Structured Robust Covariance Estimation

Ami Wiesel
School of Computer Science and Engineering
The Hebrew University of Jerusalem
amiw@cs.huji.ac.il

Teng Zhang
Department of Mathematics
University of Central Florida
Teng.Zhang@ucf.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

A. Wiesel and T. Zhang. *Structured Robust Covariance Estimation.* Foundations and Trends® in Signal Processing, vol. 8, no. 3, pp. 127–216, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-095-8

Foundations and Trends® in Signal Processing
Volume 8, Issue 3, 2014

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing

Information for Librarians

Foundations and Trends® in Signal Processing, 2014, Volume 8, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in Signal Processing
Vol. 8, No. 3 (2014) 127–216

DOI: 10.1561/2000000053

the essence of knowledge

Structured Robust Covariance Estimation

Ami Wiesel
School of Computer Science and Engineering
The Hebrew University of Jerusalem
amiw@cs.huji.ac.il

Teng Zhang
Department of Mathematics
University of Central Florida
Teng.Zhang@ucf.edu

Contents

Abstract

We consider robust covariance estimation with an emphasis on Tyler's M-estimator. This method provides accurate inference of an unknown covariance in non-standard settings, including heavy-tailed distributions and outlier contaminated scenarios. We begin with a survey of the estimator and its various derivations in the classical unconstrained settings. The latter rely on the theory of g-convex analysis which we briefly review.

Building on this background, we enhance robust covariance estimation via g-convex regularization, and allow accurate inference using a smaller number of samples. We consider shrinkage, diagonal loading, and prior knowledge in the form of symmetry and Kronecker structures. We introduce these concepts to the world of robust covariance estimation, and demonstrate how to exploit them in a computationally and statistically efficient manner.

A. Wiesel and T. Zhang. *Structured Robust Covariance Estimation*. Foundations and Trends® in Signal Processing, vol. 8, no. 3, pp. 127–216, 2014.

DOI: 10.1561/2000000053.

Notations and Acronyms

We summarize here the notation and acronyms used throughout the survey.

We denote vectors by boldface lowercase letters, e.g., $\mathbf{x} \in \mathbf{R}^n$, and matrices by boldface uppercase letters, e.g., $\mathbf{A} \in \mathbf{R}^{n,m}$. The identity matrix of appropriate dimension is written as $\mathbf{I}$. For a square matrix $\mathbf{A}$, $\text{Tr}\{\mathbf{A}\}$ is the trace, $|\mathbf{A}|$ is the determinant, $\mathbf{A} \succ \mathbf{0}$ ($\mathbf{A} \succeq \mathbf{0}$) means that $\mathbf{A}$ is symmetric and positive (nonnegative) definite, and $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B} \succeq \mathbf{0}$. We denote the ordered eigenvalues of $\mathbf{A} \in \mathbf{R}^{p,p}$ by $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_p(\mathbf{A})$. The standard Euclidean norm is denoted $\|\mathbf{x}\|$. The operator $\text{vec}(\mathbf{X})$ stacks the columns of the matrix $\mathbf{X}$ one over the other and outputs a vector. The Kronecker product is denoted by $\otimes$. We often denote the set of vectors $\{\mathbf{x}_i\}_{i=1}^n$ by $\mathcal{X}$. For a subspace $L \in \mathbf{R}^n$ of dimension $\dim(L)$, we denote the number of vectors in $\mathcal{X}$ lying on it by $N(L)$.

Following is a list of the most frequently used acronyms:

- LMMSE – Linear minimum mean squared error.
- MLE – Maximum Likelihood estimate.
- i.i.d. – independent and identically distributed.
- RMT – Random matrix theory.

- MM – Majorization minimization.
- FIM – Fisher Information matrix.
- CRB – Cramer Rao bound.

1

Preliminaries

In this chapter, we introduce the theory of g-convexity which will be used throughout the monograph. We begin with a brief review of related results from linear algebra. Next, we define the abstract theory of geodesic convexity over Riemannian manifolds. Finally, we particularize it to the case of positive definite matrices.

1.1 Positive definite matrices

The main object of interest in this monograph is the covariance matrix. Its most obvious properties are that it is symmetric and positive definite. Thus, we begin by reviewing these concepts.

Definition 1.1. A square matrix $\mathbf{Q}$ is positive definite, denoted by $\mathbf{Q} \succ \mathbf{0}$, if it is symmetric and satisfies

$$\mathbf{z}^T\mathbf{Q}\mathbf{z} > 0 \quad \forall \quad \mathbf{z} \neq \mathbf{0}. \tag{1.1}$$

Similarly, a matrix $\mathbf{Q}$ is positive semidefinite, denoted by $\mathbf{Q} \succeq \mathbf{0}$, if it is symmetric and satisfies

$$\mathbf{z}^T\mathbf{Q}\mathbf{z} \geq 0 \quad \forall \quad \mathbf{z}. \tag{1.2}$$

A few equivalent characterizations of positive definiteness are:

- A symmetric matrix $\mathbf{Q}$ is positive definite if and only if its eigenvalues are positive. Thus, it can be decomposed as

$$\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{1.3}$$

  where $\mathbf{U}$ is an orthogonal matrix, and $\mathbf{D}$ is a diagonal matrix with positive elements.

- A matrix $\mathbf{Q}$ is positive definite if and only if it has a real square root, i.e., it can be decomposed as

$$\mathbf{Q} = \mathbf{R}\mathbf{R}^T \tag{1.4}$$

  where $\mathbf{R}$ is a square invertible matrix. Two constructive choices for computing $\mathbf{R}$ are the Cholesky and eigenvalue decompositions.

In general, two symmetric matrices cannot always be simultaneously diagonalized. However, things simplify when they are positive definite.

Lemma 1.1 (Simultaneous diagonalization [40])**.** Let $\mathbf{Q}_0 \succ \mathbf{0}$ and $\mathbf{Q}_1 \succeq \mathbf{0}$ be two matrices. Then, there exist a joint diagonalization decomposition

$$\begin{aligned} \mathbf{Q}_0 &= \mathbf{V}\mathbf{V}^T \\ \mathbf{Q}_1 &= \mathbf{V}\mathbf{D}\mathbf{V}^T \end{aligned} \tag{1.5}$$

where $\mathbf{V}$ is square and invertible, and $\mathbf{D}$ is a diagonal matrix with non-negative elements.

Proof. Due to its positivity, we decompose $\mathbf{Q}_0$ as $\mathbf{Q}_0 = \mathbf{R}\mathbf{R}^T$. We define $\mathbf{Z} = \mathbf{R}^{-1}\mathbf{Q}_1(\mathbf{R}^{-1})^T$ and note that $\mathbf{Z}$ is positive semidefinite. We decompose $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonal matrix with non-negative elements. Finally, we define $\mathbf{V} = \mathbf{R}\mathbf{U}$ and obtain the required result. □

Another important result on positive definiteness addresses block partitioned matrices.

Lemma 1.2 (Schur's Complement [40]). Partition a symmetric matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \tag{1.6}$$

with $\mathbf{C} \succ 0$. Define Schur's complement as

$$\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T. \tag{1.7}$$

Then, $\mathbf{X} \succeq 0$ if and only if $\mathbf{S} \succeq 0$.

Proof. A matrix $\mathbf{X}$ is positive semidefinite if and only if $\mathbf{T}\mathbf{X}\mathbf{T}^T$ is positive semidefinite for an invertible matrix $\mathbf{T}$. If $\mathbf{C}$ is invertible then we have the following block Cholesky decomposition

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^T. \tag{1.8}$$

The matrix $\begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$ is invertible and a block diagonal matrix is positive semidefinite if and only if its blocks are positive semidefinite. □

Finally, in some derivations it is convenient to represent matrices using their vectorized version:

Definition 1.2. Let $\mathbf{A}$ be an $m \times n$ matrix. Then $\text{vec}(\mathbf{A})$ is a length mn vector with the columns of $\mathbf{A}$ stacked one over the other.

A related notion is the Kronecker product of two matrices. It is a generalization of the outer product between two vectors to matrices.

Definition 1.3. Let $\mathbf{A}$ be an $m \times n$ matrix with the elements $\mathbf{a}_{ij}$, and let $\mathbf{B}$ be a $p \times q$ matrix, then their Kronecker product is the $mp \times nq$ block matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}. \tag{1.9}$$

An important identity relating the vec and Kronecker product operators is

$$\mathrm{Tr}\left\{\mathbf{A}^T\mathbf{B}\mathbf{C}\mathbf{D}^T\right\} = \mathrm{vec}\left(\mathbf{A}\right)^T\left(\mathbf{D}\otimes\mathbf{B}\right)\mathrm{vec}\left(\mathbf{C}\right). \tag{1.10}$$

Other properties of Kronecker products include (see [61] for more details):

$$\left(\mathbf{A}\otimes\mathbf{B}\right)\left(\mathbf{C}\otimes\mathbf{D}\right) = \left(\mathbf{AC}\otimes\mathbf{BD}\right) \tag{1.11}$$

$$\left(\mathbf{A}\otimes\mathbf{B}\right)^t = \left(\mathbf{A}^t\otimes\mathbf{B}^t\right) \tag{1.12}$$

$$\left|\mathbf{A}\otimes\mathbf{B}\right| = \left|\mathbf{A}\right|^p\left|\mathbf{B}\right|^p. \tag{1.13}$$

In the first identity, we assume the matrices are conforming, in the second we assume they are positive definite and in the last identity, we assume that both are of size $p\times p$.

1.2 G-convexity

We begin with a brief review on general g-convexity on Riemannian manifolds $\mathcal{M}$. More details on this topic can be found in [70, 51, 11].

Definition 1.4. For each pair $q_0, q_1 \in \mathcal{M}$ we define a geodesic $q_t^{q_0,q_1} \in \mathcal{M}$ for $t \in [0,1]$ as a continuous path connecting the pair[1]. For simplicity, we omit the superscripts and assume q_0 and q_1 are understood from the context.

Definition 1.5. A set $\mathcal{S} \subseteq \mathcal{M}$ is g-convex if $q_t^{q_0,q_1} \in \mathcal{S}$ for any $q_0, q_1 \in \mathcal{S}$ and $t \in [0,1]$.

Definition 1.6. A real-valued function f is g-convex on a g-convex set $\mathcal{S}$ if $f(q_t) \leq tf(q_1) + (1-t)f(q_0)$ for any $q_0, q_1 \in \mathcal{S}$ and $t \in [0,1]$. The function is strictly g-convex if $f(q_t) < tf(q_1) + (1-t)f(q_0)$ for all $q_0 \neq q_1 \in \mathcal{S}$ and $t \in (0,1)$.

[1] A more rigorous definition of a geodesic requires a metric and is associated with the unique path of minimal length, but is not necessary for our exposition.

The most important property of g-convexity is the following theorem.

Theorem 1.3 ([18]). Any local minimum of a g-convex function over a g-convex set is a global minimum. The global minimizer of a strictly g-convex function is unique.

Proof. Assume $q_0 \neq q_1 \in \mathcal{S}$ are local minimizers of a g-convex function $f(q)$ over a g-convex set $\mathcal{S}$. Assume in contradiction that only q_1 is a global minimizer. Let $q_t^{q_0,q_1} \in \mathcal{S}$ be the geodesic between them. Then,

$$\begin{aligned} f\left(q_t^{q_0,q_1}\right) &\leq tf(q_1) + (1-t) f(q_0) \\ &< f(q_0), \quad \forall \quad t \in (0,1], \end{aligned} \tag{1.14}$$

where the first inequality is due to geodesic convexity and the second due to $f(q_1) < f(q_0)$. For sufficiently small t, this is a contradiction to local optimality of q_0.

For the second part of the statement, we assume in contradiction that $q_0 \neq q_1 \in \mathcal{S}$ are both global minimizers of a strictly g-convex function $f(q)$ over a g-convex set $\mathcal{S}$. Let $q_t^{q_0,q_1} \in \mathcal{S}$ be the geodesic between them. Then,

$$f\left(q_t^{q_0,q_1}\right) < tf(q_1) + (1-t) f(q_0) \tag{1.15}$$

$$= f(q_0), \quad \forall \quad t \in (0,1], \tag{1.16}$$

which is a contradiction to the global optimality of q_0. □

Theorem 1.3 is of paramount importance. Its application to classical convexity led to the overwhelming interest in convex optimization in almost all fields of engineering. Finding local minima of well-behaved functions is a tractable task via simple descent algorithms, whereas finding global minima is typically a much harder problem. Thus, in some sense, convexity has become a synonym for tractability. When one encounters an optimization problem, it is standard to check whether it is convex and if it is not then to try and find a convex approximation. But in fact Theorem 1.3 is more general and holds also for g-convex sets and functions. This generalization is less known and has only attracted attention in the last years. Specifically, in this chapter, we will show that the optimization problems associated with Tyler's M-estimator are all g-convex rather than classically convex.

The above definitions and results are general for arbitrary manifolds. The most famous use of g-convexity is classical convexity on Euclidean manifolds. In this setting, the geodesic is a simple segment

$$q_t^{q_0,q_1} = (1-t)q_0 + tq_1 \tag{1.17}$$

and there is a great body of knowledge on its associated convex sets and functions, e.g. [18].

Two intuitive results allow us to easily identify g-convex functions:

Lemma 1.4 (Convexity with respect to t [70])**.** A function f on a g-convex set $\mathcal{S}$ is g-convex if $f(q_t)$ is classically convex in $t \in [0,1]$ for any $q_0, q_1 \in \mathcal{S}$.

Lemma 1.5. [Midpoint convexity] A continuous function f on a g-convex set $\mathcal{S}$ is g-convex if $f(q_{\frac{1}{2}}) \leq \frac{1}{2}f(q_1) + \frac{1}{2}f(q_0)$ for any $q_0, q_1 \in \mathcal{S}$.

Proof. By applying midpoint convexity to $q_0 = 0$ and $q_1 = 1$, Definition 1.6 holds for $t = \frac{1}{2}$. Applying midpoint convexity again to $(q_0, q_1) = (0, \frac{1}{2})$ and $(\frac{1}{2}, 1)$, Definition 1.6 holds for $t = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$. Applying the midpoint convexity repeatedly we obtain Definition 1.6 for any $t = \frac{m}{2^n}$ for integers $m, n > 0$ and $m < 2^n$. By the continuity of f, Definition 1.6 holds for any $0 < t < 1$. □

1.3 G-convexity for positive definite matrices

In this section, we restrict the attention to g-convexity on a specific manifold, the cone of positive definite matrices. With each $\mathbf{Q}_0 \succ \mathbf{0}, \mathbf{Q}_1 \succ \mathbf{0}$ we associate the geodesic

$$\mathbf{Q}_t = \mathbf{Q}_0^{\frac{1}{2}} \left(\mathbf{Q}_0^{-\frac{1}{2}} \mathbf{Q}_1 \mathbf{Q}_0^{-\frac{1}{2}} \right)^t \mathbf{Q}_0^{\frac{1}{2}}, \quad t \in [0,1]. \tag{1.18}$$

The derivation of this fact can be found at [15, Section 6.1.6]. For simplicity, hereinafter we define g-convexity as g-convexity on the positive definite cone using the above geodesic.

To get more insight into the form of this geodesic, it is instructive to consider the special case in which $\mathbf{Q}_0$ and $\mathbf{Q}_1$ are positive scalars denoted q_0 and q_1. In this case, the geodesic reduces to

$$q_t = q_0^{1-t} q_1^t \tag{1.19}$$

which is quite intuitive and is simply a regular line after an exponential change of variable. Throughout this chapter, we will follow each result by considering its special scalar case. This will provide more intuition and is important for testing the validity of the results. Note that this scalar case is in fact the workhorse behind the successful Geometric Programming (GP) framework [17]. In some sense, one may interpret the results below as a matrix extension of the GP framework.

The scalar intuition can be formally extended to the matrix case via joint diagonalization. Using Lemma 1.1, we apply the decomposition

$$\begin{aligned} \mathbf{Q}_0 &= \mathbf{V}\mathbf{V}^T \\ \mathbf{Q}_1 &= \mathbf{V}\mathbf{D}\mathbf{V}^T \end{aligned} \tag{1.20}$$

where $\mathbf{V}$ is square and invertible, and $\mathbf{D}$ is diagonal with positive elements. It is straightforward to show that the geodesic between them is simply

$$\mathbf{Q}_t = \mathbf{V}\mathbf{D}^t\mathbf{V}^T \tag{1.21}$$

There is an interesting relation between the geodesic in (1.18) and the arithmetic-geometric mean inequality. In scalars, this seminal inequality states that

$$q_t = q_0^{1-t} q_1^t \leq (1-t)q_0 + tq_1 \tag{1.22}$$

The geodesic in (1.18) can be interpreted as the natural matrix extension and follows a similar matrix inequality.

Theorem 1.6. The matrix geodesic satisfies the arithmetic-geometric inequality

$$\mathbf{Q}_t = \mathbf{Q}_0^{\frac{1}{2}} \left(\mathbf{Q}_0^{-\frac{1}{2}} \mathbf{Q}_1 \mathbf{Q}_0^{-\frac{1}{2}}\right)^t \mathbf{Q}_0^{\frac{1}{2}} \preceq (1-t)\mathbf{Q}_0 + t\mathbf{Q}_1 \tag{1.23}$$

Proof. Using the simultaneous diagonalization definition of the geodesic, we need to show that

$$\mathbf{R}\mathbf{D}^t\mathbf{R}^T \preceq \mathbf{R}\left[(1-t)\,\mathbf{I} + t\mathbf{D}\right]\mathbf{R}^T \tag{1.24}$$

The matrix $\mathbf{R}$ is invertible, and the inequality reduces to

$$\mathbf{I}^{1-t}\mathbf{D}^t \preceq (1-t)\,\mathbf{I} + t\mathbf{D} \tag{1.25}$$

which, due to the diagonal structure, is simply multiple scalar arithmetic-geometric inequalities. □

The midpoint of the geodesic, denoted by $\mathbf{Q}_{\frac{1}{2}}$, is typically interpreted as the matrix geometric mean [60]. It has an elegant characterization via its extremal properties.

Theorem 1.7 (Extremal characterization of geometric mean)**.** The positive definite geometric mean satisfies

$$\mathbf{Q}_{\frac{1}{2}} \succeq \mathbf{Z} \tag{1.26}$$

for any symmetric $\mathbf{Z}$ that satisfies

$$\begin{bmatrix} \mathbf{Q}_0 & \mathbf{Z} \\ \mathbf{Z} & \mathbf{Q}_1 \end{bmatrix} \succeq \mathbf{0}. \tag{1.27}$$

Scalar intuition: In the scalar case, we have

$$\begin{bmatrix} q_0 & z \\ z & q_1 \end{bmatrix} \succeq \mathbf{0} \quad \Leftrightarrow \quad |z| \leq \sqrt{q_0 q_1} \tag{1.28}$$

and the maximum value of z is the well known scalar geometric mean.

Proof. Using the simultaneous diagonalization definition of the geodesic, we need to show that

$$\mathbf{D}^{\frac{1}{2}} \succeq \mathbf{Z}\ \forall\ \mathbf{Z} = \mathbf{Z}^T \ : \begin{bmatrix} \mathbf{I} & \mathbf{Z} \\ \mathbf{Z} & \mathbf{D} \end{bmatrix} \succeq \mathbf{0}. \tag{1.29}$$

Using Schur's Lemma 1.2, the condition is equivalent to $\mathbf{D} \succeq \mathbf{Z}\mathbf{Z}^T$. Both sides of this matrix inequality are positive definite, thus we can take their square roots and obtain the required result. □

In the sequel, the following properties of the geodesic will be useful.

Lemma 1.8. The inverse operator commutes with the geodesic

$$\left([\mathbf{Q}]_t\right)^{-1} = \left[\mathbf{Q}^{-1}\right]_t. \tag{1.30}$$

Scalar intuition: This lemma is trivial in the scalar case.

Proof. The proof is straightforward by the fact that the inverse operator commutes with the matrix power operation in the definition of the geodesic. □

Lemma 1.9. The matrix Kronecker product commutes with the geodesic

$$[\mathbf{Q}_1]_t \otimes \cdots \otimes [\mathbf{Q}_K]_t = [\mathbf{Q}_1 \otimes \cdots \otimes \mathbf{Q}_K]_t. \tag{1.31}$$

Proof. The identity holds due to properties (1.11)-(1.12). □

Lemma 1.10 (Positive linear maps [15])**.** Let $\mathbf{B} \succeq \mathbf{0}$ and define the positive linear map

$$\phi(\mathbf{Q}) = \mathbf{A}\mathbf{Q}\mathbf{A}^T + \mathbf{B} \tag{1.32}$$

then the following inequality holds

$$\phi\left([\mathbf{Q}]_{\frac{1}{2}}\right) \preceq [\phi(\mathbf{Q})]_{\frac{1}{2}}. \tag{1.33}$$

Equality holds when $\mathbf{B} = \mathbf{0}$ and $\mathbf{A}$ is square and invertible .

Proof. By the extremal characterization of $[\phi(\mathbf{Q})]_{\frac{1}{2}}$ we have

$$[\phi(\mathbf{Q})]_{\frac{1}{2}} \succeq \mathbf{Z} \; \forall \; \mathbf{Z} = \mathbf{Z}^T \; : \; \begin{bmatrix} \phi(\mathbf{Q}_0) & \mathbf{Z} \\ \mathbf{Z} & \phi(\mathbf{Q}_1) \end{bmatrix} \succeq \mathbf{0} \tag{1.34}$$

Thus, we need to show that $\mathbf{Z} = \phi\left([\mathbf{Q}]_{\frac{1}{2}}\right)$ satisfies the condition on $\mathbf{Z}$. By extreme characterization of the $\mathbf{Q}_{\frac{1}{2}}$ we know that

$$\begin{bmatrix} \mathbf{Q}_0 & \mathbf{Q}_{\frac{1}{2}} \\ \mathbf{Q}_{\frac{1}{2}} & \mathbf{Q}_1 \end{bmatrix} \succeq \mathbf{0}. \tag{1.35}$$

Define

$$\widetilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix}, \qquad \widetilde{\mathbf{B}} = \begin{bmatrix} \mathbf{B} & \mathbf{B} \\ \mathbf{B} & \mathbf{B} \end{bmatrix} \tag{1.36}$$

and note that $\widetilde{\mathbf{B}}$ is positive semidefinite. Multiplying both sides by $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{A}}^T$ and adding $\widetilde{\mathbf{B}} \succeq \mathbf{0}$ will not change the inequality, and we obtain

$$\begin{bmatrix} \mathbf{A}\mathbf{Q}_0\mathbf{A}^T + \mathbf{B} & \mathbf{A}\mathbf{Q}_{\frac{1}{2}}\mathbf{A}^T + \mathbf{B} \\ \mathbf{A}\mathbf{Q}_{\frac{1}{2}}\mathbf{A}^T + \mathbf{B} & \mathbf{A}\mathbf{Q}_1\mathbf{A}^T + \mathbf{B} \end{bmatrix} = \begin{bmatrix} \phi(\mathbf{Q}_0) & \phi(\mathbf{Z}) \\ \phi(\mathbf{Z}) & \phi(\mathbf{Q}_1) \end{bmatrix} \succeq \mathbf{0} \tag{1.37}$$

which is exactly what we needed to show. When $\mathbf{B} = \mathbf{0}$ and $\mathbf{A}$ is invertible, equality holds since the product of invertible matrices commutes with the inverse and matrix square root operation. □

It is straightforward to consider joint g-convexity on multiple positive definite matrices. The joint geodesic between multiple pairs of matrices is simply the multiple individual geodesics. Actually, it can be conveniently expressed using a single large block-diagonal and positive definite matrix.

The geodesic in (1.18) is the starting point for the following g-convex analysis. We now review its fundamental g-convex sets, g-convex functions and the operations that preserve g-convexity.

1.3.1 G-convex sets

The most obvious g-convex set is the manifold itself, i.e., the cone of positive definite matrices. The Cartesian product of a few such cones is also g-convex. The canonical approach to characterize this manifold is via a block diagonal matrix which consists of the various positive definite matrices in its diagonal blocks.

Theorem 1.11. The set of block diagonal positive definite matrices (with prescribed and known blocks) is g-convex. A special case is the set of diagonal positive definite matrices.

The proof is trivial and omitted.

Another g-convex set is the set of matrices which are invariant to congruence transformations:

Theorem 1.12. Let $\mathcal{U}$ be a set of orthogonal matrices, then the set $\mathcal{F} = \left\{\mathbf{Q} \succ \mathbf{0} \;:\; \mathbf{Q} = \mathbf{U}\mathbf{Q}\mathbf{U}^T \;\forall\, \mathbf{U} \in \mathcal{U}\right\}$ is g-convex.

Proof. We assume that $\mathbf{Q}_0 = \mathbf{U}\mathbf{Q}_0\mathbf{U}^T$ and $\mathbf{Q}_1 = \mathbf{U}\mathbf{Q}_1\mathbf{U}^T$. By definition, $\mathbf{Q}_t$ is the geodesic between $\mathbf{Q}_0$ and $\mathbf{Q}_1$. Due to the assumption, it is also the geodesic between $\mathbf{U}\mathbf{Q}_0\mathbf{U}^T$ and $\mathbf{U}\mathbf{Q}_1\mathbf{U}^T$. Therefore, $\mathbf{Q}_t = \left[\mathbf{U}\mathbf{Q}\mathbf{U}^T\right]_t$ and applying Lemma 1.10 yields $\mathbf{Q}_t = \mathbf{U}\mathbf{Q}_t\mathbf{U}^T$ as required.

□

Surprisingly, Theorems 1.11 and 1.12 are highly related. Group representation theory shows that if $\mathcal{U}$ is a unitary group[2], then the set $\mathcal{F}$ can be characterized as the set of matrices that can be "rotated" into a block diagonal form using a known and prescribed basis, e.g., [73, 74]. A well known example is the set of circular positive definite matrices. It is invariant to shifts, and can be rotated into a diagonal form using the Fourier transform.

1.3.2 G-convex functions

Next, we turn to the basic g-convex functions. First, we introduce the fundamental g-linear function which is both g-convex and g-concave (i.e., its negative is g-convex). In the scalar case, g-convexity is simply convexity after an exponential change of variables. Thus, the scalar g-linear function is the logarithm. The natural multidimensional extension is the log-determinant.

Lemma 1.13. The functions

$$f(\mathbf{Q}) = \pm\log|\mathbf{Q}| \tag{1.38}$$

are g-convex.

Proof. Plugging the geodesic in (1.21) into the function yields

$$\begin{aligned} f(\mathbf{Q}_t) &= \pm\log\left|\mathbf{R}\mathbf{D}^t\mathbf{R}^T\right| \\ &= \pm2\log|\mathbf{R}| \pm t\log|\mathbf{D}| \end{aligned} \tag{1.39}$$

[2] A unitary group is a set of unitary matrices including the identity matrix and closed under multiplication and inversion

which is clearly a linear (and convex function) in t:

$$f(\mathbf{Q}_t) = tf(\mathbf{Q}_0) + (1-t)f(\mathbf{Q}_0).$$

□

This result is counterintuitive. In classical convexity the log determinant is a concave function whereas in our manifold it is g-convex.

Lemma 1.14. Let $\mathbf{h} \in \mathrm{R}^m$. The function

$$f(\mathbf{Q}) = \mathbf{h}^T\mathbf{Q}\mathbf{h} \tag{1.40}$$

is strictly g-convex (unless $\mathbf{h} = \mathbf{0}$).

Scalar intuition: In this case, the function reduces to h^2q. After a change of variable $q = e^z$, we obtain a simple convex function h^2e^z.

Proof. Substituting $\mathbf{Q}_t$ in (1.21) instead of $\mathbf{Q}$ yields

$$\begin{aligned} f(\mathbf{Q}_t) &= \mathbf{h}^T\mathbf{R}\mathbf{D}^t\mathbf{R}^T\mathbf{h} \\ &= \sum_{i=1}^m \left[\mathbf{R}^T\mathbf{h}\right]_i^2 \mathbf{D}_{ii}^t \\ &= \sum_{i=1}^m \left[\mathbf{R}^T\mathbf{h}\right]_i^2 e^{t\log \mathbf{D}_{ii}} \end{aligned} \tag{1.41}$$

which is strictly convex in t since it is a positively weighted sum of strictly convex exponential functions. Strictness is due to the full rank property of $\mathbf{R}$ and $\mathbf{R}^T\mathbf{h} \neq \mathbf{0}$. Strict g-convexity of $f(\mathbf{Q})$ follows from the definition and the strict convexity in t. □

A direct consequence is the following result.

Lemma 1.15. The function $g(\mathbf{Q}) = \mathrm{Tr}\{\mathbf{Q}\}$ is strictly g-convex.

Proof. The trace is the sum of (1.40) with $\mathbf{h}_i$ being the unit vectors. Thus, the proof is a direct application of Lemma (1.14). □

Lemma 1.16. The condition number

$$f(\mathbf{Q}) = \frac{\lambda_{\max}(\mathbf{Q})}{\lambda_{\min}(\mathbf{Q})}. \tag{1.42}$$

is g-convex.

Proof. We use the variational characterization of extreme eigenvalues:

$$\lambda_{\max}(\mathbf{Q}) = \max_{\mathbf{u}:\|\mathbf{u}\|=1} \mathbf{u}^T\mathbf{Q}\mathbf{u} \tag{1.43}$$

$$\frac{1}{\lambda_{\min}(\mathbf{Q})} = \lambda_{\max}\left(\mathbf{Q}^{-1}\right) = \max_{\mathbf{v}:\|\mathbf{v}\|=1} \mathbf{v}^T\mathbf{Q}^{-1}\mathbf{v} \tag{1.44}$$

Due to the monotonicity of the logarithm, we obtain

$$f(\mathbf{Q}) = e^{\max_{\mathbf{u}:\|\mathbf{u}\|=1}\log\left(\mathbf{u}^T\mathbf{Q}\mathbf{u}\right)+\max_{\mathbf{v}:\|\mathbf{v}\|=1}\log\left(\mathbf{v}^T\mathbf{Q}^{-1}\mathbf{v}\right)}. \tag{1.45}$$

Plugging in the geodesic in (1.18) yields convex log-sum-exp functions in the maximizations objective. Finally, the point-wise maximum of a set of convex functions is convex, and the exponent of a convex function is also convex, e.g., [18]. □

1.3.3 Operations that preserve g-convexity

To enrich the class of g-convex sets and functions, it is instructive to consider operations that preserve g-convexity. See also [77] for more results and details.

Lemma 1.17. Let $f(\mathbf{Q})$ be a g-convex function. Then so is $g(\mathbf{Q}) = f(\mathbf{Q}^{-1})$.

Proof. We use the following chain of inequalities

$$\begin{aligned} g(\mathbf{Q}_t) &= f\left(([\mathbf{Q}]_t)^{-1}\right) \\ &= f\left(\left[\mathbf{Q}^{-1}\right]_t\right) \qquad \text{Lemma 1.8} \\ &\leq (1-t)f\left(\left[\mathbf{Q}^{-1}\right]_0\right) + tf\left(\left[\mathbf{Q}^{-1}\right]_1\right) \\ &= (1-t)g(\mathbf{Q}_0) + tg(\mathbf{Q}_1) \end{aligned} \tag{1.46}$$

□

Scalar intuition: Thus, $q = e^z$ and its inverse is given by $q^{-1} = e^{-z}$. If $f(e^z)$ is convex then $f(e^{-z})$ is convex too since affine transformations preserve convexity.

In the classical sense, the most important operation that preserves convexity is affine transformations. In the g-convexity counterpart, these transformations are more complex as we must remain within the symmetric positive definite cone.

Lemma 1.18 ([77])**.** Let $f(\mathbf{Q})$ be a continuous, monotonically increasing in the sense that $f(\mathbf{Q}_1) \leq f(\mathbf{Q}_2)$ for $\mathbf{Q}_1 \preceq \mathbf{Q}_2$, and g-convex function. Let $\mathbf{A}$ and $\mathbf{B} \succ \mathbf{0}$ be fixed matrices. Then $g(\mathbf{Q}) = f\left(\mathbf{AQA}^T + \mathbf{B}\right)$ is also g-convex.

Proof. We use the following chain of inequalities

$$\begin{aligned} g\left(\mathbf{Q}_{\frac{1}{2}}\right) &= f\left(\mathbf{A}\left[\mathbf{Q}\right]_{\frac{1}{2}}\mathbf{A}^T + \mathbf{B}\right) \\ &\leq f\left(\left[\mathbf{AQA}^T + \mathbf{B}\right]_{\frac{1}{2}}\right) \\ &\leq \frac{1}{2}f\left(\left[\mathbf{AQA}^T + \mathbf{B}\right]_0\right) + \frac{1}{2}f\left(\left[\mathbf{AQA}^T + \mathbf{B}\right]_1\right) \\ &= \frac{1}{2}g(\mathbf{Q}_0) + \frac{1}{2}g(\mathbf{Q}_1) \end{aligned} \tag{1.47}$$

where the first inequality is due to monotonicity and Lemma 1.10, and the second due to g-convexity. Applying Lemma 1.5, the geodesic midpoint convexity (1.47) of a continuous function implies its geodesic convexity. □

Scalar intuition: Specializing $\mathbf{AQA}^T + \mathbf{B}$ to the scalar case yields $e^{z\log a^2 + \log b}$. This is an affine transformation which preserves convexity.

Note the expressive power of Lemmas 1.17 and 1.18. The following result is a direct corollary.

Lemma 1.19. Let $\mathbf{H}_i$ for $i = 1, \cdots, n$ be a set of fixed matrices whose columns span the real space. The function $f(\mathbf{Q}) = \log\left|\sum_{i=1}^n \mathbf{H}_i\mathbf{Q}^{\pm 1}\mathbf{H}_i^T\right|$ is g-convex.

Scalar intuition: In the scalar case, the logdet function is a simple logarithm and, after a change of variables, its argument is a sum of exponents. Indeed, it is well known that the log-sum-exp function is convex.

In the special case when $\mathbf{H}_i$ are vectors, we can also examine strict g-convexity.

Lemma 1.20. Let $\mathbf{h}_i \in \mathrm{R}^m$ be nonzero vectors for $i = 1, \cdots, n$. The function

$$f(\mathbf{Q}) = \log\left(\sum_{i=1}^n \mathbf{h}_i^T\mathbf{Q}\mathbf{h}_i\right) \tag{1.48}$$

is g-convex. Equality holds in $f([\mathbf{Q}]_{\frac{1}{2}}) \leq \frac{1}{2}(f(\mathbf{Q}_0)+f(\mathbf{Q}_1))$ if and only if $\{\mathbf{Q}_0^{\frac{1}{2}}\mathbf{h}_i\}_{i=1}^n$ spans an eigenspace of $\mathbf{Q}_0^{-\frac{1}{2}}\mathbf{Q}_1\mathbf{Q}_0^{-\frac{1}{2}}$.

Proof. G-convexity can be proved as a special case of Lemma 1.19. To analyze the strictness condition, we use a different proof. Eliminating the logarithms, we need to show that

$$\left(\sum_{i=1}^n \mathbf{h}_i^T \left[\mathbf{Q}\right]_{\frac{1}{2}} \mathbf{h}_i\right)^2 \leq \left(\sum_{i=1}^n \mathbf{h}_i^T \mathbf{Q}_0 \mathbf{h}_i\right)\left(\sum_{i=1}^n \mathbf{h}_i^T \mathbf{Q}_1 \mathbf{h}_i\right) \tag{1.49}$$

To simplify the notation, we define

$$\begin{aligned} \mathbf{u}_i &= \mathbf{Q}_0^{\frac{1}{2}}\mathbf{h}_i \\ \mathbf{v}_i &= \left(\mathbf{Q}_0^{-\frac{1}{2}}\mathbf{Q}_1\mathbf{Q}_0^{-\frac{1}{2}}\right)^{\frac{1}{2}}\mathbf{Q}_0^{\frac{1}{2}}\mathbf{h}_i \end{aligned} \tag{1.50}$$

and (1.49) is equivalent to

$$\left(\sum_{i=1}^n \mathbf{u}_i^T\mathbf{v}_i\right)^2 \leq \left(\sum_{i=1}^n \|\mathbf{u}_i\|^2\right)\left(\sum_{i=1}^n \|\mathbf{v}_i\|^2\right). \tag{1.51}$$

We prove this using the Cauchy-Schwartz inequality twice

$$\begin{aligned} \left(\sum_{i=1}^n \mathbf{u}_i^T\mathbf{v}_i\right)^2 &= \left(\sum_{i=1}^n |\mathbf{u}_i^T\mathbf{v}_i|\right)^2 \\ &\leq \left(\sum_{i=1}^n \|\mathbf{u}_i\|\|\mathbf{v}_i\|\right)^2 \\ &\leq \left(\sum_{i=1}^n \|\mathbf{u}_i\|^2\right)\left(\sum_{i=1}^n \|\mathbf{v}_i\|^2\right) \end{aligned} \tag{1.52}$$

In the first inequality, we bound each bilinear term independently. In the second inequality, we bound their sum. Equalities hold if and only if $\mathbf{u}_i = c_i\mathbf{v}_i$ for some c_i and for all i, and $\|\mathbf{u}_i\| = d\|\mathbf{v}_i\|$ for some d and for all i. Together, c_i must all be identical. In terms of $\mathbf{Q}_0$, $\mathbf{Q}_1$ and $\mathbf{h}_i$, this means that $\{\mathbf{Q}_0^{\frac{1}{2}}\mathbf{h}_i\}_{i=1}^n$ are all eigenvectors of $\mathbf{Q}_0^{-\frac{1}{2}}\mathbf{Q}_1\mathbf{Q}_0^{-\frac{1}{2}}$ and share the same eigenvalue. □

Lemma 1.21. Let $f(\mathbf{Q})$ be a g-convex function, then $g(\mathbf{Q}_1, \cdots, \mathbf{Q}_K) = f(\mathbf{Q}_1 \otimes \cdots \otimes \mathbf{Q}_K)$ is jointly g-convex in all of its arguments.

Proof.

$$\begin{aligned} & tf\left([\mathbf{Q}_1]_1 \otimes \cdots \otimes [\mathbf{Q}_J]_1\right) + (1-t)f\left([\mathbf{Q}_1]_0 \otimes \cdots \otimes [\mathbf{Q}_J]_0\right) \\ & = tf\left([\mathbf{Q}_1 \otimes \cdots \otimes \mathbf{Q}_J]_1\right) + (1-t)f\left([\mathbf{Q}_1 \otimes \cdots \otimes \mathbf{Q}_J]_0\right) \\ & \geq f\left([\mathbf{Q}_1 \otimes \cdots \otimes \mathbf{Q}_J]_t\right) \qquad \text{g}-\text{convexity} \\ & = f\left([\mathbf{Q}_1]_t \otimes \cdots \otimes [\mathbf{Q}_J]_t\right) \qquad \text{Lemma 1.9} \end{aligned} \tag{1.53}$$

□

Scalar intuition: In the scalar case, the Kronecker product is a regular product. After an exponential change of variables, products become sum. It is well known that regular convexity is preserved under sums.

1.4 Majorization-minimization algorithm

In this section, we provide an introduction to the majorization-minimization (MM) algorithm. More details on the method and its analysis are available in [41, 31, 71]. The approach seeks to minimize a difficult objective function by iteratively minimizing "easier" upper bounds. Formally, suppose we want to find the minimizer of $f(x)$ in a set $\mathcal{D}$, denoted by

$$\arg\min_{x\in\mathcal{D}} \; f(x). \tag{1.54}$$

The iterations are defined as

$$x_{k+1} = T(x_k), \quad T(x_k) = \arg\min_{x\in\mathcal{D}} U(x, x_k), \tag{1.55}$$

where the majorization surrogate function (the upper bound) satisfies

$$\begin{aligned} U(x, x_k) &\geq f(x) \quad \forall \quad x, x_k, \\ U(x_k, x_k) &= f(x_k) \quad \forall \quad x_k. \end{aligned} \tag{1.56}$$

Under technical conditions formally described below, these properties ensure monotonicity of the algorithm and attainment of a local minimum. The convergence of algorithms in the rest of the book are proved by combining this theorem with the properties of the functions f and U used in the various algorithms.

Theorem 1.22. When f and T are continuous functions and f is bounded from below, any accumulation point of the sequence x_k, $\hat{x}$, is a minimizer of $U(x, \hat{x})$ if it lies in the interior of $\mathcal{D}$. In particular:

- If the minimizer of $U(x, \hat{x})$ is unique, then $\hat{x} = T(\hat{x})$, that is, $\hat{x}$ is the fixed point of the mapping T.
- If U and f are differentiable, then $\hat{x}$ is a stationary point of $f(x)$.

Proof. First of all, $f(x_k)$ is a nonincreasing sequence:

$$f(T(x_k)) = f(x_{k+1}) \leq U(x_{k+1}, x_k) \leq U(x_k, x_k) = f(x_k). \quad (1.57)$$

Since f is bounded from below, $f(x_k)$ converges. Therefore, for the converging subsequence $\{x_{m_k}\}_k \to \hat{x}$, $\lim_{k\to\infty} f(T(x_{m_k})) - f(x_{m_k}) = 0$. Applying the continuity of f and T, we have $f(T(\hat{x})) = f(\hat{x})$, and the equality in (1.57) holds if x_k and x_{k+1} are replaced by $\hat{x}$ and $T(\hat{x})$. Since the second inequality in (1.57) achieves equality, $\hat{x}$ is a minimizer of $U(x, \hat{x})$.

The proof of the special cases are as follows:

- When the minimizer of $U(x, \hat{x})$ is unique, by definition it is $T(\hat{x})$, and we have $T(\hat{x}) = \hat{x}$.
- If $\hat{x}$ is not a stationary point of f, then $U'(x, \hat{x})|_{x=\hat{x}} = f'(\hat{x}) \neq 0$, and we have $U(T(\hat{x}), \hat{x}) = \min_x U(x, \hat{x}) < U(\hat{x}, \hat{x})$, where the first equality follows from (1.56). Applying the same argument as in (1.57), it contradicts the conclusion that $f(\hat{x}) = f(T(\hat{x}))$.

□

References

[1] Y. I. Abramovich and N. K. Spencer. Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007*, volume 3. IEEE, 2007.

[2] Y.I Abramovich and O. Besson. Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach; part 1: The over-sampled case. *IEEE Transactions on Signal Processing*, 61(23):5807–5818, Dec 2013.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[4] G. I. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, 2010.

[5] T. W. Anderson. Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, pages 135–141, 1973.

[6] O. Arslan. Convergence behavior of an iterative reweighting algorithm to compute multivariate M-estimates for location and scatter. *Journal of Statistical Planning and Inference*, 118(1):115–128, 2004.

[7] A. Aubry, A. De Maio, L. Pallotta, and A. Farina. Maximum likelihood estimation of a structured covariance matrix with a condition number constraint. *IEEE Transactions on Signal Processing*, 60(6):3004–3021, 2012.

[8] C. Auderset, C. Mazza, and E. Ruh. Grassmannian Estimation. *Arxiv preprint arXiv:0809.3697*, 2008.

[9] C. Auderset, C. Mazza, and E.A. Ruh. Angular Gaussian and Cauchy estimation. *Journal of Multivariate Analysis*, 93(1):180–197, 2005.

[10] F. Bandiera, O. Besson, and G. Ricci. Knowledge-aided covariance matrix estimation and adaptive detection in compound-Gaussian noise. *IEEE Trans. on Signal Processing*, 58(10):5391 –5396, oct. 2010.

[11] A. Ben-Tal. On generalized means and generalized convex functions. *Journal of Optimization Theory and Applications*, 21(1):1–13, 1977.

[12] O. Besson and Y. Abramovich. On the Fisher information matrix for multivariate elliptically contoured distributions. *Signal Processing Letters, IEEE*, 20(11):1130–1133, 2013.

[13] O. Besson and Y. I. Abramovich. Expected likelihood approach for low sample support covariance matrix estimation in Angular central Gaussian distributions. In *Asilomar Conference on Signals, Systems and Computers*, pages 682–686. IEEE, 2013.

[14] O. Besson and Y.I Abramovich. Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach; part 2: The under-sampled case. *IEEE Transactions on Signal Processing*, 61(23):5819–5829, Dec 2013.

[15] R. Bhatia. *Positive definite matrices.* Princeton University Press, 2009.

[16] S. Bhattacharyya and P. J. Bickel. Adaptive estimation in elliptical distributions with extensions to high dimensions. *Preprint*, 2014.

[17] S. Boyd, S.J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. *Optimization and Engineering*, 8(1):67–127, 2007.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

[19] J. P. Burg, D. G. Luenberger, and D. L. Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, 1982.

[20] J. A. Cadzow. Signal enhancement-a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1):49–62, 1988.

[21] B. D. Carlson. Covariance matrix estimation errors and diagonal loading in adaptive arrays. *IEEE Transactions on Aerospace and Electronic Systems*, 24(4):397–401, 1988.

[22] Y. Chen, A. Wiesel, Y.C. Eldar, and A.O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

[23] Y. Chen, A. Wiesel, and A.O. Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107, 2011.

[24] Y. Chitour, R. Couillet, and F. Pascal. Uniqueness of Maronna's M-estimators of scatter. *arXiv preprint arXiv:1403.5977*, 2014.

[25] Y. Chitour and F. Pascal. Exact maximum likelihood estimates for SIRV covariance matrix: Existence and algorithm analysis. *IEEE Transactions on Signal Processing*, 56(10):4563–4573, 2008.

[26] E. Conte, A. De Maio, and G. Ricci. Recursive estimation of the covariance matrix of a compound-Gaussian process and its application to adaptive CFAR detection. *IEEE Transactions on Signal Processing*, 50(8):1908–1915, 2002.

[27] E. Conte, M. Lops, and G. Ricci. Asymptotically optimum radar detection in compound-Gaussian clutter. *IEEE Transactions on Aerospace and Electronic Systems*, 31(2):617–625, 2002.

[28] R. Couillet and M. R. McKay. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131(0):99 – 120, 2014.

[29] L. Dumbgen. On Tyler's M-functional of scatter in high dimension. *Annals of the Institute of Statistical Mathematics*, 50(3):471–491, 1998.

[30] P. Dutilleul. The MLE algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, 64(2):105–123, 1999.

[31] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak. Majorization–minimization algorithms for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 16(12):2980–2991, 2007.

[32] G. Frahm. Generalized elliptical distributions: theory and applications. *Unpublished Ph. D. thesis, University of Cologne*, 2004.

[33] G. Frahm and K. Glombek. Semicircle Law for Tyler's M-Estimator. *Arxiv preprint arXiv:1004.3938*, 2010.

[34] G. Frahm and U. Jaekel. A generalization of Tyler's M-estimators to the case of incomplete data. *Computational Statistics and Data Analysis*, 54(2):374–393, 2010.

[35] D. R. Fuhrmann and M. Miller. On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices. *IEEE Transactions on Information Theory*, 34(4):722–729, 1988.

[36] F. Gini and A. Farina. Vector subspace detection in compound-Gaussian clutter. Part I: survey and new results. *IEEE Transactions on Aerospace and Electronic Systems*, 38(4):1295–1311, 2002.

[37] F. Gini and M. Greco. Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter. *Signal Processing*, 82(12):1847–1859, 2002.

[38] M. Greco and F. Gini. Cramer-Rao lower bounds on covariance matrix estimation for complex elliptically symmetric distributions. *IEEE Trans. on Signal Processing*, 61(24), Dec. 2013.

[39] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104. Chapman & Hall/CRC, 2000.

[40] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge university press, 2012.

[41] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[42] S. M. Kay. *Fundamentals of Statistical Signal Processing - Estimation Theory.* Prentice Hall, 1993.

[43] J. T. Kent and D. E. Tyler. Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics*, 15(2):247–254, 1988.

[44] J. T. Kent and D. E. Tyler. Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics*, pages 2102–2119, 1991.

[45] U. Krause. Concave Perron–Frobenius theory and applications. *Nonlinear Analysis: Theory, Methods & Applications*, 47(3):1457–1466, 2001.

[46] O. Ledoit and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, pages 1081–1102, 2002.

[47] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[48] C. H. Lee, P. Dutilleul, and A. Roy. Comment on “models with a Kronecker product covariance structure: Estimation and testing” by M. Srivastava, T. von Rosen, and D. von Rosen, mathematical methods of statistics, 17 (2008), pp. 357–370. *Mathematical Methods of Statistics*, 19(1):88–90, 2010.

[49] H. Li, P. Stoica, and J. Li. Computationally efficient maximum likelihood estimation of structured covariance matrices. *IEEE Transactions on Signal Processing*, 47(5):1314–1323, 1999.

[50] J. Li, P. Stoica, and Z. Wang. On robust capon beamforming and diagonal loading. *IEEE Transactions on Signal Processing*, 51(7):1702–1715, 2003.

[51] L. Liberti. On a class of nonconvex problems where all local minima are global. *Publications de l’Institut Mathematique*, 76(90):101–109, 2004.

[52] N. Lu and D. L. Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statistics and Probability Letters*, 73(4):449 – 457, 2005.

[53] M. Mahot, F. Pascal, P. Forster, and J. Ovarlez. Asymptotic properties of robust complex covariance matrix estimates. *IEEE Transactions on Signal Processing*, 61(13):3348–3356, July 2013.

[54] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483, 1967.

[55] K. V. Mardia and C. Goodall. *Multivariate Environmental Statistics, vol. 6*, chapter Spatial-temporal analysis of multivariate environmental monitoring data. Elsevier, North-Holland, New York, 1993.

[56] R. Maronna, D. Martin, and V. Yohai. *Robust statistics*. John Wiley & Sons, Chichester. ISBN, 2006.

[57] R. A. Maronna and V. J. Yohai. Robust estimation of multivariate location and scatter. *Encyclopedia of Statistical Sciences*, 1998.

[58] R. A. Maronna and R. H. Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 2002.

[59] X. Mestre and M. A. Lagunas. Finite sample size effect on minimum variance beamformers: Optimum diagonal loading factor for large arrays. *IEEE Transactions on Signal Processing*, 54(1):69–82, 2006.

[60] M. Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005.

[61] T.K. Moon and W.C. Stirling. *Mathematical methods and algorithms for signal processing*, volume 204. Prentice hall, 2000.

[62] E. Ollila and V. Koivunen. Robust antenna array processing using M-estimators of pseudo-covariance. In *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, PIMRC 2003*, volume 3, pages 2659–2663. IEEE, 2003.

[63] E. Ollila and D. E. Tyler. Distribution-free detection under complex elliptically symmetric clutter distribution. In *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM'12), Hoboken, NJ, USA*, volume 144. IET, June 17-20 2012.

[64] E. Ollila and D. E. Tyler. Regularized M-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, 62(22):6059–6070, 2014.

[65] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor. Complex elliptically symmetric distributions: survey, new results and applications. *IEEE Transactions on Signal Processing*, 60(11):5597–5625, 2012.

[66] F. Pascal, L. Bombrun, J. Y. Tourneret, and Y. Berthoumieu. Parameter estimation for multivariate generalized Gaussian distributions. *IEEE Trans. on Signal Processing*, 61(23), Dec. 2013.

[67] F. Pascal, Y. Chitour, J. Ovarlez, P. Forster, and P. Larzabal. Covariance structure maximum-likelihood estimates in compound Gaussian noise: Existence and algorithm analysis. *IEEE Transactions on Signal Processing*, 56(1):34–48, 2008.

[68] F. Pascal, Y. Chitour, and Y. Quek. Generalized robust shrinkage estimator and its application to STAP detection problem. *IEEE Transactions on Signal Processing*, 62(21):5640–5651, 2014.

[69] F. Pascal, P. Forster, J.P. Ovarlez, and P. Larzabal. Performance analysis of covariance matrix estimates in impulsive noise. *IEEE Transactions on Signal Processing*, 56(6):2206–2217, 2008.

[70] T. Rapcsak. Geodesic convexity in nonlinear optimization. *Journal of Optimization Theory and Applications*, 69(1):169–183, 1991.

[71] M. Razaviyayn, M. Hong, and Z. Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[72] J. Schafer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):1175, 2005.

[73] P. Shah and V. Chandrasekaran. Group symmetry and covariance regularization. *Electron. J. Statist.*, 6:1600–1640, 2012.

[74] I. Soloveychik, D. Trushin, and A. Wiesel. Group symmetric robust covariance estimation. *IEEE Transactions on Signal Processing*, PP(99):1–1, 2015.

[75] I. Soloveychik and A. Wiesel. Performance analysis of Tyler's covariance estimator. *IEEE Transactions on Signal Processing*, 63(2):418–426, Jan 2015.

[76] S. Sra and R. Hosseini. Geometric optimization on positive definite matrices for elliptically contoured distributions. In *Advances in Neural Information Processing Systems*, pages 2562–2570, 2013.

[77] S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

[78] M. Srivastava, T. von Rosen, and D. von Rosen. Models with a Kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics*, 17:357–370, 2008.

[79] C. Stein. Estimation of a covariance matrix. *Rietz Lecture*, 1975.

[80] P. Stoica, L. Jian, Z. Xumin, and J.R. Guerci. On using a priori knowledge in space-time adaptive processing. *IEEE transactions on signal processing*, 56(6):2598–2602, 2008.

[81] Y. Sun, P. Babu, and D. P. Palomar. Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms. *IEEE Transactions on Signal Processing*, 62(19):5143–5156, 2014.

[82] Y. Sun, P. Babu, and D. P. Palomar. Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions. *IEEE Transactions on Signal Processing*, 63(12):3096–3109, 2015.

[83] T. Tsiligkaridis, A.O. Hero III, and S. Zhou. Convergence properties of kronecker graphical Lasso algorithms. *Arxiv preprint arXiv:1204.0585*, 2012.

[84] D. E. Tyler. Robustness and efficiency properties of scatter matrices. *Biometrika*, 70(2):411–420, 1983.

[85] D. E. Tyler. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 15(1):234–251, 1987.

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

[87] S. Vorobyov, A. B. Gershman, and Z. Q. Luo. Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem. *IEEE Transactions on Signal Processing*, 51(2):313–324, 2003.

[88] K. Werner, M. Jansson, and P. Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491, 2008.

[89] A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.

[90] A. Wiesel. Unified framework to regularized covariance estimation in scaled Gaussian models. *IEEE Transactions on Signal Processing*, 60(1):29 –38, 2012.

[91] J. H. Won, J. Lim, S. J. Kim, and B. Rajaratnam. Maximum likelihood covariance estimation with a condition number constraint. *Technical Report, Department of Statistics, Stanford University*, Aug. 2009.

[92] T. Zhang, X. Cheng, and A. Singer. Marchenko-Pastur law for Tyler's and Maronna's M-estimators. *arXiv preprint arXiv:1401.3424*, 2014.

[93] T. Zhang, A. Wiesel, and M. S. Greco. Multivariate generalized Gaussian distribution: Convexity and graphical models. *IEEE Transactions on Signal Processing*, 61(16):4141–4148, 2013.

[94] Y. Zhang and J. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. *Advances in Neural Information Processing Systems*, 2010.