Stat Comput
DOI 10.1007/s11222-016-9701-y

Approximations for weighted Kolmogorov–Smirnov distributions via boundary crossing probabilities

Nino Kordzakhia[1] · **Alexander Novikov**[2,3] · **Bernard Ycart**[4]

Received: 3 March 2016 / Accepted: 2 September 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract A statistical application to Gene Set Enrichment Analysis implies calculating the distribution of the maximum of a certain Gaussian process, which is a modification of the standard Brownian bridge. Using the transformation into a boundary crossing problem for the Brownian motion and a piecewise linear boundary, it is proved that the desired distribution can be approximated by an n-dimensional Gaussian integral. Fast approximations are defined and validated by Monte Carlo simulation. The performance of the method for the genomics application is discussed.

Keywords Boundary crossing · P value approximation · Gene set enrichment analysis

Mathematics Subject Classification 62G10 · 60J65

✉ Bernard Ycart
bernard.ycart@imag.fr

Nino Kordzakhia
nino.kordzakhia@mq.edu.au

Alexander Novikov
Alex.Novikov@uts.edu.au

1 Macquarie University, Balaclava Road, North Ryde, NSW 2109, Australia

2 School of Mathematical and Physical Sciences, University of Technology, Broadway, PO Box 123, Sydney, NSW 2007, Australia

3 Steklov Institute of Mathematics, Gubkina str. 8, 119991 Moscow, Russia

4 Laboratoire Jean Kuntzmann, Université Grenoble Alpes, 51 rue des mathématiques, 38041 Grenoble cedex, France

1 Introduction

The subject of this paper is the computation of the distribution of the following random variable:

$$D_g = \max_{0 \leqslant t \leqslant 1} X_t \ , \tag{1}$$

where $\{X_t \, , 0 \leqslant t \leqslant 1\}$ is a continuous centered Gaussian process with covariance function:

$$R_X(t,s) = \min(t,s) - ts + g(t)g(s) \ , \tag{2}$$

the function g being continuous on the interval $[0,1]$ and such that $g(0) = g(1) = 0$.

This type of problem recently arose in a statistical application to Gene Set Enrichment Analysis (Charmpi and Ycart 2015). It is important to note that it is different from similar problems with the look-alike covariance function:

$$R_Y(t,s) = \min(t,s) - ts - g(t)g(s) \ . \tag{3}$$

The latter appears in the vast literature devoted to goodness-of-fit tests when parameters are estimated: see del Barrio (2007), Parker (2013), and references therein.

Throughout the paper, $W = \{W_t \, , t \geqslant 0\}$ denotes a standard one-dimensional Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$, $B = \{B_t = W_t - tW_1 \, , \ 0 \leqslant t \leqslant 1\}$ is the corresponding Brownian bridge, and ξ is a standard Gaussian random variable, independent from B. A centered Gaussian process X with covariance function (2) can be represented on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ as follows:

$$X = \{X_t = B_t - g(t)\xi \ , \ 0 \leqslant t \leqslant 1\} \ . \tag{4}$$

Observe that $X_0 = X_1 = 0$. The tail probability of D_g at $x \geqslant 0$ will be denoted by $p_g(x)$.

Published online: 15 September 2016

$$p_g(x) = \mathbb{P}[D_g > x] = \mathbb{P}\left[\max_{0\leqslant t\leqslant 1} X_t > x\right] . \tag{5}$$

The following family of functions g is of special relevance to the genomics application:

$$g_a(t) = (t^a - t) , \quad 1/2 < a < 1 . \tag{6}$$

In particular, $a = \frac{2}{3}$ corresponds to the case where gene expression ranks are tested against a given gene set: see Sect 4 for more details.

The case $g \equiv 0$ is that of the classical Kolmogorov–Smirnov test: see Durbin (1973) and Stephens (1992) for historical aspects;

$$p_0(x) = \mathrm{e}^{-2x^2} .$$

This explicit formula can be found in a personal letter from A. Kolmogorov to P. Aleksandrov written in 1931 (Shiryaev 2003, p. 436), where Kolmogorov states that he nearly proved the result. A complete derivation appeared in Smirnov (1939).

Apart from the case $g \equiv 0$, no explicit expression exists for $p_g(x)$. Our method to approximating it has already been used in the context of nonparametric testing. It consists in:

1. Reducing the problem to a nonlinear boundary crossing problem (BCP) for the Brownian motion W. This is the classical approach to extrema of modified Brownian bridges: see Durbin (1971), Krumbholz (1976), and Bischoff et al. (2003); but analytic results for nonlinear boundaries are scarce (Kahale 2008).
2. Replacing the nonlinear boundary by a piecewice linear approximation. This has been used in many papers, including Novikov et al. (1999), Pötzelberger and Wang (2001), Hashorva (2005), and Borovkov and Novikov (2005).

Other approaches include the martingale transformation method proposed by Khmaladze (1981) and of course Monte Carlo simulation. The martingale transformation method requires calculations of compensators, which are difficult to obtain in analytical form. Monte Carlo simulation was used in Charmpi and Ycart (2015). However, for both reasons of accuracy and computing cost, it cannot be considered as an efficient method, in particular in view of the genomics application, where a high throughput and a good accuracy for very small p-values are both requested.

The paper is organized as follows. Section 2 contains the theoretical results. The reduction to a nonlinear BCP and the bounding inequalities are stated as Lemmas 1 and 2. Our main result, Theorem 1 gives an explicit bound on the approximating error. An exact computing algorithm for a piecewise linear boundary is described by Proposition 1. Explicit expressions are given for the one-node case (Propositions 2 and 3). Section 3 addresses the practical issue. Two fast approximation schemes are proposed and compared to Monte Carlo simulations. Section 4 describes the statistical application which motivated the present study. Propositions 4 and 5 show that computing p-values for Gene Set Enrichment Analysis amounts to computing $p_g(x)$ for some function g depending on the genes to be tested. An example of application to real genomic data is given.

2 Theoretical results

Throughout the paper, ϕ and Φ denote the pdf and cdf of the standard Gaussian distribution, respectively;

$$\phi(y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-y^2/2} \quad \text{and} \quad \Phi(y) = \int_{-\infty}^{y} \phi(z)\,\mathrm{d}z .$$

We begin with the transformation into a boundary crossing problem.

Lemma 1 *Denote by G the function defined on $(0, +\infty)$ by:*

$$G(s) = (s+1)\, g\left(\frac{s}{s+1}\right) , \quad 0 \leqslant s < \infty . \tag{7}$$

For $x \geqslant 0$ and $y \in \mathbb{R}$, denote by $S(x, y, G)$ the kernel:

$$S(x, y, G) = \mathbb{P}\left[\sup_{0\leqslant s<\infty} \frac{W_s - G(s)y}{s+1} > x\right] . \tag{8}$$

Then:

$$p_g(x) = \int_{-\infty}^{+\infty} S(x, y, G)\,\phi(y)\mathrm{d}y . \tag{9}$$

Proof Using the representation (4),

$$p_g(x) = \mathbb{P}\left[\max_{0\leqslant t\leqslant 1} B_t - g(t)\xi > x\right] .$$

The standard Brownian bridge has the following well-known representation:

$$\{B_t, 0 \leqslant t < 1\} \stackrel{d}{=} \left\{(1-t)W_{t/(1-t)}, 0 \leqslant t < 1\right\} .$$

Hence:

$$\begin{aligned} p_g(x) &= \mathbb{P}\left[\sup_{0\leqslant t<1} (1-t)W_{t/(1-t)} - g(t)\xi > x\right] \\ &= \mathbb{P}\left[\sup_{0\leqslant s<+\infty} \frac{W_s - G(s)\xi}{s+1} > x\right] , \end{aligned}$$

from which (9) follows, choosing ξ independent from W. □

Observe that the correspondance between g and G is one-to-one. For $0 \leqslant t < 1$:

$$g(t) = (1-t)\, G\left(\frac{t}{1-t}\right) . \tag{10}$$

In the particular case where g_a is defined by (6), one gets:

$$G_a(s) = (s+1)^{1-a} s^a - s . \tag{11}$$

Obviously, the definition of $S(x, y, G)$ is monotonic in G: for given x and y, raising the boundary can only decrease the crossing probability. This translates into the following inequalities.

Lemma 2 *Let G_l and G_u be two continuous functions defined on $[0, +\infty)$, such that for $0 \leqslant s < \infty$,*

$$G_l(s) \leqslant G(s) \leqslant G_u(s) .$$

Then:

$$p_g(x) \geqslant \int_{-\infty}^{0} S(x, y, G_l)\, \phi(y) \mathrm{d}y + \int_{0}^{+\infty} S(x, y, G_u)\, \phi(y) \mathrm{d}y ,$$

and

$$p_g(x) \leqslant \int_{-\infty}^{0} S(x, y, G_u)\, \phi(y) \mathrm{d}y + \int_{0}^{+\infty} S(x, y, G_l)\, \phi(y) \mathrm{d}y .$$

Proof For $0 \leqslant s < \infty$ and $y \leqslant 0$, $yG_u(s) \leqslant yG(s) \leqslant yG_l$. Hence for all $x \geqslant 0$,

$$S(x, y, G_l) \leqslant S(x, y, G) \leqslant S(x, y, G_u) .$$

The inequalities above are reversed for $y \geqslant 0$. Hence the result. □

Once a lower bound and an upper bound are given, the question arises naturally to control the approximation error in terms of a certain distance. This is the object of the following theorem.

Theorem 1 *For $i = 1, 2$, let g_i be a continuous function defined on $[0,1]$, derivable on $(0,1)$, such that $g_i(0) = g_i(1) = 0$. Denote by G_i its transform through the time change $t \mapsto s = \frac{t}{1-t}$ (formula (7)). Denote by $\Delta(G_1, G_2)$ the following distance:*

$$\Delta(G_1, G_2) = \int_0^{+\infty} \left(\frac{\mathrm{d}}{\mathrm{d}s}(G_1(s) - G_2(s))\right)^2 \mathrm{d}s . \tag{12}$$

Then for all $x \in \mathbb{R}$,

$$|S(x, y, G_1) - S(x, y, G_2)| \leqslant 4\Phi\left(\tfrac{|y|}{2}\sqrt{\Delta(G_1, G_2)}\right) - 2 , \tag{13}$$

and:

$$\left|p_{g_1}(x) - p_{g_2}(x)\right| \leqslant \frac{4}{\pi} \arctan\left(\frac{1}{2}\sqrt{\Delta(G_1, G_2)}\right) \leqslant \frac{2}{\pi}\sqrt{\Delta(G_1, G_2)} . \tag{14}$$

Proof Although the setting is different from that of Theorem 1 in Novikov et al. (1999), the proof is similar. It uses a representation of the kernel $S(x, y, G_1)$ in terms of the BCP from G_2, through Girsanov's theorem. Define the random variable ζ as

$$\zeta = \int_0^{+\infty} \left(\frac{\mathrm{d}}{\mathrm{d}s}(G_1(s) - G_2(s))\right)^2 \mathrm{d}W_s . \tag{15}$$

It turns out that

$$S(x, y, G_1) = \mathbb{E}\left[\mathbb{I}\left(\sup_{0 \leqslant s < \infty} \frac{W_s - G_2(s) y}{s+1} > x\right) \mathrm{e}^{-y\zeta - y^2 \Delta(G_1, G_2)/2}\right] , \tag{16}$$

where $\mathbb{E}$ denotes the mathematical expectation with respect to $\mathbb{P}$ and $\mathbb{I}$ the indicator of an event.

To prove (16), consider the probability measure on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\})$ defined by:

$$\widetilde{\mathbb{P}}[A] = \mathbb{I}(A)\ \mathrm{e}^{-y\zeta - y^2 \Delta(G_1, G_2)/2} . \tag{17}$$

By Girsanov's theorem, the Brownian motion $\{W_t\ ,\ t \geqslant 0\}$ has drift $y(G_2(t) - G_1(t))$ with respect to $\widetilde{\mathbb{P}}$. This implies:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}\left(\sup_{0 \leqslant s < \infty} \frac{W_s - G_2(s) y}{s+1} > x\right) \mathrm{e}^{-y\zeta - y^2 \Delta(G_1, G_2)/2}\right] \\
&= \widetilde{\mathbb{E}}\left[\mathbb{I}\left(\sup_{0 \leqslant s < \infty} \frac{\widetilde{W}_s + y(G_2(s) - G_1(s)) - G_2(s) y}{s+1} > x\right)\right] \\
&= \widetilde{\mathbb{E}}\left[\mathbb{I}\left(\sup_{0 \leqslant s < \infty} \frac{\widetilde{W}_s - G_1(s) y}{s+1} > x\right)\right] \\
&= \mathbb{E}\left[\mathbb{I}\left(\sup_{0 \leqslant s < \infty} \frac{W_s - G_1(s) y}{s+1} > x\right)\right] \\
&= S(x, y, G_1) ,
\end{aligned}$$

denoting by $\widetilde{\mathbb{E}}$ the mathematical expectation and by $\widetilde{W}$ the standard Brownian motion with respect to $\widetilde{\mathbb{P}}$.

The representation (16) will now be used to bound the difference between $S(x, y, G_1)$ and $S(x, y, G_2)$. Indeed:

$$\begin{aligned}
&|S(x, y, G_1) - S(x, y, G_2)| \\
&= \left|\mathbb{E}\left[\mathbb{I}\left(\sup_{0\leqslant s<\infty} \frac{W_s - G_2(s)y}{s+1} > x\right)\right.\right. \\
&\quad \left.\left.\left(\mathrm{e}^{-y\zeta - y^2\Delta(G_1,G_2)/2} - 1\right)\right]\right| \\
&\leqslant \mathbb{E}\left[\mathbb{I}\left(\sup_{0\leqslant s<\infty} \frac{W_s - G_2(s)y}{s+1} > x\right)\right. \\
&\quad \left.\left|\mathrm{e}^{-y\zeta - y^2\Delta(G_1,G_2)/2} - 1\right|\right] \\
&\leqslant \mathbb{E}\left[\left|\mathrm{e}^{-y\zeta - y^2\Delta(G_1,G_2)/2} - 1\right|\right].
\end{aligned}$$

To compute the last expectation, observe that the random variable ζ is normally distributed, with mean 0 and variance $\Delta(G_1, G_2)$. Therefore:

$$\mathbb{E}\left[\mathrm{e}^{-y\zeta - y^2\Delta(G_1,G_2)/2} - 1\right] = 0 .$$

Denote by $z^+ = z\mathbb{I}(z > 0)$ the positive part. Since $|z| = 2z^+ - z$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathrm{e}^{-y\zeta - y^2\Delta(G_1,G_2)/2} - 1\right|\right] \\
&= 2\,\mathbb{E}\left[\left(\mathrm{e}^{-y\zeta - y^2\Delta(G_1,G_2)/2} - 1\right)^+\right] .
\end{aligned}$$

A straightforward calculation shows that

$$\mathbb{E}\left[\left(\mathrm{e}^{-y\zeta - y^2\Delta(G_1,G_2)/2} - 1\right)^+\right] = 2\Phi\left(|y|\tfrac{\sqrt{\Delta(G_1,G_2)}}{2}\right) - 1.$$

Hence (13), from which (14) follows because for $c > 0$,

$$\begin{aligned}
&\int_{-\infty}^{+\infty} (\Phi(c|y|) - \Phi(-c|y|))\ \phi(y)\mathrm{d}y \\
&= 2\int_0^{+\infty}\int_{-cy}^{+cy} \phi(z)\phi(y)\,\mathrm{d}z\mathrm{d}y \\
&= \frac{1}{\pi}\int_0^{+\infty}\int_{-\arctan(c)}^{+\arctan(c)} r\mathrm{e}^{-r^2/2}\mathrm{d}\theta\mathrm{d}r \\
&= \frac{2}{\pi}\arctan(c) .
\end{aligned}$$

□

Piecewise linear boundaries will now be considered.

Definition 1 Let $\mathbf{s} = (s_i)_{i=0,\ldots,n}$ be a tuple of reals such that $0 = s_0 < s_1 < \cdots < s_n$. Let $\mathbf{b} = (b_i)_{i=0,\ldots,n}$ be a tuple of reals. We call *n-node piecewise linear boundary* the function $G_{n,\mathbf{s},\mathbf{b}}$, defined on $[0, +\infty)$ by:

$$\begin{aligned}
G_{n,\mathbf{s},\mathbf{b}}(s) = &\left(\sum_{i=1}^{n}\left(b_{i-1} + \frac{b_i - b_{i-1}}{s_i - s_{i-1}}(s - s_{i-1})\right)\right. \\
&\left.\mathbb{I}(s_{i-1} \leqslant s < s_i)\right) + b_n\mathbb{I}(s_n \leqslant s) . \qquad (18)
\end{aligned}$$

An obvious choice for approximating a given function G is to define $b_i = G(s_i)$, for $i = 0, \ldots, n$. If G is concave, then $G_{n,\mathbf{s},\mathbf{b}}(s) \leqslant G(s)$, for all s; this is the case for G_a defined by (11). Assuming moreover that G has a continuous second derivative such that

$$\sup_{0<s<+\infty} |G''(s)| = M < +\infty ,$$

the distance $\Delta(G, G_{n,\mathbf{s},\mathbf{b}})$ can be bounded as follows.

$$\begin{aligned}
\Delta(G, G_{n,\mathbf{s},\mathbf{b}}) \leqslant 4M^2\sum_{i=1}^{n}(s_i - s_{i-1})^3 \\
+\int_{s_n}^{+\infty}\left(\frac{\mathrm{d}}{\mathrm{d}s}(G(s))\right)^2 \mathrm{d}s .
\end{aligned}$$

Provided the derivative of G is square-integrable, it follows from Theorem 1 that the approximation is numerically consistent. Indeed taking for instance $s_i - s_{i-1} = \log(n)/n$, one gets that $\Delta(G, G_{n,\mathbf{s},\mathbf{b}})$ tends to 0 as n tends to infinity. The interest of piecewise linear boundaries for our problem lies in the following result.

Proposition 1 *Let $G_{n,\mathbf{s},\mathbf{b}}$ be defined by (18). For all $x > 0$, and $y \in \mathbb{R}$,*

$$\begin{aligned}
&S(x, y, G_{n,\mathbf{s},\mathbf{b}}) \\
&= 1 - \mathbb{E}\left[\left(\prod_{i=1}^{n}\left(1 - \exp\left(-\frac{2}{s_{i+1} - s_i}(b_{i-1}y + x(1 + s_{i-1})\right.\right.\right.\right. \\
&\quad \left.\left.\left. -W_{s_{i-1}})^+((b_i - b_{i-1})y + x(1 + s_i) - W_{s_i})^+\right)\right)\right) \\
&\quad \left.\left(1 - \exp\left(-2x\left(b_n y + x(1 + s_n) - W_{s_n}\right)^+\right)\right)\right].
\end{aligned}$$

Proof A more detailed derivation of a similar formula can be found in Novikov et al. (1999). The following ingredients are used.

1. Given the values of $(W_{s_i})_{i=1,\ldots,n}$, the processes $\{W_s - W_{s_{i-1}},\ s_{i-1} \leqslant s \leqslant s_i\}$ for $i = 1, \ldots, n$, and $\{W_s - W_{s_n},\ s_n \leqslant s\}$, are conditionally independent; the conditional distribution of $\{W_s - W_{s_{i-1}},\ s_{i-1} \leqslant s \leqslant s_i\}$ is that of a Brownian bridge; and the conditional distribution of $\{W_s - W_{s_n},\ s_n \leqslant s\}$ is that of a Brownian motion.
2. For $\alpha, \beta \in \mathbb{R}$,

$$\mathbb{P}\left[\sup_{0\leqslant s<\infty}\{W_s - \alpha - \beta s\} > 0\right] = \mathrm{e}^{-2\alpha^+\beta^+} . \qquad (19)$$

Formula (19) is usually credited to Bachelier: see (Doob 1949, p. 397). □

Proposition 1 expresses $S(x, y, G_{n,\mathbf{s},\mathbf{b}})$ as an expectation with respect to the joint distribution of the Gaussian vector $(W_{s_i})_{i=1,\ldots,n}$. Using the independent increment property,

it is easy to rewrite it as an integral with respect to the n-dimensional standard Gaussian density. Denote by g_n the transform of $G_{n,\mathbf{s},\mathbf{b}}$ through (10). From Lemma 1, $p_{g_n}(x)$ has an explicit expression in terms of the $(n+1)$-dimensional standard Gaussian density. In view of Theorem 1, it can be considered that the problem is solved, at least in theory: an arbitrary close approximation of $p_g(x)$ by a an n-dimensional Gaussian integral can be computed. This is not quite so in practice, because of the computational cost of Gaussian integrals: see Gentz and Bretz (2009) as a general reference. It is therefore of interest to obtain expressions as explicit as possible, in order to reduce computing costs. Two results deduced from Proposition 1 for one-node piecewise linear boundaries follow.

Proposition 2 *Let $s_1, b_0, and b_1$ be three positive reals. Let $G_{1,\mathbf{s},\mathbf{b}}$ be defined by (18) with $\mathbf{s} = (0, s_1) and \mathbf{b} = (b_0, b_1)$. Then:*

$$\begin{aligned} S(x, y, G_{1,\mathbf{s},\mathbf{b}}) = \mathbb{I}(b_0 y + x \leqslant 0) + \mathbb{I}(b_0 y + x > 0)\Bigg(\Phi\left(-\frac{\nu_1}{\mu_1}\right) \\ + \mathrm{e}^{-\nu_1+\mu_1^2/2}\,\Phi\left(\frac{\nu_1}{\mu_1} - \mu_1\right) \\ + \mathrm{e}^{-\nu_2+\mu_2^2/2}\,\Phi\left(\frac{\nu_1}{\mu_1} - \mu_2\right) \\ - \mathrm{e}^{-(\nu_1+\nu_2)+(\mu_1+\mu_2)^2/2}\,\Phi\left(\frac{\nu_1}{\mu_1} - \mu_1 - \mu_2\right)\Bigg), \end{aligned}$$

with:

$$\begin{aligned} \mu_1 &= \sqrt{s_1}\left(\frac{2(b_0 y + x)}{s_1}\right), \\ \nu_1 &= ((b_1 - b_0)y + x(1 + s_1))\left(\frac{2(b_0 y + x)}{s_1}\right), \\ \mu_2 &= \sqrt{s_1}(2x), \\ \nu_2 &= (b_1 y + x(1 + s_1))(2x). \end{aligned}$$

Proof From Proposition 1, $S(x, y, G_{1,\mathbf{s},\mathbf{b}})$ can be written as follows:

$$S(x, y, G_{1,\mathbf{s},\mathbf{b}}) = \int_{-\infty}^{+\infty} \widetilde{S}(x, y, z)\,\phi(z)\,\mathrm{d}z,$$

with:

$$\begin{aligned} \widetilde{S}(x, y, z) = 1 - \Big(1 - \exp\big(-\frac{2}{s_1}(b_0 y + x)^+ \\ \big((b_1 - b_0)y + x(1 + s_1) + \sqrt{s_1}z)^+\big)\Big) \\ \Big(1 - \exp\big(-2x(b_1 y + x(1 + s_1) + \sqrt{s_1}z)^+\big)\Big). \end{aligned}$$

If $b_0 y + x \leqslant 0$, $\widetilde{S}(x, y, z) = 1$ for all z. Assume now $b_0 y + x > 0$;

$$\begin{aligned} &\widetilde{S}(x, y, z) \\ &= \exp\left(-\frac{2(b_0 y + x)}{s_1}((b_1 - b_0)y + x(1 + s_1) + \sqrt{s_1}z)^+\right) \\ &\quad + \exp\left(-2x(b_1 y + x(1 + s_1) + \sqrt{s_1}z)^+\right) \\ &\quad - \exp\left(-\frac{2(b_0 y + x)}{s_1}((b_1 - b_0)y + x(1 + s_1) + \sqrt{s_1}z)^+\right) \\ &\quad \exp\left(-2x(b_1 y + x(1 + s_1) + \sqrt{s_1}z)^+\right). \end{aligned}$$

Or else:

$$\begin{aligned} \widetilde{S}(x, y, z) = \exp\left(-(\mu_1 z + \nu_1)^+\right) \\ + \exp\left(-(\mu_2 z + \nu_2)^+\right) \\ - \exp\left(-(\mu_1 z + \nu_1)^+\right)\exp\left(-(\mu_2 z + \nu_2)^+\right). \end{aligned}$$

Observe that $-\frac{\nu_2}{\mu_2} < -\frac{\nu_1}{\mu_1}$. For $i = 1, 2$:

$$\begin{aligned} &\int_{-\infty}^{+\infty} \exp\{-(\mu_i z + \nu_i)^+\}\,\phi(z)\,\mathrm{d}z \\ &= \Phi\left(-\frac{\nu_i}{\mu_i}\right) + \mathrm{e}^{-\nu_i+\mu_i^2/2}\,\Phi\left(\frac{\nu_i}{\mu_i} - \mu_i\right). \end{aligned}$$

Moreover:

$$\begin{aligned} &\int_{-\infty}^{+\infty} \exp\left(-(\mu_1 z + \nu_1)^+\right)\exp\left(-(\mu_2 z + \nu_2)^+\right)\,\phi(z)\,\mathrm{d}z \\ &= \Phi\left(-\frac{\nu_2}{\mu_2}\right) + \mathrm{e}^{-\nu_2+\mu_2^2/2}\left(\Phi\left(\frac{\nu_2}{\mu_2} - \mu_2\right) - \Phi\left(\frac{\nu_1}{\mu_1} - \mu_2\right)\right) \\ &\quad + \mathrm{e}^{-(\nu_1+\nu_2)+(\mu_1+\mu_2)^2/2}\Phi\left(\frac{\nu_1}{\mu_1} - (\mu_1 + \mu_2)\right). \end{aligned}$$

Hence the result. □

Integrating $S(x, y, G_{1,\mathbf{s},\mathbf{b}})$ with respect to y against the standard Gaussian distribution can be done with reasonable precision and computing time using the Gauss-Hermite quadrature. However, the calculation for many different values of x can hardly be vectorized, which makes the whole algorithm relatively slow. It turns out that in the particular case $b_0 = 0$, the integral has an explicit expression in terms of Φ. Thus it can be computed with high accuracy in virtually null computing time, for a whole range of different values of x.

Proposition 3 *With the notations of Proposition 2 assume $b_0 = 0$. Let g_1 be the transform of $G_{1,\mathbf{s},\mathbf{b}}$ through (10). Then:*

$$\begin{aligned} p_{g_1}(x) = \Phi\left(-x\,\frac{1+s_1}{\sqrt{s_1+b_1^2}}\right) \\ + \mathrm{e}^{2x^2(b_1^2/s_1^2-1)}\,\Phi\left(-x\,\frac{1-s_1+2b_1^2/s_1}{\sqrt{s_1+b_1^2}}\right) \end{aligned}$$

$$+ \mathrm{e}^{2x^2(b_1^2-1)}\, \Phi\left(-x\,\frac{-1+s_1+2b_1^2}{\sqrt{s_1+b_1^2}}\right)$$
$$- \mathrm{e}^{2x^2(1+s_1)^2 b_1^2/s_1^2}\, \Phi\left(-x\,\frac{(1+s_1)(1+2b_1^2/s_1)}{\sqrt{s_1+b_1^2}}\right).$$

Proof Using again Proposition 1, $p_{g_1}(x)$ can be written as follows:

$$p_{g_1}(x) = \int_{\mathbb{R}^2} \widetilde{S}(x, y, z)\, \phi(y)\phi(z)\, \mathrm{d}y\mathrm{d}z\ ,$$

with:

$$\begin{aligned}\widetilde{S}(x, y, z) &= 1 - \left(1 - \exp\left(-\frac{2x}{s_1}(b_1 y + x(1+s_1) - \sqrt{s_1}z)^+\right)\right)\\ &\quad \left(1 - \exp\left(-2x(b_1 y + x(1+s_1) - \sqrt{s_1}z)^+\right)\right)\\ &= \exp\left(-\frac{2x}{s_1}(b_1 y + x(1+s_1) - \sqrt{s_1}z)^+\right)\\ &\quad + \exp\left(-2x(b_1 y + x(1+s_1) - \sqrt{s_1}z)^+\right)\\ &\quad - \exp\left(-\frac{2x(1+s_1)}{s_1}(b_1 y + x(1+s_1) - \sqrt{s_1}z)^+\right).\end{aligned}$$

Thus $p_{g_1}(x)$ is a linear combination of three integrals of the following type:

$$\int_{\mathbb{R}^2} \exp\{-(\lambda y + \mu z + \nu)^+\}\, \phi(y)\phi(z)\, \mathrm{d}y\mathrm{d}z\ ,$$

That integral is easily computed using the change of variables $(y, z) \mapsto (\lambda y + \mu z, -\mu y + \lambda z)$;

$$\begin{aligned}&\int_{\mathbb{R}^2} \exp\{-(\lambda y + \mu z + \nu)^+\}\, \phi(y)\phi(z)\, \mathrm{d}y\mathrm{d}z\\ &= \Phi\left(-\tfrac{\nu}{\sqrt{\lambda^2+\mu^2}}\right) + \mathrm{e}^{-\nu+(\lambda^2+\mu^2)/2}\, \Phi\left(\tfrac{\nu-(\lambda^2+\mu^2)}{\sqrt{\lambda^2+\mu^2}}\right)\ .\end{aligned}$$

Hence the result. □

3 Fast approximation schemes

Numerical experiments were made in R (R Development Core Team 2008). At first, a simulation procedure for the trajectories of X was implemented. A regular mesh of 10^4 discretization points in $[0,1]$ was fixed. Brownian trajectories were simulated by iteratively adding Gaussian random values along the mesh. A Brownian bridge correction for the discretization bias was applied: see Sect. 6.4 of Glasserman (2004), in particular formula (6.50) p. 367. Borovkov and Novikov (2005) give a precise evaluation of the error in Monte Carlo computation of boundary crossing probabilities.

Over 10^6 simulated trajectories, the maxima and minima were recorded, thus leading to a sample of size 2×10^6 for the variable of interest. For a given function g, we denote by $\widehat{p}_g(x)$ the empirical p-value at x calculated from the sample. For a sample size of 2×10^6, the maximal absolute difference between the empirical and the theoretical cdf's should remain below 10^{-3} to be accepted by the Kolmogorov–Smirnov test at threshold 5 %. Therefore, the target precision is

$$\|p_g - \widehat{p}_g\|_\infty = \sup_{x>0} \left|p_g(x) - \widehat{p}_g(x)\right| < 10^{-3}\ .$$

In order to validate the simulation procedure, different one-node piecewise linear boundaries were chosen; the exact p values computed from Propositions 2 and 3 were compared to the empirical p values from the sample. The absolute difference remained below 10^{-3} in all experiments, which validated both the simulation procedure, and the implementation of Propositions 2 and 3.

Two approximations of $p_g(x)$ were considered. The first one used Proposition 3. With the notations of the previous section, let s_1 and b_1 be two positive reals, $\mathbf{s} = (0, s_1) and \mathbf{b} = (0, b_1)$. Denote by g_{s_1,b_1} the transform of $G_{1,\mathbf{s},\mathbf{b}}$ through (10). The intention being to approximate $p_g(x)$ by $p_{g_{s_1,b_1}}(x)$, it is natural to choose for s_1 and b_1 the values that minimize a certain distance between g and $g_{s,b}$. Five distances were tried, among which:

$$\Delta(G, G_{1,\mathbf{s},\mathbf{b}})\ ,$$

and

$$\int_0^1 \left|g(t) - g_{s,b}(t)\right|\, \mathrm{d}t\ .$$

In view of Theorem 1, one could expect the first choice to be the best. However, experimental evidence pointed at the second choice instead. Hence the value of (s_1, b_1) was fixed at

$$(s_1, b_1) = \arg\min_{(s,b)} \int_0^1 \left|g(t) - g_{s,b}(t)\right|\, \mathrm{d}t\ . \tag{20}$$

We denote by $p_{1,g}(x)$ the p value at x calculated from Proposition 3, with (s_1, b_1) defined by (20).

$$p_{1,g}(x) = p_{g_{s_1,b_1}}(x)\ . \tag{21}$$

Our second approximation scheme relied on Lemma 2 and Proposition 2. Only one parameter had to be chosen, s_1. After numeric trials, s_1 was fixed at the point such that $g(\frac{s_1}{s_1-1})$ is maximal. Here G is assumed to be increasing, concave, and bounded. Let $b_1 = G(s_1)$, $\mathbf{s} = c(0, s_1)$, $\mathbf{b} = (0, b_1)$. Since G is concave, $G_l = G_{1,\mathbf{s},\mathbf{b}}$ is such that for all $s > 0$,

$$G_l(s) \leqslant G(s)\ .$$

For the same value of s_1, let $b_1 = \sup G$ and b_0 be such that the line from $(0, b_0)$ to (s_1, b_1) is tangent to the graph of G. Let $\mathbf{s} = (0, s_1)$ and $\mathbf{b} = (b_0, b_1)$. Then $G_u = G_{1,\mathbf{s},\mathbf{b}}$ is such that for all $s > 0$,

$$G(s) \leqslant G_u(s) \ .$$

From Lemma 2, combining the integrals of $S(x, y, G_l)$ and $S(x, y, G_u)$ over $(-\infty, 0]$ and $[0, +\infty)$ against the Gaussian distribution leads to a lower bound and an upper bound for $p_g(x)$. It is a natural choice for an approximation to use the midpoint between the lower bound and the upper bound. We denote by $p_{2,g}(x)$ the p value at x calculated as that midpoint, from Propositions 2 and 3.

$$p_{2,g}(x) = \frac{1}{2}\left(\int_{-\infty}^{+\infty} S(x, y, G_l)\,\phi(y)\mathrm{d}y + \int_{-\infty}^{+\infty} S(x, y, G_u)\,\phi(y)\mathrm{d}y\right). \tag{22}$$

The family of functions g_a from (6) was considered: $g_a(t) = t^a - t$. The values of a ranged from 0.55 to 0.95 by step 0.05. The corresponding boundaries G_a defined by (11) are increasing and concave, with $1 - a$ as a limit at $+\infty$.

$$\lim_{s\to+\infty} G_a(s) = 1 - a \ .$$

The array below gives the L_∞-distances between approximated and empirical p values, for different values of a.

a	$\lVert p_{1,g_a} - \widehat{p}_{g_a}\rVert_\infty$	$\lVert p_{2,g_a} - \widehat{p}_{g_a}\rVert_\infty$
0.55	0.00665	0.00488
0.60	0.00532	0.00362
0.65	0.00449	0.00321
0.70	0.00280	0.00161
0.75	0.00222	0.00138
0.80	0.00148	0.00096
0.85	0.00097	0.00070
0.90	0.00063	0.00067
0.95	0.00046	0.00043

Several remarks must be made. That the errors decrease as a increases to 1 was expected, since g_a becomes closer to 0. The errors are above the target 10^{-3} for a <0.85: the approximations are not perfect. However, the errors consistently remain below 10^{-2}. This may be considered acceptable, especially as the largest errors concern p values which are not statistically significant. The midpoint approximation $p_{2,g}$ is definitely better than the one-node approximation $p_{1,g}$, but not by much. A trade-off with computing time must be considered. The calculation of $p_{2,g}$ was done from Proposition 2 with a Gauss-Hermite quadrature over 64 nodes. The running time for 10^5 values of x was 18.5 s, whereas the running time for the calculation of $p_{1,g}$ is negligible (0.07 s for 10^5 values of x).

The Gauss-Hermite quadrature, even with a large number of nodes, fails to output precise evaluations of the midpoint approximation $p_{2,g}(x)$ for large values of x. On the contrary $p_{1,g}(x)$, which is a linear combination of values of Φ is accurate even for very large values of x. Another calculation can be done for large x: Durbin's approximation (see Durbin (1985) and Parker (2013) for a useful review). Let $v(t)$ denote the variance function of X: $v(t) = R_X(t, t)$, where R_X is defined by (2). Assume $v(t)$ has a unique maximum over $[0,1]$ and denote by t_0 the point at which that maximum is reached. Assume v has a continuous second derivative v''. From formula (33) of Parker (2013), Durbin's approximation is

$$p_{d,g}(x) = \frac{1}{\sqrt{2v(t_0)v''(t_0)}}\exp\left(-\frac{d^2}{2v(t_0)}\right) \ . \tag{23}$$

For the same values of a as above, Durbin's approximation $p_{d,g_a}(x)$ was compared to the one-point approximation $p_{1,g_a}(x)$ and to the empirical p values $\widehat{p}_{g_a}(x)$, for values of x such that all three p values are below 5 %. It turned out that Durbin's approximation p_{d,g_a} performed slightly better than p_{1,g_a}. For each of the two approximations, the relative error, calculated as the absolute difference with $\widehat{p}_g(x)$ divided by the same, remained smaller than 5 %, over the range of values $10^{-4} < \widehat{p}_g(x) < 10^{-2}$, where $\widehat{p}_g(x)$ could be used as an estimate of $p_g(x)$.

4 Gene set enrichment analysis

This section describes the statistical application to genomics that motivated the present work. It generalizes the description of the Weighted Kolmogorov–Smirnov test that was given in Charmpi and Ycart (2015).

Gene Set Enrichment Analysis (GSEA) was introduced in Subramanian et al. (2005). It is now generally considered as a basic tool of genomic data treatment: see Huang et al. (2009) for a review. GSEA aims at comparing a vector of numeric data indexed by the set of all genes, to the genes contained in a given smaller gene set. The numeric data are typically obtained from a microarray experiment. They may consist in expression levels, p values, correlations, fold-changes, t-statistics, signal-to-noise ratios, etc. The number associated to any given gene will be referred to as its *weight*. Many examples of such data can be downloaded from the Gene Expression Omnibus (GEO) repository (Edgar et al. (2002)). The gene set may contain genes known to be associated to a given biological process, a cellular component, a type of cancer, etc. Thematic lists of such gene sets are given in the Molecular Signature (MSig) database (Subramanian

et al. 2005). The word *enrichment* refers to the question: are the weights inside the gene set significantly larger than the weights in a random gene set of the same size?

Denote by N the total number of genes ($N \simeq 20{,}000$ for the human genome). It is convenient to identify the genes to N points on the interval $[0,1]$, and their weights to the values of some function h defined on $[0,1]$: gene number i corresponds to point i/N and its weight w_i to $h(i/N)$. Traditionally, the numbering of the genes is chosen so that weights are ranked in decreasing order. Thus, the weights usually appear to vary smoothly between consecutive genes, and the function h can be assumed to be continuously decreasing.

The gene set is included in the set of all genes. Let n be its size. In practice, n ranges from a few tens to a few hundreds: n is much smaller than N. With the identification above, it is considered as a subset of size n of the interval $[0,1]$, say $\{U_1, \ldots, U_n\}$. If there is no particular relation between the weights and the gene set (null hypothesis), then the gene set must be considered as a random sample without replacement from the set of all genes. The fact that the gene set size n is much smaller than N justifies identifying the distribution of a n-sample without replacement of $\{1/N, \ldots, N/N\}$, to that of a n-sample of i.i.d. points on $[0,1]$. The commonly accepted null hypothesis is that the gene set is uniformly distributed over all subsets of the same size, which amounts to assuming that $(U_1, \ldots, U_n)$ are i.i.d. with uniform distribution on $[0,1]$. This was the setting of Charmpi and Ycart (2015). We extend it here to the following null hypothesis.

> H_0: The gene set is a tuple $(U_1, \ldots, U_n)$ of i.i.d. random variables on $[0,1]$, with common cdf F.

The interest of this generalization is the following. It is a common place observation that genes in databases have quite different frequencies. A typical gene set contains several of those ubiquitous genes that are detected as overexpressed in most situations, thus are likely to be found also at the top of the weight vector. Due to those genes stating, as a null hypothesis that the gene set is a uniformly distributed sample leads to an excessive False Discovery Rate, as explained in Ycart et al. (2014). Taking into account, differential gene frequencies through the distribution F solve the problem.

The basis of the test statistic in GSEA is the following step function that cumulates the proportion of weights inside the gene set, along the interval $[0,1]$. It is defined for all t between 0 and 1 by:

$$S_n(t) = \frac{\sum_{k=1}^{n} h(U_k)\, \mathbb{I}_{U_k \leqslant t}}{\sum_{k=1}^{n} h(U_k)} \ . \tag{24}$$

Testing enrichment amounts to testing whether the difference between S_n and its expectation under H_0 has a high maximum. The test statistic is

$$D_n = \max_{0 \leqslant t \leqslant 1} \sqrt{n}\left(S_n(t) - \mathbb{E}_{H_0}[S_n(t)]\right) \ .$$

The procedure was called *Weighted Kolmogorov–Smirnov test* (WKS) in Charmpi and Ycart (2015). Observe that the meaning of "Weighted" is different from that of Csörgő et al. (1986), although some techniques used here are similar.

Except in the case where h is constant, the exact distribution of D_n for finite n cannot be expressed simply. Its numerical computation is out of the scope of this article: see Simard and L'Ecuyer (2011) for the classical Kolmogorov-Smirnov test. However, an asymptotic approximation can be obtained for large n. The proof of the following convergence result is a simple application of well-known techniques of empirical processes: see Kosorok (2008) as a general reference. It can be easily reduced to the uniformly distributed case $F(t) = t$ detailed in Sect. 2 of Charmpi and Ycart (2015).

Proposition 4 *Let:*

$$Z_n(t) = \sqrt{n}\left(S_n(t) - \frac{\int_0^t h(u)\,\mathrm{d}F(u)}{\int_0^1 h(u)\,\mathrm{d}F(u)}\right) .$$

Under H_0, as n tends to infinity, the stochastic process $\{Z_n(t)\,,\ 0 \leqslant t \leqslant 1\}$ converges weakly in $\ell^\infty([0,1])$ to the process $\{Z_t\,,\ 0 \leqslant t \leqslant 1\}$, where:

$$Z_t = \frac{1}{\int_0^1 h(u)\,\mathrm{d}F(u)}\left(\int_0^t h(u)\,\mathrm{d}W_{F(u)} - \frac{\int_0^t h(u)\,\mathrm{d}F(u)}{\int_0^1 h(u))\,\mathrm{d}F(u)}\int_0^1 h(u)\,\mathrm{d}W_{F(u)}\right) . \tag{25}$$

The convergence in distribution of the extrema of $Z_n(t)$ to those of Z_t is an easy application of the continuous mapping theorem (Kosorok 2008, p. 109). Therefore, the distribution under H_0 of the test statistic D_n converges to that of

$$D = \max_{0 \leqslant t \leqslant 1} Z_t \ .$$

Replacing the distribution of D_n by that of D implies an approximation error which could be minimized by a small sample correction (Stephens 1970); this has not been attempted yet.

It will now be shown that computing asymptotic p values for the WKS test reduces to computing $p_g(x)$ for some function g related to h and F.

Proposition 5 *For $0 \leqslant t \leqslant 1$ denote by $H_1(t)$, $H_2(t)$, the following integrals:*

$$H_1(t) = \int_0^t h(u)\,\mathrm{d}F(u) \ ,$$

and

$$H_2(t) = \int_0^t h^2(u)\,\mathrm{d}F(u)\ .$$

With no loss of generality assume that $H_1(1) = 1$*, and set* $\gamma_2 = H_2(1)$*. Assume that h does not vanish on any interval, hence* H_2 *is strictly increasing and its inverse* H_2^{-1} *is uniquely defined. Let:*

$$g(t) = H_1(H_2^{-1}(\gamma_2 t)) - t\ .$$

Then:

$$D \stackrel{d}{=} \sqrt{\gamma_2}\, D_g\ .$$

Proof With $H_1(1) = 1$, the definition of Z_t becomes:

$$Z_t = \int_0^t h(u)\,\mathrm{d}W_{F(u)} - \int_0^t h(u)\,\mathrm{d}F(u) \int_0^1 h(u)\,\mathrm{d}W_{F(u)}\ .$$

Observe that $\{Z_t\,,\ 0 \leqslant t \leqslant 1\}$ is a centered Gaussian process, with $Z(0) = Z(1) = 0$. The covariance function is

$$\begin{aligned}\mathbb{E}[Z_s\, Z_t] = \min\{&H_2(s), H_2(t)\}\\ &-H_1(s)H_2(t) - H_1(t)H_2(s)\\ &+\gamma_2 H_1(s)H_1(t)\ .\end{aligned}$$

The following identities hold for the distribution of D;

$$\begin{aligned}D &\stackrel{d}{=} \max_{0\leqslant t\leqslant 1} W_{H_2(t)} - H_1(t)\, W_{\gamma_2}\\ &\stackrel{d}{=} \max_{0\leqslant s\leqslant \gamma_2} W_s - H_1(H_2^{-1}(s))\, W_{\gamma_2}\\ &\stackrel{d}{=} \sqrt{\gamma_2} \max_{0\leqslant t\leqslant 1} W_t - H_1(H_2^{-1}(\gamma_2 t))\, W_1\\ &\stackrel{d}{=} \sqrt{\gamma_2} \max_{0\leqslant t\leqslant 1} B_t - g(t)\,\xi = \sqrt{\gamma_2}\, D_g\ .\end{aligned}$$

To justify the first identity, it suffices to observe that Z_t and $W_{H_2(t)} - H_1(t)W_{\gamma_2}$ are two centered Gaussian processes, with the same covariance function. The second identity is obtained through the change of time $H_2(t) \mapsto s$, which does not modify ordinates of trajectories. The third identity is the invariance of Brownian motion through scaling. The last identity follows again by comparing covariance functions. □

Here is an example, which turns out to be a frequently encountered particular case. Take $F(t) = t$. Assume that the weights are replaced by their ranks, as usual in robust statistics. Thus the weight function is $h(t) = 2(1-t)$ if weights are ranked in decreasing order, or $h(t) = 2t$ in increasing order. Observe that the distribution of $\{Z_t\,,\ 0 \leqslant t \leqslant 1\}$ is invariant through time reversal $t \mapsto 1-t$. With $h(t) = 2t$,

$$\begin{aligned}&H_1(t) = t^2\ ,\\ &H_2(t) = \frac{4}{3}t^3\ ,\\ &\gamma_2 = \frac{4}{3}\ ,\\ &H_1(H_2^{-1}(\gamma_2 t)) = t^{2/3}\ ,\\ &g(t) = t^{2/3} - t\ .\end{aligned}$$

More generally, with $b > 1/2$ and $h(t) = bt^{b-1}$,

$$\begin{aligned}&H_1(t) = t^b\ ,\\ &H_2(t) = \frac{b^2}{2b-1}t^{2b-1}\ ,\\ &\gamma_2 = \frac{b^2}{2b-1}\ ,\\ &H_1(H_2^{-1}(\gamma_2 t)) = t^{b/(2b-1)}\ ,\\ &g(t) = t^{b/(2b-1)} - t\ .\end{aligned}$$

Hence the definition (6) of $g_a(t)$, with $a > 1/2$. From our observations of real data, it appears that the weight functions h encountered in practice often lead to functions g resembling g_a for $0.6 < a < 0.8$.

In Charmpi and Ycart (2015), it had been proposed to evaluate the distribution of D by Monte Carlo simulation. Although it is a commonly used method in many statistical applications including classical GSEA, Monte Carlo simulation is not acceptable, for both precision and computing cost reasons. In real applications, the test must often be applied to several hundred vectors, each tested against several thousand gene sets. The number of values of $p_g(x)$ to be computed can be of order 10^7. Thus a running time of more than 10^{-3} s per test cannot be accepted. Moreover, the most significant gene sets, which are of greatest biological relevance, often have very small p values ($<10^{-10}$), which must be accurately calculated. The Monte Carlo method proposed in Charmpi and Ycart (2015) takes about 10^{-2} s per test, for only 10^4 simulated trajectories of Z. On such a small number, the smallest p values that can be returned are of order 10^{-3}. The conclusion is that neither the computing cost nor the precision on the results match the needs of the real application. On the contrary, the approximation schemes described in Sect. 3 are both computationally efficient and precise enough for the application.

The remarks above will be illustrated on a typical example of application. We have considered the Cancer Cell Line Encyclopedia of Barretina et al. (2012) (GEO data set GSE36133, Edgar et al. 2002). It contains RNA expression

data for 917 tumor cell lines. The data was reduced to 16775 protein coding genes; thus 917 vectors of length 16775 were considered. The rank statistics of each vector was tested for enrichment in the gene sets of the MSig C2 database (version 5.1, Subramanian et al. 2005). The database was reduced to the same protein coding genes and comprised 3751 gene sets. Thus $919 \times 3751 = 3.44 \times 10^6$ p values were computed. The calculation was made using the one-node approximation $p_{1,g}$ and frequency correction; it took 3412 s, i.e., 10^{-3} s per p value. Denote by P the 3751×917 matrix of p values so obtained. The test being repeated for each vector over 3751 gene sets, a multiple testing adjustment has to be applied on the columns of P. Dependencies in the data suggest choosing the method of Benjamini and Yekutieli (2001). After multiple testing adjustment, the number of p values smaller than 5 % among the 3751 was counted for each of the 917 columns of P: these numbers ranged from 297 to 450, with a mean of 394. The numbers of p values smaller than 10^{-10} (still after multiple testing adjustment) ranged from 32 to 128 with a mean of 76. Interestingly enough, there were 17 gene sets whose p value was smaller than 10^{-10} for all 917 vectors. All 17 gene sets had biological connections with cancer.

In order to evaluate the effect of multiple testing adjustment on Monte Carlo estimated p values, all columns of P were Winsorized replacing any p value smaller than 10^{-3} by 10^{-3}. After applying multiple testing adjustment to each Winsorized column, no p value smaller than 0.05 remained. This implies that the Monte Carlo method would have missed all significant gene sets. Of course, one could consider improving Monte Carlo accuracy by speeding it up, for instance using parallelization. However, a 100- fold gain in speed is equivalent to a 10-fold gain in accuracy for a given computing time: speeding up the Monte Carlo method will not allow it to accurately estimate p values smaller than 10^{-10}, precisely those detecting relevant biological information.

Acknowledgments We are grateful to Albert Shiryaev, Marina Kleptsyna, and Alain Le Breton for helpful and pleasant discussions. The editor and reviewers made very useful remarks. Research supported by Laboratoire d'Excellence TOUCAN (Toulouse Cancer). A. Novikov was supported by the Russian Science Foundation under Grant 14-21-00162. N. Kordzakhia was supported by the Australian Research Council Grant DP150102758.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Barretina, J., Caponigro, G., Stransky, N., et al.: The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature **483**(7391), 603–607 (2012)

Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Statist. **29**(4), 1165–1188 (2001)

Bischoff, W., Hashorva, E., Hüsler, J., Miller, F.: Exact asymptotics for boundary crossings of the Brownian bridge with applications to the kolmogorov test. Ann. Inst. Statist. Math. **55**(4), 849–864 (2003)

Borovkov, K., Novikov, A.: Explicit bounds for appproximation rates of boundary crossing probabilities for the Wiener process. J. Appl. Probab. **42**(1), 85–92 (2005)

Charmpi, K., Ycart, B.: Weighted Kolmogorov-Smirnov testing: an alternative for gene set enrichment analysis. Statist. Appl. Genet. Mol. Biol. **14**(3), 279–295 (2015)

Csörgő, M., Csörgő, S., Horváth, L., Mason, D.M.: Weighted empirical and quantile processes. Ann. Probab. **14**(1), 31–85 (1986)

del Barrio, E.: Empirical and quantile processes in the asymptotic theory of goodness-of-fit tests. In: del Barrio, E., Deheuvels, P., van de Geer, S. (eds.) Lectures on empirical processes: theory and statistical applications, EMS series of lectures in Mathematics, pp. 1–92. European Mathematical Society, Zürich (2007)

Doob, J.L.: Heuristic approach to the Kolmogorov-Smirnov theorems. Ann. Math. Statist. **20**(3), 393–403 (1949)

Durbin, J.: Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test. J. Appl. Probab. **8**(3), 431–453 (1971)

Durbin, J.: Distribution theory for tests based on the sample distribution function, SIAM CBMS-NSF Regional conference series in applied mathematics, vol. 9. SIAM, Philadelphia (1973)

Durbin, J.: The first-passage density of a continuous Gaussian process to a general boundary. J. Appl. Probab. **22**(1), 99–122 (1985)

Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucl. Acids Res. **30**(1), 207–210 (2002)

Gentz, A., Bretz, F.: Computation of multivariate normal and t probabilities. In: Chan, H.P. (ed.) Lecture notes in statistics. Springer, New York (2009)

Glasserman, P.: Monte carlo methods in financial engineering. Springer, New York (2004)

Hashorva, E.: Exact asymptotics for boundary crossing probabilities of Brownian motion with piecewise linear trend. Elect. Comm. Probab. **10**, 207–217 (2005)

Huang, D.W., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucl. Acids Res. **37**(1), 1–13 (2009)

Kahale, N.: Analytic crossing probabilities for certain barriers by Brownian motion. Ann. Appl. Probab. **18**(4), 1424–1440 (2008)

Khmaladze, E.: Martingale approach in the theory of goodness-of-fit tests. Theory Probab. Appl. **26**(2), 240–257 (1981)

Kosorok, M.R.: Introduction to empirical processes and semiparametric inference. Springer, New York (2008)

Krumbholz, W.: On large deviations of Kolmogorov-Smirnov-Renyi type statistics. J. Multivar. Anal. **6**(4), 644–652 (1976)

Novikov, A., Frishling, V., Korzakhia, N.: Approximations of boundary crossing probabilities for a Brownian motion. J. Appl. Probab. **36**(4), 1019–1030 (1999)

Parker, T.: A comparison of alternative approaches to supremum-norm goodness of fit tests with estimated parameters. Econom. Theory **29**(5), 969–1008 (2013)

Pötzelberger, K., Wang, L.: Boundary crossing probability for Brownian motion. J. Appl. Probab. **38**(1), 152–164 (2001)

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org, ISBN 3-900051-07-0

Shiryaev, A.N.: Kolmogorov, Volume 2: Selecta from the correspondence between A. N. Kolmogorov and P. S. Aleksandrov. Moscow. 2003

Simard, R., L'Ecuyer, P.L.: Computing the two-sided Kolmogorov-Smirnov distribution. J. Statist. Softw. **39**(11), 1–18 (2011)

Smirnov, N.V.: On deviations of the empirical distribution curves (Russian). Mat. Sb. **6**(48), 3–26 (1939)

Stephens, M.A.: Use of the Kolmogorov-Smirnov, Cramer-von Mises and related statistics without extensive tables. J. R. Statist. Soc. B **32**(1), 115–122 (1970)

Stephens, M.A.: Introduction to Kolmogorov (1933) on the empirical determination of a distribution. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in statistics, springer series in statistics, vol. II, pp. 93–105. Springer, New York (1992)

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA **102**(43), 15545–15550 (2005)

Ycart, B., Pont, F., Fournié, J.J.: Curbing false discovery rates in interpretation of genome-wide expression profiles. J. Biomed. Inform. **47**, 58–61 (2014)