Convex Parameter Estimation of Perturbed Multivariate Generalized Gaussian Distributions

Nora Ouzir, Frédéric Pascal, Jean-Christophe Pesquet

Convex Parameter Estimation of Perturbed Multivariate Generalized Gaussian Distributions

Nora Ouzir, *Member, IEEE,* Frédéric Pascal, *Senior member, IEEE,* and Jean-Christophe Pesquet, *IEEE Fellow member*

***Abstract*—The multivariate generalized Gaussian distribution (MGGD), also known as the multivariate exponential power (MEP) distribution, is widely used in signal and image processing. However, estimating MGGD parameters, which is required in practical applications, still faces specific theoretical challenges. In particular, establishing convergence properties for the standard fixed-point approach when both the distribution mean and the scatter (or the precision) matrix are unknown is still an open problem. In robust estimation, imposing classical constraints on the precision matrix, such as sparsity, has been limited by the non-convexity of the resulting cost function. This paper tackles these issues from an optimization viewpoint by proposing a convex formulation with well-established convergence properties. We embed our analysis in a noisy scenario where robustness is induced by modelling multiplicative perturbations. The resulting framework is flexible as it combines a variety of regularizations for the precision matrix, the mean and model perturbations. This paper presents proof of the desired theoretical properties, specifies the conditions preserving these properties for different regularization choices and designs a general proximal primal-dual optimization strategy. The experiments show a more accurate precision and covariance matrix estimation with similar performance for the mean vector parameter compared to Tyler's M-estimator. In a high-dimensional setting, the proposed method outperforms the classical GLASSO, one of its robust extensions, and the regularized Tyler's estimator.**

***Index Terms*—Parameter Estimation, Convex Optimization, Robust Estimation, Precision Matrix, Multivariate Generalized Gaussian Distribution, Multivariate Exponential Power Distributions.**

I. Introduction

PROBABILISTIC distribution parameters, particularly the mean and covariance –or scatter/precision matrix–, are central in statistical signal processing. In machine learning applications, they are building blocks of clustering, classification, dimension reduction, or detection models [1], [2], [3]. These parameters are rarely known in practice, and estimating them has been a long-standing statistical problem. A Maximum Likelihood (ML) approach relying on the standard Gaussian assumption has long been the solution of choice, bringing simplicity and tractability to performance analysis. The statistical properties of the Gaussian ML estimators are also well-known: the estimated mean is Gaussian, while the covariance matrix estimator follows a Wishart distribution.

Nora Ouzir and Jean-Christophe Pesquet are with Université Paris-Saclay, CentraleSupélec, Inria OPIS, Centre de Vision Numérique, 91190, Gif-sur-Yvette, France. Frédéric Pascal is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, 91190, Gif-sur-Yvette, France. The work of Jean-Christophe Pesquet was supported by BRIDGE-ABLE ANR Chair in AI. DGA has partially funded the research of Frédéric Pascal under grant ANR-17-ASTR-0015.

The standard Gaussian assumption is no longer sufficient to handle the complexity of real-world datasets today. With datasets becoming increasingly large, high-dimensional, and heterogeneous, and often containing significant outliers, the limitations of the Gaussian probabilistic model are becoming more apparent. In domains such as radar [4], financial signal [5], and image processing [6], where these limitations are well-known, alternative models are being sought to address these challenges. Robust estimation theory is one of those alternatives that can provably deal with data perturbations [7]. It models observations with elliptically-contoured (EC) (complex and real Elliptically Symmetric (ES)) distributions. This broad-ranging model encompasses well-known multivariate distributions such as the Gaussian, Generalized Gaussian, t-, or k-distributions (see [8] for a review).

Regularizing the parameter estimation problem has been another way of tackling current data challenges, such as high dimensionality. For example, sparse ℓ_1-regularization of the precision matrix is a popular choice for Gaussian distributions [9]. Works in the standard Gaussian framework have also laid the grounds for regularizing robust estimators of EC models. In [10], the scatter matrix has been regularized through the trace of its inverse. At the same time, closely related shrinkage Tyler's estimators have been studied in [11], [12]. Although these works have addressed the problem in the general statistical context of EC models, they remain limited from the regularization viewpoint. For example, unlike existing works in the Gaussian case, they have not considered ℓ_1-regularization of the precision matrix due to the non-convexity of the resulting problem in robust estimation. It is also worth noting that mostly plug-in (as opposed to model-based) robust estimation strategies have been proposed in the Gaussian case with ℓ_1-regularization [13], [14], [15]. In [16], a framework based on geodesic convexity has been proposed to regularize the perturbed covariance matrix of a zero-mean Gaussian distribution. To circumvent non-convexity in the regularized robust case, similar geodesic convexity-based strategies have been explored for MGGDs in the particular context of covariance estimation with a chordal sparsity pattern [17]. Furthermore, previous works have not explored regularizing other parameters, such as the mean or data perturbations. As explained in the following, this work proposes an original model-based approach that copes with the non-convexity of the original problem and provides a general framework for exploring a larger variety of regularizations, along with associated theoretical conditions.

This work tackles the above issues for a popular subclass of EC models: the MGGD [6], also called MEP distribution [18].

This distribution family is instrumental for image processing because of its ability to model a wide range of features (*e.g.*, wavelet coefficients, gradients) [6], [19]. It is also increasingly used to model weight distributions in (deep) neural networks [20], [21]. From a statistical point of view, its much-needed flexibility is obtained thanks to an extra shape parameter β modelling lighter ($\beta > 2$) and heavier than the Gaussian ($\beta < 2$) tails.[1] The multivariate Gaussian distribution is a special case of MGGDs when $\beta = 2$. Furthermore, the parameter estimation problem is addressed in this work in the particular context of perturbed MGGDs. We are specifically interested in jointly estimating the mean vector, covariance matrix, and perturbation parameters, a complex problem introduced in [22][2], with broad applicability in real-world applications. Our main contribution is a convex reformulation of the resulting statistically robust joint parameter estimation problem. The proposed formulation offers all estimators theoretical guarantees of existence and establishes conditions for their uniqueness. In addition to modelling heterogeneity and outliers, the proposed formulation is flexible and can adapt to various real-life data structures. For example, the sparsity of the precision matrix can be incorporated through ℓ_1- regularization to deal with high dimensional data with relatively few observations. Unlike previous approaches, the proposed framework enables regularizing all the involved parameters beyond the precision matrix. We also study the impact of these regularizations and specify the theoretical conditions they must satisfy to preserve convexity. Our second contribution is designing a proximal primal-dual algorithm tailored to solve the parameter estimation problem with guaranteed convergence [23], [24]. The proposed algorithm is flexible, and associated proximity operators can easily incorporate various regularization choices.

The paper is organized as follows. Section II introduces the proposed perturbed MGGD model, starting with the underlying statistical framework. Section III presents the proposed convex parameter estimation approach, which is the main contribution of this paper. The proposed proximal primal-dual algorithm is detailed in Section IV. In Section V, we compare the performance of the proposed approach to different state-of-the-art robust and non-robust estimators in various experimental scenarios. Finally, concluding remarks and perspectives are provided in Section VI.

Notation: $\mathcal{S}_K$ denotes the space of symmetric real matrices of size $K \times K$, $\mathcal{S}_K^+$ is the cone of positive semi-definite matrices, and $\mathcal{S}_K^{++}$ the cone of positive definite matrices. $\mathbf{Id}$ denotes the identity matrix (whatever its size), $\mathbf{1}$ the vector (the dimension of which is understood from the context) whose components are all equal to 1, and $\mathrm{tr}(.)$ the trace of a matrix. $\|\cdot\|$ denotes the Euclidean norm, $\|.\|_{\mathrm{S}}$ is the operator norm, and $\|\cdot\|_r$ with $r \in [1, +\infty[$ denotes the element-wise ℓ^r norm (the same notation will be used for a matrix or for a vector whatever the dimension). $\Gamma_0(\mathcal{H})$ denotes the class of lower-semicontinuous convex functions from some Hilbert space $\mathcal{H}$ to $]-\infty, +\infty]$ which are proper (*i.e.*, finite at least at one point). The domain of f is $\mathrm{dom}\, f = \left\{x \in \mathcal{H} \;\middle|\; f(x) < +\infty\right\}$. $\iota_{\mathcal{D}}$ denotes the indicator function of $\mathcal{D} \subset \mathcal{H}$, which is equal to 0 on this set and $+\infty$ out of it. Finally, a function $f\colon \mathcal{H} \to]-\infty, +\infty]$ is coercive if $\lim_{\|x\|\to+\infty} f(x) = +\infty$. A comprehensive table of notations is provided in Appendix F.

[1]Note that β in this work corresponds to 2β in most MGGD-related literature. This choice allows us to simplify notations.

[2]A preliminary version of this work was presented in [22] with reduced theoretical results and limited experimental validation.

Table I
TABLE OF NOTATION.

Cost function and regularizations	
$\mathbf{Q} \in \mathcal{S}_K^{++}$	reparametrization of covariance matrix $\mathbf{C}$
$\mathbf{m} \in \mathbb{R}^K$	reparametrization of mean $\boldsymbol{\mu}$
$\boldsymbol{\theta} \in \mathbb{R}^N$	reparametrization of multiplicative perturbation $\boldsymbol{\tau}$
$\mathcal{L}$	original neg-likelihood function
$\widetilde{\mathcal{L}}$	reparametrized neg-likelihood function
$g_{\mathbf{Q}}$	regularization function on $\mathbf{Q}$
$g_{\mathbf{m}}$	regularization function on $\mathbf{m}$
g_{ϑ}	regularization function on $\boldsymbol{\theta}$
f	global regularized cost function
$\widetilde{g}_{\vartheta}$	function gathering the terms in $\boldsymbol{\theta}$
$\psi(\xi)$	$\begin{cases} -\log \xi, & \text{if } \xi > 0 \\ +\infty, & \text{otherwise.} \end{cases}$
$\Psi(\mathbf{Q})$	$\begin{cases} -N \log\det \mathbf{Q}, & \text{if } \mathbf{Q} \in \mathcal{S}_K^{++}, \\ +\infty, & \text{otherwise,} \end{cases}$
$\widetilde{g}_{\mathbf{Q}}$	$\widetilde{g}_{\mathbf{Q}} = g_{\mathbf{Q}} + \Psi$ function gathering the terms in $\mathbf{Q}$
$\widetilde{g}_{\vartheta,0}$	$\widetilde{g}_{\vartheta} - \widetilde{g}_{\vartheta,1}$
$\widetilde{g}_{\vartheta,1}$	convex part of $\widetilde{g}_{\vartheta}$ finite on $]0, +\infty[^N$, exploding on the border of this domain
$\overline{\mathbf{m}} \in \mathbb{R}^K$	known mean vector
$\lambda \in]0, +\infty[$	sparsity parameter
$\epsilon \in]0, +\infty[$	elastic net regularization parameter
$\eta \in]0, +\infty[$	first parameter of the regularization g_{ϑ} (scale parameter of a generalized Gamma distribution)
$\kappa + 1$	second parameter of the regularization g_{ϑ}, with $\kappa \in]K(1-1/\beta), +\infty[$ (shape parameter of a generalized Gamma distribution)
$\alpha \in [1, +\infty[$	third parameter of the regularization g_{ϑ} (exponent parameter of a generalized Gamma distribution)
$\overline{\boldsymbol{\tau}}$	true target value of $\boldsymbol{\tau}$
$\overline{\boldsymbol{\theta}}$	true target value of $\boldsymbol{\theta}$
$\overline{f}$	average behaviour (up to a constant factor) of f with respect θ_n
$\overline{\kappa} > 1$	$\overline{\kappa} = \kappa\beta/(K(\beta-1))$
$\overline{f}'(\theta_n), \overline{f}''(\theta_n)$	first and second derivatives of $\overline{f}$ at θ_n
$\widehat{\theta}_n$	minimizer of $\overline{f}$

II. PROBLEM FORMULATION

The MGGD [18] belongs to a broad subclass of EC distributions that can model various uni-modal probability density functions (p.d.f) with heavier or lighter tails than the Gaussian one. A random vector following an EC distribution can be defined using its stochastic representation [25]:

$$\mathbf{x} \overset{d}{=} \boldsymbol{\mu} + \mathcal{R}\, \mathbf{A}\, \mathbf{u}, \tag{1}$$

where $\overset{d}{=}$ stands for "is distributed as". The *modular variate* $\mathcal{R}$ is a positive random variable (with an unknown p.d.f), $\mathbf{u}$ is a random vector uniformly distributed on the unit-hypersphere $\{\mathbf{u} \in \mathbb{R}^K | \, \|\mathbf{u}\| = 1\}$, with $\mathcal{R}$ and $\mathbf{u}$ (statistically) independent. In (1), the mean vector $\boldsymbol{\mu}$ and the scatter matrix factorization $\mathbf{A}$ (and the associated matrix $\mathbf{C} = \mathbf{A}\,\mathbf{A}^\top$) are the unknown parameters of the EC model. The stochastic representation described above is of practical interest for simulating EC-distributed random vectors. In particular, for the MGGD subclass, $\mathcal{R}^\beta \sim \Gamma(1/2, K/\beta)$, where $\Gamma(a, b)$ is the univariate Gamma distribution with parameters a and b

(see [26] for a definition).

The zero-mean MGGD of dimension K, denoted by $\mathrm{MGGD}_K(\beta, \mathbf{0}, \mathbf{C})$, is characterized by its p.d.f

$$\mathsf{p}_{\mathsf{x}}(\cdot) = C_{K,\beta}(\det \mathbf{C})^{-1/2} \exp\left(-\frac{1}{2}\left[(\cdot)^\top \mathbf{C}^{-1}(\cdot)\right]^{\beta/2}\right), \quad (2)$$

where $C_{K,\beta} = \dfrac{K\,\Gamma(K/2)}{\pi^{K/2}\Gamma(1+K/\beta)2^{1+K/\beta}}$ is a constant, $\Gamma(\cdot)$ denotes the Gamma function, and $\mathbf{C} \in \mathcal{S}_K^{++}$ is (up to a multiplicative factor) the associated covariance matrix.[3] With this notation, when $\beta = 2$, (2) corresponds to the standard multivariate Gaussian distribution. When $\beta < 2$, distributions with heavier tails than the Gaussian one are obtained. In the remainder of this paper, we assume that the exponent $\beta > 1$ is known (see Section II-C).

A. Observation Model with Multiplicative Perturbations

This subsection introduces the proposed MGGD model under multiplicative perturbations. Let $(\mathbf{x}_n)_{1\leqslant n\leqslant N}$ be N realizations of independent and identically distributed (*i.i.d.*) random K-dimensional vectors generated according to a zero-mean MGGD with shape parameter β. Let us now consider a scenario where noisy observations $(\mathbf{y}_n)_{1\leqslant n\leqslant N}$ result from a deterministic multiplicative perturbation $\boldsymbol{\tau}$ of the unperturbed $(\mathbf{x}_n)_{1\leqslant n\leqslant N}$. The corresponding observation model reads

$$(\forall n \in \{1, \ldots, N\}) \quad \mathbf{y}_n = \tau_n \mathbf{x}_n + \boldsymbol{\mu}, \quad (3)$$

where $\boldsymbol{\tau} = (\tau_n)_{1\leqslant n\leqslant N} \in]0, +\infty[^N$ and $\beta \in]1, +\infty[$. The $\boldsymbol{\tau}$-perturbed model (3) is equivalent to the assumption that the $\mathbf{y}_n$ samples follow a $\mathrm{MGGD}_K(\beta, \boldsymbol{\mu}, \tau_n^2\,\mathbf{C})$. As (3) accounts for possible outliers in the observations (*i.e.,* τ_n values larger than 1), it can also be seen as a general EC distribution where τ_n's are realizations of an unknown positive p.d.f. Interestingly, similar models have been widely studied in the particular context of perturbed Gaussian distributions (see, *e.g.*, [27]). Starting from (3), we aim to jointly estimate the perturbations $\boldsymbol{\tau}$ and the unknown distribution parameters $\boldsymbol{\mu}$ and $\mathbf{C}$. Thus, N scalar parameters $(\tau_n)_{1\leqslant n\leqslant N}$ need to be estimated in addition to the mean and covariance matrix. This problem has been addressed in [27] for the particular case of a centred Gaussian distribution ($\boldsymbol{\mu} = \mathbf{0}$ and $\beta = 2$.) The following subsection recalls the existing ML-based approaches in the more general case.

B. Previous Work on Estimating $(\mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\tau})$

ML estimation of the unknown parameters $(\mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\tau})$ has been broadly studied for EC models, including the MGGD or compound Gaussian models. (For non-perturbed EC models, one can refer to [7], [28].) For the proposed $\boldsymbol{\tau}$-perturbed MGGD model, estimating $\mathbf{C}$, $\boldsymbol{\mu}$, and $\boldsymbol{\tau}$ can be achieved by minimizing the negative log-likelihood function arising from (3) (up to the normalizing constant that does not depend on the unknown parameters), *i.e.*,

$$\mathcal{L}(\mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \frac{1}{2}\sum_{n=1}^{N} \frac{\left[(\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})\right]^{\beta/2}}{\tau_n^\beta} + \frac{N}{2}\log\det\mathbf{C} + K\sum_{n=1}^{N}\log\tau_n. \quad (4)$$

Function (4) is non-convex, and its minimization has been intensively investigated [29], [6]: the standard approach being to first minimize with respect to $\boldsymbol{\tau}$, then plug the optimal value $\widehat{\boldsymbol{\tau}}(\mathbf{C}, \boldsymbol{\mu})$ into (4). Precisely, the first step of the standard approach yields

$$\widehat{\boldsymbol{\tau}}(\mathbf{C}, \boldsymbol{\mu}) = \left[\left(\frac{\beta[(\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})]^{\beta/2}}{2K}\right)^{\frac{1}{\beta}}\right]_{1\leqslant n\leqslant N}, \quad (5)$$

and by assuming none of the vectors $(\mathbf{y}_n)_{1\leqslant n\leqslant N}$ is equal to $\boldsymbol{\mu}$ (which is true with probability 1 for continuous random vectors for a given $\boldsymbol{\mu}$), plugging (5) into (4) leads to

$$\mathcal{L}(\mathbf{C}, \boldsymbol{\mu}, \widehat{\boldsymbol{\tau}}(\mathbf{C}, \boldsymbol{\mu})) = \frac{K}{2}\sum_{n=1}^{N}\log\left[(\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})\right] + \frac{N}{2}\log\det\mathbf{C}, \quad (6)$$

where the constant term $\frac{KN}{\beta}\left[1+\log\left(\frac{2K}{\beta}\right)\right]$ has been omitted. Note that (6) is (up to constant terms) the log-likelihood function of the (central) angular Gaussian distribution [30]. For a given value of $\boldsymbol{\mu}$, minimizing (6) with respect to $\mathbf{C}$ (or $\mathbf{C}^{-1}$) on $\mathcal{S}_K^{++}$ thus leads to Tyler's estimator [31], defined as the unique (up to a scale factor) solution to the fixed-point equation:

$$\mathbf{C} = \frac{K}{N}\sum_{n=1}^{N}\frac{(\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^\top}{(\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})}. \quad (7)$$

Considering both $\mathbf{C}$ and $\boldsymbol{\mu}$ unknown leads to the joint fixed-point equations [32]:

$$\begin{cases} \boldsymbol{\mu} = & \dfrac{\displaystyle\sum_{n=1}^{N}\frac{\mathbf{y}_n}{(\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})}}{\displaystyle\sum_{i=1}^{N}\frac{1}{(\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})}}, \\ \mathbf{C} = & \dfrac{K}{N}\displaystyle\sum_{n=1}^{N}\frac{(\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^\top}{(\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})}, \end{cases} \quad (8)$$

or equivalently,

$$\begin{cases} \boldsymbol{\mu} = & \displaystyle\sum_{n=1}^{N} u(s_n)\,\mathbf{y}_n \Big/ \sum_{i=1}^{N} u(s_n), \\ \mathbf{C} = & \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} u(s_n)\,(\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^\top, \end{cases} \quad (9)$$

where $u(s) = K/s$ and $s_n = (\mathbf{y}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})$ Tyler's estimator has been extensively studied from a computational and statistical perspective ([31], [27], [33]). Note that when

[3] The covariance matrix of the MGGD is equal to $2^{2/\beta}\Gamma((K+2)/\beta)(K\,\Gamma(K/\beta))^{-1}\,\mathbf{C}$. For large values of K, the following approximation can be used (for computational purposes): $\Gamma(a\,K + b) \sim \sqrt{2\,\pi}\exp(-a\,K)(a\,K)^{a\,K+b-1/2}$, leading to $2^{2/\beta}\Gamma((K+2)/\beta)(K\,\Gamma(K/\beta))^{-1}\,\mathbf{C} \sim 1/K\,(2\,K/\beta)^{2/\beta}\,\mathbf{C}$.

one redefines Tyler's estimator as a Huber-type estimator [31], an exponent $1/2$ is added to the denominator in (8). The following remarks are of importance.

Remark II.1 On the joint fixed-point algorithm:

(i) First, the existence, uniqueness, and convergence of the recursive algorithm associated with (8) have been studied in [31], [34] for a fixed mean. Although results have been shown for redescending M-estimators [35], establishing convergence properties when both $\boldsymbol{\mu}$ and $\mathbf{C}$ are unknown is still an open issue for the fixed-point algorithm associated with (8).[4] While combining the results of [35] and [31] could serve as a good starting point to analyse the existence and convergence of (8), this paper proposes a general framework where the mean and covariance are jointly estimated and provides proof of the properties above within this context. If existence/uniqueness is assumed, then the asymptotic joint distribution of the corresponding estimators $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{C}})$ is entirely characterized [31].

(ii) The resulting estimators $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{C}})$ **do not depend on** β. It can be shown (see [1]) that this is even true for any $\boldsymbol{\tau}$-perturbed EC distribution expressed by (3), where *extra*-parameters, such as shape, degree of freedom, or scale, only impact the estimation of $\boldsymbol{\tau}$.

This work introduces an alternative formulation where (4) is made convex through regularization. We prove the existence, uniqueness, and convergence of all three $\boldsymbol{\mu}$, $\mathbf{C}$, and $\boldsymbol{\tau}$ estimators in the proposed regularized setting.

C. Shape Parameter β

As mentioned previously, the shape parameter β only plays a role in estimating the perturbation $\boldsymbol{\tau}$. One could compensate for an approximate β value and retrieve the original distribution by varying $\boldsymbol{\tau}$, regardless of the true shape of the MGGD. Furthermore, estimating the (*extra*-)shape parameter of the EC model is a difficult problem that has been studied intensively in the non-perturbed MGGD case [36], [37], [38], [39]. Other *extra*-parameters, such as the degree of freedom of the t-distribution [40] or the shape parameter of a k-distribution [41] have also been studied.

With this in mind, estimating β for the presented $\boldsymbol{\tau}$-perturbed MGGD model and/or considering a misspecified model in terms of β appear out of the scope of this work. We will assume that the shape parameter is known or previously estimated (without misspecification) in a non-perturbed MGGD context for the experiments.

III. Proposed Convex Formulation

This section introduces a convex alternative to minimizing (4). The main idea of the proposed approach is to convexify the cost function through suitable variable changes and regularization. The purpose of regularizing the cost function is two-fold; first, prior information about the sought variables may be used to improve their estimation (*e.g.*, sparsity of the precision matrix $\mathbf{C}^{-1}$ [9]). Secondly, regularization is critical in tackling the non-convexity of the cost function. The following subsection explains the details of these transformations.

A. A Regularized Cost Function

As explained above, convexifying (4) requires a series of transformations. First, let us re-parameterize $\mathcal{L}$ by setting

$$\mathbf{Q} = \mathbf{C}^{-1/2} \tag{10}$$

$$\mathbf{m} = \mathbf{Q}\boldsymbol{\mu} \tag{11}$$

$$\boldsymbol{\theta} = (\theta_n)_{1\leqslant n\leqslant N} = (\tau_n^{\beta/(\beta-1)})_{1\leqslant n\leqslant N}, \tag{12}$$

where $\mathbf{Q} \in \mathcal{S}_K^{++}$ and $\beta \neq 1$. Note that the variable change (10) was also made in [17]. Using the re-parametrization (10)-(12), we can rewrite the original neg-likelihood function (4) as

$$\begin{aligned}&\mathcal{L}(\mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\tau})\\ &= \widetilde{\mathcal{L}}(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta})\\ &= \frac{1}{2}\sum_{n=1}^{N} \frac{\|\mathbf{Q}\mathbf{y}_n - \mathbf{m}\|^\beta}{\theta_n^{\beta-1}} - N\log\det\mathbf{Q} + K(1-1/\beta)\sum_{n=1}^{N}\log\theta_n.\end{aligned} \tag{13}$$

If we exclude the last term, the obtained function becomes convex with respect to $(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta})$ for $\beta > 1$. In particular, the first term fits the form of a perspective function, which is known to be convex with respect to its argument (see the proof outlines in Subsection III-C and Appendix A). In Subsections III-B and III-E, we show that carefully choosing a regularization function for $\boldsymbol{\theta}$ counteracts the concavity of the last term in (13) and leads to a fully convex formulation. More generally, we consider the addition of regularization functions g_{Q}, g_{m}, and g_{ϑ} on $\mathbf{Q}$, $\mathbf{m}$, and $\boldsymbol{\theta}$, respectively.[5] Thus, we introduce the general regularized form of (13) as follows:

$$f(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) = \begin{cases} \widetilde{\mathcal{L}}(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) + g_{\mathsf{Q}}(\mathbf{Q}) + g_{\mathsf{m}}(\mathbf{m}) + g_{\vartheta}(\boldsymbol{\theta}) & \text{if } \boldsymbol{\theta} \in]0, +\infty[^N \\ & \text{and } \mathbf{Q} \in \mathcal{S}_K^{++} \\ +\infty & \text{otherwise,} \end{cases} \tag{14}$$

From a Bayesian perspective, minimizing the regularized cost function (14) amounts to a Maximum A Posteriori (MAP) estimation of $\mathbf{Q}$, $\mathbf{m}$, and $\boldsymbol{\theta}$.

The regularized cost function f is suitable for different robust estimation problems depending on the choice of the regularization functions. For example, in graph processing applications, it is known that the precision matrix $\mathbf{C}^{-1}$ is sparse [9]. Prior information on the nature of the perturbation

[4] Considering function $s^{1/2}\,u(s)$ as in [35], the case in Eq. (9) corresponds to $s^{1/2}\,u(s) = K\,s^{-1/2}$, which is not an increasing function of s near 0 (*i.e.,* it is not redescending). Eqs. (8) correspond to the limiting case of redescending M-estimators.

[5] These functions can take infinity values to model potential hard constraints on the variables. For example, if one seeks to restrict the vector $\mathbf{m}$ to some set $\mathcal{D} \subset \mathbb{R}^K$ (*e.g.*, some hypercube or some ball) a suitable choice for g_{m} is the indicator function $\iota_{\mathcal{D}}$ of $\mathcal{D}$, equal to 0 on $\mathcal{D}$ and $+\infty$ elsewhere. These regularization functions can be viewed as the potentials associated with (possibly improper) prior probability density functions proportional to $\exp(-g_{\mathsf{Q}}(\cdot))$, $\exp(-g_{\mathsf{m}}(\cdot))$, and $\exp(-g_{\vartheta}(\cdot))$, respectively.

$\boldsymbol{\tau}$ may also be accessible (*e.g.,* bounds), or one may seek to restrict the mean to a specific set (*e.g.,* known values for a restricted subset of components). Examples of different practical regularizations will be discussed in more detail in Subsections III-D and III-E.

B. Study of Cost Function f

This subsection discusses key properties of the proposed regularized cost function f. We will rely on these properties to choose the regularization functions and develop a suitable minimization strategy. Namely, we introduce two central propositions describing the assumptions that will ensure convexity and the existence of a minimizer. (Proof outlines are provided in Subsection III-C, and detailed proofs of these propositions can be found in Appendices A–B.) To simplify the analysis, we start by introducing a function gathering the terms in $\boldsymbol{\theta}$ in (13)-(14), if we except those in the first summation:

$$(\forall \boldsymbol{\theta} \in \mathbb{R}^N) \quad \widetilde{g}_\vartheta(\boldsymbol{\theta}) = g_\vartheta(\boldsymbol{\theta}) - K(1-1/\beta)\sum_{n=1}^{N}\psi(\theta_n), \tag{15}$$

where

$$(\forall \xi \in \mathbb{R}) \quad \psi(\xi) = \begin{cases} -\log \xi & \text{if } \xi > 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{16}$$

Similarly, by defining

$$(\forall \mathbf{Q} \in \mathcal{S}_K) \quad \Psi(\mathbf{Q}) = \begin{cases} -N \log\det \mathbf{Q} & \text{if } \mathbf{Q} \in \mathcal{S}_K^{++} \\ +\infty & \text{otherwise,} \end{cases} \tag{17}$$

function $\widetilde{g}_\mathsf{Q} = g_\mathsf{Q} + \Psi$ gathers the terms in $\mathbf{Q}$ in (13)-(14), excluding those in the first summation. Then, we have the following result with detailed proofs in Appendices A and B. For readers who prefer to avoid technical details in convex analysis, an outline of the proofs is given in Section III-C.

Proposition III.1 *Assume that $g_\mathsf{Q} \in \Gamma_0(\mathcal{S}_K)$ and that there exists an invertible matrix $\overline{\mathbf{Q}} \in \mathcal{S}_K^{++}$ such that $g_\mathsf{Q}(\overline{\mathbf{Q}}) < +\infty$. Assume that $g_\mathsf{m} \in \Gamma_0(\mathbb{R}^K)$ and $\widetilde{g}_\vartheta \in \Gamma_0(\mathbb{R}^N)$. Then, f is a proper lower-semicontinuous convex function on $\mathcal{S}_K \times \mathbb{R}^K \times \mathbb{R}^N$.*

Proposition III.2 *In addition to the conditions stated in Proposition III.1, let us make the following assumptions:*

(i) $g_\mathsf{m} \geqslant 0$;

(ii) $\widetilde{g}_\vartheta = \widetilde{g}_{\vartheta,0} + \widetilde{g}_{\vartheta,1}$ *where* $\widetilde{g}_{\vartheta,0} \in \Gamma_0(\mathbb{R}^N)$, $\widetilde{g}_{\vartheta,1} \in \Gamma_0(\mathbb{R}^N)$, $\operatorname{dom} \widetilde{g}_{\vartheta,1} =]0,+\infty[^N$, *and*

$$(\forall \check{\boldsymbol{\theta}} \in [0,+\infty[^N \setminus]0,+\infty[^N) \quad \lim_{\boldsymbol{\theta}\to\check{\boldsymbol{\theta}}} \widetilde{g}_{\vartheta,1}(\boldsymbol{\theta}) = +\infty; \tag{18}$$

(iii) $\widetilde{g}_\mathsf{Q}$ *and* $\widetilde{g}_\vartheta$ *are coercive functions.*

Then f admits a minimizer. Such a minimizer is unique if $\widetilde{g}_\mathsf{Q}$ and $\widetilde{g}_\vartheta$ are strictly convex.

It stems from Propositions III.1 and III.2 that the convexity of f and the existence of a minimizer depend on the choice of the regularizations. First, one can easily notice that the conditions on g_m and g_Q are satisfied by a wide range of typical regularization functions. Secondly, the assumptions on function $\widetilde{g}_\vartheta$, gathering the terms in $\boldsymbol{\theta}$, are less usual but critical regarding the above results. More precisely:

- The assumptions on g_m are mild. They are satisfied both when no information is available on parameter $\mathbf{m}$ ($g_\mathsf{m} = 0$) or when this vector is known ($g_\mathsf{m} = \iota_{\{\overline{\mathbf{m}}\}}$ with $\overline{\mathbf{m}} \in \mathbb{R}^K$). In practice, restricting the mean to a specific set (*e.g.*, known values for a restricted subset of components) satisfies these conditions.
- The assumptions on g_Q are rather classical. In particular, the core assumptions are satisfied by standard convex penalization promoting sparsity or group sparsity [9]. The strict convexity assumptions can be satisfied by adding a quadratic term leading to an elastic net-like penalization [42]. Subsection III-D will explicit some useful regularization functions for $\mathbf{Q}$.
- The key conditions for convexity are related to the choice of $\widetilde{g}_\vartheta$. Note that this function contains the regularization that acts implicitly on the perturbation parameter $\boldsymbol{\tau}$. In Subsection III-E, we propose a regularization satisfying the required conditions for $\widetilde{g}_\vartheta$ and analyse its impact on estimating MGGD parameters.

In the following, we present a sketch of the proofs and then discuss some illustrative examples of regularization functions satisfying the above conditions for $\mathbf{m}, \mathbf{Q}$ and $\boldsymbol{\theta}$. As discussed above, the regularization choice for the perturbations is central to preserving convexity properties. Therefore, we will analyse the impact of the proposed regularization $\widetilde{g}_\vartheta$ on the cost function f and propose guidelines for choosing associated hyperparameters in practical scenarios.

C. Proof Outlines

1) Proof of Proposition III.1: The proof first shows that f is a sum of proper lower-semicontinuous convex functions, which makes it a lower-semicontinuous convex function. More precisely, because of the form of the first term in $\widetilde{\mathcal{L}}$, the first part of the function f is lower-semicontinuous and convex according to [43, Proposition 9.42]. In addition, the assumptions made on the different regularizations in Proposition III.1 allow us to deduce that (the entire) function f is convex lower-semicontinuous. In a second step, based on the assumptions made on g_Q and the form of the function $\widetilde{g}_\vartheta$, we show that f is also proper (*i.e.,* its domain is non-empty), which completes the proof.

2) Proof of Proposition III.2: The proof uses three main steps. First, we show the existence of a unique infimum $\widehat{\mathbf{m}}$ based on strict convexity and coercivity of f with respect to $\mathbf{m}$ (only). Then, using [43, Proposition 8.35], we restrict the analysis to a new function composed of the marginal function and remaining parts of f independent of $\mathbf{m}$. Writing the domain of this new function explicitly, we show that the necessary properties for [43, Proposition 9.33] are satisfied; thus, it is proper convex lower-semicontinuous. Based on Assumption (iii), we finally show that this new function is coercive. Thus, a minimizer exists and is unique if strict convexity is achieved for $\widetilde{g}_\mathsf{Q}$ and $\widetilde{g}_\vartheta$.

D. Examples of Regularization Choices for $\mathbf{m}$ and $\mathbf{Q}$

As seen in Subsection III-B, the assumptions on the regularizations g_{m} and g_{Q} are relatively straightforward to satisfy. In the following, we let $g_{\mathsf{m}} = \iota_{\{\mathbf{0}\}}$ and consider a frequently encountered scenario where the precision matrix $\mathbf{C}^{-1}$ is sparse. It is common in sparse estimation to enforce most entries of $\mathbf{C}^{-1}$ to zero using the ℓ_1 norm (as a surrogate for the ℓ_0 pseudo-norm, which counts the number of zero entries in a matrix or vector). More precisely, the following ways of addressing sparsity can be considered:

- The standard graphic LASSO (GLASSO) problem [9] can be obtained in the particular case where $\beta = 2$ by setting $g_{\mathsf{m}} = \iota_{\{\mathbf{0}\}}$, $g_\vartheta = \iota_{\{\mathbf{1}\}}$, and $g_{\mathsf{Q}} \colon \mathbf{Q} \mapsto \lambda\|\mathbf{Q}^2\|_1$ (or, equivalently, $\lambda\|\mathbf{C}^{-1}\|_1$) with $\lambda \in]0, +\infty[$.
- With the proposed re-parametrization, it is more direct to impose sparsity on the (new) variable $\mathbf{Q}$ (square root of $\mathbf{C}^{-1}$) by choosing
$$g_{\mathsf{Q}} \colon \mathbf{Q} \mapsto \lambda\|\mathbf{Q}\|_1. \tag{19}$$
Note that choosing (19) for $\mathbf{Q}$ (and indirectly for the precision matrix $\mathbf{C}^{-1}$), without any regularization on $\mathbf{m}$ and $\boldsymbol{\theta}$, a zero mean, and no multiplicative perturbations ($g_\vartheta = \iota_{\{\mathbf{1}\}}$), leads to a particular case of the proposed formulation studied in [17]. We will also use (19) in our experimental study in Section V.
- Another typical choice for enforcing sparsity is elastic net regularization where $g_{\mathsf{Q}} = \lambda\|\cdot\|_1 + \frac{\epsilon}{2}\|\cdot\|_{\mathrm{F}}^2$ with λ and $\epsilon \in]0, +\infty[$. It should be noted that elastic net regularization is strictly convex, and if $\widetilde{g}_\vartheta$ is also strictly convex, f will admit a unique minimizer, as stated in Proposition III.2.

E. Regularization Choice for $\boldsymbol{\theta}$ and Impact on f

1) Choice of g_ϑ: The choice of the regularization on $\boldsymbol{\theta}$ is central to our analysis. The goal is to ensure the convexity of the global cost function f by compensating for the last term in (13). The concavity of this term will be counterbalanced by choosing a regularization leading to a coercive function. A simple choice for the regularization function g_ϑ allowing us to achieve this goal (while satisfying the requirements in Propositions III.1 and III.2) is the potential of a generalized Gamma distribution with scale parameter η, shape parameter $(\kappa+1)$, and exponent parameter α, *i.e.,*
$$(\forall \boldsymbol{\theta} \in \mathbb{R}^N) \quad g_\vartheta(\boldsymbol{\theta}) = \frac{1}{\eta^\alpha}\|\boldsymbol{\theta}\|_\alpha^\alpha + \kappa \sum_{n=1}^N \psi(\theta_n) \tag{20}$$
where ψ has been defined in (16), $\alpha \in [1, +\infty[$, $\eta \in]0, +\infty[$, and $\kappa \in]K(1-1/\beta), +\infty[$. We can then set
$$\widetilde{g}_{\vartheta,0}(\boldsymbol{\theta}) = \frac{1}{\eta^\alpha}\|\boldsymbol{\theta}\|_\alpha^\alpha$$
$$\widetilde{g}_{\vartheta,1}(\boldsymbol{\theta}) = \big(\kappa - K(1-1/\beta)\big)\sum_{n=1}^N \psi(\theta_n),$$
which are functions satisfying Assumptions (ii) and (iii) in Proposition III.2. As emphasized above, coercivity is achieved by adjusting κ such that the weight $\kappa - K(1-1/\beta)$ of the logarithmic term is positive. In addition, since $\widetilde{g}_{\vartheta,1}$ is then strictly convex, $\widetilde{g}_\vartheta$ is also strictly convex. It follows that the resulting cost function has a unique minimizer, provided that the chosen g_{Q} is also strictly convex.

2) Impact on Cost Function f: To gain better insight into the impact of regularization (20), we study the average behaviour of the cost function f with respect to $\boldsymbol{\theta}$ and the associated regularization parameters η, κ, and α.[6] Let us first isolate the terms of interest by introducing functions $(f_n)_{1\leqslant n\leqslant N}$ such that, for every $(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) \in \mathcal{S}_K^{++} \times \mathbb{R}^K \times]0, +\infty[^N$,
$$f(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) = \sum_{n=1}^N f_n(\theta_n, \mathbf{Q}, \mathbf{m}) + \widetilde{g}_{\mathsf{Q}}(\mathbf{Q}) + g_{\mathsf{m}}(\mathbf{m}) \tag{21}$$
and, for every $n \in \{1, \ldots, N\}$,
$$f_n(\theta_n, \mathbf{Q}, \mathbf{m}) = \frac{\|\mathbf{Q}\mathbf{y}_n - \mathbf{m}\|^\beta}{2\theta_n^{\beta-1}} + \frac{\theta_n^\alpha}{\eta^\alpha} - \big(\kappa - K(1-1/\beta)\big)\log\theta_n. \tag{22}$$
To focus our analysis on the variable $\boldsymbol{\theta}$, we make the simplifying assumption that $\mathbf{Q}$ and $\mathbf{m}$ have been perfectly identified with respect to the original statistical model. We can then restrict the analysis to the above function f_n by looking at its average behaviour:
$$\mathsf{E}\{f_n(\theta_n, \mathbf{Q}, \mathbf{m})\} = \frac{\mathsf{E}\{\|\mathbf{Q}\mathbf{y}_n - \mathbf{m}\|^\beta\}}{2\theta_n^{\beta-1}} + \frac{\theta_n^\alpha}{\eta^\alpha} - \big(\kappa - K(1-1/\beta)\big)\log\theta_n. \tag{23}$$
In (23), $(\mathbf{Q}\mathbf{y}_n - \mathbf{m})$ follows an $\mathrm{MGGD}_K(\beta, \mathbf{0}, \overline{\tau}_n^2\,\mathbf{Id})$, where $\overline{\tau}_n > 0$ is the true value of the multiplicative perturbation factor in Model (3) (with re-parametrizations (10)–(12)). We then have
$$\mathsf{E}\{\|\mathbf{Q}\mathbf{y}_n - \mathbf{m}\|^\beta\} = \frac{2K}{\beta}\,\overline{\theta}_n^{\beta-1} \tag{24}$$
with $\overline{\theta}_n = \overline{\tau}_n^{\beta/(\beta-1)}$, and
$$\mathsf{E}\{f_n(\theta_n, \mathbf{Q}, \mathbf{m})\} = K(1-1/\beta)\overline{f}(\theta_n), \tag{25}$$
where
$$\overline{f}(\theta_n) = \frac{\overline{\theta}_n^{\beta-1}}{(\beta-1)\theta_n^{\beta-1}} + \frac{\beta\theta_n^\alpha}{K(\beta-1)\eta^\alpha} - \big(\overline{\kappa} - 1\big)\log\theta_n \tag{26}$$
and
$$\overline{\kappa} = \kappa\beta/(K(\beta-1)). \tag{27}$$
With the conditions imposed on κ, one has $\overline{\kappa} > 1$. The derivative of $\overline{f}$ at θ_n is
$$\overline{f}'(\theta_n) = \frac{1}{\theta_n}\left(1 - \left(\frac{\overline{\theta}_n}{\theta_n}\right)^{\beta-1} + \frac{\alpha\beta\theta_n^\alpha}{K(\beta-1)\eta^\alpha} - \overline{\kappa}\right). \tag{28}$$
Moreover, one has $\overline{\theta}_n = 1$ for unperturbed data. In this case, if the scale parameter
$$\eta = \left(\frac{\alpha\beta}{K(\beta-1)\overline{\kappa}}\right)^{1/\alpha} \text{ or equivalently, } \eta = \left(\frac{\alpha}{\kappa}\right)^{1/\alpha}, \tag{29}$$

[6] We will denote the true target value in this subsection by $\bar{\boldsymbol{\theta}}$ (resp. $\bar{\boldsymbol{\tau}}$) to distinguish it from the parameter or variable $\boldsymbol{\theta}$ (resp. $\boldsymbol{\tau}$).

then $\overline{f}$ is decreasing over $]0,1]$ and increasing over $[1,+\infty[$ (see Appendix C). As expected, the minimum of this function is thus attained at $\widehat{\theta}_n = 1$. For a given value of $\overline{\kappa}$, the choice for η in (29) makes function $\overline{f}$ independent of the vector dimension K, and we have, in this case,

$$\overline{f}(\theta_n) = \frac{\overline{\theta}_n^{\beta-1}}{(\beta-1)\theta_n^{\beta-1}} + \frac{\overline{\kappa}\theta_n^{\alpha}}{\alpha} - \big(\overline{\kappa}-1\big)\log\theta_n. \tag{30}$$

In Appendix C, we show that the above function has a unique minimizer $\widehat{\theta}_n$. If $\overline{\theta}_n > 1$, $\widehat{\theta}_n \in]1, \overline{\theta}_n[$, which means that a bias is introduced by the regularization. As $\widehat{\theta}_n$ is an increasing function of $\overline{\theta}_n$, the method remains effective in reducing the influence of outliers. In addition, when $\overline{\theta}_n > 1$, $\widehat{\theta}_n$ is a decaying function of α (resp. $\overline{\kappa}$). In other words, a stronger regularization comes at the expense of a larger bias in estimating the perturbation parameter τ_n. This suggests choosing both α and $\overline{\kappa}$ close to 1.

3) Choice of Parameters: Plots of the graph of the function $\overline{f}$ for different values of α are shown in Figure. 1. As expected, when α increases, stronger perturbations θ_n are penalized more heavily, which suggests the use of higher α values when more robustness is desired (*i.e.,* in heavily perturbed scenarios). Figure. 2 shows the variations of $\widehat{\theta}_n$ as a function of $\overline{\theta}_n$ for the same parameter values. The introduced bias can be seen in the increasing gap between the target $\overline{\theta}_n$ and $\widehat{\theta}_n$ as α increases. We can also notice that the bias increases as the perturbations become larger for a given value of α. Note that, although a larger value of α introduces more bias, it can also benefit the convergence speed of optimization algorithms in addition to more penalization. For example, when $\alpha = 2$, $\overline{f}$ is a strongly convex function for which we can typically expect a linear convergence rate of optimization algorithms. This suggests that a trade-off between acceptable bias, robustness to perturbations, and reasonable computational speed is to be considered in practice.

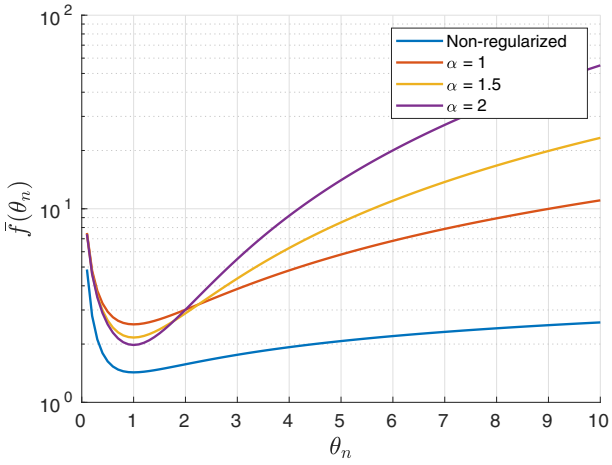

Figure 1. Graph of function $\overline{f}$ in (26) when η is given by (29), $\beta = 1.7$, and $\overline{\theta}_n = 1$. In blue, the non-regularized case when $\eta \to +\infty$ and $\kappa = 0$. Other plots correspond to $\overline{\kappa} = 1.1$ with increasing values of α, *i.e.,* $\alpha = 1$ in red, $\alpha = 1.5$ in yellow, and $\alpha = 2$ in purple. The ordinate axis is in the log scale.

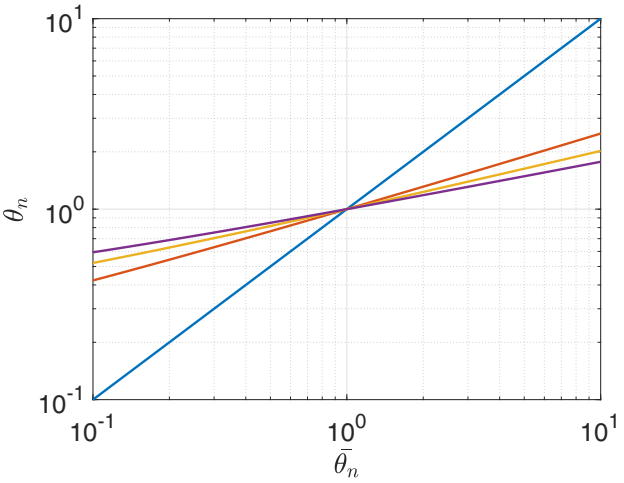

Figure 2. Variations of $\widehat{\theta}_n$ versus $\overline{\theta}_n$ based on a numerical solution of (63) when $\beta = 1.7$. In blue, the non-regularized case where $\kappa = 0$. Other plots correspond to $\overline{\kappa} = 1.1$ with increasing values of α, *i.e.,* $\alpha = 1$ in red, $\alpha = 1.5$ in yellow, and $\alpha = 2$ in purple. Both axes are in the log scale.

IV. Proximal Primal-Dual Algorithm

Minimizing cost function (14) requires handling various regularizations. Proximal algorithms can efficiently tackle such minimization problems: regularization terms are dealt with using proximity operators, usually in closed form [43], [44], [45]. In this section, we will show that Problem (14) can be reformulated as the minimization of a sum of compositions of convex functions and linear operators, which can be solved with convergence guarantees using a proximal primal-dual algorithm. Furthermore, considering a suitable norm on the dual space allows us to make the convergence of the proximal primal-dual algorithm faster.

A. Primal and Dual Spaces

The primal space is here $\mathcal{H} = \mathcal{S}_K \times \mathbb{R}^K \times \mathbb{R}^N$, endowed with the norm:

$$(\forall \mathbf{p} = (\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) \in \mathcal{H}) \quad \|\mathbf{p}\|_{\mathcal{H}} = \sqrt{\|\mathbf{Q}\|_{\mathrm{F}}^2 + \|\mathbf{m}\|^2 + \|\boldsymbol{\theta}\|^2}. \tag{31}$$

The dual space is $\mathcal{G}_1 \times \mathcal{G}_2$ with $\mathcal{G}_1 = (\mathbb{R}^K)^N \times \mathbb{R}^N$ and $\mathcal{G}_2 = \mathcal{S}_K \times \mathbb{R}^N$. The first space is endowed with the norm

$$\big(\forall \mathbf{v}_1 = \big((\mathbf{u}_n)_{1\leqslant n\leqslant N}, \boldsymbol{\theta}\big) \in \mathcal{G}_1\big) \quad \|\mathbf{v}_1\|_{\mathcal{G}_1} = \sqrt{\sum_{n=1}^{N} \|\mathbf{u}_n\|^2 + \omega_1\|\boldsymbol{\theta}\|^2}, \tag{32}$$

while the second one is equipped with the norm

$$\big(\forall \mathbf{v}_2 = \big(\mathbf{Q}, \boldsymbol{\theta}\big) \in \mathcal{G}_2\big) \quad \|\mathbf{v}_2\|_{\mathcal{G}_2} = \sqrt{\|\mathbf{Q}\|_{\mathrm{F}}^2 + \omega_2\|\boldsymbol{\theta}\|^2}, \tag{33}$$

where $(\omega_1, \omega_2) \in]0,+\infty[^2$. In the proposed algorithm, the latter parameters add more flexibility and allow to make the update of the $\boldsymbol{\theta}$ estimates faster than the standard choice $\omega_1 =$

$\omega_2 = 1$. Note that all the described product spaces are finite-dimensional Hilbert spaces.

B. Cost Function f as a Sum of Elementary Convex Functions

To design a proximal primal-dual algorithm, we first reformulate the optimization problem as a minimization of a sum of compositions of convex functions and linear operators, which is required to guarantee convergence. Recall that, when $\boldsymbol{\theta} \in]0,+\infty[^N$ and $\mathbf{Q} \in \mathcal{S}_K^{++}$,

$$f(\mathbf{Q},\mathbf{m},\boldsymbol{\theta}) = \widetilde{\mathcal{L}}(\mathbf{Q},\mathbf{m},\boldsymbol{\theta}) + g_{\mathsf{Q}}(\mathbf{Q}) + g_{\mathsf{m}}(\mathbf{m}) + g_{\vartheta}(\boldsymbol{\theta}). \quad (34)$$

Thus, by introducing (i) a new variable $\mathbf{p} = (\mathbf{Q},\mathbf{m},\boldsymbol{\theta})$, (ii) three convex functions F, G_1 and G_2 encompassing parts of the likelihood term $\widetilde{\mathcal{L}}$ and the regularizations $g_{\mathsf{Q}}, g_{\mathsf{m}}$ and g_{ϑ}, and (iii) linear operators $\mathbf{L}_1$ and $\mathbf{L}_2$, we can rewrite the optimization problem as follows

$$\underset{\mathbf{p}\in\mathcal{H}}{\text{minimize}} \;\; F(\mathbf{p}) + G_1(\mathbf{L}_1\mathbf{p}) + G_2(\mathbf{L}_2\mathbf{p}). \quad (35)$$

Precisely, with the choice of the regularisation functions made in Sections III-D and and III-E, F and G_2 are the following convex functions:

$$\begin{aligned} F(\mathbf{Q},\mathbf{m},\boldsymbol{\theta}) &= \Psi(\mathbf{Q}) + g_{\mathsf{m}}(\mathbf{m}) + \widetilde{g}_{\vartheta,1}(\boldsymbol{\theta}) \\ G_2(\mathbf{Q},\boldsymbol{\theta}) &= g_{\mathsf{Q}}(\mathbf{Q}) + \widetilde{g}_{\vartheta,0}(\boldsymbol{\theta}). \end{aligned} \quad (36)$$

Since F and G_2 are separable functions, their proximity operators can be calculated componentwise. In order to define G_1, we introduce the following functions:

$$\Phi\colon (\mathbb{R}^K)^N \times \mathbb{R}^N \colon ((\mathbf{v}_n)_{1\leqslant n\leqslant N}, \boldsymbol{\theta}) \mapsto \sum_{n=1}^N \varphi(\mathbf{v}_n, \theta_n), \quad (37)$$

$$\varphi\colon \mathbb{R}^K \times \mathbb{R} \colon (\mathbf{u},\xi) \mapsto \begin{cases} \dfrac{\|\mathbf{u}\|^\beta}{2\xi^{\beta-1}} & \text{if } \mathbf{u}\neq\mathbf{0} \text{ and } \xi > 0 \\ 0 & \text{if } \mathbf{u}=\mathbf{0} \text{ and } \xi = 0 \\ +\infty & \text{otherwise,} \end{cases} \quad (38)$$

and choose $G_1 = \Phi$. Finally, $\mathbf{L}_1$ and $\mathbf{L}_2$ are linear operators such that

$$\begin{aligned} \mathbf{L}_1 &\colon \mathcal{H} \to \mathcal{G}_1 \colon (\mathbf{Q},\mathbf{m},\boldsymbol{\theta}) \mapsto (\mathcal{T}(\mathbf{Q},\mathbf{m}),\boldsymbol{\theta}) \\ \mathbf{L}_2 &\colon \mathcal{H} \to \mathcal{G}_2 \colon (\mathbf{Q},\mathbf{m},\boldsymbol{\theta}) \mapsto (\mathbf{Q},\boldsymbol{\theta}), \end{aligned} \quad (39)$$

where $\mathcal{T}$ is the linear operator

$$\mathcal{T}\colon \mathcal{S}_K \times \mathbb{R}^K \to (\mathbb{R}^K)^N \colon (\mathbf{Q},\mathbf{m}) \mapsto (\mathbf{Q}\mathbf{y}_n - \mathbf{m})_{1\leqslant n\leqslant N}. \quad (40)$$

The existence of a solution to the minimization problem (35) follows from Proposition III.2.

C. Proposed Iterative Algorithm

To minimize (34), and equivalently, solve the reformulation (35), a standard solution is the Chambolle-Pock algorithm [46]. The algorithm iterates between a proximal step on F (in the primal space) and an update of the dual variables $(\mathbf{v}_{1,k})_{k\geqslant 0}$ and $(\mathbf{v}_{2,k})_{k\geqslant 0}$ (related to G_1 and G_2). Starting from initial values $\mathbf{p}_0 \in \mathcal{H}$, $\mathbf{v}_{1,0} \in \mathcal{G}_1$, and $\mathbf{v}_{2,0} \in \mathcal{G}_2$, the algorithm reads

$$\begin{aligned} &\text{For } k = 0, 1, \ldots \\ &\quad \mathbf{p}_{k+1} = \operatorname{prox}_{\gamma F}(\mathbf{p}_k - \gamma(\zeta_1 \mathbf{L}_1^* \mathbf{v}_{1,k} + \zeta_2 \mathbf{L}_2^* \mathbf{v}_{2,k})) \\ &\quad \widetilde{\mathbf{p}}_k = 2\mathbf{p}_{k+1} - \mathbf{p}_k \\ &\quad \mathbf{v}_{1,k+1} = \Big(\mathbf{Id} - \operatorname{prox}_{\zeta_1^{-1} G_1}\Big)(\mathbf{v}_{1,k} + \mathbf{L}_1 \widetilde{\mathbf{p}}_k) \\ &\quad \mathbf{v}_{2,k+1} = \Big(\mathbf{Id} - \operatorname{prox}_{\zeta_2^{-1} G_2}\Big)(\mathbf{v}_{2,k} + \mathbf{L}_2 \widetilde{\mathbf{p}}_k), \end{aligned} \quad (41)$$

where the adjoints of $\mathbf{L}_1$ and $\mathbf{L}_2$ are denoted by $\mathbf{L}_1^*$ and $\mathbf{L}_2^*$, and (γ,ζ_1,ζ_2) are positive scalar parameters.

To make the role of each involved variable more explicit, we can rewrite Algorithm (41) in terms of $(\mathbf{Q},\mathbf{m},\boldsymbol{\theta})$. Let us set, at each iteration number k, $\mathbf{p}_k = (\mathbf{Q}_k,\mathbf{m}_k,\boldsymbol{\theta}_k)$, $\widetilde{\mathbf{p}}_k = (\widetilde{\mathbf{Q}}_k,\widetilde{\mathbf{m}}_k,\widetilde{\boldsymbol{\theta}}_k)$, $\mathbf{v}_{1,k} = ((\mathbf{u}_{k,n})_{1\leqslant n\leqslant N}, \boldsymbol{\theta}_{1,k}^{\#})$, and $\mathbf{v}_{2,k} = (\mathbf{Q}_k^{\#}, \boldsymbol{\theta}_{2,k}^{\#})$, where $\boldsymbol{\theta}_{1,k}^{\#}$ (resp. $\boldsymbol{\theta}_{2,k}^{\#}$) corresponds to the dual variable of $\boldsymbol{\theta}_k$ in the space $\mathcal{G}_1$ (resp. $\mathcal{G}_2$), while $\mathbf{Q}_k^{\#}$ stands for the dual variable of $\mathbf{Q}_k$ in the space $\mathcal{G}_2$. Let us also define $\zeta_3 = \omega_1\zeta_1$ and $\zeta_4 = \omega_2\zeta_2$. From the expressions of the adjoint operators of $\mathbf{L}_1$ and $\mathbf{L}_2$ derived in Appendix D, we obtain the following iterative algorithm:

$$\begin{aligned} &\text{For } k = 0, 1, \ldots \\ &\quad \widehat{\mathbf{Q}}_k = \mathbf{Q}_k - \gamma\Big(\zeta_1 \sum_{n=1}^N \mathbf{u}_{k,n}\mathbf{y}_n^\top + \zeta_2 \mathbf{Q}_k^{\#}\Big) \\ &\quad \mathbf{Q}_{k+1} = \operatorname{prox}_{\gamma\Psi}\big((\widehat{\mathbf{Q}}_k + \widehat{\mathbf{Q}}_k^\top)/2\big) \\ &\quad \mathbf{m}_{k+1} = \operatorname{prox}_{\gamma g_{\mathsf{m}}}\Big(\mathbf{m}_k + \gamma\zeta_1 \sum_{n=1}^N \mathbf{u}_{k,n}\Big) \\ &\quad \boldsymbol{\theta}_{k+1} = \operatorname{prox}_{\gamma\widetilde{g}_{\vartheta,1}}\big(\boldsymbol{\theta}_k - \gamma(\zeta_3\boldsymbol{\theta}_{1,k}^{\#} + \zeta_4\boldsymbol{\theta}_{2,k}^{\#})\big) \\ &\quad \widetilde{\mathbf{Q}}_k = 2\mathbf{Q}_{k+1} - \mathbf{Q}_k \\ &\quad \widetilde{\mathbf{m}}_k = 2\mathbf{m}_{k+1} - \mathbf{m}_k \\ &\quad \widetilde{\boldsymbol{\theta}}_k = 2\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \\ &\quad \text{For } n = 1, \ldots, N \\ &\quad\quad (\mathbf{u}_{k+1,n}, \theta_{1,k+1,n}^{\#}) \\ &\quad\quad\quad = \Big(\mathbf{Id} - \operatorname{prox}_{\varphi}^{\zeta_1^{-1},\zeta_3^{-1}}\Big)(\mathbf{u}_{k,n} + \widetilde{\mathbf{Q}}_k\mathbf{y}_n - \widetilde{\mathbf{m}}_k, \theta_{1,k,n}^{\#} + \widetilde{\theta}_{k,n}) \\ &\quad \mathbf{Q}_{k+1}^{\#} = \Big(\mathbf{Id} - \operatorname{prox}_{\zeta_2^{-1} g_{\mathsf{Q}}}\Big)(\mathbf{Q}_k^{\#} + \widetilde{\mathbf{Q}}_k) \\ &\quad \boldsymbol{\theta}_{2,k+1}^{\#} = \Big(\mathbf{Id} - \operatorname{prox}_{\zeta_4^{-1}\widetilde{g}_{\vartheta,0}}\Big)(\boldsymbol{\theta}_{2,k}^{\#} + \widetilde{\boldsymbol{\theta}}_k), \end{aligned}$$

where the required proximity operators are specified in Appendix E.

D. Choice of the Step Parameters (γ,ζ_1,ζ_2)

To ensure the algorithm convergence, the scalar parameters (γ,ζ_1,ζ_2) must be chosen such that

$$\gamma(\zeta_1\|\mathbf{L}_1\|_{\mathrm{S}}^2 + \zeta_2\|\mathbf{L}_2\|_{\mathrm{S}}^2) < 1, \quad (42)$$

where $\|\cdot\|_{\mathrm{S}}$ is the operator norm. From the calculations of the norms of $\mathbf{L}_1$ and $\mathbf{L}_2$ in Appendix D, we deduce that a sufficient condition for (42) to be satisfied is

$$\zeta_1 \max\{\|\mathbf{Y}\|_{\mathrm{S}}^2, \omega_1\} + \zeta_2 \max\{1, \omega_2\} < \gamma^{-1}, \quad (43)$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_N \\ 1 & \dots & 1 \end{bmatrix}. \tag{44}$$

Based on (43), we chose $\omega_1 = \|\mathbf{Y}\|_{\mathrm{S}}^2$ and $\omega_2 = 1$ in our implementation.

V. Experiments

This section evaluates the proposed method in comparison with various state-of-the-art approaches. For the proposed method, the experiments are carried out by imposing sparsity on the precision matrix and regularization (20) on the perturbations (see Subsections III-D and III-E). We present results with varying N (sample size) and K (dimension) values. In particular, we study the standard case $N > K$ and the high-dimensional setting where $N \ll K$. The impact of sparse regularization on the precision matrix is investigated by considering dense and sparse matrices. Finally, we study the behaviour of the estimators with varying perturbation levels.

Our comparisons aim to evaluate the estimators' consistency and mean square error (MSE) with respect to the true parameters. The consistency of an estimated matrix $\hat{\boldsymbol{A}}$ is quantified by $\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{F}}$, whereas the MSE is $\mathsf{E}\left(\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{F}}^2\right)$, where $\boldsymbol{A}$ is the true matrix. In our experiments, the expectation is computed empirically by averaging $N_{\mathrm{MC}} = 10^4$ Monte Carlo runs except for the high-dimensional case where we use $N_{\mathrm{MC}} = 10^3$.

A. State-Of-The-Art Methods

In the standard case when $N > K$, the proposed method is compared to i) the empirical statistics corresponding to the classical sample mean and covariance and ii) robust Tyler's estimates obtained with (8) in both perturbed and ideal non-perturbed ($(\forall n)\ \tau_n = 1$) scenarios. Note that a fixed-point update estimates the mean in Tyler's method. Although this approach has no convergence guarantees, it is frequently used in practice [32]. In the high-dimensional setting ($N \ll K$), where Tyler's estimates do not exist, we show the results of the regularized Tyler's method [11], [10], where the covariance $\mathbf{C}$ is obtained by solving

$$\mathbf{C}(\lambda) = (1-\lambda)\frac{K}{N}\sum_{n=1}^{N}\frac{(\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^T}{(\mathbf{y}_n - \boldsymbol{\mu})^T\mathbf{C}(\lambda)^{-1}(\mathbf{y}_n - \boldsymbol{\mu})} + \lambda\mathbf{Id}\,, \tag{45}$$

with $\lambda \in]0,1[$ a fixed scalar regularization parameter. In addition, we compare the proposed method to the classical GLASSO [9] (discussed in Subsection III-D), and its robust extension GlassoGaussQn proposed in [15]. Both use a sparse regularization on the precision matrix, and the corresponding estimators are obtained by maximizing the penalized log-likelihood (with respect to $\mathbf{C}^{-1}$)

$$\log\det\mathbf{C}^{-1} - \mathrm{tr}(\widehat{\mathbf{C}}\mathbf{C}^{-1}) - \lambda\|\mathbf{C}^{-1}\|_1, \tag{46}$$

where the regularization parameter controlling sparsity is again denoted by λ, and $\widehat{\mathbf{C}} = \widehat{\mathbf{C}}_{\mathrm{SCM}}$ is the sample covariance matrix in GLASSO , which is replaced by a robust estimate of the covariance $\widehat{\mathbf{C}} = \widehat{\mathbf{C}}_{\mathrm{GlassoGaussQn}}$ in GlassoGaussQn. In [15], several robust GLASSO extensions are proposed by replacing the classical GLASSO input with various robust covariance matrix estimates. Among the estimators presented in [15], we investigate GlassoGaussQn (the best-performing one), which computes the robust covariance matrix using the Qn-estimator of scale and a Gaussian rank correlation matrix.

B. Simulation Parameters

Unless indicated otherwise, $\beta = 1.5$, the perturbations are uniformly distributed in the interval $[1, \boldsymbol{\tau}_{\max}]$, with a maximum perturbation level $\tau_{\max} = 5$, and the proportion of corrupted data is set to $p_\tau = 0.3$. (We end the following subsection with experiments analysing the performance for larger p_τ values.) The mean vector is drawn from the standard normal distribution (zero-mean and identity covariance matrix) in all experiments. The gamma prior parameters are set to $\kappa = 1.1\,K(1-1/\beta)$ and $\alpha = 1$, which are parameters satisfying the conditions in Subsection III-E. Note that we do not fine-tune κ (nor η) in our experiments, but performance could be improved by adjusting the amount of regularization. The sparse regularization parameter λ is a user-defined parameter; in this work, it is automatically adjusted by providing the desired sparsity level (that is, the proportion of zero entries in the generated matrices).

C. Experiments with $N > K$

The first experiments study two different cases of *i.i.d.* $K = 20$-dimensional data vectors $(\mathbf{x}_n)_{1\leqslant n\leqslant N}$ distributed first according to an $\mathrm{MGGD}_K(\beta, \mathbf{0}, \mathbf{C}_1)$ where $\mathbf{C}_1^{-1}$ is a *sparse* precision matrix, and then, an $\mathrm{MGGD}_K(\beta, \mathbf{0}, \mathbf{C}_2)$ with a *dense* precision matrix $\mathbf{C}_2^{-1}$.[7] The precision matrix $\mathbf{C}_1^{-1}$ is modelled by an auto-regressive (AR) process of order 3 such that

$$\mathbf{C}_1^{-1}(i,j) = \begin{cases} \rho^{|i-j|}, & \text{for } |i-j| = 0 \text{ and } i \in \{1,\dots,K\} \\ & \text{for } |i-j| = 1 \text{ and } i \in \{2,\dots,K-1\} \\ & \text{for } |i-j| = 2 \text{ and } i \in \{3,\dots,K-2\} \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

resulting in a tri-diagonal sparse matrix. The entries of the dense matrix $\mathbf{C}_2^{-1}$ are given by

$$(\forall (i,j) \in \{1,\dots,K\}^2), \quad \mathbf{C}_2^{-1}(i,j) = \rho^{|i-j|}, \tag{48}$$

where $\rho = 0.5$ is the correlation coefficient used to generate both precision matrices. Figure. 3 shows the evolution of the MSE and consistency of the obtained mean, covariance, and precision matrices for different values of N in the case of a sparse precision matrix. As expected, the best mean estimates correspond to the empirical ones in the ideal case. The least accurate/consistent are those obtained with the empirical estimator in the perturbed case. Comparing the proposed method and Tyler's estimates reveals an almost identical performance. This is interesting because, unlike Tyler's method, the proposed

[7] The observed samples $(\mathbf{y}_n)_{1\leqslant n\leqslant N}$ follow an $\mathrm{MGGD}_K(\beta, \boldsymbol{\mu}, \tau_n^2\,\mathbf{C}_1)$ (or an $\mathrm{MGGD}_K(\beta, \boldsymbol{\mu}, \tau_n^2\,\mathbf{C}_2)$) according to (3).

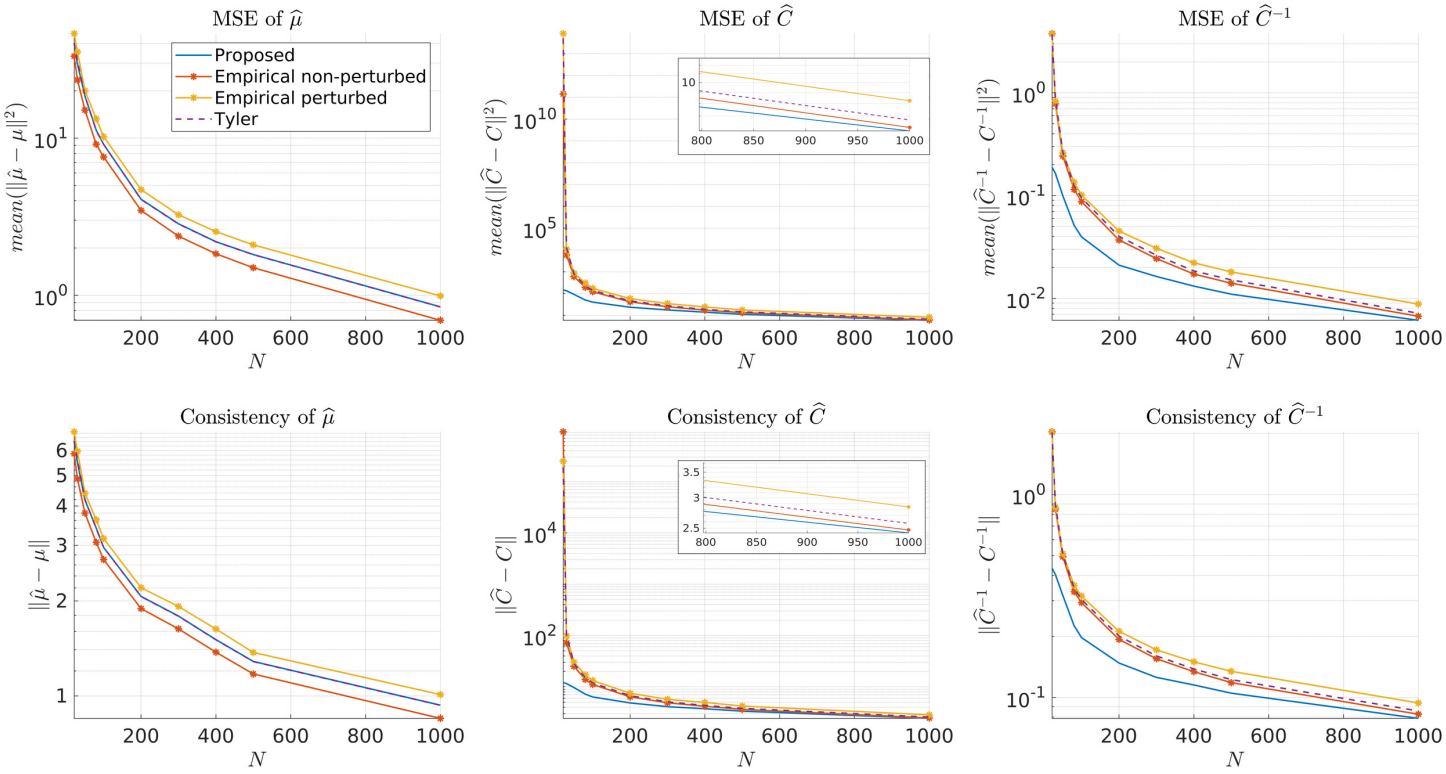

Figure 3. MSEs and consistencies of different mean, covariance and precision matrix estimators for the experiments with a sparse AR precision matrix.

method guarantees convergence in the unknown mean case. On the other hand, the results obtained by the proposed method for covariance and precision matrix estimation show a notable improvement over all other approaches. These improvements in MSE and consistency are even more noticeable for small sample sizes, indicating that the proposed method performs relatively well with a limited number of observations. The superior performance of the proposed method can be explained by the direct estimation of the precision matrix, strengthened by an appropriate sparse regularization. The benefits of sparse regularization in this first experimental scenario can be understood by studying the performance of dense precision matrix estimation (Table II). Here, the MSE gap between different methods is less pronounced, probably due to the minor impact of sparse regularization. However, it is worth noting that the performance of the proposed method remains superior to the other approaches for covariance and precision matrix estimation. As in the sparse case, the MSEs of the mean are similar for the proposed and Tyler's methods.

Table II
MSEs OBTAINED FOR THE EXPERIMENTS WITH A DENSE PRECISION MATRIX FOR INCREASING SAMPLE SIZES N.

N	Empirical			Tyler			Proposed		
	μ	$\mathbf{C}_2$	$\mathbf{C}_2^{-1}$	μ	$\mathbf{C}_2$	$\mathbf{C}_2^{-1}$	μ	$\mathbf{C}_2$	$\mathbf{C}_2^{-1}$
21	42.51	1.29e12	3.925	**37.94**	1.10e12	3.950	40.54	**1.90e4**	**0.629**
30	33.23	7.03e3	0.732	**29.37**	8.05e3	0.791	29.98	**1.65e3**	**0.330**
50	16.73	766.25	0.262	**14.75**	651.59	0.265	14.85	**419.14**	**0.191**
80	11.43	235.91	0.127	**9.99**	177.13	0.118	10.04	**147.80**	**0.100**
100	8.74	145.11	0.100	**7.77**	119.10	0.092	7.80	**99.53**	**0.079**
200	4.40	47.14	0.045	**4.07**	38.55	0.040	4.06	**35.51**	**0.037**
300	2.99	27.85	0.029	**2.59**	22.31	0.025	2.60	**21.50**	**0.024**
400	2.23	20.14	0.022	**1.98**	15.49	0.019	**1.98**	**15.25**	**0.018**
500	1.93	15.00	0.018	**1.66**	12.00	**0.015**	1.67	**11.85**	**0.015**
1000	0.97	6.89	0.009	**0.84**	5.43	**0.007**	**0.84**	**5.42**	**0.007**

We investigated the different estimator behaviours with varying perturbation levels and fixed $K = 20$ and $N = 100$. Figure. 4 shows the MSEs of the empirical (perturbed when $p_\tau > 0$), Tyler's, and proposed dense covariance matrix estimators[8] with different τ proportions. The experiments are conducted with $\beta = 2$ and $\beta = 1.5$, corresponding to a Gaussian and heavier-tailed distribution. Note that the empirical statistics are expected to perform ideally in the non-perturbed Gaussian case. As expected, the results show the robustness of the proposed method and Tyler's estimates compared to the empirical statistics when the perturbations increase. We can also see that the proposed method outperforms Tyler's estimates in all cases. Interestingly, the proposed method achieves the ideal performance of the empirical statistics in the non-perturbed Gaussian case ($p_\tau = 0$, $\beta = 2$). In the non-perturbed heavy-tailed scenario ($p_\tau = 0$, $\beta = 1.5$), the proposed method outperforms the empirical statistics. In contrast, Tyler's estimates are the least accurate when both distributions are not perturbed.

[8]Similar behaviour is observed for the precision matrix estimates, while the mean estimates follow previous findings and are comparable to Tyler's ones for varying perturbation levels.

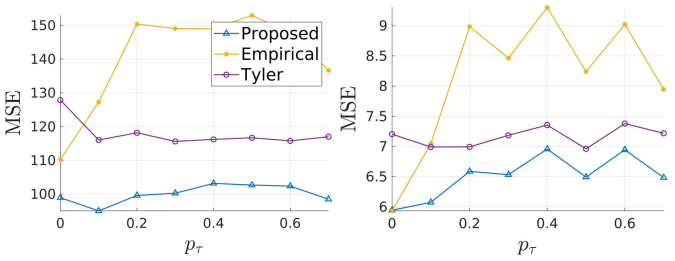

Figure 4. MSEs of the estimated dense covariance matrices for varying perturbation proportions p_τ for (a) an MGGD ($\beta = 1.5$) and (b) Gaussian distribution ($\beta = 2$).

Table III
MSEs OBTAINED FOR SPARSE PRECISION MATRIX ESTIMATION WHEN $K = 100$ AND $N \ll K$. EXECUTION TIMES (IN SECONDS) ARE PROVIDED FOR $N = 10$.

N	Regularized Tyler	GLASSO	GlassoGaussQn	Proposed
10	0.80 (**0.08s**)	0.77 (60.18s)	0.76 (34.85s)	**0.64** (0.95s)
20	0.88	0.74	0.76	**0.61**
40	0.65	0.48	0.48	**0.42**
60	0.78	0.51	0.48	**0.44**
80	0.89	0.57	0.53	**0.44**

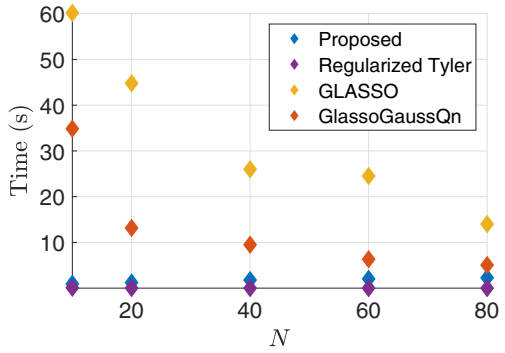

Figure 5. Execution times (in seconds) for the experiments with $N \ll K$.

D. High-Dimensional Case: $N \ll K$

For the higher dimensional experiments, we run $N_{\text{MC}} = 10^3$ Monte Carlo runs for $K = 100$ and N-values between 10 and 80. We focus on an experimental scenario for which sparse regularization is suitable. A uniformly sparse precision matrix is generated with approximately 90% of the off-diagonal elements equal 0. The sparse regularization parameters providing the best MSEs are used for all methods, and we use the same values as before for all other parameters of the proposed method. Table III shows the obtained MSEs for the precision matrices estimated with the regularized Tyler, GLASSO, GlassoGaussQn, and proposed methods. Recall that GLASSO, its robust version GlassoGaussQn, and the proposed method employ a sparse ℓ_1-regularization of the precision matrix (see (46)). This could explain the gap in performance between these methods and the regularized Tyler's method. The performance of GLASSO is similar to its robust version GlassoGaussQn, although slightly improved for higher N-values. We can see that the proposed method results in the smallest errors for all experiments.

Finally, the average execution times per Monte Carlo run are displayed in Figure. 5 for different values of N. The times corresponding to $N = 10$ are also reported in Table III. [9] For all values of N, the regularized Tyler is the fastest method, while GLASSO is the most time-consuming. The proposed method is notably faster than GLASSO and GlassoGaussQn, but remains slower than Tyler's. However, it is worth noting that Tyler's method is the least accurate estimator in this experiment, where regularization is clearly advantageous. Notice also that the times associated with the proposed and Tyler's methods increase slightly with sample sizes. Nonetheless, they are less affected across different experiments when compared to GLASSO and GlassoGaussQn, which converge much slower for small N-values.

VI. CONCLUSION

This paper has introduced a novel robust Bayesian-like estimator for the perturbed multivariate Generalized Gaussian distribution, which is based on a convex parameter estimation method. This optimization-based approach judiciously combines a reparametrization of the original likelihood with regularization. Unlike existing methods, theoretical convergence guarantees are obtained for all parameters in the perturbed case, and experiments show improvements in performance in various simulation scenarios, including the high-dimensional case. Sparse precision matrix estimation is notably improved in accuracy and speed compared to regularized robust and non-robust plugin methods. These improvements can be useful for various (high-dimensional) machine learning and graph processing applications. Forthcoming work will investigate such applications in the context of real noisy data.

APPENDICES

A. Proof of Proposition III.1

Let us define again the following functions:

$$\Phi\colon (\mathbb{R}^K)^N \times \mathbb{R}^N\colon ((\mathbf{v}_n)_{1\leqslant n\leqslant N}, \boldsymbol{\theta}) \mapsto \sum_{n=1}^{N} \varphi(\mathbf{v}_n, \theta_n), \quad (49)$$

$$\varphi\colon \mathbb{R}^K \times \mathbb{R}\colon (\mathbf{u}, \xi) \mapsto \begin{cases} \dfrac{\|\mathbf{u}\|^\beta}{2\xi^{\beta-1}} & \text{if } \mathbf{u} \neq \mathbf{0} \text{ and } \xi > 0 \\ 0 & \text{if } \mathbf{u} = \mathbf{0} \text{ and } \xi = 0 \\ +\infty & \text{otherwise,} \end{cases} \quad (50)$$

and the linear operator

$$\mathcal{T}\colon \mathcal{S}_K \times \mathbb{R}^K \to (\mathbb{R}^K)^N\colon (\mathbf{Q}, \mathbf{m}) \mapsto (\mathbf{Q}\mathbf{y}_n - \mathbf{m})_{1\leqslant n\leqslant N}. \quad (51)$$

Function φ fits the form of a so-called perspective function as defined in [43, Example 9.43]. Thus, function Φ defined in (49) belongs to $\Gamma_0\big((\mathbb{R}^K)^N \times \mathbb{R}^N\big)$ and the function

$$(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) \mapsto \Phi\big(\mathcal{T}(\mathbf{Q}, \mathbf{m}), \boldsymbol{\theta}\big),$$

[9] All codes are implemented in MATLAB and run on a computer with an Intel(R) Xeon(R) W-2235 CPU processor @3.8MHz using 251GB of RAM, running Linux v5.4.0.

which corresponds to the composition of a linear operator with a convex function is also proper, lower-semicontinuous, and convex. In addition,

$$(\forall \mathbf{Q} \in \mathcal{S}_K) \quad \Psi(\mathbf{Q}) = N \sum_{k=1}^{K} \psi(\sigma_k), \tag{52}$$

where $(\sigma_k)_{1 \leqslant k \leqslant K}$ are the eigenvalues of $\mathbf{Q}$.r Ψ is a spectral function (*i.e.*, a symmetric function of the eigenvalues of its matrix argument in $\mathcal{S}_K$), which is associated with $\psi \in \Gamma_0(\mathbb{R})$. Therefore $\Psi \in \Gamma_0(\mathcal{S}_K)$.

The considered cost function can be re-expressed as

$$f(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) = \Phi\big(\mathcal{T}(\mathbf{Q}, \mathbf{m}), \boldsymbol{\theta}\big) + \Psi(\mathbf{Q}) + g_{\mathsf{Q}}(\mathbf{Q}) + g_{\mathsf{m}}(\mathbf{m}) + \widetilde{g}_{\vartheta}(\boldsymbol{\theta}) \tag{53}$$

and it follows from the previous observations on the two first functions $\Phi\big(\mathcal{T}\cdot, \cdot\big)$ and Ψ, and the assumptions made on the three last ones that f is lower-semicontinuous and convex. Then, it is left to show that the domain of f is non-empty, *i.e.,* f is proper. First, $\varnothing \neq \operatorname{dom} \widetilde{g}_{\vartheta}$ and, according to (15), $\widetilde{g}_{\vartheta}(\boldsymbol{\theta})$ is not defined if $\boldsymbol{\theta} \notin]0, +\infty[^K$, which means that $\operatorname{dom} \widetilde{g}_{\vartheta} \subset]0, +\infty[^N$. Since we have assumed that g_{Q} is also finite at least for one symmetric positive definite matrix, f is proper. Thus, we have shown that f is a proper lower-semicontinuous convex function on $\mathcal{S}_K \times \mathbb{R}^K \times \mathbb{R}^N$.

B. Proof of Proposition III.2

Having shown that f is proper, lower-semicontinuous and convex, we will prove that a minimizer exists and provide conditions for its uniqueness.

Since $\|\cdot\|^\beta$ with $\beta > 1$ is a strictly convex function, for given values of $(\mathbf{Q}, \boldsymbol{\theta}) \in \mathcal{S}_K^{++} \times]0, +\infty[^N$,

$$\mathbf{m} \mapsto \Phi\big(\mathcal{T}(\mathbf{Q}, \mathbf{m}), \boldsymbol{\theta}\big) + g_{\mathsf{m}}(\mathbf{m}) \tag{54}$$

is a strictly convex function. Assumption (i) yields

$$(\forall \mathbf{m} \in \mathbb{R}^K) \quad \Phi\big(\mathcal{T}(\mathbf{Q}, \mathbf{m}), \boldsymbol{\theta}\big) + g_{\mathsf{m}}(\mathbf{m}) \geqslant \frac{1}{2} \sum_{n=1}^{N} \frac{\|\mathbf{Q}\mathbf{y}_n - \mathbf{m}\|^\beta}{\theta_n^{\beta-1}}. \tag{55}$$

This allows us to deduce from the coercivity of $\|\cdot\|^\beta$ that the function in (54) is a coercive lower-semicontinuous strictly convex function. Its infimum is thus reached for a unique vector $\widehat{\mathbf{m}}(\mathbf{Q}, \boldsymbol{\theta})$.

Let function Θ be defined as

$$(\forall (\mathbf{Q}, \boldsymbol{\theta}) \in \mathcal{S}_K \times \mathbb{R}^N) \quad \Theta(\mathbf{Q}, \boldsymbol{\theta}) = \inf_{\mathbf{m} \in \mathbb{R}^K} \Phi\big(\mathcal{T}(\mathbf{Q}, \mathbf{m}), \boldsymbol{\theta}\big) + g_{\mathsf{m}}(\mathbf{m}) + \Psi(\mathbf{Q}) + \widetilde{g}_{\vartheta,1}(\boldsymbol{\theta}). \tag{56}$$

It follows from [43, Proposition 8.35] that the marginal function $(\mathbf{Q}, \boldsymbol{\theta}) \mapsto \inf_{\mathbf{m} \in \mathbb{R}^K} \Phi\big(\mathcal{T}(\mathbf{Q}, \mathbf{m}), \boldsymbol{\theta}\big) + g_{\mathsf{m}}(\mathbf{m})$ is convex and from the convexity of the last two terms that Θ is a convex function. In addition, by incorporating the infimum $\widehat{\mathbf{m}}(\mathbf{Q}, \boldsymbol{\theta})$, and by using Definition (17) and Assumption (ii), for every $(\mathbf{Q}, \boldsymbol{\theta}) \in \mathcal{S}_K^{++} \times]0, +\infty[^N$ we have

$$\Theta(\mathbf{Q}, \boldsymbol{\theta}) = \Phi\big(\mathcal{T}(\mathbf{Q}, \widehat{\mathbf{m}}(\mathbf{Q}, \boldsymbol{\theta})), \boldsymbol{\theta}\big) + g_{\mathsf{m}}(\widehat{\mathbf{m}}(\mathbf{Q}, \boldsymbol{\theta})) + \Psi(\mathbf{Q}) + \widetilde{g}_{\vartheta,1}(\boldsymbol{\theta}) < +\infty. \tag{57}$$

This shows that the domain of Θ is the open set $\mathcal{S}_K^{++} \times]0, +\infty[^N$. It thus follows from [43, Corollary 8.39] that Θ is continuous on this domain. In addition let $(\check{\mathbf{Q}}, \check{\boldsymbol{\theta}}) \in \mathcal{S}_K^{+} \times [0, +\infty[^N$ be a point on the border of $\operatorname{dom} \Theta$. One of the eigenvalues of $\check{\mathbf{Q}}$ or one of the components of $\check{\boldsymbol{\theta}}$ is thus equal to zero. As a consequence of the form of Ψ in (17) and Assumption (ii),

$$\lim_{\substack{(\mathbf{Q}, \boldsymbol{\theta}) \to (\check{\mathbf{Q}}, \check{\boldsymbol{\theta}}) \\ (\mathbf{Q}, \boldsymbol{\theta}) \in \operatorname{dom} \Theta}} \Theta(\mathbf{Q}, \boldsymbol{\theta}) \geqslant \lim_{\substack{(\mathbf{Q}, \boldsymbol{\theta}) \to (\check{\mathbf{Q}}, \check{\boldsymbol{\theta}}) \\ \mathbf{Q} \in \mathcal{S}_K^{++}, \boldsymbol{\theta} \in]0, +\infty[^N}} \Psi(\mathbf{Q}) + \widetilde{g}_{\vartheta,1}(\boldsymbol{\theta}) = +\infty. \tag{58}$$

It thus follows from [43, Proposition 9.33] that Θ is a proper lower-semicontinuous convex function. Besides,

$$\inf_{\mathbf{Q} \in \mathcal{S}_K, \mathbf{m} \in \mathbb{R}^K, \boldsymbol{\theta} \in \mathbb{R}^N} f(\mathbf{Q}, \mathbf{m}, \boldsymbol{\theta}) = \inf_{\mathbf{Q} \in \mathcal{S}_K, \boldsymbol{\theta} \in \mathbb{R}^N} \Theta(\mathbf{Q}, \boldsymbol{\theta}) + g_{\mathsf{Q}}(\mathbf{Q}) + \widetilde{g}_{\vartheta,0}(\boldsymbol{\theta}). \tag{59}$$

The function

$$(\mathbf{Q}, \boldsymbol{\theta}) \mapsto \Theta(\mathbf{Q}, \boldsymbol{\theta}) + g_{\mathsf{Q}}(\mathbf{Q}) + \widetilde{g}_{\vartheta,0}(\boldsymbol{\theta}) \tag{60}$$

is lower-semicontinuous and convex as a finite sum of lower-semicontinuous convex functions. It is also proper since, according to the assumptions in Proposition III.1 and Assumption (ii), there exists at least one point $(\overline{\mathbf{Q}}, \overline{\boldsymbol{\theta}}) \in \mathcal{S}_K^{++} \times]0, +\infty[^N$ such that $g_{\mathsf{Q}}(\overline{\mathbf{Q}})$, $+\widetilde{g}_{\vartheta,0}(\overline{\boldsymbol{\theta}})$ is finite. In addition, for every $(\mathbf{Q}, \boldsymbol{\theta}) \in \mathcal{S}_K \times]0, +\infty[^N$,

$$\Theta(\mathbf{Q}, \boldsymbol{\theta}) + g_{\mathsf{Q}}(\mathbf{Q}) + \widetilde{g}_{\vartheta,0}(\boldsymbol{\theta}) \geqslant \widetilde{g}_{\mathsf{Q}(\mathbf{Q})} + \widetilde{g}_{\vartheta}(\boldsymbol{\theta}). \tag{61}$$

Using now Assumption (iii), we deduce that function (60) is coercive. Since it is a coercive proper lower-semicontinuous convex function, it admits a minimizer $(\widehat{\mathbf{Q}}, \widehat{\boldsymbol{\theta}})$. Hence, a minimizer of f is $\big(\widehat{\mathbf{Q}}, \widehat{\mathbf{m}}(\widehat{\mathbf{Q}}, \widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}\big)$. In addition, if $\widetilde{g}_{\mathsf{Q}}$ and $\widetilde{g}_{\vartheta}$ are strictly convex, function (60) is strictly convex, which ensures that it has a unique minimizer.

C. Study of Function $\overline{f}$

The first and second-order derivatives of $\overline{f}$ in (30) read

$$(\forall \theta_n \in]0, +\infty[) \quad \overline{f}'(\theta_n) = \frac{1}{\theta_n}\left(1 - \Big(\frac{\overline{\theta}_n}{\theta_n}\Big)^{\beta-1} + \overline{\kappa}(\theta_n^\alpha - 1)\right) \quad \overline{f}''(\theta_n) = \frac{1}{\theta_n^2}\left(\beta\Big(\frac{\overline{\theta}_n}{\theta_n}\Big)^{\beta-1} + (\alpha-1)\overline{\kappa}\theta_n^\alpha + \overline{\kappa} - 1\right). \tag{62}$$

The second derivative being positive, $\overline{f}$ is a strictly convex function. Since $\overline{f}(\theta_n) \to +\infty$ as $\theta_n \to 0$ and $\theta_n \to +\infty$, $\overline{f}$ has a unique minimizer. This minimizer is a function of $\overline{\theta}_n$, $\overline{\kappa}$, α, and β but, for simplicity's sake, we will not make it

explicit in our notation. The minimizer $\widehat{\theta}_n$ satisfies the first-order optimality condition

$$1 - \Big(\frac{\overline{\theta}_n}{\widehat{\theta}_n}\Big)^{\beta-1} + \overline{\kappa}(\widehat{\theta}_n^{\alpha} - 1) = 0, \tag{63}$$

and we have

$$\overline{f}'(1) = 1 - \overline{\theta}_n^{\beta-1} \tag{64}$$

$$\overline{f}'(\overline{\theta}_n) = \frac{\overline{\kappa}}{\overline{\theta}_n}(\overline{\theta}_n^{\alpha} - 1). \tag{65}$$

We deduce that, if $\overline{\theta}_n > 1$, then $\overline{f}'(1) < 0$ and $\overline{f}'(\overline{\theta}_n) > 0$, which shows that $\widehat{\theta}_n \in]1, \overline{\theta}_n[$. Similarly, if $\overline{\theta}_n < 1$, then $\widehat{\theta}_n \in]\overline{\theta}_n, 1[$. In addition, by implicit derivation of (63),

$$q(\widehat{\theta}_n)\frac{\partial\widehat{\theta}_n}{\partial\overline{\theta}_n} = (\beta-1)\overline{\theta}_n^{\beta-2}\widehat{\theta}_n^{1-\beta} \tag{66}$$

$$q(\widehat{\theta}_n)\frac{\partial\widehat{\theta}_n}{\partial\alpha} = -\overline{\kappa}\widehat{\theta}_n^{\alpha}\log\widehat{\theta}_n \tag{67}$$

$$q(\widehat{\theta}_n)\frac{\partial\widehat{\theta}_n}{\partial\overline{\kappa}} = 1 - \widehat{\theta}_n^{\alpha} \tag{68}$$

$$q(\widehat{\theta}_n)\frac{\partial\widehat{\theta}_n}{\partial\beta} = \ln\Big(\frac{\overline{\theta}_n}{\widehat{\theta}_n}\Big)\Big(\frac{\overline{\theta}_n}{\widehat{\theta}_n}\Big)^{\beta-1} \tag{69}$$

where $q(\widehat{\theta}_n) = (\beta-1)\overline{\theta}_n^{\beta-1}\widehat{\theta}_n^{-\beta} + \overline{\kappa}\alpha\widehat{\theta}_n^{\alpha-1}$. Since $\partial\widehat{\theta}_n/\partial\overline{\theta}_n > 0$, $\widehat{\theta}_n$ is an increasing function of $\overline{\theta}_n$. If $\overline{\theta}_n > 1$, then $\partial\widehat{\theta}_n/\partial\alpha < 0$, $\partial\widehat{\theta}_n/\partial\overline{\kappa} < 0$, and $\partial\widehat{\theta}_n/\partial\beta > 0$, which shows that $\widehat{\theta}_n$ decreases w.r.t. α and $\overline{\kappa}$, and increases w.r.t. β. Conversely, when $\overline{\theta}_n < 1$, $\widehat{\theta}_n$ is an increasing function w.r.t. α and $\overline{\kappa}$, and a decreasing one with respect to β.

Note that, since $\widehat{\theta}_n$ increases w.r.t. $\overline{\theta}_n$, $\widehat{\theta}_n \to \widehat{\theta}_\infty$ as $\overline{\theta}_n \to +\infty$, where the limit $\widehat{\theta}_\infty$ is either finite or $+\infty$. Eq. (63) allows us to discard the former case and to conclude that $\lim_{\overline{\theta}_n\to+\infty}\widehat{\theta}_n = +\infty$.

D. Properties of Operators $\mathbf{L}_1$ and $\mathbf{L}_2$

From the definitions of $\mathbf{L}_1$, $\mathbf{L}_2$, and the norms equipping the primal and dual spaces $\mathcal{H}$, $\mathcal{G}_1$, and $\mathcal{G}_2$, we deduce that

$$\|\mathbf{L}_1\|_{\mathrm{S}} = \sup_{\mathbf{p}\in\mathcal{H}\setminus\{0\}} \frac{\|\mathbf{L}_1\mathbf{p}\|_{\mathcal{G}_1}}{\|\mathbf{p}\|_{\mathcal{H}}} = \max\{\|\mathcal{T}\|_{\mathrm{S}}, \sqrt{\omega_1}\} \tag{70}$$

$$\|\mathbf{L}_2\|_{\mathrm{S}} = \sup_{\mathbf{p}\in\mathcal{H}\setminus\{0\}} \frac{\|\mathbf{L}_2\mathbf{p}\|_{\mathcal{G}_2}}{\|\mathbf{p}\|_{\mathcal{H}}} = \max\{1, \sqrt{\omega_2}\}. \tag{71}$$

Let us now evaluate the norm of $\mathcal{T}$. For every $\mathbf{Q} \in \mathcal{S}_K$ and $\mathbf{m} \in \mathbb{R}^K$,

$$\begin{aligned}\|\mathcal{T}(\mathbf{Q},\mathbf{m})\|^2 &= \sum_{n=1}^{N} \|\mathbf{Q}\mathbf{y}_n - \mathbf{m}\|^2 \\ &= \sum_{n=1}^{N} \Big\|[\mathbf{Q} \ -\mathbf{m}]\begin{bmatrix}\mathbf{y}_n\\ 1\end{bmatrix}\Big\|^2 \\ &= \mathrm{tr}([\mathbf{Q} \ -\mathbf{m}]\mathbf{Y}\mathbf{Y}^\top[\mathbf{Q} \ -\mathbf{m}]^\top),\end{aligned} \tag{72}$$

where $\mathbf{Y}$ is given by (44). We also have

$$\begin{aligned}&[\mathbf{Q} \ -\mathbf{m}]\mathbf{Y}\mathbf{Y}^\top[\mathbf{Q} \ -\mathbf{m}]^\top \\ &\preceq \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{S}}[\mathbf{Q} \ -\mathbf{m}][\mathbf{Q} \ -\mathbf{m}]^\top \\ &= \|\mathbf{Y}\|_{\mathrm{S}}^2[\mathbf{Q} \ -\mathbf{m}][\mathbf{Q} \ -\mathbf{m}]^\top.\end{aligned} \tag{73}$$

Combining (72) and (73) yields

$$\begin{aligned}\|\mathcal{T}(\mathbf{Q},\mathbf{m})\|^2 &\leqslant \|\mathbf{Y}\|_{\mathrm{S}}^2\,\mathrm{tr}([\mathbf{Q} \ -\mathbf{m}][\mathbf{Q} \ -\mathbf{m}]^\top) \\ &= \|\mathbf{Y}\|_{\mathrm{S}}^2\|[\mathbf{Q} \ -\mathbf{m}]\|_{\mathrm{F}}^2 \\ &= \|\mathbf{Y}\|_{\mathrm{S}}^2(\|\mathbf{Q}\|_{\mathrm{F}}^2 + \|\mathbf{m}\|^2).\end{aligned} \tag{74}$$

This shows that $\|\mathcal{T}\|_{\mathrm{S}} \leqslant \|\mathbf{Y}\|_{\mathrm{S}}$. Using (70) and (71), we deduce that (43) is a sufficient condition for (42) to be satisfied.

The adjoint of $\mathbf{L}_1$ is

$$\mathbf{L}_1^*\colon \mathcal{G}_1 \to \mathcal{H}\colon \big((\mathbf{u}_n)_{1\leqslant n\leqslant N}, \boldsymbol{\theta}\big) \mapsto (\mathbf{Q}, \mathbf{m}, \omega_1\boldsymbol{\theta}) \tag{75}$$

with $(\mathbf{Q},\mathbf{m}) = \mathcal{T}^*\big((\mathbf{u}_n)_{1\leqslant n\leqslant N}\big)$, where $\mathcal{T}^*$ is the adjoint of $\mathcal{T}$. The adjoint of $\mathcal{T}$ can be deduced from the identity

$$\begin{aligned}&(\forall \mathbf{Q}\in\mathcal{S}_K)(\forall \mathbf{m}\in\mathbb{R}^K)(\forall (\mathbf{u}_n)_{1\leqslant n\leqslant N} \in (\mathbb{R}^K)^N) \\ &\quad \sum_{n=1}^{N} \mathbf{u}_n^\top[\mathcal{T}(\mathbf{Q},\mathbf{m})]_n \\ &\quad = \sum_{n=1}^{N} \mathbf{u}_n^\top(\mathbf{Q}\mathbf{y}_n - \mathbf{m}) \\ &\quad = \mathrm{tr}\Big(\mathbf{Q}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{u}_n^\top\Big) - \mathbf{m}^\top\sum_{n=1}^{N}\mathbf{u}_n,\end{aligned} \tag{76, 77}$$

which shows that

$$\mathcal{T}^*((\mathbf{u}_n)_{1\leqslant n\leqslant N}) = \Big(\frac{1}{2}\sum_{n=1}^{N}(\mathbf{u}_n\mathbf{y}_n^\top + \mathbf{y}_n\mathbf{u}_n^\top), -\sum_{n=1}^{N}\mathbf{u}_n\Big). \tag{78}$$

In turn, the adjoint of $\mathbf{L}_2$ is simply expressed as

$$\mathbf{L}_2^*\colon (\mathbf{Q},\boldsymbol{\theta}) \mapsto (\mathbf{Q}, \mathbf{0}, \omega_2\boldsymbol{\theta}). \tag{79}$$

E. Proximity Operators

The expressions of the proximity operators of the functions involved in (53), up to a positive scaling parameter γ, are provided below.[10]

- Function ψ: For every $\xi \in \mathbb{R}$,

$$\mathrm{prox}_{\gamma\psi}(\xi) = \frac{\xi + \sqrt{\xi^2 + 4\gamma}}{2}. \tag{80}$$

- Function Ψ: For every $\mathbf{Q} \in \mathcal{S}_K$, let us perform the eigenvalue decomposition of $\mathbf{Q}$ as $\mathbf{U}\,\mathrm{Diag}(\sigma_1,\ldots,\sigma_K)\mathbf{U}^\top$ where $(\sigma_k)_{1\leqslant k\leqslant K}$ are the eigenvalues of $\mathbf{Q}$ and the columns of $\mathbf{U}$ form the associated orthonormal basis of eigenvectors. Then,

$$\begin{aligned}&\mathrm{prox}_{\gamma\Psi}(\mathbf{Q}) \\ &\quad = \mathbf{U}\,\mathrm{Diag}\big(\mathrm{prox}_{N\gamma\psi}(\sigma_1),\ldots,\mathrm{prox}_{N\gamma\psi}(\sigma_K)\big)\mathbf{U}^\top.\end{aligned} \tag{81}$$

- Function Φ: As shown by (49), Φ is a separable function of the components of its arguments. Thus,

$$\begin{aligned}&\big(\forall (\mathbf{v}_n)_{1\leqslant n\leqslant N} \in (\mathbb{R}^K)^N\big)(\forall\boldsymbol{\theta}\in\mathbb{R}^N) \\ &\mathrm{prox}_{\gamma\Phi}((\mathbf{v}_n)_{1\leqslant n\leqslant N},\boldsymbol{\theta}) = \big(\mathrm{prox}_{\gamma\varphi}(\mathbf{v}_n,\theta_n)\big)_{1\leqslant n\leqslant N}.\end{aligned} \tag{82}$$

[10] See http://proximity-operator.net.

The expression of the proximity operator of the perspective function φ has been derived in [44, Example 3.7]. Let

$$\beta^* = \frac{\beta}{\beta-1}, \qquad \varrho = \left(2\Big(1-\frac{1}{\beta^*}\Big)\right)^{\beta^*-1} \tag{83}$$

and, for every $\mathbf{u} \in \mathbb{R}^K \setminus \{\mathbf{0}\}$ and $\xi \in \mathbb{R}$ such that $\beta^*\gamma^{\beta^*-1}\xi + \varrho\|\mathbf{u}\|^{\beta^*} > 0$, let

$$\mathbf{p}(\mathbf{u},\xi) = \frac{t(\mathbf{u},\xi)}{\|\mathbf{u}\|}\mathbf{u} \tag{84}$$

where $t(\mathbf{u},\xi)$ is the unique solution in $]0,+\infty[$ of the equation

$$s^{2\beta^*-1} + \frac{\beta^*\xi}{\gamma\varrho}s^{\beta^*-1} + \frac{\beta^*}{\varrho^2}s - \frac{\beta^*}{\gamma\varrho^2}\|\mathbf{u}\| = 0. \tag{85}$$

Then, for every $\mathbf{u} \in \mathbb{R}^K$ and $\xi \in \mathbb{R}$,

$$\operatorname{prox}_{\gamma\varphi}(\mathbf{u},\xi) = \tag{86}$$

$$\begin{cases} \Big(\mathbf{u} - \gamma\mathbf{p}(\mathbf{u},\xi), \xi + \dfrac{\gamma\varrho}{\beta^*}t(\mathbf{u},\xi)^{\beta^*}\Big) & \text{if } \mathbf{u} \neq \mathbf{0} \text{ and} \\ & \beta^*\gamma^{\beta^*-1}\xi + \varrho\|\mathbf{u}\|^{\beta^*} > 0 \\ (\mathbf{0},\xi) & \text{if } \mathbf{u} = \mathbf{0} \text{ and } \xi > 0 \\ (\mathbf{0},0) & \text{otherwise,} \end{cases}$$

$$= \big(\operatorname{prox}^{(1)}_{\gamma\varphi}(\mathbf{u},\xi), \operatorname{prox}^{(2)}_{\gamma\varphi}(\mathbf{u},\xi)\big). \tag{87}$$

As shown by the developments in Section IV (see (32) – (33)), one may be interested in deriving the expression of the proximity operator in a metric that weights the two variables in an unbalanced manner. This means that, for every $(\gamma_1,\gamma_2) \in [0,+\infty[^2$, $\mathbf{u} \in \mathbb{R}^K \setminus \{\mathbf{0}\}$ and $\xi \in \mathbb{R}$, one would seek

$$\operatorname{prox}^{\gamma_1,\gamma_2}_{\varphi}(\mathbf{u},\xi) = \operatorname*{argmin}_{\substack{\mathbf{u}\in\mathbb{R}^K,\\ \xi'\in[0,+\infty[}} \varphi(\mathbf{u}',\xi') + \frac{1}{2\gamma_1}\|\mathbf{u}'-\mathbf{u}\|^2 + \frac{1}{2\gamma_2}(\xi'-\xi)^2. \tag{88}$$

Performing the variable change $\tilde{\mathbf{u}} = \mathbf{u}'/\sqrt{\gamma_1}$ and $\tilde{\xi} = \xi'/\sqrt{\gamma_2}$ and using the definition of φ in (50) yields

$$\operatorname{prox}^{\gamma_1,\gamma_2}_{\varphi}(\mathbf{u},\xi) = \Bigg(\sqrt{\gamma_1}\operatorname{prox}^{(1)}_{\gamma\varphi}\Big(\frac{\mathbf{u}}{\sqrt{\gamma_1}}, \frac{\xi}{\sqrt{\gamma_2}}\Big), \sqrt{\gamma_2}\operatorname{prox}^{(2)}_{\gamma\varphi}\Big(\frac{\mathbf{u}}{\sqrt{\gamma_1}}, \frac{\xi}{\sqrt{\gamma_2}}\Big)\Bigg), \tag{89}$$

with $\gamma = \sqrt{\gamma_1^{\beta}/\gamma_2^{\beta-1}}$.

REFERENCES

[1] Violeta Roizman, Matthieu Jonckheere, and Frédéric Pascal, "A flexible EM-like clustering algorithm for noisy data," *arXiv preprint arXiv:1907.01660*, 2019.

[2] Pierre Houdouin, Andrew Wang, Matthieu Jonckheere, and Frédéric Pascal, "Robust classification with flexible discriminant analysis in heterogeneous data," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5717–5721.

[3] Teng Zhang and Gilad Lerman, "A novel M-estimator for robust pca," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 749–808, 2014.

[4] Antonio De Maio, Maria Sabrina Greco, et al., "Modern radar detection theory," Tech. Rep., SciTech Publishing Inc, 2015.

[5] Yiyong Feng, Daniel P. Palomar, et al., "A signal processing perspective on financial engineering," *Foundations and Trends® in Signal Processing*, vol. 9, no. 1–2, pp. 1–231, 2016.

[6] Frédéric Pascal, Lionel Bombrun, Jean-Yves Tourneret, and Yannick Berthoumieu, "Parameter estimation for Multivariate Generalized Gaussian Distributions," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5960–5971, 2013.

[7] Ricardo A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Stat.*, vol. 5, no. 1, pp. 51–67, 1976.

[8] Esa Ollila, David E. Tyler, Visa Koivunen, and H. Vincent Poor, "Complex elliptically symmetric distributions: survey, new results and applications," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5597–5625, 2012.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, July 2008.

[10] Esa Ollila and David E. Tyler, "Regularized M-estimators of scatter matrix," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 6059–6070, 2014.

[11] Frédéric Pascal, Yacine Chitour, and Yihui Quek, "Generalized robust shrinkage estimator and its application to stap detection problem," *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5640–5651, 2014.

[12] Ying Sun, Prabhu Babu, and Daniel P. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5143–5156, 2014.

[13] Ginette Lafit, Javier Nogales, Marcelo Ruiz, and Ruben Zamar, *Robust and Sparse Estimation of Graphical Models Based on Multivariate Winsorization*, pp. 249–275, Springer International Publishing, Cham, 2023.

[14] Garth Tarr, Samuel Müller, and Neville C Weber, "Robust estimation of precision matrices under cellwise contamination," *Computational Statistics & Data Analysis*, vol. 93, pp. 404–420, 2016.

[15] Viktoria Öllerer and Christophe Croux, *Robust High-Dimensional Precision Matrix Estimation*, pp. 325–350, Springer International Publishing, Cham, 2015.

[16] Ami Wiesel, "Geodesic convexity and covariance estimation," *IEEE transactions on signal processing*, vol. 60, no. 12, pp. 6182–6189, 2012.

[17] Teng Zhang, Ami Wiesel, and Maria Sabrina Greco, "Multivariate generalized gaussian distribution: Convexity and graphical models," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4141–4148, 2013.

[18] Eusebio Gómez, MA Gomez-Viilegas, and J Miguel Marín, "A multivariate generalization of the power exponential family of distributions," *Communications in Statistics-Theory and Methods*, vol. 27, no. 3, pp. 589–600, 1998.

[19] Nitin Kumar and Suyash P Awate, "Semi-supervised robust mixture models in rkhs for abnormality detection in medical images," *IEEE Transactions on Image Processing*, vol. 29, pp. 4772–4787, 2020.

[20] Qilong Wang, Peihua Li, Qinghua Hu, Pengfei Zhu, and Wangmeng Zuo, "Deep global generalized Gaussian networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5080–5088.

[21] Li Chai, Jun Du, Qing-Feng Liu, and Chin-Hui Lee, "Using generalized gaussian distributions to improve regression error modeling for deep learning-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1919–1931, 2019.

[22] Nora Ouzir, Jean-Christophe Pesquet, and Frédéric Pascal, "A convex formulation for the robust estimation of multivariate exponential power models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5772–5776.

[23] Nikos Komodakis and Jean-Christophe Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, 2015.

[24] Patrick L. Combettes and Jean-Christophe Pesquet, "Fixed point strategies in data science," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3878–3905, 2021.

[25] Kung Yao, "A representation theorem and its applications to spherically-invariant random processes," *IEEE Transactions on Information Theory*, vol. 19, no. 5, pp. 600–608, 1973.

[26] Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan, *Continuous univariate distributions, volume 2*, vol. 289, John wiley & sons, 1995.
[27] Frédéric Pascal, Yacine Chitour, Jean-Philippe Ovarlez, Philippe Forster, and Pascal Larzabal, "Covariance structure maximum-likelihood estimates in compound gaussian noise: Existence and algorithm analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 34–48, 2007.
[28] Yacine Chitour and Frédéric Pascal, "Exact maximum likelihood estimates for sirv covariance matrix: Existence and algorithm analysis," *IEEE Transactions on signal processing*, vol. 56, no. 10, pp. 4563–4573, 2008.
[29] Yilun Chen, Ami Wiesel, and Alfred O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4097–4107, 2011.
[30] David E. Tyler, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, 1987.
[31] David E. Tyler, "A distribution-free M-estimator of multivariate scatter," *The annals of Statistics*, pp. 234–251, 1987.
[32] Joana Frontera-Pons, Miguel Angel Veganzones, Frédéric Pascal, and Jean-Philippe Ovarlez, "Hyperspectral anomaly detectors using robust estimators," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 720–731, 2015.
[33] Frédéric Pascal, Philippe Forster, Jean-Philippe Ovarlez, and Pascal Larzabal, "Performance analysis of covariance matrix estimates in impulsive noise," *IEEE Transactions on signal processing*, vol. 56, no. 6, pp. 2206–2217, 2008.
[34] Ami Wiesel and Teng Zhang, "Structured robust covariance estimation," *Foundations and Trends® in Signal Processing*, vol. 8, no. 3, pp. 127–216, 2015.
[35] John T. Kent and David E. Tyler, "Redescending m-estimates of multivariate location and scatter," *The Annals of Statistics*, vol. 19, no. 4, pp. 2102–2119, 1991.
[36] Karnran Sharifi and Alberto Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.
[37] Kai-Sheng Song, "A globally convergent and consistent method for estimating the shape parameter of a generalized gaussian distribution," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 510–527, 2006.
[38] J. Armando Dominguez-Molina, Graciela González-Farías, Ramón M Rodríguez-Dagnino, and ITESM Campus Monterrey, "A practical procedure to estimate the shape parameter in the generalized gaussian distribution," *technique report I-01-18_eng. pdf, available through http://www. cimat. mx/reportes/enlinea/I-01-18_eng. pdf*, vol. 1, 2003.
[39] Robert Krupiński and Jan Purczyński, "Approximated fast estimator for the shape parameter of generalized gaussian distribution," *Signal processing*, vol. 86, no. 2, pp. 205–211, 2006.
[40] Frédéric Pascal, Esa Ollila, and Daniel P. Palomar, "Improved estimation of the degree of freedom parameter of multivariate t-distribution," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 860–864.
[41] Douglas A Abraham and Anthony P Lyons, "Reliable methods for estimating the k-distribution shape parameter," *IEEE Journal of Oceanic Engineering*, vol. 35, no. 2, pp. 288–302, 2010.
[42] Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
[43] Heinz Bauschke and Patrick L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Space*, CMS Books in Mathematics. Springer International Publishing, 2017.
[44] Patrick L. Combettes and Christian L. Müller, "Perspective functions: Proximal calculus and applications in high-dimensional statistics," *Journal of Mathematical Analysis and Applications*, vol. 457, no. 2, pp. 1283–1306, 2018, Special Issue on Convex Analysis and Optimization: New Trends in Theory and Applications.
[45] Patrick L. Combettes and Jean-Christophe Pesquet, *Proximal Splitting Methods in Signal Processing*, pp. 185–212, Springer New York, New York, NY, 2011.
[46] Antonin Chambolle and Thomas Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, pp. 120–145, 2011.