ARTICLE IN PRESS

Appl. Numer. Math. xxx (xxxx) xxx

Contents lists available at ScienceDirect

Applied Numerical Mathematics

journal homepage: www.elsevier.com/locate/apnum

Research Paper

Efficient estimates for matrix-inverse quadratic forms

Emmanouil Bizas [a], Marilena Mitrouli [a], Ondřej Turek [b,*]

[a] Department of Mathematics, National and Kapodistrian University of Athens, Panepistimiopolis, Athens, 157 84, Greece
[b] Department of Mathematics, University of Ostrava, Ostrava, 701 03, Czech Republic

ARTICLE INFO

Keywords:
Inverse matrix
Quadratic form
Approximation
Computational complexity
Heuristic technique
Condition number

ABSTRACT

In this paper we present two approaches for estimating matrix-inverse quadratic forms $x^T A^{-1} x$, where A is a symmetric positive definite matrix of order n, and $x \in \mathbb{R}^n$. Using the first, analytic approach, we establish two families of estimates which are convenient for matrices with small condition number. Based on the second, heuristic approach, we derive two families of estimates which are suitable for matrices when vector x is close enough to an eigenvector. The low complexity and stability of the estimates is proved. Several numerical results illustrating the effectiveness of the methods are presented.

1. Introduction

We consider the quadratic form $x^T A^{-1} x$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite (spd) matrix of order n, and $x \in \mathbb{R}^n$. The purpose of this work is to establish a direct heuristic approach for the estimation of this quadratic form avoiding the explicit computation of the matrix inverse, which may be prohibitive when the matrix A is large.

Once the quadratic form $x^T A^{-1} x$ is estimated, the bilinear form $x^T A^{-1} y$, where $x, y \in \mathbb{R}^n$, can be estimated too by employing the polarization identity

$$x^T A^{-1} y = \frac{1}{4}(w^T A^{-1} w - v^T A^{-1} v)$$

where $w = x + y$ and $v = x - y$.

Quadratic forms of the type $x^T A^{-q} x,\ q = 1, 2$ arise in many applications for a suitably selected vector $x \in \mathbb{R}^n$. Specifically, in Statistics and Uncertainty Quantification, they are required for approximating the inverse of covariance matrices [3,14]. In Network Analysis, they are useful for the specification of diagonal elements of the adjacency matrix [5]. In Numerical Analysis, quadratic forms are generally applied for the specification of the regularization parameter in the Tikhonov regularization method for the solution of ill-posed problems [6].

The exact computation of the quadratic form $x^T A^{-1} x$ requires the direct evaluation of the matrix A^{-1} which results to $\mathcal{O}(n^3)$ required floating point operations. In practice, this exact computation can be replaced by an estimate that is cheaper and faster to evaluate. This important problem has attracted a lot of attention and several approaches have been proposed [10,9,12]. In [1] the authors propose simplified anti-Gauss quadrature rules, whereas in [2] multiple orthogonal polynomials are employed. Moreover in [4] the global block Lanczos method is adopted and in [7] the authors apply extrapolation techniques for the estimation of bilinear forms of Hermitian matrices.

* Corresponding author.
E-mail addresses: manolisdiog@gmail.com (E. Bizas), mmitroul@math.uoa.gr (M. Mitrouli), ondrej.turek@osu.cz (O. Turek).

https://doi.org/10.1016/j.apnum.2024.01.013
Received 28 October 2023; Received in revised form 12 January 2024; Accepted 17 January 2024

Available online 23 January 2024

In this work we will present two approaches to estimating the quadratic form $x^T A^{-1}x$. In the analytical approach, we will employ the Taylor expansion. We will propose certain easily evaluable expressions (of quadratic complexity) for which we demonstrate that they have the same leading terms as the original quadratic form. Those expressions appear to be efficient estimates of $x^T A^{-1}x$. In the heuristic approach, we will employ the Cauchy-Schwarz inequality, and express the quadratic form as a function of the so called index of proximity, which can be heuristically estimated. This method extends the approach initially proposed in [8]. We will derive two-parameter families of estimates which require only matrix-vector products for their computation. All the obtained formulae of estimates are stable and fast to compute thanks to their quadratic complexity. The accuracy of the estimates will be assessed by the index of proximity: the closer is its value to one, the better are achieved estimates. The challenge in the heuristic estimation is that the whole process can be guided from the original data. By checking the index of proximity of our initial data set, we can predict if the heuristic estimates are appropriate to use.

The paper is structured as follows: In Section 2, we use asymptotic methods to establish efficient estimates of the quadratic form $x^T A^{-1}x$ suitable primarily for matrices with small condition numbers. Section 3 describes the heuristic method and proposes two approximation schemes leading to the derivation of two families of estimates. In Section 4 we present a detailed backward error analysis of the estimates proving their stability and providing error bounds. Section 5 reports numerical examples that illustrate the performance of the proposed heuristic estimates. Finally, conclusions are drawn in Section 6.

Throughout the paper $\|\cdot\|$ is the 2-norm of a vector or matrix, $(\cdot,\cdot)$ is a bilinear form, the superscript T denotes the transpose, the vector δ_i stands for the ith column of the identity matrix, and $\mathbb{N}_{\geq 1} = \{1, 2, 3, \dots\}$. The symbol O is used as the "big O notation"; i.e., $f(t) = O(g(t))$ as $t \to a$ if there is a constant M such that $|f(t)| \leq M g(t)$ in a certain neighborhood of a.

2. Analytic approach

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and $x \in \mathbb{R}^n$, $x \neq 0$. Let $\lambda_1, \dots, \lambda_n$ denote the eigenvalues of A. Then for every $k \in \mathbb{R}$ we have

$$(x, A^k x) = \|x\|^2 \sum_{j=1}^{n} \lambda_j^k \alpha_j^2 , \tag{1}$$

where $(\alpha_1, \dots, \alpha_n)^T$ are the coordinates of the normalized vector $\frac{1}{\|x\|}x$ in the eigenbasis of A associated with the eigenvalues $\lambda_1, \dots, \lambda_n$; i.e.,

$$\sum_{j=1}^{n} \alpha_j^2 = 1. \tag{2}$$

In particular, equation (1) with the choice $k = -1$ acquires the form

$$(x, A^{-1}x) = \|x\|^2 \sum_{j=1}^{n} \lambda_j^{-1} \alpha_j^2 . \tag{3}$$

Let $\bar{\lambda} = \frac{1}{n}\sum_{j=1}^{n} \lambda_j$ be the arithmetic mean of the eigenvalues of A, and

$$\varepsilon_j = \lambda_j - \bar{\lambda}$$

for every $j = 1, \dots, n$. Then

$$|\varepsilon_j| \leq \lambda_{\max} - \lambda_{\min}$$

and

$$\bar{\lambda} \geq \lambda_{\min} ;$$

hence

$$\frac{\varepsilon_j}{\bar{\lambda}} \leq \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\min}} = \kappa - 1,$$

where κ is the condition number of A. Formula (1) gives

$$(x, A^{-1}x) = \|x\|^2 \sum_{j=1}^{n} \lambda_j^{-1}\alpha_j^2 = \|x\|^2 \sum_{j=1}^{n} (\bar{\lambda} + \varepsilon_j)^{-1}\alpha_j^2 = \|x\|^2 \sum_{j=1}^{n} \frac{1}{\bar{\lambda}} \left(1 + \frac{\varepsilon_j}{\bar{\lambda}}\right)^{-1} \alpha_j^2.$$

By Taylor's theorem, we have

$$(1+t)^{-1} = 1 - t + t^2 + \dots + (-1)^{\ell-1} t^{\ell} + R_j(t),$$

and the Lagrange form of the remainder leads to the estimate

$$|R_\ell(t)| \le \left|\frac{t}{1-t}\right|^{\ell+1}.$$

In particular, using (2) the choice $\ell = 2$ gives

$$\begin{aligned}(x, A^{-1}x) &= \|x\|^2 \sum_{j=1}^{n} \frac{1}{\bar\lambda}\left[1 - \frac{\varepsilon_j}{\bar\lambda} + \left(\frac{\varepsilon_j}{\bar\lambda}\right)^2 + R_2\left(\frac{\varepsilon_j}{\bar\lambda}\right)\right]\alpha_j^2 \\ &= \frac{\|x\|^2}{\bar\lambda}\left[1 - \sum_{j=1}^{n}\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2 + \sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2 + \underbrace{\sum_{j=1}^{n} R_2\left(\frac{\varepsilon_j}{\bar\lambda}\right)\alpha_j^2}_{r_2}\right].\end{aligned}$$

The term $r_2 = \sum_{j=1}^{n} R_2\left(\frac{\varepsilon_j}{\bar\lambda}\right)\alpha_j^2$ can be estimated as

$$|r_2| = \left|\sum_{j=1}^{n} R_2\left(\frac{\varepsilon_j}{\bar\lambda}\right)\alpha_j^2\right| \le \sum_{j=1}^{n}\left|\frac{\frac{\varepsilon_j}{\bar\lambda}}{1-\frac{\varepsilon_j}{\bar\lambda}}\right|^3 \alpha_j^2 \le \sum_{j=1}^{n}\left|\frac{\kappa-1}{1-(\kappa-1)}\right|^3\alpha_j^2 = \left|\frac{\kappa-1}{2-\kappa}\right|^3;$$

thus $r_2 = O\left((\kappa-1)^3\right)$ as $\kappa \to 1$.

The goal of the forthcoming section 2.1 is to derive estimates that coincide with the expansion (3) up to the second powers of $\frac{\varepsilon_j}{\bar\lambda}$, while their evaluation uses only inner products and matrix vector products.

2.1. Estimates of $(x, A^{-1}x)$ of the 2nd order

We search for an estimate of the form

$$(x, A^{-1}x) \approx (x,x)^{\ell_0} \cdot (x, Ax)^{\ell_1} \cdot (Ax, Ax)^{\ell_2}$$

for $\ell_0, \ell_1, \ell_2 \in \mathbb{Z}$. Obviously, the exponents must satisfy

$$\ell_0 + \ell_1 + \ell_2 = 1 \qquad \text{and} \qquad \ell_1 + 2\ell_2 = -1;$$

hence, by choosing ℓ_2, the other two exponents are fixed as $\ell_1 = -1 - 2\ell_2$ and $\ell_0 = 2 + \ell_2$.

The simplest choice $\ell_2 = 0$ gives the estimate $(x, A^{-1}x) \approx \frac{(x,x)^2}{(x,Ax)}$, but one can check that it does not have the required precision. Therefore, let us proceed to the natural next choice, $\ell_2 = 1$, corresponding to the estimate

$$(x, A^{-1}x) \approx (x,x)^3 \cdot (x,Ax)^{-3} \cdot (Ax,Ax)^1 = \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3}.$$

We get

$$\begin{aligned}\frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} &= \frac{\|x\|^8 \sum_{j=1}^{n}\lambda_j^2\alpha_j^2}{\left(\|x\|^2\sum_{j=1}^{n}\lambda_j\alpha_j^2\right)^3} = \|x\|^2 \frac{\sum_{j=1}^{n}(\bar\lambda+\varepsilon_j)^2\alpha_j^2}{\left(\sum_{j=1}^{n}(\bar\lambda+\varepsilon_j)\alpha_j^2\right)^3} \\ &= \frac{\|x\|^2}{\bar\lambda}\cdot\frac{\sum_{j=1}^{n}\left(1+\frac{\varepsilon_j}{\bar\lambda}\right)^2\alpha_j^2}{\left(\sum_{j=1}^{n}\left(1+\frac{\varepsilon_j}{\bar\lambda}\right)\alpha_j^2\right)^3} = \frac{\|x\|^2}{\bar\lambda}\cdot\frac{1 + 2\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2 + \sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2}{\left(1+\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^3}.\end{aligned} \tag{4}$$

As $\kappa \to 1$, we have

$$\left|\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right| \le \sum_{j=1}^{n}\left|\frac{\varepsilon_j}{\bar\lambda}\right|\alpha_j^2 \le \sum_{j=1}^{n}(\kappa-1)\alpha_j^2 = \kappa - 1 = O(\kappa-1);$$

$$\left|\sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2\right| = \sum_{j=1}^{n}\left|\frac{\varepsilon_j^2}{\bar\lambda^2}\right|\alpha_j^2 \le \sum_{j=1}^{n}(\kappa-1)^2\alpha_j^2 = (\kappa-1)^2 = O(\kappa-1)^2$$

Let us compare the expression (4) with $(x, A^{-1}x)$. A straightforward calculation (which can be carried out using a computer) gives the following result:

$$\frac{\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}}{(x,A^{-1}x)}=\frac{\frac{1+2\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2+\sum_{j=1}^n\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2}{\left(1+\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^3}}{1-\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2+\sum_{j=1}^n\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2+r_2}$$

$$=1+2\left(\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^3-3\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\cdot\sum_{j=1}^n\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2+r_2+O\left((\kappa-1)^4\right)$$

$$=1+O\left((\kappa-1)^3\right)+O\left((\kappa-1)^4\right)$$

$$=1+O\left((\kappa-1)^3\right)\quad\text{as}\quad\kappa\to1. \tag{5}$$

To sum up, we have obtained an estimate

$$(x,A^{-1}x)\approx\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3} \tag{6}$$

with the relative error satisfying

$$Rel=\frac{\left|\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}-(x,A^{-1}x)\right|}{|(x,A^{-1}x)|}=O\left((\kappa-1)^3\right)\quad\text{as}\quad\kappa\to1.$$

Let us develop another estimate coinciding with $(x,A^{-1}x)$ up to the second powers of $\frac{\varepsilon_j}{\bar\lambda}$. It will be constructed as a suitably chosen combination of $\frac{\|x\|^4}{(x,Ax)}$ and $\frac{(x,Ax)^3}{\|Ax\|^4}$.

We start from the expansion

$$\frac{\|x\|^4}{(x,Ax)}=\frac{\|x\|^4}{\|x\|^2\sum_{j=1}^n\lambda_j\alpha_j^2}=\frac{\|x\|^2}{\sum_{j=1}^n(\bar\lambda+\varepsilon_j)\alpha_j^2}=\frac{\|x\|^2}{\bar\lambda+\sum_{j=1}^n\varepsilon_j\alpha_j^2}=\frac{\|x\|^2}{\bar\lambda}\left(1+\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^{-1}$$

$$=\frac{\|x\|^2}{\bar\lambda}\left[1-\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2+\left(\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^2+R_2\left(\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)\right].$$

Since

$$\left|R_2\left(\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)\right|\le\left|\frac{\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2}{1-\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2}\right|^3\le\left|\frac{\sum_{j=1}^n(\kappa-1)\alpha_j^2}{1-\sum_{j=1}^n(1-\kappa)\alpha_j^2}\right|^3$$
$$=\left|\frac{\kappa-1}{1-(\kappa-1)}\right|^3=\left|\frac{\kappa-1}{2-\kappa}\right|^3,$$

we have

$$\frac{\|x\|^4}{(x,Ax)}=\frac{\|x\|^2}{\bar\lambda}\left[1-\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2+\left(\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^2+O\left((\kappa-1)^3\right)\right] \tag{7}$$

as $\kappa\to1$.

Similarly, the expansion

$$\frac{(x,Ax)^3}{\|Ax\|^4} = \frac{\left(\|x\|^2 \sum_{j=1}^{n} \lambda_j \alpha_j^2\right)^3}{\left(\|x\|^2 \sum_{j=1}^{n} \lambda_j^2 \alpha_j^2\right)^2} = \|x\|^2 \frac{\left(\sum_{j=1}^{n} (\bar{\lambda}+\varepsilon_j)\alpha_j^2\right)^3}{\left(\sum_{j=1}^{n} (\bar{\lambda}+\varepsilon_j)^2\alpha_j^2\right)^2}$$
$$= \frac{\|x\|^2}{\bar{\lambda}} \cdot \frac{\left(\sum_{j=1}^{n}\left(1+\frac{\varepsilon_j}{\bar{\lambda}}\right)\alpha_j^2\right)^3}{\left(\sum_{j=1}^{n}\left(1+\frac{\varepsilon_j}{\bar{\lambda}}\right)^2\alpha_j^2\right)^2} = \frac{\|x\|^2}{\bar{\lambda}} \cdot \frac{\left(1+\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2\right)^3}{\left(1+2\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2+\sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar{\lambda}^2}\alpha_j^2\right)^2}$$

leads to

$$\frac{(x,Ax)^3}{\|Ax\|^4} = \frac{\|x\|^2}{\bar{\lambda}} \cdot \left[1-\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2+3\left(\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2\right)^2-2\sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar{\lambda}^2}\alpha_j^2+O\left((\kappa-1)^3\right)\right]. \tag{8}$$

A suitable combination of expressions (7) and (8) gives

$$\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}$$
$$=\frac{3}{2}\cdot\frac{\|x\|^2}{\bar{\lambda}}\left[1-\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2+\left(\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2\right)^2+O\left((\kappa-1)^3\right)\right]$$
$$-\frac{1}{2}\cdot\frac{\|x\|^2}{\bar{\lambda}}\left[1-\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2+3\left(\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2\right)^2-2\sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar{\lambda}^2}\alpha_j^2+O\left((\kappa-1)^3\right)\right]$$
$$=\frac{\|x\|^2}{\bar{\lambda}}\left[1-\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2+\sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar{\lambda}^2}\alpha_j^2+O\left((\kappa-1)^3\right)\right].$$

Comparing this result with (3), we conclude that

$$\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}=(x,A^{-1}x)+\frac{\|x\|^2}{\bar{\lambda}}O\left((\kappa-1)^3\right)$$

as $\kappa\to 1$. A more precise calculation gives the following formula, which is close to the result obtained in equation (5):

$$\frac{\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}}{(x,A^{-1}x)}$$
$$=1+2\left(\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2\right)^3-3\sum_{j=1}^{n}\frac{\varepsilon_j}{\bar{\lambda}}\alpha_j^2\cdot\sum_{j=1}^{n}\frac{\varepsilon_j^2}{\bar{\lambda}^2}\alpha_j^2+r_2+O\left((\kappa-1)^4\right)$$
$$=1+O\left((\kappa-1)^3\right). \tag{9}$$

Therefore, the estimate

$$(x,A^{-1}x)\approx\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4} \tag{10}$$

has the relative error satisfying

$$Rel=\frac{\left|\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}-(x,A^{-1}x)\right|}{|(x,A^{-1}x)|}=O\left((\kappa-1)^3\right).$$

Observe that both estimates (6) and (10) derived above have the relative error $O\left((\kappa-1)^3\right)$ as $\kappa\to 1$. Consequently, they can be combined as follows, generating a parametric family of estimates

$$(x,A^{-1}x)\approx(1-p)\cdot\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}+p\cdot\left(\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}\right) \tag{11}$$

Table 1
Formulae of second-order estimates.

$est\_2(-1)$	$(x,A^{-1}x)\approx 2\cdot\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}-\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}+\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}$
$est\_2(0)$	$(x,A^{-1}x)\approx\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}$
$est\_2(\frac{1}{2})$	$(x,A^{-1}x)\approx\frac{1}{2}\cdot\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}+\frac{3}{4}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{4}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}$
$est\_2(1)$	$(x,A^{-1}x)\approx\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}$
$est\_2(2)$	$(x,A^{-1}x)\approx-\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}+3\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{(x,Ax)^3}{\|Ax\|^4}$

with a parameter $p\in\mathbb{R}$ such that all those estimates have the relative error of the same order

$$Rel=\frac{\left|(1-p)\cdot\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}+p\cdot\left(\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}\right)-(x,A^{-1}x)\right|}{|(x,A^{-1}x)|}=O\left((\kappa-1)^3\right).$$

In particular, the choice $p=0$ gives the estimate (6), and the choice $p=1$ corresponds to the estimate (10). Since for each p, these estimates coincide with the exact value of $(x,A^{-1}x)$ up to the $O((\kappa-1)^2)$-terms, we will denote these estimates as $est\_2(p)$.

Table 1 expresses some of the estimates that we can build.

2.2. Estimates of $(x,A^{-1}x)$ of the 3rd order

A detailed calculation shows that the estimates (11) satisfy

$$\begin{aligned}&\frac{(1-p)\cdot\frac{\|x\|^6\cdot\|Ax\|^2}{(x,Ax)^3}+p\cdot\left(\frac{3}{2}\cdot\frac{\|x\|^4}{(x,Ax)}-\frac{1}{2}\cdot\frac{(x,Ax)^3}{\|Ax\|^4}\right)-(x,A^{-1}x)}{(x,A^{-1}x)}\\&=1+\underbrace{2\left(\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^3-3\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\cdot\sum_{j=1}^n\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2+\sum_{j=1}^n\frac{\varepsilon_j^3}{\bar\lambda^3}\alpha_j^2}_{O((\kappa-1)^3)}+O\left((\kappa-1)^4\right).\end{aligned}\tag{12}$$

The goal of this section is to amend the estimates (11) with the view of eliminating the error terms $O\left((\kappa-1)^3\right)$. In this way, we will find a family of estimates of $(x,A^{-1}x)$ coinciding with $(x,A^{-1}x)$ up to the terms of order $(\kappa-1)^3$.

Let us begin from expanding $(x,A^{-1}x)$ up to the 3rd order terms:

$$\begin{aligned}(x,A^{-1}x)&=\|x\|^2\sum_{j=1}^n\lambda_j^{-1}\alpha_j^2=\frac{\|x\|^2}{\bar\lambda}\sum_{j=1}^n\left(1+\frac{\varepsilon_j}{\bar\lambda}\right)^{-1}\alpha_j^2\\&=\frac{\|x\|^2}{\bar\lambda}\left[1-\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2+\sum_{j=1}^n\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2-\sum_{j=1}^n\frac{\varepsilon_j^3}{\bar\lambda^3}\alpha_j^2+\underbrace{\sum_{j=1}^nR_3\left(\frac{\varepsilon_j}{\bar\lambda}\right)\alpha_j^2}_{r_3}\right].\end{aligned}$$

Using the Lagrange form of the remainder, we easily obtain

$$|r_3|\le\sum_{j=1}^n\left|\frac{\frac{\varepsilon_j}{\bar\lambda}}{1-\frac{\varepsilon_j}{\bar\lambda}}\right|^4\alpha_j^2\le\sum_{j=1}^n\left|\frac{\kappa-1}{1-(\kappa-1)}\right|^4\alpha_j^2=\left|\frac{\kappa-1}{2-\kappa}\right|^4=O\left((\kappa-1)^4\right)$$

as $\kappa\to1$. Now let us focus on the quantity $\frac{\|x\|^4(Ax,A^2x)}{\|Ax\|^4}$. (As we will show below, combining (11) with this quantity allows to achieve a 3rd order precision.) Performing an expansion in the same manner as in the previous section, one gets

$$\begin{aligned}\frac{\frac{\|x\|^4(Ax,A^2x)}{\|Ax\|^4}}{(x,A^{-1}x)}&=1+4\left(\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\right)^3-6\sum_{j=1}^n\frac{\varepsilon_j}{\bar\lambda}\alpha_j^2\cdot\sum_{j=1}^n\frac{\varepsilon_j^2}{\bar\lambda^2}\alpha_j^2+2\sum_{j=1}^n\frac{\varepsilon_j^3}{\bar\lambda^3}\alpha_j^2\\&\quad+O\left((\kappa-1)^4\right)\quad\text{as}\quad\kappa\to1.\end{aligned}\tag{13}$$

Comparing (13) with (11), we observe that

Table 2
Formulae of third-order estimates.

*est*_3(−1)	$(x, A^{-1}x) \approx 4 \cdot \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} - 3 \cdot \frac{\|x\|^4}{(x,Ax)} + \frac{(x,Ax)^3}{\|Ax\|^4} - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4}$
*est*_3(0)	$(x, A^{-1}x) \approx 2 \cdot \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4}$
*est*_3($\frac{1}{2}$)	$(x, A^{-1}x) \approx \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} + \frac{3}{2} \cdot \frac{\|x\|^4}{(x,Ax)} - \frac{1}{2} \cdot \frac{(x,Ax)^3}{\|Ax\|^4} - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4}$
*est*_3(1)	$(x, A^{-1}x) \approx 3 \cdot \frac{\|x\|^4}{(x,Ax)} - \frac{(x,Ax)^3}{\|Ax\|^4} - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4}$
*est*_3(2)	$(x, A^{-1}x) \approx -2 \cdot \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} + 6 \cdot \frac{\|x\|^4}{(x,Ax)} - 2 \cdot \frac{(x,Ax)^3}{\|Ax\|^4} - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4}$

$$\frac{2 \cdot \left[(1-p) \cdot \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} + p \cdot \left(\frac{3}{2} \cdot \frac{\|x\|^4}{(x,Ax)} - \frac{1}{2} \cdot \frac{(x,Ax)^3}{\|Ax\|^4}\right)\right] - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4}}{(x, A^{-1}x)}$$
$$= 1 + O\left((\kappa - 1)^4\right).$$

Therefore,

$$2\left[(1-p)\frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} + p\left(\frac{3}{2} \cdot \frac{\|x\|^4}{(x,Ax)} - \frac{1}{2} \cdot \frac{(x,Ax)^3}{\|Ax\|^4}\right)\right] - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4}$$
$$= (x, A^{-1}x) \cdot \left[1 + O\left((\kappa-1)^4\right)\right]$$

as $\kappa \to 1$. Consequently, for each $p \in \mathbb{R}$, the estimate

$$(x, A^{-1}x) \approx 2(1-p) \cdot \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} + 3p \cdot \frac{\|x\|^4}{(x,Ax)} - p \cdot \frac{(x,Ax)^3}{\|Ax\|^4} - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4} \tag{14}$$

has the relative error satisfying

$$Rel = \frac{\left|2(1-p) \cdot \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3} + 3p \cdot \frac{\|x\|^4}{(x,Ax)} - p \cdot \frac{(x,Ax)^3}{\|Ax\|^4} - \frac{\|x\|^4 (Ax,A^2x)}{\|Ax\|^4} - (x, A^{-1}x)\right|}{|(x, A^{-1}x)|} = O\left((\kappa-1)^4\right).$$

Estimates (14) coincide for each p with the exact value of $(x, A^{-1}x)$ up to the terms of order $(\kappa - 1)^3$. We will denote these estimates as *est*_3(p).

Table 2 expresses some of the estimates that we can build.

3. Heuristic approach

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite (spd) matrix and $x \in \mathbb{R}^n$, $x \neq 0$. From the Cauchy-Schwarz inequality [11], we get:

$$|(x, Ax)| \leq \|x\| \cdot \|Ax\| \Rightarrow (x, Ax)^2 \leq \|x\|^2 \cdot \|Ax\|^2$$
$$\Rightarrow (x^T Ax)^2 \leq \|x\|^2 \cdot (x^T A^2 x)$$
$$\Rightarrow 1 \leq \frac{\|x\|^2 \cdot (x^T A^2 x)}{(x^T Ax)^2}$$

We introduce the following definition:

Definition 1. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and $x \in \mathbb{R}^n$, $x \neq 0$. The quantity

$$\rho(x) = \frac{\|x\|^2 (x^T A^2 x)}{(x^T Ax)^2}$$

is called *index of proximity*.

The index of proximity can be regarded as an easily computable measure of closeness of x to an eigenvector of A, in accordance with the following lemma:

Lemma 1. *Let $A \in \mathbb{R}^{n\times n}$ be a positive definite matrix and $x \in \mathbb{R}^n$ be a nonzero vector. Then $\rho(x) \geq 1$, and furthermore $\rho(x) = 1$ if and only if x is an eigenvector of A.*

Proof. The inequality $\rho(x) \geq 1$ follows from the Cauchy–Schwarz inequality. Furthermore, we have;

$$\rho(x) = 1 \quad \Leftrightarrow \quad \frac{\|x\|^2 \cdot \|Ax\|^2}{(x^T Ax)^2} = 1 \quad \Leftrightarrow \quad |x^T Ax| = \|x\| \cdot \|Ax\|.$$

Table 3
Multi-parameter formulae of estimates.

$hest\_m(1,0,0)$	$(x, A^{-1}x) \approx \frac{\|x\|^4}{x^T Ax}\rho(x)^0 = \frac{\|x\|^4}{x^T Ax}$
$hest\_m(1,0,-2)$	$(x, A^{-1}x) \approx \frac{\|x\|^4}{x^T Ax}\rho(x)^{-2} = \frac{(x,Ax)^3}{\|Ax\|^4}$
$hest\_m(2,1,\frac{1}{2},0,-1)$	$(x, A^{-1}x) \approx \frac{\|x\|^4}{x^T Ax}\rho(Ax)^{\frac{1}{2}}\rho(x)^{-1} = \frac{\|x\|^2\cdot\|A^2x\|\cdot(x,Ax)}{\|Ax\|\cdot(x,A^3x)}$
$hest\_m(3,1,\frac{1}{2},0,-1,1,-1)$	$(x, A^{-1}x) \approx \frac{\|x\|^4}{x^T Ax}\rho(Ax)^{-\frac{1}{2}}\rho(x)^{-1} = \frac{\|x\|^2\cdot(x,Ax)\cdot(x,A^3x)}{\|A^2x\|\cdot\|Ax\|^3}$
$hest\_m(4,1,\frac{1}{2},0,-1,1,-1,0,2)$	$(x, A^{-1}x) \approx \frac{\|x\|^4}{x^T Ax}\rho(Ax)^{-\frac{1}{2}}\rho(x) = \frac{\|x\|^6\cdot\|Ax\|\cdot(x,A^3x)}{\|A^2x\|\cdot(x,Ax)^3}$

By the Cauchy–Schwarz inequality, $|x^T Ax| = \|x\| \cdot \|Ax\|$ if and only if the vectors x and Ax are linearly dependent, i.e., $x = 0$ or $Ax = \lambda x$ for some $\lambda \in \mathbb{R}$. But the case $x = 0$ is excluded by assumption. Hence, $\rho(x) = 1$ iff x is an eigenvector of A. □

Furthermore, the index of proximity is characterized by the following properties:

Lemma 2. *Let $A \in \mathbb{R}^{n\times n}$ be a positive definite matrix and $x \in \mathbb{R}^n$ be a nonzero vector. The following statements are equivalent:*

- $\rho(x) = 1$*;*
- $\rho(A^{-\frac{1}{2}}x) = 1$*;*
- $\rho(A^k x) = 1$ *for any* $k \in \mathbb{R}$.

Proof. By Lemma 1, $\rho(x) = 1$ iff x is an eigenvector of A. This is further equivalent with $A^{-\frac{1}{2}}x$ and $A^k x$ being eigenvectors of A, and so with $\rho(A^{-\frac{1}{2}}x) = 1$ and $\rho(A^k x) = 1$ for any $k \in \mathbb{R}$. □

By the following proposition, a quadratic form can be expressed as a function of the index of proximity:

Proposition 1. *Let $A \in \mathbb{R}^{n\times n}$ be an spd matrix and $x \in \mathbb{R}^n$, $x \neq 0$. It holds that*

$$(x, A^{-1}x) = \frac{\|x\|^4}{x^T Ax}\rho(A^{-\frac{1}{2}}x)$$

Proof. From an explicit evaluation of $\rho(A^{-\frac{1}{2}}x)$, we get:

$$\rho(A^{-\frac{1}{2}}x) = \frac{\|A^{-\frac{1}{2}}x\|^2 \cdot \|A^{\frac{1}{2}}x\|^2}{((A^{-\frac{1}{2}}x)^T A^{\frac{1}{2}}x)^2} = \frac{((A^{-\frac{1}{2}}x)^T A^{-\frac{1}{2}}x)((A^{\frac{1}{2}}x)^T A^{\frac{1}{2}}x)}{(x^T x)^2} = \frac{(x^T A^{-1}x)(x^T Ax)}{\|x\|^4} \quad \square$$

Thus, by specifying an appropriate approximation of $\rho(A^{-\frac{1}{2}}x)$, it would be possible to derive an estimation formula of the quadratic form $(x, A^{-1}x)$.

Since the lowest value that the index of proximity can take is $\rho(x) = 1$, we propose the following approximation scheme, which becomes exact for vectors and matrices having index of proximity equal to 1:

Approximation Scheme I

Let $A \in \mathbb{R}^{n\times n}$ be an spd matrix and $x \in \mathbb{R}^n$, $x \neq 0$. Let $k_1, \ldots, k_l$ non negative integers and $p_1, \ldots, p_l \in \mathbb{R}$. The quantity $\rho(A^{-\frac{1}{2}}x)$ can be approximated by the expression

$$\rho(A^{-\frac{1}{2}}x) \approx \rho(A^{k_1}x)^{p_1} \cdots \rho(A^{k_l}x)^{p_l}$$

Since the integers k_i and p_i can be arbitrary selected, from the above scheme we can derive a *multi parameter family* of heuristic estimates which will be denoted as $hest\_m(l, k_1, p_1, \ldots, k_m, p_m)$.

Table 3 expresses some of the estimates that we can build.

Remark 1. Since the approximation scheme I becomes exact when $\rho(x) = 1$, the accuracy of the estimates *hest_m* will be satisfactory if the index of proximity of the selected vectors x is close to 1. In section 4, we will examine if given data sets will have index of proximity close to 1 and thus the heuristic estimates will be appropriate for their process.

Table 4
Single-parameter formulae of estimates.

$hest\_s(0)$	$(x, A^{-1}x) \approx \frac{\|x\|^6 \cdot \|Ax\|^2}{(x,Ax)^3}$
$hest\_s(1)$	$(x, A^{-1}x) \approx \frac{\|x\|^8 \cdot \|Ax\|^8}{(x,Ax)^6 \cdot (x,A^3x)}$
$hest\_s(-1)$	$(x, A^{-1}x) \approx \frac{\|x\|^4 \cdot (x,A^3x)}{\|Ax\|^4}$
$hest\_s(2)$	$(x, A^{-1}x) \approx \frac{\|x\|^{10} \cdot \|Ax\|^{14}}{(x,Ax)^9 \cdot (x,A^3x)^2}$
$hest\_s(-2)$	$(x, A^{-1}x) \approx \frac{\|x\|^2 \cdot (x,Ax)^3 \cdot (x,A^3x)^2}{\|Ax\|^{10}}$

Another approximation scheme can be derived on the basis of the previous one:

Approximation Scheme II

Let $A \in \mathbb{R}^{n\times n}$ be an spd matrix and $x \in \mathbb{R}^n$, $x \neq 0$. Let $\kappa \in \mathbb{Z}$. The quantity $\rho(A^{-\frac{1}{2}}x)$ can be approximated by the expression

$$\rho(A^{-\frac{1}{2}}x) \approx \rho(x)^{1+\kappa}\rho(A^{\frac{1}{2}}x)^{-\kappa}$$

According to the arbitrary selection of κ, we can derive a *single-parameter family* of heuristic estimates which will be denoted as $hest\_s(\kappa)$.

Table 4 shows some of the estimates that can be built. Notice that the estimate $hest\_s(0)$ coincides with $est\_2(0)$, established in Section 2.1.

Finally, let us observe that the analytic estimates are related to the heuristic estimates, as shown in the following lemma.

Lemma 3. *We have:*

$$est\_2(p) = (1-p)\cdot hest\_s(0) + \frac{3p}{2}\cdot hest\_m0 - \frac{p}{2}\cdot hest\_m1 \tag{15}$$

$$est\_3(p) = 2\cdot est\_2(p) - hest\_s(-1) \tag{16}$$

4. Complexity and stability of the estimates

Let $A \in \mathbb{R}^{n\times n}$ be an spd matrix and $x \in \mathbb{R}^n$, $x \neq 0$. Then it holds $(Ax, y) = (x, Ay)$. Using this property of spd matrices, all the derived formulae of estimates can be reformed to minimize the required number of matrix vector products. Tables 5, 6 outline the modified formulae of estimates which were adapted for a more efficient computation. The estimates $hest\_m0, hest\_m1, hest\_m2, hest\_m3, hest\_m4$ denote the estimates of the multi-parameter family derived from the approximation scheme I.

4.1. Complexity

We notice that the dominant operations involved in the computation of the estimates are inner products (ip's) and matrix vector products (mvp's). Let $A \in \mathbb{R}^{n\times n}$ be an spd matrix and $x \in \mathbb{R}^n$. The computation of an ip requires $\mathcal{O}(n)$ flops, whereas the mvp requires $\mathcal{O}(n^2)$ flops. We consider that multiplications and divisions between inner products, as well as square roots, cost $\mathcal{O}(1)$ time. Also, we assume that the vector x is not normalized when we count the required ip's in every formula. Table 7 and Table 8 show the computational complexity for each estimator.

Remark 2. It can be seen from Tables 7 and 8 that the evaluation of the estimates is computationally very cheap. The maximum complexity of all the formulas is 2 mvp's and 5 ip's. This guarantees their fast computation.

4.2. Stability

4.2.1. Stability of the heuristic estimates

We present the error analysis of the computation

$$fl(hest\_m1) = fl(\frac{(x,Ax)^3}{(Ax,Ax)^2})$$

We apply the backward error analysis approach as it was introduced by J. Wilkinson in [15], as well as the comprehension analysis in [16].

We adopt the following assumptions: In the inner product operation, the error will be equally splitted to the two vectors. In the matrix vector product operation, all the error will be accumulated to the matrix.

Table 5
Modified formulae of analytic estimates.

ESTIMATES	FORMULA
$est\_2(-1)$	$2\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3} - \frac{3}{2}\frac{(x,x)^2}{(x,Ax)} + \frac{1}{2}\frac{(x,Ax)^3}{(Ax,Ax)^2}$
$est\_2(0)$	$\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3}$
$est\_2(\frac{1}{2})$	$\frac{1}{2}\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3} + \frac{3}{4}\frac{(x,x)^2}{(x,Ax)} - \frac{1}{4}\frac{(x,Ax)^3}{(Ax,Ax)^2}$
$est\_2(1)$	$\frac{3}{2}\frac{(x,x)^2}{(x,Ax)} - \frac{1}{2}\frac{(x,Ax)^3}{(Ax,Ax)^2}$
$est\_2(2)$	$-\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3} + 3\frac{(x,x)^2}{(x,Ax)} - \frac{(x,Ax)^3}{(Ax,Ax)^2}$
– – –	– – –
$est\_3(-1)$	$4\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3} - 3\frac{(x,x)^2}{(x,Ax)} + \frac{(x,Ax)^3}{(Ax,Ax)^2} - \frac{(x,x)^2(Ax,A^2x)}{(Ax,Ax)^2}$
$est\_3(0)$	$2\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3} - \frac{(x,x)^2(Ax,A^2x)}{(Ax,Ax)^2}$
$est\_3(\frac{1}{2})$	$\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3} + \frac{3}{2}\frac{(x,x)^2}{(x,Ax)} - \frac{1}{2}\frac{(x,Ax)^3}{(Ax,Ax)^2} - \frac{(x,x)^2(Ax,A^2x)}{(Ax,Ax)^2}$
$est\_3(1)$	$3\frac{(x,x)^2}{(x,Ax)} - \frac{(x,Ax)^3}{(Ax,Ax)^2} - \frac{(x,x)^2(Ax,A^2x)}{(Ax,Ax)^2}$
$est\_3(2)$	$-2\frac{(x,x)^3(Ax,Ax)}{(x,Ax)^3} + 6\frac{(x,x)^2}{(x,Ax)} - 2\frac{(x,Ax)^3}{(Ax,Ax)^2} - \frac{(x,x)^2(Ax,A^2x)}{(Ax,Ax)^2}$

Table 6
Modified formulae of heuristic estimates.

ESTIMATES	FORMULA
$hest\_m0$	$\frac{(x,x)^2}{(x,Ax)}$
$hest\_m1$	$(x,Ax)^3 \cdot \frac{1}{(Ax,Ax)^2}$
$hest\_m2$	$(x,x)(x,Ax) \cdot \sqrt{\frac{(A^2x,A^2x)}{(Ax,Ax)}} \cdot \frac{1}{(Ax,A^2x)}$
$hest\_m3$	$(x,x)(x,Ax) \cdot \frac{1}{(Ax,Ax)\sqrt{(A^2x,A^2x)(Ax,Ax)}} \cdot (Ax,A^2x)$
$hest\_m4$	$\frac{(x,x)^3}{(x,Ax)^3} \cdot \sqrt{\frac{(Ax,Ax)}{(A^2x,A^2x)}} \cdot (Ax,A^2x)$
– – –	– – –
$hest\_s(0)$	$\frac{(x,x)^3}{(x,Ax)^3} \cdot (Ax,Ax)$
$hest\_s(1)$	$\frac{(x,x)^4}{(x,Ax)^6} \cdot (Ax,Ax)^4 \cdot \frac{1}{(Ax,A^2x)}$
$hest\_s(-1)$	$(x,x)^2 \cdot \frac{1}{(Ax,Ax)^2} \cdot (Ax,A^2x)$
$hest\_s(2)$	$\frac{(x,x)^5}{(x,Ax)^9} \cdot (Ax,Ax)^7 \cdot \frac{1}{(Ax,A^2x)^2}$
$hest\_s(-2)$	$(x,x)(x,Ax)^3 \cdot \frac{1}{(Ax,Ax)^5} \cdot (Ax,A^2x)^2$

Table 7
Computational complexity of analytic estimates for $p \in [0 : \frac{1}{4} : 1]$.

ESTIMATES	mvp's	ip's	COMPLEXITY
$est\_2(p)$	1	3	$n^2 + 3n + \mathcal{O}(1)$
– – –	– – –	– – –	– – –
$est\_3(p)$	2	4	$2n^2 + 4n + \mathcal{O}(1)$

Table 8
Computational complexity of heuristic estimates where $i_1 = 0, 1$, $i_2 = 2, 3, 4$, and $i_3 = -2, 1, 2$.

ESTIMATES	mvp's	ip's	COMPLEXITY
hest_m i_1	1	2	$n^2 + 2n + \mathcal{O}(1)$
hest_m i_2	2	5	$2n^2 + 5n + \mathcal{O}(1)$
− − −	− − −	− − −	− − −
hest_s(0)	1	3	$n^2 + 3n + \mathcal{O}(1)$
hest_s(−1)	2	3	$2n^2 + 3n + \mathcal{O}(1)$
hest_s(i_3)	2	4	$2n^2 + 4n + \mathcal{O}(1)$

We introduce two more notations. Let $x \in \mathbb{R}^n$ be a vector. Then the symbol $\tilde{x}^{(k)}$ stands for a vector that was obtained from x by k floating point operations; in other words, $\tilde{x}^{(k)}$ has been burdened k times with some error during floating point operations. Similarly, if $A \in \mathbb{R}^{n\times n}$ is a matrix, then $\tilde{A}^{(k)}$ stands for a matrix obtained by contaminating A with round off error; the superscript (k) denotes the number of times that the matrix has been burdened with rounding error during the matrix vector product floating point operations.

Theorem 1 *(Backward stability of* hest_m1*). Let $A \in \mathbb{R}^{n\times n}$ be an spd matrix and $x \in \mathbb{R}^n$, $x \neq 0$. Then the heuristic estimate* hest_m1 *computed using floating point arithmetic with round off error u_1, satisfies*

$$fl\Big(\frac{(x, Ax)^3}{(Ax, Ax)^2}\Big) = \frac{(\tilde{x}^{(1)}, \tilde{A}^{(1)}\tilde{x}^{(1)})^3}{(\tilde{A}^{(1)}\tilde{x}^{(2)}, \tilde{A}^{(1)}\tilde{x}^{(2)})^2}$$

with the following error-bound characteristics:

- $\tilde{x}^{(1)} = x(1 + \epsilon), \quad |\epsilon| \leq \frac{n+1}{2}u_1$
- $\tilde{x}^{(2)} = x(1 + \delta), \quad |\delta| \leq \frac{n+3}{2}u_1$
- $\tilde{A}^{(1)} = A + \delta A^{(1)}, \quad \|\delta A^{(1)}\|_\infty \leq n u_1 \cdot \|A\|_\infty$

Proof. We have that:

$$\begin{aligned} fl\Big(\frac{(x, Ax)^3}{(Ax, Ax)^2}\Big) &= fl\Big(\frac{fl((x, Ax)^3)}{fl((Ax, Ax)^2)}\Big) \\ &= \frac{fl((x, Ax)^3)}{fl((Ax, Ax)^2)} \cdot (1 + E_4) \\ &= \frac{(fl((x, Ax)))^3 \cdot (1 + E_1)(1 + E_2)}{(fl((Ax, Ax)))^2 \cdot (1 + \Delta_3)} \cdot (1 + E_4) \\ &= \frac{(fl((x, Ax)))^3}{(fl((Ax, Ax)))^2} \cdot (1 + E_1)(1 + E_2)\frac{1}{(1 + \Delta_3)}(1 + E_4) \end{aligned}$$

where $|E_1|, |E_2|, |\Delta_3|, |E_4| \leq u_1$. Rewriting $1 + E_i = \frac{1}{1+\Delta i}$ with $|\Delta_i| \leq u_1$, we get:

$$fl\Big(\frac{(x, Ax)^3}{(Ax, Ax)^2}\Big) = \frac{(fl((x, Ax)))^3}{(fl((Ax, Ax)))^2} \cdot \frac{1}{(1 + \Delta_1)(1 + \Delta_2)(1 + \Delta_3)(1 + \Delta_4)}$$

where:

- $fl((x, Ax)) = fl((x, fl(Ax))) = fl((x, \tilde{A}^{(1)}x)) = (\tilde{x}^{(1)}, \tilde{A}^{(1)}\tilde{x}^{(1)})$
- $fl((Ax, Ax)) = fl((fl(Ax), fl(Ax))) = fl((\tilde{A}^{(1)}x, \tilde{A}^{(1)}x)) = (\tilde{A}^{(1)}\tilde{x}^{(1)}, \tilde{A}^{(1)}\tilde{x}^{(1)})$

So:

$$\begin{aligned} fl\Big(\frac{(x, Ax)^3}{(Ax, Ax)^2}\Big) &= \frac{(\tilde{x}^{(1)}, \tilde{A}^{(1)}\tilde{x}^{(1)})^3}{(\tilde{A}^{(1)}\tilde{x}^{(1)}, \tilde{A}^{(1)}\tilde{x}^{(1)})^2} \cdot \frac{1}{(1 + \Delta_1)(1 + \Delta_2)(1 + \Delta_3)(1 + \Delta_4)} \\ &= \frac{(\tilde{x}^{(1)}, \tilde{A}^{(1)}\tilde{x}^{(1)})^3}{(\tilde{A}^{(1)}\tilde{x}^{(2)}, \tilde{A}^{(1)}\tilde{x}^{(2)})^2} \end{aligned}$$

Table 9
Estimation of a quadratic form for Heatflow matrix.

$n = 10000$, $u = 0.2$ $x = \delta_1 - 2\delta_2 + \delta_{20}$ $\kappa = 2.5986$ exact value $(x, A^{-1}x) = 3.1963$		
Analytic estimates	Relative error	Time (sec.)
$est\_2(-1)$	0.00331	0.07691
$est\_2(0)$	0.00423	0.08288
$est\_2(1/2)$	0.00469	0.10753
$est\_2(1)$	0.00515	0.09782
$est\_2(2)$	0.00607	0.09505
$est\_3(-1)$	0.00122	0.16973
$est\_3(0)$	0.00062	0.16961
$est\_3(1/2)$	0.00154	0.16715
$est\_3(1)$	0.00246	0.17344
$est\_3(2)$	0.00430	0.16456

where:

- $\tilde{x}^{(1)} = x(1+\epsilon), \quad |\epsilon| \leq \frac{n+1}{2}u_1$
- $\tilde{x}^{(2)} = x(1+\delta), \quad |\delta| \leq \frac{n+3}{2}u_1$
- $\tilde{A}^{(1)} = A + \delta A^{(1)}, \quad \|\delta A^{(1)}\|_\infty \leq nu_1 \cdot \|A\|_\infty$

Since the relative errors $Rel(\tilde{x}^{(1)}), Rel(\tilde{x}^{(2)}), Rel(\tilde{A}^{(1)})$ are of order $\mathcal{O}(n)$, we conclude that $hest\_m1$ is backward stable. □

Remark 3. The error analysis of all the other heuristic estimates can be performed in a similar way and thus all of them are stable.

4.2.2. Stability of the analytic estimates

The analytic estimates $est\_2(p)$ and $est\_3(p)$ are related to the heuristic estimates by the equations (15) and (16). In view of the proven backward stability of heuristic estimates, we have: If the following conditions are satisfied, then the possibility of subtracting approximate equal terms in the algebraic expressions (15) and (16) is eliminated.

$$est\_2(-1) \text{ stable} \Leftrightarrow hest\_m0 \not\approx \frac{4}{3} hest\_s(0)$$

$$\forall 0 \leq p \leq 1 \quad [est\_2(p) \text{ stable} \Leftrightarrow hest\_m0 \not\approx \frac{1}{3} hest\_m1]$$

$$est\_2(2) \text{ stable} \Leftrightarrow hest\_m0 \not\approx \frac{1}{3}(hest\_s(0) + hest\_m1)$$

and

$$\forall p \in \mathbb{R} \quad [est\_3(p) \text{ stable} \Leftrightarrow (est\_2(p) \text{ stable}) \text{ and } (est\_2(p) \not\approx \frac{1}{2} hest\_s(-1))]$$

The above conditions are realistic and hold for almost all the estimates. Thus, the stability of the analytic estimates can be guaranteed.

5. Numerical implementation

In this section, numerical experiments are presented. We will use the aforementioned methods for estimating quadratic forms, bilinear forms, and the matrix diagonal of various spd matrices which appear in applications, and compare the results in terms of accuracy and execution time. All computations were performed in JULIA [17] 64-bit on an Intel Core i5-1135G7 computer with 8 GB RAM. The exact values reported in this section are those given by the function *inv* of JULIA.

Example 1 *(The Heatflow matrix)*. Let us consider the *Heatflow* matrix. This matrix is symmetric, block tridiagonal (sparse) and comes from the discretization of the linear heat flow problem using the simplest implicit finite difference method. The coefficient matrix A of the resulted linear system of equations is a $m^2 \times m^2$ block tridiagonal matrix $A = tridiag(C, D, C)$, where D is a $m \times m$ tridiagonal matrix $D = tridiag(-u, 1+4u, -u)$, $C = diag([-u, -u, \dots, -u])$, $u = \Delta t/h^2$, Δt is the timestep, and h is the spacing interval. The *Heatflow* matrix A is symmetric positive definite for $u > 0$. We tested this matrix for $u = 0.2$. In Table 9 we see the results for estimating $(x, A^{-1}x)$, for $x = \delta_1 - 2\delta_2 + \delta_{20}$, while Table 10 shows the results for approximating the same quadratic form for $x = \delta_{256}$.

Table 10
Estimation of a diagonal element for Heatflow matrix.

$n = 10000$, $u = 0.2$ $x = \delta_{256}$ $\kappa = 2.5986$ exact value $(x, A^{-1}x) = 0.5865$		
Analytic estimates	Relative error	Time (sec.)
$est\_2(-1)$	0.00278	0.09160
$est\_2(0)$	0.00603	0.08726
$est\_2(1/2)$	0.00766	0.08865
$est\_2(1)$	0.00928	0.08928
$est\_2(2)$	0.01253	0.08793
$est\_3(-1)$	0.00686	0.17081
$est\_3(0)$	0.00036	0.17394
$est\_3(1/2)$	0.00289	0.17428
$est\_3(1)$	0.00614	0.16158
$est\_3(2)$	0.01264	0.16303

Table 11
Estimation of a quadratic form for KMS matrix.

$n = 5000$, $r = 0.2$ $x = rand(5000)$ $\kappa = 2.25$ exact value $(x, A^{-1}x) = 1290.47873$		
Analytic estimates	Relative error	Time (sec.)
$est\_2(-1)$	0.02071	0.01699
$est\_2(0)$	0.02235	0.01748
$est\_2(1/2)$	0.02318	0.01614
$est\_2(1)$	0.02400	0.01957
$est\_2(2)$	0.02565	0.01816
$est\_3(-1)$	0.00496	0.03949
$est\_3(0)$	0.00826	0.04456
$est\_3(1/2)$	0.00991	0.03424
$est\_3(1)$	0.01155	0.03722
$est\_3(2)$	0.01485	0.03571

Example 2 *(The Kac–Murdock–Szegö matrix).* We consider the Kac-Murdock-Szegö (KMS) matrix $A = [a_{ij}]$ which is symmetric positive-definite and Toeplitz. The elements of this matrix are $a_{ij} = r^{|i-j|}, 0 < r < 1$. The matrix is important within the context of digital signal processing because it characterizes first-order stationary Markov random signals, which can be decorrelated via the Karhunen–Loéve Transform.

We test this matrix for $r = 0.2$ and dimension $n = 3000, 5000$. In Table 11 we see the relative error, as well as the execution times, that were obtained by the est estimations of the quadratic form $(x, A^{-1}x)$ for random vector x with entries in the interval $(0, 1)$. Table 12 summarizes the relative errors, as well as the execution times, that were obtained by the $hest$ estimations of the bilinear form $(x, A^{-1}y)$ for random vectors x and y with entries in the interval $(0, 1)$.

Example 3 *(The covariance matrix).* Let us consider a symmetric positive definite matrix $A = [a_{ij}]$, whose entries are computed via a decaying positive definite covariance function, i.e. $a_{ii} = 1 + i^a$ and $a_{ij} = \dfrac{1}{|i-j|^b}$ for $i \neq j$, where $a, b \in \mathbb{R}$ and $b \geq 1$. The elements of this matrix show a decaying behavior away from the main diagonal, which simulates the decreasing correlation of high dimensional data samples in covariance matrix analysis.

We consider simulations of covariance matrices $A = [a_{ij}]$ for $a = 1$ and $b = 1$. Tables 13 and 14 report the mean relative error and the execution time in seconds for the estimation of the whole diagonal of the matrix A^{-1} of dimension $n = 10000$.

Example 4 *(The Poisson matrix).* Let us consider the Poisson matrix which is symmetric positive-definite, block tridiagonal (sparse) and arises from the five-point finite difference approximation of the Poisson equation in a unit square with an $m \times m$ mesh. The Poisson matrix is of the form $A = tridiag(-I_m, T, -I_m)$, where each block $T = tridiag(-1, 4, -1)$ has dimension m. We test this matrix for $m = 100$. In Table 15 we see the relative errors, as well as the execution times, that were obtained by the estimations of the quadratic form $(x, A^{-1}x)$ for vector $x = \sum_{i=1}^{10000}(-1)^i\delta_i$.

Example 5 *(Random matrix).* In this example we generate a random spd matrix $A = Q\Lambda Q^T$, by selecting random eigenvalues from the interval $(0, 1]$ (in order to set a diagonal matrix Λ) and extracting an orthogonal matrix Q from the QR-analysis of a random

Table 12
Estimation of a bilinear form for KMS matrix.

$n = 3000$, $r = 0.2$ $x, y = rand(3000)$ $\rho(x+y) = 1.02016$, $\rho(x-y) = 1.08303$ $\kappa = 2.25$ exact value $(x, A^{-1}y) = 505.6259540$		
Heuristic estimates	Relative error	Time (sec.)
*hest_m*0	0.02296	0.01442
*hest_m*1	0.03403	0.01417
*hest_m*2	0.03127	0.03490
*hest_m*3	0.02530	0.04508
*hest_m*4	0.01580	0.04381
hest_s(0)	0.01968	0.02342
hest_s(1)	0.01091	0.03950
hest_s(−1)	0.02837	0.04527
hest_s(2)	0.00207	0.04194
hest_s(−2)	0.03698	0.03760

Table 13
Analytic estimation of the whole diagonal for covariance matrices of order 10000.

$n = 10000, (a, b) = (1, 1)$ $\kappa = 7274.5915$		
Analytic estimates	Relative error	Time (sec.)
*est*_2(−1)	$\mathcal{O}(10^{-5})$	0.76354
*est*_2(0)	$\mathcal{O}(10^{-5})$	1.22598
*est*_2(1/2)	$\mathcal{O}(10^{-5})$	0.73289
*est*_2(1)	$\mathcal{O}(10^{-5})$	0.78936
*est*_2(2)	$\mathcal{O}(10^{-5})$	0.77095
*est*_3(−1)	$\mathcal{O}(10^{-5})$	21.2278
*est*_3(0)	$\mathcal{O}(10^{-5})$	21.6131
*est*_3(1/2)	$\mathcal{O}(10^{-5})$	22.5864
*est*_3(1)	0.00013	21.7935
*est*_3(2)	0.00018	21.6341
Mean time of $diag(inv(\cdot))$		35.7707

Table 14
Heuristic estimation of the whole diagonal for covariance matrices of order 10000.

$n = 10000, (a, b) = (1, 1)$ $\kappa = 7274.5915$		
Heuristic estimates	Mean Relative Error	Time (sec.)
*hest_m*0	0.00011	0.00457
*hest_m*1	0.00033	1.46202
*hest_m*2	0.00133	54.6736
*hest_m*3	0.00076	56.3275
*hest_m*4	0.00643	55.3931
hest_s(0)	$\mathcal{O}(10^{-5})$	2.13158
hest_s(1)	$\mathcal{O}(10^{-5})$	31.1815
hest_s(−1)	0.00014	26.6270
hest_s(2)	0.00012	27.9058
hest_s(−2)	0.00027	28.1999
Mean time of $diag(inv(\cdot))$		35.7707

Table 15
Estimation of a quadratic form for Poisson matrix.

$n = 10000$, $x = -\delta_1 + \delta_2 - \cdots + \delta_{10000}$, $\rho(x) = 1.00245$, $\kappa = 4133.6429$, exact value $(x, A^{-1}x) = 2513.7515245$		
Heuristic estimates	Relative error	Time (sec.)
*hest_m*0	0.00547	0.06886
*hest_m*1	0.01033	0.11449
*hest_m*2	0.00634	0.20192
*hest_m*3	0.00946	0.19586
*hest_m*4	0.00460	0.18515
hest_s(0)	0.00303	0.10505
hest_s(1)	0.00302	0.21474
hest_s(−1)	0.00305	0.19147
hest_s(2)	0.00300	0.18929
hest_s(−2)	0.00307	0.19829

Table 16
Analytic estimation of the whole diagonal for a random symmetric positive definite matrix of order 3000.

$n = 3000$, $\kappa = 2341.50$		
Analytic estimates	Relative error	Time (sec.)
*est*_2(−1)	0.61783	0.10575
*est*_2(0)	0.64796	0.08495
*est*_2(1/2)	0.66303	0.11080
*est*_2(1)	0.67810	0.08748
*est*_2(2)	0.70823	0.10614
*est*_3(−1)	0.53299	0.75705
*est*_3(0)	0.59326	0.74724
*est*_3(1/2)	0.62339	0.75358
*est*_3(1)	0.65352	0.73928
*est*_3(2)	0.71379	0.75002
Mean time of $diag(inv(\cdot))$		0.77821

Table 17
Heuristic estimation of the whole diagonal for a random symmetric positive definite matrix of order 3000.

$n = 3000$, $\kappa = 2341.50$		
Heuristic estimates	Mean Relative Error	Time (sec.)
*hest_m*0	0.73577	0.00012
*hest_m*1	0.85110	0.09751
*hest_m*2	0.79509	1.22826
*hest_m*3	0.80801	1.26786
*hest_m*4	0.65925	1.30938
hest_s(0)	0.64796	0.10991
hest_s(1)	0.58317	0.66231
hest_s(−1)	0.70267	0.65829
hest_s(2)	0.50643	0.68983
hest_s(−2)	0.74887	0.67497
Mean time of $diag(inv(\cdot))$		0.77821

real matrix. Tables 16 and 17 outline the mean relative error, as well as the execution time, for estimating the whole diagonal of a random spd matrix A being in moderate condition.

6. Conclusions

In this work we derived four parametric families of analytic and heuristic estimates for the approximation of the quadratic form $x^T A^{-1} x$, where A is a symmetric positive definite matrix of order n, and $x \in \mathbb{R}^n$. The proposed estimates are easily implemented and require only the computation of few mvp's, therefore the whole diagonal of a matrix inverse can be heuristically estimated in a few seconds. A detailed error analysis of the proposed formulae of estimates is presented and establishes their stability. All the numerical results supported the basic idea of machine learning, i.e., to learn from the data [13]. Indeed, all the tested data sets indicated that they can be processed either analytically or heuristically in a fast and stable way, producing very satisfactory estimations of the required quadratic forms. It is fascinating that all the matrices tested so far and for various choices of vectors x, y, have the index of proximity close to one, which indicates that the heuristic estimates are ideal for a stable and fast approximation of related quadratic forms. Furthermore, even though the construction of estimates est_2 and est_3 employs the assumption that the matrix A has a small condition number, numerical experiments show that the estimates work well for large matrices with a large condition number too. This is visible from Tables 14 and 15: the estimate $hest\_s(0)$, which coincides with $est\_2(0)$, gives good results even if κ goes far beyond 1000. Further research concerning the extension of the heuristic approach for estimating other quantities useful in applications is in process.

Acknowledgements

We would like to thank the reviewers for providing us with useful comments that helped us to improve the quality of the paper.

References

[1] H. Alqahtani, L. Reichel, Simplified anti-Gauss quadrature rules with applications in linear algebra, Numer. Algorithms 77 (2018) 577–602.
[2] H. Alqahtani, L. Reichel, Multiplied orthogonal polynomials applied to matrix function evaluation, BIT Numer. Math. 58 (2018) 835–849.
[3] C. Bekas, A. Curioni, I. Fedulova, Low-cost data uncertainty quantification, Concurr. Comput., Pract. Exp. 24 (8) (2012) 908–920.
[4] M. Bellalij, L. Reichel, G. Rodriguez, H. Sadok, Bounding matrix functionals via partial global Lanczos decomposition, Appl. Numer. Math. 94 (2015) 127–139.
[5] M. Benzi, C. Klymco, Communicability as a centrality measure, J. Complex Netw. 1 (2013) 124–149.
[6] C. Fenu, L. Reichel, G. Rodriguez, GCV for Tikhonov regularization via Golub-Kahan decomposition, Numer. Linear Algebra Appl. 23 (2016) 467–484.
[7] P. Fika, M. Mitrouli, Estimation of the bilinear form $y^* f(A)x$ for Hermitian matrices, Linear Algebra Appl. 502 (2016) 140–158.
[8] P. Fika, M. Mitrouli, O. Turek, On the estimation of $x^T A^{-1} x$ for symmetric matrices, Electron. J. Linear Algebra 37 (2021) 549–561.
[9] M. Mitrouli, A. Polychronou, P. Roupa, O. Turek, Estimating the quadratic form $x^T A^{-m} x$ for symmetric matrices: further progress and numerical computations, Mathematics 9 (2021) 1432.
[10] G.H. Golub, G. Meurant, Matrices, Moments and Quadrature with Applications, Princeton University Press, Princeton, 2010.
[11] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1985.
[12] L. Reichel, G. Rodriguez, T. Tang, New block quadrature rules for the approximation of matrix functions, Linear Algebra Appl. 502 (2016) 299–326.
[13] G. Strang, Linear Algebra and Learning from Data, Wellesley-Cambridge Press, Wellesley, 2019.
[14] J. Tang, Y. Saad, A probing method for computing the diagonal of a matrix inverse, Numer. Linear Algebra Appl. 19 (2012) 485–501.
[15] J.H. Wilkinson, Rounding Errors in Algebraic Processes, Prentice-Hall, N.J., 1963.
[16] P. Fika, M. Mitrouli, P. Roupa, D. Triantafyllou, The e-MoM approach for approximating matrix functionals, J. Comput. Appl. Math. 373 (2020) 112243.
[17] B. Lauwens, A.B. Downey, Think Julia: How to Think Like a Computer Scientist, O'Reilly Media, 2019.