Improved bounds on sample size for implicit matrix trace estimators

Farbod Roosta-Khorasani and Uri Ascher *

July 23, 2014

Abstract

This article is concerned with Monte-Carlo methods for the estimation of the trace of an implicitly given matrix A whose information is only available through matrix-vector products. Such a method approximates the trace by an average of N expressions of the form $\mathbf{w}^t(A\mathbf{w})$, with random vectors $\mathbf{w}$ drawn from an appropriate distribution. We prove, discuss and experiment with bounds on the number of realizations N required in order to guarantee a probabilistic bound on the relative error of the trace estimation upon employing Rademacher (Hutchinson), Gaussian and uniform unit vector (with and without replacement) probability distributions.

In total, one necessary bound and six sufficient bounds are proved, improving upon and extending similar estimates obtained in the seminal work of Avron and Toledo (2011) in several dimensions. We first improve their bound on N for the Hutchinson method, dropping a term that relates to $rank(A)$ and making the bound comparable with that for the Gaussian estimator.

We further prove new sufficient bounds for the Hutchinson, Gaussian and the unit vector estimators, as well as a necessary bound for the Gaussian estimator, which depend more specifically on properties of the matrix A. As such they may suggest for what type of matrices one distribution or another provides a particularly effective or relatively ineffective stochastic estimation method.

Keywords: randomized algorithms, trace estimation, Monte-Carlo methods, implicit linear operators

Mathematics Subject Classification (2010): 65C20, 65C05, 68W20

Communicated by Emmanuel Candès.

1 Introduction

The need to estimate the trace of an implicit square matrix is of fundamental importance [15] and arises in many applications; see for instance [10, 5, 4, 9, 7, 18, 13, 11, 8, 3] and references

*Dept. of Computer Science, University of British Columbia, Vancouver, Canada farbod/ascher@cs.ubc.ca . This work was supported in part by NSERC Discovery Grant 84306.

therein. By "implicit" we mean that the matrix of interest is not available explicitly: only probes in the form of matrix-vector products for any appropriate vector are available. The standard approach for estimating the trace of such a matrix A is based on a Monte-Carlo method, where one generates N random vector realizations $\mathbf{w}_i$ from a suitable probability distribution D and computes

$$tr_D^N(A) := \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_i^t A \mathbf{w}_i. \tag{1}$$

For the popular case where A is symmetric positive semi-definite (SPSD), the original method for estimating its trace, $tr(A)$, is due to Hutchinson [10] and uses the Rademacher distribution for D.

Until the work by Avron and Toledo [4], the main analysis and comparison of such methods was based on the variance of one sample. It is known that compared to other methods the Hutchinson method has the smallest variance, and as such it has been extensively used in many applications. In [4] so-called (ε, δ) bounds are derived in which, using Chernoff-like analysis, a lower bound is obtained on the number of samples required to achieve a probabilistically guaranteed relative error of the estimated trace. More specifically, for a given pair (ε, δ) of small (say, < 1) positive values and an appropriate probability distribution D, a lower bound on N is provided such that

$$Pr\left(|tr_D^N(A) - tr(A)| \le \varepsilon\ tr(A)\right) \ge 1 - \delta. \tag{2}$$

These authors further suggest that minimum-variance estimators may not be practically best, and conclude based on their analysis that the method with the best bound is the one using the Gaussian distribution. Let us denote

$$c = c(\varepsilon, \delta) := \varepsilon^{-2} \ln(2/\delta), \tag{3a}$$
$$r = rank(A). \tag{3b}$$

Then [4] showed that, provided A is real SPSD, (2) holds for the Hutchinson method if $N \ge 6(c + \varepsilon^{-2} \ln r)$ and for the Gaussian distribution if $N \ge 20c$.

In the present paper we continue to consider the same objective as in [4], and our first task is to improve on these bounds. Specifically, in Theorems 1 and 3 we show that (2) holds for the Hutchinson method if

$$N \ge 6c(\varepsilon, \delta), \tag{4}$$

and for the Gaussian distribution if

$$N \ge 8c(\varepsilon, \delta). \tag{5}$$

The bound (4) removes a previous factor involving the rank of the matrix A, conjectured in [4] to be indeed redundant. Note that these two bounds are astoundingly simple and general: they hold for any SPSD matrix, regardless of size or any other matrix property. Thus, we cannot expect them to be tight in practice for many specific instances of A that arise in applications.

Although practically useful, the bounds on N given in (4) and (5) do not provide insight into how different types of matrices are handled with each probability distribution. Our next contribution is to provide different bounds for the Gaussian and Hutchinson trace estimators which, though generally not computable for implicit matrices, do shed light on this question.

Furthermore, for the Gaussian estimator we prove a practically useful *necessary lower* bound on N, for a given pair (ε, δ).

A third probability distribution we consider was called the unit vector distribution in [4]. Here, the vectors $\mathbf{w}_i$ in (1) are uniformly drawn from the columns of a scaled identity matrix, $\sqrt{n}I$, and A need not be SPSD. We slightly generalize the bound in [4], obtained for the case where the sampling is done with replacement. Our bound, although not as simply computed as (4) or (5), can be useful in determining which types of matrices this distribution works best on. We then give a tighter bound for the case where the sampling is done without replacement, suggesting that when the difference between the bounds is significant (which happens when N is large), a uniform random sampling of unit vectors without replacement may be a more advisable distribution to estimate the trace with.

This paper is organized as follows. Section 2 gives two bounds for the Hutchinson method as advertised above, namely the improved bound (4) and a more involved but potentially more informative bound. Section 3 deals likewise with the Gaussian method and adds also a necessary lower bound, while Section 4 is devoted to the unit vector sampling methods.

In Section 5 we give some numerical examples verifying that the trends predicted by the theory are indeed realized. Conclusions and further thoughts are gathered in Section 6.

In what follows we use the notation $tr_H^N(A)$, $tr_G^N(A)$, $tr_{U_1}^N(A)$, and $tr_{U_2}^N(A)$ to refer, respectively, to the trace estimators using Hutchinson, Gaussian, and uniform unit vector with and without replacement, in lieu of the generic notation $tr_D^N(A)$ in (1) and (2). We also denote for any given random vector of size n, $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in})^t$. We restrict attention to real-valued matrices, although extensions to complex-valued ones are possible, and employ the 2-norm by default.

2 Hutchinson estimator bounds

In this section we consider the Hutchinson trace estimator, $tr_H^N(A)$, obtained by setting $D = H$ in (1), where the components of the random vectors $\mathbf{w}_i$ are i.i.d Rademacher random variables (i.e., $Pr(w_{ij} = 1) = Pr(w_{ij} = -1) = \frac{1}{2}$).

2.1 Improving the bound in [4]

Theorem 1 *Let A be an $n \times n$ SPSD matrix. Given a pair (ε, δ), the inequality (2) holds with $D = H$ if N satisfies (4).*

Proof Since A is SPSD, it can be diagonalized by a unitary similarity transformation as $A = U^t \Lambda U$. Consider N random vectors $\mathbf{w}_i$, $i = 1, \ldots, N$, whose components are i.i.d and drawn from the Rademacher distribution, and define $\mathbf{z}_i = U\mathbf{w}_i$ for each. We have

$$
\begin{aligned}
Pr\left(tr_H^N(A) \leq (1-\varepsilon)tr(A)\right) &= Pr\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i^t A\mathbf{w}_i \leq (1-\varepsilon)tr(A)\right) \\
&= Pr\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i^t \Lambda\mathbf{z}_i \leq (1-\varepsilon)tr(A)\right) \\
&= Pr\left(\sum_{i=1}^{N}\sum_{j=1}^{r}\lambda_j z_{ij}^2 \leq N(1-\varepsilon)tr(A)\right) \\
&= Pr\left(\sum_{j=1}^{r}\frac{\lambda_j}{tr(A)}\sum_{i=1}^{N} z_{ij}^2 \leq N(1-\varepsilon)\right) \\
&\leq \exp\{tN(1-\varepsilon)\}\mathbb{E}\left(\exp\{\sum_{j=1}^{r}\frac{\lambda_j}{tr(A)}\sum_{i=1}^{N}-tz_{ij}^2\}\right),
\end{aligned}
$$

where the last inequality holds for any $t > 0$ by Markov's inequality.

Next, using the convexity of the exp function and the linearity of expectation, we obtain

$$
\begin{aligned}
\mathbb{E}\left(\exp\{\sum_{j=1}^{r}\frac{\lambda_j}{tr(A)}\sum_{i=1}^{N}-tz_{ij}^2\}\right) &\leq \sum_{j=1}^{r}\frac{\lambda_j}{tr(A)}\mathbb{E}\left(\exp\{\sum_{i=1}^{N}-tz_{ij}^2\}\right) \\
&= \sum_{j=1}^{r}\frac{\lambda_j}{tr(A)}\mathbb{E}\left(\prod_{i=1}^{N}\exp\{-tz_{ij}^2\}\right) \\
&= \sum_{j=1}^{r}\frac{\lambda_j}{tr(A)}\prod_{i=1}^{N}\mathbb{E}\left(\exp\{-tz_{ij}^2\}\right),
\end{aligned}
$$

where the last equality holds since, for a given j, z_{ij}'s are independent with respect to i.

Now, we want to have that $\exp\{tN(1-\varepsilon)\}\prod_{i=1}^{N}\mathbb{E}\left(\exp\{-tz_{ij}^2\}\right) \leq \delta/2$. For this we make use of the inequalities in the end of the proof of Lemma 5.1 of [2]. Following inequalities (15)–(19) in [2] and letting $t = \varepsilon/(2(1+\varepsilon))$, we get

$$
\exp\{tN(1-\varepsilon)\}\prod_{i=1}^{N}\mathbb{E}\left(\exp\{-tz_{ij}^2\}\right) < \exp\{-\frac{N}{2}(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3})\}.
$$

Next, if N satisfies (4) then $\exp\{-\frac{N}{2}(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3})\} < \delta/2$, and thus it follows that

$$
Pr\left(tr_H^N(A) \leq (1-\varepsilon)tr(A)\right) < \delta/2.
$$

By a similar argument, making use of inequalities (11)–(14) in [2] with the same t as above, we also obtain with the same bound for N so that $Pr\left(tr_H^N(A) \geq (1+\varepsilon)tr(A)\right) \leq \delta/2$. So finally using the union bound yields the desired result. ■

It can be seen that (4) is the same bound as the one in [4] with the important exception that the factor $r = rank(A)$ does not appear in the bound. Furthermore, the same bound on N holds for any SPSD matrix.

2.2 A matrix-dependent bound

Here we derive another bound for the Hutchinson trace estimator which may shed light as to what type of matrices the Hutchinson method is best suited for.

Let us denote by $a_{k,j}$ the (k,j)th element of A and by $\mathbf{a}_j$ its jth column, $k,j = 1,\ldots,n$.

Theorem 2 *Let A be an $n \times n$ symmetric positive semi-definite matrix, and define*

$$\mathcal{K}_H^j := \frac{\|\mathbf{a}_j\|^2 - a_{j,j}^2}{a_{j,j}^2} = \sum_{k\neq j} a_{k,j}^2 \;/\; a_{j,j}^2, \quad \mathcal{K}_H := \max_j \mathcal{K}_H^j. \tag{6}$$

Given a pair of positive small values (ε,δ), the inequality (2) holds with $D = H$ if

$$N > 2\mathcal{K}_H c(\varepsilon,\delta). \tag{7}$$

Proof Elementary linear algebra implies that since A is SPSD, $a_{j,j} \geq 0$ for each j. Furthermore, if $a_{j,j} = 0$ then the jth row and column of A identically vanish, so we may assume below that $a_{j,j} > 0$ for all $j = 1,\ldots,n$. Note that

$$tr_H^N(A) - tr(A) = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{n} a_{j,k}w_{ij}w_{ik}.$$

Hence

$$\begin{aligned}
Pr\left(tr_H^N(A) \leq (1-\varepsilon)tr(A)\right) &= Pr\left(\sum_{j=1}^{n}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{n} -a_{j,k}w_{ij}w_{ik} \geq N\varepsilon\; tr(A)\right)\\
&= Pr\left(\sum_{j=1}^{n}\frac{a_{j,j}}{tr(A)}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{n} -\frac{a_{j,k}}{a_{j,j}}w_{ij}w_{ik} \geq N\varepsilon\right)\\
&\leq \exp\{-tN\varepsilon\}\mathbb{E}\left(\exp\{\sum_{j=1}^{n}\frac{a_{j,j}}{tr(A)}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{n} -\frac{a_{j,k}t}{a_{j,j}}w_{ij}w_{ik}\}\right),
\end{aligned}$$

where the last inequality is again obtained for any $t > 0$ by using Markov's inequality. Now, again using the convexity of the exp function and the linearity of expectation, we obtain

$$\begin{aligned}
Pr\left(tr_H^N(A) \leq (1-\varepsilon)tr(A)\right) &\leq \exp\{-tN\varepsilon\}\sum_{j=1}^{n}\frac{a_{j,j}}{tr(A)}\mathbb{E}\left(\exp\{\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{n} -\frac{a_{j,k}t}{a_{j,j}}w_{ij}w_{ik}\}\right)\\
&= \exp\{-tN\varepsilon\}\sum_{j=1}^{n}\frac{a_{j,j}}{tr(A)}\prod_{i=1}^{N}\mathbb{E}\left(\exp\{\sum_{\substack{k=1\\k\neq j}}^{n} -\frac{a_{j,k}t}{a_{j,j}}w_{ij}w_{ik}\}\right)
\end{aligned}$$

by independence of $w_{ij}w_{ik}$ with respect to the index i.

Next, note that

$$\mathbb{E}\left(\exp\{\sum_{\substack{k=1\\k\neq j}}^{n}\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right)=\mathbb{E}\left(\exp\{\sum_{\substack{k=1\\k\neq j}}^{n}-\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right).$$

Furthermore, since $Pr(w_{ij}=-1)=Pr(w_{ij}=1)=\frac{1}{2}$, and using the law of total expectation, we have

$$\mathbb{E}\left(\exp\{\sum_{\substack{k=1\\k\neq j}}^{n}-\frac{a_{j,k}t}{a_{j,j}}w_{ij}w_{ik}\}\right)=\mathbb{E}\left(\exp\{\sum_{\substack{k=1\\k\neq j}}^{n}\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right)=\prod_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}\left(\exp\{\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right),$$

so

$$Pr\left(tr_H^N(A)\leq(1-\varepsilon)tr(A)\right)\quad\leq\quad\exp\{-tN\varepsilon\}\sum_{j=1}^{n}\frac{a_{j,j}}{tr(A)}\prod_{i=1}^{N}\prod_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}\left(\exp\{\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right).$$

We want to have the right hand side expression bounded by $\delta/2$.

Applying Hoeffding's lemma we get $\mathbb{E}\left(\exp\{\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right)\leq\exp\{\frac{a_{j,k}^2t^2}{2a_{j,j}^2}\}$, hence

$$\exp\{-tN\varepsilon\}\prod_{i=1}^{N}\prod_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}\left(\exp\{\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right)\quad\leq\quad\exp\{-tN\varepsilon+\mathcal{K}_H^jNt^2/2\}\tag{8a}$$

$$\leq\quad\exp\{-tN\varepsilon+\mathcal{K}_HNt^2/2\}.\tag{8b}$$

The choice

$$t=\varepsilon/\mathcal{K}_H$$

minimizes the right hand side. Now if (7) holds then

$$\exp(-tN\varepsilon)\prod_{i=1}^{N}\prod_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}\left(\exp\{\frac{a_{j,k}t}{a_{j,j}}w_{ik}\}\right)\leq\delta/2,$$

hence we have

$$Pr(tr_H^N(A)\leq(1-\varepsilon)tr(A))\quad\leq\quad\delta/2.$$

Similarly, we obtain that

$$Pr(tr_H^N(A)\geq(1+\varepsilon)tr(A))\quad\leq\quad\delta/2,$$

and using the union bound finally gives desired result. ■

Comparing (7) to (4), it is clear that the bound of the present subsection is only worthy of consideration if $\mathcal{K}_H < 3$. Note that Theorem 2 emphasizes the relative ℓ_2 energy of the off-diagonals: the matrix does not necessarily have to be diagonally dominant (i.e., where a similar relationship holds in the ℓ_1 norm) for the bound on N to be moderate. Furthermore, a matrix need not be "nearly" diagonal for this method to require small sample size. In fact a matrix can have off-diagonal elements of significant size that are far away from the main diagonal without automatically affecting the performance of the Hutchinson method. However, note also that our bound can be pessimistic, especially if the average value or the mode of $\mathcal{K}_H^j$ in (6) is far lower than its maximum, $\mathcal{K}_H$. This can be seen in the above proof where the estimate (8b) is obtained from (8a). Simulations in Section 5 show that the Hutchinson method can be a very efficient estimator even in the presence of large outliers, so long as the bulk of the distribution is concentrated near small values.

The case $\mathcal{K}_H = 0$ corresponds to a diagonal matrix, for which the Hutchinson method yields the trace with one shot, $N = 1$. In agreement with the bound (7), we expect the actual required N to grow when a sequence of otherwise similar matrices A is envisioned in which $\mathcal{K}_H$ grows away from 0, as the energy in the off-diagonal elements grows relatively to that in the diagonal elements.

3 Gaussian estimator bounds

In this section we consider the Gaussian trace estimator, $tr_G^N(A)$, obtained by setting $D = G$ in (1), where the components of the random vectors $\mathbf{w}_i$ are i.i.d standard normal random variables. We give two sufficient and one necessary lower bounds for the number of Gaussian samples required to achieve an (ε, δ) trace estimate. The first sufficient bound (5) improves the result in [4] by a factor of 2.5. Our bound is only worse than (4) by a fraction, and it is an upper limit of the potentially more informative (if less available) bound (10), which relates to the properties of the matrix A. The bound (10) provides an indication as to what matrices may be suitable candidates for the Gaussian method. Then we present a practically computable, necessary bound for the sample size N.

3.1 Sufficient bounds

The proof of the following theorem closely follows the approach in [4].

Theorem 3 *Let A be an $n \times n$ SPSD matrix and denote its eigenvalues by $\lambda_1, \ldots, \lambda_n$. Further, define*

$$\mathcal{K}_G^j := \frac{\lambda_j}{tr(A)}, \quad \mathcal{K}_G := \max_j \mathcal{K}_G^j = \frac{\|A\|}{tr(A)}. \tag{9}$$

Then, given a pair of positive small values (ε, δ), the inequality (2) holds with $D = G$ provided that (5) holds. This estimate also holds provided that

$$N > 8\mathcal{K}_G c(\varepsilon, \delta). \tag{10}$$

Proof Since A is SPSD, we have $\|A\| \leq tr(A)$, so if (5) holds then so does (10). We next concentrate on proving the result assuming the tighter bound (10) on the actual N required in a given instance.

Writing as in the previous section $A = U^t \Lambda U$, consider N random vectors $\mathbf{w}_i$, $i = 1, \ldots, N$, whose components are i.i.d and drawn from the normal distribution, and define $\mathbf{z}_i = U\mathbf{w}_i$. Since U is orthogonal, the elements z_{ij} of $\mathbf{z}_i$ are i.i.d Gaussian random variables. We have as before,

$$\begin{aligned} Pr\left(tr_G^N(A) \leq (1-\varepsilon)tr(A)\right) &= Pr\left(\sum_{i=1}^{N}\sum_{j=1}^{r}\lambda_j z_{ij}^2 \leq N(1-\varepsilon)tr(A)\right) \\ &\leq \exp\{tN(1-\varepsilon)tr(A)\}\mathbb{E}\left(\exp\{\sum_{i=1}^{N}\sum_{j=1}^{r} -t\lambda_j z_{ij}^2\}\right) \\ &\leq \exp\{tN(1-\varepsilon)tr(A)\}\prod_{i=1}^{N}\prod_{j=1}^{r}\mathbb{E}\left(\exp\{-t\lambda_j z_{ij}^2\}\right). \end{aligned}$$

Here z_{ij}^2 is a χ^2 random variable of degree 1 (see [12]), and hence for the characteristics we have

$$\mathbb{E}\left(\exp\{-t\lambda_j z_{ij}^2\}\right) = (1+2\lambda_j t)^{-\frac{1}{2}}.$$

This yields the bound

$$Pr\left(tr_G^N(A) \leq (1-\varepsilon)tr(A)\right) \leq \exp\{tN(1-\varepsilon)tr(A)\}\prod_{j=1}^{r}(1+2\lambda_j t)^{-\frac{N}{2}}.$$

Next, it is easy to prove by elementary calculus that given any $0 < \alpha < 1$, the following holds for all $0 \leq x \leq \frac{1-\alpha}{\alpha}$,

$$\ln(1+x) \geq \alpha x. \tag{11}$$

Setting $\alpha = 1-\varepsilon/2$, then by (11) and for all $t \leq (1-\alpha)/(2\alpha\|A\|)$, we have that $(1+2\lambda_j t) > \exp\{2\alpha\lambda_j\}t$, so

$$\begin{aligned} Pr\left(tr_G^N(A) \leq (1-\varepsilon)tr(A)\right) &\leq \exp\{tN(1-\varepsilon)tr(A)\}\prod_{j=1}^{r}\exp(-N\alpha\lambda_j t) \\ &= \exp\{tN(1-\varepsilon-\alpha)tr(A)\}. \end{aligned}$$

We want the latter right hand side to be bounded by $\delta/2$, i.e., we want to have

$$N \geq \frac{\ln\left(2/\delta\right)}{(\alpha-(1-\varepsilon))tr(A)t} = \frac{2\varepsilon c(\varepsilon,\delta)}{tr(A)t},$$

where $t \leq (1-\alpha)/(2\alpha\|A\|)$. Now, setting $t = (1-\alpha)/(2\alpha\|A\|) = \varepsilon/(2(2-\varepsilon)\|A\|)$, we obtain

$$N \geq 4(2-\varepsilon)c(\varepsilon,\delta)\mathcal{K}_G,$$

so if (10) holds then

$$Pr\left(tr_G^N(A) \leq (1-\varepsilon)tr(A)\right) \leq \delta/2.$$

Using a similar argument we also obtain

$$Pr\left(tr_G^N(A) \geq (1+\varepsilon)tr(A)\right) \leq \delta/2,$$

and subsequently the union bound yields the desire result. ■

The matrix-dependent bound (10), proved to be sufficient in Theorem 3, provides additional information over (5) about the type of matrices for which the Gaussian estimator is (probabilistically) guaranteed to require only a small sample size: if the eigenvalues of an SPSD matrix are distributed such that the ratio $\|A\|/tr(A)$ is small (e.g., if they are all of approximately the same size), then the Gaussian estimator bound requires a small number of realizations. This observation is reaffirmed by looking at the variance of this estimator, namely $2\|A\|_F^2$. It is easy to show that among all the matrices with a fixed trace and rank, those with equal eigenvalues have the smallest Frobenius norm.

Furthermore, it is easy to see that the *stable rank* (see [17] and references therein) of any real rectangular matrix C which satisfies $A = C^tC$ equals $1/\mathcal{K}_G$. Thus, the bound constant in (10) is inversely proportional to this stable rank, suggesting that estimating the trace using the Gaussian distribution may become inefficient if the stable rank of the matrix is low. Theorem 5 in Section 3.2 below further substantiates this intuition.

As an example of an application of the above results, let us consider finding the minimum number of samples required to compute the rank of a projection matrix using the Gaussian estimator [4, 6]. Recall that a projection matrix is SPSD with only 0 and 1 eigenvalues. Compared to the derivation in [4], here we use Theorem 3 directly to obtain a similar bound with a slightly better constant.

Corollary 4 *Let A be an $n \times n$ projection matrix with rank $r > 0$, and denote the rounding of any real scalar x to the nearest integer by $round(x)$. Then, given a positive small value δ, the estimate*

$$Pr\left(round(tr_G^N(A)) \neq r\right) \leq \delta \tag{12a}$$

holds if

$$N \geq 8\, r \ln\left(2/\delta\right). \tag{12b}$$

Proof The result immediately follows using Theorem 3 upon setting $\varepsilon = 1/r, \|A\| = 1$ and $tr(A) = r$. ■

3.2 A necessary bound

Below we provide a rank-dependent, almost tight necessary condition for the minimum sample size required to obtain (2). This bound is easily computable in case that $r = rank(A)$ is known.

Before we proceed, recall the definition of the regularized Gamma functions

$$P\left(\alpha,\beta\right):=\frac{\gamma\left(\alpha,\beta\right)}{\Gamma\left(\alpha\right)},\quad Q\left(\alpha,\beta\right):=\frac{\Gamma\left(\alpha,\beta\right)}{\Gamma\left(\alpha\right)},$$

where $\gamma\left(\alpha,\beta\right),\Gamma\left(\alpha,\beta\right)$ and $\Gamma\left(\alpha\right)$ are, respectively, the lower incomplete, the upper incomplete and the complete Gamma functions, see [1]. We also have that $\Gamma\left(\alpha\right)=\Gamma\left(\alpha,\beta\right)+\gamma\left(\alpha,\beta\right)$. Further, define

$$\Phi_{\theta}(x):=P\left(\frac{x}{2},\frac{\tau(1-\theta)x}{2}\right)+Q\left(\frac{x}{2},\frac{\tau(1+\theta)x}{2}\right),\tag{13a}$$

where

$$\tau=\frac{\ln(1+\theta)-\ln(1-\theta)}{2\theta}.\tag{13b}$$

Theorem 5 *Let A be a rank-r SPSD $n\times n$ matrix, and let (ε,δ) be a tolerance pair. If the inequality (2) with $D=G$ holds for some N, then necessarily*

$$\Phi_{\varepsilon}(Nr)\leq\delta.\tag{14}$$

Proof As in the proof of Theorem 3 we have

$$\begin{aligned}Pr\left(|tr_{G}^{N}(A)-tr(A)|\leq\varepsilon\ tr(A)\right) &= Pr\left(|\sum_{i=1}^{N}\sum_{j=1}^{r}\lambda_{j}z_{ij}^{2}-Ntr(A)|\leq\varepsilon Ntr(A)\right)\\ &= Pr\left((1-\varepsilon)\leq\sum_{i=1}^{N}\sum_{j=1}^{r}\frac{\lambda_{j}}{tr(A)\ N}z_{ij}^{2}\leq(1+\varepsilon)\right).\end{aligned}$$

Next, applying Theorem 3 of [16] gives

$$Pr\left(|tr_{G}^{N}(A)-tr(A)|\leq\varepsilon\ tr(A)\right)\leq Pr\left(c(1-\varepsilon)\leq\frac{1}{Nr}\mathcal{X}_{Nr}^{2}\leq c(1+\varepsilon)\right),$$

where $\mathcal{X}_{M}^{2}$ denotes a chi-squared random variable of degree M with the cumulative distribution function

$$CDF_{\mathcal{X}_{M}^{2}}(x)=Pr\left(\mathcal{X}_{M}^{2}\leq x\right)=\frac{\gamma\left(\frac{M}{2},\frac{x}{2}\right)}{\Gamma\left(\frac{M}{2}\right)}.$$

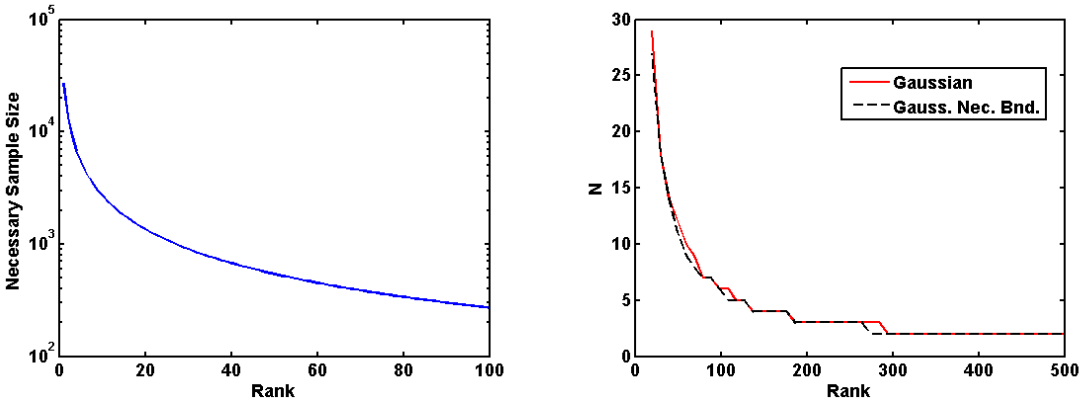
A further straightforward manipulation yields the stated result. ■

Having a computable necessary condition is practically useful: given a pair of fixed sample size N and error tolerance ε, the failure probability δ cannot be smaller than $\delta_{0}=\Phi_{\varepsilon}(Nr)$.

Since our sufficient bounds are not tight, it is not possible to make a direct comparison between the Hutchinson and Gaussian methods based on them. However, using this necessary condition can help for certain matrices. Consider a low rank matrix with a rather small $\mathcal{K}_H$ in (7). For such a matrix and a given pair (ε, δ), the condition (14) will probabilistically necessitate a rather large N, while (7) may give a much smaller sufficient bound for N. In this situation, using Theorem 5, the Hutchinson method is indeed guaranteed to require a smaller sample size than the Gaussian method.

The condition in Theorem 5 is almost tight in the following sense. Note that in (13b), $\tau \approx 1$ for $\theta = \varepsilon$ sufficiently small. So, $1 - \Phi_\varepsilon(Nr)$ would be very close to $Pr\left((1-\varepsilon) \le tr_G^N(A^*) \le (1+\varepsilon)\right)$, where A^* is an SPD matrix of the same rank as A whose eigenvalues are all equal to $1/r$. Next note that the condition (14) should hold for all matrices of the same rank; hence it is almost tight. Figures 1 and 4 demonstrate this effect.

Notice that for a very low rank matrix and a reasonable pair (ε, δ), the necessary N given by (14) could be even larger than the matrix size n, rendering the Gaussian method useless for such instances; see Figure 1.



(a) $\varepsilon = \delta = 0.02$ (b) $n = 10,000,\ \varepsilon = \delta = 0.1$

Figure 1: Necessary bound for Gaussian estimator: (a) the log-scale of N according to (14) as a function of $r = rank(A)$: larger ranks yield smaller necessary sample size. For very low rank matrices, the necessary bound grows significantly: for $n = 1000$ and $r \le 30$, necessarily $N > n$ and the Gaussian method is practically useless; (b) tightness of the necessary bound demonstrated by an actual run as described for Example 4 in Section 5 where A has all eigenvalues equal.

4 Random unit vector bounds, with and without replacement, for general square matrices

An alternative to the Hutchinson and Gaussian estimators is to draw the vectors $\mathbf{w}_i$ from among the n columns of the scaled identity matrix $\sqrt{n}I$. Note that if $\mathbf{w}_i$ is the ith (scaled) unit vector then $\mathbf{w}_i^t A \mathbf{w}_i = n a_{ii}$. Hence the trace can be recovered in $N = n$ deterministic steps upon setting in (1) $i = j,\ j = 1, 2, \ldots, n$. However, our hope is that for some matrices

a good approximation for the trace can be recovered in $N \ll n$ such steps, with $\mathbf{w}_i$'s drawn as mentioned above.

There are typically two ways one can go about drawing such samples: with or without replacement. The first of these has been studied in [4]. However, in view of the exact procedure, we may expect to occasionally require smaller sample sizes by using the strategy of sampling without replacement. In this section we make this intuitive observation more rigorous.

In what follows, U_1 and U_2 refer to the uniform distribution of unit vectors with and without replacement, respectively. We first find expressions for the mean and variance of both strategies, obtaining a smaller variance for U_2.

Lemma 6 *Let A be an $n \times n$ matrix and let N denote the sample size. Then*

$$\mathbb{E}\left(tr^N_{U_1}(A)\right) = \mathbb{E}\left(tr^N_{U_2}(A)\right) = tr(A), \tag{15a}$$

$$Var\left(tr^N_{U_1}(A)\right) = \frac{1}{N}\left(n\sum_{j=1}^{n} a_{jj}^2 - tr(A)^2\right), \tag{15b}$$

$$Var\left(tr^N_{U_2}(A)\right) = \frac{(n-N)}{N(n-1)}\left(n\sum_{j=1}^{n} a_{jj}^2 - tr(A)^2\right), \quad N \leq n. \tag{15c}$$

Proof The results for U_1 are proved in [4]. Let us next concentrate on U_2, and group the randomly selected unit vectors into an $n \times N$ matrix W. Then

$$\mathbb{E}\left(tr^N_{U_2}(A)\right) = \frac{1}{N}\mathbb{E}\left(tr\left(W^tAW\right)\right) = \frac{1}{N}\mathbb{E}\left(tr\left(A\,WW^t\right)\right) = \frac{1}{N}tr\left(A\,\mathbb{E}\left(WW^t\right)\right).$$

Let y_{ij} denote the (i,j)th element of the random matrix WW^t. Clearly, $y_{ij} = 0$ if $i \neq j$. It is also easily seen that y_{ii} can only take on the values 0 or n. We have

$$\mathbb{E}\left(y_{ii}\right) = nPr\left(y_{ii} = n\right) = n\frac{\binom{n-1}{N-1}}{\binom{n}{N}} = N,$$

so $\mathbb{E}(WW^t) = N \cdot I$, where I stands for the identity matrix. This, in turn, gives $\mathbb{E}\left(tr^N_{U_2}(A)\right) = tr(A)$.

For the variance, we first calculate

$$\mathbb{E}\left[\left(tr^N_{U_2}(A)\right)^2\right] = \frac{1}{N^2}\mathbb{E}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbf{w}_i^tA\mathbf{w}_i\right)\left(\mathbf{w}_j^tA\mathbf{w}_j\right)\right)$$

$$= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[\left(\mathbf{w}_i^tA\mathbf{w}_i\right)^2\right] + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\mathbb{E}\left[\left(\mathbf{w}_i^tA\mathbf{w}_i\right)\left(\mathbf{w}_j^tA\mathbf{w}_j\right)\right]\right). \tag{16}$$

Let $\mathbf{e}_j$ denote the j^{th} column of the scaled identity matrix, $\sqrt{n}I$. Using the law of total expectation (i.e., the tower rule), we have for any two random vectors $\mathbf{w}_i$ and $\mathbf{w}_j$ with $i \neq j$,

$$\begin{aligned}
\mathbb{E}\left[\left(\mathbf{w}_i^t A \mathbf{w}_i\right)\left(\mathbf{w}_j^t A \mathbf{w}_j\right)\right] &= \sum_{k=1}^{n} \mathbb{E}\left[\left(\mathbf{w}_i^t A \mathbf{w}_i\right)\left(\mathbf{w}_j^t A \mathbf{w}_j\right) | \mathbf{w}_i = \mathbf{e}_k\right] \cdot Pr\left(\mathbf{w}_i = \mathbf{e}_k\right) \\
&= \sum_{k=1}^{n} n a_{kk} \cdot \mathbb{E}\left[\left(\mathbf{w}_j^t A \mathbf{w}_j\right) | \mathbf{w}_i = \mathbf{e}_k\right] \cdot \frac{1}{n} \\
&= \sum_{k=1}^{n} a_{kk} \sum_{\substack{l=1 \\ l \neq k}}^{n} \mathbb{E}\left[\left(\mathbf{w}_j^t A \mathbf{w}_j\right) | \mathbf{w}_j = \mathbf{e}_l\right] \cdot Pr\left(\mathbf{w}_j = \mathbf{e}_l | \, \mathbf{w}_i = \mathbf{e}_k\right) \\
&= \sum_{k=1}^{n} a_{kk} \sum_{\substack{l=1 \\ l \neq k}}^{n} n a_{ll} \frac{1}{n-1} = \frac{n}{n-1} \sum_{k=1}^{n} \sum_{\substack{l=1 \\ k \neq l}}^{n} a_{kk} a_{ll} \\
&= \frac{n}{n-1} (tr(A)^2 - \sum_{j=1}^{n} a_{jj}^2).
\end{aligned}$$

Substituting this in (16) gives

$$\mathbb{E}\left[\left(tr_{U_2}^N(A)\right)^2\right] = \frac{1}{N^2}\left(nN \sum_{j=1}^{n} a_{jj}^2 + \frac{nN(N-1)}{n-1}(tr(A)^2 - \sum_{j=1}^{n} a_{jj}^2)\right).$$

Next, the variance is

$$Var\left(tr_{U_2}^N(A)\right) = \mathbb{E}\left[\left(tr_{U_2}^N(A)\right)^2\right] - \left[\mathbb{E}\left(tr_{U_2}^N(A)\right)\right]^2,$$

which gives (15c). ■

Note that $Var\left(tr_{U_2}^N(A)\right) = \frac{n-N}{n-1} Var\left(tr_{U_1}^N(A)\right)$. The difference in variance between these sampling strategies is small for $N \ll n$, and they coincide if $N = 1$. Moreover, in case that the diagonal entries of the matrix are all equal, the variance for both sampling strategies vanishes.

We now turn to the analysis of the sample size required to ensure (2) and find a slight improvement over the bound given in [4] for U_1. A similar analysis for the case of sampling without replacement shows that the latter may generally be a somewhat better strategy.

Theorem 7 *Let A be a real $n \times n$ matrix, and denote*

$$\mathcal{K}_U^{(i,j)} = \frac{n}{tr(A)} |a_{ii} - a_{jj}|, \quad \mathcal{K}_U = \max_{\substack{1 \leq i,j \leq n \\ i \neq j}} \mathcal{K}_U^{(i,j)}. \tag{17}$$

Given a pair of positive small values (ε, δ), the inequality (2) holds with $D = U_1$ if

$$N > \frac{\mathcal{K}_U^2}{2} c(\varepsilon, \delta) \equiv \mathcal{F}, \tag{18}$$

and with $D = U_2$ *if*

$$N \geq \frac{n+1}{1+\frac{n-1}{\mathcal{F}}}. \tag{19}$$

Proof This proof is refreshingly short. Note first that every sample of these estimators takes on a Rayleigh value in $[n \min_j a_{jj}, n \max_j a_{jj}]$.

The proof of (18), for the case with replacement, uses Hoeffding's inequality in exactly the same way as the corresponding theorem in [4]. We obtain directly that if (18) is satisfied then (2) holds with $D = U_1$.

For the case without replacement we use Serfling's inequality [14] to obtain

$$Pr\left(|tr^N_{U_2}(A) - tr(A)| \geq \varepsilon tr(A)\right) \leq 2\exp\left\{\frac{-2N\varepsilon^2}{(1-f_{N-1})\,\mathcal{K}_U^2}\right\},$$

where f_N is the sampling fraction defined as

$$f_N = \frac{N-1}{n-1}.$$

Now, for the inequality (2) to hold, we need

$$2\exp\left\{\frac{-2N\varepsilon^2}{(1-f_{N-1})\,\mathcal{K}_U^2}\right\} \leq \delta,$$

so we require that

$$\frac{N}{1-f_{N-1}} \geq \mathcal{F}.$$

The stated result (19) is obtained following some straightforward algebraic manipulation. ■

Looking at the bounds (18) for U_1 and (19) for U_2 and observing the expression (17) for $\mathcal{K}_U$, one can gain insight as to the type of matrices which are handled efficiently using this estimator: this would be the case if the diagonal elements of the matrix all have similar values. In the extreme case where they are all the same, we only need one sample. The corresponding expression in [4] does not reflect this result.

An illustration of the relative behaviour of the two bounds is given in Figure 2.

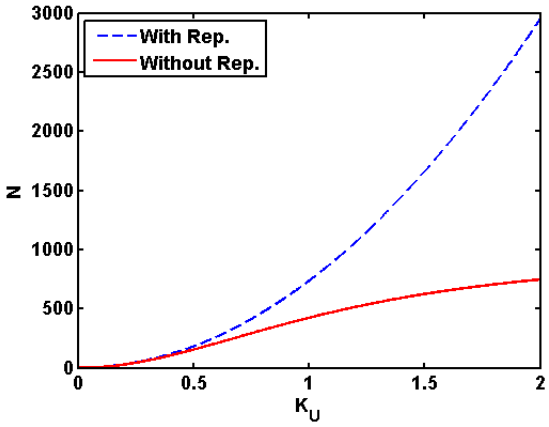


Figure 2: The behaviour of the bounds (18) and (19) with respect to the factor $K = \mathcal{K}_U$ for $n = 1000$ and $\varepsilon = \delta = 0.05$. The bound for U_2 is much more resilient to the distribution of the diagonal values than that of U_1. For very small values of $\mathcal{K}_U$, there is no major difference between the bounds.

5 Numerical Examples

In this section we experiment with several examples, comparing the performance of different methods with regards to various matrix properties and verifying that the bounds obtained in our theorems indeed agree with the numerical experiments.

Example 1 *In this example we do not consider δ at all. Rather, we check numerically for various values of ε what value of N is required to achieve a result respecting this relative tolerance. We have calculated maximum and average values for N over 100 trials for several special examples, verifying numerically the following considerations.*

- *The matrix of all 1s (in* MATLAB, `A=ones(n,n)`*) has been considered in [4]. Here $tr(A) = n$, $\mathcal{K}_H = n - 1$, and a very large N is often required if ε is small for both Hutchinson and Gauss methods. For the unit vector method, however, $\mathcal{K}_U = 0$ in (17), so the latter method converges in one iteration, $N = 1$. This fact yields an example where the unit vector estimator is far better than either Hutchinson or Gaussian estimators; see Figure 3.*

- *Another extreme example, where this time it is the Hutchinson estimator which requires only one sample whereas the other methods may require many more, is the case of a diagonal matrix A. For a diagonal matrix, $\mathcal{K}_H = 0$, and the result follows from Theorem 2.*

- *If A is a multiple of the identity then, since $\mathcal{K}_U = \mathcal{K}_H = 0$, only the Gaussian estimator from among the methods considered requires more than one sample; thus, it is worst.*

- *Examples where the unit vector estimator is consistently (and significantly) worst are obtained by defining $A = Q^t D Q$ for a diagonal matrix D with different positive ele-*

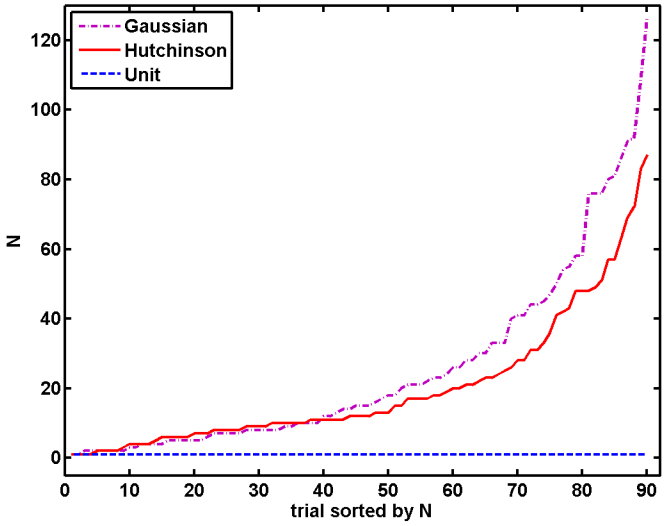


Figure 3: Example 1. For the matrix of all 1s with $n = 10,000$, the plot depicts the numbers of samples in 100 trials required to satisfy the relative tolerance $\varepsilon = .05$, sorted by increasing N. The average N for both Hutchinson and Gauss estimators was around 50, while for the uniform unit vector estimator always $N = 1$. Only the best 90 results (i.e., lowest resulting values of N) are shown for reasons of scaling. Clearly, the unit vector method is superior here.

ments which are of the same order of magnitude and a nontrivial orthogonal matrix Q.

- *We have not been able to come up with a simple example of the above sort where the Gaussian estimator shines over both others, although we have seen many occasions in practice where it slightly outperforms the Hutchinson estimator with both being significantly better than the unit vector estimators.*

Example 2 *Consider the matrix $A = \mathbf{x}\mathbf{x}^t/\|\mathbf{x}\|^2$, where $\mathbf{x} \in \mathbb{R}^n$, and for some $\theta > 0$, $x_j = \exp(-j\theta)$, $1 \le j \le n$. This extends the example of all 1s of Figure 3 (for which $\theta = 0$) to instances with rapidly decaying elements.*

It is easy to verify that

$$
\begin{aligned}
&tr(A) = 1, \quad r = 1, \quad \mathcal{K}_G = 1,\\
&\mathcal{K}_H^j = \|\mathbf{x}\|^2 x_j^{-2} - 1, \quad \mathcal{K}_H = \|\mathbf{x}\|^2 x_n^{-2} - 1,\\
&\mathcal{K}_U^{(i,j)} = \frac{n}{\|\mathbf{x}\|^2}|x_i^2 - x_j^2|, \quad \mathcal{K}_U = \frac{n}{\|\mathbf{x}\|^2}(x_1^2 - x_n^2),\\
&\|\mathbf{x}\|^2 = \frac{\exp(-2\theta) - \exp(-2(n+1)\theta)}{1 - \exp(-2\theta)}.
\end{aligned}
$$

Figure 4 displays the "actual sample size" N for a particular pair (ε, δ) as a function of θ for the three distributions. The values N were obtained by running the code 100 times for each θ to calculate the empirical probability of success.

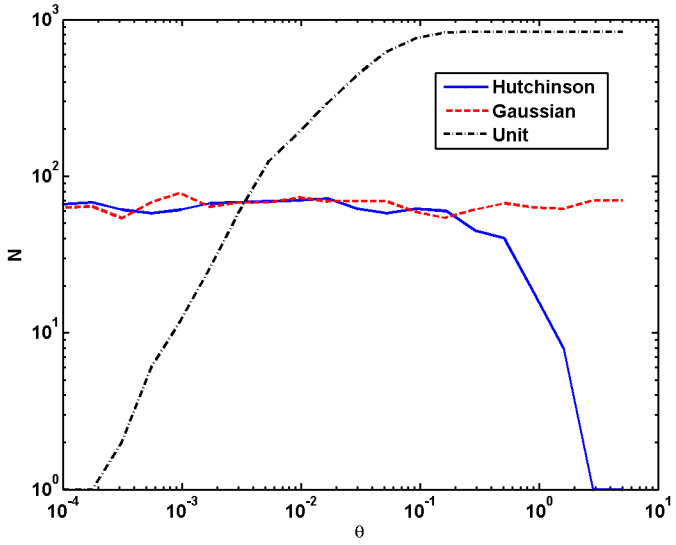


Figure 4: Example 2. For the rank-1 matrix arising from a rapidly-decaying vector with $n = 1000$, this log-log plot depicts the actual sample size N required for (2) to hold with $\varepsilon = \delta = 0.2$, vs. various values of θ. In the legend, "Unit" refers to the random sampling method without replacement.

In this example the distribution of $\mathcal{K}_H^j$ values gets progressively worse with heavier tail values as θ gets larger. However, recall that this matters in terms of the sufficient bounds (4) and (7) only so long as $\mathcal{K}_H < 3$. Here the crossover point happens roughly when $\theta \sim 1/(2n)$. Indeed, for large values of θ the required sample size actually drops when using the Hutchinson method: Theorem 2, being only a sufficient condition, merely distinguishes types of matrices for which Hutchinson is expected to be efficient*, while making no claim regarding those matrices for which it is an inefficient estimator.*

On the other hand, Theorem 5 clearly distinguishes the types of matrices for which the Gaussian method is expected to be inefficient*, because its condition is necessary rather than sufficient. Note that N (the red curve in Figure 4) does not change much as a function of θ, which agrees with the fact that the matrix rank stays fixed and low at $r = 1$.*

The unit vector estimator, unlike Hutchinson, deteriorates steadily as θ is increased, because this estimator ignores off-diagonal elements. However, for small enough values of θ the $\mathcal{K}_U^{(i,j)}$'s are spread tightly near zero, and the unit vector method, as predicted by Theorem 7, requires a very small sample size.

For Examples 3 and 5 below, given (ε, δ), we plot the probability of success, i.e., $Pr\left(|tr_D^N(A) - tr(A)| \leq \varepsilon\ tr(A)\right)$ for increasing values of N, starting from $N = 1$. We stop when for a given N, the probability of success is greater than or equal to $1 - \delta$. In order to evaluate this for each N, we run the experiments 500 times and calculate the empirical probability.

In the figures below, 'With Rep.' and 'Without Rep.' refer to uniform unit sampling with and without replacement, respectively. In all cases, by default, $\varepsilon = \delta = .05$. We also

provide distribution plots of the quantities $\mathcal{K}_H^j$, $\mathcal{K}_G^j$ and $\mathcal{K}_U^{(i,j)}$ appearing in (6), (9) and (17), respectively. These quantities are indicators for the performance of the Hutchinson, Gaussian and unit vector estimators, respectively, as evidenced not only by Theorems 2, 3 and 7, but also in Examples 1 and 2, and by the fact that the performance of the Gaussian and unit vector estimators is not affected by the energy of the off-diagonal matrix elements.

Example 3 (Data fitting with many experiments) *A major source of applications where trace estimation is central arises in problems involving least squares data fitting with many experiments. In its simplest, linear form, we look for* $\mathbf{m} \in \mathbb{R}^l$ *so that the misfit function*

$$\phi(\mathbf{m}) = \sum_{i=1}^{n} \|J_i\mathbf{m} - \mathbf{d}_i\|^2, \tag{20a}$$

for given data sets $\mathbf{d}_i$ *and sensitivity matrices* J_i*, is either minimized or reduced below some tolerance level. The* $m \times l$ *matrices* J_i *are very expensive to calculate and store, so this is avoided altogether, but evaluating* $J_i\mathbf{m}$ *for any suitable vector* $\mathbf{m}$ *is manageable. Moreover,* n *is large. Next, writing* (20a) *using the Frobenius norm as*

$$\phi(\mathbf{m}) = \|C\|_F^2, \tag{20b}$$

where C *is* $m \times n$ *with the* j*th column* $C_j = J_j\mathbf{m} - \mathbf{d}_j$*, and defining the SPSD matrix* $A = C^tC$*, we have*

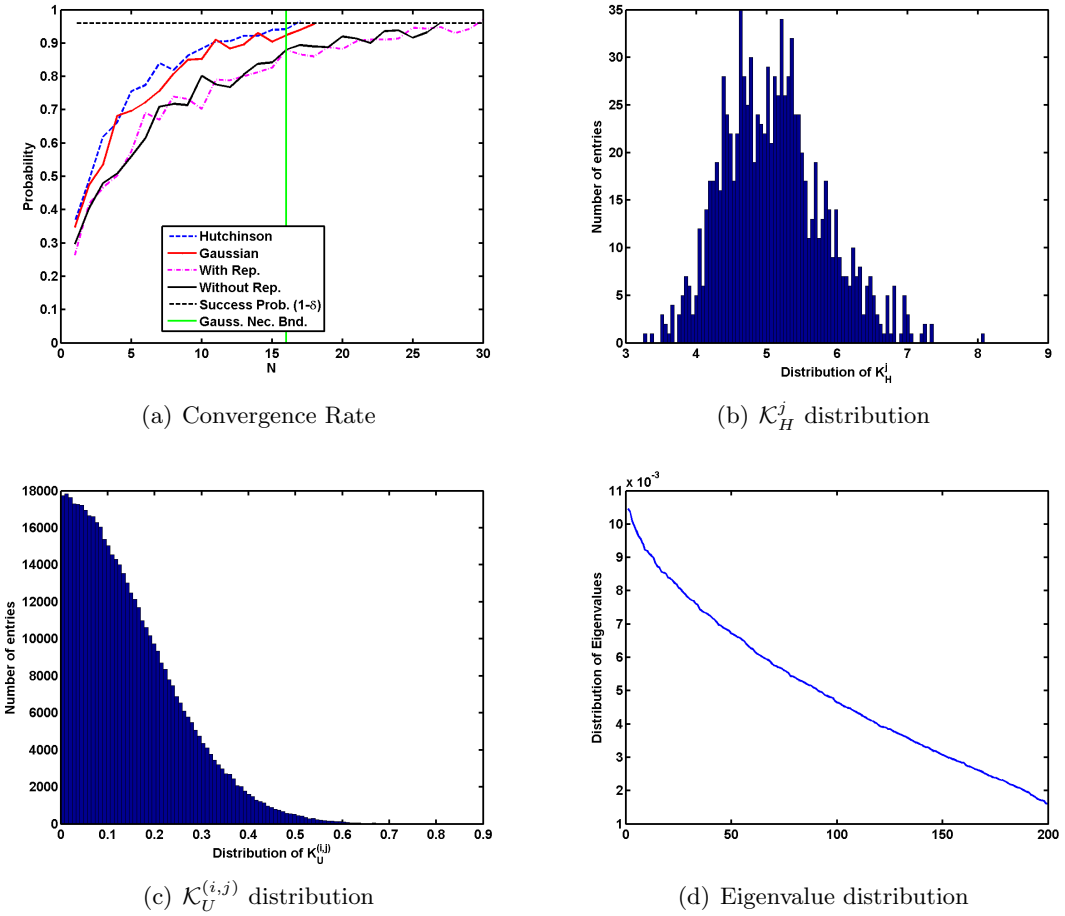
$$\phi(\mathbf{m}) = tr(A). \tag{20c}$$

Cheap estimates of the misfit function $\phi(\mathbf{m})$ *are then sought by approximating the trace in* (20c) *using only* N *(rather than* n*) linear combinations of the columns of* C*, which naturally leads to expressions of the form* (1)*. Hutchinson and Gaussian estimators in a similar or more complex context were considered in [9, 11, 18].*

Drawing the $\mathbf{w}_i$ *as random unit vectors instead is a method proposed in [7] and compared to others in [13], where it is called "random subset": this latter method can have efficiency advantages that are beyond the scope of the presentation here. Typically,* $m \ll n$*, and thus the matrix* A *is dense and often has low rank.*

Furthermore, the signs of the entries in C *can be, at least to some extent, considered random. Hence we consider below matrices* $A = C^tC$ *whose entries are Gaussian random variables, obtained using the* MATLAB *command* `C = randn(m,n)`*. We use* $m = 200$ *and hence the rank is, almost surely,* $r = 200$*.*

It can be seen from Figure 5(a) that the Hutchinson and the Gaussian methods perform similarly here. The sample size required by both unit vector estimators is approximately twice that of the Gaussian and Hutchinson methods. This relative behaviour agrees with our observations in the context of actual application as described above, see [13]. From Figure 5(d), the eigenvalue distribution of the matrix is not very badly skewed, which helps the Gaussian method perform relatively well for this sort of matrix. On the other hand, by Figure 5(b) the relative ℓ_2 *energies of the off-diagonals are far from being small, which is not favourable for the Hutchinson method. These two properties, in combination, result*



(a) Convergence Rate

(b) $\mathcal{K}_H^j$ distribution

(c) $\mathcal{K}_U^{(i,j)}$ distribution

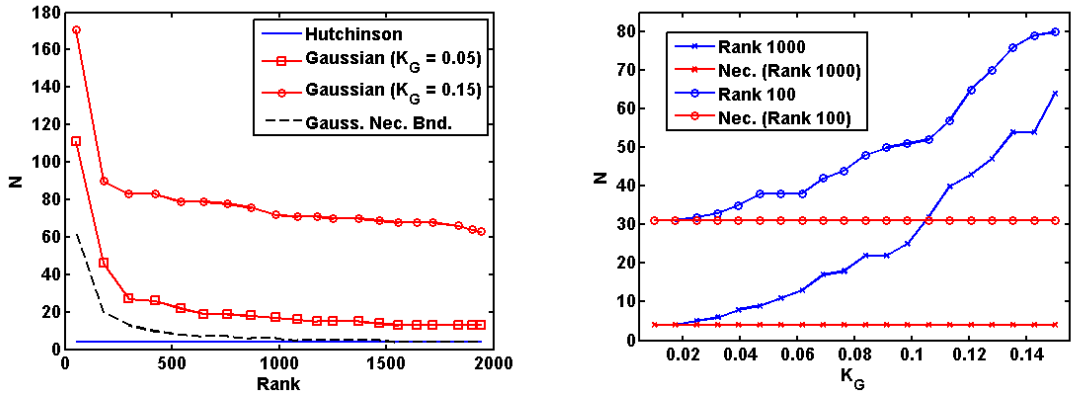
(d) Eigenvalue distribution

Figure 5: Example 3. A dense SPSD matrix A is constructed using MATLAB's `randn`. Here $n = 1000$, $r = 200$, $tr(A) = 1, \mathcal{K}_G = 0.0105$, $\mathcal{K}_H = 8.4669$ and $\mathcal{K}_U = 0.8553$. The method convergence plots in (a) are for $\varepsilon = \delta = .05$.

in the similar performance of the Hutchinson and Gaussian methods despite the relatively low rank. The contrast between $\mathcal{K}_U^{(i,j)}$'s is not too large according to Figure 5(c), hence a relatively decent performance of both unit vector (or, random sampling) methods is observed. There is no reason to insist on avoiding repetition here either.

Example 4 (Effect of rank and $\mathcal{K}_G$ on the Gaussian estimator) *In this example we plot the actual sample size N required for (2) to hold. In order to evaluate (2), we repeat the experiments 500 times and calculate the empirical probability. In all experiments, the sample sizes predicted by (4) and (5) were so pessimistic compared with the true N that we simply did not include them in the plots.*

In order to concentrate only on rank and $\mathcal{K}_G$ variation, we make sure that in all experiments $\mathcal{K}_H \ll 1$. For the results displayed in Figure 6(a), where r is varied for each of two values of $\mathcal{K}_G$, this is achieved by playing with MATLAB*'s normal random generator function* `sprandn`*. For Figure 6(b), where $\mathcal{K}_G$ is varied for each of two values of r, diagonal*



(a) `sprandn`, $n = 5,000$ (b) diagonal, $n = 10,000$

Figure 6: Example 4. The behaviour of the Gaussian method with respect to rank and $\mathcal{K}_G$. We set $\varepsilon = \delta = .05$ and display the necessary condition (14) as well.

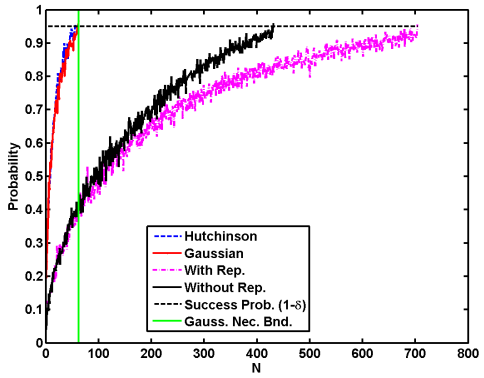
matrices are utilized: we start with a uniform distribution of the eigenvalues and gradually make this distribution more skewed, resulting in an increased $\mathcal{K}_G$. The low $\mathcal{K}_H$ values cause the Hutchinson method to look very good, but that is not our focus here.

It can be clearly seen from Figure 6(a) that as the matrix rank gets lower, the sample size required for the Gaussian method grows significantly. For a given rank, the matrix with a smaller $\mathcal{K}_G$ requires smaller sample size. From Figure 6(b) it can also be seen that for a fixed rank, the matrix with more skewed $\mathcal{K}_G^j$'s distribution (marked here by a larger $\mathcal{K}_G$) requires a larger sample size.

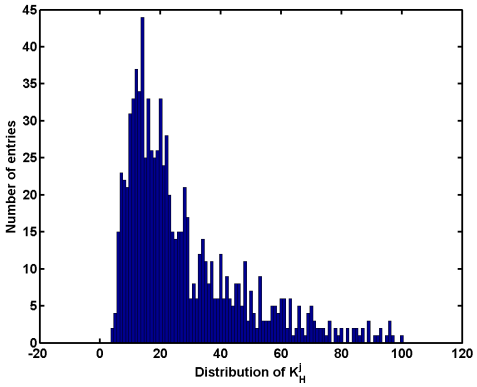
Example 5 (Method performance for different matrix properties) *Next we consider a much more general setting than that in Example 4, and compare the performance of different methods with respect to various matrix properties. The matrix A is constructed as in Example 3, except that also a uniform distribution is used. Furthermore, a parameter d controlling denseness of the created matrix is utilized. This is achieved in* MATLAB *using the commands* `C=sprandn(m,n,d)` *or* `C=sprand(m,n,d)`*. By changing* `m` *and* `d` *we can change the matrix properties $\mathcal{K}_H$, $\mathcal{K}_G$ and $\mathcal{K}_U$ while keeping the rank r fixed across experiments. We maintain $n = 1000$, $tr(A) = 1$ and $\varepsilon = \delta = .05$ throughout. In particular, the four figures related to this example are comparable to Figure 5 but for a lower rank.*

By comparing Figures 7 and 8, as well as 9 and 10, we can see how not only the values of $\mathcal{K}_H$, $\mathcal{K}_G$ and $\mathcal{K}_U$, but also the distribution of the quantities they maximize matters. Note how the performance of both unit vector strategies is negatively affected with increasing average values of $\mathcal{K}_U^{(i,j)}$'s. From the eigenvalue (or $\mathcal{K}_G^j$) distribution of the matrix, it can also be seen that the Gaussian estimator is heavily affected by the skewness of the distribution of the eigenvalues (or $\mathcal{K}_G^j$'s): given the same r and n, as this eigenvalue distribution becomes increasingly uneven, the Gaussian method requires larger sample size.

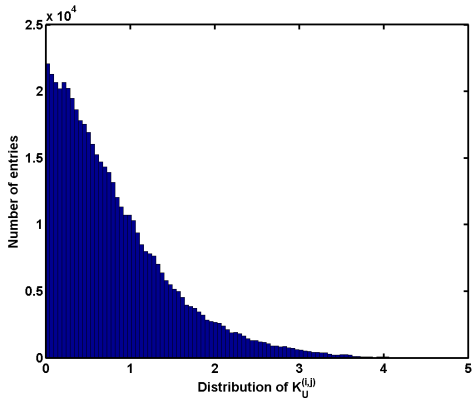
Note that comparing the performance of the methods on different matrices solely based on their values $\mathcal{K}_H$, $\mathcal{K}_G$ or $\mathcal{K}_U$ can be misleading. This can be seen for instance by considering



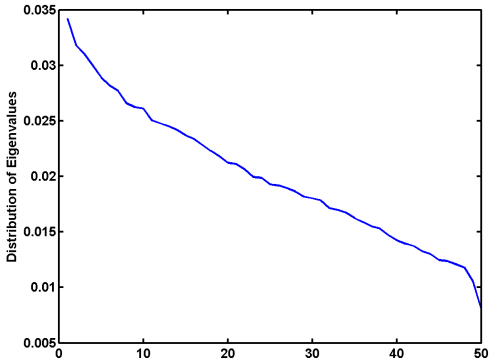
(a) Convergence Rate



(b) $\mathcal{K}_H^j$ distribution ($\mathcal{K}_H^j \leq 100$)



(c) $\mathcal{K}_U^{(i,j)}$ distribution

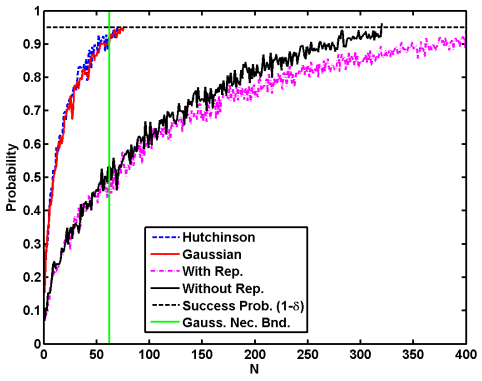


(d) Eigenvalue distribution

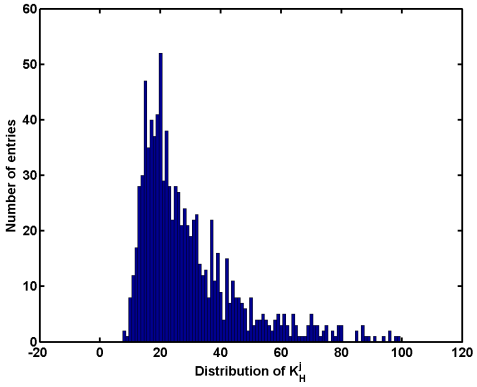
Figure 7: Example 5. A sparse matrix `(d = 0.1)` is formed using `sprandn`. Here $r = 50$, $\mathcal{K}_G = 0.0342$, $\mathcal{K}_H = 15977.194$ and $\mathcal{K}_U = 4.8350$.

the performance of the Hutchinson method in Figures 7, 8, 9 and 10 and comparing their respective $\mathcal{K}_H^j$ distributions as well as $\mathcal{K}_H$ values. Indeed, none of our 6 *sufficient bounds can be guaranteed to be generally tight. As remarked also earlier, this is an artifact of the generality of the proved results.*

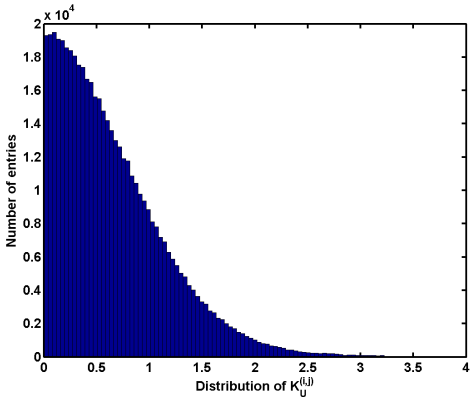
Note also that rank and eigenvalue distribution of a matrix have no direct effect on the performance of the Hutchinson method: by Figures 9 and 10 it appears to only depend on the $\mathcal{K}_H^j$ distribution. In these figures, one can observe that the Gaussian method is heavily affected by the low rank and the skewness of the eigenvalues. Thus, if the distribution of $\mathcal{K}_H^j$'s is favourable to the Hutchinson method and yet the eigenvalue distribution is rather skewed, we can expect a significant difference between the performance of the Gaussian and Hutchinson methods.



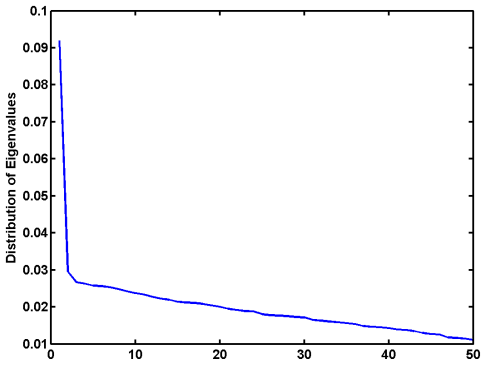
(a) Convergence Rate



(b) $\mathcal{K}_H^j$ distribution ($\mathcal{K}_H^j \leq 100$)

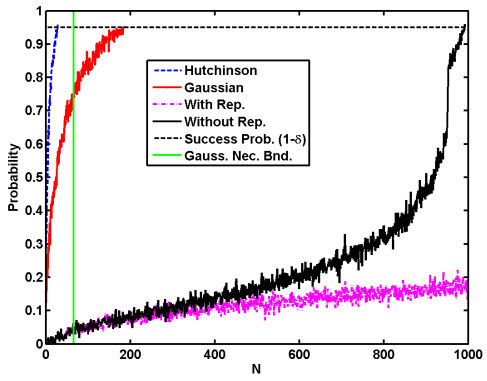


(c) $\mathcal{K}_U^{(i,j)}$ distribution

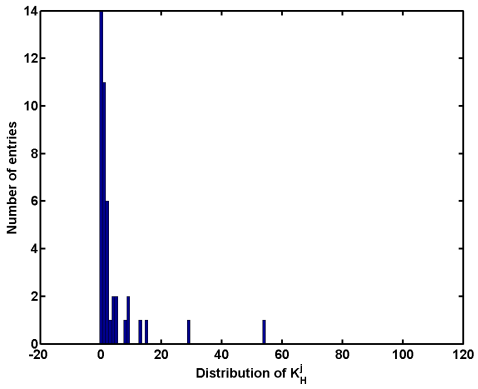


(d) Eigenvalue distribution

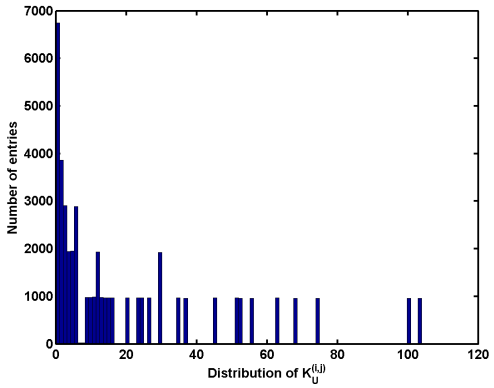
Figure 8: Example 5. A sparse matrix `(d = 0.1)` is formed using `sprand`. Here $r = 50, \mathcal{K}_G = 0.0919$, $\mathcal{K}_H = 11624.58$ and $\mathcal{K}_U = 3.8823$.



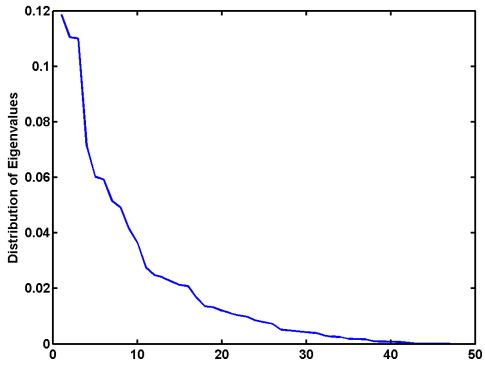
(a) Convergence Rate



(b) $\mathcal{K}_H^j$ distribution ($\mathcal{K}_H^j \leq 50$)



(c) $\mathcal{K}_U^{(i,j)}$ distribution

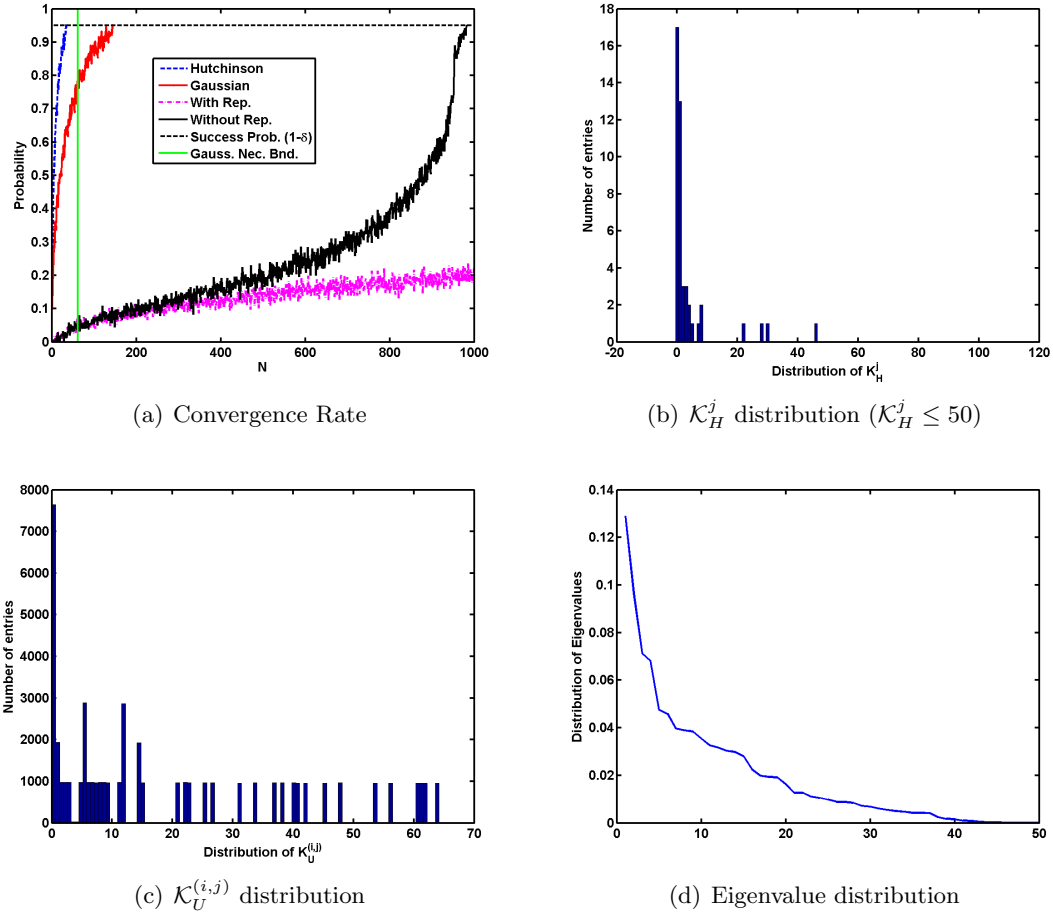


(d) Eigenvalue distribution

Figure 9: Example 5. A very sparse matrix `(d = 0.01)` is formed using `sprandn`. Here $r = 50$, $\mathcal{K}_G = 0.1186$, $\mathcal{K}_H = 8851.8$ and $\mathcal{K}_U = 103.9593$.

6 Conclusions and further thoughts

In this article we have proved six sufficient bounds for the minimum sample size N required to reach, with probability $1-\delta$, an approximation for $tr(A)$ to within a relative tolerance ε. Two such bounds apply to each of the three estimators considered in Sections 2, 3 and 4, respectively. In Section 3 we have also proved a necessary bound for the Gaussian estimator. These bounds have all been verified numerically through many examples, some of which are summarized in Section 5.



(a) Convergence Rate

(b) $\mathcal{K}_H^j$ distribution ($\mathcal{K}_H^j \leq 50$)

(c) $\mathcal{K}_U^{(i,j)}$ distribution

(d) Eigenvalue distribution

Figure 10: Example 5. A very sparse matrix `(d = 0.01)` is formed using `sprand`. Here $r = 50$, $\mathcal{K}_G = 0.1290$, $\mathcal{K}_H = 1611.34$ and $\mathcal{K}_U = 64.1707$.

Two of these bounds, namely, (4) for Hutchinson and (5) for Gaussian, are immediately computable without knowing anything else about the SPSD matrix A. In particular, they are independent of the matrix size n. As such they may be very pessimistic. And yet, in some applications (for instance, in exploration geophysics) where n can be very large and ε need not be very small due to uncertainty, these bounds may indeed provide the comforting assurance that $N \ll n$ suffices (say, n is in the millions and N in the thousands). Generally,

these two bounds have the same quality.

The underlying objective in this work, which is to seek a small N satisfying (2), is a natural one for many applications and follows that of other works. But when it comes to comparing different methods, it is by no means the only performance indicator. For example, variance can also be considered as a ground to compare different methods. However, one needs to exercise caution to avoid basing the entire comparison solely on variance: it is possible to generate examples where a linear combination of $\mathcal{X}^2$ random variables has smaller variance, yet higher tail probability.

The lower bound (14) that is available only for the Gaussian estimator may allow better prediction of the actual required N, in cases where the rank r is known. At the same time it also implies that the Gaussian estimator can be inferior in cases where r is small. The Hutchinson estimator does not enjoy a similar theory, but empirically does not suffer from the same disadvantage either.

The matrix-dependent quantities $\mathcal{K}_H$, $\mathcal{K}_G$ and $\mathcal{K}_U$, defined in (6), (9) and (17), respectively, are not easily computable for any given implicit matrix A. However, the results of Theorems 2, 3 and 7 that depend on them can be more indicative than the general bounds. In particular, examples where one method is clearly better than the others can be isolated in this way. At the same time, the sufficient conditions in Theorems 2, 3 and 7, merely distinguish the types of matrices for which the respective methods are expected to be efficient, and make no claims regarding those matrices for which they are inefficient estimators. This is in direct contrast with the necessary condition in Theorem 5.

It is certainly possible in some cases for the required N to go over n. In this connection it is important to always remember the deterministic method which obtains $tr(A)$ in N applications of unit vectors: if N grows above n in a particular stochastic setting then it may be best to abandon ship and choose the safe, deterministic way.

Acknowledgment We thank our three anonymous referees for several valuable comments which have helped to improve the text.
Part of this work was completed while the second author was visiting IMPA, Rio de Janeiro, supported by a Brazilian "Science Without Borders" grant and hosted by Prof. J. Zubelli. Thank you all.

References

[1] M. Abramowitz. *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover, 1974.

[2] D. Achlioptas. Database-friendly random projections. In *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 01*, volume 20, pages 274–281, 2001.

[3] H. Avron. Counting triangles in large graphs using randomized matrix trace estimation. *Workshop on Large-scale Data Mining: Theory and Applications*, 2010.

[4] H. Avron and S. Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *JACM*, 58(2), 2011. Article 8.

[5] Z. Bai, M. Fahey, and G. Golub. Some large scale matrix computation problems. *J. Comput. Appl. Math.*, 74:71–89, 1996.

[6] C. Bekas, E. Kokiopoulou, and Y. Saad. An estimator for the diagonal of a matrix. *Appl. Numer. Math.*, 57:12141229, 2007.

[7] K. van den Doel and U. Ascher. Adaptive and stochastic algorithms for EIT and DC resistivity problems with piecewise constant solutions and many measurements. *SIAM J. Scient. Comput.*, 34:DOI: 10.1137/110826692, 2012.

[8] G. H. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

[9] E. Haber, M. Chung, and F. Herrmann. An effective method for parameter estimation with PDE constraints with multiple right-hand sides. *SIAM J. Optimization*, 22:739–757, 2012.

[10] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *J. Comm. Stat. Simul.*, 19:433–450, 1990.

[11] T. Van Leeuwen, S. Aravkin, and F. Herrmann. Seismic waveform inversion by stochastic optimization. *Hindawi Intl. J. Geophysics*, 2011:doi:10.1155/2011/689041, 2012.

[12] A. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill; 3rd edition, 1974.

[13] F. Roosta-Khorasani, K. van den Doel, and U. Ascher. Stochastic algorithms for inverse problems involving PDEs and many measurements. *SIAM J. Scient. Comput.*, 2014. To appear.

[14] R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, 2:39–48, 1974.

[15] A. Shapiro, D. Dentcheva, and D. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. Piladelphia: SIAM, 2009.

[16] G. J. Székely and N. K. Bakirov. Extremal probabilities for Gaussian quadratic forms. *Probab. Theory Related Fields*, 126:184–202, 2003.

[17] J. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. *SODA*, pages 978–986, 2009. SIAM.

[18] J. Young and D. Ridzal. An application of random projection to parameter estimation in partial differential equations. *SIAM J. Scient. Comput.*, 34:A2344–A2365, 2012.