Numer. Math. (2017) 137:353–395
DOI 10.1007/s00211-017-0880-z

Numerische Mathematik

Randomized matrix-free trace and log-determinant estimators

Arvind K. Saibaba[1] · **Alen Alexanderian**[1] · **Ilse C. F. Ipsen**[1]

Received: 14 May 2016 / Revised: 20 January 2017 / Published online: 6 April 2017

Abstract We present randomized algorithms for estimating the trace and determinant of Hermitian positive semi-definite matrices. The algorithms are based on subspace iteration, and access the matrix only through matrix vector products. We analyse the error due to randomization, for starting guesses whose elements are Gaussian or Rademacher random variables. The analysis is cleanly separated into a structural (deterministic) part followed by a probabilistic part. Our absolute bounds for the expectation and concentration of the estimators are non-asymptotic and informative even for matrices of low dimension. For the trace estimators, we also present asymptotic bounds on the number of samples (columns of the starting guess) required to achieve a user-specified relative error. Numerical experiments illustrate the performance of the estimators and the tightness of the bounds on low-dimensional matrices, and on a challenging application in uncertainty quantification arising from Bayesian optimal experimental design.

The third author acknowledges the support from the XDATA Program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323 FA8750-12-C-0323.

✉ Arvind K. Saibaba
asaibab@ncsu.edu
http://www4.ncsu.edu/~asaibab/

Alen Alexanderian
alexanderian@ncsu.edu
http://www4.ncsu.edu/~aalexan3/

Ilse C. F. Ipsen
ipsen@ncsu.edu
http://www4.ncsu.edu/~ipsen/

[1] Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA

Mathematics Subject Classification 68W20 · 65F15 · 65F40 · 65F25 · 65F35 · 15B52 · 62F15

1 Introduction

Computing the trace of high-dimensional matrices is a common problem in various areas of applied mathematics, such as evaluation of uncertainty quantification measures in parameter estimation and inverse problems [3,17,18,38], and generalized cross validation (GCV) [15,46,47].

Our original motivation came from trace and log-determinant computations of high-dimensional operators in Bayesian optimal experimental design (OED) [11]. Of particular interest is OED for Bayesian inverse problems that are constrained by partial differential equations (PDEs) with high-dimensional parameters. In Sect. 6 we give an example of such a Bayesian inverse problem and illustrate the evaluation of OED criteria with our algorithms.

Trace and determinant computations are straightforward if the matrices are explicitly defined, and one has direct access to individual matrix entries. The trace is computed as the sum of the diagonal elements, while the determinant can be computed as the product of the diagonal elements from a triangular factor [21, Section 14.6]. However, if the matrix dimension is large, or explicit access to individual entries is expensive, alternative methods are needed.

Here we focus on computing the trace and log-determinant of implicitly defined matrices, where the matrix can be accessed only through matrix vector products. We present randomized estimators for $\mathrm{trace}(\mathbf{A})$ and[1] $\mathrm{logdet}(\mathbf{I}+\mathbf{A})$ for Hermitian, or real symmetric, positive semi-definite matrices $\mathbf{A} \in \mathbb{C}^{n\times n}$.

1.1 Main features of our estimator

Our estimators are efficient and easy to implement, as they are based on randomized subspace iteration; and they are accurate for many matrices of interest. Unlike Monte Carlo estimators, see Sect. 1.3, whose variance depends on individual matrix entries, our error bounds rely on eigenvalues. To this end we need to assume that the matrix has a well-defined dominant eigenspace, with a large eigenvalue gap whose location is known. Our bounds quantify the effect of the starting guess on the dominant eigenspace, and are informative even in the non-asymptotic regime, for matrices of low dimension. Our estimators, although biased, can be much more accurate than Monte Carlo estimators.

1.2 Contributions

Our paper makes the following four contributions.

[1] The square matrix $\mathbf{I}$ denotes the identity, with ones on the diagonal and zeros everywhere else.

1.2.1 Randomized estimators

Assume that the Hermitian positive semi-definite matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ has k dominant eigenvalues separated by a gap from the remaining $n-k$ sub-dominant eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_k \gg \lambda_{k+1} \geq \cdots \geq \lambda_n$. The idea is to capture the dominant eigenspace associated with $\lambda_1, \ldots, \lambda_k$ via a low-rank approximation $\mathbf{T}$ of $\mathbf{A}$. Our estimators (Sect. 2.1) for $\mathsf{trace}(\mathbf{T}) \approx \mathsf{trace}(\mathbf{A})$ and $\log\det(\mathbf{I}+\mathbf{T}) \approx \log\det(\mathbf{I}+\mathbf{A})$ appear to be new. Here $\mathbf{T} \equiv \mathbf{Q}^*\mathbf{A}\mathbf{Q} \in \mathbb{C}^{\ell\times\ell}$ where $k \leq \ell \ll n$. The matrix $\mathbf{Q}$ approximates the dominant eigenspace of $\mathbf{A}$, and is computed from q iterations of subspace iteration applied to a starting guess $\boldsymbol{\Omega}$, followed by the thin QR factorization of $\mathbf{A}^q\boldsymbol{\Omega}$.

1.2.2 Structural and probabilistic error analysis

We derive absolute error bounds for $\mathsf{trace}(\mathbf{T})$ and $\log\det(\mathbf{I}+\mathbf{T})$, for starting guesses that are Gaussian random variables (Sect. 2.2.2), and Rademacher random variables (Sect. 2.2.3) The derivations are cleanly separated into a "structural" (deterministic) part, followed by a probabilistic part.

Structural analysis (Sect. 3) These are perturbation bounds that apply to any matrix $\boldsymbol{\Omega}$, be it random or deterministic. The resulting absolute error bounds for $\mathsf{trace}(\mathbf{T})$ and $\log\det(\mathbf{I}+\mathbf{T})$ imply that the estimators are accurate if: (1) the starting guess $\boldsymbol{\Omega}$ has a large contribution in the dominant eigenspace; (2) the eigenvalue gap is large; and (3) the sub-dominant eigenvalues are negligible.

The novelty of our analysis is the focus on the eigendecomposition of $\mathbf{A}$. In contrast, as discussed in Sect. 2.3, the analyses of Monte Carlo estimators depend on the matrix entries, and do not take into account the spectral properties of $\mathbf{A}$.

To understand the contribution of the random starting guess $\boldsymbol{\Omega}$, let the columns of $\mathbf{U}_1 \in \mathbb{C}^{n\times k}$ represent an orthonormal basis for the dominant eigenspace, while the columns of $\mathbf{U}_2 \in \mathbb{C}^{n\times(n-k)}$ represent an orthonormal basis associated with the $n-k$ sub-dominant eigenvalues. The "projections" of the starting guess on the respective eigenspaces are are $\boldsymbol{\Omega}_1 \equiv \mathbf{U}_1^*\boldsymbol{\Omega} \in \mathbb{C}^{k\times\ell}$ and $\boldsymbol{\Omega}_2 \equiv \mathbf{U}_2^*\boldsymbol{\Omega} \in \mathbb{C}^{(n-k)\times\ell}$.

The success of $\mathbf{T}$ in capturing the dominant subspace $\mathsf{range}(\mathbf{U}_1)$ depends on the quantity[2] $\|\boldsymbol{\Omega}_2\|_2\|\boldsymbol{\Omega}_1^\dagger\|_2$.

Probabilistic analysis (Sect. 4). We bound the norms of the projections $\|\boldsymbol{\Omega}_2\|_2$ and $\|\boldsymbol{\Omega}_1^\dagger\|_2$ for starting guesses $\boldsymbol{\Omega}$ that are Gaussian or Rademacher random matrices.

For Gaussian starting guesses, we present bounds for the mean (or expectation), and concentration about the mean, based on existing bounds for the spectral norms of Gaussian random matrices and their pseudo-inverse.

For Rademacher starting guesses, we present Chernoff-type concentration inequalities, and show that $\ell \sim (k+\log n)\log k$ samples are required to guarantee $\mathsf{rank}(\boldsymbol{\Omega}_1) = k$ with high probability.

[2] The superscript † denotes the Moore–Penrose inverse.

1.2.3 Asymptotic efficiency

One way to quantify the efficiency of a Monte Carlo estimator is a so-called (ϵ, δ) estimator [6], which bounds the number of samples required to achieve a relative error of at most ϵ with probability at least $1 - \delta$. Our asymptotic (ϵ, δ) bounds (Theorem 4) show that our trace estimator can require significantly fewer samples than Monte Carlo estimators.

1.2.4 Numerical experiments

Comprehensive numerical experiments corroborate the performance of our estimators, and illustrate that our error bounds hold even in the non-asymptotic regime, for matrices of small dimension (Sect. 5). Motivated by our desire for fast and accurate estimation of uncertainty measures in Bayesian inverse problems, we present a challenging application from Bayesian OED (Sect. 6).

1.3 Related work

We demonstrate that the novelty of our paper lies in both, the estimators and their analysis.

There are several popular estimators for the trace of an implicit, Hermitian positive semi-definite matrix $\mathbf{A}$, the simplest one being a Monte Carlo estimator. It requires only matrix vector products with N independently generated random vectors $\mathbf{z}_j$ and computes

$$\mathsf{trace}(\mathbf{A}) \approx \frac{1}{N} \sum_{j=1}^{N} \mathbf{z}_j^* \mathbf{A} \mathbf{z}_j .$$

The original algorithm, proposed by Hutchinson [24], uses Rademacher random vectors and produces an unbiased estimator. Unbiased estimators can also be produced with other distributions, such as Gaussian random vectors, or columns of the identity matrix that are sampled uniformly with or without replacement [6,35], see the detailed comparison in Sect. 2.3.

Randomized matrix algorithms [19,28] could furnish a potential alternative for trace estimation. Low-rank approximations of $\mathbf{A}$ can be efficiently computed with randomized subspace iteration [26,29] or Nyström methods [14], and their accuracy is quantified by probabilistic error bounds in the spectral and Frobenius norms. Yet we were not able to find error bounds for the corresponding trace estimator in the literature.

Like our estimators, the spectrum-sweeping method [27, Algorithm 5] is based on a randomized low-rank approximation of $\mathbf{A}$. However, it is designed to compute the trace of smooth functions of Hermitian matrices in the context of *density of state* estimations in quantum physics. Numerical experiments illustrate that the method can be much faster than Hutchinson's estimator, but there is no formal convergence analysis.

A related problem is the trace computation of the matrix inverse. One can combine a Hutchinson estimator $\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i^* \mathbf{A}^{-1}\mathbf{z}_i$ with quadrature rules for approximating the bilinear forms $\mathbf{z}_i^* \mathbf{A}^{-1}\mathbf{z}_i$ [7,8]. For matrices $\mathbf{A}$ that are sparse, banded, or whose off-diagonal entries decay away from the main diagonal, one can use a probing method [41] to estimate the diagonal of $\mathbf{A}^{-1}$ with carefully selected vectors that exploit structure and sparsity.

Computation of the log-determinant is required for maximum likelihood estimation in areas like machine learning, robotics and spatial statistics [48]. This can be achieved by applying a Monte Carlo algorithm to the log-determinant directly [9], or to an expansion [32,48].

Alternatively one can combine the identity $\log\det(\mathbf{A}) = \mathsf{trace}(\log(\mathbf{A}))$ [7, Section 3.1.4] with a Monte Carlo estimator for the trace. Since computation of $\log(\mathbf{A})$, whether with direct or matrix-free methods, is expensive for large $\mathbf{A}$, the logarithm can be expanded into a Taylor series [10,32,48], a Chebyshev polynomial [20], or a spline [4,12].

2 Algorithms and main results

We present the algorithm for randomized subspace iteration (Sect. 2.1), followed by the main error bounds for the trace and logdet estimators (Sect. 2.2), and conclude with a discussion of Monte Carlo estimators (Sect. 2.3).

2.1 The algorithm

We sketch the estimators for $\mathsf{trace}(\mathbf{A})$ and $\log\det(\mathbf{I}_n+\mathbf{A})$, for Hermitian positive semi-definite matrices $\mathbf{A} \in \mathbb{C}^{n\times n}$ with k dominant eigenvalues. The estimators relinquish the matrix $\mathbf{A}$ of order n for a matrices $\mathbf{T}$ of smaller dimension $\ell \ll n$ computed with Algorithm 1, so that $\mathsf{trace}(\mathbf{T})$ is an estimator for $\mathsf{trace}(\mathbf{A})$, and $\log\det(\mathbf{I}_\ell + \mathbf{T})$ an estimator for $\log\det(\mathbf{I}_n + \mathbf{A})$.

Algorithm 1 is an idealized version of randomized subspace iteration. Its starting guess is a random matrix $\mathbf{\Omega}$ with $k \le \ell \ll n$ columns, sampled from a fixed distribution, that is then subjected to q power iterations with $\mathbf{A}$. A thin QR decomposition of the resulting product $\mathbf{A}^q\mathbf{\Omega}$ produces a matrix $\mathbf{Q}$ with orthonormal columns. The output of Algorithm 1 is the $\ell \times \ell$ restriction $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ of $\mathbf{A}$ to $\mathsf{span}(\mathbf{Q})$.

Algorithm 1 Randomized subspace iteration (idealized version)

Input: Hermitian positive semi-definite matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ with target rank k,
Number of subspace iterations $q \ge 1$
Starting guess $\mathbf{\Omega} \in \mathbb{C}^{n\times \ell}$ with $k \le \ell \le n-k$ columns
Output: Matrix $\mathbf{T} \in \mathbb{C}^{\ell\times\ell}$
1: Multiply $\mathbf{Y} = \mathbf{A}^q\mathbf{\Omega}$
2: Thin QR factorization $\mathbf{Y} = \mathbf{Q}\mathbf{R}$
3: Compute $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$.

The idealized subspace iteration in Algorithm 1 can be numerically unstable. The standard remedy is to alternate matrix products and QR factorizations [37, Algorithm 5.2]. In practice, one can trade off numerical stability and efficiency by computing the QR factorization once every few steps [37, Algorithm 5.2]. Throughout this paper, we assume exact arithmetic and do not take into account finite precision effects.

Random starting guess The entries of $\mathbf{\Omega}$ are i.i.d.[3] variables from one of the two distributions: standard normal (zero mean and variance 1), or Rademacher (values ± 1 with equal probability).

As in Sect. 1.2.2, let $\mathbf{\Omega}_1 \equiv \mathbf{U}_1^* \mathbf{\Omega}$ and $\mathbf{\Omega}_2 \equiv \mathbf{U}_2^* \mathbf{\Omega}$ be the respective "projections" of the starting guess on the dominant and sub-dominant eigenspaces. The success of $\mathbf{T}$ in capturing the dominant subspace depends on the quantity $\|\mathbf{\Omega}_2\|_2 \|\mathbf{\Omega}_1^\dagger\|_2$. We make the reasonable assumption $\mathsf{rank}(\mathbf{\Omega}_1) = k$, so that $\mathbf{\Omega}_1^\dagger$ is a right inverse. Asymptotically, for both Gaussian [19, Propositions A.2 and A.4] and Rademacher random matrices [36, Theorem 1.1], $\|\mathbf{\Omega}_2\|_2$ grows like $\sqrt{n-k} + \sqrt{\ell}$, and $1/\|\mathbf{\Omega}_1^\dagger\|_2$ like $\sqrt{\ell} - \sqrt{k}$.

Other than that, however, there are major differences. For Gaussian random matrices, the number columns in $\mathbf{\Omega}$ is $\ell = k + p$, where p is a user-specified oversampling parameter. The discussion in [16, Section 5.3] indicates that the bounds in Sect. 2.2.2 should hold with high probability for $p \lesssim 20$. Asymptotically, the required number of columns in a Gaussian starting guess is $\ell \sim k$.

In contrast, the number of columns in a Rademacher random matrix cannot simply be relegated, once and for all, to a fixed oversampling parameter, but instead show a strong dependence on the dimension k of the dominant subspace and the matrix dimension n. We show (Sect. 4) that the error bounds in Sect. 2.2.3 hold with high probability, if the number of columns in $\mathbf{\Omega}$ is $\ell \sim (k + \log n) \log k$. This behavior is similar to that of structured random matrices from sub-sampled random Fourier transforms and sub-sampled random Hadamard transforms [42]. It is not yet clear, though, whether the asymptotic factor $(k + \log n) \log k$ is tight, or whether it is merely an artifact of the analysis.

2.2 Main results

We clarify our assumptions (Sect. 2.2.1), before presenting the main error bounds for the trace and logdet estimators, when the random matrices for the starting guess are Gaussian (Sect. 2.2.2) and Rademacher (Sect. 2.2.3).

2.2.1 Assumptions

Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a Hermitian positive semi-definite matrix with eigenvalue decomposition

$$\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^*, \qquad \mathbf{\Lambda} = \mathsf{diag}\left(\lambda_1 \cdots \lambda_n\right) \in \mathbb{R}^{n \times n},$$

[3] independent and identically distributed.

where the eigenvector matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$ is unitary, and the eigenvalues are ordered, $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$.

We assume that the eigenvalues of $\mathbf{A}$ have a gap $\lambda_k > \lambda_{k+1}$ for some $1 \leq k < n$, and distinguish the dominant eigenvalues from the sub-dominant ones by partitioning

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_1 & \\ & \mathbf{\Lambda}_2 \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{pmatrix},$$

where $\mathbf{\Lambda}_1 = \mathsf{diag}\begin{pmatrix} \lambda_1 & \cdots & \lambda_k \end{pmatrix} \in \mathbb{R}^{k\times k}$ is nonsingular, and $\mathbf{U}_1 \in \mathbb{C}^{n\times k}$. The size of the gap is inversely proportional to

$$\gamma \equiv \lambda_{k+1}/\lambda_k = \|\mathbf{\Lambda}_2\|_2 \, \|\mathbf{\Lambda}_1^{-1}\|_2 < 1.$$

Given a number of power iterations $q \geq 1$, and a starting guess $\mathbf{\Omega} \in \mathbb{C}^{n\times \ell}$ with $k \leq \ell \leq n$ columns, we assume that the product has full column rank,

$$\mathsf{rank}(\mathbf{A}^q \, \mathbf{\Omega}) = \ell. \tag{1}$$

Extract an orthonormal basis for $\mathsf{range}(\mathbf{A}^q \, \mathbf{\Omega})$ with a thin QR decomposition $\mathbf{A}^q \, \mathbf{\Omega} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{C}^{n\times \ell}$ with $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_\ell$, and the matrix $\mathbf{R} \in \mathbb{C}^{\ell\times \ell}$ nonsingular.

To distinguish of the effect of the dominant subspace on the starting guess from that of the sub-dominant space, partition

$$\mathbf{U}^*\mathbf{\Omega} = \begin{pmatrix} \mathbf{U}_1^*\mathbf{\Omega} \\ \mathbf{U}_2^*\mathbf{\Omega} \end{pmatrix} = \begin{pmatrix} \mathbf{\Omega}_1 \\ \mathbf{\Omega}_2 \end{pmatrix},$$

where $\mathbf{\Omega}_1 \equiv \mathbf{U}_1^*\mathbf{\Omega} \in \mathbb{C}^{k\times \ell}$ and $\mathbf{\Omega}_2 \equiv \mathbf{U}_2^*\mathbf{\Omega} \in \mathbb{C}^{(n-k)\times \ell}$. We assume that $\mathbf{\Omega}$ has a sufficient contribution in the dominant subspace of $\mathbf{A}$,

$$\mathsf{rank}(\mathbf{\Omega}_1) = k. \tag{2}$$

2.2.2 Gaussian random matrices

We present absolute error bounds for the trace and logdet estimators when the random starting guess $\mathbf{\Omega}$ in Algorithm 1 is a Gaussian. The bounds come in two flavors: expectation, or mean (Theorem 1); and concentration around the mean (Theorem 2). We argue that for matrices with sufficiently dominant eigenvalues, the bounds are close.

The number of columns in $\mathbf{\Omega}$ is equal to

$$\ell = k + p,$$

where $0 \leq p < n - k$ is a user-specified oversampling parameter. We abbreviate

$$\mu \equiv \sqrt{n-k} + \sqrt{k+p}. \tag{3}$$

Theorem 1 (Expectation) *With the assumptions in Sect. 2.2.1, let* $\mathbf{T}$ *be computed by Algorithm 1 with a Gaussian starting guess* $\mathbf{\Omega}^{n\times(k+p)}$ *and furthermore, let* $p \geq 2$. *Then*

$$0 \leq \mathbb{E}\left[\,\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T})\right] \leq \left(1 + \gamma^{2q-1}\, C_{ge}\right)\, \mathsf{trace}(\mathbf{\Lambda}_2),$$

and

$$0 \leq \mathbb{E}\left[\,\log\det(\mathbf{I}+\mathbf{A}) - \log\det(\mathbf{I}+\mathbf{T})\right] \leq \log\det\left(\mathbf{I}+\mathbf{\Lambda}_2\right) + \log\det\left(\mathbf{I} + \gamma^{2q-1}\, C_{ge}\, \mathbf{\Lambda}_2\right),$$

where

$$C_{ge} \equiv \frac{e^2\,(k+p)}{(p+1)^2}\,\left(\frac{1}{2\pi\,(p+1)}\right)^{\frac{2}{p+1}}\,\left(\mu+\sqrt{2}\right)^2\left(\frac{p+1}{p-1}\right).$$

Proof See Sect. 4.1.1. □

Theorem 1 demonstrates that Algorithm 1 with a Gaussian starting guess produces a biased estimator. However, when $\mathbf{\Lambda}_2 = \mathbf{0}$, then Algorithm 1 produces an unbiased estimator.

In the special case when $\mathsf{rank}(\mathbf{A}) = k$, the assumption (2) guarantees exact computation, $\mathsf{trace}(\mathbf{T}) = \mathsf{trace}(\mathbf{A})$ and $\log\det(\mathbf{I}+\mathbf{T}) = \log\det(\mathbf{I}+\mathbf{A})$. Hence the bounds are zero, and hold with equality. If $\mathbf{A}$ has $n-k$ eigenvalues close to zero, i.e. $\mathbf{\Lambda}_2 \approx 0$, the upper bounds in Theorem 1 are small, implying that the estimators are accurate in the absolute sense. If $\mathbf{A}$ has k dominant eigenvalues that are very well separated from the remaining eigenvalues, i.e. $\gamma \ll 1$, then Theorem 1 implies that the absolute error in the estimators depends on the mass of the neglected eigenvalues $\mathbf{\Lambda}_2$. The above is true also for the following concentration bounds, which have the same form as the expectation bounds.

Theorem 2 (Concentration) *With the assumptions in Sect. 2.2.1, let* $\mathbf{T}$ *be computed by Algorithm 1 with a Gaussian starting guess guess* $\mathbf{\Omega}^{n\times(k+p)}$ *where* $p \geq 2$. *If* $0 < \delta < 1$, *then with probability at least* $1-\delta$

$$0 \leq \mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) \leq \left(1 + \gamma^{2q-1}\, C_{g}\right)\, \mathsf{trace}(\mathbf{\Lambda}_2),$$

and

$$0 \leq \log\det(\mathbf{I}+\mathbf{A}) - \log\det(\mathbf{I}+\mathbf{T}) \leq \log\det\left(\mathbf{I}+\mathbf{\Lambda}_2\right) + \log\det\left(\mathbf{I} + \gamma^{2q-1}\, C_{g}\, \mathbf{\Lambda}_2\right),$$

where

$$C_g \equiv \frac{e^2\,(k+p)}{(p+1)^2}\,\left(\frac{2}{\delta}\right)^{\frac{2}{p+1}}\left(\mu+\sqrt{2\log\frac{2}{\delta}}\right)^2.$$

Proof Substitute Lemma 5 into Theorems 6 and 8. □

The expectation and concentration bounds in Theorems 1 and 2 are the same save for the constants C_{ge} and C_g. For matrices $\mathbf{A}$ with sufficiently well separated eigenvalues, i.e. $\gamma \ll 1$, and sufficiently many power iterations q in Algorithm 1, the factor γ^{2q-1} subdues the effect of C_{ge} and C_g, so that Theorems 1 and 2 are effectively the same.

Nevertheless, we can still compare Theorems 1 and 2 by comparing their constants. To this end we take advantage of the natural logarithm, and consider two cases. For a high failure probability $\delta = 2/e$, the ratio is

$$\frac{C_g}{C_{ge}} = (2\,e\,\pi\,(p+1))^{\frac{2}{p+1}}\,\left(\frac{p-1}{p+1}\right) \to 1 \qquad \text{as} \quad p \to \infty.$$

Hence the concentration bound approaches the expectation bound as the oversampling increases. Note, though, that the rank assumptions for the bounds impose the limit $p < n - k$. However, for the practical value $p = 20$, the ratio $C_g/C_{ge} \approx 1.6$, so that the constants differ by a factor less than 2.

For a lower failure probability $\delta < 2/e$, we have $C_g > C_{ge}$. Hence the concentration bound in Theorem 2 has a higher constant.

2.2.3 Rademacher random matrices

We present absolute error bounds for the trace and logdet estimators when the random starting guess $\mathbf{\Omega}$ in Algorithm 1 is a Rademacher random matrix. In contrast to Gaussian starting guesses, the number of columns in the Rademacher guess reflects the dimension of the dominant subspace.

The error bounds contain a parameter $0 < \rho < 1$ that controls the magnitude of $\|\mathbf{\Omega}_1^{\dagger}\|_2$. The bound below has the same form as the error bound in Theorem 2 with Gaussian starting guesses; the only difference being the constant.

Theorem 3 *With the assumptions in Sect. 2.2.1, let* $0 < \delta < 1$ *be a given failure probability, and let* $\mathbf{T}$ *be computed by Algorithm 1 with a Rademacher starting guess* $\mathbf{\Omega} \in \mathbb{R}^{n\times\ell}$. *If the number of columns in* $\mathbf{\Omega}$ *satisfies*

$$\ell \geq 2\rho^{-2}\left(\sqrt{k}+\sqrt{8\log\frac{4n}{\delta}}\right)^2 \log\left(\frac{4k}{\delta}\right),$$

then with probability at least $1-\delta$

$$0 \leq \mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) \leq \left(1+\gamma^{2q-1}\,C_r\right)\,\mathsf{trace}(\mathbf{\Lambda}_2),$$

and

$$0 \leq \log\det(\mathbf{I}+\mathbf{A}) - \log\det(\mathbf{I}+\mathbf{T}) \leq \log\det\left(\mathbf{I}+\mathbf{\Lambda}_2\right) + \log\det\left(\mathbf{I}+\gamma^{2q-1}\, C_r\, \mathbf{\Lambda}_2\right),$$

where

$$C_r \equiv \frac{1}{(1-\rho)}\left[1 + 3\ell^{-1}\left(\sqrt{n-k} + \sqrt{8\log\frac{4\ell}{\delta}}\right)^2 \log\frac{4(n-k)}{\delta}\right].$$

Proof Substitute the bound for $\|\mathbf{\Omega}_2\|_2^2\|\mathbf{\Omega}_1^\dagger\|_2^2$ from Theorem 5 into Theorems 6 and 8. □

The interpretation of Theorem is the same as that of Theorems 1 and 2. In contrast to Gaussian starting guesses, whose number of columns depends on a fixed oversampling parameter p, the columns of the Rademacher guess increase with the dimension of the dominant subspace.

Theorem 1 shows that when Algorithm 1 is run with a Gaussian starting guess, the resulting estimators for the trace and determinant are biased. We are not able to provide a similar result for the expectation of the estimators for the Rademacher starting guess. However, we conjecture that the estimators for trace and determinant are biased even when the Rademacher starting guess is used in Algorithm 1.

2.3 Comparison with Monte Carlo estimators

The reliability of Monte Carlo estimators is judged by the variance of a single sample. This variance is $2(\|\mathbf{A}\|_F^2 - \sum_{j=1}^n \mathbf{A}_{jj}^2)$ for the Hutchinson estimator, and $2\|\mathbf{A}\|_F^2$ for the Gaussian estimator.

Avron and Toledo [6] were the first to determine the number of Monte Carlo samples N required to achieve a *relative* error ϵ with probability $1-\delta$, and defined an (ϵ, δ) estimator

$$\mathbb{P}\left[\left|\mathsf{trace}(\mathbf{A}) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{z}_j^*\mathbf{A}\mathbf{z}_j\right| \leq \epsilon\,\mathsf{trace}(\mathbf{A})\right] \geq 1-\delta.$$

An (ϵ, δ) estimator based on Gaussian vectors $\mathbf{z}_j$ requires $N \geq 20\,\epsilon^{-2}\log(2/\delta)$ samples. In contrast, the Hutchinson estimator, which is based on Rademacher vectors, requires $N \geq 6\,\epsilon^{-2}\log(2\mathsf{rank}(\mathbf{A})/\delta)$ samples.

Roosta-Khorasani and Ascher [35] improve the above bounds for Gaussian estimators to $N \geq 8\,\epsilon^{-2}\log(2/\delta)$; and for the Hutchinson estimator to $N \geq 6\,\epsilon^{-2}\log(2/\delta)$, thus removing the dependence on the rank. They also derived bounds on the number of samples required for an (ϵ, δ) estimator, using the Hutchinson, Gaussian and the unit vector random samples, which depend on specific properties of $\mathbf{A}$. All bounds

retain the ϵ^{-2} factor, though, which means that an accurate trace estimate requires many samples in practice. In fact, even for small matrices, while a few samples can estimate the trace up to one digit of accuracy, many samples are needed in practice to estimate the trace to machine precision.

To facilitate comparison between our estimators and the Monte Carlo estimators, we derive the number of iterations needed for an (ϵ, δ) estimator. Define the relative error

$$\Delta \equiv \mathsf{trace}(\mathbf{\Lambda}_2)/\mathsf{trace}(\mathbf{\Lambda}). \tag{4}$$

In practice, the relative error Δ is not known. Instead, it can be estimated as follows: the bounds $\mathsf{trace}(\mathbf{\Lambda}_2) \leq (n-k)\lambda_{k+1}$, $\mathsf{trace}(\mathbf{\Lambda}_1) \geq k\lambda_k$, can be combined to give us the upper bound

$$\Delta \leq \frac{(n-k)\gamma}{n\gamma + k(1-\gamma)}.$$

Assuming that $\Delta > 0$, abbreviate $\epsilon_\Delta \equiv \epsilon/\Delta$. If $\Delta = 0$, then we have achieved our desired relative error, i.e., the relative error is less than ϵ.

We present the following theorem that gives the asymptotic bound on the number of matrix–vector products needed for an (ϵ, δ) trace estimator.

Theorem 4 (Asymptotic bounds) *With the assumptions in Sect. 2.2.1, let ϵ be the desired accuracy and let $0 < \Delta < \epsilon \leq 1$. The number of matrix–vector products for an (ϵ, δ) estimator is asymptotically*

$$k\left(\log\frac{1}{\epsilon_\Delta - 1} + \log\frac{2}{\delta}\right), \tag{5}$$

for Gaussian starting guess, whereas for a Rademacher starting guess the number of matrix–vector products is asymptotically

$$(k + \log n)\log k\left(\log\frac{1}{\epsilon_\Delta - 1} + \log\left\lceil (n-k)\log\frac{4n}{\delta}\right\rceil\right). \tag{6}$$

Proof The number of matrix–vector products in Algorithm 1 is $\ell(q+1)$. Recall that the number of samples required for a Gaussian starting guess are $\ell \sim k$; whereas for a Rademacher starting guess $\ell \sim (k + \log n)\log k$. With probability of failure at most δ, for an (ϵ, δ) estimator

$$\frac{\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T})}{\mathsf{trace}(\mathbf{A})} \leq (1 + \gamma^{2q-1}C)\Delta.$$

Here C can either take values C_g for standard Gaussian matrices or C_r for standard Rademacher matrices. Equating the right hand side to ϵ gives us $(1+\gamma^{2q-1}C)\Delta = \epsilon$. Assuming $\epsilon > \Delta$, we can solve for q to obtain

$$q = \left\lceil \frac{1}{2}\left(1 + \log\left(\frac{C\Delta}{\epsilon - \Delta}\right) \Big/ \log\gamma^{-1}\right)\right\rceil.$$

Asymptotically, $\log C_g$ behaves like $\log 2/\delta$ and $\log C_r$ behaves like $\log\left[(n-k)\log 4n/\delta\right]$. This proves the desired result. □

Theorem 4 demonstrates both estimators are computationally efficient compared to the Monte Carlo estimators if Δ is sufficiently small.

3 Structural analysis

We defer the probabilistic part of the analysis as long as possible, and start with deterministic error bounds for $\mathsf{trace}(\mathbf{T})$ (Sect. 3.1) and $\log\det(\mathbf{T})$ (Sect. 3.2), where $\mathbf{T}$ is the restriction of $\mathbf{A}$ computed by Algorithm 1. These deterministic bounds are called "structural" because they hold for *all* matrices $\mathbf{\Omega}$ that satisfy the rank conditions (1) and (2).

3.1 Trace estimator

We derive the following absolute error bounds for Hermitian positive semi-definite matrices $\mathbf{A}$ and matrices $\mathbf{T}$ computed by Algorithm 1.

Theorem 5 *With the assumptions in Sect. 2.2.1, let* $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ *be computed by Algorithm 1. Then*

$$0 \leq \mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) \leq (1+\theta_1)\ \mathsf{trace}(\mathbf{\Lambda}_2)$$

where $\theta_1 \equiv \min\{\gamma^{q-1}\,\|\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\|_2,\ \gamma^{2q-1}\,\|\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\|_2^2\}$.

Proof The lower bound is derived in Lemma 1, and the upper bounds in Theorem 6. □

Theorem 5 implies that $\mathsf{trace}(\mathbf{T})$ has a small absolute error if Algorithm 1 applies a sufficient number q of power iterations. More specifically, only a few iterations are required if the eigenvalue gap is large and $\gamma \ll 1$. The term θ_1 quantifies the contribution of the starting guess $\mathbf{\Omega}$ in the dominant subspace $\mathbf{U}_1$. The minimum in θ_1 is attained by $\gamma^{q-1}\,\|\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\|_2$ when, relative to the eigenvalue gap and the iteration count q, the starting guess $\mathbf{\Omega}$ has only a "weak" contribution in the dominant subspace.

We start with the derivation of the lower bound, which relies on the variational inequalities for Hermitian matrices, and shows that the trace of a restriction can never exceed that of the original matrix.

Lemma 1 *With the assumptions in Sect. 2.2.1, let* $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ *be computed by Algorithm 1. Then*

$$\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) \geq \mathsf{trace}(\mathbf{\Lambda}_2) - (\lambda_{k+1} + \cdots + \lambda_{\ell}) \geq 0.$$

Proof Choose $\mathbf{Q}_\perp \in \mathbb{C}^{n\times(n-\ell)}$ so that $\hat{\mathbf{Q}} \equiv \begin{pmatrix}\mathbf{Q} & \mathbf{Q}_\perp\end{pmatrix} \in \mathbb{C}^{n\times n}$ is unitary, and partition

$$\hat{\mathbf{Q}}^*\mathbf{A}\hat{\mathbf{Q}} = \begin{pmatrix}\mathbf{T} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^* & \mathbf{A}_{22}\end{pmatrix},$$

where $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ is the important submatrix. The matrices $\hat{\mathbf{Q}}^*\mathbf{A}\hat{\mathbf{Q}}$ and $\mathbf{A}$ have the same eigenvalues $0 \le \lambda_n \le \cdots \le \lambda_1$. With $\lambda_\ell(\mathbf{T}) \le \cdots \le \lambda_1(\mathbf{T})$ being the eigenvalues of $\mathbf{T}$, the Cauchy-interlace theorem [33, Section 10-1] implies

$$0 \le \lambda_{(n-\ell)+j} \le \lambda_j(\mathbf{T}) \le \lambda_j, \qquad 1 \le j \le \ell.$$

Since $\lambda_j \ge 0$, this implies (for $\ell = k$ we interpret $\sum_{j=k+1}^{\ell} \lambda_j = 0$)

$$\begin{aligned}\mathsf{trace}(\mathbf{T}) &\le \sum_{j=1}^{\ell} \lambda_j = \mathsf{trace}(\mathbf{\Lambda}_1) + \sum_{j=k+1}^{\ell} \lambda_j \\ &= \mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{\Lambda}_2) + \sum_{j=k+1}^{\ell} \lambda_j \le \mathsf{trace}(\mathbf{A}),\end{aligned}$$

where the last inequality follows from $\sum_{j=k+1}^{\ell} \lambda_j \le \sum_{j=k+1}^{n} \lambda_j = \mathsf{trace}(\mathbf{\Lambda}_2)$. □

Next we derive the two upper bounds. The first one, (7), is preferable when, relative to the eigenvalue gap and the iteration count q, the starting guess $\mathbf{\Omega}$ has only a "weak" contribution in the dominant subspace.

Theorem 6 *With the assumptions in Sect. 2.2.1, let* $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ *be computed by Algorithm 1. Then*

$$\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) \le \left(1 + \gamma^{q-1}\,\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2\right)\,\mathsf{trace}(\mathbf{\Lambda}_2). \tag{7}$$

If $0 < \|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2 \le \gamma^{-q}$*, then the following bound is tighter,*

$$\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) \le \left(1 + \gamma^{2q-1}\,\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\right)\,\mathsf{trace}(\mathbf{\Lambda}_2). \tag{8}$$

Proof The proof proceeds in six steps. The first five steps are the same for both bounds.

1. Shrinking the space from ℓ to k dimensions If $\mathbf{W} \in \mathbb{C}^{\ell\times k}$ is any matrix with orthonormal columns, then Lemma 1 implies

$$\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{Q}^*\mathbf{A}\mathbf{Q}) \le \mathcal{U} \equiv \mathsf{trace}(\mathbf{A}) - \mathsf{trace}\left((\mathbf{Q}\mathbf{W})^*\,\mathbf{A}\,(\mathbf{Q}\mathbf{W})\right).$$

The upper bound $\mathcal{U}$ replaces the matrix $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$ of order ℓ by the matrix $(\mathbf{Q}\mathbf{W})^*\,\mathbf{A}\,(\mathbf{Q}\mathbf{W})$ of order $k \le \ell$. The eigendecomposition of $\mathbf{A}$ yields

$$\mathsf{trace}\left((\mathbf{Q}\mathbf{W})^*\,\mathbf{A}\,(\mathbf{Q}\mathbf{W})\right) = t_1 + t_2,$$

where dominant eigenvalues are distinguished from sub-dominant ones by

$$t_1 \equiv \mathsf{trace}\left((\mathbf{U}_1^*\mathbf{Q}\mathbf{W})^*\,\mathbf{\Lambda}_1\,(\mathbf{U}_1^*\mathbf{Q}\mathbf{W})\right), \quad t_2 \equiv \mathsf{trace}\left((\mathbf{U}_2^*\mathbf{Q}\mathbf{W})^*\,\mathbf{\Lambda}_2\,(\mathbf{U}_2^*\mathbf{Q}\mathbf{W}\right).$$

Note that t_1 and t_2 are real. Now we can write the upper bound as

$$\mathcal{U} = \mathsf{trace}(\mathbf{A}) - t_1 - t_2. \tag{9}$$

2. Exploiting the structure of $\mathbf{Q}$ Assumption (1) implies that $\mathbf{R}$ is nonsingular, hence we can solve for $\mathbf{Q}$ in $\mathbf{A}^q\,\mathbf{\Omega} = \mathbf{Q}\mathbf{R}$, to obtain

$$\mathbf{Q} = (\mathbf{A}^q\,\mathbf{\Omega})\,\mathbf{R}^{-1} = \mathbf{U}\mathbf{\Lambda}^q\,\mathbf{U}^*\,\mathbf{\Omega}\mathbf{R}^{-1} = \mathbf{U}\begin{pmatrix}\mathbf{\Lambda}_1^q\,\mathbf{\Omega}_1\\ \mathbf{\Lambda}_2^q\,\mathbf{\Omega}_2\end{pmatrix}\mathbf{R}^{-1}. \tag{10}$$

3. Choosing $\mathbf{W}$ Assumption (2) implies that the $k \times \ell$ matrix $\mathbf{\Omega}_1$ has full row rank, and a right inverse $\mathbf{\Omega}_1^\dagger = \mathbf{\Omega}_1^*\,(\mathbf{\Omega}_1\mathbf{\Omega}_1^*)^{-1}$. Our choice for $\mathbf{W}$ is

$$\mathbf{W} \equiv \mathbf{R}\,\mathbf{\Omega}_1^\dagger\mathbf{\Lambda}_1^{-q}\,(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2} \qquad \text{where} \qquad \mathbf{F} \equiv \mathbf{\Lambda}_2^q\,\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\,\mathbf{\Lambda}_1^{-q},$$

so that we can express (10) as

$$\mathbf{Q}\mathbf{W} = \mathbf{U}\begin{pmatrix}\mathbf{\Lambda}_1^q\,\mathbf{\Omega}_1\\ \mathbf{\Lambda}_2^q\,\mathbf{\Omega}_2\end{pmatrix}\mathbf{R}^{-1}\,\mathbf{W} = \mathbf{U}\begin{pmatrix}\mathbf{I}_k\\ \mathbf{F}\end{pmatrix}(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}. \tag{11}$$

The rightmost expression shows that $\mathbf{Q}\mathbf{W}$ has orthonormal columns. To see that $\mathbf{W}$ itself also has orthonormal columns, show that $\mathbf{W}^*\mathbf{W} = \mathbf{I}_k$ with the help of

$$\mathbf{R}^*\mathbf{R} = (\mathbf{Q}\mathbf{R})^*(\mathbf{Q}\mathbf{R}) = (\mathbf{\Lambda}_1^q\,\mathbf{\Omega}_1)^*(\mathbf{\Lambda}_1^q\,\mathbf{\Omega}_1) + (\mathbf{\Lambda}_2^q\,\mathbf{\Omega}_2)^*(\mathbf{\Lambda}_2^q\,\mathbf{\Omega}_2).$$

4. Determining $\mathcal{U}$ *in* (9) From $\mathbf{U}_1^*\mathbf{Q}\mathbf{W} = (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1/2}$ in (11) follows

$$t_1 = \mathsf{trace}\left((\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1/2}\,\mathbf{\Lambda}_1\,(\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1/2}\right) = \mathsf{trace}\left(\mathbf{\Lambda}_1\,(\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1}\right).$$

From (11) also follows $\mathbf{U}_2^*\mathbf{Q}\mathbf{W} = \mathbf{F}\,(\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1/2}$, so that

$$t_2 = \mathsf{trace}\left((\mathbf{I}+\mathbf{F}^*\mathbf{F})^{-1/2}\,\mathbf{F}^*\,\mathbf{\Lambda}_2\,\mathbf{F}\,(\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1/2}\right) = \mathsf{trace}\left(\mathbf{\Lambda}_2\,\mathbf{F}\,(\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1}\,\mathbf{F}^*\right).$$

Distinguish dominant from sub-dominant eigenvalues in $\mathcal{U}$ via $\mathcal{U} = \mathcal{U}_1 + \mathcal{U}_2$, where

$$\mathcal{U}_1 \equiv \mathsf{trace}(\mathbf{\Lambda}_1) - t_1, \qquad \mathcal{U}_2 \equiv \mathsf{trace}(\mathbf{\Lambda}_2) - t_2.$$

Since t_1 and t_2 are real, so are $\mathcal{U}_1$ and $\mathcal{U}_2$. With the identity

$$\mathbf{\Lambda}_1\,\left(\mathbf{I} - (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1}\right) = \mathbf{\Lambda}_1\,\mathbf{F}^*\mathbf{F}\,(\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1},$$

and remembering that $\mathbf{F} = \mathbf{\Lambda}_2^q \, \mathbf{Z} \, \mathbf{\Lambda}_1^{-q}$ with $\mathbf{Z} \equiv \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger$, we obtain $\mathcal{U} = \mathcal{U}_1 + \mathcal{U}_2$ with

$$\begin{aligned} \mathcal{U}_1 &= \mathsf{trace}\left(\mathbf{\Lambda}_1^{1-q} \, \mathbf{Z}^* \mathbf{\Lambda}_2^q \, \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \right) \\ \mathcal{U}_2 &= \mathsf{trace}\left(\mathbf{\Lambda}_2 \, (\mathbf{I} - \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \mathbf{F}^*) \right). \end{aligned}$$

5. Bounding $\mathcal{U}$ Since $\mathbf{\Lambda}_2$ and $\mathbf{I} - \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \mathbf{F}^*$ both have dimension $(n-k) \times (n-k)$, the *von Neumann trace theorem* [23, Theorem 7.4.11] can be applied,

$$\mathcal{U}_2 \leq \sum_j \sigma_j(\mathbf{\Lambda}_2) \, \sigma_j(\mathbf{I} - \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \mathbf{F}^*) \leq \sum_j \sigma_j(\mathbf{\Lambda}_2) = \mathsf{trace}(\mathbf{\Lambda}_2).$$

The last equality is true because the singular values of a Hermitian positive semi-definite matrix are also the eigenvalues. Analogously, $\mathbf{\Lambda}_2^q \, \mathbf{Z} \, \mathbf{\Lambda}_1^{1-q}$ and $\mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1}$ both have dimension $(n-k) \times k$, so that

$$\begin{aligned} \mathcal{U}_1 &\leq \sum_j \sigma_j \left(\mathbf{\Lambda}_2^q \, \mathbf{Z} \, \mathbf{\Lambda}_1^{1-q} \right) \sigma_j \left(\mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \right) \\ &\leq \| \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \|_2 \sum_j \sigma_j \left(\mathbf{\Lambda}_2^q \, \mathbf{Z} \, \mathbf{\Lambda}_1^{1-q} \right). \end{aligned}$$

Repeated applications of the singular value inequalities [22, Theorem 3.314] for the second factor yield

$$\begin{aligned} \sum_j \sigma_j \left(\mathbf{\Lambda}_2^q \, \mathbf{Z} \, \mathbf{\Lambda}_1^{1-q} \right) &\leq \| \mathbf{Z} \|_2 \, \| \mathbf{\Lambda}_1 \|_2^{1-q} \sum_j \sigma_j(\mathbf{\Lambda}_2^q) \\ &\leq \| \mathbf{Z} \|_2 \, \| \mathbf{\Lambda}_1 \|_2^{1-q} \, \| \mathbf{\Lambda}_2 \|_2^{q-1} \sum_j \sigma_j(\mathbf{\Lambda}_2) = \gamma^{q-1} \, \| \mathbf{Z} \|_2 \, \mathsf{trace}(\mathbf{\Lambda}_2). \end{aligned}$$

Substituting this into the bound for $\mathcal{U}_1$ gives

$$\mathcal{U}_1 \leq \| \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \|_2 \, \gamma^{q-1} \, \| \mathbf{Z} \|_2 \, \mathsf{trace}(\mathbf{\Lambda}_2).$$

6. Bounding $\| \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \|_2$ For (7) we bound $\| \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \|_2 \leq 1$, which yields $\mathcal{U}_1 \leq \gamma^{q-1} \, \| \mathbf{Z} \|_2 \, \mathsf{trace}(\mathbf{\Lambda}_2)$. For (8) we use

$$\| \mathbf{F} \, (\mathbf{I} + \mathbf{F}^*\mathbf{F})^{-1} \|_2 \leq \| \mathbf{F} \|_2 \leq \| \mathbf{\Lambda}_2 \|_2^q \, \| \mathbf{Z} \|_2 \, \| \mathbf{\Lambda}_1 \|_2^{-q} = \gamma^q \, \| \mathbf{Z} \|_2,$$

which yields $\mathcal{U}_1 \leq \gamma^{2q-1} \| \mathbf{Z} \|_2^2 \, \mathsf{trace}(\mathbf{\Lambda}_2)$.

Comparing the two preceding bounds for $\mathcal{U}_1$ shows that (8) is tighter than (7) if $\gamma^{2q-1} \, \| \mathbf{Z} \|_2^2 \leq \gamma^{q-1} \, \| \mathbf{Z} \|_2$, that is $\| \mathbf{Z} \|_2 \leq \gamma^{-q}$. □

Remark 1 The two special cases below illustrate that, even in a best-case scenario, the accuracy of $\mathsf{trace}(\mathbf{T})$ is limited by $\mathsf{trace}(\mathbf{\Lambda}_2)$.

– If $\ell = k$ and $\mathbf{\Omega} = \mathbf{U}_1$ then

$$\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) = \mathsf{trace}(\mathbf{\Lambda}_2).$$

This follows from Lemma 1, and from both bounds in Theorem 6 with $\mathbf{\Omega}_1 = \mathbf{I}_k$ and $\mathbf{\Omega}_2 = \mathbf{0}$.

– If $\ell > k$ and $\mathbf{\Omega}$ consists of the columns of $\mathbf{U}$ associated with the dominant eigenvalues $\lambda_1, \ldots, \lambda_\ell$ of $\mathbf{A}$, then

$$\mathsf{trace}(\mathbf{\Lambda}_2) - (\lambda_{k+1} + \cdots + \lambda_\ell) \leq \mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T}) \leq \mathsf{trace}(\mathbf{\Lambda}_2).$$

This follows from Lemma 1, and from both bounds in Theorem 6 with $\mathbf{\Omega}_1 = \begin{pmatrix}\mathbf{I}_k & \mathbf{0}_{k\times(k-\ell)}\end{pmatrix}$ and $\mathbf{\Omega}_2 = \begin{pmatrix}\mathbf{0}_{(n-k)\times k} & *\end{pmatrix}$.
Theorem 6 cannot be tight for $\ell > k$ because step 3 of the proof deliberately transitions to a matrix with k columns. Hence the eigenvalues $\lambda_{k+1}, \ldots, \lambda_\ell$ do not appear in the bounds of Theorem 6.

3.2 Log determinant estimator

Subject to the Assumptions in Sect. 2.2.1, we derive the following absolute error bounds for Hermitian positive semi-definite matrices $\mathbf{A}$ and matrices $\mathbf{T}$ computed by Algorithm 1.

Theorem 7 *With the assumptions in Sect. 2.2.1, let* $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ *be computed by Algorithm 1. Then*

$$0 \leq \log\det(\mathbf{I}_n + \mathbf{A}) - \log\det(\mathbf{I}_\ell + \mathbf{T}) \leq \log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) + \log\det(\mathbf{I}_{n-k} + \theta_2\,\mathbf{\Lambda}_2)$$

where $\theta_2 \equiv \gamma^{2q-1}\,\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\,\min\{1, \frac{1}{\lambda_k}\}$.

Proof The lower bound is derived in Lemma 2, and the upper bounds in Theorem 8. □

Theorem 7 implies that $\log\det(\mathbf{I}_\ell + \mathbf{T})$ has a small absolute error if Algorithm 1 applies a sufficient number q of power iterations. As in Theorem 6, only a few iterations are required if the eigenvalue gap is large and $\gamma \ll 1$. The term θ_2 quantifies the contribution of the starting guess $\mathbf{\Omega}$ in the dominant subspace $\mathbf{U}_1$. The two alternatives differ by a factor of only λ_k^{-1}. The second one is smaller if $\lambda_k > 1$.

Theorem 9 extends Theorem 7 to $\log\det(\mathbf{A})$ for positive definite $\mathbf{A}$.

As before, we start with the derivation of the lower bound, which is the counter part of Lemma 1 and shows that the log determinant of the restriction can never exceed that of the original matrix.

Lemma 2 *With the assumptions in Sect. 2.2.1, let* $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ *be computed by Algorithm 1. Then*

$$\log\det(\mathbf{I}_n + \mathbf{A}) - \log\det(\mathbf{I}_\ell + \mathbf{T}) \geq \log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) - \log\prod_{j=k+1}^{\ell}(1 + \lambda_j) \geq 0.$$

Proof Choose the unitary matrix $\hat{\mathbf{Q}}$ as in the proof of Lemma 1,

$$\hat{\mathbf{Q}}^*(\mathbf{I}_n + \mathbf{A})\hat{\mathbf{Q}} = \begin{pmatrix} \mathbf{I}_\ell + \mathbf{T} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^* & \mathbf{I}_{n-\ell} + \mathbf{A}_{22} \end{pmatrix},$$

and proceed likewise with the Cauchy-interlace theorem [33, Section 10-1] to conclude

$$\begin{aligned} \det(\mathbf{I}_\ell + \mathbf{T}) &\leq \det(\mathbf{I}_k + \mathbf{\Lambda}_1) \prod_{j=k+1}^{\ell} (1 + \lambda_j) \\ &= \det(\mathbf{I}_n + \mathbf{A}) \prod_{j=k+1}^{\ell} (1 + \lambda_j) / \det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) \leq \det(\mathbf{I}_n + \mathbf{A}), \end{aligned}$$

where for $\ell = k$ we interpret $\prod_{j=k+1}^{\ell} (1 + \lambda_j) = 1$. The monotonicity of the logarithm implies

$$\begin{aligned} \log\det(\mathbf{I}_\ell + \mathbf{T}) &\leq \log\det(\mathbf{I}_n + \mathbf{A}) - \log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) + \log \prod_{j=k+1}^{\ell} (1 + \lambda_j) \\ &\leq \log\det(\mathbf{I}_n + \mathbf{A}). \end{aligned}$$

□

The following auxiliary result, often called *Sylvester's determinant identity*, is required for the derivation of the upper bound.

Lemma 3 (Corollary 2.1 in [31]) *If* $\mathbf{B} \in \mathbb{C}^{m\times n}$ *and* $\mathbf{C} \in \mathbb{C}^{n\times m}$ *then*

$$\det(\mathbf{I}_m \pm \mathbf{BC}) = \det(\mathbf{I}_n \pm \mathbf{CB}).$$

Next we derive two upper bounds. The second one, (13), is preferable when $\lambda_k > 1$ because it reduces the extraneous term.

Theorem 8 *With the assumptions in Sect. 2.2.1, let* $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ *be computed by Algorithm 1. Then*

$$\begin{aligned} &\log\det(\mathbf{I}_n + \mathbf{A}) - \log\det(\mathbf{I}_\ell + \mathbf{T}) \leq \\ &\qquad \log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) + \log\det\left(\mathbf{I}_{n-k} + \gamma^{2q-1}\, \|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\, \mathbf{\Lambda}_2\right). \end{aligned} \tag{12}$$

If $\lambda_k > 1$ *then the following bound is tighter*

$$\begin{aligned} &\log\det(\mathbf{I}_n + \mathbf{A}) - \log\det(\mathbf{I}_\ell + \mathbf{T}) \leq \\ &\qquad \log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) + \log\det\left(\mathbf{I}_{n-k} + \tfrac{\gamma^{2q-1}}{\lambda_k}\, \|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\, \mathbf{\Lambda}_2\right). \end{aligned} \tag{13}$$

Proof The structure of the proof is analogous to that of Theorem 6, and the first three steps are the same for (12) and (13). Abbreviate $f(\cdot) \equiv \log\det(\cdot)$.

1. Shrinking the space Lemma 2 implies

$$f(\mathbf{I}_n + \mathbf{A}) - f(\mathbf{I}_\ell + \mathbf{Q}^*\mathbf{A}\mathbf{Q}) \leq f(\mathbf{I}_n + \mathbf{A}) - f(\mathbf{I}_k + \mathbf{H}),$$

where

$$\mathbf{H} \equiv \mathbf{W}^*\mathbf{T}\mathbf{W} = \mathbf{W}^*\,\mathbf{Q}^*\mathbf{A}\mathbf{Q}\,\mathbf{W} = (\mathbf{U}^*\mathbf{Q}\mathbf{W})^* \begin{pmatrix} \mathbf{\Lambda}_1 & \\ & \mathbf{\Lambda}_2 \end{pmatrix} (\mathbf{U}^*\mathbf{Q}\mathbf{W}).$$

The upper bound for the absolute error equals

$$\begin{aligned} f(\mathbf{I}_n + \mathbf{A}) - f(\mathbf{I}_k + \mathbf{H}) &= f(\mathbf{I}_k + \mathbf{\Lambda}_1) + f(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) - f(\mathbf{I}_k + \mathbf{H}) \\ &= f(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) + \mathcal{E}. \end{aligned}$$

Since nothing can be done about $f(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2)$, it suffices to bound

$$\mathcal{E} \equiv f(\mathbf{I}_k + \mathbf{\Lambda}_1) - f(\mathbf{I}_k + \mathbf{H}). \tag{14}$$

2. Exploiting the structure of $\mathbf{Q}$ *and choosing* $\mathbf{W}$ To simplify the expression for $\mathbf{H}$, we exploit the structure of $\mathbf{Q}$ and choose $\mathbf{W}$ as in the proof of Theorem 6. From (11) follows

$$\mathbf{U}^*\mathbf{Q}\mathbf{W} = \begin{pmatrix} \mathbf{I}_k \\ \mathbf{F} \end{pmatrix} (\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}, \qquad \text{where} \quad \mathbf{F} \equiv \mathbf{\Lambda}_2^q\,\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\,\mathbf{\Lambda}_1^{-q}.$$

Substituting this into the eigendecomposition of $\mathbf{H}$ gives

$$\mathbf{H} = (\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}\,\left(\mathbf{\Lambda}_1 + \mathbf{F}^*\mathbf{\Lambda}_2\mathbf{F}\right)\,(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}.$$

3. Lower bound for $f(\mathbf{I}_k + \mathbf{H})$ We use the Loewner partial order [23, Definition 7.7.1] to represent positive semi-definiteness, $\mathbf{F}^*\mathbf{\Lambda}_2\mathbf{F} \succeq \mathbf{0}$, which implies

$$\mathbf{H} \;\succeq\; \mathbf{H}_1 \equiv (\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}\,\mathbf{\Lambda}_1\,(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}. \tag{15}$$

The properties of the Loewner partial order [23, Corollary 7.7.4] imply

$$f(\mathbf{I}_k + \mathbf{H}) \geq f(\mathbf{I}_k + \mathbf{H}_1).$$

We first derive (12) and then (13).

Derivation of (12) *in Steps 4a–6a*

4a. Sylvester's determinant identity Applying Lemma 3 to $\mathbf{H}_1$ in (15) gives

$$f(\mathbf{I}_k + \mathbf{H}_1) = f(\mathbf{I}_k + \mathbf{H}_2) \qquad \text{where} \quad \mathbf{H}_2 \equiv \mathbf{\Lambda}_1^{1/2}(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1}\mathbf{\Lambda}_1^{1/2}.$$

5a. Upper bound for $\mathcal{E}$ in (14) Steps 3 and 4a imply

$$\mathcal{E} \leq f(\mathbf{I}_k + \mathbf{\Lambda}_1) - f(\mathbf{I}_k + \mathbf{H}_2) = f(\mathbf{H}_3),$$

where

$$\begin{aligned}\mathbf{H}_3 &\equiv (\mathbf{I}_k + \mathbf{H}_2)^{-1/2}(\mathbf{I}_k + \mathbf{\Lambda}_1)(\mathbf{I}_k + \mathbf{H}_2)^{-1/2} \\ &= (\mathbf{I}_k + \mathbf{H}_2)^{-1} + (\mathbf{I}_k + \mathbf{H}_2)^{-1/2}\mathbf{\Lambda}_1(\mathbf{I}_k + \mathbf{H}_2)^{-1/2}.\end{aligned}$$

Expanding the first summand into $(\mathbf{I}_k+\mathbf{H}_2)^{-1} = \mathbf{I}-(\mathbf{I}_k+\mathbf{H}_2)^{-1/2}\mathbf{H}_2(\mathbf{I}_k+\mathbf{H}_2)^{-1/2}$ gives

$$\mathbf{H}_3 = \mathbf{I}_k + (\mathbf{I}_k + \mathbf{H}_2)^{-1/2}(\mathbf{\Lambda}_1 - \mathbf{H}_2)(\mathbf{I}_k + \mathbf{H}_2)^{-1/2}.$$

Since $\mathbf{I}_k - (\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1} \preceq \mathbf{F}^*\mathbf{F}$, the center term can be bounded by

$$\mathbf{\Lambda}_1 - \mathbf{H}_2 = \mathbf{\Lambda}_1^{1/2}\left(\mathbf{I}_k - (\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1}\right)\mathbf{\Lambda}_1^{1/2} \preceq \mathbf{K} \equiv \mathbf{\Lambda}_1^{1/2}\mathbf{F}^*\mathbf{F}\mathbf{\Lambda}_1^{1/2}.$$

Because the singular values of $(\mathbf{I}_k+\mathbf{H}_2)^{-1/2}$ are less than 1, Ostrowski's Theorem [23, Theorem 4.5.9] implies

$$\mathbf{H}_3 \preceq \mathbf{I}_k + (\mathbf{I}_k + \mathbf{H}_2)^{-1/2}\mathbf{K}(\mathbf{I}_k + \mathbf{H}_2)^{-1/2} \preceq \mathbf{I}_k + \mathbf{K}.$$

Thus $\mathcal{E} \leq f(\mathbf{H}_3) \leq f(\mathbf{I}_k + \mathbf{K})$.
6a. Bounding $f(\mathbf{I}_k + \mathbf{K})$ Abbreviate $\mathbf{G}_1 \equiv \mathbf{\Lambda}_2^{q-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\mathbf{\Lambda}_1^{-q+1/2}$ and expand $\mathbf{K}$,

$$\mathbf{I}_k + \mathbf{K} = \mathbf{I}_k + \mathbf{\Lambda}_1^{1/2}\mathbf{F}^*\mathbf{F}\mathbf{\Lambda}_1^{1/2} = \mathbf{I}_k + \mathbf{G}_1^*\mathbf{\Lambda}_2\mathbf{G}_1.$$

Applying Lemma 3, gives

$$\det(\mathbf{I}_k + \mathbf{G}_1^*\mathbf{\Lambda}_2\mathbf{G}_1) = \det(\mathbf{I}_{n-k} + \mathbf{G}_1\mathbf{G}_1^*\mathbf{\Lambda}_2).$$

The fact that the absolute value of a determinant is the product of the singular values, and the inequalities for sums of singular values [22, Theorem 3.3.16] implies

$$\begin{aligned}\det(\mathbf{I}_{n-k} + \mathbf{G}_1\mathbf{G}_1^*\mathbf{\Lambda}_2) &\leq \prod_{j=1}^{n-k}\sigma_j(\mathbf{I}_{n-k} + \mathbf{G}_1\mathbf{G}_1^*\mathbf{\Lambda}_2) \\ &\leq \prod_{j=1}^{n-k}(1 + \sigma_j(\mathbf{G}_1\mathbf{G}_1^*\mathbf{\Lambda}_2)).\end{aligned}$$

Observe that $\|\mathbf{G}_1\|_2 \leq \gamma^{q-1/2}\,\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2$ and apply the singular value product inequalities [22, Theorem 3.3.16(d)]

$$\sigma_j(\mathbf{G}_1\mathbf{G}_1^*\boldsymbol{\Lambda}_2) \leq \sigma_1(\mathbf{G}_1\mathbf{G}_1^*)\sigma_j(\boldsymbol{\Lambda}_2) \leq \gamma^{2q-1}\,\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2^2\,\lambda_{k+j}, \qquad 1 \leq j \leq n-k,$$

and therefore

$$\det(\mathbf{I}_{n-k} + \mathbf{G}_1\mathbf{G}_1^*\boldsymbol{\Lambda}_2) \leq \det\left(\mathbf{I}_{n-k} + \gamma^{2q-1}\,\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2^2\,\boldsymbol{\Lambda}_2\right).$$

Thus

$$\mathcal{E} \leq f(\mathbf{I}_k + \mathbf{K}) \leq f\left(\mathbf{I}_{n-k} + \gamma^{2q-1}\,\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2^2\,\boldsymbol{\Lambda}_2\right).$$

Derivation of (13) *in Steps 4b–5b*

4b. Upper bound for $\mathcal{E}$ *in* (14) For the matrix $\mathbf{H}_1$ in (15) write

$$\begin{aligned}\mathbf{I}_k + \mathbf{H}_1 &= (\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}\,\left((\mathbf{I}_k + \mathbf{F}^*\mathbf{F}) + \boldsymbol{\Lambda}_1\right)\,(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}\\ &\succeq \mathbf{H}_4 \equiv (\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}\,(\mathbf{I}_k + \boldsymbol{\Lambda}_1)\,(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})^{-1/2}.\end{aligned}$$

Thus $f(\mathbf{H}_4) = f(\mathbf{I}_k + \boldsymbol{\Lambda}_1) - f(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})$. The properties of the Loewner partial order [23, Corollary 7.7.4] imply

$$\mathcal{E} \leq f(\mathbf{I}_k + \boldsymbol{\Lambda}_1) - f(\mathbf{H}_4) \leq f(\mathbf{I}_k + \mathbf{F}^*\mathbf{F}).$$

5b. Bounding $f(\mathbf{I}_k + \mathbf{F}^*\mathbf{F})$ Write

$$\mathbf{I}_k + \mathbf{F}^*\mathbf{F} = \mathbf{I}_k + \mathbf{G}_2^*\boldsymbol{\Lambda}_2\mathbf{G}_2 \qquad \text{where} \quad \mathbf{G}_2 \equiv \boldsymbol{\Lambda}_2^{q-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\boldsymbol{\Lambda}_1^{-q}.$$

Observe that, $\mathbf{G}_2 = \mathbf{G}_1\boldsymbol{\Lambda}_1^{-1/2}$ and therefore, $\|\mathbf{G}_2\|_2 \leq \frac{\gamma^{q-1/2}}{\sqrt{\lambda_k}}\,\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2$. The rest of the proof follows the same steps as Step 6a and will not be repeated here. □

The following discussion mirrors that in Remark 1.

Remark 2 The two special cases below illustrate that, even in a best-case scenario, the accuracy of $\log\det(\mathbf{I}_\ell + \mathbf{T})$ is limited by $\log\det(\mathbf{I}_{n-k} + \boldsymbol{\Lambda}_2)$.

– If $\ell = k$ and $\boldsymbol{\Omega} = \mathbf{U}_1$ then

$$\log\det(\mathbf{I}_n + \mathbf{A}) - \log\det(\mathbf{I}_\ell + \mathbf{T}) = \log\det(\mathbf{I}_{n-k} + \boldsymbol{\Lambda}_2).$$

This follows from Lemma 2, and from both bounds in Theorem 8 with $\boldsymbol{\Omega}_1 = \mathbf{I}_k$ and $\boldsymbol{\Omega}_2 = \mathbf{0}$.

– If $\ell > k$ and $\mathbf{\Omega}$ consists of the columns of $\mathbf{U}$ associated with the dominant eigenvalues $\lambda_1, \ldots, \lambda_\ell$ of $\mathbf{A}$, then

$$\log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2) - \log \prod_{j=k+1}^{\ell} (1+\lambda_j) \leq \log\det(\mathbf{I}_n + \mathbf{A}) - \log\det(\mathbf{I}_\ell + \mathbf{T}) \leq \log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2).$$

This follows from Lemma 2, and from both bounds in Theorem 8 with $\mathbf{\Omega}_1 = \begin{pmatrix}\mathbf{I}_k & \mathbf{0}_{k\times(k-\ell)}\end{pmatrix}$ and $\mathbf{\Omega}_2 = \begin{pmatrix}\mathbf{0}_{(n-k)\times k} & *\end{pmatrix}$.
Theorem 8 cannot be tight for the same reasons as in Remark 1.

The proof for the following bounds is very similar to that of Lemma 2 and Theorem 8.

Theorem 9 *In addition to the assumptions in Sect. 2.2.1, let* $\mathbf{A}$ *be positive definite; and let* $\mathbf{T} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ *be computed by Algorithm 1. Then*

$$0 \leq \log(\det\mathbf{A}) - \log\det(\mathbf{T}) \leq \log\det(\mathbf{\Lambda}_2) + \log\det\left(\mathbf{I}_{n-k} + \frac{\gamma^{2q-1}}{\lambda_k}\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2\,\mathbf{\Lambda}_2\right).$$

4 Probabilistic analysis

We derive probabilistic bounds for $\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2$, a term that represents the contribution of the starting guess $\mathbf{\Omega}$ in the dominant eigenspace $\mathbf{U}_1$ of $\mathbf{A}$, when the elements of $\mathbf{\Omega}$ are either Gaussian random variables (Sect. 4.1) or Rademacher random variables (Sect. 4.2).

The theory for Gaussian random matrices suggests the value $p \lesssim 20$, whereas theory for Rademacher random matrices suggests that $\ell \sim (k + \log n)\log k$ samples need to be taken to ensure $\mathsf{rank}(\mathbf{\Omega}_1) = k$. However, the theory for Rademacher random matrices is pessimistic, and numerical experiments demonstrate that a practical value of $p \lesssim 20$ is sufficient.

4.1 Gaussian random matrices

For the Gaussian starting guess, we present bounds for expectation

We split our analysis into two parts: an average case analysis (Sect. 4.1.1) and a concentration inequality (Sect. 4.1.1), and prove Theorem 1.

Definition 1 A "standard" Gaussian matrix has elements that are independently and identically distributed random $\mathcal{N}(0, 1)$ variables, that is normal random variables with mean 0 and variance 1.

"Appendix 1" summarizes the required results for standard Gaussian matrices. In particular, we will need the following.

Remark 3 The distribution of a standard Gaussian matrix $\mathbf{G}$ is rotationally invariant. That is, if $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices, then $\mathbf{U}^*\mathbf{G}\mathbf{V}$ has the same distribution as $\mathbf{G}$. Due to this property, the contribution of the starting guess on the dominant eigenspace does not appear in the bounds below.

4.1.1 Average case analysis

We present bounds for the expected values of $\|\mathbf{G}\|_2^2$ and $\|\mathbf{G}^\dagger\|_2^2$ for standard Gaussian matrices $\mathbf{G}$, and then prove Theorem 1.

Lemma 4 *Draw two Gaussian random matrices* $\mathbf{G}_2 \in \mathbb{R}^{(n-k)\times(k+p)}$ *and* $\mathbf{G}_1 \in \mathbb{R}^{k\times(k+p)}$ *and let* $p \geq 2$. *With* μ *defined in* (3), *then*

$$\mathbb{E}\left[\|\mathbf{G}_2\|_2^2\right] \leq \mu^2 + 2\left(\sqrt{\frac{\pi}{2}}\mu + 1\right). \tag{16}$$

If, in addition, also $k \geq 2$, *then*

$$\mathbb{E}\left[\|\mathbf{G}_1^\dagger\|_2^2\right] \leq \frac{p+1}{p-1}\left(\frac{1}{2\pi\,(p+1)}\right)^{2/(p+1)}\left(\frac{e\sqrt{k+p}}{p+1}\right)^2. \tag{17}$$

Proof See "Appendix 1".

We are ready to derive the main theorem on the expectation of standard Gaussian matrices.

Proof of Theorem 1 We start as in the proof of [19, Theorem 10.5]. The assumptions in Sect. 2.2.1 and Remark 3 imply that $\mathbf{U}^*\boldsymbol{\Omega}$ is a standard Gaussian matrix. Since $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are non-overlapping submatrices of $\mathbf{U}^*\boldsymbol{\Omega}$, they are both standard Gaussian and stochastically independent. The sub-multiplicative property implies $\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2 \leq \|\boldsymbol{\Omega}_2\|_2\|\boldsymbol{\Omega}_1^\dagger\|_2$.

We use the independence of $\boldsymbol{\Omega}_2$ and $\boldsymbol{\Omega}_1$ and apply both parts of (17); with $\mu \equiv \sqrt{n-k} + \sqrt{k+p}$ this gives

$$\mathbb{E}\left[\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2^2\right] \leq \left[\mu^2 + 2\left(\sqrt{\frac{\pi}{2}}\mu + 1\right)\right]\frac{p+1}{p-1}\left(\frac{1}{2\pi(p+1)}\right)^{\frac{2}{(p+1)}}\left(\frac{e\sqrt{k+p}}{p+1}\right)^2. \tag{18}$$

Bounding

$$\mu^2 + 2\left(\sqrt{\frac{\pi}{2}}\mu + 1\right) \leq (\mu + \sqrt{2})^2.$$

gives $\mathbb{E}\left[\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2^2\right] \leq C_{ge}$. Substituting into the result of Theorem 6 gives the desired result in Theorem 1.

For a positive definite matrix $\log\det(\mathbf{A}) = \mathsf{trace}(\log(\mathbf{A}))$, therefore

$$\log\det(\mathbf{I} + \gamma^{2q-1}\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2^2\boldsymbol{\Lambda}_2) = \sum_{j=k+1}^{n} \log\left(1 + \gamma^{2q-1}\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2^2\lambda_j\right).$$

Observe that $\log(1+\alpha x)$ is a concave function. Using Jensen's inequality, for $j = k+1, \ldots, n$

$$\mathbb{E}\left[\log(1 + \gamma^{2q-1}\lambda_j\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2^2)\right] \leq \log\left(1 + \gamma^{2q-1}\lambda_j\mathbb{E}\left[\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_2^2\right]\right).$$

Since $\log(1+\alpha x)$ is a monotonically increasing function, the second result in Theorem 1 follows by substituting the above equation into the bounds from Theorem 8 and simplifying the resulting expressions. □

4.1.2 Concentration inequality

As with the expectation bounds, it is clear that we must focus our attention on the term $\|\boldsymbol{\Omega}_2\|_2\|\boldsymbol{\Omega}_1^{\dagger}\|_2$. We reproduce here a result on the concentration bound of this term. The proof is provided in [16, Theorem 5.8].

Lemma 5 *Let $\boldsymbol{\Omega}_2 \in \mathbb{R}^{(n-k)\times(k+p)}$ and $\boldsymbol{\Omega}_1 \in \mathbb{R}^{k\times(k+p)}$ be two independent Gaussian random matrices and let $p \geq 4$. Then for $0 < \delta < 1$,*

$$\mathbb{P}\left[\|\boldsymbol{\Omega}_2\|_2^2\|\boldsymbol{\Omega}_1^{\dagger}\|_2^2 \geq C_g\right] \leq \delta, \tag{19}$$

where C_g is defined in Theorem 2.

The following statements mirror the discussion in [16, Section 5.3]. While the oversampling parameter p does not significantly affect the expectation bounds as long as $p \geq 2$, it seems to affect the concentration bounds significantly. The oversampling parameter p can be chosen in order to make $(2/\delta)^{2/(p+1)}$ a modest number, say less than equal to 10. Choosing

$$p = \left\lceil 2\log_{10}\left(\frac{2}{\delta}\right)\right\rceil - 1,$$

for $\delta = 10^{-16}$ gives us $p = 32$. In our experiments, we choose the value for the oversampling parameter to be $p = 20$.

4.2 Rademacher random matrices

We present results for the concentration bounds when $\boldsymbol{\Omega}$ is a Rademacher random matrix. We start with the following definition.

Definition 2 A Rademacher random matrix has elements that are independently and identically distributed and take on values ± 1 with equal probability.

Note that unlike standard Gaussian matrices, the distribution of a Rademacher matrix is not rotationally invariant.

As before we partition $\mathbf{U} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}$ and let $\boldsymbol{\Omega}_1 = \mathbf{U}_1^*\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_2 = \mathbf{U}_2^*\boldsymbol{\Omega}$. The following result bounds the tail of $\|\boldsymbol{\Omega}_2\|_2^2\|\boldsymbol{\Omega}_1^\dagger\|_2^2$. This result can be used to readily prove Theorem 3.

Theorem 10 *Let $\rho \in (0, 1)$ and $0 < \delta < 1$ and integers $n, k \geq 1$. Let the number of samples ℓ be defined as in Theorem 3. Draw a random Rademacher matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n\times\ell}$. Then*

$$\mathbb{P}\left[\|\boldsymbol{\Omega}_2\|_2^2\|\boldsymbol{\Omega}_1^\dagger\|_2^2 \geq C_r\right] \leq \delta,$$

where C_r is defined in Theorem 3.

Proof See "Appendix 2". □

Remark 4 From the proof of Theorem 10, to achieve $\|\boldsymbol{\Omega}_1^\dagger\|_2 \geq 3/\sqrt{\ell}$ with at least 99.5% probability and $n = 1024$, the number of samples required is

$$\ell \geq 2.54(\sqrt{k} + 11)^2(\log(4k) + 4.7).$$

Here $\rho = 8/9$ is chosen to be so that $1/(1 - \rho) = 9$.

The imposition that $\ell \leq n$ implies that the bound may only be informative for k small enough. Theorem 10 suggests that the number of samples $\ell \sim (k + \log n)\log k$ to ensure that $\|\boldsymbol{\Omega}_1^\dagger\|_2$ is small and $\mathsf{rank}(\boldsymbol{\Omega}_1) = k$.

We investigate this issue numerically. We generate random Rademacher matrices $\boldsymbol{\Omega}_1 \in \mathbb{R}^{k\times\ell}$; here we assume $\mathbf{U} = \mathbf{I}_n$. Here we choose three different values for k, namely $k = 10, 75, 500$. For each value of k, the oversampling ℓ varies from $\ell = k+1$ to $\ell = k + 40$. We generate 500 runs for each value of ℓ. In Fig. 1 we plot $\|\boldsymbol{\Omega}_1^\dagger\|_2^2$ and the percentage of matrices that are rank deficient. For $k = 10$, a few samples were rank deficient but the percentage of rank deficient matrices dropped significantly; after $p = 20$ there were no rank deficient matrices. For larger values of k we observed that none of the sampled matrices were rank deficient and $p = 20$ was sufficient to ensure that $\|\boldsymbol{\Omega}_1^\dagger\|_2^2 \lesssim 10$. Similar results were observed for randomly generated orthogonal matrices $\mathbf{U}$. In summary, a modest amount of oversampling $p \lesssim 20$ is sufficient to ensure that $\mathsf{rank}(\boldsymbol{\Omega}_1) = k$ for the Rademacher random matrices, similar to Gaussian random matrices. In further numerical experiments we shall use this particular choice of oversampling parameter p.

5 Numerical experiments

In this section, we demonstrate the performance of our algorithm and bounds on two different kinds of examples. In the first example, we focus on small matrices (with

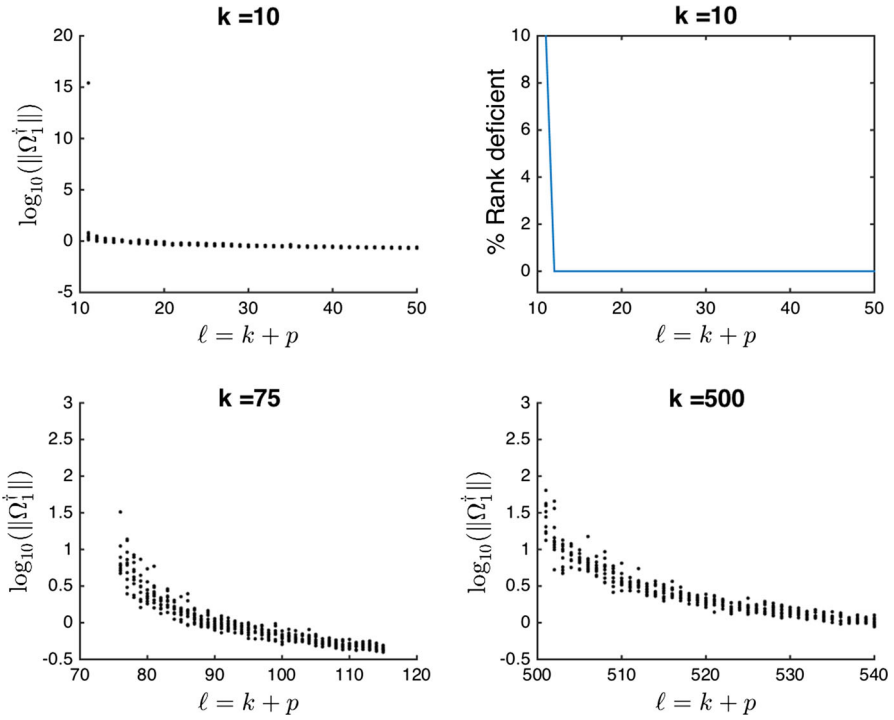

Fig. 1 The quantity $\log_{10}(\|\mathbf{\Omega}_1^\dagger\|_2)$ for different amounts of sampling $\ell = k+1$ to $\ell = k+40$. We consider $k = 10$ (*top left*), $k = 75$ (*bottom left*), and $k = 500$ (*bottom right*). The *top right plot* shows the percentage of numerically rank deficient sampled matrices

dimension 128) in the non-asymptotic regime. We show that our bounds are informative and our estimators are accurate with high probability. In the second examples, we look at medium sized matrices (of dimension 5000) and demonstrate the behavior of our estimators.

5.1 Small matrices

In this section we study the performance of the proposed algorithms on small test examples.

The matrix $\mathbf{A}$ is chosen to be of size 128×128 and its eigenvalues satisfy $\lambda_{j+1} = \gamma^j \lambda_1$ for $j = 1, \ldots, n-1$. To help interpret the results of Theorems 1–3, we provide simplified versions of the bounds. The relative error in the trace estimator can be bounded as

$$\Delta_t \equiv \frac{\mathsf{trace}(\mathbf{A}) - \mathsf{trace}(\mathbf{T})}{\mathsf{trace}(\mathbf{A})} \leq (1 + \gamma^{2q-1} C) \frac{\gamma^k (1 - \gamma^{n-k})}{1 - \gamma^n}. \tag{20}$$

Here C can take the value C_g for a Gaussian starting guess and C_r for a Rademacher starting guess.

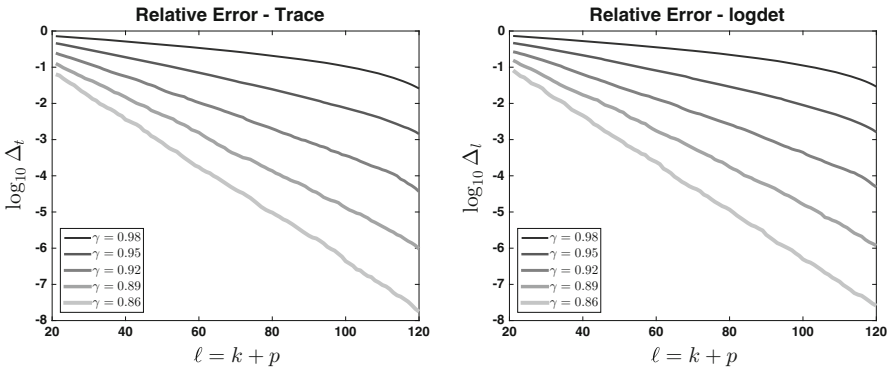

Fig. 2 Accuracy of proposed estimators on a matrix with geometrically decaying eigenvalues. The relative error is plotted against the sample size. Accuracy of (*left*) trace and (*right*) logdet estimators. Here, a Gaussian starting guess was used

For the logdet estimator, we observe that $\log(1 + x) \leq x$. Using the relation $\log\det(\mathbf{I} + \mathbf{\Lambda}) = \mathsf{trace}\log(\mathbf{I} + \mathbf{A})$, it is reasonable to bound $\log\det(\mathbf{I} + \mathbf{\Lambda})$ by $\mathsf{trace}(\mathbf{\Lambda})$. With the abbreviation $f(\cdot) = \log\det(\cdot)$ we can bound the relative error of the logdet estimator as

$$\Delta_l \equiv \frac{f(\mathbf{I}+\mathbf{A}) - f(\mathbf{I}+\mathbf{T})}{f(\mathbf{I}+\mathbf{A})} \leq (1 + \gamma^{2q-1}C)\frac{\gamma^k(1-\gamma^{n-k})}{1-\gamma^n}.$$

Consequently, the error in the trace and logdet approximations approaches 0 as $k \to n$ and is equal to 0 if $k = n$.

In the following examples, we study the performance of the algorithms with increasing sample size. It should be noted here that, since $\ell = k + p$ and p is fixed, increasing the sample size corresponds to increasing the dimension k; consequently, the location of the gap is changing, as is the residual error $\Delta = \mathsf{trace}(\mathbf{\Lambda}_2)/\mathsf{trace}(\mathbf{\Lambda})$.

1. Effect of eigenvalue gap Matrices $\mathbf{A}$ are generated with different eigenvalue distributions. The eigenvalue gap parameter γ varies from 0.98 to 0.86. We consider sampling from both Gaussian random matrices. The oversampling was set to be $p = 20$ for both distributions. The subspace iteration parameter q was set to be 1. The results are displayed in Fig. 2. Clearly, both the trace and logdet become increasingly accurate as the eigenvalue gap increases. This confirms the theoretical estimate in (20) since the error goes to zero as $k \to n$. The behavior of the error with both Gaussian and Rademacher starting guesses is very similar and is not displayed here.

2. Comparison with Monte Carlo estimators We fix the eigenvalue gap to $\gamma = 0.9$, sampling parameter $p = 20$ and subspace iteration parameter $q = 1$. We consider sampling from both Gaussian and Rademacher random matrices and consider their accuracy against their Monte Carlo counterparts. As mentioned earlier, the Monte Carlo estimator cannot be directly applied to the logdet estimator; however, using the following identity

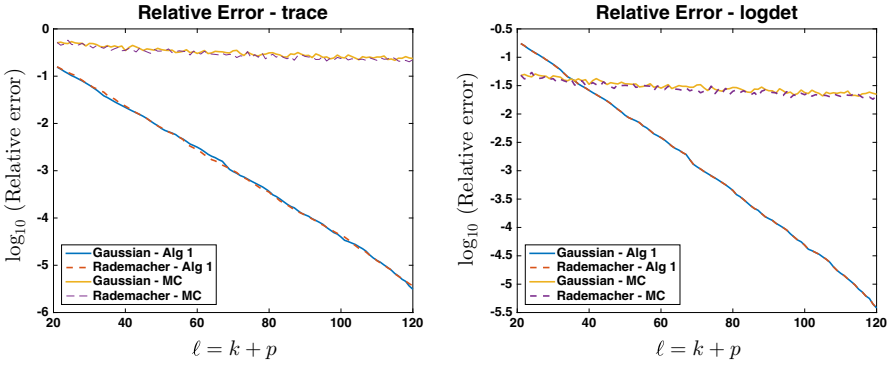

Fig. 3 Comparison against Monte Carlo estimators for (*left*) trace and (*right*) logdet computations. The relative error is plotted against the sample size

$$\log\det(\mathbf{I}_n + \mathbf{A}) = \mathsf{trace}\log(\mathbf{I}_n + \mathbf{A}),$$

the Monte Carlo estimators can be applied to the matrix $\log(\mathbf{I}_n + \mathbf{A})$. For a fair comparison with the estimators proposed in this paper, the number of samples used equals the target rank plus the oversampling parameter $p = 20$, i.e., $(k + p)$ samples. We averaged the Monte Carlo estimators over 100 independent runs. The results are illustrated in Fig. 3. It can be readily seen that when the matrix has rapidly decaying eigenvalues, our estimators are much more accurate than the Monte Carlo estimators. The number of samples required for the Monte Carlo methods for a relative accuracy ϵ depends on ϵ^{-2}, so the number of samples required for an accurate computation can be large. For the logdet estimator, initially the Monte Carlo estimator seems to outperform our method for small sample sizes; however, the error in our estimators decays sharply. It should be noted that for this small problem one can compute $\log(\mathbf{I} + \mathbf{A})$ but for a larger problem it may be costly, even prohibitively expensive. For all the cases described here, Gaussian and Rademacher random matrices seem to have very similar behavior.

3. Effect of subspace iteration parameter The matrix $\mathbf{A}$ is the same as in the previous experiment but p is chosen to be 0. The subspace iteration parameter is varied from $q = 1$ to $q = 5$. The results of the relative error as a function of ℓ are displayed in Fig. 4. The behavior is similar for both Gaussian and Rademacher starting guesses, therefore we only display results for Gaussian starting guess. We would like to emphasize that Algorithm 1 is not implemented *as is* since it is numerically unstable and susceptible to round-off error pollution; instead a numerically stable version is implemented based on [16, Algorithm A.1]. As can be seen, increasing the parameter improves the accuracy for a fixed target rank k. However, both from the analysis and the numerical results, this is clearly a case of diminishing returns. This is because the overall error is dominated by $\mathsf{trace}(\mathbf{\Lambda}_2)$ and $\log\det(\mathbf{I}_{n-k} + \mathbf{\Lambda}_2)$. Increasing the subspace iteration parameter q only improves the multiplicative factor in front of one of the terms. Moreover, in the case that the eigenvalues are decaying rapidly, one iteration, i.e., $q = 1$ is adequate to get an accurate estimator.

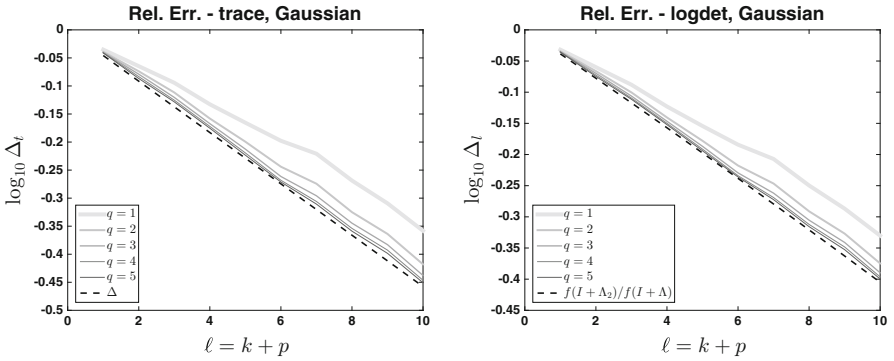

Fig. 4 Effect of subspace iteration parameter q on the (*left*) trace estimator and (*right*) logdet estimator. The relative error is plotted against the sample size. A Gaussian starting guess was used. The behavior is similar for Rademacher starting guesses

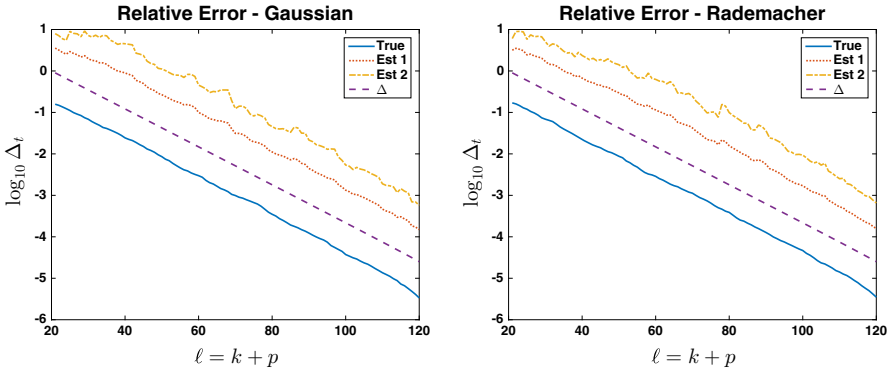

Fig. 5 Accuracy of the error bounds for the trace estimators. The relative error is plotted against the sample size. (*left*) Gaussian and (*right*) Rademacher random matrices. 'Est 1' and 'Est 2' refers to the bounds in Theorem 6. For comparison, we also plot $\mathsf{trace}(\mathbf{\Lambda}_2)/\mathsf{trace}(\mathbf{A})$

4. How descriptive are the bounds? In this experiment we demonstrate the accuracy of the bounds derived in Sect. 3. The matrix is chosen to be the same as the one in Experiment 2. In Fig. 5 we consider the bounds in the trace estimator derived in Theorem 6. We consider both the Gaussian (left panel) and Rademacher distributions (right panel). For comparison we also plot the term Δ, which is the theoretical optimum. 'Est 1' refers to the first bound in (7) and 'Est 2' refers to the second bound in (8). Both the bounds are qualitatively similar to both the true error and the theoretical estimate Δ, and also quantitatively within a factor of 10 of the theoretical estimate Δ. Since γ is close to 1 and $\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2 > 1$, $\gamma\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2 > 1$ and therefore 'Est 1' is a more accurate estimator. The error of the logdet estimator is plotted against the theoretical bounds (see Theorem 8) in Fig. 6; as before, our estimator is both qualitatively and quantitatively accurate. The conclusions are identical for both Gaussian and Rademacher matrices. The empirical performance of this behavior is studied in the next experiment.

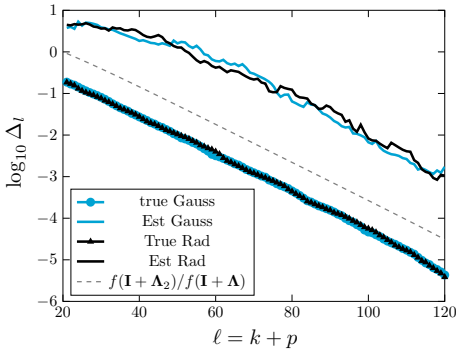

Fig. 6 Accuracy of the error bounds for the logdet estimators. The relative error is plotted against the sample size. Both Gaussian and Rademacher starting guesses are used. For comparison, we also plot $\log\det(\mathbf{I}+\mathbf{\Lambda}_2)/\log\det(\mathbf{I}+\mathbf{A})$

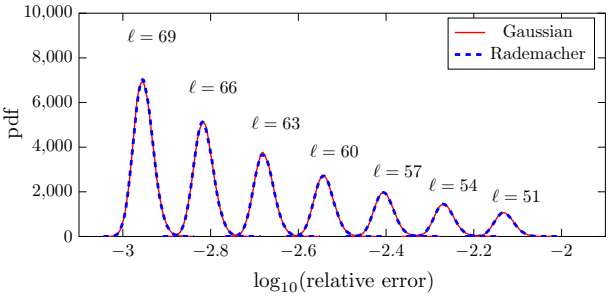

Fig. 7 Empirical distribution of the relative error for the trace estimator. 10^5 samples were used for the distribution

5. Concentration of measure We choose the same matrix as in Experiment 2. We generate 10^5 starting guesses (both Gaussian and Rademacher) and compute the distribution of relative errors for the trace (quantified by Δ_t) and logdet (quantified by Δ_l). Figures 7 and 8 show the empirical probability density function for the relative errors in the trace and logdet respectively. We observe that the two distributions are nearly identical and that the empirical density is concentrated about the mean. Furthermore, as the sample size ℓ increases, both the mean and variance of the empirical distribution decrease. These results demonstrate that the randomized methods are indeed effective with high probability.

5.2 Medium sized example

This example is inspired by a test case from Sorensen and Embree [39]. Consider the matrix $\mathbf{A} \in \mathbb{R}^{5000\times 5000}$ defined as

$$\mathbf{A} \equiv \sum_{j=1}^{40} \frac{h}{j^2}\mathbf{x}_j\mathbf{x}_j^T + \sum_{j=41}^{300} \frac{l}{j^2}\mathbf{x}_j\mathbf{x}_j^T, \tag{21}$$

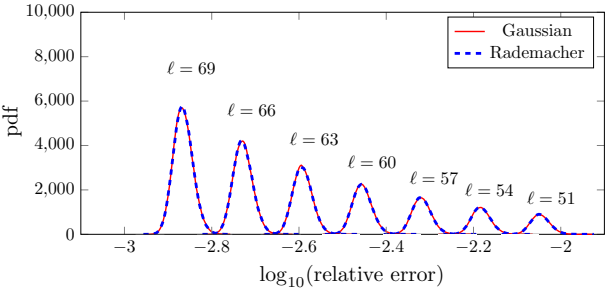

Fig. 8 Empirical distribution of the relative error for the logdet estimator. 10^5 samples were used for the distribution

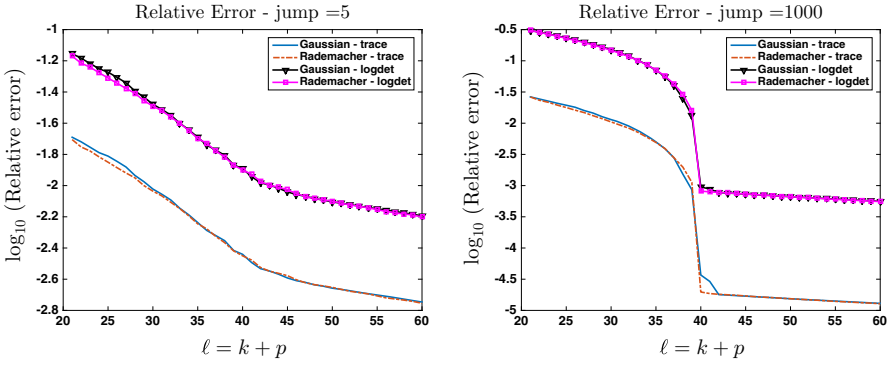

Fig. 9 Accuracy of trace and logdet computations for the matrix in (21). (*left*) $l = 1, h = 5$ and (*right*) $l = 1, h = 1000$

where $\mathbf{x}_j \in \mathbb{R}^{5000}$ are sparse vectors with random nonnegative entries. In MATLAB this can be generated using the command `xj = sprand(5000,1,0.025)`. It should be noted the vectors $\mathbf{x}_j$ are not orthonormal; therefore, the outer product form is not the eigenvalue decomposition of the matrix $\mathbf{A}$. However, the eigenvalues decay like $1/j^2$ with a gap at index 40, and its magnitude depends on the ratio h/l. The exact rank of this matrix is 300.

First we fix $l = 1$ and consider two different cases $h = 5$, and 1000. The oversampling parameter $p = 20$ and the subspace iteration parameter is $q = 1$. The results are displayed in Fig. 9. The accuracy of both the trace and the logdet estimators improves considerably around the sample size $\ell = 40$ mark, when the eigenvalues undergo the large jump for $h = 1000$; the transition is less sharp when $h = 5$. This demonstrates the benefit of having a large eigenvalue gap for the accuracy of the estimators. As an extreme case, consider $l = 0$ and $h = 2$. In this example, the matrix $\mathbf{A}$ has exactly rank 40, and therefore 40 matrix–vector products with $\mathbf{A}$ are enough to recover the trace and logdet to machine precision (see Fig. 10). This result highlights the power of our proposed estimators.

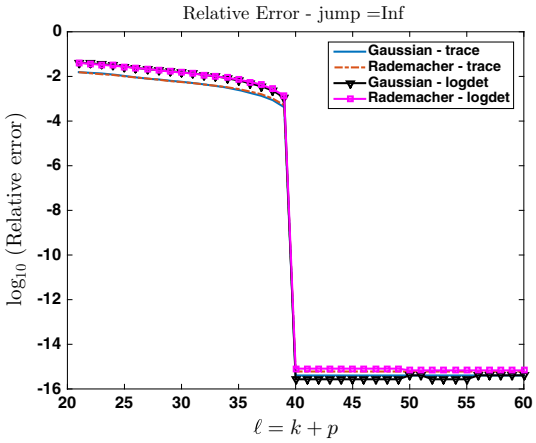

Fig. 10 Accuracy of trace and logdet computations for the matrix in (21) with $l = 0, h = 2$

6 Applications to evaluation of uncertainty quantification measures

As mentioned in the introduction, the computation of traces and log-determinants of high-dimensional operators is essential is the emerging field of uncertainty quantification. In this section, we use the methods developed in this article to compute some common statistical quantities that appear in the context of Bayesian inverse problems. In particular, we focus on a time-dependent advection–diffusion equation in which we seek to infer an uncertain initial condition from measurements of the concentration at discrete points in space/time; This is a commonly used example in the inverse problem community; see e.g., [1,3,13,34]. Below, we briefly outline the components of the Bayesian inverse problem. The model problem used here is adaptated from [3], and therefore, we refer the readers to that paper for further details.

The forward problem The forward problem models diffusive transport of a contaminant in a domain $\mathcal{D} \subset \mathbb{R}^2$, which is depicted in Fig. 11 (left). The domain boundary $\partial\mathcal{D}$ is a combination of the outer edges of the domain as well as the internal boundaries of the rectangles that model buildings. The *forward operator* maps an initial condition θ to space/time observations of the contaminant concentration, by solving the advection diffusion equation,

$$\begin{aligned} &u_t - \kappa \Delta u + \mathbf{v} \cdot \nabla u = 0 \text{ in } \mathcal{D} \times (0, T), \\ &u(\cdot, 0) = \theta \qquad\qquad\quad \text{in } \mathcal{D}, \\ &\kappa \nabla u \cdot \mathbf{n} = 0 \qquad\qquad\quad \text{on } \partial\mathcal{D} \times (0, T), \end{aligned} \tag{22}$$

and extracting solution values at spatial points [sensor locations as indicated in Fig. 11 (left)] and at pre-specified times. Here, $\kappa > 0$ is the diffusion coefficient and $T > 0$ is the final time. In our numerical experiments, we use $\kappa = 0.001$. The velocity field $\mathbf{v}$, shown in Fig. 11, is obtained by solving a steady Navier-Stokes equation with the side walls driving the flow; see [3] for details. The discretized equations give rise to a discretized linear solution operator for the forward problem, which is composed

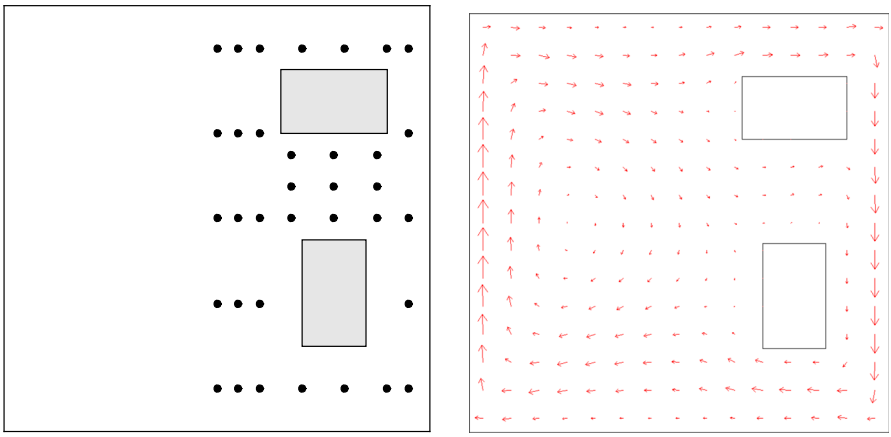

Fig. 11 *Left* the computational domain $\mathcal{D}$ is the region $[0, 1]^2$ with two rectangular regions (representing buildings) removed. The *black dots* indicate the locations of sensors where observations are recorded. *Right* the velocity field

with an observation operator to extract the space-time observations. We denote this discretized forward operator by $\mathbf{F}$.

The Bayesian inverse problem The inverse problem aims to use a vector of observed data $\mathbf{d}$, which consists of sensor measurements at discrete points in time, to reconstruct the uncertain initial condition. The dimension of $\mathbf{d}$, which we denote by N_{obs}, is given by the product of the number of sensors and the number of points in time where observations are recorded. In the present example, we use 35 sensors and record measurements at $t = 1, t = 2$, and $t = 3.5$. Therefore, we have $\mathbf{d} \in \mathbb{R}^{N_{\text{obs}}}$, with $N_{\text{obs}} = 105, n = 1018$, and that $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^{105}$. We use a Gaussian prior measure $\mathcal{N}(\boldsymbol{\theta}_0, \mathbf{C}_0)$, and use an additive Gaussian noise model. Following [3], the prior covariance is chosen to be the discretized biharmonic operator. The solution of the Bayesian inverse problem is the posterior measure, $\mathcal{N}(\boldsymbol{\theta}_{\text{post}}, \mathbf{C}_{\text{post}})$ with

$$\mathbf{C}_{\text{post}} = (\mathbf{F}^* \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{F} + \mathbf{C}_0^{-1})^{-1}, \qquad \boldsymbol{\theta}_{\text{post}} = \mathbf{C}_{\text{post}} (\mathbf{F}^* \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{d} + \mathbf{C}_0^{-1} \boldsymbol{\theta}_0),$$

We denote by $\mathbf{H} \equiv \mathbf{F}^* \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{F}$ the Fisher information matrix. In many applications $\mathbf{H}$ has a rapidly decaying spectrum; see Fig. 12 (left). Moreover, in the present setup, the rank of $\mathbf{H}$ is bounded by the dimension of the observations, which in our example is given by $N_{\text{obs}} = 105$. The prior-preconditioned Fisher information matrix

$$\mathbf{H}_0 = \mathbf{C}_0^{1/2} \mathbf{H} \mathbf{C}_0^{1/2}$$

is also of importance in what follows. Notice that preconditioning of $\mathbf{H}$ by the prior, due to the smoothing properties of the priors employed in the present example, results in a more rapid spectral decay; see Fig. 12 (right).

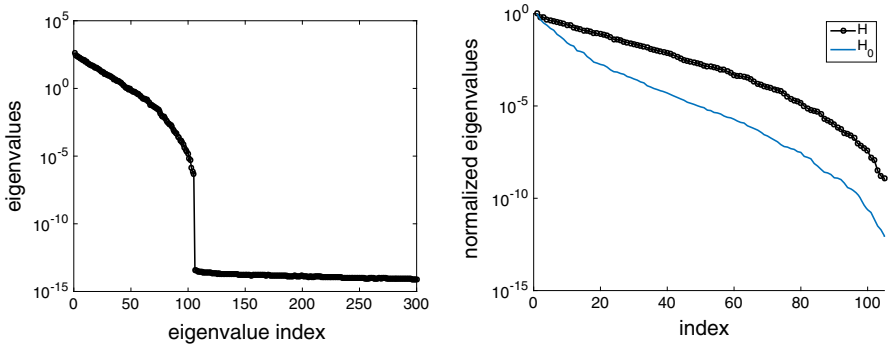

Fig. 12 *Left* first 300 eigenvalues of $\mathbf{H}$; *right* normalized nonzero eigenvalues of $\mathbf{H}$ and $\mathbf{H}_0$

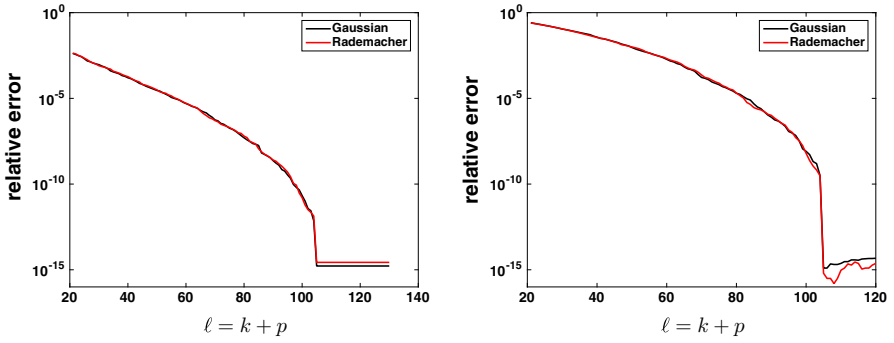

Fig. 13 *Left* error in estimation of $trace(\mathbf{H}_0)$; *right* error in estimation of $\log\det(\mathbf{H}_0 + \mathbf{I})$. Computations were done using $p = 20$ and the randomized estimators use $\ell = k + p$ random vectors with increasing values of k

We point out that the quantity $\mathsf{trace}(\mathbf{H}_0)$ is related to the *sensitivity criterion* in optimal experimental design (OED) theory [44]. On the other hand, $\log\det(\mathbf{I}+\mathbf{H}_0)$ is related to Bayesian D-optimal design criterion [11]. As shown in [2], $\log\det(\mathbf{I}+\mathbf{H}_0)$ is the expected information gain from the posterior measure to the prior measure in a Bayesian linear inverse problem with Gaussian prior and noise distributions, and with an inversion parameter that belongs to a Hilbert space. Note that in the present context, information gain is quantified by the Kullback–Leibler divergence from posterior measure to prior measure. A detailed discussion of uncertainty measures is also provided in [38].

In Fig. 13, we report the error in the approximation of $\mathsf{trace}(\mathbf{H})$ and $\log\det(\mathbf{I}+\mathbf{H}_0)$. Both of these quantities are of interest in OED theory, where one is interested in measures of uncertainty in reconstructed parameters [5,44]. Such statistical measures are then used to guide the experimental configurations used to collect experimental data so as to maximize the statistical quality of the reconstructed/inferred parameters. Note that, in the present example, an experimental configuration is given by the placement of sensors [black dots in Fig. 11 (left) where concentration data is recorded].

7 Conclusion

We present randomized estimators for the trace and log-determinant of implicitly defined Hermitian positive semi-definite matrices. The estimators are low-rank approximations computed with subspace iteration. We show, theoretically and numerically, that our estimators are effective for matrices with a large eigenvalue gap or rapidly decaying eigenvalues.

Our error analyses for the estimators are cleanly separated into two parts: A structural analysis, which is applicable to any choice of a starting guess, paves the way for a probabilistic analysis, in this case for Gaussian and Rademacher starting guesses. In addition, we derive asymptotic bounds on the number of random vectors required to guarantee a specified accuracy with low probability of failure. We present comprehensive numerical experiments to illustrate the performance of the estimators, and demonstrate their suitability for challenging application problems, such as the computation of the expected information gain in a Bayesian linear inverse problem governed by a time-dependent PDE.

Future work will evolve around two main issues.

Rademacher random matrices Our analysis implies that a Gaussian starting guess can do with a fixed oversampling parameter, while the oversampling amount for a Rademacher starting guess depends on the dimension of the dominant eigenspace and the dimension of the matrix. However, the numerical experiments indicate that, for both types of starting guesses, an oversampling parameter of 20 leads to accurate estimators. We plan to further investigate estimators with Rademacher starting guesses, and specifically to derive error bounds for the expectation of the corresponding estimators. Another issue to be explored is the tightness of the bound $\ell \sim (k + \log n) \log k$ for Rademacher starting guesses.

Applications We plan to integrate our estimators into computational methods for large-scale uncertainty quantification. Our main goal is the computation of OED for large-scale inverse problems. This can be posed as an optimization problem, where the objective function is the trace or log-determinant of a high-dimensional operator. Due to their efficiency and high accuracy, we expect that our estimators are well suited for OED.

Appendix 1: Gaussian random matrices

In this section, we state a lemma on the pseudo-inverse of a rectangular Gaussian random matrix, and use this result to prove both parts of Lemma 4.

Pseudo-inverse of a Gaussian random matrix

We state a result on the large deviation bound of the pseudo-inverse of a Gaussian random matrix [19, Proposition 10.4].

Lemma 6 *Let* $\mathbf{G} \in \mathbb{R}^{k\times(k+p)}$ *be a random Gaussian matrix and let* $p \geq 2$. *For all* $t \geq 1$,

$$\mathbb{P}\left[\, \|\mathbf{G}^{\dagger}\|_2 \geq \frac{e\sqrt{k+p}}{p+1} \cdot t \right] \leq t^{-(p+1)}. \tag{23}$$

Proof of Lemma 4

Proof From [45, Corollary 5.35] we have

$$\mathbb{P}\left[\, \|\mathbf{G}_2\|_2 > \sqrt{n-k} + \sqrt{k+p} + t \right] \leq \exp(-t^2/2).$$

Recall from (3) $\mu = \sqrt{n-k} + \sqrt{k+p}$. From the *law of the unconscious statistician* [16, Proposition S4.2],

$$\begin{aligned}
\mathbb{E}\left[\, \|\mathbf{G}_2\|_2^2 \right] &= \int_0^{\infty} 2t\mathbb{P}\left[\, \|\mathbf{G}_2\|_2 > t \right] dt \\
&\leq \int_0^{\mu} 2t\,dt + \int_{\mu}^{\infty} 2t\mathbb{P}\left[\, \|\mathbf{G}_2\|_2 > t \right] dt \\
&\leq \mu^2 + \int_0^{\infty} 2(u+\mu)\exp(-u^2/2)du = \mu^2 + 2\left(1 + \mu\sqrt{\frac{\pi}{2}}\right).
\end{aligned}$$

This concludes the proof for (16).

Next consider (17). Using Lemma 6, we have for $t > 0$

$$\mathbb{P}\left[\, \|\mathbf{G}_1^{\dagger}\|_2 \geq t \right] \leq Dt^{-(p+1)} \qquad D \equiv \frac{1}{\sqrt{2\pi(p+1)}}\left(\frac{e\sqrt{k+p}}{p+1}\right). \tag{24}$$

As before, we have

$$\begin{aligned}
\mathbb{E}\left[\, \|\mathbf{G}_1^{\dagger}\|_2^2 \right] &= \int_0^{\infty} 2t\mathbb{P}\left[\, \|\mathbf{G}_1^{\dagger}\|_2 > t \right] dt \\
&\leq \int_0^{\beta} 2t\,dt + \int_{\beta}^{\infty} 2t\mathbb{P}\left[\, \|\mathbf{G}_1^{\dagger}\|_2 > t \right] dt \\
&\leq \beta^2 + \int_{\beta}^{\infty} 2t\,Dt^{-(p+1)}dt = \beta^2 + 2D\frac{\beta^{1-p}}{p-1}.
\end{aligned}$$

Minimizing w.r.t. β, we get $\beta = (D)^{1/(p+1)}$. Substitute this value for β and simplify. □

Appendix 2: Rademacher random matrices

In this section, we state the matrix Chernoff inequalities [43] and other useful concentration inequalities and use these results to prove Theorem 10.

Useful concentration inequalities

The proof of Theorem 10 relies on the matrix concentration inequalities developed in [43]. We will need the following result [43, Theorem 5.1.1] in what follows.

Theorem 11 (Matrix Chernoff) *Let* $\{\mathbf{X}_k\}$ *be finite sequence of independent, random,* $d \times d$ *Hermitian matrices. Assume that* $0 \leq \lambda_{\min}(\mathbf{X}_k)$ *and* $\lambda_{\max}(\mathbf{X}_k) \leq L$ *for each index* k. *Let us define*

$$\mu_{\min} \equiv \lambda_{\min}\left(\sum_k \mathbb{E}[\mathbf{X}_k]\right) \qquad \mu_{\max} \equiv \lambda_{\max}\left(\sum_k \mathbb{E}[\mathbf{X}_k]\right),$$

and let $g(x) \equiv e^x(1+x)^{-(1+x)}$. *Then for any* $\epsilon > 0$

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq (1+\epsilon)\mu_{\max}\right] \leq d\, g(\epsilon)^{\mu_{\max}/L},$$

and for any $0 \leq \epsilon < 1$

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_k \mathbf{X}_k\right) \leq (1-\epsilon)\mu_{\min}\right] \leq d\, g(-\epsilon)^{\mu_{\min}/L}.$$

The following result was first proved by Ledoux [25] but we reproduce the statement from [42, Proposition 2.1].

Lemma 7 *Suppose* $f : \mathbb{R}^n \to \mathbb{R}$ *is a convex function that satisfies the following Lipschitz bound*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Let $\mathbf{z} \in \mathbb{R}^n$ *be a random vector with entries drawn from an i.i.d. Rademacher distribution. Then, for all* $t \geq 0$,

$$\mathbb{P}[f(\mathbf{z}) \geq \mathbb{E}[f(\mathbf{z})] + Lt] \leq e^{-t^2/8}.$$

Lemma 8 *Let* $\mathbf{V}$ *be a* $n \times r$ *matrix with orthonormal columns and let* $n \geq r$. *Let* $\mathbf{z}$ *be an* $n \times 1$ *vector with entries drawn from an i.i.d. Rademacher distribution. Then, for* $0 < \delta < 1$,

$$\mathbb{P}\left[\ \|\mathbf{V}^*\mathbf{z}\|_2 \geq \sqrt{r} + \sqrt{8\log\left(\frac{1}{\delta}\right)}\right] \leq \delta.$$

Proof Our proof follows the strategy in [42, Lemma 3.3]. Define the function $f(\mathbf{x}) = \|\mathbf{V}^*\mathbf{x}\|_2$. We observe that f satisfies the assumptions of Lemma 7, with Lipschitz constant $L = 1$; the latter follows from

$$|\|\mathbf{V}^*\mathbf{x}\|_2 - \|\mathbf{V}^*\mathbf{y}\|_2| \leq \|\mathbf{V}^*(\mathbf{x}-\mathbf{y})\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2.$$

Furthermore, using Hölder's inequality

$$\mathbb{E}[\,f(\mathbf{z})] \leq [\mathbb{E}\left[\,f(\mathbf{z})^2\right]]^{1/2} = \|\mathbf{V}\|_F = \sqrt{r}.$$

Using Lemma 7 with $t_\delta = \sqrt{8\log(1/\delta)}$ we have

$$\mathbb{P}\left[\,f(\mathbf{z}) \geq \sqrt{r} + t_\delta\right] \leq \mathbb{P}[\,f(\mathbf{z}) \geq \mathbb{E}[\,f(\mathbf{z})] + t_\delta] \leq e^{-t_\delta^2/8} = \delta.$$

□

Lemma 9 *Let X_i for $i = 1, \ldots, n$ be a sequence of i.i.d. random variables. If for each $i = 1, \ldots, n$, $\mathbb{P}[\,X_i \geq a] \leq \xi$ holds, where $\xi \in (0, 1]$, then*

$$\mathbb{P}\left[\max_{i=1,\cdots,n} X_i \geq a\right] \leq n\xi.$$

Proof Since $\mathbb{P}[\,X_i \geq a] \leq \xi$ then $\mathbb{P}[\,X_i < a] \geq 1 - \xi$. We can bound

$$\begin{aligned}\mathbb{P}\left[\max_{i=1,\cdots,n} X_i \geq a\right] &= \left(1 - \mathbb{P}\left[\max_{i=1,\cdots,n} X_i < a\right]\right) \\ &= \left(1 - \prod_{i=1}^{n} \mathbb{P}[\,X_i < a]\right) \leq 1 - (1-\xi)^n.\end{aligned}$$

The proof follows from Bernoulli's inequality [40, Theorem 5.1] which states $(1 - \xi)^n \geq 1 - n\xi$ for $\xi \in [0, 1]$ and $n \geq 1$. □

Proof of Theorem 10

Proof Recall that $\mathbf{\Omega}_1 = \mathbf{U}_1^*\mathbf{\Omega}$ and $\mathbf{\Omega}_2 = \mathbf{U}_2^*\mathbf{\Omega}$ where $\mathbf{\Omega}$ is random matrix with entries chosen from an i.i.d. Rademacher distribution. The proof proceeds in three steps.

1. Bound for $\|\boldsymbol{\Omega}_2\|_2^2$ The proof uses the matrix Chernoff concentration inequality. Let $\boldsymbol{\omega}_i \in \mathbb{R}^{n\times 1}$ be the i-th column of $\boldsymbol{\Omega}$. Note $\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^* \in \mathbb{C}^{(n-k)\times(n-k)}$ and

$$\mathbb{E}\left[\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*\right] = \sum_{i=1}^{\ell} \mathbf{U}_2^*\mathbb{E}\left[\boldsymbol{\omega}_i\boldsymbol{\omega}_i^*\right]\mathbf{U}_2 = \ell\mathbf{I}_{n-k}.$$

Furthermore, define $\mu_{\min}(\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*) \equiv \lambda_{\min}(\mathbb{E}\left[\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*\right])$ and $\mu_{\max}(\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*) \equiv \lambda_{\max}(\mathbb{E}\left[\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*\right])$. Clearly $\mu_{\min} = \mu_{\max} = \ell$. Note that here we have expressed $\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*$ as a finite sum of ℓ rank-1 matrices, each with a single nonzero eigenvalue $\boldsymbol{\omega}_i^*\mathbf{U}_2\mathbf{U}_2^*\boldsymbol{\omega}_i$. We want to obtain a probabilistic bound for the maximum eigenvalue i.e., $L_2 = \max_{i=1,\cdots,\ell}\|\mathbf{U}_2^*\boldsymbol{\omega}_i\|_2^2$. Using Lemma 8 we can write with probability at most $e^{-t^2/8}$

$$\left(\sqrt{n-k}+t\right)^2 \leq \|\mathbf{U}_2^*\boldsymbol{\omega}_i\|_2^2 = \boldsymbol{\omega}_i^*\mathbf{U}_2\mathbf{U}_2^*\boldsymbol{\omega}_i.$$

Since $\|\mathbf{U}_2^*\boldsymbol{\omega}_i\|_2^2$ are i.i.d., applying Lemma 9 gives

$$\mathbb{P}\left[\max_{i=1,\cdots,\ell}\|\mathbf{U}_2^*\boldsymbol{\omega}_i\|_2 \geq \sqrt{n-k}+t\right] \leq \ell e^{-t^2/8}.$$

Take $t = \sqrt{8\log(4\ell/\delta)}$ to obtain

$$\mathbb{P}\left[L_2 \geq C_u^2\right] \leq \delta/4, \qquad C_u \equiv \sqrt{n-k}+\sqrt{8\log\left(\frac{4\ell}{\delta}\right)}. \tag{25}$$

The matrix $\boldsymbol{\Omega}_2$ satisfies the conditions of the matrix Chernoff theorem 11; for $\eta \geq 0$ we have

$$\mathbb{P}\left[\lambda_{\max}(\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*) \geq (1+\eta)\ell\right] \leq (n-k)g(\eta)^{\frac{\ell}{L_2}},$$

where the function $g(\eta)$ is defined in Theorem 11. For $\eta > 1$ the Chernoff bounds can be simplified [30, Section 4.3] since $g(\eta) \leq e^{-\eta/3}$, to obtain

$$\mathbb{P}\left[\lambda_{\max}(\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^*) \geq (1+\eta)\ell\right] \leq (n-k)\exp\left(-\frac{\eta\ell}{3L_2}\right).$$

Choose the parameter

$$\eta_\delta = C_{\ell,\delta}C_u^2 = 3\ell^{-1}C_u^2\log\left(\frac{4(n-k)}{\delta}\right),$$

so that

$$\mathbb{P}\left[\ \|\mathbf{\Omega}_2\|_2^2 \geq (1+\eta_\delta)\ell\right] \leq (n-k)\exp\left(-\frac{C_u^2}{L_2}\log\frac{4(n-k)}{\delta}\right) = (n-k)\left(\frac{\delta}{4(n-k)}\right)^{C_u^2/L_2}.$$

Finally, we want to find a lower bound for $\|\mathbf{\Omega}_2\|_2^2$. Define the events

$$A = \left\{\mathbf{\Omega}_2 \mid L_2 < C_u^2\right\}, \qquad B = \left\{\mathbf{\Omega}_2 \mid \|\mathbf{\Omega}_2\|_2^2 \geq (1+\eta_\delta)\ell\right\}.$$

Note that $\mathbb{P}\left[\ A^c\right] \leq \delta/4$ and under event A we have $C_u^2 > L_2$ so that

$$\mathbb{P}[\ B \mid A] \leq (n-k)\left(\frac{\delta}{4(n-k)}\right)^{C_u^2/L_2} \leq \delta/4.$$

Using the law of total probability

$$\begin{aligned}\mathbb{P}[\ B] &= \mathbb{P}[\ B \mid A]\,\mathbb{P}[\ A] + \mathbb{P}\left[\ B \mid A^c\right]\mathbb{P}\left[\ A^c\right] \\ &\leq \mathbb{P}[\ B \mid A] + \mathbb{P}\left[\ A^c\right],\end{aligned}$$

we can obtain a bound for $\mathbb{P}[\ B]$ as

$$\mathbb{P}\left[\ \|\mathbf{\Omega}_2\|_2^2 \geq \ell\left(1 + C_u^2 C_{\ell,\delta}\right)\right] \leq \delta/2.$$

2. Bound for $\|\mathbf{\Omega}_1^\dagger\|_2^2$ The steps are similar and we again use the matrix Chernoff concentration inequality. Consider $\mathbf{\Omega}_1\mathbf{\Omega}_1^* \in \mathbb{C}^{k\times k}$, and as before, write this matrix as the sum of rank-1 matrices to obtain

$$\mathbb{E}\left[\ \mathbf{\Omega}_1\mathbf{\Omega}_1^*\right] = \sum_{i=1}^{\ell}\mathbf{U}_1^*\mathbb{E}\left[\ \boldsymbol{\omega}_i\boldsymbol{\omega}_i^*\right]\mathbf{U}_1 = \ell\mathbf{I}_k,$$

and $\mu_{\min}(\mathbf{\Omega}_1\mathbf{\Omega}_1^*) = \ell$. Each summand in the above decomposition of $\mathbf{\Omega}_1\mathbf{\Omega}_1^*$ has one nonzero eigenvalue $\boldsymbol{\omega}_i^*\mathbf{U}_1\mathbf{U}_1^*\boldsymbol{\omega}_i$. Following the same strategy as in Step 1, we define $L_1 \equiv \max_{i=1,\ldots,\ell}\|\mathbf{U}_1^*\boldsymbol{\omega}_i\|_2^2$ and apply Lemma 8 to obtain

$$\mathbb{P}\left[\max_{i=1,\cdots,\ell}\|\mathbf{U}_1^*\boldsymbol{\omega}_i\|_2 \geq \sqrt{k}+t\right] \leq \ell e^{-t^2/8} \leq n e^{-t^2/8}.$$

Take $t = \sqrt{8\log(4n/\delta)}$ to obtain

$$\mathbb{P}\left[\ L_1 \geq C_l^2\right] \leq \delta/4, \qquad C_l \equiv \sqrt{k} + \sqrt{8\log\left(\frac{4n}{\delta}\right)}. \tag{26}$$

A straightforward application of the Chernoff bound in Theorem 11 gives us

$$\mathbb{P}\left[\lambda_{\min}(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^*) \leq (1-\rho)\ell\right] \leq kg(-\rho)^{\frac{\ell}{L_1}}.$$

Next, observe that $-\log g(-\rho)$ has the Taylor series expansion in the region $0<\rho<1$

$$-\log g(-\rho) = \rho + (1-\rho)\log(1-\rho) = \frac{\rho^2}{2} + \frac{\rho^3}{6} + \frac{\rho^4}{12} + \cdots$$

so that $-\log g(-\rho) \geq \rho^2/2$ for $0 < \rho < 1$ or $g(-\rho) \leq e^{-\rho^2/2}$. This gives us

$$\mathbb{P}\left[\|\boldsymbol{\Omega}_1^\dagger\|_2^2 \geq \frac{1}{(1-\rho)\ell}\right] \leq k\exp\left(-\frac{\rho^2\ell}{2L_1}\right), \tag{27}$$

where we have used $\lambda_{\min}(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^*) = 1/\|\boldsymbol{\Omega}_1^\dagger\|_2^2$ assuming $\mathsf{rank}(\boldsymbol{\Omega}_1) = k$.

With the number of samples as defined in Theorem 3

$$\ell \geq 2\rho^{-2}C_l^2\log\left(\frac{4k}{\delta}\right),$$

the Chernoff bound (27) becomes

$$\mathbb{P}\left[\|\boldsymbol{\Omega}_1^\dagger\|_2^2 \geq \frac{1}{(1-\rho)\ell}\right] \leq k\left(\frac{\delta}{4k}\right)^{C_l^2/L_1}.$$

Define the events

$$C = \left\{\boldsymbol{\Omega}_1 \mid \|\boldsymbol{\Omega}_1^\dagger\|_2^2 \geq \frac{1}{(1-\rho)\ell}\right\}, \qquad D = \{\boldsymbol{\Omega}_1 \mid L_1 < C_\ell^2\}.$$

Note that $\mathbb{P}\left[D^c\right] \leq \delta/4$ from (26). Then since the exponent is strictly greater than 1, we have

$$\mathbb{P}[C \mid D] \leq k\left(\frac{\delta}{4k}\right)^{C_l^2/L_1} \leq \delta/4.$$

Using the conditioning argument as before gives $\mathbb{P}[C] \leq \delta/2$.

3. Combining bounds Define the event

$$E = \left\{\boldsymbol{\Omega} \mid \|\boldsymbol{\Omega}_1^\dagger\|_2^2 \geq \frac{1}{(1-\rho)\ell}\right\}, \qquad F = \left\{\boldsymbol{\Omega} \mid \|\boldsymbol{\Omega}_2\|_2^2 \geq (1 + C_{\ell,\delta}C_u^2)\ell\right\},$$

where $C_{\ell,\delta}$ is defined in Step 1, $\mathbb{P}[E] \leq \delta/2$ and from Step 2, $\mathbb{P}[F] \leq \delta/2$. It can be verified that

$$\left\{ \boldsymbol{\Omega} \mid \|\boldsymbol{\Omega}_2\|_2^2 \|\boldsymbol{\Omega}_1^\dagger\|_2^2 \geq \frac{1}{1-\rho}(1 + C_{\ell,\delta} C_u^2) \right\} \subseteq E \cup F,$$

and therefore, we can use the union bound

$$\mathbb{P}\left[\|\boldsymbol{\Omega}_2\|_2^2 \|\boldsymbol{\Omega}_1^\dagger\|_2^2 \geq \frac{1}{1-\rho}(1 + C_{\ell,\delta} C_u^2) \right] \leq \mathbb{P}[E] + \mathbb{P}[F] \leq \delta.$$

Plugging in the value of $C_{\ell,\delta}$ and C_u^2 from Step 1 gives the desired result. □

References

1. Akçelik, V., Biros, G., Draganescu, A., Ghattas, O., Hill, J., Van Bloemen Waanders, B.: Dynamic data-driven inversion for terascale simulations: real-time identification of airborne contaminants. In: Supercomputing, 2005. Proceedings of the ACM/IEEE SC 2005 Conference, pp. 43–43 (2005)
2. Alexanderian, A., Gloor, P.J., Ghattas, O.: On Bayesian *A*-and *D*-optimal experimental designs in infinite dimensions. Bayesian Anal. **11**(3), 671–695 (2016)
3. Alexanderian, A., Petra, N., Stadler, G., Ghattas, O.: A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification. SIAM J. Sci. Comput. **36**(5), A2122–A2148 (2014)
4. Anitescu, M., Chen, J., Wang, L.: A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem. SIAM J. Sci. Comput. **34**(1), A240–A262 (2012)
5. Atkinson, A.C., Donev, A.N.: Optimum Experimental Designs. Oxford University Press, Oxford (1992)
6. Avron, H., Toledo, S.: Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. J. ACM **58**(2), Art. 8, 17 (2011)
7. Bai, Z., Fahey, M., Golub, G.: Some large-scale matrix computation problems. J. Comput. Appl. Math. **74**(1), 71–89 (1996)
8. Bai, Z., Golub, G.H.: Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices. Ann. Numer. Math. **4**(1-4), 29–38 (1997). The heritage of P.L. Chebyshev: a Festschrift in honor of the 70th birthday of T.J. Rivlin
9. Barry, R.P., Pace, R.K.: Monte Carlo estimates of the log determinant of large sparse matrices. Linear Algebra Appl. **289**(1-3), 41–54 (1999). Linear algebra and statistics (Istanbul, 1997)
10. Boutsidis, C., Drineas, P., Kambadur, P., Zouzias, A.: A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. arXiv:1503.00374 (2015)
11. Chaloner, K., Verdinelli, I.: Bayesian experimental design: a review. Stat. Sci. **10**(3), 273–304 (1995)
12. Chen, J., Anitescu, M., Saad, Y.: Computing $f(A)b$ via least squares polynomial approximations. SIAM J. Sci. Comput. **33**(1), 195–222 (2011)
13. Flath, P.H., Wilcox, L.C., Akçelik, V., Hill, J., van Bloemen Waanders, B., Ghattas, O.: Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations. SIAM J. Sci. Comput. **33**(1), 407–432 (2011)
14. Gittens, A., Mahoney, M.W.: Revisiting the Nystrom method for improved large-scale machine learning. arXiv:1303.1849 (2013)
15. Golub, G.H., Von Matt, U.: Generalized cross-validation for large-scale problems. J. Comput. Graph. Stat. **6**(1), 1–34 (1997)
16. Gu, M.: Subspace iteration randomization and singular value problems. SIAM J. Sci. Comput. **37**(3), A1139–A1173 (2015)

17. Haber, E., Horesh, L., Tenorio, L.: Numerical methods for experimental design of large-scale linear ill-posed inverse problems. Inverse Probl. **24**(5), 055012–055017 (2008)
18. Haber, E., Magnant, Z., Lucero, C., Tenorio, L.: Numerical methods for *A*-optimal designs with a sparsity constraint for ill-posed inverse problems. Comput. Optim. Appl. **52**, 293–314 (2012)
19. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011)
20. Han, I., Malioutov, D., Shin, J.: Large-scale log-determinant computation through stochastic Chebyshev expansions. arXiv:1503.06394 (2015)
21. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002)
22. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
23. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge (2013)
24. Hutchinson, M.F.: A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. Commun. Stat. Simul. Comput. **18**(3), 1059–1076 (1989)
25. Ledoux, M.: On Talagrand's deviation inequalities for product measures. ESAIM Probab. Statist. **1**, 63–87 (1995/1997)
26. Liberty, E., Woolfe, F., Martinsson, P.G., Rokhlin, V., Tygert, M.: Randomized algorithms for the low-rank approximation of matrices. Proc. Natl. Acad. Sci. USA **104**(51), 20167–20172 (2007)
27. Lin, L.: Randomized estimation of spectral densities of large matrices made accurate. Numer. Math. 1–31 (2016). doi:10.1007/s00211-016-0837-7
28. Mahoney, M.W.: Randomized Algorithms for Matrices and Data. Now Publishers Inc, Hanover (2011)
29. Martinsson, P.G., Rokhlin, V., Tygert, M.: A randomized algorithm for the decomposition of matrices. Appl. Comput. Harmon. Anal. **30**(1), 47–68 (2011)
30. Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, Cambridge (2005)
31. Ouellette, D.V.: Schur complements and statistics. Linear Algebra Appl. **36**, 187–295 (1981)
32. Pace, R.K., LeSage, J.P.: Chebyshev approximation of log-determinants of spatial weight matrices. Comput. Stat. Data Anal. **45**(2), 179–196 (2004)
33. Parlett, B.N.: The Symmetric Eigenvalue Problem. Prentice Hall Inc, Englewood Cliffs (1980)
34. Petra, N., Stadler, G.: Model variational inverse problems governed by partial differential equations. Tech. Rep. 11-05, The Institute for Computational Engineering and Sciences, The University of Texas at Austin (2011)
35. Roosta-Khorasani, F., Ascher, U.: Improved bounds on sample size for implicit matrix trace estimators. Found. Comput. Math. **15**(5), 1187–1212 (2015)
36. Rudelson, M., Vershynin, R.: Smallest singular value of a random rectangular matrix. Commun. Pure Appl. Math. **62**(12), 1707–1739 (2009)
37. Saad, Y.: Numerical methods for large eigenvalue problems. In: Classics in Applied Mathematics, vol. 66. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2011). Revised edition of the 1992 original
38. Saibaba, A.K., Kitanidis, P.K.: Fast computation of uncertainty quantification measures in the geostatistical approach to solve inverse problems. Adv. Water Resour. **82**, 124–138 (2015)
39. Sorensen, D.C., Embree, M.: A DEIM induced CUR factorization. SIAM J. Sci. Comput. **38**(3), A1454–A1482 (2016)
40. Stirling, D.S.G.: Mathematical Analysis and Proof, 2nd edn. Horwood Publishing Limited, Chichester (2009)
41. Tang, J.M., Saad, Y.: A probing method for computing the diagonal of a matrix inverse. Numer. Linear Algebra Appl. **19**(3), 485–501 (2012)
42. Tropp, J.A.: Improved analysis of the subsampled randomized Hadamard transform. Adv. Adapt. Data Anal. **3**(1–2), 115–126 (2011)
43. Tropp, J.A.: An introduction to matrix concentration inequalities. Found. Trends Mach. Learn. **8**(1–2), 1–230 (2015)
44. Uciński, D.: Optimal Measurement Methods for Distributed Parameter System Identification. CRC Press, Boca Raton (2005)
45. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y.C., Kutyniok, G. (eds.) Compressed Sensing, pp. 210–268. Cambridge University Press, Cambridge (2012)
46. Wahba, G.: Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal. **14**(4), 651–667 (1977)

47. Wahba, G.: Spline Models for Observational Data, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1990)
48. Zhang, Y., Leithead, W.E., Leith, D.J., Walshe, L.: Log-det approximation based on uniformly distributed seeds and its application to Gaussian process regression. J. Comput. Appl. Math. **220**(1–2), 198–214 (2008)