theorem. Please note that the integrand in the following theorem decays reasonably fast when the right-hand side of (31) is large.

Theorem 3.7. *Suppose that ℓ is a positive integer, Z_1, Z_2, ..., Z_ℓ are i.i.d. Gaussian random variables of zero mean and unit variance, and σ_1, σ_2, ..., σ_ℓ and ζ_1, ζ_2, ..., ζ_ℓ are real numbers. Suppose in addition that X is the random variable*

$$X = \sum_{k=1}^{\ell} \sigma_k^2 \, (Z_k + \zeta_k)^2. \tag{33}$$

Then, the cdf F of X is

$$F(x) = \frac{1}{2} - \int_0^\infty \mathrm{Im} \left(\frac{e^{-iy} \prod_{k=1}^{\ell} e^{\zeta_k^2 (1 - v_k(y))/(2v_k(y))}}{\pi y \prod_{k=1}^{\ell} \sqrt{v_k(y)}} \right) dy \tag{34}$$

for any positive real number x, where

$$v_k(y) = 1 - 2iy\sigma_k^2/x, \tag{35}$$

and $F(x) = 0$ for any nonpositive real number x. The square roots in (34) denote the principal branch, and Im *takes the imaginary part.*

Remark 3.8. The integrand in (34) is not too large (except for values of y that are closer to 0 than are typical quadrature nodes), since the real part of $(1 - v_k(y))/(2v_k(y))$ is always nonpositive, so that

$$\left| e^{\zeta_k^2 (1 - v_k(y))/(2v_k(y))} \right| \leq 1. \tag{36}$$

Moreover, the numerator in (34) decays reasonably fast (it is sub-Gaussian) when the right-hand side of (31) is large.

4. Numerical method

Combining Sections 2 and 3 yields an efficient method for calculating the cdf F of n times the square of the Euclidean distance between the model and empirical distributions, in the limit that n is large, when the n observed draws are taken i.i.d. from an alternative distribution $p_a = p_0 + a/\sqrt{n}$ (as always, p_0 is the model — a probability distribution over m bins — and a is a vector whose m entries satisfy $\sum_{k=1}^{m} a_k = 0$). Indeed, Theorem 2.1 shows that the desired F is the same as that in (16) and (34), with the real numbers σ_1, σ_2, ..., σ_ℓ and ζ_1, ζ_2, ..., ζ_ℓ

calculated as detailed in Section 2 (identifying $\ell = m-1$). Remark 3.2 describes an efficient means of evaluating $F(x)$ in (16) that is numerically stable when the right-hand side of (31) is not too many orders of magnitude in size. When the right-hand side of (31) is many orders of magnitude in size, we can apply quadratures to the representation of $F(x)$ in (34) instead (see Remark 3.8).

5. Plotting the asymptotic statistical power

Let us denote by π the cdf of the P-values for the Euclidean distance (or, equivalently, for any positive multiple of the square of the Euclidean distance); π is also the statistical power function of the hypothesis test based on the Euclidean distance (as a function of the significance level). The method of Section 4 is sufficient for plotting π in the limit that the number of draws is large. Indeed, suppose that X denotes n times the square of the Euclidean distance between the model and empirical distributions, F_0 denotes the cdf for X when taking n draws i.i.d. from the model probability distribution p_0, and F_a denotes the cdf for X when taking n draws i.i.d. from $p_a = p_0 + a/\sqrt{n}$, where a is a vector whose m entries satisfy $\sum_{k=1}^m a_k = 0$. The P-value P equals $1 - F_0(X)$, in the limit that n is large, and then the cdf π of the P-values for draws from p_a is

$$\begin{aligned}\pi(1 - F_0(x)) = \text{Prob}\{P \le 1 - F_0(x)\} &= \text{Prob}\{1 - F_0(X) \le 1 - F_0(x)\} \\ &= \text{Prob}\{X \ge x\} = 1 - F_a(x) \quad (37)\end{aligned}$$

for any nonnegative real number x; thus, the graph of all points $(\alpha, \pi(\alpha))$ with α ranging from 0 to 1 is the same as the graph of all points $(1-F_0(x), 1-F_a(x))$ with x ranging from 0 to ∞, in the limit that n is large. Section 4 describes how to evaluate $F_0(x)$ and $F_a(x)$ for any real number x, in the limit that the number n of draws is large; note that $F_0(x) = F_a(x)$ when the entries of a are all zeros, so the procedure of Section 4 can evaluate $F_0(x)$ as well as $F_a(x)$. When the entries of a are all zeros, $\zeta_1 = \zeta_2 = \cdots = \zeta_\ell = 0$ in the method of Section 4, and then the right-hand side of (31) is exactly 1.

6. Numerical examples

This section illustrates the algorithms of the present paper via several numerical examples. As detailed in the subsections below, we consider three examples for the model p_0 (as always, p_0 is a probability distribution over m bins, that is, a vector whose entries are all positive and satisfy $\sum_{k=1}^m (p_0)_k = 1$), taking n i.i.d. draws from the alternative probability distribution

$$p_a = p_0 + a/\sqrt{n}, \quad (38)$$

where a is a vector whose m entries satisfy $\sum_{k=1}^m a_k = 0$ (the subsections below detail several examples for a). Figure 1 plots the cdf π of the P-values for n i.i.d. draws taken from the alternative distribution p_a, when n is large; π is also the

statistical power function of the hypothesis test based on the Euclidean distance (as a function of the significance level). For each of the examples, Figure 1 plots the cdf π both for $n = 1{,}000{,}000$ draws (computed via Monte-Carlo simulations) and in the limit that n is large (computed via the algorithms of the present paper); not surprisingly, there is little difference between the plots for $n = 1{,}000{,}000$ and for the limit that n is large. The lines in Figure 1 corresponding to $n = 1{,}000{,}000$ draws are colored green; the lines corresponding to the limit of large n are black.

Remark 6.1. For each example, we computed the cdf π for $n = 1{,}000{,}000$ draws via 40,000 Monte-Carlo simulations. A straightforward argument based on the binomial distribution, detailed in Remark 3.4 of Perkins, Tygert, and Ward (2011a), shows that the standard errors of the resulting estimates of the P-values P are equal to $\sqrt{P(1-P)/40000} \leq 0.0025$, ensuring that the standard errors of the plotted abscissae α for the green points in Figure 1 are approximately $\sqrt{\alpha(1-\alpha)/40000} \leq 0.0025$ (roughly the size of the radii of the plotted points).

Remark 6.2. For each example, we plotted the cdf π in the limit of a large number n of draws via the scheme of Section 5. Figure 1 displays the points $(\alpha, \pi(\alpha)) = (1 - F_0(x), 1 - F_a(x))$ for the 10000 values $x = 1/2000, 2/2000, \ldots, 10000/2000$, in the limit that the number n of draws is large, where $F_0(x)$ and $F_a(x)$ are defined in Section 5 and computed to at least 6-digit accuracy via the method of Section 4.

Table 1 summarizes computational costs of the procedure described in Section 4. The headings of Table 1 have the following meanings:

- m is the number of bins in the probability distributions p_0 and p_a.
- q_0 is the maximum number of quadrature nodes required in any of the 10000 evaluations of F_0 plotted in Figure 1 (Section 5 defines F_0), using adaptive Gaussian quadratures as described in Remark 3.2.
- q_a is the maximum number of quadrature nodes required in any of the 10000 evaluations of F_a plotted in Figure 1 (Section 5 defines F_a), using adaptive Gaussian quadratures as described in Remark 3.2.
- t is the time in seconds required to perform the quadratures for both $F_0(x)$ and $F_a(x)$ at a single value of x, amortized over the 10000 pairs $(1 - F_0(x), 1 - F_a(x))$ plotted in Figure 1 (Section 5 defines F_0 and F_a).

6.1. Uniform model

For our first example, we take

$$(p_0)_k = 1/10 \tag{39}$$

for $k = 1, 2, \ldots, 10$, and take

$$a_k = (-1)^k/5 \tag{40}$$

for $k = 1, 2, \ldots, 10$. The Euclidean distance is equivalent to the canonical χ^2 statistic for this example, since p_0 is a uniform distribution.

6.2. Nonuniform model

For our second example, we take

$$(p_0)_k = \begin{cases} 1/2, & k = 1 \\ 1/198, & k = 2, 3, \ldots, 100 \end{cases} \tag{41}$$

for $k = 1, 2, \ldots, 100$, and take

$$a_k = \begin{cases} 2/3, & k = 1 \\ -2/297, & k = 2, 3, \ldots, 100 \end{cases} \tag{42}$$

for $k = 1, 2, \ldots, 100$.

6.3. Poisson model

For our third example, we take

$$(p_0)_k = e^{-3}\, 3^{k-1}/(k-1)! \tag{43}$$

for $k = 1, 2, 3, \ldots$, and take

$$a_k = \begin{cases} (-1)^k/4, & k = 1, 2, 3, 4 \\ (-1)^k/2, & k = 5, 6 \\ 0, & k = 7, 8, 9, \ldots \end{cases} \tag{44}$$

for $k = 1, 2, 3, \ldots$. For all numerical computations associated with this example, we can truncate to the first 20 bins, since $\sum_{k=21}^{\infty} (p_0)_k < 10^{-10}$.

6.4. Poisson model with a different alternative

For our fourth example, we again take

$$(p_0)_k = e^{-3}\, 3^{k-1}/(k-1)! \tag{45}$$

for $k = 1, 2, 3, \ldots$, but now take

$$a_k = \begin{cases} 1, & k = 1 \\ -1/11, & k = 2, 3, \ldots, 12 \\ 0, & k = 13, 14, 15, \ldots \end{cases} \tag{46}$$

for $k = 1, 2, 3, \ldots$. For all numerical computations associated with this example, we can truncate to the first 20 bins, since $\sum_{k=21}^{\infty} (p_0)_k < 10^{-10}$.

Remark 6.3. The right-hand side of (31) is 8.233 for Subsection 6.1, 2.443 for Subsection 6.2, and 24.05 for Subsection 6.3. As discussed in Remark 3.5, roundoff errors in the numerical evaluation of (16) are therefore guaranteed to be negligible for the standard floating-point arithmetic (the mantissa in the standard, double-precision arithmetic has a dynamic range of about $5 \cdot 10^{15} \gg 24.05$). The right-hand side of (31) is $1.478 \cdot 10^{16}$ for Subsection 6.4, so we used (34) rather than (16) for the last example (Remark 3.8 explains why).

TABLE 1
Computational costs

	m	q_0	q_a	t
example 1	10	230	230	0.006
example 2	100	530	550	0.090
example 3	20	250	330	0.013
example 4	20	350	350	0.010

We used Fortran 77 and ran all examples on one core of a 2.2 GHz Intel Core 2 Duo microprocessor with 2 MB of L2 cache. Our code is compliant with the IEEE double-precision standard (so that the mantissas of variables have approximately one bit of precision less than 16 digits, yielding a relative precision of about $2 \cdot 10^{-16}$). We diagonalized the matrix B defined in (7) using the Jacobi algorithm (see, for example, Chapter 8 of Golub and Van Loan (1996)), not taking advantage of Remark 2.2; explicitly forming the entries of the matrix B defined in (7) can incur a numerical error of at most the machine precision (about $2 \cdot 10^{-16}$) times $\max_{1 \le k \le m} (p_0)_k / \min_{1 \le k \le m} (p_0)_k$, yielding 6-digit accuracy or better for all our examples. A future article will exploit the interlacing properties of eigenvalues, following Gu and Eisenstat (1994), to obtain higher precision. Of course, even 4-digit precision would suffice for most statistical applications; however, modern computers can produce high accuracy very fast, as the examples in this section illustrate.

Acknowledgements

We would like to thank Jim Berger, Tony Cai, Jianqing Fan, Andrew Gelman, Peter W. Jones, Ron Peled, Vladimir Rokhlin, and Rachel Ward for many helpful discussions.

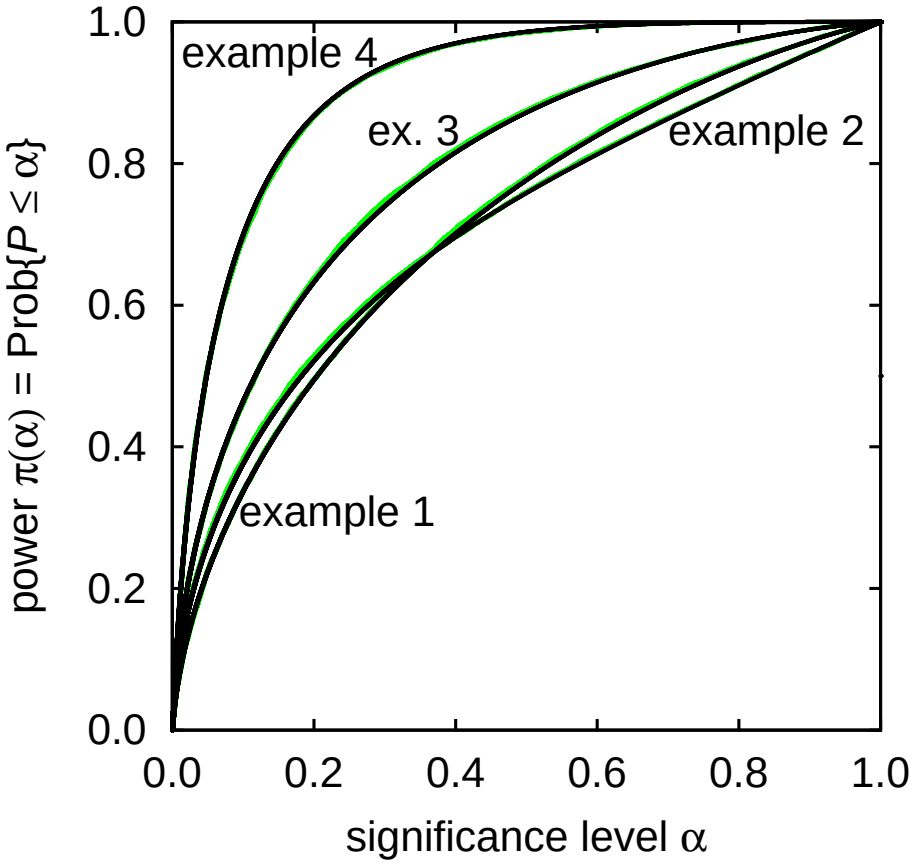

Fig. 1: Cumulative distribution functions of the P-values P for draws from the alternative distributions defined in Subsections 6.1–6.4

References

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *J. Roy. Statist. Soc. Ser. C* **29** 323–333.

Duchesne, P. and de Micheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Comput. Statist. Data Anal.* **54** 858–862.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins University Press, Baltimore, Maryland.

Gu, M. and Eisenstat, S. C. (1994). A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem. *SIAM J. Matrix Anal. Appl.* **15** 1266–1276.

Gu, M. and Eisenstat, S. C. (1995). A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM J. Matrix Anal. Appl.* **16** 172–191.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 419–426.

Kendall, M. G., Stuart, A., Ord, K. and Arnold, S. (2009). *Kendall's Advanced Theory of Statistics* **2A**, 6th ed. Wiley.

Moore, D. S. and Spruill, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Ann. Statist.* **3** 599–616.

Perkins, W., Tygert, M. and Ward, R. (2011a). χ^2 and classical exact tests often wildly misreport significance; the remedy lies in computers. Technical Report No. 1108.4126, arXiv. http://cims.nyu.edu/∼tygert/abbreviated.pdf.

Perkins, W., Tygert, M. and Ward, R. (2011b). Computing the confidence levels for a root-mean-square test of goodness-of-fit. *Appl. Math. Comput.* **217** 9072–9084.

Perkins, W., Tygert, M. and Ward, R. (2011c). Computing the confidence levels for a root-mean-square test of goodness-of-fit, II Technical Report No. 1009.2260, arXiv.

Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. (2007). *Numerical Recipes*, 3rd ed. Cambridge University Press, Cambridge, UK.

Rao, C. R. (2002). Karl Pearson chi-square test: The dawn of statistical inference. In *Goodness-of-Fit Tests and Model Validity* (C. Huber-Carol, N. Balakrishnan, M. S. Nikulin and M. Mesbah, eds.) 9–24. Birkhäuser, Boston.

Rice, S. O. (1980). Distribution of quadratic forms in normal random variables — Evaluation by numerical integration. *SIAM J. Sci. Stat. Comput.* **1** 438–448.