Some flavours of PCA

Arnaud Breloy

Université Paris Nanterre
École Doctorale Connaissance, Langage et Modélisation

Habilitation à Diriger des Recherches

SOME FLAVOURS OF PCA

Arnaud BRELOY
LEME, EA 4416

Soutenue le 20 Novembre 2020, devant le jury composé de :

Rapporteurs :	Cédric Richard	-	PR Université Côte d'Azur
	Jean-Yves Tourneret	-	PR INP-ENSEEIHT
Examinateurs :	Olivier Besson	-	PR ISAE-SUPAERO
	Emilie Chouzenoux	-	CR Inria Saclay (OPIS)
	Nicolas Le Bihan	-	DR GIPSA-lab
Garant :	Mohammed Nabil El Korso	-	MCF Université Paris Nanterre

Contents

List of Figures

Curriculum vitae

Arnaud Breloy

31 years old (29/05/1989), French
Associate professor at University Paris Nanterre
Electrical Engineering department & LEME (EA4416) laboaraory

50, rue de Sèvres, 92410 Ville d'Avray
pers. : +33.06.74.51.65.34., pro. : +33.01.40.97.48.15.
abreloy@parisnanterre.fr, a.breloy@gmail.com
[parisnanterre.fr/arnaud-breloy–702124.kjsp]

Previous work experience

- *2015-2016* full-time **temporary assistant professor** (ATER)
  University Paris Nanterre, Electrical Engineering department
- *2012-2015* **Ph.D student** CNRS (DGA grant) and **Teaching Assistant** (Monitorat)
  SATIE (ENS-Cachan) and SONDRA (CentraleSupélec)
  University Paris Nanterre, Electrical Engineering department

Education

- *2015* **Ph.D of University Paris Saclay**
  Title: "*Estimation/detection algorithms in low-rank heterogeneous context*"
  Laboratories: SATIE (ENS-Cachan) and SONDRA (CentraleSupélec)
  Supervisors: Guillaume Ginolhac (director), Frédéric Pascal and Philippe Forster
  Jury :
  - ⋆ Pierre Comon, CNRS research director, GIPSA-Lab (Chairman)
  - ⋆ Olivier Besson, Professor, ISAE-SUPAERO (Reviewer)
  - ⋆ Pascal Chevalier, Professor, CNAM (Reviewer)
  - ⋆ Chin Yuan Chong, Research Engineer, DSO (Examinator)
  - ⋆ Guillaume Ginolhac, Professor, University Savoie-Mont-Blanc (Director)
  - ⋆ Frédéric Pascal, Professor, CentraleSupélec (Advisor)
  - ⋆ Philippe Forster, Professor, University Paris Nanterre (Advisor)
- *2013* **Engineer degree from Ecole Centrale Marseille (ECM)**
- *2012* **Master's degree of University Aix-Marseille**

Teaching experience

Degrees: I mostly teach for the first and second years of D.U.T. of the Electrical engineering (GEII) department (I.U.T. de Ville d'Avray). I also periodically participate to several courses for Master's degree (E2SC) and the FIPMECA formation (engineering degree).

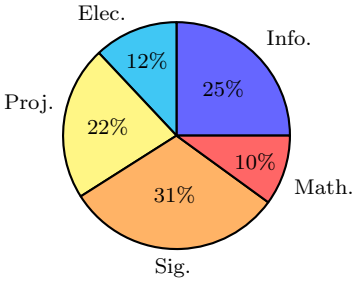
Courses: My courses consist mostly in practicals (TP) and supervised group work (TD). I also supervise several student projects with various formats: individual or group projects, either autonomous or fully supervised on a series of practicals.

Topics : My teachings concern analog electronics, programming, microcontrollers, and digital signal processing. Below is a list of the courses I was involved in

- **Analog Electronics** (1^{st} and 2^{nd} year D.U.T GEII).
- **C/C++/Matlab programming** (1^{st} and 2^{nd} year D.U.T GEII).
- **Mathematics of signal analysis** (1^{st} and 2^{nd} year D.U.T GEII).
- **Microcontrollers** (1^{st} year D.U.T GEII).
- **Digital signal processing** (2^{nd} year D.U.T GEII, FIPMECA, M2-E2SC).
- **Automation, Control theory** (2^{nd} year, D.U.T GEII).
- **Probability and statistics** (2^{nd} year D.U.T GEII, M2-E2SC).
- **Statistical signal processing** (M2-E2SC).
- **Students projects**
  - 1^{st} **year D.U.T. GEII** (semi-autonomous, $\sim 30h$)
    - Labyrinth challenge (robotC programming, LEGO-NXT)
    - Line tracking robot challenge (arduino programming)
  - 2^{nd} **year D.U.T. GEII** (autonomous, $\sim 4h$/week for 8 months)
    - GEII Robotics cup, organized by Cachan I.U.T..
    - BB-8 robot prototype.
  - **M2-E2SC and FIPMECA** (autonomous, $\sim 70h$)
    - Shazam algorithm.
    - "Audio beat tracking" (IEEE signal processing cup 2017)
    - Audio sources localization.

Synthesis: In order to shorten the exposition, my teaching hours are simply reported below:

Teaching service synthesis				
Year	CM	TD	TP	Total (eqTD)
2015-2016	26	55.5	229.5	324 h
2016-2017	0	79	148	227 h
2017-2018	4	108	168	282 h
2018-2019	8	115	201	328 h
2019-2020	0	64.5	179.5	244 h



Research themes

My research activities revolve around multivariate analysis, robust statistics, optimization, and elements of Riemannian geometry with applications to various signal processing and machine learning problems, notably in array processing and remote sensing. Stepping back, we can say that I am quite focused on the problem of subspace recovery and principal component analysis (PCA), tackled from various perspectives. These approaches can be summed as follows:

• **Probabilistic/Bayesian PCA**: My thesis work addressed the problem of signal subspace estimation from a (structured) mixture of elliptical distributions. For this problem, we proposed several algorithms with relaxed/robust variations. Following from this work, we extended the approach to a Bayesian setting in the thesis of R. Ben Abdallah. This lead to new subspace estimation algorithms accounting for a possible prior on the signal subspace distribution. The practical interest of all these PCA algorithms was illustrated in low-rank filtering and radar detection applications.

• **M-estimation**: When it comes to second order moment estimation, M-estimators of the scatter constitute a robust alternative to the traditional sample covariance matrix. The computation of these robust estimators is challenging when some structural prior is involved, as well as an insufficient sample support occurs. To overcome these issues, we proposed several regularized or structured estimation algorithms. Notably, we derived and studied the performance of 2-steps approaches for robust structured covriance matrix estimation in the thesis of B. Meriaux. In parallel, we also proposed new statistical characterizations for detection processes and PCA based on M-estimators in the thesis of G. Drašković.

• **Majorization-Minimization (MM) for subspace bases**: MM is an optimization technique that we leveraged extensively in the aforementioned works. In collaboration with Hong Kong University of Science and Technology (HKUST), we proposed a unified approach to MM for the Stiefel manifold (set of tall orthonormal matrices). This lead to a practical framework to deal with the constrained optimization of a large familly cost functions. This framework drove the formulation of a novel sparse PCA algorithm, combining a robust subspace recovery cost and sparsity promoting penalties. We currently explore extensions to other applications, as well as for optimization in distributed settings.

• **Riemannian geometry**: For subspace and covariance estimation problems, the Riemannian point of view offered interesting perspectives on to two complementary aspects. First, the so-called *intrinsic* Cramér-Rao theory allowed us to derive performance bounds on natural Riemannian distances for covariance estimation problems in elliptical distribution. This study was extended to derive a performance bound for subspace recovery problems (i.e., on the Grassmann manifold) in spiked elliptical models. Second, we proposed several algorithms for optimizing costs functions of low-rank structures (and structured structured covariance matrices) within the framework of Riemannian optimization.

• **Sparse subspace clustering in array processing**: We started to explore the use of robust robust subspace clustering algorithms after initiating a collaboration with North Carolina Sate University (NCSU). In this scope, we extended existing algorithms to include a dictionary of prior targets, and analyzed their use for radar detection applications.

• **Change detection in satellite image time series**: Local covariance matrix equality testing is a popular approach to detect changes in SAR image times series. In this scope, we derived a familly of new detectors to extend this approach to principal components, i.e. testing changes within low rank signal structures in covariance matrices. This approach, coupled with robust estimation was shown to yield improved change detection performance on real data (notably illustrated by the work of A. Mian).

For the sake of storytelling, only students and international collaborations were mentioned above. However, I cannot conclude this overview without thanking my regular co-authors, who also deserve a lot of credit for these works: G. Ginolhac, F. Pascal, P. Forster, J-P. Ovarlez, M.N. El Korso, C. Ren, A. Renaux, and F. Bouchard.

Supervision experience

Ph.D. Students (past)

• **Rayen Ben Abdallah** (co-supervisor, 35%), LEME, defended the 4/11/2019
Title: "*Statistical signal processing exploiting low-rank priors with applications to detection in Heterogeneous Environment*"
Supervision team: D. Lautru (director, 30%), M.N. El Korso (co-supervisor, 35%)
Funding: University Paris Nanterre, ED 139
Current situation: Post-doc, ISAE-Supaéro

• **Gordana Drašković** (co-supervisor, 30%), L2S, defended the 27/09/2019
Title : "*Robust estimation analysis for signal and image processing*"
Supervision team: F. Pascal (director, 40%), F. Tupin (co-supervisor, 30%)
Funding: DigiCosme grant, Paris-Saclay
Current situation : Artificial intelligence researcher engineer, Schlumberger

Ph.D. Students (current)

• **Bruno Meriaux** (co-supervisor, 30%), SONDRA, defense scheduled for October 2020
Title : "*Robust adaptive signal processing without secondary data*"
Supervision team: P. Forster (director, 10%), M.N. El Korso (co-supervisor, 25%), C. Ren (co-supervisor, 35%)
Funding: 1/2 DGA grant completed by SONDRA

• **Antoine Collas** (co-supervisor, 25%), SONDRA, defense expected for 2022
Title: "*Robust clustering for satellite image time series*"
Supervision team: J-P. Ovarlez (director, 25%), G. Ginolhac (co-director, 25%), C. Ren (co-supervisor, 25%)
Funding: SONDRA

Master's degree students

• **Taha Essalih** (co-supervisor, 35%), April-September 2017
Internship of Ecole Centrale Marseille Master's degree
Topic: "Robust calibration of large radio-interferometers"
Supervision team: M.N. El Korso (35%), Rémi Flamary (15%), Franck Iutzeler (15%)

• **Bruno Meriaux** (co-supervisor, 30%), April-September 2017
Internship for ENS Paris Saclay Master's degree
Topic : "Robust estimation of structured scatter matrices"
Supervision team: P. Forster (10%), M.N. El Korso (25%), C. Ren (35%)

• **Hugo Brehier** (co-supervisor, 70%), March-September 2020
Internship for ENSAI Master's degree
Topic : "Robust sparse PCA"
Supervision team: M.N. El Korso (30%)

Research conferences activities

Tutorials

• **IEEE RadarConf 2020**: "Robust statistical framework for radar change detection applications". with G. Ginolhac.
• **EUSIPCO 2018**: "Robust Covariance and Subspace Learning: dealing with high-dimensionality and small sample support". with F. Pascal and G. Ginolhac.

Special sessions organization and invitations

• **EUSIPCO 2020**: "Recent advances in differential geometry for signal and image processing"
• **EUSIPCO 2018**: "Emerging Data Structure Paradigms for Subspace Estimation"
• 6 invited conference papers in special sessions.

Committees

• **NCMIP 2019**: Scientific committee member
• **EUSIPCO 2018, 2020**: TPC member

Reviewing service

• **International journals** : IEEE Trans. on Signal Processing, IEEE Signal Processing letters, IEEE Trans. on Aerospace And Electronic Systems, Signal Processing ELSEVIER, Digital Signal Processing ELSEVIER, EURASIP Journal on Advances in Signal Processing.
• **International conferences** : EUSIPCO, SSP, ISIT, IEEE RadarConf.

International collaborations

• **HKUST, Hong Kong** : invited in the group of Prof. Daniel P. Palomar (Department of Electronic and Computer Engineering of HKUST) the month of July 2015. This invited stay resulted in the publication of one journal and three conferences articles:

> [HK1] Y. Sun, A. Breloy, P. Babu, D.P. Palomar, F. Pascal, G. Ginolhac, "Low-Complexity Algorithms for Low Rank Clutter Parameters Estimation in Radar Systems," IEEE Transactions on Signal Processing, vol. 64, no. 8, pp. 1986-1998, 2016.
>
> [HK2] A. Breloy, Y. Sun, P. Babu, G. Ginolhac, D.P. Palomar, "Robust Rank Constrained Kronecker Covariance Matrix Estimation," Asilomar Conf. 2016.
>
> [HK3] A. Breloy, Y. Sun, P. Babu, D.P. Palomar, "Low-Complexity Algorithms for Low Rank Clutter Parameters Estimation in Radar Systems," EUSIPCO, 2016.
>
> [HK4] A. Breloy, Y. Sun, P. Babu, D.P. Palomar, F. Pascal, G. Ginolhac, "A robust signal subspace estimator," SSP 2016.

This collaboration is still active, and we submitted a journal article in 2020:

> [HK5] A. Breloy, S. Kumar, Y. Sun, D.P. Palomar, "Majorization-Minimization on the Stiefel Manifold with application to Robust Sparse PCA," submitted.

• **NC State University (NCSU), Raleigh, USA** : invited in the group of Prof. Hamid Krim (Electrical and Computer Engineering Department of NCSU) for two weeks in October 2016. This invited stay resulted in the publication of a conference article:

[US1] A. Breloy, M.N. El Korso, A. Panahi, H. Krim, "Robust Subspace Clustering for Radar Detection," EUSIPCO 2018.

as well as the organization of a special session at the conference EUSIPCO 2018.

• **Aalto University, Helsinki, Finlande** : invited by Prof. Esa Ollila for the first "Finnish-French Statistical Learning for Signal Processing Mini-workshop" in December 2018. This workshop initiated a collaboration that resulted in the publication of a conference article:

[FI1] A. Breloy, E. Ollila, F. Pascal, "Spectral Shrinkage of Tyler's M-Estimator of Covariance Matrix", IEEE CAMSAP 2019

This collaboration is still ongoing and was the starting point of a CRCT project (2021 campaign) for a 6 month mobility at Aalto University.

Research projects and grants

1/2 CRCT	**PI**	**2020-21**	
Semester of remunerated sabbatical leave for research purpose (french CRCT) awarded by National Council of Universities (CNU). Invited stay of 6 months at Aalto University, Helsinki.			

PNTS	**Collaborator**	**2019-20**	**15k€**
PNTS-2019-11: Statistical learning in SAR image time series with missing data PI : Y. Yan (PI)			

PHC-PROCORE	**Co-PI**	**2019-20**	**14k€**
Project members : M.N. El Korso (PI-fr), M. Pesavento (PI-ger), C. Ren, P. Forster.			

DGA 1/2 thesis grant	**Co-PI**	**2017-20**	**50k€**
Thesis of Bruno Meriaux, "*Robust adaptive signal processing without secondary data*" Project members: P. Forster (PI), M.N. El Korso, C. Ren.			

ANR-ASTRID	**Sub-tasks manager**	**2017-22**	**78k€**
MARGARITA (Modern Adaptive Radar), Ref. ANR-17-ASTR-0015 [link1] [link2] PI: Guillaume Ginolhac.			

Digiteo-DigiCosme grant	**Co-PI**	**2016-19**	**100k€**
Thesis of Gordana Drašković, "*Robust estimation analysis for signal and image processing* Project members : F. Pascal (PI), F. Tupin.			

Young researcher GDR-ISIS	**Co-PI**	**2016-18**	**7k€**
Project ON FIRE (robust calibration of future large interferometers) Project members : M.N. El Korso (PI), Rémi Flamary, Franck Iutzeler.			

Publications

My publication record can be summed up by the following table:

Format	Total
Journals	12
International conferences	23
French conferences	13

Below is reported the list of corresponding articles. Underlined co-authors correspond to interns/Ph.D. students that I supervised (with attested percentage). *Underlined-italics* co-authors correspond to interns/Ph.D. students I worked with, without attributed supervision percentage. For these co-authors, an eventual switch back to the standard typography indicates that the work was conducted after their graduation.

Journal publications

Ongoing (submitted):

[X6] A. Collas, F. Bouchard, **A. Breloy**, G. Ginolhac, C. Ren, J.P. Ovarlez, "A Riemannian Geometry for Probabilistic PCA with Compound Gaussian Signals"

[X5] B. Mériaux, C. Ren, **A. Breloy**, M. N. El Korso, P. Forster, "Mismatched Robust Estimation of Kronecker Product of Linearly Structured Scatter Matrices"

[X4] **A. Breloy**, , G Ginolhac, Y. Gao, F. Pascal, "MIMO Filters based on Robust Rank-Constrained Kronecker Covariance Matrix Estimation"

[X3] F. Bouchard, **A. Breloy**, G. Ginolhac, A. Renaux, F. Pascal, "A Riemannian Framework for Low-Rank Structured Elliptical Models," [arxiv.org/abs/2001.01141]

[X2] A. Mian, G. Ginolhac, J.P. Ovarlez, **A. Breloy**, F. Pascal, "An overview of covariance-based change detection methodologies in multivariate SAR image time series," Chapter 2 from the book "Change Detection and Image Time-Series Analysis", ISTE WILEY, 2020.

[X1] **A. Breloy**, S. Kumar, Y. Sun, D.P. Palomar, "Majorization-Minimization on the Stiefel Manifold with application to Robust Sparse PCA"

Post-thesis publications :

[J12] *A. Mian*, A. Collas, **A. Breloy**, G. Ginolhac, J-P. Ovarlez, "Robust Low-rank Change Detection for Multivariate SAR Image Time Series," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, to appear.

[J11] *A. Bouiba*, M. N. El Korso, **A. Breloy**, P. Forster, M. Hamadouche, M. Lagha, "Two dimensional robust source localization under non-Gaussian noise," in Circuits, Systems & Signal Processing, Springer, pp. 1-22., 2020.
doi: [10.1007/s00034-020-01381-2]

[J10] R. Ben Abdallah, **A. Breloy**, M. N. El Korso, D. Lautru, "Bayesian Signal Subspace Estimation with Compound Gaussian Sources," in Signal Processing, vol. 167, 2020.
doi: [10.1016/j.sigpro.2019.107310]

[J9] G. Draškovic, **A. Breloy**, F. Pascal, "On the performance of robust plug-in detectors using M-estimators," in Signal Processing Journal, Elsevier, vol. 167, 2020
doi: [10.1016/j.sigpro.2019.107282]

[J8] G. Draškovic, **A. Breloy**, F. Pascal, "On the asymptotics of Maronna's robust PCA" in IEEE Transactions on Signal Processing, vol. 67, no. 19, pp. 4964-4975, 2019.
doi: [10.1109/TSP.2019.2932877]

[J7] B. Mériaux, C. Ren, M. N. El Korso, **A. Breloy**, P. Forster, "Robust estimation of structured scatter matrices in (mis)matched models," in Signal Processing, vol. 165, pp. 163-174, Dec 2019.

doi: [10.1016/j.sigpro.2019.06.030]

[J6] R. Ben Abdallah, *A. Mian*, **A. Breloy**, A. Taylor, M. N. El Korso, D. Lautru, "Detection Methods Based on Structured Covariance Matrices for Multivariate SAR Images Processing," in IEEE Geoscience and Remote Sensing Letters, vol. 16 , no. 7, pp. 1160-1164, Jul 2019.
doi: [10.1109/LGRS.2018.2890155]

[J5] B. Mériaux, C. Ren, M. N. El Korso, **A. Breloy**, P. Forster, "Asymptotic Performance of Complex M-Estimators for Multivariate Location and Scatter Estimation," in IEEE Signal Processing Letters, vol. 26, no. 2, pp. 367-371, Feb. 2019.
doi: [10.1109/LSP.2019.2891201].

[J4] **A. Breloy**, G. Ginolhac, A. Renaux, *F. Bouchard*, "Intrinsic Cramér–Rao Bounds for Scatter and Shape Matrices Estimation in CES Distributions," in IEEE Signal Processing Letters, vol. 26, no. 2, pp. 262-266, Feb. 2019.
doi: [10.1109/LSP.2018.2886700]

Thesis publications:

[J3] **A. Breloy**, G. Ginolhac, F. Pascal, P. Forster, "Robust Covariance Matrix estimation in Low-Rank Heterogeneous Context", IEEE Transactions on Signal Processing, vol. 64, no. 22, pp. 5794-5806, Nov.15, 15 2016.
doi : [10.1109/TSP.2016.2599494]

[J2] Y. Sun, **A. Breloy**, P. Babu, D.P. Palomar, F. Pascal, G. Ginolhac, "Low-Complexity Algorithms for Low Rank Clutter Parameters Estimation in Radar Systems", IEEE Transactions on Signal Processing, vol. 64, no. 8, pp. 1986-1998, 2016.
doi: [10.1109/TSP.2015.2512535]

[J1] **A. Breloy**, G. Ginolhac, F. Pascal, P. Forster, "Clutter Subspace Estimation in Low Rank Heterogeneous Noise Context," in IEEE Transactions on Signal Processing, vol. 63, no. 9, pp. 2173-2182, 2015.
doi: [10.1109/TSP.2015.2403284].

International conferences

Post-thesis publications:

[C23] F. Bouchard, **A. Breloy**, G. Ginolhac, F. Pascal, "Riemannian framework for robust covariance matrix estimation in spiked models," IEEE ICASSP 2020

[C22] F. Bouchard, **A. Breloy**, A. Renaux, G. Ginolhac, "Riemannian geometry and Cramér-Rao bound for blind separation of Gaussian sources," IEEE ICASSP 2020.

[C21] **A. Breloy**, E. Ollila, F. Pascal, "Spectral Shrinkage of Tyler's M-Estimator of Covariance Matrix," IEEE CAMSAP 2019.

[C20] B. Mériaux, C. Ren, **A. Breloy**, M. N. El Korso, P. Forster, "Modified Sparse Subspace Clustering for Radar Detection in Non-Stationary Clutter," IEEE CAMSAP 2019.

[C19] R. Ben Abdallah, A. Breloy, M. N. El Korso, D. Lautru, "Bayesian Robust Signal Subspace Estimation in Non-Gaussian Environment," EUSIPCO 2019.

[C18] R. Ben Abdallah, **A. Breloy**, A. Taylor, M. N. El Korso, D. Lautru, "Signal Subspace Change Detection in Structured Covariance Matrices," EUSIPCO 2019.

[C17] B. Mériaux, C. Ren, **A. Breloy**, M. N. El Korso, P. Forster, J.-P. Ovarlez, "On the Recursions of Robust COMET Algorithm for Convexly Structured Shape Matrix," EUSIPCO 2019.

[C16] *A. Mian*, **A. Breloy**, G. Ginolhac, J-P. Ovarlez, "Robust Low-rank Change Detection for SAR Image Time Series," IEEE IGARSS 2019.

[C15] B. Mériaux, C. Ren, M.N. El Korso, **A. Breloy**, P. Forster, "Efficient Estimation of Scatter Matrix with Convex Structure under t-distribution", IEEE ICASSP 2018.

[C14] **A. Breloy**, M. N. El Korso, A. Panahi, H. Krim, "Robust Subspace Clustering for Radar Detection," EUSIPCO 2018.

[C13] B. Meriaux, C. Ren, M. N. El Korso, **A. Breloy**, P. Forster, "Robust-COMET for Covariance Estimation in Convex Structures: Algorithm and Statistical Properties," IEEE CAMSAP 2017.

[C12] R. Ben Abdallah, **A. Breloy**, M. N. El Korso, D. Lautru, H. Ouslimani, "Minimum Mean Square Distance Estimation of Subspaces in presence of Gaussian sources with application to STAP detection", International Conference on New Computational Methods for Inverse Problems (NCMIP), IOP publishing in the series "Journal of Physics : Conference Series," 2017.

[C11] *Q. Hoarau*, **A. Breloy**, G. Ginolhac, A.M. Atto, J.M. Nicolas, "A subspace approach for shrinkage parameter selection in undersampled configuration for regularized Tyler estimators," IEEE ICASSP 2017.

[C10] G. Drašković, F. Pascal, **A. Breloy**, J-Y. Tourneret, "New asymptotic properties for the Robust ANMF," IEEE ICASSP 2017.

[C9] T. Bao, **A. Breloy**, M.N. El Korso, K. Abed-Meraim, H.H. Ouslimani, "Performance analysis of direction-of-arrival and polarization estimation using a non-uniform linear COLD array," Seminar on Detection Systems: Architectures and Technologies 2017.

[C8] **A. Breloy**, Y. Sun, P. Babu, G. Ginolhac, D.P. Palomar, "Robust Rank Constrained Kronecker Covariance Matrix Estimation," IEEE Asilomar Conference on Signals, Systems, and Computers 2016.

Thesis publications:

[C7] **A. Breloy**, Y. Sun, P. Babu, D.P. Palomar, F. Pascal, G. Ginolhac, "A robust signal subspace estimator," IEEE Workshop on Statistical Signal Processing 2016.

[C6] **A. Breloy**, Y. Sun, P. Babu, D.P. Palomar, "Low-Complexity Algorithms for Low Rank Clutter Parameters Estimation in Radar Systems," EUSIPCO, 2016.

[C5] J-P. Ovarlez, F. Pascal, **A. Breloy**, "Asymptotic Detection Performance Analysis of the Robust Adaptive Normalized Matched Filter," IEEE CAMSAP 2015.

[C4] **A. Breloy**, G. Ginolhac, F. Pascal, P. Forster, "Robust estimation of the clutter subspace for a low rank heterogeneous noise under high clutter to noise ratio assumption," IEEE ICASSP 2014.

[C3] **A. Breloy**, G. Ginolhac, F. Pascal, P. Forster, "CFAR property and robustness of the low rank adaptive normalized matched filters detectors in low rank compound Gaussian context," IEEE SAM 2014.

[C2] **A. Breloy**, L. Le Magoarou, G. Ginolhac, F. Pascal, P. Forster, "Numerical performances of low rank STAP based on different heterogeneous clutter subspace estimators," International RADAR Conf. 2014.

[C1] **A. Breloy**, L. Le Magoarou, G. Ginolhac, F. Pascal , P. Forster, "Maximum likelihood estimation of clutter subspace in non-homogeneous noise context," EUSIPCO 2013.

French conferences

Post-Thesis publications:

[FC13] R. Ben Abdallah, **A. Breloy**, M. N. El Korso, D. Lautru, "Détection de changement de sous-espace signal dans des matrices de covariance structurées," GRETSI 2019.

[FC12] F. Bouchard, **A. Breloy**, A. Renaux, G. Ginolhac, "Bornes de Cramér-Rao Intrinsèques pour l'estimation de la matrice de dispersion normalisée dans les distributions elliptiques," GRETSI 2019.

[FC11] G. Drašković, A. Breloy, Frédéric Pascal, "Caractérisations asymptotiques pour les composantes principales des M-estimateurs," GRETSI 2019.

[FC10] B. Mériaux, C. Ren, M.N. El Korso, **A. Breloy**, P. Forster, "Estimation robuste de matrices de dispersion structurées pour des modèles bien/mal spécifiés," GRETSI 2019.

[FC9] B. Mériaux, C. Ren, **A. Breloy**, M.N. El Korso, P. Forster, J.-P. Ovarlez, "Une version récursive de RCOMET pour l'estimation robuste de matrices de forme à structure convexe," GRETSI 2019.

[FC8] *A. Mian*, **A. Breloy**, G. Ginolhac, J-P. Ovarlez, "Détection de Changement Robuste en Rang Faible pour les Séries Temporelles d'Images SAR," GRETSI 2019.

[FC7] G. Drašković, F. Pascal, **A. Breloy**, J-Y. Tourneret, "Nouvelles propriétés asymptotiques de détecteurs robustes," GRETSI 2017.

[FC6] *Q. Hoarau*, **A. Breloy**, G. Ginolhac, A. Atto, J-M. Nicolas, "Estimateur de Tyler régularisé dans le cas sous-déterminé. Application à la détection d'objets enfouis," GRETSI 2017.

[FC5] A. Taylor, **A. Breloy**, M. N. El-Korso, "Détection d'anomalie de composantes principales pour des cibles mobiles étendues en SAR," GRETSI 2017.

[FC4] R. Ben Abdallah, **A. Breloy**, M. N. El Korso, D. Lautru, H. Ouslimani, "Estimation de sous-espaces en présence de sources gaussiennes avec application à la détection STAP," GRETSI 2017.

[FC3] **A. Breloy**, A. Renaux, G. Ginolhac, *F. Bouchard*, "Borne de Cramér-Rao intrinsèque pour la matrice de covariance des distributions elliptiques complexes," GRETSI 2017.

Thesis publications:

[FC2] A. Breloy, G. Ginolhac, F. Pascal, P. Forster, "Estimation Robuste de la Matrice de Covariance en contexte Hétérogène Rang Faible," GRETSI 2015.

[FC1] A. Breloy, L. Le Magoarou, G. Ginolhac, F. Pascal, P. Forster, "Estimation par maximum de vraisemblance du sous-espace clutter dans un bruit hétérogène rang faible avec application au STAP," GRETSI 2013.

1 | Introduction

Contents

1.1 The subspace recovery problem

In many data sets, the relevant information often lies in a subspace of much lower dimension than the ambient observation space. Thus, the goal of many learning algorithms can be broadly interpreted as trying to find, or exploit, this underlying structure. In this scope, a recurring idea is that the whole data can generally be projected with no (or minimal) loss on a linear subspace of low-dimension. This principle is illustrated in Figure 1.1, where we can grasp that (most of) the 3D data information can simply be contained in the represented 2D plan.

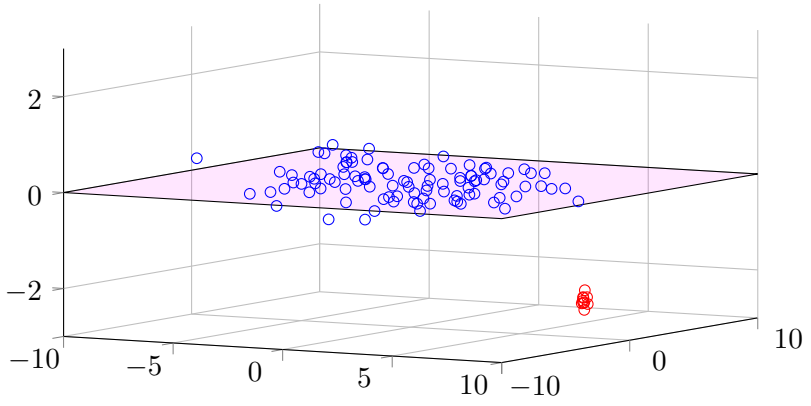


Figure 1.1: Illustration of data contained in a low-dimensional subspace (outliers represented in red).

The issue of estimating this unknown subspace directly from the data will be referred to as the *subspace recovery* problem. The most celebrated solution for this problem is the principal component

analysis (PCA), which is briefly reviewed below. Nevertheless, the subspace recovery problem still raises some challenges and theoretical questions, that will be discussed afterward.

1.1.1 From principal component analysis...

An insightful overview of PCA can be found in [Jolliffe, 1986], which constitutes a solid reference on the topic. This short introduction is inspired by the two PCA's historical roots, with a personal take on the reformulation of the problems so that the document feels consistent.

The seminal idea originates from [Pearson, 1901], which proposed to find the subspace that minimizes the sum of the residual errors on the centered (or demeaned) samples $\{\mathbf{z}_i\}_{i=1}^n$ (with $\mathbf{z}_i \in \mathbb{C}^p \ \forall i \in [\![1, n]\!]$) after projection. A k-dimensional subspace can be represented in many ways, but we will focus on the use of an orthonormal basis $\mathbf{U} \in \mathrm{St}(p, k)$, where $\mathrm{St}(p, k)$ denotes the Stiefel manifold:

$$\mathrm{St}(p, k) = \{\mathbf{U} \in \mathbb{C}^{p \times k} \mid \mathbf{U}^H\mathbf{U} = \mathbf{I}\}. \tag{1.1}$$

Defining the projection error as $\mathrm{d}^2(\mathbf{U}, \mathbf{z}) = ||(\mathbf{I} - \mathbf{U}\mathbf{U}^H)\mathbf{z}||_F^2$, the expressed problem can then be straightforwardly rewritten as

$$\underset{\mathbf{U} \in \mathrm{St}(p,k)}{\text{maximize}} \quad \mathrm{Tr}\{\mathbf{U}^H\mathbf{S}\mathbf{U}\}, \tag{1.2}$$

with $\mathbf{S} = \sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^H$. A solution to this problem given by the k leftmost eigenvectors of $\mathbf{S}$ (where leftmost stands for "associated with the largest eigenvalues"). Note that this solution can also be obtained with the k leftmost singular vectors of the singular value decomposition (SVD) of the data matrix $\mathbf{Z} = [\mathbf{z}_1, \ \ldots, \ \mathbf{z}_n]$, which is why we often rely on the shortcut "PCA = SVD".

This concept was then mainly re-developed by [Hotelling, 1933], who proposed to find the subspace in which the projected data has maximal variance. Assuming that the covariance matrix[1] of the data is known, and denoted $\mathbf{\Sigma} = \mathbb{E}[\mathbf{z}\mathbf{z}^H]$, this problem is expressed

$$\underset{\mathbf{U} \in \mathrm{St}(p,k)}{\text{maximize}} \quad \mathrm{Tr}\{\mathbf{U}^H\mathbf{\Sigma}\mathbf{U}\}. \tag{1.3}$$

Of course $\mathbf{\Sigma}$ is generally unknown in practice, but it can be estimated with the sample covariance matrix, i.e. $\hat{\mathbf{\Sigma}}_{\mathrm{SCM}} = \mathbf{S}/n$. Plugging $\hat{\mathbf{\Sigma}}_{\mathrm{SCM}}$ instead of $\mathbf{\Sigma}$ in (1.3) then yields the same solution as previously. Nevertheless, the leap from geometrical to statistical interpretation offers interesting perspectives. A main example is that if the columns of $\mathbf{U}$ correspond to eigenvectors of the covariance matrix, the representation $\tilde{\mathbf{z}}_i = \mathbf{U}^H\mathbf{z}_i$ produces uncorrelated entries. This aspect, coupled with the dimension reduction, can be extremely useful in terms of interpretation/analysis of the data.

1.1.2 ... to current challenges

Unfortunately, the SVD of the data matrix does not solve every aspect of the subspace recovery problem. This is especially true for modern data sets, that gather complex information for which a single linear subspace may be too restrictive. Some of these current challenges are discussed below:

• **Robustness**: A common problem in many applications is the presence of outliers in the dataset (e.g., as seen in red in Figure 1.1). These outliers can cause a subspace swap phenomenon [Thomas et al., 1995], meaning that the standard PCA does not recover the relevant signal subspace accurately. The issue is conceptually complex to tackle, as solving it requires to characterize what is an outlier, which implies knowing the subspace we aim to recover in the first place. Various approaches exist to address the problem of robustness to outliers in subspace recovery [Lerman and Maunu, 2018]. We can notably mention: redefining the cost of "best fitting" in PCA [De La Torre and Black, 2003, Lerman and Maunu, 2017], using of robust statistics to estimate the covariance matrix [Croux and Haesbroeck, 2000, Drašković et al., 2019],

[1] The initial work rather uses the correlation matrix, but both approaches have their merits, as discussed in [Jolliffe, 1986, Sec. 2.3].

or expressing of the problem as a low-rank plus sparse decomposition of the data matrix [Chandrasekaran et al., 2011, Candès et al., 2011]. Each of these solution has their own merits, depending on the context, meaning that other options still remain to explore. Some interesting leads can also come from formulation of recovery problems with alternate representation of a subspace in order to overcome limitations of existing methods. An example would be sparse subspace clustering, where the subspaces are recovered through the factors of linear combinations linking the samples [Elhamifar and Vidal, 2009].

• **Structure and priors**: Some applications can benefit from available prior information on the subspace to be estimated (e.g., a rough knowledge from previous estimations). Some structural information on the covariance matrix can also be obtained from physical considerations on the observed phenomenon (e.g., Toeplitz for signals measured with uniform linear arrays) [Forster, 2001]. In some other cases, a sparse structure can be desired within the principal components in order to simplify the variable selection and the statistical analysis [Hu et al., 2016]. Several approaches exist to leverage these priors: Bayesian estimation [Besson et al., 2011], regularization penalties [Benidis et al., 2016], structured parameterizations [Sun et al., 2016], etc., each with own benefits and associated bottlenecks. As for the robustness issue, the challenge lies in the formulation of a (solvable) estimation problem that both promote the desired structures, while ensuring good performance for the process.

• **Optimization**: Once a new relevant optimization problem is formulated to respond to the two aforementioned issues, it generally does not have an analytic solution as for the standard PCA. Thus, the actual computation of such newly formulated subspace estimators is also at stake. Indeed, subspaces are generally represented by objects with complex constraints, such as the orthonormality of a basis, or the rank-deficiency of a matrix. The subspace recovery problem hence motivates the development of efficient constrained optimization methods, for which we can, e.g., mention Riemannian optimization [Absil et al., 2009], and majorization-minimization techniques [Sun et al., 2016].

• **Performance analysis**: Once a subspace recovery problem is cast within the prism of a statistical model for the data, it is possible to theoretically study the ultimate estimation performance (Cramér-Rao bounds) [Kay, 1993] or the statistical characterization of an estimator [Krim et al., 1992]. These derivations are generally not trivial, either because the considered model is complex, or because the studied estimator is not obtained in closed-form, on top of satisfying inherent constraints that should be acknowledged in the analysis [Gorman and Hero, 1990]. Additionally, we can question the performance criterion that should be used when dealing with subspaces represented by structured parameters. An interesting option is to turn to the natural Riemannian distance between two subspaces, which can take into account the aforementioned characteristics, as well as provide insightful results. However, this approach generally implies technical difficulties in the analysis [Smith, 2005].

It goes without saying that this list is set as a motivation for the following chapters, but is not meant to be exhaustive. There exists many other challenges related to subspace recovery, such as: developing efficient on-line [Feng et al., 2013] or distributed methods [Huroyan and Lerman, 2018], performing dimension reduction with non-linear representations [Pennec et al., 2018], recovering a mixture of subspaces [Vidal, 2011], reducing the dimension of multilinear data [Lu et al., 2016]... Some of these issues will be evoked as perspectives in the concluding chapter of this document.

1.2 Manuscript organization and contributions

This section presents the outline of the rest of this manuscript. Each chapter is centered around a theme and was aimed to be readable as a stand alone. Nevertheless, there are of course some overlaps and connections between these chapters (common motivation, models/formulations, or optimization techniques). These links are made as explicit as possible in the "big picture" presented at the end of this section.

Chapter 2: Bayesian PCA with Compound Gaussian signals

This chapter introduces the model of a structured mixture of compound Gaussian distributions. This statistical model appears as an interesting alternative to the multivariate Gaussian one, used in standard probabilistic PCA. It notably allows to account for signals with fluctuating power, and possibly outliers. We then present a family of distributions for orthonormal bases, that can be used as Bayesian priors for the signal subspace. We then discuss the derivation of estimation processes for the corresponding models.

Related publications in the cv (page xi)

- Non-Bayesian: [J1-J3], [C1-C5], [C7], [FC1, FC2].
- Bayesian models: [J10], [C12], [C19], [FC4].

Chapter 3: Robust covariance matrix estimation in elliptical models

This chapter presents the general framework of elliptical distributions (a family that encompasses the compound Gaussians) and associated M-estimators of the scatter. First, we present new asymptotic characterizations for PCA built with M-estimators and samples following an elliptical distribution. Second, we discuss robust structured covariance matrix estimation methods adapted to this context: a new algorithm is presented and statistically characterized. In a third part, *intrinsic* (manifold oriented) Cramér-Rao bounds are derived, notably for covariance matrix estimation problems, as well as subspace recovery problems in spiked (low-rank structured) elliptical models.

Related publications in the cv (page xi)

- Performance analysis [J5, J8, J9], [C10], [FC7, FC11].
- Structured covariance matrix estimation [J7], [C13, C15, C17], [FC9, FC10].
- Intrinsic Cramér-Rao bounds [J4], [C22], [X3].

Chapter 4: Change detection in satellite image time series

Classical statistical change detection methodologies based on covariance matrix analysis are usually built upon the (unstructured) Gaussian assumption. In order to refine this approach, we discuss how the aforementioned PCA models (compound Gaussian mixture and structured elliptical) can be integrated in the formulation of change detection tests, and how to compute the corresponding detectors. The idea is then applied to change detection in multivariate synthetic aperture radar image time series.

Related publications in the cv (page xi)

- Change detection [J6, J12], [C16, C18], [FC8, FC13], [X2].

Chapter 5: Geometric approaches for subspace recovery

This chapter presents some of my recent works, leaping from "statistically" to "geometrically" oriented formulations for subspace recovery. These approaches consist in designing optimization problems that are not necessarily linked to a statistical distribution, but rather geometric insights and functions/constraints that promote certain structures on the solution. First, we present a new class of sparse PCA algorithms, for which the objective function is composed of an M-estimation type subspace fitting term plus a regularizer that promotes sparsity in the principal components. Second, we present a prospective reformulation of radar detection problem as a robust subspace clustering one (i.e. recovering multiple linear subspaces from a heterogeneous data set.).

Related publications in the cv (page x)

- Sparse PCA [X1].
- Subspace clustering [C14], [C20].

Chapter 6: Perspectives

This chapter concludes by drawing some research perspectives brought by the presented work.

Appendix A: Majorization-Minimization on the Stiefel manifold

Along this manuscript, it will often be mentioned that the occurring optimization problems can be solved by tailoring dedicated majorization-minimization algorithms. The detailed derivation of these algorithms will be voluntarily omitted in order to lighten the presentation. Nevertheless, to fill this gap, this appendix presents a synthetic and generic framework for deriving majorization-minimization algorithm for a variable $\mathbf{U} \in \mathrm{St}(p, k)$, which constitutes a major part of the considered subspace recovery problems. For other specific problems (side parameters estimation), the full derivations can be found in the auxiliary annexes.

Annexes (auxiliary document)

The annexes consist in a side document, gathering a selection of my articles. It is not meant to be extensively studied, as this main document already presents their core results in a synthetic manner. However, interested readers can find here most of the technical details (algorithm derivations, proofs, etc.) that were omitted in this synthesis.

Links with Ph.D. students and international collaborations

The following table details the implication of Ph.D. students and international collaborators in the contributions of each chapter:

Chapter	Ben Abdallah	Drašković	Mériaux	Mian	Bouchard	Intl. collab.
2	✓					✓
3		✓	✓		✓	
4	✓			✓		
5			✓			✓

We also mention the following clarifications:
• Ph.D. students I officially supervised: R. Ben Abdallah, G. Drašković, B. Mériaux.
• Ph.D. students I worked with: A. Mian, F. Bouchard (now post-doc in our team).
• International collaborations: HKUST (Honk Kong), NCSU (USA).

Big pictures

Since the section 1.1.2 was set as a motivation, we can report on which of the aforementioned challenges the presented contributions bring an element of response to:

Links with challenges of section 1.1.2				
Chapter	Robustness	Structure and priors	Optimization	Performance analysis
2	✓	✓	∼	
3	✓	✓	∼	✓
4	✓	✓		
5	✓	✓	✓	

We can add several remarks to this table:

- The word "robustness" appears in each chapter, however it can refer to various meanings that will be specified. The three main examples are: to outliers, to various underlying distributions and model mismatches, to heterogeneity in the data set.
- Some contributions in Chapter 2-4 involve the development of majorization-minimization algorithms. However, these chapters rather focus on the problem formulation. For more details on majorization-minimization, the reader will be referred to the first part of chapter 5. Some of my recent work addresses Riemannian optimization algorithms. These are discussed in the perspectives of chapter 3, but not presented in details.
- Chapter 3 focuses on elliptically distributed samples, while Chapter 2 uses structured mixtures of compound Gaussian distributions. Chapter 4 discusses both models in the context of a specific application, however the presented contribution focuses on the first one.

To conclude this synthesis, Figure 1.2 presents the global picture, i.e. details the topics (and corresponding chapters) evoked in this manuscript.

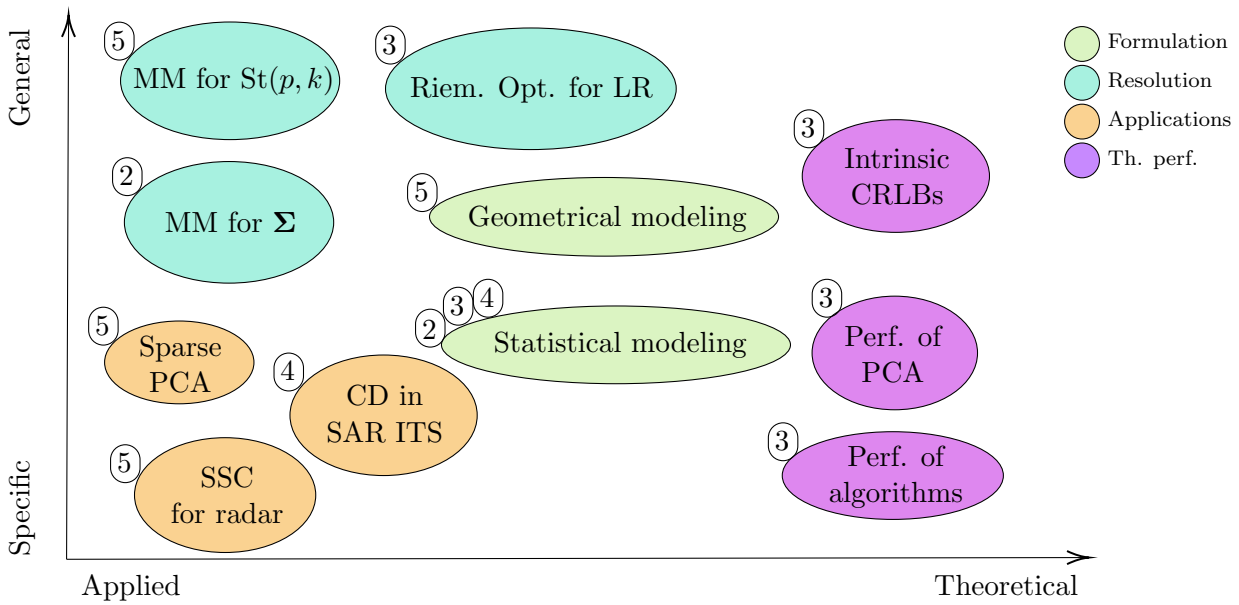


Figure 1.2: Overview of the topics considered in this manuscript, with corresponding chapters.

1.3 Can we recover a subspace from all this?

Finally, it is hard not to reduce the dimension of this work... Consider the following data matrix:

	[J1]	[J2]	[J3]	[J4]	[J5]	[J6]	[J7]	[J8]	[J9]	[J10]	[X1]	[X2]	[X3]	[C1]	[C4]	[C7]	[C8]	[C9]	[C10]	[C11]	[C14]	[C17]	[C18]	[C19]	[C20]	[C21]	[C22]	[C23]
Sta	✓	✓	✓			✓		✓		✓		✓	✓	✓	✓	✓							✓	✓				✓
Geo		✓									✓										✓				✓			
MM		✓				✓				✓	✓						✓						✓			✓		
Rie				✓									✓														✓	✓
CRB				✓			✓						✓					✓									✓	
Perf					✓		✓	✓	✓										✓			✓						
Bay										✓														✓				
Reg										✓	✓									✓						✓		
Cov	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓		✓	✓		✓				✓	✓	✓
App		✓	✓			✓			✓	✓	✓	✓					✓	✓	✓		✓	✓	✓		✓		✓	
Rob			✓		✓		✓	✓	✓		✓	✓	✓		✓	✓	✓		✓		✓	✓		✓	✓	✓		✓

where each sample represents a paper shortlisted from page xi (28 in total), and measurements consists in a list of 11 yes-or-no (0 or 1) keywords: statistical PCA, geometric PCA, MM, Riemannian geometry, CRLB, performance analysis, Bayesian, regularization, covariance, application, robustness.

PCA is applied on the correlation matrix of the data (which appeared more relevant for binary entries): the left side of Figure 1.3 represents the data points projected on the subspace spanned by the 3 leading principal components, containing 54% of the information. From this representation, we can interpret 3 clusters: *a*) the group [X1, C14, C20], which gather recent works, and illustrates a shift from statistical to sparse/geometrical subspace representations; *b*) the group [J4, C9, C22], which focus only on Cramér-Rao bounds derivations; *c*) the rest, which is mostly about statistics-based approaches (either covariance matrix estimation or probabilistic PCA).

The right side of Figure 1.3 displays 2D representation on the data projected onto the 2 leading principal components, for which we can give the following interpretation: The component $\mathbf{u}_1$ separates [J10, X1, C18] (estimation algorithms) from [J7, X3, J4] (performance and Cramér-Rao analyses), which seem to reflect whether or not a statistical analysis is present in the contribution. The component $\mathbf{u}_2$ opposes [X1, C14, C20, J9, C10, C17] to [J10, J4, X3], which appears to indicate if an application is targeted or not. Additionally the line following $(\mathbf{u}_1 - \mathbf{u}_2)$ splits the contributions between "PCA only" and "covariance only" approaches. The points in the center contain some contributions about low-rank structured covariance matrices, which is a bridge between the two approaches. The plane $\{\mathbf{u}_3, \mathbf{u}_4\}$ (not displayed) was hard to interpret as clearly, except that $\mathbf{u}_3$ separates the clusters *a*) and *b*) from *c*), as seen in the 3D representation in the left.

Of course, this analysis was made with no pretension outside of being a fun pretext to apply PCA on papers about PCA, before moving to the technicalities of the upcoming chapters.

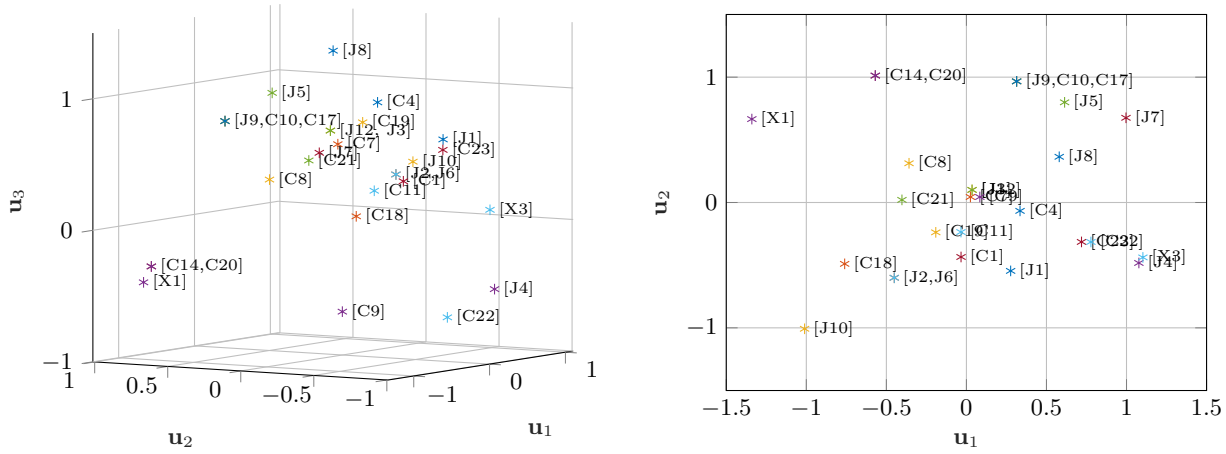


Figure 1.3: PCA applied on the keywords: data projected on $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ (left), and $\{\mathbf{u}_1, \mathbf{u}_2\}$ (right).

2 | Bayesian PCA with Compound Gaussian signals

Contents

2.1 Contributions of the chapter

This chapter concerns subspace recovery with a statistical approach. We present a series of statistical model (based on structured mixtures of compound Gaussian distributions) suited to this problem and discuss their relevance in signal processing applications. We then extend these models to a Bayesian context, where a prior distribution is assumed on the signal subspace basis in order to account for potential available information. We then present several estimation methods driven by this model, and finally illustrate their interest on simulations and a radar detection problem.

The presented work consists in the follow-up of my thesis results, which we generalized to a Bayesian setting in the thesis of Rayen Ben Abdallah.
Related publications in the cv (page xi): Non-Bayesian (my thesis): [J1-J3], [C1-C5], [C7], [FC1], [FC2]. Bayesian models (Thesis of Rayen Ben Abdallah): [J10], [C12], [C19], [FC4].

2.2 Context overview

2.2.1 Probabilistic PCA

Interestingly, the initial derivations of PCA [Pearson, 1901, Hotelling, 1933] are obtained without any assumption on the probabilistic model of the data (only the second-order moment is assumed

to exist). The term "probabilistic PCA" (PPCA) was coined in [Tipping and Bishop, 1999b], who demonstrated that PCA can also be derived within a density estimation framework. Indeed, consider that the samples are drawn as the sum of a low-rank structured signal plus noise

$$\mathbf{z} \overset{d}{=} \mathbf{W}\mathbf{s} + \mathbf{n} \tag{2.1}$$

with $\mathbf{W} \in \mathbb{C}^{p\times k}$ a full rank matrix, $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}) \in \mathbb{C}^k$ and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}) \in \mathbb{C}^p$. We then have the representation $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^H + \sigma^2\mathbf{I}$. The maximum likelihood estimator (MLE) for this model corresponds to the regularization of the sample covariance matrix, where the last $p - k$ eigenvalues are averaged. Hence, from the point of view of subspace recovery (i.e., estimating span($\mathbf{W}$)), this model still leads to the standard PCA since eigenvectors of the sample covariance matrix are kept as estimators. However this formulation of the problem opens the path for numerous extensions. For example, [Tipping and Bishop, 1999b] exploits this model to propose a computationally efficient expectation-maximization (EM) algorithm (close to the power iteration method [Golub and Van Loan, 2012]), as well as a generalization to the case of data with missing entries [Little and Rubin, 2019]. This approach was generalized to Gaussian mixture models for clustering and/or multiple subspace recovery in [Tipping and Bishop, 1999a]. Finally, a link with the so-called factor model in econometrics [Ruppert, 2011] can be pointed out since it generally considers (2.1) with $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is diagonal (i.e. a different variance for each entries).

2.2.2 Beyond Gaussian models

Non-Gaussian signals

In the scope of PPCA, a main angle that still draws research attention is to consider models that go beyond the Gaussian distribution for the two signal components in the representation (2.1). The rationale is that if an estimation algorithm accounts for a statistical model with a better empirical fit to the data, we can expect a more accurate subspace recovery. The presented contributions will follow this idea and consider the use of compound Gaussian distributions (denoted $\mathcal{CG}$) for the signal component. The approach was mainly motivated by high-resolution array processing application, as these distributions are suited to model sources with fluctuating power [Greco et al., 2006, Ollila et al., 2012b], and where the columns of $\mathbf{W}$ represent steering vectors. Embedding compound Gaussian signals in white Gaussian noise then reflects the thermal noise of the system. In these applications, the signal subspace requires to be estimated to apply adaptive subspace methods [Haardt et al., 2014] or low-rank filters (interference cancelation) [Rangaswamy et al., 2004, Ginolhac and Forster, 2016]. However, mixtures of compound Gaussian distributions are quite general and were also used in the statistical literature, for which we can point some references:

Model in (2.1)	$\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$	$\mathbf{n} \sim \mathcal{CG}(\mathbf{0}, \sigma^2\mathbf{I})$
$\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$	PPCA [Tipping and Bishop, 1999b]	Heteroscedastic PPCA [Hong et al., 2018]
$\mathbf{s} \sim \mathcal{CG}(\mathbf{0}, \mathbf{I})$	CG subspace estimation (detailed section 2.3)	Robust PPCA [Archambeau et al., 2006, Chen et al., 2009]

Another approach that we can mention consists in simply assuming a general non-Gaussian multivariate distribution on $\mathbf{z}$ with the second order moment directly structured as $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^H + \sigma^2\mathbf{I}$. This approach, rather than refining the structure of multiple-contribution models (2.1), offers generally less physical interpretation. Yet, it can lead to good estimation performance in practice: robust estimation approaches within this context will be studied in Chapter 3 and 4.

Bayesian priors in PCA

In some applications, we might want to account for some available prior knowledge on the subspace to be recovered. This prior information can be for example drawn from previous estimates in a

sequential process, or from a physical model of the data. From the Bayesian perspective, accounting for this information can be done by assigning a prior distribution on the orthonormal subspace basis and deriving corresponding Bayesian estimators (e.g., maximum a posteriori and minimum mean squared distance). This approach was initially proposed to develop Bayesian PCA algorithms in [Srivastava, 2000, Besson et al., 2011]. In [Besson et al., 2012], these results have been extended to a subspace parameterized by its CS decomposition. In [Elvira et al., 2017a, Elvira et al., 2017b], these concepts were extended to the Bayesian non-parametric framework in order to adaptively select the rank of the subspace to be estimated. In this scope, this chapter will present the derivation of Bayesian PCA algorithms for compound Gaussian sources.

2.3 Data models

2.3.1 Compound Gaussian distributions

Compound Gaussian distributions, also referred to as spherically invariant random vectors [Yao, 1973], have been widely employed in the statistical signal processing literature e.g., in image processing [Shi and Selesnick, 2007, Zozor and Vignat, 2010, Portilla et al., 2003], and for modeling radar clutter [Greco et al., 2006, Ollila et al., 2012b]. A main interest is that they encompass a large family of multivariate distributions, notably heavy-tailed ones. These distributions are a sub-family of the wider class of elliptical distributions [Ollila et al., 2012a] (that will be presented in Chapter 3). However, we focus here on compound Gaussian distributions only because of a practical representation theorem, that will be instrumental to handle mixtures models and Bayesian priors:

Definition 2.3.1. ***Compound Gaussian distribution***
A p-dimensional CG vector is represented as a product of two statistically independent components, i.e., if $\mathbf{z} \in \mathbb{C}^p$ *follows a compound Gaussian distribution, denoted* $\mathbf{z} \sim \mathcal{CG}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f_\tau)$*, it has the following stochastic representation*

$$\mathbf{z} \overset{d}{=} \boldsymbol{\mu} + \sqrt{\tau}\mathbf{d}, \tag{2.2}$$

where

i) $\boldsymbol{\mu} \in \mathbb{C}^p$ *is the center of distribution, which coincide with* $\mathbb{E}\{\mathbf{z}\} = \boldsymbol{\mu}$ *when existing.*

ii) $\tau \in \mathbb{R}^+$ *is a positive random scalar, called texture, of p.d.f.* f_τ*. This parameter is statistically independent of* $\mathbf{d}$*.*

iii) $\mathbf{d} \in \mathbb{C}^p$ *follows a zero-mean multivariate complex Gaussian distribution of covariance matrix* $\boldsymbol{\Sigma}$*, denoted,* $\mathbf{d} \sim \mathcal{CN}(0, \boldsymbol{\Sigma})$*. The matrix* $\boldsymbol{\Sigma} \in \mathcal{H}_p^{++}$ *(*$\mathcal{H}_p^{++}$ *denotes the set of* $p \times p$ *positive definite Hermitian matrices) is referred to as the scatter matrix. Notice that if* $\mathbb{E}\{\tau\} < \infty$*, the covariance matrix of* $\mathbf{z}$ *exists and is proportional to the scatter matrix, i.e.,* $\mathbb{E}\{(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^H\} = \mathbb{E}\{\tau\}\boldsymbol{\Sigma}$*.*

The p.d.f. of a random vector $\mathbf{z} \sim \mathcal{CG}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f_\tau)$ *is thus defined by*

$$f(\mathbf{z}) = \pi^{-M}|\boldsymbol{\Sigma}|^{-1} \int_0^\infty \tau^{-M} \exp\left\{\frac{-(\mathbf{z}-\boldsymbol{\mu})^H\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})}{\tau}\right\} f_\tau(\tau)d\tau, \tag{2.3}$$

and conditionally to the texture, the random vector $\mathbf{z}$ *has the distribution* $\mathbf{z}|\tau \sim \mathcal{CN}(\boldsymbol{\mu}, \tau\boldsymbol{\Sigma})$*.*

Note that compound Gaussian distributions are not uniquely defined, as the vectors $\mathbf{z} \sim \mathcal{CG}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f_\tau)$ and $\mathbf{z}' \sim \mathcal{CG}(\boldsymbol{\mu}, c\boldsymbol{\Sigma}, f_{\tau'})$, with $\tau' = \frac{\tau}{c}$ and $c \in \mathbb{R}_+^*$, satisfy $\mathbf{z} \overset{d}{=} \mathbf{z}'$. In order to avoid any ambiguity, we may impose an arbitrary scaling constraint, such as $\mathrm{Tr}\{\boldsymbol{\Sigma}\} = 1$, $|\boldsymbol{\Sigma}| = 1$, or $\mathbb{E}\{\tau\} = 1$ (when $\mathbb{E}\{\tau\} < \infty$).

Depending on the choice of f_τ, the compound Gaussian representation can lead to numerous usual multivariate distributions, such as Gaussian, Weibull, K-, and Student t- distributions (c.f.

[Ollila et al., 2012a] for details). In most applications, a strong prior information on the texture distribution is not actually available. In order to design algorithms that are robust to the whole family compound Gaussian distributions, a common approach is to consider that the texture τ is unknown and deterministic for each realization, which offers an interesting robustness-performance trade off [Tyler, 1987, Pascal et al., 2008b]. This model will be denoted as below:

Definition 2.3.2. ***Compound Gaussian distribution with deterministic texture***
Consider a n-sample $\{\mathbf{z}_i\}_{i=1}^n$ *following a compound Gaussian distribution while assuming a deterministic texture* τ_i *for each sample* $\mathbf{z}_i$*. We denote this model by* $\mathbf{z}_i \sim \mathcal{CG}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_{\tau_i})$*, or equivalently,* $\mathbf{z}_i|\tau_i \sim \mathcal{CN}(\boldsymbol{\mu}, \tau_i\boldsymbol{\Sigma})$*,* $\forall i$*.*

2.3.2 Generalized Bingham-Langevin distributions

Definition

First, recall that we denote the Stiefel manifold $\mathrm{St}(p,k) = \{\mathbf{U} \in \mathbb{C}^{p\times k} \mid \mathbf{U}^H\mathbf{U} = \mathbf{I}\}$. In this section, we present a familly of distributions for orthonormal bases $\mathbf{U} \in \mathrm{St}(p,k)$ from the field of directional data analysis. Probability distributions and statistical inference on $\mathrm{St}(p,k)$ have been developed primarily in the spatial statistics literature [Chikuse, 2003, Mardia and Jupp, 2009], starting with the real circle and real sphere, then extended to higher dimensions (and the complex case). These distributions were notably used for the formulation of Bayesian PCA algorithms [Srivastava, 2000, Besson et al., 2011]. In [Ben Abdallah et al., 2020], we introduced the complex generalized Bingham Langevin (CGBL) distributions as a generalization of these directional statistics to the case of matrix variables with complex entries: the CGBL is a probability distribution on $\mathrm{St}(p,k)$ which combines linear and quadratic terms. CGBL are defined as follows:

Definition 2.3.3. ***Complex generalized Bingham Langevin (CGBL) distribution***
The CGBL is parametrized by the set of Hermitian matrices $\{\mathbf{A}_r\} \in \mathcal{H}_p^+$ *and the matrix* $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_k] \in \mathbb{C}^{p\times k}$*. We denote* $\mathbf{U} \sim \mathrm{CGBL}(\mathbf{C}, \{\mathbf{A}_r\})$ *when the p.d.f. of* $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ *on* $\mathrm{St}(p,k)$ *reads*

$$p_{CGBL}(\mathbf{U}) = c_{\mathrm{CGBL}}(\mathbf{C}, \{\mathbf{A}_r\}) \exp\left\{\sum_{r=1}^{k} \mathrm{Re}\{\mathbf{c}_r^H\mathbf{u}_r\} + \mathbf{u}_r^H\mathbf{A}_r\mathbf{u}_r\right\} \tag{2.4}$$

where $c_{\mathrm{CGBL}}(\mathbf{C}, \{\mathbf{A}_r\})$ *is a normalizing constant.*

From (2.4), we can interpret that p_{CGBL} promotes the concentration of each vector $\mathbf{u}_r$ around $\mathbf{c}_r$ and each range space $\mathbf{u}_r\mathbf{u}_r^H$ around the subspace associated to the strongest eigenvalues of the Hermitian matrix $\mathbf{A}_r$. Typically, if $\mathbf{A}_r = \mathbf{A}$, $\forall r \in [\![1,k]\!]$, the range space $\mathbf{U}\mathbf{U}^H$ tends to be close to the dominant eigenspace of $\mathbf{A}$. In the following, we list some usual special cases of the CGBL distribution.

Example 2.3.1. ***Complex invariant Bingham (CIB) distribution***
The CIB is a special case of the CGBL where $\mathbf{C} = \mathbf{0}$ *and* $\mathbf{A}_r = \kappa\bar{\mathbf{U}}\bar{\mathbf{U}}^H$*,* $\forall r \in [\![1,k]\!]$ *where* $\bar{\mathbf{U}} \in \mathrm{St}(p,k)$ *represents the center of the distribution and* κ *denotes the concentration parameter. We denote* $\mathbf{U} \sim CIB(\kappa, \bar{\mathbf{U}}\bar{\mathbf{U}}^H)$ *when* $\mathbf{U}$ *has as p.d.f. of the form*

$$p_{CIB}(\mathbf{U}) = c_{CIB}(\kappa, \bar{\mathbf{U}}) \mathrm{etr}\left\{\kappa\mathbf{U}^H\bar{\mathbf{U}}\bar{\mathbf{U}}^H\mathbf{U}\right\} \tag{2.5}$$

in which $c_{CIB}(\kappa, \bar{\mathbf{U}})$ *denotes the normalizing constant.*

Note that the $p_{\mathrm{CIB}}(\mathbf{U})$ is invariant by rotation $\mathbf{U}' = \mathbf{U}\mathbf{Q}$, $\forall \mathbf{Q} \in \mathrm{St}(p,k)$. This means that p_{CIB} characterizes a distribution for the subspace represented by the orthogonal projector $\mathbf{U}\mathbf{U}^H$, which will be its main interest.

Example 2.3.2. ***Complex Langevin (CL) distribution***
The CL is a special case of the CGBL for which $\{\mathbf{A}_r = \mathbf{0}\}$ *and* $\mathbf{C} = \kappa\bar{\mathbf{U}}$*, where* $\bar{\mathbf{U}} \in \mathrm{St}(p,k)$ *is the center of distribution and* κ *is the concentration parameter. We denote* $\mathbf{U} \sim \mathrm{CL}(\kappa, \bar{\mathbf{U}})$ *when* $\mathbf{U}$ *has as p.d.f. of the following form*

$$p_{\mathrm{CL}}(\mathbf{U}) = {}_0F_1(\frac{1}{2}N, \frac{1}{4}\bar{\mathbf{U}}^H\bar{\mathbf{U}})^{-1}\mathrm{etr}\left\{\kappa\mathrm{Re}\{\bar{\mathbf{U}}^H\mathbf{U}\}\right\} \tag{2.6}$$

with ${}_0F_1(\frac{1}{2}N, \frac{1}{4}\bar{\mathbf{U}}^H\bar{\mathbf{U}})^{-1}$ *is the normalizing constant for this distribution.*

To illustrate these distributions, Table 2.1 displays the p.d.f. of the real Bingham distribution $\mathbf{u} \sim \mathrm{B}(\kappa\bar{\mathbf{u}}\bar{\mathbf{u}}^T)$ and the real Langevin distributions $\mathbf{u} \sim \mathrm{L}(\kappa\bar{\mathbf{u}})$ on $\mathrm{St}(2,1)$ (unit circle in $\mathbb{R}^2$) where $\bar{\mathbf{u}} = [1/\sqrt{2}, 1/\sqrt{2}]$ defines the center of distribution, and κ is a concentration parameter. We note that for high value of $\kappa \in \mathbb{R}_+^*$, i.e., $\kappa = 50$, the generated samples $\mathbf{u} \sim \mathrm{L}(\kappa\bar{\mathbf{u}})$ are more concentrated around the center $\bar{\mathbf{u}}$. For real Bingham distribution, the samples are gathered on both sides $\bar{\mathbf{u}}$ and $-\bar{\mathbf{u}}$, since, this distribution characterizes the quantity $\mathbf{u}\mathbf{u}^T$.

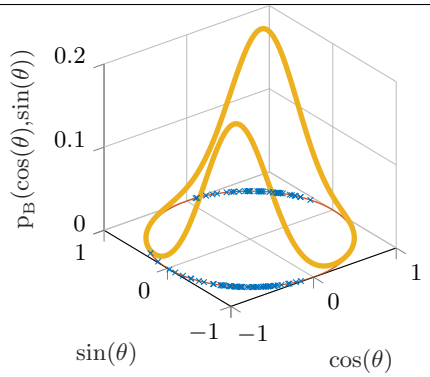
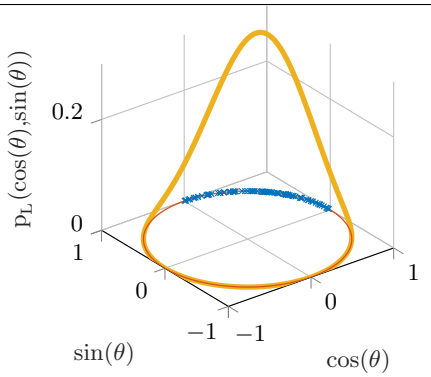
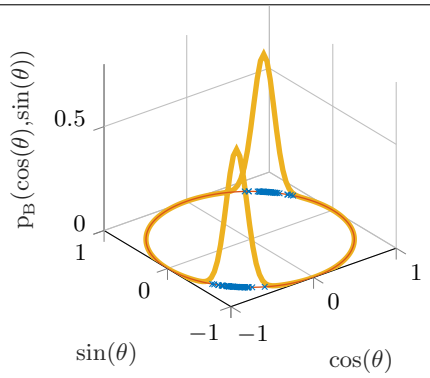
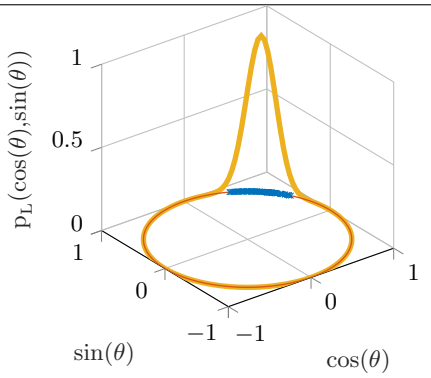
	Bingham distribution	Langevin distribution
$\kappa = 5$		
$\kappa = 50$		

Table 2.1: The p.d.f. of Bingham and Langevin distributions for various values of concentration parameter κ and for a prior center $\bar{\mathbf{u}} = [1/\sqrt{2}, 1/\sqrt{2}]$ on $\mathrm{St}(2,1)$ (represented by the unit circle) and 100 samples generated according to these distributions

Sampling from CGBL distribution

Several sampling methods were proposed in order to simulate random matrices drawn from the aforementioned distribution. Such approaches are based on Markov Chain Monte Carlo (MCMC) methods [Hoff, 2009] and/or acceptance rejection schemes [Kent et al., 2013]. In [Ben Abdallah et al., 2020], we proposed a general method to draw samples as $\mathbf{U} \sim \mathrm{CGBL}(\mathbf{C}, \{\mathbf{A}_r\})$. The proposed sampling technique is based on previous results, and summarized as follows:

i) The generation of $\mathbf{U}$ is obtained as a Markov chain on the columns $\{\mathbf{u}_r\}$, as proposed in [Hoff, 2009, Besson et al., 2011].

ii) The generation of each $\mathbf{u}_r$ is obtained by using the results of [Kent et al., 2013], which proposed an acceptance rejection method to sample from a vector Bingham-Langevin distribution. For this problem [Hoff, 2009] proposed a Markov chain on the entries of the vectors. Both methods allow to sample the desired distribution, however, the acceptance rejection scheme from [Kent et al., 2013] allows us to significantly reduce the computation time.

Notice that [Hoff, 2009] and [Kent et al., 2013] proposed methods adapted to distributions of real variables. In order to generalize these sampling techniques to the case of complex distributions, we resort to the change of variables proposed in [Mardia and Jupp, 2009].

2.3.3 Structured mixtures of compound Gaussian models

This section details several models that we used to propose signal subspace estimation methods. These models are variations around (2.1) using compound Gaussian signals with deterministic textures. An important note is that we resort to the reparameterization $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^H = \mathbf{U}\mathbf{D}\mathbf{U}^H$, with $\mathbf{U} \in \mathrm{St}(p,k)$ and $\mathbf{D}$ a diagonal matrix with positive entries. The main goal is to explicitly focus on the subspace recovery problem, e.g., in the structure of the MLE. Another interest of this parameterization is the possibility to include Bayesian priors directly on the parameter of interest $\mathbf{U}$, rather than on the intermediary $\mathbf{W}$ [Bishop, 1999].

Definition 2.3.4. ***Noisy low-rank compound Gaussian (LRCG)***
Consider a n-sample $\{\mathbf{z}_i\}_{i=1}^n$ *having the stochastic representation*

$$\mathbf{z}_i \overset{d}{=} \sqrt{\tau_i}\mathbf{U}\mathbf{D}^{1/2}\mathbf{s}_i + \mathbf{n}_i, \tag{2.7}$$

with $\mathbf{n}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, $\mathbf{s}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, $\mathbf{U} \in \mathrm{St}(p,k)$ *and* $\mathbf{D} = \mathrm{diag}(\mathbf{d}) \in \mathbb{R}_+^{k\times k}$ *diagonal with positive entries, and where* τ_i *is an unknown positive deterministic texture. We denote this LRCG model by its conditional representation* $\mathbf{z}_i|\tau_i \sim \mathcal{CN}(\mathbf{0},\tau_i\mathbf{\Sigma}+\mathbf{I})$, $\forall i$, *with* $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^H$.

Again, LRCG corresponds to a model for low-rank non-Gaussian signal embedded in white Gaussian noise, which offers a very general framework for generic PPCA, as discussed in [Besson, 2016]. Note that we assumed knowledge of the white Gaussian noise power, or equivalently $\sigma^2 = 1$ for the ease of exposition. However, the extension to unknown σ^2 can be done trivially.

Definition 2.3.5. ***Simplified LRCG (LRCGs)***
Consider a n-sample $\{\mathbf{z}_i\}_{i=1}^n$ *having the stochastic representation*

$$\mathbf{z}_i \overset{d}{=} \sqrt{\tau_i}\mathbf{U}\mathbf{s}_i + \mathbf{n}_i, \tag{2.8}$$

with $\mathbf{n}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, $\mathbf{s}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$ *with* $\mathbf{U} \in \mathrm{St}(p,k)$, *and where* τ_i *is an unknown positive deterministic texture. We denote this LRCGs model by its conditional representation* $\mathbf{z}_i|\tau_i \sim \mathcal{CN}(\mathbf{0},\tau_i\mathbf{U}\mathbf{U}^H+\mathbf{I})$, $\forall i$.

The model LRCGs was initially introduced as a relaxation of LRCG (identical eigenvalues for the low-rank signal covariance matrix), which turned out to be difficult to solve from the point of view of ML estimation [Raghavan, 2012, Breloy et al., 2013]. However, as we will see, this approach leads to good performance, even in the mismatched case. This observation suggests that it is reasonable to neglect the variations of eigenvalues if the main goal is only to estimate the signal subspace.

Definition 2.3.6. ***Simplified LRCG with outliers (LRCGo)***
Consider a n-sample $\{\mathbf{z}_i\}_{i=1}^n$ *having the stochastic representation*

$$\mathbf{z}_i \overset{d}{=} \sqrt{\tau_i}\mathbf{U}\mathbf{s}_i + \sqrt{\beta_i}\mathbf{U}_\perp\mathbf{o}_i + \mathbf{n}_i, \tag{2.9}$$

with $\mathbf{n}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, $\mathbf{s}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, $\mathbf{U} \in \mathrm{St}(p,k)$, $\mathbf{o}_i \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, $\mathbf{U}_\perp \in \mathrm{St}(p,p-k)$ *an orthogonal complement of* $\mathbf{U}$ *(i.e.,* $[\mathbf{U},\mathbf{U}_\perp]^H[\mathbf{U},\mathbf{U}_\perp]=\mathbf{I}$*), and where* τ_i *and* β_i *are unknown positive deterministic textures. We denote this LRCGo model by its conditional representation* $\mathbf{z}_i|\tau_i,\beta_i \sim \mathcal{CN}(\mathbf{0},\tau_i\mathbf{U}\mathbf{U}^H + \beta_i\mathbf{U}_\perp^H\mathbf{U}_\perp^H + \mathbf{I})$, $\forall i$.

The model LRCGo considers potential outliers as orthogonal contributions to the signal subspace. It was introduced in [Breloy et al., 2016] in order to propose a robust subspace recovery algorithm. Interestingly, mixtures of orthogonal elliptical distributions are also referred to as haystack (or needle-haystack) models the machine learning community [Lerman et al., 2015, Lerman and Maunu, 2017]. However, these are more generally used for assessing the performance of robust subspace recovery algorithms, rather than for driving a parametric estimation problem.

Finally, we denote with B-LRCG, B-LRCGs, and B-LRCGo, the corresponding counterparts of these three models when a prior $\mathbf{U} \sim \mathrm{CGBL}(\mathbf{C},\{\mathbf{A}_r\})$ is assumed on the signal subspace basis. We can sum-up these models and (a selected) corresponding reference in the following table:

Model	LRCG	LRCGs	LRCGo
PPCA	[Sun et al., 2016]	[Raghavan, 2012]	[Breloy et al., 2016]
Bayesian PCA	[Ben Abdallah et al., 2020]	[Ben Abdallah et al., 2020]	[Ben Abdallah et al., 2019a]

2.4 Algorithms

2.4.1 Maximum a posteriori (MAP)

For the LRCG model from definition 2.3.4, we denote the likelihood of a n-sample by $\mathcal{L}(\{\mathbf{z}_i\}_{i=1}^n|\mathbf{U},\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ aggregates side parameters. Assuming a CGBL distribution as in (2.4) for $\mathbf{U}$, the posterior probability of this parameter is then given in

$$p_{\mathbf{U}}(\mathbf{U}|\{\mathbf{z}_i\}_{i=1}^n) \propto \mathcal{L}(\{\mathbf{z}_i\}_{i=1}^n|\mathbf{U},\boldsymbol{\theta})p_{\mathrm{CGBL}}(\mathbf{U}). \tag{2.10}$$

After several manipulations, the maximum a posteriori (MAP) of this B-LRCG model can be expressed as the solution of the problem

$$\begin{aligned} \underset{\{\tau_i\},\{d_r\},\{\mathbf{u}_r\}}{\text{minimize}} \quad & \sum_{i=1}^{n}\sum_{r=1}^{k}\left[\ln\left(1+\tau_i d_r\right) - \frac{\tau_i d_r}{\tau_i d_r+1}\mathbf{z}_i^H\mathbf{u}_r\mathbf{u}_r^H\mathbf{z}_i\right] + \ln(p_{\mathrm{CGBL}}(\mathbf{U})) \\ \text{subject to} \quad & \tau_i \geq 0,\ d_r \geq 0, \\ & \mathbf{U}=[\mathbf{u}_1,\ldots,\mathbf{u}_k] \in \mathrm{St}(p,k) \end{aligned} \tag{2.11}$$

where $\ln(p_{\mathrm{CGBL}}(\mathbf{U}))$ consists in linear and quadratic terms in $\mathbf{U}$. This problem has no closed-form solutions and requires the use of iterative algorithms. In [Ben Abdallah et al., 2020], we proposed to leverage the majorization-minimization approach for this problem. The majorization-minimization algorithm proceeds with two steps: *i*) (majorization) finding a function that locally upperbounds the objective function up to a constant, referred to as surrogate function; *ii*) (minimization) minimizing this surrogate function. This procedure generates a sequence that monotonically decreases the objective value, and its main interest is that it can yield a sequence of subproblems that are easy to deal with. Following this procedure, we derived appropriate surrogate functions in order to obtain an algorithm with closed-form updates of the variables, and theoretical convergence guarantees. In order to lighten the presentation, these tedious derivations are not reported in this chapter[1]. For more information on the matter, Appendix A also details a generic majorization-minimization framework for a parameter $\mathbf{U} \in \mathrm{St}(p,k)$ (notably applicable to the update of $\mathbf{U}$ in (2.11)). Note that, when setting $\ln(p_{\mathrm{CGBL}}(\mathbf{U})) = 0$ (i.e., no prior), we also recover the algorithms to compute the maximum

[1]Interested readers can find the details in the sections 2.1 and 2.2 of the annexes

likelihood from [Sun et al., 2016].

Interestingly, the MAP for B-LRCGo (resp. B-LRCGs with $\beta_i = 0,\ \forall i \in [\![1, n]\!]$) admits an insightful structure in some specific cases. First, define $\mathcal{P}_k$ as the operator that extracts the k leftmost eigenvectors of a matrix

$$\begin{array}{rlll} \mathcal{P}_k: & \mathcal{H}_p^+ & \longrightarrow & \mathrm{St}(p,k) \\ & \mathbf{M} \overset{\text{EVD}}{=} [\mathbf{U}_k|\mathbf{U}_k^\perp]\ \mathbf{D}\ [\mathbf{U}_k|\mathbf{U}_k^\perp]^H & \longmapsto & \mathcal{P}_k\{\mathbf{M}\} = \mathbf{U}_k. \end{array} \tag{2.12}$$

We have the following theorem:

Theorem 2.4.1. ***MAP structure for B-LRCGo*** *[Ben Abdallah et al., 2019a]*
Let $\{\mathbf{z}_i\}_{i=1}^n$ be a n-sample following the B-LRCGo model in definition 2.3.6, with $\mathbf{U} \sim CIB(\kappa, \bar{\mathbf{U}}\bar{\mathbf{U}}^H)$ (cf. example 2.3.1). The MAP of the signal subspace basis $\hat{\mathbf{U}}$ satisfies the fixed-point equation

$$\hat{\mathbf{U}} = \mathcal{P}_k \left[\sum_{i=1}^{n} \rho\left(\hat{\mathbf{U}}, \mathbf{z}_i\right) \mathbf{z}_i \mathbf{z}_i^H + \kappa \bar{\mathbf{U}}\bar{\mathbf{U}}^H \right], \tag{2.13}$$

where the function ρ is defined by

$$\rho\left(\hat{\mathbf{U}}, \mathbf{z}_i\right) = \frac{\max(0,\ \hat{\tau}_i - \hat{\beta}_i\ )}{(\hat{\beta}_i + 1)\,(\hat{\tau}_i + 1)},\ \textit{with}\ \left\{ \begin{array}{ll} \hat{\tau}_i = & \max(0,\ \mathbf{z}_i^H \hat{\mathbf{U}}\hat{\mathbf{U}}^H \mathbf{z}_i/k - 1), \\ \hat{\beta}_i = & \max(0,\ \mathbf{z}_i^H (\mathbf{I} - \hat{\mathbf{U}}\hat{\mathbf{U}}^H)\mathbf{z}_i/(p-k) - 1). \end{array} \right. \tag{2.14}$$

Hence, this subspace estimator is contained in the dominant eigenspace of an intermediary sample covariance matrix (built from adaptively weighted samples) plus a weighted projector on the center of the prior. The adaptive weights tend to reject (resp. promote) samples if they are perceived as outliers (resp. inliners), which occurs when $\hat{\beta}_i > \hat{\tau}_i$ (resp. $\hat{\beta}_i \ll \hat{\tau}_i$). Such estimate can be obtained through fixed point iterations, which also correspond to a block-coordinate descent algorithm with closed form updates on each parameters.

2.4.2 Minimum mean square distance (MMSD)

The MMSD estimator minimizes the expected Euclidean distance between the true range space $\mathcal{R}(\mathbf{U}) = \mathbf{U}\mathbf{U}^H$ and its estimate $\mathcal{R}(\hat{\mathbf{U}}) = \hat{\mathbf{U}}\hat{\mathbf{U}}^H$. This formulation was proposed in [Besson et al., 2011], in which its practical expression is obtained as follows:

$$\begin{aligned} \hat{\mathbf{U}}_{\text{MMSD}} &= \arg\min_{\hat{\mathbf{U}}}\ \mathbb{E}_{\mathbf{U},\mathbf{Z}} \left\{ \| \hat{\mathbf{U}}\hat{\mathbf{U}}^H - \mathbf{U}\mathbf{U}^H \|_F^2 \right\} = \arg\max_{\hat{\mathbf{U}}}\ \mathbb{E}_{\mathbf{U},\mathbf{Z}}\{\mathrm{Tr}\{\hat{\mathbf{U}}^H \mathbf{U}\mathbf{U}^H \hat{\mathbf{U}}\}\} \\ &= \arg\max_{\hat{\mathbf{U}}}\ \int \left[\int \mathrm{Tr}\{\hat{\mathbf{U}}^H \mathbf{U}\mathbf{U}^H \hat{\mathbf{U}}\} p(\mathbf{U}|\mathbf{Z}) d\mathbf{U} \right] p(\mathbf{Z}) d\mathbf{Z} \\ &= \arg\max_{\hat{\mathbf{U}}}\ \mathrm{Tr}\left\{ \hat{\mathbf{U}}^H \left[\int \mathbf{U}\mathbf{U}^H p(\mathbf{U}|\mathbf{Z}) d\mathbf{U} \right] \hat{\mathbf{U}} \right\} \\ &= \mathcal{P}_k \left\{ \int \mathbf{U}\mathbf{U}^H p(\mathbf{U}|\mathbf{Z}) d\mathbf{U} \right\} \triangleq \mathcal{P}_k \left\{ \mathbf{M}(p(\mathbf{U}|\mathbf{Z})) \right\} \end{aligned} \tag{2.15}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{C}^{p \times n}$ denote the data matrix, and where $\mathcal{P}_k$ is defined in (2.12). Hence, the MMSD depends on $p(\mathbf{U}|\mathbf{Z})$, which is specified from both the data model and the prior distribution assigned to the parameters. Except from several special cases, there is no closed-form solutions to compute $\mathbf{M}(p(\mathbf{U}|\mathbf{Z}))$. However, the MMSD can still be evaluated using the so-called induced arithmetic mean

$$\hat{\mathbf{U}}_{\text{MMSD}} \approx \mathcal{P}_P \left\{ \frac{1}{N_r} \sum_{n=N_{bi}+1}^{N_{bi}+N_r} \mathbf{U}_{(n)} {\mathbf{U}_{(n)}}^H \right\}, \tag{2.16}$$

where $\mathbf{U}_{(n)}$ are sampled from $p(\mathbf{U}|\mathbf{Z})$, N_{bi} stands for the burn-in samples (number of thrown samples from the Markov chain), and N_r is the number of samples used to evaluate the integral.

Combining this methodology with the previous majorization-minimization approach, we proposed several algorithms to evaluate the MMSD of the models LRCG and LRCGs in [Ben Abdallah et al., 2020]. Again, the details are omitted for the sake of conciseness.

2.5 Simulations and application

2.5.1 Simulations

This section displays some simulations examples that allow to draw general conclusions on the developed methodologies. MLE, MAP and MMSD denote the estimators build using the true data model. The suffixes "s" (resp."o") indicate the use of the LRCGs (resp. LRCGo) model for the estimation process, even for actually LRCG distributed data (mismatch). Thus, we study the impact of assuming a simplified model (equal eigenvalues) in this case. $\bar{\mathbf{U}}$ denotes the "prior only" estimator. We consider the the average fraction of recovered energy (AFE), defined as

$$\mathrm{AFE}(\hat{\mathbf{U}}) = \mathbb{E}_{\mathbf{U},\mathbf{Z}}\left\{\mathrm{Tr}\left\{\hat{\mathbf{U}}^H\mathbf{U}\mathbf{U}^H\hat{\mathbf{U}}\right\}\right\}/k, \tag{2.17}$$

for a given estimator $\hat{\mathbf{U}}$. Figure 2.1 displays the AFE of several algorithms for scenarios corresponding to a general B-LRCG. Figure 2.2 displays the AFE of several algorithms for scenarios corresponding to B-LRCGo.

• For LRCG scenarios, the MLE approach generally yields the same performance as the standard PCA. The observation was also made in [Besson, 2016], where PCA is observed to reach an accuracy close to the Cramér-Rao bound. This disappointing result can be explained by the structure of the ML problem in (2.11): for high signal to noise ratio (high values of $\tau_k d_r$), the weights applied on the samples tend to be close to 1, meaning that the signal subspace MLE is close to the one from the standard PCA. This result is unexpected since it does not involve robust estimate of the covariance matrix, even in the presence of non-Gaussian observations. The MLE still offers a gain when highly impulsive signals occur, and when n is large enough to benefit from the natural "sample selection" of the MLE's structure, as observed in [Breloy et al., 2015]. Additionally, the ML approach remains useful for other applications, e.g., when requiring the estimation of the covariance matrix parameters [Breloy et al., 2016a].

• Bayesian estimators achieve better estimation accuracy thanks to the inclusion of prior knowledge. In practice, such prior may not be fully available. However, the approach seems robust to slight mismatches on the parameters selection. This can be observed e.g., in Figure 2.1, where the MMSD of an approximated model reaches performance close the the actual MMSD. The MAP represents an interesting trade-off between the MMSD and MLE, notably because of the computational cost associated to the computation of the MMSD. Additionally, we can generally observe that the approximation of equals eigenvalues (i.e., assuming LRCGs) is quite harmless if the only focus is the estimation of the signal subspace.

• The most encouraging results are obtained when investigating the robustness of the developed subspace recovery methods. In this scope, Figure 2.2 illustrates that (B-)LRCGo can yield good recovery in practice. Interestingly, the approach is also robust to wrong assumptions (simplifying priors, equals eigenvalues), which permits the derivation of efficient algorithms with a low computational cost.

2.5.2 Application to radar detection

The inclusion of a Bayesian prior can significantly improve the performance of an estimation process. However, the design of this prior depends on the considered application and comes from ap-

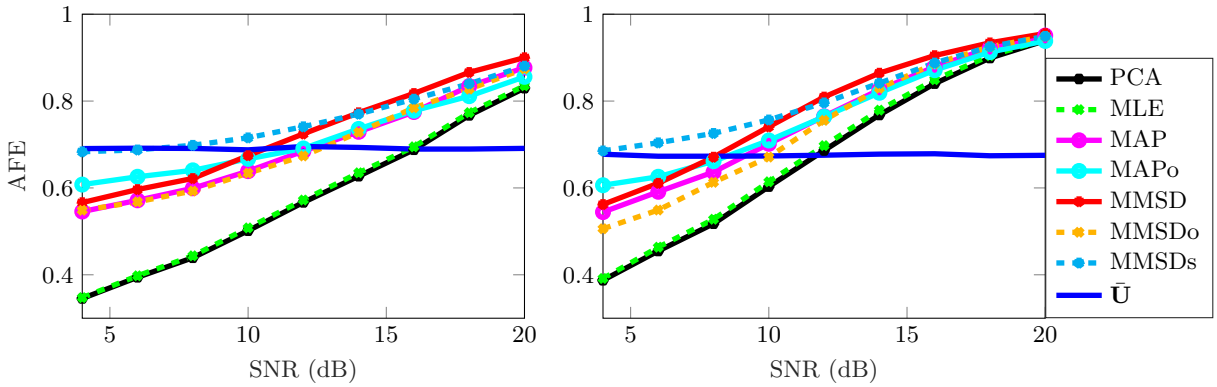


Figure 2.1: AFE w.r.t. signal to noise ratio (SNR) for various estimators. B-LRCG model $\mathbf{z}_k|\tau_k \sim \mathcal{CN}(0, \tau_k\mathbf{U}\mathbf{D}\mathbf{U}^H + \sigma^2\mathbf{I})$, with $\tau_k \sim \Gamma(\nu, \frac{1}{\nu})$, $\forall k$, and $\nu = 0.5$. $[\mathbf{D}]_{r,r} = (k+1-r)/(\sum_{i=1}^{k} i)$ and σ^2 to fix the SNR as SNR$= \mathrm{Tr}\{\mathbf{\Lambda}\}/\sigma^2$. $\mathbf{U} \sim \mathrm{CGBL}(\mathbf{0}, \{\kappa_0\phi_r\bar{\mathbf{U}}\bar{\mathbf{U}}^H\}_{r=1}^{k})$, $\phi_r = (k+1-r)/(\sum_{i=1}^{k} i)$, $\kappa_0 = 300$, $\bar{\mathbf{U}}$ is the first vectors of the canonical basis. $k = 5$, $p = 20$, $n = 3k$ (left) and $n = 6k$ (right).

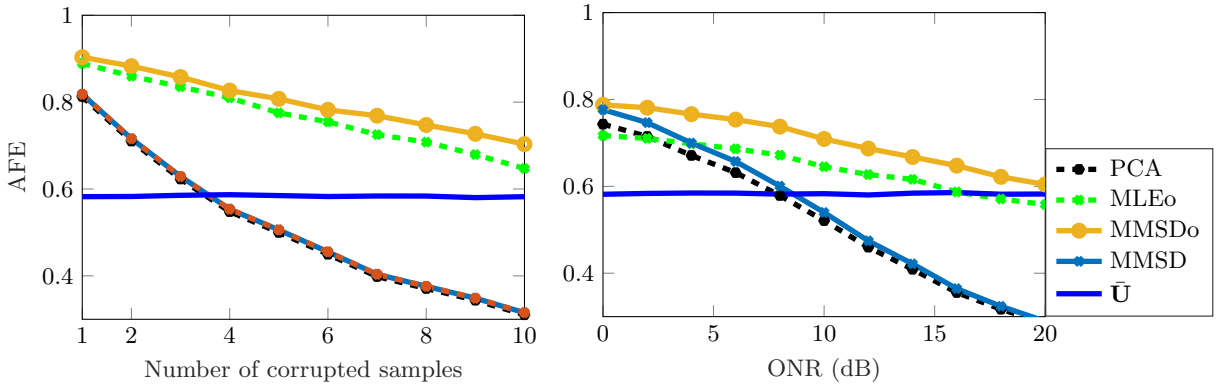


Figure 2.2: AFE w.r.t. number of corrupted samples for outlier to noise ratio ONR=SNR=15dB (left), and w.r.t. ONR for SNR=10dB (right). B-LRCGo model $\mathbf{z}_k|\tau_k, \beta_k \sim \mathcal{CN}(0, \tau_k\mathbf{U}\mathbf{U}^H + \beta_k\mathbf{U}_\perp\mathbf{U}_\perp^H + \sigma^2\mathbf{I})$, with $\tau_k \sim \Gamma(1,1)$ and $\beta_k \sim \Gamma(1,1)$, $\forall k$. $\mathbf{U} \sim \mathrm{CIB}(\kappa, \bar{\mathbf{U}}\bar{\mathbf{U}}^H)$, $\kappa = 60$, $\bar{\mathbf{U}}$ is the first vectors of the canonical basis. $k = 5$, $p = 30$, $n = 20$. For this scenario, the MAP and MMSD coincide.

propriate physical considerations/models on the system. In the following, we illustrate the practical use of the proposed methods for the airborne STAP detection [Ward, 1994]. In this application, the clutter (response of the environment) lies in a low dimensional subspace that needs to be estimated to perform adaptive interference cancellation. We consider the approach that directly leverages the physical model of [Ward, 1994] in order to improve the performance of low-rank detectors [Rangaswamy et al., 2004, Ginolhac and Forster, 2016] on a real dataset provided by the French agency DGA/MI [Ovarlez et al., 2011]. The STAP detection problem is a binary hypothesis test:

$$\begin{cases} \mathcal{H}_0 : \mathbf{z}_i = \mathbf{c}_i + \mathbf{n}_i, \ \forall i \in [\![0, n]\!] \\ \mathcal{H}_1 : \mathbf{z}_0 = \mathbf{d} + \mathbf{c}_0 + \mathbf{n}_0, \quad \mathbf{z}_i = \mathbf{c}_i + \mathbf{n}_i, \ \forall i \in [\![1, n]\!] \end{cases} \tag{2.18}$$

where the secondary data $\mathbf{z}_i \in \mathbb{C}^p$, $\forall i \in [\![1, n]\!]$ are assumed to be i.i.d.. The additive noise in each sample is the sum of clutter (ground response) $\mathbf{c}_i$ plus white Gaussian noise $\mathbf{n}_i$. The tested cell $\mathbf{z}_0$ may potentially contain a moving target $\mathbf{d} = \alpha_0\mathbf{p}$ where α_0 is the amplitude and $\mathbf{p}$ is the steering vector. From the Brennan's rule [Brennan and Staudaher, 1992], we know that the clutter lies in an unknown low-dimensional subspace represented by the orthogonal projector $\mathbf{\Pi}_c = \mathbf{U}_c\mathbf{U}_c^H$ (CSP) of known rank k. In this context, we can use the LR-ANMF detector [Ginolhac and Forster, 2016] to

assess the performance of various CSP estimation methods:

$$\hat{\Lambda} = \frac{|\mathbf{d}^H \hat{\mathbf{\Pi}}_c^{\perp} \mathbf{z}_0|^2}{|\mathbf{d}^H \hat{\mathbf{\Pi}}_c^{\perp} \mathbf{d}||\mathbf{z}_0^H \hat{\mathbf{\Pi}}_c^{\perp} \mathbf{z}_0|} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \delta \tag{2.19}$$

where $\hat{\mathbf{\Pi}}_c^{\perp} = \mathbf{I} - \hat{\mathbf{\Pi}}_c$ is an estimate of the orthogonal complement of the CSP. We compare the following detectors: *i*) $\hat{\Lambda}^{\text{SCM}}$ is the LR-ANMF where the CSP is built from the standard PCA; *ii*) $\hat{\Lambda}^{\text{SFPE}}$ is the LR-ANMF built from the SVD of regularized Tyler's estimator [Pascal et al., 2014a] with regularization parameter γ selected manually to obtain the best results.; *iii*) $\hat{\Lambda}^{\text{G-MUSIC}}$ is the LR-ANMF using [Mestre and Lagunas, 2008] to estimate the quadratic forms associated to the CSP; *iv*) $\hat{\Lambda}^{\text{sMMSD}}$ denotes the LR-ANMF built from the MMSD estimator of LRCGs model; *v*) $\hat{\Lambda}^{\text{MMSD-OM}}$ stands for the LR-ANMF built from the MMSD estimator of LRCGo model; *vi*) $\hat{\Lambda}^{\text{MMSD-G}}$ is the LR-ANMF where the CSP is built from the MMSD estimator assuming a Gaussian model [Ben Abdallah et al., 2017]. For all the Bayesian estimators, we leverage the physical model of [Ward, 1994], that allows us to build a prior of the CSP basis $\bar{\mathbf{U}}$ from the SVD of the STAP covariance matrix model. We then consider a CIB prior (cf. definition 2.3.1) where the concentration parameter κ is set manually.

The tested cell contains the response of 10 moving targets to be detected in presence of ground clutter response (and eventually in presence of outliers in the secondary data). We test the aforementioned detectors on two scenarios: Figure 2.3 displays the output of the detectors in the standard situation ($n \gg p$), and at low sample support ($n = 2k \ll p$) with outliers in the secondary data. In the standard case, all of the detectors allow for target detection with apparently low false alarm rates. In the challenging setting, SCM and SFPE detectors are not able to correctly detect the targets. The G-MUSIC detector appears robust to outliers in terms of detection but leads to a visually higher noise floor (false alarm rate). Conversely, the Bayesian detectors still achieve interference rejection and reliable target detection, which illustrates the interest of introducing some prior information in an adaptive subspace estimation process. Notably, the MMSD-OM detector yields the cleanest detection map, probably thanks to its robustness to sample corruption by outliers.

2.6 Perspectives

Finally, we can point out some direct perspectives from this work:

- The parameters of the CGBL distributions were used to gather prior information in the context of Bayesian PCA. However, we did do not address their estimation, nor the question of their automatic selection (e.g. using mini-batches), which could be an interesting direction.
- Inspired by the M-estimation framework, we could propose robust subspace estimators following the structure from Theorem 2.4.1 with various function ρ. An approach following this direction is discussed in chapter 5. It would then be interesting to characterize the existence and performance of these generic subspace M-estimators in various settings.
- The LRCG model still leaves opened theoretical questions. For example, the phase transition could be predicted by following [Hong et al., 2018]. The study of intrinsic Cramér-Rao bounds could follow from [Besson et al., 2011]. A more complex issue concerns the characterization of the robustness of an estimation process when outliers are present [Lerman and Maunu, 2017].

2.7 Onward to the next chapter

This chapter was focused on the subspace recovery problem, while relying on specific statistical models to perform PCA with a Bayesian approach. These models notably involved mixtures of (conditional) Gaussian distributions. Starting from these two concluding remarks, we can give the following introduction to the next chapter:

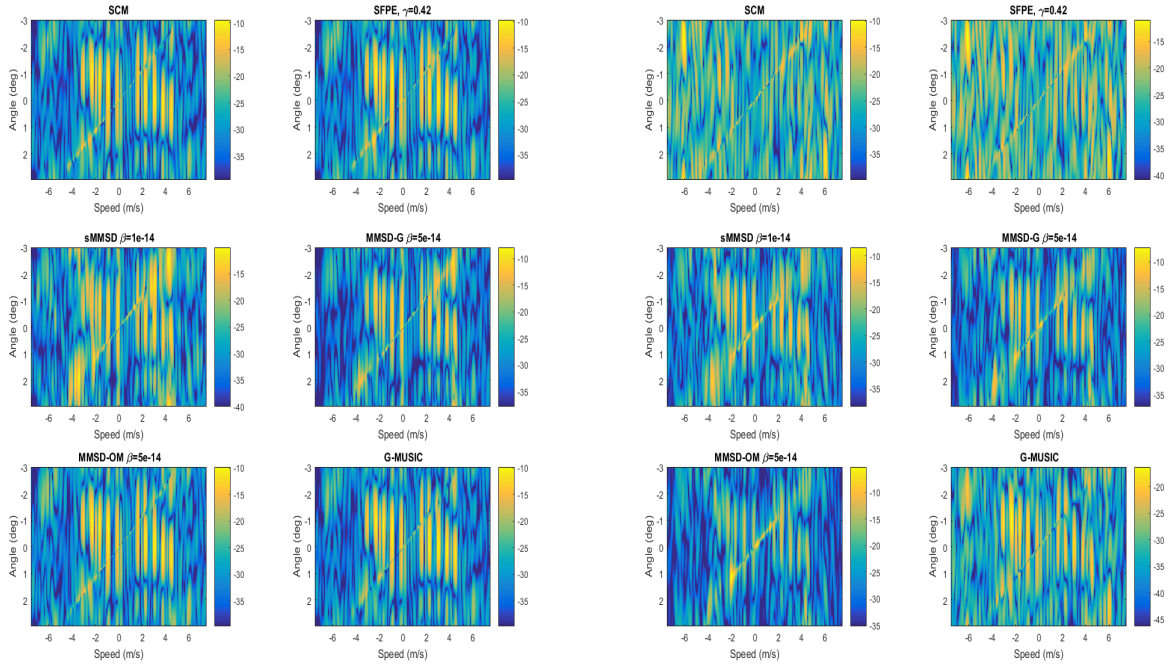


Figure 2.3: Output of various low-rank detectors on STAP data for $n = 397$ ($n > p$) (left), and $n = 2k$ ($n \ll p$) in presence of outliers (right). $k = 46$, $p = 256$.

- Chapter 3 will focus on covariance matrix problem estimation. The issue is still linked with subspace recovery, as we will notably conduct a statistical analysis for the EVD parameters of covariance matrices estimators (probabilistic PCA), as well as performance bounds for subspace estimation. However, we will also discuss some other problems, such as structured covariance matrix estimation.

- The statistical model considered in chapter 3 is the class of complex elliptically symetric distributions, which encompass compound Gaussian distributions as a special case. However, it will consider a single signal contribution, which cannot yield the mixture models from this chapter. Also, no priors on the parameter distribution will be involved (non-Bayesian setting).

- Though some estimation methods will be proposed, chapter 3 is quite focused on statistical analysis and performance bounds derivations.

3 | Robust covariance matrix estimation in elliptical models

Contents

3.1 Contributions of the chapter

This chapter focuses on covariance matrix estimation problems with a focus on statistical performance characterization. As discussed in the global picture, covariance matrix estimation can be a prelude to a PCA approach, i.e., performing a subspace recovery from the principal subspace of the estimate. However, it is worth mentioning that this step is also a fundamental problem on its own, as the covariance matrix is a core component of many statistical signal processing and machine learning methods. In this scope, we present several statistical analysis for a class of robust covariance matrix estimators (M-estimators) in a general model of elliptical distributions.

The first contribution details a new asymptotic characterization for the eigendecomposition of M-estimators. This work was conducted during the thesis of Gordana Drašković.
Related publications in the cv (page xi): [J5], [J8], [J9], [C10], [FC7], [FC11].

The second contribution concerns robust estimation methods for structured covariance matrices and their statistical analysis in the mismatched case. This work was conducted during the thesis of Bruno Mériaux.

Related publications in the cv (page xi): [J7], [C17], [C15], [C13], [FC9], [FC10].

The third contribution presented in this chapter concerns intrinsic Cramér-Rao bounds, i.e., performance bounds on Riemannian distances, for covariance matrix estimation in elliptical distributions.
Related publications in the cv (page xi): [J4], [X3].

3.2 Context overview

3.2.1 Complex elliptically symmetric distributions and M-estimators

There are many approaches to derive robust parametric estimation schemes. One of them is to consider a statistical model that is general enough to yield an accurate estimation, independently from the "true" underlying distribution of the data. A main challenge is then to manage the trade-off between the generality of the model and its practical representation (i.e., being simple enough to be handled). In this scope, complex elliptically symmetric (CES) distributions form a general family of circular multivariate distributions [Ollila et al., 2012a], parameterized by a mean vector $\boldsymbol{\mu}$ and a scatter matrix $\boldsymbol{\Sigma}$, which describes the correlations between the entries. This family encompasses notably the compound Gaussian distributions (used in chapter 2) and generalized Gaussian ones [Zhang et al., 2013, Pascal et al., 2013] as special cases. Another main interest is that CES includes many heavy-tailed distributions, which is useful to account for potential outliers. Thus we can expect these elliptical models to yield both robustness to model mismatches, and outliers in the sample set.

Definition 3.2.1. ***Elliptical models (CES distributions)***
Let $\boldsymbol{\Sigma} \in \mathcal{H}_p^{++}$ and $\boldsymbol{\mu} \in \mathbb{C}^p$. A vector $\mathbf{z}$ follows a (absolute continuous) CES distribution of center $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$, denoted $\mathcal{CES}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g_{\mathbf{z}})$, if it has the following p.d.f.:

$$f_{\mathbf{z}}(\mathbf{z}) \quad = \quad C|\boldsymbol{\Sigma}|^{-1}\, g_{\mathbf{z}}((\mathbf{z}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})) \tag{3.1}$$

where C is a normalization constant and $g_{\mathbf{z}} : [0,\infty) \to [0,\infty)$ is any function (called the density generator), ensuring that (3.1) defines a p.d.f.. Moreover, this vector admits the following stochastic representation

$$\mathbf{z} \overset{d}{=} \sqrt{\mathcal{Q}}\mathbf{A}\mathbf{u} + \boldsymbol{\mu} \tag{3.2}$$

where $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^H$, $\mathbf{u}$ is uniformly distributed on the complex sphere $\mathcal{U}_1^p$, and $\mathcal{Q}$ is a non-negative real random variable, called the modular variate, independent of $\mathbf{u}$ with a p.d.f. depending only on $g_{\mathbf{z}}$.

Note that from this definition, the Gaussian distribution $\mathbf{z} \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ appears as a special case with $g_{\mathbf{z}}(z) = e^{-z}$ and $C = \pi^{-p}$. The density generator $g_{\mathbf{z}}$ allows heavier or lighter tailed distributions to be described (cf. [Ollila et al., 2012a] for examples).

In the following, we will focus on the known mean case, which allows us to set $\boldsymbol{\mu} = \mathbf{0}$. In the general case, also note that (when existing) the covariance matrix $\mathbb{E}[\mathbf{z}\mathbf{z}^H]$ is proportional to the scatter matrix $\boldsymbol{\Sigma}$ (which always exists). This scaling mismatch is generally not an issue because most processes are insensitive to it (e.g., eigenvectors extracted for PCA). For this reason, we adopt the common abuse of denomination "covariance matrix estimation", while we technically estimate the scatter up to a scale factor.

When dealing with heavy-tailed distributed samples, it is well known that the traditional sample covariance matrix (SCM) usually fails to provide an accurate estimate. A solution to this problem is brought by M-estimators [Maronna, 1976, Tyler, 1987], that appear as generalized maximum likelihood estimators (MLE).

Definition 3.2.2. ***M-estimators of the scatter***
Let $\{\mathbf{z}_i\}_{i=1}^n$ be an n-sample of p-dimensional complex i.i.d. vectors with $\mathbf{z}_i \sim \mathcal{CES}(\mathbf{0}, \boldsymbol{\Sigma}, g_{\mathbf{z}})$. An

M-estimator of $\boldsymbol{\Sigma}$*, denoted by* $\hat{\boldsymbol{\Sigma}}$*, is defined by the solution of the following fixed-point equation*

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{z}_i^H \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^H \triangleq \mathcal{H}(\hat{\boldsymbol{\Sigma}}). \tag{3.3}$$

where u *is any real-valued weight function on* $[0, \infty)$ *that respects Maronna's conditions, ensuring existence and uniqueness of* (3.3) *[Maronna, 1976]. When these conditions are met, and for* $n > p$*, this estimator can be computed using the fixed-point algorithm* $\boldsymbol{\Sigma}_{t+1} = \mathcal{H}(\boldsymbol{\Sigma}_t)$ *(where* t *refers to the iteration index).*

When $u(t) = -g'_{\mathbf{z}}(t)/g_{\mathbf{z}}(t)$, (3.3) corresponds to the MLE of the scatter matrix parameter for $\mathbf{z} \sim \mathcal{CES}(\mathbf{0}, \boldsymbol{\Sigma}, g_{\mathbf{z}})$. However, u may not be related to $g_{\mathbf{z}}$, which is generally unknown in practice. A popular example is Tyler's estimator [Tyler, 1987, Pascal et al., 2008a], obtained with $u(t) = p/t$. Despite the potential mismatch, M-estimators ensure good performance in terms of estimation accuracy in the whole CES family (formally characterized in the following sections). Additionally, M-estimators present robustness to contamination by outliers [Maronna, 1976], which is why they are also usually referred to as robust estimators.

3.2.2 Current issues

M-estimators offer an interesting solution to robust covariance matrix estimation issues. Yet, there still remains some open problems, from which we can mention:

• **Statistical analysis**: The statistical characterization of the M-estimators is a complex issue because they are defined by fixed-point equations. While the SCM in a Gaussian setting follows a well-known Wishart distribution [Muirhead, 1982], the true distribution of the M-estimators remains unknown. Therefore, several works derived various characterizations for these estimators. Their asymptotic Gaussian distribution was derived in [Maronna, 1976, Tyler, 1982] and extended to the complex case in [Ollila et al., 2012a, Mahot et al., 2013]. Probably approximately correct (PAC) error bounds have been studied in [Soloveychik and Wiesel, 2015b]. Their analysis in the large random matrix regime (i.e. when both the number of samples and the dimension tends to infinity at the same rate) has been established in [Zhang et al., 2014, Couillet et al., 2015]. Yet, we still aim for a characterization that is as handy as the Wishart distribution in order to tune statistical processes. In this direction, an axis of response will be explored in section 3.3.

• **Low and insufficient support**: M-estimators require at least $n > p$ samples to be computed. Following [Reed et al., 1974], the general rule of thumb even suggests that $n > 2p$ is required in order to reach an accurate estimation. Both of these conditions can be difficult to meet for high dimensional data. An approach to overcome low sample support issues is to account for prior knowledge on the covariance matrix structure in the estimation scheme, i.e., reducing the degree of freedom of the estimation problem. Variations around this method were recently proposed for M-estimators [Soloveychik and Wiesel, 2014, Wiesel and Zhang, 2015, Soloveychik et al., 2016, Sun et al., 2016]. Section 3.4 will present our contributions in this scope. For insufficient sample support scenarios ($n < p$), a solution is brought by regularization methods, which will be discussed in the perspectives.

3.3 On the asymptotics of PCA with M-estimators

The parameters of an eigenvalue decomposition (EVD) of the second order statistics are ubiquitous in statistical analysis and signal processing. Notably, the eigenvalue decomposition (EVD) of M-estimators is involved in numerous processes, such as robust probabilistic PCA algorithms [Croux and Haesbroeck, 2000, Zhao and Jiang, 2006], as well as in the derivation of robust counterparts of low rank filters or detectors [Rangaswamy et al., 2004]. The eigenvalues of the scatter ma-

trix are also used in model order selection [Stoica and Selen, 2004, Terreaux et al., 2018], and functions of eigenvalues are involved in various applications such as regularization parameter selection [Ollila and Tyler, 2014b, Kammoun et al., 2018], detection [Ciuonzo et al., 2017], and classification [Bouveyron and Brunet-Saumard, 2014]. Hence, accurately characterizing the distribution of the M-estimators EVD represents an interest, both from the points of view of performance analysis and optimal process design. Towards the goal of characterizing these objects, we derived new asymptotics for the EVD parameters of M-estimators in elliptical models in [Drašković et al., 2019], from which the main results are summarized below.

3.3.1 Standard Asymptotic Regime

This section extends the analysis of [Kollo and Neudecker, 1993] (for the SCM) to the complex M-estimators in a general elliptical model. This asymptotic analysis also provides an extension of the results obtained in [Tyler, 1981, Boente, 1987, Croux and Haesbroeck, 2000] to the complex case, with some additional characterizations (cf. discussion following Theorem 3.3.1).

First, let us denote the EVD of the scatter matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} \overset{\text{EVD}}{=} \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \quad \text{with} \quad \begin{array}{l} \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathrm{St}(p,k), \\ \mathbf{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda}), \\ \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]\,. \end{array} \tag{3.4}$$

In order to avoid ambiguity in this definition, we assume ordered eigenvalues as $\lambda_1 > \ldots > \lambda_p > 0$, and an element of each $\mathbf{u}_j$ (e.g., the first entry) for $j = 1, \ldots, n$, can be assumed to be real positive. We then have the following Theorem:

Theorem 3.3.1. ***Standard asymptotics for M-estimators's EVD*** *[Drašković et al., 2019] Let $\hat{\mathbf{\Sigma}} \overset{\text{EVD}}{=} \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^H$ be an M-estimator (defined as the solution of the fixed point equation (3.3)) built from n samples drawn as $\mathbf{z} \sim \mathcal{CES}(\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{z}})$. The asymptotic distribution of the EVD of $\hat{\mathbf{\Sigma}}$ is given by*

$$\begin{cases} \sqrt{n}\left(\sigma\hat{\boldsymbol{\lambda}}^M - \boldsymbol{\lambda}\right) \overset{d}{\to} \mathcal{N}\left(\mathbf{0}, \vartheta_1\mathbf{\Lambda}^2 + \vartheta_2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right), \\ \sqrt{n}\mathbf{\Pi}_j^{\perp}\hat{\mathbf{u}}_j^M \overset{d}{\to} \mathcal{CN}\left(\mathbf{0}, \mathbf{\Xi}_j\right). \end{cases} \tag{3.5}$$

where

$$\mathbf{\Xi}_j = \vartheta_1\lambda_j\left(\mathbf{U}\mathbf{\Lambda}(\lambda_j\mathbf{I} - \mathbf{\Lambda})^+\right)^2\mathbf{U}^H \tag{3.6}$$

with $\mathbf{\Pi}_j^{\perp} = \mathbf{I} - \mathbf{u}_j\mathbf{u}_j^H$, the scalar factor $\sigma > 0$ it the solution of $\mathbb{E}[\Psi(\sigma t)] = p$ with $\Psi(\sigma t) = u(\sigma t)\sigma t$ and $t = \mathbf{z}^H\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}$, and the constants $\vartheta_1 > 0$ and $\vartheta_2 > -\vartheta_1/p$ are given by

$$\begin{array}{l} \vartheta_1 = c_M^{-2}a_M p(p+1), \\ \vartheta_2 = (c_M - p^2)^{-2}(a_M - p^2) - c_M^{-2}a_M(p+1), \end{array} \tag{3.7}$$

where $a_M = E[\Psi^2(\sigma\mathcal{Q})]$ and $c_M = E[\Psi'(\sigma\mathcal{Q})\sigma\mathcal{Q}] + p^2$.

The results given in Theorem 3.3.1 are interesting since, besides the variance of each eigenvalue, they provide the correlation between them. Note that for a Wishart-distributed matrix this correlation is equal to zero, as shown in [Kollo and Neudecker, 1993] for the real case. Conversely, Theorem 3.3.1 shows that the eigenvalues of an M-estimator are asymptotically correlated, as stated in [Croux and Haesbroeck, 2000] (but not explicitly characterized). This correlation depends on the second scale parameter ϑ_2. Concerning the eigenvectors, note that the covariance depends only on ϑ_1 since $\mathbf{u}_j$ is scale invariant w.r.t. to the covariance matrix (cf. [Mahot et al., 2013] for more details).

3.3.2 Gaussian core Wishart equivalent

This second result represents a new approach based on [Drašković and Pascal, 2018]. A new characterization is proposed to show that the EVD parameters of M-estimators are asymptotically concentrated around a Wishart equivalent, with a variance that is significantly lower than the one of the standard asymptotic regime (i.e., derived around the true expected values). Thus, it quantifies when it is acceptable to directly rely on well-established results on the EVD of Wishart-distributed matrices for characterizing the EVD of M-estimators. First, let us define two quantities related to the hidden Gaussian cores of CES vectors.

Definition 3.3.1. ***Gaussian cores of CES vectors*** *[Drašković and Pascal, 2018]*
Let $\mathbf{z} \sim \mathcal{CES}(\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{z}})$. This vector has a representation analogous to (3.2)*, given as*

$$\mathbf{z} \overset{d}{=} \sqrt{\mathcal{Q}}\mathbf{A}\mathbf{g}/\|\mathbf{g}\|, \tag{3.8}$$

where $\mathbf{g} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. The vector $\mathbf{x} = \mathbf{A}\mathbf{g}$ is referred to as the Gaussian-core of $\mathbf{z}$.

Definition 3.3.2. ***Gaussian cores Wishart equivalent (GCWE)*** *[Drašković and Pascal, 2018]*
Let $\{\mathbf{z}_i\}_{i=1}^n$ be a n-sample drawn as $\mathbf{z} \sim \mathcal{CES}(\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{z}})$ and denote $\{\mathbf{x}_i\}_{i=1}^n$ their Gaussian cores from the representation $\mathbf{z}_i = \sqrt{\mathcal{Q}_i}/\|\mathbf{x}_i\|\mathbf{A}\mathbf{x}_i$ (cf. Definition 3.3.1). Let $\hat{\mathbf{\Sigma}}$ be an M-estimator built with $\{\mathbf{z}_i\}_{i=1}^n$ using (3.3)*. The SCM built from the Gaussian cores, i.e.*

$$\hat{\mathbf{\Sigma}}_{\text{GCWE}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^H \tag{3.9}$$

is referred to as Gaussian Core Wishart Equivalent (GCWE) of $\hat{\mathbf{\Sigma}}$.

Note that the GCWE cannot be computed in practice. It is a theoretical quantity distributed according to a Wishart distribution. The asymptotic distribution of the difference between an M-estimator and its GCWE is derived in [Drašković and Pascal, 2018]. Following this result, we derived the following theorem:

Theorem 3.3.2. ***GCWE for M-estimators's EVD*** *[Drašković et al., 2019]*
Let $\hat{\mathbf{\Sigma}} \overset{\text{EVD}}{=} \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^H$ be an M-estimator as in (3.3) *built from n samples drawn as $\mathbf{z} \sim \mathcal{CES}(\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{z}})$. Let $\hat{\mathbf{\Sigma}}_{\text{GCWE}} \overset{\text{EVD}}{=} \hat{\mathbf{U}}^{\text{GCWE}}\hat{\mathbf{\Lambda}}^{\text{GCWE}}\left(\hat{\mathbf{U}}^{\text{GCWE}}\right)^H$ be its GCWE (Definition 3.3.2). The asymptotic distribution of the difference between the EVD parameters of $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Sigma}}_{\text{GCWE}}$ is given by*

$$\begin{cases} \sqrt{n}\left(\sigma\hat{\boldsymbol{\lambda}}^M - \hat{\boldsymbol{\lambda}}^{\text{GCWE}}\right) \overset{d}{\to} \mathcal{N}\left(\mathbf{0}, \sigma_1\mathbf{\Lambda}^2 + \sigma_2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right), \\ \sqrt{n}\mathbf{\Pi}_j^{\perp}\left(\hat{\mathbf{u}}_j^M - \hat{\mathbf{u}}_j^{\text{GCWE}}\right) \overset{d}{\to} \mathcal{CN}\left(\mathbf{0}, \sigma_1/\vartheta_1\mathbf{\Xi}_j\right), \end{cases} \tag{3.10}$$

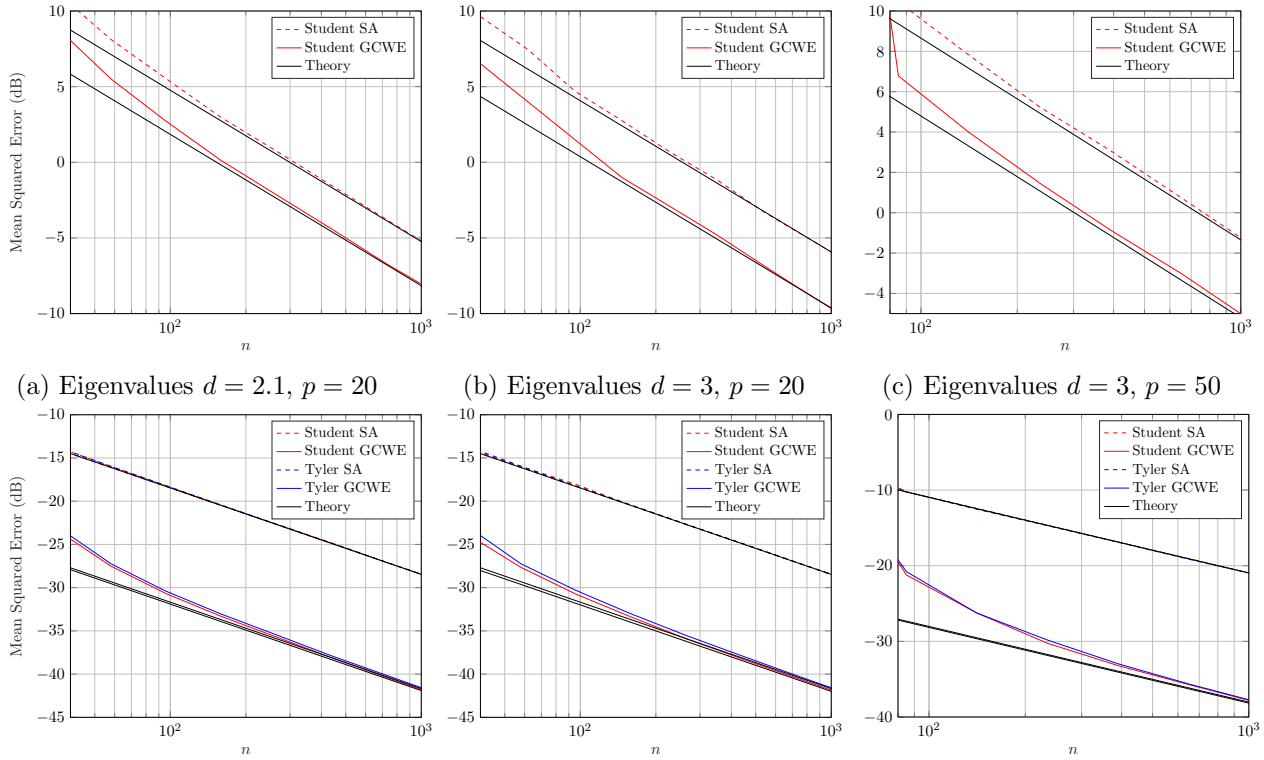
with $\mathbf{\Xi}_j$, σ, a_M and c_M are defined in Theorem 3.3.1, and $b_M = E[\Psi(\sigma\mathcal{Q})\|\mathbf{g}\|^2]$, and where the coefficients σ_1 and σ_2 are given by

$$\begin{aligned} \sigma_1 &= (a_M p(p+1) + c(c - 2b_M))/c_M^2, \\ \sigma_2 &= \vartheta_2 - 2p(b_M - c_M)/(c_M(c_M - p^2)). \end{aligned} \tag{3.11}$$

This theorem characterizes the asymptotic variance of the EVD of an M-estimator compared to the one of its GCWE. Interestingly, it shows that their covariance structure is the same as in the standard asymptotic regime, and differs only through the variance scales (σ_1, σ_2) (instead of $(\vartheta_1, \vartheta_2)$). As noted in [Drašković and Pascal, 2018], the total variance captured by the GCWE factors is much smaller than the standard one. For example, Table 3.1 displays these factors for two M-estimators, assuming Student t-distributed data with degree of freedom (DoF) parameter d (whose p.d.f. is given by (3.1) with $g_{\mathbf{z}}(x) = (1 + 2x/d)^{-(p+d/2)}$). In this case, the factors σ_1 and ϑ_1 differ by an order $1/p$. This observation is also confirmed by the validation simulations displayed in Figure 3.1. In conclusion, these results support the idea that an underlying Wishart distribution can offer a better approximation for characterizing the distribution of the M-estimator's EVD.

	Student's M-estimator (MLE)	Tyler's M-estimator
SA	$\vartheta_1 = (p + d/2 + 1)/(p + d/2)$	$\vartheta_1 = (p+1)/p$
	$\vartheta_2 = 2(p + d/2 + 1)/(d(p + d/2))$	$\vartheta_2 = -(p+1)/p^2$
GCWE	$\sigma_1 = 1/(p + d/2)$	$\sigma_1 = 1/p$
	$\sigma_2 = 2(p + d/2 + 1)/(d(p + d/2))$	$\sigma_2 = (p-1)/p^2$

Table 3.1: Coefficients ϑ_1, ϑ_2, σ_1 and σ_2 for Student's and Tyler's M-estimator with t-distributed data (SA stands for Standard asymptotic while GCWE refers to as GCWE asymptotic).



(a) Eigenvalues $d = 2.1$, $p = 20$ (b) Eigenvalues $d = 3$, $p = 20$ (c) Eigenvalues $d = 3$, $p = 50$

(d) Eigenvectors $d = 2.1$, $p = 20$ (e) Eigenvectors $d = 3$, $p = 20$ (f) Eigenvectors $d = 3$, $p = 50$

Figure 3.1: Validation of theoretical results on eigenvalues (top) and eigenvectors (bottom) for Student's and Tyler's M-estimator built with Student t-distributed data with various DoF d and dimensions p.

3.3.3 Practical use of the result

By establishing two different asymptotic regimes, we have shown that the behavior of the EVD parameters of M-estimators can be more accurately characterized by an equivalent Wishart model than by their standard asymptotic Gaussian distribution. This approximation allows us to leverage well established results (e.g., [Muirhead, 1982, Zanella et al., 2009]), and offers a thinner analysis compared to the asymptotic Gaussian results. In [Drašković et al., 2019], this idea was illustrated on several

examples[1]:

1. We addressed the complex issue of characterizing the intrinsic bias [Smith, 2005] of M-estimators in the CES context. This quantity has been studied in [Smith, 2005] for the SCM in the Gaussian context thanks to the distribution of the eigenvalues of a Wishart matrix [Muirhead, 1982]. Extending this analysis to M-estimators in the general CES context represents, at first sight, an intractable problem because of their unknown exact distribution. However, the established convergence of the eigenvalues of an M-estimator toward their GCWE counterpart allows an accurate approximation of this intrinsic bias to be derived;

2. In the context of model order selection (i.e., rank estimation) from non-Gaussian samples, we showed that the use of M-estimators (rather than the SCM) in theoretic criteria derived for Gaussian models [Akaike, 1974, Wax and Ziskind, 1989] yields the same results as the one obtained with the theoretical GCWE. Again, this justifies a plug-in approach (using M-estimators in processes derived under the Wishart assumption), instead of a complete re-derivation that would require to assume an exact CES distribution;

3. The performance of low rank filters [Ginolhac and Forster, 2010] built from M-estimators were derived in the same way (i.e., approached by the one of their GCWE) to illustrate that the method also holds for adaptive processes based on the eigenvectors.

We can also mention that an identical approach was conducted in [Drašković et al., 2020] to study robust detection methods[2]. We showed that, from CES distributed samples, the distribution of a detection statistic (adaptive normalized matched filter, Kelly's GLRT, and Rao test) built with plug-in M-estimators can be accurately approximated by the distribution of the same statistic built from the SCM of an equivalent Gaussian core. The loss due to this approximation was theoretically derived and shown to be negligible in most cases. This explicit equivalent statistic is especially interesting since it allowed us to tune robust plug-in detectors with well established results from the Gaussian detection framework.

3.4 Robust estimation of structured covariance matrices

Besides being Hermitian positive (semi-)definite, the covariance matrix can exhibit a specific structure depending on the considered application. This structure can often be determined by physical considerations (e.g. symmetries) on the data acquisition system. We can give the following taxonomy of usual covariance matrices structures:

- Linear structures, i.e., when the covariance matrix belongs to a set of the form $\mathcal{S} = \{\mathbf{\Sigma} = \sum_{r=1}^{d} \alpha_r \mathbf{B}_r, \alpha_r \in \mathbb{R}\}$, where $\{\mathbf{B}_r\}_{r=1}^{d}$ is a known basis of the considered set. These structures encompass notably Toeplitz, banded, and sum-of-rank-1 matrices (i.e. factor models with known factors).

- Group symmetric structures, i.e., when the covariance matrix belongs to a set in of form $\mathcal{S}_{\mathcal{H}} = \{\mathbf{\Sigma} | \mathbf{\Sigma} = \mathbf{H}\mathbf{\Sigma}\mathbf{H}^H, \forall \mathbf{H} \in \mathcal{H}\}$, where $\mathcal{H} = \{\mathbf{H}_k\}_{i=1}^{d}$ is a multiplicative group of orthogonal matrices. These structures encompass notably persymmetric and circulant matrices.

- Spectral structures, i.e., when the eigenvalues of the covariance matrix satisfy a certain property. These structures encompass notably positive plus scaled identity (lower-bounded eigenvalues), low-rank plus scaled identity (identical eigenvalues after certain index), and matrices with a constrained condition number.

[1] Interested readers can find the details in section 2.3 of the annexes.
[2] Interested readers can find the details in section 2.4 of the annexes.

- Kronecker products, i.e., when the covariance matrix is expressed as $\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B}$. In some cases, the covariance matrix can also be expressed as a Kronecker products of structured matrices.

Taking this prior knowledge on the covariance matrix structure in the estimation process reduces the degree of freedom of the estimation problem, which is especially interesting when the sample size is low.

An approach to generalize M-estimators to the case of structured matrices is solve for the generalized log-likelihood function under structural constraint, i.e. focus on the problem

$$\begin{array}{ll} \underset{\mathbf{\Sigma}}{\text{minimize}} & \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{z}_i \mathbf{\Sigma}^{-1} \mathbf{z}_i) + \ln|\mathbf{\Sigma}| \\ \text{subject to} & \mathbf{\Sigma} \in \mathcal{S} \end{array} \tag{3.12}$$

with $\rho'(t) = -u(t)$ (to match definition 3.2.2), and where $\mathcal{S}$ is the considered set of structured Hermitian positive definite matrices. This formulation leads to non-trivial issues because the problem (3.12) is non-convex, even in the unstructured case. Nevertheless, this objective function holds hidden (*geodesic*) convexity properties [Wiesel, 2012a], which is preserved on some specific structure sets (e.g., group symmetric ones [Pailloux et al., 2011, Soloveychik et al., 2016]). This property can help to establish existence/uniqueness of specific solutions, but the general problem of robust structured covariance matrix estimation still requires a case by case study. Several methodologies were considered to tackle this issue:

- Reparameterization: reformulating the problem (3.12) using a parameter $\boldsymbol{\theta}$ and a corresponding mapping $\mathbf{\Sigma} \triangleq \boldsymbol{\mathcal{R}}(\boldsymbol{\theta}) \in \mathcal{S}$, then leveraging an optimization method directly on $\boldsymbol{\theta}$. In this scope, [Sun et al., 2016] proposed majorization-minimization algorithms for various structures. A similar methodology was applied for structures involving Kronecker product of low rank matrices in [Breloy et al., 2016b]. This solution has few theoretical guarantees (apart from convergence to a local minimum), but was shown to be quite accurate in practice.

- Relaxations: considering an alternative to (3.12) that still allows some properties (accuracy, robustness,...) to be ensured. In this scope, [Soloveychik and Wiesel, 2014] proposed a convex relaxation of Tyler's cost function, which is suited to matrices belonging to convex sets.

- Projections: projecting an M-estimator on the set of interest. This 2-step procedure is generally sub-optimal, but can benefit from a low computational cost, and offers some theoretical guarantees.

In the following we present a projection-based method for convex sets. The estimation procedure was proposed in [Meriaux et al., 2019], and has the advantage of having nice asymptotic properties.

3.4.1 Structured scatter matrix estimator (SESAME)

In this section, we assume that the scatter matrix belongs to a convex subset $\mathcal{S}$ of Hermitian matrices (e.g., Toeplitz, persymmetric, or banded), for which there exists a one-to-one differentiable mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\mathcal{R}}(\boldsymbol{\theta})$ from $\mathbb{R}^d$ to $\mathcal{S}$. We propose a robust two-step estimation procedure of $\boldsymbol{\theta}$ inspired by [Ottersten et al., 1998]. An important note is that we will theoretically study the robustness of the method to a mismatch scenario, i.e.:

- The data set $\{\mathbf{z}_i\}_{i=1}^n$ is distributed according to $\mathbf{z} \sim \mathcal{CES}(\mathbf{0}, \mathbf{\Sigma}_e, g)$, with $\mathbf{\Sigma}_e = \boldsymbol{\mathcal{R}}(\boldsymbol{\theta}_e) \in \mathcal{S}$, its corresponding p.d.f. is denoted $f_{\mathbf{z}}$.

- The estimation procedure is conducted assuming the model $\mathbf{z} \sim \mathcal{CES}(\mathbf{0}, \boldsymbol{\mathcal{R}}(\boldsymbol{\theta}), g_{\text{mod}})$, the corresponding p.d.f. is denoted f_{mod}.

Indeed, the assumed model (g_{mod}) can be different from the true one (g) in practice. A main example would be assuming Gaussian distributed samples, when they are actually not. Other incorrect

assumptions can also come from the selection of shape/scale parameters in the density generator (e.g. the selected degrees of freedom for a t-distribution).

The SESAME algorithm proceeds as follows:

Step 1: compute $\hat{\boldsymbol{\Sigma}}_m$, the M-estimator corresponding to the MLE of the assumed model (cf. definition 3.3 using the function $u_{\text{mod}}(s) = -g'_{\text{mod}}(s)/g_{\text{mod}}(s)$). Notice that this provides a consistent estimator of $\sigma^{-1}\boldsymbol{\Sigma}_e$, with σ defined in Theorem 3.3.1 [Ollila et al., 2012a].

Step 2: estimate $\boldsymbol{\theta}$ by solving

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{J}_{\hat{\boldsymbol{\Sigma}}_m,\hat{\boldsymbol{\Sigma}}}(\boldsymbol{\theta}) \tag{3.13}$$

with

$$\mathcal{J}_{\hat{\boldsymbol{\Sigma}}_m,\hat{\boldsymbol{\Sigma}}}(\boldsymbol{\theta}) \;=\; \kappa_1 \mathrm{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\Sigma}}_m - \boldsymbol{\mathcal{R}}(\boldsymbol{\theta})\right)\hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\Sigma}}_m - \boldsymbol{\mathcal{R}}(\boldsymbol{\theta})\right)\right) + \kappa_2\left[\mathrm{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\Sigma}}_m - \boldsymbol{\mathcal{R}}(\boldsymbol{\theta})\right)\right)\right]^2, \tag{3.14}$$

where $\hat{\boldsymbol{\Sigma}}$ refers to any consistent estimator of $\boldsymbol{\Sigma}_{\text{e}}$ up to a scale factor, such as for instance $\hat{\boldsymbol{\Sigma}}_m$, $\kappa_2 = \kappa_1 - 1$, and $\kappa_1 = \mathbb{E}_{f_{\text{mod}}}\left[\Psi^2_{\text{mod}}\left(|\mathbf{t}_{\text{mod}}|^2\right)\right]/(m(m+1)) \neq 0$ where $\mathbf{t}_{\text{mod}} \sim \mathbb{C}\mathcal{ES}_m(\mathbf{0}, \mathbf{I}, g_{\text{mod}})$, and where f_{mod} is the *assumed* p.d.f. of the data. The criterion $\mathcal{J}_{\hat{\boldsymbol{\Sigma}}_m,\hat{\boldsymbol{\Sigma}}}(\boldsymbol{\theta})$ is strongly related to the Fisher information metric derived for CES distributions in [Besson and Abramovich, 2013]. Given $\hat{\boldsymbol{\Sigma}}_m$ and $\hat{\boldsymbol{\Sigma}}$, the function $\mathcal{J}_{\hat{\boldsymbol{\Sigma}}_m,\hat{\boldsymbol{\Sigma}}}(\boldsymbol{\theta})$ is convex w.r.t $\boldsymbol{\mathcal{R}}(\boldsymbol{\theta})$. Therefore, for the desired convex set $\boldsymbol{\Sigma} \in \mathcal{S}$, the minimization of (3.13) w.r.t. $\boldsymbol{\mathcal{R}}(\boldsymbol{\theta})$ is a convex problem that admits a global minimizer, yielding a solution $\hat{\boldsymbol{\theta}}$ thanks to the one-to-one mapping. Some practical implementations for holding the positiveness constraint $\boldsymbol{\mathcal{R}}(\boldsymbol{\theta}) \succcurlyeq \mathbf{0}$ are discussed in [Meriaux et al., 2019][3].

Step 3 (optional): Perform the recursions

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \arg\min_{\boldsymbol{\theta}} \mathcal{J}_{\hat{\boldsymbol{\Sigma}}_m,\hat{\boldsymbol{\Sigma}}^{(k)}}(\boldsymbol{\theta}) \quad \text{with } \hat{\boldsymbol{\Sigma}}^{(k)} = \boldsymbol{\mathcal{R}}\left(\hat{\boldsymbol{\theta}}^{(k)}\right), \quad \text{for } k = 1, \ldots, N_{\text{it}}, \tag{3.15}$$

The intuition behind this refinement is the following: The estimate $\hat{\boldsymbol{\Sigma}}$ can be any consistent estimate up to a scaling factor. Intuitively (and also experienced in practice), the more accurate the estimator $\hat{\boldsymbol{\Sigma}}$, the more accurate the solution $\hat{\boldsymbol{\theta}}$. Since SESAME will be shown to be consistent, this leads naturally to a recursive procedure, where the minimized norm is refined at each step by updating $\hat{\boldsymbol{\Sigma}}$ with the previously computed $\boldsymbol{\mathcal{R}}(\hat{\boldsymbol{\theta}})$.

3.4.2 Asymptotic analysis of SESAME

Pseudo-parameter and consistency

Theorem 3.4.1. ***Consistency of SESAME*** *[Meriaux et al., 2019]*
The SESAME estimate $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_c$ such that $\mathrm{vec}\left(\boldsymbol{\mathcal{R}}(\boldsymbol{\theta}_c)\right) = \boldsymbol{\sigma}_c \triangleq \sigma^{-1}\boldsymbol{\sigma}_e = \sigma^{-1}\mathrm{vec}\left(\boldsymbol{\mathcal{R}}(\boldsymbol{\theta}_e)\right)$. *Likewise, $\boldsymbol{\mathcal{R}}\left(\hat{\boldsymbol{\theta}}\right)$ is a consistent estimator of $\sigma^{-1}\boldsymbol{\mathcal{R}}(\boldsymbol{\theta}_e)$.*

With a potential model misspecification, the so-called *pseudo-true* parameter vector, $\boldsymbol{\theta}_0$, is classically introduced for an asymptotic analysis [White, 1982, Richmond and Horowitz, 2015, Fortunati et al., 2016, Mennad et al., 2018]. The latter is defined as the minimizer of the Kullback-Leibler divergence (KLD) between the *true* and the *assumed* models, i.e.,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \mathcal{D}\left(f_{\text{mod}} \| f_{\mathbf{z}}\right) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{f_{\text{mod}}}\left[\log f_{\mathbf{z}}(\mathbf{z}; \boldsymbol{\theta})\right], \tag{3.16}$$

where $\mathcal{D}\left(f_{\text{mod}} \| f_{\mathbf{z}}\right) \triangleq \mathbb{E}_{f_{\text{mod}}}\left[\log \dfrac{f_{\text{mod}}(\mathbf{z}; \boldsymbol{\theta}_e)}{f_{\mathbf{z}}(\mathbf{z}; \boldsymbol{\theta})}\right]$. In the following, we always assume the existence and the uniqueness of the *pseudo-true* parameter vector, $\boldsymbol{\theta}_0$ (the reader is referred to [Fortunati et al., 2016] for necessary and sufficient conditions).

[3] Interested readers can find the details in section 2.5 of the annexes.

Proposition 3.4.1. *The pseudo-true parameter vector, $\boldsymbol{\theta}_0$, is equal to $\boldsymbol{\theta}_c$. Thus, the SESAME estimate, $\hat{\boldsymbol{\theta}}$, given by* (3.13)*, is a consistent estimator of $\boldsymbol{\theta}_0$ such that $\boldsymbol{\theta}_0 = arg\min_{\boldsymbol{\theta}} \mathcal{D}\left(f_{mod} \| f_{\mathbf{z}}\right)$.*

Efficiency in the mismatched framework

In the matched context, the Cramér-Rao Bound (CRB) is a lower bound of the variance of any unbiased estimator (which corresponds then to the Mean Square Error) of a deterministic parameter. Such an estimator is said to be (asymptotically) efficient if its variance reaches the CRB for an (in)finite number of samples. Likewise, under misspecified models, the Misspecified Cramér-Rao Bound (MCRB) is defined as a lower bound of the variance of any unbiased estimator $\hat{\boldsymbol{\theta}}_{g_{\text{mod}}}$ of $\boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is actually the *pseudo-true* parameter vector [Fortunati et al., 2016, Mennad et al., 2018]. Specifically, we have

$$\operatorname{Var}\left(\hat{\boldsymbol{\theta}}_{g_{\text{mod}}}\right) \succeq \frac{1}{n}\mathbf{A}^{-1}\left(\boldsymbol{\theta}_0\right)\mathbf{B}\left(\boldsymbol{\theta}_0\right)\mathbf{A}^{-1}\left(\boldsymbol{\theta}_0\right) \triangleq \frac{1}{n}\mathbf{MCRB}, \tag{3.17}$$

where, for all $k,\ell = 1,\ldots,P$, $\left[\mathbf{B}\left(\boldsymbol{\theta}_0\right)\right]_{k,\ell} = \mathbb{E}_{f_{\mathbf{z}}}\left[\left.\frac{\partial \log f_{\text{mod}}\left(\mathbf{z};\boldsymbol{\theta}\right)}{\partial\theta_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left.\frac{\partial \log f_{\text{mod}}\left(\mathbf{z};\boldsymbol{\theta}\right)}{\partial\theta_\ell}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$ and $\left[\mathbf{A}\left(\boldsymbol{\theta}_0\right)\right]_{k,\ell} = \mathbb{E}_{f_{\mathbf{z}}}\left[\left.\frac{\partial^2 \log f_{\text{mod}}\left(\mathbf{z};\boldsymbol{\theta}\right)}{\partial\theta_k\partial\theta_\ell}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$. For the proposed estimation method, we have the following theorem:

Theorem 3.4.2. ***(mismatched-)Efficiency of SESAME*** *[Meriaux et al., 2019]*
Let $\hat{\boldsymbol{\theta}}_n$ be the SESAME estimate computed from n i.i.d. observations, then $\hat{\boldsymbol{\theta}}_n$ is asymptotically efficient in the mismatched sense, i.e.,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{MCRB}\right), \tag{3.18}$$

with

$$\mathbf{MCRB} = \vartheta_1\mathbf{C}^{-1} + \vartheta_2\mathbf{C}^{-1}\mathbf{D}\mathbf{C}^{-1} = \left(\vartheta_1^{-1}\mathbf{C} - \frac{\vartheta_2}{\vartheta_1(\vartheta_1 + m\vartheta_2)}\mathbf{D}\right)^{-1}, \tag{3.19}$$

where

$$\begin{cases} \mathbf{C} = \boldsymbol{\mathcal{J}}\left(\boldsymbol{\theta}_0\right)^H \mathbf{W}_0^{-1}\boldsymbol{\mathcal{J}}\left(\boldsymbol{\theta}_0\right), \\ \mathbf{D} = \boldsymbol{\mathcal{J}}\left(\boldsymbol{\theta}_0\right)^H \mathbf{U}_0\boldsymbol{\mathcal{J}}\left(\boldsymbol{\theta}_0\right), \end{cases} \tag{3.20}$$

in which $\mathbf{W}_0 = \boldsymbol{\Sigma}_0^T \otimes \boldsymbol{\Sigma}_0$, $\mathbf{U}_0 = \operatorname{vec}\left(\boldsymbol{\Sigma}_0^{-1}\right)\operatorname{vec}\left(\boldsymbol{\Sigma}_0^{-1}\right)^H$, $\beta_1 = \vartheta_1\kappa_1^2$, $\beta_2 = \vartheta_1\kappa_2\left(2\kappa_1 + m\kappa_2\right) + \vartheta_2\left(\kappa_1 + m\kappa_2\right)^2$, $\left.\frac{\partial\boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}} \triangleq \boldsymbol{\mathcal{J}}\left(\boldsymbol{\theta}\right)$ refers to the Jacobian matrix of $\boldsymbol{\sigma}(\boldsymbol{\theta})$ evaluated in $\boldsymbol{\theta}$, and where the coefficients ϑ_1 and ϑ_2 are given in (3.7)*.*

Notice that, as a corollary from this theorem, the SESAME estimator is efficient in the matched case, i.e., when $g = g_{\text{mod}}$.

3.4.3 Simulations

First, we illustrate the theoretical analysis on SESAME performance under misspecifications. For the simulations, we choose a scatter matrix with an Hermitian Toeplitz structure. We consider a scenario where the *true* p.d.f is a Weibull distribution and we assume a Gaussian model for the estimation. Thus, we have

$$\begin{cases} g\left(t\right) = t^{s-1}\exp\left(-\frac{t^s}{b}\right) \quad , \quad \text{with} \;\; b = 2 \qquad \text{and} \;\; s = 0.8, \\ g_{\text{mod}}\left(t\right) = \exp\left(-t\right). \end{cases}$$

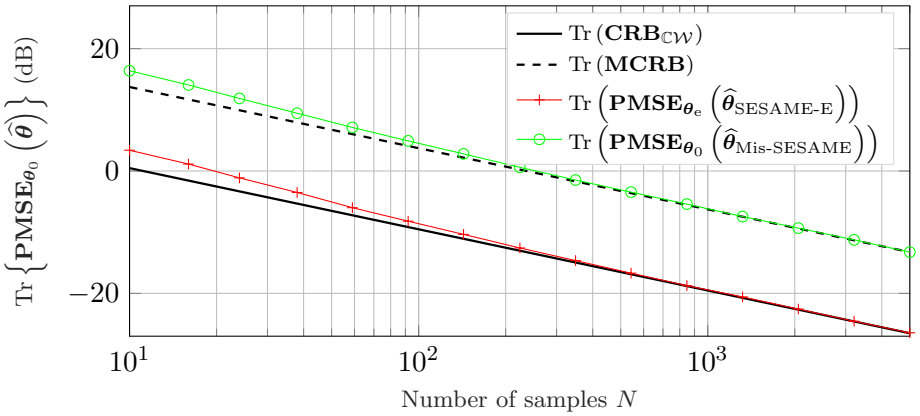


Figure 3.2: PMSE of SESAME procedures, *true* p.d.f. is Weibull distribution with $b = 2$ and $s = 0.8$ and *assumed* model is Gaussian. For $m = 5$, the *true* scatter matrix has is Toeplitz and its first row is $\left[2, \rho, \rho^2, \ldots, \rho^{m-1}\right]$, with $\rho = 0.8 + 0.3i$.

The estimate obtained by SESAME under misspecification is referred to as $\widehat{\boldsymbol{\theta}}_{\text{Mis-SESAME}}$ whereas the one computed in the matched case is denoted by $\widehat{\boldsymbol{\theta}}_{\text{SESAME-E}}$. We also compare the performance of SESAME under both matched and mismatched models with the CRB and the MCRB. To draw the comparison, we define the Pseudo Mean Square Error (PMSE) w.r.t the *pseudo*-parameter $\boldsymbol{\theta}_0$ by $\mathbf{PMSE}_{\boldsymbol{\theta}_0}\left(\widehat{\boldsymbol{\theta}}\right) = \mathbb{E}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T\right]$. In Figure 3.2, the asymptotic variance of the SESAME estimates under both matched and mismatched models reach the corresponding CRB derived in either matched or mismatched scenarios, i.e., the (mismatched-)efficiency of the algorithm is verified. The unbiasedness as well as the consistency can be also indirectly observed.

Second, we consider the matched case ($g = g_{\text{mod}}$) where the covariance matrix is Toeplitz and the data follows a t-distribution. In Figure 3.3, we compare the performance of SESAME to the state of the art: RCOMET from [Meriaux et al., 2017], COCA from [Soloveychik and Wiesel, 2014] and Constrained Tyler from [Sun et al., 2016]. The three methods are based on the Tyler's scatter estimator [Tyler, 1987] using normalized observations $\tilde{\mathbf{z}}_n = \mathbf{z}_n / \|\mathbf{z}_n\|$. It should be noted that, for Constrained Tyler, Algorithm 3 in [Sun et al., 2016] derived for real-valued PSD Toeplitz matrices cannot be directly applied. However, the Vandermonde factorization of PSD Toeplitz matrices allows us to use Algorithm 2 of [Sun et al., 2016]. In this algorithm, the set of PSD Toeplitz matrices is parameterized by $\mathcal{S} = \left\{\mathbf{R} \mid \mathbf{R} = \mathbf{APA}^H\right\}$ through the unknown diagonal matrix $\mathbf{P} \succeq \mathbf{0}$ and with $\mathbf{A} = [\mathbf{a}(-90°), \mathbf{a}(-88°), \cdots, \mathbf{a}(86°), \mathbf{a}(88°)]$, where $\mathbf{a}(\theta) = \left[1, e^{-j\pi\sin(\theta)}, \cdots, e^{-j\pi(m-1)\sin(\theta)}\right]^T$. Finally, we also study the estimate obtained by averaging the real and imaginary parts of diagonals of the unstructured ML estimator, which corresponds to the Euclidean projection onto the Toeplitz set. The asymptotic efficiency of our estimator is checked in Figure 3.3: its MSE reaches the CRB as n increases. RCOMET, Constrained Tyler and COCA do not reach this bound since they do not take into account the underlying distribution of the data. In addition, the asymptotic unbiasedness of SESAME as well as those of the other algorithms can be observed.

3.5 Intrinsic Cramér-Rao bounds in elliptical models

Cramér-Rao lower bounds (CRLBs) are ubiquitous tools in statistical signal processing, as they characterize the optimum performances in terms of mean squared error that can be achieved for a given parametric estimation problem. However, the classical Cramér-Rao analysis provides a lower bound on the mean squared error, while this criterion may not be the most appropriate for characterizing the performance in a given context. Especially, for parameters living in a mani-

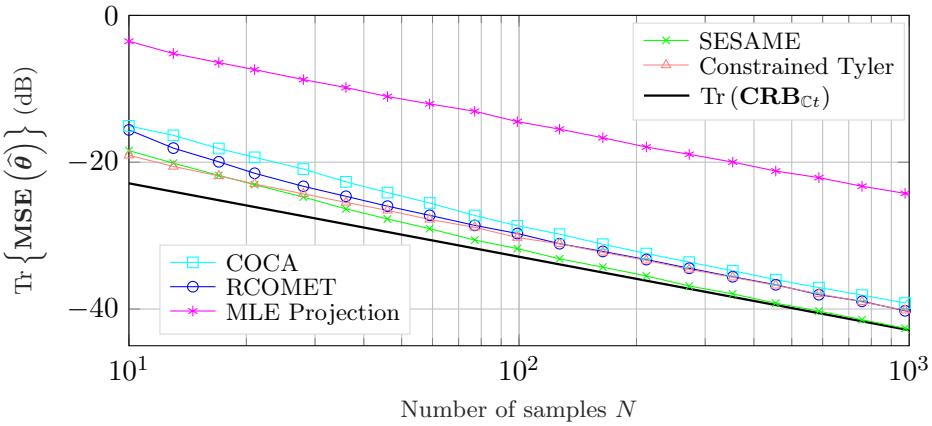


Figure 3.3: Comparison of several structured robust estimators. Samples follow a t-distribution with $d = 5$ DoF. $m = 4$, and the Toeplitz scatter matrix is defined by its first row: $[1, -0.83 - 0.20i, 0.78 + 0.37i, -0.66 - 0.70i]$.

fold (which is the case for covariance matrices and subspaces), it can be more relevant to caracterize a lower bound on the mean natural Riemannian distance between the true parameters and the estimators, which can also reveal hidden properties of estimators. To overcome this issue, intrinsic (i.e. in a manifold setting) versions of the Cramér-Rao inequality have been established in [Hendriks, 1991, Xavier and Barroso, 2002, Xavier and Barroso, 2005, Smith, 2005, Boumal, 2013, Bonnabel and Barrau, 2015]. Notably, in [Smith, 2005] intrinsic CRLBs are expressed in the form of a matrix inequality that is valid for any chosen Riemannian metric. Hence, it allows us to obtain Cramér-Rao bounds for various distances (depending on the chosen metric). This section presents the application of this framework to elliptical distributions.

3.5.1 Intrinsic CRLB (ICRLB)

The intrinsic Cramér-Rao bound extends the traditional Cramér-Rao bound for parameters living in a manifold and for an arbitrary chosen Riemannian metric. Indeed, the traditional estimation error (Euclidean distance) is defined through the difference between the true parameter and its estimate, which is not always properly defined (e.g., for subspaces). To deal with this issue [Smith, 2005] derived a Cramér-Rao type theorem for parameters living in a manifold by bounding the expected intrinsic distance between the estimate and the true parameter. Eventually, this theorem retrieves the well-known inequality "$\mathbf{C} \succcurlyeq \mathbf{F}^{-1}$", with $\mathbf{C}$ being the covariance matrix of the estimation error and $\mathbf{F}$ being the Fisher information matrix. However, these parameters have a different definition due to the specific nature of the considered quantities. This section briefly presents those definitions and the essential tools needed for the derivation of our contributions. We also refer the reader to the Chapter 6 of [Boumal, 2014] and the reference [Barrau and Bonnabel, 2013], which provide good introductions to the topic.

First, we list some definition used afterward

- $\mathcal{M}$ denotes a manifold, i.e. a space in which each point has a neighborhood that is homeomorphic to the Euclidean space.
- $\boldsymbol{\theta} \in \mathcal{M}$ denotes the parameter of interest, a point in the manifold $\mathcal{M}$.
- $T_{\boldsymbol{\theta}}\mathcal{M}$ is the tangent space at point $\boldsymbol{\theta}$, which is a vector space that conceptually contains the possible directions in which one can tangentially pass through the point $\boldsymbol{\theta}$.
- $g_{\boldsymbol{\theta}} : T_{\boldsymbol{\theta}}\mathcal{M} \times T_{\boldsymbol{\theta}}\mathcal{M} \to \mathbb{R}^+$ is a Riemannian metric: a scalar product on $T_{\boldsymbol{\theta}}\mathcal{M}$. The pair $(\mathcal{M}, g_{\boldsymbol{\theta}})$ is a Riemannian manifold. At each point, we denote by $\{\boldsymbol{\Omega}_i\}$ a basis of the tangent space $T_{\boldsymbol{\theta}}\mathcal{M}$

that is orthonormal w.r.t. $g_{\boldsymbol{\theta}}$.

- $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ is the geodesic distance induced by $g_{\boldsymbol{\theta}}$ on $\mathcal{M}$ (length of the shortest path between two points when integrating w.r.t. $g_{\boldsymbol{\theta}}$).

Our aim is to obtain a lower bound on the expected error of an estimator $\hat{\boldsymbol{\theta}}$ when measuring with d rather that the Euclidean squared norm $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||_F^2$. This requires to redefine the error measure, as the difference between two points on a manifold is not properly defined:

Definition 3.5.1. ***Riemannian estimation error***
Let $\hat{\boldsymbol{\theta}}$ be an estimator of the parameter $\boldsymbol{\theta} \in \mathcal{M}$. The estimation error $\mathbf{X}_{\boldsymbol{\theta}} \in T_{\boldsymbol{\theta}}\mathcal{M}$ is given by inverse exponential map (or logarithmic map):

$$\mathbf{X}_{\boldsymbol{\theta}} = \exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}}. \tag{3.21}$$

where $\exp_{\boldsymbol{\theta}}^{-1}$ denotes the operator that creates a vector "pointing toward" $\hat{\boldsymbol{\theta}}$ from $\boldsymbol{\theta}$, whose length coincides with the geodesic distance $d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$. Let the tangent space $T_{\boldsymbol{\theta}}\mathcal{M}$ be endowed with any metric (inner product) $g_{\boldsymbol{\theta}}$ and $\{\boldsymbol{\Omega}_i\}$ be a basis of this space. We have the coordinate vector $\mathbf{x}(\boldsymbol{\theta})$ with entries $[\mathbf{x}(\boldsymbol{\theta})]_i = g_{\boldsymbol{\theta}}(\mathbf{X}_{\boldsymbol{\theta}}, \boldsymbol{\Omega}_i)$, and the squared magnitude of estimation error is

$$\|\mathbf{x}(\boldsymbol{\theta})\|_F^2 = \mathbf{x}(\boldsymbol{\theta})^H \mathbf{x}(\boldsymbol{\theta}) = \left\| \exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}} \right\|_{\boldsymbol{\theta}}^2 = d^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \ , \tag{3.22}$$

where d is the distance defined w.r.t. the chosen Riemannian metric $g_{\boldsymbol{\theta}}$.

All these definitions are summarized in the Figure 3.4:

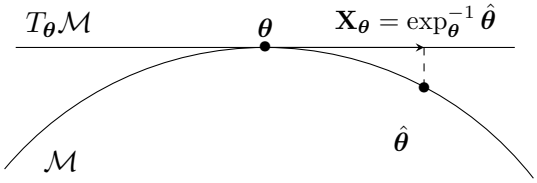


Figure 3.4: Illustration of the Riemannian estimation error $\mathbf{X}_{\boldsymbol{\theta}} = \exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}} \in T_{\boldsymbol{\theta}}\mathcal{M}$ between the parameter $\boldsymbol{\theta} \in \mathcal{M}$ and its estimate $\hat{\boldsymbol{\theta}} \in \mathcal{M}$.

We can finally state the two following theorems:

Theorem 3.5.1. ***Fisher information metric*** *[Smith, 2005]*
Let $f(\{\mathbf{z}_i\}_{i=1}^n | \boldsymbol{\theta})$ be a family of probability density function parameterized by $\boldsymbol{\theta}$ living in a manifold $\mathcal{M}$, $l = \log f$ be the log-likelihood function, and g_{fim} be the Fisher information metric. Let $\{\boldsymbol{\Omega}\}$ be an element of the tangent space $T_{\boldsymbol{\theta}}\mathcal{M}$ of the manifold $\mathcal{M}$ at point $\boldsymbol{\theta}$. We have the relation

$$g_{fim}(\boldsymbol{\Omega}, \boldsymbol{\Omega}) = -\mathbb{E}\left[\left. \frac{d^2}{dt^2} l\left(\{\mathbf{z}\}_{i=1}^n | \boldsymbol{\theta} + t\boldsymbol{\Omega}\right) \right|_{t=0} \right]. \tag{3.23}$$

Let $\{\boldsymbol{\Omega}_i\}$ be a basis of $T_{\boldsymbol{\theta}}\mathcal{M}$. The Fisher information matrix $\mathbf{F}$ is defined as

$$[\mathbf{F}]_{i,j} = g_{fim}(\boldsymbol{\Omega}_i, \boldsymbol{\Omega}_j), \tag{3.24}$$

where $g_{fim}(\boldsymbol{\Omega}_i, \boldsymbol{\Omega}_j)$ can be obtained from (3.23) using a polarization formula.

Theorem 3.5.2. ***Intrinsic Cramér-Rao bound*** *[Smith, 2005]*
Let $f(\{\mathbf{z}_i\}_{i=1}^n|\boldsymbol{\theta})$ be a family of probability density function parameterized by $\boldsymbol{\theta} \in \mathcal{M}$, $\mathbf{F}$ be the Fisher information matrix, and d be the distance associated with $\mathcal{M}$ and chosen Riemannian metric $g_{\boldsymbol{\theta}}$. Assume that $\hat{\boldsymbol{\theta}}$ is an unbiased (cf. definitions 1 and 2 of [Smith, 2005]) estimator of $\boldsymbol{\theta}$, then the covariance of the estimation error $\exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}}$ satisfies the matrix inequality

$$\mathbb{E}\left[\left(\exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}}\right)\left(\exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}}\right)^H\right] \succeq \mathbf{F}^{-1} - \frac{1}{3}\left(\mathbf{F}^{-1}\mathbf{R}_m\left(\mathbf{F}^{-1}\right) + \mathbf{R}_m\left(\mathbf{F}^{-1}\right)\mathbf{F}^{-1}\right) + \mathcal{O}(\lambda_{\max}(\mathbf{F}^{-1})^{2+1/2}) \quad (3.25)$$

where $\mathbf{R}_m()$ defines a Riemannian curvature term (cf. [Boumal, 2014, Eq. 6.6]).

This theorem generalizes the standard Cramér-Rao inequality with additional Riemanian curvature terms. These terms reflect the impact of the intrinsic structure of the parameter space (natural constraints satisfied within the manifold) in the estimation problem. However, if we neglect them (relevant for large n), the inequality translates in

$$\mathbb{E}\left[d^2\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\right)\right] \geq \mathrm{Tr}\left\{\mathbf{F}^{-1}\right\} \ , \quad (3.26)$$

where d corresponds to the distance associated with the chosen Riemannian metric (in which (3.22) was used to recover this term). Notice that intrinsic Cramér-Rao bounds in (3.26) are defined relatively to a Riemannian metric to be chosen, which allows for bounding a distance (performance criterion) that is considered to be meaningful for the addressed estimation problem.

3.5.2 ICRLB for scatter matrix estimation

We consider the problem of scatter matrix[4] estimation from elliptically distributed samples $\mathbf{z} \sim \mathcal{CES}\left(\mathbf{0}, \boldsymbol{\Sigma}, g_{\mathbf{z}}\right)$ from definition 3.2.1. The parameter $\boldsymbol{\Sigma}$ naturally belongs to the space of Hermitian positive definite matrices $\mathcal{H}_p^{++}$, which is indeed a Riemannian manifold (though the choice of the metric will be specified afterward) [Bhatia, 2009]. Notably, its tangent space $T_{\boldsymbol{\Sigma}}\mathcal{H}_p^{++}$ at each point $\boldsymbol{\Sigma}$ can be identified as $\mathcal{H}_p$, the space of Hermitian matrices. The related Fisher information metric is given in the following theorem:

Theorem 3.5.3. ***Fisher information metric for CES*** *[Breloy et al., 2019a]*
Let $\{\mathbf{z}_i\}_{i=1}^n$ be a n-sample of $\mathbf{z} \sim \mathcal{CES}\left(\mathbf{0}, \boldsymbol{\Sigma}, g_{\mathbf{z}}\right)$. Let $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ be two elements of $T_{\boldsymbol{\Sigma}}\mathcal{H}_p^{++}$. Then, the Fisher information metric associated with $\boldsymbol{\Sigma}$ is given as

$$g_{\boldsymbol{\Sigma}}^{fim}\left(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2\right) = n g_{\boldsymbol{\Sigma}}^{ces}\left(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2\right) \quad (3.27)$$

with

$$g_{\boldsymbol{\Sigma}}^{ces}\left(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2\right) = \alpha \mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}_2\right\} + \beta \mathrm{Tr}\left\{\boldsymbol{\Omega}_1\boldsymbol{\Sigma}^{-1}\right\}\mathrm{Tr}\left\{\boldsymbol{\Omega}_2\boldsymbol{\Sigma}^{-1}\right\}, \quad (3.28)$$

where the coefficients α and β are defined as

$$\alpha = 1 - \mathbb{E}\left[\mathcal{Q}^2\phi'\left(\mathcal{Q}\right)\right]/(p(p+1)) \quad \text{and} \quad \beta = \alpha - 1 \quad (3.29)$$

with $\phi = g_{\mathbf{z}}'(t)/g_{\mathbf{z}}(t)$ and where $\mathcal{Q}$ is the second order modular variate of the considered distribution.

This metric corresponds to the Rao-Fisher metric [Micchelli and Noakes, 2005] (also referred to as affine-invariant metric) with specific scalings (α,β) depending on the underlying distribution. We now turn to the choice of the metric used to measure the estimation error. Two popular choices are considered:

[4]The problem of shape (normalized scatter) matrix estimation is also studied in [Breloy et al., 2019a]. Interested readers can find the details in section 2.6 of the annexes.

	Euclidean	Natural Riemannian
Metric	$g^{Eucl}(\mathbf{\Omega},\mathbf{\Omega}) = \mathrm{Tr}\left\{\mathbf{\Omega}^2\right\}$	$g^{nat}_{\mathbf{\Sigma}}(\mathbf{\Omega},\mathbf{\Omega}) = \mathrm{Tr}\left\{\left(\mathbf{\Sigma}^{-1}\mathbf{\Omega}\right)^2\right\}$
Distance	$d^2_{Eucl}(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2) = \|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|^2_F$	$d^2_{nat}(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2) = \|\log(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2})\|^2_F$

Notice that $g^{nat}_{\mathbf{\Sigma}}$ corresponds to a special case of $g^{ces}_{\mathbf{\Sigma}}$ for which $\alpha = 1$ and $\beta = 0$. For these two metrics we have the following results:

Theorem 3.5.4. ***Euclidean Cramér-Rao bound on*** $\mathbf{\Sigma}$ *[Breloy et al., 2019a]*
Let $\{\mathbf{z}_i\}_{i=1}^n$ *be a* n*-sample of* $\mathbf{z} \sim \mathcal{CES}(\mathbf{0},\mathbf{\Sigma},g_{\mathbf{z}})$*. The Cramér-Rao bound on the Euclidean distance between an unbisaed estimator* $\hat{\mathbf{\Sigma}}$ *and* $\mathbf{\Sigma}$ *is*

$$\mathbb{E}\left[d^2_{Eucl}\left(\hat{\mathbf{\Sigma}},\mathbf{\Sigma}\right)\right] \geq \mathrm{Tr}\left\{\mathbf{F}^{-1}_{Eucl}\right\} \tag{3.30}$$

with

$$[\mathbf{F}_{Eucl}]_{i,j} = n\alpha\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\mathbf{\Omega}_i^{Eucl}\mathbf{\Sigma}^{-1}\mathbf{\Omega}_j^{Eucl}\right\} + n\beta\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\mathbf{\Omega}_i^{Eucl}\right\}\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\mathbf{\Omega}_j^{Eucl}\right\} \tag{3.31}$$

where $\{\mathbf{\Omega}_i^{Eucl}\}_{i=1}^{p^2}$ *is an orthonormal basis of* $\mathcal{H}_p$ *(e.g., given in [Smith, 2005]).*

Remark that this corresponds well to the Euclidean Cramér-Rao bounds obtained for several distributions in [Greco and Gini, 2013, Pascal and Renaux, 2010, Besson and Abramovich, 2013, Mitchell, 1989]. Also notice that we retrieve the same result as Theorem 5 of [Smith, 2005] for the special case of Gaussian distribution, obtained with $\alpha = 1$ and $\beta = 0$.

Theorem 3.5.5. ***Natural Cramér-Rao bound on*** $\mathbf{\Sigma}$ *[Breloy et al., 2019a]*
Let $\{\mathbf{z}_i\}_{i=1}^n$ *be a* n*-sample of* $\mathbf{z} \sim \mathcal{CES}(\mathbf{0},\mathbf{\Sigma},g_{\mathbf{z}})$*. The Cramér-Rao bound on the natural Riemannian distance between an unbiased estimator* $\hat{\mathbf{\Sigma}}$ *and* $\mathbf{\Sigma}$ *is*

$$\mathbb{E}\left[d^2_{nat}\left(\hat{\mathbf{\Sigma}},\mathbf{\Sigma}\right)\right] \geq \frac{p^2-1}{n\alpha} + (n(\alpha+p\beta))^{-1}\ . \tag{3.32}$$

Besides providing a more interpretable result, the Riemannian analysis can also reveal interesting hidden properties of estimators. A main example from [Smith, 2005] is that the SCM in the Gaussian case appears efficient w.r.t. d_{Eucl}, while not being so w.r.t. d^2_{nat} at small sample size. This better reflects results from empirical processes, meaning that there is still room for improvement when $n \sim p$, which motivated recent works such as [Tiomoko et al., 2019]. These insights are confirmed by the validation simulation presented in Figure 3.5.

3.5.3 ICRLB for subspace estimation in spiked models

This section briefly presents recent results derived in [Bouchard et al., 2020]. We consider spiked models (also referred to as factor models), which correspond to the case of covariance matrices structured as

$$\mathbf{\Sigma} = \mathbf{\Sigma}_k + \mathbf{I} \tag{3.33}$$

where $\mathrm{rank}(\mathbf{\Sigma}_k) = k < p$. In the following, the distributions $\mathbf{z} \sim \mathcal{CES}(\mathbf{0},\mathbf{\Sigma}_k + \mathbf{I},g_{\mathbf{z}})$ are referred to as *spiked elliptical models*, which correspond to a generalization of low-rank structured Gaussian models (e.g., [Tipping and Bishop, 1999a]) to the case of CES distributions.

Dealing with this model requires to chose a geometry for the manifold $\mathcal{H}^+_{p,k}$ of $p \times p$ Hermitian positive semi-definite matrices of rank k, which has recently attracted much attention; see *e.g.*, [Bonnabel and Sepulchre, 2009, Vandereycken and Vandewalle, 2010, Meyer et al., 2011, Vandereycken et al., 2012, Massart and Absil, 2018]. In [Bouchard et al., 2020], we opted for the parameterization

$$\mathbf{\Sigma}_k = \mathbf{U}\mathbf{\Lambda}_k\mathbf{U}^H \tag{3.34}$$

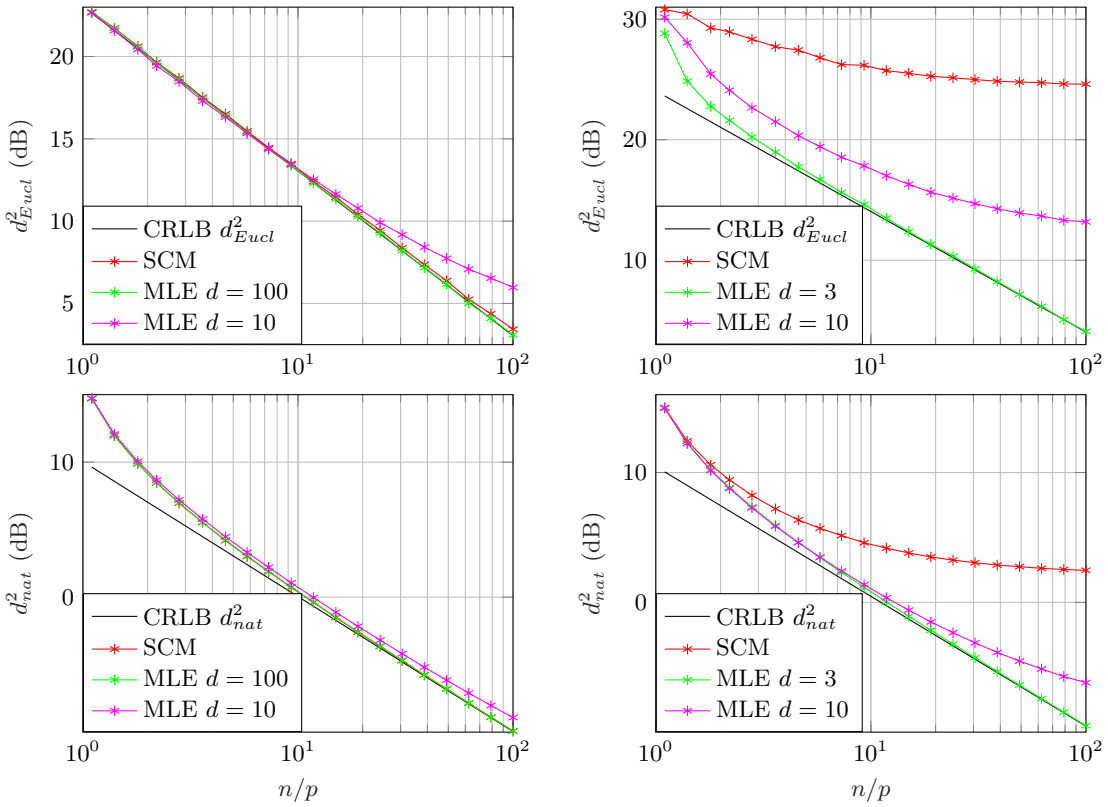


Figure 3.5: Performance of several M-estimators compared to the Euclidean (top) and natural (bottom) Cramér-Rao bounds on scatter estimation. $\mathbf{z} \sim (\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{z}})$ with a t-distribution ($g_{\mathbf{z}}(t) = \left(1 + d^{-1}t\right)^{-(d+M)}$) and $[\mathbf{\Sigma}]_{i,j} = \rho^{|i-j|}$ with $\rho = 0.9\sqrt{1/2}\,(1+i)$. $p = 10$, left: $d = 100$ (similar to Gaussian case), right: $d = 3$.

with $\mathbf{U} \in \mathrm{St}_{p,k}$ and $\mathbf{\Lambda}_k \in \mathcal{H}_k^{++}$. Notice that this representation is invariant by the unitary transformation $(\mathbf{U}\mathbf{O}^H)(\mathbf{O}\mathbf{\Lambda}_k\mathbf{O}^H)(\mathbf{O}\mathbf{U}^H)$ with $\mathbf{O} \in \mathcal{U}_k$, so we consider the geometry induced by the quotient $(\mathrm{St}_{p,k} \times \mathcal{H}_k^{++})/\mathcal{U}_k$. We then derived geometric tools induced by a new Riemannian metric on the product $\mathrm{St}_{p,k} \times \mathcal{H}_k^{++}$: the part in $\mathrm{St}_{p,k}$ is the so-called canonical metric on Stiefel [Edelman et al., 1998] while the part in $\mathcal{H}_k^{++}$ is a general form of the Fisher information metric of elliptical distributions in $\mathcal{H}_k^{++}$ (cf. Theorem 3.5.3). It is of particular interest in our context because the principal subspace of the covariance matrix is directly obtained from this parametrization and a closed-form divergence function, which can be exploited to measure estimation errors [Bonnabel and Sepulchre, 2009]. Tedious calculations details are omitted for clarity of exposition[5]. However we can still report the following result:

Theorem 3.5.6. ***Cramér-Rao bound for subspace estimation in spiked elliptical models*** *Let $\{\mathbf{z}_i\}_{i=1}^n$ be a n-sample of $\mathbf{z} \sim \mathcal{CES}\,(\mathbf{0}, \mathbf{\Sigma}_k + \mathbf{I}, g_{\mathbf{z}})$, with $\mathbf{\Sigma}_k = \mathbf{U}\mathbf{\Lambda}_k\mathbf{U}^H$ and $\mathbf{\Lambda}_k = \mathrm{diag}(\{\sigma_i\})$. Let $d^2_{\mathcal{G}_{p,k}}(\mathrm{span}(\mathbf{U}), \mathrm{span}(\hat{\mathbf{U}}))$ be the estimation error between the subspaces spanned by $\mathbf{U}$ and an unbiased estimator $\hat{\mathbf{U}}$, defined as*

$$d^2_{\mathcal{G}_{p,k}}(\mathrm{span}(\mathbf{U}), \mathrm{span}(\hat{\mathbf{U}})) = ||\mathbf{\Theta}||_F^2, \tag{3.35}$$

where $\mathbf{\Theta}$ is derived from the singular value decomposition $\mathbf{U}^H\hat{\mathbf{U}} = \mathbf{O}\cos(\mathbf{\Theta})\hat{\mathbf{O}}^H$. We have the following inequality

$$\mathbb{E}\left[d^2_{\mathcal{G}_{p,k}}(\mathrm{span}(\mathbf{U}), \mathrm{span}(\hat{\mathbf{U}}))\right] \geq \frac{(p-k)}{n\alpha}\sum_{i=1}^{k}\frac{1+\sigma_i}{\sigma_i^2}. \tag{3.36}$$

[5]Interested readers can find the details in section 2.7 of the annexes

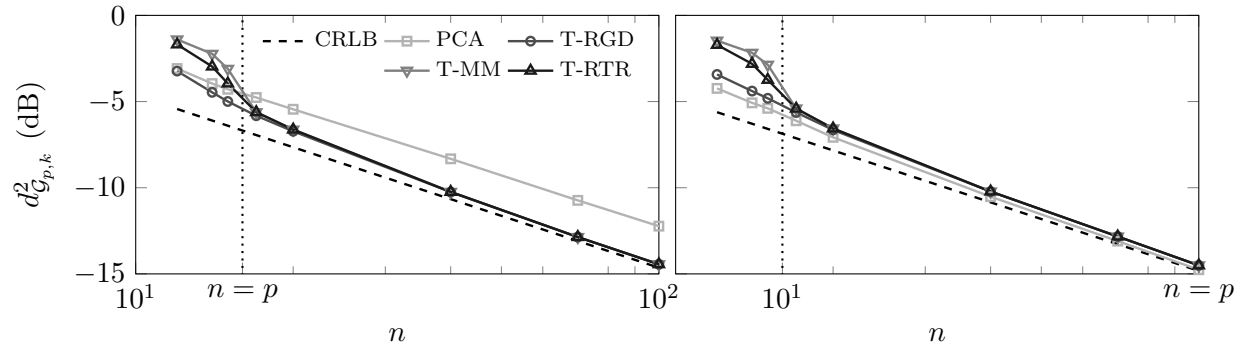


Figure 3.6: Performance of several subspace estimators from [Sun et al., 2016, Bouchard et al., 2020] compared the intrinsic Cramér-Rao bounds. $\mathbf{z} \sim \left(\mathbf{0}, \sigma \mathbf{U}\mathbf{\Lambda}_k\mathbf{U}^H + \mathbf{I}, g_{\mathbf{z}}\right)$ with a t-distribution ($g_{\mathbf{z}}(t) = \left(1 + d^{-1}t\right)^{-(d+M)}$), $\mathbf{U}$ is a random matrix in $\mathrm{St}(p,k)$, $\mathbf{\Lambda}_k$ is a diagonal matrix whose minimal and maximal elements are $1/\sqrt{c}$ and $\sqrt{c}$ ($c = 20$ is the condition number); its other elements are randomly drawn from the uniform distribution between $1/\sqrt{c}$ and $\sqrt{c}$; its trace is then normalized as $\mathrm{Tr}(\mathbf{\Lambda}_k) = k$, $\sigma = 50$ is a free parameter corresponding to the signal to noise ratio. $p = 16$ and $k = 4$, left: $d = 3$ (heavy tailed), right: $d = 100$ (similar to Gaussian).

Interestingly, this theorem leads to an interpretable lowerbound for subspace recovery in terms of problem dimensions and signal to noise ratio. We also note that it coincides with the Gaussian signal case studied in [Smith, 2005] for $\alpha = 1$ and $\sigma_i = \mathrm{SNR}, \forall i \in [\![1, k]\!]$. Finally, Figure 3.6 illustrates that this lower bound can asymptotically be achieved by the algorithms proposed in [Sun et al., 2016, Bouchard et al., 2020].

3.6 Related works and perspectives

Insufficient sample size

In this chapter, we focused on the case $n > p$. In order to enjoy the robustness properties of M-estimators at insufficient sample support ($n < p$), it is possible to leverage regularization methods. A regularized M-estimator can be expressed as the minimizer of the cost function $\hat{\mathbf{\Sigma}}(\alpha) = \mathrm{argmin}_{\mathbf{\Sigma}}\ \mathcal{L}(\mathbf{\Sigma}) + \alpha\mathcal{P}(\mathbf{\Sigma})$, where $\mathcal{L}$ is the objective in (3.12), $\mathcal{P}$ is a penalty function, and α is a regularization parameter. Recent works [Wiesel, 2012b, Sun et al., 2014, Pascal et al., 2014b, Ollila and Tyler, 2014a] focused on penalties of the form $\mathcal{P}(\mathbf{\Sigma}) = \mathrm{Tr}\{\mathbf{\Sigma}^{-1}\mathbf{T}\} + \ln|\mathbf{\Sigma}|$ that shrink the estimate towards a prior $\mathbf{T}$. These regularized M-estimators satisfy the fixed-point equation

$$\hat{\mathbf{\Sigma}}(\alpha) = \frac{1}{1+\alpha}\sum_{i=1}^{n} \rho'(\mathbf{z}_i^H\hat{\mathbf{\Sigma}}^{-1}(\alpha)\mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^H + \frac{\alpha}{1+\alpha}\mathbf{T} \tag{3.37}$$

and can exist for $n < p$ [Ollila and Tyler, 2014a]. In this topic, several perspectives of the presented work can be mentioned:

- Numerous works considered the problem of optimal α parameter selection, with respect to several applicative criterions (cf. [Hoarau et al., 2017] and refs. therein). The issue is quite complex, therefore it would be interesting to simplify it by establishing Wishart equivalents of regularized M-estimators (e.g. in the form of shrinked SCM [Ledoit and Wolf, 2004]).

- In order to promote certain spectral structures [Aubry et al., 2018, Tyler and Yi, 2018, Basiri et al., 2019], it would be interesting to develop EVD-based regularization penalties, e.g., $\mathcal{P}(\mathbf{\Sigma}) = \mathcal{P}_\lambda(\mathbf{\Lambda}) + \mathcal{P}_\mathrm{v}(\mathbf{V})$ for $\mathbf{\Sigma} \overset{\mathrm{EVD}}{=} \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$. We started working on this approach by first considering unitary invariant shrinkage ($\mathcal{P}_\mathrm{v}(\mathbf{V}) = 0$) in [Breloy et al., 2019b]. An interesting lead

would be to promote sparsity in the eigenvectors, e.g., using the tools that will be presented in chapter 4.

Information geometry and Riemannian approaches

Beyond performance bounds, the Riemannian point of view offers a unified approach to tackle statistical estimation problems involving structured matrices. Indeed, the information geometry derived from the considered model can guide the choice of relevant metrics for the derivation of the geometry of the parameter space. This geometry can be exploited by both deriving performance bounds, as well as optimization algorithms on Riemannian manifolds [Absil et al., 2009]. For example, we also proposed new estimation algorithms in [Bouchard et al., 2020], with promising results at insufficient sample support. We also started addressing robust blind source separation with this perspective in [Bouchard et al., 2020]. Finally, the derivation of information geometry for the considered parameter space and distribution may yield new Riemannian distances that can be leveraged for other applications such as clustering, which is a lead we started to explore in the thesis of Antoine Collas (started in september 2019).

3.7 Onward to the next chapter

This chapter presented several theoretical derivations (within the elliptical model framework) for covariance matrix estimation related problems. The chapter 4 will be more application-centered, as it will discuss the application of spiked elliptical models (cf. sec 3.5.3) to the problem of change detection in muntivariate image time series.

4 | Change detection in SAR image time series

Contents

4.1 Contributions of the chapter

This chapter addresses the problem of change detection in multivariate image times series. In this context, testing the similarity of covariance matrices from groups of observations has been shown to be a relevant approach. The corresponding classical statistical methodologies are usually built upon the Gaussian assumption, as well as an unstructured signal model. Both of these hypotheses may be inaccurate for high-dimension/resolution images (non-Gaussian observations), and where all channels are not always informative (structured signals). The problem will be tackled by deriving new detection methods using signal models from the previous chapters, that alleviate the aforementioned limitations. While the term "testing similarity" usually refers to equality, we also discuss the testing of shared properties in the eigendecomposition (e.g., principal components) of groups of covariance matrices structures.

The contributions presented in this chapter detail works conducted during the theses of Ammar Mian and Rayen Ben Abdallah.
Related publications in the cv (page xi): [J6], [C16], [C18], [FC13], [FC8], [J12], [X2].

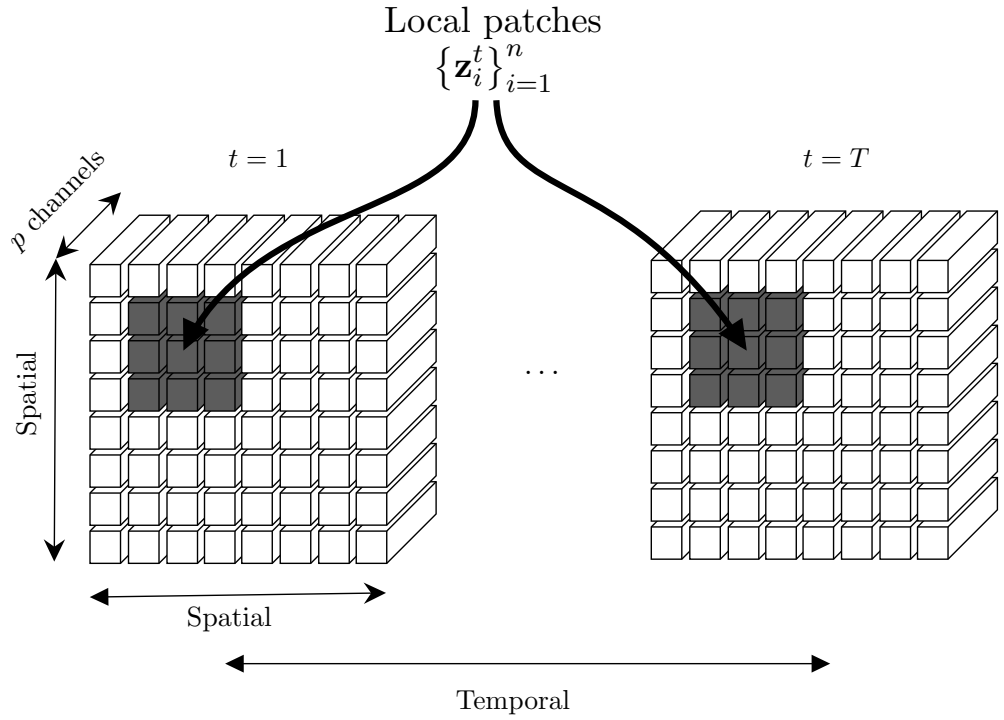


Figure 4.1: Representation of p-variate SAR-ITS data set. The pixels highlighted in black correspond to the local observation window (here, $n = 9$).

4.2 Context overview

4.2.1 Change detection in SAR image time series

The analysis of Synthetic Aperture Radar (SAR) Image Time Series (ITS) has become a popular topic of research since it has many practical applications for Earth monitoring, such as disaster assessment, infrastructure monitoring or land-cover analysis. Over the past years, SAR-ITS have been made more widely available thanks to various missions such as Sentinel-1, TerraSAR-X, or UAVSAR. As a consequence, an active topic of research addresses the development of reliable automatic Change Detection (CD) methodologies in order to efficiently process this large amount of data. The CD problem is indeed challenging due to the lack of available ground truths, which does not allow applying supervised methods from the image processing literature. Moreover, it is well known that SAR images are subjected to speckle noise, which makes traditional image-based approaches prone to high false alarm rates. In this case, unsupervised methodologies, often based on statistical tools, have yield interesting approaches in recent decades [Hussian et al., 2013].

In the following, we consider a multitemporal time-series of T multivariate SAR images as described in Figure 4.1. Each pixel of a SAR image at a given date t corresponds to a vector of dimension p, denoted $\mathbf{z} \in \mathbb{C}^p$. The p channels can correspond to a polarimetric diversity ($p = 3$), or to another kind of diversity such as a spectro-angular one, obtained through wavelet transforms [Mian et al., 2019]. The change detection process is applied using a local window around the pixel of interest, including n pixels. Locally, the whole data set is denoted $\{\{\mathbf{z}_i^t\}_{i=1}^n\}_{t=1}^T$, which corresponds to the aggregation of all pixels at spatial indexes $i \in [\![1, n]\!]$ and dates $t \in [\![1, T]\!]$.

4.2.2 Statistical change detection with the GLRT

For a given time t, the local observation $\{\mathbf{z}_i^t\}_{i=1}^n$ is assumed to be distributed according to a fixed parametric distribution, of parameter $\boldsymbol{\theta}_t$. The corresponding likelihood is denoted $\mathcal{L}\left(\{\mathbf{z}_i^t\}_{i=1}^n|\boldsymbol{\theta}_t\right)$. The parameters $\{\boldsymbol{\theta}_t\}$ are feature that characterize the local data at each date t: if a local change occurs, this parameter is expected to vary. The CD problem can thus be formulated as a binary hypothesis test:

$$\begin{cases} \mathrm{H}_0: & \boldsymbol{\theta}_i = \boldsymbol{\theta}_0,\ \forall i \in [\![1,T]\!] \quad \text{(no change)}, \\ \mathrm{H}_1: & \boldsymbol{\theta}_i \neq \boldsymbol{\theta}_j,\ \text{for } i \neq j \qquad \text{(change)}. \end{cases} \tag{4.1}$$

Notice that we will consider an *omnibus* CD problem, i.e. we do not test for a change at a specific date. Conversely, the *sequential* test focusing on the date t_0 is expressed as:

$$\begin{cases} \mathrm{H}_0: & \boldsymbol{\theta}_i = \boldsymbol{\theta}_1,\ \forall i \in [\![1,T]\!] & \text{(no change)}, \\ \mathrm{H}_1: & \boldsymbol{\theta}_i = \boldsymbol{\theta}_1,\ \forall i \in [\![1,t_0-1]\!] \\ & \boldsymbol{\theta}_i = \boldsymbol{\theta}_{t_0},\ \forall i \in [\![t_0,T]\!],\ \text{with } \boldsymbol{\theta}_{t_0} \neq \boldsymbol{\theta}_1 & \text{(change at } t_0). \end{cases} \tag{4.2}$$

which can often be recasted as an omnibus one with $T = 2$ by re-partitioning the data set if there is a single change (yet, this property is not always true depending on the chosen model).

In order to derive a metric of decision, we consider the use of the generalized likelihood ratio test (GLRT) for the hypothesis test (4.1). This test consists in computing the following quantity:

$$\hat{\Lambda}_{\mathrm{GLRT}} = \frac{\max\limits_{\{\boldsymbol{\theta}_t\}_{t=1}^T} \prod_{t=1}^T \mathcal{L}^{\mathrm{H}_1}\left(\{\mathbf{z}_i^t\}_{i=1}^n|\boldsymbol{\theta}_t\right)}{\max\limits_{\boldsymbol{\theta}_0} \prod_{t=1}^T \mathcal{L}^{\mathrm{H}_0}\left(\{\mathbf{z}_i^t\}_{i=1}^n|\boldsymbol{\theta}_0\right)}, \tag{4.3}$$

where $\mathcal{L}^{\mathrm{H}_1}$ (resp. $\mathcal{L}^{\mathrm{H}_0}$) denotes the likelihood function and $\{\boldsymbol{\theta}_t\}_{t=1}^T$ (resp. $\boldsymbol{\theta}_0$) corresponds to the parameters of the distribution, both under H_1 (resp. H_0).

Additionally, a possible parameter splitting $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{\mathrm{test}}, \boldsymbol{\theta}^{\mathrm{side}}\}$ can lead to the following hypothesis test:

$$\begin{cases} \mathrm{H}_0: & \boldsymbol{\theta}_i^{\mathrm{test}} = \boldsymbol{\theta}_0^{\mathrm{test}},\ \forall i \in [\![1,T]\!] \\ & \boldsymbol{\theta}_i^{\mathrm{side}} \neq \boldsymbol{\theta}_j^{\mathrm{side}},\ \textbf{for } i \neq j & \text{(no change in } \boldsymbol{\theta}^{\mathrm{test}}), \\ \mathrm{H}_1: & \boldsymbol{\theta}_i^{\mathrm{test}} \neq \boldsymbol{\theta}_j^{\mathrm{test}},\ \textbf{for } i \neq j \\ & \boldsymbol{\theta}_i^{\mathrm{side}} \neq \boldsymbol{\theta}_j^{\mathrm{side}},\ \textbf{for } i \neq j & \text{(change in } \boldsymbol{\theta}^{\mathrm{test}}). \end{cases} \tag{4.4}$$

Modified accordingly, the GLRT formulation of (4.3) still allows us to test for specific changes within the parameters. Thus, one statistical model can yield several tests, whose relevance will depend on the phenomenon we aim to test. This perspective will be discussed in the following, but not fully detailed because our experiments showed that the considered CD application (i.e., detecting any change) favors detectors that test for all the parameters.

In conclusion: to develop efficient GLRT detectors, the problem remains to select an assumed distribution (and corresponding parameters) that accurately reflects the behavior of the data and the phenomenon we aim to test.

4.2.3 Current issues

The CD problem still generates several challenges, from which we can stress the following:

- A first issue concerns the modeling of the data. The CD processes are built upon the Gaussian assumption, which can be inaccurate for high-dimension/resolution images (i.e., non-Gaussian observation). An element of response is brought by the use or more general families of distributions, such a compound Gaussian ones (cf. chapters 2-3). For a chosen distribution, the choice

of parameters to test is also not trivial, as it should leverage prior physical considerations on the acquisition system. In this scope, we will consider the use of spiked models (i.e., covariance matrices with a low rank structure), which is often appropriate in radar signal processing.

- It is also worth mentioning that, depending on the chosen model, the evaluation of the GLRT may lead to complex optimization problems, which boils down to the computation of maximum likelihood estimators. In the following presentation, we will elude this question by simply noticing that the proposed tests can be computed using the majorization-minimization techniques evoked in the previous chapters and Appendix A[1].

4.3 GLRTs based on the covariance matrix

Testing the similarity of covariances matrices between groups of observations is a well-established topic in the statistical literature [Nagao, 1973, Schott, 2001, Anderson, 2003, Hallin and Paindaveine, 2009], which has also been considered for CD in time-series in, e.g. [Galeano and Peña, 2007, Aue et al., 2009]. More specifically for SAR-ITS applications, various test statistics based on covariance matrix equality testing from Gaussian samples have been proposed within statistical detection framework [Conradsen et al., 2003, Novak, 2005, Barber, 2015, Carotenuto et al., 2015, Maio et al., 2017], or using information theory [Nascimento et al., 2019, Ratha et al., 2017] . A good review of these Gaussian detectors can be found in [Ciuonzo et al., 2017].

In the following, we will consider generalizations of the Gaussian GLRT approach: we present the GLRTs corresponding to four models, which can be splitted between two distributions

- Gaussian, i.e., assuming $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$,
- Compound Gaussian with deterministic textures, i.e. assuming $\mathbf{z}_i|\tau_i \sim \mathcal{CN}(\boldsymbol{\mu}, \tau_i\mathbf{\Sigma})$, $\forall i$, where τ_i is unknown deterministic,

and two parameterizations for the covariance matrix

- Unstructured, i.e., no specific structure assumed on $\mathbf{\Sigma}$,
- Spiked (low-rank structured), i.e., assuming $\mathbf{\Sigma} = \mathbf{\Sigma}_k + \sigma^2\mathbf{I}$, where $\text{rank}(\mathbf{\Sigma}_k) = k$ and k is fixed,

which are summed-up (with a corresponding reference) in the following table:

Model	Gaussian	Compound Gaussian
Unstructured	[Conradsen et al., 2003]	[Mian et al., 2019]
Spiked	[Ben Abdallah et al., 2019c]	[Mian et al., 2019]

4.3.1 Gaussian models

Unstructured Gaussian

Assuming Gaussian distributed samples, the CD can be performed by testing a change in the covariance matrix [Conradsen et al., 2003, Novak, 2005]. The corresponding GLRT, denoted $\hat{\Lambda}_{\text{G}}$, corresponds to (4.1) and (4.3) with the following distribution/parameters:

$$\begin{aligned} &\text{Model:} && \mathbf{z}_i^t \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_t) \\ &\text{Param.:} && \text{H}_0 : \boldsymbol{\theta}_0 = \mathbf{\Sigma}_0 \\ & && \text{H}_1 : \{\boldsymbol{\theta}_t\}_{t=1}^T = \{\mathbf{\Sigma}_t\}_{t=1}^T. \end{aligned} \tag{4.5}$$

[1] Interested readers can find the details in sections 2.8 and 2.9 of the annexes.

This test has well established statistical properties (cf. eg. [Ciuonzo et al., 2017]) and admits the closed-form expression

$$\hat{\Lambda}_{\mathrm{G}} = |\hat{\mathbf{\Sigma}}_0^{\mathrm{SCM}}| / \prod_{t=1}^{T} |\hat{\mathbf{\Sigma}}_t^{\mathrm{SCM}}|^{1/T}, \tag{4.6}$$

with the Sample Covariance Matrices (SCMs) defined by

$$\hat{\mathbf{\Sigma}}_0^{\mathrm{SCM}} = \frac{1}{Tn} \sum_{t,i} \mathbf{z}_i^t (\mathbf{z}_i^t)^H \quad \text{and} \quad \hat{\mathbf{\Sigma}}_t^{\mathrm{SCM}} = \frac{1}{n} \sum_{i} \mathbf{z}_i^t (\mathbf{z}_i^t)^H. \tag{4.7}$$

Spiked Gaussian (and extensions)

The performance of the CD methods is tightly linked to the accuracy of the covariance matrix estimation. The general rule-of-thumb suggests that $n > 2p$ samples are required in order to obtain a correct estimation. However, spiked models are very common structure in radar due to signals lying in a lower dimensional subspace. In order to lower n (i.e., reduce the local window size), we proposed in [Ben Abdallah et al., 2019c] to extend the Gaussian GLRT to the spiked model. The resulting test corresponds to (4.1) and (4.3) with the following distribution/parameters:

$$\begin{aligned} &\text{Model:} && \mathbf{z}_i^t \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{\Sigma}_k^t + \sigma_t^2 \mathbf{I}\right) \\ &\text{Param.:} && \mathrm{H}_0 : \boldsymbol{\theta}_0 = \{\mathbf{\Sigma}_k^0, \sigma_0^2\} \\ &&& \mathrm{H}_1 : \{\boldsymbol{\theta}_t\}_{t=1}^T = \{\mathbf{\Sigma}_k^t\}_{t=1}^T \end{aligned} \tag{4.8}$$

where $\mathbf{\Sigma}_k$ belongs to the set of $p \times p$ Hermitian positive semi-definite matrices of rank k. Following from [Tipping and Bishop, 1999a], this test consists in evaluating the ratio (4.3) with the MLEs

$$\begin{aligned} \hat{\mathbf{\Sigma}}_k^0 + \hat{\sigma}_0^2 \mathbf{I} &= \mathcal{T}_{\mathrm{k}}\{\hat{\mathbf{\Sigma}}_0^{\mathrm{SCM}}\} \\ \hat{\mathbf{\Sigma}}_k^t + \hat{\sigma}_t^2 \mathbf{I} &= \mathcal{T}_{\mathrm{k}}\{\hat{\mathbf{\Sigma}}_t^{\mathrm{SCM}}\} \end{aligned} \tag{4.9}$$

with the operator $\mathcal{T}_{\mathrm{k}}$ (projection on the set of low-rank plus scaled identity) is defined by

$$\mathcal{T}_{\mathrm{R}} : \mathbf{\Sigma} = \mathbf{U} \operatorname{diag}(\mathbf{d}) \mathbf{U}^H \quad \mapsto \quad \mathcal{T}_{\mathrm{R}}(\mathbf{\Sigma}) \triangleq \mathbf{U} \operatorname{diag}(\tilde{\mathbf{d}}) \mathbf{U}^H \tag{4.10}$$

with

$$\begin{aligned} \mathbf{d} &= [d_1, \ldots, d_k, d_{k+1}, \ldots, d_p] \\ \tilde{\mathbf{d}} &= \left[d_1, \ldots, d_k, \hat{\sigma}^2, \ldots, \hat{\sigma}^2\right] \\ \hat{\sigma}^2 &= \frac{1}{p-k} \sum_{r=k+1}^{p} d_r \,. \end{aligned} \tag{4.11}$$

We also note that if the noise variance is assumed to be known and equal to σ^2, the solution is given in [Kang et al., 2014], and consists in replacing $\tilde{\mathbf{d}}$ by $\bar{\mathbf{d}} = \left[\max(d_1, \sigma^2), \ldots, \max(d_R, \sigma^2), \sigma^2, \ldots, \sigma^2\right]$.

Several generalizations of this approach can be found in [Ben Abdallah et al., 2019c, Ben Abdallah et al., 2019b], where the considered model is

$$\mathbf{z}_i^t \overset{d}{=} \mathbf{U}_t (\mathbf{D}_i^t)^{1/2} \mathbf{s}_i^t + \sigma_t \mathbf{n}_i^t \tag{4.12}$$

with $\mathbf{U}_t \in \mathrm{St}(p,k)$, $\mathbf{D}_i^t \in \mathcal{H}_k^{++}$, $\mathbf{s}_i^t \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ and $\mathbf{n}_i^t \mathcal{CN}(\mathbf{0}, \mathbf{I})$. This model encompasses (4.8), but also allows for extensions echoing to the non-Gaussian models in chapter 2, such as LRCG in definition 2.3.4. Most interestingly, splitting the parameters as in (4.4) can yield detectors that test for a specific feature change. For example [Ben Abdallah et al., 2019b] proposed a detector for a change in the signal subspace, while being insensitive to variations of the other parameters (power fluctuations, inner correlations, etc.)[2]. The approach is an interesting prospect for other applications that aim to test for specific physical phenomenons. However, these formulations are not specifically efficient for standard CD where we aim to detect any change.

[2]Interested readers can find the details in section 2.9 of the annexes.

4.3.2 Compound Gaussian models

Unstructured compound Gaussian

As stated previously, the standard Gaussian assumption is no longer valid for high-resolution, or heterogeneous SAR images. This mismodeling induces a strong reduction of the CD performance when using $\hat{\Lambda}_{\mathrm{G}}$, notably caused by the inaccuracy of the SCM computed from non-Gaussian observations. This issue can be alleviated by assuming a compound Gaussian model. Under this assumption, the CD can be performed by testing a change in both the covariance matrix and the texture parameters [Mian et al., 2019]. The corresponding GLRT, denoted $\hat{\Lambda}_{\mathrm{CG}}$, corresponds to (4.1) and (4.3) with the following distribution/parameters:

$$\begin{aligned} &\text{Model:} && \mathbf{z}_i^t \sim \mathcal{CN}\left(\mathbf{0}, \tau_i^t \boldsymbol{\Sigma}_t\right) \\ &\text{Param.:} && \mathrm{H}_0 : \boldsymbol{\theta}_0 = \left\{\boldsymbol{\Sigma}_0, \{\tau_i^0\}_{i=1}^n\right\} \\ & && \mathrm{H}_1 : \{\boldsymbol{\theta}_t\}_{t=1}^T = \left\{\boldsymbol{\Sigma}_t, \{\tau_i^t\}_{i=1}^n\right\}_{t=1}^T \end{aligned} \tag{4.13}$$

The evaluation of the quantity $\hat{\Lambda}_{\mathrm{CG}}$ involves fixed-point equations that can be computed numerically. A study and generalizations (testing for textures or covariance matrices individually) of this approach can be found in [Mian et al., 2019].

Spiked compound Gaussian

In [Mian et al., 2019], we proposed to combine the advantages of both the low-rank structure and the CG distribution. Thus, we considered a model of CG distributed samples with a spiked covariance matrix. The corresponding GLRT for CD, denoted $\hat{\Lambda}_{\mathrm{LRCG}}$, corresponds to (4.1) and (4.3) with the following distribution/parameters:

$$\begin{aligned} &\text{Model:} && \mathbf{z}_i^t \sim \mathcal{CN}\left(\mathbf{0}, \tau_i^t \left(\boldsymbol{\Sigma}_k^t + \sigma_t^2 \mathbf{I}\right)\right) \\ &\text{Param.:} && \mathrm{H}_0 : \boldsymbol{\theta}_0 = \left\{\boldsymbol{\Sigma}_k^0, \sigma_0^2, \{\tau_i^0\}_{i=1}^n\right\} \\ & && \mathrm{H}_1 : \{\boldsymbol{\theta}_t\}_{t=1}^T = \left\{\boldsymbol{\Sigma}_k^t, \sigma_t^2, \{\tau_i^t\}_{i=1}^n\right\}_{t=1}^T \end{aligned} \tag{4.14}$$

Here, the test accounts for a possible change of both the covariance matrix and the textures between acquisitions, as it was shown to be the most relevant approach for SAR-ITS [Mian et al., 2019]. This GLRT can be computed with appropriate modifications of the spiked covariance matrix estimation algorithm in [Sun et al., 2016].

4.4 Application to real data

The performance of the proposed change detector is tested on a SAR ITS dataset, and assessed with ROC curves (displaying the probability of detection versus the false alarm rate). As a mean to assess the effectiveness of combining LR structure with a robust model, it is compared to the following detectors:

i) the unstructured Gaussian detector proposed in [Conradsen et al., 2003, Ciuonzo et al., 2017], denoted $\hat{\Lambda}_G$,

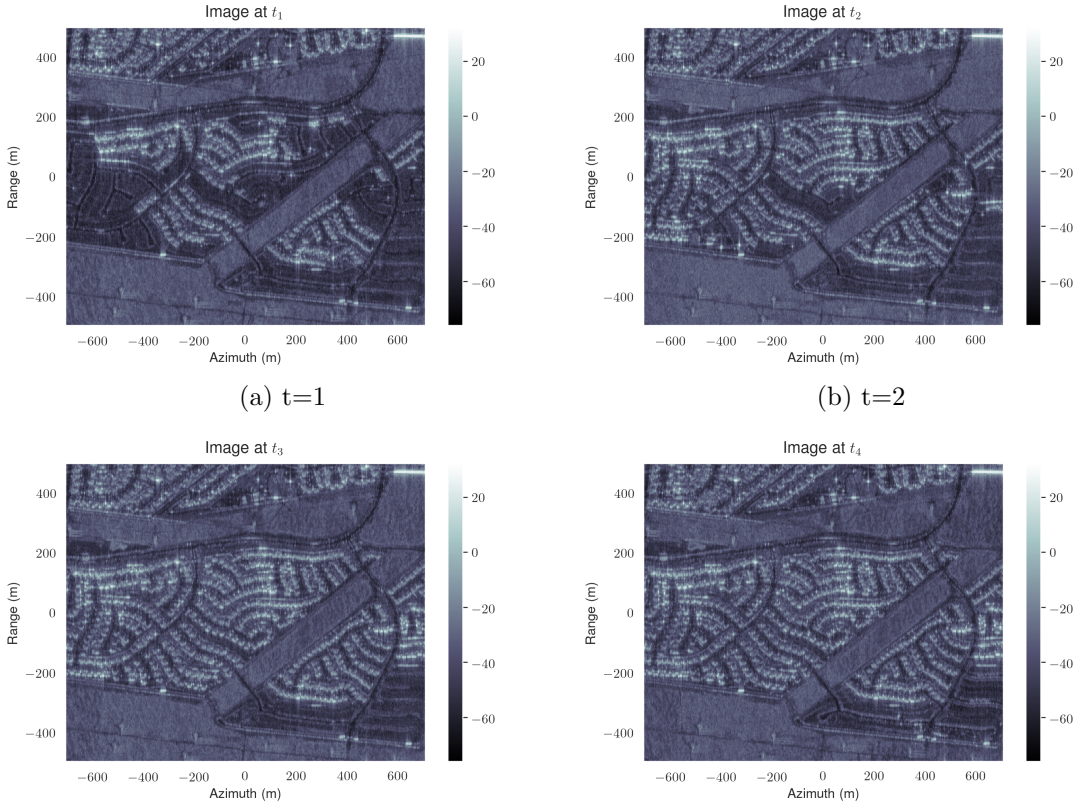
ii) the spiked Gaussian detector of [Ben Abdallah et al., 2019c], denoted $\hat{\Lambda}_{LRG}$

iii) the compound Gaussian detector proposed in [Mian et al., 2019], denoted $\hat{\Lambda}_{CG}$

iv) the spiked compound Gaussian detector proposed in [Mian et al., 2019], denoted Λ_{LRCG}.

Table 4.1: Description of SAR data used

Dataset	Url	Resolution	Scene	p	T	Size	Coordinates (top-left px)
UAVSAR SanAnd_26524_03 Segment 4 April 23, 2009 - May 15, 2011	https://uavsar.jpl.nasa.gov	Rg: 1.67m Az: 0.6m	Scene 1	12	4	2360×600 px	[Rg, Az] = [2891, 28891]
			Scene 2	12	2	2300×600 px	[Rg, Az] = [3236,25601]



(a) t=1 (b) t=2

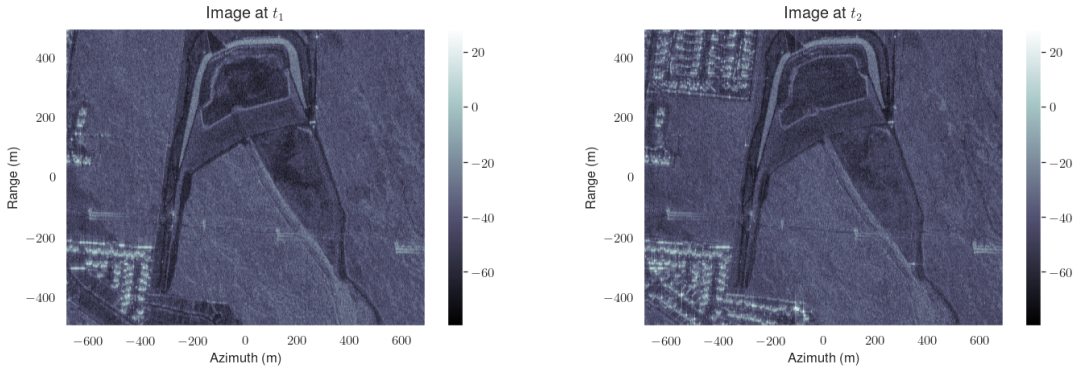
(c) t=3 (d) t=4

Figure 4.2: UAVSAR Dataset used in this study for the scene 1. Four dates are available between April 23, 2009 and May 15, 2011.

4.4.1 Datasets description

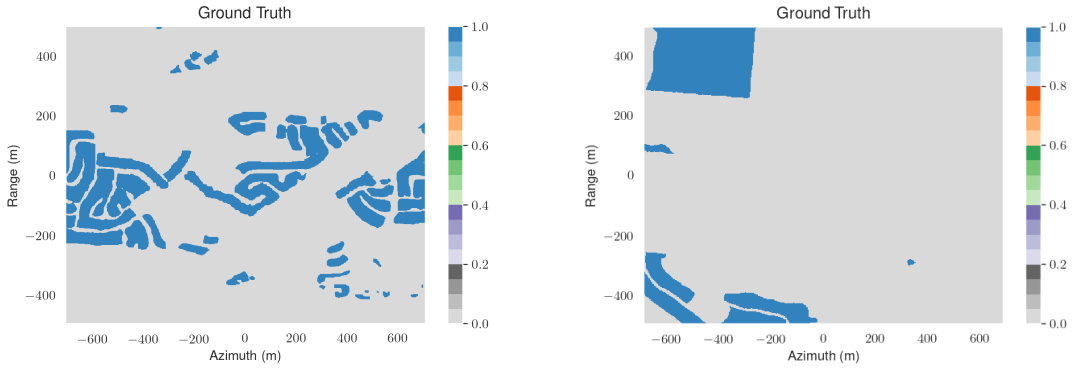
The SAR ITS data set is taken from UAVSAR (courtesy of NASA/JPL-Caltech). Two scenes with respectively 4 and 2 images are used. They are displayed in Figure 4.2 for the scene 1 (4 images) and 4.3 for the scene 2 (2 images). The CD ground truths are collected from [Ratha et al., 2017, Nascimento et al., 2018] and are shown in figures 4.4. Table 4.1 gives an overall perspective of the scenes used in the study. The SAR images correspond to full-polarisation data with a resolution of 1.67m in range and 0.6m in azimuth. Since the scatterers present in this scene exhibit an interesting spectro-angular behavior, each polarization of these images has been subjected to the wavelet transform presented in [Mian et al., 2019], allowing to obtain images of dimension $p = 12$.

The rank k is chosen according to Figure 4.5, which displays the eigenvalues of the total sample covariance matrix. For this data set, $k = 3$ appears to be an interesting value to separate signal from noise components. Notably, this rank gathers 81% of the total variance.



(a) t=1 (b) t=2

Figure 4.3: UAVSAR Dataset used in this study for the scene 2. Four dates are available between April 23, 2009 and May 15, 2011.



(a) Scene 1 (b) Scene 2

Figure 4.4: Ground truth for scenes 1 and 2

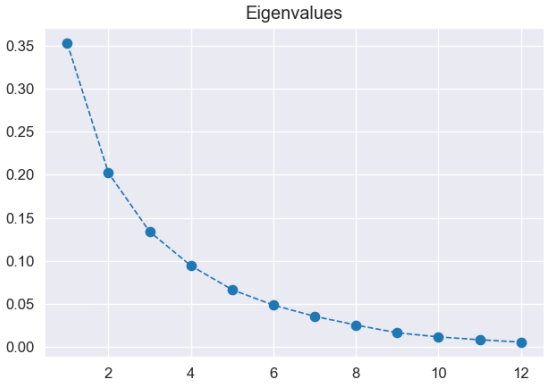
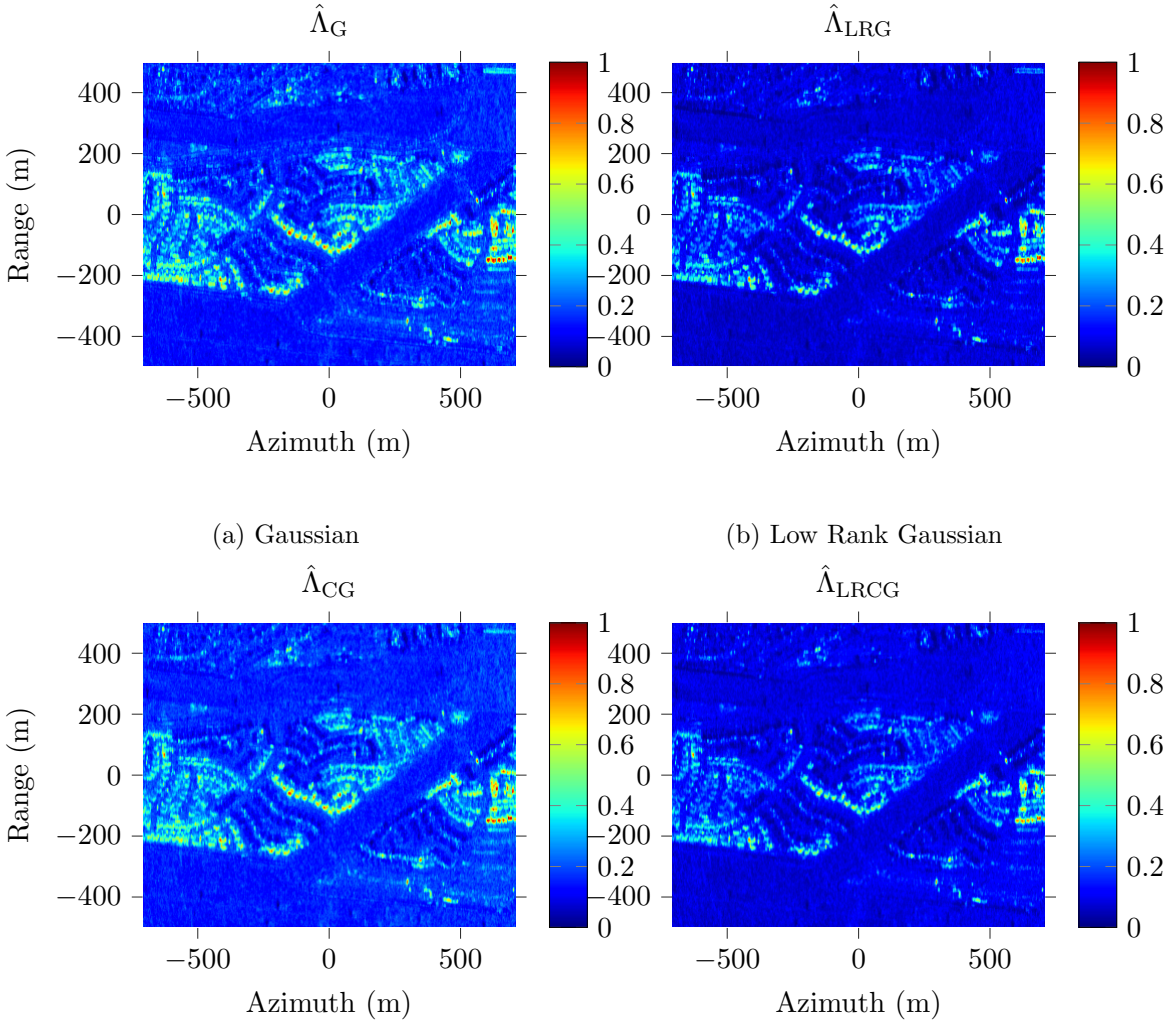


Figure 4.5: Repartition of eigenvalues mean over the ITS for the scene 1.

4.4.2 Results

Comparison between the four methods

Figure 4.6 displays the outputs of the 4 detectors applied to scene 1. From visual inspection, the levels of the false alarms appear lower for the low-rank based detectors. Figure 4.8 confirms this insight, and also assesses that the LRCG method achieves the best performance in terms of probability of



(a) Gaussian

(b) Low Rank Gaussian

(c) Compound Gaussian

(d) Low Rank Compound Gaussian

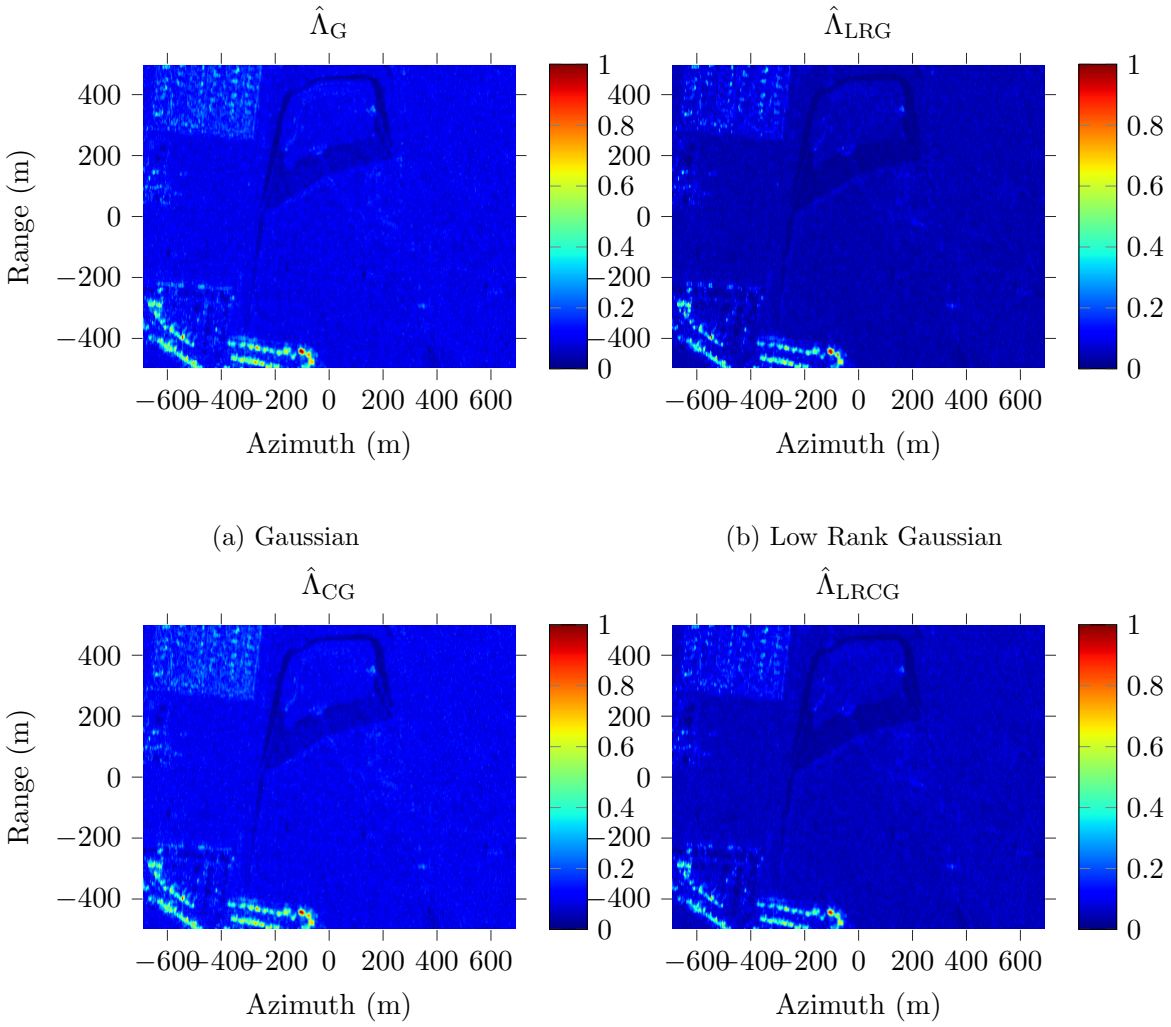
Figure 4.6: Outputs of the 4 methods for the scene 1: Gaussian, Low Rank Gaussian, Compound Gaussian (CG) and Low Rank Compound Gaussian (LRCG). Rank is fixed as 3, the window size is 7×7 and σ^2 is assumed unknown for low rank models.

detection versus false alarm rate. For the scene 2, the same conclusions can be drawn from Figure 4.7 (detectors output) and Figure 4.8 (ROC curves).

Robustness to parameter selection

Figure 4.9 (left) displays the ROC curves of Λ_{LRCG} on the scene 1 for three different values of the rank k. It is interesting to notice that these curves do not vary significantly with respect to this parameter. Therefore, we can expect that a slight error in the rank estimation will not lead to a significant drop in CD performance.

In [Ben Abdallah et al., 2019b], the noise variance σ^2 is pre-estimated locally with the mean of the $(p-k)$ lowest eigenvalues of the sample covariance matrix of all samples in the patch. This value is then used as a known parameter in the detector. Figure 4.9 (right) compares this method with the fully adaptive one (cf. (4.10)-(4.11)). It shows that the parameter σ^2 can be left as a degree of freedom at each t without loosing in terms of CD performance.



(a) Gaussian

(b) Low Rank Gaussian

(c) Compound Gaussian

(d) Low Rank Compound Gaussian

Figure 4.7: Outputs of the 4 methods for the scene 2: Gaussian, Low Rank Gaussian, Compound Gaussian (CG) and Low Rank Compound Gaussian (LRCG). Rank is fixed as 3, the window size is 7×7 and σ^2 is assumed unknown for low rank models.

4.5 Perspectives

Finally, we can list several direct perspectives following from this work:

- The presented framework assumed zero-mean samples. The approach is relevant for SAR imaging, but not in general for satellite image time series. The extension the case of unknown mean is therefore an interesting prospect as it would allow us to apply the proposed methodologies to other data-sets. Especially, the spiked models appear suited to hyperspectral imaging, where the sample dimension p is generally too high to apply local covariance-based processes.

- The considered formulation requires to fix the rank k in spiked models prior to any derivation. A potential extension of our work would be to investigate the change of the rank within a CD process.

- Detecting specific phenomenons through structural changes in the covariance matrix can be an interesting approach, but it requires a strong prior on the physics of the data. Some promising

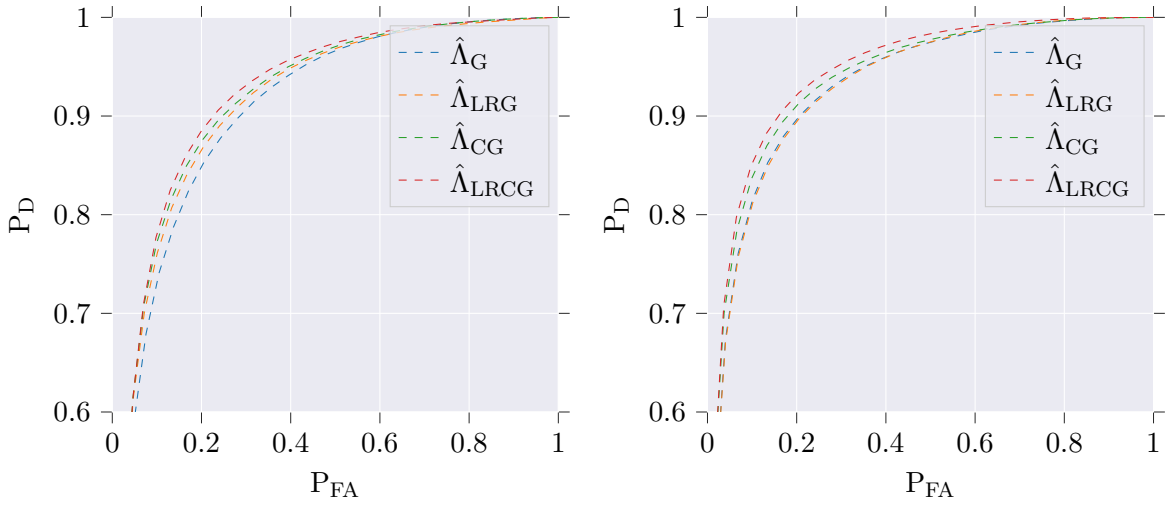


Figure 4.8: Comparison between 4 methods for the scene 1 (left) and 2 (right): Gaussian, Low Rank Gaussian, Compound Gaussian (CG) and Low Rank Compound Gaussian (LRCG). Rank is fixed as 3, the window size is 7×7 and σ^2 is assumed unknown for both low rank models.

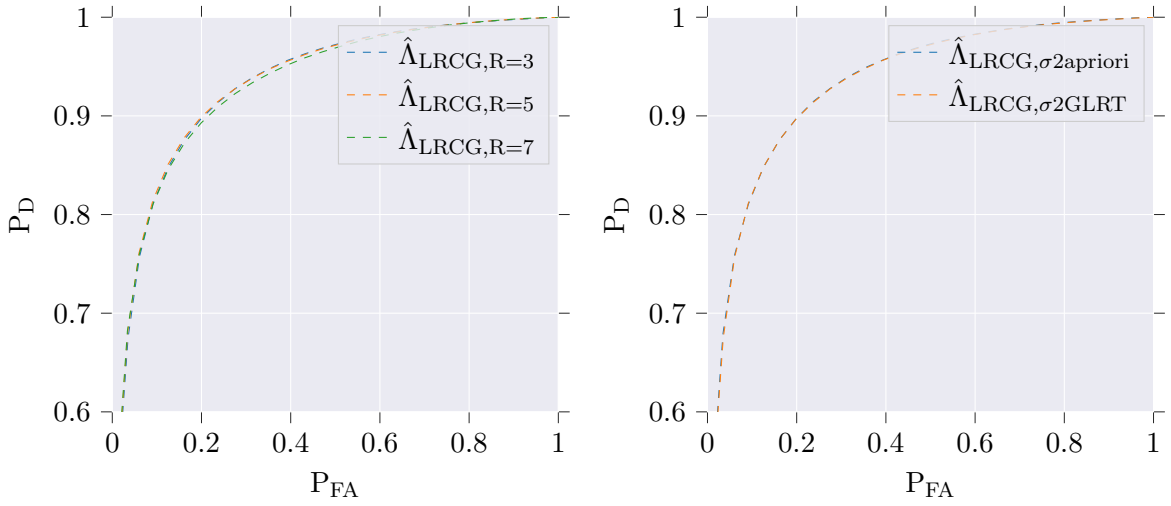


Figure 4.9: Robustness to parameter selection for LRCG: (left) Comparison of several LRCG on the scene 1 for different rank values. The window size is fixed at 7×7 and σ^2 is assumed unknown. (right) Comparison of the ROC curves between assumed known σ^2 and assumed unknown σ^2 (integrated in GLRT). $R = 3$, Window size is equal to 7×7.

results were obtained in SAR tomography in collaboration with A. Taylor. However, the lead remains to be further explored.

4.6 Onward to the next chapter

This chapter presented an application-oriented conclusion to the statistical model discussed in the chapters 2 and 3. Following from this, the last chapter will still address problems related to subspace recovery and detection. However, it will not involve statistical models, but rather consider recovery/detection problems through geometric formulations.

5 | Geometric approaches for subspace recovery

Contents

5.1 Contributions of the chapter

This chapter presents some of my recent works, focused on geometric formulations for subspace recovery. First, we present a new class of majorization-minimization algorithms for sparse PCA, where the objective function is composed of an M-estimation type subspace fitting term plus a regularizer that promotes sparsity in the principal components. Second, we present a prospective work, where we introduce a reformulation of the radar detection problem as a robust subspace clustering problem (i.e. recovering multiple linear subspaces from a heterogeneous data set).

The first contribution presented in this chapter details works conducted in collaboration with Hong Kong University of Science and technology (HKUST).
Related publications in the cv (page xi): [X1].

The second contribution presented in this chapter details works initiated in collaboration with North Carolina State University (NCSU), that has been further developed in the thesis of Bruno Mériaux.
Related publications in the cv (page xi): [C14], [C20].

5.2 Sparse PCA with majorization-minimization

5.2.1 Motivations

In standard PCA, the estimated principal components are usually dense (i.e., a linear combinations of all entries of the variables). Since the principal components have an actual physical meaning in many applications, estimating sparse principal components can significantly help the interpretation, as well as the variable selection process. As an example, we can cite the analysis of gene

expression data, where the aim is to identify only a few relevant genes (generally among thousands) within the principal components. Many algorithms have been proposed to perform this task [Chen et al., 2012, Chen and Huang, 2012, Bunea et al., 2012, Hu et al., 2016, Benidis et al., 2016, Uematsu et al., 2017]. Most of the proposed methods involve an orthonormal basis $\mathbf{U} \in \mathrm{St}(p,k)$, whose columns represent the principal components, and can be generically formulated through the problem

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\text{minimize}} \quad \mathcal{L}(\mathbf{U}, \{\mathbf{z}_i\}_{i=1}^n) + \lambda\xi\left(\mathbf{U}\right), \tag{5.1}$$

where $\mathcal{L}$ is a data fitting term (orthogonal regression on the dataset $\{\mathbf{z}_i\}_{i=1}^n$), ξ is a sparsity promoting penalty (e.g., the ℓ_1-norm), and λ is a regularization parameter. The introduction of the sparse penalty usually makes the minimization in (5.1) hard to tackle under the orthonormality constraint. Thus, most algorithms in the literature relax this constraint and resort to a trade-off [Chen et al., 2012, Chen and Huang, 2012, Bunea et al., 2012, Hu et al., 2016]. A natural way to handle the orthonormality constraint is to turn to the framework of Riemannian optimization for $\mathrm{St}(p,k)$ [Edelman et al., 1998, Manton, 2002, Absil et al., 2009]. However, these methods can be computationally expensive, and handling sparsity promoting penalties in this framework has only recently been addressed [Chen et al., 2018, Huang and Wei, 2019]. In the following, these issues will be addressed by using the majorization-minimization framework. The contribution is twofold:

- We proposed a framework to deal with the orthonormality constraint in a systematic manner, which is fully detailed in Appendix A. First, we present a majorization rule that transforms the initial problem in a series of orthogonal Procrustes ones, and analyze the corresponding generic algorithm. Second, we derive a catalog of corresponding majorizers for standard cost function.
- Driven by this methodology, we proposed a new set of algorithms for sparse PCA by leveraging proxies of ℓ_0-norm proposed from [Song et al., 2015]. Interestingly, the approach allows us to combine M-estimation type data fitting terms (robustness) and sparsity promoting penalties, while still ensuring orthonormality of the principal components.

5.2.2 Robust Sparse PCA with MM

We consider the generic formulation in (5.1). The following subsections detail options for the data fitting term and penalties that can be managed with the presented MM framework.

Robust data fitting term

In the context of probabilistic PCA, the data fitting term can be chosen as the log-likelihood of a statistical model, such as the ones presented in chapters 2 and 3. For the corresponding objective functions, one can leverage the linear surrogates from [Sun et al., 2016, Breloy et al., 2016, Ben Abdallah et al., 2020] to fit the proposed framework. In this work, we will rather focus on a geometric approach inspired by the robust subspace recovery (RSR) algorithms of [De La Torre and Black, 2003, Maronna, 2005, Ding et al., 2006, Lerman and Maunu, 2017]. Its main interest is to offer a flexible formulation that does not involve side parameters, while being robust to potential outliers within this set [Lerman and Maunu, 2018]. Thus, we consider the function

$$\mathcal{L}(\mathbf{U}, \{\mathbf{z}_i\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\mathrm{d}^2\left(\mathbf{U}, \mathbf{z}_i\right)\right), \tag{5.2}$$

where

$$\mathrm{d}^2\left(\mathbf{U}, \mathbf{z}\right) = \|(\mathbf{U}\mathbf{U}^H)^{\perp}\mathbf{z}\|_F^2 = \mathbf{z}^H\mathbf{z} - \mathrm{Tr}\left(\mathbf{U}^H\mathbf{z}\mathbf{z}^H\mathbf{U}\right) \tag{5.3}$$

is the Euclidean distance between a vector $\mathbf{z} \in \mathbb{C}^p$ and the subspace spanned by $\mathbf{U} \in \mathrm{St}(p,k)$, and where $\rho : \mathbb{R} \to \mathbb{R}$ is a function that ensures the robustness to outliers. Here, ρ is assumed to be a concave nondecreasing function, which holds for a wide variety of usual robust formulations from the literature, as illustrated by the following examples:

Example 5.2.1. ***(ℓ_p-norm)*** *For $p > 0$, ℓ_p-norm nonconvex RSR estimators are defined as in* (5.2) *by using the function*

$$\rho_p(t) = t^{p/2}. \tag{5.4}$$

The least-square estimator (standard PCA) is obtained for $p = 2$.

Example 5.2.2. ***(Huber-type)*** *For a parameter $T \geq 0$, Huber-type nonconvex RSR estimators are defined as in* (5.2) *by using the function*

$$\rho_{\mathrm{H}}(t) = \begin{cases} t/\sqrt{T} & \text{if } t \leq T, \\ \sqrt{t} & \text{if } t > T. \end{cases} \tag{5.5}$$

The median estimator, e.g. considered in [Ding et al., 2006, Lerman and Maunu, 2017], corresponds to the case limit case $T \to 0$.

Example 5.2.3. ***(Cauchy–Lorentz-type)*** *For a parameter $T \geq 1$, Cauchy–Lorentz-type nonconvex RSR estimators are defined as in* (5.2) *by using the function*

$$\rho_{\mathrm{CL}}(t) = T \ln(T + t). \tag{5.6}$$

Example 5.2.4. ***(Geman-McClure-type)*** *For a parameter $T \geq 0$, Geman-McClure-type nonconvex RSR estimators are defined as in* (5.2) *by using the function*

$$\rho_{\mathrm{GMC}}(t) = t/(T + t), \tag{5.7}$$

which has been used in, e.g., [De La Torre and Black, 2003].

Sparse penalties with linear surrogates on $\mathrm{St}(p, k)$

In order to design sparsity promoting penalties suited to the MM framework, we follow the approach proposed in [Song et al., 2015], i.e., approximating the ℓ_0-norm by a smooth function denoted l_γ^ϵ, and defined as

$$l_\gamma^\epsilon(x) = \begin{cases} a|x|^2, & \text{if } |x| \leq \epsilon \\ l_\gamma(x) - b, & \text{if } |x| > \epsilon, \end{cases} \tag{5.8}$$

with appropriate constants a and b so that the approximations l_γ^ϵ are continuous and differentiable (cf. [Song et al., 2015]), and where l_γ belongs to the family of functions (involving a tuning parameter γ):

a) ℓ_γ-norm [Gorodnitsky and Rao, 1997, Chartrand and Yin, 2008, Lai et al., 2013]:
$l_\gamma(x) = |x|^\gamma,\ \gamma \in (0, 1]$

b) ℓ_1-norm approximation from [Sriperumbudur et al., 2011, Candès et al., 2008]:
$l_\gamma(x) = \ln(1 + |x|/\gamma) \ln(1 + 1/\gamma),\ \gamma > 0$

c) lower bound of sign function from [Mangasarian, 1996]:
$l_\gamma(x) = 1 - e^{-|x|/\gamma},\ \gamma > 0.$

Thus, this class covers most of standard 1-dimensional sparsity forcing penalties (i.e., a proxy of the sign function). In order to mimic a weighted ℓ_0-norm, we consider the cost function:

$$r_0(\mathbf{U}) = \sum_{r=1}^{R} \omega_r \sum_{n=1}^{N} l_\gamma^\epsilon\left([\mathbf{u}_r]_n\right), \tag{5.9}$$

where ω_r are positive weights and with l_γ^ϵ in (A.29), and where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$, and $[\cdot]_n$ denotes the n^{th} element of a vector. Such type of penalty was initially proposed for Sparse PCA in [Benidis et al., 2016] using the ℓ_1-norm approximation [Sriperumbudur et al., 2011, Candès et al., 2008] (i.e., l_γ^ϵ in the class *b*) of the considered family).

Algorithm derivation

In this section, an MM algorithm is derived in order to solve problem (5.10). First, recall that we consider the problem

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}\rho\left(\mathrm{d}^2\left(\mathbf{U},\mathbf{z}_i\right)\right)+\lambda r_0\left(\mathbf{U}\right). \tag{5.10}$$

Following the surrogates given in [Breloy et al., 2020], the objective of this problem admits a linear surrogate function on $\mathrm{St}(p,k)$, given in the following proposition:

Proposition 5.2.1. *The objective in* (5.10) *can be majorized on* $\mathrm{St}(p,k)$ *at point* $\mathbf{U}^t$ *by a surrogate satisfying* (A.7) *with*

$$\mathbf{R}(\mathbf{U^t})=\frac{1}{n}\mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}^t-\lambda\mathbf{H}\left(\mathbf{U}^t\right) \tag{5.11}$$

with

$$\mathbf{M}\left(\mathbf{U}\right)=\sum_{i=1}^{n}\rho'\left(\mathrm{d}^2\left(\mathbf{U},\mathbf{z}_i\right)\right)\mathbf{z}_i\mathbf{z}_i^H \qquad \textit{and} \qquad \begin{cases}\mathbf{H}\left(\mathbf{U}^t\right)=\left[\ \mathbf{W}_1\mathbf{u}_1^t\ \middle|\ \cdots\ \middle|\ \mathbf{W}_R\mathbf{u}_R^t\ \right]\\ \mathbf{W}_r=\mathrm{diag}(\ \mathbf{w}_r\ -\ \mathrm{w}_r^{\max}\ \mathbf{1}_N\ )\\ \left[\mathbf{w}_r\right]_n=\omega_r\phi\left(\left[\mathbf{u}_r^t\right]_n\right)\\ \mathrm{w}_r^{\max}=\max(\mathbf{w}_r),\end{cases} \tag{5.12}$$

and where the function ϕ *is a majorizing function that depends on the chosen* l_γ^ϵ *(full expressions are detailed in Proposition A.3.5 of Appendix A).*

Therefore, we can apply Algorithm 2 of Appendix A, which leads to the following MM iterations:

$$\mathbf{U}^{t+1}=\mathcal{P}_{\mathrm{Proc}}\left\{\frac{1}{n}\mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}^t-\lambda\mathbf{H}\left(\mathbf{U}^t\right)\right\}. \tag{5.13}$$

This algorithm will be referred to as *Robust-Sparse PCA* (RSPCA).

5.2.3 Numerical validations

First, we study the robustness of RSPCA to corruption by outliers in the sample set. The aim is to illustrate the interest of both the robust fitting objective and the sparsity promoting penalty. The simulation setup is built around the so-called *haystack* model [Lerman et al., 2015], which corresponds here to a mixture of orthogonal Gaussian distributions plus additive noise:

$$\begin{aligned}\{\mathbf{z}_i\}_{i=1}^n&=\{\{\mathbf{z}_i^{\mathrm{in}}\}_{i=1}^{n_{\mathrm{in}}},\{\mathbf{z}_i^{\mathrm{out}}\}_{i=n_{\mathrm{in}}+1}^{n}\}\\ \mathbf{z}^{\mathrm{in}}&\sim\mathcal{CN}(\mathbf{0},\mathrm{SNR}\times\mathbf{U}\mathbf{U}^H+\mathbf{I})\\ \mathbf{z}^{\mathrm{out}}&\sim\mathcal{CN}(\mathbf{0},\mathrm{ONR}\times\mathbf{U}_\perp\mathbf{U}_\perp^H+\mathbf{I}),\end{aligned} \tag{5.14}$$

where $\mathbf{U}$ and $\mathbf{U}_\perp$ are built from the canonical basis such that $\mathbf{U}^H\mathbf{U}_\perp=\mathbf{0}$ (hence $\mathbf{U}$'s sparsity is maximal), SNR is signal to noise ratio, and ONR is outlier to noise ratio. We will compare four algorithms: *i*) RSPCA with least square fitting (cf. Example 5.2.1) and $\lambda=0$ (i.e., the standard PCA); *ii*) RSPCA with least square fitting and $\lambda=100$; *iv*) RSPCA with Huber cost (cf. Example 5.2.2) and $\lambda=0$ (i.e., RSR); *iii*) RSPCA with Huber cost and $\lambda=1000$. Figure 5.1 displays the average fraction of recovered energy of each algorithm with respect to ONR and the fraction of outliers in the sample set. We can notice that the use of a robust cost improves the performance compared to the standard PCA. Moreover, the introduction of the sparse penalty improves the results in terms of AFE, but also interestingly improves the robustness of the estimation process. Here, it is worth mentioning two critical points: *a*) RSPCA appears very robust when a valid signal subspace basis is actually sparse, which is probably because the sparse penalty contributes to naturally discard dense outliers. If the true subspace basis is dense, results can be degraded in practice. *b*) The starting

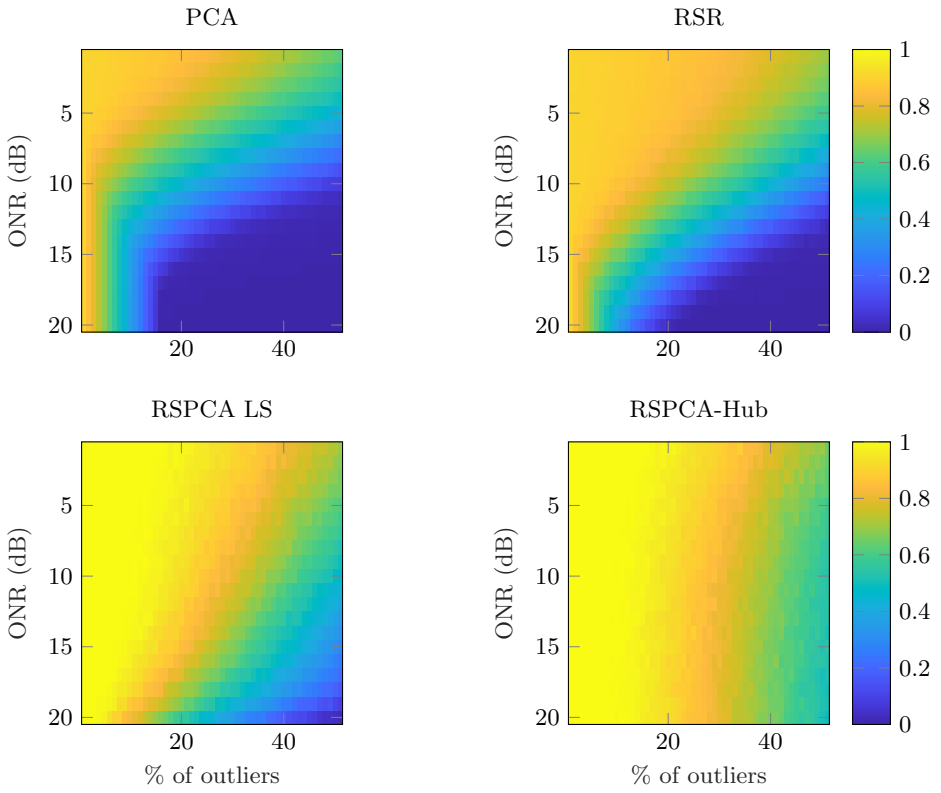


Figure 5.1: AFE versus ONR and number of outliers for various algorithms. $p = 100$, $k = 15$, $n = p$, SNR $= 10$. RSPCA built with r_0 penalty.

point plays an important role in the achieved robustness of all iterative algorithms (even without regularization). In these simulations, we used the spherical PCA (PCA applied on the normalized samples) as starting point. It has been observed that the achieved robustness can be lowered by using the standard PCA instead.

Second, we compare the performance of RSPCA with state of the art sparse PCA algorithms on the Leukemia data set [Golub et al., 1999][1]. The data consists in gene expression measurements from RNA micro-array: $p = 7129$ gene are studied for $n = 72$ patients. Remark that $n \ll p$, which suggests the interest for both dimension reduction and variable selection. The studied performance criteria are the following

$$\begin{aligned} \mathrm{SP}(\hat{\mathbf{U}}) &= 1 - ||\hat{\mathbf{U}}||_0/(pk) \\ \widetilde{\mathrm{CPEV}}(\hat{\mathbf{U}}) &= \mathrm{Tr}\{(\mathrm{span}(\hat{\mathbf{U}}))^H \mathbf{Z}\mathbf{Z}^H \mathrm{span}(\hat{\mathbf{U}})\}/\mathrm{Tr}\{\mathbf{Z}\mathbf{Z}^H\} \\ \mathrm{NOR}(\hat{\mathbf{U}}) &= ||\hat{\mathbf{U}}^H\hat{\mathbf{U}} - \mathbf{I}||_F^2 \end{aligned} \tag{5.15}$$

which measure respectively the sparsity (expected to be close to 1, i.e., 100%), the explained variance (expected to be close to 1, i.e., 100%), and the non-orthonormality (expected to be low). As the goal is to explain as much as possible, with less entries, we are interested in studying the explained variance-sparsity trade-off. In this case, notice that the $\widetilde{\mathrm{CPEV}}$ criterion may slightly favor the algorithms that relax the orthonormality constraint in the principal components, so we also check this property with respect to the sparsity. The RSPCA estimator is build with a GMC cost (cf. Example 5.2.4), r_0 penalty and l_ϵ^γ is from the proxy of the lower bound of the sign function (i.e., c) in the familly of proposed proxies). We notice that similar conclusions can be drawn with other objectives, up to minor changes of the parameters. RSPCA is computed with Algorithm 2 using an outer loop, decreasing ϵ

[1]Described in https://github.com/ramhiser/datamicroarray/wiki/Golub-(1999)

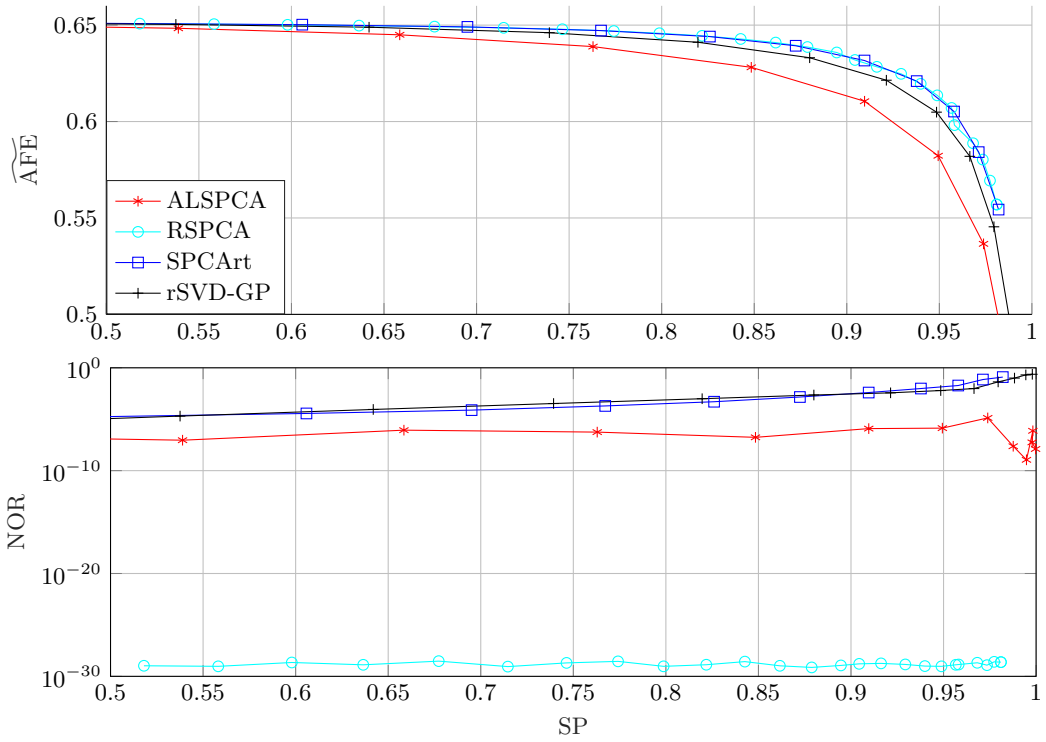


Figure 5.2: $\widetilde{\text{CPEV}}$ and NOR versus SP for various sparse PCA algorithms on the Leukemia data set. $p = 7129$, $k = 10$, $n = 72$.

from 10^{-1} to 10^{-7} in order to avoid potential local minima. This algorithm is compared to ALSPCA [Lu and Zhang, 2012], SPCArt and rSVD-GP from [Hu et al., 2016]. Figure 5.2 displays the $\widetilde{\text{CPEV}}$ and NOR versus SP for the studied sparse PCA algorithms on the Leukemia data set. Interestingly, we can notice that RSPCA achieves state of the art performance when it comes to the explained variance-sparsity trade-off, but without relaxing the orthonormality constraint, as done by the other algorithms.

5.3 Robust subspace clustering for radar detection

5.3.1 Motivations

Statistical radar detection

Adaptive detection of targets embedded in a complex environment (strong clutter, jammers, etc.) is a major issue in array processing. This topic has been extensively —and is still actively— studied in the signal processing literature for a plethora of signal models and assumed noise distributions. Following the classical statistical paradigm [Kelly, 1986, Kraut et al., 2001, Kraut and Scharf, 1999, Kraut et al., 2005], the detection problem can be formulated as a binary hypothesis test (target present or not), with unknown statistical parameters (e.g., the disturbance covariance matrix). Formally, we consider the following binary hypothesis test:

$$\begin{cases} H_0 : \ \mathbf{z}_0 = \mathbf{c}_0 + \mathbf{n}_0 & ; \mathbf{z}_i = \mathbf{c}_i + \mathbf{n}_i \ , \ \forall i \in [\![1, n]\!] \\ H_1 : \ \mathbf{z}_0 = \alpha_0 \mathbf{p} + \mathbf{c}_0 + \mathbf{n}_0 & ; \mathbf{z}_i = \mathbf{c}_i + \mathbf{n}_i \ , \ \forall i \in [\![1, n]\!] \end{cases}$$

where:

- $\mathbf{z}_0$ is referred to as primary sample (tested cell) and the $\{\mathbf{z}_k\}_{i=1}^{n}$ is the secondary data set, in which samples are assumed to be i.i.d. and target-free.

- $\mathbf{p}$ is the signal to be detected, and α_0 is an unknown power/phase-shift coefficient.
- $\mathbf{c}_i,\ \forall i \in [\![0, n]\!]$ represents the interference: clutter (response of the environment) and/or jammers.
- $\mathbf{n}_i,\ \forall i \in [\![0, n]\!]$ represents the additive thermal noise of the system.

Depending on the assumed noise plus interference model, various detection schemes and likelihood ratios can be envisioned. From a robust and practical point of view, one can rely on the adaptive coherence estimator (ACE) [Kraut and Scharf, 1999, Kraut et al., 2005], also referred to as ANMF detector, which is defined as:

$$\hat{\Lambda}(\hat{\mathbf{\Sigma}}) = \frac{|\mathbf{p}^H \hat{\mathbf{\Sigma}}^{-1} \mathbf{z}_0|^2}{|\mathbf{p}^H \hat{\mathbf{\Sigma}}^{-1} \mathbf{p}||\mathbf{z}_0^H \hat{\mathbf{\Sigma}}^{-1} \mathbf{z}_0|} \underset{H_0}{\overset{H_1}{\gtrless}} \delta_{\hat{\mathbf{\Sigma}}}, \tag{5.16}$$

for a given plug-in estimator $\mathbf{\Sigma}$ of the interference-plus-noise covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_c + \sigma^2 \mathbf{I}$, computed from the secondary data $\{\mathbf{z}_k\}_{k \in [\![1,K]\!]}$ (excluding $\mathbf{z}_0$). To sum up, a classical 2-step detection process is performed as follows:

$$\begin{array}{lccc} \text{Step 1:} & \{\mathbf{z}_i\}_{i=1}^n & \xrightarrow{\text{estimation}} & \hat{\mathbf{\Sigma}} \\ \text{Step 2:} & \left\{\mathbf{z}_0, \hat{\mathbf{\Sigma}}\right\} & \xrightarrow{\text{plug-in detector}} & \text{detection} \end{array} \tag{5.17}$$

In order to improve the performance of this detection process, the estimation of the interference plus noise covariance matrix (or interference subspace) represents a crucial step. This problem thus still drives a lot of current research, notably for dealing with the problems of robustness and low sample supports (cf. chapters 2 and 3).

Limitation of statistical based methods

Most of the aforementioned detection methodologies have been built upon the availability of a homogeneous secondary data set, i.e., i.i.d. and target-free samples, that are used to estimate the unknown statistical parameters. From a practical point of view, the scanned environment can indeed be assumed stationary for a given amount of observations. However embedded systems encounter non-stationarity due to varying environment and/or switching jammers. Dealing with change points upstream is not a trivial task, which often leads to heterogeneous secondary data sets. Moreover, the secondary data are also potentially corrupted by outliers, such as targets. Generally speaking, statistical-based methods may suffer from an important performance loss if the assumed hypothesis are not met (so called mismatched situations). To alleviate this issue, the sample selection/partition could be performed and checked using a more complex estimation chain. Though, efficient in practice, this process may be tedious and computationally expensive, as it involves numerous unknown parameters. While recent works keep robustness to heterogeneity/corruption in mind, it seems interesting to explore new methodologies, such as geometrical formulations, in order to tackle these problems.

In this scope, we will present a recent exploration, where the detection problem is reformulated as a union-of-subsapces recovery. Of course, this reformulation will eventually not solve every aspect of the question, and will bring its own inherent issues. Yet, it appeared as an interesting prospect that remained to be delved into.

Heterogeneous interferences modeled as an union-of-subspaces

This section motivates the reformulation of the detection problem as a union-of-subsapces recovery from the whole data. This approach is justified by the fact that the radar clutter (and/or jamming) interference is often contained in a subspace of low dimension compared to the size of the data [Brennan and Staudaher, 1992, Goodman and Stiles, 2007]. Hence, the background of a piecewise

stationary environment can be modeled as a union-of-subspaces. Additionally, the present sources can be modeled as a known dictionary of steering-vectors multiplied by a sparse matrix of power and phase-shift coefficients. Recovering these two components from a noisy observation (the sample set) is referred to as a *robust subspace clustering* problem [Vidal, 2011, Elhamifar and Vidal, 2013, Bian et al., 2018].

We consider that the whole collected sample set $\{\mathbf{z}_i\}_{i=0}^{n}$ is not necessarily homogeneous: the interference (clutter and/or jammers) covariance and distribution may change at certain points of the acquisition, with J unknown homogeneous sub-partitions (or clusters). Denote the partitioned sample set $\{\mathbf{z}_i^j\}_{i=1}^{n_j}$ with $j \in [\![1, J]\!]$, and $\sum_{j=1}^{J} n_j = n + 1$. These samples are modeled as

$$\mathbf{z}_i^j = \mathbf{v}_i^j + \mathbf{c}_i^j + \mathbf{n}_i^j \tag{5.18}$$

with the following additive contributions:

- $\mathbf{v}_k$ is the sum of target responses, expressed as:

$$\mathbf{v}_i^j = [\mathbf{p}_1, \ \cdots \ , \mathbf{p}_P]\, \boldsymbol{\alpha}_k = \mathbf{P}\boldsymbol{\alpha}_i^j \tag{5.19}$$

  where $\{\mathbf{p}_i\}_{i=1}^{P}$ is a dictionary of known steering vectors (targets we seek to detect), and where the vector $\boldsymbol{\alpha}_i^j$ contains power/phase-shifts coefficients. Under the realistic assumption that there are few targets to be detected, only few entries in $\boldsymbol{\alpha}_i^j$ are non-zero. Therefore, these vectors are expected to be sparse.

- $\mathbf{c}_i^j$ represents the interference, such as ground clutter (response of the scanned environment) and/or jammers. Such contribution is commonly assumed to be zero-mean with an assumed existing covariance matrix $\boldsymbol{\Sigma}_c$ and following a given (possibly heavy-tailed [Ollila et al., 2012a]) distribution. In this work, the underlying distribution will be considered unknown and unspecified. A crucial point is however that, from physical considerations on the system [Brennan and Staudaher, 1992, Goodman and Stiles, 2007], we can assume that the covariance matrix $\boldsymbol{\Sigma}_c^j$ of the interference in each cluster reads

$$\boldsymbol{\Sigma}_c^j = \sum_{r=1}^{k_j} c_r^j \mathbf{u}_r^j \mathbf{u}_r^{jH} \tag{5.20}$$

  with $k_j < p$ (low-rank). Thus, in a given cluster $j \in [\![1, J]\!]$, the interference realizations lie in a low dimensional subspace and satisfy $\mathbf{c}_i^j = \boldsymbol{\Pi}_c^j \mathbf{c}_i^j \ \ \forall\ i \in [\![1, n_j]\!]$, with the rank k_j orthogonal projector $\boldsymbol{\Pi}_c^j = \sum_{r=1}^{k_j} \mathbf{u}_r^j \mathbf{u}_r^{jH}$.

- $\mathbf{n}_i^j$ represents the thermal noise, assumed to be white Gaussian with a covariance matrix $\sigma^2\mathbf{I}$.

Note that, since the interference is heterogeneous with respect to each cluster, the whole set $\{\{\mathbf{c}_i^j\}_{i=1}^{n_j}\}_{j=1}^{J}$ can be represented by a union-of-subspaces, as illustrated in Figure 5.3. From this union-of-subspaces representation, an instrumental geometric relation will be given by noticing that each interference realization lying in the hyperplane spanned by $\boldsymbol{\Pi}_c^j$ can be obtained as a linear combination of the others (when $n_j > k_j$), i.e., $\mathbf{c}_i^j = \sum_{l \neq i} \gamma_l^j \mathbf{c}_l^j$, or in matrix form

$$\mathbf{C}_j = \mathbf{C}_j \mathbf{W}_j, \ \text{with} \ [\mathbf{W}_j]_{i,i} = 0 \tag{5.21}$$

with $\mathbf{C}_j = [\mathbf{c}_1^j, \cdots, \mathbf{c}_{n_j}^j]$, and where $\mathbf{W}_j$ is the matrix containing the coefficients γ_l^j. This formulation is also referred to as a *self-representative* property of the data.

Finally, if we denote the concatenation operator $\Cup$, and the corresponding matrices

$$\begin{cases} \mathbf{Z} = \Cup\{\mathbf{z}_k^j\}, & \mathbf{C} = \Cup\{\mathbf{c}_k^j\} = \Cup\{\mathbf{C}_j\}, & \mathbf{V} = \Cup\{\mathbf{v}_k^j\}, \\ \mathbf{N} = \Cup\{\mathbf{n}_k^j\}, & \mathbf{A} = \Cup\{\boldsymbol{\alpha}_k^j\}, & \mathbf{W} = \Cup\{\mathbf{W}_j\}, \end{cases}$$

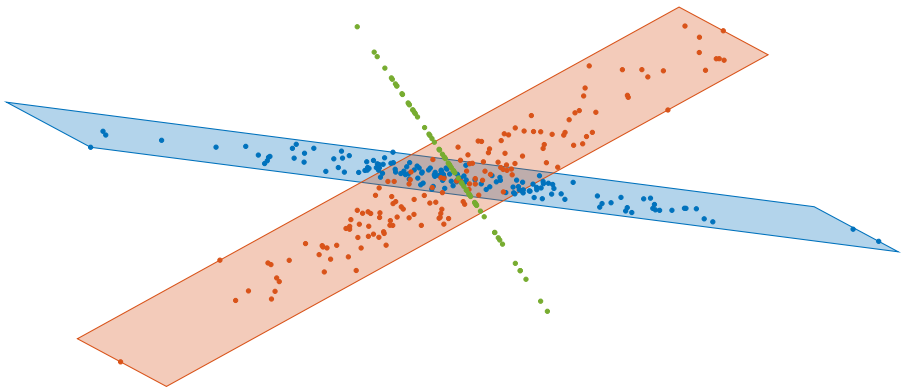


Figure 5.3: Illustration of data contained in a union of three low-dimensional subspaces

we then obtain the expression of the data matrix as:

$$\mathbf{Z} = \mathbf{PA} + \mathbf{C} + \mathbf{N}, \text{ with } \mathbf{C} = \mathbf{CW}, \text{ and } [\mathbf{W}]_{i,i} = 0 \tag{5.22}$$

To sum up on this formulation, most of the power of the samples is contained in the union of unknown subspace $\mathbf{\Pi}_c^j$, and the matrix $\mathbf{A}$ is a sparse matrix that contain the information about the present targets (outliers with respect to the low-rank subspaces).

Note that, as evoked previously, this global model is quite complex to deal with within a statistical framework. Indeed, the number of partitions J, the different distributions of interferences, their corresponding covariance matrices $\mathbf{\Sigma}_c^j$ and ranks k_j, the index of target-free samples (thus requiring simultaneous estimation and detection), are unknown. From a practical point of view, this model probably involves too many unknown parameters for deriving an efficient statistical estimation procedure such as an Expectation-Maximization algorithm. Thus, we will turn to sparse recovery algorithms in order to propose a solution. Interestingly, this formulation will not involve any assumptions about homogeneous and signal-free secondary data as in the statistical approach, and will allow the whole data to be processed in a single step.

5.3.2 Recovery algorithms

From the data model in (5.22), the problem is to recover a union of low-rank subspaces (interferences) and a sparse matrix (target responses) from a noisy observation of the matrix $\mathbf{PA} + \mathbf{C}$. Note that, in a detection application, we are primarily interested in the recovery of $\mathbf{A}$, which informs on the presence (or not) of targets in each samples. This approach, while —to the best of our knowledge— unconventional for radar detection, is under a lot of ongoing investigations for machine learning and computer vision problems [Elhamifar and Vidal, 2013]. In the following, we will thus consider the use of several sparse recovery algorithms, which are described thereafter. The technical core of the optimization algorithms will not be detailed. However, it is worth mentioning that our contribution consisted in the adaptation of existing algorithms to include the dictionary $\mathbf{P}$.

Robust subspace recovery via bi-sparsity pursuit (RoSuRe)

In order to recover the components from the data model in (5.22), we consider the following minimization problem:

$$\underset{\mathbf{W},\mathbf{A},\mathbf{C}}{\text{minimize}} \, ||\mathbf{Z} - \mathbf{PA} - \mathbf{C}||_F^2 + \lambda_1 ||\mathbf{W}||_1 + \lambda_2 ||\mathbf{A}||_1 \text{ subject to } \begin{cases} \text{(ii)} & \mathbf{C} = \mathbf{CW}, \\ \text{(iii)} & \text{diag}\,(\mathbf{W}) = \mathbf{0} \end{cases} \tag{5.23}$$

Intuitively, the ℓ_1-norm promotes sparsity of the matrices $\mathbf{W}$ and $\mathbf{A}$, which allows to recover the low dimensional structures within the self-representative relation of the data, as well as the active targets within the observations. Due to the bilinear constraint (ii), this problems is non-convex. The optimization can be achieved by the linearized version of the Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011, Lin et al., 2011], which was initially proposed in [Bian et al., 2018] (then with dictionary in [Breloy et al., 2018]). The corresponding algorithm will be referred to as RoSuRe.

Sparse subspace clustering (SSC)

Based on a rewriting introduced in [Elhamifar and Vidal, 2013], we also proposed to study a convexified modification of the problem (5.23) in [Mériaux et al., 2019]. Indeed, starting from $\mathbf{Z} = \mathbf{PA}+\mathbf{C}+\mathbf{N}$ and $\mathbf{C} = \mathbf{CW}$, we obtain

$$\begin{aligned} & \mathbf{ZW} = \mathbf{PAW} + \mathbf{CW} + \mathbf{NW} = \mathbf{PAW} + \mathbf{Z} - \mathbf{PA} - \mathbf{N} + \mathbf{NW} \\ \text{Consequently,} \quad & \mathbf{Z} = \mathbf{ZW} + \mathbf{PA}\left(\mathbf{I} - \mathbf{W}\right) + \mathbf{N}\left(\mathbf{I} - \mathbf{W}\right) \\ & \mathbf{Z} = \mathbf{ZW} + \mathbf{P}\tilde{\mathbf{A}} + \tilde{\mathbf{N}} \end{aligned} \tag{5.24}$$

Thus, we consider the following modified problem

$$\underset{\mathbf{W},\tilde{\mathbf{A}}}{\text{minimize}} \, ||\mathbf{Z} - \mathbf{P}\tilde{\mathbf{A}} - \mathbf{ZW}||_F^2 + \lambda_1||\mathbf{W}||_1 + \lambda_2||\tilde{\mathbf{A}}||_1 \text{ subject to } \text{diag}\left(\mathbf{W}\right) = \mathbf{0} \tag{5.25}$$

where the ℓ_1-norm promotes the sparsity of the matrices $\mathbf{W}$ and $\tilde{\mathbf{A}}$ and the parameter λ balances the two terms in the criterion. The problem (5.25) being convex, it can be efficiently solved using convex programming tools [Boyd and Vandenberghe, 2004]. However, the reformulation trick sightly degrades the original problem, which can lead to a biased estimate of the matrix $\mathbf{A}$. The corresponding algorithm will be referred to as SSC.

Principal component pursuit (PCP)

Another popular approach consists in recovering a "low-rank plus sparse" decomposition of the data matrix. In our context, this can be performed by the pincipal component pursuit algorithm [Chandrasekaran et al., 2011, Candès et al., 2011], that solves the problem

$$\underset{\mathbf{L},\mathbf{A}}{\text{minimize}} \, ||\mathbf{Z} - \mathbf{PA} - \mathbf{L}||_F^2 + \lambda_1||\mathbf{L}||_* + \lambda_2||\mathbf{A}||_1 \tag{5.26}$$

where $||\cdot||_*$ is the nuclear norm. In this recovery, the union-of-subspace representation is factorized in a single low-rank matrix $\mathbf{L}$. Yet, since we are mostly interested in the matrix $\mathbf{A}$, it can still provide an interesting solution when the union-of-subspaces still lies in a single low dimensional one.

5.3.3 Application to detection in non-stationary jammers

In this section, we consider the problem of target detection where the interferences are due to non-stationary jammers. Consider a uniform linear array with $p = 8$ sensors, spaced each other of $\lambda/2$, collecting $n = 50$ samples. The steering vector is given by $\mathbf{d}\left(\theta\right) = \left[1, e^{-i\pi\sin\theta}, \ldots, e^{-i\pi(p-1)\sin\theta}\right]^T$. The dictionary $\mathbf{D}$ is built from $\mathbf{d}\left(\theta\right)$, with $\theta \in [\![-90^o, 90^o[\![$. The targets of interest, which are located at $\theta_t = 40^o, 10^o, -10^o$ and -60^o at different sample times, are represented by the white markers in Figure 5.4. We consider in addition the presence of jammers, through a stochastic model, in the observation scenario, which are similar to fake targets $\mathbf{d}\left(\theta_j\right)$. For this non-stationary scenario, there exists $J = 3$ homogeneous clusters, bounded by the white dotted line in Figure 5.4. In the first sub-partition, the jammers are in $\theta_j = 20^o$ and -20^o with $\gamma_1 = \gamma_2 = 3$. In the second one, we have $\theta_j = 20^o, -45^o$ and

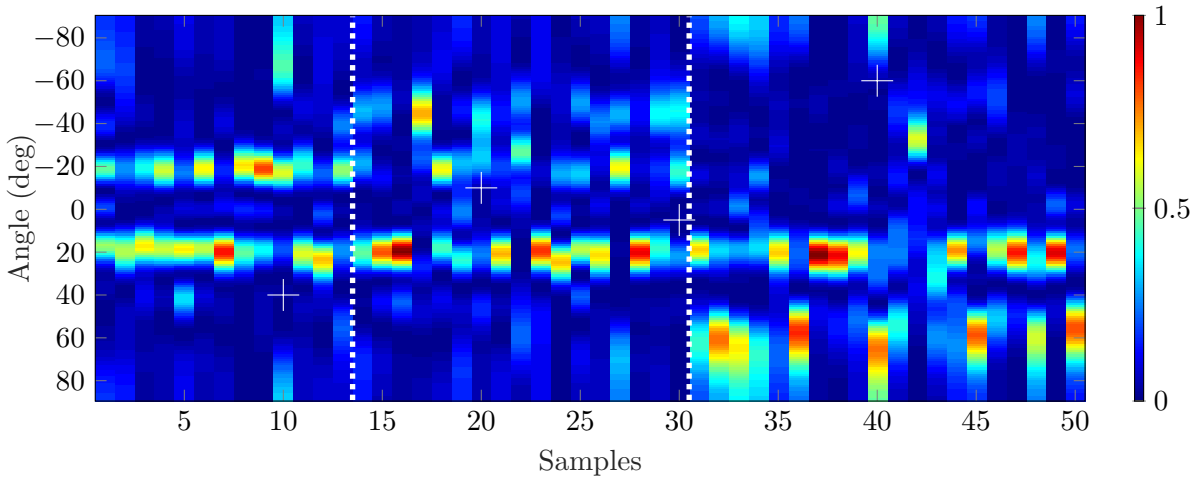


Figure 5.4: Scenario with non-stationary jammer

-25^o with $\gamma_1 = \gamma_3 = 2 = \gamma_2/2$. In the last one, we set $\theta_j = 60^o$ and 20^o with $\gamma_1 = \gamma_3 = 3$. Thus, the low-rank covariance matrix of the jammers is given in each cluster by:

$$\mathbf{R}_{\text{jam}}^{j} = \sum_{i=1}^{k_j} \mathbf{d}\left(\theta_i\right) \mathbf{d}\left(\theta_i\right)^H = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \tag{5.27}$$

where $\mathbf{U}$ and $\mathbf{\Lambda}$ are the eigen-decomposition of $\mathbf{R}_{\text{jam}}$. Then, we define the covariance matrix of the total noise by $\mathbf{R} = (\text{JNR}/\text{Tr}\{\mathbf{\Lambda}\}))\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H + \sigma^2\mathbf{I}_M$, with JNR is the Jammer to Noise Ratio. Analogously, we define the Signal to Noise Ratio (SNR) by $\text{SNR} = \|\mathbf{V}\|_2/\sigma^2$, where σ^2 is fixed equal to 1. Finally, the clutter plus noise is sampled from a Gaussian model $\mathcal{CN}\left(\mathbf{0}, \mathbf{R}\right)$.

In the considered application, we consider the standard adaptive detection methods: *i*) the Adaptive Normalized Matched Filter (ANMF), where the covariance learning is based on the sample covariance matrix, computed from the $2M$ samples surrounding the tested sample, which will serve as secondary data; *ii*) the ANMF, where the covariance learning is using the Tyler's estimator [Tyler, 1987] on the $2M$ samples surrounding the tested sample. These methods are compared to the following sparse recovery approaches: *a*) the RoSuRe-detector, which solves the problem (5.23) from the given observations $\mathbf{Z}$ and the considered dictionary $\mathbf{D}$; *b*) the SSC-detector, denoted by SSC and which is obtained by solving the problem (5.25) from $\mathbf{Z}$ and $\mathbf{D}$; *c*) the PCP-detector, denoted by m-RoSuRe and which is obtained by solving the problem (5.25) from $\mathbf{Z}$ and $\mathbf{D}$. The tuning parameters (λ_1, λ_2) for each recovery algorithms have been selected by scanning a grid prior to the simulation, then fixed in order to fairly compare the result.

Figure 5.5, displays the probability of detection with respect to the signal to noise ratio (for a probability of false alarm set to 10^{-3} for each detector). Interestingly, we can observe that the sparse recovery approach can outperform the ANMF-based methods. However, it is still hard to grasp a global pattern, as the performance of each of these method depends on the observed target.

5.4 Perspectives

MM framework for the Stiefel manifold

The framework used in section A.2 (detailed in appendix A) appears to be a practical solution to derive optimization algorithms involving the orthonormality constraints. In terms of applications, we can point the following immediate perspectives:

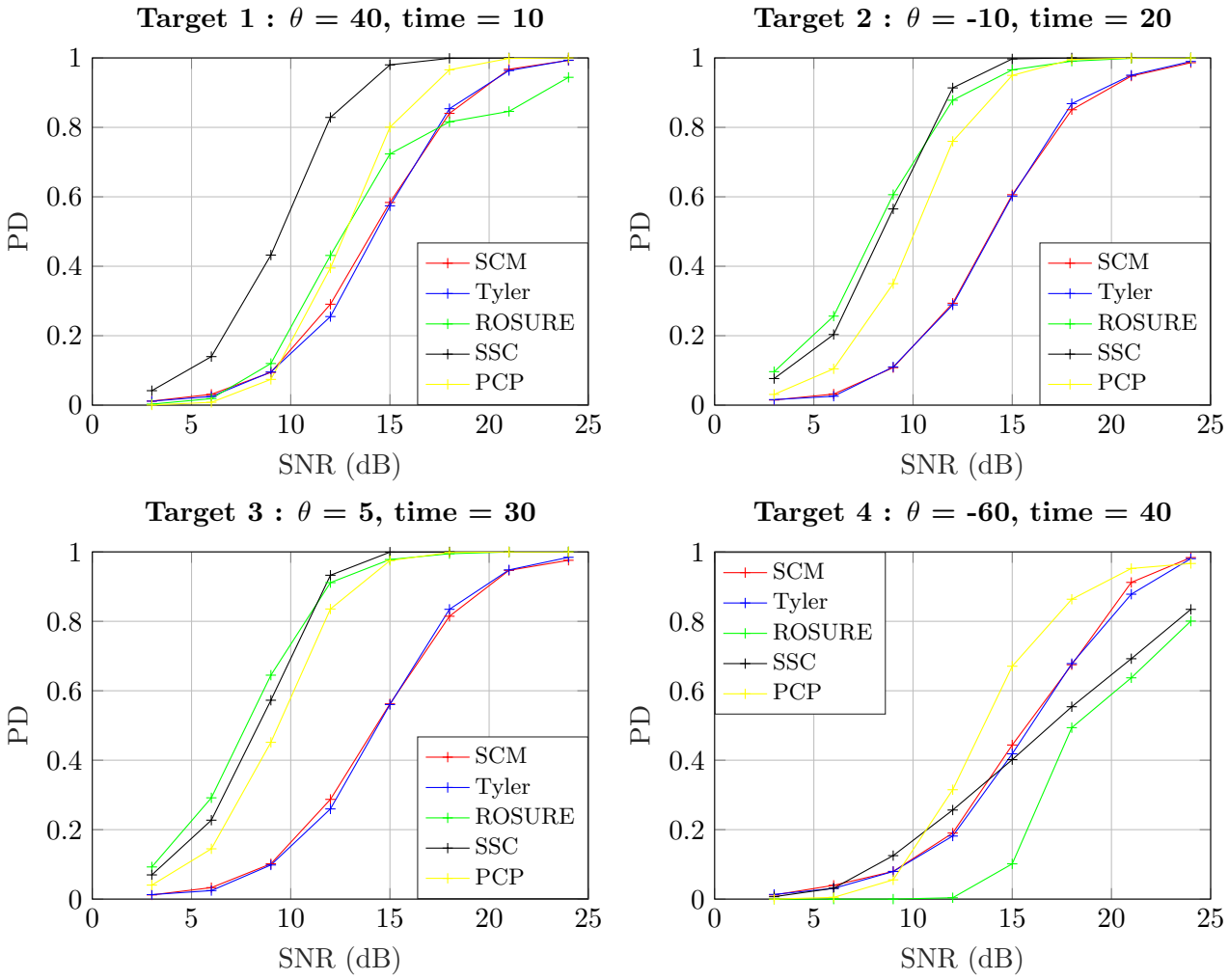


Figure 5.5: PD versus SNR for each target in the scenario. The probability of false alarm is set to 10^{-3} for each detector.

- The regularization of the eigenvectors of positive semi-definite Hermitian matrices, which can be useful for covariance matrices (discussed in section 3.6), as well as graph Laplacian matrices (discussed in section 6.2).
- The computation of robust tensor decomposition, such as the higher-order SVD (HOSVD) [Chachlakis et al., 2019].

Robust subspace clustering

Sparse subspace recovery methods leveraging the self-representativity of the data represent a new perspective that I aim to further explore, notably from a more theoretical aspect such as optimization. The radar detection problem offered a good entry point to this topic, as it brought some interesting preliminary results. The considered methodology can be applied for two purposes: either performing the detection itself by looking at the recovered sparse error matrix (revealing present targets), and/or for doing a first step clustering of homogeneous samples, that can be then used in a traditional statistical detection process. Eventually, this second option (that we did not study yet) appear probably more suited to detection problems. Indeed, such application is quite demanding in guarantees (e.g., the constant false alarm property w.r.t. to the interference), which seems hard to ensure with sparse recovery approaches.

6 | Perspectives

Contents

Wrapping up, the previous chapters presented my research output around subspace recovery and its applications. My first works adopted a statistical approach (probabilistic PCA and covariance matrix estimation), and I recently moved to geometrical formulations. In this scope, my main contributions concern the derivation of optimisation methods (majorization-minimization and Riemannian optimization) to compute robust estimators related to various problems. A part of my work is also focused on theoretical performance analysis (asymptotic characterizations and intrinsic Cramér-Rao bounds). On the practical side, these methods/studies were mainly motivated by array processing and teledetection applications (airborne radar and satellite image time series).

To conclude this synthesis, this chapter will now detail perspectives opened by these works.

6.1 Perspectives by chapters

Several immediate perspectives (i.e., extensions of the presented work) were evoked at the end of each chapter. This section rather focuses on non-trivial, "medium/long-run", leads that I aim to explore.

- **Chapter 2** presented Bayesian subspace estimation methods for mixtures of low-rank compound Gaussian models. The Bayesian priors for the subspace were drawn from directional statistics (Bingham-Langevin distributions). A main remaining issue concerns the "fast" generation of samples according to these distributions, which could be addressed by exploiting modern Langevin and Hamiltonian Monte Carlo methods for manifolds [Girolami and Calderhead, 2011]. It would also be interesting to study new distributions of subspaces from a Riemannian point of view, directly involving the natural Riemannian distance on the Grassmann manifold [Edelman et al., 1998].

  On a second note, the problem of missing data should also be addressed. This can be efficiently done within the expectation-maximization framework [Little and Rubin, 2019]. Hence, we can generalize our previous algorithms for data with missing entries. This mostly requires tedious technical derivations, but appears very important from a practical/applied point of view. We started exploring this lead in the thesis work of Alexandre Hippert-Ferrer.

- **Chapter 3** presented several theoretical performance characterizations (asymptotic analysis or Cramér-Rao bounds) related to M-estimators and complex elliptically symmetric distributions. I am now interested in studying non-asymptotic characterizations [Vershynin, 2010], in order to obtain concentration bounds for subspace recovery algorithms (as e.g. in [Uematsu et al., 2017]). Another puzzling question concerns the Riemanian intrinsic bias of maximum likelihood estimators exhibited in [Smith, 2005]: it seems quite counter-intuitive that estimates could be biased or not, depending on the chosen metric. Should the intrinsic bias be corrected? The answer

probably lies in practical/applied results, since the correction interest cannot be validated by simply measuring an error (which is metric dependent).

- **Chapter 4** presented a framework to test for shared properties in the eigendecomposition of covariance matrices from groups of observations. The approach is an interesting prospect for applications that aim to test for specific physical phenomenons, but it requires a strong knowledge of the considered application (and its underlying physics). I am currently learning about EEG signals (mostly for blind source separation), where several shared formulations indicate that the approach could be useful. Another interest of this framework is that it allows us to extract features that can be leveraged for clustering within time-series, which is a lead currently explored in the thesis of Antoine Collas.

- **Chapter 5** presented a majorization-minimization framework for the Stiefel manifold. Notably, we proposed cost functions and tricks to promote sparsity in orthonormal bases. Being sparse on a smooth manifold is a complex issue with numerous applications: sparse PCA, sparse subspace representations, sparse covariance/precision matrix estimation... A lead that I aim to explore is brought by recent proximal gradient methods for Riemannian optimization [Chen et al., 2018, Huang and Wei, 2019]. In the second part, chapter 5 presented an application of robust sparse subspace clustering methods to radar detection. In this scope, the subspace recovery problem through self-representative linear combinations could also be generalized to formulations that naturally promote inherent symmetries (such as persymetric ones), that are relevant in some applications.

6.2 Perspectives on new themes

This section details some topics I want to explore in a relatively near future. Though some connections exist with the previous works, these topics are not directly related to any chapter in particular and mostly represent new perspectives.

On Spectral regularization for graphs and covariances

Graph Laplacian learning shares a lot of common formulation with covariance matrix estimation. Thus, the regularization and optimization techniques evoked for covariance matrices can be leveraged in this context. Specifically, I want to study estimators formulated as

$$\begin{aligned} &\boldsymbol{\Sigma}(\alpha) = \operatorname{argmin}_{\boldsymbol{\Sigma}} \ \mathcal{L}(\boldsymbol{\Sigma}) + \alpha \mathcal{P}(\boldsymbol{\Sigma}), \\ &\mathcal{P}(\boldsymbol{\Sigma}) = \mathcal{P}_{\lambda}(\boldsymbol{\Lambda}) + \mathcal{P}_{\mathrm{v}}(\mathbf{V}) \text{ for } \boldsymbol{\Sigma} \overset{\mathrm{EVD}}{=} \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{H}, \end{aligned} \tag{6.1}$$

where $\mathcal{L}$ is an objective function (either for graph or covariance learning), α is a regularization parameter, and $\mathcal{P}$ is a penalty function with separate actions on the EVD parameters of $\boldsymbol{\Sigma}$. This formulation can be used in order to promote certain spectral structures, as e.g. considered in [Kumar et al., 2020a] (through constraints rather than penalties). Moreover, eigenvectors are also essential in numerous processes such as PCA and graph Fourier transform. It will therefore be interesting to control separately their behavior in a regularization process. Notably, the use of sparsity promoting penalties on eigenvectors from chapter 5 can be useful for sparse precision matrices estimation (of primary interest in graph analysis). This lead will be my main focus for my 1/2 CRCT in 2021 (mobility at Aalto University, Helsinki).

On probabilistic PCA

I am also looking for new models that go beyond mixtures of elliptical distributions. An interesting perspective is brought by the theory of copulas using a Gaussian kernel from [Woodbridge et al., 2017],

expressed as

$$\mathbf{z} \overset{d}{=} f(\mathbf{x}) \tag{6.2}$$

with $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$, and where f is an element-wise operator, i.e., $[\mathbf{z}]_i = f_i([\mathbf{x}]_i)$. This formulation allows us to change the marginal distribution of each entries, while controlling the correlation between them through $\mathbf{\Sigma}$, which appears especially interesting to model heterogeneous data. This model first needs more empirical validation (i.e., exhibiting a fit to real data) to be motivated. However, it also opens the door to numerous generalizations, as well as to the development of new subspace estimation methods.

On dimension reduction

Still concerning subspace learning, I am also interested in studying low-dimensional structures in non-linear spaces/representations. In the context of Riemannian geometry, a main question would be: what is a subspace on a manifold? Elements of response are brought by tangent PCA, principal geodesic analysis [Fletcher et al., 2004] and recent developments on the matter [Pennec et al., 2018]. A good entry point related to my work would be to study the problem of robust structured covariance estimation when the structure is not actually known. Some solutions were proposed for linearly structured covariance matrices [Soloveychik and Wiesel, 2015a], but the Riemannian point of view is —to the best of my knowledge— not yet explored in this context. In this scope, Riemannian dimensionality reduction techniques, e.g. inspired from [Harandi et al., 2018], would be interesting to explore.

Appendices

A | Majorization-minimization on the Stiefel manifold

A.1 Majorization-minimization (MM) algorithms

The general MM framework is briefly reviewed below. For more complete information, we refer the reader to [Sun et al., 2016]. Consider the following optimization problem:

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}}\ f(\mathbf{x}), \tag{A.1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a continuous function and $\mathcal{X}$ is a closed set. Given an initial point $\mathbf{x}^0 \in \mathcal{X}$, the MM procedure minimizes f over $\mathcal{X}$ by updating $\mathbf{x}$ iteratively as

$$\mathbf{x}^{t+1} \in \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}}\ g(\mathbf{x}|\mathbf{x}^t), \tag{A.2}$$

where $g(\cdot|\mathbf{x}^t) : \mathcal{X} \to \mathbb{R}$ is a *surrogate function* of f satisfying the following property:

$$\mathbf{x}^t \in \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}}\ g(\mathbf{x}|\mathbf{x}^t) - f(\mathbf{x}). \tag{A.3}$$

In other words, $g(\cdot|\mathbf{x}^t)$ upperbounds f globally over set $\mathcal{X}$ up to a constant:

$$g(\mathbf{x}|\mathbf{x}^t) - f(\mathbf{x}) \geq c^t \triangleq \{g(\mathbf{x}^t|\mathbf{x}^t) - f(\mathbf{x}^t)\}, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{A.4}$$

The sequence $\{f(\mathbf{x}^t)\}_{t\in\mathbb{N}}$ generated by (A.2) is non-increasing since

$$f(\mathbf{x}^{t+1}) \overset{(A.4)}{\leq} g(\mathbf{x}^{t+1}|\mathbf{x}^t) - c^t \overset{(A.2)}{\leq} g(\mathbf{x}^t|\mathbf{x}^t) - c^t = f(\mathbf{x}^t). \tag{A.5}$$

The MM procedure suggests thus the possibility of minimizing f by iteratively seeking for a sequence of surrogate functions $\{g(\cdot|\mathbf{U}^t)\}_{t\in\mathbb{N}}$ that are easy to minimize over the feasible set. This procedure in recapped in Figure A.1.

A.2 Systematic Procrustes reformulations for the Stiefel manifold

A.2.1 Generic algorithm

Consider a generic optimization problem where the variable $\mathbf{U}$ is constrained to the Stiefel manifold $\mathrm{St}(p,k) = \left\{\mathbf{U} \in \mathbb{C}^{p\times k} \mid \mathbf{U}^H\mathbf{U} = \mathbf{I}\right\}$:

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\text{minimize}}\quad f\left(\mathbf{U}\right), \tag{A.6}$$

where $f : \mathbb{C}^{p\times k} \to \mathbb{R}$ is a smooth differentiable objective function suited to an application of interest. Notice that optimization problems over the Stiefel manifold $\mathrm{St}(p,k)$ are nonconvex due to the orthonormality constraint. Hence, they are usually hard to deal with, even for apparently simple

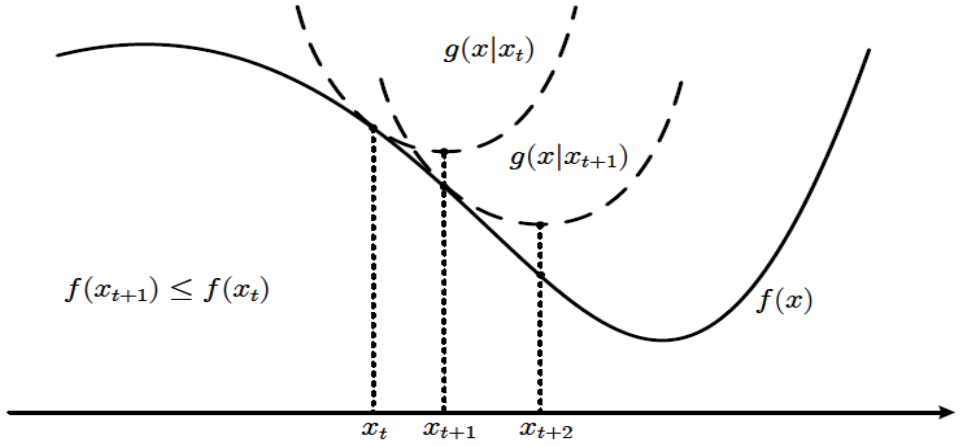


Figure A.1: MM principle: "Iteratively minmizing a smooth local tight upperbound of the objective".

objective functions f. Therefore, we consider applying the MM framework (cf. section A.1) and minimizing f by solving a sequence of simpler problems: in short, we will construct surrogate functions that are linear when restricted to the feasible set $\mathrm{St}(p,k)$. The corresponding subproblems can then be recast as an orthogonal Procrustes ones (detailed below), leading to simpled updates of the variable.

In the following, we assume that the objective in (A.6) is majorized at point $\mathbf{U}^t$ by a surrogate $g(\mathbf{U}|\mathbf{U}^t)$ that satisfies the following properties:

Assumption A.2.1. *The surrogate function $g : \mathbb{C}^{p\times k} \times \mathbb{C}^{p\times k} \to \mathbb{R}$ satisfies the following conditions:*

i) Tightness: $g(\mathbf{U}^t|\mathbf{U}^t) = f(\mathbf{U}^t)$,

ii) Continuity: $g(\cdot\,|\,\cdot)$ *is continuous on* $\mathbb{C}^{p\times k} \times \mathbb{C}^{p\times k}$,

iii) Upperbound: $g(\mathbf{U}|\mathbf{U}^t) \geq f(\mathbf{U}),\ \forall \mathbf{U} \in \mathrm{St}(p,k)$,

iv) Linearity: restricting to $\mathrm{St}(p,k)$, g *can be expressed as*

$$\begin{aligned} g\left(\mathbf{U}|\mathbf{U}^t\right) &= -\operatorname{Tr}\left\{\mathbf{R}^H\left(\mathbf{U}^t\right)\mathbf{U}\right\} - \operatorname{Tr}\left\{\mathbf{U}^H\mathbf{R}\left(\mathbf{U}^t\right)\right\} + \text{const.}, \\ &= -2\mathfrak{Re}\left\{\operatorname{Tr}\{\mathbf{U}^H\mathbf{R}\left(\mathbf{U}^t\right)\}\right\} + \text{const.}, \end{aligned} \tag{A.7}$$

where $\mathbf{R} : \mathbb{C}^{p\times k} \to \mathbb{C}^{p\times k}$ *is a matrix function of* $\mathbf{U}^t$.

Following the MM procedure described in section A.1, an update of the parameter $\mathbf{U}$ is given by

$$\mathbf{U}^{t+1} \in \underset{\mathbf{U}\in\mathrm{St}(p,k)}{\operatorname{argmin}}\ g\left(\mathbf{U}|\mathbf{U}^t\right). \tag{A.8}$$

Since g is linear (cf. (A.7)) and $\mathbf{U} \in \mathrm{St}(p,k)$, it is not hard to see that obtaining this update is equivalent to solving

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\text{minimize}}\quad ||\mathbf{R}\left(\mathbf{U}^t\right) - \mathbf{U}||_F^2, \tag{A.9}$$

which is referred to as an orthogonal Procrustes problem. When $\mathbf{R}\left(\mathbf{U}^t\right)$ is full rank[1], the problem (A.9) admits a unique solution [Manton, 2002], leading to the MM update:

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathrm{Proc}}\left\{\mathbf{R}\left(\mathbf{U}^t\right)\right\}, \tag{A.10}$$

[1]Rank deficiency is a case we do not focus on: some pathological counter-examples can be build but they rely on either *i*) a cost function f that does not satisfy the initial regularity assumptions; *ii*) a subspace within $\mathbf{U}^t$ that has reached a local stationary point. For the second point, the proposed method can still be applied by setting the stable subspace fixed and updating only the remaining portion of $\mathbf{U}^t$ (i.e., recasting the problem with $k' < k$). In practice, the issue has not been experienced with the considered cost functions.

Algorithm 1 Computation of $\mathcal{P}_{\text{Proc}}$ (projection on $\text{St}(p,k)$)

Entry: $\mathbf{R} \in \mathbb{C}^{p\times k}$
Compute the thin-SVD: $\mathbf{R} \overset{\text{TSVD}}{=} \mathbf{V}_{\text{left}}\mathbf{P}\mathbf{V}_{\text{right}}^H$
Set $\mathbf{U}$=$\mathbf{V}_{\text{left}}\mathbf{V}_{\text{right}}^H$
Output: $\mathbf{U} = \mathcal{P}_{\text{Proc}}(\mathbf{R}) \in \text{St}(p,k)$

Algorithm 2 Generic Procrustes-MM Algorithm

Entry $t=0$, $\mathbf{U}^{(0)} \in \mathbb{C}^{N\times R}$
repeat
    Compute $\mathbf{R}\left(\mathbf{U}^t\right)$ from surrogate (A.7)
    Update $\mathbf{U}^{t+1} = \mathcal{P}_{\text{Proc}}\left\{\mathbf{R}\left(\mathbf{U}^t\right)\right\}$ with Algorithm 1
    $t = t+1$
until convergence criterion is met
Output $\{\mathbf{U}^t\}$

where the operator $\mathcal{P}_{\text{Proc}}$ is defined in Algorithm 1. Eventually, solving the sequence of orthogonal Procrustes problems results in a MM procedure to optimize f under the orthonormality constraint, which is summarized in Algorithm 2.

Remark A.2.1. *Importantly, the MM approach is also applicable to objective function consisting in a sum of functions of the form*

$$f(\mathbf{U}) = \sum_{i=1}^{I} f_i(\mathbf{U}). \tag{A.11}$$

Then, if each f_i can be majorized by a linear surrogate g_i of the form

$$g_i\left(\mathbf{U}|\mathbf{U}^t\right) = -\text{Tr}\left\{\mathbf{R}_i^H\left(\mathbf{U}^t\right)\mathbf{U}\right\} - \text{Tr}\left\{\mathbf{U}^H\mathbf{R}_i\left(\mathbf{U}^t\right)\right\} + \text{const.}, \tag{A.12}$$

following the same steps as (A.8)-(A.10)*, the MM updates can simply be obtained as*

$$\mathbf{U}^{t+1} = \mathcal{P}_{\text{Proc}}\left\{\sum_{i=1}^{I}\mathbf{R}_i\left(\mathbf{U}^t\right)\right\}. \tag{A.13}$$

Obviously, this methodology cannot be applied to any arbitrary cost function. Still, Section A.3 presents a catalog of surrogate functions satisfying (A.7) for a large set of standard cost functions that can be used as building blocks. The method also suggests that simple algorithms can be obtained by designing meaningful proxies of the desired function that can be majorized by a linear surrogate on $\text{St}(p,k)$. A practical example for a proxies of ℓ_0-norm is proposed in section 5.2.2 .

A.2.2 Computational complexity

Interestingly, a single iteration in Algortithm 2 essentially involves two operations:

- The computation of the matrix $\mathbf{R}(\mathbf{U})$: this step usually involves functions of the $p \times n$ data matrix and/or multiplying this matrix with the current point $\mathbf{U}_t$. Thus, this step is generally $\mathcal{O}(npk)$. Also notice that this computation can most of the time, be parallelized. Hence it does not represent the major bottleneck of Algorithm 2, contrarily to the second step.

- The computation of $\mathcal{P}_{\text{Proc}}$: this step requires to compute thin-SVD of a tall matrix $\mathbf{R} \in \mathbb{C}^{p\times k}$ which is $\mathcal{O}(pk^2 + k^3)$.

Comparing to the existing approaches, e.g., the steepest descent on the Stiefel manifold [Absil et al., 2009], an iteration requires computing the gradient (also generally $\mathcal{O}(npk)$) and a retraction (local mapping between a point in $\mathrm{St}(p,k)$ and its tangent space). The choice of the retraction is not unique, which leads to several options, e.g., based on geodesic paths [Edelman et al., 1998], Procrustes projection [Manton, 2002], or QR decomposition [Absil et al., 2009]. Nevertheless, for all of the corresponding algorithms, the retraction step is $\mathcal{O}(pk^2+k^3)$. Hence, the computational complexity of an iteration of Algortithm 2 is on par with standard first-order based methods. However, this MM procedure is step-size free, thus it does not require the knowledge of any global parameter (such as the Lipschitz constant), or its adaptive estimation using a line search-type method. Compared to the latter option, this property effectively reduces the computational burden of each iteration, as it does not involve multiple computations of the retraction step (the computational bottleneck).

A.2.3 Convergence analysis

The convergence analysis of Algorithm 2 can be obtained by following the one of the successive upper-bound minimization (SUM) algorithm in [Razaviyayn et al., 2013]. Note that the result of [Razaviyayn et al., 2013] does not hold directly for Algorithm 2, as the SUM framework does not cover non-convex constraints. Nevertheless, this result can be adapted to $\mathrm{St}(p,k)$ as in [Benidis et al., 2016, Fu et al., 2017, Kumar et al., 2020b], leading to the following proposition:

Proposition A.2.1. *Let $\{\mathbf{U}^t\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 2. Then the following holds:*

1. *The sequence $\{f(\mathbf{U}^t)\}_{t\in\mathbb{N}}$ converges.*
2. *Every limit point $\mathbf{U}^*$ of the sequence is a critical (also referred to as Karush-Kuhn-Tucker, or KKT) point of the problem* (A.6).
3. *The whole sequence converges to $\mathcal{K}$, the set of KKT point of the problem* (A.6).

Note that the convergence to $\mathcal{K}$ does not imply the convergence of Algorithm 2 in terms of the variable $\mathbf{U}$. Establishing this property requires a case-by-case analysis: in some cases the monotonic decrement of the objective can directly imply the convergence in terms of variable [Kiers, 1995]. For the case of rotation invariant costs, this convergence in variable requires to be expressed in terms of subspace, e.g. as in [Lerman and Maunu, 2017].

A.3 Standard cost functions and their surrogates

The key to apply Algorithm 2 is to obtain a linear surrogate of the objective on the set $\mathrm{St}(p,k)$. In this section, we gather (and generalized) such surrogates functions from the literature (cf. e.g., [Kiers, 1995, Kiers, 2002]) for several standard minimizing problems: convex/concave quadratic forms, concave functions, quotients of quadratic forms, and a class of tailored sparse penalties. This catalog offers then practical building-blocks to tackle a large class of objective functions through systematic Procrustes reformulations.

A.3.1 Quadratic forms (QFs)

First, define the Brockett function [Absil et al., 2009, Sec. 4.8] for $\mathbf{U}=[\mathbf{u}_1|\cdots|\mathbf{u}_k]\in\mathrm{St}(p,k)$ as

$$f_{\mathrm{B}}(\mathbf{U}) = \sum_{r=1}^{k} d_r \mathbf{u}_r^H \mathbf{M}\mathbf{u}_r = \mathrm{Tr}\left\{\mathbf{U}^H\mathbf{M}\mathbf{U}\mathbf{D}\right\}, \tag{A.14}$$

with $\mathbf{M}\in\mathcal{H}_p^+$, and $\mathbf{D}\in\mathbb{R}^{k\times k}$ a diagonal matrix with $[\mathbf{D}]_{r,r}=d_r$ satisfying $0\leq d_1\leq\cdots\leq d_k$. In the following, functions of the form f_{B} (resp. $-f_{\mathrm{B}}$) are referred to as convex (resp. concave) QFs. Note that some other expressions of QFs exist, but they can usually be rewritten as special cases or combinations (e.g., sums) of Brockett functions.

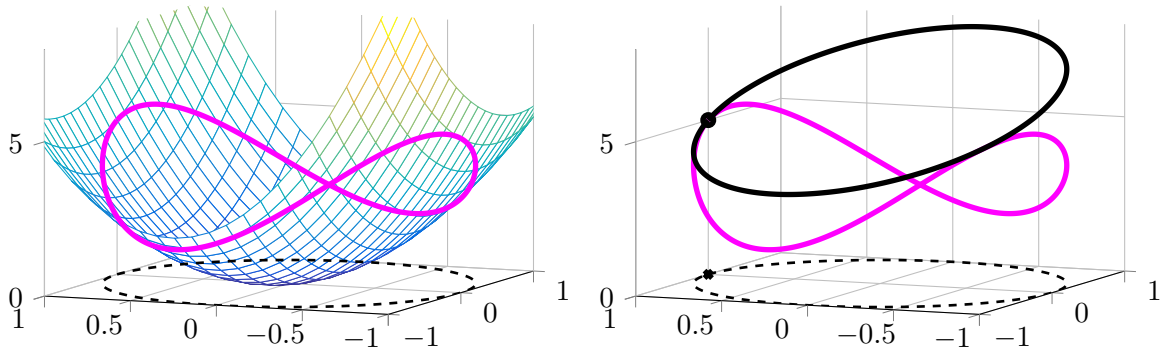


Figure A.2: Quadratic form on $\mathbb{R}^2$ and its restriction to $\mathcal{U}_1^2$ (left). Linear majorization of this quadratic form at a given point over the set $\mathcal{U}_1^2$ (right).

Proposition A.3.1. ***(Majorization of concave QF)*** *The function $-f_{\mathrm{B}}$ as in* (A.14) *admits at point $\mathbf{U}_R^t$ a linear majorizing surrogate in the form of* (A.7)*, with*

$$\mathbf{R}(\mathbf{U}^t) = \mathbf{M}\mathbf{U}^t\mathbf{D}. \tag{A.15}$$

Equality holds at $\mathbf{U}^t$.

Proof. The function $-f_{\mathrm{B}}$ as in (A.14) is concave, so it can be majorized at point $\mathbf{U}^t$ by its first order Taylor expansion (cf. [Sun et al., 2016] section III.A), i.e.,

$$-f_{\mathrm{B}}(\mathbf{U}) \leq -\mathrm{Tr}\left\{\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)^H \mathbf{U}\right\} - \mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)\right\} + \text{const.}. \tag{A.16}$$

□

Proposition A.3.2. ***(Majorization of convex QF)*** *The function f_{B} in* (A.14) *admits on* $\mathrm{St}(p,k)$ *and at point $\mathbf{U}^t$ a linear majorizing surrogate in the form of* (A.7)*, with*

$$\mathbf{R}(\mathbf{U}^t) = -\mathbf{K}\mathbf{U}^t\mathbf{D}, \tag{A.17}$$

where $\mathbf{K} = \mathbf{M} - \lambda_{\mathbf{M}}^{\max}\mathbf{I}$ and $\lambda_{\mathbf{M}}^{\max}$ is the largest eigenvalue of $\mathbf{M}$. Equality holds at $\mathbf{U}^t$.

Proof. The function f_{B} in (A.14) can be expressed as

$$\mathrm{Tr}\left\{\mathbf{U}^H\mathbf{M}\mathbf{U}\mathbf{D}\right\} = \mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{M} - \lambda_{\mathbf{M}}^{\max}\mathbf{I}\right)\mathbf{U}\mathbf{D}\right\} + \mathrm{Tr}\left\{\mathbf{U}^H\left(\lambda_{\mathbf{M}}^{\max}\mathbf{I}\right)\mathbf{U}\mathbf{D}\right\}, \tag{A.18}$$

where the second term is constant and equal to $\lambda_{\mathbf{M}}^{\max}\mathrm{Tr}\{\mathbf{D}\}$ for the restriction $\mathbf{U} \in \mathrm{St}(p,k)$. The first term of this expression is concave in $\mathbf{U}$ ($\mathbf{U} \in \mathbb{C}^{M\times R}$) so it can be upper-bounded by its first order Taylor expansion, thus

$$f_{\mathrm{B}}(\mathbf{U}) \leq +\mathrm{Tr}\left\{\left(\mathbf{K}\mathbf{U}^t\mathbf{D}\right)^H \mathbf{U}\right\} + \mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{K}\mathbf{U}^t\mathbf{D}\right)\right\} + \text{const.}, \tag{A.19}$$

with $\mathbf{K}$ defined as in Proposition A.3.2. □

Remark A.3.1. *Majorizing a convex QF of $\mathbf{U}$ by a linear one seems counter-intuitive since it is not possible on the entire Euclidean space $\mathbb{C}^{N\times R}$. Nevertheless, the restriction to the set* $\mathrm{St}(p,k)$ *makes the upperbound in Propositon A.3.2 possible. In order to give some insight, a visual example on $\mathbb{R}^2$ is presented in Figure A.2.*

A.3.2 Concave compositions of quadratic forms

Compositions involving inner QFs that yield concave functions are often used in order to build robust loss functions (examples are given in section 5.2.2). The following proposition gives a linear majorizer of concave functions composed from the Brockett function.

Proposition A.3.3. ***(Majorization of concave function composed from concave QF)*** *Let $\rho : \mathbb{R} \to \mathbb{R}$ be a concave non-decreasing function. For f_{B} as in* (A.14)*, the function $\rho(-f_{\mathrm{B}})$ admits at point $\mathbf{U}^t$ a linear majorizing surrogate in the form of* (A.7)*, with*

$$\mathbf{R}(\mathbf{U}^t) = \rho'\left(-f_{\mathrm{B}}\left(\mathbf{U}^t\right)\right)\mathbf{M}\mathbf{U}^t\mathbf{D}. \tag{A.20}$$

Equality holds at $\mathbf{U}^t$.

Proof. The function $-f_{\mathrm{B}}$ is concave in $\mathbf{U}$ and ρ is concave non-decreasing. It follows that the function $\rho(-f_{\mathrm{B}})$ is concave, so it can be upper-bounded at point $\mathbf{U}^t$ by its first order Taylor expansion, i.e.,

$$\rho\left(-f_{\mathrm{B}}\left(\mathbf{U}\right)\right) \leq -\rho'\left(-f_{\mathrm{B}}\left(\mathbf{U}^t\right)\right)\mathrm{Tr}\left\{\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)^H\mathbf{U}\right\} - \rho'\left(-f_{\mathrm{B}}\left(\mathbf{U}^t\right)\right)\mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)\right\} + \text{const.} \tag{A.21}$$

□

Following this proof, other linear surrogates can be derived using compositions of concave (non-decreasing/non-increasing) functions and the chain rule. It is also worth noting that one can apply the reformulation of Proposition A.3.2 to express a quadratic QF as a concave term plus a constant in order to do so (the obtained majoration is then only valid on $\mathrm{St}(p,k)$).

A.3.3 Quotients of quadratic forms

Various formulations of quotients of quadratic forms arise in generalized versions of PCA [Kiers, 1995]. Most of them can be obtained as linear combinations of functions of the form

$$f_{\mathrm{q}}(\mathbf{U}) = -\mathrm{Tr}\left\{\left(\mathbf{U}^H\mathbf{C}\mathbf{U}\right)^{-1}\mathbf{U}^H\mathbf{A}\mathbf{U}\right\}, \tag{A.22}$$

where $\mathbf{C}$ is positive definite and $\mathbf{A}$ is positive semi-definite.

Proposition A.3.4. ***(Majorization of quotient of QFs)*** *The function f_{q} as in* (A.22) *admits on $\mathrm{St}(p,k)$ and at point $\mathbf{U}^t$ a linear majorizing surrogate in the form of* (A.7)*, with*

$$\mathbf{R}(\mathbf{U}^t) = \mathbf{T}(\mathbf{U}^t) - \left(\mathbf{K}\mathbf{U}^t\widetilde{\mathbf{T}}(\mathbf{U}^t)\right), \tag{A.23}$$

in which

$$\mathbf{T}(\mathbf{U}^t) = \mathbf{A}\mathbf{U}^t\left((\mathbf{U}^t)^H\mathbf{C}\mathbf{U}^t\right)^{-1} \quad \text{and} \quad \widetilde{\mathbf{T}}(\mathbf{U}^t) = \left(\mathbf{A}^{-1/2}\mathbf{T}(\mathbf{U}^t)\right)^H\left(\mathbf{A}^{-1/2}\mathbf{T}(\mathbf{U}^t)\right), \tag{A.24}$$

with $\mathbf{K} = \mathbf{C} - \lambda_{\mathbf{C}}^{\max}\mathbf{I}$, and where $\lambda_{\mathbf{C}}^{\max}$ is the largest eigenvalue of $\mathbf{C}$. Equality holds at $\mathbf{U}^t$.

Proof. Starting from the inequality

$$\left\|\left(\mathbf{U}^H\mathbf{C}\mathbf{U}\right)^{-1/2}\mathbf{U}^H\mathbf{A}^{1/2} - \left(\mathbf{U}^H\mathbf{C}\mathbf{U}\right)^{1/2}\left((\mathbf{U}^t)^H\mathbf{C}\mathbf{U}^t\right)^{-1}(\mathbf{U}^t)^H\mathbf{A}^{1/2}\right\|^2 \geq 0, \tag{A.25}$$

we obtain

$$f_{\mathrm{q}}(\mathbf{U}) \leq -2\mathfrak{Re}\left\{\left(\mathbf{T}(\mathbf{U}^t)\right)^H\mathbf{U}\right\} + \mathrm{Tr}\left\{\mathbf{U}^H\mathbf{C}\mathbf{U}\widetilde{\mathbf{T}}(\mathbf{U}^t)\right\}, \tag{A.26}$$

with $\mathbf{T}(\mathbf{U}^t)$ and $\widetilde{\mathbf{T}}(\mathbf{U}^t)$ as in (A.24), and where equality holds at $\mathbf{U}^t$. Following the proof of Proposition A.3.2, we can majorize on $\mathrm{St}(p,k)$ the quadratic term in (A.26) as

$$\mathrm{Tr}\left\{\mathbf{U}^H\mathbf{C}\mathbf{U}\widetilde{\mathbf{T}}(\mathbf{U}^t)\right\} \leq +\mathrm{Tr}\left\{\left(\mathbf{K}\mathbf{U}^t\widetilde{\mathbf{T}}(\mathbf{U}^t)\right)^H\mathbf{U}\right\} + \mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{K}\mathbf{U}^t\widetilde{\mathbf{T}}(\mathbf{U}^t)\right)\right\} + \text{const.}, \tag{A.27}$$

with $\mathbf{K}$ as in (A.24), and where equality holds again at $\mathbf{U}^t$. Combining the inequalities (A.26) and (A.27) concludes the proof. □

A.3.4 Proxies of element-wise sign function

The ℓ_0-norm of a complex number can be expressed as

$$||x||_0 = \text{sgn}(|x|) \tag{A.28}$$

where $|\cdot|$ stands for the modulus, and sgn is the sign function. This function would serve as an ideal for the formulation of sparsity promoting penalty, however, it is too complex to deal with due to its discontinuity. To alleviate this issue, we follow the approach proposed in [Song et al., 2015], i.e., approximating the absolute sign function by a smooth function denoted l_γ^ϵ, and defined as

$$l_\gamma^\epsilon(x) = \begin{cases} a|x|^2, & \text{if } |x| \leq \epsilon \\ l_\gamma(x) - b, & \text{if } |x| > \epsilon, \end{cases} \tag{A.29}$$

with appropriate constants a and b so that the approximations l_γ^ϵ are continuous and differentiable (cf. [Song et al., 2015]), and where l_γ belongs to the family of functions:

a) ℓ_γ-norm [Gorodnitsky and Rao, 1997, Chartrand and Yin, 2008, Lai et al., 2013]:
$l_\gamma(x) = |x|^\gamma,\ \gamma \in (0, 1]$

b) ℓ_1-norm approximation from [Sriperumbudur et al., 2011, Candès et al., 2008]:
$l_\gamma(x) = \ln(1 + |x|/\gamma)\ln(1 + 1/\gamma),\ \gamma > 0$

c) lower bound of sign function from [Mangasarian, 1996]:
$l_\gamma(x) = 1 - e^{-|x|/\gamma},\ \gamma > 0,$

involving a tuning parameter γ for each case. Thus, this class covers most of standard proxies of the sign function. Notice that this class is still valid for data with complex entries by reading $|\cdot|$ as the modulus function. We have the following proposition [Song et al., 2015, Section III]:

Proposition A.3.5. *The function l_γ^ϵ in* (A.29) *is majorized at point x_t by the following quadratic surrogate:*

$$l_\gamma^\epsilon(x|x_t) \leq \phi(x_t)|x|^2 + \text{const.} \tag{A.30}$$

where the function ϕ depends on l_γ^ϵ as:

a) *ℓ_γ-norm [Gorodnitsky and Rao, 1997, Chartrand and Yin, 2008, Lai et al., 2013]:*
$$\phi(x_t) = \begin{cases} (\gamma/2)\epsilon^{\gamma-2}, & |x_t| \leq \epsilon \\ (\gamma/2)\,|x_t|^{\gamma/2}, & |x_t| > \epsilon \end{cases}$$

b) *ℓ_1-norm approximation [Sriperumbudur et al., 2011, Candès et al., 2008]:*
$$\phi(x_t) = \begin{cases} (\ 2\epsilon(\gamma+\epsilon)\ln(1+1/\gamma)\ )^{-1}, & |x_t| \leq \epsilon \\ (\ 2\ln(1+1/\gamma)\,|x_t|\,(|x_t|+\gamma)\ )^{-1}, & |x_t| > \epsilon \end{cases}$$

c) *lower bound of sign function [Mangasarian, 1996]:*
$$\phi(x_t) = \begin{cases} e^{-\epsilon/\gamma}/2\gamma\epsilon, & |x_t| \leq \epsilon \\ e^{-|x_t|/\gamma}/2\gamma\,|x_t|, & |x_t| > \epsilon \end{cases}$$

Equality is achieved at x_t.

The functions l_γ^ϵ can now serve as basic building blocks to build sparsity promoting penalties for $\mathbf{U} \in \text{St}(p, k)$. A main example following from [Benidis et al., 2016] (weighted sum) is presented in section 5.2.2. The key trick to obtain a linear majorizer is to obtain the following series of inequalities:

- Using Proposition A.3.5, to obtain a convex quadratic surrogate function of $\mathbf{U}$ over $\text{St}(p, k)$.
- Applying Proposition A.3.2 to obtain a linear majorizer on $\text{St}(p, k)$.

Bibliography

[Absil et al., 2009] Absil, P.-A., Mahony, R., and Sepulchre, R., *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009. (Cited on pages 3, 36, 49, and 67.)

[Akaike, 1974] Akaike, H., "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974. (Cited on page 25.)

[Anderson, 2003] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics. Wiley, 2003. (Cited on page 40.)

[Archambeau et al., 2006] Archambeau, C., Delannay, N., and Verleysen, M., "Robust probabilistic projections," In *Proceedings of the 23rd International conference on machine learning*, pages 33–40, 2006. (Cited on page 9.)

[Aubry et al., 2018] Aubry, A., De Maio, A., and Pallotta, L., "A geometric approach to covariance matrix estimation and its applications to radar problems," *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 907–922, 2018. (Cited on page 36.)

[Aue et al., 2009] Aue, A., Hörmann, S., Horváth, L., and Reimherr, M., "Break detection in the covariance structure of multivariate time series models," *Ann. Statist.*, vol. 37, no. 6B, pp. 4046–4087, 2009. (Cited on page 40.)

[Barber, 2015] Barber, J., "A Generalized Likelihood Ratio Test for Coherent Change Detection in Polarimetric SAR," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 9, pp. 1873–1877, 2015. (Cited on page 40.)

[Barrau and Bonnabel, 2013] Barrau, A. and Bonnabel, S., "A note on the intrinsic Cramér-Rao bound," In *Geometric Science of Information*, pages 377–386. Springer, 2013. (Cited on page 31.)

[Basiri et al., 2019] Basiri, S., Ollila, E., Drašković, G., and Pascal, F., "Fusing Eigenvalues," In *Proc. of ICASSP*, pages 4968–4972, 2019. (Cited on page 36.)

[Ben Abdallah et al., 2019a] Ben Abdallah, R., Breloy, A., El Korso, M., and Lautru, D., "Bayesian Robust Signal Subspace Estimation in Non-Gaussian Environment," In *27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019a. (Cited on pages 14 and 15.)

[Ben Abdallah et al., 2020] Ben Abdallah, R., Breloy, A., El Korso, M., and Lautru, D., "Bayesian signal subspace estimation with compound Gaussian sources," *Signal Processing*, vol. 167, pp. 107310, 2020. (Cited on pages 11, 12, 14, 15, and 49.)

[Ben Abdallah et al., 2017] Ben Abdallah, R., Breloy, A., El Korso, M. N., Lautru, D., and Ouslimani, H. H., "Minimum Mean Square Distance Estimation of Subspaces in presence of Gaussian sources with application to STAP detection," *Journal of Physics: Conference Series, IOP Publishing*, vol. 904, no. 1, pp. 012010, 2017. (Cited on page 17.)

[Ben Abdallah et al., 2019b] Ben Abdallah, R., Breloy, A., Taylor, A., El Korso, M., and Lautru, D., "Signal subspace change detection in structured covariance matrices," In *27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019b. (Cited on pages 41 and 45.)

[Ben Abdallah et al., 2019c] Ben Abdallah, R., Mian, A., Breloy, A., El Korso, M. N., and Lautru, D., "Detection Methods Based on Structured Covariance Matrices for Multivariate SAR Images Processing," *IEEE Geoscience and Remote Sensing Letters*, 2019c. (Cited on pages 40, 41, and 42.)

[Benidis et al., 2016] Benidis, K., Sun, Y., Babu, P., and Palomar, D. P., "Orthogonal sparse pca and covariance estimation via procrustes reformulation," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6211–6226, 2016. (Cited on pages 3, 49, 50, 67, and 70.)

[Besson, 2016] Besson, O., "Bounds for a mixture of low-rank compound-Gaussian and white Gaussian noises," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5723–5732, 2016. (Cited on pages 13 and 16.)

[Besson and Abramovich, 2013] Besson, O. and Abramovich, Y. I., "On the Fisher information matrix for multivariate elliptically contoured distributions," *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 1130–1133, 2013. (Cited on pages 27 and 33.)

[Besson et al., 2011] Besson, O., Dobigeon, N., and Tourneret, J.-Y., "Minimum mean square distance estimation of a subspace," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5709–5720, 2011. (Cited on pages 3, 10, 11, 12, 15, and 19.)

[Besson et al., 2012] Besson, O., Dobigeon, N., and Tourneret, J.-Y., "CS decomposition based Bayesian subspace estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4210–4218, 2012. (Cited on page 10.)

[Bhatia, 2009] Bhatia, R., *Positive definite matrices*, Princeton university press, 2009. (Cited on page 33.)

[Bian et al., 2018] Bian, X., Panahi, A., and Krim, H., "Bi-sparsity pursuit: A paradigm for robust subspace recovery," *Signal Processing*, vol. 152, pp. 148–159, 2018. (Cited on pages 55 and 57.)

[Bishop, 1999] Bishop, C. M., "Bayesian pca," In *Advances in neural information processing systems*, pages 382–388, 1999. (Cited on page 13.)

[Boente, 1987] Boente, G., "Asymptotic theory for robust principal components," *Journal of Multivariate Analysis*, vol. 21, no. 1, pp. 67 – 78, 1987. (Cited on page 22.)

[Bonnabel and Barrau, 2015] Bonnabel, S. and Barrau, A., "An Intrinsic Cramér-Rao Bound on Lie Groups," In Nielsen, F. and Barbaresco, F., editors, *Geometric Science of Information*, pages 664–672. Springer International Publishing, 2015. (Cited on page 31.)

[Bonnabel and Sepulchre, 2009] Bonnabel, S. and Sepulchre, R., "Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1055–1070, 2009. (Cited on pages 34 and 35.)

[Bouchard et al., 2020] Bouchard, F., Breloy, A., Ginolhac, G., Renaux, A., and Pascal, F., "A Riemannian Framework for Low-Rank Structured Elliptical Models," *arXiv preprint arXiv:2001.01141*, 2020. (Cited on pages iv, 34, 35, and 36.)

[Bouchard et al., 2020] Bouchard, F., Breloy, A., Renaux, A., and Ginolhac, G., "Riemannian Geometry and Cramér-rao Bound for Blind Separation of Gaussian Sources," In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4717–4721, 2020. (Cited on page 36.)

[Boumal, 2013] Boumal, N., "On intrinsic Cramér-Rao bounds for Riemannian submanifolds and quotient manifolds," *IEEE transactions on signal processing*, vol. 61, no. 7, pp. 1809–1821, 2013. (Cited on page 31.)

[Boumal, 2014] Boumal, N., *Optimization and estimation on manifolds*, PhD thesis, Université catholique de Louvain, 2014. (Cited on pages 31 and 32.)

[Bouveyron and Brunet-Saumard, 2014] Bouveyron, C. and Brunet-Saumard, C., "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014. (Cited on page 22.)

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011. (Cited on page 57.)

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L., *Convex optimization*, Cambridge university press, 2004. (Cited on page 57.)

[Breloy et al., 2018] Breloy, A., El Korso, M. N., Panahi, A., and Krim, H., "Robust Subspace Clustering for Radar Detection," In *26th European Signal Processing Conference (EUSIPCO 2018)*. IEEE, 2018. (Cited on page 57.)

[Breloy et al., 2015] Breloy, A., Ginolhac, G., Pascal, F., and Forster, P., "Clutter subspace estimation in low rank heterogeneous noise context," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2173–2182, 2015. (Cited on page 16.)

[Breloy et al., 2016a] Breloy, A., Ginolhac, G., Pascal, F., and Forster, P., "Robust covariance matrix estimation in heterogeneous low rank context," *IEEE Transactions on Signal Processing*, vol. 64, no. 22, pp. 5794–5806, 2016a. (Cited on page 16.)

[Breloy et al., 2019a] Breloy, A., Ginolhac, G., Renaux, A., and Bouchard, F., "Intrinsic Cramér–Rao Bounds for Scatter and Shape Matrices Estimation in CES Distributions," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 262–266, 2019a. (Cited on page 33.)

[Breloy et al., 2020] Breloy, A., Kumar, S., Sun, Y., and Palomar, D., "Majorization-Minimization on the Stiefel Manifold with Application to Robust Sparse PCA," *(submitted)*, 2020. (Cited on page 51.)

[Breloy et al., 2013] Breloy, A., Le Magoarou, L., Ginolhac, G., Pascal, F., and Forster, P., "Maximum likelihood estimation of clutter subspace in non homogeneous noise context," In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5. IEEE, 2013. (Cited on page 13.)

[Breloy et al., 2019b] Breloy, A., Ollila, E., and Pascal, F., "Spectral Shrinkage of Tyler's M-Estimator of Covariance Matrix," In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 535–538, 2019b. (Cited on page 36.)

[Breloy et al., 2016b] Breloy, A., Sun, Y., Babu, P., Ginolhac, G., and Palomar, D. P., "Robust rank constrained kronecker covariance matrix estimation," In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 810–814. IEEE, 2016b. (Cited on page 27.)

[Breloy et al., 2016] Breloy, A., Sun, Y., Babu, P., Ginolhac, G., and Palomar, D. P., "Robust rank constrained kronecker covariance matrix estimation," In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 810–814, 2016. (Cited on page 49.)

[Breloy et al., 2016] Breloy, A., Sun, Y., Babu, P., Ginolhac, G., Palomar, D. P., and Pascal, F., "A robust signal subspace estimator," In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–4. IEEE, 2016. (Cited on pages 13 and 14.)

[Brennan and Staudaher, 1992] Brennan, L. and Staudaher, F., "Subclutter visibility demonstration," Technical report, Tech. Rep., RL-TR-92-21, Adaptive Sensors Incorporated, 1992. (Cited on pages 17, 54, and 55.)

[Bunea et al., 2012] Bunea, F., She, Y., Wegkamp, M. H., et al., "Joint variable and rank selection for parsimonious estimation of high-dimensional matrices," *The Annals of Statistics*, vol. 40, no. 5, pp. 2359–2388, 2012. (Cited on page 49.)

[Candès et al., 2011] Candès, E. J., Li, X., Ma, Y., and Wright, J., "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011. (Cited on pages 2 and 57.)

[Candès et al., 2008] Candès, E. J., Wakin, M. B., and Boyd, S. P., "Enhancing Sparsity by Reweighted $\ell_1 1$ Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008. (Cited on pages 50 and 70.)

[Carotenuto et al., 2015] Carotenuto, V., Maio, A. D., Clemente, C., and Soraghan, J., "Unstructured Versus Structured GLRT for Multipolarization SAR Change Detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 8, pp. 1665–1669, 2015. (Cited on page 40.)

[Chachlakis et al., 2019] Chachlakis, D. G., Prater-Bennette, A., and Markopoulos, P. P., "L1-norm Tucker Tensor Decomposition," *IEEE Access*, vol. 7, pp. 178454–178465, 2019. (Cited on page 58.)

[Chandrasekaran et al., 2011] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S., "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011. (Cited on pages 2 and 57.)

[Chartrand and Yin, 2008] Chartrand, R. and Yin, W., "Iteratively reweighted algorithms for compressive sensing," In *IEEE Intl. Conf. on Acoustics, speech and signal processing (ICASSP) 2008*, pages 3869–3872. IEEE, 2008. (Cited on pages 50 and 70.)

[Chen et al., 2012] Chen, K., Chan, K.-S., and Stenseth, N. C., "Reduced rank stochastic regression with a sparse singular value decomposition," *J. of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 203–221, 2012. (Cited on page 49.)

[Chen and Huang, 2012] Chen, L. and Huang, J. Z., "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *J. of the American Statistical Association*, vol. 107, no. 500, pp. 1533–1545, 2012. (Cited on page 49.)

[Chen et al., 2018] Chen, S., Ma, S., So, A. M.-C., and Zhang, T., "Proximal gradient method for manifold optimization," *arXiv preprint arXiv:1811.00980*, vol. 5, no. 6, pp. 8, 2018. (Cited on pages 49 and 61.)

[Chen et al., 2009] Chen, T., Martin, E., and Montague, G., "Robust probabilistic PCA with missing data and contribution analysis for outlier detection," *Computational Statistics & Data Analysis*, vol. 53, no. 10, pp. 3706–3716, 2009. (Cited on page 9.)

[Chikuse, 2003] Chikuse, Y., *Statistics on Special Manifold*, New York: Springer-Verlag, 2003. (Cited on page 11.)

[Ciuonzo et al., 2017] Ciuonzo, D., Carotenuto, V., and Maio, A. D., "On Multiple Covariance Equality Testing with Application to SAR Change Detection," *IEEE Transactions on Signal Processing*, vol. 65, no. 19, pp. 5078–5091, 2017. (Cited on pages 22, 40, and 42.)

[Conradsen et al., 2003] Conradsen, K., Nielsen, A. A., Schou, J., and Skriver, H., "A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 1, pp. 4–19, 2003. (Cited on pages 40 and 42.)

[Couillet et al., 2015] Couillet, R., Pascal, F., and Silverstein, J. W., "The Random Matrix Regime of Maronna's M-estimator with elliptically distributed samples," *Journal of Multivariate Analysis*, vol. 139, pp. 56–78, 2015. (Cited on page 22.)

[Croux and Haesbroeck, 2000] Croux, C. and Haesbroeck, G., "Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies," *Biometrika*, vol. 87, pp. 603–618, 2000. (Cited on pages 2, 22, and 23.)

[De La Torre and Black, 2003] De La Torre, F. and Black, M. J., "A framework for robust subspace learning," *Intl. J. of Computer Vision*, vol. 54, no. 1-3, pp. 117–142, 2003. (Cited on pages 2, 49, and 50.)

[Ding et al., 2006] Ding, C., Zhou, D., He, X., and Zha, H., "R1-PCA: rotational invariant L 1-norm principal component analysis for robust subspace factorization," In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288. ACM, 2006. (Cited on pages 49 and 50.)

[Drašković et al., 2019] Drašković, G., Breloy, A., and Pascal, F., "On the Asymptotics of Maronna's Robust PCA," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4964–4975, 2019. (Cited on pages 2, 22, 23, and 24.)

[Drašković et al., 2020] Drašković, G., Breloy, A., and Pascal, F., "On the performance of robust plug-in detectors using M-estimators," *Signal Processing*, vol. 167, pp. 107282, 2020. (Cited on page 26.)

[Drašković and Pascal, 2018] Drašković, G. and Pascal, F., "New Insights Into the Statistical Properties of M-Estimators," *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4253–4263, 2018. (Cited on pages 23 and 24.)

[Edelman et al., 1998] Edelman, A., Arias, T., and Smith, S., "The Geometry of Algorithms with Orthogonality Constraints," *SIAM J. on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998. (Cited on pages 35, 49, 60, and 67.)

[Elhamifar and Vidal, 2009] Elhamifar, E. and Vidal, R., "Sparse subspace clustering," In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797. IEEE, 2009. (Cited on page 3.)

[Elhamifar and Vidal, 2013] Elhamifar, E. and Vidal, R., "Sparse Subspace Clustering: Algorithm, Theory, and Applications," *EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013. (Cited on pages 55, 56, and 57.)

[Elvira et al., 2017a] Elvira, C., Chainais, P., and Dobigeon, N., "Bayesian nonparametric Principal Component Analysis," *arXiv:1709.05667*, 2017a. (Cited on page 10.)

[Elvira et al., 2017b] Elvira, C., Chainais, P., and Dobigeon, N., "Bayesian nonparametric subspace estimation," *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2247–2251, 2017b. (Cited on page 10.)

[Feng et al., 2013] Feng, J., Xu, H., and Yan, S., "Online robust PCA via stochastic optimization," In *Advances in Neural Information Processing Systems*, pages 404–412, 2013. (Cited on page 3.)

[Fletcher et al., 2004] Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S., "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE transactions on medical imaging*, vol. 23, no. 8, pp. 995–1005, 2004. (Cited on page 62.)

[Forster, 2001] Forster, P., "Generalized rectification of cross spectral matrices for arrays of arbitrary geometry," *IEEE Transactions on Signal Processing*, vol. 49, no. 5, pp. 972–978, 2001. (Cited on page 3.)

[Fortunati et al., 2016] Fortunati, S., Gini, F., and Greco, M. S., "The misspecified Cramér-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions," *IEEE Transactions on Signal Processing*, vol. 64, no. 9, pp. 2387–2399, 2016. (Cited on page 28.)

[Fu et al., 2017] Fu, X., Huang, K., Hong, M., Sidiropoulos, N. D., and So, A. M., "Scalable and flexible Max-Var generalized canonical correlation analysis via alternating optimization," In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5855–5859, 2017. (Cited on page 67.)

[Galeano and Peña, 2007] Galeano, P. and Peña, D., "Covariance changes detection in multivariate time series," *Journal of Statistical Planning and Inference*, vol. 137, no. 1, pp. 194 – 211, 2007. (Cited on page 40.)

[Ginolhac and Forster, 2010] Ginolhac, G. and Forster, P., "Performance analysis of a robust low-rank STAP filter in low-rank Gaussian clutter," In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2746–2749, 2010. (Cited on page 26.)

[Ginolhac and Forster, 2016] Ginolhac, G. and Forster, P., "Approximate distribution of the low-rank adaptive normalized matched filter test statistic under the null hypothesis," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 2016–2023, 2016. (Cited on pages 9, 16, and 17.)

[Girolami and Calderhead, 2011] Girolami, M. and Calderhead, B., "Riemann manifold langevin and hamiltonian monte carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011. (Cited on page 60.)

[Golub and Van Loan, 2012] Golub, G. H. and Van Loan, C. F., *Matrix computations*, volume 3, JHU press, 2012. (Cited on page 9.)

[Golub et al., 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.," *Science*, vol. 286, no. 5439, pp. 531–537, 1999. (Cited on page 52.)

[Goodman and Stiles, 2007] Goodman, N. A. and Stiles, J. M., "On Clutter Rank Observed by Arbitrary Arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 178–186, 2007. (Cited on pages 54 and 55.)

[Gorman and Hero, 1990] Gorman, J. D. and Hero, A. O., "Lower bounds for parametric estimation with constraints," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1285–1301, 1990. (Cited on page 3.)

[Gorodnitsky and Rao, 1997] Gorodnitsky, I. F. and Rao, B. D., "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Transactions on signal processing*, vol. 45, no. 3, pp. 600–616, 1997. (Cited on pages 50 and 70.)

[Greco and Gini, 2013] Greco, M. and Gini, F., "Cramér-Rao lower bounds on covariance matrix estimation for complex elliptically symmetric distributions," *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6401–6409, 2013. (Cited on page 33.)

[Greco et al., 2006] Greco, M., Gini, F., and Rangaswamy, M., "Statistical analysis of measured polarimetric clutter data at different range resolutions," *IEE Proceedings-RADAR, Sonar and Navigation*, vol. 153, no. 6, pp. 473–481, 2006. (Cited on pages 9 and 10.)

[Haardt et al., 2014] Haardt, M., Pesavento, M., Roemer, F., and El Korso, M. N., "Subspace methods and exploitation of special array structures," In *Electronic Reference in Signal Processing: Array and Statistical Signal Processing*, chapter 2.5, pages 651–717. Academic Press Library in Signal Processing, Elsevier Ltd., 2014. (Cited on page 9.)

[Hallin and Paindaveine, 2009] Hallin, M. and Paindaveine, D., "Optimal tests for homogeneity of covariance, scale, and shape," *Journal of Multivariate Analysis*, vol. 100, no. 3, pp. 422 – 444, 2009. (Cited on page 40.)

[Harandi et al., 2018] Harandi, M., Salzmann, M., and Hartley, R., "Dimensionality Reduction on SPD Manifolds: The Emergence of Geometry-Aware Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 48–62, 2018. (Cited on page 62.)

[Hendriks, 1991] Hendriks, H., "A Cramér-Rao type lower bound for estimators with values in a manifold," *Journal of Multivariate Analysis*, vol. 38, no. 2, pp. 245–261, 1991. (Cited on page 31.)

[Hoarau et al., 2017] Hoarau, Q., Breloy, A., Ginolhac, G., Atto, A. M., and Nicolas, J. M., "A subspace approach for shrinkage parameter selection in undersampled configuration for Regularised Tyler Estimators," In *Proc. of ICASSP*, pages 3291–3295, 2017. (Cited on page 36.)

[Hoff, 2009] Hoff, P. D., "Simulation of the matrix Bingham-von Mises-Fisher distribution with applications to multivariate and relational data," *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 438–456, 2009. (Cited on page 12.)

[Hong et al., 2018] Hong, D., Balzano, L., and Fessler, J. A., "Asymptotic performance of PCA for high-dimensional heteroscedastic data," *Journal of multivariate analysis*, vol. 167, pp. 435–452, 2018. (Cited on pages 9 and 18.)

[Hotelling, 1933] Hotelling, H., "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology*, vol. 24, no. 6, pp. 417, 1933. (Cited on pages 2 and 8.)

[Hu et al., 2016] Hu, Z., Pan, G., Wang, Y., and Wu, Z., "Sparse principal component analysis via rotation and truncation," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 875–890, 2016. (Cited on pages 3, 49, and 53.)

[Huang and Wei, 2019] Huang, W. and Wei, K., "Riemannian Proximal Gradient Methods," *arXiv preprint arXiv:1909.06065*, 2019. (Cited on pages 49 and 61.)

[Huroyan and Lerman, 2018] Huroyan, V. and Lerman, G., "Distributed robust subspace recovery," *SIAM Journal on Scientific Computing*, vol. 40, no. 5, pp. A3067–A3090, 2018. (Cited on page 3.)

[Hussian et al., 2013] Hussian, M., Chen, D., Cheng, A., Wei, H., and Stanley, D., "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS Journal of photogrammetry and remote sensing*, vol. 80, pp. 91–106, 2013. (Cited on page 38.)

[Jolliffe, 1986] Jolliffe, I. T., *Principal component analysis*, Springer, 1986. (Cited on page 2.)

[Kammoun et al., 2018] Kammoun, A., Couillet, R., Pascal, F., and Alouini, M., "Optimal Design of the Adaptive Normalized Matched Filter Detector Using Regularized Tyler Estimators," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 2, pp. 755–769, 2018. (Cited on page 22.)

[Kang et al., 2014] Kang, B., Monga, V., and Rangaswamy, M., "Rank-Constrained Maximum Likelihood Estimation of Structured Covariance Matrices," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 50, no. 1, pp. 501–515, 2014. (Cited on page 41.)

[Kay, 1993] Kay, S. M., *Fundamentals of statistical signal processing*, Prentice Hall PTR, 1993. (Cited on page 3.)

[Kelly, 1986] Kelly, E. J., "An Adaptive Detection Algorithm," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22, no. 2, pp. 115–127, 1986. (Cited on page 53.)

[Kent et al., 2013] Kent, J. T., Ganeiber, A. M., and Mardia, K. V., "A new method to simulate the Bingham and related distributions in directional data analysis with applications," *arXiv preprint arXiv:1310.8110*, 2013. (Cited on page 12.)

[Kiers, 2002] Kiers, H. A., "Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems," *Computational statistics & data analysis*, vol. 41, no. 1, pp. 157–170, 2002. (Cited on page 67.)

[Kiers, 1995] Kiers, H. A. L., "Maximization of sums of quotients of quadratic forms and some generalizations," *Psychometrika*, vol. 60, no. 2, pp. 221–245, 1995. (Cited on pages 67 and 69.)

[Kollo and Neudecker, 1993] Kollo, T. and Neudecker, H., "Asymptotics of Eigenvalues and Unit-Length Eigenvectors of Sample Variance and Correlation Matrices," *Journal of Multivariate Analysis*, vol. 47, no. 2, pp. 283 – 300, 1993. (Cited on pages 22 and 23.)

[Kraut et al., 2001] Kraut, S., Scharf, L., and McWhorter, L., "Adaptive subspace detectors," *IEEE Trans. on Sig. Proc.*, vol. 49, no. 1, pp. 1–16, 2001. (Cited on page 53.)

[Kraut and Scharf, 1999] Kraut, S. and Scharf, L. L., "The CFAR adaptive subspace detector is a scale-invariant GLRT," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2538–2541, 1999. (Cited on pages 53 and 54.)

[Kraut et al., 2005] Kraut, S., Scharf, L. L., and Butler, R. W., "The adaptive coherence estimator: a uniformly most-powerful-invariant adaptive detection statistic," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 427–438, 2005. (Cited on pages 53 and 54.)

[Krim et al., 1992] Krim, H., Forster, P., and Proakis, J. G., "Operator approach to performance analysis of root-MUSIC and root-min-norm," *IEEE Transactions on Signal Processing*, vol. 40, no. 7, pp. 1687–1696, 1992. (Cited on page 3.)

[Kumar et al., 2020a] Kumar, S., Ying, J., de M. Cardoso, J. V., and Palomar, D. P., "A Unified Framework for Structured Graph Learning via Spectral Constraints," *Journal of Machine Learning Research*, vol. 21, no. 22, pp. 1–60, 2020a. (Cited on page 61.)

[Kumar et al., 2020b] Kumar, S., Ying, J., de M. Cardoso, J. V., and Palomar, D. P., "A Unified Framework for Structured Graph Learning via Spectral Constraints," *Journal of Machine Learning Research*, vol. 21, no. 22, pp. 1–60, 2020b. (Cited on page 67.)

[Lai et al., 2013] Lai, M., Xu, Y., and Yin, W., "Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization," *SIAM J. on Numerical Analysis*, vol. 51, no. 2, pp. 927–957, 2013. (Cited on pages 50 and 70.)

[Ledoit and Wolf, 2004] Ledoit, O. and Wolf, M., "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365 – 411, 2004. (Cited on page 36.)

[Lerman and Maunu, 2017] Lerman, G. and Maunu, T., "Fast, robust and non-convex subspace recovery," *Information and Inference: A Journal of the IMA*, vol. 7, no. 2, pp. 277–336, 2017. (Cited on pages 2, 13, 19, 49, 50, and 67.)

[Lerman and Maunu, 2018] Lerman, G. and Maunu, T., "An Overview of Robust Subspace Recovery," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1380–1410, 2018. (Cited on pages 2 and 49.)

[Lerman et al., 2015] Lerman, G., McCoy, M. B., Tropp, J. A., and Zhang, T., "Robust computation of linear models by convex relaxation," *Foundations of Computational Mathematics*, vol. 15, no. 2, pp. 363–410, 2015. (Cited on pages 13 and 51.)

[Lin et al., 2011] Lin, Z., Liu, R., and Su, Z., "Linearized alternating direction method with adaptive penalty for low-rank representation," In *Advances in neural information processing systems*, pages 612–620, 2011. (Cited on page 57.)

[Little and Rubin, 2019] Little, R. J. and Rubin, D. B., *Statistical analysis with missing data*, volume 793, John Wiley & Sons, 2019. (Cited on pages 9 and 60.)

[Lu et al., 2016] Lu, C., Feng, J., Chen, Y., Liu, W., Lin, Z., and Yan, S., "Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5249–5257, 2016. (Cited on page 3.)

[Lu and Zhang, 2012] Lu, Z. and Zhang, Y., "An augmented Lagrangian approach for sparse principal component analysis," *Mathematical Programming*, vol. 135, no. 1-2, pp. 149–193, 2012. (Cited on page 53.)

[Mahot et al., 2013] Mahot, M., Pascal, F., Forster, P., and Ovarlez, J.-P., "Asymptotic Properties of Robust Complex Covariance Matrix Estimates," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3348–3356, 2013. (Cited on pages 22 and 23.)

[Maio et al., 2017] Maio, A. D., Orlando, D., Pallotta, L., and Clemente, C., "A Multifamily GLRT for Oil Spill Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 63–79, 2017. (Cited on page 40.)

[Mangasarian, 1996] Mangasarian, O., "Machine learning via polyhedral concave minimization," In *Applied Mathematics and parallel computing*, pages 175–188. Springer, 1996. (Cited on pages 50 and 70.)

[Manton, 2002] Manton, J., "Optimization algorithms exploiting unitary constraints," *IEEE Trans. on Signal Process.*, vol. 50, no. 3, pp. 635–650, 2002. (Cited on pages 49, 65, and 67.)

[Mardia and Jupp, 2009] Mardia, K. V. and Jupp, P. E., *Directional statistics*, volume 494, John Wiley & Sons, 2009. (Cited on pages 11 and 13.)

[Maronna, 2005] Maronna, R., "Principal Components and Orthogonal Regression Based on Robust Scales," *Technometrics*, vol. 47, no. 3, pp. 264–273, 2005. (Cited on page 49.)

[Maronna, 1976] Maronna, R. A., "Robust M-Estimators of Multivariate Location and Scatter," *Annals of Statistics*, vol. 4, no. 1, pp. 51–67, 1976. (Cited on pages 21 and 22.)

[Massart and Absil, 2018] Massart, E. and Absil, P.-A., "Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices," *Technical Report UCL-INMA-2018.06*, 2018. (Cited on page 34.)

[Mennad et al., 2018] Mennad, A., Fortunati, S., El Korso, M. N., Younsi, A., Zoubir, A. M., and Renaux, A., "Slepian-Bangs-type formulas and the related misspecified Cramér-Rao bounds for complex elliptically symmetric distributions," *Signal Processing*, vol. 142, pp. 320–329, 2018. (Cited on page 28.)

[Meriaux et al., 2017] Meriaux, B., Ren, C., El Korso, M. N., Breloy, A., and Forster, P., "Robust-COMET for covariance estimation in convex structures: algorithm and statistical properties," In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE, 2017. (Cited on page 29.)

[Meriaux et al., 2019] Meriaux, B., Ren, C., El Korso, M. N., Breloy, A., and Forster, P., "Robust Estimation of Structured Scatter Matrices in (Mis) matched Models," *Signal Processing*, vol. 165, pp. 163–174, 2019. (Cited on pages 27, 28, and 29.)

[Mestre and Lagunas, 2008] Mestre, X. and Lagunas, M. A., "Modified subspace algorithms for DoA estimation with large arrays," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 598–614, 2008. (Cited on page 17.)

[Meyer et al., 2011] Meyer, G., Bonnabel, S., and Sepulchre, R., "Regression on fixed-rank positive semidefinite matrices: a Riemannian approach," *Journal of Machine Learning Research*, vol. 12, pp. 593–625, 2011. (Cited on page 34.)

[Mian et al., 2019] Mian, A., Breloy, A., Ginolhac, G., and Ovarlez, J.-P., "Robust Low-Rank Change Detection for SAR Image Time Series," In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 10079–10082, 2019. (Cited on pages 40 and 42.)

[Mian et al., 2019] Mian, A., Ginolhac, G., Ovarlez, J.-P., and Atto, A. M., "New Robust Statistics for Change Detection in Time Series of Multivariate SAR Images," *IEEE Transactions on Signal Processing*, vol. 67, no. 2, pp. 520–534, 2019. (Cited on pages 40 and 42.)

[Mian et al., 2019] Mian, A., Ovarlez, J., Atto, A. M., and Ginolhac, G., "Design of New Wavelet Packets Adapted to High-Resolution SAR Images With an Application to Target Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3919–3932, 2019. (Cited on pages 38 and 43.)

[Micchelli and Noakes, 2005] Micchelli, C. A. and Noakes, L., "Rao distances," *Journal of Multivariate Analysis*, vol. 92, no. 1, pp. 97–115, 2005. (Cited on page 33.)

[Mitchell, 1989] Mitchell, A. E., "The information matrix, skewness tensor and a-connections for the general multivariate elliptic distribution," *Annals of the Institute of Statistical Mathematics*, vol. 41, no. 2, pp. 289–304, 1989. (Cited on page 33.)

[Mériaux et al., 2019] Mériaux, B., Breloy, A., Ren, C., El Korso, M. N., and Forster, P., "Modified Sparse Subspace Clustering for Radar Detection in Non-stationary Clutter," In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 669–673, 2019. (Cited on page 57.)

[Muirhead, 1982] Muirhead, R. J., "Aspects of multivariate statistical analysis.," *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1982, 656*, 1982. (Cited on pages 22 and 24.)

[Nagao, 1973] Nagao, H., "On Some Test Criteria for Covariance Matrix," *Ann. Statist.*, vol. 1, no. 4, pp. 700–709, 1973. (Cited on page 40.)

[Nascimento et al., 2018] Nascimento, A. D. C., Frery, A. C., and Cintra, R. J., "Detecting Changes in Fully Polarimetric SAR Imagery with Statistical Information Theory," *arXiv:1801.08901*, 2018. (Cited on page 43.)

[Nascimento et al., 2019] Nascimento, A. D. C., Frery, A. C., and Cintra, R. J., "Detecting Changes in Fully Polarimetric SAR Imagery With Statistical Information Theory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1380–1392, 2019. (Cited on page 40.)

[Novak, 2005] Novak, L. M., "Coherent Change Detection for Multi-Polarization SAR," In *Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.*, pages 568–573, 2005. (Cited on page 40.)

[Ollila and Tyler, 2014a] Ollila, E. and Tyler, D., "Regularized M -Estimators of Scatter Matrix," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 6059–6070, 2014a. (Cited on page 36.)

[Ollila and Tyler, 2014b] Ollila, E. and Tyler, D. E., "Regularized M-estimators of scatter matrix," *Signal Processing, IEEE Transactions on*, vol. 62, no. 22, pp. 6059–6070, 2014b. (Cited on page 22.)

[Ollila et al., 2012a] Ollila, E., Tyler, D. E., Koivunen, V., and Poor, H. V., "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5597–5625, 2012a. (Cited on pages 10, 21, 22, 27, and 55.)

[Ollila et al., 2012b] Ollila, E., Tyler, D. E., Koivunen, V., and Vincent, H. P., "Compound Gaussian clutter modeling with an inverse Gaussian texture distribution," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 876–879, 2012b. (Cited on pages 9 and 10.)

[Ottersten et al., 1998] Ottersten, B., Stoica, P., and Roy, R., "Covariance matching estimation techniques for array signal processing applications," *Digital Signal Processing*, vol. 8, no. 3, pp. 185–210, 1998. (Cited on page 27.)

[Ovarlez et al., 2011] Ovarlez, J.-P., Le Chevalier, F., and Bidon, S., "Les données de club STAP, Introduction au STAP," *Traitement de Signal*, vol. 28, 2011. (Cited on page 16.)

[Pailloux et al., 2011] Pailloux, G., Forster, P., Ovarlez, J.-P., and Pascal, F., "Persymmetric adaptive radar detectors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 4, pp. 2376–2390, 2011. (Cited on page 26.)

[Pascal et al., 2013] Pascal, F., Bombrun, L., Tourneret, J.-Y., and Berthoumieu, Y., "Parameter estimation for multivariate generalized Gaussian distributions," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5960–5971, 2013. (Cited on page 21.)

[Pascal et al., 2008a] Pascal, F., Chitour, Y., Ovarlez, J.-P., Forster, P., and Larzabal, P., "Covariance structure maximum-likelihood estimates in Compound-Gaussian noise: existence and algorithm analysis," *Signal Processing, IEEE Transactions on*, vol. 56, no. 1, pp. 34–48, 2008a. (Cited on page 22.)

[Pascal et al., 2014a] Pascal, F., Chitour, Y., and Quek, Y., "Generalized robust shrinkage estimator and its application to STAP detection problem," *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5640–5651, 2014a. (Cited on page 17.)

[Pascal et al., 2014b] Pascal, F., Chitour, Y., and Quek, Y., "Generalized Robust Shrinkage Estimator and Its Application to STAP Detection Problem," *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5640–5651, 2014b. (Cited on page 36.)

[Pascal et al., 2008b] Pascal, F., Forster, P., Ovarlez, J.-P., and Larzabal, P., "Performance analysis of covariance matrix estimates in impulsive noise," *IEEE Transactions on signal processing*, vol. 56, no. 6, pp. 2206–2217, 2008b. (Cited on page 10.)

[Pascal and Renaux, 2010] Pascal, F. and Renaux, A., "Statistical analysis of the covariance matrix MLE in K-distributed clutter," *Signal Processing*, vol. 90, no. 4, pp. 1165–1175, 2010. (Cited on page 33.)

[Pearson, 1901] Pearson, K., "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, vol. 6, no. 2, pp. 559–572, 1901. (Cited on pages 2 and 8.)

[Pennec et al., 2018] Pennec, X. et al., "Barycentric subspace analysis on manifolds," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2711–2746, 2018. (Cited on pages 3 and 62.)

[Portilla et al., 2003] Portilla, J., Strela, V., Wainwright, M. J., and Simoncelli, E. P., "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Transactions Image Processing*, vol. 12, no. 11, 2003. (Cited on page 10.)

[Raghavan, 2012] Raghavan, R. S., "Statistical Interpretation of a Data Adaptive Clutter Subspace Estimation Algorithm," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 2, pp. 1370–1384, 2012. (Cited on pages 13 and 14.)

[Rangaswamy et al., 2004] Rangaswamy, M., Lin, F. C., and Gerlach, K. R., "Robust adaptive signal processing methods for heterogeneous RADAR clutter scenarios," *Signal Processing*, vol. 84, no. 9, pp. 1653–1665, 2004. (Cited on pages 9, 16, and 22.)

[Ratha et al., 2017] Ratha, D., De, S., Celik, T., and Bhattacharya, A., "Change Detection in Polarimetric SAR Images Using a Geodesic Distance Between Scattering Mechanisms," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1066–1070, 2017. (Cited on pages 40 and 43.)

[Razaviyayn et al., 2013] Razaviyayn, M., Hong, M., and Luo, Z., "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization," *SIAM on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013. (Cited on page 67.)

[Reed et al., 1974] Reed, I. S., Mallett, J. D., and Brennan, L. E., "Rapid convergence rate in adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems*, vol. , no. 6, pp. 853–863, 1974. (Cited on page 22.)

[Richmond and Horowitz, 2015] Richmond, C. D. and Horowitz, L. L., "Parameter bounds on estimation accuracy under model misspecification," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2263–2278, 2015. (Cited on page 28.)

[Ruppert, 2011] Ruppert, D., *Statistics and data analysis for financial engineering*, volume 13, Springer, 2011. (Cited on page 9.)

[Schott, 2001] Schott, J. R., "Some tests for the equality of covariance matrices," *Journal of Statistical Planning and Inference*, vol. 94, no. 1, pp. 25 – 36, 2001. (Cited on page 40.)

[Shi and Selesnick, 2007] Shi, F. and Selesnick, I. W., "An elliptically contoured exponential mixture model for wavelet based image denoising," *Applied and Computational Harmonic Analysis*, vol. 23, no. 1, pp. 131–151, 2007. (Cited on page 10.)

[Smith, 2005] Smith, S. T., "Covariance, subspace, and intrinsic Cramér-Rao bounds," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1610–1630, 2005. (Cited on pages 3, 24, 31, 32, 33, 34, 35, and 60.)

[Soloveychik et al., 2016] Soloveychik, I., Trushin, D., and Wiesel, A., "Group Symmetric Robust Covariance Estimation," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 244–257, 2016. (Cited on pages 22 and 26.)

[Soloveychik and Wiesel, 2014] Soloveychik, I. and Wiesel, A., "Tyler's Covariance Matrix Estimator in Elliptical Models With Convex Structure," *IEEE Transactions on Signal Processing*, vol. 62, no. 20, pp. 5251–5259, 2014. (Cited on pages 22, 27, and 29.)

[Soloveychik and Wiesel, 2015a] Soloveychik, I. and Wiesel, A., "Joint covariance estimation with mutual linear structure," *IEEE Transactions on Signal Processing*, vol. 64, no. 6, pp. 1550–1561, 2015a. (Cited on page 62.)

[Soloveychik and Wiesel, 2015b] Soloveychik, I. and Wiesel, A., "Performance Analysis of Tyler's Covariance Estimator," *Signal Processing, IEEE Transactions on*, vol. 63, no. 2, pp. 418–426, 2015b. (Cited on page 22.)

[Song et al., 2015] Song, J., Babu, P., and Palomar, D. P., "Sparse Generalized Eigenvalue Problem Via Smooth Optimization," *IEEE Trans. on Signal Process.*, vol. 63, no. 7, pp. 1627–1642, 2015. (Cited on pages 49, 50, and 70.)

[Sriperumbudur et al., 2011] Sriperumbudur, B. K., Torres, D. A., and Lanckriet, G. R., "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Machine learning*, vol. 85, no. 1, pp. 3–39, 2011. (Cited on pages 50 and 70.)

[Srivastava, 2000] Srivastava, A., "A Bayesian approach to geometric subspace estimation," *IEEE Transaction on Signal Processing*, vol. 48, no. 5, pp. 1390–1400, 2000. (Cited on pages 10 and 11.)

[Stoica and Selen, 2004] Stoica, P. and Selen, Y., "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004. (Cited on page 22.)

[Sun et al., 2014] Sun, Y., Babu, P., and Palomar, D., "Regularized Tyler's Scatter Estimator: Existence, Uniqueness, and Algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5143–5156, 2014. (Cited on page 36.)

[Sun et al., 2016] Sun, Y., Babu, P., and Palomar, D., "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," *IEEE Trans. on Signal Process.*, vol. PP, no. 99, pp. 1–1, 2016. (Cited on pages 3, 64, and 68.)

[Sun et al., 2016] Sun, Y., Babu, P., and Palomar, D. P., "Robust Estimation of Structured Covariance Matrix for Heavy-Tailed Elliptical Distributions," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3576–3590, 2016. (Cited on pages iv, 3, 22, 27, 29, 35, and 42.)

[Sun et al., 2016] Sun, Y., Breloy, A., Babu, P., Palomar, D. P., Pascal, F., and Ginolhac, G., "Low-complexity algorithms for low rank clutter parameters estimation in RADAR systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1986–1998, 2016. (Cited on pages 14 and 49.)

[Terreaux et al., 2018] Terreaux, E., Ovarlez, J.-P., and Pascal, F., "Robust Model Order Selection in Large Dimensional Elliptically Symmetric Noise," *arXiv:1710.06735*, 2018. (Cited on page 22.)

[Thomas et al., 1995] Thomas, J. K., Scharf, L. L., and Tufts, D. W., "The probability of a subspace swap in the SVD," *IEEE Transactions on Signal Processing*, vol. 43, no. 3, pp. 730–736, 1995. (Cited on page 2.)

[Tiomoko et al., 2019] Tiomoko, M., Bouchard, F., Ginholac, G., and Couillet, R., "Random matrix improved covariance estimation for a large class of metrics," *arXiv preprint arXiv:1902.02554*, 2019. (Cited on page 34.)

[Tipping and Bishop, 1999a] Tipping, M. E. and Bishop, C. M., "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999a. (Cited on pages 9, 34, and 41.)

[Tipping and Bishop, 1999b] Tipping, M. E. and Bishop, C. M., "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999b. (Cited on page 9.)

[Tyler, 1981] Tyler, D., "Asymptotic Inference for Eigenvectors," *The Annals of Statistics*, pages 725–736, 1981. (Cited on page 22.)

[Tyler, 1982] Tyler, D. E., "Radial estimates and the test for sphericity," *Biometrika*, vol. 69, no. 2, pp. 429, 1982. (Cited on page 22.)

[Tyler, 1987] Tyler, D. E., "A distribution-free M-estimator of multivariate scatter," *The annals of Statistics*, pages 234–251, 1987. (Cited on pages 10, 21, 22, 29, and 58.)

[Tyler and Yi, 2018] Tyler, D. E. and Yi, M., "Lassoing Eigenvalues," *arXiv preprint arXiv:1805.08300*, 2018. (Cited on page 36.)

[Uematsu et al., 2017] Uematsu, Y., Fan, Y., Chen, K., Lv, J., and Lin, W., "SOFAR: large-scale association network learning," *arXiv preprint arXiv:1704.08349*, 2017. (Cited on pages 49 and 60.)

[Vandereycken et al., 2012] Vandereycken, B., Absil, P.-A., and Vandewalle, S., "A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank," *IMA Journal of Numerical Analysis*, vol. 33, no. 2, pp. 481–514, 2012. (Cited on page 34.)

[Vandereycken and Vandewalle, 2010] Vandereycken, B. and Vandewalle, S., "A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 5, pp. 2553–2579, 2010. (Cited on page 34.)

[Vershynin, 2010] Vershynin, R., "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010. (Cited on page 60.)

[Vidal, 2011] Vidal, R., "Subspace Clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011. (Cited on pages 3 and 55.)

[Ward, 1994] Ward, J., "Space time adaptive processing for airborne RADAR," Technical report, MIT, Lexington, Mass., USA, 1994. (Cited on pages 16 and 17.)

[Wax and Ziskind, 1989] Wax, M. and Ziskind, I., "Detection of the number of coherent signals by the MDL principle," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 8, pp. 1190–1196, 1989. (Cited on page 25.)

[White, 1982] White, H., "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982. (Cited on page 28.)

[Wiesel, 2012a] Wiesel, A., "Geodesic Convexity and Covariance Estimation," *Signal Processing, IEEE Transactions on*, vol. 60, no. 12, pp. 6182–6189, 2012a. (Cited on page 26.)

[Wiesel, 2012b] Wiesel, A., "Unified Framework to Regularized Covariance Estimation in Scaled Gaussian Models," *Signal Processing, IEEE Transactions on*, vol. 60, no. 1, pp. 29–38, 2012b. (Cited on page 36.)

[Wiesel and Zhang, 2015] Wiesel, A. and Zhang, T., "Structured Robust Covariance Estimation," *Founds. and Trends in Signal Process.*, vol. 8, no. 3, pp. 127–216, 2015. (Cited on page 22.)

[Woodbridge et al., 2017] Woodbridge, Y., Elidan, G., and Wiesel, A., "Signal detection in complex structured para normal noise," *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2306–2316, 2017. (Cited on page 61.)

[Xavier and Barroso, 2002] Xavier, J. and Barroso, V., "Intrinsic distance lower bound for unbiased estimators on Riemannian manifolds," In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–1141. IEEE, 2002. (Cited on page 31.)

[Xavier and Barroso, 2005] Xavier, J. and Barroso, V., "Intrinsic Variance Lower Bound (IVLB): an extension of the Cramér-Rao bound to Riemannian manifolds," In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages v–1033. IEEE, 2005. (Cited on page 31.)

[Yao, 1973] Yao, K., "A representation theorem and its applications to spherically invariant random processes," *IEEE Transactions on Information Theory*, vol. 19, no. 5, pp. 600–608, 1973. (Cited on page 10.)

[Zanella et al., 2009] Zanella, A., Chiani, M., and Win, M. Z., "On the marginal distribution of the eigenvalues of Wishart matrices," *IEEE Transactions on Communications*, vol. 57, no. 4, pp. 1050–1060, 2009. (Cited on page 24.)

[Zhang et al., 2014] Zhang, T., Cheng, X., and Singer, A., "Marchenko-Pastur Law for Tyler's and Maronna's M-estimators," *arXiv preprint*, 2014. (Cited on page 22.)

[Zhang et al., 2013] Zhang, T., Wiesel, A., and Greco, M. S., "Multivariate generalized Gaussian distribution: Convexity and graphical models," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4141–4148, 2013. (Cited on page 21.)

[Zhao and Jiang, 2006] Zhao, J. and Jiang, Q., "Probabilistic PCA for t distributions," *Neurocomputing*, vol. 69, no. 16-18, pp. 2217–2226, 2006. (Cited on page 22.)

[Zozor and Vignat, 2010] Zozor, S. and Vignat, C., "Some results on the denoising problem in the elliptically distributed context," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 134–150, 2010. (Cited on page 10.)