Studies in Matrix Perturbation and Robust Statistics

by

Yanyuan Ma

B.S., Peking University, 1990-1994

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 1999

Signature of Author: ....
Department of Mathematics
April 30, 1999

Certified by ....
Alan Edelman
Associate Professor of Applied Mathematics
Thesis Supervisor

Accepted by ....
Michael Sipser
Chair, Applied Mathematics Committee

Accepted by ....
Richard Melrose
Chair, Departmental Committee on Graduate Students

Studies in Matrix Perturbation and Robust Statistics

by

Yanyuan Ma

Submitted to the Department of Mathematics
on April 30, 1999 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ABSTRACT

This thesis has three parts. The first part concentrates on the matrix eigenvalue perturbation theory. It discusses the non-generic eigenvalue behavior under a perturbation on a matrix in its Jordan form. Comparing to the widely known generic behavior, where all eigenvalues form a ring, the non-generic behavior gives several sets of eigenvalues that form different rings. The second part discusses the stability of the Jordan form of matrices and the Kronecker form of pencils. It analyzes the widely used algorithm of determining the Canonical form, the staircase algorithm, and explains the reason that causes the failure of the algorithm. The methods used in the two parts are mainly based on the geometrical view point of looking at an $n \times n$ matrix as a point in an n^2 dimensional space, and the set of similar matrices as an orbit in the space. The third part focuses on robust statistics. A new robust estimator on covariance is given and later it is generalized to autocovariance estimator and dispersion matrix estimator. The statistical properties of the estimator are studied and simulations are also given to test the estimator. The estimator is mainly based on the fact that $4\text{cov}(X, Y) = \text{var}(X + Y) - \text{var}(X - Y)$.

Thesis Supervisor: Alan Edelman
Title: Associate Professor

Acknowledgment

I would first like to thank my advisor, Professor Alan Edelman, for making this thesis possible. I am deeply grateful for his constant encouragement and enthusiasm, personal attention, and sharing of his perspectives on numerical analysis, mathematical understanding and effective presentation. I will never forget the late night rides after conquering another difficulty.

I would also like to thank Marc Genton for guiding me in the robust statistics area and showing me the general technologies in statistics.

Thanks also to Professor Gilbert Strang for teaching me wavelets and filter banks.

For generous financial support during my time at MIT, I am grateful for teaching assistantships and a fellowship from mathematics department, and research assistantships funded by National Science Foundation grants 9501278-DMS to Professor Edelman.

I am thankful for both technical help and companionship from colleagues in the math department, particularly the numerics group of Peter McCorquadale, Ross Lippert and Parry Husbands. Finally, thanks to my family and my husband for their devotion and support during these years.

Contents

List of Figures

List of Tables

1 Introduction

1.1 The Eigenvalue Perturbation Problem

Perturb an $n \times n$ Jordan block by order ϵ mathematically or through rounding errors on a computer, and typically the eigenvalues split up into a ring of radius $O(\epsilon^{1/n})$. In this thesis, we study the non-typical behavior. We stifle the matrix's ability to form large eigenvalue rings by only allowing perturbations that are upper k-Hessenberg, meaning a matrix containing exactly k subdiagonals below and including the diagonal. The obvious question to ask is what is the typical behavior under this assumption. The result we will show is that the eigenvalue perturbations will then follow the greediest possible pattern consistent with forming no rings bigger than k. We then generalize and examine some multiple Jordan block cases.

Our interest in this problem came from a perturbation study of Ruhe's matrix [35] using the qualitative approach proposed by Chatelin and Frayssé [13]. We found that non-generic behaviors occurred some small percentage of the time. Chatelin and Frayssé themselves point out in one example [13, page 192] that only 97% of their examples follow the expected behavior. We also became interested in this problem because we wanted to understand how eigenvalues perturb if we move in some, but not all normal directions to the orbit of a matrix with a particular Jordan form such as in Arnold's versal deformation [1, 33]. Such information may be of value in identifying the nearest matrix with a given Jordan structure. Finally, we point out, that the ϵ-pseudo-spectra of a matrix can depend very much on the sparsity structure of the allowed perturbations. Following an example from Trefethen [84], if we take a Jordan block J and then compute in the presence of roundoff error, $A = Q^T J Q$, where Q is a banded orthogonal matrix, then the behavior of $||A^k||$, is quite different from what would happen if Q were dense.

It is generally known [2, page 109],[66, page 65] that if a matrix A is perturbed by any matrix ϵB, then any multiple eigenvalue splits into rings, and their expansion in ϵ is a Puiseux series since it is a branch of the solution of a polynomial with analytic coefficients.

Unfortunately, the classical references give little information as to how the eigenvalues split as a function of the sparsity structure of the perturbation matrix. Without loss of generality, we will focus on one multiple eigenvalue. Associated with any perturbation B, we may define a partition $\pi(B)$ which contains the sizes (number of eigenvalues) of the rings.[1]

We can quickly summarize most of what is known about the Puiseux series. If A is a single Jordan block of size n, then $\pi(B)$ is almost always $\{n\}$. This happens if and only if the lower left element is non-zero. For more complicated Jordan structures, say A is a nilpotent matrix, $\pi(B)$ is almost always the Segré characteristics of A, i.e, the sizes of the Jordan block structure of A. Lidskii explicitly determined the coefficients of the first order term, and Newton diagram approaches may also be used (See [81] for a discussion).

We used the words "almost always" in the above paragraph. There is an algebraic variety on which different behavior occurs. An ideal mathematical treatment would conveniently categorize all possible behaviors as a function of the perturbation B. This is a very difficult open problem. The only result of which we are aware is given by Burke and Overton [10] and Moro, Burke and Overton [81] . The former studied when perturbations only yield periods of size 1 and 2 as part of a study of when the perturbations fall to one side, and the latter studied the first order perturbations under generic conditions and addressed some nongeneric situations.

Our approach is to try to identify classes of non-generic situations where we can explain the typical behavior. We set up hypotheses on the structure of the perturbation, thereby creating non-generic perturbations. We then ask what is the generic behavior of the eigenvalues given these hypotheses. (To be more precise, unless the perturbation satisfies certain algebraic conditions, the behavior occurs.)

[1] The size of a ring is denoted its "period" in [66, 2]. The eigenvalue functions of ϵ in the same ring constitute a "cycle" in the terminology of these references.

1.2 The Staircase Algorithm Problem

The problem of accurately computing Jordan and Kronecker canonical structures of matrices and pencils has captured the attention of many specialists in numerical linear algebra. Standard algorithms for this process are denoted "staircase algorithms" because of the shape of the resulting matrices [47, Page 370], but understanding of how and why they fail is incomplete. In this paper, we study the geometry of matrices in n^2 dimensional space and pencils in $2mn$ dimensional space to explain these failures. This follows a geometrical program to complement and perhaps replace traditional numerical concepts associated with matrix subspaces that are usually viewed in n dimensional space.

This section targets expert readers who are already familiar with the staircase algorithm. We refer readers to [47, Page 370] and [21] for excellent background material and we also list other literature for the reader wishing a comprehensive understanding of the algorithm. On the mathematical side, it is also helpful if the reader has some knowledge of Arnold's theory of versal forms, though a dedicated reader should be able to read this paper without such knowledge, perhaps skipping Section 3.3.2.

The first staircase algorithm was given by Kublanovskaya for Jordan structure in 1966 [68], where a normalized QR factorization is used for rank determination and nullspace separation. Ruhe [91] first introduced the use of the SVD into the algorithm in 1970. The SVD idea is further developed by Golub and Wilkinson [48, Section 10]. Kågström and Ruhe [62, 63] wrote the first library quality software for the complete JNF reduction, with the capability of returning after different steps in the reduction. Recently, Chatitin-Chatelin and Frayssé [14] developed a non-staircase "qualitative" approach.

The staircase algorithm for the Kronecker structure of pencils is given by Van Dooren [28, 29, 30] and Kågström and Ruhe [64]. Kublanovskaya [69] fully analyzed the AB algorithm, however, earlier work on the AB algorithm goes back to the 1970s. Kågström [60, 61] gave a RGDSVD/RGQZD algorithm and this provided a base for later work on software. Error bounds for this algorithm are given by Demmel and Kågström [19, 20]. Beelen and Van Dooren [3] gave an improved algorithm which requires $O(m^2n)$ operations for $m \times n$ pencils. Boley [5] studied the sensitivity of the algebraic structure. Error bounds are given by Demmel and Kågström [21, 22].

Staircase algorithms are used both theoretically and practically. Elmroth and Kågström [36] use the staircase algorithm to test the set of 2-by-3 pencils hence to analyze the algorithm, Demmel and Edelman [18] use the algorithm to calculate the dimension of matrices and pencils with a given form. Van Dooren [29, 37, 67, 7], Emami-Naeini [37], Kautsky and Nichols [67], Boley [7], Wicks and DeCarlo [94] consider systems and control applications. Software for control theory is provided by Demmel and Kågström [23].

A number of papers use geometry to understand Jordan and Kronecker structure problems. Fairgrieve [38] regularizes by taking the most degenerate matrix in a neighborhood, Edelman, Elmroth and Kågström [34, 35] study versality and stratifications, and Boley [6] concentrates on stratifications.

1.3 The Dispersion Estimation Problem

Dispersion matrices, i.e. covariance and correlation matrices, play an important role in many methods of multivariate statistics. For instance, they are the cornerstones of principal component analysis, discriminant analysis, factor analysis, canonical correlation analysis, and many others [76]. Moreover, dispersion matrices are themselves quantities of interest since they represent a measure of association or interdependence between several characteristics. They provide information about the shape of the ellipsoid of the data cloud in a multidimensional space. Therefore, reliable estimators of dispersion matrices are of prime importance.

Unfortunately, classical sample dispersion matrices are known to be very sensitive to outlying values in the data, which can typically be hidden in the high dimensionality of the space of variables. As a consequence, eigenvalues and eigenvectors of the dispersion matrix inherit this sensitivity. A principal component analysis could thus reveal an artificial structure in the data, that does not really exist but is merely created by a few outliers.

We describe some commonly used estimators for the dispersion matrix, as well as some recent robust proposals. We focus on the estimation of covariance matrices, since estimation of correlation matrices can be derived in the same way.

Suppose that the sample $\mathbf{x}_1, \dots, \mathbf{x}_n$, with $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \dots, n$, is independently and identically distributed according to a multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Note that estimation of the correlation matrix R can always be derived from the relation $R = D\Sigma D$, where $D = \text{diag}(1/\sqrt{\Sigma_{11}}, \dots, 1/\sqrt{\Sigma_{pp}})$. The maximum likelihood estimator (MLE) of the covariance matrix Σ is:

$$\hat{\Sigma}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T, \tag{1}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$.

The breakdown point is an important feature of reliability of an estimator. It indicates, roughly speaking, the largest proportion of data that can be replaced by arbitrary values to bring the estimator to the boundaries of the parameter space. More details can be found in [27, 58, 59, 53]. The breakdown point of the maximum likelihood estimator (1) is zero, indicating its very poor resistance.

Affine equivariant M-estimators for dispersion matrices were first suggested [50], and studied by [77, 57, 58]. Unfortunately, their breakdown point is at most $1/(p+1)$. This is not satisfactory, because it means that the breakdown point becomes smaller with increasing dimension, where there are more opportunities for outliers to occur. The performance of some M-estimators were studied by mean of a Monte Carlo study by [25, 26].

[92] and [27] were first to independently propose robust affine equivariant estimators of multivariate location and dispersion having a high breakdown point (asymptotically 1/2) for any dimension. They are defined as weighted mean and weighted dispersion, where the weights are functions of a measure of "outlyingness" obtained by considering all univariate projections of the data. Subsequently, other high breakdown point equivariant

multivariate estimators have been introduced. The most well known is probably the Minimum Volume Ellipsoid (MVE) estimator, introduced by [85], and discussed in [88, 90]. The method seeks an ellipsoid of minimum volume, containing $m = \lfloor (n+p+1)/2 \rfloor$ points, where $\lfloor \cdot \rfloor$ denotes the integer part. More precisely, it consists in finding $\hat{\boldsymbol{\mu}}_{MVE}$ and $\hat{\Sigma}_{MVE}$ such that the determinant of Σ is minimized subject to

$$\#\{i \big| (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \leq a^2\} \geq m \tag{2}$$

where a^2 is a fixed constant, for example χ^2_p in the case of Gaussian data. The MVE has a finite sample breakdown point of m, i.e. 50% asymptotically. Two algorithm (resampling and projection) to compute an approximate solution of MVE can be found in [90].

The MVE estimator has been generalized to multivariate S-estimators [17, 74, 75]. [72] proposed a dispersion matrix estimator based on robustifying principal components via projection pursuit techniques. A class of projection estimators for dispersion matrices were studied by [78]. [93] discusses finite sample breakdown point of projection based estimators, in particular the Stahel-Donoho estimator. Recently, [79] studied asymptotic and finite-sample behaviors of the Stahel-Donoho robust multivariate estimators. From a simulation study, they concluded that they compare favorably with other proposals like multivariate M- or S-estimators, and Rousseeuw's MVE. However, the main drawback remains the lack of feasible methods to compute the estimators for dimensions larger than $p = 2$.

In the three last decades, many attempts to overcome the poor resistance properties of the classical sample dispersion matrix have been made. The robust proposals can be classified in two main categories: robust componentwise estimation and robust global estimation of the dispersion matrix. The first one can be approached via location estimation, or scale estimation. It has the advantage of being able to deal with missing values in the data, but is not affine invariant and does not provide a positive definite matrix directly. The second category usually insures affine invariance and positive definiteness, but is less appropriate to deal with missing data.

We propose the use of a highly robust estimator of scale, denoted by Q_n, in the componentwise approach. In fact, we show that it is the best robust choice available at the present time in the componentwise approach. The highly robust estimator of scale Q_n has already been successfully used in the context of regression [55, 16], as well as for variogram estimation [41] in spatial statistics.

We proceed to form a highly robust autocovariance estimator in time series based on the dispersion extimator. Autocovariance is often used to study the underlying dependence structure of the process [8, 9], it serves as an important step towards constructing an appropriate mathematical model for the data. To have a sample autocovariance function which remains close to the true underlying autocovariance function, even when outliers, i.e. faulty observations, are present in the data is of crucial meaning. Otherwise, important goals of the time series analysis such as inference or forecasting can be non-informative. In fact, experience from a broad spectrum of applied sciences shows that measured data may contain between 10-15% of outlying values [50] due to gross errors, round-off errors, measurement mistakes, faulty recording, etc, and this proportion can even go up to 30% [57]. The estimator we introduce has a temporal breakdown point of 15% at the worst case.

2 Non-generic Eigenvalue Perturbations of Jordan Blocks

2.1 Introduction

We know that diagonalizable matrices form a dense set in the matrix space, and for diagonalizable matrices, an order ϵ perturbation on the matrix will lead to an order ϵ perturbation on the eigenvalues. For non-diagonalizable matrices, the eigenvalue behavior will depend on the Jordan form of the matrix. More precisely, if the Jordan form of a matrix A contains an $n \times n$ Jordan block with eigenvalue λ, then an order ϵ dense perturbation ϵB on A will produce n eigenvalues $\lambda_1, \ldots, \lambda_n$, which spread out on a circle centered at λ with radius $O(\epsilon^{1/n})$. This classical result [2, page 109], [66, page 65] is known for a random dense B generically. For the non-generic case, when we impose special forms on B, the situation is very complicated. We proceed to study the various non-generic situations.

In Section 2.2 we explore the case when A is a single Jordan block and B is upper k-Hessenberg. For example, suppose that we perturb an $n \times n$ Jordan block J with a matrix ϵB, where B has the form:

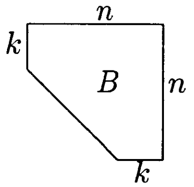


Figure 1: $B =$ upper k-Hessenberg matrix

We assume that k denotes the number of subdiagonals (including the main diagonal itself) that is not set to zero. If B were dense ($B_{71} \neq 0$), the eigenvalues of $J + \epsilon B$ would split uniformly onto a ring of size $n = 7$ and radius $O(\epsilon^{\frac{1}{7}})$. However, if $k = 4$, we obtain one ring of size 4 with radius $O(\epsilon^{\frac{1}{4}})$ and one ring of size 3 with radius $O(\epsilon^{\frac{1}{3}})$ as illustrated

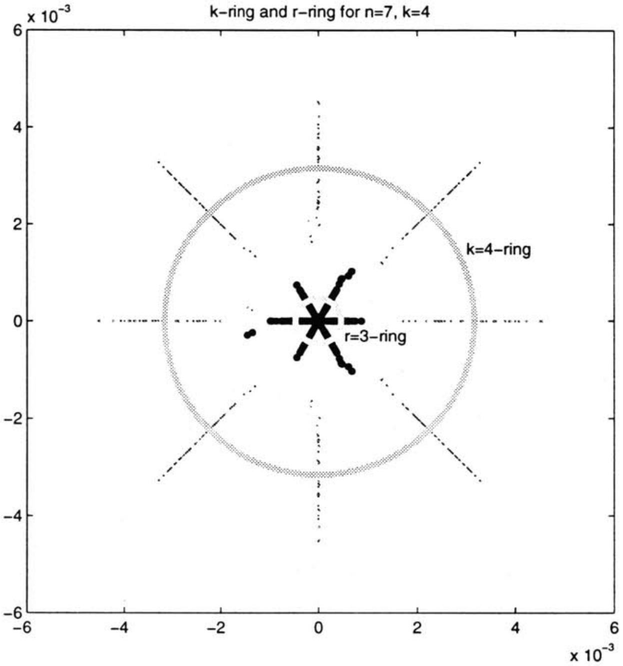


Figure 2: Example rings for $n = 7$ and $k = 4$. We collected eigenvalues of 50 different random $J + \epsilon B$, $\epsilon = 10^{-12}$. The figure represents 50 different copies of one 4-ring (thin dots) and 50 copies of one 3-ring (thick dots). The two circles have radii $O(10^{-3})$ and $O(10^{-4})$. If B were a random dense matrix, there would be only one 7-ring with radius $O(10^{-\frac{12}{7}})$.

in Figure 2. Table 1 contains a table of possible ring sizes when $n = 7$ for $k = 1, \ldots, 7$.

Our main result is that if a Jordan block of size n is perturbed by an upper k-Hessenberg matrix, then the eigenvalues typically split into $\lceil \frac{n}{k} \rceil$ rings, where $p \equiv \lfloor \frac{n}{k} \rfloor$ of them are k-rings with radius $O(\epsilon^{\frac{1}{k}})$, and if k does not divide n, there is typically one remaining r-ring with radius $O(\epsilon^{\frac{1}{r}})$, where $r \equiv n \mod k$. Moreover, the first order perturbation of the pk eigenvalues in the k-rings only depends on the kth diagonal of B.

In Section 2.4, we extend these results to the case of t equally sized Jordan blocks. We only concentrate on the case where all the t blocks have the same eigenvalue λ, since it is well known (see [81]) that the behavior of the perturbation on different eigenvalues splits.

Let $J = \text{Diag}[J_1, J_2, \ldots J_t]$, where the J_i's are $n \times n$ Jordan blocks, and we conformally

		ring size						
		1	2	3	4	5	6	7
k	1	7						
	2	1	3					
	3	1		2				
	4			1	1			
	5		1			1		
	6	1					1	
	7							1

Table 1: Table for one Jordan block of size 7. The entries in each row are the number of rings of a given size when the perturbation is upper k-Hessenberg.

partition

$$B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1t} \\ B_{21} & B_{22} & \dots & B_{2t} \\ \dots & \dots & \dots & \dots \\ B_{t1} & B_{t2} & \dots & B_{tt} \end{bmatrix}.$$

Suppose every B_{ij} is an upper k-Hessenberg matrix. We will show in Theorem 2.2 that generically, the eigenvalues break into $t\lceil\frac{n}{k}\rceil$ rings, tp of them are k-rings and the remaining t are r-rings if k does not divide n. Here, p and r has the same meaning as before. Again, the first order perturbation of the first tpk eigenvalues only depends on the kth diagonal of every B_{ij}.

For example, if

$$J = J_7(\lambda) \oplus J_7(\lambda),$$

	ring size						
k	1	2	3	4	5	6	7
1	14						
2	2	6					
3	2		4				
4			2	2			
5		2			2		
6	2					2	
7							2

Table 2: Table for two blocks, column index represents size of rings and row index value of k. Entries are numbers of rings.

so that $n = 7$ and $t = 2$, our block upper k-Hessenberg matrices have the form

$$B = \left[\begin{array}{ccccccc|ccccccc}
* & * & * & * & * & * & * & * & * & * & * & * & * & * \\
* & * & * & * & * & * & * & * & * & * & * & * & * & * \\
* & * & * & * & * & * & * & * & * & * & * & * & * & * \\
0 & * & * & * & * & * & * & 0 & * & * & * & * & * & * \\
0 & 0 & * & * & * & * & * & 0 & 0 & * & * & * & * & * \\
0 & 0 & 0 & * & * & * & * & 0 & 0 & 0 & * & * & * & * \\
0 & 0 & 0 & 0 & * & * & * & 0 & 0 & 0 & 0 & * & * & * \\
\hline
* & * & * & * & * & * & * & * & * & * & * & * & * & * \\
* & * & * & * & * & * & * & * & * & * & * & * & * & * \\
* & * & * & * & * & * & * & * & * & * & * & * & * & * \\
0 & * & * & * & * & * & * & 0 & * & * & * & * & * & * \\
0 & 0 & * & * & * & * & * & 0 & 0 & * & * & * & * & * \\
0 & 0 & 0 & * & * & * & * & 0 & 0 & 0 & * & * & * & * \\
0 & 0 & 0 & 0 & * & * & * & 0 & 0 & 0 & 0 & * & * & *
\end{array}\right],$$

i.e. $k = 3$, hence $r = 1$.

In this case, the eigenvalues of $J + \epsilon B$ will split into four 3-rings centered at λ with radii $O(\epsilon^{\frac{1}{3}})$, two 1-rings centered at λ with radius $O(\epsilon)$, See Table 2 for a list of possible rings when $k = 1, \dots, 7$ and $n = 7$.

2.2 One Block Case

Suppose that the Jordan form of J is simply one Jordan block. We assume that $J = J_n(0)$, which we will perturb with ϵB, where B has the sparsity structure given in Figure 1.

Definition 2.1 *Suppose a matrix has k subdiagonals that are closest to the main diagonal (including the main diagonal), not zero, then we call the matrix an* ***upper k-Hessenberg matrix****.*

Definition 2.2 *Suppose for ϵ sufficiently small,*

$$\lambda_j = \lambda + c\epsilon^{\frac{1}{k}}\omega^j + o(\epsilon^{\frac{1}{k}}),$$

for $j = 0, 1, \dots, k-1$ and $c \neq 0$. We then refer to the set $\{\lambda_1(\epsilon), \dots, \lambda_k(\epsilon)\}$ as a ***k-ring****. Here $\omega = e^{\frac{2\pi i}{k}}$ and we refer to c as the* ***ring constant****.*

Lemma 2.1 *[66, page 65] Let λ be a multiple eigenvalue of J with multiplicity s, then there will be s eigenvalues of $J+\epsilon B$ grouped in the manner $\{\lambda_{11}(\epsilon), \dots, \lambda_{1s_1}(\epsilon)\}$, $\{\lambda_{21}(\epsilon), \dots, \lambda_{2s_2}(\epsilon)\}$, $\dots$, and in each group i, the eigenvalues admit the Puiseux series*

$$\lambda_{ih}(\epsilon) = \lambda + \alpha_{i1}\omega_i^h \epsilon^{\frac{1}{s_i}} + \alpha_{i2}\omega_i^{2h}\epsilon^{\frac{2}{s_i}} + \dots$$

for $h = 1, \dots, s_i$. Here $\omega_i = e^{\frac{2\pi i}{s_i}}$.

Our Theorem 2.1 shows how the eigenvalues split into rings, and in Corollary 2.1 and 2.2 we analyze the ring constant c. The main idea of the proofs is that only certain terms in the characteristic polynomial of $J + \epsilon B$ influence the ring constants. For the

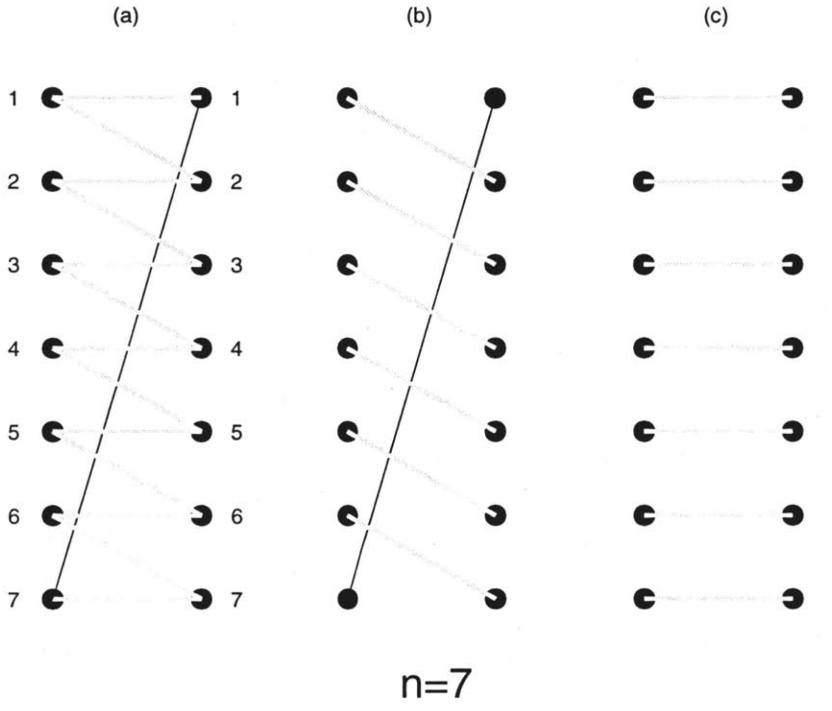


Figure 3: (a) Bipartite graph of $\lambda I - (J_n + \epsilon B)$ (where $J = J_n(0)$). (b) Perfect matching defining $\epsilon B_{n,1}$ term. (c) Perfect matching defining λ^n term.

k-rings, we are interested in the terms from $\det(\lambda I - J - \epsilon B)$ of the form $\sum \alpha_i \epsilon^i \lambda^{n-ki}$ and no higher order terms in ϵ. For the r-ring, we are interested in the $O(\epsilon^{p+1})$ term in $\det(A + \epsilon B)$ and the $O(\epsilon^p)$ term multiplying λ^r in the characteristic polynomial of $A + \lambda B$. All of these may be viewed as determinants with entries removed or as bipartite matchings.

The **bipartite-graph** associated with an $n \times n$ sparse matrix A is a graph on n **left** vertices and n **right** vertices such that non-zero elements a_{ij} are associated with an edge between left node i and right node j. We find it convenient to associate terms in the determinental expansion of $\det(\lambda I - A)$ with subgraphs of the bipartite graph that are perfect matchings.

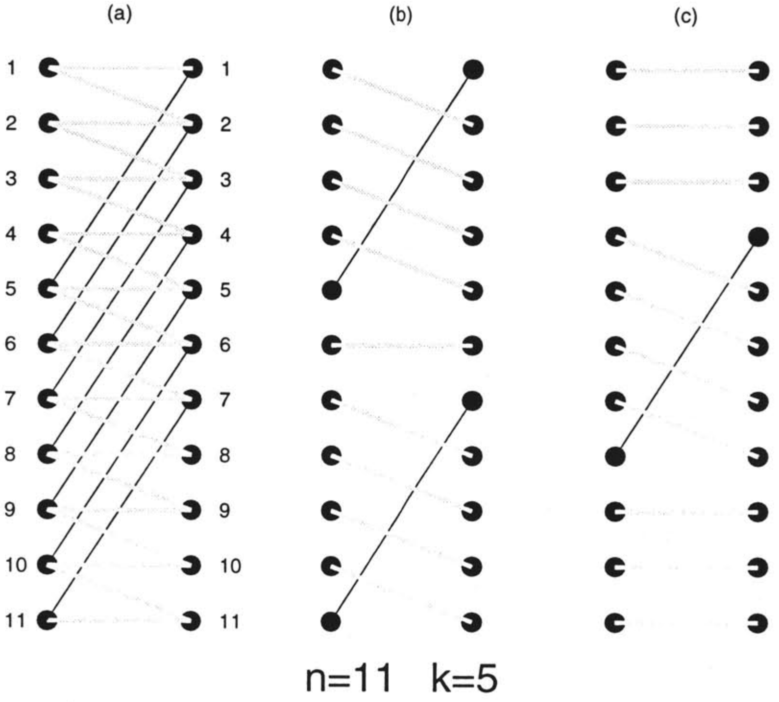


Figure 4: (a) Bipartite Graph of $\lambda I - J - \epsilon B$. (b) Perfect Matching with two laced sections. (c) Perfect Matching with one laced section.

A simple example where $J = J_n(0)$ and B is non-zero only in the $(n, 1)$ entry is plotted in Figure 3. We denote the set in the second column a **laced section**.

Theorem 2.1 *Let J, B, n and k be given as above. Let r be the remainder of n divided by k, i.e. $n = pk + r$, $0 \leq r < k$. The eigenvalues of $J + \epsilon B$ will then generically split into a) p k-rings and b) one r-ring if $r \neq 0$.*

Remark 2.1 *Here, generic means that α_p from Equation (13) and γ from Equation (14) are both not zero. If only $\alpha_p \neq 0$, we have p k-rings, but the r-ring is not guaranteed. Some pathological examples that violate the two generic conditions are given in Section 3.*

Proof:

If B only has elements on the kth subdiagonal, it is easy to study the p k-rings. The situation is illustrated in Figure 4. Every term in the characteristic polynomial must correspond to a union of laced sections of size k and horizontal lines.

We therefore have that

$$\det(\lambda I - (J + \epsilon B)) = \lambda^n + \alpha_1 \epsilon \lambda^{n-k} + \alpha_2 \epsilon^2 \lambda^{n-2k} + \cdots + \alpha_p \epsilon^p \lambda^{n-pk}$$

where

$$\alpha_i = (-1)^i \sum_{l_{j+1} - l_j \geq k} B_{l_1+k-1,l_1} B_{l_2+k-1,l_2} \cdots B_{l_i+k-1,l_i} \tag{3}$$

for $i = 1, \ldots, p$, and $B_{k,1}, \ldots, B_{n,n-k+1}$ denote the elements on the kth diagonal of B. Therefore $J + \epsilon B$ has r eigenvalues equal to 0 up to $\mathrm{O}(\epsilon^{1/k})$ and p k-rings with radii the kth powers of the zeros of

$$q(z) = z^p + \alpha_1 z^{p-1} + \cdots + \alpha_p.$$

Now if B is upper k-Hessenberg and we wish to study the $\mathrm{O}(\epsilon^{1/k})$ eigenvalues, only the lowest subdiagonal elements matter to first order. To see this, it is clear that $\det(\lambda I - (J + \epsilon B))$ has no laced sections of size $> k$ and any of size $< k$ has too many λ's for dominant balance. Therefore, only the laced sections of size k remain.

Alternatively this result may be obtained following the Lidskii approach of letting $\lambda = \mu \epsilon^{1/k}$, $z = \epsilon^{1/k}$, $L_1 = \mathrm{diag}[z^{-1}, z^{-2}, \ldots, z^{-n}]$, $R_1 = \mathrm{diag}[1, z, \ldots, z^{n-1}]$ and studying the limit of $L_1(\mu z I - J - z^k B)R_1$ as $z \to 0$. We write the proof based on this technique in part (a) of the following alternative proof.

We now turn to the r-ring. Readers familiar with the Newton diagram [2, 81] can easily see that one r-ring remains, because the Newton diagram consists of one line segment from $(0, 0)$ to (pk, p) and a second from (pk, p) to $(n, p + 1)$ when $r \neq 0$ (See

Corollary 3 in [81]). Using the bipartite graph approach, it is easy to see that the typical term consists of $p+1$ laced sections each of size at most k. This may also be obtained from a Lidskii style argument as the determinant in Figure 5 and is written out in detail in part (b) of the following alternative proof. □

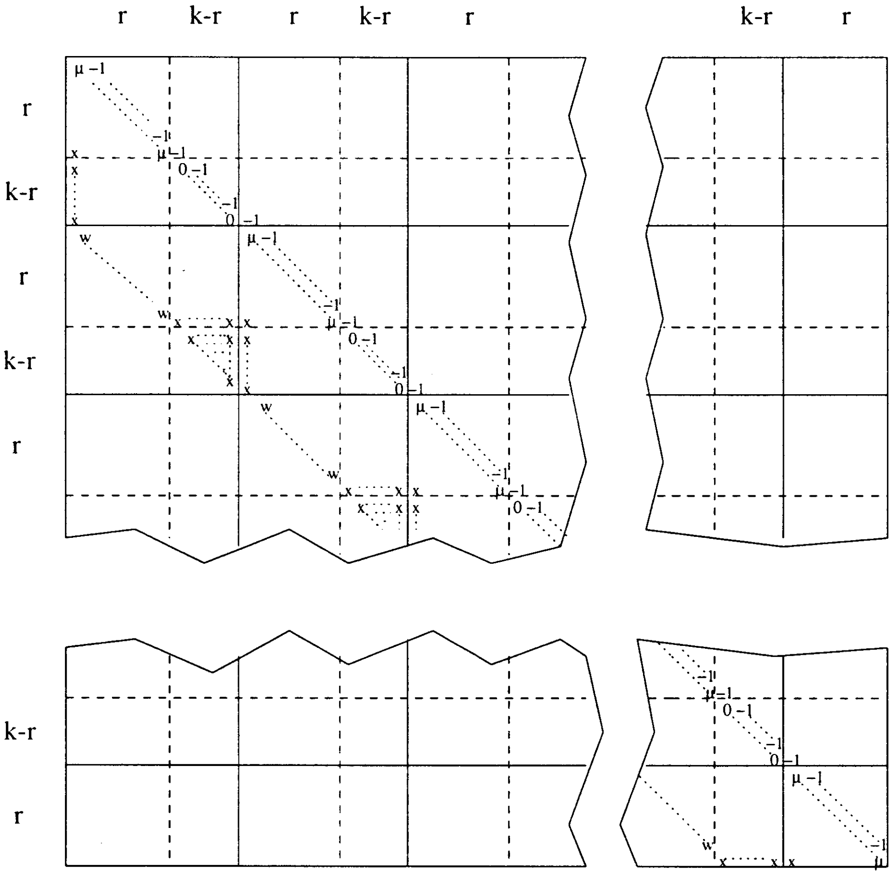


Figure 5: The picture of the effective matrix in Lidskii approach, the picture of $N(0)$ in the alternative proof of Theorem 2.1. The blocks with μ's on the diagonals have sizes $r \times r$. The blocks with 0's on the diagonals have sizes $(k-r) \times (k-r)$. x's and w's represent the original entries of the matrix B at the same position. ...'s represent a repetition of the format.

Alternative Proof:

In part (a) of our proof, we show that the eigenvalues split into p k-rings. In part (b), we prove the statement about the possible existence of one r-ring.

Part (a):

First, we study the p k-rings. In this case, we proceed to show by a change of variables that in fact, only the lowest subdiagonal plays a role in the first order perturbation theory.

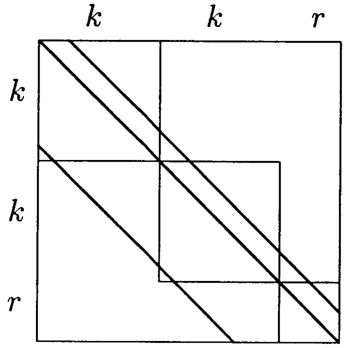


Figure 6: $M(0)$ after it is divided

Let $\lambda = \mu\epsilon^{\frac{1}{k}}$ and $z = \epsilon^{\frac{1}{k}}$. Let

$$L_1 = \mathrm{diag}[z^{-1}, z^{-2}, \ldots, z^{-n}] \tag{4}$$

and

$$R_1 = \mathrm{diag}[1, z^1, \ldots, z^{n-1}] \tag{5}$$

be scaling matrices. Consider $M(z) \equiv L_1(\lambda I - J - \epsilon B)R_1 = (L_1(\mu z I - J - z^k B)R_1)$. At $z = 0$, it has the form

$$M(0) = \begin{bmatrix} \mu & -1 & & & & & \\ & \mu & -1 & & & & \\ & & \cdot & \cdot & & & \\ \ast & & & \cdot & \cdot & & \\ & \ast & & & \cdot & \cdot & \\ & & \ast & & & \mu & -1 \\ & & & \ast & & & \mu \end{bmatrix}. \tag{6}$$

$M(0)$ has only three diagonals, and the kth subdiagonal has the negative of the original entries of B on it.

We claim that $f(\mu) \equiv \det(M(0))$ has the form $\mu^r q(\mu^k)$, where $q(\cdot)$ is a polynomial of order p and its constant term does not vanish generically. Let

$$\omega = e^{2i\pi/k}. \tag{7}$$

Let L_1' be L_1 with z replaced by ω, i.e.

$$L_1' = \text{diag}[\omega^{-1}, \omega^{-2}, \dots, \omega^{-n}],$$

and let R_1' be R_1 with z replaced by ω, i.e.

$$R_1' = \text{diag}[1, \omega^1, \dots, \omega^{n-1}],$$

then

$$\omega^{-n} f(\omega\mu) = \omega^{-1-2-\dots-n} f(\omega\mu)\omega^{0+1+\dots+(n-1)}$$

$$= \det\left(L_1' \begin{bmatrix} \omega\mu & -1 & & & & & \\ & \omega\mu & -1 & & & & \\ & & \ddots & \ddots & & & \\ * & & & \ddots & \ddots & & \\ & * & & & \ddots & \ddots & \\ & & * & & & \omega\mu & -1 \\ & & & * & & & \omega\mu \end{bmatrix} R_1' \right) = \det \begin{bmatrix} \mu & -1 & & & & & \\ & \mu & -1 & & & & \\ & & \ddots & \ddots & & & \\ * & & & \ddots & \ddots & & \\ & * & & & \ddots & \ddots & \\ & & * & & & \mu & -1 \\ & & & * & & & \mu \end{bmatrix} = f(\mu).$$

Therefore $f(\mu) = \omega^{-n} f(\omega\mu) = \omega^{-r} f(\omega\mu)$, from which we can see $f(\mu)$ must be of the form $\mu^r q(\mu^k)$.

We now check that the extreme terms of $f(\cdot)$, of degree n and r, do not vanish. The product of the diagonal entries gives the highest order term μ^n in $f(\mu)$. Now consider the μ^r term. Divide the matrix $M(0)$ into p $k \times k$ diagonal blocks and one $r \times r$ block as in Figure 6. We show that the μ^r term generically does not vanish by considering the coefficient of μ^r term of $f(\mu)$ as a polynomial in the $*$'s entries. In the first p blocks, we take all the -1's and the one element $B_{jk(j-1)k+1}$ at the left bottom corner of the j-th block; in the last block, we take all the μ's. Thus, generically, i.e. unless one of the $B_{jk(j-1)k+1}$ is zero, the constant term of the polynomial q does not vanish. This proves the claim.

Since the polynomial $q(\cdot)$ has p nonzero roots, which we denote $c_1, c_2, \ldots, c_p$, then the polynomial $f(\mu) = \mu^r q(\mu^k)$ has pk non-zero roots distributed evenly on p circles. From the implicit function theorem, there are pk roots of the determinant of the original matrix near $z = 0$ that have the form $\sqrt[k]{c_i} + o(z)$, for $i = 1, 2 \ldots, p$. Note that $\sqrt[k]{c_i}$ yields k different values $\omega^0, \ldots, \omega^{k-1}$ for every i. This shows the pk eigenvalues form p k-rings, this completes the first half of the proof.

Part (b):

Let

$$L_2 = \text{diag}[D_L, z^{-r}I_{k-r}, z^{-r}D_L, z^{-2r}I_{k-r}, \ldots, z^{-pr}D_L], \tag{8}$$

where

$$D_L = \text{diag}[z^{-1}, \ldots, z^{-r}].$$

Let

$$R_2 = \text{diag}[D_R, z^r I_{k-r}, z^r D_R, z^{2r} I_{k-r}, \ldots, z^{pr} D_R], \tag{9}$$

where

$$D_R = \text{diag}[z^0, \ldots, z^{r-1}].$$

Also as with our original proof, we make a change of variables by setting $\lambda = \mu z$ and $z = \epsilon^{\frac{1}{r}}$. Let

$$N(z) = L_2(\lambda I - J - \epsilon B)R_2 = L_2(\mu z I - J - z^r B)R_2.$$

Figure 5 illustrates $N(0)$.

Again we claim that

$$g(\mu) \equiv \det(N(0)) \tag{10}$$

is a polynomial of μ^r. ¡Let

$$\omega = e^{\frac{2\pi i}{r}} \tag{11}$$

and let $L_2'{=}L_2$ with z replaced by ω and $R_2'{=}R_2$ with z replaced by ω. Replace μ in $g(\mu)$ by $\omega\mu$, then we get

$$g(\omega\mu) = \omega^{-r(p+1)} g(\omega\mu) = \det(L_2' N_{\omega\mu}(0) R_2') = g(\mu).$$

Here $N_{\omega\mu}(0)$ represents the matrix $N(0)$ with $\omega\mu$ instead of μ on the main diagonal. Therefore, g is a polynomial of μ^r, say $g(\mu) = h(\mu^r)$. By taking the left bottom element of each of the $k \times k$ diagonal blocks, the left bottom element of the $r \times r$ diagonal block and all the -1's in the remaining rows and columns, we can see that the constant term is generically not zero. By taking the same entries of the first $k \times k$ blocks and all the μ's of the last $r \times r$ block, we generically obtain a nonzero μ^r term. Hence there are at least r roots of $h(\mu^r)$, and they are the rth root of some constant c. By the implicit function theorem there are at least r eigenvalues having the expression $\sqrt[r]{c}\omega^j\epsilon^{\frac{1}{r}} + o(\epsilon^{\frac{1}{r}})$, with $j = 0, \ldots, r-1$. They form an r-ring. This is as many as we can get since we already have pk of the eigenvalues from Part (a). □

Corollary 2.1 *The kth power of the ring constants for the k-rings are the roots of $q(z)$, where*

$$q(z) = z^p + \alpha_1 z^{p-1} + \cdots + \alpha_i z^{p-i} + \cdots + \alpha_p, \tag{12}$$

and

$$\alpha_i = (-1)^i \sum_{l_{j+1}-l_j \geq k} B_{l_1+k-1,l_1} B_{l_2+k-1,l_2} \ldots B_{l_i+k-1,l_i} \tag{13}$$

for $i = 1, \ldots, p$, where $B_{k,1}, \ldots, B_{n,n-k+1}$ denote the elements on the kth diagonal of B. So long as $\alpha_p \neq 0$, we obtain the generic behavior described in Theorem 2.1.

A careful looking at the bipartite graph or Figure 5 also yields the following Corollary.

Corollary 2.2 *The rth power of the ring constant for the r-ring is the root of*

$$\alpha_p z + (-1)^{p+1}\gamma = 0$$

Here, α_p is defined in equation (13), and

$$\gamma = \sum B_{i_1,i_0+1} B_{i_2,i_1+1} \ldots B_{i_{p+1},i_p+1}. \tag{14}$$

$B_{i_1,i_0+1}, B_{i_2,i_1+1}, \ldots, B_{i_{p+1},i_p+1}$ are the entries of B in the x position of figure 5, $i_0 = 0$, $i_{p+1} = n$ and they satisfy

$$r \leq i_{m+1} - i_m \leq k \tag{15}$$

for

$$m = 0, 1, 2, \ldots, p$$

2.3 Pathological Exceptions

When $k > \frac{n}{2}$, $p = 1$, the $q(z)$ in Corollary 2.1 is $q(z) = z + \alpha_1$, where $\alpha_1 = -\sum_{l_1} B_{l_1+k-1,l_1}$. Therefore $\alpha_1 \neq 0$ is the generic case. In such cases, J. Burke and M. Overton([10, Theorem 4]) gave a general result on the characteristic polynomial of $A + \epsilon B$: the coefficient of every term $\epsilon\lambda^i$ is the sum of the elements on the $(n-i)$th subdiagonal for $i = 0, 1, \ldots, n-1$. From this theorem, if we assume that the last subdiagonal that does not sum up to zero is the kth subdiagonal, for $k > \frac{n}{2}$, using a Newton diagram [81, 2](see Figure 7 for an example of a Newton diagram), it can be easily seen that the eigenvalues split into one k-ring and one $(n-k)$-ring.

We can argue similarly for $k < \frac{n}{2}$. When $\alpha_p = 0$, we generically lose one k-ring and the r-ring. Consider the Newton diagram: the (pk, p) point moves up and the whole

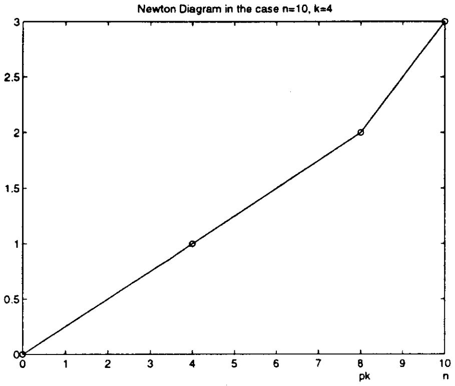


Figure 7: Newton diagram

diagram generically breaks into three segments, one with slope $\frac{1}{k}$, of length $(p-1)k$ and one with slope $\frac{1}{k-1}$, of length $k-1$ and one with slope $\frac{1}{r+1}$, of length $r+1$. This means it has $(p-1)k$ eigenvalues forming $p-1$ k-rings and $k-1$ eigenvalues forming one $(k-1)$-ring and $r+1$ eigenvalues forming an $(r+1)$-ring. There are two special cases when this does not happen. One is when $k-1=r+1$, then the last two segments combine into one segment. The other is when $k-1<r+1$ which can happen when $k=r+1$, and the whole diagram breaks into only two segments, the first one remains untouched, and the second one has slope $\frac{2}{k+r}$, length $k+r$. When γ in Equation (14) is zero, the r-ring will be lost.

The following are three examples that violate the two generic conditions.

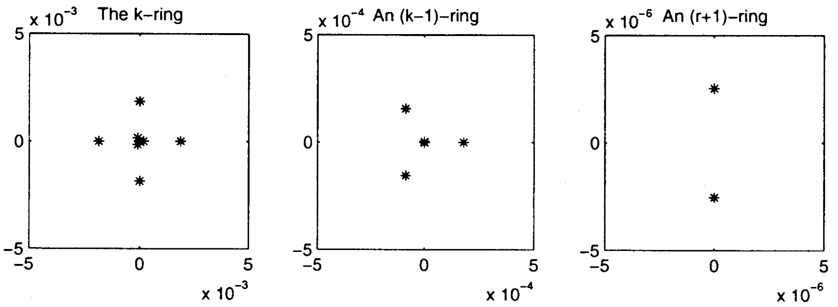


Figure 8: Example 2.1: $\alpha_p = 0$, we lose one k-ring and the r-ring. The last k-ring becomes an $(k-1)$-ring and the remaining $r+1$ eigenvalues form an $(r+1)$-ring.

Example 2.1 $n = 9$, $k = 4$, $p = 2$, $r = 1$,

$$B = \begin{bmatrix} -11 & 1 & -1 & -14 & 22 & 2 & -6 & -13 & -9 \\ -8 & -2 & 2 & -4 & 3 & 10 & -15 & 7 & -10 \\ 4 & -3 & -1 & -5 & 9 & 12 & -1 & -14 & -1 \\ 1 & -7 & 17 & 18 & 7 & -5 & 6 & -13 & -24 \\ 0 & -1 & 16 & 8 & 6 & 9 & 1 & -6 & -7 \\ 0 & 0 & 6 & 1 & 10 & -2 & 16 & -15 & -14 \\ 0 & 0 & 0 & -3 & 13 & -3 & -3 & 6 & 3 \\ 0 & 0 & 0 & 0 & 0 & 5 & 8 & -3 & 6 \\ 0 & 0 & 0 & 0 & 0 & 9 & -8 & -13 & 1 \end{bmatrix}.$$

Example 2.2 $n = 5,\ k = 2,\ p = 2,\ r = 1,$

$$B = \begin{bmatrix} 16 & 2 & 4 & -6 & -10 \\ 4 & -7 & 2 & -7 & 8 \\ 0 & 8 & -17 & -2 & -1 \\ 0 & 0 & -3 & -6 & -4 \\ 0 & 0 & 0 & 1 & -10 \end{bmatrix}.$$

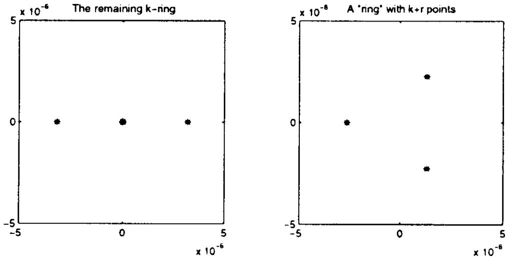


Figure 9: Example 2.2: $\alpha_p = 0$, we lose one k-ring and the r-ring. The remaining $k + r$ eigenvalues still spread evenly on a circle, but they do not form a ring.

Examples 2.1 and 2.2 violate the generic condition of the k-rings, while Example 2.3 violates the generic condition of the r-ring.

Example 2.3 $n = 5,\ k = 3,\ p = 1,\ r = 2,$

$$B = \begin{bmatrix} 1 & 10 & 3 & -5 & 5 \\ -72 & 19 & -6 & -1 & 9 \\ -12 & 17 & -9 & -4 & 26 \\ 0 & 7 & 5 & 6 & -10 \\ 0 & 0 & 1 & -6 & -10 \end{bmatrix}.$$

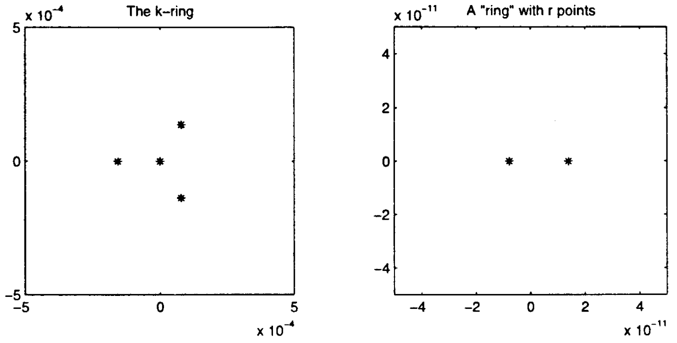


Figure 10: Example 2.3: $\gamma = 0$, the r eigenvalues still spread evenly on a circle, but they do not form a ring.

2.4 t Block Case (All B_{ij}'s are upper k-Hessenberg matrices)

We now study the case when the Jordan form of J has t blocks all with the same size n. We found the case when J has a Jordan structure of different size blocks too complicated for general analysis, though individual cases are easily examined. In this section, we only consider the admittedly special case where the perturbation matrix B has the block upper k-Hessenberg form obtained by dividing B into $n \times n$ blocks and every B_{ij} is an upper k-Hessenberg matrix. In this special case, we have

Theorem 2.2 *Let J, B, n and k be given as above and let r be the remainder of n divided by k, i.e. $n = pk + r$, $0 \leq r < k$. The eigenvalues of $J + \epsilon B$ will then split into tp k-rings and t r-rings if $r \neq 0$.*

Proof: The proof follows closely that of the proof of Theorem 2.1, but we now imagine that B has only elements on the kth subdiagonal of each block. Every term in the characteristic polynomial must correspond to a union of possibly deformed laced sections of size k (see Figure 11) and horizontal lines.

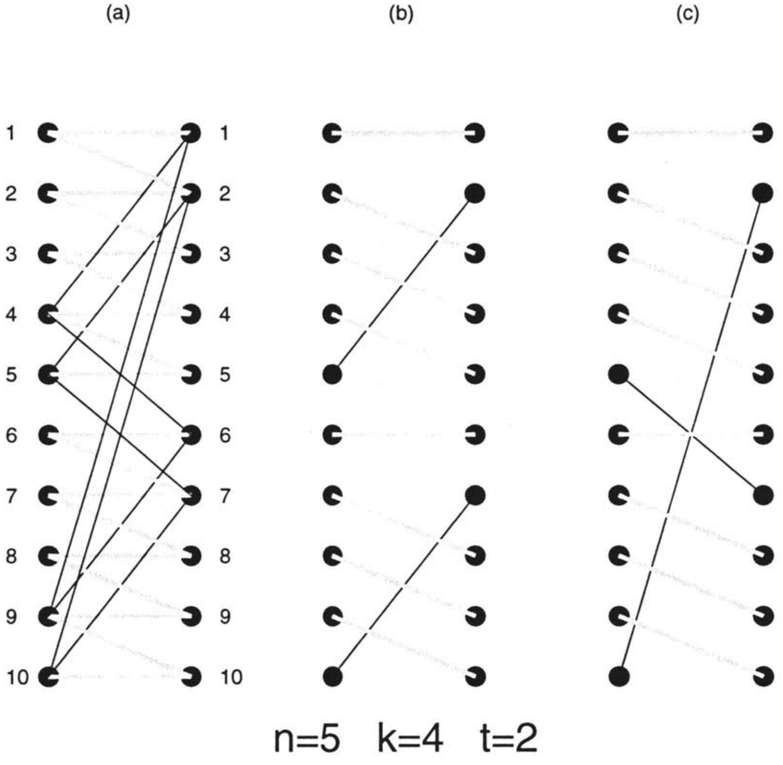


Figure 11: (a) Bipartite graph of $\lambda I - (J + \epsilon B)$ (b) Perfect matching with two laced sections of size 4 contributing to $\epsilon^2\lambda^2$ term. (c) Perfect matching with two deformed laced sections of size 4 contributing to $\epsilon^2\lambda^2$ term.

We therefore have that

$$\det(\lambda I - (J + \epsilon B)) = \lambda^{nt} + \alpha_1 \epsilon \lambda^{nt-k} + \alpha_2 \epsilon^2 \lambda^{nt-2k} + \cdots + \alpha_{pt} \epsilon^{pt} \lambda^{nt-pkt}$$

where instead of Equation (13) in Section 2.2, we have

$$\alpha_i = (-1)^i \sum \det(B_{l_1, l_2, \ldots, l_i})$$

Here, $B_{l_1,l_2,\ldots,l_i}$ represents the matrix formed by extracting the entries at rows $l_1+k-1, l_2+k-1, \ldots, l_i+k-1$ and columns $l_1, l_2, \ldots, l_i$, with $l_{j+1}-l_j \geq k$.

Therefore $J+\epsilon B$ has rt eigenvalues equal to 0 up to $\mathrm{O}(\epsilon^{1/k})$ and pt k-rings with radii the kth powers of the zeros of

$$q(z) = z^{pt} + \alpha_1 z^{pt-1} + \cdots + \alpha_{pt}.$$

Now if B has the block upper k-Hessenberg form and we wish to study the $\mathrm{O}(\epsilon^{1/k})$ eigenvalues, only the lowest subdiagonal elements matter to first order for the same reason as in the proof of Theorem 2.1.

Alternatively, this result may be obtained following the Lidskii approach of letting $\lambda = \mu\epsilon^{1/k}$, $z = \epsilon^{1/k}$,

$$L_1 = \mathrm{diag}[\mathrm{diag}[z^{-1}, z^{-2} \ldots, z^{-n}], \ldots, \mathrm{diag}[z^{-1}, z^{-2}, \ldots, z^{-n}]],$$

$$R_1 = \mathrm{diag}[\mathrm{diag}[1, z, \ldots, z^{n-1}], \ldots, \mathrm{diag}[1, z, \ldots, z^{n-1}]]$$

and studying the limit of $L_1(\mu z I - J - z^k B)R_1$ as $z \to 0$. We write the proof based on this technique in part (a) of the following alternative proof.

We now turn to the r-rings. Although the Newton diagram approach can not be applied in an obvious way here, with the bipartite graph approach or the Lidskii approach, it can be seen that the typical terms consist of $tp+i$ possibly deformed laced sections each of size between r and k, with $i = 0, 1, \ldots, t$. We write out the Lidskii approach proof in part (b) of the following alternative proof. □

Alternative Proof:

The proof follows closely the alternative proof of Theorem 2.1.

Part a:

Let L_1 be a block diagonal matrix with t blocks and every block has the form as in

Equation (4). Let R_1 be a block diagonal matrix with t blocks and every block has the form in Equation (5). Let

$$M(z) = L_1(\lambda I - J - \epsilon B)R_1.$$

Then $M(0)$ breaks into t^2 $n \times n$ blocks. All of the diagonal blocks have the same form as in Equation (6) and the form of the off diagonal blocks results from replacing the μ's and -1's with 0's in the diagonal blocks. Call the resulting matrix $M(0)$. We can reach the same claim that

$$f(\mu) \equiv \det(M(0)) = \mu^{r_0} q(\mu^k), \tag{16}$$

where $r_0 \equiv nt \mod k$. By considering the diagonal blocks we can see that generically the terms μ^{nt} and μ^{rt} appear. This can be shown simply by using the same ω as in Equation (7) and constructing L_1' and R_1' by replacing the z's in L_1 and R_1 with ω's, and going through exactly the same procedure. Thus, we will have at least $nt - rt = tpk$ eigenvalues yielding the form $\sqrt[k]{c_i}\epsilon^{\frac{1}{k}} + o(\epsilon^{\frac{1}{k}})$, $i = 1, ...tp$. Note that every $\sqrt[k]{c_i}$ gives k values. They form pt k-rings.

Part b:

Let L_2 be a block diagonal matrix with t blocks and every block has the form as in Equation (8). Let R_2 be a block diagonal matrix with t blocks and every block has the form as in Equation (9). Let

$$N(z) = L_2(\lambda I - J - \epsilon B)R_2.$$

Then $N(0)$ breaks into t^2 $n \times n$ blocks. All of the diagonal blocks have the same form as $N(0)$ in Figure 5 and the form of the off diagonal blocks results from replacing the μ's and -1's with 0's in the diagonal blocks. We can reach the same claim that

$$g(\mu) \equiv \det(N(0)) = h(\mu^r) \tag{17}$$

and by considering the diagonal blocks we can see that generically the term μ^0 and μ^{rt} appear. This can be shown simply by using the same ω as in Equation (11) and construct L_2' and R_2' by replacing the z's in L_2 and R_2 with ω's, and going through exactly the same procedure. Thus, we will have at least rt eigenvalues yielding the form $\sqrt[r]{c_l}\epsilon^{\frac{1}{r}} + o(\epsilon^{\frac{1}{r}})$, here $l = 1, ...t$. Note that every $\sqrt[r]{c_l}$ gives r values. They form t r-rings.

Since the matrix $J + \epsilon B$ has only nt eigenvalues, it must have exactly tpk and tr of each. This completes the proof of the theorem. □

Corollary 2.3 *The kth power of the ring constants for the k-rings are the roots of* $q(z)$*, where*

$$q(z) = z^{pt} + \alpha_1 z^{pt-1} + \alpha_2 z^{pt-2} + \cdots + \alpha_i z^{pt-i} + \cdots + \alpha_{pt}$$

and

$$\alpha_i = (-1)^i \sum \det(B_{l_1,l_2,\ldots,l_i})$$

where $B_{l_1,l_2,\ldots,l_i}$ *represents the matrix formed by extracting the entries at rows* $l_1 + k - 1, l_2 + k - 1, \ldots, l_i + k - 1$ *and columns* $l_1, l_2, \ldots, l_i$*, with* $l_{j+1} - l_j \geq k$*. So long as* $\alpha_{pt} \neq 0$*, we obtain the generic behavior described in Theorem 2.2.*

Corollary 2.4 *The rth power of the ring constants for the r-rings are the roots of* $g(z)$*, where*

$$g(z) = \gamma_0 z^t + \gamma_1 z^{t-1} + \gamma_2 z^{t-2} + \cdots + \gamma_i z^{t-i} + \cdots + \gamma_t$$

and

$$\gamma_i = (-1)^{pt+i} \sum \det(B_{l_0,l_1,l_2,\ldots,l_{pt+i}})$$

with

$$r \leq i_{m+1} - i_m \leq k$$

for $i = 0, 1, \ldots, t$. *Here,* $B_{l_0,l_1,l_2,\ldots,l_{pt+i}}$ *represents a matrix obtained by extracting the entries on rows* $l_1, l_2, \ldots, l_{pt+i}$ *and columns* $l_0 + 1, l_1 + 1, \ldots, l_{pt+i-1} + 1$ *from the matrix formed by repeating Figure 5* t *times on the diagonal and Figure 5 with* μ*'s and* -1*'s replaced by* 0*'s on the off diagonals, and we have* $l_0 = 0$, $l_{pt+t} = nt$.

2.5 t Block Case (Every B_{ij} is an upper K_{ij}-Hessenberg matrix)

When the number of subdiagonals in each B_{ij} differs, the situation becomes much more complicated, the general problem remains open. We have some observations in two special cases. Let K_{ij} = the number of subdiagonals of B_{ij}, for $1 \leq i, j \leq n$, i.e., B_{ij} is an upper K_{ij}-Hessenberg matrix.

Theorem 2.3

Case 1:

Let $K_{\max} = \max(K_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n$. *If* $K_{11} = K_{22} = \ldots = K_{tt} = K_{\max}$ *then Theorem 2.2 holds upon taking* $k = K_{\max}$.

Case 2:

Let $K_{\max} = \max(K_{ij})$. *If we can find* t K_{ij}*'s equal to* $K_{\max}$ *s.t. no two of them are in the same row or column, then the result from Theorem 2.2 holds upon taking* $k = K_{\max}$.

Case 3:

When $K_{ii} \geq K_{ij}$, $K_{ii} \geq K_{ji}$ for all i and j, and $K_{ii} \geq \frac{n}{2}$, then the resulting eigenvalue behavior looks like putting the t diagonal blocks together, i.e, $J + \epsilon B$ has K_{ii} eigenvalues that form one K_{ii}-ring for $i = 1, \ldots, t$. It also has $n - K_{ii}$ eigenvalues that form one $(n - K_{ii})$-ring for $i = 1, \ldots, t$.

Case 4:

If we can find t numbers $K_{i_1 j_1}, K_{i_2 j_2}, \ldots, K_{i_t j_t}$, all $\geq \frac{n}{2}$, such that $K_{i_s j_s} \geq K_{i_s l}$ and $K_{i_s j_s} \geq K_{m j_s}$ for any l and m, $s = 1, \ldots, t$, and $i_s \neq i_{s'}$, $j_s \neq j_{s'}$, when $s \neq s'$, then $J + \epsilon B$ has $K_{i_s j_s}$ eigenvalues that form one $K_{i_s j_s}$-ring for $s = 1, \ldots, t$. The remaining $n - K_{i_s j_s}$ eigenvalues form an $(n - K_{i_s j_s})$-ring for $s = 1, \ldots, t$.

Proof of Case 1:

This can be checked simply by replacing all the K_{ij}'s with $K_{\max}$ and noticing that the proof of Theorem 2.2 is still valid with k replaced by $K_{\max}$, in that the genericity condition is the same even if some of the off diagonal entries are zero. □

Proof of Case 2:

We also replace all the K_{ij}'s with $K_{\max}$. The proof of Theorem 2.2 with k replaced by $K_{\max}$ remains valid with a minor modification. While some terms in $f(\mu)$ and $g(\mu)$ in equations (16) and (17) as defined in the proof of Theorem 2.2 may be 0 in one block, one can always obtain non-zero terms in each block row and column in the block with $K_{ij} = K_{\max}$. This will guarantee the same nonzero terms generically. □

The following is an example where $t = 2$:

$$\left[\begin{array}{cccccccc|cccccccc}
\mu & -1 & & & & & & & & & & & & & & \\
 & \mu & -1 & & & & & & & & & & & & & \\
\clubsuit & & \mu & -1 & & & & & & \heartsuit & & & & & & \\
 & \clubsuit & & \mu & -1 & & & & & & \heartsuit & & & & & \\
 & & \clubsuit & & \mu & -1 & & & & & & \heartsuit & & & & \\
 & & & \clubsuit & & \mu & -1 & & & & & & \heartsuit & & & \\
 & & & & \clubsuit & & \mu & -1 & & & & & & \heartsuit & & \\
 & & & & & \clubsuit & & \mu & & & & & & & \heartsuit & \\
\hline
 & & & & & & & & \mu & -1 & & & & & & \\
 & & & & & & & & & \mu & -1 & & & & & \\
 & \heartsuit & & & & & & & \clubsuit & & \mu & -1 & & & & \\
 & & \heartsuit & & & & & & & \clubsuit & & \mu & -1 & & & \\
 & & & \heartsuit & & & & & & & \clubsuit & & \mu & -1 & & \\
 & & & & \heartsuit & & & & & & & \clubsuit & & \mu & -1 & \\
 & & & & & \heartsuit & & & & & & & \clubsuit & & \mu & -1 \\
 & & & & & & \heartsuit & & & & & & & \clubsuit & & \mu
\end{array}\right]$$

This is an example with $t = 2$ and $K_{\max} = K_{12} = K_{21}$, in which instead of taking $\clubsuit$'s which may be all zeros, we take $\heartsuit$'s which are nonzero generically.

Proof of Case 3:

For any K_{ii}, let L_{1_i} be a diagonal matrix formed by t blocks of size $n \times n$. For block j, if $K_{jj} \leq K_{ii}$, then the block will be

$$\text{diag}[z^{-1}, z^{-2}, ...z^{-n}],$$

if $K_{jj} \geq K_{ii}$, then the block will be

$$\operatorname{diag}[D_{l_i}, z^{-n+K_{ii}} I_{K_{ii}-n+K_{jj}}, z^{-n+K_{ii}} D_{l_j}], \tag{18}$$

where

$$D_{l_i} = \operatorname{diag}[z^{-1}, \dots, z^{-n+K_{ii}}] \tag{19}$$

and

$$D_{l_j} = \operatorname{diag}[z^{-1}, \dots, z^{-n+K_{jj}}].$$

Let R_{1_i} be a diagonal matrix formed by t blocks of size $n \times n$. For block j, if $K_{jj} \leq K_{ii}$, then the block will be

$$\operatorname{diag}[z^0, z^1, \dots z^{n-1}],$$

if $K_{jj} \geq K_{ii}$, then the block will be

$$\operatorname{diag}[D_{r_i}, z^{K_{ii}} I_{K_{ii}+K_{jj}-n}, z^{K_{ii}} D_{r_{ij}}], \tag{20}$$

where

$$D_{r_i} = \operatorname{diag}[z^0, \dots, z^{K_{ii}-1}], \tag{21}$$

and

$$D_{r_{ij}} = \operatorname{diag}[z^0, \dots, z^{2n-K_{ii}-K_{jj}-1}].$$

Let $\lambda = \mu z$, $z = \epsilon^{\frac{1}{K_{ii}}}$ and $M_i(z) = L_{1_i}(\lambda I - J - \epsilon B) R_{1_i}$. Then $M_i(0)$ is a $t \times t$ block matrix where the jth diagonal block looks like either the $M(0)$ in Equation (6) for $k = K_{ii}$ or the block has the form

$$\begin{bmatrix}
\mu & -1 & & & & & & & & & & \\
 & \cdot & \cdot & & & & & & & & & \\
 & & \mu & -1 & & & & & & & & \\
* & & & 0 & -1 & & & & & & & \\
* & & & & 0 & -1 & & & & & & \\
\cdot & & & & & \cdot & \cdot & & & & & \\
\cdot & & & & & & \cdot & \cdot & & & & \\
* & & & & & & & 0 & 1 & & & \\
 & * & & & & & & & \mu & -1 & & \\
 & & \cdot & & & & & & & \cdot & \cdot & \\
 & & & \cdot & & & & & & & \cdot & \cdot \\
 & & & & * & & & & & & & \mu
\end{bmatrix}$$

Here, the $*$'s on the first column appear from the K_{ii}th row to K_{jj}th row. The off diagonal block $M(0)_{i_{lm}}$ looks the same as in Equation (6) with $k = \max(l, m)$. Replacing z in L_{1_i} and R_{1_i} with ω, which is $e^{\frac{1}{K_{ii}}}$, to get L'_{1_i} and R'_{1_i}, we get the same conclusion that $\det(M_i(0))$ is of the form $\mu^{n-K_{ii}}p(\mu^{K_{ii}})$ and by extracting the constant terms from the diagonal blocks with the new form above and the $\mu^{n-K_{ii}}$ terms and the μ^n terms from all the other diagonal blocks, we get the result that there are at least $t_i K_{ii}$ eigenvalues forming t_i K_{ii}-rings. Here, t_i is the number of times K_{ii} appears on the diagonal.

For any K_{ii}, let L_{2_i} be a diagonal matrix formed by t blocks of size $n \times n$. For block j, if $K_{jj} \leq K_{ii}$, then the block will be

$$\text{diag}[D_{l_i}, z^{-n+K_{ii}} I_{2K_{ii}-n}, z^{-n+K_{ii}} D_{l_i}],$$

where D_{l_i} is given by Equation (19), if $K_{jj} \geq K_{ii}$, then the block will be the same as in Equation (18). Let R_{2_i} be a diagonal matrix formed by t blocks of size $n \times n$. Let

$$D_{ni} = \mathrm{diag}[z^0, \ldots, z^{n-K_{ii}-1}]. \tag{22}$$

For block j, if $K_{jj} \leq K_{ii}$, then the block will be

$$\mathrm{diag}[D_{ni}, z^{n-K_{ii}} I_{2K_{ii}-n}, z^{n-K_{ii}} D_{ni}],$$

if $K_{jj} \geq K_{ii}$, then the block will be

$$\mathrm{diag}[D_{ni}, z^{n-K_{ii}} I_{K_{jj}+K_{ii}-n}, z^{n-K_{ii}} D_{nj}],$$

where D_{nj} and D_{ni} follows the definition in Equation (22). Let $\lambda = \mu z$ and $z = \epsilon^{\frac{1}{n-K_{ii}}}$. It can be checked that $L_{2_i}(\lambda I - J - \epsilon B)R_{2_i}$ at $z = 0$ is a $t \times t$ block matrix $N(0)_i$ while the jth diagonal block looks like

$$\begin{bmatrix}
\mu & -1 & & & & & & & & & \\
 & \cdot & \cdot & & & & & & & & \\
 & & \cdot & \cdot & & & & & & & \\
* & & \mu & -1 & & & & & & & \\
* & & & 0 & -1 & & & & & & \\
\cdot & & & & \cdot & \cdot & & & & & \\
\cdot & & & & & \cdot & \cdot & & & & \\
* & & & & & & 0 & -1 & & & \\
 & * & & & & & & \mu & -1 & & \\
 & & \cdot & & & & & & \cdot & \cdot & \\
 & & & \cdot & & & & & & \cdot & \cdot \\
 & & & & * & & & & & & \mu
\end{bmatrix}.$$

For $K_{jj} \leq K_{ii}$, the $*$ in the first column goes from the $(n - K_{ii})$'th row to the K_{ii}th row, while for $K_{jj} \geq K_{ii}$, the $*$ in first column goes from the $(n - K_{ii})$th row to the K_{jj}th row. For the off diagonal blocks, if $l < m$, then $N(0)_{lm}$ has exactly the same form as $N(0)_{ll}$ with μ and -1 replaced by 0. If $l > m$, then it has the form $M(0)_{mm}$ with only the $*$'s on the first column remaining. Taking L'_{2_i} and R'_{2_i} as L_{2_i} and R_{2_i} with z replaced by $\omega = e^{\frac{2\pi i}{n-K_{ii}}}$, we find that $\det(N_i(0))$ is $f(\mu^{n-K_{ii}})$ and the constant term and $\mu^{(n-K_{ii})t_i}$ term appear generically by inspecting the diagonal blocks only. So $J + \epsilon B$ has at least $(n - K_{ii})t_i$ eigenvalues forming t_i $(n - K_{ii})$-rings. Comparing the total number of eigenvalues of $J + \epsilon B$, we reach the conclusion. □

Proof of Case 4:

This can be proved by treating the $K_{i_1j_1}, K_{i_2j_2}, ...K_{i_tj_t}$ as $K_{11}, K_{22}, ...K_{tt}$'s as in Case 3 and going through the same proof, applying the same permutation as in Case 2. □

Proof of Case 3 and 4:

A proof on a higher level can be given for both Case 3 and Case 4 at the same time. Imagine all $K_{i_1j_1}, K_{i_2j_2}, \ldots, K_{i_tj_t}$ are obtained from the matrices on the main block diagonal and the matrices on the off diagonals are all zero matrices, then the results hold obviously. Now assume the off diagonal matrices are the ones satisfying the conditions in case 2. The way to change the Newton diagram is through getting a new nonzero term from the off-diagonal elements, which is not possible in this case. The result still holds without the assumption that $K_{i_1j_1}, K_{i_2j_2}, \ldots, K_{i_tj_t}$ are obtained from the main block diagonal matrices as long as we notice that we still get the same nonzero terms generically from those off diagonal matrices where the $K_{i_1j_1}, K_{i_2j_2}, \ldots, K_{i_tj_t}$ instead come from diagonal matrices. □

3 Staircase Failures Explained by Orthogonal Versal Forms

3.1 Introduction

Accurately computing Jordan and Kronecker canonical structures of matrices and pencils is an inportant problem in numerical linear algebra. Staircase algorithms regularize the ill-posed problem of computing the Jordan structure of a matrix by attempting to find a nearby matrix with an "interesting" Jordan structure. It does this by making a sequence of rank decisions. The algorithm may also be directed towards a particular structure. We study the failure of the staircase algorithms. we take the geometry approach to view matrices in n^2 dimensional space and pencils in $2mn$ dimensional space to explain these failures. This follows a geometrical program to complement and perhaps replace traditional numerical concepts associated with matrix subspaces.

The most important contributions of this section may be summarized:

- A geometrical explanation of staircase algorithm failures
- Identification of three significant subspaces that decompose matrix or pencil space: $\mathcal{T}_b$,$\mathcal{R}$, $\mathcal{S}$. The most important of these spaces is $\mathcal{S}$, which we choose to call the "staircase invariant space".
- The idea that the staircase algorithm computes an Arnold normal form that is numerically more appropriate than Arnold's "matrices depending on parameters".
- A first order perturbation theory for the staircase algorithm
- Illustration of the theory using an example by Boley [5]

The section is organized as follows: In Section 3.1.1 we introduce concepts that we call **pure, greedy** and **directed** staircase to emphasize subtle distinctions on how the algorithm might be used. Section 3.1.2 contains some important messages that result from the theory to follow.

Section 3.2 presents two similar looking matrices with very different staircase behavior. Section 3.3 studies the relevant n^2 dimensional geometry of matrix space while Section 3.4 applies this theory to the staircase algorithm. The main result may be found in Theorem 3.6.

Sections 3.5, 3.6 and 3.7 mimic Sections 3.2, 3.3 and 3.4 for matrix pencils. Section 3.8 applies the theory towards special cases introduced by Boley [5] and Demmel and B. Kågström [23].

3.1.1 The Staircase Algorithms

Staircase algorithms for the Jordan and Kronecker form work by making sequences of rank decisions in combination with eigenvalue computations. We wish to emphasize a few variations on how the algorithm might be used by coining the terms **pure staircase**, **greedy staircase**, and **directed staircase**. Pseudocode for the Jordan versions appear near the end of this subsection. In combination with these three choices, one can choose an option of **zeroing** or not. These choices are explained below.

The three variations for purposes of discussion are considered in exact arithmetic. The **pure** version is the pure mathematician's algorithm: it gives precisely the Jordan structure of a given matrix. The **greedy** version (also useful for a pure mathematician!) attempts to find the most "interesting" Jordan structure near the given matrix. The **directed** staircase attempts to find a nearby matrix with a preconceived Jordan structure. Roughly speaking, the difference between pure, greedy, and directed is whether the Jordan structure is determined by the matrix, a user controlled neighborhood of the matrix, or directly by the user respectively.

In the **pure** staircase algorithm, rank decisions are made using the singular value decomposition. An explicit distinction is made between zero singular values and nonzero singular values. This determines the exact Jordan form of the input matrix.

The **greedy** staircase algorithm attempts to find the most interesting Jordan structure nearby the given matrix. Here the word "interesting" (or degenerate) is used in the sense of precious gems, the rarer, the more interesting. Algorithmically, as many singular values as possible are thresholded to zero with a user defined threshold. The more singular values that are set to 0, the rarer in the sense of codimension (see [18, 34, 35]).

The **directed** staircase algorithm allows the user to decide in advance what Jordan structure is desired. The Jordan structure dictates which singular values are set to 0. Directed staircase is used in a few special circumstances. For example, it is used when separating the zero Jordan structure from the right singular structure (used in `GUPTRI` [21, 22]). Moreover, Elmroth and Kågström imposed structures by the staircase algorithm in their investigation of the set of 2 × 3 pencils [36]. Recently, Lippert and Edelman [73] use directed staircase to compute an initial guess for a Newton minimization approach to computing the nearest matrix with a given form in the Frobenius norm.

In the greedy and directed modes if we explicitly **zero** the singular values, we end up computing a new matrix in staircase form that has the same Jordan structure as a matrix *near* the original one. If we do not explicitly **zero** the singular values, we end up computing a matrix that is orthogonally similar to the original one (in the absence of roundoff errors), that is *nearly* in staircase form. For example, in `GUPTRI` [22], the choice of whether to zero the singular values is made by the user with an input parameter named `zero` which may be true or false.

To summarize the many choices associated with a staircase algorithm, there are really five distinct algorithms worth considering: the pure algorithm stands on its own, otherwise the two choices of combinatorial structure (greedy and directed) may be paired with the choice to zero or not. Thereby we have the five algorithms:

1. pure staircase
2. greedy staircase with zeroing
3. greedy staircase without zeroing
4. directed staircase with zeroing

5. directed staircase without zeroing

Notice that in the pure staircase, we do not specify zeroing or not, since both will give the same result vacuously.

Of course algorithms run in finite precision. One further detail is that there is some freedom in the singular value calculations which lead to an ambiguity in the staircase form: in the case of unequal singular values, an order must be specified, and when singular values are equal, there is a choice of basis to be made. We will not specify any order for the SVD, except that all singular values considered to be zero appear first.

In the ith loop iteration, we use w_i to denote the number of singular values that are considered to be 0. For the directed algorithm, w_i are input, otherwise, w_i are computed. In pseudocode, we have the following staircase algorithms for computing the Jordan form corresponding to eigenvalue λ.

```
INPUT:
    1) matrix $A$
    2) specify pure, greedy, or direct mode
    3) specify zeroing or not zeroing
OUTPUT:
    1) matrix $A$ that may or may not be in staircase form
    2) $Q$ (optional)
```

```
$i = 0, \quad Q = I$
$A_{tmp} = A - \lambda I$
while $A_{tmp}$ not full rank
    $i = i + 1$
    Let $n' = \sum_{j=1}^{i-1} w_j$ and $n_{tmp} = n - n' = dim(A_{tmp})$
    Use the SVD to compute an $n_{tmp}$ by $n_{tmp}$ unitary matrix $V$ whose leading $w_i$ columns
            span the nullspace or an approximation
        Choice I: Pure: Use the SVD algorithm to compute $w_i$ and the exact nullspace
        Choice II: Greedy: Use the SVD algorithm and threshold the small singular values with
            a user specified tolerance, thereby defining $w_i$. The corresponding singular vectors
            become the first $w_i$ vectors of $V$.
        Choice III: Directed: Use the SVD algorithm, the $w_i$ are defined from the input Jordan
            structure. The $w_i$ singular vectors are the first $w_i$ columns of $V$.
    $A = \mathrm{diag}(I_{n'}, V^*) \cdot A \cdot \mathrm{diag}(I_{n'}, V), \quad Q = Q \cdot \mathrm{diag}(I_{n'}, V)$
    Let $A_{tmp}$ be the lower right $n_{tmp} - w_i$ by $n_{tmp} - w_i$ corner of $A$
    $A_{tmp} = A_{tmp} - \lambda I$
endwhile
If zeroing, return $A$ in the form $\lambda I +$ a block strictly upper triangular matrix.
```

While the staircase algorithm often works very well, it has been known to fail. We can say that the greedy algorithm fails if it does not detect a matrix with the least generic form [18] possible within a given tolerance. We say that the directed algorithm fails if the staircase form it produces is very far (orders of magnitude, in terms of the usual Frobenious norm of matrix space) from the staircase form of the nearest matrix with the intended structure. In this paper, we mainly concentrate on the greedy staircase algorithm and its failure, but the theory is applicable to both approaches. We emphasize that we are intentionally vague about how "far" is "far" as this may be application dependent, but we will consider several orders of magnitude to constitute the notion of "far".

3.1.2 Geometry of Staircase and Arnold Forms

Our geometrical approach is inspired by Arnold's theory of versality [1]. For readers already familiar with Arnold's theory, we point out that we have a new normal form that enjoys the same properties as Arnold's original form, but is more useful numerically. For numerical analysts, we point out that these ideas are important for understanding the staircase algorithm. Perhaps it is safe to say that numerical analysts have had an "Arnold Normal Form" for years, but we did not recognize as such – the computer was doing it for us automatically.

The power of the normal form that we introduce in Section 3.3 is that it provides a first order rounding theory of the staircase algorithm. We will show that instead of decomposing the perturbation space into the normal space and a tangent space at a matrix A, the algorithm chooses a so called staircase invariant space to take the place of the normal space. When some directions in the staircase invariant space are very close to the tangent space, the algorithm can fail.

From the theory, we decompose the matrix space into three subspaces that we call $\mathcal{T}_b$, $\mathcal{R}$ and $\mathcal{S}$, the precise definitions of the three spaces are given in Definitions 3.1 and 3.3. Here, $\mathcal{T}_b$ and $\mathcal{R}$ are two subspaces of the tangent space, and $\mathcal{S}$ is a certain complimentary space of the tangent space in the matrix space. For the impatient reader, we point out that angles between these spaces are related to the behavior of the staircase algorithm; note that $\mathcal{R}$ is always orthogonal to $\mathcal{S}$. (We use $< \cdot, \cdot >$ to represent the angle between two spaces.)

		angles			components	
A	Staircase fails	$< \mathcal{S}, \mathcal{T}_b \oplus \mathcal{R} >$	$< \mathcal{T}_b, \mathcal{R} >$	$< \mathcal{S}, \mathcal{R} >$	$\mathcal{S}$	$\mathcal{R}$
no weak stair	no	large	large	$\pi/2$	small	small
weak stair	no	large	small	$\pi/2$	small	large
weak stair	yes	small	small	$\pi/2$	large	large

Here, by a weak stair [31], we mean the near rank deficiency of any superdiagonal block of the strictly block upper triangular matrix A.

3.2 A Staircase Algorithm Failure to Motivate the Theory

Consider the two matrices

$$A_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & \delta \\ 0 & 0 & 0 \end{pmatrix} \text{ and } A_2 = \begin{pmatrix} 0 & \delta & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

where δ =`1.5e-9` is approximately on the order of the square root of the double precision machine $\epsilon = 2^{-52}$, roughly `2.2e-16`. Both of these matrices clearly have the Jordan structure $J_3(0)$, but the staircase algorithm on A_1 and A_2 can behave very differently.

To test this, we used the `GUPTRI` [22] algorithm. `GUPTRI` [2] requires an input matrix A and two tolerance parameters `EPSU` and `GAP`. We ran `GUPTRI` on $\tilde{A}_1 \equiv A_1 + \epsilon \mathrm{E}$ and $\tilde{A}_2 \equiv A_2 + \epsilon \mathrm{E}$, where

$$\mathrm{E} = \begin{pmatrix} .3 & .4 & .2 \\ .8 & .3 & .6 \\ .4 & .9 & .6 \end{pmatrix}$$

and ϵ = `2.2e-14` is roughly 100 times the double precision machine ϵ. The singular values of each of the two matrices $\tilde{A}_1$ and $\tilde{A}_2$ are σ_1 = `1.0000e00`, σ_2 = `1.4901e-09` and σ_3 = `8.8816e-15`. We set `GAP` to be always ≥ 1, and let `EPSU` $= a/(\|\tilde{A}_i\| *$ `GAP`), where we vary the value of a (The tolerance is effectively a). Our observations are tabulated below.

[2] GUPTRI [21, 22] is a "greedy" algorithm with a sophisticated thresholding procedure based on two input parameters EPSU and GAP ≥ 1. We threshold σ_{k-1} if $\sigma_{k-1} < GAP \times \max(\sigma_k, EPSU \times \|A\|)$ (Defining $\sigma_{n+1} \equiv 0$). The first argument of the maximum σ_k ensures a large gap between thresholded and non-thresholded singular values. The second argument ensures that σ_{k-1} is small. Readers who look at the GUPTRI software should note that singular values are ordered from smallest to largest, contrary to modern convention.

a	computed Jordan Structure for $\tilde{A}_1$	computed Jordan Structure for $\tilde{A}_2$
$a \geq \sigma_2$	$J_2(0) \oplus J_1(0) + O(10^{-9})$	$J_2(0) \oplus J_1(0) + O(10^{-9})$
$\gamma \leq a < \sigma_2$	$J_3(0) + O(10^{-6})$ ☹	$J_3(0) + O(10^{-14})$
$a < \gamma$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$

Here, we use $J_k(\lambda)$ to represent a $k \times k$ Jordan block with eigenvalue λ. In the table, typically $\alpha \neq \beta \neq 0$. Setting a small (smaller than $\gamma =$ 1.9985e-14 here, which is the smaller singular value in the second stage), the software returns two nonzero singular values in the first and second stages of the algorithm and one nonzero singular value in the third stage. Setting EPSU × GAP large (larger than σ_2 here), we zero two singular values in the first stage and one in the second stage giving the structure $J_2(0) \oplus J_1(0)$ for both $\tilde{A}_1$ and $\tilde{A}_2$ (There is a matrix within $O(10^{-9})$ of A_1 and A_2 of the form $J_2(0) \oplus J_1(0)$). The most interesting case is in between. For appropriate EPSU × GAP $\approx a$ (between γ and σ_2 here), we zero one singular value in each of the three stages, getting a $J_3(0)$ which is $O(10^{-14})$ away for A_2, while we can only get a $J_3(0)$ which is $O(10^{-6})$ away for A_1. In other words, the staircase algorithm fails for A_1 but not for A_2. As pictured in Figure 12, the A_1 example indicates that a matrix of the correct Jordan structure may be within the specified tolerance, but the staircase algorithm may fail to find it.

Consider the situation when A_1 and A_2 are transformed using a random orthogonal matrix Q. As a second experiment, we pick

$$Q \approx \begin{pmatrix} -.39878 & .20047 & -.89487 \\ -.84538 & -.45853 & .27400 \\ -.35540 & .86577 & .35233 \end{pmatrix},$$

and take $\tilde{A}_1 = Q(A_1 + \epsilon E)Q^T$, $\tilde{A}_2 = Q(A_2 + \epsilon E)Q^T$. This will impose a perturbation of order ϵ. We ran GUPTRI on these two matrices; the following is the result:

a	computed Jordan Structure for $\tilde{A}_1$	computed Jordan Structure for $\tilde{A}_2$
$a \geq \sigma_2$	$J_2(0) \oplus J_1(0) + O(10^{-5})$	$J_2(0) \oplus J_1(0) + O(10^{-6})$
$\gamma \leq a < \sigma_2$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta)$ ☹	$J_3(0) + O(10^{-6})$
$a < \gamma$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$

In the table, $\gamma =$ `2.6980e-14`, all other values are the same as in the previous table.

In this case, `GUPTRI` is still able to detect a J_3 structure for $\tilde{A}_2$, although the one it finds is $O(10^{-6})$ away. But it fails to find any J_3 structure at all for $\tilde{A}_1$. The comparison of A_1 and A_2 in the two experiments indicates that the explanation is more subtle than the notion of a weak stair (a superdiagonal block that is almost column rank deficient) [31].

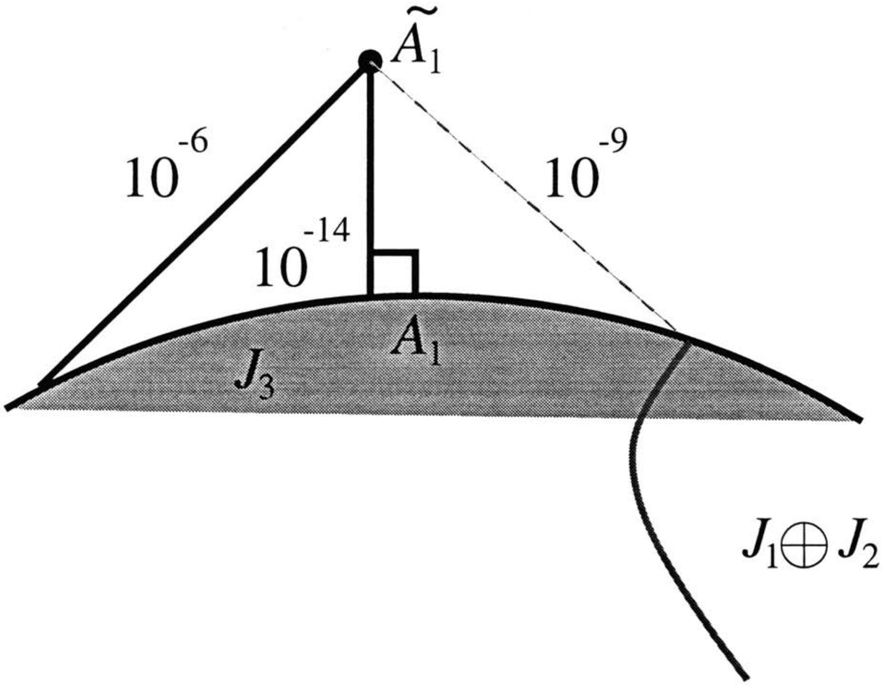


Figure 12: The staircase algorithm fails to find A_1 at distance 2.2e-14 from $\tilde{A}_1$ but does find a $J_3(0)$ or a $J_2(0) \oplus J_1(0)$ if given a much larger tolerance. (The latter is δ away from $\tilde{A}_1$.)

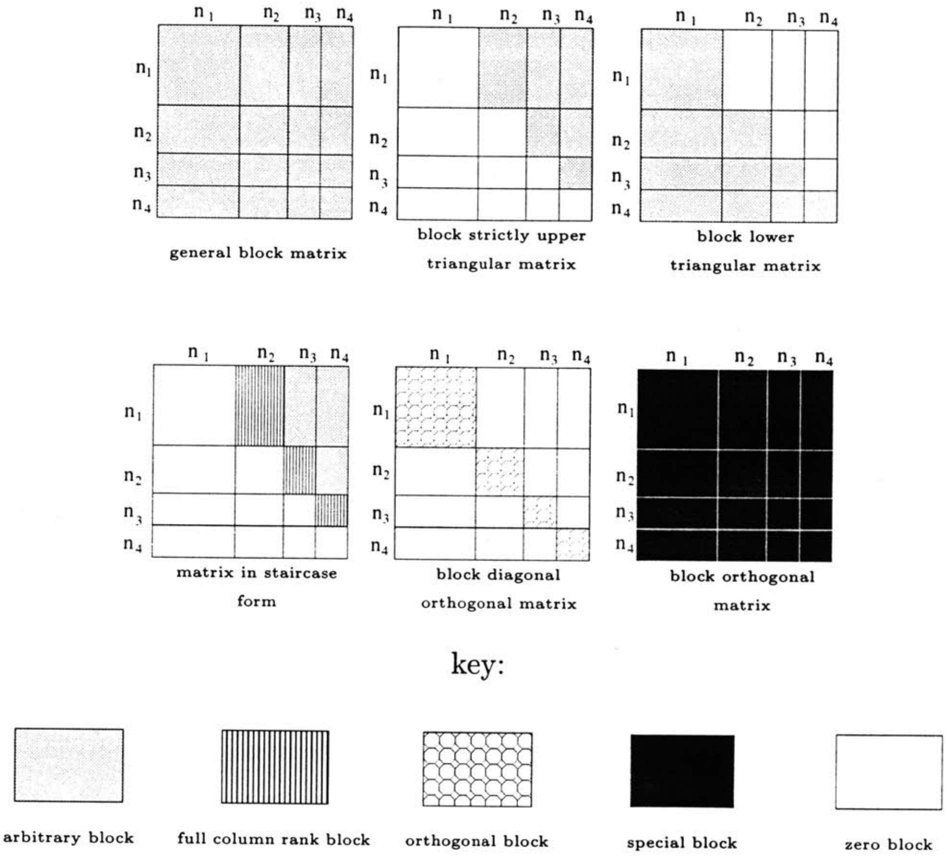
In this paper we present a geometrical theory that clearly predicts the difference between A_1 and A_2. The theory is based on how close certain directions that we will denote **staircase invariant directions** are to the tangent space of the manifold of matrices similar to the matrix with specified canonical form. It turns out that for A_1, these directions are nearly in the tangent space, but not for A_2. This is the crucial difference!

The tangent directions and the staircase invariant directions combine to form a "versal deformation" in the sense of Arnold [1], but one with more useful properties for our purposes.

3.3 Staircase Invariant Space and Versal Deformations

3.3.1 The Staircase Invariant Space and Related Subspaces

We consider block matrices as in Figure 13. Dividing a matrix A into blocks of row and column sizes $n_1, \ldots, n_k$, we obtain a **general block matrix**. A block matrix is **conforming to** A if it is also partitioned into blocks of size $n_1, \ldots, n_k$ in the same manner as A. If a general block matrix has non-zero entries only in the upper triangular blocks excluding the diagonal blocks, we call it a **block strictly upper triangular matrix**. If a general block matrix has non-zero entries only in the lower triangular blocks including the diagonal blocks, we call it a **block lower triangular matrix**. A matrix A is in **staircase form** if we can divide A into blocks of sizes $n_1 \geq n_2 \geq \cdots \geq n_k$ s.t. A is a strictly block upper triangular matrix and every superdiagonal block has full column rank. If a general block matrix only has nonzero entries on its diagonal blocks, and each diagonal block is an orthogonal matrix, we call it a **block diagonal orthogonal matrix**. We call the matrix e^B a **block orthogonal matrix (conforming to** A**)** if B is a block anti-symmetric matrix (conforming to A) (i.e. B is anti-symmetric with zero diagonal blocks. Here, we abuse the word "conforming" since e^B does not have a block structure.)



general block matrix

block strictly upper triangular matrix

block lower triangular matrix

matrix in staircase form

block diagonal orthogonal matrix

block orthogonal matrix

Figure 13: A schematic of the block matrices defined in the text.

Definition 3.1 *Suppose A is a matrix in staircase form. We call S a* **staircase invariant matrix** *of A if $S^TA = 0$ and S is block lower triangular. We call the space of matrices consisting of all such S the* ***staircase invariant space of*** *A, and denote it by $\mathcal{S}$.*

We remark that the columns of S will not be independent except possibly when $A = 0$; S can be the zero matrix as an extreme case. However the generic sparsity structure of S may be determined by the sizes of the blocks. For example, let A have the staircase form

$$A = \begin{pmatrix} 0&0&0&\times&\times&\times&\times&\times \\ 0&0&0&\times&\times&\times&\times&\times \\ 0&0&0&\times&\times&\times&\times&\times \\ &&&0&0&\times&\times&\times \\ &&&0&0&\times&\times&\times \\ &&&&&0&0&\times \\ &&&&&0&0&\times \\ &&&&&&&0 \end{pmatrix}, \text{then } S = \begin{pmatrix} \times&\times&\times&&&&& \\ \times&\times&\times&&&&& \\ \times&\times&\times&&&&& \\ \times&\times&\times&\circ&\circ&&& \\ \times&\times&\times&\circ&\circ&&& \\ \times&\times&\times&\times&\times&\times&\times& \\ \times&\times&\times&\times&\times&\times&\times& \\ \times&\times&\times&\times&\times&\times&\times&\times \end{pmatrix}$$

is a staircase invariant matrix of A if every column of S is a left eigenvector of A. Here, the $\circ$ notation indicates 0 entries in the block lower triangular part of S that are a consequence of the requirement that every column be a left eigenvector. This may be formulated as a general rule: if we find more than one block of size $n_i \times n_i$ then only those blocks on the lowest block row appear in the sparsity structure of S. For example, the $\circ$ do not appear because they are above another block of size 2. As a special case, if A is strictly upper triangular, then S is 0 above the bottom row as is shown below. Readers familiar with Arnold's normal form will notice that if A is a given single Jordan block in normal form, then S contains the versal directions.

$$A = \begin{pmatrix} &\times&\times&\times&\times&\times&\times \\ &&\times&\times&\times&\times&\times \\ &&&\times&\times&\times&\times \\ &&&&\times&\times&\times \\ &&&&&\times&\times \\ &&&&&&\times \\ &&&&&&& \end{pmatrix}, \quad S = \begin{pmatrix} &&&&&& \\ &&&&&& \\ &&&&&& \\ &&&&&& \\ &&&&&& \\ &&&&&& \\ \times&\times&\times&\times&\times&\times&\times \end{pmatrix}.$$

Definition 3.2 *Suppose A is a matrix. We call $\mathcal{O}(A) \equiv \{XAX^{-1} : X$ is a non-singular matrix$\}$ the* **orbit** *of a matrix A. We call $\mathcal{T} \equiv \{AX - XA : X$ is any matrix$\}$ the* **tangent space** *of $\mathcal{O}(A)$ at A.*

Theorem 3.1 *Let A be an $n \times n$ matrix in staircase form, then the staircase invariant space $\mathcal{S}$ of A and the tangent space $\mathcal{T}$ form an oblique decomposition of $n \times n$ matrix space, i.e. $\mathbb{R}^{n^2} = \mathcal{S} \oplus \mathcal{T}$.*

Proof:
Assume that $A_{i,j}$, the (i,j) block of A, is $n_i \times n_j$ for $i,j = 1,\dots,k$ and of course $A_{i,j} = 0$ for all $i \leq j$.

There are n_1^2 degrees of freedom in the first block column of S because there are n_1 columns and each column may be chosen from the n_1 dimensional space of left eigenvectors of A. Indeed there are n_i^2 degrees of freedom in the ith block, because each of the n_i columns may be chosen from the n_i dimensional space of left eigenvectors of the matrix obtained from A by deleting the first $i-1$ block rows and columns. The total number of degrees of freedom is $\sum_{i=1}^{k} n_i^2$, which combined with $\dim(\mathcal{T}) = n^2 - \sum_{i=1}^{k} n_i^2$ [18], gives the dimension of the whole space n^2.

If $S \in \mathcal{S}$ is also in $\mathcal{T}$ then S has the form $AX - XA$ for some matrix X. Our first step will be to show that X must have block upper triangular form after which we will conclude that $AX - XA$ is strictly block upper triangular. Since S is block lower triangular, it will then follow that if it is also in $\mathcal{T}$, it must be 0.

Let i be the first block column of X which does not have block upper triangular structure. Clearly the ith block column of XA is 0 below the diagonal block, so that the ith block column of $S = AX - XA$ contains vectors in the column space of A. However every column of S is a left eigenvector of A from the definition (notice that we do not require these column vectors of S to be independent), and therefore orthogonal to the column space of A. Thus the ith block column of S is 0, and from the full column rank conditions on the superdiagonal blocks of A, we conclude that X is 0 below the block diagonal. □

Definition 3.3 *Suppose A is a matrix. We call $\mathcal{O}_b(A) \equiv \{Q^T A Q : Q = e^B, B$ is a block anti-symmetric matrix conforming to $A\}$ the* **block orthogonal-orbit** *of a matrix A. We call $\mathcal{T}_b \equiv \{AX - XA : X$ is a block anti-symmetric matrix conforming to $A\}$ the* **block tangent space** *of the block orthogonal orbit $\mathcal{O}_b(A)$ at A. We call $\mathcal{R} \equiv \{$ block strictly upper triangular matrix conforming to $A\}$ the* **strictly upper block space** *of A.*

Note that because of the complementary structure of the two matrices R and S, we can see that $\mathcal{S}$ is always orthogonal to $\mathcal{R}$.

Theorem 3.2 *Let A be an $n \times n$ matrix in staircase form, then the tangent space $\mathcal{T}$ of the orbit $\mathcal{O}(A)$ can be split into the block tangent space $\mathcal{T}_b$ of the orbit $\mathcal{O}_b(A)$ and the strictly upper block space $\mathcal{R}$, i.e. $\mathcal{T} = \mathcal{T}_b \oplus \mathcal{R}$.*

Proof:
We know that the tangent space $\mathcal{T}$ of the orbit at A has dimension $n^2 - \sum_{i=1}^{k} n_i^2$. If we decompose X into a block upper triangular matrix and a block anti-symmetric matrix, we can decompose every $AX - XA$ into a block strictly upper triangular matrix and a matrix in $\mathcal{T}_b$. Since $\mathcal{T} = \mathcal{T}_b + \mathcal{R}$, each of $\mathcal{T}_b$ and $\mathcal{R}$ has dimension $\leq 1/2(n^2 - \sum_{i=1}^{k} n_i^2)$, they must both be exactly of dimension $1/2(n^2 - \sum_{i=1}^{k} n_i^2)$. Thus we know that they actually form a decomposition of $\mathcal{T}$, and the strictly upper block space $\mathcal{R}$ can also be represented as $\mathcal{R} \equiv \{AX - XA : X \text{ is block upper triangular matrix conforming to } A\}$.
□

Corollary 3.1 $\mathbb{R}^{n^2} = \mathcal{T}_b \oplus \mathcal{R} \oplus \mathcal{S}$. *See Figure 14.*

In Definition 3.3, we really do not need the whole set $\{e^B : B$ is block antisymmetric $\} \equiv \{e^B\}$, we merely need a small neighborhood around $B = 0$. Readers may well wish to skip ahead to Section 3.4, but for those interested in mathematical technicalities we review a few simple concepts. Suppose that we have partitioned $n = n_1 + \ldots + n_k$. An orthogonal decomposition of n-dimensional space into k mutually orthogonal subspaces of dimensions $n_1, n_2, \ldots, n_k$ is a point on the **flag manifold**. (When $k = 2$ this is the **Grassmann manifold**). Equivalently, a point on the flag manifold is specified by a filtration, i.e., a nested sequence of subspaces V_i of dimension $n_1 + \ldots + n_i$ $(i = 1, \ldots, k)$:

$$0 \subset V_1 \subset \cdots \subset V_k = \mathbb{C}^n.$$

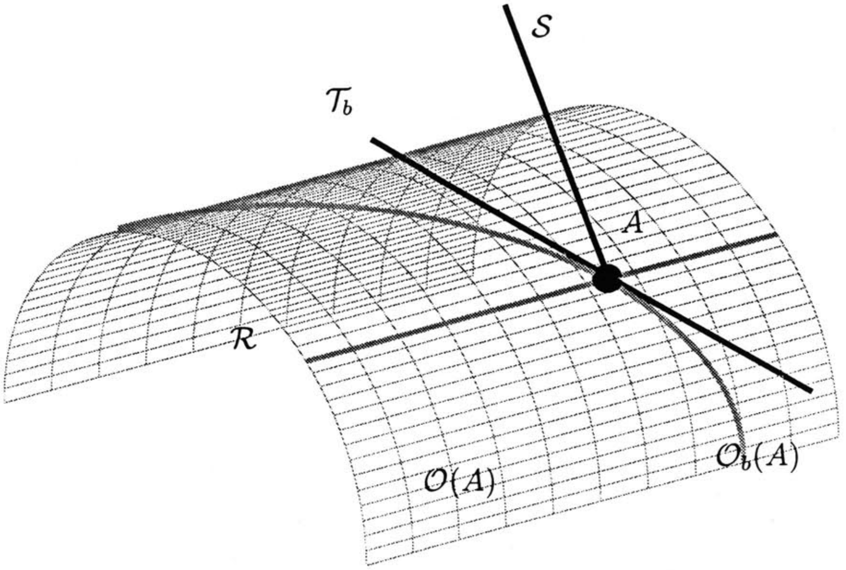


Figure 14: A diagram of the orbits and related spaces. The similarity orbit at A is indicated by a surface $\mathcal{O}(A)$, the block orthogonal orbit is indicated by a curve $\mathcal{O}_b(A)$ on the surface, the tangent space of $\mathcal{O}_b(A)$, $\mathcal{T}_b$ is indicated by a line, $\mathcal{R}$ which lies on $\mathcal{O}(A)$ is pictured as a line too, and the staircase invariant space $\mathcal{S}$ is represented by a line pointing away from the plane.

The corresponding decomposition can be written as

$$\mathbb{C}^n = V_k = V_1 \oplus V_2 \backslash V_1 \oplus \cdots \oplus V_k \backslash V_{k-1}.$$

This may be expressed concretely. If from a unitary matrix U, we only define V_i for $i = 1, \ldots, k$ as the span of the first $n_1 + n_2 + \ldots + n_i$ columns, then we have $V_1 \subset \cdots \subset V_k$, i.e., a point on the flag manifold. Of course many unitary matrices U will correspond to the same flag manifold point. In an open neighborhood of $\{e^B\}$, near the point $e^0 = I$, the map between $\{e^B\}$ and an open subset of the flag manifold is a one to one homeomorphism. The former set is referred to as a local cross section [54, Lemma 4.1, page 123] in Lie algebra. No two unitary matrices in a local cross section would have the same sequence of subspaces $V_i, i = 1, \dots, k$.

3.3.2 Staircase as a Versal Deformation

Next, we are going to build up the theory of our versal form. Following Arnold [1], a **deformation** of a matrix A, is a matrix $A(\lambda)$ with entries that are power series in the complex variables λ_i, where $\lambda = (\lambda_1, \dots, \lambda_k)^T \in \mathbb{C}^k$, convergent in a neighborhood of $\lambda = 0$, with $A(0) = A$.

A good introduction to versal deformations may be found in [1, Section 2.4] or [34]. The key property of a versal deformation is that it has enough parameters so that no matter how the matrix is perturbed, it may be made equivalent by analytic transformations to the versal deformation with some choice of parameters. The advantage of this concept for a numerical analyst is that we might make a rounding error in any direction and yet still think of this as a perturbation to a standard canonical form.

Let $N \subset M$ be a smooth submanifold of a manifold M. We consider a smooth mapping $A : \Lambda \to M$ of another manifold Λ into M, and let λ be a point in Λ such that $A(\lambda) \in N$. The mapping A is called **transversal** to N at λ if the tangent space to M at $A(\lambda)$ is the direct sum

$$TM_{A(\lambda)} = A_* T\Lambda_\lambda \oplus TN_{A(\lambda)}.$$

Here, $TM_{A(\lambda)}$ is the tangent space of M at $A(\lambda)$, $TN_{A(\lambda)}$ is the tangent space of N at $A(\lambda)$, $T\Lambda_\lambda$ is the tangent space of Λ at λ and A_* is the mapping from $T\Lambda_\lambda$ to $TM_{A(\lambda)}$ induced by A (It is the Jacobian).

Theorem 3.3 *Suppose A is in staircase form. Fix $S_i \in \mathcal{S}$, $i = 1, \dots, k$ s.t. span$\{S_i\} = \mathcal{S}$ and $k \geq$ dim$(\mathcal{S})$. It follows that*

$$A(\lambda) \equiv A + \sum_i \lambda_i S_i \tag{23}$$

is a versal deformation of every particular $A(\lambda)$ for λ small enough. $A(\lambda)$ is **mini***versal at $\lambda = 0$ if $\{S_i\}$ is a basis of $\mathcal{S}$.*

Proof:

Theorem 3.1 tells us the mapping $A(\lambda)$ is transversal to the orbit at A. From the equivalence of transversality and versality [1], we know that $A(\lambda)$ is a versal deformation of A. Since the dimension of the staircase invariant space $\mathcal{S}$ is the codimension of the orbit, $A(\lambda)$ given by Equation (23) is a miniversal deformation if the S_i are a basis for $\mathcal{S}$ (i.e. $k = \dim(\mathcal{S})$). More is true, $A(\lambda)$ is a versal deformation of every matrix in a neighborhood of A, in other words, the space $\mathcal{S}$ is transversal to the orbit of every $A(\lambda)$. Take a set of matrices X_i s.t. the $X_iA - AX_i$ form a basis of the tangent space $\mathcal{T}$ of the orbit at A. We know $\mathcal{T} \oplus \mathcal{S} = \mathbb{R}^{n^2}$, here $\oplus$ implies $\mathcal{T} \cap \mathcal{S} = 0$ so there is a fixed minimum angle θ between $\mathcal{T}$ and $\mathcal{S}$. For small enough λ, we can guarantee that the $X_iA(\lambda) - A(\lambda)X_i$ are still linearly independent of each other and they span a subspace of the tangent space at $A(\lambda)$ that is at least, say, $\theta/2$ away from $\mathcal{S}$. This means that the tangent space at $A(\lambda)$ is transversal to $\mathcal{S}$. □

Arnold's theory concentrates on general similarity transformations. As we have seen above, the staircase invariant directions are a perfect versal deformation. This idea can be refined to consider similarity transformations that are block orthogonal. Everything is the same as above, except that we add the block strictly upper triangular matrices R to compensate for the restriction to block orthogonal matrices. We now spell this out in detail:

Definition 3.4 *If the matrix $C(\lambda)$ is block orthogonal for every λ, then we refer to the deformation as a* **block orthogonal deformation.**

We say that two deformations $A(\lambda)$ and $B(\lambda)$ are **block orthogonally-equivalent** *if there exists a block orthogonal deformation $C(\lambda)$ of the identity matrix such that $A(\lambda) = C(\lambda)B(\lambda)C(\lambda)^{-1}$.*

We say that a deformation $A(\lambda)$ is **block orthogonally-versal** *if any other defor-*

mation $B(\mu)$ is block orthogonally-equivalent to the deformation $A(\phi(\mu))$. Here, ϕ is a mapping analytic at 0 with $\phi(0) = 0$.

Theorem 3.4 *A deformation $A(\lambda)$ of A is block orthogonally-versal iff the mapping $A(\lambda)$ is transversal to the block orthogonal-orbit of A at $\lambda = 0$.*

Proof:

The proof follows Arnold [1, Sections 2.3 and 2.4] except that we use the block orthogonal version of the relevant notions, and we remember that the tangents to the block orthogonal group are the commutators of A with the block anti-symmetric matrices. □

Since we know that $\mathcal{T}$ can be decomposed into $\mathcal{T}_b \oplus \mathcal{R}$, we get:

Theorem 3.5 *Suppose a matrix A is in staircase form. Fix $S_i \in \mathcal{S}, i = 1, \ldots, k$ s.t. span$\{S_i\} = \mathcal{S}$ and $k \geq$ dim$(\mathcal{S})$. Fix $R_j \in \mathcal{R}, j = 1, \ldots, l$ s.t. span$\{R_j\} = \mathcal{R}$ and $l \geq$ dim$(\mathcal{R})$. It follows that*

$$A(\lambda) \equiv A + \sum_i \lambda_i S_i + \sum_j \lambda_j R_j$$

*is a block orthogonally-versal deformation of every particular $A(\lambda)$ for λ small enough. $A(\lambda)$ is block orthogonally-***mini***versal at A if $\{S_i\}$, $\{R_j\}$ are bases of $\mathcal{S}$ and $\mathcal{R}$.*

It is not hard to see that the theory we set up for matrices with all eigenvalues 0 can be generalized to a matrix A with different eigenvalues. The staircase form is a block upper triangular matrix, each of its diagonal blocks of the form $\lambda_i I + A_i$, with A_i in staircase form defined at the beginning of this chapter, and superdiagonal blocks arbitrary matrices. Its staircase invariant space is spanned by the block diagonal matrices, each diagonal block being in the staircase invariant space of the corresponding diagonal block A_i. $\mathcal{R}$ space is spanned by the block strictly upper triangular matrices s.t. every diagonal block is in the $\mathcal{R}$ space of the corresponding A_i. $\mathcal{T}_b$ is defined exactly the same as in the one eigenvalue case. All our theorems are still valid. When we give the definitions or apply the theorems, we do not really use the values of the eigenvalues, all that is important is how many different eigenvalues A has. In other words, we are working with bundle instead of orbit.

These forms are normal forms that have the same property as the Arnold's normal form: they are continuous under perturbation. The reason that we introduce block orthogonal notation is that the staircase algorithm is a realization to first order of the block orthogonally-versal deformation, as we will see in the next section.

3.4 Application to Matrix Staircase Forms

We are ready to understand the staircase algorithm described in Section 3.1.1. We concentrate on matrices with all eigenvalues 0, since otherwise, the staircase algorithm will separate other structures and continue recursively.

We use the notation $\mathrm{stair}(A)$ to denote the output A of the staircase algorithm as described in Section 3.1.1. Now suppose that we have a matrix A which is in staircase form. To zeroth order, any instance of the staircase algorithm replaces A with $\hat{A} = Q_0^T A Q_0$, where Q_0 is block diagonal orthogonal. Of course this does not change the staircase structure of A; the Q_0 represents the arbitrary rotations within the subspaces, and can depend on how the software is written, and the subtlety of roundoff errors when many singular values are 0. Next, suppose that we perturb A by ϵE. According to Corollary 3.1, we can decompose the perturbation matrix uniquely as $E = S + R + T_b$, with $S \in \mathcal{S}$, $R \in \mathcal{R}$ and $T_b \in \mathcal{T}_b$. Theorem 3.6 states that in addition to some block diagonal matrix Q_0, the staircase algorithm will apply a block orthogonal similarity transformation $Q_1 = I + \epsilon X + o(\epsilon)$ to $A + \epsilon E$ to kill the perturbation in T_b.

Theorem 3.6 *Suppose that A is a matrix in staircase form and E is any perturbation matrix. The staircase algorithm (without zeroing) on $A + \epsilon E$ will produce an orthogonal matrix Q (depending on ϵ) and the output matrix* $\mathrm{stair}(A + \epsilon E) = Q^T(A + \epsilon E)Q =$

$\hat{A} + \epsilon(\hat{S} + \hat{R}) + o(\epsilon)$, where $\hat{A}$ has the same staircase structure as A, $\hat{S}$ is a staircase invariant matrix of $\hat{A}$ and $\hat{R}$ is a block strictly upper triangular matrix. If singular values are zeroed out, then the algorithm further kills $\hat{S}$ and outputs $\hat{A} + \epsilon\hat{R} + o(\epsilon)$.

Proof:

After the first stage of the staircase algorithm, the first block column is orthogonal to the other columns, and this property is preserved through the completion of the algorithm. Generally, after the ith iteration, the ith block column below (including) the diagonal block is orthogonal to all other columns to its right, and this property is preserved all through. So when the algorithm terminates, we will have a matrix whose columns below (including) the diagonal block are orthogonal to all the columns to the right, in other words, it is a matrix in staircase form plus a staircase invariant matrix.

We can always write the similarity transformation matrix as $Q = Q_0(I + \epsilon X + \mathrm{o}(\epsilon))$, where Q_0 is a block diagonal orthogonal matrix and X is a block anti-symmetric matrix that does not depend on ϵ because of the local cross section property that we mentioned at the beginning of Section 3.3. Notice that Q_0 is not a constant matrix decided by A, it depends on ϵE to its first order, we should have written $(Q_0)_0 + \epsilon(Q_0)_1 + \mathrm{o}(\epsilon)$ instead of Q_0 . However, we do not expand Q_0 since as long as it is a block diagonal orthogonal transformation, it does not change the staircase structure of the matrix. Hence, we get

$$\begin{aligned}
\text{stair}(A + \epsilon E) =& \text{stair}(A + \epsilon S + \epsilon R + \epsilon T_b) \\
=&(I + \epsilon X^T + o(\epsilon))Q_0^T(A + \epsilon S + \epsilon R + \epsilon T_b)Q_0(I + \epsilon X + o(\epsilon)) \\
=&(I + \epsilon X^T + o(\epsilon))(\hat{A} + \epsilon\hat{S} + \epsilon\hat{R} + \epsilon\hat{T_b})(I + \epsilon X + o(\epsilon)) \qquad (24)\\
=&\hat{A} + \epsilon(\hat{S} + \hat{R} + \hat{T_b} + \hat{A}X - X\hat{A}) + o(\epsilon) \\
=&\hat{A} + \epsilon(\hat{S} + \hat{R}) + o(\epsilon).
\end{aligned}$$

Here, $\hat{A}, \hat{S}, \hat{R}$ and $\hat{T_b}$ are respectively $Q_0^TAQ_0, Q_0^TSQ_0, Q_0^TRQ_0$ and $Q_0^TT_bQ_0$. It is easy to check that $\hat{S}, \hat{R}, \hat{T_b}$ is still in the $\mathcal{S}, \mathcal{R}, \mathcal{T}_b$ space of $\hat{A}$. X is a block anti-symmetric matrix satisfying $\hat{T_b} = X\hat{A} - \hat{A}X$. We know that X is uniquely determined because the dimensions of $\hat{T}_b$ and the block anti-symmetric matrix space are the same. The reason that $\hat{T}_b = X\hat{A} - \hat{A}X$ hence the last equality in (24) holds is because the algorithm forces the output form as described in the first paragraph of this proof: $\hat{A} + \epsilon\hat{R}$ is in staircase form and $\epsilon\hat{S}$ is a staircase invariant matrix. Since $(\mathcal{S} \oplus \mathcal{R}) \cap \mathcal{T}_b$ is the zero matrix, the T_b term must vanish. □

To understand more clearly what this observation tells us, let us check some simple situations. If the matrix A is only perturbed in the direction $\mathcal{S}$ or $\mathcal{R}$, then the similarity transformation will be simply a block diagonal orthogonal matrix Q_0. If we ignore this transformation which does not change any structure, we can think of the output to be unchanged from the input, this is the reason we call $\mathcal{S}$ the staircase invariant space. The reason we did not include R into the staircase invariant space is that $A + \epsilon R$ is still within $O_b(A)$. If the matrix A is only perturbed along the block tangent direction $\mathcal{T}_b$, then the staircase algorithm will kill the perturbation and do a block diagonal orthogonal similarity transformation.

Although the staircase algorithm decides this Q_0 step by step all through the algorithm (due to SVD rank decisions), we can actually think of the Q_0 as decided at the first step. We can even ignore this Q_0 because the only reason it comes up is that the `svd` we use follows a specific way to sort singular values when they are different, and to choose the basis of the singular vector space when the same singular values appear.

We know that every matrix A can be reduced to a staircase form under an orthogonal transformation, in other words, we can always think of any general matrix M as P^TAP, where A is in staircase form. Thus in general, the staircase algorithm always introduces an orthogonal transformation and returns a matrix in staircase form and a first order perturbation in its staircase invariant direction, i.e. stair$(M + \epsilon E)$=stair$(P^TAP + \epsilon E)$=stair$(A + \epsilon PEP^T)$.

It is now obvious that if a staircase form matrix A has its $\mathcal{S}$ and $\mathcal{T}$ almost normal to each other, then the staircase algorithm will behave very well. On the other hand, if $\mathcal{S}$ is very close to $\mathcal{T}$ then it will fail. To emphasize this, we write it as a conclusion.

Conclusion 1 *The angle between the staircase invariant space $\mathcal{S}$ and the tangent space $\mathcal{T}$ decides the behavior of the staircase algorithm. The smaller the angle, the worse the algorithm behaves.*

In the one Jordan block case, we have an if-and-only-if condition for $\mathcal{S}$ to be near $\mathcal{T}$.

Theorem 3.7 *Let A be an $n \times n$ matrix in staircase form and suppose that all of its block sizes are 1×1, then $\mathcal{S}(A)$ is close to $\mathcal{T}(A)$ iff the following two conditions hold:*
(1)(row condition) there exists a non-zero row in A s.t. every entry on this row is o(1);
(2)(chain condition) there exists a chain of length $n-k$ with the chain value O(1), where k is the lowest row satisfying (1).
Here, we call $A_{i_1,i_2}, A_{i_2,i_3}, \ldots, A_{i_t,i_{t+1}}$ a chain of length t and the product $A_{i_1,i_2}A_{i_2,i_3}\cdots A_{i_t,i_{t+1}}$ is the chain value.

Proof Sketch:
Notice that $\mathcal{S}$ being close to $\mathcal{T}$ is equivalent to $\mathcal{S}$ being almost perpendicular to $\mathcal{N}$, the normal space of A. In this case, $\mathcal{N}$ is spanned by $\{I, A^T, A^{T2}, \ldots, A^{T(n-1)}\}$, $\mathcal{S}$ consists of matrices with nonzero entries only in the last row. Considering the angle between any two matrices from the two spaces, it is straightforward to show that $\mathcal{S}$ is almost perpendicular to $\mathcal{N}$ is equivalent to
(1) there exists a k s.t. the (n, k) entry of each of the matrices $I, A^T, \ldots, A^{T(n-1)}$ is $o(1)$ or 0;
(2) if the entry is $o(1)$, then it must have some other O(1) entry in the same matrix. Assume k is the largest choice if there are different k's. By a combinatorial argument, we can show that these two conditions are equivalent to the row and chain conditions respectively in our theorem. □

Remark 3.1 *Note that there exists an O(1) entry in a matrix is equivalent to say that there exists a singular value of the matrix of O(1). So, the chain condition is the same as saying that the singular values of A^{n-k} are not all $O(\epsilon)$ or smaller.*

Generally, we do not have an if-and-only-if condition for $\mathcal{S}$ to be close to $\mathcal{T}$, we only have a necessary condition, that is, only if at least one of the superdiagonal blocks of the original unperturbed matrix has a singular value almost 0, i.e. it has a weak stair, will $\mathcal{S}$ be close to $\mathcal{T}$. Actually, it is not hard to show that the angle between $\mathcal{T}_b$ and $\mathcal{R}$ is at most in the same order as the smallest singular value of the weak stair. So, when the perturbation matrix E is decomposed into $R + S + T_b$, R and T_b are typically very large, but whether S is large or not depends on whether $\mathcal{S}$ is close to $\mathcal{T}$ or not.

Notice that equation (24) is valid for sufficiently small ϵ. What range of ϵ is "sufficiently small"? Clearly, ϵ has to be smaller than the smallest singular value δ of the weak stairs. Moreover, the algorithm requires the perturbation along $\mathcal{T}$ and $\mathcal{S}$ to be both smaller than δ. Assume the angle between $\mathcal{T}$ and $\mathcal{S}$ is θ, then generally, when θ is large, we would expect an ϵ smaller than δ to be sufficiently small. However, when θ is close to 0, for a random perturbation, we would expect an ϵ in the order of δ/θ to be sufficiently small. Here, again, we can see that the angle between $\mathcal{S}$ and $\mathcal{T}$ decides the range of effective ϵ. For small θ, when ϵ is not sufficiently small, we observed some discontinuity in the 0th order term in Equation (24) caused by the ordering of singular values during certain stages of the algorithm. Thus, instead of the identity matrix, we get a permutation matrix in the 0th order term.

The theory explains why the staircase algorithm behaves so differently on the two matrices A_1 and A_2 in Section 2. Using Theorem 3.7, we can see that A_1 is a staircase failure ($k = 2$), while A_2 is not ($k = 1$). By a direct calculation, we find that the tangent space and the staircase invariant space of A_1 is very close ($\sin(<\mathcal{S}, \mathcal{T}>) = \delta/\sqrt{1+\delta^2}$), while this is not the situation for A_2 ($\sin(<\mathcal{S}, \mathcal{T}>) = 1/\sqrt{3}$). When transforming to get $\tilde{A}_1$ and $\tilde{A}_2$ with Q, which is an approximate orthogonal matrix up to the order of square root of machine precision ϵ_m, another error in the order of $\sqrt{\epsilon_m}$ (10^{-7}) is introduced, it is comparable with δ in our experiment, so the staircase algorithm actually runs on a shifted version $A_1 + \delta E_1$ and $A_2 + \delta E_1$. That is why we see R as large as an $O(10^{-6})$ added to J_3 in the second table for $\tilde{A}_2$. We might as well call A_2 a staircase failure in this situation, but A_1 suffers a much worse failure under the same situation, in that the staircase algorithm fails to detect a J_3 structure at all. This is because the tangent space and the staircase invariant space are so close that the S and T component are very large hence Equation (24) does not apply any more.

3.5 A Staircase Algorithm Failure to Motivate the Theory for Pencils

The pencil analog to the staircase failure in Section 3.2 is

$$(A_1, B_1) = \left(\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \delta & 0 & 0 & 0 \\ 0 & \delta & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \right),$$

where $\delta =$ `1.5e-8`. This is a pencil with the structure $L_1 \oplus J_2(0)$. After we add a random perturbation of size `1e-14` to this pencil, `GUPTRI` fails to return back the original pencil no matter which `EPSU` we choose. Instead, it returns back a more generic $L_2 \oplus J_1(0)$ pencil $O(\epsilon)$ away.

On the other hand, for another pencil with the same $L_1 \oplus J_2(0)$ structure:

$$(A_2, B_2) = \left(\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \delta & 0 \end{bmatrix} \right),$$

`GUPTRI` returns an $L_1 \oplus J_2(0)$ pencil $O(\epsilon)$ away.

At this point, readers may correctly expect that the reason behind this is again the angle between two certain spaces as in the matrix case.

3.6 Matrix Pencils

Parallel to the matrix case, we can set up a similar theory for the pencil case. For simplicity, we concentrate on the case when a pencil only has L-blocks and $J(0)$-blocks. Pencils containing L^T-blocks and non-zero (including ∞) eigenvalue blocks can always be reduced to the previous case by transposing and exchanging the two matrices of the pencil and/or shifting.

3.6.1 The Staircase Invariant Space and Related Subspaces for Pencils

A pencil (A, B) is in **staircase form** if we can divide both A and B into block rows of sizes $r_1, \ldots, r_k$ and block columns of sizes $s_1, \ldots, s_{k+1}$, s.t. A is strictly block upper triangular with every superdiagonal block having full column rank and B is block upper triangular with every diagonal block having full row rank and the rows orthogonal to each other. Here we allow s_{k+1} to be zero. A pencil is called **conforming to** (A, B) if it has the same block structure as (A, B). A square matrix is called **row (column) conforming to** (A, B) if it has diagonal block sizes the same as the row (column) sizes of (A, B).

Definition 3.5 *Suppose (A, B) is a pencil in staircase form and B_d is the block diagonal part of B. We call (S_A, S_B) a* **staircase invariant pencil of** (A, B) *if $S_A^T A = 0$, $S_B B_d^T = 0$ and (S_A, S_B) has complimentary structure to (A, B). We call the space consisting of all such (S_A, S_B) the* **staircase invariant space of** (A, B), *and denote it by $\mathcal{S}$.*

For example, let (A,B) have the staircase form

$(A,B) =$

$$\left(\begin{bmatrix} 0&0&0 & \times&\times & \times&\times & \times \\ 0&0&0 & \times&\times & \times&\times & \times \\ &&& 0&0 & \times&\times & \times \\ &&& 0&0 & \times&\times & \times \\ &&&&& 0&0 & \times \\ &&&&& 0&0 & \times \\ &&&&&&& 0 \end{bmatrix}, \begin{bmatrix} \times&\times&\times & \times&\times & \times&\times & \times \\ \times&\times&\times & \times&\times & \times&\times & \times \\ &&& \times&\times & \times&\times & \times \\ &&& \times&\times & \times&\times & \times \\ &&&&& \times&\times & \times \\ &&&&& \times&\times & \times \\ &&&&&&& \times \end{bmatrix} \right),$$

then $(S_A, S_B) =$

$$\left(\begin{bmatrix} \circ&\circ&\circ &&&&& \\ \circ&\circ&\circ &&&&& \\ \circ&\circ&\circ & \circ&\circ &&& \\ \circ&\circ&\circ & \circ&\circ &&& \\ \times&\times&\times & \times&\times & \times&\times & \\ \times&\times&\times & \times&\times & \times&\times & \\ \times&\times&\times & \times&\times & \times&\times & \times \end{bmatrix}, \begin{bmatrix} &&&&&&& \\ &&&&&&& \\ \times&\times&\times &&&&& \\ \times&\times&\times &&&&& \\ \times&\times&\times & \circ&\circ &&& \\ \times&\times&\times & \circ&\circ &&& \\ \times&\times&\times & \circ&\circ & \circ&\circ & \end{bmatrix} \right)$$

is a staircase invariant pencil of (A,B) if every column of S_A is in the left null space of A and every row of S_B is in the right null space of B. Notice that the sparsity structure of S_A and S_B is at most complimentary to that of A and B respectively, but S_A and S_B are often less sparse, because of the requirement on the nullspace. To be precise, if we find more than one diagonal block with the same size, then among the blocks of this size, only the blocks on the lowest block row appear in the sparsity structure of S_A. If any of the diagonal blocks of B is a square block, then S_B has all zero entries throughout the corresponding block column.

As special cases, if A is a strictly upper triangular square matrix and B is an upper triangular square matrix with diagonal entries nonzero, then S_A only has nonzero entries in the bottom row and S_B is simply a zero matrix. If A is a strictly upper triangular $n \times (n+1)$ matrix and B is an upper triangular $n \times (n+1)$ matrix with diagonal entries nonzero, then (S_A, S_B) is the zero pencil.

Definition 3.6 *Suppose (A, B) is a pencil. We call $\mathcal{O}(A, B) \equiv \{X(A, B)Y : X, Y$ are non-singular square matrices $\}$ the* **orbit** *of a pencil (A, B). We call $\mathcal{T} \equiv \{X(A, B) - (A, B)Y : X, Y$ are any square matrices $\}$ the* **tangent space** *of $\mathcal{O}(A, B)$ at (A, B).*

Theorem 3.8 *Let (A, B) be an $m \times n$ pencil in staircase form, then the staircase invariant space $\mathcal{S}$ of (A, B) and the tangent space $\mathcal{T}$ form an oblique decomposition of $m \times n$ pencil space, i.e. $\mathbb{R}^{2mn} = \mathcal{S} + \mathcal{T}$.*

Proof:

The proof of the theorem is similar to that of Theorem 3.1; first we prove the dimension of $\mathcal{S}(A, B)$ is the same as the codimension of $\mathcal{T}(A, B)$, then we prove $\mathcal{S} \cap \mathcal{T} = \{0\}$ by induction. The readers may try to fill out the details. □

Definition 3.7 *Suppose (A, B) is a pencil. We call $\mathcal{O}_b(A, B) \equiv \{P(A, B)Q : P = e^X, X$ is a block anti-symmetric matrix row conforming to $(A, B), Q = e^Y, Y$ is a block anti-symmetric matrix column conforming to (A, B) the* **block orthogonal-orbit** *of a pencil (A, B). We call $\mathcal{T}_b \equiv \{X(A, B) - (A, B)Y$: X is a block anti-symmetric matrix row conforming to (A, B), Y is a block anti-symmetric matrix column conforming to $(A, B)\}$ the* **block tangent space** *of the block orthogonal-orbit $\mathcal{O}_b(A, B)$ at (A, B). We call $\mathcal{R} \equiv \{U(A, B) - (A, B)V$: U is a block upper triangular matrix row conforming to (A, B), V is a block upper triangular matrix column conforming to $(A, B)\}$ the* **block upper pencil space** *of (A, B).*

Theorem 3.9 *Let (A, B) be an $m \times n$ pencil in staircase form, then the tangent space $\mathcal{T}$ of the orbit $\mathcal{O}(A, B)$ can be split into the block tangent space $\mathcal{T}_b$ of the orbit $\mathcal{O}_b(A, B)$ and the block upper pencil space $\mathcal{R}$, i.e. $\mathcal{T} = \mathcal{T}_b \oplus \mathcal{R}$.*

Proof:

This can be proved by a very similar argument concerning the dimensions as for matrix, in which the dimension of $\mathcal{R}$ is $2\sum_{i<j} r_i s_j + \sum r_i s_i$, the dimension of $\mathcal{T}_b$ is $\sum_{i<j} r_i r_j +$

$\sum_{i<j} s_i s_j$, the codimension of the orbit $\mathcal{O}(A,B)$ (or $\mathcal{T}$) is $\sum s_i r_i - \sum_{j>i} s_i s_j + 2\sum_{j>i} s_i r_j - \sum_{j>i} r_i r_j$ [18]. □

Corollary 3.2 $\mathbb{R}^{2mn} = \mathcal{T}_b \oplus \mathcal{R} \oplus \mathcal{S}$.

3.6.2 Staircase as a Versal Deformation for pencils

The theory of versal forms for pencils [34] is similar to the one for matrices. A **deformation** of a pencil (A,B) is a pencil $(A,B)(\lambda)$ with entries power series in the real variables λ_i. We say that two deformations $(A,B)(\lambda)$ and $(C,D)(\lambda)$ are **equivalent** if there exist two deformations $P(\lambda)$ and $Q(\lambda)$ of identity matrices such that $(A,B)(\lambda) = P(\lambda)(C,D)(\lambda)Q(\lambda)$.

Theorem 3.10 *Suppose (A,B) is in staircase form. Fix $S_i \in \mathcal{S}$, $i = 1,\dots,k$ s.t. span$\{S_i\} = \mathcal{S}$ and $k \geq$ dim$(\mathcal{S})$. It follows that*

$$(A,B)(\lambda) \equiv (A,B) + \sum_i \lambda_i S_i \tag{25}$$

is a versal deformation of every particular $(A,B)(\lambda)$ for λ small enough. $(A,B)(\lambda)$ is **mini***versal at $\lambda = 0$ if $\{S_i\}$ is a basis of $\mathcal{S}$.*

Definition 3.8 *We say two deformations $(A,B)(\lambda)$ and $(C,D)(\lambda)$ are* **block orthogonally-equivalent** *if there exist two block orthogonal deformations $P(\lambda)$ and $Q(\lambda)$ of the identity matrix such that $(A,B)(\lambda) = P(\lambda)(C,D)(\lambda)Q(\lambda)$. Here, $P(\lambda)$ and $Q(\lambda)$ are exponentials of matrices which are conforming to (A,B) in row and column respectively.*

We say that a deformation $(A,B)(\lambda)$ is **block orthogonally-versal** *if any other deformation $(C,D)(\mu)$ is block orthogonally-equivalent to the deformation $(A,B)(\phi(\mu))$. Here, ϕ is a mapping holomorphic at 0 with $\phi(0) = 0$.*

Theorem 3.11 *A deformation $(A,B)(\lambda)$ of (A,B) is block orthogonally-versal iff the mapping $(A,B)(\lambda)$ is transversal to the block orthogonal-orbit of (A,B) at $\lambda = 0$.*

This is the corresponding result to Theorem 3.4.

Since we know that $\mathcal{T}$ can be decomposed into $\mathcal{T}_b \oplus \mathcal{R}$, we get:

Theorem 3.12 *Suppose a pencil (A, B) is in staircase form. Fix $S_i \in \mathcal{S}, i = 1, \ldots, k$ s.t. $span\{S_i\} = \mathcal{S}$ and $k \geq dim(\mathcal{S})$. Fix $R_j \in \mathcal{R}, j = 1, \ldots, l$ s.t. $span\{R_j\} = \mathcal{R}$ and $l \geq dim(\mathcal{R})$. It follows that*

$$(A, B)(\lambda) \equiv (A, B) + \sum_i \lambda_i S_i + \sum_j \lambda_j R_j$$

*is a block orthogonally-versal deformation of every particular $(A, B)(\lambda)$ for λ small enough. $(A, B)(\lambda)$ is block orthogonally-***mini***versal at (A, B) if $\{S_i\}$, $\{R_j\}$ are bases of $\mathcal{S}$ and $\mathcal{R}$.*

Notice that as in the matrix case, we can also extend our definitions and theorems to the general form containing L^T-blocks and non-zero eigenvalue blocks, and again, we will not specify what eigenvalues they are and hence get into the bundle case. We only want to point out one particular example here. If (A, B) is in the staircase form of $L_n + J_1(\cdot)$, then, A will be a strictly upper triangular matrix with nonzero entries on the super diagonal and B will be a triangular matrix with nonzero entries on the diagonal except the $(n + 1, n + 1)$ entry. S_A will be the zero matrix and S_B will be a matrix with the only nonzero entry on its $(n + 1, n + 1)$ entry.

3.7 Application to Pencil Staircase Forms

We concentrate on $L \oplus J(0)$ structures only, since otherwise, the staircase algorithm will separate all other structures and continue similarly after a shift and/or transpose on that part only. As in the matrix case, the staircase algorithm basically decomposes the perturbation pencil into three spaces $\mathcal{T}_b$, $\mathcal{R}$, and $\mathcal{S}$ and kills the perturbation in $\mathcal{T}_b$.

Theorem 3.13 *Suppose that (A, B) is a pencil in staircase form and E is any perturbation pencil. The staircase algorithm (without zeroing) on $(A, B) + \epsilon E$ will produce two* *orthogonal matrices P and Q (depending on ϵ) and the output pencil* $\mathrm{stair}((A,B)+\epsilon E) = P^T((A,B)+\epsilon E)Q = (\hat{A},\hat{B}) + \epsilon(\hat{S}+\hat{R}) + o(\epsilon)$, *where $(\hat{A},\hat{B})$ has the sane staircase structure as (A,B), $\hat{S}$ is a staircase invariant pencil of $(\hat{A},\hat{B})$ and $\hat{R}$ is in the block upper pencil space $\mathcal{R}$. If singular values are zeroed out, then the algorithm further kills $\hat{S}$ and output $(\hat{A},\hat{B}) + \epsilon\hat{R} + o(\epsilon)$.*

We use a formula to explain the statement more clearly:

$$\begin{aligned}
&(I+\epsilon X+o(\epsilon))P_1((A,B)+\epsilon S+\epsilon R+\epsilon T_b)Q_1(I-\epsilon Y+o(\epsilon))\\
=&(I+\epsilon X+o(\epsilon))((\hat{A},\hat{B})+\epsilon\hat{S}+\epsilon\hat{R}+\epsilon\hat{T_b})(I-\epsilon Y+o(\epsilon))\\
=&(\hat{A},\hat{B})+\epsilon(\hat{S}+\hat{R}+\hat{T_b}+X(\hat{A},\hat{B})-(\hat{A},\hat{B})Y)+o(\epsilon)\\
=&(\hat{A},\hat{B})+\epsilon(\hat{S}+\hat{R})+o(\epsilon).
\end{aligned} \tag{26}$$

Similarly, we can see that when a pencil has its $\mathcal{T}$ and $\mathcal{S}$ almost normal to each other, the staircase algorithm will behave well. On the other hand, if $\mathcal{S}$ is very close to $\mathcal{T}$, then it will behave badly. This is exactly the situation in the two pencil examples in Section 5. Although the two pencils are both ill conditioned, a direct calculation shows that the first pencil has its staircase invariant space very close to the tangent space (the angle $<\mathcal{S},\mathcal{T}>=\delta/\sqrt{\delta^2+2}$) while the second one does not (the angle $<\mathcal{S},\mathcal{T}>=1/\sqrt{2+\delta^2}$).

The if-and-only-if condition for $\mathcal{S}$ to be close to $\mathcal{T}$ is more difficult than in the matrix case. One necessary condition is that one super diagonal block of A is almost of not full column rank or one diagonal block of B is almost not full row rank. This is usually referred to as **weak coupling**.

3.8 Examples: The geometry of the Boley pencil and others

Boley [5, Example 2, Page 639] presents an example of a 7×8 pencil (A,B) that is controllable (has generic Kronecker structure) yet it is known that an uncontrollable system (non-generic Kronecker structure) is nearby at a distance 6e-4. What makes the example interesting is that the staircase algorithm fails to find this nearby uncontrollable system while other methods succeed. Our theory provides a geometrical understanding of why this famous example leads to staircase failure: the staircase invariant space is very close to the tangent space.

The pencil that we refer to is $(A, B(\epsilon))$, where

$$A = \begin{pmatrix} \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix} \text{ and } B(\epsilon) = \begin{pmatrix} 1 & -1 & -1 & -1 & -1 & -1 & -1 & 7 \\ \cdot & 1 & -1 & -1 & -1 & -1 & -1 & 6 \\ \cdot & \cdot & 1 & -1 & -1 & -1 & -1 & 5 \\ \cdot & \cdot & \cdot & 1 & -1 & -1 & -1 & 4 \\ \cdot & \cdot & \cdot & \cdot & 1 & -1 & -1 & 3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & -1 & 2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \epsilon & 1 \end{pmatrix}.$$

(The dots refer to zeros, and in the original Boley example $\epsilon = 1$.)

When $\epsilon = 1$, the staircase algorithm predicts a distance of 1, and is therefore off by nearly four orders of magnitude. To understand the failure, our theory works best for smaller values of ϵ, but it is still clear that even for $\epsilon = 1$, there will continue to be difficulties.

It is useful to express the pencil $(A, B(\epsilon))$ as $P_0 + \epsilon E$, where $P_0 = (A, B(0))$ and S is zero except for a "one" in the (7,7) entry of its B part. P_0 is in the bundle of pencils whose Kronecker form is $L_6 + J_1(\cdot)$ and the perturbation E is exactly in the unique staircase invariant direction (hence the notation "S") as we pointed out at the end of Section 3.6.

The relevant quantity is then the angle between the staircase invariant space and the pencil space. An easy calculation reveals that the angle is very small: $\theta_S = 0.0028$ radians. In order to get a feeling for what range of ϵ first order theory applies, we calculated the exact distance $d(\epsilon) \equiv d(P(\epsilon), \text{bundle})$ using the nonlinear eigenvalue template software [73]. To first order, $d(\epsilon) = \theta_S \cdot \epsilon$. Figure 15 plots the distances first for $\epsilon \in [0, 2]$ and then a closeup for $\epsilon = [0, 0.02]$.

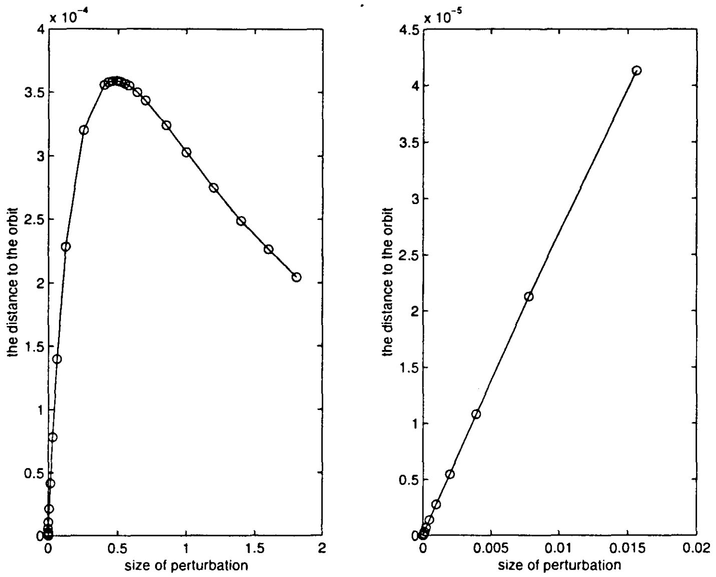


Figure 15: The picture to explain the change of the distance of the pencils $P_0 + \epsilon E$ to the bundle of $L_6 + J(\cdot)$as ϵ changes. The second subplot is part of the first one at the points near $\epsilon = 0$.

Our observation based on this data suggests that first order theory is good to two decimal places for $\epsilon \leq 10^{-4}$ and one place for $\epsilon \leq 10^{-2}$. To understand the geometry of staircase algorithmic failure, one decimal place or even merely an order of magnitude is quite sufficient.

In summary, we see clearly that the staircase invariant direction is at a small angle to the tangent space, and therefore the staircase algorithm will have difficulty finding the nearest pencil on the bundle or predicting the distance. This difficulty is quantified by the angle θ_S.

Since the Boley example is for $\epsilon = 1$, we computed the distance well past $\epsilon = 1$. The breakdown of first order theory is attributed to the curving of the bundle towards S. A three dimensional schematic is portrayed in Figure 16.

The relevant picture for control theory is a planar intersection of the above picture. In control theory, we set the special requirement that the "A" matrix has the form $[0 \;\; I]$. Pencils on the intersection of this hyperplane and the bundle are termed "uncontrollable."

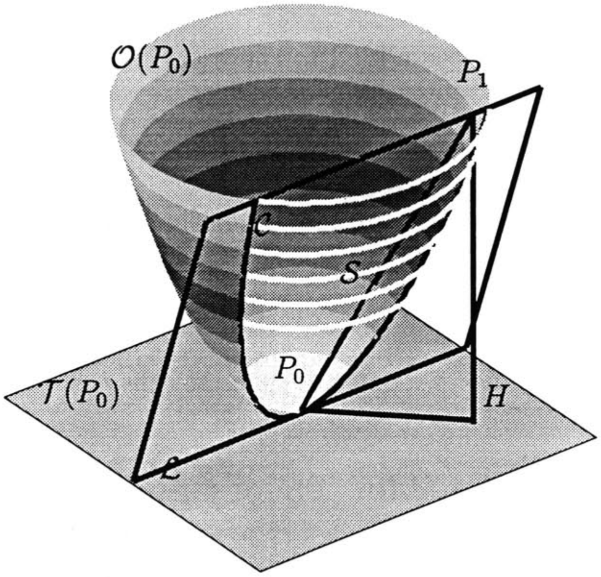


Figure 16: The staircase algorithm on the Boley example. The surface represents the orbit $\mathcal{O}(P_0)$. Its tangent space at the pencil P_0, $\mathcal{T}(P_0)$, is represented by the plane on the bottom. P_1 lies on the staircase invariant space $\mathcal{S}$ inside the "bowl". The hyperplane of uncontrollable pencils is represented by the plane cutting through the surface along the curve $\mathcal{C}$. It intersects $\mathcal{T}(P_0)$ along $\mathcal{L}$. The angle between $\mathcal{L}$ and $\mathcal{S}$ is θ_c. The angle between $\mathcal{S}$ and $\mathcal{T}(P_0)$, θ_S, is represented by the angle $\angle HP_0P_1$.

We analytically calculated the angle θ_c between S and the tangent space for the "uncontrollable surfaces." We found that $\theta_c = 0.0040$. Using the nonlinear eigenvalue template software [73], we numerically computed the true distance from $P_0 + \epsilon E$ to the "uncontrollable surfaces" and calculated the ratio of this distance to ϵ, we found that for $\epsilon < 8e - 4$, the ratio agrees with $\theta_c = 0.0040$ very well.

We did a similar analysis on the three pencils C_1, C_2 C_3 given by J. Demmel and B. Kågström [23]. We found that the *sin* values of the angles between $\mathcal{S}$ and $\mathcal{T}$ are respectively 2.4325e-02, 3.4198e-02 and 8.8139e-03, and the *sin* values between $\mathcal{T}_b$ and $\mathcal{R}$ are respectively 1.7957e-02 7.3751e-03 and 3.3320e-06. This explains why we saw the staircase algorithm behave progressively worse on them. Especially, it explains why when a perturbation about 10^{-3} is added to these pencils, C_3 behaves dramatically worse then C_1 and C_2. The component in $\mathcal{S}$ is almost of the same order as the entries of the original pencil.

So we conclude that the reason the staircase algorithm does not work well on this example is because $P_0 = (A, B(0))$ is actually a staircase failure, in that its tangent space is very close to its staircase invariant space and also the perturbation is so large that even if we know the angle in advance we can not estimate the distance well.

3.9 Valid Region of First Order Theory and Discussion on Discontinuity

3.9.1 An Interesting Phenomenon

Concerning the staircase algorithm, we mentioned the discontinuity in the 0th order term $Q_0 I$ of $Q_0(I + \epsilon X + o(\epsilon))$ in the equation:

$$\begin{aligned} \text{stair}(A + \epsilon E) =& \text{stair}(A + \epsilon S + \epsilon R + \epsilon T_b) \\ =& (I + \epsilon X^T + o(\epsilon))Q_0^T(A + \epsilon S + \epsilon R + \epsilon T_b)Q_0(I + \epsilon X + o(\epsilon)) \end{aligned} \tag{27}$$

We noticed that instead of $Q_0 I$, the 0th order becomes the product of Q_0 and a permutation matrix. This is caused by the ordering of singular values during certain stages of the algorithm. To explain this phenomenon more clearly, we take a simplest example:

Let

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & \delta \\ 0 & 0 & 0 \end{pmatrix} \tag{28}$$

where $\delta = 1e - 6$. Let E be a random dense matrix with Frobenious norm 1. We show the distance between Q and the identity matrix up to an equivalent class (See Section 3.3) in the first plot of Figure 17.

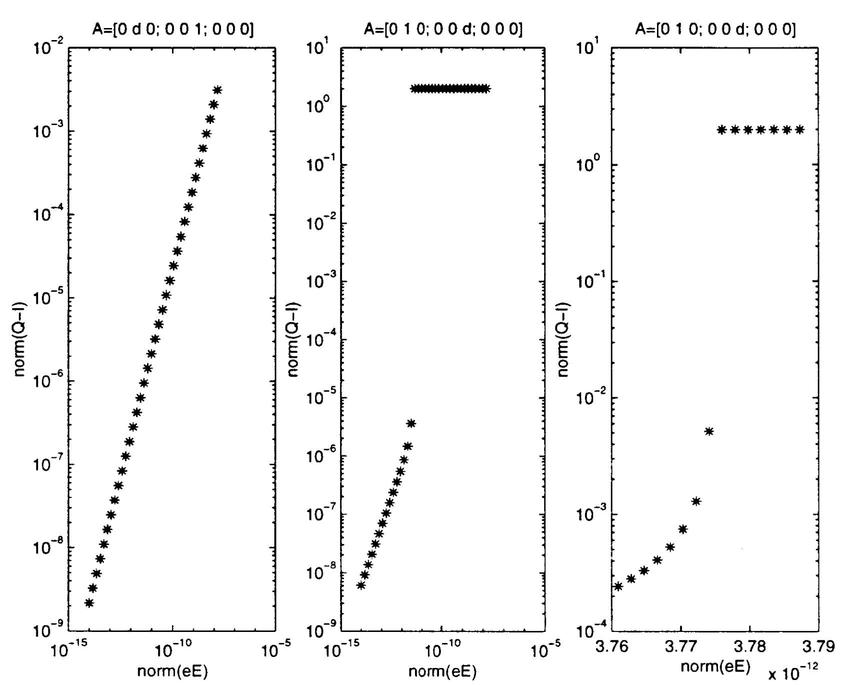


Figure 17: The plot of the distance between the orthogonal matrix $Q_0(I + \epsilon X + o(\epsilon))$ in Equation 27 and Q_0. The perturbation matrix E is a same fixed random dense matrix for both subplots. We vary the norm of the perturbation by multiplying to it an ϵ which changes from $\delta^2/100$ to $\delta/100$. The third subplot is a closer look near the jumping point in the second subplot with adjusted difference between every two neighboring ϵ's. The unperturbed matrix A is as in Equation 28 in the second and third subplot, by exchanging the position of 1 and δ we get the unperturbed matrix in the first subplot.

When we apply staircase algorithm on $A + \epsilon E$, we first do an SVD on $A + \epsilon E$ and get $A + \epsilon E = U \Sigma V^T$, where the singular values are in decreasing order along the diagonal of Σ. Then we let B be the right lower 2×2 block of $V^T(A + \epsilon E)V = V^T U \Sigma$ and do another SVD on B, i.e. $B = U_1 \Sigma_1 V_1^T$ and continue with the operation $V_1^T B V_1 = V_1^T U_1 \Sigma_1$. Typically, for sufficiently small ϵ, we would see both V and V_1 close to identity. However, as we increase ϵ from 0 to a certain value, the singular values of B become equal and if we continue to increase ϵ, the two singular values will switch their positions and hence the V_1, which contains the two right singular vectors will change positions too. This is exactly how the identity matrix in the 0th order changes to the permutation matrix in our example.

The switching of singular vectors at a certain stage of SVD during the staircase algorithm explains the similar discontinuity phenomenon for other matrices.

However, remember that as we increase ϵ, there is something else going on. That is, ϵ may go out of the valid region of our first order theory. When X is large enough, sometimes the discontinuous phenomenon is not visible because X itself is large enough to overwhelm the 0th order totally.

Numerical experiments seem to agree that most time, when we have small angle between $\mathcal{S}$ and $\mathcal{T}$ spaces of A, we will be able to observe the discontinuity, i.e. the discontinuity point is reached within the valid region of the first order theory. Very often, when $\mathcal{S}$ and $\mathcal{T}$ are not close, we can not observe this discontinuity (second subplot of Figure 17). We are hoping that we can find an if-and-only-if condition for the discontinuity to happen within the valid region of the first order theory in terms of the S and T spaces. It is not achieved.

3.9.2 Discussion

For possible future work, we write a summary of what we conclude and what the major difficulty is:

Consider the $\mathcal{R}$, $\mathcal{T}_b$ and $\mathcal{S}$ spaces of a given matrix A, we further let $\mathcal{T} = \mathcal{T}_b + \mathcal{R}$. When $\mathcal{R}$ and $\mathcal{T}_b$ are close, it means large $\mathcal{R}$ components and $\mathcal{T}_b$ components when we decompose ϵE, although the component in $\mathcal{R}+\mathcal{T}_b$ may be small. So, when ϵ grows larger, the resulting $A + R$ changes quickly.

When not only $\mathcal{R}$ and $\mathcal{T}_b$ are close, but also $\mathcal{T}$ and $\mathcal{S}$ are close, sometimes, a discontinuity can be observed. That is, when ϵ small, Equation (27) is still valid. But when ϵ reaches a particular value, the 0th order in Equation (27) no longer consists of the identity matrix. It becomes a certain permutation matrix.

We could have hoped that this kind of discontinuity applies to all matrices with small angle between $\mathcal{T}$ and $\mathcal{S}$. However, remember that if $\mathcal{T}$ and $\mathcal{S}$ are close to each other, it must imply that $\mathcal{T}_b$ and $\mathcal{R}$ are close too. So, sometimes this discontinuity is not observable because before the discontinuity is reached, $\mathcal{R}$ component is already big enough to make first order theory meaningless at all.

The key problem is then the comparison between the speed on reaching the discontinuity point and reaching the boundary of the valid region of first order theory. It is actually the question of comparing the two angles $\angle_{\mathcal{T}_b,\mathcal{R}}$ and $\angle_{\mathcal{T},\mathcal{S}}$, which remains to be answered.

One thing we ought to clarify is on "discontinuity". We did a careful investigation on several examples and believe that instead of discontinuity mathematically, it is rather an extremely "sharp change". Figure 17 demonstrates why it is more likely not discontinuity.

However we need to emphasize that whenever we saw a "sharp change", there is actually a swap of singular values, and that is why the "discontinuity" or "sharp change" phenomenon is interesting.

4 A Highly Robust Dispersion Estimator

4.1 Introduction

Many dispersion matrix estimators exist in the literature. We propose a new componentwise estimator, based on a highly robust estimator of scale Q_n and the simple fact that $4\text{cov}(X, Y) = \text{var}(X + Y) - \text{var}(X - Y)$. We study its robustness properties by means of the influence function and the breakdown point. Further characteristics like asymptotic variance and efficiency were also analyzed. A major advantage of the novel estimator is that its behavior is close to the maximum likelihood estimator in noncontaminated situations, whereas it is highly robust in contaminated situations. We show that in the componentwise approach, for multivariate Gaussian distributions, covariance matrix estimation is more difficult than correlation matrix estimation, because the asymptotic variance of the covariance estimator increases with increasing dependence, whereas it decreases with increasing dependence for correlation estimators. We also proved that the asymptotic variance of covariance estimators for multivariate Gaussian distributions is proportional to the asymptotic variance of the underlying scale estimator. The proportionality value depends only on the underlying dependence. Therefore, our highly robust dispersion estimator is the best robust choice at the present time in the componentwise approach, because it combines small variability and robustness properties like high breakdown point and bounded influence function. A simulation study was carried out in order to assess the behavior of the new estimator. First, a comparison with another robust componentwise estimator based on the median absolute deviation (MAD) scale estimator was performed. The highly robust properties of the new estimator were confirmed. Moreover, it is shown that the behavior of the new estimator is better than the one based on the MAD, although the latter is the most B-robust componentwise dispersion estimator. A second comparison with global estimators like the maximum likelihood estimator or the minimum volume ellipsoid estimator has also been performed, with two types of outliers. In this case, the highly robust dispersion matrix estimator turns out to be a compromise between the high efficiency of the maximum likelihood estimator in noncontaminated situations and the highly robust properties of the minimum volume ellipsoid estimator in contaminated situations, with exploding type of outliers.

Furthermore, we apply the method to estimate autocovariance. Consider a time series $\{X_t : t \in \mathbb{Z}\}$ and assume that it satisfies the hypothesis of second-order stationarity:

(i) $\mathrm{E}(X_t^2) < \infty, \quad \forall t \in \mathbb{Z},$

(ii) $\mathrm{E}(X_t) = \mu = \text{constant}, \quad \forall t \in \mathbb{Z},$

(iii) $\mathrm{Cov}(X_{t+h}, X_t) = \gamma(h), \quad \forall t, h \in \mathbb{Z},$

where $\gamma(h)$ is the autocovariance function of X_t at lag h. The classical estimator for the autocovariance function, based on the method of moments, on a sample $\mathbf{x} = (X_1, \ldots, X_n)^T$, is

$$\hat{\gamma}_M(h, \mathbf{x}) = \frac{1}{n-h} \sum_{i=1}^{n-h} (X_{i+h} - \bar{X})(X_i - \bar{X}), \quad 0 \leq h \leq n-1, \tag{29}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Note that $\frac{n-h}{n}\hat{\gamma}_M(h, \mathbf{x})$ is often used in order to ensure positive definiteness of the estimated covariance matrix.

Applying the dispersion estimator in a time series, we can get a new robust autocovariance estimator. Section 4.5.1 introduces a concept of temporal breakdown point of an autocovariance estimator and discusses its link with the classical breakdown point. The influence function for autocovariance estimators is computed in Section 4.5.2, and the formulas for their asymptotic variance is given in Section 4.5.3. These results are completed with a simulation study in Section 4.5.4, on AR(1) and MA(1) models. The behavior of the classical and highly robust autocovariance estimator in presence of outliers is also studied. In Section 4.5.5, a time series of monthly interest rates of an Austrian bank is analyzed.

4.2 The Highly Robust Dispersion Estimator

4.2.1 Dispersion between two random variables

Traditionally, covariance estimation between two random variables X and Y is based on a location approach, since $\mathrm{Cov}(X,Y) = \mathrm{E}\big[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\big]$, yielding for example the maximum likelihood estimator $\hat{\gamma}_{MLE} = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \hat{\mu_X})(y_i - \hat{\mu_Y})$. However, covariance estimation can also be based on a scale approach, by means of the following identity ([45, 58]):

$$\mathrm{Cov}(X,Y) = \frac{\alpha\beta}{4}\Big[\mathrm{Var}(X/\alpha + Y/\beta) - \mathrm{Var}(X/\alpha - Y/\beta)\Big], \quad \forall\alpha,\beta \in \mathbb{R}. \tag{30}$$

In general, X and Y may be measured in different units, and the choice $\alpha = \sigma_X \equiv \sqrt{\mathrm{Var}(X)}$, $\beta = \sigma_Y \equiv \sqrt{\mathrm{Var}(Y)}$ is recommended by Rousseeuw and Croux [46]. The choice of a robust estimator of the variance in Equation (30) produces a robust estimator of the covariance between X and Y.

In the context of scale estimation, Rousseeuw and Croux [86, 87] proposed a simple, explicit and highly robust scale estimator Q_n:

$$Q_n(\mathbf{z}) = d\Big\{|z_i - z_j|; i < j,\, i,j = 1,2,\ldots,n\Big\}_{(k)}, \tag{31}$$

where $\mathbf{z} = (z_1, \ldots, z_n)^T$ is a sample of a random variable Z, $k = \lfloor(\binom{n}{2}+2)/4\rfloor+1$ and $\lfloor\cdot\rfloor$ denotes the integer part. The factor d is for consistency: for the Gaussian distribution, $d = 2.2191$. This means that we sort the set of all absolute differences $|z_i - z_j|$ in increasing order for $i < j$, $i,j = 1,2,\ldots,n$ and then compute its k-th order statistic (approximately the 1/4 quantile for large n). This value is multiplied by d, thus yielding Q_n. Note that this estimator computes the k-th order statistic of the $\binom{n}{2}$ interpoint distances. It is of interest to remark that Q_n does not rely on any location knowledge and is therefore said to be location-free. This is in contrast to the classical sample covariance matrix estimation, which can be obtained by inserting the classical sample variance estimator in Equation (30). At first sight, the estimator Q_n appears to need $O(n^2)$ computation time, which would be a disadvantage. However, it can be computed using no more than $O(n \log n)$ time and $O(n)$ storage, by means of the fast algorithm described by Croux and Rousseeuw [15].

Using the identity (30) and the definition (31) of the scale estimator Q_n, we propose the following highly robust estimator to compute the covariance between two random variables X and Y. First, use Q_n to estimate the standard deviations σ_X and σ_Y of X and Y. Then, use Q_n again to estimate the standard deviations σ_+ and σ_- of $X/\sigma_X + Y/\sigma_Y$ and $X/\sigma_X - Y/\sigma_Y$. The covariance θ between X and Y is $\sigma_X \sigma_Y (\sigma_+^2 - \sigma_-^2)/4$. Therefore, the highly robust estimator $\hat{\gamma}_Q$ of the covariance is:

$$\hat{\gamma}_Q(\mathbf{x}, \mathbf{y}) = \frac{\alpha\beta}{4}\left[Q_n^2(\mathbf{x}/\alpha + \mathbf{y}/\beta) - Q_n^2(\mathbf{x}/\alpha - \mathbf{y}/\beta)\right], \tag{32}$$

where $\alpha = Q_n(\mathbf{x})$, $\beta = Q_n(\mathbf{y})$. Note that the highly robust covariance estimator $\hat{\gamma}_Q$ can also be carried out with $O(n \log n)$ time and $O(n)$ storage.

In order to obtain the estimator $\hat{\rho}_Q$ for the correlation between two random variables X and Y, we divide the estimator $\hat{\gamma}_Q(\mathbf{x}, \mathbf{y})$ in Equation (32) by $Q_n(\mathbf{x})$ and $Q_n(\mathbf{y})$:

$$\hat{\rho}_Q(\mathbf{x}, \mathbf{y}) = \frac{1}{4}\left[Q_n^2(\mathbf{x}/\alpha + \mathbf{y}/\beta) - Q_n^2(\mathbf{x}/\alpha - \mathbf{y}/\beta)\right], \tag{33}$$

where $\alpha = Q_n(\mathbf{x})$, $\beta = Q_n(\mathbf{y})$.

4.2.2 Dispersion between p random variables

In the case of n observations of a p-dimensional random vector X, we use the estimator $\hat{\gamma}_Q$ to estimate every covariance between X_i and X_j ($i, j = 1, \ldots, p$, $i \neq j$) to get the (i, j) entry of the covariance matrix Σ. The diagonal entries are estimated using Q_n^2 directly on the X_i's ($i = 1, \ldots, p$). This provides a highly robust componentwise estimator $\hat{\Sigma}_Q$ of the covariance matrix Σ.

Using $\hat{\rho}_Q$, we can estimate the entries of the correlation matrix R similarly as in the covariance matrix case, thus yielding a highly robust componentwise estimator $\hat{R}_Q$. We set all the diagonal entries of $\hat{R}_Q$ to 1's.

Note that since the method we propose is componentwise instead of global, there is no guarantee that we get a positive definite matrix at the end of the estimation. Rousseeuw and Molenberghs [89] proposed three kinds of methods to transform the estimated matrix to a positive definite matrix. They are respectively the shrinking method, the eigenvalue method, and the scaling method. When the dispersion matrix itself is the quantity of interest, one should transform it to a positive definite matrix using one of these methods, while if some particular entries in the matrix are the values of interest, then the estimated values should provide a good estimation of the real values.

4.3 Properties of the Estimator

4.3.1 Breakdown point

In the context of robust statistics, the breakdown point of an estimator is an important feature of reliability. It indicates how many data points need to be replaced by arbitrary values to destroy the estimator. The classical notion of breakdown point of a scale estimator is given in the following definition.

Definition 4.1 *Let $\mathbf{z} = (z_1, \ldots, z_n)^T$ be a sample of size n and $\tilde{\mathbf{z}}$ is obtained by replacing any m observations of $\mathbf{z}$ by arbitrary values. The sample breakdown point of a scale estimator $S_n(\mathbf{z})$ is:*

$$\varepsilon_n^*(S_n(\mathbf{z})) = \max\left\{\frac{m}{n} \ : \ \sup_{\tilde{\mathbf{z}}} S_n(\tilde{\mathbf{z}}) < \infty \textit{ and } \inf_{\tilde{\mathbf{z}}} S_n(\tilde{\mathbf{z}}) > 0\right\}.$$

Roughly speaking, the classical breakdown point gives the maximum fraction of outliers that the scale estimator can cope with. It indicates how many data points can be replaced by arbitrary values before the scale estimator explodes (tends to infinity) or implodes (tends to zero). Further discussions of this concept can be found in [49, 51, 52, 58, 59, 27, 42]. The sample breakdown point ε_n^* of most scale estimators is known, or can be computed. However, by using a scale estimator to compute the covariance, it is on the level of sums $\mathbf{x}+\mathbf{y}$ and differences $\mathbf{x}-\mathbf{y}$ that the estimator is applied. Similarly, we can define the sample breakdown point of a scale based covariance estimator, using Equation (30).

Definition 4.2 *Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$ be two samples of size n. Let $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ and $\tilde{\mathbf{z}}$ is obtained by replacing any m pairs of $\mathbf{z}$ by arbitrary values. The sample breakdown point of a covariance estimator $\hat{\gamma}_{S_n}(h, \mathbf{z}) = \frac{1}{4\alpha\beta}[S_n^2(\alpha\mathbf{x}+\beta\mathbf{y}) - S_n^2(\alpha\mathbf{x}-\beta\mathbf{y})]$ based on a scale estimator S_n is:*

$$\varepsilon_n^*(\hat{\gamma}_{S_n}(h, \mathbf{z})) = \max\left\{\frac{m}{n} : \sup_{\tilde{\mathbf{z}}} \hat{\gamma}_{S_n}(h, \tilde{\mathbf{z}}) < \infty \textit{ and } \inf_{\tilde{\mathbf{z}}} \hat{\gamma}_{S_n}(h, \tilde{\mathbf{z}}) > 0\right\}.$$

It is known that the breakdown point of Q_n is 50% ([87]). Inspecting $X/\alpha + Y/\beta$ (or $X/\alpha - Y/\beta$), we can see that as long as x_i (or y_i) is contaminated, then $x_i/\alpha + y_i/\beta$ (or $x_i/\alpha - y_i/\beta$) is contaminated. So in the pairs $(x_1, y_1), \ldots, (x_n, y_n)$, we can at most have half of the pairs containing contaminated data. If we look at one pair as one observation, then the estimators $\hat{\gamma}_Q$ and $\hat{\rho}_Q$ are robust against at most half of the contaminated observations. So, they have breakdown point of 50%. In estimating the covariance matrix Σ and the correlation matrix R, we form pairs of all the observations of X_i and X_j $(i, j = 1, \ldots, p)$, and the estimator allows at most half of the pairs to be contaminated. Therefore, among the n observation vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, at most half of them can contain contaminated data. In other words, the breakdown point of the highly robust componentwise estimators $\hat{\Sigma}_Q$ and $\hat{R}_Q$ is 50%.

4.3.2 Influence function

We denote by γ_Q, ρ_Q, and Q the statistical functional ([53, 58]) corresponding to the estimators $\hat{\gamma}_Q$, $\hat{\rho}_Q$, and Q_n respectively. Consider a sample $Z_1, \ldots, Z_n$ and a scale estimator $S_n(Z_1, \ldots, Z_n)$, i.e. it satisfies $S_n(aZ_1 + b, \ldots, aZ_n + b) = |a|S_n(Z_1, \ldots, Z_n)$, $\forall a, b \in \mathbb{R}$.

We write $S_n(Z_1, \dots, Z_n) = S_n(F_n)$, where $F_n(z) = \frac{1}{n}\sum_{i=1}^{n} \Delta_{Z_i}(z)$ is the empirical distribution, and Δ_{Z_i} is the Dirac function with jump at Z_i. Let $S(F)$ be the corresponding statistical functional of scale such that $S(F_n) = S_n(F_n)$. The influence function ([51]) of S at a distribution F is defined by:

$$IF(u; S, F) = \lim_{\varepsilon \to 0^+} \frac{S((1-\varepsilon)F + \varepsilon\Delta_u) - S(F)}{\varepsilon}, \tag{34}$$

in those u where this limit exists. The importance of the influence function lies in its heuristic interpretation: it describes the effect of an infinitesimal contamination at the point u on the estimate, standardized by the mass of the contamination, i.e. it measures the asymptotic bias caused by the contamination in the observations. The gross-error sensitivity ([53]) defined by $\gamma^*(S, F) = \sup_u |IF(u; S, F)|$, measures the worst asymptotic bias due to the contamination. If $\gamma^*(S, F) < \infty$, the estimator is said to be B-robust, i.e. robust with respect to the bias.

Let Θ be a statistical functional of covariance corresponding to a covariance estimator $\hat{\theta}$ based on Equation (30):

$$\Theta(\mathbf{F}) = \frac{\alpha\beta}{4}\left[S^2(F_+) - S^2(F_-)\right], \tag{35}$$

where $\mathbf{F}$ is a bivariate distribution with marginal distributions F_X and F_Y, and F_+ and F_- denote the distributions of $X/\alpha + Y/\beta$ and $X/\alpha - Y/\beta$ respectively. For simplicity, we assume that F_X and F_Y both have mean zero. A natural way to define the influence function of Θ is through the influence function of S. Note that the influence function describes the first order sensitivity of the estimator to contamination, and thus has similar properties as the usual first derivative.

Proposition 4.1 *Suppose F is a distribution with variance σ, $\tilde{F}$ is the standardized* *distribution of F, i.e.* $\tilde{F}(x) = F(\sigma x)$*, and h is a real differentiable function. Then*

$$\begin{aligned} IF(x, S, F) =& \sigma IF(\frac{x}{\sigma}, S, \tilde{F}) \\ IF(x, S^2, F) =& \sigma^2 IF(\frac{x}{\sigma}, S^2, \tilde{F}) \\ IF(x, h(S), F) =& h'(S(F)) IF(x, S, F) \end{aligned} \tag{36}$$

Moreover, the following equalities on the asymptotic variance hold for independent observations:

$$\begin{aligned} Var(S, F) =& \sigma^2 \, Var(S, \tilde{F}) \\ Var(S^2, F) =& \sigma^4 \, Var(S^2, \tilde{F}) \end{aligned} \tag{37}$$

Proof:

We prove the first equation in detail and briefly explain the latter ones.

$$\begin{aligned} IF(x, S, F) =& \frac{\partial}{\partial \varepsilon} S((1-\varepsilon)F(u) + \varepsilon \Delta_x(u))|_{\varepsilon=0} \\ =& \frac{\partial}{\partial \varepsilon} S((1-\varepsilon)\tilde{F}(\frac{u}{\sigma}) + \varepsilon \Delta_{\frac{x}{\sigma}}(\frac{u}{\sigma}))|_{\varepsilon=0} \end{aligned}$$

We use $P(\frac{u}{\sigma})$ to denote the function $(1-\varepsilon)\tilde{F}(\frac{u}{\sigma}) + \varepsilon \Delta_{\frac{x}{\sigma}}(\frac{u}{\sigma})$, and by the property of equivariance of the scale estimator, we know that $S(P(\frac{u}{\sigma})) = \sigma S(P(u))$. So, we get

$$\begin{aligned} IF(x, S, F) =& \sigma \frac{\partial}{\partial \varepsilon} (S((1-\varepsilon)\tilde{F}(u) + \varepsilon \Delta_{\frac{x}{\sigma}}(u)))|_{\varepsilon=0} \\ =& \sigma IF(\frac{x}{\sigma}, S, \tilde{F}) \end{aligned}$$

The second equality in Equation (36) can be proved similarly as the first one. The third equality is obvious. To prove the equalities in Equation (37), we use the formula $\text{Var}(S, F) = \int |IF(x, S, F)|^2 dF(x)$ and $\text{Var}(S^2, F) = \int |IF(x, S^2, F)|^2 dF(x)$ and the equalities in Equation (36). Note that this result is valid only when the estimator is carried on independent observations. □

From Equation (35), we define the following influence function for Θ:

Definition 4.3

$$\begin{aligned} IF((u,v);\Theta,\mathbf{F}) =& \frac{\alpha\beta}{4}\left[IF(\frac{u}{\alpha}+\frac{v}{\beta};S^2,F_+) - IF(\frac{u}{\alpha}-\frac{v}{\beta};S^2,F_-)\right] \\ =& \frac{\alpha\beta}{2}\left[S(F_+)IF(\frac{u}{\alpha}+\frac{v}{\beta};S,F_+) - S(F_-)IF(\frac{u}{\alpha}-\frac{v}{\beta};S,F_-)\right]. \end{aligned} \tag{38}$$

Defining the influence function of a bivariate estimator through the influence function of a univariate estimator, as in Equation (38), provides a way to generalize the unidimensional Dirac function Δ_u to a bidimensional Dirac function. Note that in this definition, the perturbations we consider depend on the choice of the covariance estimator: they are respectively perturbations along $u/\alpha + v/\beta$ and $u/\alpha - v/\beta$ directions. In fact, this is a typical method to reduce a higher dimensional problem to a lower dimensional one already known. Using $\alpha = \sigma_X$, $\beta = \sigma_Y$ and Proposition 4.1, Equation (38) becomes:

$$\begin{aligned} & IF((u,v);\Theta,\mathbf{F}) \\ =& \frac{\sigma_X\sigma_Y}{2}\left[IF\Big((\frac{u}{\sigma_X}+\frac{v}{\sigma_Y})/\sigma_+;S,\tilde{F}_+\Big)\sigma_+^2 - IF\Big((\frac{u}{\sigma_X}-\frac{v}{\sigma_Y})/\sigma_-;S,\tilde{F}_-\Big)\sigma_-^2\right], \end{aligned} \tag{39}$$

where $\sigma_+ = S(F_+)$ and $\sigma_- = S(F_-)$.

Let R be a statistical functional of correlation corresponding to a correlation estimator $\hat{\rho}$ based on Equation (33):

$$R(\mathbf{F}) = \frac{\alpha\beta}{4S(F_X)S(F_Y)}\left[S^2(F_+) - S^2(F_-)\right]. \tag{40}$$

Similar to the covariance case, the influence function of R is:

$$\begin{aligned} IF((u,v);R,\mathbf{F}) =& IF((u,v);\Theta,\mathbf{F})\frac{1}{S(F_X)S(F_Y)} \\ & - \frac{\Theta(\mathbf{F})}{S^2(F_X)S^2(F_Y)}\big(IF(u;S,F_X)S(F_Y) + IF(v;S,F_Y)S(F_X)\big). \end{aligned} \tag{41}$$

Using Equations (35), (38), as well as $\alpha = \sigma_X$ and $\beta = \sigma_Y$, Equation (41) becomes:

$$IF((u,v);R,\mathbf{F}) = \frac{1}{\sigma_X\sigma_Y} IF((u,v);\Theta,\mathbf{F}) - \rho\big(IF(u/\sigma_X;S,\tilde{F}_X) + IF(v/\sigma_Y;S,\tilde{F}_Y)\big). \quad (42)$$

The links between the gross-error sensitivities for scale and for dispersion estimators are given in the next two propositions. Let us define $\gamma_+^*(S,F) = \sup_u IF(u;S,F)$, $\gamma_-^*(S,F) = -\inf_u IF(u;S,F)$, $\gamma^*(\Theta,\mathbf{F}) = \sup_{u,v} |IF((u,v);\Theta,\mathbf{F})|$, $\gamma^*(R,\mathbf{F}) = \sup_{u,v} |IF((u,v);R,\mathbf{F})|$.

Proposition 4.2 *Let θ be the covariance between two random variables X and Y, and Θ be a statistical functional of covariance based on a statistical functional S of scale. The gross-error sensitivity of Θ is:*

$$\gamma^*(\Theta,\mathbf{F}) = \frac{\sigma_X\sigma_Y}{2} \max\big(\sigma_+^2\gamma_+^*(S,\tilde{F}_+) + \sigma_-^2\gamma_-^*(S,\tilde{F}_-), \sigma_+^2\gamma_-^*(S,\tilde{F}_+) + \sigma_-^2\gamma_+^*(S,\tilde{F}_-)\big).$$

In particular, when $\gamma_+^*(S,\tilde{F}_\pm) = \gamma_-^*(S,\tilde{F}_\pm) = \gamma^*(S,\tilde{F}_\pm)$:

$$\gamma^*(\Theta,\mathbf{F}) = (\sigma_X\sigma_Y + \theta)\gamma^*(S,\tilde{F}_+) + (\sigma_X\sigma_Y - \theta)\gamma^*(S,\tilde{F}_-).$$

Proof:

From Equation (39), the influence function $IF((u,v);\Theta,\mathbf{F})$ must be bounded between $-\frac{\sigma_X\sigma_Y}{2}(\sigma_+^2\gamma_-^*(S,\tilde{F}_+) + \sigma_-^2\gamma_+^*(S,\tilde{F}_-))$ and $\frac{\sigma_X\sigma_Y}{2}(\sigma_+^2\gamma_+^*(S,\tilde{F}_+) + \sigma_-^2\gamma_-^*(S,\tilde{F}_-))$. Because the supremum and infimum of the influence function of S can be reached simultaneously, i.e. at the same (u,v), the two bounds are tight. In particular, when $\gamma_+^*(S,\tilde{F}_\pm) = \gamma_-^*(S,\tilde{F}_\pm) = \gamma^*(S,\tilde{F}_\pm)$, the two bounds have the same absolute value $\frac{\sigma_X\sigma_Y}{2}(\sigma_+^2\gamma^*(S,\tilde{F}_+) + \sigma_-^2\gamma^*(S,\tilde{F}_-)) = (\sigma_X\sigma_Y + \theta)\gamma^*(S,\tilde{F}_+) + (\sigma_X\sigma_Y - \theta)\gamma^*(S,\tilde{F}_-)$. □

Proposition 4.3 *Let ρ be the correlation between two random variables X and Y, and R be a statistical functional of correlation based on a stqtistical functional S of scale.*

The gross-error sensitivity of R is:

$$\begin{aligned}\gamma^*(R,\mathbf{F}) = \max\Big(&\frac{1}{2}\sigma_+^2\gamma_+^*(S,\tilde{F}_+) + \frac{1}{2}\sigma_-^2\gamma_-^*(S,\tilde{F}_-) + \rho\big(\gamma_-^*(S,\tilde{F}_X) + \gamma_-^*(S,\tilde{F}_Y)\big),\\ &\frac{1}{2}\sigma_+^2\gamma_-^*(S,\tilde{F}_+) + \frac{1}{2}\sigma_-^2\gamma_+^*(S,\tilde{F}_-) + \rho\big(\gamma_+^*(S,\tilde{F}_X) + \gamma_+^*(S,\tilde{F}_Y)\big)\Big), \mathit{for}\, \rho \geq 0;\\ \gamma^*(R,\mathbf{F}) = \max\Big(&\frac{1}{2}\sigma_+^2\gamma_+^*(S,\tilde{F}_+) + \frac{1}{2}\sigma_-^2\gamma_-^*(S,\tilde{F}_-) - \rho\big(\gamma_+^*(S,\tilde{F}_X) + \gamma_+^*(S,\tilde{F}_Y)\big),\\ &\frac{1}{2}\sigma_+^2\gamma_-^*(S,\tilde{F}_+) + \frac{1}{2}\sigma_-^2\gamma_+^*(S,\tilde{F}_-) - \rho\big(\gamma_-^*(S,\tilde{F}_X) + \gamma_-^*(S,\tilde{F}_Y)\big)\Big), \mathit{for}\, \rho < 0.\end{aligned}$$

In particular, when $\gamma_+^*(S,\tilde{F}_\pm) = \gamma_-^*(S,\tilde{F}_\pm) = \gamma^*(S,\tilde{F}_\pm)$:

$$\gamma^*(R,\mathbf{F}) = \frac{\sigma_+^2}{2}\gamma^*(S,\tilde{F}_+) + \frac{\sigma_-^2}{2}\gamma^*(S,\tilde{F}_-) + |\rho|\big(\gamma^*(S,\tilde{F}_X) + \gamma^*(S,\tilde{F}_Y)\big).$$

Proof: Similar to Proposition 4.2. □

Propositions 4.2 and 4.3 tell us that the dispersion estimators are B-robust if the underlying scale estimators are B-robust. The most interesting M-estimators of scale satisfy $\gamma_+^*(S,\tilde{F}_\pm) \geq \gamma_-^*(S,\tilde{F}_\pm)$, with equality when they have 50% breakdown point ([44, 58]). The opposite situation leads to implosion of the scale estimator as well as lower efficiency. Observe that often $\tilde{F}_+ = \tilde{F}_- = \tilde{F}_X = \tilde{F}_Y$ in Propositions 4.2 and 4.3, yielding further simplifications. For instance, this is the case for multivariate Gaussian distributions, and even for some specific members of the more general class of elliptically contoured distributions ([39]) like multivariate t or multivariate Cauchy distributions. Conditions for this property to hold is given by Fang and Zhang [65]. Note that in order to compare the gross-error sensitivities of two dispersion estimators, one should standardize them [53, page 228-229], for example with respect to their variances (self-standardized), or to the Fisher information (information-standardized).

It can be checked that the influence functions of both the covariance estimator and the correlation estimator satisfy $\int IFd\mathbf{F} = 0$.

4.3.3 Asymptotic variance

Under regularity conditions, both $\hat{\gamma}_Q$ and $\hat{\rho}_Q$ are consistent estimators, since Q_n is consistent ([87]). Moreover, they are asymptotically normal with asymptotic variance given by:

$$\begin{aligned} V(\gamma_Q, \mathbf{F}) &= \int IF((u,v); \gamma_Q, \mathbf{F})^2 d\mathbf{F}(u,v), \\ V(\rho_Q, \mathbf{F}) &= \int IF((u,v); \rho_Q, \mathbf{F})^2 d\mathbf{F}(u,v), \\ V(Q,F) &= \int IF(u; Q, F)^2 dF(u). \end{aligned} \tag{43}$$

Subsequently, we assume a bivariate Gaussian distribution $\mathbf{F} = \vec{\Phi}$ for $(X,Y)^T$, i.e.:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \theta \\ \theta & \sigma_Y^2 \end{pmatrix} \right) = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \tau\sigma_X\sigma_Y \\ \tau\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right),$$

where θ is the covariance and τ is the correlation between X and Y. We have:

Proposition 4.4 *The asymptotic variance of the estimator $\hat{\gamma}_Q$ is*

$$V(\gamma_Q, \vec{\Phi}) = 2V(Q, \Phi)(\sigma_X^2\sigma_Y^2 + \theta^2) = 1.215(\sigma_X^2\sigma_Y^2 + \theta^2). \tag{44}$$

Here, Φ represents the standard Gaussian distribution function, i.e. with mean zero and variance one.

Remark 4.1 *A much more general result can be parallelly proven to be true: Let Θ be a statistical functional of covariance based on a statistical functional S of scale. The asymptotic variance of Θ at the bivariate Gaussian distribution Φ_τ is:*

$$V(\Theta, \Phi_\tau) = 2(\sigma_X^2\sigma_Y^2 + \theta^2)V(S, \Phi),$$

where $V(S,\Phi)$ is the asymptotic variance of S at Φ.

Proof:

The asymptotic variance of $\hat{\gamma}_Q$ at $\vec{\Phi}$ is

$$\begin{aligned} V(\gamma_Q,\vec{\Phi}) &= \iint IF^2((u,v);\gamma_Q,\vec{\Phi})d\vec{\Phi}(u,v) \\ &= \frac{\sigma_X^2\sigma_Y^2}{4}\iint \left[\sigma_+ IF(\frac{u}{\sigma_X}+\frac{v}{\sigma_Y};Q,\Phi_+) - \sigma_- IF(\frac{u}{\sigma_X}-\frac{v}{\sigma_Y};Q,\Phi_-)\right]^2 d\vec{\Phi}(u,v). \end{aligned}$$

The change of variables

$$\begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_+\sigma_X} & \frac{1}{\sigma_+\sigma_Y} \\ \frac{1}{\sigma_-\sigma_X} & \frac{-1}{\sigma_-\sigma_Y} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix},$$

yields

$$dsdt = \frac{2}{\sigma_+\sigma_-\sigma_X\sigma_Y}dudv,$$

and corresponds to the random variables $\frac{X}{\sigma_+\sigma_X}+\frac{Y}{\sigma_+\sigma_Y}$ and $\frac{X}{\sigma_-\sigma_X}-\frac{Y}{\sigma_-\sigma_Y}$, each of which follows the standard normal distribution Φ and is independent of each other. Therefore

$$V(\gamma_Q,\vec{\Phi}) = \frac{\sigma_X^2\sigma_Y^2}{4}\left[\sigma_+^4\iint IF^2(s;Q,\Phi)d\Phi(s)d\Phi(t) + 0 + \sigma_-^4\iint IF^2(t;Q,\Phi)d\Phi(s)d\Phi(t)\right].$$

Note that we use the linear property of the influence function given in Proposition 4.1: $IF(\alpha x; Q, \Phi_{\alpha X}) = \alpha IF(x; Q, \Phi_X), \forall \alpha \in \mathbb{R}$. Thus:

$$\begin{aligned} V(\gamma_Q, \vec{\Phi}) &= \frac{\sigma_X^2 \sigma_Y^2}{4}\left[\sigma_+^4 + 0 + \sigma_-^4\right] V(Q, \Phi) \\ &= \frac{\sigma_X^2 \sigma_Y^2}{4}\left[(2 + 2\frac{\theta}{\sigma_X \sigma_Y})^2 + (2 - 2\frac{\theta}{\sigma_X \sigma_Y})^2\right] V(Q, \Phi) \\ &= 2V(Q, \Phi)(\sigma_X^2 \sigma_Y^2 + \theta^2) \\ &= 1.215(\sigma_X^2 \sigma_Y^2 + \theta^2) \end{aligned}$$

□

Note that due to the form of Equation (41), a closed form of the asymptotic variance of the correlation estimator $\hat{\rho}_Q$ is not available. However, following the formulas in Equation (43), we calculate numerically (i.e. by numerical integration) the variance of the covariance estimator and the correlation estimator for various underlying variances and covariances. The results are presented in the fourth and fifth columns of Table 4.3.3. The numerical results for the covariance estimator agree with the theoretical result given in Proposition 4.4 very well.

Following Remark 4.1, we can replace the Q_n estimator in Proposition 4.4 with the maximum likelihood estimator of scale MLE, and calculate the closed form of the variance of the covariance estimator $\hat{\gamma}_{MLE}$ and the correlation estimator $\hat{\rho}_{MLE}$:

Proposition 4.5 *The influence function of the maximum likelihood estimator of covariance $\hat{\gamma}_{MLE}$ is $IF((u, v); \gamma_{MLE}, \vec{\Phi}) = uv - \theta$.*

The asymptotic variance of the estimator $\hat{\gamma}_{MLE}$ is

$$V(\gamma_{MLE}, \vec{\Phi}) = \sigma_X^2 \sigma_Y^2 + \theta^2. \tag{45}$$

The influence function of the maximum likelihood estimator of correlation $\hat{\rho}_{MLE}$ is

$$IF((u, v); \rho_{MLE}, \vec{\Phi}) = \frac{uv}{\sigma_X \sigma_Y} - \frac{\tau u^2}{2{\sigma_X}^2} - \frac{\tau v^2}{2{\sigma_Y}^2}.$$

The asymptotic variance of the estimator $\hat{\rho}_{MLE}$ is

$$V(\rho_{MLE}, \vec{\Phi}) = (1-\tau^2)^2. \tag{46}$$

Proof:

The influence function of the MLE of scale, S_{MLE}, is $IF(x; S_{MLE}, \Phi) = \frac{1}{2}(x^2 - 1)$. So

$$\begin{aligned} IF((u,v); \gamma_{MLE}, \vec{\Phi}) &= \frac{\sigma_X\sigma_Y}{2}\left[\sigma_+ IF(\frac{u}{\sigma_X} + \frac{v}{\sigma_Y}; S_{MLE}, \Phi_+) - \sigma_- IF(\frac{u}{\sigma_X} - \frac{v}{\sigma_Y}; S_{MLE}, \Phi_-)\right] \\ &= \frac{\sigma_X\sigma_Y}{2}\left[\sigma_+^2(\frac{1}{2}(\frac{u}{\sigma_+\sigma_X} + \frac{v}{\sigma_+\sigma_Y})^2 - 1) - \sigma_-^2(\frac{1}{2}(\frac{u}{\sigma_-\sigma_X} - \frac{v}{\sigma_-\sigma_Y})^2 - 1)\right] \\ &= uv - \theta \end{aligned}$$

Using Remark 4.1, the asymptotic variance of $\hat{\gamma}_{MLE}$ is given by

$$V(\gamma_{MLE}, \vec{\Phi}) = 2V(S_{MLE}, \Phi)(\sigma_X^2\sigma_Y^2 + \theta^2) = \sigma_X^2\sigma_Y^2 + \theta^2.$$

For correlation, we use Equation (41) and get

$$\begin{aligned} &IF((u,v); \rho_{MLE}, \vec{\Phi}) \\ &= IF((u,v); \gamma_{MLE}, \vec{\Phi})/(\sigma_X\sigma_Y) - \theta IF(u; S_{MLE}, \Phi_X)/(\sigma_X^2\sigma_Y) - \theta IF(v; M, \Phi_Y)/(\sigma_X\sigma_Y^2) \\ &= \frac{uv-\theta}{\sigma_X\sigma_Y} - \theta\frac{(u/\sigma_X)^2 - 1}{2\sigma_X\sigma_Y} - \theta\frac{(v/\sigma_Y)^2 - 1}{2\sigma_X\sigma_Y} \\ &= \frac{uv}{\sigma_X\sigma_Y} - \frac{\tau u^2}{2\sigma_X^2} - \frac{\tau v^2}{2\sigma_Y^2} \end{aligned}$$

The asymptotic variance of this estimator is

$$V(\rho_{MLE}, \vec{\Phi}) = \iint (\frac{uv}{\sigma_X\sigma_Y} - \frac{\tau u^2}{2\sigma_X^2} - \frac{\tau v^2}{2\sigma_Y^2})^2 d\vec{\Phi}(u,v) = 1 - 2\tau^2 + \tau^4.$$

□

σ_X^2	σ_Y^2	θ	$V(\gamma_Q, \vec{\Phi})$	$V(\rho_Q, \vec{\Phi})$	$\mathrm{Eff}(\gamma_Q, \vec{\Phi})$	$\mathrm{Eff}(\rho_Q, \vec{\Phi})$
1	1	0	1.215	1.215	0.823	0.823
1	1	0.2	1.264	1.132	0.701	0.783
1	1	0.5	1.519	0.725	0.296	0.621
1	1	0.8	1.993	0.183	0.040	0.431
1	2	0.5	2.735	1.044	0.498	0.652
1	3	0.5	3.950	1.133	0.589	0.685
1	10	0.5	12.458	1.215	0.745	0.763

Table 3: Asymptotic variance and efficiency (calculated by numerical integration) of the dispersion estimators $\hat{\gamma}_Q$ and $\hat{\rho}_Q$, in the case of Gaussian distributions. The numerical values of $V(\gamma_Q, \vec{\Phi})$ agree with the result in Proposition 4.4 and the values of $\mathrm{Eff}(\gamma_Q, \vec{\Phi})$ agree with the result in Proposition 4.8.

We observe that the behavior of the asymptotic variance of correlation estimators at bivariate Gaussian distributions is opposite to the one for covariance estimators. It seems that it is maximal in the independent case, and decreases strictly with the absolute value of the underlying correlation. However, no simple proof is available, due to the much more complicated form of the influence function of correlation estimators.

4.3.4 Fisher information

For Gaussian distributions, a closed form of the Fisher information of both covariance and correlation can be obtained:

Proposition 4.6 *The Fisher information of the covariance θ is*

$$I(\theta, \vec{\Phi}) = \frac{\sigma_X^2 \sigma_Y^2 + \theta^2}{(\sigma_X^2 \sigma_Y^2 - \theta^2)^2}. \tag{47}$$

Proof:

We write out the probability density function of the bivariate Gaussian distribution:

$$\begin{aligned}\vec{\Phi}_\theta(u,v) &= \frac{1}{2\pi\sqrt{ab-\theta^2}}\exp\left(-\frac{1}{2}(u\ v)\begin{pmatrix} a & \theta \\ \theta & b\end{pmatrix}^{-1}\begin{pmatrix} u \\ v\end{pmatrix}\right)\\ &= \frac{1}{2\pi\sqrt{ab-\theta^2}}\exp\Big(\frac{bu^2+av^2-2\theta uv}{-2(ab-\theta^2)}\Big)\\ &= \frac{1}{\pi\sqrt{2B}}\exp\Big(-\frac{A}{B}\Big),\end{aligned}$$

where $A = bu^2 + av^2 - 2\theta uv$ and $B = 2ab - 2\theta^2$. Following the definition of the Fisher information, we have

$$\begin{aligned}I(\theta) &= \iint (\frac{\partial}{\partial\theta}\log\vec{\Phi}_\theta(u,v))^2\vec{\Phi}_\theta(u,v)dudv\\ &= \iint 2\Big(\frac{\theta e^{-\frac{A}{B}}}{2\pi(ab-\theta^2)^{\frac{3}{2}}} + \frac{(\frac{2uv}{B} - \frac{4\theta A}{B^2})e^{-\frac{A}{B}}}{2\pi(ab-\theta^2)^{\frac{1}{2}}}\Big)^2\frac{\pi(ab-\theta^2)^{\frac{1}{2}}}{e^{-\frac{A}{B}}}dudv \qquad (48)\\ &= \iint \frac{e^{-\frac{A}{B}}}{2\pi}(ab-\theta^2)^{-\frac{9}{2}}\big[(ab-\theta^2)\theta - \theta(bu^2+av^2) + (ab+\theta^2)uv\big]^2 dudv.\end{aligned}$$

Let

$$\begin{cases} s = \sqrt{b}u + \sqrt{a}v, \\ t = \sqrt{b}u - \sqrt{a}v. \end{cases}$$

Then, we have:

$$\begin{aligned}bu^2 + av^2 &= \frac{s^2+t^2}{2},\\ uv &= \frac{s^2-t^2}{2},\\ dudv &= \frac{1}{2\sqrt{ab}}dsdt,\end{aligned}$$

and Equation (48) becomes

$$\begin{aligned}I(\theta) =& \iint \frac{e^{-\frac{A}{B}}}{2\pi}(ab-\theta^2)^{-\frac{9}{2}}\big[(ab-\theta^2)\theta - \theta\frac{s^2+t^2}{2} + (ab+\theta^2)\frac{s^2-t^2}{4\sqrt{ab}}\big]^2\frac{1}{2\sqrt{ab}}dsdt \\ =& \iint e^{-\frac{(\sqrt{ab}-\theta)s^2+(\sqrt{ab}+\theta)t^2}{4(ab-\theta^2)\sqrt{ab}}}\big[(ab-\theta^2)\theta + \frac{(\sqrt{ab}-\theta)^2s^2}{4\sqrt{ab}} - \frac{(\sqrt{ab}+\theta)^2t^2}{4\sqrt{ab}}\big]^2 dsdt \qquad (49) \\ & \cdot\frac{(ab-\theta^2)^{-\frac{9}{2}}}{4\pi\sqrt{ab}}\end{aligned}$$

Let $p = \sqrt{\frac{s^2}{4\sqrt{ab}(\sqrt{ab}+\theta)}}$ and $q = \sqrt{\frac{t^2}{4\sqrt{ab}(\sqrt{ab}-\theta)}}$. Then Equation (49) becomes

$$\begin{aligned}I(\theta) =& \iint e^{-p^2-q^2}\big[(ab-\theta^2)\theta + (\sqrt{ab}+\theta)(\sqrt{ab}-\theta)^2p^2 - (\sqrt{ab}-\theta)(\sqrt{ab}+\theta)^2q^2\big]^2 \\ & \cdot 4\sqrt{ab}\sqrt{ab-\theta^2}dpdq\frac{(ab-\theta^2)^{-\frac{9}{2}}}{4\pi\sqrt{ab}} \\ =&(ab-\theta^2)^{-2}(ab+\theta^2)\end{aligned}$$

□

From the Fisher information for the covariance θ, it is straightforward to get the Fisher information for the correlation, since the correlation τ is simply $\frac{\theta}{\sigma_X\sigma_Y}$. Thus, we obtain:

Proposition 4.7 *The Fisher information of the correlation τ is*

$$I(\tau, \vec{\Phi}) = \frac{1+\tau^2}{(1-\tau^2)^2}. \qquad (50)$$

Proof:

By the definition of the Fisher information, we know

$$I(\tau) = I(\theta)(\frac{d\theta}{d\tau})^2,$$

where $\theta = \sqrt{ab}\tau$ in this case. Using Equation (47), we get

$$\begin{aligned} I(\tau) =& (ab - ab\tau^2)^{-2}(ab + ab\tau^2)(ab) \\ =& (1 - \tau^2)^{-2}(1 + \tau^2). \end{aligned}$$

□

4.3.5 Efficiency

Efficiency is defined as the inverse of the product of the Fisher information and the asymptotic variance of the estimator. For Gaussian distributions, we can calculate the efficiency of γ_Q.

Proposition 4.8 *The efficiency of the covariance estimator $\hat{\gamma}_Q$ is*

$$\mathit{Eff}(\gamma_Q, \vec{\Phi}) = \frac{(\sigma_X^2\sigma_Y^2 - \theta^2)^2}{2V(Q,\Phi)(\sigma_X^2\sigma_Y^2 + \theta^2)^2} = 0.823\frac{(\sigma_X^2\sigma_Y^2 - \theta^2)^2}{(\sigma_X^2\sigma_Y^2 + \theta^2)^2}. \tag{51}$$

We present the efficiency of both the covariance and the correlation estimators in the sixth and seventh column of Table 4.3.3, calculated by numerical integration of the asymptotic variance. The numerical results of the covariance estimator are very close to the theoretical result given in Proposition 4.8.

In comparison with the estimator based on Q_n, we calculate the efficiency of covariance and correlation estimators based on MLE. As expected, the efficiency of γ_{MLE} is higher.

Proposition 4.9 *The efficiency of the maximum likelihood estimator of the covariance $\hat{\gamma}_{MLE}$ is*

$$\mathit{Eff}(\gamma_{MLE}, \vec{\Phi}) = \frac{(\sigma_X^2\sigma_Y^2 - \theta^2)^2}{(\sigma_X^2\sigma_Y^2 + \theta^2)^2}. \tag{52}$$

The efficiency of the maximum likelihood estimator of the correlation $\hat{\rho}_{MLE}$ is

$$\mathit{Eff}(\rho_{MLE}, \vec{\Phi}) = \frac{1}{1+\tau^2}. \tag{53}$$

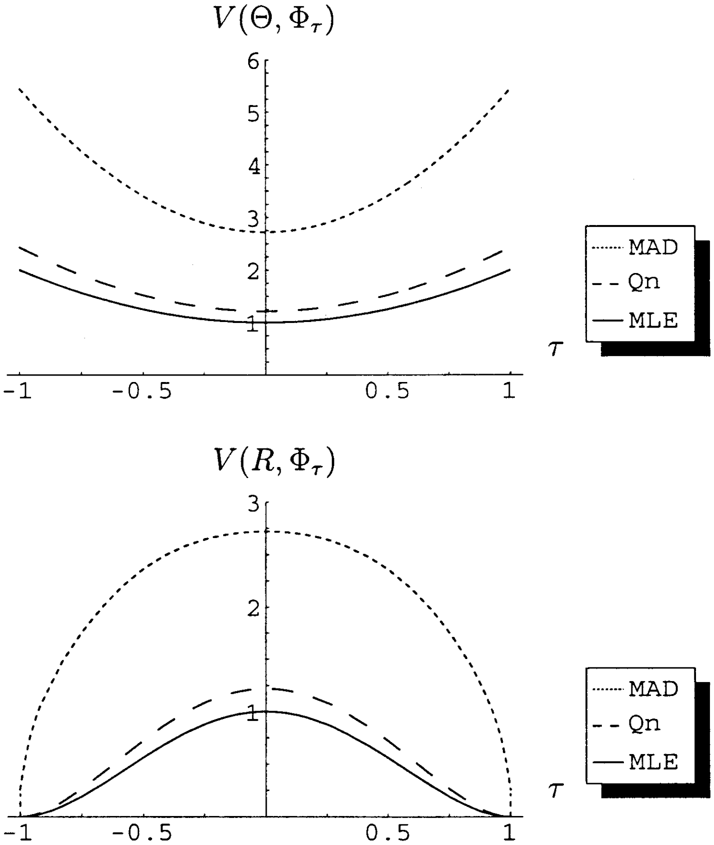


Figure 18: The asymptotic variance of the covariance (top) and correlation (bottom) estimators based on MLE, MAD, and Q_n respectively, for a standardized bivariate Gaussian distribution with covariance τ. The $\hat{\gamma}_{MLE}$ estimator has the smallest asymptotic variance, the asymptotic variance of the $\hat{\gamma}_Q$ estimator is slightly larger, whereas $\hat{\gamma}_{MAD}$ has an asymptotic variance much larger than the other two. For all three covariance estimators, the asymptotic variance increases when the covariance between the two random variables increases. For all three correlation estimators, the asymptotic variance decreases when the covariance between the two random variables increases.

4.4 Comparisons

We first compare the estimator $\hat{\gamma}_Q$ with the maximum likelihood estimator $\hat{\gamma}_{MLE}$ and another componentwise robust estimator $\hat{\gamma}_{MAD}$, based on the median absolute deviation ([53]). Next we compare $\hat{\Sigma}_Q$ with the global estimator $\hat{\Sigma}_{MVE}$ and with the maximum likelihood estimator $\hat{\Sigma}_{MLE}$. We focus on covariance estimation here since as we will point out in Section 4.4, it is more difficult than correlation estimation.

4.4.1 Comparison with MLE and MAD

As we have pointed out, Proposition 4.4 is valid for any covariance estimator based on an M-estimator of scale ([43]). In Figure 18, we plot the asymptotic variance of the three covariance and correlation estimators $\hat{\gamma}_Q$, $\hat{\gamma}_{MLE}$ and $\hat{\gamma}_{MAD}$, for a standardized Gaussian distribution with covariance τ.

The three curves for the variance of the covariance estimators in Figure 18 are computed with the formula in Proposition 4.4, whereas the three curves of the correlation estimators are computed numerically with formula in Equations (41) and (43) (except for the MLE where formula (46) is used). We can see that when the covariance between two random variables increases, the variance of the covariance estimator increases, while the variance of the correlation estimator decreases. As a consequence, correlation estimation is easier than covariance estimation, in the sense that it has smaller variability. In the independent standard Gaussian distribution case, the variance of the covariance estimator and the correlation estimator have the same value.

We carry out some simulations to test the mean and variance of the dispersion estimators based on the MLE, MAD and Q_n estimators. The simulation is on two standardized Gaussian random variables with covariance 0 and 0.5, and based on 1000 samples. The sample sizes are 20, 100 and 200. The results are presented in Table 4. We can see that the estimators are unbiased and the variance of the estimators increases as the variance between the two random variables increases.

	sample size	mean			variance		
		$\hat{\gamma}_Q$	$\hat{\gamma}_{MLE}$	$\hat{\gamma}_{MAD}$	$\hat{\gamma}_Q$	$\hat{\gamma}_{MLE}$	$\hat{\gamma}_{MAD}$
covariance=0	20	-0.007	0.005	-0.002	1.630	0.966	2.684
	100	-0.002	-0.002	-0.002	1.257	0.988	2.865
	200	-0.003	-0.003	-0.003	1.320	1.057	2.794
covariance=0.5	20	0.526	0.477	0.506	2.018	1.163	3.497
	100	0.499	0.493	0.496	1.715	1.302	3.477
	200	0.504	0.500	0.500	1.649	1.258	3.254

Table 4: The mean and variance of the covariance estimators $\hat{\gamma}_Q$, $\hat{\gamma}_{MLE}$ and $\hat{\gamma}_{MAD}$. The data followed an independent standard Gaussian distribution, and a Gaussian distribution with means zero, variances one, covariance between the two random variables 0.5 respective. We calculated the mean and variance after running 1000 samples. The three estimators are all unbiased, and the variance of the $\hat{\gamma}_{MAD}$ is significantly larger than the other two.

4.4.2 Comparison with MVE and MLE

In order to compare the highly robust componentwise estimator $\hat{\Sigma}_Q$ with the minimum volume ellipsoid estimator $\hat{\Sigma}_{MVE}$ and the global maximum likelihood estimator $\hat{\Sigma}_{MLE}$, we carry out some simulations on three variables, i.e. Σ is a 3×3 matrix. In Table 5,

$$\Sigma = \begin{pmatrix} 1.0 & 0.9 & -0.5 \\ 0.9 & 2.0 & 0.2 \\ -0.5 & 0.2 & 3.0 \end{pmatrix}, \tag{54}$$

and in Table 6,

$$\Sigma = \begin{pmatrix} 0.050 & -0.010 & 0.005 \\ -0.001 & 1.520 & 0 \\ 0.005 & 0 & 1 \end{pmatrix}. \tag{55}$$

We generate 1000 sets of data, each with sample size 100 and we use the three estimators to calculate the covariance matrix Σ. In the first columns, the data do not contain any outliers, in the second column, 10% of the data have a covariance matrix 9Σ (explode type outliers), in the third column, 10% of the data have a covariance matrix $\Sigma/9$ (implode type outliers). Based on the 1000 estimated covariance matrices, we compute the mean and the variance of the estimations. The results are presented in Table 5 and 6. In these examples, the matrices $\hat{\Sigma}_Q$ are positive definite. In case there are not positive definite, a transformation as described at the end of Section 4.2.2 must be applied. For convenience, we call the sum of the absolute values of all the entries of a matrix the 1-norm of the matrix. The smallest 1-norm in each column is emphasized by boldface font. From the tables, we can see that when there is no outliers, $\hat{\Sigma}_{MLE}$ behaves the best, $\hat{\Sigma}_Q$ is slightly worse, while $\hat{\Sigma}_{MVE}$ behaves the worst. When the outliers are of explode type (the observation tends to be much larger than the true value), $\hat{\Sigma}_{MVE}$ has the best estimation, whereas $\hat{\Sigma}_{MLE}$ gives the worst result. For outliers that are of implode type (the observation tends to be much smaller than the true value), $\hat{\Sigma}_Q$ and $\hat{\Sigma}_{MLE}$ both give relatively good estimation, whereas $\hat{\Sigma}_{MVE}$ gives the worst result.

This can be understood if we notice that the estimator $\hat{\Sigma}_{MVE}$ only takes into account half of the observations which are distributed nearest to an estimated center. Thus exploding outliers will not have much effect on the estimator, whereas imploding outliers can bring significant challenge to the estimator. In other words, $\hat{\Sigma}_{MVE}$ is robust only against exploding outliers, not imploding outliers. $\hat{\Sigma}_{MLE}$ gives very good results in the imploding case because the implode values we tested are not extreme case and they only take 10% of the data, so under the averaging procedure, the effect of imploding is very small. $\hat{\Sigma}_Q$ is not the best in any of the three simulations, but it is relatively good in all three simulations. So, in practice when one does not really know what kind of outliers exist and how many percentage of the data are contaminated, $\hat{\Sigma}_Q$ is a suitable estimator to use.

	no outliers	10% explode	10% implode
bias of $\hat{\Sigma}_Q$	$\begin{pmatrix} -0.007 & -0.012 & -0.004 \\ -0.012 & -0.011 & -0.003 \\ -0.004 & -0.003 & 0.044 \end{pmatrix}$	$\begin{pmatrix} 0.270 & 0.232 & -0.136 \\ 0.232 & 0.515 & 0.052 \\ -0.136 & 0.052 & 0.853 \end{pmatrix}$	$\begin{pmatrix} -0.130 & -0.119 & 0.068 \\ -0.119 & -0.269 & -0.031 \\ 0.068 & -0.031 & -0.403 \end{pmatrix}$
1-norm of bias of $\hat{\Sigma}_Q$	0.100	2.478	1.237
Variance of $\hat{\Sigma}_Q$	$\begin{pmatrix} 0.027 & 0.036 & 0.046 \\ 0.036 & 0.097 & 0.080 \\ 0.046 & 0.080 & 0.236 \end{pmatrix}$	$\begin{pmatrix} 0.118 & 0.116 & 0.100 \\ 0.116 & 0.438 & 0.140 \\ 0.100 & 0.140 & 1.144 \end{pmatrix}$	$\begin{pmatrix} 0.040 & 0.045 & 0.038 \\ 0.045 & 0.170 & 0.062 \\ 0.038 & 0.062 & 0.361 \end{pmatrix}$
1-norm of variance of $\hat{\Sigma}_Q$	0.683	2.411	0.860
bias of $\hat{\Sigma}_{MLE}$	$\begin{pmatrix} -0.004 & -0.005 & -0.007 \\ -0.005 & -0.013 & -0.011 \\ -0.007 & -0.011 & 0.020 \end{pmatrix}$	$\begin{pmatrix} 0.806 & 0.722 & -0.407 \\ 0.722 & 1.580 & 0.165 \\ -0.407 & 0.165 & 2.501 \end{pmatrix}$	$\begin{pmatrix} -0.092 & -0.082 & 0.045 \\ -0.082 & -0.179 & -0.020 \\ 0.045 & -0.020 & -0.272 \end{pmatrix}$
1-norm of bias of $\hat{\Sigma}_{MLE}$	**0.081**	7.476	**0.837**
Variance of $\hat{\Sigma}_{MLE}$	$\begin{pmatrix} 0.020 & 0.028 & 0.035 \\ 0.028 & 0.075 & 0.060 \\ 0.035 & 0.060 & 0.182 \end{pmatrix}$	$\begin{pmatrix} 0.832 & 0.778 & 0.483 \\ 0.778 & 3.163 & 0.581 \\ 0.483 & 0.581 & 8.112 \end{pmatrix}$	$\begin{pmatrix} 0.028 & 0.032 & 0.029 \\ 0.032 & 0.109 & 0.052 \\ 0.029 & 0.052 & 0.235 \end{pmatrix}$
1-norm of variance of $\hat{\Sigma}_{MLE}$	**0.522**	15.792	**0.599**
bias of $\hat{\Sigma}_{MVE}$	$\begin{pmatrix} -0.162 & -0.148 & 0.071 \\ -0.148 & -0.315 & -0.035 \\ 0.071 & -0.035 & -0.415 \end{pmatrix}$	$\begin{pmatrix} -0.093 & -0.093 & 0.043 \\ -0.093 & -0.207 & -0.017 \\ 0.043 & -0.017 & -0.248 \end{pmatrix}$	$\begin{pmatrix} -0.272 & -0.243 & 0.134 \\ -0.243 & -0.550 & -0.067 \\ 0.134 & -0.067 & -0.818 \end{pmatrix}$
1-norm of bias of $\hat{\Sigma}_{MVE}$	1.399	**0.853**	2.527
Variance of $\hat{\Sigma}_{MVE}$	$\begin{pmatrix} 0.053 & 0.059 & 0.054 \\ 0.059 & 0.205 & 0.083 \\ 0.054 & 0.083 & 0.416 \end{pmatrix}$	$\begin{pmatrix} 0.038 & 0.050 & 0.058 \\ 0.050 & 0.159 & 0.104 \\ 0.058 & 0.104 & 0.336 \end{pmatrix}$	$\begin{pmatrix} 0.099 & 0.093 & 0.059 \\ 0.093 & 0.405 & 0.083 \\ 0.059 & 0.083 & 0.903 \end{pmatrix}$
1-norm of variance of $\hat{\Sigma}_{MVE}$	1.066	**0.958**	1.877

Table 5: The biases and variances of the estimators $\hat{\Sigma}_Q$, $\hat{\Sigma}_{MLE}$ and $\hat{\Sigma}_{MVE}$, Σ given in Equation (54).

4.5 The Robust Autocovariance Estimator: an Application

The autocovariance function describes the covariance between observations at different time lag distances h. Just like in the covariance context, we define the highly robust autocovariance function estimator as follows. Extract the first $n-h$ observations of $\mathbf{x} = (X_1, \ldots, X_n)^T$ to produce a vector $\mathbf{u}$ with length $n-h$ and the last $n-h$ observation of $\mathbf{x}$ to produce a vector $\mathbf{v}$ of length $n-h$, as shown in Figure 19. Then:

$$\hat{\gamma}_Q(h, \mathbf{x}) = \frac{1}{4}\Big[Q_{n-h}^2(\mathbf{u}+\mathbf{v}) - Q_{n-h}^2(\mathbf{u}-\mathbf{v})\Big]. \tag{56}$$

	no outliers	10% explode	10% implode
bias of $\hat{\Sigma}_Q$	$\begin{pmatrix} 0.001 & -0.001 & -0.000 \\ -0.001 & 0.011 & -0.010 \\ -0.000 & -0.010 & -0.009 \end{pmatrix}$	$\begin{pmatrix} 0.014 & -0.004 & 0.001 \\ -0.004 & 0.396 & 0.007 \\ 0.001 & 0.007 & 0.269 \end{pmatrix}$	$\begin{pmatrix} -0.007 & 0.002 & -0.001 \\ 0.002 & -0.208 & -0.001 \\ -0.001 & -0.001 & -0.133 \end{pmatrix}$
1-norm of bias of $\hat{\Sigma}_Q$	0.041	0.703	0.354
Variance of $\hat{\Sigma}_Q$	$\begin{pmatrix} 0.000 & 0.001 & 0.001 \\ 0.001 & 0.058 & 0.020 \\ 0.001 & 0.020 & 0.025 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.002 & 0.001 \\ 0.002 & 0.258 & 0.035 \\ 0.001 & 0.035 & 0.119 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.001 & 0.001 \\ 0.001 & 0.092 & 0.018 \\ 0.001 & 0.018 & 0.039 \end{pmatrix}$
1-norm of variance of $\hat{\Sigma}_Q$	0.127	0.452	0.171
bias of $\hat{\Sigma}_{MLE}$	$\begin{pmatrix} 0.000 & -0.000 & -0.000 \\ -0.000 & -0.000 & -0.007 \\ -0.000 & -0.007 & -0.010 \end{pmatrix}$	$\begin{pmatrix} 0.040 & -0.008 & 0.006 \\ -0.008 & 1.182 & 0.014 \\ 0.006 & 0.014 & 0.817 \end{pmatrix}$	$\begin{pmatrix} -0.004 & 0.000 & -0.001 \\ 0.000 & -0.142 & 0.002 \\ -0.001 & 0.002 & -0.090 \end{pmatrix}$
1-norm of bias of $\hat{\Sigma}_{MLE}$	**0.025**	2.097	**0.242**
Variance of $\hat{\Sigma}_{MLE}$	$\begin{pmatrix} 0.000 & 0.001 & 0.001 \\ 0.001 & 0.047 & 0.015 \\ 0.001 & 0.015 & 0.019 \end{pmatrix}$	$\begin{pmatrix} 0.002 & 0.007 & 0.005 \\ 0.007 & 1.812 & 0.138 \\ 0.005 & 0.138 & 0.852 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.001 & 0.000 \\ 0.001 & 0.059 & 0.014 \\ 0.000 & 0.014 & 0.027 \end{pmatrix}$
1-norm of variance of $\hat{\Sigma}_{MLE}$	**0.099**	2.967	**0.117**
bias of $\hat{\Sigma}_{MVE}$	$\begin{pmatrix} -0.007 & 0.002 & -0.001 \\ 0.002 & -0.220 & -0.005 \\ -0.001 & -0.005 & -0.162 \end{pmatrix}$	$\begin{pmatrix} -0.005 & -0.000 & -0.000 \\ -0.000 & -0.138 & 0.004 \\ -0.000 & 0.004 & -0.089 \end{pmatrix}$	$\begin{pmatrix} -0.014 & 0.003 & -0.002 \\ 0.003 & -0.420 & -0.003 \\ -0.002 & -0.003 & -0.272 \end{pmatrix}$
1-norm of bias of $\hat{\Sigma}_{MVE}$	0.404	**0.240**	0.721
Variance of $\hat{\Sigma}_{MVE}$	$\begin{pmatrix} 0.000 & 0.001 & 0.001 \\ 0.001 & 0.101 & 0.021 \\ 0.001 & 0.021 & 0.052 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.001 & 0.001 \\ 0.001 & 0.098 & 0.024 \\ 0.001 & 0.024 & 0.041 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.001 & 0.001 \\ 0.001 & 0.231 & 0.021 \\ 0.001 & 0.021 & 0.098 \end{pmatrix}$
1-norm of variance of $\hat{\Sigma}_{MVE}$	0.208	**0.191**	0.373

Table 6: The biases and variances of the estimators $\hat{\Sigma}_Q$, $\hat{\Sigma}_{MLE}$ and $\hat{\Sigma}_{MVE}$, Σ given in Equation (55).

This turns out to be a highly robust estimator of autocovariance. As shown at the end of Section 4.5.1, it has a temporal breakdown point of 25%, which is the highest possible value in the autocovariance case. Note that the highly robust autocovariance estimator $\hat{\gamma}_Q(h, \mathbf{x})$ can also be carried out with $O(n \log n)$ time and $O(n)$ storage.

Another approach to obtain a robust estimator for the autocovariance is by truncating large terms in the sum of Equation (29). However, we prefer the scale approach suggested by Equation (30), because it allows the use of the highly robust estimator of scale Q_n, which has a remarkably high asymptotic Gaussian efficiency of 82.27%. For instance, Q_n has already been successfully used in the context of regression ([16, 55]), as well as for variogram estimation ([41]) in spatial statistics.

4.5.1 Temporal Breakdown Point

Outliers in time series can seriously affect the estimation and inference of parameters ([12, 80]). The main problem is that estimators which take account of the time series structure are not invariant under permutation of the data, as in the case of estimators for i.i.d. observations. Consequently, distinction between outliers occuring in isolation, in patches, or periodicly, becomes important. Three types of outliers are generally considered ([24]): innovation outliers (IO), which affect all subsequent observations, and additive outliers (AO) or replacement outliers (RO), which have no effect on subsequent observations. Consider a second-order stationary ARMA(p, q) process $\{X_t : t \in \mathbb{Z}\}$ such that for every t:

$$X_t - \rho_1 X_{t-1} - \cdots - \rho_p X_{t-p} = Z_t + \theta_1 Z_{t-1}, + \cdots + \theta_q Z_{t-q} \tag{57}$$

where $\rho_1, \ldots, \rho_p$ and $\theta_1, \ldots, \theta_q$ are real parameters, and the innovations are white noise $\{Z_t\} \sim WN(0, \sigma^2)$. Subsequently, we assume that the parameters of the ARMA process are defined such that the process is causal and invertible. More details on these notions, as well as necessary and sufficient conditions for causality and invertibility are given by Brockwell and Davis [9].

The ARMA(p, q) process $\{X_t : t \in \mathbb{Z}\}$ is said to have innovation outliers (IO) if it satisfies Equation (57), but the innovations $\{Z_t\}$ have a heavy-tailed distribution, for instance $F_\varepsilon = (1 - \varepsilon)F + \varepsilon H$, where ε is small and H is an arbitrary distribution with greater dispersion than F. The important characteristic of this kind of outliers is that even when the Z_t have outliers, Equation (57) is satisfied and therefore $\{X_t : t \in \mathbb{Z}\}$ is a perfectly observed ARMA(p, q) process. Robust estimators, like M-estimators, can typically cope with IO ([12]).

The process $\{X_t : t \in \mathbb{Z}\}$ is said to have additive outliers (AO) if it is not itself an ARMA(p, q) process, but rather defined by $X_t = V_t + B_t W_t$, where V_t is an ARMA(p, q) process satisfying Equation (57), B_t is a Bernoulli process with $P(B_t = 1) = \varepsilon$, $P(B_t =$

$0) = 1 - \varepsilon$, and W_t is an independent sequence of variables, independent of the sequences V_t and B_t. Therefore, the ARMA(p,q) process V_t is observed with probability $1 - \varepsilon$, whereas the ARMA(p,q) process V_t plus an error W_t is observed with probability ε. AO are known to be much more dangerous than IO. Note also that additive outliers have the same effect as replacement outliers (RO), where $X_t = (1 - B_t)V_t + B_t W_t$. This means that the ARMA(p,q) process V_t is observed with probability $1 - \varepsilon$, and replaced by an error W_t with probability ε. In the sequel, we consider RO.

In time series, one is much more interested in the breakdown point related to the initial data, which are located in time. Therefore, the classical definition loses its meaning because the time location of the outlier becomes important. In fact, the effect of the perturbation of a point located close to the boundary of the time domain can be quite different from one located in the middle of the time domain, and the effect depends notably on the time lag distance h. Therefore, we introduce the following definition of a temporal sample breakdown point of an autocovariance estimator based on Equation (30).

Definition 4.4 *Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ be a sample of size n and $\tilde{\mathbf{x}}$ is obtained by replacing any m observations of $\mathbf{x}$ by arbitrary values. Denote by I_m a subset of size m of $\{1, \ldots, n\}$. The temporal sample breakdown point of an autocovariance estimator $\hat{\gamma}(h, \mathbf{x})$ is:*

$$\varepsilon_n^t(\hat{\gamma}(h,\mathbf{x})) = \max\Big\{\frac{m}{n} : \sup_{I_m}\sup_{\tilde{\mathbf{x}}} S_{n-h}(\tilde{\mathbf{u}} + \tilde{\mathbf{v}}) < \infty \, and \, \inf_{I_m}\inf_{\tilde{\mathbf{x}}} S_{n-h}(\tilde{\mathbf{u}} + \tilde{\mathbf{v}}) > 0$$
$$and \; \sup_{I_m}\sup_{\tilde{\mathbf{x}}} S_{n-h}(\tilde{\mathbf{u}} - \tilde{\mathbf{v}}) < \infty \, and \, \inf_{I_m}\inf_{\tilde{\mathbf{x}}} S_{n-h}(\tilde{\mathbf{u}} - \tilde{\mathbf{v}}) > 0\Big\},$$

where $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are derived from $\tilde{\mathbf{x}}$ (t is used to emphasize temporal).

Note that in opposition to Definition 4.2, the configuration (i.e. the temporal location) of the perturbation is now taken into account, by adding the supremum and infimum on I_m. This definition is justified by the fact that an autocovariance estimator can be destroyed by a single configuration of perturbation, indexed in I_m. Therefore, it is quite possible to find other configurations, with more than $\varepsilon_n^t(\hat{\gamma}(h, \mathbf{x}))$ of perturbations, which do not demolish the estimator. Notice furthermore that this definition is local, in the sense that it is valid for a fixed h.

Consider a fixed temporal lag distance $h \in \mathbb{R}$. For $m = 1$ perturbed data point, it follows that, if $h < \frac{n}{2}$, one perturbation at time i, with $h < i \leq n - h$, generates the perturbation of two sums $\mathbf{u} + \mathbf{v}$ and two differences $\mathbf{u} - \mathbf{v}$, whereas for $0 < i \leq h$ or $n - h < i \leq n$, a single sum (difference) is perturbed. Finally, if $h \geq \frac{n}{2}$, one perturbation at time i, with $0 < i \leq n - h$ or $h < i \leq n$, affects one sum (difference), and none in the other cases. Therefore, to one perturbed observation corresponds at most two perturbed sums (differences). For general $m \geq 1$, we are interested in finding the most unfavorable configuration of perturbed data for a fixed h. Such a configuration is shown in Figure 19 for the case $h = 3$, $m = 7$ and $n = 21$. White points represent unperturbed observations, whereas black points represent perturbed observations. There are m black points. Construction of this configuration consists in placing h unperturbed observations, followed by h perturbed observations, followed by h unperturbed observations, and so on until exhaustion of the m black points. This configuration ensures that the most possible

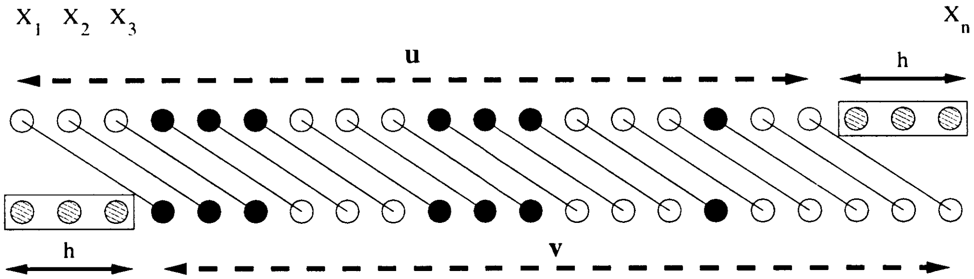


Figure 19: The most unfavorable configuration of perturbation for the case $h = 3$, $m = 7$ and $n = 21$. White points represent unperturbed observations, whereas black points represent perturbed observations.

sums (differences) are perturbed (i.e. each black point perturbes two sums (differences)). Moreover, perturbations do not overlap for a given lag distance h, which means that no sum (difference) between two perturbed observations is ever taken. Let $v_{max}(h, m, n)$ be the maximal number of perturbed sums (differences) for given h, m and n. This function depends on the relation between m and h. Let p and q be the two non-negative integers such that $m = ph + q$ and $q < h$. By disjunction of cases, it is then possible to compute the function $v_{max}(h, m, n)$ explicitly:

$$v_{max}(h,m,n) = \begin{cases} n-h & \text{if } m = \frac{n}{2}, \\ & \text{or } \frac{n}{2} > m \geq h,\ q = 0,\ n - 2m < h, \\ & \text{or } \frac{n}{2} > m \geq h,\ q \geq 1,\ h + q > n - 2ph \geq 0, \\ & \text{or } m < h,\ m + 2h > n,\ m \geq n - h, \\ 2m & \text{if } \frac{n}{2} > m \geq h,\ q = 0,\ n - 2m \geq h, \\ & \text{or } \frac{n}{2} > m \geq h,\ q \geq 1,\ n - 2ph \geq 2h + q, \\ & \text{or } m < h,\ m + 2h \leq n,\ n - 2m < h, \\ n - 2h + q & \text{if } \frac{n}{2} > m \geq h,\ q \geq 1,\ 2h + q > n - 2ph \geq 2h, \\ 2ph + q & \text{if } \frac{n}{2} > m \geq h,\ q \geq 1,\ 2h > n - 2ph \geq h + q, \\ m + n - 2h & \text{if } m < h,\ m + 2h > n,\ m < n - h,\ h < \frac{n}{2}, \\ m & \text{if } m < h,\ m + 2h > n,\ m < n - h,\ h \geq \frac{n}{2}. \end{cases}$$

Notice that the case $m > \frac{n}{2}$ makes no sense because it implies that more than half of the differences are perturbed. No equivariant scale estimator can be that resistant ([58]). Proposition 4.10 examines the relation between the classical sample breakdown point (usually known) and the temporal one.

Proposition 4.10 *For each $h \in \{0, \ldots, n-1\}$ and for each integer $M = n\varepsilon_n^*(\hat{\gamma}(h, \mathbf{x})) \leq \frac{n}{2}$, the sample breakdown point and the temporal sample breakdown point of an autocovariance estimator $\hat{\gamma}(h, \mathbf{x})$ satisfy the double inequality*

$$2\,\varepsilon_n^t(\hat{\gamma}(h,\mathbf{x})) \leq \varepsilon_n^*(\hat{\gamma}(h,\mathbf{x})) \leq \frac{2n}{n-h}\,\varepsilon_n^t(\hat{\gamma}(h,\mathbf{x})).$$

The first equality holds if and only if $h = \frac{n}{2}$ or $M = \frac{n}{2}$, and the second equality holds if *and only if* $v_{max}(h, M, n) = 2M$.

Remark 4.2 *By writing the inequality in Proposition 4.10 slightly differently, we can bound the temporal sample breakdown point with the classical sample breakdown point:*

$$\frac{n-h}{2n}\,\varepsilon_n^*(\hat{\gamma}(h,\mathbf{x})) \le \varepsilon_n^t(\hat{\gamma}(h,\mathbf{x})) \le \frac{1}{2}\,\varepsilon_n^*(\hat{\gamma}(h,\mathbf{x})).$$

Proof:

In order to prove the first inequality, consider the function

$$\delta(h,m,n) = \frac{v_{max}(h,m,n)}{n-h} - 2\,\frac{m}{n}.$$

We have to show that the function δ is non negative for all possible integers m.
If $v_{max}(h,m,n) = n-h$, then $\delta(h,m,n) = 1 - \frac{2m}{n} \ge 0$ (because $\frac{n}{2} > m$).
If $v_{max}(h,m,n) = 2m$, then $\delta(h,m,n) = \frac{2m}{n-h} - \frac{2m}{n} \ge 0$ (because $n-h < n$).
If $v_{max}(h,m,n) = n - 2h + q$, then

$$\begin{aligned}
\delta(h,m,n) &= \frac{n-2h+q}{n-h} - \frac{2m}{n} = \frac{n^2-(p+2)hn+m(2h-n)}{n(n-h)} \\
&\ge \frac{n^2-(p+2)hn+h(p+1)(2h-n)}{n(n-h)} \\
&\quad \text{(because} \quad m < h(p+1) \quad \text{and} \quad 2h-n<0) \\
&= \frac{n-2(p+1)h}{n} \ge 0 \quad \text{(because} \quad n-2ph \ge 2h).
\end{aligned} \tag{58}$$

If $v_{max}(h,m,n) = 2ph + q$, then

$$\begin{aligned}
\delta(h,m,n) &= \frac{2ph+q}{n-h} - \frac{2m}{n} = \frac{2mh-nq}{n(n-h)} \\
&\ge \frac{2mh-2hq(p+1)}{n(n-h)} \quad \text{(because} \quad n-2h(p+1)<0) \\
&= \frac{2hp(h-q)}{n(n-h)} \ge 0 \quad \text{(because} \quad h>q).
\end{aligned}$$

If $v_{max}(h,m,n) = m+n-2h$, then

$$\begin{aligned}\delta(h,m,n) &= \frac{m+n-2h}{n-h} - \frac{2m}{n} = \frac{m(2h-n)+n(n-2h)}{n(n-h)}\\ &\geq \frac{-h(n-2h)+n(n-2h)}{n(n-h)} \quad \text{(because} \quad m<h)\\ &= \frac{n-2h}{n} \geq 0 \quad \text{(because} \quad h<\frac{n}{2}).\end{aligned}$$

If $v_{max}(h,m,n) = m$, then

$$\delta(h,m,n) = \frac{m}{n-h} - \frac{2m}{n} = \frac{m(2h-n)}{n(n-h)} \geq 0 \quad \text{(because} \quad h<\frac{n}{2}).$$

Finally, if $h = \frac{n}{2}$ or $m = \frac{n}{2}$, then $\delta(h,m,n) = 0$ and therefore equality is reached. The second inequality follows from the fact that a perturbation on a single observation generates the perturbation of at most two sums (differences). Thus, the perturbation of m observations generates the perturbation of at most $2m$ sums (differences), i.e. $v_{max}(h,m,n) \leq 2m$. Consequently, we have the inequality

$$\varepsilon_n^*(\hat{\gamma}(h,\mathbf{x})) = \frac{v_{max}(h,M,n)}{n-h} \leq \frac{2M}{n-h} = \frac{2n}{n-h}\frac{M}{n} = \frac{2n}{n-h}\varepsilon_n^t(\hat{\gamma}(h,\mathbf{x})),$$

with equality if and only if $v_{max}(h,M,n) = 2M$. □

The classical sample autocovariance function is based on the classical scale estimator (standard deviation) whose sample breakdown point is zero. Therefore, by Remark 4.2, the temporal sample breakdown point of this estimator is also zero, for every lag h. This means that a single outlier in the data can destroy it. Figure 19 shows the temporal sample breakdown point $\varepsilon_{100}^t(\hat{\gamma}_Q(h,\mathbf{x}))$ of the highly robust autocovariance estimator, for each lag distance h, represented by the black curve. The upper and lower bounds given in Remark 4.2 are represented by the light grey curves. As it was stated, the temporal sample breakdown point equals its lower bound as long as $v_{max}(h,M,n) = 2M$, and equals its upper bound if $h = \frac{n}{2}$. The interpretation of Figure 20 is as follows.

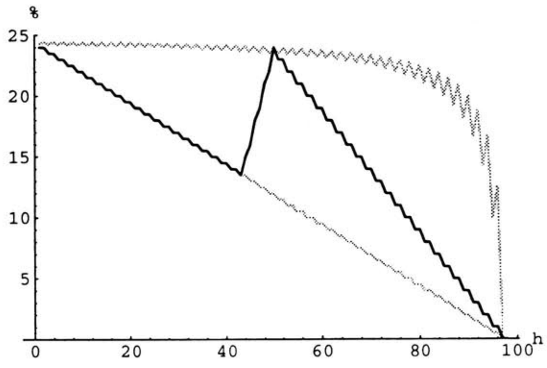


Figure 20: The temporal sample breakdown point (in black) as a function of the temporal lag distance h, for the highly robust sample autocovariance estimator $\hat{\gamma}_Q(h, \mathbf{x})$. The upper and lower bounds are drawn in light grey.

For a fixed h, if the percentage of perturbed observations is below the black curve, the estimator is never destroyed. If the percentage is above the black curve, there exists at least one configuration which destroys the estimator. This implies that the highly robust autocovariance estimator is more resistant at small time lags h or around $h = n/2$, than at large time lags h or before $h = n/2$, according to Figure 19. Note that from Remark 4.2, asymptotically, the temporal breakdown point of the autocovariance estimator is half the classical breakdown point.

Recall that Q_n has classical asymptotic breakdown point 50%, which means we can contaminate half of the observations yet still get reasonable estimate. Therefore, the highly robust autocovariance estimator has breakdown point 25%. This is because in forming $\mathbf{u}$ and $\mathbf{v}$ from the observation $\mathbf{x}$, most data will appear twice and hence in the worst case, the number of pairs $(\mathbf{u}_i, \mathbf{v}_i)$ that contain outliers will be twice the number of original outliers in $\mathbf{x}$. Note that this is the highest possible breakdown point for an autocovariance estimator. We can give up such high breakdown point by choosing different quantile from 1/4 in Equation (31), with the benefit of higher efficiency ([86]).

For example, if we choose the 0.91 quantile, we will reach the highest efficiency ($\approx 99\%$) for Q_n estimator, hence reach the highest efficiency for our estimator too. For the 0.91 quantile, the classical breakdown point is approximately 4.6% for Q_n, and therefore the temporal breakdown point of $\hat{\gamma}_Q$ is 2.3%.

4.5.2 Influence Function

We notice that in the autocovariance case, we can choose $\alpha = \beta = 1$ in Equation (38). Under a bivariate Gaussian distribution $\mathbf{F}$, the influence function of the γ_Q autocovariance estimator is:

$$IF\big((u,v);\gamma_Q,\mathbf{F}\big) = \frac{1}{2}\left[\sigma_+^2 IF\Big(\frac{u+v}{\sigma_+},Q,\Phi\Big) - \sigma_-^2 IF\Big(\frac{u-v}{\sigma_-},Q,\Phi\Big)\right], \tag{59}$$

where the influence function of Q_n at Φ ([87]) is:

$$IF(x,Q,\Phi) = c\frac{1/4 - \Phi(x+1/c) + \Phi(x-1/c)}{\int \phi(y+1/c)\phi(y)dy}, \tag{60}$$

with $c = 2.2191$. Figure 21 shows the plot of the influence function of γ_Q when the covariance is zero. The cases of non-zero covariance yield similar graphs. Note that the influence function of γ_Q is bounded between $\pm[(a-b)(\sigma_U^2+\sigma_V^2)+2|a+b|\sigma_U\sigma_V]/2$, where $a = \max IF(x,Q,\Phi), b = \min IF(x,Q,\Phi)$. The bounds can be computed by writing $\sigma_\pm^2$ as $\sigma_U^2 + \sigma_V^2 \pm 2\mathrm{Cov}(U,V)$ and noticing that $\mathrm{Cov}(U,V)$ is bounded between $\pm\sigma_U\sigma_V$. To the contrary, the influence function of γ_M is proportional to uv when the covariance is zero, and therefore unbounded.

4.5.3 Asymptotic Variance

Under regularity conditions, $\hat{\gamma}_Q$ is consistent, i.e. $\hat{\gamma}_Q \longrightarrow \gamma_Q$ in probability as $n \to \infty$, since Q_n is consistent ([87]). Moreover, $\sqrt{n}(\hat{\gamma}_Q - \gamma_Q)$ is asymptotically normal with zero

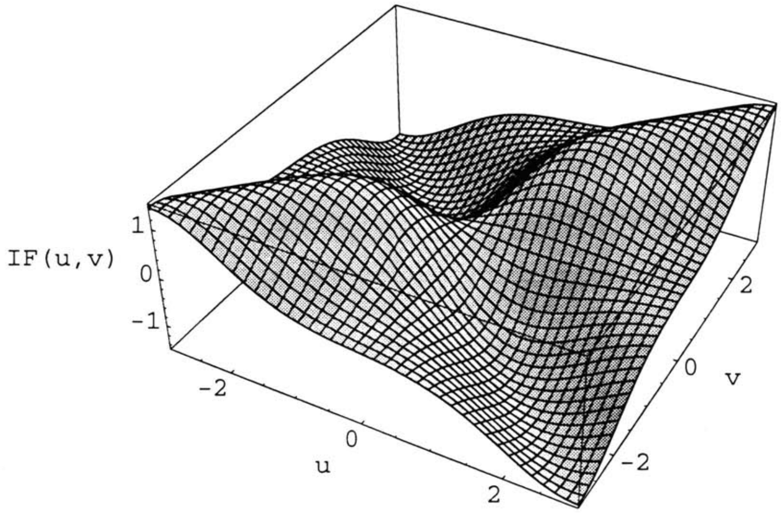


Figure 21: The influence function $IF((u,v),\gamma_Q,F)$, where U and V are independent and have identical standard Gaussian distribution.

expectation and variance given by Portnoy and Genton [82, 83, 40]

$$\begin{aligned}\mathrm{Var}(\gamma_Q,F) &= \iint |IF((u,v),\gamma_Q,F)|^2 dF(u,v)\\ +2\sum_{k=1}^{\infty}\int\dots\int & IF((u_1,v_1),\gamma_Q,F)IF((u_{1+k},v_{1+k}),\gamma_Q,F)dF((u_1,v_1),(u_{1+k},v_{1+k})).\end{aligned} \tag{61}$$

Regularity conditions for consistency and asymptotic normality are given by Huber [56] for the independent case and by Portnoy and Bustos [82, 83, 11] for the dependent case. In the latter situation, mixing conditions like α-mixing or ϕ-mixing are sufficient ([4, 32]). Note that Equation (61) is valid for any consistent estimator. In particular, for the classical autocovariance estimator γ_M, the equation is equivalent to Bartlett's formula [9, page 222].

The computation of the variance turns out to be tedious and very often impossible to get a closed form. For the classical autocovariance estimator on MA(1) model and AR(1) model, we have an explicit formula for the variance given in Proposition 4.11. Here, MA(1) and AR(1) are special cases of the ARMA(p, q) model, i.e. MA(1)=ARMA(0,1) and AR(1)=ARMA(1,0).

Proposition 4.11 *For an MA(1) model with variance matrix*

$$\Sigma_\theta = \begin{bmatrix} 1+\theta^2 & \theta & & & \\ \theta & & \ddots & & \\ & & \ddots & & \\ & & \ddots & & \theta \\ & & & \theta & 1+\theta^2 \end{bmatrix},$$

the asymptotic variance of the classical autocovariance estimator of the (i,i), $(i, i\pm 1)$ *and* $(i, i\pm h)$ *(for* $h \geq 2$*) entries of* Σ_θ *are respectively*

$$\begin{aligned} Var(\gamma_M(0), MA(1)) &= 2(1+\theta^2)(1+2\theta^2) \\ Var(\gamma_M(1), MA(1)) &= (1+\theta^2)^2 + 3\theta^2 \\ Var(\gamma_M(h), MA(1)) &= (1+\theta^2)^2 + 2\theta^2. \end{aligned} \tag{62}$$

For an AR(1) model with variance matrix

$$\Sigma_\rho = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & \rho & & \dots & \rho^{n-1} \\ \rho & & \ddots & & \vdots \\ & & \ddots & & \\ \vdots & & \ddots & & \rho \\ \rho^{n-1} & \dots & & \rho & 1 \end{bmatrix}$$

the asymptotic variance of the classical autocovariance estimator of the (i,i) *and* $(i, i\pm h)$

for $(h \geq 1)$ entries of Σ_ρ are respectively

$$\begin{aligned} Var(\gamma_M(0), AR(1)) &= \frac{2+2\rho^2}{(1-\rho^2)^3} \\ Var(\gamma_M(h), AR(1)) &= \frac{1+2\rho^2+(1+2h)\rho^{2h}-(2h-1)\rho^{2h+2}}{(1-\rho^2)^3} \end{aligned} \tag{63}$$

Proof:

The calculation of these asymptotic variances are of the same style. For simplicity, we explain the first equality only. We use the fact that $IF(x, S, \Phi) = \frac{1}{2}(x^2-1)$ and $\int |IF(x, S, \Phi)|^2 d\Phi(x) = 1/2$, where S is the maximum likelihood estimator of scale, i.e. the standard deviation. Using Equation (61), we get

$$\begin{aligned} &\mathrm{Var}(\gamma_M(0), \mathrm{MA}(1)) \\ =&4(1+\theta^2)^2\frac{1}{2} + 2\int\int 4(1+\theta^2)^2\frac{1}{2}((\frac{x}{\sqrt{1+\theta^2}})^2-1)\frac{1}{2}(1+\theta^2)((\frac{y}{\sqrt{1+\theta^2}})^2-1)dF(x,y) \\ =&2(1+\theta^2)^2 + 2(1+\theta^2)\int\int(x^2-(1+\theta^2))(y^2-(1+\theta^2))dF(x,y) \\ =&2(1+\theta^2)^2 + 2(1+\theta^2)((1+\theta^2)^2+2\theta^2-(1+\theta^2)^2) \\ =&2(1+\theta^2)(1+2\theta^2) \end{aligned}$$

□

In order to calculate the efficiency, we need to calculate the Fisher information of the estimator, and this is also computationally difficult. For the AR(1) model, we get the Fisher information $(n-2)/(1-\rho^2) + (1+\rho^2)((1-\rho^2)^2)$. So in this particular case, we can calculate the efficiency symbolically since efficiency is just the inverse of the product of Fisher information and the variance of the estimator.

4.5.4 Simulations

In this section, we present some simulations in order to compare the $\hat{\gamma}_M$ and $\hat{\gamma}_Q$ autocovariance estimator on MA(1) and AR(1) models, with and without replacement outliers. We start with a brief description of the experiment.

The standard Gaussian AR(1) and MA(1) models $\{V_t\}$ are considered, with or without replacement outliers (RO) defined by $X_t = (1 - B_t)V_t + B_t W_t$. The Bernoulli process satisfies $P(B_t = 1) = \varepsilon$ and $P(B_t = 0) = 1 - \varepsilon$, with $\varepsilon = 0$ and $\varepsilon = 10\%$. The distribution of W_t is chosen to be $N(0, \tau^2)$, where $\tau^2 = k^2 \text{Var}(V_t)$ with $k = 3$ and $k = 10$. We generate 1000 samples of sizes 20, 50 and 100 for each model with parameters θ (respectively ρ) equal to 0 and 0.5. The mean of $\hat{\gamma}_M$ and $\hat{\gamma}_Q$ are computed over the 1000 replications, as well as the relative efficiency (REF) of $\hat{\gamma}_Q$ to $\hat{\gamma}_M$. We built an `S-Plus` function to compute $\hat{\gamma}_Q$, which is available on the Web. The results are presented in Table 7.

From the simulation, we can see that when there is no outliers, both estimators yield a mean that is close to the true autocovariance, i.e. unbiased. The REF is around 80% for large n. This is considered high for a highly robust autocovariance estimator. In the presence of outliers, the classical autocovariance estimator shows a weak resistance in terms of the mean value, and it also has smaller efficiency than the robust estimator. This is particularly clear when the outliers are large ($k = 10$). One can also check that the asymptotic variances of $\hat{\gamma}_M$ given in Equations (62) and (63) agrees with the ones find in the simulations. Moreover, for the MA(1) model with $\theta = 0$, i.e. the i.i.d. case, the asymptotic variance of $\hat{\gamma}_Q$ can be computed numerically from Equation (61), which yields 2.482 (to be compared with 2 for $\hat{\gamma}_M$). This yields an asymptotic relative efficiency of 80.6%, which is close to the one find by simulation in Table 7.

Note that the classical estimator we took is not modified to ensure positive definiteness of the covariance matrix. If we use the modified version (divided by n instead of $n - h$), then we should also ensure positive definiteness for the highly robust autocovariance estimator. This can be done by the shrinking, the eigenvalue or the scaling method

([89]). A typical application is then to use the highly robust autocovariance estimator in the Yule-Walker equations ([9]) in order to estimate the parameters of an AR model robustly.

4.5.5 Example

We carry out the classical autocovariance estimator $\hat{\gamma}_M$ and the highly robust autocovariance estimator $\hat{\gamma}_Q$ on 91 monthly interest rates of an Austrian bank (see Figure 22 for the data). This data set has already been analyzed by Kunsch [70, 71]. He pointed out the presence of three outliers for the months number 18, 28, 29. In Figure 23, we run $\hat{\gamma}_M$ and $\hat{\gamma}_Q$ on the original data in (a) and (b). Then we replace the three outliers by 9.85 as suggested by Künsch in (c) and (d). Looking at (c) and (d), we can see that the new estimator $\hat{\gamma}_Q$ behaves similarly to $\hat{\gamma}_M$ when no outliers are present. Comparing the difference between (a) and (c) with the difference between (b) and (d), we can see that $\hat{\gamma}_Q$ has better resistance to the outliers than $\hat{\gamma}_M$. This effect is particularly visible for small time lags.

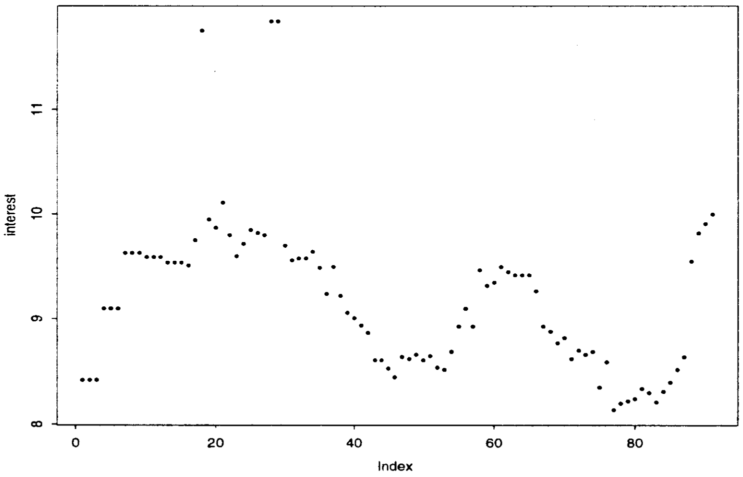


Figure 22: Monthly interest rates of an Austrian bank during 91 months.

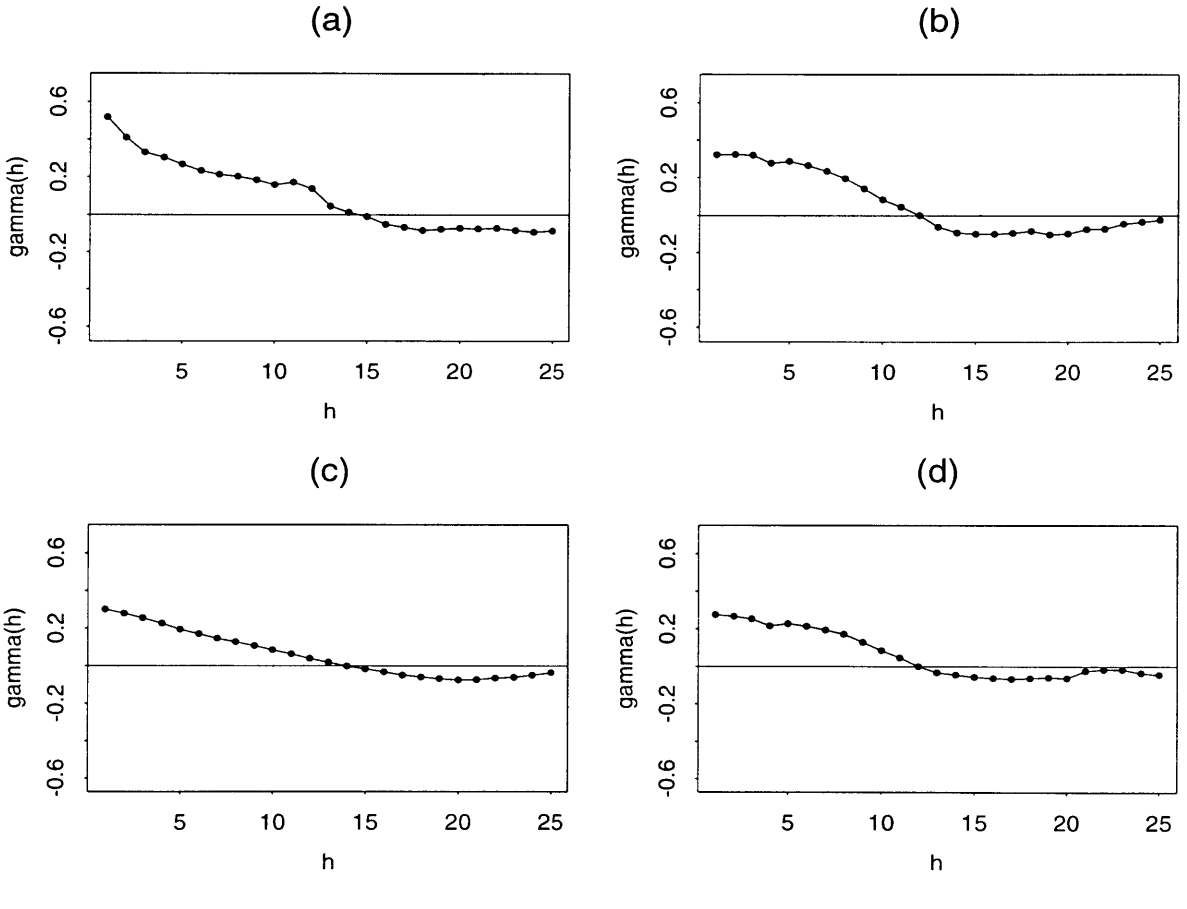


Figure 23: Autocovariance estimator on monthly interest rates of an Austrian bank: (a) classical $\hat{\gamma}_M$ on original data (b) highly robust $\hat{\gamma}_Q$ on original data (c) classical $\hat{\gamma}_M$ on corrected data (d) highly robust $\hat{\gamma}_Q$ on corrected data

Table 7: The mean $\mathrm{m}(\hat{\gamma}_M)$, $\mathrm{m}(\hat{\gamma}_Q)$ and the relative efficiency REF of the autocovariance estimators at time lag $h = 0$, $h = 1$ and $h = 2$ on MA(1) model with $\theta = 0$, MA(1) model with $\theta = 0.5$ and AR(1) model with $\theta = 0.5$ respectively, with and without RO outliers.

			$\varepsilon = 0$			$\varepsilon = 10\%$, $k = 3$			$\varepsilon = 10\%$, $k = 10$		
h	$\gamma(h)$	n	$\mathrm{m}(\hat{\gamma}_M)$	$\mathrm{m}(\hat{\gamma}_Q)$	REF	$\mathrm{m}(\hat{\gamma}_M)$	$\mathrm{m}(\hat{\gamma}_Q)$	REF	$\mathrm{m}(\hat{\gamma}_M)$	$\mathrm{m}(\hat{\gamma}_Q)$	REF
MA(1), $\theta = 0$											
0	1	50	0.983	1.012	0.721	1.784	1.290	4.015	10.842	1.514	262.2
		100	0.983	1.000	0.762	1.766	1.268	4.035	10.612	1.467	277.0
1	0	50	-0.020	-0.027	0.712	-0.030	-0.026	1.006	-0.231	-0.048	12.9
		100	-0.005	-0.010	0.728	-0.018	-0.028	1.100	-0.093	-0.044	14.9
2	0	50	-0.019	-0.019	0.680	-0.049	-0.061	1.039	-0.271	-0.100	14.6
		100	-0.018	-0.020	0.741	-0.023	-0.021	1.043	-0.142	-0.034	13.3
MA(1), $\theta = 0.5$											
0	1.25	50	1.214	1.254	0.754	2.156	1.589	3.135	13.075	1.855	225.0
		100	1.228	1.244	0.801	2.227	1.575	3.787	13.798	1.838	226.1
1	0.5	50	0.464	0.481	0.704	0.338	0.515	1.025	0.106	0.882	10.2
		100	0.481	0.487	0.800	0.380	0.535	1.061	0.300	0.888	10.8
2	0	50	-0.046	-0.049	0.711	-0.068	-0.087	0.930	-0.331	-0.115	12.9
		100	-0.016	-0.018	0.809	-0.027	-0.037	1.174	-0.147	-0.041	19.1
AR(1), $\theta = 0.5$											
0	1.33	50	1.252	1.293	0.772	2.292	1.647	3.5	14.433	1.933	199.0
		100	1.291	1.313	0.803	2.378	1.669	3.3	14.803	1.940	235.1
1	0.67	50	0.586	0.609	0.749	0.434	0.668	1.0	0.271	1.126	7.2
		100	0.628	0.638	0.804	0.501	0.724	1.0	0.404	1.159	7.6
2	0.33	50	0.246	0.254	0.763	0.180	0.258	1.0	0.082	0.497	9.6
		100	0.291	0.296	0.778	0.227	0.334	1.0	0.093	0.543	12.0

References

[1] V. Arnold. On matrices depending on parameters. *Russian Math. Surveys*, 26:29–43, 1971.

[2] H. Baumgärtel. *Analytic Perturbation Theory for Matrices and Operators.* Birkhäuser, Basel, 1985.

[3] T. Beelen and P. Van Dooren. An improved algorithm for the computation of Kronecker's canonical form of a singular pencil. *Lin. Alg. Appl.*, 105:9–65, 1988.

[4] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

[5] D. Boley. Estimating the sensitivity of the algebraic structure of pencils with simple eigenvalue estimates. *SIAM J. Matrix Anal. Appl.*, 11(4):632–643, 1990.

[6] D. Boley. The algebraic structure of pencils and block Toeplitz matrices. *Lin. Alg. Appl.*, 279(1-3):255–279, 1998.

[7] D. Boley and P. Van Dooren. Placing zeroes and the Kronecker canonical form. *Circuits Systems Signal Process*, 13(6):783–802, 1994.

[8] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden Day, 1976.

[9] P. J. Brockwell and R. A. Davis. *Times Series: Theory and Methods.* Springer, New York, 1991.

[10] J. V. Burke and M. L. Overton. Stable Perturbations of Nonsymmetric Matrices. *Linear Algebra Appl.*, pages 249–273, 1992.

[11] O. H. Bustos. General M-estimates for contaminated p-th order autoregressive processes: Consistency and asymptotic normality. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 59:491–504, 1982.

[12] O. H. Bustos and V. J. Yohai. Robust estimates for ARMA models. *J. Am. Stat. Assoc.*, 81:155–168, 1986.

[13] F. Chaitin-Chatelin and V. Frayssé. *Lectures on Finite Precision Computations.* SIAM, Philadelphia, 1996.

[14] F. Chaitin-Chatelin and V. Frayssé. *Lectures on Finite Precision Computations.* SIAM, Philadelphia, 1996.

[15] C. Croux and P. J. Rousseeuw. Time-efficient algorithms for two highly robust estimators of scale. *Computational Statistics*, 2:411–428, 1992.

[16] C. Croux, P. J. Rousseeuw, and O. Hössjer. Generalized S-estimators. *J. Am. Stat. Assoc.*, 89:1271–1281, 1994.

[17] P. L. Davies. Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Annal. of Stat.*, 15:1269–1292, 1987.

[18] J. Demmel and A. Edelman. The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms. *Lin. Alg. Appl.*, 230:61–87, 1995.

[19] J. Demmel and B. Kågström. Stably computing the Kronecker structure and reducing subspace of singular pencils $A - \lambda B$ for uncertain data. In J. Cullum and R. Willoughby, editors, *Large Scale Eigenvalue Problems*, volume 127 of *North-Holland Mathematics Studies*, pages 283–323, 1986.

[20] J. Demmel and B. Kågström. Computing stable eigendecompositions of matrix pencils. *Lin. Alg. Appl.*, 88/89:139–186, 1987.

[21] J. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: robust software with error bounds and applications. Part I: theory and algorithms. *ACM Trans. Math. Softw.*, 19(2):160–174, 1993.

[22] J. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: robust software with error bounds and applications. Part II: software and applications. *ACM Trans. Math. Softw.*, 19(2):175–201, 1993.

[23] J. Demmel and B. Kågström. Accurage solutions of ill-posed problems in control theory. *SIAM J. Matrix Anal. Appl.*, 9(1):126–145, 1988.

[24] J. Denby and R. D. Martin. Robust estimation of the first-order autoregressive parameter. *J. Am. Stat. Assoc.*, 74:140–146, 1979.

[25] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62:531–545, 1975.

[26] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Jour American Stat Asso*, 76:354–362, 1981.

[27] D. L. Donoho and P. J. Huber. The notion of breakdown point. In P. J. Bickel, K. A. Doksum, and J. L. Hodges Jr., editors, *A Festschrift for Erich L. Lehmann*, pages 157–184, 1983.

[28] P. Van Dooren. The computation of Kronecker's canonical form of a singular pencil. *Lin. Alg. Appl.*, 27:103–140, 1979.

[29] P. Van Dooren. The generalized eigenstructure problem in linear system theory. *IEEE Trans. Autom. Contr.*, 26(1):111–129, 1981.

[30] P. Van Dooren. Reducing subspaces: definitions, properties and algorithms. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, volume 973 of *Lecture Notes in Mathematics*, pages 58–73, Berlin, 1983. Springer-Verlag.

[31] P. Van Dooren. Oral communication. October 1996.

[32] P. Doukhan. *Mixing: Properties and Examples.* Springer, New York, 1994.

[33] A. Edelman, E. Elmroth, and B. Kågström. A Geometric Approach To Perturbation Theory of Matrices and Matrix Pencils: Part 1: Versal Deformations. *SIAM J. Mat. Anal Appl*, pages 653–692, 1997.

[34] A. Edelman, E. Elmroth, and B. Kågström. A geometric approach to perturbation theory of matrices and matrix pencils: Part 1: versal deformations. *SIAM J. Mat. Anal Appl*, 18:653–692, 1997.

[35] A. Edelman, E. Elmroth, and B. Kågström. A geometric approach to perturbation theory of matrices and matrix pencils: Part 2: stratification-enhanced staircase algorithm. *SIAM J. Mat. Anal Appl.*, 20(3):667–699, 1999.

[36] E. Elmroth and B. Kågström. The set of 2-by-3 matrix pencils — Kronecker structures and their transitions under perturbations. *SIAM J. Matrix Anal. Appl.*, 17(1):1–34, 1996.

[37] A. Emami-Naeini and P. Van Dooren. Computation of zeros of linear multivariable systems. *Automatica*, 18:415–430, 1982.

[38] T. F. Fairgrieve. The application of singularity theory to the computation of Jordan Canonical Form. *Master Thesis at Computer Science in Univ. of Toronto*, 1986.

[39] K. Fang and Y. Zhang. *Generalized Multivariate Analysis.* Springer, New York, 1990.

[40] M. G. Genton. Asymptotic variance of M-estimators for dependent Gaussian random variables. *Statistics and Probability Letters*, 38:255–261, 1998.

[41] M. G. Genton. Highly robust variogram estimation. *Mathematical Geology*, 30:213–221, 1998.

[42] M. G. Genton. Spatial Breakdown Point of Variogram Estimators. *Math. Geology*, 30:853–871, 1998.

[43] M. G. Genton and Y. Ma. Robustness properties of dispersion estimators. *Statistics & Probability Letters, to appear.*

[44] M. G. Genton and P. J. Rousseeuw. The change-of-variance function of M-estimators of scale under general contamination. *Journal of Computational and Applied Mathematics*, 64:69–80, 1995.

[45] R. Gnanadesikan. *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley, New York, 2 edition, 1997.

[46] R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multireponse data. *Biometrics*, 28:81–124, 1972.

[47] G. Golub and C. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore and London, 3 edition, 1996.

[48] G. Golub and J. Wilkinson. Ill-conditioned eigensystems and the computation of the Jordan canonical form. *SIAM Rev.*, 18(4):578–619, 1976.

[49] F. R. Hampel. A general qualitative definition of robustness. *Ann. Math. Stat.*, 42:1887–1896, 1971.

[50] F. R. Hampel. Robust estimation: a condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 27:87–104, 1973.

[51] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.

[52] F. R. Hampel. On the breakdown points of some rejection rules with mean. *Research Report 11, Fachgruppe für Statistik, ETHZ*, 1976.

[53] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics, the Approach based on Influence Functions.* Wiley, New York, 1986.

[54] S. Helgason. *Differential Geometry, Lie Groups, and Symmetric Spaces.* Academic Press, New York, San Francisco and London, 1978.

[55] O. Hössjer, C. Croux, and P. J. Rousseeuw. Asymptotics of generalized S-estimators. *J. Multivariate Analysis*, 51:148–177, 1994.

[56] P. J. Huber. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233, 1967.

[57] P. J. Huber. *Robust Statistics Procedures.* SIAM, Philadelphia, 1977.

[58] P. J. Huber. *Robust Statistics.* Wiley, New York, 1981.

[59] P. J. Huber. Finite sample breakdown of M- and P-estimators. *Ann. Math. Stat.*, 12:119–126, 1984.

[60] B. Kågström. The generalized singular value decomposition and the general $A - \lambda B$ problem. *BIT*, 24:568–583, 1984.

[61] B. Kågström. RGSVD - An algorithm for computing the Kronecker structure and reducing subspaces of singular $A - \lambda B$ pencils. *SIAM J. Sci. Stat. Comput.*, 7(1):185–211, 1986.

[62] B. Kågström and A. Ruhe. ALGORITHM 560: JNF, An algorithm for numerical computation of the Jordan normal form of a complex matrix [F2]. *ACM Trans. Math. Softw.*, 6(3):437–443, 1980.

[63] B. Kågström and A. Ruhe. An algorithm for numerical computation of the Jordan normal form of a complex matrix. *ACM Trans. Math. Softw.*, 6(3):398–419, 1980.

[64] B. Kågström and A. Ruhe, editors. *Matrix Pencils*, volume 973 of *Lecture Notes in Mathematics*, New York, 1982. Swedish Institute of Applied Mathematics, Springer. Pite Havsbad, 1982.

[65] Y. Kano. Consistency property of elliptical probability density functions. *Journal of Multivariate Analysis*, 51:139–147, 1994.

[66] T. Kato. *Perturbation Theory for Linear Operators.* Springer, Berlin, 1980.

[67] J. Kautsky, N. K. Nichols, and P. Van Dooren. Robust pole assignment in linear state feedback. *Int. J. Contr.*, 41(5):1129–1155, 1985.

[68] V. Kublanovskaya. On a method of solving the complete eigenvalue problem of a degenerate matrix. *USSR Comput. Math. Phys.*, 6(4):1–14, 1966.

[69] V. Kublanovskaya. AB-algorithm and its modifications for the spectral problem of linear pencils of matrices. *Numer. Math.*, 43:329–342, 1984.

[70] H. Künsch. Robust estimation for autoregressive processes. *Proc. Institute for Statist. Math.*, 31:51–64, 1983.

[71] H. Künsch. Infinitesimal robustness for autoregressive processes. *Ann. Stat.*, 12:843–863, 1984.

[72] G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association*, 80:759–766, 1985.

[73] R. Lippert and A. Edelman. Nonlinear eigenvalue problems. In Z. Bai, editor, *Templates for Eigenvalue Problems*, to appear.

[74] H. Lopuhaä. On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics*, 17:1662–1683, 1989.

[75] H. Lopuhaä and P. J. Rousseeuw. Breakdown point of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19:229–248, 1991.

[76] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, New York, San Francisco and London, 1979.

[77] R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67, 1976.

[78] R. A. Maronna, W. A. Stahel, and V. J. Yohai. Bias-robust estimation of multivariate scatter based on projections. *Journal of Multivariate Analysis*, 42:141–161, 1992.

[79] R. A. Maronna and V. J. Yohai. The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341, 1995.

[80] R. D. Martin and V. J. Yohai. Robustness in time series and estimating ARMA models. In E. J. Hannan, P. R. Krishnaiah, and M. M. Rao, editors, *Handbook of Statistics*, volume 5, pages 119–155, 1985.

[81] J. Moro, J. V. Burke, and M. L. Overton. On the Lidskii-Vishik-Lyusternik Perturbation Theory for Eigenvalues of Matrices with Arbitrary Jordan Structure. *SIAM J. Matrix Anal. Appl.*, pages 793–817, 1997.

[82] S. Portnoy. Robust estimation in dependent situations. *Ann. Stat.*, 5:22–43, 1977.

[83] S. Portnoy. Further remarks on robust estimation in dependent situations. *Ann. Stat.*, 7:224–231, 1979.

[84] L. Reichel and L. N. Trefethen. Eigenvalues and Pseudo-Eigenvalues of Toeplitz Matrices. *Linear Algebra Appl.*, pages 153–185, 1992.

[85] P. J. Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W.Wertz, editors, *Mathematical Statistics and Applications*, pages 283–297, Dodrecht, 1985. Reidel Publishing.

[86] P. J. Rousseeuw and C. Croux. Explicit scale estimators with high breakdown point. *L_1 Statistical Analyses and Related Methods*, pages 77–92, 1992.

[87] P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283, 1993.

[88] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[89] P. J. Rousseeuw and G. Molenberghs. Transformation of non positive semidefinite correlation matrices. *Commun. Statist.-Theory Meth.*, 22:965–984, 1993.

[90] P. J. Rousseeuw and B. C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–651, 1990.

[91] A. Ruhe. An algorithm for numerical determination of the structure of a general matrix. *BIT*, 10:196–216, 1970.

[92] W. A. Stahel. Breakdown of covariance estimators. *Res. Rep. achgruppe für Statistik ETH*, 31, 1981.

[93] D. E. Tyler. Finite sample breakdown points of projection based multivariate location and scatter statistics. *The Annals of Statistics*, 22:1024–1044, 1994.

[94] M. Wicks and R. DeCarlo. Computing the distance to an uncontrollable system. *IEEE Trans. Autom. Contr.*, 36(1):39–49, 1991.