DCA based approaches for bi-level variable selection and application for estimate multiple sparse covariance matrices

Hoai An Le Thi, Duy Nhat Phan, Tao Pham Dinh

Version of Record: https://www.sciencedirect.com/science/article/pii/S0925231221013977
Manuscript_d94b40e5c65e1035e27aa8f7ce2a5e45

DCA based approaches for Bi-level variable selection and application for estimate multiple sparse covariance matrices

Hoai An Le Thi[a,*], Duy Nhat Phan[a], Tao Pham Dinh[b]

[a] *Université de Lorraine, LGIPM, Département IA, F-57000 Metz, France*
[b] *Laboratory of Mathematics, INSA–Rouen, University of Normandie, France*

Abstract

Variable selection plays an important role in analyzing high dimensional data and is a fundamental problem in machine learning. When the data possesses certain group structures in which individual variables are also meaningful scientifically, we are naturally interested in selecting important groups as well as important variables within the selected groups. This is referred to as the bi-level variable selection which is much more complex than the selection of individual variables. In recent years, research on the topic of variable selection is very active, but the majority of the work is focused on the individual variable selection. There is therefore a need to further develop more effective approaches for bi-level variable selection. Since DC (Difference of Convex functions) programming and DCA (DC Algorithm), powerful tools in nonconvex programming framework, have been successfully investigated for individual variable selection, we believe that they could be efficiently exploited for the more difficult bi-level variable selection task. In that direction, we investigate in this work DC approximations of the mixed zero norm ($\ell_{0,0}$) and the combined norm ($\ell_0 + \ell_{q,0}$). The resulting approximate problems are then formulated as DC programs for which DCA based algorithms are introduced. As an application, these DCA schemes

*Corresponding author
Email addresses: hoai-an.le-thi@univ-lorraine.fr (Hoai An Le Thi), duy-nhat.phan@univ-lorrainre.fr (Duy Nhat Phan), pham@insa-rouen.fr (Tao Pham Dinh)

are developed for estimating multiple sparse covariance matrices sharing some common structures such as the locations or weights of non-zero elements. The experimental results on both simulated and real datasets indicate the efficiency of our algorithms.

Keywords: Bi-Level Variable Selection, Mixed norm, Nonconvex approximation, Multiple Covariance Matrices, DC Programming, DCA

1. Introduction

Variable selection plays a pivotal role in various domains such as machine learning, statistics, computational biology, signal processing and other related areas. When the data possesses certain group structures in which individual variables are also meaningful scientifically, we are naturally interested in selecting important groups as well as important variables within the selected groups. This is referred to as the bi-level variable selection. For example, in genomic data analysis, the correlations between genes sharing the biological pathway can be high. Hence these genes should be considered as a group. Moreover, we also would like to identify particularly important genes in pathways of interest. In recent years, research on the topic of variable selection is very active (see [1, 2] and references quoted therein), but the majority of the work is focused on the individual variable selection. There is therefore a need to further develop more effective approaches for bi-level variable selection. The passage from individual variable selection to bi-level variable selection is not evident, since the regularization term dealing with group sparsity is more complex, and therefore the resulting optimization problem is more difficult. Since DC (Difference of Convex functions) programming and DCA (DC Algorithm), powerful tools in nonconvex programming framework, have been successfully investigated for individual variable selection, we believe that they could be efficiently exploited for the more difficult bi-level variable selection task. In this paper, we introduce new DCA based approaches for enforcing sparsity of groups and within each group by using the $\ell_{0,0}$ regularization and $\ell_0 + \ell_{q,0}$ regularization with $q \geq 1$.

Problem statement. Before formulating the problem, let us give some basic notations.

For a vector $\mathbf{x} \in \mathbb{R}^d$ the ℓ_q norm of $\mathbf{x}$, with $1 \leq q < \infty$, is given by

$$\|\mathbf{x}\|_q = (|\mathbf{x}_1|^q + ... + |\mathbf{x}_d|^q)^{1/q},$$

while $\|\mathbf{x}\|_\infty$ is defined as $\max(|\mathbf{x}_1|, ..., |\mathbf{x}_d|)$.

The ℓ_0 norm of $\mathbf{x}$, denoted by $\|\mathbf{x}\|_0$, is the number of non-zero components of $\mathbf{x}$. Let $s : \mathbb{R} \to \mathbb{R}$ be the step function given by $s(t) = 1$ if $t \neq 0$ and $s(t) = 0$ otherwise. Then the ℓ_0 norm of $\mathbf{x}$ can be expressed as

$$\|\mathbf{x}\|_0 = \sum_{i=1}^{d} s(\mathbf{x}_i).$$

The ℓ_0-norm is an important concept for modelling the sparsity of data and plays a crucial role in optimization problems where one has to select representative variables. The function ℓ_0, apparently very simple, is lower-semicontinuous on $\mathbb{R}^d$, but its discontinuity at the origin makes nonconvex programs involving $\|.\|_0$ challenging.

Assume that the elements $\mathbf{x}_1, ..., \mathbf{x}_d$ of $\mathbf{x}$ are partitioned into J non-overlapping groups $\mathbf{x}^{(j)} \in \mathbb{R}^{d_j}$ for $j = 1, ..., J$. The mixed $\ell_{q,0}$ norm of $\mathbf{x}$, for $q \geq 0$, denoted by $\|\mathbf{x}\|_{q,0}$, is the number of groups whose ℓ_q norm is non-zero:

$$\|\mathbf{x}\|_{q,0} = \sum_{j=1}^{J} s\left(\|\mathbf{x}^{(j)}\|_q\right),$$

in particular

$$\|\mathbf{x}\|_{0,0} = \sum_{j=1}^{J} s(\sum_{i=1}^{d_j} s(\mathbf{x}_i^{(j)})).$$

Whilst the ℓ_0 norm encourages individual sparsity, the $\ell_{q,0}$ norm supports group sparsity, i.e., it can selects important groups. Especially, the $\ell_{0,0}$ norm promotes both individual and group sparsity: $\ell_{0,0}$ is a composite function of the step functions in which the inner step function relates to the selection of individual variables while the outer step function selects groups of variables. Hence, the two most natural ways for enforcing simultaneously sparsity of groups and within

each group is, either combining the ℓ_0 norm and $\ell_{q,0}$ norm with $q \geq 1$, or using the $\ell_{0,0}$ norm.

Variable selection is often done during a learning task (e.g. classification, regression, ...). Formally, the mathematical formulation of such a learning problem takes the form

$$\min \left\{ f(\mathbf{x}) + \lambda r(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d \right\}, \tag{1}$$

where $f(\mathbf{x})$ is a loss function related to the learning task while $r(\mathbf{x})$ is a regularization term dealing with sparsity, and λ is a positive number, called the regularization parameter, that makes the trade-off between the criterion f and the sparsity.

In this paper, we consider the problem (1) in which $f(\mathbf{x}) = g(\mathbf{x}) + f_1(\mathbf{x}) - h(\mathbf{x})$, where g, h are convex functions and f_1 is a continuously differentiable function with L-Lipschitz continuous gradient, and $r(\mathbf{x})$ is the $\ell_0 + \ell_{q,0}$ regularization ($q \geq 1$) or the $\ell_{0,0}$ regularization. More precisely

$$r(\mathbf{x}) = \alpha \|\mathbf{x}\|_0 + (1-\alpha)\|\mathbf{x}\|_{q,0}, \tag{2}$$

where α is a tradeoff parameter between the two regularization terms, and alternatively,

$$r(\mathbf{x}) = \|\mathbf{x}\|_{0,0}. \tag{3}$$

Many statistical modeling problems take the form of (1) [3, 4, 5, 6, 7, 8, 9, 10], for example, the multiple linear/logistic/Cox regression, the multiple graphical model, the multiple covariance matrices estimation, and the compressed sensing, etc. In particular, let us mention an important application of bi-level variable selection corresponding to the model (1). We consider a dataset with Q classes. For the k-th class, let $\mathbf{X}^{(k)}$ be an $n_k \times d$ matrix consisting of n_k observations with the number of features d being common to all classes. Furthermore, we assume that the observations within each class are independently and identically sampled from the distribution $\mathcal{N}(0, \mathbf{\Sigma}^{(k)})$. Let $\mathbf{S}^{(k)} = \frac{1}{n_k}(\mathbf{X}^{(k)})^T\mathbf{X}^{(k)}$ be the sample covariance matrix for the k-th class. The Q covariance matrices

$\{\mathbf{\Sigma}\} = \{\mathbf{\Sigma}^{(1)}, ..., \mathbf{\Sigma}^{(Q)}\}$ are estimated via minimizing the negative log-likelihood

$$\mathcal{L}(\{\mathbf{\Sigma}\}) = \sum_{k=1}^{Q} n_k \left[\log\det \mathbf{\Sigma}^{(k)} + \mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})\right].$$

When the multiple covariance matrices share some common structures such as the locations and weights of non-zero elements, it seems reasonable to consider the elements (i, j) across all Q covariance matrices as groups. Moreover, each element within each group also has different roles in its covariance matrix. Hence, we consider the bi-level variable selection problem in estimating multiple sparse covariance matrices which takes the form of (1):

$$\min_{\mathbf{\Sigma}^{(k)} \succ 0} \{\mathcal{L}(\{\mathbf{\Sigma}\}) + r(\{\mathbf{\Sigma}\})\}. \tag{4}$$

Here the notation $\mathbf{\Sigma}^{(k)} \succ 0$ means that $\mathbf{\Sigma}^{(k)}$ is symmetric positive definite.

Our contributions. Existing approaches for bi-level variable selection can be divided into two groups: convex (sparse group lasso [3, 4, 11]) and non-convex approximation approaches (minimax concave penalty (MCP) or exponential approximation [12, 13, 14, 15], see Section 2 for more details). Our work is within the nonconvex approximation framework. We investigate DC approximation approaches for the mixed $\ell_{0,0}$ norm and the combined $\ell_0 + \ell_{q,0}$ norm with $q \geq 1$. Our contributions are multiple.

Firstly, motivated by the great success of the ℓ_0-norm for individual variable selection in recent years (e.g. [2, 16, 17, 18, 19, 20, 21, 22]), we design new models using the $\ell_0 + \ell_{q,0}$ and $\ell_{0,0}$ regularizations for bi-level variable selection. As indicated above, these regularizations are the two most natural ways to tackle simultaneously individual and group sparsity. We approximate the step function by common DC functions and show that the resulting approximations of the $\ell_0 + \ell_{q,0}$ and the $\ell_{0,0}$ regularizations can be expressed as DC functions. The idea of using a mixed norm to deal with group variable sparsity is not new in the literature. However the existing works only concern some very special problems (linear regression, logistic regression, Cox regression, multiple inverse covariance matrices) where the loss function is convex (see Section 2 for more

details). Hence, the resulting optimization in convex approximation approaches (using the convex $\ell_1+\ell_{2,1}$ regularization to approximate $\ell_0+\ell_{2,0}$) is convex and therefore it is so far easy to solve. As for the works in nonconvex approximation approaches [12, 15, 14], they can be regarded as using the MCP (resp. exponential) function to approximate the mixed $\ell_{0,0}$ (resp. $\ell_{1,0}$) norm. The limitations of these works are that they deal only with the continuous differentiable convex loss functions (multiple linear regression) and require continuous differentiability of the approximate functions on $[0;1)$. In contrast, the optimization problem studied in this paper is much more general and more difficult in a double sense. On one hand, the loss function is nonconvex, nondifferentiable, which includes a large class of learning problems. On another hand, we consider a general DC approximate function (possibly nondifferentiable) of the step function that covers all existing approximations as special cases. We study general models involving $\ell_0+\ell_{q,0}$ ($q\geq 1$) and/or $\ell_{0,0}$ in a unifying DC programming framework, with the aim of offering efficient DCA based approaches for tackling several difficult problems/real applications.

It is worth mentioning that the approximate term of $\ell_0+\ell_{q,0}$ and/or $\ell_{0,0}$ is not a *simple* DC function but a DC composite function. More precisely, the $\ell_0+\ell_{q,0}$ approximate regularization is a composite function of a convex and a DC function while the $\ell_{0,0}$ one is a composite of two DC functions. Hence, their DC structures are not evident. This is the main difference that causes more difficulties when moving from individual variable selection to bi-level variable selection. Fortunately, with a careful study, we can prove that our approximate terms are DC functions, since then DCA based algorithms can be investigated.

Secondly, we develop solution methods based on DC programming and DCA (DC Algorithms), a powerful technique in nonconvex optimization [23, 24, 25, 26, 27], for solving the resulting approximate problems. Although DCA is a standard approach, how to design an efficient DCA scheme for a concrete problem is still a challenge in nonconvex programming framework. Indeed, the general DCA scheme is rather a philosophy than an algorithm. There is not only one DCA but infinitely many DCAs for a considered problem (see

[23, 24, 25, 26, 27]). The elaboration of suitable DC decompositions such that the convex subproblems can be efficiently solved is a crucial question to be studied while using DCA. It is worth noting that the $\ell_0 + \ell_{q,0}$ and $\ell_{0,0}$ approximate regularizations are not separable in their variables, so the minimization problems involving these regularizations are more challenging than those related to the ℓ_0-norm. Exploiting special structures of the considered problems, we propose in this work appropriate DC decompositions and corresponding DCA schemes which require solving a strongly convex subproblem at each iteration. This strongly convex problem can be solved in closed form or by an inexpensive algorithm, hence the proposed algorithms are very useful in many real application problems.

Thirdly, we consider an application of the problem (4), namely the bi-level variable selection in estimating multiple sparse covariance matrices. We extend the interesting results concerning the estimation of a single sparse covariance matrix obtained in [22] to estimate multiple sparse covariance matrices. More precisely, we consider equivalent problems of (4) in which the constraints $\mathbf{\Sigma}^{(k)} \succ 0$ are replaced by $\mathbf{\Sigma}^{(k)} \succeq \delta \mathbf{I}_d$ after showing that there exists a lower bound δ of the all covariance matrices $\mathbf{\Sigma}^{(k)}$ for $k = 1, ..., Q$. Among several existing approximations of the step function, we use the Capped-ℓ_1 function [28] for implementing our algorithms. This choice is motivated by the fact that the Capped-ℓ_1 approximation has been shown to be efficient in the individual variable selection problem (see e.g. [18, 2, 21, 20, 22]). We develop three DCA based schemes for the approximate problems of the $\ell_0 + \ell_{1,0}$, $\ell_0 + \ell_{2,0}$ and $\ell_{0,0}$ regularizations. Thanks to suitable DC decompositions, the convex subproblems in these algorithms are easier to solve. Especially, in the two out of three DCAs the convex sub-problem can be decomposed into separable smaller sub-problems. We provide special convergence properties of the proposed algorithms and perform a careful empirical experiment on both simulated and real datasets to study their performances.

The rest of the paper is organized as follows. The related works on bi-level variable selection is mentioned in Section 2.

DC approximations and DCA based algorithms for bi-level variable selection using the $\ell_0 + \ell_{q,0}$ and $\ell_{0,0}$ regularizations are described in Section 3. Section 4 shows how to apply the proposed algorithms to bi-level variable selection in estimating of multiple sparse covariance matrices. The numerical experiments are reported in Section 5 and Section 6 concludes the paper.

2. Related work on bi-level variable selection

The first approach for bi-level variable selection, named sparse group lasso ($\ell_1 + \ell_{2,1}$ regularization) [3], uses the convex approximation $\ell_1 + \ell_{2,1}$ of the $\ell_0 + \ell_{2,0}$ regularization. More precisely, the sparse group lasso was added to the multiple linear regression [3] in which its convex loss function is given by

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \sum_{j=1}^{J} \mathbf{A}_j \mathbf{x}^{(j)}\|^2, \tag{5}$$

where $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A}_j \in \mathbb{R}^{n \times d_j}$, and $\mathbf{x}^{(j)} \in \mathbb{R}^{d_j}$ are vectors of coefficients for the j-th group. The authors proposed a coordinate gradient descent based algorithm to solve the convex problem (5). Later, an accelerated generalized gradient descent was developed in [11]. The authors also extended the algorithm to the sparse-group logistic regression and sparse-group Cox regression. Although these models are more complex than the multiple linear regression, they are still convex. The sparse group lasso was widely used in many other convex problems such as the multiclass classification problem [5], the estimation of multiple sparse inverse matrices [6, 7, 8, 9, 10].

The second approach deals with nonconvex approximations of the $\ell_{0,0}$ and the $\ell_{1,0}$ regularizations. In [12], the authors considered a multiple linear regression function $f(\mathbf{x})$ defined in (5) and replaced both inner and outer step functions in the $\ell_{0,0}$ regularization by the following minimax concave penalty (MCP) [13]

$$\eta_{\theta,a}(t) = \begin{cases} \theta|t| - \frac{t^2}{2a} \text{ if } |t| \leq a\theta \\ \frac{1}{2}a\theta^2 \text{ if } |t| > a\theta. \end{cases} . \tag{6}$$

Recently, Breheny [15] also considered the generalized linear regression (5) and replaced the step functions in the $\ell_{1,0}$ regularization by the following exponential approximation

$$\eta_\theta(t) = 1 - \exp(-\theta|t|). \tag{7}$$

The two mentioned studies developed local coordinate descent methods using first-order Taylor series approximations of these functions. The limitation of these methods is to requite continuous differentiability of the approximate function on $[0, \infty)$.

In [14], the authors considered a multiple linear regression function $f(\mathbf{x})$ in (5) and a group bridge regularization which simply replaced the step function in the $\ell_{1,0}$ regularization by the function $\eta(t) = t^\gamma$ for $t \geq 0$, where $0 < \gamma < 1$. The authors proposed an alternative algorithm for the considered problem.

3. DCA based algorithms for bi-level variable selection via $\ell_0 + \ell_{q,0}$ and $\ell_{0,0}$ regularizations

3.1. Outline of DC programming and DCA

DC programming and DCA constitute a quite logical and natural extension of modern convex analysis/programming to nonsmooth nonconvex analysis/programming, sufficiently large to cover most real-world nonsmooth nonconvex programs, but not too broad in order to explore/exploit the powerful arsenal of convex analysis/programming. This theoretical and algorithmic philosophy was first introduced in 1985 by Pham Dinh Tao, and widely developed by Le Thi Hoai An and Pham Dinh Tao since 1993 to become now classic and increasingly popular (see [23, 24, 25, 27] and a comprehensible review on thirty years of developments of DC programming and DCA in [26]).

This section provides a brief introduction to DC programming and DCA, which is essential for describing DCA to model and solve the problem (1). First, let us recall some basic notions of the modern convex analysis needed for the DC framework ([29]).

In the sequel, the space $\mathbb{X} = \mathbb{R}^n$ is endowed with the canonical scalar product $\langle \cdot, \cdot \rangle$. Its dual space $\mathbb{Y}$ is identified with $\mathbb{X}$ itself. Given an extended-real-valued function $\psi : \mathbb{X} \to \mathbb{R} \cup \{+\infty\}$, its effective domain dom ψ is dom $\psi := \{\mathbf{x} \in \mathbb{X} : \psi(\mathbf{x}) < +\infty\}$, and it is called proper if dom $\psi \neq \emptyset$. We say that ψ is lower semicontinuous (lsc) at $\mathbf{x} \in$ dom ψ iff $\lim_{k \to +\infty} \psi(\mathbf{x}^k) \geq \psi(\mathbf{x})$ for any sequence $\{\mathbf{x}^k\}$ converging to $\mathbf{x}$ such that the sequence $\{\psi(\mathbf{x}^k)\}$ be convergent. Throughout the paper $\Gamma_0(\mathbb{X})$ denotes the set of all lower semicontinuous proper convex functions on $\mathbb{X}$.

Let $\rho \geq 0$ and X be a nonempty convex set of dom ψ. One says that ψ is strictly convex on X if

$$\psi(\lambda \mathbf{x} + (1-\lambda)u) < \lambda \psi(\mathbf{x}) + (1-\lambda)\psi(u), \; 0 < \lambda < 1 \tag{8}$$

for any two different points $\mathbf{x}$ and u in X. The function ψ is called ρ-convex, with a nonnegative scalar ρ, on X if

$$\psi(\lambda \mathbf{x} + (1-\lambda)u) \leq \lambda \psi(\mathbf{x}) + (1-\lambda)\psi(u) - \frac{\rho}{2}\lambda(1-\lambda)\|\mathbf{x} - u\|^2, \; 0 < \lambda < 1$$

for any two different points $\mathbf{x}$ and u in X. It amounts to saying that $\psi - (\rho/2)\|\cdot\|^2$ is convex on X. The modulus of convexity of ψ on X, denoted by $\rho(\psi, X)$ or $\rho(\psi)$ if $X = \mathbb{X}$, is given by

$$\rho(\psi, X) = \sup\{\rho \geq 0 : \psi - (\rho/2)\|\cdot\|^2 \text{ be convex on } X\}. \tag{9}$$

By definition, the function ψ is strongly convex on X iff $\rho(\psi, X) > 0$.
The conjugate ψ^* of $\psi \in \Gamma_0(\mathbb{X})$ is defined on $\mathbb{Y}$ by

$$\psi^*(\mathbf{y}) := \sup\{\langle x, \mathbf{y} \rangle - \psi(x) : x \in \mathbb{X}\}, \qquad \forall \mathbf{y} \in \mathbb{Y}. \tag{10}$$

For $\psi \in \Gamma_0(\mathbb{X})$ and $\mathbf{x} \in$ dom ψ, the subdifferential $\partial\psi$ of ψ at $\mathbf{x}$, is defined by

$$\partial\psi(\mathbf{x}) := \{\mathbf{y} \in \mathbb{Y} : \psi(u) \geq \psi(\mathbf{x}) + \langle \mathbf{y}, u - \mathbf{x} \rangle, \quad \forall u \in \mathbb{X}\},$$

which is a closed convex set of $\mathbb{Y}$. Its effective domain dom $\partial\psi$ is

$$\text{dom } \partial\psi := \{\mathbf{x} \in \mathbb{X} : \partial\psi(\mathbf{x}) \neq \emptyset\}$$

The subdifferential $\partial\psi$ of $\psi \in \Gamma_0(\mathbb{X})$ is the tightest extension of its derivative $\nabla\psi$, tailored for convex functions in the sense that ψ is differentiable at $\mathbf{x} \in \text{dom}\ \partial\psi$ iff $\partial\psi(\mathbf{x})$ is reduced to the singleton $\{\nabla\psi(\mathbf{x})\}$.
The vector space of DC functions, denoted by $DC(\mathbb{X})$, is the smallest vector space generated by the convex cone $\Gamma_0(\mathbb{X}) \cup \{\omega\} : DC(\mathbb{X}) := [\Gamma_0(\mathbb{X}) \cup \{\omega\}] - [\Gamma_0(\mathbb{X}) \cup \{\omega\}]$, with $\omega(\mathbf{x}) := +\infty$ for $\mathbf{x} \in \mathbb{X}$. It contains most real-world objective functions (see [23, 24, 25, 27] and additional references in [26]) and is closed under usual operations encountered in optimization [24].

Standard primal DC program is of the form

$$(P_{dc}) \quad \alpha := \inf\{\mathbf{f}(\mathbf{x}) := \mathbf{g}(\mathbf{x}) - \mathbf{h}(\mathbf{x}) : \mathbf{x} \in \mathbb{X}\} \tag{11}$$

with $\mathbf{g}$, $\mathbf{h} \in \Gamma_0(\mathbb{X})$. The function $\mathbf{f}$ is called a DC function on $\mathbb{X}$, $\mathbf{g} - \mathbf{h}$ a DC decomposition of $\mathbf{f}$, while $\mathbf{g}$ and $\mathbf{h}$ are its DC components. According to the usual convention $+\infty - (+\infty) = +\infty$, the finiteness of the optimal value α implies $\text{dom}\,\mathbf{f} = \text{dom}\,\mathbf{g} \subset \text{dom}\,\mathbf{h}$. Clearly, for a nonempty closed convex set X of $\mathbb{X}$, the constrained DC program

$$\alpha := \inf\{\mathbf{f}(\mathbf{x}) := \mathbf{g}(\mathbf{x}) - \mathbf{h}(\mathbf{x}) : \mathbf{x} \in X\} \tag{12}$$

can be reformulated as (11)

$$\alpha := \inf\{\mathbf{f}(\mathbf{x}) := \mathbf{g}(\mathbf{x}) - \mathbf{h}(\mathbf{x}) : \mathbf{x} \in X\} = \inf\{(\mathbf{g} + \chi_X)(\mathbf{x}) - \mathbf{h}(\mathbf{x}) : \mathbf{x} \in \mathbb{X}\} \tag{13}$$

where $\chi_X \in \Gamma_0(\mathbb{X})$ is the indicator function of X defined by

$$\chi_X(\mathbf{x}) := 0 \text{ if } \mathbf{x} \in X, +\infty, \text{ otherwise.}$$

The dual DC program (D_{dc}) of (P_{dc}) is defined by

$$(D_{dc}) \quad \alpha := \inf\{\mathbf{h}^*(\mathbf{y}) - \mathbf{g}^*(\mathbf{y}) : \mathbf{y} \in \mathbb{Y}\} \tag{14}$$

and the dual of (D_{dc}) is exactly (P_{dc}) (see [24]).

3.1.1. Duality, criticality and local optimality for (11).

By definition, $\mathbf{x}^* \in \text{dom}\,\mathbf{g}$ is a DC-critical point of $\mathbf{g} - \mathbf{h}$ if

$$\partial\mathbf{g}(\mathbf{x}^*) \cap \partial\mathbf{h}(\mathbf{x}^*) \neq \emptyset, \text{or equivalently } 0 \in [\partial\mathbf{g}(\mathbf{x}^*) - \partial\mathbf{h}(\mathbf{x}^*)], \tag{15}$$

i.e., a zero of the difference of two cyclically maximal monotone $\partial \mathbf{g}$ and $\partial \mathbf{h}$ ([29, 30, 31]). It is called strongly DC critical point of $\mathbf{g} - \mathbf{h}$ if

$$\emptyset \neq \partial \mathbf{h}(\mathbf{x}^*) \subset \partial \mathbf{g}(\mathbf{x}^*). \tag{16}$$

These two basic notions depend on DC components $\mathbf{g}$ and $\mathbf{h}$ of DC function $f = \mathbf{g} - \mathbf{h}$. They are equivalent if $\mathbf{g}$ or $\mathbf{h}$ are differentiable at $\mathbf{x}^*$. Strong DC-criticality is equivalent to d-(irectional) stationarity, while it coincides with Clarke (resp. Fréchet) stationarity if $\mathbf{g}$ or $\mathbf{h}$ (resp. $\mathbf{h}$) is differentiable at $\mathbf{x}^*$ ([32]. Strong DC-criticality prominently features local optimality in DC programming, where it the necessary (resp. sufficient) DC local optimality conditions (resp. under additional assumptions). Moreover, the transportation between primal DC criticality (resp. primal DC globality) and dual DC duality (resp. dual DC globality) by is performed by ∂h and $\partial \mathbf{g}^*$ respectively.

3.1.2. Philosophy of DCA for solving the standard DC program (11)

DCA is based on local optimality conditions and duality in DC programming, which introduces the nice and elegant concept of approximating a DC program by a sequence of convex ones: at each iteration l, DCA approximates the second DC component $\mathbf{h}$ by its affine minorization $\mathbf{h}_l(\mathbf{x}) := \mathbf{h}(\mathbf{x}^l) + \langle \mathbf{x} - \mathbf{x}^l, \mathbf{y}^l \rangle$, with $\mathbf{y}^l \in \partial \mathbf{h}(\mathbf{x}^l)$, and solves the resulting convex subprogram

$$(P_l) \quad \alpha_l := \inf\{\mathbf{f}(\mathbf{x}) := \mathbf{g}(\mathbf{x}) - \mathbf{h}_l(\mathbf{x}) : \mathbf{x} \in \mathbb{X}\} \tag{17}$$

whose $\mathbf{x}^{l+1}$ is a solution, i.e., $\mathbf{x}^{l+1} \in \partial \mathbf{g}^*(\mathbf{y}^l)$ (see [23, 24]).
The fundamental properties presented above inspired the following construction of the standard DCA:

Standard DCA scheme

Initialization: Let $\mathbf{x}^0 \in \text{dom } \partial \mathbf{h}$, $l = 0$.

repeat

1. Compute $\mathbf{y}^l \in \partial \mathbf{h}(\mathbf{x}^l)$.

2. Compute $\mathbf{x}^{l+1} \in \arg\min_{\mathbf{x}\in\mathbb{X}}\{\mathbf{g}(\mathbf{x}) - \langle \mathbf{x}, \mathbf{y}^l\rangle\}$ (P_l).

until Stopping criterion

DCA's stopping rules will be specified later in remark 1.

The above construction of the standard DCA for solving (11) and (14) can be rewritten as (see [23, 24] for more details): from a given point $\mathbf{x}^0 \in \text{dom } \mathbf{g}$,

$$\text{for } l \geq 0, \text{ set } \mathbf{x}^l \to \mathbf{y}^l \in \partial\mathbf{h}(\mathbf{x}^l) \to \mathbf{x}^{l+1} \in \partial g^*(\mathbf{y}^l) \to \mathbf{y}^{l+1} \in \partial\mathbf{h}(\mathbf{x}^{l+1}). \quad (18)$$

We thus recognize the complete symmetry between the sequences $\{\mathbf{x}^l\}$ and $\{\mathbf{y}^l\}$ with respect to the DC duality, and DCA can be seen a primal dual subgradient method for DC programs.

It is clear that DCAs strongly depend on both the equivalent formulations of the original DC program and the corresponding DC decompositions. As a result, there are infinitely many DCAs for a given DC program. This basic DC flexibility implies the universality of DCA: with appropriate DC decompositions and suitably equivalent DC reformulations, DCA permits to recover standard methods in convex (including projected gradient, proximal among others) and nonconvex programming (see [26], Section 1.2) and many nonconvex optimization algorithms recently developed can be seen as a special version of DCA-based algorithms (see [26], Section 3.3).

3.1.3. Convergence of DCA for standard DC programs

Convergences properties of the standard DCA and its complete theoretical foundation in the DC programming framework can be found in [23, 24, 25]. We will report below the DCA's convergence theorem extracted from [24], which are needed for proving the convergence properties of our algorithms in Section 4. Its proof was given in [24].

The whole DCA's convergence theorem in [24] concerns both DC primal and dual problems (P_{dc}) and (D_{dc}). Here, to simplify the presentation we draw only the part dealing with the primal problem. First, a point $\bar{x}$ is called a limit point of the sequence $\{\mathbf{x}^l\}$ if and only if $\bar{x}$ is a limit of some subsequence of $\{\mathbf{x}^l\}$.

Let ρ_i and ρ_i^*, $(i = 1, 2)$ be real nonnegative numbers such that $0 \leq \rho_i < \rho(f_i)$ (resp. $0 \leq \rho_i^* < \rho(f_i^*)$) where $\rho_i = 0$ (resp. $\rho_i^* = 0$) if $\rho(f_i) = 0$ (resp. $\rho(f_i^*) = 0$) and ρ_i (resp. ρ_i^*) may take the value $\rho(f_i)$ (resp. $\rho(f_i^*)$) if it is attained in (9). We next set $f_1 = \mathbf{g}$ and $f_2 = \mathbf{h}$. Also let $d\mathbf{x}^l := \mathbf{x}^{l+1} - \mathbf{x}^l$ and $d\mathbf{y}^l := \mathbf{y}^{l+1} - \mathbf{y}^l$.

Theorem 1. *(from Theorem 3 in [24]) Let $\{\mathbf{x}^l\}$ and $\{\mathbf{y}^l\}$ be the two sequences generated by the standard DCA. The DCA is a descent method without line-search, which enjoys the following key properties (for the primal DC program):*
i) The decrease of the sequence $\{(\mathbf{g} - h)(\mathbf{x}^l)\}$ is expressed by

$$\begin{aligned}(\mathbf{g} - \mathbf{h})(\mathbf{x}^{l+1}) &\leq (\mathbf{h}^* - \mathbf{g}^*)(\mathbf{y}^l) - \frac{\rho_2}{2}\|d\mathbf{x}^l\|^2 \\ &\leq (\mathbf{g} - \mathbf{h})(\mathbf{x}^l) - \frac{\rho_1 + \rho_2}{2}\|d\mathbf{x}^l\|^2, \quad \forall l\end{aligned} \tag{19}$$

where the equality

$$(\mathbf{g} - \mathbf{h})(\mathbf{x}^{l+1}) = (\mathbf{g} - \mathbf{h})(\mathbf{x}^l) \tag{20}$$

holds iff $\mathbf{x}^l \in \partial\mathbf{g}^*(\mathbf{y}^l), \mathbf{y}^l \in \partial\mathbf{h}(\mathbf{x}^{l+1})$ and $(\rho_1 + \rho_2)d\mathbf{x}^l = 0$. In this case the following main statements are established:
i1. *$\mathbf{x}^l, \mathbf{x}^{l+1}$ are the critical points of $\mathbf{g} - \mathbf{h}$ satisfying*
$\mathbf{y}^l \in (\partial\mathbf{g}(\mathbf{x}^l) \cap \partial\mathbf{h}(\mathbf{x}^l))$ and $\mathbf{y}^l \in (\partial\mathbf{g}(\mathbf{x}^{l+1}) \cap \partial\mathbf{h}(\mathbf{x}^{l+1}))$,
i2. *$\mathbf{y}^l$ is a critical point of $\mathbf{h}^* - \mathbf{g}^*$ satisfying $[\mathbf{x}^l, \mathbf{x}^{l+1}] \subset ((\partial\mathbf{g}^*(\mathbf{y}^l) \cap \partial\mathbf{h}^*(\mathbf{y}^l))$.*
i3. *If $\rho_1 + \rho_2 > 0$ then $\mathbf{x}^{l+1} = \mathbf{x}^l$, if $\rho_1^* > 0$ then $\mathbf{y}^l = \mathbf{y}^{l-1}$, if $\rho_2^* > 0$ then $\mathbf{y}^l = \mathbf{y}^{l+1}$.*
(ii) If α^* is finite, then the two sequences $\{(\mathbf{g} - \mathbf{h})(\mathbf{x}^l)\}$ and $\{(\mathbf{h}^* - \mathbf{g}^*)(\mathbf{y}^l)\}$ decrease and converge to the same limit $\bar{\alpha} \geq \alpha^*$, i.e.,

$$\lim_{k \to +\infty} (\mathbf{g} - \mathbf{h})(\mathbf{x}^l) = \lim_{k \to +\infty} (\mathbf{h}^* - \mathbf{g}^*)(\mathbf{y}^l) = \bar{\alpha}.$$

If $\rho_1 + \rho_2 > 0$ then $\lim_{k \to +\infty}\{\mathbf{x}^{l+1} - \mathbf{x}^l\} = 0$.
(iii) If α^* is finite and the sequences $\{\mathbf{x}^l\}$ and $\{\mathbf{y}^l\}$ are bounded, then for every limit point $\mathbf{x}^*$ of $\{\mathbf{x}^l\}$ (resp. $\mathbf{y}^*$ of $\{\mathbf{y}^l\}$) there exists a limit point $\mathbf{y}^*$ of $\{\mathbf{y}^l\}$ (resp. $\mathbf{x}^*$ of $\{\mathbf{x}^l\}$) such that

$$(\mathbf{x}^*, \mathbf{y}^*) \in [\partial\mathbf{g}^*(\mathbf{y}^*) \cap \partial\mathbf{h}^*(\mathbf{y}^*)] \times [\partial\mathbf{g}(x^*) \cap \partial\mathbf{h}(x^*)], (\mathbf{g} - \mathbf{h})(\mathbf{x}^*) = (\mathbf{h}^* - \mathbf{g}^*)(\mathbf{y}^*) = \bar{\alpha}.$$

Such a point $\mathbf{x}^*$ (resp. $\mathbf{y}^*$) is a critical point of $\mathbf{g} - \mathbf{h}$ (resp. $\mathbf{h}^* - \mathbf{g}^*$).

Proof. See the proof of Theorem 3 in [24]. □

Note that, if X and Y are convex sets such that $\{\mathbf{x}^l\} \subset X$ and $\{\mathbf{y}^l\} \subset Y$, then the above theorem remains valid if we replace $\rho(f_i)$ by $\rho(f_i, X)$ and $\rho(f_i^*)$ by $\rho(f_i^*, Y)$, $i = 1, 2$.

From the above theorem 1 we immediately deduce the following two interesting DCA's convergence properties in case either $\mathbf{g}$ or $\mathbf{h}$ is strongly convex.

Corollary 1. *Let $\mathbf{x}^*$ be a limit point of the sequence $\{\mathbf{x}^l\}$ computed by the standard DCA - and so a critical point of the DC program (P_{dc}). If $(\rho_1+\rho_2) > 0$ (this is equivalent to $\rho(\mathbf{g})+\rho(\mathbf{h}) > 0$, i.e. either $\mathbf{g}$ or $\mathbf{h}$ is strongly convex), then*

i) The series $\{\|\mathbf{x}^{l+1} - \mathbf{x}^l\|^2\}$ is convergent with its limit bounded above by

$$\frac{\rho_1 + \rho_2}{2} \sum_{l=0}^{k} \|\mathbf{x}^{l+1} - \mathbf{x}^l\|^2 \leq \mathbf{f}(\mathbf{x}^0) - \alpha^*, \ \forall k.$$

ii) Moreover,

$$\min\{\|\mathbf{x}^{l+1} - \mathbf{x}^l\| : l = 0, ..., k\} \leq \frac{2^{1/2}[\mathbf{f}(x^0) - \mathbf{f}(x^*)]^{1/2}}{(\rho_1+\rho_2)^{1/2}(k+1)^{1/2}}. \tag{21}$$

Thus, DCA has a complexity $O(1/\sqrt{k})$, i.e., the primal critical point x^* can be approximated within the tolerance error $\varepsilon > 0$ after $k \geq \frac{2[\mathbf{f}(\mathbf{x}^0)-\mathbf{f}(\mathbf{x}^*)]}{(\rho_1+\rho_2)\varepsilon^2}$ iterations.

Proof. i) From (19) we obtain $\frac{\rho_1+\rho_2}{2}\|d\mathbf{x}^l\|^2 \leq \mathbf{f}(\mathbf{x}^l) - \mathbf{f}(\mathbf{x}^{l+1}), \forall l$, this results in

$$\frac{\rho_1 + \rho_2}{2} \sum_{l=0}^{k} \|\mathbf{x}^{l+1} - \mathbf{x}^l\|^2 \leq \mathbf{f}(\mathbf{x}^0) - \mathbf{f}(\mathbf{x}^{k+1}) \leq \mathbf{f}(\mathbf{x}^0) - \bar{\alpha} \leq \mathbf{f}(\mathbf{x}^0) - \alpha^*, \ \forall k. \tag{22}$$

Hence the primal DC series $\{\|\mathbf{x}^{l+1} - \mathbf{x}^l\|^2\}$ is convergent.

ii) If $(\rho_1+\rho_2) > 0$, one deduces from (22) that $\mathbf{f}(\mathbf{x}^*) = \bar{\alpha} := \lim_{l\to+\infty} \mathbf{f}(\mathbf{x}^l)$ and

$$(k+1)\frac{\rho_1+\rho_2}{2} \min\{\|\mathbf{x}^{l+1} - \mathbf{x}^l\|^2 : l = 0, ..., k\} \leq [\mathbf{f}(\mathbf{x}^0) - \mathbf{f}(\mathbf{x}^*)], \forall k$$

$$i.e., \ \min\{\|x^{l+1} - x^l\| : l = 0, ..., k\} \leq \frac{2^{1/2}[\mathbf{f}(\mathbf{x}^0) - \mathbf{f}(\mathbf{x}^*)]^{1/2}}{(\rho_1+\rho_2)^{1/2}(k+1)^{1/2}}.$$

Hence, DCA has a complexity $O(1/\sqrt{k})$. □

Remark 1. *From Theorem 1 we see that*

i) If $\mathbf{f}(\mathbf{x}^l) = \mathbf{f}(\mathbf{x}^{l+1})$ *then* $\mathbf{x}^l$ *and* $\mathbf{x}^{l+1}$ *are critical points of* (P_{dc})*, and DCA has a finite convergence.*

ii) If $\rho_1 + \rho_2 > 0$*, then* $\mathbf{f}(\mathbf{x}^l) = \mathbf{f}(\mathbf{x}^{l+1})$ *and* $\mathbf{x}^{l+1} = \mathbf{x}^l$ *are equivalent.*

These remarkable results proper to DCA naturally lead to the the stopping rule for DCA in the standard DCA scheme described above:

$$\textit{Either } |\mathbf{f}(\mathbf{x}^l) - \mathbf{f}(\mathbf{x}^{l+1})| \leq \epsilon_1 \textit{ or } \|\mathbf{x}^l - \mathbf{x}^{l+1}\| \leq \epsilon_2.$$

DC programming and DCA are known as powerful nonconvex optimization tools due to their robustness and performance compared to existing methods, their rapidity and scalability, and the flexibility of DC decompositions. Their efficiency and popularity rely on their remarkable simplicity, their rigorous mathematical foundations and their pervasiveness in optimization.

3.2. DCA for solving $\ell_0 + \ell_{q,0}$ approximate problem

The discontinuity of the step functions can be overcome by using nonconvex approximation functions. Specifically, we consider approximations of the ℓ_0 and $\ell_{q,0}$ norms respectively defined by

$$\|\mathbf{x}\|_0 \approx \sum_{i=1}^{d} \eta_I(x_i), \text{ and } \|\mathbf{x}\|_{q,0} \approx \sum_{j=1}^{J} \eta_O\left(\|\mathbf{x}^{(j)}\|_q\right), \tag{23}$$

where η_I and η_O are nonconvex penalty functions such as the exponential function [40], Smoothly Clipped Absolute Deviation (SCAD, [41]), Minimax Concave Penalty (MCP, [13]), Capped-ℓ_1 [28], etc. A common property of the nonconvex approximation functions η_I and η_O is that they are even and concave increasing on $[0, +\infty)$ (see [2] for more details).

By using (23), we get the following $\ell_0 + \ell_{q,0}$ approximate problem.

$$\min_{x \in \mathbb{R}^d} \{f(\mathbf{x}) + \lambda\alpha \sum_{i=1}^{d} \eta_I(x_i) + \lambda(1-\alpha) \sum_{j=1}^{J} \eta_O(\|\mathbf{x}^{(j)}\|_q)\}. \tag{24}$$

The nonconvex penalty function η_I can be expressed as a DC function $\eta_I(t) = \theta_I|t| - [\theta_I|t| - \eta_I(t)]$, where $\theta_I \geq \eta'(0)$ with $\eta'(0)$ being the right derivative at 0

of η. It has been proved in [2] that $\theta_I|t|-\eta_I(t)$ is convex for all $\theta_I \geq \eta'(0)$. Thus, the ℓ_0-approximation, which is a sum of DC functions, is also a DC function

$$\sum_{i=1}^{d} \eta_I(x_i) = \theta_I \|\mathbf{x}\|_1 - \sum_{i=1}^{d} \left[\theta_I |x_i| - \eta_I(x_i)\right]. \tag{25}$$

The following result shows that the $\ell_{q,0}$-approximation is a DC function.

Proposition 1. *We assume that the function η_O is concave on $[0,+\infty)$ and there exists the right derivative at 0 of η_O denoted by $\eta'_O(0)$. Hence, the function $\mathbf{p}(\mathbf{y}) = \theta_O \|\mathbf{y}\|_q - \eta_O\left(\|\mathbf{y}\|_q\right)$ is convex on $\mathbb{R}^m$ for $\theta_O \geq \eta'_O(0)$.*

Proof. The proposition will be proved once we prove the inequality below.

$$\mathbf{p}(\beta\mathbf{y} + \gamma\mathbf{z}) \leq \beta\mathbf{p}(\mathbf{y}) + \gamma\mathbf{p}(\mathbf{z}), \tag{26}$$

for any $\mathbf{y}, \mathbf{z} \in \mathbb{R}^m$ and $\beta, \gamma > 0$ such that $\beta + \gamma = 1$. The inequality (26) is equivalent to

$$\beta\eta_O\left(\|\mathbf{y}\|_q\right) + \gamma\eta_O\left(\|\mathbf{z}\|_q\right) - \eta_O(v) \leq \theta_O w, \tag{27}$$

where $v = \|\beta\mathbf{y} + \gamma\mathbf{z}\|_q \geq 0$ and $w = \beta\|\mathbf{y}\|_q + \gamma\|\mathbf{z}\|_q - \|\beta\mathbf{y} + \gamma\mathbf{z}\|_q \geq 0$. Since η_O is concave function on $\mathbb{R}_+$, we get

$$\beta\eta_O\left(\|\mathbf{y}\|_q\right) + \gamma\eta_O\left(\|\mathbf{z}\|_q\right) \leq \eta_O(\beta\|\mathbf{y}\|_q + \gamma\|\mathbf{z}\|_q) \tag{28}$$

Let $-t \in \partial\left(-\eta_O\right)(v)$. From the concavity of η_O, we have

$$\eta_O(\beta\|\mathbf{y}\|_q + \gamma\|\mathbf{z}\|_q) \leq \eta_O(v) + tw. \tag{29}$$

Combining (28) with (29), we get

$$\beta\eta_O\left(\|\mathbf{y}\|_q\right) + \gamma\eta_O\left(\|\mathbf{z}\|_q\right) - \eta_O(v) \leq tw. \tag{30}$$

The concavity of η_O implies that

$$\eta_O(\frac{v}{2}) \leq \eta_O(0) + \eta'_O(0)\frac{v}{2} \text{ and } \eta_O(\frac{v}{2}) \leq \eta_O(v) - t\frac{v}{2}.$$

Combining these two inequalities, we have

$$\left[t - \eta'_O(0)\right] v \leq 2\left[\eta_O(v)) + \eta_O(0) - 2\eta_O(\frac{v}{2})\right] \leq 0.$$

Therefore

$$[t - \eta'_O(0)]\, v \leq 0.$$

This and the condition $\eta'_O(0) \leq \theta_O$ imply $t \leq \eta'_O(0) \leq \theta_O$. From this and (30), we conclude that the inequality (27) holds.

□

Thanks to Proposition 1, we propose the following DC decomposition of the $\ell_{q,0}$-approximation:

$$\sum_{j=1}^{J} \eta_O(\|\mathbf{x}^{(j)}\|_q) = \theta_O \|\mathbf{x}\|_{q,1} - [\theta_O \|\mathbf{x}\|_{q,1} - \sum_{j=1}^{J} \eta_O(\|\mathbf{x}^{(j)}\|_q)], \tag{31}$$

where $\|\mathbf{x}\|_{q,1} := \sum_{j=1}^{J} \|\mathbf{x}^{(j)}\|_q$.

On another hand, since f_1 has the L-Lipschitz continuous gradient, $\frac{\mu}{2}\|\cdot\|^2 - f_1$ is convex for $\mu \geq L$. Thus, a DC decomposition of $f := g + f_1 - h$ can be chosen as (with $\mu \geq L$)

$$f(\mathbf{x}) = \left[g(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2\right] - \left[h(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2 - f_1(\mathbf{x})\right]. \tag{32}$$

Finally, by using (25), (31), and (32), the $\ell_0 + \ell_{q,0}$ approximate problem (24) can be reformulated as a DC program

$$\min_{\mathbf{x}\in\mathbb{R}^d} \{\mathbf{F}_q(\mathbf{x}) := \mathbf{g}_q(\mathbf{x}) - \mathbf{h}_q(\mathbf{x})\}, \tag{33}$$

where $\mathbf{g}_q(\mathbf{x})$ and $\mathbf{h}_q(\mathbf{x})$ are convex functions respectively defined by

$$\mathbf{g}_q(\mathbf{x}) = g(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2 + \lambda\alpha\theta_I\|\mathbf{x}\|_1 + \lambda(1-\alpha)\theta_O\|\mathbf{x}\|_{q,1},$$

and

$$\begin{aligned}\mathbf{h}_q(\mathbf{x}) = \quad & h(\mathbf{x}) + \tfrac{\mu}{2}\|\mathbf{x}\|^2 - f_1(\mathbf{x}) + \lambda\alpha[\theta_I\|\mathbf{x}\|_1 - \sum_{i=1}^{d} \eta_I(x_i)] \\ & + \lambda(1-\alpha)[\theta_O\|\mathbf{x}\|_{q,1} - \textstyle\sum_{j=1}^{J} \eta_O(\|\mathbf{x}^{(j)}\|_q)].\end{aligned}$$

According to the generic DCA scheme, DCA-$\ell_0 - \ell_{q,0}$ can be described as follows.

DCA-$\ell_0 - \ell_{q,0}$

Initialization: Choose $\mathbf{x}^0$, $l \leftarrow 0$ and a tolerance $\tau > 0$.

repeat

1. Compute $\mathbf{v}^l \in \mu\mathbf{x}^l - \nabla f_1(\mathbf{x}^l) + \partial[h(\mathbf{x}^l) + \lambda\alpha(\theta_I\|\mathbf{x}^l\|_1 - \sum_{i=1}^{d}\eta_I(x_i^l)) + \lambda(1-\alpha)[\theta_O\|\mathbf{x}^l\|_{q,1} - \sum_{j=1}^{J}\eta_O(\|(\mathbf{x}^{(j)})^l\|_q)]$.
2. Compute $\mathbf{x}^{l+1}$ by solving the following strongly convex problem

$$\min_{\mathbf{x}}\{g(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2 + \lambda\alpha\theta_I\|\mathbf{x}\|_1 + \lambda(1-\alpha)\theta_O\|\mathbf{x}\|_{q,1} - \langle\mathbf{v}^l, \mathbf{x}\rangle\}. \tag{34}$$

until Stopping criterion

3.3. DCA for solving $\ell_{0,0}$ approximate problem

For dealing with $\ell_{0,0}$, we also approximate the inner and outer step functions by the nonconvex penalty functions η_I and η_O, respectively. Thus, the approximation of the $\ell_{0,0}$ norm is given by

$$\|\mathbf{x}\|_{0,0} \approx \sum_{j=1}^{J}\eta_O(\sum_{i=1}^{d_j}\eta_I(x_i^{(j)})). \tag{35}$$

Using this approximation gives us the following $\ell_{0,0}$ approximate problem

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left\{\mathbf{F}(\mathbf{x}) := f(\mathbf{x}) + \lambda\sum_{j=1}^{J}\eta_O(\sum_{i=1}^{d_j}\eta_I(x_i^{(j)}))\right\}. \tag{36}$$

We will propose a DC decomposition of the $\ell_{0,0}$-approximation (35) by using Proposition 2 below.

Proposition 2. *We assume that η_I and η_O are concave and increasing on $[0, +\infty)$, and there exists the right derivatives at 0 of η_I and η_O, respectively denoted by $\eta_I'(0)$ and $\eta_O'(0)$. Then, the function $\varphi(\mathbf{y}) = \kappa\|\mathbf{y}\|_1 - \eta_O\left(\sum_{i=1}^{m}\eta_I\left(y_i\right)\right)$ is convex on $\mathbb{R}^m$ for $\kappa \geq \eta_I'(0)\eta_O'(0)$.*

Proof. We need to prove that

$$\varphi(\beta\mathbf{y} + \gamma\mathbf{z}) \leq \beta\varphi(\mathbf{y}) + \gamma\varphi(\mathbf{z}), \tag{37}$$

for any $\mathbf{y}, \mathbf{z} \in \mathbb{R}^m$ and $\beta, \gamma > 0$ such that $\beta + \gamma = 1$. We denote the vector $(|y_1|, ..., |y_m|)$ by $|\mathbf{y}|$. Let $\nu = |\beta\mathbf{y} + \gamma\mathbf{z}|$, i.e., $\nu_i = |\beta y_i + \gamma z_i| \geq 0$ for all

$i = 1, ..., m$ and $\pi = \beta|\mathbf{y}| + \gamma|\mathbf{z}| - |\beta\mathbf{y} + \gamma\mathbf{z}|$. It follows from the triangle inequality that $\pi_i = \beta|y_i| + \gamma|z_i| - |\beta y_i + \gamma z_i| \geq 0$ for $i = 1, ..., m$. Thus, we have $\|\pi\|_1 = \sum_{i=1}^m (\beta|y_i| + \gamma|z_i| - |\beta y_i + \gamma z_i|) = \beta\|\mathbf{y}\|_1 + \gamma\|\mathbf{z}\|_1 - \|\beta\mathbf{y} + \gamma\mathbf{z}\|_1$. Therefore, the inequality (37) is equivalent to

$$\beta\psi(|\mathbf{y}|) + \gamma\psi(|\mathbf{z}|) - \psi(\nu) \leq \kappa\|\pi\|_1, \tag{38}$$

where $\psi(\mathbf{y}) = \eta_O\left(\sum_{i=1}^m \eta_I(y_i)\right)$. Since the function $\eta_I(y_i)$ is concave on $[0, +\infty)$, the sum of the concave functions $\sum_{i=1}^m \eta_I(y_i)$ is also concave on $\mathbb{R}_+^m$. Hence, $\psi(\mathbf{y}) = \eta_O\left(\sum_{i=1}^m \eta_I(y_i)\right)$, which is the composition of a concave and an increasing concave function, is concave on $\mathbb{R}_+^m$. This implies

$$\beta\psi(|\mathbf{y}|) + \gamma\psi(|\mathbf{z}|) \leq \psi(\beta|\mathbf{y}| + \gamma|\mathbf{z}|). \tag{39}$$

Let $-t$ be a subgradient of $-\eta_O$ at $\sum_{i=1}^m \eta_I(\nu_i)$ and $-\xi_i$ be a subgradient of $-\eta_I$ at ν_i for $i = 1, ..., m$. From the concavity and the increase of η_O, we have

$$0 \leq \eta_O\left(\sum_{i=1}^m \eta_I(\nu_i) + 1\right) - \eta_O\left(\sum_{i=1}^m \eta_I(\nu_i)\right) \leq t. \tag{40}$$

It also follows from the concavity of η_O that

$$\begin{aligned}\psi(\beta|\mathbf{y}| + \gamma|\mathbf{z}|) - \psi(\nu) &\leq t\left(\sum_{i=1}^m \eta_I(\beta|y_i| + \gamma|z_i|) - \sum_{i=1}^m \eta_I(\nu_i)\right) \\ &\leq t\sum_{i=1} \xi_i\pi_i = t\langle\xi, \pi\rangle,\end{aligned} \tag{41}$$

where the second inequality holds by the concavity of η_I and (40). Combining (39) with (41), we get

$$\beta\psi(|\mathbf{y}|) + \gamma\psi(|\mathbf{z}|) - \psi(\nu) \leq t\langle\xi, \pi\rangle. \tag{42}$$

Similar to the proof of Proposition 1, since η_O and η_I are concave, we can show that $t \leq \eta_O'(0)$ and $\xi_i \leq \eta_I'(0)$ for all $i = 1, ..., m$. From this and the condition $\kappa \geq \eta_I'(0)\eta_O'(0)$, we have

$$t\langle\xi, \pi\rangle \leq \sum_{i=1}^m \eta_O'(0)\eta_I'(0)\pi_i \leq \kappa\|\pi\|_1. \tag{43}$$

By combining (42) with (43), we conclude that the inequality (38) holds.

□

From Proposition 2, we propose the following DC decomposition of the $\ell_{0,0}$-approximation:

$$\sum_{j=1}^{J}\eta_O(\sum_{i=1}^{d_j}\eta_I(x_i^{(j)})) = \kappa\|\mathbf{x}\|_1 - [\kappa\|\mathbf{x}\|_1 - \sum_{j=1}^{J}\eta_O(\sum_{i=1}^{d_j}\eta_I(x_i^{(j)}))]. \tag{44}$$

Using (32) and (44), we get a DC formulation of the $\ell_{0,0}$ approximate problem (36) as follows.

$$\min_{\mathbf{x}\in\mathbb{R}^d}\{\mathbf{F}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) - \mathbf{h}(\mathbf{x})\}, \tag{45}$$

where $\mathbf{g}(\mathbf{x})$ and $\mathbf{h}(\mathbf{x})$ are convex functions respectively defined by

$$\begin{aligned}\mathbf{g}(\mathbf{x}) &= g(\mathbf{x}) + \tfrac{\mu}{2}\|\mathbf{x}\|^2 + \lambda\kappa\|\mathbf{x}\|_1,\\ \mathbf{h}(\mathbf{x}) &= h(\mathbf{x}) + \tfrac{\mu}{2}\|\mathbf{x}\|^2 - f_1(\mathbf{x}) + \lambda[\kappa\|\mathbf{x}\|_1 - \sum_{j=1}^{J}\eta_O(\sum_{i=1}^{d_j}\eta_I(x_i^{(j)}))].\end{aligned}$$

The corresponding DCA-$\ell_{0,0}$ is described in the algorithm below.

DCA-$\ell_{0,0}$

Initialization: Choose $\mathbf{x}^0$, $l \leftarrow 0$ and a tolerance $\tau > 0$.

repeat

1. Compute $\mathbf{v}^l \in \mu\mathbf{x}^l - \nabla f_1(\mathbf{x}^l) + \partial[h(\mathbf{x}^l) + \lambda\kappa\|\mathbf{x}^l\|_1 - \lambda\sum_{j=1}^{J}\eta_O(\sum_{i=1}^{d_j}\eta_I((x_i^{(j)})^l))]$.
2. Compute $\mathbf{x}^{l+1}$ by solving the following strongly convex problem

$$\min_{\mathbf{x}}\left\{g(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2 + \lambda\kappa\|\mathbf{x}\|_1 - \langle\mathbf{v}^l, \mathbf{x}\rangle\right\}. \tag{46}$$

until Stopping criterion

---

4. Application in estimating multiple sparse covariance matrices

The estimation of sparse covariance matrices plays an important role in various areas of statistical analysis such as portfolio management and risk assessment, high dimensional classification, analysis of independence and conditional independence relationships between components in graphical models, etc

[42, 43, 44, 45, 46]. In recent years, much interest has focused on estimating a covariance matrix on the basis of a $n \times d$ data matrix $\mathbf{X}$, where n is the number of observations and d is the number of features. Suppose that the observations $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ are independent and identically distributed $\mathcal{N}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a positive definite $d \times d$ matrix. A natural way to estimate the covariance matrix $\mathbf{\Sigma}$ is via minimizing negative log-likelihood. The resulting optimization problem is

$$\min_{\mathbf{\Sigma} \succ 0} \left\{ \log \det \mathbf{\Sigma} + \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) \right\}, \tag{47}$$

where $\mathbf{S} = 1/n \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ is the sample covariance matrix.

Sparse estimation of a covariance matrix or its inverse is particularly relevant in the study of graphical models. Various proposals aim to estimate a single sparse covariance matrix or its inverse in the high dimensional setting. The first direction is to estimate a single sparse inverse covariance matrix [47, 48, 49, 50]. These methods added the ℓ_1 regularization to the objective function of the log-likelihood minimization problem. The second direction is to directly estimate a single covariance matrix. The most of the works in this direction added the ℓ_1 regularization to the objective function of the problem (47) or its surrogate convex loss function (see e.g. [51, 52, 53, 54, 55, 56]). In [22], the authors studied DC approximation approaches for estimating a single covariance matrix. The standard methods for estimating a single covariance matrix or its inverse assume that observations are drawn from the same distribution. However, in many datasets, the observations of each class have different distributions. In such cases, the covariance matrices share some common structures such as the weights and locations of non-zero elements. Therefore, these matrices need to be jointly estimated. In the literature, there are some works which extended the first direction on the estimation of multiple sparse inverse covariance matrices (see e.g. [6, 57, 7, 8, 9, 10]). [57] used a hierarchical penalty for group variable selection in estimating multiple sparse inverse covariance matrices. [6, 7, 8, 9, 10] used the sparse group lasso or generalized fused lasso for simultaneous group and individual variable selection in estimating multiple sparse inverse covariance

matrices. In this paper, we introduce novel approaches for bi-level variable selection in directly estimating multiple sparse covariance matrices using the $\ell_0 + \ell_{q,0}$ and $\ell_{0,0}$ regularization.

By using the $\ell_0 + \ell_{q,0}$-regularization, the problem (4) becomes

$$\min_{\mathbf{\Sigma}^{(k)} \succ 0} \left\{ \mathcal{L}(\{\mathbf{\Sigma}\}) + \lambda\alpha \|\{\mathbf{\Sigma}\}\|_0 + \lambda(1-\alpha)\|\{\mathbf{\Sigma}\}\|_{q,0} \right\}, \tag{48}$$

where the ℓ_0 and $\ell_{q,0}$ norm of $\{\mathbf{\Sigma}\}$ respectively defined by

$$\|\{\mathbf{\Sigma}\}\|_0 = \sum_{k=1}^{Q} \sum_{i,j} s(\Sigma_{ij}^{(k)}) \text{ and } \|\{\mathbf{\Sigma}\}\|_{q,0} = \sum_{i,j} s(\|(\Sigma_{ij}^{(1)}, ..., \Sigma_{ij}^{(Q)})\|_q).$$

We also consider the $\ell_{0,0}$-regularization problem of (4) given by

$$\min_{\mathbf{\Sigma}^{(k)} \succ 0} \left\{ \mathcal{L}(\{\mathbf{\Sigma}\}) + \lambda \|\{\mathbf{\Sigma}\}\|_{0,0} \right\}, \tag{49}$$

where the $\ell_{0,0}$ norm of $\{\mathbf{\Sigma}\}$ is defined by

$$\|\{\mathbf{\Sigma}\}\|_{0,0} = \sum_{i,j} s(\sum_{k=1}^{Q} s(\Sigma_{ij}^{(k)})).$$

First of all, we will show that both problems (48) and (49) are unbounded below if there exists a singular sample covariance matrix $\mathbf{S}^{(k_0)}$. Indeed, if $\mathbf{S}^{(k_0)}$ is singular, there exists a vector $\mathbf{v} \neq 0$ such that $\mathbf{S}^{(k_0)}\mathbf{v} = 0$ and $\|\mathbf{v}\|_2 = 1$. Let $\mathbf{v}_2, ..., \mathbf{v}_d$ be vectors such that $\{\mathbf{v}, \mathbf{v}_2, ..., \mathbf{v}_d\}$ is an orthogonal basis of $\mathbb{R}^d$. Let $\{\mathbf{\Sigma}_\gamma\} = \{\mathbf{\Sigma}_\gamma^{(1)}, ..., \mathbf{\Sigma}_\gamma^{(Q)}\}$, where $\mathbf{\Sigma}_\gamma^{(k)}$ is given by

$$\mathbf{\Sigma}_\gamma^{(k)} = \begin{cases} \gamma \mathbf{v}\mathbf{v}^T + \sum_{i=2}^{d} \mathbf{v}_i \mathbf{v}_i^T & \text{if } k = k_0, \\ \mathbf{I}_d & \text{otherwise,} \end{cases} \tag{50}$$

with a parameter $\gamma > 0$. We have

$$\log\det(\mathbf{\Sigma}_\gamma^{(k_0)}) = \log\gamma, \text{ and } \operatorname{tr}(\mathbf{\Sigma}_\gamma^{(k_0)})^{-1}\mathbf{S}^{(k_0)} = \sum_{i=2}^{d} \mathbf{v}_i^T \mathbf{S}^{(k_0)} \mathbf{v}_i.$$

Thus, we obtain

$$\mathcal{L}(\{\mathbf{\Sigma}_\gamma\}) = n_{k_0} \log\gamma + n_{k_0} \sum_{i=2}^{d} \mathbf{v}_i^T \mathbf{S}^{(k_0)} \mathbf{v}_i + \sum_{k \neq k_0} n_k \operatorname{tr}(\mathbf{S}^{(k)}) \to -\infty,$$

as $\gamma \to 0^+$. Furthermore, the $\ell_0 + \ell_{q,0}$ and $\ell_{0,0}$ norm of $\{\mathbf{\Sigma}_\gamma\}$ are bounded. Hence, we have

$$\mathcal{L}(\{\mathbf{\Sigma}_\gamma\}) + \lambda\alpha\|\{\mathbf{\Sigma}_\gamma\}\|_0 + \lambda(1-\alpha)\|\{\mathbf{\Sigma}_\gamma\}\|_{q,0} \to -\infty$$

and

$$\mathcal{L}(\{\mathbf{\Sigma}_\gamma\}) + \lambda\|\{\mathbf{\Sigma}_\gamma\}\|_{0,0} \to -\infty,$$

as $\gamma \to 0^+$. From this it follows that the objective functions of the problems (48) and (49) are unbounded below if there exists a singular sample covariance matrix $\mathbf{S}^{(k_0)}$. In order to be sure about the positive definiteness, we replace $\mathbf{S}^{(k)}$ with $\mathbf{S}^{(k)} + \epsilon\mathbf{I}_d$ for some $\epsilon > 0$ when $\mathbf{S}^{(k)}$ is singular.

The proposition below shows that there exists a lower bound of $\mathbf{\Sigma}^{(k)}$ for both the problems (48) and (49).

Proposition 3. *If $\mathbf{S}^{(k)}$ is nonsingular for all k, the following statements hold. a) There exists $\delta > 0$ such that the problem* (48) *is equivalent to*

$$\min_{\mathbf{\Sigma}^{(k)} \succeq \delta\mathbf{I}_d} \{\mathcal{L}(\{\mathbf{\Sigma}\}) + \lambda\alpha\|\{\mathbf{\Sigma}\}\|_0 + \lambda(1-\alpha)\|\{\mathbf{\Sigma}\}\|_{q,0}\}. \tag{51}$$

b) There exists $\delta > 0$ such that the problem (49) *is equivalent to*

$$\min_{\mathbf{\Sigma}^{(k)} \succeq \delta\mathbf{I}_d} \{\mathcal{L}(\{\mathbf{\Sigma}\}) + \lambda\|\{\mathbf{\Sigma}\}\|_{0,0}\}. \tag{52}$$

Here the notation $\mathbf{\Sigma}^{(k)} \succeq \delta\mathbf{I}_d$ means that $\mathbf{\Sigma}^{(k)} - \delta\mathbf{I}_d$ is symmetric positive semidefinite.

Proof. The proof of properties (a) and (b) are similar. We give here the proof of (a).

Let $\{\mathbf{\Sigma}\} = \{\mathbf{\Sigma}^{(1)}, ..., \mathbf{\Sigma}^{(Q)}\}$ be the solution to the problem (48). We need to show that there exist $\delta > 0$ such that $\mathbf{\Sigma}^{(k)} \succeq \delta\mathbf{I}_d$ for $k = 1, ..., Q$. The eigen decomposition of $\mathbf{\Sigma}^{(k)}$ is given by

$$\mathbf{\Sigma}^{(k)} = \sum_{i=1}^{d} \lambda_i^{(k)} \mathbf{u}_i \mathbf{u}_i^T,$$

where $\lambda_1^{(k)} \geq ... \geq \lambda_d^{(k)} > 0$. Thus, we have

$$\mathcal{L}(\{\mathbf{\Sigma}\} = \sum_{k=1}^{Q}\sum_{i=1}^{d} n_k(\log \lambda_i^{(k)} + \frac{\mathbf{u}_i^T \mathbf{S}^{(k)} \mathbf{u}_i}{\lambda_i^{(k)}}) = \sum_{k=1}^{Q}\sum_{i=1}^{d} n_k h(\lambda_i^{(k)}, \mathbf{u}_i^T \mathbf{S}^{(k)} \mathbf{u}_i), \quad (53)$$

where $h(x, a) = \log x + \frac{a}{x}$. Denote the smallest eigenvalue of $\mathbf{S}^{(k)}$ by $\lambda_{\min}(\mathbf{S}^{(k)})$. We note that $\lambda_{\min}(\mathbf{S}^{(k)}) = \min_{\|\mathbf{u}\|=1} \mathbf{u}^T \mathbf{S}^{(k)} \mathbf{u}$. Let $\lambda_{\min}^S = \min_k \lambda_{\min}(\mathbf{S}^{(k)}) > 0$. For $a > 0$, the function h has the minimum value $\log a + 1$. Moreover, $h(x, a)$ is increasing in a for all $x > 0$. From these properties and (53), we get

$$\begin{aligned} \mathcal{L}(\{\mathbf{\Sigma}\}) &\geq \textstyle\sum_{k=1}^{Q} \sum_{i=1}^{d} n_k h(\lambda_i^{(k)}, \lambda_{\min}^S) \\ &\geq n_{k_0} h(\lambda_d^{k_0}, \lambda_{\min}^S) + (n - n_{k_0})(d-1)\left(\log \lambda_{\min}^S + 1\right), \end{aligned} \quad (54)$$

where $k_0 = \arg\max_k \lambda_1^{(k)}$. Denote the objective function of (48) by $F(\{\mathbf{\Sigma}\})$. Since $\lambda\alpha\|\{\mathbf{\Sigma}\}\|_0 + \lambda(1-\alpha)\|\{\mathbf{\Sigma}\}\|_{q,0} \geq 0$, $\mathcal{L}(\{\mathbf{\Sigma}\}) \leq F(\{\mathbf{\Sigma}\})$. Moreover, we have $F(\{\mathbf{\Sigma}\}) \leq F(\{\tilde{\mathbf{\Sigma}}\})$ for arbitrary positive definite matrices $\{\tilde{\mathbf{\Sigma}}\} = \{(\tilde{\mathbf{\Sigma}}^{(1)}), ..., (\tilde{\mathbf{\Sigma}}^{(Q)})\}$. Thus $\mathcal{L}(\{\mathbf{\Sigma}\}) \leq F(\{\tilde{\mathbf{\Sigma}}\})$ and so we obtain

$$n_{k_0} h(\lambda_d^{k_0}, \lambda_{\min}^S) + (n - n_{k_0})(d-1)\left(\log \lambda_{\min}^S + 1\right) \leq F(\{\tilde{\mathbf{\Sigma}}\}).$$

This is equivalently rewritten as

$$h(\lambda_d^{k_0}, \lambda_{\min}^S) \leq \frac{F(\{\tilde{\mathbf{\Sigma}}\}) - ((n - n_{k_0})(d-1)\left(\log \lambda_{\min}^S + 1\right)}{n_{k_0}}.$$

Denote a set C by

$$C = \{x \in \mathbb{R}_{++} : h(x, \lambda_{\min}^S) \leq c\}, \quad (55)$$

where c is defined by

$$c = \frac{F(\{\tilde{\mathbf{\Sigma}}\}) - ((n - n_{k_0})(d-1)\left(\log \lambda_{\min}^S + 1\right)}{n_{k_0}}.$$

Since $h(x, \lambda_{\min}^S)$ is continuous on $\mathbb{R}_{++}$ and

$$\lim_{x \to 0^+} h(x, \lambda_{\min}^S) = \lim_{x \to +\infty} h(x, \lambda_{\min}^S) = +\infty,$$

C is a compact set. Hence there exists $0 < \delta \leq \lambda_{\min}^S$ such that $\delta = \min_{x \in C} x$. In other words, $\mathbf{\Sigma}^{(k)} \succeq \delta \mathbf{I}_d$ for all $k = 1, ..., Q$. The proof is then complete. □

We observe that the problems (51) and (52) take the form of (1) where the function f is given by

$$f(\{\mathbf{\Sigma}\}) = \mathcal{L}(\{\mathbf{\Sigma}\}) + \chi_K(\{\mathbf{\Sigma}\}),$$

where $K = \{\{\mathbf{\Sigma}\} = \{\mathbf{\Sigma}^{(1)}, ..., \mathbf{\Sigma}^{(Q)}\} : \mathbf{\Sigma}^{(k)} \succeq \delta \mathbf{I}_d\}$ and χ_K is the indicator function defined by $\chi_K(\{\mathbf{\Sigma}\}) = 0$ if $\{\mathbf{\Sigma}\} \in K$ and $+\infty$ otherwise.

Among several existing approximations of the step function, we use the Capped-ℓ_1 function [28] which is defined by

$$\eta(t) = \min\{1, \theta|t|\}, \tag{56}$$

where θ is a tuning parameter such that $\eta(t)$ approximates the step function $s(t)$ as θ tends to $+\infty$. This choice is motivated by the fact that the Capped-ℓ_1 approximation has been showed to be efficient in the individual variable selection problem (see e.g. [18, 2, 21, 20, 22]).

Now, we consider approximate problems of both the problems (51) and (52) and develop DCA based algorithms for solving them.

4.1. DCA-$\ell_0 - \ell_{q,0}$ for solving the approximate problem of (51)

Using the Capped-ℓ_1 function (56), the corresponding $\ell_0 + \ell_{q,0}$ approximate problem of (51) is

$$\min_{\{\mathbf{\Sigma}\}}\{\mathbf{F}_q(\{\mathbf{\Sigma}\}) := f(\{\mathbf{\Sigma}\}) + \lambda\alpha \sum_{k=1}^{Q}\sum_{ij} \eta(\Sigma_{ij}^{(k)}) + \lambda(1-\alpha)\sum_{ij} \eta(\|(\Sigma_{ij}^{(1)}, ..., \mathbf{\Sigma}_{ij}^{(Q)})\|_q)\}. \tag{57}$$

We note that $\log\det \mathbf{\Sigma}^{(k)}$ is concave while $\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})$ is convex in $\mathbf{\Sigma}^{(k)}$. Hence, we have a natural DC decomposition of f as follows.

$$f(\{\mathbf{\Sigma}\}) = [g(\{\mathbf{\Sigma}\}) + f_1(\{\mathbf{\Sigma}\})] - h(\{\mathbf{\Sigma}\}), \tag{58}$$

where $g(\{\mathbf{\Sigma}\})$, $f_1(\{\mathbf{\Sigma}\})$ and $h(\{\mathbf{\Sigma}\})$ are convex functions respectively given by $g(\{\mathbf{\Sigma}\}) = \chi_K(\{\mathbf{\Sigma}\})$, $f_1(\{\mathbf{\Sigma}\}) = \sum_{k=1}^{Q} n_k \mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})$, and $h(\{\mathbf{\Sigma}\}) = \sum_{k=1}^{Q} -n_k \log\det \mathbf{\Sigma}^{(k)}$. However, following Section 3, we propose a special DC decomposition of f by moving $f_1(\{\mathbf{\Sigma}\})$ into the second DC component.

$$f(\{\mathbf{\Sigma}\}) = g(\{\mathbf{\Sigma}\}) + \frac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 - [\frac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 - f_1(\{\mathbf{\Sigma}\}) + h(\{\mathbf{\Sigma}\})], \tag{59}$$

where the norm $\|\{\mathbf{\Sigma}\}\|$ of $\{\mathbf{\Sigma}\}$ is defined by

$$\|\{\mathbf{\Sigma}\}\| = \sqrt{\sum_{k=1}^{Q} \|\mathbf{\Sigma}^{(k)}\|_F^2}.$$

For estimating μ such that $\frac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 - f_1(\{\mathbf{\Sigma}\})$ is convex, we provide the following proposition.

Proposition 4. *If* $\mu \geq \max_k 2n_k\|\mathbf{S}^{(k)}\|_2\delta^{-3}$, *then* $\frac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 - f_1(\{\mathbf{\Sigma}\})$ *is convex in* $\{\mathbf{\Sigma}\}$.

Proof. We have

$$\tfrac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 - f_1(\{\mathbf{\Sigma}\}) = \sum_{k=1}^{Q} [\tfrac{\mu}{2}\|\mathbf{\Sigma}^{(k)}\|_F^2 - n_k \mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})].$$

Since the sum of convex functions is also convex, it is sufficient to show that $\frac{\mu}{2}\|\mathbf{\Sigma}^{(k)}\|_F^2 - n_k\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})$ is convex. From the proof of Lemma 1 in [22], we obtain

$$\rho(\nabla^2\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})) \leq \|\nabla^2\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})\|_2 \leq 2\|\mathbf{S}^{(k)}\|_2\delta^{-3}, \tag{60}$$

where $\rho(\nabla^2\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)}))$ and $\|\nabla^2\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})\|_2$ are the spectral radius and spectral norm of the Hessian matrix of $\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})$, respectively. Moreover, for $k = 1, ..., Q$ we have

$$2n_k\|\mathbf{S}^{(k)}\|_2\delta^{-3} \leq \max_k 2n_k\|\mathbf{S}^{(k)}\|_2\delta^{-3} \leq \mu. \tag{61}$$

It follows from (60) and (61) that μ is greater than the spectral radius of the Hessian matrix of $n_k\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})$, i.e., the function $\frac{\mu}{2}\|\mathbf{\Sigma}^{(k)}\|_F^2 - n_k\mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})$ is convex. The proof is then complete. □

The Capped-ℓ_1 function can be expressed as a DC function as follows:

$$\eta(t) = \theta|t| - \phi(t), \tag{62}$$

where $\phi(t) = -1 + \max\{1, \theta|t|\}$. In addition, the Capped-ℓ_1 function verifies the assumptions in Proposition 1 with $\eta'(0) = \theta$, hence we have DC decompositions

of the ℓ_0-approximation and $\ell_{q,0}$-approximation as follows:

$$\sum_{k=1}^{Q}\sum_{ij}\eta\left(\Sigma_{ij}^{(k)}\right) = \theta\|\{\mathbf{\Sigma}\})\|_1 - \sum_{k=1}^{Q}\sum_{ij}\phi(\Sigma_{ij}^{(k)}), \tag{63}$$

and

$$\sum_{ij}\eta\left(\|(\Sigma_{ij}^{(1)},...,\Sigma_{ij}^{(Q)})\|_q\right) = \theta\sum_{ij}\|(\Sigma_{ij}^{(1)},...,\Sigma_{ij}^{(Q)})\|_q - \sum_{ij}\phi(\|(\Sigma_{ij}^{(1)},...,\Sigma_{ij}^{(Q)})\|_q). \tag{64}$$

Finally, by using (59), (63) and (64), we propose a DC formulation of the approximate problem (57) as follows:

$$\min_{\{\mathbf{\Sigma}\}}\left\{\mathbf{F}_q(\{\mathbf{\Sigma}\}) = \mathbf{g}_q(\{\mathbf{\Sigma}\}) - \mathbf{h}_q(\{\mathbf{\Sigma}\})\right\}, \tag{65}$$

where

$$\begin{aligned}\mathbf{g}_q(\{\mathbf{\Sigma}\}) = \quad & \tfrac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 + \chi_K(\{\mathbf{\Sigma}\}) \\ & +\lambda\alpha\theta\|\{\mathbf{\Sigma}\})\|_1 + \lambda(1-\alpha)\theta\textstyle\sum_{ij}\|(\Sigma_{ij}^{(1)},...,\Sigma_{ij}^{(Q)})\|_q,\end{aligned}$$

and

$$\begin{aligned}\mathbf{h}_q(\{\mathbf{\Sigma}\}) = \quad & \tfrac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 - \textstyle\sum_{k=1}^{Q} n_k[\log\det\mathbf{\Sigma}^{(k)} + \mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})] \\ & +\lambda\alpha\textstyle\sum_{k=1}^{Q}\sum_{ij}\phi(\Sigma_{ij}^{(k)}) + \lambda(1-\alpha)\sum_{ij}\phi(\|(\Sigma_{ij}^{(1)},...,\Sigma_{ij}^{(Q)})\|_q).\end{aligned}$$

Note that (65) is a family of DC programs depending on the values of q. In our application, we consider two special values of q ($q = 1$ and $q = 2$).

4.1.1. ***DCA-***$\ell_0 - \ell_{1,0}$

According to DCA-$\ell_0 - \ell_{q,0}$ with $q = 1$, at each iteration l, we have to compute $\{\mathbf{V}^l\} \in \partial\mathbf{h}_1(\{\mathbf{\Sigma}^l\})$, and then compute $\{\mathbf{\Sigma}^{l+1}\}$ by solving the following convex problem

$$\min_{\{\mathbf{\Sigma}\}}\left\{\mathbf{g}_1(\{\mathbf{\Sigma}\}) - \langle\{\mathbf{V}^l\},\{\mathbf{\Sigma}\}\rangle\right\}. \tag{66}$$

Computing $\{\mathbf{V}^l\}$ can be explicitly given by $\{\mathbf{V}^l\} = \{\mathbf{A}^l\} + \{\mathbf{B}^l\} + \{\mathbf{C}^l\}$, where

$$(\mathbf{A}^{(k)})^l = \mu(\mathbf{\Sigma}^{(k)})^l + n_k[(\mathbf{\Sigma}^{(k)})^l]^{-1}S^{(k)}[(\mathbf{\Sigma}^{(k)})^l]^{-1} - n_k[(\mathbf{\Sigma}^{(k)})^l]^{-1}, \tag{67}$$

$$(B_{ij}^{(k)})^l = \begin{cases} \text{sgn}(\Sigma_{ij}^{(k)})^l \lambda\alpha\theta & \text{if } \theta|(\Sigma_{ij}^{(k)})^l| \geq 1 \\ 0 & \text{otherwise} \end{cases}, \tag{68}$$

$$(C_{ij}^{(k)})^l = \begin{cases} \text{sgn}(\Sigma_{ij}^{(k)})^l \lambda(1-\alpha)\theta & \text{if } \theta\|(\Sigma_{ij}^{(1)})^l, ..., (\Sigma_{ij}^{(Q)})^l\|_1 \geq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{69}$$

Since $\|\{\mathbf{\Sigma}\}\|_{1,1} = \|\{\mathbf{\Sigma}\}\|_1$ is separable, the convex problem (66) can be decomposed into Q independent sub-problems of the same form:

$$\min_{\mathbf{\Sigma}^{(k)} \succeq \delta \mathbf{I}_d} \left\{ \frac{\mu}{2}\|\mathbf{\Sigma}^{(k)}\|_F^2 + \lambda\theta\|\mathbf{\Sigma}^{(k)}\|_1 - \langle (\mathbf{V}^{(k)})^l, \mathbf{\Sigma}^{(k)} \rangle \right\}. \tag{70}$$

For solving each convex sub-problem (70), we use the alternating direction method of multipliers (ADMM) [58]. At each iteration m of ADMM, we have to compute $\mathbf{X}^{m+1}, \mathbf{Y}^{m+1}$, and update $\mathbf{Z}^{m+1} = \mathbf{Z}^m + \rho(\mathbf{X}^{m+1} - \mathbf{Y}^{m+1})$ as follows. The solution $\mathbf{X}^{m+1}$ is computed by

$$\mathbf{X}^{m+1} = \mathbf{U}\mathbf{D}_\delta\mathbf{U}^T, \tag{71}$$

where $\mathbf{D}_\delta = \text{diag}(\max(D_{ii}, \delta))$ with $\mathbf{U}\mathbf{D}\mathbf{U}^T = ((\mathbf{V}^{(k)})^l - \mathbf{Z}^m + \rho\mathbf{Y}^m)/(\mu + \rho)$. $\mathbf{Y}^{m+1}$ is given by

$$\mathbf{Y}^{m+1} = \mathcal{S}\left(\mathbf{X}^{m+1} + \frac{\mathbf{Z}^m}{\rho}, \frac{\lambda\theta}{\rho}\right), \tag{72}$$

where $\mathcal{S}$ is the elementwise soft-thresholding operator defined by $\mathcal{S}(\mathbf{A}, \mathbf{B})_{ij} = \text{sgn}(A_{ij})(|A_{ij}| - B_{ij})_+$.

4.1.2. ***DCA-*** $\ell_0 - \ell_{2,0}$

Each iteration of DCA-$\ell_0 - \ell_{2,0}$ consists in computing a subgradient $\{\mathbf{V}^l\} \in \partial\mathbf{h}_2(\{\mathbf{\Sigma}^l\})$ and solving the following convex problem

$$\min_{\{\mathbf{\Sigma}\}} \left\{ \mathbf{g}_2(\{\mathbf{\Sigma}\}) - \langle \{\mathbf{V}^l\}, \{\mathbf{\Sigma}\} \rangle \right\}. \tag{73}$$

Computing $\{\mathbf{V}^l\}$ can be explicitly computed as $\{\mathbf{V}^l\} = \{\mathbf{A}^l\} + \{\mathbf{B}^l\} + \{\mathbf{E}^l\}$, where $\{\mathbf{A}^l\}$ and $\{\mathbf{B}^l\}$ are respectively computed by using (67) and (68) while

$\{\mathbf{E}^l\}$ is computed by $(E^{(k)})^l_{ij} = \lambda(1-\alpha)\theta(\Sigma^{(k)})^l_{ij}/\|(\Sigma^{(1)})^l_{ij},...,(\Sigma^{(Q)})^l_{ij}\|_2$ if $\theta\|(\Sigma^{(1)})^l_{ij},...,(\Sigma^{(Q)})^l_{ij}\|_2 \geq 1$ and 0 otherwise. The convex problem (73) can be rewritten as

$$\min_{\mathbf{\Sigma}^{(k)} \succeq \delta \mathbf{I}_d} \{\frac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 + \lambda\alpha\theta\|\{\mathbf{\Sigma}\})\|_1 - \langle\{\mathbf{V}^l\},\{\mathbf{\Sigma}\}\rangle + \lambda(1-\alpha)\theta\sum_{ij}\|(\Sigma^{(1)}_{ij},...,\Sigma^{(Q)}_{ij})\|_2\}.$$

This problem cannot be decomposed into smaller sub-problems as the previous case. Here we also apply the ADMM algorithm for computing this problem. At each iteration m, ADMM consists in computing $\{\mathbf{X}^{m+1}\}$, $\{\mathbf{Y}^{m+1}\}$, and updating $\{\mathbf{Z}^{m+1}\} = \{\mathbf{Z}^m\} + \rho(\{\mathbf{X}^{m+1}\} - \{\mathbf{Y}^{m+1}\})$ as follows. $\{\mathbf{X}^{m+1}\}$ is computed by

$$(\mathbf{X}^{(k)})^{m+1} = \mathbf{U}\mathbf{D}_\delta\mathbf{U}^T, \tag{74}$$

where $\mathbf{D}_\delta = \text{diag}(\max(D_{ii},\delta))$ with

$$\mathbf{U}\mathbf{D}\mathbf{U}^T = \left((\mathbf{V}^{(k)})^l - (\mathbf{Z}^{(k)})^m + \rho(\mathbf{Y}^{(k)})^m\right)/(\mu+\rho).$$

We compute $\{\mathbf{Y}^{m+1}\}$ as

$$\{Y^{m+1}_{ij}\} = [\|\mathbf{R}\|_2 - \lambda(1-\alpha)\theta/\rho]_+ \frac{\mathbf{R}}{\|\mathbf{R}\|_2}, \tag{75}$$

where $\mathbf{R} = \mathcal{S}\left(\{X^{m+1}_{ij}\} + \{Z^m_{ij}\}/\rho, \lambda\alpha\theta/\rho\right)$. Here $\{Y^{m+1}_{ij}\}$ denotes

$$\{Y^{m+1}_{ij}\} = \{(Y^{(1)}_{ij})^{m+1},...,(Y^{(Q)}_{ij})^{m+1}\}.$$

4.2. ***DCA-$\ell_{0,0}$*** *for solving the approximate problem of* (52)

We continue using the Capped-ℓ_1 function for both the inner and outer step function. The resulting $\ell_{0,0}$ approximate problem of (52) is given by

$$\min_{\{\mathbf{\Sigma}\}}\{f(\{\mathbf{\Sigma}\}) + \lambda\sum_{ij}\eta(\sum_{k=1}^{Q}\eta(\Sigma^{(k)}_{ij}))\}. \tag{76}$$

We observe that η verifies the assumptions in Proposition 2 with the directional derivative $\psi'_+(0,\mathbf{e}_k) = \theta^2$ for $k = 1,...,Q$, where $\psi(\{\Sigma_{ij}\}) = \eta(\sum_{k=1}^{Q}\eta(\Sigma^{(k)}_{ij}))$. Hence, we propose a DC decomposition of the $\ell_{0,0}$-approximation below.

$$\sum_{ij}\eta(\sum_{k=1}^{Q}\eta(\Sigma^{(k)}_{ij})) = \theta^2\|\{\mathbf{\Sigma}\}\|_1 - [\theta^2\|\{\mathbf{\Sigma}\}\|_1 - \sum_{ij}\eta(\sum_{k=1}^{Q}\eta(\Sigma^{(k)}_{ij}))]. \tag{77}$$

By using (59) and (77), the approximate problem (76) can be reformulated as a DC program:

$$\min_{\{\mathbf{\Sigma}\}} \{\mathbf{F}(\{\mathbf{\Sigma}\}) := \mathbf{g}(\{\mathbf{\Sigma}\}) - \mathbf{h}(\{\mathbf{\Sigma}\})\}, \tag{78}$$

where

$$\mathbf{g}(\{\mathbf{\Sigma}\}) = \frac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 + \chi_K(\{\mathbf{\Sigma}\}) + \lambda\theta^2\|\{\mathbf{\Sigma}\})\|_1,$$

and

$$\begin{aligned}\mathbf{h}(\{\mathbf{\Sigma}\}) = \quad & -\sum_{k=1}^{Q} n_k \left[\log\det \mathbf{\Sigma}^{(k)} + \mathrm{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})\right] \\ & +\tfrac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 + \lambda[\theta^2\|\{\mathbf{\Sigma}\}\|_1 - \textstyle\sum_{ij}\eta(\sum_{k=1}^{Q}\eta(\Sigma_{ij}^{(k)}))].\end{aligned}$$

According to DCA-$\ell_{0,0}$, at each iteration l, we have to compute $\{\mathbf{V}^l\} \in \partial\mathbf{h}(\{\mathbf{\Sigma}^l\})$, and then compute $\{\mathbf{\Sigma}^{l+1}\}$ by solving the following convex problem:

$$\min_{\{\mathbf{\Sigma}\}} \left\{\mathbf{g}(\{\mathbf{\Sigma}\}) - \langle\{\mathbf{V}^l\}, \{\mathbf{\Sigma}\}\rangle\right\}. \tag{79}$$

Computing $\{\mathbf{V}^l\}$ can be explicitly given by $\{\mathbf{V}^l\} = \{\mathbf{A}^l\} + \{\mathbf{T}^l\}$, where $\{\mathbf{A}^l\}$ is computed by using (67) while $\{\mathbf{T}^l\}$ is computed by

$$(T_{ij}^{(k)})^l = \lambda\theta^2(1 - e^{-\theta|(\Sigma_{ij}^{(k)})^l| + \sum_{h=1}^{Q}\eta(|(\Sigma_{ij}^{(h)})^l|)|})\mathrm{sgn}((\Sigma_{ij}^{(k)})^l).$$

The convex problem (79) is also decomposed into Q independent sub-problems of the same form:

$$\min_{\mathbf{\Sigma}^{(k)} \succeq \delta\mathbf{I}_d} \{\frac{\mu}{2}\|\mathbf{\Sigma}^{(k)}\|_F^2 + \lambda\theta^2\|\mathbf{\Sigma}^{(k)}\|_1 - \langle(\mathbf{V}^{(k)})^l, \mathbf{\Sigma}^{(k)}\rangle\}, \tag{80}$$

for which ADMM is investigated as similarly as it is done in DCA-$\ell_0 + \ell_{1,0}$.

4.3. Convergence analysis

We provide the convergence properties of DCA-$\ell_0 - \ell_{q,0}$ and DCA-$\ell_{0,0}$ in the following theorem. Let us use the common function φ to denote either $\mathbf{F}_q$ or $\mathbf{F}$.

Theorem 2. *Let $\left\{\{\mathbf{\Sigma}^l\}\right\}$ be the sequence generated by DCA-$\ell_0 - \ell_{q,0}$ (resp. DCA-$\ell_{0,0}$), the following statements hold.*

a) The sequence $\left\{\varphi(\{\mathbf{\Sigma}^l\})\right\}$ is decreasing and its decrease is expressed by

$$\varphi(\{\mathbf{\Sigma}^l\}) - \varphi(\{\mathbf{\Sigma}^{l+1}\}) \geq \frac{\mu}{2}\|\{\mathbf{\Sigma}^l\} - \{\mathbf{\Sigma}^{l+1}\}\|^2. \tag{81}$$

Therefore, the equality $\varphi(\{\mathbf{\Sigma}^l\}) = \varphi(\{\mathbf{\Sigma}^{l+1}\})$ *holds if and only if* $\{\mathbf{\Sigma}^l\} = \{\mathbf{\Sigma}^{l+1}\}$. *In such a case, DCA-*$\ell_0 - \ell_{q,0}$ *(resp. DCA-*$\ell_{0,0}$*) has a finite convergence and* $\{\mathbf{\Sigma}^l\} = \{\mathbf{\Sigma}^{l+1}\}$ *is a critical points of* (65) *(resp.* (78)*).*

b) The sequence $\{\{\mathbf{\Sigma}^l\}\}$ *is bounded.*

c) The sequence $\big\{\{\mathbf{\Sigma}^l\}\big\}$ *has at least one limit point and every its limit point is a critical point of* (65) *(resp.* (78)*).*

d) The series $\|\{\mathbf{\Sigma}^l\} - \{\mathbf{\Sigma}^{l+1}\}\|^2$ *is convergent.*

e) Let $\{\mathbf{\Sigma}^*\}$ *be a limit point of* $\big\{\{\mathbf{\Sigma}^l\}\big\}$ *(which is a critical point of* (65) *(resp.* (78)*) according to c)). One has*

$$\min_{l=0,\ldots,k} \|\{\mathbf{\Sigma}^l\} - \{\mathbf{\Sigma}^{l+1}\}\|^2 \leq \frac{2^{1/2}[\varphi(\{\mathbf{\Sigma}^0\}) - \varphi(\{\mathbf{\Sigma}^*\})]^{1/2}}{(\rho_1+\rho_2)^{1/2}(k+1)^{1/2}}. \tag{82}$$

Hence, the complexity of DCA-$\ell_0 - \ell_{q,0}$ *and DCA-*$\ell_{0,0}$ *is* $O(1/\sqrt{k})$.

Proof. The convergence properties of DCA-$\ell_0 - \ell_{q,0}$ and DCA-$\ell_{0,0}$ are proved analogously. Therefore, we give here the proof for DCA-$\ell_0 - \ell_{q,0}$ only. It is based on the convergence properties of the generic DCA scheme mentioned in Section 3.1 (Theorem 1).

First of all, we remark that the first DC component of $\mathbf{F}_q$ (resp. $\mathbf{F}$) in the DC program (65) (resp. (78)) is strongly convex with $\rho(\mathbf{g}_q) > \mu$ (resp. $\rho(\mathbf{g}) > \mu$).

a) The inequality (81) in a) is immediately deduced from (19) with $\rho_1 = \mu$ and $\rho_2 = 0$. The remaining statements of a) are direct consequences of (81) and the property i) of Theorem 1.

b) First, we will show that the level set $\mathbf{L}$ defined by

$$\mathbf{L} = \big\{\{\mathbf{\Sigma}\} \in K : \mathbf{F}_q(\{\mathbf{\Sigma}\}) \leq \mathbf{F}_q(\{\mathbf{\Sigma}^0\})\big\}$$

is bounded. Suppose on contrary that $\mathbf{L}$ is not bounded. Then there exists a sequence $\big\{\{\bar{\mathbf{\Sigma}}^l\}\big\} \subset \mathbf{L}$ such that $\|\{\bar{\mathbf{\Sigma}}^l\}\| \to +\infty$ as $l \to +\infty$. Let $\lambda_1((\bar{\mathbf{\Sigma}}^{(k)})^l), ..., \lambda_d((\bar{\mathbf{\Sigma}}^{(k)})^l)$ be the eigenvalues of $(\bar{\mathbf{\Sigma}}^{(k)})^l$ such that

$$\lambda_{\max}((\bar{\mathbf{\Sigma}}^{(k)})^l) := \lambda_1((\bar{\mathbf{\Sigma}}^{(k)})^l) \geq ... \geq \lambda_d((\bar{\mathbf{\Sigma}}^{(k)})^l) \geq \delta.$$

From the properties of the matrix norms, we have

$$\|\{\bar{\mathbf{\Sigma}}^l\}\|^2 = \sum_{k=1}^{Q} \|(\bar{\mathbf{\Sigma}}^{(k)})^l\|_F^2 \leq \sum_{k=1}^{Q} d\lambda_{\max}^2((\bar{\mathbf{\Sigma}}^{(k)})^l) \leq Qd\lambda_{\max}^2(\{\bar{\mathbf{\Sigma}}^l\}), \tag{83}$$

where $\lambda_{\max}(\{\bar{\mathbf{\Sigma}}^l\}) := \max_k \lambda_{\max}((\bar{\mathbf{\Sigma}}^{(k)})^l)$. It follows from (83) that

$$\|\{\bar{\mathbf{\Sigma}}^l\}\| \leq \sqrt{Qd}\lambda_{\max}(\{\bar{\mathbf{\Sigma}}^l\}). \tag{84}$$

Combining (84) and $\|\{\bar{\mathbf{\Sigma}}^l\}\| \to +\infty$ as $l \to +\infty$, we get

$$\lambda_{\max}(\{\bar{\mathbf{\Sigma}}^l\}) \to +\infty, \text{ as } l \to +\infty. \tag{85}$$

Since $\text{tr}(((\bar{\mathbf{\Sigma}}^{(k)})^l)^{-1}\mathbf{S}^{(k)})$ is always nonnegative for $k = 1, ..., Q$ and $\eta(t) \geq 0$, we obtain

$$\begin{aligned}\mathbf{F}_q(\{\bar{\mathbf{\Sigma}}^l\}) &\geq \sum_{k=1}^{Q} n_k \log\det(\bar{\mathbf{\Sigma}}^{(k)})^l = \sum_{k=1}^{Q} n_k \sum_{i=1}^{d} \log\lambda_i((\bar{\mathbf{\Sigma}}^{(k)})^l) \\ &\geq \log\lambda_{\max}(\{\bar{\mathbf{\Sigma}}^l\}) + (nd-1)\log\delta,\end{aligned} \tag{86}$$

where $n = \sum_{k=1}^{Q} n_k$. It follows from (85) and (86) that $\mathbf{F}_q(\{\bar{\mathbf{\Sigma}}^l\}) \to +\infty$ as $l \to +\infty$. But, the fact that $\{\bar{\mathbf{\Sigma}}^l\} \in \mathbf{L}$ implies $\mathbf{F}_q(\{\bar{\mathbf{\Sigma}}^l\}) \leq \mathbf{F}_q(\{\bar{\mathbf{\Sigma}}^0\})$ for all l, which is a contradiction. Hence the level set $\mathbf{L}$ is bounded.

Since the sequence $\{\mathbf{F}_q(\{\mathbf{\Sigma}^l\})\}$ is monotonically decreasing, we have $\left\{\{\mathbf{\Sigma}^l\}\right\} \subseteq \mathbf{L}$. This and the boundness of the level set $\mathbf{L}$ imply that $\{\{\mathbf{\Sigma}^l\}\}$ is bounded.

c) Basing on the property iii) of the DCA's convergence theorem (Theorem 1) we see that c) will be verified when the sequence $\{\{\mathbf{\Sigma}^l\}\}$ is bounded (the above property b)) and the minimum value of Problem (65), say $\alpha^* = \inf_{\{\mathbf{\Sigma}\}} \mathbf{F}_q(\{\mathbf{\Sigma}\})$, is finite. Hence, to prove c) it remains to show that α^* is finite.

Indeed, we have

$$\mathbf{F}_q(\{\mathbf{\Sigma}\}) \geq \sum_{k=1}^{Q} n_k \log\det\mathbf{\Sigma}^{(k)} \geq nd\log\delta, \tag{87}$$

where $n = \sum_{k=1}^{Q} n_k$. The first inequality comes from the fact that $\text{tr}((\mathbf{\Sigma}^{(k)})^{-1}\mathbf{S}^{(k)})$ is always nonnegative for $k = 1, ..., Q$ and

$$\lambda\alpha\sum_{k=1}^{Q}\sum_{ij}\eta\left(\Sigma_{ij}^{(k)}\right) + \lambda(1-\alpha)\sum_{ij}\eta\left(\|(\Sigma_{ij}^{(1)}, ..., \Sigma_{ij}^{(Q)})\|_q\right) \geq 0,$$

while the second one is verified because $\log\det\mathbf{\Sigma}^{(k)} \geq d\log\delta$ for all $\mathbf{\Sigma}^{(k)} \succeq \delta\mathbf{I}_d$. It follows from (87) that $\alpha^* > -\infty$.

Finally, d) and e) are immediate consequences of the properties i) and ii) of Corollary 1 respectively, and the fact that function $\mathbf{g}_q$ is strongly convex.
The proof is then complete.

□

5. Numerical experiments

5.1. Comparative algorithms

We will compare the proposed algorithms (DCA-$\ell_0 - \ell_{1,0}$, DCA-$\ell_0 - \ell_{2,0}$ and DCA-$\ell_{0,0}$) with four methods.

• The convex approximation approach of the $\ell_0 + \ell_{2,0}$ regularization replaced by the $\ell_1 + \ell_{2,1}$ regularization [3, 4]. The resulting problem of estimating multiple sparse covariance matrices is

$$\min_{\{\mathbf{\Sigma}\}} \{f(\{\mathbf{\Sigma}\}) + \lambda\alpha\|\{\mathbf{\Sigma}\}\|_1 + \lambda(1-\alpha)\|\{\mathbf{\Sigma}\}\|_{2,1}\}. \tag{88}$$

This problem is nonconvex and then still difficult. We propose a DCA based algorithm (DCA-$\ell_1 - \ell_{2,1}$) for solving the problem (88). The DCA-$\ell_1 - \ell_{2,1}$ is similar to DCA-$\ell_0 - \ell_{2,0}$. We simply replace the computation of $\{\mathbf{V}^l\} \in \partial\mathbf{h}_2(\{\mathbf{\Sigma}^l\})$ with $\{\mathbf{V}^l\} = \{\mathbf{A}^l\}$ computed by (67) in DCA-$\ell_0 - \ell_{2,0}$.

• The bi-level-weighted method: we consider another equivalent form of the problem (76) as follows

$$\min_{\{\mathbf{\Sigma}\},\{\mathbf{Y}\}} \{f(\{\mathbf{\Sigma}\}) + \lambda \sum_{ij} \eta(\sum_{k=1}^{Q} \eta(Y_{ij})) : |\mathbf{\Sigma}_{ij}^{(k)}| \leq Y_{ij}^{(k)}\}. \tag{89}$$

We can show that $\psi(\mathbf{Y}) = \sum_{ij} \eta(\sum_{k=1}^{Q} \eta(Y_{ij}))$ is concave on $\mathbb{R}_+^Q$. Hence, this problem is a DC program for which DCA is investigated. DCA (DCA-w-$\ell_{0,0}$) consists of solving the following weighted problems

$$\min_{\mathbf{\Sigma}^{(k)} \succeq \delta\mathbf{I}_d} \{\frac{\mu}{2}\|\{\mathbf{\Sigma}\}\|^2 - \langle\{A^l\},\{\mathbf{\Sigma}\}\rangle + \sum_{k=1}^{Q}\sum_{ij}(W_{ij}^{(k)})^l|\Sigma_{ij}^{(k)}|\}, \tag{90}$$

where the weights $(W_{ij}^{(k)})^l$ are computed by

$$(W_{ij}^{(k)})^l = \begin{cases} \lambda\alpha\theta^2 & \text{if } \theta\sum_{k=1}^{Q}\eta(\Sigma_{ij}^{(k)})^l) \leq 1 \text{ and } \theta|(\Sigma_{ij}^{(k)})^l| \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

• The proximal gradient based algorithm [59] (PGD): at each iteration l, PGD for solving (76) requires the computation of the following proximal operator

$$\{\mathbf{\Sigma}^{l+1}\} = \arg\min \frac{\mu}{2}\|\{\mathbf{\Sigma}\}-\{\mathbf{\Sigma}^l\}+\frac{1}{\mu}\nabla\mathcal{L}(\{\mathbf{\Sigma}^l\})\|^2+\chi_K(\{\mathbf{\Sigma}\})+\lambda\sum_{ij}\eta(\sum_{k=1}^{Q}\eta(\Sigma_{ij}^{(k)})).$$

Unfortunately, this proximal operator is non-convex and does not have a closed form. We therefore use DCA for computing it.

• Moreover, for the real application in classification problems we consider in addition a method using the multiple sample covariance matrices: the Quadratic Discriminant Algorithm (QDA).

5.2. Experimental setups

The proposed algorithms are implemented in R software and all algorithms are performed on a PC Intel i7 CPU3770, 3.40 GHz of 8GB RAM. In experiments, we set the stopping tolerance $\tau = 10^{-5}$ for DCA based algorithms and ADMM. The starting point $\{\mathbf{\Sigma}^0\}$ of DCA is the sample covariance matrices $\{\mathbf{S}^{(1)}, ..., \mathbf{S}^{(Q)}\}$. The values of parameters λ, α, θ and ϵ are chosen through a 5-fold cross-validation procedure from the sets of candidates $\Lambda = \{0.04, 0.06, 0.08, 0.1, 1.2, 1.4, 1.6, 1.8, 2\}, \Gamma = \{0.1, 0.5, 0.8\}$, and $\Theta = \{0.5, 1, 3, 5, 7\}$, respectively. Since the $\ell_{1,0}$ regularization term can simultaneously encourage sparsity at the level of both groups and individual variables in each group [60, 15], we set $\alpha = 0$ in DCA-$\ell_0/\ell_{1,0}$ to avoid performing tuning this parameter. From the proof of Proposition 3, δ can be computed by $\delta = \min_{x\in C} x$, where C is defined by (55) with $\{\tilde{\mathbf{\Sigma}}\} = \{\text{diag}(\mathbf{S})\}$. We therefore compute δ using the Newton method.

5.3. Experiments on synthetic datasets

We evaluate the performance of the proposed algorithms on four synthetic networks whose the details are given as follows. For the first class, the covariance matrix is a block diagonal matrix $\mathbf{\Sigma}^{(1)} = \text{diag}(\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_{10})$ with 10 blocks $\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_{10}$. For class 2, we create the covariance matrix $\mathbf{\Sigma}^{(2)}$ by setting

$\mathbf{\Sigma}^{(2)} = \mathbf{\Sigma}^{(1)}$ and resetting one of its sub-network blocks to the identity matrix. Similarly, for the k-th class with $3 \leq k \leq Q$, we set $\mathbf{\Sigma}^{(k)} = \mathbf{\Sigma}^{(k-1)}$ and reset an additional sub-network block to the identity matrix. We consider four types of sub-network blocks corresponding to the four synthetic networks.

M1: each block $\mathbf{\Sigma}_m$ is zero except elements in the last row and the last column.

M2: the blocks $\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_{10}$ are dense matrices.

M3: for $m = 1, ..., 10$ we take $(\mathbf{\Sigma}_m)_{ij} = (\mathbf{\Sigma}_m)_{ji}$ to be nonzero with the probability 0.02, independently of the other elements.

M4: for $m = 1, ..., 10$ we generate $(\mathbf{\Sigma}_m)_{i,i-1} = (\mathbf{\Sigma}_m)_{i-1,i}$ to be nonzero for $i = 2, ..., d$.

In the first three cases, the nonzero entries of matrices $\mathbf{\Sigma}^{(k)}, k = 1, ..., Q$ are randomly drawn in the set $\{+1, -1\}$. In Model 4, all nonzero values set to be 0.4. Finally, for each class k we generate independently, identically distributed observations $\mathbf{X}^{(k)} = [\mathbf{x}_1^k, ..., \mathbf{x}_{n_k}^k]$ from an $\mathcal{N}(0, \mathbf{\Sigma}^{(k)})$ distribution. In this experiment, for each model, we generate a training set and a validation set with size $n_1 = n_2 = n_3 = 150$, $d = 100$ and $Q = 3, 5$. We also consider other sizes $n_1 = n_2 = n_3 = 150 < d = 200$ and $Q = 3, 5$ in the first model.

The selection procedure of parameters on synthetic datasets is described as follows. We denote by $\{\hat{\mathbf{\Sigma}}_{\lambda,\alpha,\theta,\epsilon}\} = \{\hat{\mathbf{\Sigma}}^{(1)}_{\lambda,\alpha,\theta,\epsilon}, ..., \hat{\mathbf{\Sigma}}^{(Q)}_{\lambda,\alpha,\theta,\epsilon}\}$ the estimates of the multiple covariance matrices $\{\mathbf{\Sigma}\} = \{\mathbf{\Sigma}^{(1)}, ..., \mathbf{\Sigma}^{(Q)}\}$ from the training set with parameters $\lambda, \alpha, \theta, \epsilon$. Let $\{\mathbf{S}^{(1)}_{\text{valid}}, ..., \mathbf{S}^{(Q)}_{\text{valid}}\}$ be the sample covariance matrices computed from the validation set. We choose the best parameters $\hat{\lambda}, \hat{\alpha}, \hat{\theta}, \hat{\epsilon}$ by

$$
\begin{aligned}
&(\hat{\lambda}, \hat{\alpha}, \hat{\theta}, \hat{\epsilon}) \\
&= \arg\max_{\lambda,\alpha,\theta,\epsilon} \sum_{k=1}^{Q} -n_k[\log\det \hat{\mathbf{\Sigma}}^{(k)}_{\lambda,\alpha,\theta,\epsilon} + \text{tr}((\hat{\mathbf{\Sigma}}^{(k)}_{\lambda,\alpha,\theta,\epsilon})^{-1}\mathbf{S}^{(k)}_{\text{valid}})].
\end{aligned}
$$

To evaluate the performance of each method, we consider three loss functions which are the average root-mean-square error (ARMSE), the average entropy

loss (AEN) and the average Kullback-Leibler loss (AKL), respectively.

$$\begin{aligned}
\text{ARMSE} &= \frac{1}{Q}\sum_{k=1}^{Q}\frac{||\hat{\mathbf{\Sigma}}^k-\mathbf{\Sigma}^{(k)}||_F}{d},\\
\text{AEN} &= \frac{1}{Q}\sum_{k=1}^{Q}[-\log\det(\hat{\mathbf{\Sigma}}^{(k)}\Theta^{(k)})+\text{tr}(\hat{\mathbf{\Sigma}}^{(k)}\Theta^{(k)})-d],\\
\text{AKL} &= \frac{1}{Q}\sum_{k=1}^{Q}[-\log\det(\hat{\Theta}^{(k)}\mathbf{\Sigma}^{(k)})+\text{tr}(\hat{\Theta}^{(k)}\mathbf{\Sigma}^{(k)})-d],
\end{aligned}$$

where $\hat{\mathbf{\Sigma}}^{(k)}$ is a sparse estimate of the covariance matrix $\mathbf{\Sigma}^{(k)}$, $\Theta^{(k)}=(\mathbf{\Sigma}^{(k)})^{-1}$, and $\hat{\Theta}^{(k)}=(\hat{\mathbf{\Sigma}}^{(k)})^{-1}$. In terms of the sparsity, we also consider three metrics including the average false positive rate (the proportion of true zero elements estimated as nonzero, denoted by FP), the average false negative rate (the proportion of nonzero elements estimated as zero, denoted by FN) and the common zeros error rate (the proportion of common zeros across $\mathbf{\Sigma}^{(1)},...,\mathbf{\Sigma}^{(Q)}$ estimated as nonzero, denoted by CZ). They are respectively defined by

$$\begin{aligned}
\text{FP} &= \frac{1}{Q}\sum_{k=1}^{Q}\frac{\sum_{ij}\mathbf{I}(\Sigma_{ij}^{(k)}=0,\hat{\Sigma}_{ij}^{(k)}\neq 0)}{\sum_{ij}\mathbf{I}(\Sigma_{ij}^{(k)}=0)},\\
\text{FN} &= \frac{1}{Q}\sum_{k=1}^{Q}\frac{\sum_{ij}\mathbf{I}(\mathbf{\Sigma}_{ij}^{(k)}\neq 0,\hat{\Sigma}_{ij}^{(k)}=0)}{\sum_{ij}\mathbf{I}(\Sigma_{ij}^{(k)}\neq 0)},\\
\text{CZ} &= \frac{\sum_{ij}\mathbf{I}(\sum_{k=1}^{Q}|\Sigma_{ij}^{(k)}|=0,\sum_{k=1}^{Q}|\hat{\Sigma}_{ij}^{(k)}|\neq 0)}{\sum_{ij}\mathbf{I}(\sum_{k=1}^{Q}|\Sigma_{ij}^{(k)}|=0)}.
\end{aligned}$$

The numerical results including ARMSE, AEN, AKL, FP, FN, CZ, the training time in seconds and their standard deviations over 10 runs are reported in Table 1-3. The numbers in parentheses are standard deviations. Bold fonts indicate the best result in each row. Tables 1 and 2 consider the data in the same dimension ($d=100$, $n_k=150$) with different numbers of classes ($Q=3$ in Table 1 and $Q=5$ in Table 2), while Table 3 consider the data with a larger dimension ($d=200>n_k=150$).

Table 1: Numerical results of six comparative algorithms on synthetic datasets with $Q = 3, d = 100$ and $n_k = 150$.

		DCA-$\ell_0 - \ell_{1,0}$	DCA-$\ell_0 - \ell_{2,0}$	DCA-$\ell_{0,0}$	DCA-$\ell_1 - \ell_{2,1}$	DCA-w-$\ell_{0,0}$	PGD-$\ell_{0,0}$
	ARMSE	**0.0381** (0.0017)	0.0597 (0.0001)	0.0385 (0.0008)	0.0715 (0.0019)	0.0471 (0.0007)	0.0406 (0.0011)
	AEN	**5.0686** (0.111)	5.7574 (0.0887)	7.5534 (0.8193)	7.9095 (0.9251)	7.48 (0.71)	7.6171 (1.1659)
	AKL	**1.9985** (0.009)	10.391 (0.392)	2.017 (0.001)	15.4362 (2.5568)	4.2695 (0.4215)	2.11 (0.1049)
M1	FP	**0.0018** (0)	0.1991 (0.0078)	0.0023 (0)	0.3257 (0.0466)	0.0788 (0.007)	0.004 (0.0002)
	FN	0.0185 (0)	**0** (0)	0.0166 (0.0026)	**0** (0)	0.0166 (0.0078)	0.0166 (0.0026)
	CZ	**0.0009** (0.0001)	0.3053 (0.0171)	0.0012 (0.0002)	0.509 (0.0689)	0.0882 (0.0078)	0.0044 (0.001)
	Time	**168.52** (12.13)	227.985 (10.26)	201.69 (28.2)	179.41 (16.45)	479.36 (55.65)	1982.8 (29.3)
	ARMSE	0.2696 (0.0045)	**0.2449** (0.0028)	0.283 (0.0061)	0.2544 (0.0018)	0.3059 (0.0002)	0.4028 (0.0114)
	AEN	19.7197 (0.079)	**8.2639** (0.3776)	15.7342 (0.9)	9.5852 (0.455)	16.0576 (0.6891)	18.1014 (1.7745)
	AKL	21.2054 (1.5861)	**10.3973** (0.2968)	23.9448 (2.1152)	16.0781 (1.1286)	28.8 (1.3757)	24.4039 (1.3741)
M2	FP	0.1321 (0.0037)	**0.0333** (0)	0.1534 (0.0023)	0.2441 (0.0191)	0.154 (0.002)	0.1448 (0.0115)
	FN	**0.0788** (0.0067)	0.363 (0.0073)	0.0801 (0.0346)	0.123 (0.0044)	0.09571 (0.0017)	0.0881 (0.004)
	CZ	0.1454 (0.0039)	**0.0618** (0.0014)	0.1677 (0.0012)	0.4013 (0.0329)	0.3461 (0.0756)	0.1635 (0.0122)
	Time	202.35 (2.69)	222.4 (1.22)	197.86 (8.76)	**175.2** (2.76)	343.93 (12.47)	2069.73 (7.09)

	ARMSE	**0.0297** (0.0002)	0.0469 (0.0026)	0.0841 (0.033)	0.0402 (0.0004)	0.0965 (0.0009)	0.0899 (0.1057)
	AEN	**4.2585** (0.2389)	11.2 (1.4542)	6.1743 (1.7627)	31.8804 (7.8063)	10.5121 (0.1209)	38.3836 (8.77)
	AKL	6.9788 (0.5305)	24.539 (4.7554)	2.4683 (0.1796)	5.7352 (0.5543)	7.0199 (0.5774)	**1.6314** (0.5153)
M3	FP	0.1688 (0.0102)	0.4943 (0.0355)	0.0002 (0)	**0** (0)	0.0055 (0.0006)	0.0005 (0.0005)
	FN	**0** (0)	**0** (0)	**0** (0)	0.0833 (0)	**0** (0)	**0** (0)
	CZ	0.4136 (0.0204)	0.717 (0.043)	**0.0002** (0)	0.0003 (0)	0.0055 (0.0006)	0.0008 (0.0005)
	Time	**116.7** (0.26)	229.74 (0.19)	188.43 (99.26)	182.63 (2.8)	141.71 (13.04)	984.8 (88.91)
	ARMSE	0.0174 (0.0002)	0.0415 (0.0001)	0.0175 (0.0012)	0.0412 (0.0002)	0.0709 (0.0028)	**0.0154** (0.001)
	AEN	7.175 (1.5333)	10.8688 (0.5784)	9.4899 (1.2251)	11.1003 (2.1296)	17.9928 (2.7433)	**2.2626** (0.1516)
	AKL	3.9272 (0.5571)	13.634 (0.2109)	4.2156 (0.5352)	13.4999 (0.3437)	9.3015 (3.1669)	**2.8178** (0.1516)
M4	FP	0.0041 (0.0001)	0.0391 (0.0012)	**0.0031** (0.0004)	0.0428 (0.0043)	0.046 (0.0003)	0.0205 (0.0018)
	FN	0.0171 (0.0026)	0.002 (0.0009)	0.0252 (0.0023)	0.0016 (0.0023)	0.0116 (0.0015)	**0** (0)
	CZ	0.0111 (0.0058)	0.0333 (0.0014)	**0.0083** (0.001)	0.0594 (0.0091)	0.0508 (0.0003)	0.0595 (0.0058)
	Time	168.64 (3.91)	226.56 (1.79)	**131.7** (1.03)	163.71 (21.76)	192.17 (14.6)	2150.57 (35)

Table 2: Numerical results of six comparative algorithms on synthetic datasets with $Q = 5, d = 100$ and $n_k = 150$.

Size			DCA-$\ell_0 - \ell_{1,0}$	DCA-$\ell_0 - \ell_{2,0}$	DCA-$\ell_{0,0}$	DCA-$\ell_1 - \ell_{2,1}$	DCA-w-$\ell_{0,0}$	PGD-$\ell_{0,0}$
		ARMSE	**0.044** (0.0018)	0.0599 (0.0001)	0.047 (0.0033)	0.0655 (0.0017)	0.0558 (0.0002)	0.0459 (0.001)
		AEN	4.8257 (0.2106)	6.609 (0.1578)	5.0285 (0.7967)	7.2929 (0.8934)	6.0004 (0.2104)	**3.9857** (1.1793)
		AKL	**1.9201** (0.041)	12.3379 (0.3393)	2.2122 (0.2547)	13.8089 (2.195)	5.8324 (0.2146)	2.0904 (0.1711)
	M1	FP	**0.011** (0.0012)	0.2854 (0.0333)	0.0127 (0.0038)	0.336 (0.0399)	0.03611 (0.0025)	0.013 (0.0017)
		FN	0.0088 (0)	**0** (0)	0.01 (0.0047)	**0** (0)	**0** (0)	0.0066 (0.0031)
		CZ	**0.0097** (0.001)	0.5234 (0.0736)	0.0114 (0.0033)	0.6221 (0.0679)	0.0348 (0.0013)	0.0137 (0.0011)
		Time	**203.75** (6.06)	386.17 (0.32)	228.43 (26.2)	302.11 (5.44)	553.13 (10.35)	2903.51 (498.58)
		ARMSE	0.2356 (0.0089)	**0.1958** (0.0056)	0.2408 (0.0028)	0.2049 (0.0076)	0.241 (0.0019)	0.2521 (0.0145)
		AEN	19.0316 (1.4859)	**7.0435** (0.3582)	19.3134 (1.0337)	8.1291 (0.1702)	21.3881 (1.3683)	14.835 (5.266)
		AKL	25.854 (1.0119)	**9.2226** (0.598)	23.5866 (3.2311)	11.5028 (0.1844)	24.7421 (2.5573)	27.7589 (1.8117)
	M2	FP	0.2474 (0.0491)	**0.0772** (0.0012)	0.2981 (0.0041)	0.1611 (0.016)	0.2812 (0.0245)	0.252 (0.0054)
		FN	0.0275 (0.0048)	0.1647 (0.0092)	**0.0188** (0.0042)	0.0861 (0.0232)	0.0213 (0.0032)	0.0346 (0.0048)
		CZ	0.2861 (0.0501)	**0.1555** (0)	0.454 (0.0122)	0.3225 (0.035)	0.3465 (0.0392)	0.2963 (0.0114)
		Time	**251.24** (6.58)	380.7 (2.69)	284.31 (3.12)	297.87 (2.6)	498.9 (118.43)	3476.26 (124.56)

	ARMSE	0.0105 (0.0002)	0.0214 (0.0004)	**0.0101** (0.0004)	0.0253 (0.0005)	0.106 (0.001)	0.0107 (0.0001)
	AEN	0.5554 (0.0282)	20.9791 (6.4251)	**0.5059** (0.0314)	6.6962 (6.3432)	0.5403 (0.0188)	0.6253 (0.03296)
	AKL	0.5782 (0.0246)	2.7441 (0.2438)	**0.523** (0.02)	3.1728 (0.5355)	0.5589 (0.0257)	0.6594 (0.0442)
M3	FP	**0.0002** (0.0001)	0.0024 (0.0003)	0.0013 (0.0001)	0.0029 (0.0002)	0.0003 (0)	0.0061 (0.0004)
	FN	**0** (0)	0.0285 (0.0001)	**0** (0)	**0** (0)	**0** (0)	**0** (0)
	CZ	0.0069 (0.0004)	0.0052 (0.0005)	**0.0041** (0.0007)	0.0061 (0.0001)	0.0549 (0.0007)	0.0256 (0.0025)
	Time	38.93 (8.44)	253. 72 (43.97)	**16.87** (3.42)	305.03 (2.01)	30.54 (2.45)	764.12 (14.46)
	ARMSE	0.0145 (0.0004)	0.0397 (0.001)	0.0128 (0.0003)	0.0386 (0.0005)	0.0281 (0.0209)	**0.0122** (0.0006)
	AEN	5.6426 (0.8228)	12.5991 (3.9518)	2.6486 (0.1344)	10.6253 (1.6804)	3.9778 (0.6436)	**1.5349** (0.0561)
	AKL	2.8214 (0.1954)	12.3653 (1.0526)	2.0312 (0.2397)	11.5881 (0.5695)	4.9378 (1.7801)	**1.7578** (0.0894)
M4	FP	**0.0026** (0.0003)	0.0398 (0.0002)	0.0037 (0.0008)	0.0358 (0.0002)	0.0574 (0.0054)	0.0148 (0.008)
	FN	0.0113 (0.0013)	0.0154 (0.0087)	0.005 (0.0012)	0.0036 (0.0022)	0.0137 (0.0029)	**0** (0)
	CZ	**0.0071** (0.0007)	0.0667 (0.002)	0.0114 (0.003)	0.0587 (0.0011)	0.0631 (0.0034)	0.0644 (0.0001)
	Time	180.36 (1.76)	320.65 (94.63)	**144.58** (4.75)	305.36 (4.29)	541.51 (64.16)	3553.1 (58.04)

Table 3: Numerical results of six comparative algorithms for the M1 model with $Q = 3$, $Q = 5$ and $d = 200 > n_k = 150$.

		DCA-$\ell_0 - \ell_{1,0}$	DCA-$\ell_0 - \ell_{2,0}$	DCA-$\ell_{0,0}$	DCA-$\ell_1 - \ell_{2,1}$	DCA-w-$\ell_{0,0}$	PGD-$\ell_{0,0}$
$Q = 3$	ARMSE	**0.0426** (0)	0.0894 (0.0321)	0.0525 (0.0158)	0.1224 (0.0111)	0.0629 (0.0216)	0.0659 (0.0351)
	AEN	**18.4655** (2.5195)	36.2179 (26.758)	26.8817 (6.8868)	32.4057 (3.6644)	30.76 (4.2259)	36.5135 (9.2809)
	AKL	**4.1365** (1.1152)	30.1647 (2.4198)	8.3445 (6.9359)	81.0891 (10.110)	10.386 (4.7861)	15.4805 (6.9634)
	FP	**0.0049** (0.0005)	0.2086 (0.0294)	0.0117 (0.0104)	0.5835 (0.0597)	0.0284 (0.004)	0.02436 (0.0026)
	FN	0.0185 (0.0113)	0.0408 (0.0246)	0.0184 (0.0012)	**0** (0)	**0** (0)	0.0114 (0.0062)
	CZ	**0.0045** (0.0003)	0.03391 (0.01479)	0.0119 (0.0112)	0.822 (0.0606)	0. 0247 (0.0072)	0.0259 (0.0028)
	Time	**847.57** (57.81)	1190.57 (67.49)	1142.47 (67.06)	1042.32 (10.87)	5551.18 (45.91)	10937.21 (389.67)
$Q = 5$	ARMSE	**0.0564** (0.0012)	0.1238 (0.027)	0.0832 (0.0132)	0.1276 (0.0055)	0.0918 (0.0193)	0.0824 (0.0282)
	AEN	**9.5855** (3.6704)	68.0489 (3.7496)	30.1689 (6.3052)	39.4 (2.4412)	30.0735 (5.1566)	22.4321 (10.1848)
	AKL	**5.1279** (0.7131)	37.3221 (13.5456)	17.1981 (5.8622)	107.1191 (7.0967)	21.0489 (2.1294)	15.3805 (5.1433)
	FP	**0.0217** (0.0009)	0.5201 (0.01666)	0.0687 (0.0061)	0.6558 (0.0243)	0.0805 (0.002)	0.0646 (0.0503)
	FN	0.0083 (0.0047)	**0** (0)	0.01 (0.0008)	**0** (0)	0.0241 (0.0001)	0.01 (0.0022)
	CZ	**0.0224** (0.0015)	0.8699 (0.01233)	0.0738 (0.0271)	0.9613 (0.01)	0.0951 (0.0038)	0.0705 (0.0159)
	Time	1779.09 (20.64)	2066.66 (6.92)	**1448.37** (187.34)	1734.31 (0.601)	1012.81 (105.27)	18334.11 (170.87)

We observe from Table 1 that DCA-$\ell_0 - \ell_{1,0}$ gives the best results in terms of all three losses on the model M1. It also provides the best results in terms of two out of three losses on the M3. However, on the M2, DCA-$\ell_0 - \ell_{2,0}$ gives the best results in terms of all three losses. PGD-$\ell_{0,0}$ obtains the best results in terms of all three losses on the model M4 and it is also the best in terms of the Kullback-Leibler loss on the M3. DCA-$\ell_{0,0}$ achieves quite good results on the all models. In general, DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ are better than DCA-$\ell_0 - \ell_{2,0}$, DCA-$\ell_1 - \ell_{2,1}$, DCA-w-$\ell_{0,0}$, and PGD-$\ell_{0,0}$ in terms of the false positive, the false negative and the common zeros rates while DCA-$\ell_0 - \ell_{2,0}$ is better than DCA-$\ell_1 - \ell_{2,1}$. In terms of running time, the DCA based algorithms are faster than PGD-$\ell_{0,0}$.

Table 2 provides numerical results with $Q = 5$ classes and the same dimension ($d = 100$ and $n_k = 150$). In Table 2, in terms of three losses, DCA-$\ell_0 - \ell_{1,0}$, DCA-$\ell_{0,0}$, DCA-w-$\ell_{0,0}$, and PGD-$\ell_{0,0}$ are comparable and better than DCA-$\ell_0 - \ell_{2,0}$ and DCA-$\ell_1 - \ell_{2,1}$ on the three models 1, 3 and 4. More precisely, DCA-$\ell_0 - \ell_{1,0}$, DCA-$\ell_{0,0}$, and PGD-$\ell_{0,0}$ are the best on the models M1, M3 and M4, respectively. DCA-$\ell_0 - \ell_{2,0}$ achieves the best results in terms of the three losses on the M2. Similar to the previous experiment, DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ are most of the time better than DCA-$\ell_0 - \ell_{2,0}$, DCA-$\ell_1 - \ell_{2,1}$, DCA-w-$\ell_{0,0}$, and PGD-$\ell_{0,0}$ in terms of the false positive, the false negative and the common zeros rates while DCA-$\ell_0 - \ell_{2,0}$ is better than DCA-$\ell_1 - \ell_{2,1}$. Regarding the training time, DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ are faster than DCA-$\ell_0 - \ell_{2,0}$ and DCA-$\ell_1 - \ell_{2,1}$ on all four models, especially, on the M3. This can be explained by the fact that the sequences of convex sub-problems in DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ can be decomposed into the independent smaller sub-problems. The DCA based algorithms run faster than PGD-$\ell_{0,0}$.

Table 3 provides numerical results on synthetic datasets for the M1 with $Q = 3$ and $Q = 5$ classes and the dimension $d = 200 > n_k = 150$. Concerning Table 3, we observe that DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ are comparable and better than DCA-$\ell_0 - \ell_{2,0}$, DCA-$\ell_1 - \ell_{2,1}$, DCA-w-$\ell_{0,0}$ and PGD-$\ell_{0,0}$ in terms of most of the comparison criteria. In particular, DCA-$\ell_0 - \ell_{1,0}$ is slightly better than

DCA-$\ell_{0,0}$.

5.4. Experiments on real datasets

We illustrate the use of the estimation of multiple sparse covariance matrices via a real application: a classification problem based Quadratic Discriminant Analysis (QDA). This application requires the estimation of the covariance matrices. We assume that n_k observations $\mathbf{x}_i^k$ $(i = 1, ..., n_k)$ within the k-th class C_k are sampled from $\mathcal{N}(\mu_\mathbf{k}, \mathbf{\Sigma}^{(k)})$. We denote the prior probability of the k-th class by π_k. The quadratic discriminant function is

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log\det\mathbf{\Sigma}^{(k)} - \frac{1}{2}(\mathbf{x}-\mu_k)^T(\mathbf{\Sigma}^{(k)})^{-1}(\mathbf{x}-\mu_k) + \log\pi_k.$$

Then the predicted class for a new observation $\mathbf{x}$ is $\arg\max_k \delta_k(\mathbf{x})$. In practice we do not know $\pi_k, \mu_k, \mathbf{\Sigma}^{(k)}$, and will need to estimate them using the training data.

We evaluate the proposed algorithms on seven datasets from UCI Machine Learning Repository[1] (Wine, Vehicle, Vowel, Satimage, Waveform 2, Optical Recognition of Handwritten Digits, and Semeion Handwritten Digit). We use the cross-validation scheme to validate the performance of various approaches on these seven datasets. Each dataset is split into a training set containing 2/3 of the samples and a test set containing 1/3 of the samples. This process is repeated 10 times, each with a random choice of training set and test set. The tuning parameters are chosen via 5-fold cross-validation to achieve the lowest testing error.

Table 4: Numerical results of 7 comparative algorithms on real datasets.

[1] `https://archive.ics.uci.edu/ml/datasets`

Data	Algorithm	Testing error (%)	Training error (%)	Time (seconds)
Wine	DCA-$\ell_0 - \ell_{1,0}$	2.15 (0.6)	1.41 (0.29)	4.99 (3.76)
	DCA-$\ell_0 - \ell_{2,0}$	2.25 (0.97)	4.2 (0.52)	5.04 (5.52)
	DCA-$\ell_{0,0}$	**2.05** (1.69)	3.36 (2.9)	**2.13** (6.02)
	DCA-$\ell_1 - \ell_{2,1}$	2.22 (0.54)	3.67 (1.29)	2.8 (0.74)
	DCA-w-$\ell_{0,0}$	3.38 (0.91)	2.52 (0.84)	6.65 (1.35)
	PGD-$\ell_{0,0}$	3.77 (0.46)	1.83 (0.84)	63.91 (10.51)
	QDA	5.55 (0.92)	**0.84** (0.47)	-
Vehicle	DCA-$\ell_0 - \ell_{1,0}$	16.31 (1.77)	11.17 (1.11)	**0.07** (0.02)
	DCA-$\ell_0 - \ell_{2,0}$	16.07 (1.63)	9.36 (1.54)	0.08 (0.01)
	DCA-$\ell_{0,0}$	**15.95** (0.38)	10.1 (0.97)	**0.07** (0.01)
	DCA-$\ell_1 - \ell_{2,1}$	16.66 (1.63)	11.23 (0.1)	**0.07** (0.01)
	DCA-w-$\ell_{0,0}$	16.64 (1.99)	9.87 (1.51)	0.11 (0.02)
	PGD-$\ell_{0,0}$	15.68 (1.86)	10 (0.98)	1.2 (0.02)
	QDA	18.33 (1.64)	**8.91** (1.08)	-
Vowel	DCA-$\ell_0 - \ell_{1,0}$	20.3 (0.23)	12.87 (0.46)	27.04 (4.13)
	DCA-$\ell_0 - \ell_{2,0}$	19.69 (0.44)	13.7 (0.78)	32.25 (3.17)
	DCA-$\ell_{0,0}$	**18.48** (1.09)	**11.96** (0.92)	31.79 (3.35)
	DCA-$\ell_1 - \ell_{2,1}$	20.81 (1.66)	13.84 (1.34)	**26.82** (2.05)
	DCA-w-$\ell_{0,0}$	19.09 (1.4)	14.09 (1.62)	30.6 (0.4)
	PGD-$\ell_{0,0}$	19.88 (1.23)	14.07 (1.18)	528.44 (10.53)
	QDA	23.03 (0.3)	13.68 (1.91)	-
Satimage	DCA-$\ell_0 - \ell_{1,0}$	23.71 (1.13)	21.33 (0.22)	**30.85** (2.21)
	DCA-$\ell_0 - \ell_{2,0}$	**18.06** (1.35)	**15.68** (1.45)	59.73 (5.43)
	DCA-$\ell_{0,0}$	22.76 (0.83)	20.6 (0.82)	70.9 (32.08)
	DCA-$\ell_1 - \ell_{2,1}$	19.21 (0.33)	17.73 (0.93)	48.67 (2.14)
	DCA-w-$\ell_{0,0}$	23.04 (2.45)	21.04 (2.45)	123.38 (14.58)
	PGD-$\ell_{0,0}$	26.87 (1.48)	23.9 (1.7)	718.54 (6.45)
	QDA	39.04 (0.87)	35.63 (0.48)	-

Waveform 2	DCA-$\ell_0 - \ell_{1,0}$	**13.01** (0.25)	11.41 (1.3)	**3.28** (1.14)
	DCA-$\ell_0 - \ell_{2,0}$	14.64 (0.38)	12.68 (0.32)	12.5 (2.48)
	DCA-$\ell_{0,0}$	14.1 (0.59)	12.78 (0.23)	3.72 (1.79)
	DCA-$\ell_1 - \ell_{2,1}$	13.42 (0.94)	12.17 (0.93)	13.98 (2.79)
	DCA-w-$\ell_{0,0}$	15.14 (0.73)	11.96 (0.61)	14.49 (0.32)
	PGD-$\ell_{0,0}$	15.97 (1.44)	14.73 (1.97)	16.64 (3.17)
	QDA	16.42 (0.76)	**9.07** (0.31)	-
Optical	DCA-$\ell_0 - \ell_{1,0}$	**2.74** (0.31)	**1.92** (0.14)	**97.9** (10.43)
	DCA-$\ell_0 - \ell_{2,0}$	3.64 (0.18)	2.05 (0.39)	582.92 (22.75)
	DCA-$\ell_{0,0}$	2.98 (0.64)	2.15 (0.38)	106.49 (10.46)
	DCA-$\ell_1 - \ell_{2,1}$	3.98 (0.41)	3.18 (0.51)	574.66 (36.81)
	DCA-w-$\ell_{0,0}$	3.78 (0.15)	2.44 (0.5)	134.74 (15.79)
	PGD-$\ell_{0,0}$	3.93 (0.45)	2.78 (0.24)	2792.21 (6.62)
	QDA	4.2 (0.47)	2.49 (0.21)	-
Semeion	DCA-$\ell_0 - \ell_{1,0}$	5.28 (0.27))	1.97 (0.76)	**167.85** (22.71)
	DCA-$\ell_0 - \ell_{2,0}$	**4.89** (0.76)	1.97 (0.67)	945.18 (68.32)
	DCA-$\ell_{0,0}$	5.24 (0.38)	2.51 (0.42)	264.94 (84.04)
	DCA-$\ell_1 - \ell_{2,1}$	6.02 (0.58)	1.01 (0.44)	947.56 (153.15)
	DCA-w-$\ell_{0,0}$	5.71 (0.93)	2.12 (0.38)	310.58 (25.12)
	PGD-$\ell_{0,0}$	5.49 (0.97)	1.78 (0.53)	9456.58 (152.46)
	QDA	6.26 (0.43)	**0** (0)	-

The computational results of the six comparative algorithms in the previous experiment and QDA are reported in Table 4. As before, the numbers in parentheses are standard deviations, and bold fonts indicate the best result in each row.

We observe from Table 4 that the DCA based algorithms and PGD-$\ell_{0,0}$ give better testing errors than QDA on all datasets and they also achieve better training errors than QDA on three out of seven datasets. DCA-$\ell_0 - \ell_{1,0}$ and

DCA-$\ell_{0,0}$ outperform DCA-$\ell_0 - \ell_{2,0}$, DCA-$\ell_1 - \ell_{2,1}$, DCA-w-$\ell_{0,0}$, and PGD-$\ell_{0,0}$ in terms of the testing error on 5/7 datasets. DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ are also better than DCA-$\ell_0 - \ell_{2,0}$, DCA-$\ell_1 - \ell_{2,1}$, DCA-w-$\ell_{0,0}$ and PGD-$\ell_{0,0}$ in terms of the training on 4/7 datasets. More precisely, in terms of the testing error, DCA-$\ell_{0,0}$ obtains better performances than DCA-$\ell_0 - \ell_{1,0}$ on 6/7 datasets while DCA-$\ell_0 - \ell_{2,0}$ is better than DCA-$\ell_1 - \ell_{2,1}$ on 5/7 datasets. As for the training time, DCA-$\ell_0 - \ell_{1,0}$ is the fastest on 5/7 datasets, and we further note that the DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ are significantly faster than the other approaches on the last three datasets (Waveform 2, Optical and Semeion).

6. Conclusion

In a unifying DC programming framework, we have studied the bi-level variable selection in a learning problem whose the loss function is DC (possibly non-differentiable) while the regularization term dealing with sparsity is the $\ell_0 + \ell_{q,0}$ and/or $\ell_{0,0}$ mixed norm. Using a general DC approximate function (possibly non-differentiable) of the step function that covers all existing approximations as special cases, we have shown that the $\ell_0 + \ell_{q,0}$ and $\ell_{0,0}$ approximations can be expressed as DC functions. Then the two resulting approximate problems have been reformulated as DC programs for which two DCA schemes have been investigated. We have offered efficient DCA based approaches for tackling several difficult problems/real applications The proposed algorithms have been successfully applied to the bi-level variable selection in multiple sparse covariance matrices estimation problem. Numerical experiments on both simulation and real datasets have shown that DCA-$\ell_0 - \ell_{1,0}$ and DCA-$\ell_{0,0}$ have obtained the best performance on most of the comparison criteria, and are the fastest algorithms.

In the future, we will study the bi-level variable selection for other applications. We also study the combination of DCA and other methods for the problem of multiple sparse covariance matrices estimation.

References

[1] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Optimization with sparsity-inducing penalties, Found. Trends Mach. Learn. 4 (1) (2012) 1–106.

[2] H. A. Le Thi, T. Pham Dinh, H. M. Le, X. Vo, DC approximation approaches for sparse optimization, European Journal of Operational Research 244 (2015) 26–44.

[3] T. T. Wu, K. Lange, Coordinate descent algorithms for lasso penalized regression, Ann. Appl. Stat. 2 (1) (2008) 224–244.

[4] J. Friedman, T. Hastie, R. Tibshirani, A note on the group lasso and sparse group lasso, arXiv:1001.0736v1 (2010) 1–8.

[5] M. Vincent, N. R. Hansen, Sparse group lasso and high dimensional multinomial classification, Computational Statistics and Data Analysis 71 (2014) 771–786.

[6] P. Danaher, P. Wang, D. M. Witten, The joint graphical lasso for inverse covariance estimation across multiple classes, J. R. Statist. Soc. B 76 (2014) 373–397.

[7] S. Hara, T. Washio, Learning a common substructure of multiple graphical gaussian models, Neural Networks 38 (2013) 23 – 38.

[8] S. Yang, Z. Lu, X. Shen, P. Wonka, J. Ye, Fused multiple graphical lasso, SIAM Journal on Optimization 25 (2) (2015) 916–943.

[9] S. Yang, Z. Lu, X. Shen, P. Wonka, J. Ye, Fused multiple graphical lasso, SIAM Journal on Optimization 25 (2) (2015) 916–943.

[10] T. Saegusa, A. Shojaie, Joint estimation of precision matrices in heterogeneous populations, Electron. J. Statist. 10 (1) (2016) 1341–1392.

[11] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse-group lasso, Journal of Computational and Graphical Statistics 22 (2013) 231–245.

[12] P. Breheny, J. Huang, Penalized methods for bi-level variable selection, Statistics and its interface 2 (2009) 369–380.

[13] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2) (2010) 894–942.

[14] J. Huang, S. Ma, H. Xie, C.-H. Zhang, A group bridge approach for variable selection, Biometrika 96 (2) (2009) 339–355.

[15] P. Breheny, The group exponential lasso for bi-level variable selection, Biometrics 71 (3) (2015) 731–740.

[16] H. A. Le Thi, H. M. Le, V. V. Nguyen, T. Pham Dinh, A DC programming approach for feature selection in support vector machines learning, Adv Data Anal Classif 2 (3) (2008) 259–278.

[17] H. A. Le Thi, M. Moeini, T. Pham Dinh, Portfolio selection under downside risk measures and cardinality constraints based on DC programming and DCA, Computational Management Science 6 (4) (2009) 459–475.

[18] H. A. Le Thi, X. Vo, T. Pham Dinh, Feature selection for linear svms under uncertain data: robust optimization based on difference of convex functions algorithms, Neural Networks 59 (2014) 36–50.

[19] C. S. Ong, H. A. Le Thi, Learning sparse classifiers with difference of convex functions algorithms, Optim Method Softw 28 (4) (2013) 830–854.

[20] H. Le Thi, D. Phan, DC programming and DCA for sparse optimal scoring problem, Neurocomputing 186 (2016) 170 – 181.

[21] H. A. Le Thi, D. N. Phan, DC programming and DCA for sparse fisher linear discriminant analysis, Neural Computing and Applications 28 (9) (2017) 2809–2822.

[22] D. N. Phan, H. A. Le Thi, T. Pham Dinh, Sparse covariance matrix estimation by DCA-based algorithms, Neural Computation 29 (11) (2017) 3040–3077.

[23] H. A. Le Thi, T. Pham Dinh, The DC (Difference of Convex functions) programming and DCA revisited with DC models of real world nonconvex pptimization problems, Annals of Operations Research 133 (2005) 23–46.

[24] T. Pham Dinh, H. A. Le Thi, Convex analysis approach to D.C. programming: Theory, algorithms and applications, Acta Mathematica Vietnamica 22 (1) (1997) 289–355.

[25] T. Pham Dinh, H. A. Le Thi, A DC optimization algorithm for solving the trust-region subproblem, SIAM Journal of Optimization 8 (2) (1998) 476–505.

[26] H. A. Le Thi, T. Pham Dinh, DC programming and DCA: thirty years of developments, Mathematical Programming 169 (1) (2018) 5–68.

[27] T. Pham Dinh, H. A. Le Thi, Recent advances in DC programming and DCA, Transactions on Computational Collective Intelligence 8342 (2014) 1–37.

[28] D. Peleg, R. Meir, A bilinear formulation for vector sparsity optimization, Signal Processing 88 (2) (2008) 375–389.

[29] R. T. Rockafellar, Convex Analysis, Princeton University Press, N.J., 1970.

[30] P. Mahey, T. Pham Dinh, Partial regularization of the sum of two maximal monotone operators, ESAIM-Math Model Num 27 (3) (1993) 375–392.

[31] P. Mahey, S. Oualibouch, T. Pham Dinh, Proximal decomposition on the graph of a maximal monotone operator, SIAM J Optimz 5 (1995) 454–466.

[32] H. A. Le Thi, V. N. Huynh, T. Pham Dinh, Convergence analysis of dc algorithm for dc programming with subanalytic data, Journal of Optimization Theory and Applications 179 (2018) 103–126.

[33] H. A. Le Thi, V. T. Ho, Online learning based on online DCA and application to online classification, Neural Computation 32 (4) (2020) 759–793.

[34] H. A. Le Thi, H. M. Le, D. N. Phan, B. Tran, Stochastic DCA for minimizing a large sum of DC functions with application to multi-class logistic regression, Neural Networks 132 (2020) 220–231.

[35] Y. Xu, Q. Qi, Q. Lin, R. Jin, T. Yang, Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 6942–6951.

[36] D. N. Phan, H. M. Le, H. A. Le Thi, Accelerated difference of convex functions algorithm and its application to sparse binary logistic regression, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 1369–1375.

[37] F. J. Aragón Artacho, P. T. Vuong, The boosted difference of convex functions algorithm for nonsmooth functions, SIAM Journal on Optimization 30 (1) (2020) 980–1006.

[38] H. A. Le Thi, H. M. Le, D. N. Phan, B. Tran, Novel DCA based algorithms for a special class of nonconvex problems with application in machine learning, Applied Mathematics and Computation (2020) 1–22doi:10.1016/j.amc.2020.125904.

[39] T. Liu, T. K. Pong, A. Takeda, A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems, Mathematical Programming 176 (1) (2019) 339–367.

[40] P. S. Bradley, O. L. Mangasarian, Feature selection via concave minimization and support vector machines, in: Proceeding of international conference on machine learning ICML'98, 1998.

[41] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Stat. Ass. 96 (456) (2001) 1348–1360.

[42] O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, Journal of Empirical Finance 10 (2003) 603–621.

[43] O. Ledoit, M. Wolf, Honey, i shrunk the sample covariance matrix, J. Portfolio Management 30 (2004) 110–119.

[44] R. Jagannathan, T. Ma, Risk reduction in large portfolios: Why imposing the wrong constraints helps, Journal of Finance 58 (2003) 1651–1684.

[45] Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, Biostatistics 8 (1) (2007) 86–100.

[46] J. Leek, J. Storey, A general framework for multiple testing dependence, Proc. Natn. Acad. Sci. USA 105 (2008) 18718–18723.

[47] J. Friedman, T. J. Hastie, R. J. Tibshirani, Sparse inverse covariance estimationwith the graphical lasso, Biostatistics 9 (2008) 432–441.

[48] A. Rothman, E. Levina, J. Zhu, Sparse permutation invariant covariance estimation, Electron. J. Statist. 2 (2008) 494–515.

[49] M. Yuan, Y. Lin, Model selection and estimation in the gaussian graphical model, Biometrika 94 (2007) 19–35.

[50] O. Banerjee, L. E. Elghaoul, A. D'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data, J. Mach. Learn. Res. 9 (2008) 485–516.

[51] J. Bien, R. Tibshirani, Sparse estimation of a covariance matrix, Biometrika 98 (4) (2011) 807–820.

[52] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, The Annals of Statistics 37 (2009) 4254–4278.

[53] H. Liu, L. Wang, T. Zhao, Sparse covariance matrix estimation with eigenvalue contraints, Journal of Computational and Graphical Statistics 23 (2) (2014) 439–459.

[54] A. J. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, J. Am. Statist. Assoc. 104 (2009) 177–186.

[55] A. Rothman, Positive definite estimators of large covariance matrices, Biometrika 99 (2012) 733–740.

[56] L. Xue, S. Ma, H. Zuo, Positive-definite ℓ_1-penalized estimation of large covariance matrices, Journal of the American Statistical Association 107 (2012) 1480–1491.

[57] J. Guo, E. Levina, G. Michailidis, J. Zhu, Joint estimation of multiple graphical models, Biometrika 98 (1) (2011) 1–15.

[58] S. Boyd, N. Parikh, E. Chu, B. Peleato, P. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundat. Trends Mach. Learn. 3 (1) (2011) 1–122.

[59] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, J. Ye, A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, in: Proceedings of ICML'13 - Volume 28, ICML'13, JMLR.org, 2013, pp. II–37–II–45.

[60] J. Huang, P. Breheny, S. Ma, A selective review of group selection in high-dimensional models, Statist Sci. 27.