Numerical Computations and the ω-Condition Number *†

Xuan Vinh Doan Henry Wolkowicz

July 4, 2011‡

Key words and phrases: Condition numbers, uniform condition number, preconditioning, roundoff-error, iterative methods

AMS subject classifications: 15A12, 65F35, 65F08, 65F10, 65G50

Contents

List of Algorithms

List of Tables

*Research Report CORR 2011-03

†This report is available at URL: orion.math.uwaterloo.ca/~hwolkowi/henry/reports/ABSTRACTS.html

‡Department of Combinatorics and Optimization, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

List of Figures

Abstract

We study the computation and properties of a new measure for the condition number of a positive definite matrix. This measure, the ratio of the arithmetic and geometric means of the eigenvalues, depends uniformly on all the eigenvalues of the matrix. Moreover, it can be evaluated easily and accurately. And, we see that: (i) it correlates better with the number of iterations in iterative methods compared to the standard condition number which depends only on the largest and smallest eigenvalues; (ii) it provides a criteria for obtaining optimal efficient preconditioners; and (iii) it presents a more average relative error measure compared to the worst case behaviour of the standard condition number.

1 Introduction

We study the properties of the following condition number measure, which can be computed accurately and efficiently and which depends uniformly on all the eigenvalues of the matrix $\boldsymbol{A}$:

$$\omega(\boldsymbol{A}) := \frac{\text{trace}(\boldsymbol{A})/n}{\det(\boldsymbol{A})^{\frac{1}{n}}} = \frac{\sum_{i=1}^{n} \lambda_i(\boldsymbol{A})/n}{\left(\prod_{i=1}^{n} \lambda_i(\boldsymbol{A})\right)^{\frac{1}{n}}}, \qquad \boldsymbol{A} \in \mathcal{S}^n_{++}, \tag{1.1}$$

where $\mathcal{S}^n_{++}$ is the cone of real symmetric positive definite matrices. The standard condition number $\kappa(\boldsymbol{A})$ depends only on the largest and smallest eigenvalues of $\boldsymbol{A}$. It is generally used as an indicator for whether the problem of solving the system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$ is well-conditioned (low condition number) or ill-conditioned (high condition number). In general, iterative algorithms used to solve the system $\boldsymbol{Ax} = \boldsymbol{b}$ require a large number of iterations to achieve a solution with high accuracy if the problem is ill-conditioned. Preconditioners are introduced to obtain a better conditioned problem. Currently, reducing $\kappa(\boldsymbol{A})$ condition number is the main aim for preconditioners. The condition number $\kappa(\boldsymbol{A})$ is also a measure of how much a solution $\boldsymbol{x}$ will change with respect to changes in the right-hand side $\boldsymbol{b}$.

Our goal in this paper is to establish the basic properties of the ω-condition number and to study whether it is a better indicator of whether the problem $\boldsymbol{Ax} = \boldsymbol{b}$ is well- or ill-conditioned. In addition, we present a procedure for efficiently calculating $\omega(\boldsymbol{A})$ and avoiding the difficulty of evaluating the determinant of large matrices.

1.1 Outline of Results

We continue in Section 2 with the basic properties of the condition number measure ω. We include Section 2.1 with Algorithm 2.1 that calculates ω efficiently and accurately. Section 3 presents results that relate the number of CG iterations for solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with the magnitudes of the two condition number measures. The better (inverse) correlation with ω is described in Items 1 to 4, page 7. Since the number of iterations depends on the number of distinct eigenvalues of A, we illustrate the seeming paradox that with the value of κ fixed, then both (relatively) large and small values of ω result in a low number of iterations. Section 4 compares how preconditioners arise from the two measures and their effectiveness. The expense of finding an optimal ω preconditioner is low compared to the optimal κ preconditioner. Section 5 discusses the role of the two condition numbers in error analysis. We provide the analytic formulation for the expected relative error under the assumption of independent normal distributions. We illustrate the known result that κ is a bound for the worst case of roundoff error, and show that ω provides a better estimate of the expected roundoff error. Concluding remarks are given in Section 6.

2 Basic Properties

The standard condition number of a nonsingular matrix $\boldsymbol{A}$ is defined as

$$\kappa(\boldsymbol{A}) = ||\boldsymbol{A}|| \left|\left|\boldsymbol{A}^{-1}\right|\right|,$$

where $||.||$ is a matrix norm. We let $\mathcal{S}^n$ denote the space of real $n \times n$ symmetric matrices equipped with the trace inner product; $\mathcal{S}^n_+$ denote the cone of positive semidefinite matrices; and $\mathcal{S}^n_{++}$ is the cone of positive definite matrices. If $||.||$ is the ℓ_2-norm, we then have: $\kappa(\boldsymbol{A}) = \dfrac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})}$, the ratio of largest and smallest singular values; and, if $\boldsymbol{A} \in \mathcal{S}^n_{++}$, then $\kappa(\boldsymbol{A}) = \dfrac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}$.

The ω-condition number has been studied in the context of scaling for quasi-Newton methods in [1]. Some basic properties of $\omega(\boldsymbol{A})$ with respect to $\kappa(\boldsymbol{A})$ are shown in the following.

Proposition 1 ([1]). *Let $A \succ 0$ be given. The measure $\omega(\boldsymbol{A})$ satisfies*

1. $1 \le \omega(\boldsymbol{A}) \le \kappa(\boldsymbol{A}) < \dfrac{(\kappa(\boldsymbol{A})+1)^2}{\kappa(\boldsymbol{A})} \le 4\omega^n(\boldsymbol{A})$,
   with equality in the first and second inequality if and only if $\boldsymbol{A}$ is a multiple of the identity and equality in the last inequality if and only if
   $$\lambda_2 = \cdots = \lambda_{n-1} = \frac{\lambda_1 + \lambda_n}{2};$$
2. $\omega(\alpha\boldsymbol{A}) = \omega(\boldsymbol{A})$, *for all $\alpha > 0$;*
3. *if $n = 2$, $\omega(\boldsymbol{A})$ is isotonic with $\kappa(\boldsymbol{A})$.* □

For a full rank matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ we let σ denote the vector of nonzero singular values. Then we can also define additional measures for general matrices.

$$\omega_2(\boldsymbol{A}) := \frac{\mathcal{A}(\sigma)}{\mathcal{G}(\sigma)} \ge \omega_c(\boldsymbol{A}) := \sqrt{\omega(\boldsymbol{A}^T\boldsymbol{A})} = \sqrt{\frac{\text{trace}(\boldsymbol{A}^T\boldsymbol{A})/n}{\det(\boldsymbol{A}^T\boldsymbol{A})^{\frac{1}{n}}}} = \sqrt{\frac{\mathcal{A}(\sigma^2)}{\mathcal{G}(\sigma^2)}},$$

where $\mathcal{A}, \mathcal{G}$, denote the arithmetic and geometric means respectively. These two condition numbers are for least squares solutions of overdetermined systems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, m > n$.

2.1 Numerical evaluation of ω

One issue with $\kappa(\boldsymbol{A})$ is how to estimate it efficiently when the size of matrix $\boldsymbol{A}$ is large since eigendecompositions can be expensive. A survey of estimates and, in particular, estimates using the ℓ_1-norm, are given in [4, 5]. Extensions to sparse matrices and block-oriented generalizations are given in [3, 6]. Results from these papers form the basis of the *condest* command in MATLAB; this illustrates the difficulty in accurately estimating $\kappa(A)$.

On the other hand, the measure $\omega(\boldsymbol{A})$ can be calculated using the trace and determinant function which do not require eigenvalue decompositions. However, for large n, the determinant is also numerically difficult to compute as it could easily result in an overflow $+\infty$ or 0 due to the limits of finite precision arithmetic. In order to overcome this problem, we use the following simple homogeneity property for the determinant: $\det(\alpha\boldsymbol{A}) = \alpha^n \det(\boldsymbol{A}), \forall \boldsymbol{A} \in \mathbb{R}^{n\times n}$. Thus we have:

$$\det(\boldsymbol{A})^{\frac{1}{n}} = \frac{\det(\alpha\boldsymbol{A})^{\frac{1}{n}}}{\alpha}, \quad \forall\, \alpha > 0.$$

Our main task is then to appropriately select $\alpha > 0$ such that $\det(\alpha\boldsymbol{A})$ is well-defined within the machine precision. Using the fact that $\det(\boldsymbol{A}) = |\prod_{i=1}^{n} u_{ii}|$, where u_{ii} are the diagonal elements of $\boldsymbol{U}$ in the LU decomposition of $\boldsymbol{A}$, we can simplify the calculation of $\det(\alpha\boldsymbol{A})$ by using the vector $\mathrm{diag}(\boldsymbol{U})$. Since $\log\det(\alpha\boldsymbol{A})$ is a nondecreasing function in $\alpha > 0$, we can use a simple binary search (not necessarily with high accuracy) to find α such that $\log\det(\alpha\boldsymbol{A}) \in (-\infty, +\infty)$, see Algorithm 2.1, page 5. This technique is used in our numerical tests throughout this paper. A MATLAB file for this evaluation is available with
URL: orion.math.uwaterloo.ca/~hwolkowi/henry/reports/ABSTRACTS.html#unifcondnumb.

3 Iterations for Krylov Subspace Methods and ω

When the matrix $\boldsymbol{A}$ is large and sparse, then efficient methods to solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ are based on Krylov subspaces. These methods perform repeated matrix-vector multiplications involving $\boldsymbol{A}$, e.g., Ipsen and Meyer [8]. The idea behind Krylov subspace methods is that after k iterations, the solution $\boldsymbol{x}_k$ lies in the Krylov space generated by a vector $\boldsymbol{c}$,

$$\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{c}) := \mathrm{span}(\boldsymbol{c}, \boldsymbol{A}\boldsymbol{c}, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{c}).$$

If $\boldsymbol{A} \in \mathcal{S}^n_{++}$, then the solution $\boldsymbol{x}$ belongs to the Krylov space $\mathcal{K}_d(\boldsymbol{A}, \boldsymbol{b})$, where $\boldsymbol{d}$ is the number of distinct eigenvalues of $\boldsymbol{A}$. This implies that, theoretically, a Krylov subspace method finds the solution $\boldsymbol{x}$ after d iterations. We would like to investigate how the ω-condition number relates to the number of iterations of Krylov subspace methods. We first show that a small number of distinct eigenvalues implies a small ω-condition number.

The ω-condition number can be defined as a function of the eigenvalue vector $\boldsymbol{\lambda}$ of the positive

Algorithm 2.1: Evaluation of $\omega_c(\boldsymbol{X}) = \sqrt{(\omega(\boldsymbol{A}))} = \sqrt{(\omega(\boldsymbol{X}^T\boldsymbol{X}))}$ with given tolerance, *tol*.

Input($\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{X}$ *full column rank*), tolerance *tol* ;
Set: $tA = \text{trace}(\boldsymbol{A})/n$, $[\boldsymbol{L}, \boldsymbol{U}] = \text{lu}(\boldsymbol{A})$, $dA = \text{prod}(\text{diag}(\boldsymbol{U}))$;
if *dA and* $\log(dA)$ *are finite* **then**
    Set: $\omega = \sqrt{\dfrac{tA}{(dA)^{\frac{1}{n}}}}$, **RETURN**
end if
if *dA is infinite* **then**
    Set: $aL = 0$, $aH = 1$;
else
    Set: $aL = 1$, $aH = 2$, $dA = \text{prod}(aH * \text{diag}(\boldsymbol{U}))$, $itercount = 1$;
    while *dA is finite &* $\log(dA)$ *& is infinite & itercount* $<$ *maxiter* **do**
        $aL = aH$, $aH = 2 * aH$, $dA = \text{prod}(aH * \text{diag}(\boldsymbol{U}))$, $index = index + 1$;
    end while
end if
while *(dA is infinite* | $\log(dA)$ *is infinite) &* $(aH - aL)/2 > tol$ **do**
    Set: $a = (aH + aL)/2$, $dA = \text{prod}(a * \text{diag}(\boldsymbol{U}))$;
    if *dA is finite &* $\log(dA)$ *is finite* **then**
        Set: $\omega = \sqrt{\dfrac{a * tA}{(dA)^{\frac{1}{n}}}}$, **RETURN**
    end if
    if *dA is infinite* **then**
        Set: $aH = a$;
    else
        Set: $aL = a$;
    end if
end while
if $(aH - aL)/2 > tol$ **then**
    Set: $\omega = \sqrt{\dfrac{a * tA}{(dA)^{\frac{1}{n}}}}$;
else
    Set: $\omega = \sqrt{tA * (aH + aL)/2}$;
end if
Output($\omega_c(\boldsymbol{X}) = \sqrt{(\omega(\boldsymbol{A}))} = \omega$) ;

definite matrix $\boldsymbol{A}$:

$$\omega(\boldsymbol{A}) = \omega(\boldsymbol{\lambda}) := \frac{\left(\sum_{i=1}^{n} \lambda_i / n\right)}{\left(\prod_{i=1}^{n} \lambda_i\right)^{\frac{1}{n}}}.$$

By abuse of notation, we allow ω to act on $\mathbb{R}^n_{++}$; and, for a subset $\mathcal{S} \subseteq \{1, \ldots, n\}$, we let $\boldsymbol{\lambda}_{\mathcal{S}} := (\lambda_i)_{i \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$.

Proposition 2. *Let the set $\emptyset \neq \mathcal{S} \subsetneq \{1, \ldots, n\}$ and the positive vector $\bar{\boldsymbol{\lambda}}_{\mathcal{S}} > 0$ be given. Then the optimal solution of*

$$\begin{array}{ll} \min & \omega(\boldsymbol{\lambda}) \\ \text{s.t.} & \lambda_i = \bar{\lambda}_i, \quad \forall\, i \in \mathcal{S}, \\ & \boldsymbol{\lambda} \in \mathbb{R}^n_{++}, \end{array} \tag{3.1}$$

is given by

$$\lambda_i = \begin{cases} \bar{\lambda}_i, & \text{if } i \in \mathcal{S}, \\ \dfrac{1}{n - |\mathcal{S}|} \displaystyle\sum_{i \in \mathcal{S}} \bar{\lambda}_i, & \text{otherwise.} \end{cases}$$

Proof. We can change the problem to the unconstrained minimization

$$\min \log\left(\sum_{i \in \mathcal{S}} \bar{\lambda}_i + \sum_{i \notin \mathcal{S}} \lambda_i\right) - \frac{1}{n} \log\left(\prod_{i \in \mathcal{S}} \bar{\lambda}_i\right)\left(\prod_{i \notin \mathcal{S}} \lambda_i\right).$$

Let $K^* = \displaystyle\sum_{i \in \mathcal{S}} \bar{\lambda}_i + \sum_{i \notin \mathcal{S}} \lambda_i^*$ at an optimal λ^*. Then stationarity implies that

$$\frac{1}{K^*} - \frac{1}{n\lambda_i^*} = 0, \; \forall\, i \notin \mathcal{S}.$$

This means that $(n - |\mathcal{S}|)\lambda_i = \displaystyle\sum_{i \in \mathcal{S}} \bar{\lambda}_i, \; \forall\, i \notin \mathcal{S}$. □

Proposition 2 shows that the ω-condition number is minimized when all free eigenvalues are equal, which means the number of distinct eigenvalues is reduced significantly. We emphasize the comparison with the standard condition number $\kappa(\boldsymbol{A})$.

Corollary 1. *Let $\mathcal{S} = \{1, n\}$ in Proposition 2. Then the optimal solution of* (3.1) *is given by*

$$\lambda_i^* = \frac{1}{n-2}\left(\bar{\lambda}_1 + \bar{\lambda}_n\right). \quad \square$$

Corollary 2. *Let $\{1, n\} \subseteq S$ in Proposition 2. And, consider the problem* (3.1) *with* min *replaced by* max *and with the added constraints that the optimal vector $\boldsymbol{\lambda}$ must maintain the ordering $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then the number of distinct numbers in the optimal solution $\boldsymbol{\lambda}^*$ is given by $|\mathcal{S}|$.*

Proof. The result follows immediately from the pseudoconvexity [1] of the function ω, i.e., the max must occur at an extreme point of the feasible set. □

[1] See Proposition 3 Item 1, below, for the statement on the pseudoconvexity of ω.

Corollary 1 indicates that the number of distinct eigenvalues reduces when the ω-condition number is minimized. It shows that the ω-condition number is likely positively correlated with the number of iterations of Krylov subspace methods for solving the system of linear equation $\boldsymbol{Ax} = \boldsymbol{b}$ when the standard condition number is not kept constant. In fact, we expect a better correlation for ω compared to κ since ω depends on all the eigenvalues. Surprisingly, Corollary 2 indicates that when $\kappa(\boldsymbol{A})$ remains the same, then we get a reduction in the number of iterations when ω is *maximized*.

We now construct numerical results to test the relationship between the number of iterations and the two condition numbers.

Random Procedure 1. *Unless specified otherwise, in the following tests for Figures 1 to 4: we generate random matrices with a particular number of distinct eigenvalues; we then solve the system of linear equations with random right-hand sides using the preconditioned conjugate gradient method,* pcg*, in MATLAB. The size of the matrices is* $n = 5000$ *with the number of distinct eigenvalues chosen randomly in the interval* $[1, 500]$*. The tolerance for pcg is set to* 10^{-10}*. We used* 5000 *random instances.*

1. The scatter plot in Figure 1 shows, as expected, that the number of distinct eigenvalues (on the horizontal axis) is positively correlated with the number of Krylov iterations to solve the linear system (on the vertical axis).

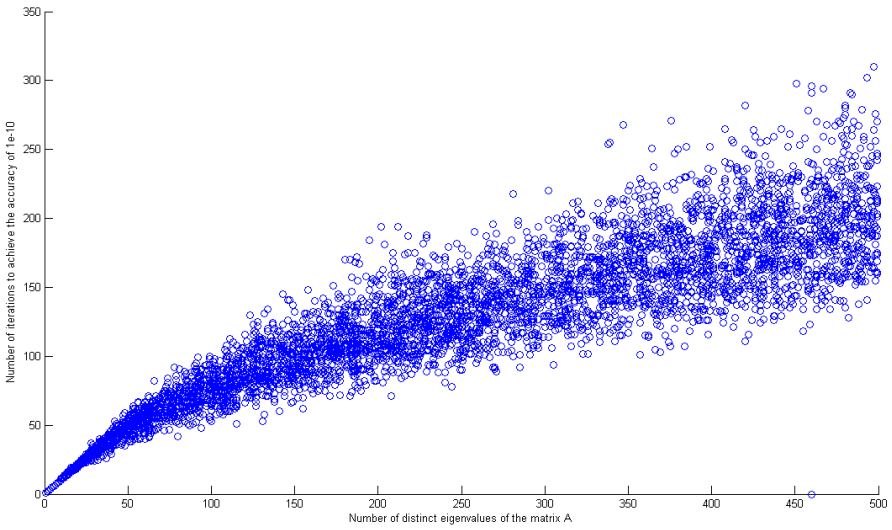

Figure 1: (# distinct λ_i) **vs** (# Krylov iterations) is pos. correlated; 5000 random instances.

2. The scatter plot in Figure 2 indicates that when the number of iterations is large enough, there is a positive correlation between the number of iterations and the ω-condition number of matrix $\boldsymbol{A}$. But this is only for small values of ω. In fact, the relatively large values of ω correspond to very low iteration counts. This confirms the results about maximizing in Corollary 2.

3. The maximization of ω observation in Corollary 2 is illustrated further in the scatter plot in Figure 3, where we fix λ_1, λ_n. We see the low number of Krylov iterations (on the vertical axis) when ω (on the horizontal axis) is small. But, in addition, as ω gets large, the number of Krylov iterations decrease dramatically.

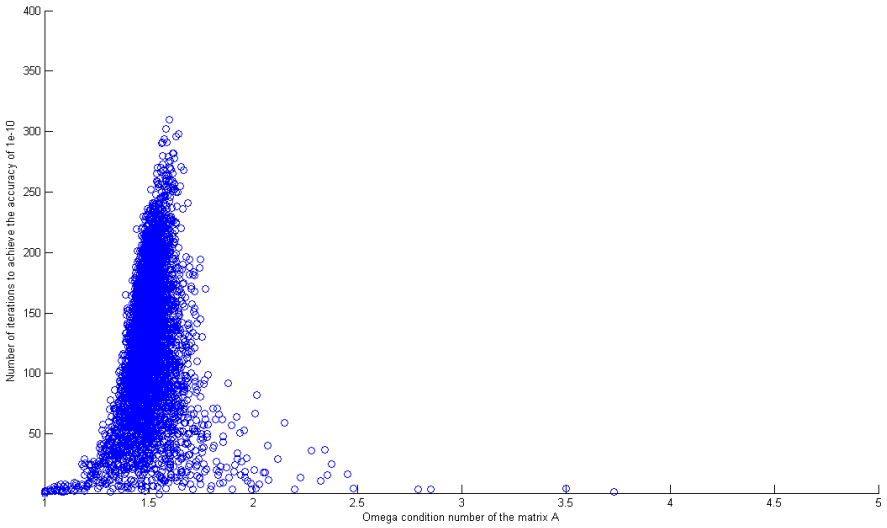

Figure 2: (ω) **vs** (# Krylov iterations) is pos. correlated for large iterations

4. Figure 4 shows the relationship between the number of iterations and the standard condition number. In sharp contrast to the results in Figures 2 and 3, We can see that there is no clear correlation between the number of Krylov iterations and the standard condition number.

4 ω-Condition Number and Preconditioning

If the matrix $\boldsymbol{A}$ is ill-conditioned, one way to achieve solutions with high accuracy is with preconditioners. Consider the overdetermined system $\boldsymbol{Ax} = \boldsymbol{b}$ with a full rank matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, where $m > n$. The least squares solution is the solution of the system $\boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{A}^T\boldsymbol{x}$. A nonsingular preconditioner $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is used to create a new overdetermined system $\boldsymbol{ADy} = \boldsymbol{b}$, where $\boldsymbol{y} = \boldsymbol{D}^{-1}\boldsymbol{x}$, such that the least squares solution $\boldsymbol{y}$ will be solved by a better-conditioned system of linear equations, $\boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{ADy} = \boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{b}$. We could try to find $\boldsymbol{D}$ that minimizes $\kappa(\boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{AD})$ since the standard condition number is an indicator of whether the problem is well- or ill-conditioned. Lu and Pong [9] shows that the diagonal preconditioner $\boldsymbol{D}$ that minimizes $\kappa(\boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{AD})$ can be compute

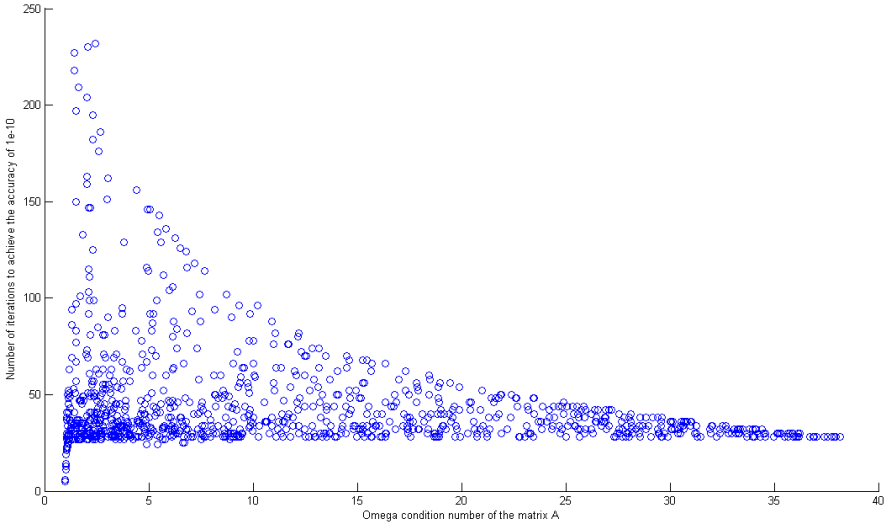

Figure 3: (ω) **vs** (# Krylov iterations) is neg. correlated for large ω; λ_1, λ_n fixed

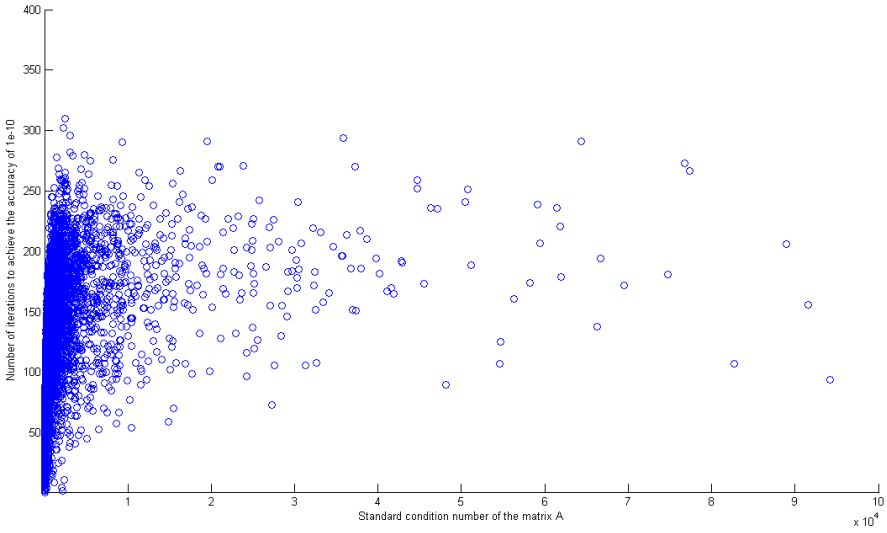

Figure 4: (κ) **vs** (# Krylov iterations) not correlated

by solving a conic optimization problem with semidefinite and second-order cone constraints:

$$\begin{array}{ll} \min & s \\ \text{s.t.} & \begin{pmatrix} \boldsymbol{I} & \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{x}) \\ \mathrm{diag}(\boldsymbol{x})\boldsymbol{A}^T & s\boldsymbol{I} \end{pmatrix} \succeq 0, \\ & \begin{pmatrix} \boldsymbol{I} & \mathrm{diag}(\boldsymbol{y}) \\ \mathrm{diag}(\boldsymbol{y}) & \boldsymbol{A}^T\boldsymbol{A} \end{pmatrix} \succeq 0, \\ & \begin{pmatrix} (x_i+y_i)/2 \\ (y_i-x_i)/2 \\ 1 \end{pmatrix} \in \mathbb{Q}^3, \qquad i=1,\ldots,n, \\ & (d_i-\eta)t \le x_i \le (d_i+\eta)t, \qquad i=1,\ldots,n, \\ & t \ge 0,\ \boldsymbol{x},\boldsymbol{y} \ge \boldsymbol{0}, \end{array} \tag{4.1}$$

where $K \succeq 0$ denotes positive semidefiniteness, $\mathbb{Q}^n$ is the n-dimensional second-order cone, $d_i = \dfrac{1}{||\boldsymbol{A}_{:,i}||}$ for all $i = 1,\ldots,n$, and $\eta > 0$ is large enough.

With respect to the ω-condition number, Dennis and Wolkowicz [1] show that the simple scaling diagonal preconditioner (see [11]) is the optimal preconditioner. Doan et al. [2] extend the result for block diagonal preconditioners. The details are shown in the following proposition.

Proposition 3. *The following statements are true.*

1. *The measure ω is pseudoconvex* [2] *on the set of s.p.d. matrices, and thus any stationary point is a global minimizer of ω.*

2. *Let $\boldsymbol{A}$ be a full rank $m \times n$ matrix, $n \le m$. Then the optimal column scaling that minimizes the measure ω, i.e.*

$$\min\ \omega((\boldsymbol{AD})^T(\boldsymbol{AD})),$$

over $\boldsymbol{D}$ positive, diagonal, is given by

$$D_{ii} = \frac{1}{\|\boldsymbol{A}_{:,i}\|}, i = 1,...,n,$$

where $\boldsymbol{A}_{:,i}$ is the i-th column of $\boldsymbol{A}$.

3. *Let $\boldsymbol{A}$ be a full rank $m \times n$ matrix, $n \le m$ with block structure $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \ldots & \boldsymbol{A}_k \end{bmatrix}, \boldsymbol{A}_i \in \mathbb{R}^{m \times n_i}$. Then* an *optimal corresponding block diagonal scaling*

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 & 0 & 0 & \ldots & 0 \\ 0 & \boldsymbol{D}_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \boldsymbol{D}_k \end{bmatrix}, \quad \boldsymbol{D}_i \in \mathbb{R}^{n_i \times n_i},$$

that minimizes the measure ω, i.e.

$$\min\ \omega((\boldsymbol{AD})^T(\boldsymbol{AD})),$$

[2] Note that a function is pseudoconvex if

$$(\boldsymbol{y}-\boldsymbol{x})^T\nabla f(\boldsymbol{x}) \ge 0 \Longrightarrow f(\boldsymbol{y}) \ge f(\boldsymbol{x}),$$

and for pseudoconvex functions, all stationary solutions are global minimizers (see for example, [10]).

over $\boldsymbol{D}$ block diagonal, is given by the factorization

$$\boldsymbol{D}_i\boldsymbol{D}_i^T = \{\boldsymbol{A}_i^T\,\boldsymbol{A}_i\}^{-1}, i = 1, ..., k.$$

Proof. Items 1 to 2 are proved in [1] and Item 3 is proved in [2]. For clarity, we present the proof again here for the general block diagonal preconditioners (which include the diagonal case).

Let the blocked $\boldsymbol{A}$ be given. Then the arithmetic-geometric mean inequality yields

$$\begin{aligned} \omega((\boldsymbol{AD})^T(\boldsymbol{AD})) &= \frac{\operatorname{trace}(\boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{AD})/n}{\det(\boldsymbol{D}^T\boldsymbol{A}^T\boldsymbol{AD})^{\frac{1}{n}}} \\ &= \frac{\operatorname{trace}(\boldsymbol{A}^T\boldsymbol{ADD}^T/n}{\det(\boldsymbol{DD}^T)^{\frac{1}{n}}\det(\boldsymbol{A}^T\boldsymbol{A})^{\frac{1}{n}}}. \end{aligned} \tag{4.2}$$

Let $\bar{\boldsymbol{D}} = \boldsymbol{DD}^T, \bar{\boldsymbol{A}}_i = \boldsymbol{A}_i^T\boldsymbol{A}_i$. We have $\bar{\boldsymbol{D}}$ is also a block diagonal matrix with $\bar{\boldsymbol{D}}_i = \boldsymbol{D}_i\boldsymbol{D}_i^T$ for all $i = 1, \ldots, k$, and $\operatorname{trace}(\boldsymbol{A}^T\boldsymbol{ADD}^T) = \operatorname{trace}\sum_{i=1}^{k}\bar{\boldsymbol{A}}_i\bar{\boldsymbol{D}}_i$. Since the minimum of the function ω in (4.2) is scale free, we can ignore the constants $1/n, \det(\boldsymbol{A}^T\boldsymbol{A})^{\frac{1}{n}}$, take logs, and equivalently solve

$$\max\left\{\sum_{i=1}^{k}\log\det\bar{\boldsymbol{D}}_i : \sum_{i=1}^{k}\bar{\boldsymbol{A}}_i\cdot\bar{\boldsymbol{D}}_i = 1\right\}.$$

Using a Lagrange multiplier λ and Cramer's rule applied to the gradient of the objective function, we get an implicit expression for the optimum in

$$\bar{\boldsymbol{D}}_i{}^{-1} - \lambda\bar{\boldsymbol{A}}_i = 0, \quad i = 1, \ldots k.$$

□

Note that the optimality condition only requires $\bar{\boldsymbol{D}}_i = \bar{\boldsymbol{A}}_i^{-1}$ for $i = 1, \ldots, k$, which means we can use the QR decomposition to find the optimal $\boldsymbol{D}_i$ instead of the matrix square root. In other words, if $\boldsymbol{A}_i$ has the QR decomposition $\boldsymbol{A}_i = \boldsymbol{Q}_i\boldsymbol{R}_i$, we can set $\boldsymbol{D}_i = \boldsymbol{R}_i^{-1}$ for all $i = 1, \ldots, k$ to minimize the measure $\omega((\boldsymbol{AD})^T(\boldsymbol{AD}))$. We let $\boldsymbol{D}_\omega, \boldsymbol{D}_\kappa$ denote the optimal column scaling for the measures ω, κ, respectively.

We now generate random matrices $\boldsymbol{A}$ of size 1000 by 500 with random singular values. Figures 5 and 6 show that $\boldsymbol{D}_\omega$ and $\boldsymbol{D}_\kappa$ indeed reduces the ω-condition number and the standard condition number, respectively. We continue and generate random *sparse* matrices with random singular values. We then consider the overdetermined system $\boldsymbol{Ax} = \boldsymbol{b}$ with no preconditioner, the diagonal preconditioner $\boldsymbol{D}_\omega$, and the diagonal precoditioner $\boldsymbol{D}_\kappa$. We solve 1000 random sparse matrix instances with the least squares solver LSQR in MATLAB to achieve the accuracy of 10^{-10}. Figure 7 shows that both preconditioners reduce the number of iterations significantly with the average change of -58.77%. The preconditioner $\boldsymbol{D}_\kappa$ is slightly better than $\boldsymbol{D}_\omega$ with about 0.84% reduction in number of iterations. However, we can see that the preconditioner $\boldsymbol{D}_\omega$ is much easier to compute as compared to the preconditioner $\boldsymbol{D}_\kappa$, which requires solving a conic optimization problem of size $2n$.

We now test the reduction in number of iterations and compare the computational times. We again generate random (sparse) matrices of size 1000 by 500 and solve the system with no preconditioner ($p = 0$), and block diagonal preconditioners with the block size $p = 1, 2, 5, 10, 20, 25, 50, 100,$

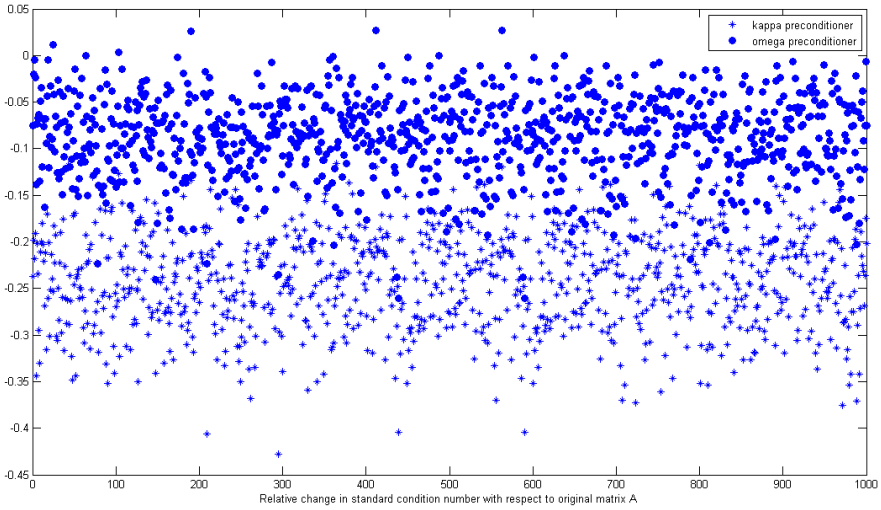

Figure 5: κ with two different diagonal preconditioners

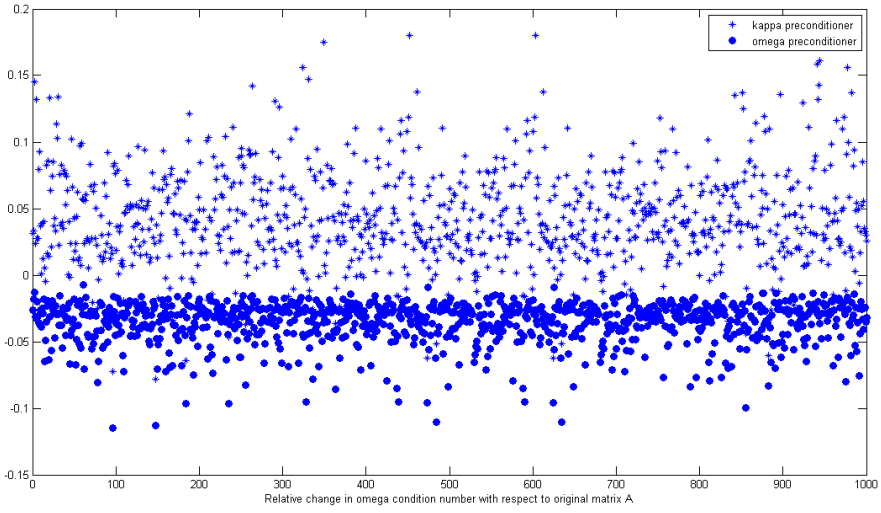

Figure 6: ω with two different diagonal preconditioners

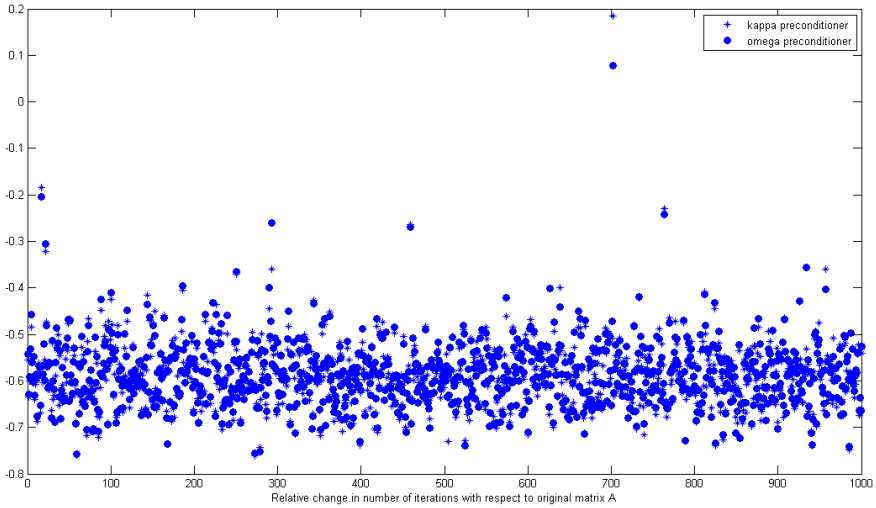

Figure 7: # Krylov iterations with two different diagonal preconditioners for sparse matrices

and 250. The results are shown in Table 4.1 for random sparse matrices, where n_i is the number of iterations, t_p is the time to compute the preconditioner, and t_t is the total time to find the least squares solution. We can see that block diagonal preconditioners improve the number of iterations and also the computation time.

p	0	1	2	5	10	20	25	50	100	250
$n_i(\times 10^4)$	5.89	1.31	1.01	0.66	0.53	0.46	0.45	0.38	0.30	0.11
t_p	-	2.46	1.36	0.70	0.50	0.42	0.42	0.42	0.44	0.26
t_t	39.00	14.50	11.15	7.94	6.76	6.28	6.20	5.89	5.60	4.14

Table 4.1: Average # iterations; computational time block preconditioners; sparse matrices

5 ω-Condition Number and Relative Error

We now focus our attention again for $\boldsymbol{A} \in \mathcal{S}^n_{++}$, positive definite, and the system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. We would like to consider the following ratio of the relative errors

$$r_e(\boldsymbol{A}, \boldsymbol{b}, \Delta\boldsymbol{b}) := \frac{\left|\left|\boldsymbol{A}^{-1}\Delta\boldsymbol{b}\right|\right|}{\left|\left|\boldsymbol{A}^{-1}\boldsymbol{b}\right|\right|} \left(\frac{||\Delta\boldsymbol{b}||}{||\boldsymbol{b}||}\right)^{-1}. \tag{5.1}$$

The condition number $\kappa(\boldsymbol{A})$ provides an upper bound of this ratio

$$\kappa(\boldsymbol{A}) = \max_{\boldsymbol{b}, \Delta\boldsymbol{b}} r_e(\boldsymbol{A}, \boldsymbol{b}, \Delta\boldsymbol{b}).$$

The bound is attained when $\boldsymbol{b} = \beta\boldsymbol{v}_1$ and $\Delta\boldsymbol{b} = \Delta\beta\boldsymbol{v}_n$, where $\boldsymbol{v}_1$ and $\boldsymbol{v}_n$ are the eigenvectors corresponding to the largest and smallest eigenvectors. Clearly, if $\boldsymbol{b}$ and $\Delta\boldsymbol{b}$ are arbitrary vectors,

this bound is unlikely to be attained. We would like to consider the expected value of $r_e(\boldsymbol{A}, \boldsymbol{b}, \Delta\boldsymbol{b})$ when $\boldsymbol{b}$ and $\Delta\boldsymbol{b}$ are random vectors. We assume that $\boldsymbol{b}$ and $\Delta\boldsymbol{b}$ are independent normal random vectors, $\boldsymbol{b}, \Delta\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Let us consider the expectation

$$r(\boldsymbol{A}) = \mathbb{E}\left[r_e(\boldsymbol{A}, \boldsymbol{b}, \Delta\boldsymbol{b})\right].$$

The following proposition provides the analytic formulation for $r(\boldsymbol{A})$.

Proposition 4. *If $\boldsymbol{b}$ and $\Delta\boldsymbol{b}$ are independent normal random vectors, then $r(\boldsymbol{A})$ can be calculated as follows:*

$$r(\boldsymbol{A}) = \left(\sum_{i=0}^{\infty} \frac{(-1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B})\right)\left(\sum_{i=0}^{\infty} \frac{(1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B})\right), \tag{5.2}$$

where $\boldsymbol{B} = (\boldsymbol{A}^{-1})^T\boldsymbol{A}^{-1}$, $\beta = \lambda^2_{\min}(\boldsymbol{A})$, $(a)_b = \dfrac{\Gamma(a+b)}{\Gamma(a)}$, and $d_k(\boldsymbol{A})$ is the normalized top-order zonal polynomial,

$$d_k(\boldsymbol{A}) = \sum_{k_i \geq 0:\, \sum_{i=1}^n k_i = k} \left(\prod_{i=1}^{n} \frac{\left(\frac{1}{2}\right)_{k_i} \lambda_i(\boldsymbol{A})^{k_i}}{k_i!}\right), \quad \forall\, k \geq 0.$$

.

Proof. We have: since $\boldsymbol{b}$ and $\Delta\boldsymbol{b}$ are independent,

$$r(\boldsymbol{A}) = \mathbb{E}\left[\frac{||\boldsymbol{A}^{-1}\Delta\boldsymbol{b}||}{||\Delta\boldsymbol{b}||}\right]\mathbb{E}\left[\frac{||\boldsymbol{b}||}{||\boldsymbol{A}^{-1}\boldsymbol{b}||}\right]. \tag{5.3}$$

These two expectations has the same form; they are the expectations of ratios of quadratic forms, $\dfrac{(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})^p}{(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x})^q}$, where p and q are positive real number, $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$. Smith [12] shows the exact formulation for the expectation with $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{Q})$ which involves top-order invariant polynomials. The formulation can be simplified if $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{B} = \boldsymbol{I}$ in which top-order invariant polynomials are replaced by simpler top-order zonal polynomials (see Smith [12] for details). In our particular case, we have: $p = q = \dfrac{1}{2}$ since for arbitrary $\boldsymbol{A}$ and $\boldsymbol{x}$, $||\boldsymbol{A}\boldsymbol{x}|| = \left(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}\right)^{\frac{1}{2}}$. In addition, there is at least an identity matrix, either in the denominator or numerator. Deriving directly from the general formulation presented in Smith [12], we obtain the following formulation for $\mathbb{E}\left[\dfrac{(\Delta\boldsymbol{b}^T\boldsymbol{B}\Delta\boldsymbol{b})^{\frac{1}{2}}}{(\Delta\boldsymbol{b}^T\Delta\boldsymbol{b})^{\frac{1}{2}}}\right]$, which only involves top-order zonal polynomials:

$$\mathbb{E}\left[\frac{(\Delta\boldsymbol{b}^T\boldsymbol{B}\Delta\boldsymbol{b})^{\frac{1}{2}}}{(\Delta\boldsymbol{b}^T\Delta\boldsymbol{b})^{\frac{1}{2}}}\right] = \beta^{-\frac{1}{2}}\sum_{i=0}^{\infty}\frac{(-1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B}),$$

where $\boldsymbol{B} = (\boldsymbol{A}^{-1})^T\boldsymbol{A}^{-1}$ and $\beta = \lambda^2_{\min}(\boldsymbol{A})$. Similarly, we have:

$$\mathbb{E}\left[\frac{(\boldsymbol{b}^T\boldsymbol{b})^{\frac{1}{2}}}{(\boldsymbol{b}^T\boldsymbol{B}\boldsymbol{b})^{\frac{1}{2}}}\right] = \beta^{\frac{1}{2}}\sum_{i=0}^{\infty}\frac{(1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B}).$$

Thus we obtain the exact formulation of $r(\boldsymbol{A})$ as shown in (5.2). $\square$

The formulation (5.2) involves infinite sums. Following the approach in Hillier et al. [7], we now attempt to find an upper bound for $r(\boldsymbol{A})$ with partial sums. We have: $(-1/2)_k < 0$ for all $k \geq 1$. In addition, $\boldsymbol{C} = \boldsymbol{I} - \beta\boldsymbol{B} \succeq 0$; therefore, $d_k(\boldsymbol{C}) \geq 0$ for all $k \geq 1$. Thus for $M \geq 0$,

$$\sum_{i=0}^{\infty} \frac{(-1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B}) \leq \sum_{i=0}^{M} \frac{(-1/2)_i}{(n/2)_i} d_i(\boldsymbol{C}).$$

We have: $n \geq 1$, thus $\dfrac{(1/2)_k}{(n/2)_k} \leq \dfrac{(1/2)_{M+1}}{(n/2)_{M+1}}$ for all $k \geq M+1$. Using the generating function of $d_k(\boldsymbol{C})$, we have:

$$[\det(\boldsymbol{I} - t\boldsymbol{C})]^{-\frac{1}{2}} = \sum_{k=0}^{\infty} d_k(\boldsymbol{C}) t^k.$$

Let $t = 1$, we have:

$$[\det(\beta\boldsymbol{B})]^{-\frac{1}{2}} = \sum_{k=0}^{\infty} d_k(\boldsymbol{C}).$$

Thus for $M \geq 0$,

$$\sum_{i=0}^{\infty} \frac{(1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B}) \leq \sum_{i=0}^{M} \frac{(1/2)_i}{(n/2)_i} d_i(\boldsymbol{C}) + \frac{(1/2)_{M+1}}{(n/2)_{M+1}} \left[\frac{1}{\beta^{\frac{n}{2}} [\det(\boldsymbol{B})]^{\frac{1}{2}}} - \sum_{i=0}^{M} d_i(\boldsymbol{C}) \right].$$

We then have the following upper bound for $r(\boldsymbol{A})$ for each $M \geq 0$:

$$r(\boldsymbol{A}) \leq \left(\sum_{i=0}^{M} \frac{(-1/2)_i}{(n/2)_i} d_i(\boldsymbol{C}) \right) \left(\sum_{i=0}^{M} \frac{(1/2)_i}{(n/2)_i} d_i(\boldsymbol{C}) + \frac{(1/2)_{M+1}}{(n/2)_{M+1}} \left[\frac{1}{\beta^{\frac{n}{2}} [\det(\boldsymbol{B})]^{\frac{1}{2}}} - \sum_{i=0}^{M} d_i(\boldsymbol{C}) \right] \right) \quad (5.4)$$

Let $M = 1$, we have: $d_0(\boldsymbol{C}) = 1$ and $d_1(\boldsymbol{C}) = \dfrac{\text{trace}(\boldsymbol{C})}{2}$. In addition, eigenvalues of $\boldsymbol{C}$ are $\gamma_i = 1 - \dfrac{\lambda_n^2}{\lambda_i^2}$. Thus

$$\sum_{i=0}^{\infty} \frac{(-1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B}) \leq 1 - \frac{\text{trace}(\boldsymbol{C})}{2n} = \frac{1}{2} \left(1 + \frac{\lambda_n^2}{n} \left(\sum_{i=1}^{n} \frac{1}{\lambda_i^2} \right) \right).$$

We also have:

$$\sum_{i=0}^{\infty} \frac{(1/2)_i}{(n/2)_i} d_i(\boldsymbol{I} - \beta\boldsymbol{B}) \leq \frac{3(n-1)}{2n} - \frac{n-1}{2(n+2)} \left(\frac{\lambda_n^2}{n} \right) \left(\sum_{i=1}^{n} \frac{1}{\lambda_i^2} \right) + \frac{3}{n(n+2)} \prod_{i=1}^{n} \left(\frac{\lambda_i}{\lambda_n} \right).$$

We obtain an upper bound for $r(\boldsymbol{A})$:

$$r(\boldsymbol{A}) \leq \frac{1}{2} \left[1 + \frac{\lambda_n^2}{n} \left(\sum_{i=1}^{n} \frac{1}{\lambda_i^2} \right) \right] \left[\frac{3(n-1)}{2n} - \frac{n-1}{2(n+2)} \left(\frac{\lambda_n^2}{n} \right) \left(\sum_{i=1}^{n} \frac{1}{\lambda_i^2} \right) + \frac{3}{n(n+2)} \prod_{i=1}^{n} \left(\frac{\lambda_i}{\lambda_n} \right) \right].$$

Unfortunately, this upper bound in general is not very tight and there are examples in which we need very large M to obtain a decent approximation of $r(\boldsymbol{A})$. We now empirically calculate the sample

approximation $\bar{r}(\boldsymbol{A})$ of $r(\boldsymbol{A})$ with a large number of samples and check the relationship between $\bar{r}(\boldsymbol{A})$ and ω-condition number of $\boldsymbol{A}$ as well as the standard condition number. We generate random matrices of the size 1000 by 1000 and we calculate $\bar{r}(\boldsymbol{A})$ with $N = 10,000$ samples. Figures 8 and 9 show the ratios of the ω-condition number and standard condition numberof $\boldsymbol{A}$, respectively, to the sample approximation of $r(\boldsymbol{A})$. The maximum value of ratios is limited to 10 in both figures for the purpose of comparison.

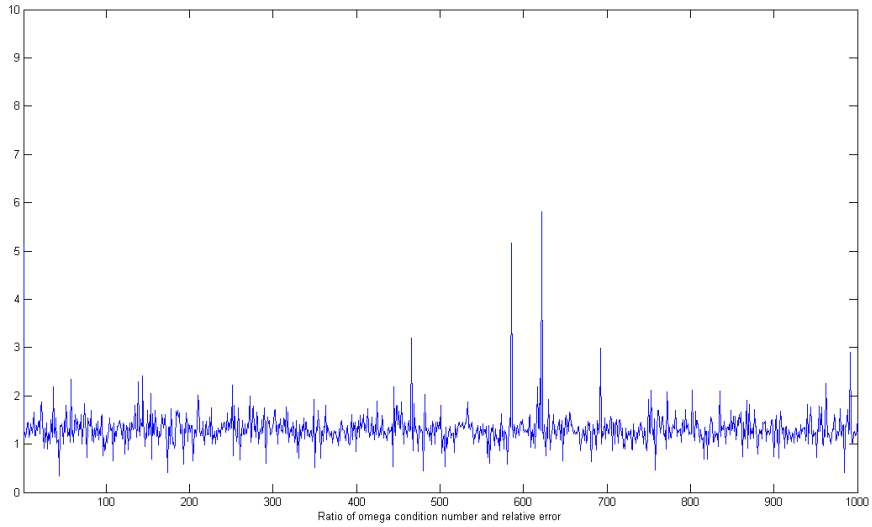

Figure 8: Ratio of ω and the relative error $\bar{r}(\boldsymbol{A})$

Clearly, the standard condition number is an upper bound of $\bar{r}(\boldsymbol{A})$; however, the average standard condition number is 1.03×10^3 while the maximum relative error is just 4.41. On the other hand, the ω-condition number is not always an upper bound of $\bar{r}(\boldsymbol{A})$ but it is in more than 90% of all instances. In addition, the average ω-condition number is 1.49, which is very close to the average relative error of 1.18. The average ratio of ω-condition number is 1.31 with the standard deviation of 0.35 while these two measures of κ-condition number are 487.92 and 3.82×10^3, respectively. We can conclude that the ω-condition number is a good indicator for the expected relative error $r(\boldsymbol{A})$ in this random setting.

6 Conclusion

We have presented computational and theoretical properties to show that the condition number measure $\omega(A)$ is a better indicator than the standard condition number κ for the conditioning of A when it comes to: (i) evaluation, (ii) predicting the number of iterations in Krylov iterative methods, (iii) finding optimal preconditioners for iterative methods, and (iv) estimating the *expected* roundoff error when solving linear equations.

In addition, we have shown that ω is a simpler function (measure) to use when finding optimal preconditioners, as it is easy to differentiate. And, motivated by the fact that the number of iterations in Krylov type methods depends directly on the number of distinct eigenvalues, we

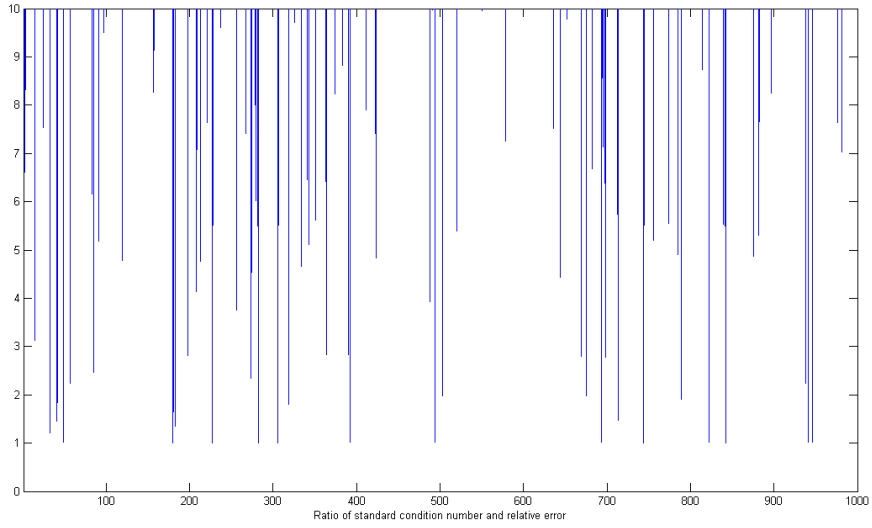

Figure 9: Ratio of κ and the relative error $\bar{r}(\boldsymbol{A})$

showed empirically that if κ is fixed, then both small and large values of ω lead to lower iteration numbers; this follows since both maximizing and minimizing ω leads to clustering of eigenvalues when κ is fixed. Therefore, an interesting conjecture is that: $m$aximizing ω subject to keeping κ constant leads to good preconditioners. However, the problem of maximizing the pseudoconvex function ω is difficult and so heuristics will have to be used.

References

[1] J.E. Dennis Jr. and H. Wolkowicz. Sizing and least-change secant methods. *SIAM J. Numer. Anal.*, 30(5):1291–1314, 1993. 3, 10, 11

[2] X.V. Doan, S. Kruk, and H. Wolkowicz. A robust algorithm for semidefinite programming. *Optim. Methods Softw.*, .(CORR 2010-09):., 2011. submitted in November, 2010, accepted June, 2011. 10, 11

[3] R.G. Grimes and J.G. Lewis. Condition number estimation for sparse matrices. *SIAM J. Sci. Statist. Comput.*, 2:384–388, 1981. 4

[4] W. W. Hager. Condition estimates. *SIAM J. Sci. Statist. Comput.*, 5(2):311–316, 1984. 4

[5] N.J. Higham. A survey of condition number estimation for triangular matrices. *SIAM Rev.*, 29:575–596, 1987. 4

[6] N.J. Higham and F. Tisseur. A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra. *SIAM J. Matrix Anal. Appl.*, 21(4):1185–1201 (electronic), 2000. 4

[7] G. Hillier, R. Kan, and X. Wang. Computationally efficient recursions for top-order invariant polynomials with applications. *Econometric Theory*, 25:211–242, 2009. 15

[8] Ilse C. F. Ipsen and Carl D. Meyer. The idea behind Krylov methods. *Amer. Math. Monthly*, 105(10):889–899, 1998. 4

[9] Z. Lu and T.K. Pong. Minimizing condition number via convex programming. Technical report, University of Washington, Seattle, WA, 2010. 8

[10] O.L. Mangasarian. *Nonlinear Programming*. McGraw-Hill, New York, NY, 1969. 10

[11] G. Pini and G. Gamboniati. Is a simple diagonal scaling the best preconditioner for conjugate gradients on supercomputers? *Advances in Water Resources*, 13(3):147–153, 1990. 10

[12] Murray D. Smith. Expectations of ratios of quadratic forms in normal variables: evaluating some top-order invariant polynomials. *Austral. J. Statist.*, 35(3):271–282, 1993. 14

Index