Nonconvex and Nonsmooth Optimization with Generalized Orthogonality Constraints

Zhu, Hong; Zhang, Xiaowei; Chu, Delin; Liao, Lizhi

Published in:
Journal of Scientific Computing

DOI:
10.1007/s10915-017-0359-1

Published: 01/07/2017

Document Version:
Peer reviewed version

J. Sci. Comput. manuscript No.
(will be inserted by the editor)

Nonconvex and Nonsmooth Optimization with Generalized Orthogonality Constraints: An Approximate Augmented Lagrangian Method

Hong Zhu · Xiaowei Zhang · Delin Chu · Li-Zhi Liao

Received: date / Accepted: date

Abstract Nonconvex and nonsmooth optimization problems with linear equation and generalized orthogonality constraints have wide applications. These problems are difficult to solve due to nonsmooth objective function and nonconvex constraints. In this paper, by introducing an extended proximal alternating linearized minimization (EPALM) method, we propose a framework based on the augmented Lagrangian scheme (EPALMAL). We also show that the EPALMAL method has global convergence in the sense that every bounded sequence generated by the EPALMAL method has at least one convergent subsequence that converges to the Karush-Kuhn-Tucker (KKT) point of the original problem. Experiments on a variety of applications, including compressed modes and multivariate data analysis, have demonstrated that the proposed method is noticeably efficient and achieves comparable performance with existing methods.

Keywords Generalized orthogonality constraints · Augmented Lagrangian scheme · Proximal alternating minimization method · Linearization.

Hong Zhu
Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.
E-mail: 13479652@life.hkbu.edu.hk

Xiaowei Zhang
Bioinformatics Institute, A*STAR, Singapore.
E-mail: zhangxw@bii.a-star.edu.sg

Delin Chu
Department of Mathematics, National University of Singapore, Singapore 19076.
E-mail: matchudl@nus.edu.sg

Li-Zhi Liao (Corresponding author)
Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.
E-mail: liliao@hkbu.edu.hk

1 Introduction

In this paper, we consider problems of the following form

$$\begin{aligned} \min_{X,Y} \ & \phi(X,Y) := f(X) + g(Y) + h(X,Y) \\ \text{s.t. } & AX + BY = C, \\ & X^T M X = I_q, \end{aligned} \tag{1}$$

where $X \in \mathbb{R}^{n\times q}$, $Y \in \mathbb{R}^{m\times q}$, $A \in \mathbb{R}^{l\times n}$, $B \in \mathbb{R}^{l\times m}$, $C \in \mathbb{R}^{l\times q}$, B has full row rank, and $M \in \mathbb{R}^{n\times n}$ is a given symmetric positive semi-definite matrix. Moreover, f and g are proper lower semicontinuous functions and h is a C^1 function (i.e., continuously differentiable).

As a special case of (1), problem

$$\min_{X\in\mathbb{R}^{n\times q}} \ h(X) \quad \text{s.t.} \quad X^T X = I, \tag{2}$$

with smooth $h(X)$ has attracted lots of research attentions, such as the orthonormal procrustes problem and the Penrose regression problem [14], linear eigenvalue problem [28, 49], nearest low-rank matrix problem [21], total energy minimization [20], etc.. Constraint set $St(q,n) := \{X \in \mathbb{R}^{n\times q} \mid X^T X = I\}$ is well-known as the Stiefel manifold. There are various optimization methods developed for problem (2), and most of them are feasible methods which preserve manifold structure of orthogonality constraints during iterations via geodesics or retractions [3]. For example, steepest descent gradient methods [1,39], conjugate gradient methods [2,3,18], Quasi-Newton methods [44], Newton methods [3,18], and trust region method [3,52], to name but a few. Efficient variants of retraction-based methods have also been developed by using properly designed retractions, see [26,39,50] and references therein. Recently, Zhang *et al*. [55,56] proposed a self-consistent-field (SCF) iteration scheme for the maximization of the sum of the trace ration on the Stiefel manifold based on the necessary condition for the global maximizers. There are also infeasible methods for problem (2), such as the augmented Lagrangian method [12] and splitting method [31].

The requirement of smoothness in the objective function of model (2) restricts its applicability, since for many applications in areas like machine learning, signal processing, and computer vision, nonsmooth regularization terms are often used to prevent overfitting, leading to problems of the form

$$\min_{X\in\mathbb{R}^{n\times q}} \ h(X) + g(X) \quad \text{s.t.} \quad X^T Q X = I, \tag{3}$$

where $g(X)$ is a convex but nonsmooth function and Q is positive definite matrix. Most of existing research focuses on the splitting method. For example, alternating direction method of multipliers (ADMM) or its variants have been proposed in [29,31,40] to solve (3). However, none of them provides convergence analysis. For the special case of problem (3), where $g(x) = 0$, $h(x)$ is twice continuously differentiable, the global convergence of ADMM was analyzed in [32]. Recently, Wang *et al*. [48] analyzed the global convergence of ADMM for nonconvex nonsmooth problem

$$\min_{x_1,\cdots,x_p,y} \ f(x_1,\cdots,x_p,y) \quad \text{s.t.} \quad A_1x_1 + \cdots + A_px_p + By = b,$$

which contains problem (3) as a special case, under some assumptions on the objective function and the coefficient matrices. Notice that the convergence result does not hold when Q in problem (3) is positive semi-definite matrix since in such case the manifold is not

compact any more. The inexact augmented Lagrangian method [8,34] is also used in some applications. By hybridizing the augmented Lagrangian method and the proximal alternating minimization scheme [6], Chen *et al*. [12] proposed a proximal alternating minimized augmented Lagrangian (PAMAL) method to solve problem

$$\min_{X\in\mathbb{R}^{n\times q}} g(X)+\mu\,\mathrm{tr}(X^THX) \quad \text{s.t.} \quad X^TQX=I_q,$$

where $g(X)$ is a nonsmooth convex function such as $\|X\|_1:=\sum_{i,j}|X_{ij}|$ or $\|X\|_{1,2}:=\sum_{i=1}^n(\sum_{j=1}^q X_{ij}^2)^{1/2}$, H is a symmetric matrix and Q is a positive definite matrix. Under mild assumptions, this method has sub-sequence convergence property in the sense that there exists at least one convergent subsequence and any convergent subsequence converges to a Karush-Kuhn-Tucker (KKT) point. However, in each inner iteration of this algorithm, one has to compute the inverse of an n-by-n matrix.

Motivated by flexibility of the augmented Lagrangian method [8,7] and convergence property of the proximal alternating linearized minimization (PALM) method [9], we propose an approximate augmented Lagrangian scheme named EPALMAL to solve optimization problem (1). Different from [4,12], we update iterates based on the scaled form of Powell-Hestenes-Rockafellar augmented Lagrangian function [23,41,42], then update the Lagrangian multipliers and the penalty parameter. In each iteration of our method, we approximately minimize the augmented Lagrangian function and preserve the generalized orthogonality constraints. For this purpose, we extend the proximal terms in the PALM method to terms with matrix norm to deal with the generalized orthogonality constraints, and name this method as EPALM. The global convergence of EPALM method can be established under similar assumptions as in the PALM method. We also investigate the convergence property of EPALMAL for problem (1), which shows that if the generated sequence is bounded, then every limit point is a KKT point. Notice that our method can be extended to handle optimization problems with multiple blocks of variables of the form

$$\begin{aligned}\min\ &\sum_{i=1}^p f_i(X_i)+\sum_{j=1}^s g_j(Y_j)+h(X_1,\ldots,X_p,Y_1,\ldots,Y_s)\\ \text{s.t. }&A_1X_1+\cdots+A_pX_p+B_1Y_1+\cdots+B_sY_s=C,\\ &X_i^TM_iX_i=I_q,\quad i=1,\ldots,p,\end{aligned}\tag{4}$$

where $\{M_i\}_{i=1}^p$ is a given set of symmetric positive semi-definite matrices.

1.1 Notations

In the rest of this paper, we denote the set of $n\times n$ symmetric matrices by S^n, and the set of $n\times n$ symmetric positive semi-definite (definite) matrices by S_+^n (S_{++}^n). Notation $M\succeq 0$ ($M\succ 0$) means $M\in S_+^n$ ($M\in S_{++}^n$). For any matrix $X\in\mathbb{R}^{m\times n}$, we define $\|X\|_\infty:=\max_{1\le i\le m,\ 1\le j\le n}\{|X_{i,j}|\}$ and the Frobenius norm $\|X\|_F:=\sqrt{\sum_{i,j}X_{i,j}^2}$. For matrices $X,Y\in\mathbb{R}^{m\times n}$, $\langle X,Y\rangle:=\mathrm{tr}(X^TY)$ denotes matrix inner product, where $\mathrm{tr}(A)=\sum_{i=1}^n A_{ii}$ ($A\in\mathbb{R}^{n\times n}$) represents the matrix trace. For any matrix $Q\succ 0$, we define norm $\|x\|_Q$ ($\|X\|_Q$) as $\|x\|_Q=\sqrt{x^TQx}$ ($\|X\|_Q=\sqrt{\mathrm{tr}(X^TQX)}$). Let $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the smallest and the largest eigenvalues of matrix A, respectively. We use $\bar{x}_i^c$ to denote point $(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_p)$, where $x_i\in\mathbb{R}^{n_i}$, $i=1,\cdots,p$; and $(x_i,\bar{x}_i^c):=(x_1,\cdots,x_{i-1},x_i,x_{i+1},\cdots,x_p)$. Let $f:\mathbb{R}^d\to(-\infty,+\infty]$ be a proper and lower semicontinuous function and $\mathrm{dom} f:=\{x\in\mathbb{R}^d: f(x)<$

$+\infty\}$ be its domain, the Fréchet subdifferential [36] of f at $x \in \mathrm{dom} f$ is defined as $\hat{\partial} f(x) := \{u \mid \lim_{y \neq x} \inf_{y \to x} \frac{f(y)-f(x)-\langle u, y-x\rangle}{\|y-x\|} \geq 0\}$ and the limiting-subdifferential of f at $x \in \mathrm{dom} f$ is defined as $\partial f(x) := \{u \in \mathbb{R}^d \mid \exists\, x^k \to x, f(x^k) \to f(x)$ and $u^k \in \hat{\partial} f(x^k) \to u$ as $k \to \infty\}$. A point $x \in \mathrm{dom} f$ is called a *critical point* of f if $0 \in \partial f(x)$. Other notations will be defined when they occur.

1.2 Organization

The rest of this paper is organized as follows. In Section 2, we provide some basic assumptions and some preliminaries on the Kurdyka-Łojasiewicz (K-L) property and the PALM method. At the end of Section 2, we extend the PALM method to the case with matrix norm proximal terms to deal with the generalized orthogonality constraints. In Section 3, we describe the derivation of algorithm EPALMAL and provide its convergence analysis. In Section 4, we show details of applying EPALM method to approximately minimize the augmented Lagrangian function in each iteration of EPALMAL. In Section 5, we test the efficiency and effectiveness of the proposed algorithm on a variety of applications in compressed modes and multivariate data analysis. Finally, some conclusions are drawn in Section 6.

2 Assumptions and Preliminaries

In this section, we outline some basic assumptions on problem (1) and introduce the Kurdyka-Łojasiewicz (K-L) property, and briefly review the PALM method.

2.1 Basic Assumptions

We make the following assumptions on problem (1).

Assumption 1 *(i) $f(X)$ and $g(Y)$ are proper and lower semicontinuous functions satisfying $\inf f(X) > -\infty$ and $\inf g(X) > -\infty$, $h(X,Y)$ is a C^1 function with $\inf h(X,Y) > -\infty$, and $\phi(X,Y)$ satisfies the K-L property.*

(ii) For any fixed Y, the function $X \to h(X,Y)$ is $C^{1,1}_{L_1(Y)}$, namely the partial gradient $\nabla_X h(X,Y)$ is globally Lipschitz continuous with moduli $L_1(Y)$, that is,

$$\|\nabla_X h(X,Y) - \nabla_X h(\widetilde{X},Y)\| \leq L_1(Y)\|X - \widetilde{X}\| \quad \forall\, X,\ \tilde{X} \in \mathbb{R}^{m\times n}.$$

Likewise, for any fixed X, the function $Y \to h(X,Y)$ is assumed to be $C^{1,1}_{L_2(X)}$.

(iii) ∇h is Lipschitz continuous on any bounded subset of $\mathbb{R}^{n\times q} \times \mathbb{R}^{m\times q}$. In other words, for each bounded subset $B_1 \times B_2$ of $\mathbb{R}^{n\times q} \times \mathbb{R}^{m\times q}$, there exists a $\vartheta > 0$ such that for all $(X_i, Y_i) \in B_1 \times B_2,\ i = 1,2$,

$$\|(\nabla_X h(X_1,Y_1) - \nabla_X h(X_2,Y_2), \nabla_Y h(X_1,Y_1) - \nabla_Y h(X_2,Y_2))\| \leq \vartheta \|(X_1 - X_2, Y_1 - Y_2)\|.$$

(iv) For any bounded set $\mathscr{B}$, $\cup_{X\in\mathscr{B}}\{\partial f(X)\}$ is bounded. So are ∂g, $\nabla_X h$, and $\nabla_Y h$.

By Proposition B.24 in [8], $\cup_{X\in\mathscr{B}}\{\partial f(X)\}$ is bounded if f is a convex function.

2.2 Preliminaries on the Kurdyka-Łojasiewicz Property

Similar to the convergence analysis in [6,9], the K-L property is needed to show that the sequence generated by the EPALM method is a Cauchy sequence. To this end, we recall some essential elements of the K-L property in this part.

Definition 1 (Kurdyka-Łojasiewicz property [5]) Let $\sigma : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

a) The function σ is said to have the Kurdyka-Lojasiewicz property at $\bar{x} \in dom\ \partial\sigma := \{x \in \mathbb{R}^n \mid \partial\sigma(x) \neq \emptyset\}$, if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\bar{x}$ and a concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:
   (i) $\varphi(0) = 0$;
   (ii) φ is C^1 on $(0, \eta)$ and continuous at 0;
   (iii) for all $s \in (0, \eta)$, $\varphi'(s) > 0$;
   (iv) for all x in $U \cap \{x \mid \sigma(\bar{x}) < \sigma(x) < \sigma(\bar{x}) + \eta\}$, the following Kurdyka-Łojasiewicz inequality holds
   $$\varphi'(\sigma(x) - \sigma(\bar{x}))dist(0, \partial\sigma(x)) \geq 1, \tag{5}$$
   where $dist(x, \mathscr{X}) := \inf\{\|y - x\| \mid y \in \mathscr{X}\}$ denotes the distance from x to $\mathscr{X}$.
b) If σ satisfies the K-L inequality (5) at each point of $dom\ \partial\sigma$, then σ is called a K-L function.

The K-L property plays an important role in nonconvex and nonsmooth analysis. According to inequality (5), functions satisfying the K-L property can avoid flatness around critical points. Moreover, the K-L property can be captured by adequate analytic assumptions, such as metric regularity, cohypomonotonicity, self-concordance and partial smoothness [6]. Functions such as semi-algebraic, subanalytic and log-exp are all K-L functions. Some fundamental works on the K-L property can be found in [30,33].

2.3 Preliminaries on PALM Methods

For nonconvex and nonsmooth problems of the form

$$\min_{x \in \mathbb{R}^n,\ y \in \mathbb{R}^m} f(x) + g(y) + H(x, y),$$

Attouch *et al*. [5] proposed a proximal alternating minimization (PAM) method, which updates x and y alternately by

$$\begin{cases} x^{k+1} \in \arg\min\limits_{x}\ f(x) + H(x, y^k) + \frac{1}{2}\|x - x^k\|^2_{B_1^k}, \\ y^{k+1} \in \arg\min\limits_{y}\ g(y) + H(x^{k+1}, y) + \frac{1}{2}\|y - y^k\|^2_{B_2^k}, \end{cases} \tag{6}$$

with $\{B_i^k \succ 0\}_{i=1,2}^{k \in \mathbb{N}}$. Global convergence for PAM method is provided based on inequality (5) and the assumption on the initial point. In [6], an inexact version, which contains formula (6) as a special case, was given. Global convergence is also provided under the assumption of boundedness of $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ by testing the sufficient decrease condition, the relative error condition, and the continuity condition. However, this method may not be practical, since exact minimization of two nonconvex and nonsmooth problems is required at each iteration, which is difficult for general $f(x)$ (or $g(y)$) and nonquadratic $H(x, y^k)$ (or $H(x^{k+1}, y)$). To overcome this drawback, Bolte *et al*. [9] considered the linearized approximation of $H(x, y)$

Algorithm 1 (**EPALM**: Extended Proximal Alternating Linearized Minimization)

Input: Initial point: $(x_1^0,\cdots,x_p^0)\in\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_p}$, and $k=0$.

Output: A sequence $\{(x_1^k,\cdots,x_p^k)\}_{k\in\mathbb{N}}$.

1: **while** stopping criterion is not satisfied **do**

2: compute x_i^{k+1} for $i=1,\cdots,p$ by solving

$$\min_{x_i} f_i(x_i)+\langle x_i-x_i^k,\nabla_{x_i}H(x_1^{k+1},\cdots,x_{i-1}^{k+1},x_i^k,\cdots,x_p^k)\rangle+\frac{1}{2}\|x_i-x_i^k\|^2_{B_i^k}, \tag{8}$$

where $B_i^k\succ 0$.

3: **end while**

4: **return** $\{(x_1^k,\cdots,x_p^k)\}_{k\in\mathbb{N}}$.

and proposed a proximal alternating linearized minimization (PALM) algorithm, which updates x and y alternately by

$$\begin{cases} x^{k+1}\in\arg\min\limits_x\ f(x)+\langle x-x^k,\nabla_x H(x^k,y^k)\rangle+\frac{d^k}{2}\|x-x^k\|_2^2\\ \qquad=\mathbf{prox}^f_{d^k}(x^k-\frac{1}{d^k}\nabla_x H(x^k,y^k)),\\ y^{k+1}\in\arg\min\limits_y\ g(y)+\langle y-y^k,\nabla_y H(x^{k+1},y^k)\rangle+\frac{e^k}{2}\|y-y^k\|_2^2\\ \qquad=\mathbf{prox}^g_{e^k}(y^k-\frac{1}{e^k}\nabla_y H(x^{k+1},y^k)),\end{cases}$$

where **prox** denotes the proximal map [38] defined as follows: Let $f:\mathbb{R}^n\to(-\infty,+\infty]$ be a proper and lower semicontinous function, $u\in\mathbb{R}^n$ and $t>0$ be given, the proximal map associated with f is defined as

$$\mathbf{prox}_t^f(u):=\arg\min_x f(x)+\frac{t}{2}\|x-u\|_2^2.$$

Under some assumptions, similar global convergence results as the PAM method can be obtained for PALM, see [9] for details.

Next, we extend the ℓ_2-norm proximal terms in PALM method to matrix norm to deal with the generalized orthogonality constraints as well as multiple variables case, which can be seen in Section 4, and name the resulting algorithm EPALM. More specifically, we consider the following optimization problem

$$\min_{x_1,\cdots,x_p}\left\{\psi(x_1,\cdots,x_p):=\sum_{i=1}^p f_i(x_i)+H(x_1,\cdots,x_p):x_i\in\mathbb{R}^{n_i}\right\}, \tag{7}$$

where $f_i:\mathbb{R}^{n_i}\to(-\infty,+\infty]$ are proper and lower semicontinous functions satisfying $\inf f_i(x_i)>-\infty$, $i=1,\cdots,p$, $H:\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_p}\to\mathbb{R}$ is a C^1 function, and $\inf H(x_1,\cdots,x_p)>-\infty$. The EPALM algorithm for problem (7) is described in Algorithm 1.

To guarantee global convergence of Algorithm 1, we need the following assumptions:

Assumption 2 *(i) For any $i\in\{1,\cdots,p\}$, the function $x_i\to H(x_i,\bar{x}_i^c)$ is $C^{1,1}_{L_i(\bar{x}_i^c)}$, namely the partial gradient $\nabla_{x_i}H(x_i,\bar{x}_i^c)$ is globally Lipschitz continuous with moduli $L_i(\bar{x}_i^c)$,*

$$\|\nabla_{x_i}H(x_i,\bar{x}_i^c)-\nabla_{x_i}H(\tilde{x}_i,\bar{x}_i^c)\|\le L_i(\bar{x}_i^c)\|x_i-\tilde{x}_i\|\quad\forall x_i,\ \tilde{x}_i\in\mathbb{R}^{n_i}.$$

(ii) For each $i\in\{1,\cdots,p\}$, there exists $-\infty<\lambda_i^-\le\lambda_i^+<+\infty$, such that

$$\inf\{L_i((\bar{x}_i^k)^c):k\in\mathbb{N}\}\ge\lambda_i^-\quad\textit{and}\quad\sup\{L_i((\bar{x}_i^k)^c):k\in\mathbb{N}\}\le\lambda_i^+.$$

(iii) *For each* $i \in \{1,\cdots,p\}$, $B_i^k - L_i((\bar{x}_i^k)^c)I \succ 0$, *and there exist* $-\infty < \underline{\lambda}_i \le \bar{\lambda}_i < +\infty$, *such that*

$$\inf\{\lambda_{min}(B_i^k) : k \in \mathbb{N}\} \ge \underline{\lambda}_i \quad \text{and} \quad \sup\{\lambda_{max}(B_i^k) : k \in \mathbb{N}\} \le \bar{\lambda}_i.$$

(iv) ∇H *is Lipschitz continuous on any bounded subset of* $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_p}$.

Notice that Algorithm 1 reduces to the PALM method if $B_i^k = c_i^k I$ and $c_i^k > L_i((\bar{x}_i^k)^c)$, and it becomes the PAM method if $H(x_i, \bar{x}_i^c)$ is quadratic on x_i for any $i \in \{1,\ldots,p\}$, due to the fact that for quadratic function $H(x,y^k)$ we can choose $B^k = B_1^k + \nabla_x^2 H(x,y^k)$ with $B_1^k \succeq 0$, and it holds that

$$\begin{aligned} x^{k+1} &\in \arg\min_x \ f(x) + \langle x - x^k, \nabla_x H(x^k,y^k)\rangle + \frac{1}{2}\|x-x^k\|^2_{B^k} \\ &= \arg\min_x \ f(x) + H(x,y^k) + \frac{1}{2}\|x-x^k\|^2_{B_1^k}, \end{aligned}$$

which exactly recovers the first subproblem of (6).

Next, we list some convergence properties of the EPALM method, whose proof is similar to that of the PALM method in [9] and hence omitted.

Lemma 1 *(Convergence properties). Suppose that Assumptions 1 (i)-(iii) and 2 hold. Let* $\{z^k := (x_1^k,\ldots,x_p^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by Algorithm 1 (EPALM). Then the following assertions hold.*

a) *The sequence* $\{\psi(z^k)\}_{k\in\mathbb{N}}$ *is nonincreasing and in particular*

$$\frac{\kappa_1}{2}\|z^{k+1}-z^k\|^2 \le \psi(z^k) - \psi(z^{k+1}) \quad \forall k \ge 0,$$

where $\kappa_1 := \min_{1\le i\le p}\{\underline{\lambda}_i - \lambda_i^-\}$.

b) *It holds that*

$$\sum_{k=1}^{\infty}\sum_{i=1}^{p}\|x_i^{k+1}-x_i^k\|^2 = \sum_{k=1}^{\infty}\|z^{k+1}-z^k\|^2 < \infty,$$

and hence $\lim_{k\to\infty}\|z^{k+1}-z^k\| = 0$.

c) *For each integer k, we define for* $i = 1,\cdots,p$

$$A_{x_i}^k := \nabla_{x_i} H(x_1^k,\ldots,x_p^k) - \nabla_{x_i} H(x_1^k,\ldots,x_{i-1}^k,x_i^{k-1},\ldots,x_p^{k-1}) + B_i^{k-1}(x_i^{k-1}-x_i^k).$$

If $\{z^k\}_{k\in\mathbb{N}}$ *is bounded, then* $(A_{x_1}^k,\ldots,A_{x_p}^k) \in \partial\psi(x_1^k,\ldots,x_p^k)$ *and there exists a* $\vartheta > 0$ *such that*

$$\|(A_{x_1}^k,\ldots,A_{x_p}^k)\| \le \|A_{x_1}^k\| + \cdots + \|A_{x_p}^k\| \le \Big(\frac{p(p+1)}{2}\vartheta + p\kappa_2\Big)\|z^k - z^{k-1}\| \quad \forall k \ge 1,$$

where $\kappa_2 := \max_{1\le i\le p}\{\bar{\lambda}_i\}$.

d) *Suppose that* ψ *is a K-L function and* $\{z^k\}_{k\in\mathbb{N}}$ *is bounded. Then*

(i) *the sequence* $\{z^k\}_{k\in\mathbb{N}}$ *has finite length, that is,*

$$\sum_{k=1}^{\infty}\|z^{k+1}-z^k\| < \infty.$$

(ii) *the sequence* $\{z^k\}_{k\in\mathbb{N}}$ *converges to a critical point* $z^* = (x_1^*,\ldots,x_p^*)$ *of* ψ.

3 Augmented Lagrangian Scheme and Convergence Analysis

In this section, we first propose the augmented Lagrangian scheme to solve problem (1), then show some convergence properties of the proposed scheme.

3.1 Augmented Lagrangian Scheme

For any closed set $\mathscr{X}$, we define the indicator function $\delta_{\mathscr{X}}$ as

$$\delta_{\mathscr{X}}(x) = \begin{cases} 0, & \text{if } x \in \mathscr{X}, \\ +\infty, & \text{otherwise,} \end{cases}$$

and let

$$\mathscr{M} = \{X \in \mathbb{R}^{n\times q} \mid X^T MX = I_q,\ M \in S_+^n\}$$

denote the set of all matrices satisfying the generalized orthogonality constraints. Then, by introducing auxiliary variables $G = X$, problem (1) is equivalent to

$$\min_{X,\,Y,\,G} \phi(X,Y) + \delta_{\mathscr{M}}(G) \quad \text{s.t.} \quad AX + BY = C,\ X - G = 0. \tag{9}$$

Denoting $\Lambda := (\Lambda_1^T,\ \Lambda_2^T)^T \in R^{(l+n)\times q}$, the classical augmented Lagrangian function associated with (9) is given by

$$\begin{aligned}\tilde{L}_\rho(X,Y,G;\Lambda) = \phi(X,Y) + \delta_{\mathscr{M}}(G) + \langle \Lambda_1, AX+BY-C\rangle + \frac{\rho}{2}\|AX+BY-C\|_F^2 \\ + \langle \Lambda_2, X-G\rangle + \frac{\rho}{2}\|X-G\|_F^2,\end{aligned}$$

where ρ is a given positive penalty parameter. In the sequel of this paper, we consider the scaled form

$$L_\rho(X,Y,G;\Lambda) = \frac{1}{\rho}\tilde{L}_\rho(X,Y,G;\Lambda). \tag{10}$$

Following the augmented Lagrangian framework [4], in the proposed scheme we alternately update (X,Y,G), the Lagrangian multiplier Λ and the penalty parameter ρ. Moreover, we apply the EPALM method to minimize the scaled augmented Lagrangian function and preserve the generalized orthogonality constraints $G^T MG = I_q$, which will be discussed in detail in Section 4. The outline of the augmented Lagrangian scheme is given in Algorithm 2.

Remark 1 In **Step 1**, the condition $\|A^k\|_\infty \le \frac{\varepsilon_{k-1}}{\rho_{k-1}}$ seems stringent. However, the problem we considered satisfies

$$L_{\rho_k}(X,Y,G;\bar{\Lambda}^k) = \frac{1}{\rho_k}\tilde{L}_{\rho_k}(X,Y,G;\bar{\Lambda}^k).$$

Therefore, condition (11) is equivalent to that there exists an $\tilde{A}^k \in \partial \tilde{L}_{\rho_{k-1}}(X^k,Y^k,G^k;\bar{\Lambda}^{k-1})$ such that

$$\|\tilde{A}^k\|_\infty \le \varepsilon_{k-1} \text{ and } \varepsilon_k \downarrow 0 \ as\ k \to +\infty,$$

for the classical form of the augment Lagrangian scheme. In addition, iterate (X^k,Y^k,G^k) can be viewed as an $\frac{\varepsilon_{k-1}}{\rho_{k-1}}$-perturbation of some critical point from the set $\{(X,Y,G) \mid 0 \in$

Algorithm 2 (**EPALMAL**: Approximate Augmented Lagrangian Method for (9))

Input: $\{\varepsilon_k\}_{k\in\mathbb{N}} \downarrow 0$, $-\infty < \bar{\Lambda}_{i,min} \le \bar{\Lambda}_{i,max} < +\infty$, $i = 1,2$, $\tau \in [0,1)$, $\mu > 1$, $k = 1$, $\rho_0 > 0$.
Output: A sequence $\{(X^k, Y^k, G^k, \Lambda_1^k, \Lambda_2^k)\}_{k\in\mathbb{N}}$.

1: **while** stopping criterion is not satisfied **do**
2: **Step 1.** For given $\rho_{k-1}, \bar{\Lambda}^{k-1}$, compute (X^k, Y^k, G^k) such that $(G^k)^T M G^k = I_q$, and there exists an $A^k \in \partial L_{\rho_{k-1}}(X^k, Y^k, G^k; \bar{\Lambda}^{k-1})$ satisfying

$$\|A^k\|_\infty \le \frac{\varepsilon_{k-1}}{\rho_{k-1}}. \tag{11}$$

3: **Step 2.** Update the Lagrangian multipliers

$$\begin{cases} \Lambda_1^k = \bar{\Lambda}_1^{k-1} + \rho_{k-1}(AX^k + BY^k - C), \\ \Lambda_2^k = \bar{\Lambda}_2^{k-1} + \rho_{k-1}(X^k - G^k), \end{cases}$$

where $\bar{\Lambda}_i^k$ is the projection of Λ_i^k on $\{\Lambda_i : \bar{\Lambda}_{i,min} \le \Lambda_i \le \bar{\Lambda}_{i,max}\}$, $i = 1,2$.
4: **Step 3.** Update the penalty parameter

$$\rho_k = \begin{cases} \rho_{k-1} & \text{if } \|R_i^k\|_\infty \le \tau \|R_i^{k-1}\|_\infty,\ i = 1,2, \\ \mu\rho_{k-1} & \text{otherwise,} \end{cases}$$

where $R_1^k := AX^k + BY^k - C$, $R_2^k := X^k - G^k$.
5: **end while**
6: **return** $\{(X^k, Y^k, G^k, \Lambda_1^k, \Lambda_2^k)\}_{k\in\mathbb{N}}$.

$\partial L_{\rho_{k-1}}(X, Y, G; \bar{\Lambda}^{k-1})\}$. In **Step 2**, the Lagrangian multipliers Λ_1^k, Λ_2^k are projected to bounded boxes to guarantee that the global minimizers of the original problem (9) are obtainable if each outer iteration computes a global minimizer of subproblem (11) [4]. In **Step 3**, the penalty parameter ρ_k is updated according to the constraint violation.

3.2 Convergence Analysis

In this subsection, we first give the optimality condition for problem (9), followed by the optimality condition for problem (1). Then, we discuss convergence properties of the augmented Lagrangian scheme EPALMAL.

Since constraints $G^T M G = I_q$ are symmetric, the corresponding Lagrangian multiplier $\Lambda_3 \in \mathbb{R}^{q\times q}$ is a symmetric matrix. The Lagrangian function associated with problem (9) is given by

$$\mathscr{L}(X, Y, G) = \phi(X, Y) + \langle \Lambda_1, AX + BY - C\rangle + \langle \Lambda_2, X - G\rangle + \langle \Lambda_3, G^T M G - I_q\rangle.$$

Moreover, we have the following lemma describing the first-order optimality condition for problem (9).

Lemma 2 *Suppose that (X^*, Y^*, G^*) is a local minimizer of problem (9). Then there exist $\Lambda_1^* \in \mathbb{R}^{l\times q}, \Lambda_2^* \in \mathbb{R}^{n\times q}, \Lambda_3^* \in \mathscr{S}^q$ such that $(X^*, Y^*, G^*; \Lambda_1^*, \Lambda_2^*, \Lambda_3^*)$ satisfies the first-order optimality conditions*

$$\begin{pmatrix} v^* + \partial_X h(X^*, Y^*) \\ w^* + \partial_Y h(X^*, Y^*) \\ 0 \end{pmatrix} + \begin{pmatrix} A^T & I & 0 \\ B^T & 0 & 0 \\ 0 & -I & 2MG^* \end{pmatrix} \begin{pmatrix} \Lambda_1^* \\ \Lambda_2^* \\ \Lambda_3^* \end{pmatrix} = 0, \tag{12}$$

where $v^* \in \partial f(X^*)$, $w^* \in \partial g(Y^*)$ *and*

$$AX^* + BY^* - C = 0, \quad X^* - G^* = 0, \quad (G^*)^T M G^* = I_q.$$

Furthermore, $(X^*, Y^*; \Lambda_1^*, \Lambda_3^*)$ *satisfies the first-order optimality conditions of problem (1), that is,*

$$\begin{pmatrix} v^* + \partial_X h(X^*, Y^*) \\ w^* + \partial_Y h(X^*, Y^*) \end{pmatrix} + \begin{pmatrix} A^T & 2MX^* \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \Lambda_1^* \\ \Lambda_3^* \end{pmatrix} = 0, \tag{13}$$

and

$$AX^* + BY^* - C = 0, \quad (X^*)^T M X^* = I_q.$$

Proof The proof is given in Appendix A.

We call any feasible point (X^*, Y^*, G^*) satisfying conditions (12) a KKT point of problem (9), and call any feasible point (X^*, Y^*) satisfying conditions (13) a KKT point of problem (1).

Next, we show that under the boundedness assumption of the sequence $\{(X^k, Y^k, G^k)\}$ generated by Algorithm 2, any limit point of the sequence satisfies the first-order optimality conditions (12).

Theorem 1 *Suppose that Assumption 1 holds. Let* $\{(X^k, Y^k, G^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by Algorithm 2. If* $\{(X^k, Y^k, G^k)\}_{k\in\mathbb{N}}$ *is bounded, then any limit point* (X^*, Y^*, G^*) *of* $\{(X^k, Y^k, G^k)\}_{k\in\mathbb{N}}$ *is a KKT point of problem (9). Moreover,* (X^*, Y^*) *is a KKT point of problem (1).*

Proof The proof is given in Appendix B.

Remark 2 Since Algorithm EPALM can handle problems with multiple blocks of variables, Algorithm EPALMAL can be extended to solve the multiple generalized orthogonality constrained problem (4) where at least one B_j has full row rank.

4 The EPALM Method

In Algorithm 2, **Step 1** is the main crucial part, which computes an approximate minimizer of the scaled augmented Lagrangian function. The effectiveness and efficiency of **Step 1** dominates Algorithm 2. In this section, we show how the EPALM method proposed in Subsection 2.3 can be employed to find an approximate minimizer (X^k, Y^k, G^k) such that (11) holds.

Define

$$\begin{aligned} H_k(X,Y,G) := {} & \frac{1}{\rho_{k-1}} h(X,Y) + \frac{1}{\rho_{k-1}} \left\langle \bar{\Lambda}_1^{k-1}, AX + BY - C \right\rangle + \frac{1}{2} \|AX + BY - C\|_F^2 \\ & + \frac{1}{\rho_{k-1}} \left\langle \bar{\Lambda}_2^{k-1}, X - G \right\rangle + \frac{1}{2} \|X - G\|_F^2 \end{aligned}$$

and

$$f_1^k(X) = \frac{1}{\rho_{k-1}} f(X), \quad f_2^k(Y) = \frac{1}{\rho_{k-1}} g(Y), \quad f_3^k(G) = \frac{1}{\rho_{k-1}} \delta_{\mathscr{M}}(G),$$

then

$$L_{\rho_{k-1}}(X, Y, G; \bar{\Lambda}^{k-1}) = f_1^k(X) + f_2^k(Y) + f_3^k(G) + H_k(X, Y, G), \tag{14}$$

which is of the form (7). Therefore, the EPALM method is applicable.

Let $M = U\Sigma U^T$, where $U \in \mathbb{R}^{n\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$ and $r = \mathrm{rank}(M)$, be the reduced SVD of M. For any fixed $k \geq 1$, we use the following inner iterations to find a critical point of $L_{\rho_{k-1}}(X,Y,G;\bar{\Lambda}^{k-1})$:

$$\begin{cases} X^{k,j} \in \arg\min\limits_X & f_1^k(X) + \left\langle X - X^{k,j-1}, \nabla_X H_k(X^{k,j-1}, Y^{k,j-1}, G^{k,j-1})\right\rangle \\ & + \frac{1}{2}\|X - X^{k,j-1}\|^2_{B_1^{k,j-1}}, \\ Y^{k,j} \in \arg\min\limits_Y & f_2^k(Y) + \left\langle Y - Y^{k,j-1}, \nabla_Y H_k(X^{k,j}, Y^{k,j-1}, G^{k,j-1})\right\rangle \\ & + \frac{1}{2}\|Y - Y^{k,j-1}\|^2_{B_2^{k,j-1}}, \\ G^{k,j} \in \arg\min\limits_G & f_3^k(G) + H_k(X^{k,j}, Y^{k,j}, G) + \frac{1}{2}\|G - G^{k,j-1}\|^2_{B_3^k}, \end{cases} \tag{15}$$

where

$$\begin{cases} B_1^{k,j-1} = \gamma_1 L_1^{k,j-1} I_n, \quad \gamma_1 > 1, \quad L_1^{k,j-1} = \frac{L_1(Y^{k,j-1})}{\rho_{k-1}} + \|A^TA\| + 1, \\ B_2^{k,j-1} = \gamma_2 L_2^{k,j-1} I_m, \quad \gamma_2 > 1, \quad L_2^{k,j-1} = \frac{L_2(X^{k,j})}{\rho_{k-1}} + \|B^TB\| + 1, \\ B_3^k = \alpha^k M - I_n + \alpha^k(I_n - UU^T), \quad \alpha^k = \frac{\gamma_3}{\min\{1, \min(\mathrm{diag}(\Sigma))\}}, \gamma_3 > 1. \end{cases}$$

Here, $L_1^{k,j-1}$ and $L_2^{k,j-1}$ are the global Lipschitz constants of $\nabla_X H(X, Y^{k,j-1}, G^{k,j-1})$ and $\nabla_Y H(X^{k,j}, Y, G^{k,j-1})$, respectively. γ_1, $\gamma_2 > 1$ ensure that $B_1^{k,j-1}$ and $B_2^{k,j-1}$ satisfy Assumption 2 (iii). Notice that $G^{k,j}$ is updated by the PAM method and α^k is selected to ensure $B_3^k \succeq 0$. Simple calculation shows that subproblems regarding variables X and Y can be solved by proximal mappings

$$X^{k,j} \in \mathbf{prox}^{f_1^k}_{\gamma_1 L_1^{k,j-1}}\left(X^{k,j-1} - \frac{1}{\gamma_1 L_1^{k,j-1}}\nabla_X H(X^{k,j-1}, Y^{k,j-1}, G^{k,j-1})\right), \tag{16}$$

and

$$Y^{k,j} \in \mathbf{prox}^{f_2^k}_{\gamma_2 L_2^{k,j-1}}\left(Y^{k,j-1} - \frac{1}{\gamma_2 L_2^{k,j-1}}\nabla_Y H(X^{k,j}, Y^{k,j-1}, G^{k,j-1})\right), \tag{17}$$

respectively. For the subproblem regarding G, we use the PAM method rather than the PALM method since $H_k(X^{k,j}, Y^{k,j}, G)$ is quadratic on G. It can be reformulated as

$$G^{k,j} = \arg\min_{G^TMG = I_q} \frac{1}{2}\|G - (X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1})\|_F^2 + \frac{1}{2}\|G - G^{k,j-1}\|^2_{B_3^k}. \tag{18}$$

It is easy to see that the generalized orthogonality constraints $G^TMG = I_q$ are equivalent to

$$(\Sigma^{\frac{1}{2}}U^TG)^T(\Sigma^{\frac{1}{2}}U^TG) = I_q.$$

Let $T = \Sigma^{\frac{1}{2}}U^TG \in \mathbb{R}^{r\times q}$ and $U_\perp$ be column orthogonal such that $[U, U_\perp]$ is an orthogonal matrix, it follows that $T^TT = I_q$ and any G satisfying the generalized orthogonality constraints can be written as

$$G = U\Sigma^{-\frac{1}{2}}T + U_\perp Z,$$

Algorithm 3 (EPALM for subproblem (11) with fixed k)

Input: $(X^{k,0}, Y^{k,0}, G^{k,0})$.
Output: A sequence (X^k, Y^k, G^k).

1: Compute $T^{k,0} = \Sigma^{-\frac{1}{2}}(U^T G^{k,0})$, $U_\perp Z^{k,0} = G^{k,0} - U(U^T G^{k,0})$, and let $j = 1$.
2: **while** $\|A^{k,j}\|_\infty > \frac{\varepsilon_{k-1}}{\rho_{k-1}}$ **do**
3: Update $X^{k,j}$ and $Y^{k,j}$ by proximal mappings (16) and (17), respectively.
4: Compute $\Delta_1 = U^T(X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1})$ and $\Delta_2 = (X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}) - U\Delta_1$.
5: Compute $\Delta = \Sigma^{-\frac{1}{2}}\Delta_1 + (\alpha^k I - \Sigma^{-1})T^{k,j-1}$ and its reduced SVD $\Delta = \tilde{U}\tilde{\Sigma}\tilde{V}^T$.
6: Compute $T^{k,j} = \tilde{U}\tilde{V}^T$ and $U_\perp Z^{k,j} = \frac{1}{\alpha^k}\Delta_2 + (1 - \frac{1}{\alpha^k})(U_\perp Z^{k,j})$.
7: Compute $A^{k,j}$ as in (19).
8: **end while**
9: **return** $X^k = X^{k,j}$, $Y^k = Y^{k,j}$ and $G^k = U\Sigma^{-\frac{1}{2}}T^{k,j} + U_\perp Z^{k,j}$.

for some $Z \in \mathbb{R}^{(n-r)\times q}$. Therefore, we can convert the optimization problem (18) into an optimization problem over T and Z. Specifically, if $G^{k,j-1} = U\Sigma^{-\frac{1}{2}}T^{k,j-1} + U_\perp Z^{k,j-1}$ with $(T^{k,j-1})^T T^{k,j-1} = I_q$, then a simple calculation yields that

$$\|G - G^{k,j-1}\|^2_{B_3^k} = \|T - T^{k,j-1}\|^2_{\alpha^k I - \Sigma^{-1}} + (\alpha^k - 1)\|Z - Z^{k,j-1}\|_F^2.$$

Moreover, (18) is equivalent to

$$\begin{aligned}(T^{k,j}, Z^{k,j}) = \arg\min_{T^T T = I_q} & -2\langle T, \Sigma^{-\frac{1}{2}}U^T(X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}) + (\alpha^k I - \Sigma^{-1})T^{k,j-1}\rangle \\ & + \alpha^k\|Z - (U_\perp^T(X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}) + (\alpha^k - 1)Z^{k,j-1})\|_F^2.\end{aligned}$$

Hence, $G^{k,j} = U\Sigma^{-\frac{1}{2}}T^{k,j} + U_\perp Z^{k,j}$, where

$$T^{k,j} = \tilde{U}\tilde{V}^T = \arg\max_{T^T T = I_q} \left\langle T, \Sigma^{-\frac{1}{2}}U^T(X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}) + (\alpha^k I - \Sigma^{-1})T^{k,j-1} \right\rangle$$

with $\tilde{U}$ and $\tilde{V}$ being the left and right singular matrices of $\Sigma^{-\frac{1}{2}}U^T(X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}) + (\alpha^k I - \Sigma^{-1})T^{k,j-1})$, respectively, and

$$\begin{aligned}Z^{k,j} =& \frac{1}{2}\|Z - \frac{1}{\alpha^k}(U_\perp^T(X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}) + (\alpha^k - 1)Z^{k,j-1})\|_F^2 \\ =& \frac{1}{\alpha^k}(U_\perp^T(X^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}) + (\alpha^k - 1)Z^{k,j-1}).\end{aligned}$$

Therefore, we can compute $T^{k,j}$ and $U_\perp Z^{k,j}$ in the inner iterations, instead of computing $G^{k,j}$. Details about the EPALM method for problem (14) are described in Algorithm 3.

In Algorithm 3, we select $(X^{k,0}, Y^{k,0}, G^{k,0})$ randomly if $k = 0$, and select $(X^{k,0}, Y^{k,0}, G^{k,0}) = (X^{k-1}, Y^{k-1}, G^{k-1})$ for $k \geq 1$. To check the convergence condition (11), we select $A^{k,j} = (A_X^{k,j}, A_Y^{k,j}, A_G^{k,j})$ as follows:

$$\begin{cases} A_X^{k,j} = \nabla_X H(X^{k,j}, Y^{k,j}, Z^{k,j}) - \nabla_X H_k(X^{k,j-1}, Y^{k,j-1}, Z^{k,j-1}) + B_1^{k,j-1}(X^{k,j-1} - X^{k,j}), \\ A_Y^{k,j} = \nabla_Y H(X^{k,j}, Y^{k,j}, Z^{k,j}) - \nabla_Y H_k(X^{k,j}, Y^{k,j-1}, Z^{k,j-1}) + B_2^{k,j-1}(Y^{k,j-1} - Y^{k,j}), \\ A_G^{k,j} = B_3^k(G^{k,j-1} - G^{k,j}). \end{cases} \tag{19}$$

In the next proposition, we show that $A^{k,j}$ indeed satisfies the convergence condition (11) if we choose ρ_0 in Algorithm 2 properly. Moreover, the sequence generated by Algorithm 2 is strongly convergence.

Proposition 1 *For each $k \geq 1$, let $\{(X^{k,j}, Y^{k,j}, G^{k,j})\}_{j\in\mathbb{N}}$ be the sequence generated by Algorithm 3. Then*

1) $A^{k,j}$ defined in (19) satisfies

$$A^{k,j} \in \partial L_{\rho_{k-1}}(X^{k,j}, Y^{k,j}, G^{k,j}; \bar{\Lambda}^{k-1}) \quad \forall j \in \mathbb{N}.$$

Moreover, suppose ρ_0 in Algorithm 2 is chosen such that $\phi(X,Y) + \frac{\rho_0}{2}||AX+BY-C||_F^2$ is a coercive function. Then

$$\|A^{k,j}\|_\infty \to 0 \text{ as } j \to \infty.$$

2) The sequence $\{X^{k,j}, Y^{k,j}, G^{k,j}\}_{j\in\mathbb{N}}$ has finite length, that is,

$$\sum_{j=1}^{\infty} \|(X^{k,j+1}, Y^{k,j+1}, G^{k,j+1}) - (X^{k,j}, Y^{k,j}, G^{k,j})\| < \infty.$$

Moreover, $\{(X^{k,j}, Y^{k,j}, G^{k,j})\}_{j\in\mathbb{N}}$ converges to a critical point $(X^{k,*}, Y^{k,*}, G^{k,*})$ of function $L_{\rho_{k-1}}(X, Y, G; \bar{\Lambda}^{k-1})$.

Proof The proof is given in Appendix C.

Remark 3 A closely related method to our method EPALMAL in Algorithm 2 is the PAMAL method proposed in [12]. Both methods rely on minimizing the augmented Lagrangian function approximately. However, there are three major differences between PAMAL and EPALMAL: (1) EPALMAL considers more general optimization problems with linear equation constraint and the generalized orthogonality constraints $\{X \mid X^T MX = I_q,\ M \in S_+^n\}$, while PAMAL considers the compact orthogonality constraints $\{X \mid X^T QX = I_q,\ Q \in S_{++}^n\}$. (2) In the subproblem of minimizing the augmented Lagrangian function, PAMAL adopts the PAM method [6] while EPALMAL adopts the EPALM method. A prominent advantage of EPALM over PAM [9] is that each step of PAM requires exact minimization of nonconvex and nonsmooth problems (6), which is challenging to solve when H is non-quadratic, while each step of EPALM only requires proximal operations (8). Moreover, as we mentioned in Subsection 2.3, EPALM reduces to PAM for quadratic function H and properly selected proximal term. (3) Computationally, when h is quadratic (e.g., the compressed modes problem in Subsection 5.1), EPALMAL requires only matrix multiplications, while each inner iteration of PAMAL needs to compute the inversion of an n-by-n matrix.

5 Numerical Experiments

In this section, we evaluate the effectiveness and efficiency of our newly proposed method EPALMAL on some applications. In Subsection 5.1, we compare EPALMAL with the PAMAL method in [11,12] and the SOC method in [31,40] on the compressed modes for variational problems. Then, we evaluate the effectiveness and efficiency of EPALMAL on dimensionality reduction of high-dimensional data, including application to sparse uncorrelated linear discriminant analysis (sparse ULDA) in Subsection 5.2 and application to sparse canonical correlation analysis (sparse CCA) in Subsection 5.3. In the following experiments, we choose $\varepsilon_k = 0.999^k$, $\tau = 0.99$, $\mu = 1.02$, $\bar{\Lambda}_{i,min} = -10^{-2}$, $\bar{\Lambda}_{i,max} = 10^2$, $i = 1,2$. Other parameters will be given when they occur. All experiments were conducted on MATLAB R2014a running on CentOS 5 with four AMD 2.4GHz Opteron 850 CPUs and 32GB RAM at the High Performance Computing (HPC) Center of National University of Singapore.

5.1 The Compressed Modes for Variational Problems in Physics

Compressed modes (CMs) [40] are spatially localized solutions to the independent-particle Schrödinger's equation

$$\hat{H}\phi(x) = \lambda\phi(x), \qquad x \in \Omega, \tag{20}$$

where $\Omega \subseteq \mathbb{R}^d$ is a bounded set, $\hat{H} = -\frac{1}{2}\Delta + V$ denotes the Hamiltonian operator with Δ being the Laplacian operator and V being the potential energy function. In [40], the authors developed a variational approach to produce CMs to eigenvalue problem (20) by considering a finite discretized system of N electrons and ignoring the electron spin. Following the settings in [40], we focus on $\Omega = [0,L]^d$ with periodic boundary conditions and n equally spaced nodes in each direction. Then, the CMs for problem (20) is given by solution Ψ^* to the following ℓ_1-regularized optimization problem

$$\min_{\Psi\in\mathbb{R}^{n\times N}} J(\Psi) := \frac{1}{\kappa}\|\Psi\|_1 + tr(\Psi^T H\Psi) \quad \text{s.t.} \quad \Psi^T\Psi = I_N, \tag{21}$$

where $\|\Psi\|_1 := \sum_{i=1}^{n}\sum_{j=1}^{N}|\Psi_{ij}|$, $\kappa > 0$ is a pre-defined parameter, and $H = H^T$ is the discreted Hamiltonian in the electronic structure context. We refer the interested reader to [40] for details.

By introducing auxiliary variable $X = \Psi$, $X^TX = I$, we can reformulate (21) as

$$\min_{\Psi,X\in\mathbb{R}^{n\times N}} \frac{1}{\kappa}\|\Psi\|_1 + tr(\Psi^T H\Psi) \quad \text{s.t.} \quad \Psi - X = 0,\ X^TX = I. \tag{22}$$

We name the algorithm obtained from applying Algorithm 2 to problem (21) as CMsAL, and outline details in Algorithm 4 described in Appendix D. In the implementation of Algorithm 4, we choose $\gamma_1^k = 0.51(\frac{2}{\rho_{k-1}}\|H\|_2 + 1)$ and $\gamma_2 = 0.5$.

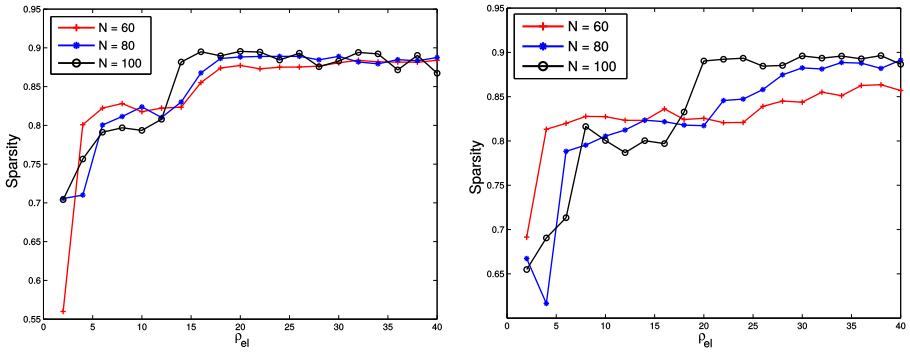
We compare Algorithm CMsAL with the PAMAL method and the SOC method for solving (21) on one-dimensional ($d = 1$) free-electron (FE) and the Kronig-Penney (KP) models. [1] For FE model, we use $V \equiv 0$. For KP model, we use the potential energy function V approximated by inverted Gaussian

$$V(x) = -\sum_{j=1}^{N_{el}} \exp\left[-\frac{(x-x_j)^2}{\gamma_{el}\delta^2}\right]. \tag{23}$$

Parameters in FE and KP models are outlined in Table 1. In all experiments, we use H as the discretized Hamiltonian operator with centered difference. Parameters of Algorithm CMsAL are set the same as those of the PAMAL method as recommended in [12] except that we use $\rho_0 = 2|\lambda_{min}(H)| + k/\rho_{el}$ in both methods, where ρ_{el} is outlined in Table 1. As shown in Fig. 1, ρ_{el} affects sparsity of the solution computed by CMsAL. We choose those values of ρ_{el} in Table 1 so that the sparsity level is stable and ρ_{el} is as small as possible on the premise of meeting the convergence condition. For the SOC method, parameters are set to be the same as in [40]. For all three methods, we use the same random orthogonal initial point and terminate the outer iteration when

$$\frac{|J(\Psi^i) - J(\Psi^{i-1})|}{\max\{1, |J(\Psi^{i-1})|\}} < 1E-5 \quad \text{and} \quad \|X^i - \Psi^i\|_F < 1E-2.$$

[1] We did not compare with the ADMM method although the convergence for compact constraints is given (e.g. [48]) since the only difference between the SOC method and the ADMM method is that the augmented penalty parameters in the latter method are generally the same for different constraints.



(a) FE model (b) KP model

Fig. 1 Influence of ρ_{el} on the sparsity of Ψ for 256 nodes in domain $[0,100]$ with $\kappa = 10$.

Table 1 Parameters in FE and KP models. Each row corresponds to one experimental setting.

n	L	N_{el}	x_j	γ_{el}	δ	ρ_{el}
128	50	5	10*j	2	3	6
256	100	10	100/11*j	1	2	20

Table 2 Results for the FE model (average over 50 repetitions).

Problems			No. of outer iterations			Total No. of inner iterations			CPU time (s)			Objective function value			Eigenvalue error		
n	k	κ	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC
128	5	30	198	182	748	204	182	748	**0.02**	0.06	0.21	1.01	0.98	0.99	1.836	1.431	1.556
	5	50	179	211	1613	185	211	1613	**0.02**	0.07	0.42	0.77	0.67	0.67	1.666	1.000	1.102
	50	10	415	332	1857	716	412	1857	0.82	**0.60**	2.52	89.91	90.13	89.95	0.033	0.032	0.030
	60	10	507	238	2264	1246	794	2264	1.82	**1.37**	3.88	141.96	142.22	142.28	0.021	0.020	0.020
256	10	30	385	290	1660	479	290	1660	**0.12**	0.26	1.19	2.06	1.99	2.01	2.113	1.393	1.550
	10	50	356	242	3487	398	242	3487	**0.11**	0.21	2.53	1.40	1.34	1.35	1.284	0.837	0.969
	50	10	332	426	1660	447	521	1660	**0.64**	1.22	3.51	37.73	37.70	37.74	0.133	0.123	0.130
	60	10	261	343	2122	318	467	2122	**0.60**	1.27	5.32	54.80	54.74	54.72	0.101	0.091	0.088
	80	10	240	255	2694	411	375	2694	**1.12**	1.41	10.63	105.09	105.13	105.20	0.065	0.056	0.050
	100	10	311	313	3465	918	714	3465	3.57	**3.47**	16.63	180.73	181.36	181.72	0.035	0.031	0.031

For each selection of (n,L), we did experiments using various N and κ.

Table 2 and Table 3 list results of all three algorithms on FE and KP models, respectively, including the number of outer iterations, total number of inner iterations, CPU times in seconds, objective function values $J(\Psi^i)$ and relative error of eigenvalues between H and $\Psi_N^T H \Psi_N$ for each method. Here, relative error of eigenvalues is computed by $\frac{\|h-p\|_2}{\|p\|_2}$, where h and p stand for the N largest eigenvalues of H and $(\Psi_N)^T H \Psi_N$ obtianned by each method, respectively. It can be seen that for most cases CMsAL requires less time than PAMAL and SOC, while all three methods obtain similar objective function values and relative errors of eigenvalue. Notice that in many cases CMsAL needs more inner iterations than PAMAL, especially with increase of k. This is mainly because CMsAL uses linearization in solving inner subproblems which usually needs more inner iterations to reach certain accuracy. We also plot the first M eigenvalues of the matrix $\Psi_N^T H \Psi_N$ obtained by each method, and the first M eigenvalues of the corresponding Schrödinger's operators H in Fig. 2 to demonstrate the approximation accuracy of these three methods. We observe that all three methods obtain comparable accuracy, and higher accuracy can be obtained by increasing the column number of Ψ_N.

Table 3 Results for the KP model (average over 50 repetitions).

Problems			No. of outer iterations			Total No. of inner iterations			CPU time (s)			Objective function value			Eigenvalue error		
n	k	κ	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC	CMsAL	PAMAL	SOC
128	5	50	340	250	877	344	250	877	**0.04**	0.08	0.23	-3.86	-3.87	-3.88	0.006	0.004	0.004
	5	300	359	119	2456	363	119	2456	**0.04**	**0.04**	0.66	-4.21	-4.21	-4.19	0.001	0.000	0.005
	50	10	556	763	1969	1356	775	1969	1.55	**1.16**	2.70	53.21	56.94	52.19	0.038	0.029	0.039
	60	10	756	640	2355	1987	705	2355	2.98	**1.31**	4.12	96.92	101.36	96.81	0.026	0.020	0.026
256	10	50	400	263	2205	402	263	2205	**0.12**	0.22	1.56	-6.14	-6.19	-6.15	0.004	0.002	0.003
	10	300	326	197	3416	329	197	3416	**0.09**	0.17	2.48	-6.67	-6.70	-6.63	0.004	0.001	0.009
	50	10	315	272	1831	317	274	1831	**0.47**	0.65	3.92	18.70	18.59	18.56	0.147	0.151	0.147
	60	10	304	299	2159	307	307	2159	**0.59**	0.86	5.40	32.28	32.21	32.22	0.108	0.113	0.110
	80	10	432	260	2837	524	308	2837	1.47	**1.15**	9.82	75.59	75.61	75.56	0.063	0.066	0.058
	100	10	260	176	3610	424	293	3610	1.72	**1.45**	18.54	144.93	145.13	145.28	0.042	0.046	0.035

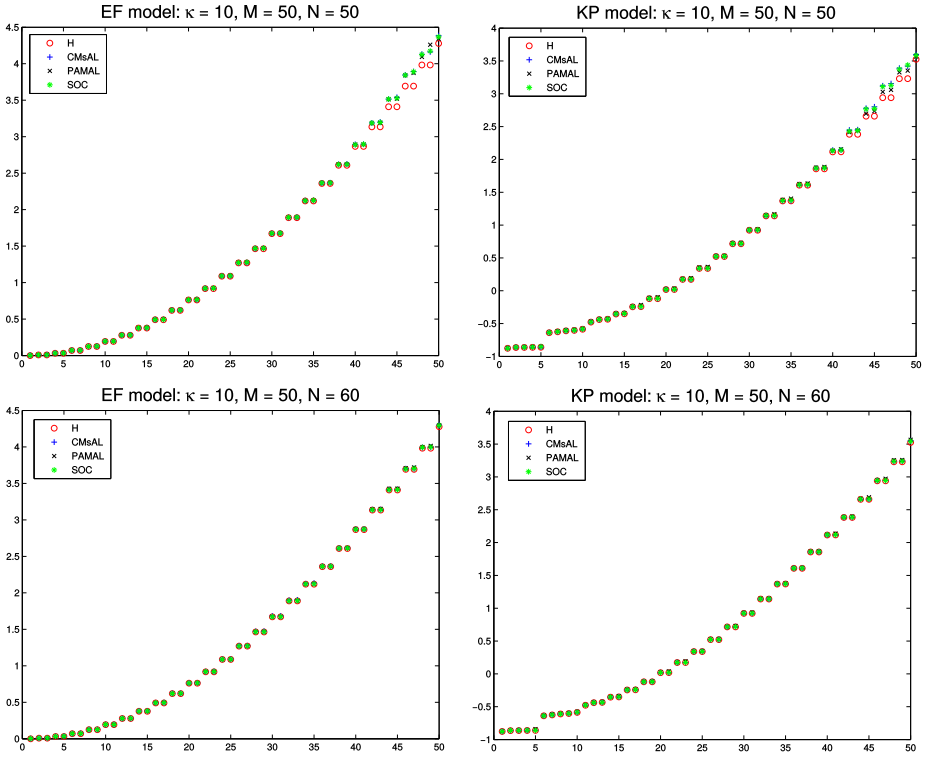


Fig. 2 The comparison of the first M eigenvalues for 128 nodes in domain [0, 50]. Left: EF model; Right: KP model.

5.2 Sparse Uncorrelated Linear Discriminant Analysis

Linear discriminant analysis (LDA) addresses dimensionality reduction problem of finding projections to map a high-dimensional data vector into the most discriminative low-dimensional subspace. This is accomplished by maximizing the between-class variance and minimizing the within-class variance in the projected space. Uncorrelated LDA (ULDA) [54] is a generalization of LDA such that the projected data are uncorrelated. Formally, given a data matrix $A \in \mathbb{R}^{n\times d}$ where n and d denote the numbers of samples and features, respectively, and each row of A is a d-dimensional sample belonging to one of K classes, and let $E \in \mathbb{R}^{n\times K}$ be an indicator matrix where $E_{ik} = 1$ if the i-th sample belongs to the k-th class and $E_{ik} = 0$ otherwise, we can define the between-class scatter matrix and total scatter

matrix [25] as

$$S_b = \frac{1}{n}(I - \frac{\mathbb{1}\mathbb{1}^T}{n})A^T E(E^T E)^{-1}E^T A(I - \frac{\mathbb{1}\mathbb{1}^T}{n}) \quad \text{and} \quad S_t = \frac{1}{n}(I - \frac{\mathbb{1}\mathbb{1}^T}{n})A^T A(I - \frac{\mathbb{1}\mathbb{1}^T}{n}),$$

where all entries of $\mathbb{1} \in \mathbb{R}^n$ are 1. ULDA finds optimal discriminant transformation G^* by

$$G^* = \arg \max_{G^T S_t G = I} \operatorname{tr}(G^T S_b G). \tag{24}$$

Sparse ULDA was proposed in [58] to promote sparsity in discriminant transformation G so that the projected data has a meaningful interpretation. This is accomplished by seeking the most sparse solution from all minimum dimension solutions of the generalized ULDA problem (24). According to [58], all minimum dimension solutions to problem (24) satisfy $U^T G^* = \Sigma_t^{-1/2} V X$, where $U \in St(d,t)$ and Σ_t are obtained from the reduced SVD of $S_t = U\Sigma_t U^T$, $V \in St(t,q)$ is the left singular matrix of $\Sigma_t^{-1/2} U^T S_b U \Sigma_t^{-1/2}$ associated with positive singular values, and $X \in \mathbb{R}^{q\times q}$ is an arbitrary orthogonal matrix. To find the sparest solution of ULDA, sparse ULDA solves the following problem

$$\min_{G,\,X} \|G\|_1 \quad \text{s.t.} \quad \mathscr{A}X + \mathscr{B}G = 0,\ X^T X = I, \tag{25}$$

where $\mathscr{A} = \Sigma_t^{-1/2} V$ and $\mathscr{B} = -U^T$ are constant matrices, $\mathscr{B}$ has full row rank since $\mathscr{B}\mathscr{B}^T = I$. It is easy to formulate sparse ULDA problem (25) as the form of general problem (1) with $f = 0$, $g(G) = \|G\|_1$, $h = 0$ and $M = I$. Thus, Algorithm 2 is applicable. Algorithm 2 can be also used to solve the following group sparse ULDA (GSULDA) problem

$$\min_{G,\,Z} \|G\|_{1,2} \quad \text{s.t.} \quad \mathscr{A}X + \mathscr{B}G = 0,\ X^T X = I, \tag{26}$$

where the $\ell_{1,2}$ norm $\|G\|_{1,2} := \sum_i \sqrt{\sum_j G_{ij}^2}$. Applying Algorithm 2 to problems (25) and (26) are the same except the difference of updating G. We name the algorithms obtained from applying Algorithm 2 to problems (25) and (26) as SULDAAL and GSULDAAL, respectively, and outline details in Algorithm 5 described in Appendix D. In the implementation of Algorithm 5 (SULDAAL or GSULDAAL), we choose $\rho_0 = \frac{1}{\|\mathscr{A}^T\mathscr{B}\|_\infty}$, $\gamma_1 = 1.01$, $\gamma_2 = 1.01$, $\gamma_3 = 0.5$. Moreover, we terminate the inner iteration when

$$\max\{\|A_X^{k,j}\|_\infty,\ \|A_Y^{k,j}\|_\infty,\ \|A_Z^{k,j}\|_\infty\} < \frac{\varepsilon_{k-1}}{\rho_{k-1}},$$

or the number of iterations reaches 50, and terminate the outer iteration when

$$\max\left\{\frac{\|X^k - X^{k-1}\|_F}{\max\{1, \|X^{k-1}\|_F\}},\ \frac{\|G^k - G^{k-1}\|_F}{\max\{1, \|G^{k-1}\|_F\}}\right\} < 1E-3,$$

and

$$\max\{\|\mathscr{A}X^k + \mathscr{B}G^k\|_F,\ \|X^k - Z^k\|_F\} < 1E-5.$$

We compare Algorithm 5 with four existing sparse LDA algorithms: SULDA_admm and SULDA_ℓ_1 [59] for problem (25), sparse discriminant analysis (SDA) [16] and group-lasso optimal scoring solver (GLOSS) [35]. Implementations of the comparing algorithms can

Table 4 Data stuctures: data dimension (d), training size (n), the number of classes (K) and the number of testing data (# Testing).

Data set	d	n	K	# Testing	Data set	d	n	K	# Testing
Lymphoma	4026	31	3	31	Palmprint	4096	300	100	300
Srbct	2308	31	4	32	tr11	6429	209	9	205
Brain	5597	21	5	21	tr23	5832	104	6	100
Carcinom	9182	85	11	89	tr41	7454	442	10	436
$\text{ORL}_{64\times64}$	4096	200	40	200	tr45	8261	347	10	343

be found on the authors' websites [2], and all algorithms are implemented on MATLAB. In SULDA_ℓ_1, we fix $Z = I$ and terminate the iteration if $\frac{\|\mathscr{A}+\mathscr{B}G^k\|_F}{\|\mathscr{A}\|_F} < 1E-5$. In SULDA_*admm*, we terminate the iteration if $\frac{\beta^k\|G^k-G^{k-1}\|_F}{\max\{1,\|G^{k-1}\|_F\}} < 1E-2$ and $\|\mathscr{A}X^k+\mathscr{B}G^k\|_F < 1E-5$. For SDA and GLOSS, we follow [59] where parameters are tuned to compute solutions having similar sparsity as other approaches. We conducted experiments on some real-world data, including four gene expression data sets [3]: Srbct, Brain, Lymphoma, and Carcinom [53]; two image data sets [4]: $\text{ORL}_{64\times64}$, and Palmprint; and four text-document data sets [47,60] [5]: tr11, tr23, tr41 tr45. For each data set, we randomly split it into training and testing data, and repeated the splitting 10 times. Table 4 lists some important statistics of these data sets.

Table 5 summarizes numerical results on all data sets. In Table 5, the following results are recorded: 'Accuracy' denotes the classification accuracy computed by nearest neighbor classifier [17], 'Orthogonality' computed by $\|G^T S_t G - I_q\|_F/\sqrt{q}$ denotes the uncorrelation conditions of solution G, 'Sparsity' computed by $\frac{number\ of\ zeroes\ in\ G}{total\ number\ of\ entries\ in\ G} \times 100\%$ denotes the percentage of zero elements in solution G, and 'Selec. Fea.' counts the number of selected features by G. Among all d features, we say one feature is selected if the corresponding row of G is nonzero. Comparing the performance of SULDAAL and GSULDAAL, we observe that both algorithms have similar classification accuracies, but SULDAAL is faster than GSULDAAL. The lower efficiency of GSULDAAL is caused by the increasing number of iterations incurred by the the usage of $\ell_{1,2}$-norm. In addition, GSULDAAL selects less features than SULDAAL due to the usage of $\ell_{1,2}$-norm. However, SULDAAL obtains higher sparsity than GSULDAAL, especially on image data sets. One reason is that for selected features, the corresponding rows of G computed by SULDAAL is sparse while those rows of G computed by GSULDAAL is dense. Next, we compare the two newly proposed algorithms with existing methods. On the side of efficiency, SULDAAL is much faster than existing methods on all data sets, and GSULDAAL is much faster than existing methods on all data sets except Brain and Srbct where it is slightly slower than SDA. On the side of effectiveness, our algorithms have competitive classification performance as the best one in the sense that SULDAAL or GSULDAAL has the highest accuracies on five data sets, and for the rest data sets they achieve comparable accuracies (within 1.5%) as the best

[2] SULDA_admm and SULDA_ℓ_1 can be found at `http://web.bii.a-star.edu.sg/~zhangxw/Publications.html`, SDA and GLOSS can be found at `http://www2.imm.dtu.dk/pubdb/views/publicationdetails.php?id=567` and `http://www.hds.utc.fr/~grandval/dokuwiki/doku.php?id=en:code`, respectively.

[3] Brain, Lymphoma and Srbct are obtained from `http://stat.ethz.ch/~dettling/bagboost.html`.

[4] $\text{ORL}_{64\times64}$ is from `http://www.cl.cam.ac.uk/Reasearch/DTG/attarchive:pub/data/attfaces/tar.Z`, and Palmprint is from `http://www4.comp.polyu.edu.hk/~biometrics/`.

[5] All data sets were downloaded from `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download`.

Table 5 Average results over 10 training-testing splittings: CPU time (CPU), classification accuracy in percentage (Accuracy), uncorrelation condition (Orthogonality), percentage or zero entries in G (Sparsity), and the number of selected features (Selec. Fea.). Values in the parentheses are standard deviations.

	Algorithm	CPU	Accuracy	Orthogonality	Sparsity	Selec. Fea.
Lymphoma	SULDAAL	2.34 (9.66E-02)	98.00 (4.89E+00)	5.45E-05 (2.16E-10)	98.91 (4.05E-07)	81.73 (2.12E+01)
	GSULDAAL	6.82 (3.56E+00)	96.67 (1.11E+01)	3.25E-04 (3.65E-09)	98.46 (8.32E-07)	62 (1.35E+01)
	SULDA_ℓ_1	31.15 (3.27E+01)	98.00 (4.89E+00)	4.82E-06 (9.16E-12)	99.23 (1.39E-09)	58.78 (1.61E+00)
	SULDA_*admm*	453.93 (2.30E+03)	97.67 (4.56E+00)	5.43E-07 (1.42E-13)	99.18 (2.09E-07)	62.40 (6.44E+00)
	SDA	18.78 (4.84E+02)	99.33 (1.78E+00)	3.10E+01 (2.74E-09)	98.07 (8.27E-06)	155 (5.28E+02)
	GLOSS	13.91 (2.69E+02)	98.00 (7.11E+00)	1.60E+02 (1.15E+03)	99.16 (2.47E-07)	33.82 (4.00E+00)
Srbct	SULDAAL	2.19 (9.06E-01)	98.71 (2.50E+00)	1.06E-05 (1.84E-11)	98.42 (9.26E-08)	92.55 (1.21E+01)
	GSULDAAL	5.20 (8.01E-01)	98.71 (2.50E+00)	1.23E-05 (2.01E-11)	97.28 (2.47E-06)	62.78 (1.32E+01)
	SULDA_ℓ_1	13.06 (7.20E+00)	98.06 (4.58E+00)	4.25E-06 (2.11E-12)	98.66 (6.05E-09)	80.32 (5.61E+00)
	SULDA_*admm*	412.52 (1.65E+03)	99.03 (4.27E+00)	2.44E-08 (1.22E-15)	98.58 (2.51E-07)	84.70 (1.64E+01)
	SDA	2.58 (1.19E+01)	99.68 (9.37E-01)	3.10E+01 (1.14E-10)	97.91 (2.11E-08)	140.10 (1.60E+00)
	GLOSS	17.47 (1.15E+02)	97.74 (6.35E+00)	1.52E+02 (2.05E+03)	98.43 (1.12E-06)	36.24 (5.96E+00)
Brain	SULDAAL	3.75 (5.8E-01)	80.48 (8.5E+00)	2.48E-05 (4.6E-06)	99.36 (3.5E-04)	134.1 (6.1E+00)
	GSULDAAL	11.39 (1.8E+00)	79.52 (6.0E+00)	6.81E-05 (3.6E-05)	98.54 (1.2E-03)	81.8 (7.0E+00)
	SULDA_ℓ_1	68.53 (7.8E+00)	75.24 (9.7E+00)	6.39E-06 (2.1E-06)	99.64 (5.6E-05)	77.4 (1.8E+00)
	SULDA_*admm*	691.16 (2.6E+01)	76.19 (1.1E+01)	9.32E-07 (8.2E-07)	99.60 (2.1E-04)	85.4 (5.0E+00)
	SDA	7.32 (1.2E+01)	75.24 (7.0E+00)	2.00E+01 (1.8E-06)	98.99 (6.0E-05)	224.4 (1.8E+00)
	GLOSS	17.18 (2.6E+01)	74.29 (1.1E+01)	2.63E+01 (8.3E+00)	99.57 (4.4E-04)	24 (2.4E+00)
Carcinom	SULDAAL	29.54 (1.1E+01)	94.24 (2.9E+00)	9.17E-06 (1.4E-06)	98.75 (2.5E-04)	802.3 (2.4E+01)
	GSULDAAL	42.79 (1.7E+01)	95.06 (2.1E+00)	9.43E-06 (2.2E-06)	96.14 (1.4E-03)	254.7 (1.3E+01)
	SULDA_ℓ_1	589.14 (1.6E+02)	93.53 (2.6E+00)	7.24E-06 (9.9E-07)	99.04 (4.8E-05)	642.3 (1.6E+01)
	SULDA_*admm*	6525.18 (1.7E+03)	94.00 (3.5E+01)	3.78E-07 (1.1E-07)	98.82 (2.6E-04)	790.7 (1.9E+01)
	SDA	2164.10 (1.2E+03)	92.59 (1.4E+00)	8.80E+01 (4.1E-06)	97.92 (1.3E-04)	1674.1 (2.3E+01)
	GLOSS	347.31 (2.8E+02)	92.94 (2.5E+00)	2.00E+02 (5.9E+00)	98.72 (1.2E-03)	117.4 (1.1E+01)
$ORL_{64\times64}$	SULDAAL	49.16 (1.4E+01)	91.80 (2.0E+00)	4.60E-04 (1.8E-05)	94.18 (2.9E-04)	3142.2 (2.4E+01)
	GSULDAAL	37.47 (1.4E+01)	92.65 (2.3E+00)	1.15E-03 (6.8E-04)	66.87 (6.7E-03)	1357.1 (2.7E+01)
	SULDA_ℓ_1	6854.57 (2.4E+03)	92.15 (2.6E+00)	2.26E-05 (2.1E-06)	95.14 (1.5E-05)	2928.1 (2.2E+01)
	SULDA_*admm*	14756.10 (5.3E+03)	91.85 (2.2E+00)	1.98E-05 (5.0E-06)	95.21 (6.8E-04)	2935.3 (2.6E+01)
	SDA	23435.93 (5.2E+03)	91.45 (2.2E+00)	1.99E+02 (3.6E-04)	93.68 (1.1E-04)	3399.4 (2.9E+01)
	GLOSS	27571.35 (1.2E+04)	93.70 (1.8E+00)	2.31E+02 (6.0E+00)	93.11 (2.9E-03)	282.3(1.2E+01)
Palmprint	SULDAAL	225.51 (8.9E+01)	98.87 (5.6E-01)	2.17E-06 (4.2E-08)	91.78 (2.1E-04)	4040 (7.3E+00)
	GSULDAAL	187.09 (7.4E+01)	98.90 (5.3E-01)	4.17E-06 (2.7E-07)	35.03 (5.9E-03)	2673.2 (2.4E+01)
	SULDA_ℓ_1	1900.59 (3.9E+02)	98.67 (5.3E-01)	1.77E-05 (5.7E-07)	92.67 (2.0E-05)	4014 (8.7E+00)
	SULDA_*admm*	44413.03 (9.4E+03)	98.80 (5.1E-01)	1.97E-07 (3.8E-08)	89.96 (1.2E-03)	4072.6 (4.0E+00)
	SDA	95638.45 (9.6E+03)	98.50 (8.6E-01)	2.99E+02 (6.0E-05)	90.71 (7.0E-05)	4081.4 (4.4E+00)
	GLOSS	23845.97 (5.2E+03)	98.47 (7.8E-01)	7.99E+01 (1.1E+00)	87.51 (2.9E-03)	507.6 (1.2E+01)
tr11	SULDAAL	34.31 (2.4E+00)	72.24 (5.0E+00)	1.61E-05 (5.8E-06)	96.45 (3.5E-04)	685.7 (1.7E+01)
	GSULDAAL	36.47 (2.4E+00)	73.37 (3.5E+00)	1.40E-05 (4.7E-06)	93.05 (9.8E-04)	444.1 (6.3E+00)
	SULDA_ℓ_1	516.04 (4.5E+01)	71.80 (4.8E+00)	2.44E-05 (1.3E-05)	96.78 (1.1E-04)	626.8 (1.3E+01)
	SULDA_*admm*	8964.04 (1.5E+03)	72.20 (5.0E+00)	3.76E-07 (4.3E-07)	96.31 (1.1E-03)	736.8 (1.6E+01)
	SDA	9462.01 (8.8E+02)	60.49 (3.3E+00)	2.08E+02 (2.2E-04)	95.63 (4.4E-04)	1272.2 (6.2E+01)
	GLOSS	1112.46 (5.0E+02)	66.34 (3.7E+00)	1.16E+03 (2.2E+02)	96.44 (1.9E-03)	229.1 (1.2E+01)
tr23	SULDAAL	15.41 (1.8E+00)	71.40 (4.1E+00)	6.05E-05 (2.7E-05)	98.05 (3.6E-04)	290.6 (8.5E+00)
	GSULDAAL	21.21 (3.3E+00)	71.30 (4.8E+00)	6.47E-05 (3.4E-05)	96.71 (1.2E-03)	191.6 (6.8E+00)
	SULDA_ℓ_1	248.59 (1.4E+02)	72.00 (3.5E+00)	9.13E-05 (5.6E-05)	98.24 (1.4E-04)	260.3 (5.8E+00)
	SULDA_*admm*	4651.27 (1.1E+03)	72.20 (4.4E+00)	6.91E-07 (5.5E-07)	98.06 (7.3E-04)	300.6 (1.5E+01)
	SDA	456.63 (8.9E+01)	67.70 (3.8E+00)	1.03E+02 (4.2E-03)	97.91 (1.4E-04)	423 (1.3E+01)
	GLOSS	148.77 (9.8E+01)	69.30 (4.8E+00)	1.20E+03 (5.5E+02)	98.15 (5.7E-04)	107.7 (3.3E+00)
tr41	SULDAAL	88.00 (8.1E+01)	85.34 (1.5E+00)	6.71E-06 (1.5E-06)	93.58 (6.3E-04)	1483.8 (2.5E+01)
	GSULDAAL	72.20 (2.2E+01)	85.53 (2.0E+00)	4.04E-06 (8.9E-07)	86.92 (2.0E-03)	974.9 (15E+01)
	SULDA_ℓ_1	886.61 (2.5E+02)	85.55 (1.8E+00)	1.35E-05 (5.0E-06)	94.09 (1.5E-04)	1366.4 (1.5E+01)
	SULDA_*admm*	17512.35 (3.2E+03)	85.34 (1.7E+00)	1.32E-07 (1.1E-07)	92.91 (1.8E-03)	1612.8 (2.3E+01)
	SDA	68172.13 (1.7E+04)	81.70 (2.1E+00)	4.41E+02 (4.6E-04)	92.09 (7.8E-04)	2869.2 (8.1E+01)
	GLOSS	3341.10 (1.2E+03)	83.88 (1.7E+00)	4.76E+03 (8.7E+02)	93.36 (4.5E-03)	495.1 (3.3E+01)
tr45	SULDAAL	72.60 (1.9E+01)	79.33 (3.3E+00)	1.62E-05 (2.8E-06)	95.41 (2.8E-04)	1180.3 (2.9E+01)
	GSULDAAL	76.58 (2.2E+01)	79.59 (1.3E+00)	1.01E-05 (1.9E-06)	90.68 (9.6E-04)	769.7 (7.9E+00)
	SULDA_ℓ_1	904.55 (3.4E+02)	79.56 (2.0E+00)	2.62E-05 (6.9E-06)	95.81 (5.4E-05)	1079.7 (1.7E+01)
	SULDA_*admm*	17957.25 (4.4E+03)	78.98 (1.4E+00)	3.55E-07 (2.4E-07)	95.05 (1.8E-03)	1282.8 (1.3E+01)
	SDA	36659.93 (1.1E+04)	75.04 (2.1E+00)	3.46E+02 (1.4E-04)	94.51 (4.6E-04)	2189.4 (5.1E+01)
	GLOSS	3052.20 (1.2E+03)	75.36 (3.5E+00)	1.79E+03 (1.8E+02)	95.24 (2.4E-03)	393.3 (2.0E+01)

one. From the 'Orthogonality' column, we see that our algorithms satisfy the orthogonality constraints very well, and the features extracted by algorithms SULDAL, GSULDAAL, SULDA_ℓ_1, and SULDA_*admm* are mutually uncorrelated. Among all algorithms, GLOSS selects the least number of features, followed by GSULDAAL, and the number of selected features of SULDAAL is similar to that of SULDA_ℓ_1 and SULDA_*admm*. Although both GSULDAAL and GLOSS use $\ell_{1,2}$-norm to promote group sparsity, the sparse solution computed by GLOSS is the solution of a penalized LDA [35] (i.e., it is an approximate solution of LDA) while the sparse solution computed by GSULDAAL is sought from the solution set of ULDA, which may explain why GLOSS selects less features than GSULDAAL.

5.3 Sparse Canonical Correlation Analysis

Canonical correlation analysis (CCA) [24] has been widely used in multivariate analysis for assessing the association between two random variables. Mathematically, given two centered data matrices $X \in \mathbb{R}^{n\times d_1}$ and $Y \in \mathbb{R}^{n\times d_2}$ corresponding to random variables $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$, respectively, CCA can be formulated as the following optimization problem:

$$\max_{W_x\in\mathbb{R}^{d_1\times l},\, W_y\in\mathbb{R}^{d_2\times l}} tr(W_x^T X^T Y W_y) \quad \text{s.t.} \quad W_x^T X^T X W_x = I,\ W_y^T Y^T Y W_y = I. \tag{27}$$

In this subsection, we apply Algorithm 2 to sparse CCA problem, which seeks sparse transformations W_x and W_y such that the projected variables in the low-dimensional space are highly correlated. We consider two different models of sparse CCA: *penalty model* (e.g. [22,51]) and *exact solution model* ([13,57]).

Penalty model computes sparse solution by penalizing the ℓ_1-norm of W_x and W_y on the basis of problem (27), leading to

$$\begin{aligned} \min_{W_x,\, W_y} \ & -tr(W_x^T X^T Y W_y) + \lambda_1\|W_x\|_1 + \lambda_2\|W_y\|_1 \\ & \text{s.t. } W_x^T X^T X W_x = I,\ W_y^T Y^T Y W_y = I, \end{aligned} \tag{28}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are parameters controlling the balance between sparsity and the correlation between XW_x and YW_y.

Exact solution model computes sparse solution by seeking the sparsest solutions from the set of minimum dimensional solutions of CCA. It has been shown [13] that all the minimum dimensional solutions of CCA satisfy

$$U_1^T W_x = \Sigma_1^{-1} P_1 W \quad \text{and} \quad V_1^T W_y = \Sigma_2^{-1} P_2 W, \tag{29}$$

where $X = Q_1\Sigma_1 U_1^T$ and $Y = Q_2\Sigma_2 V_1^T$ are the reduced SVDs of X and Y, respectively, $Q_1^T Q_2 = P_1\Sigma P_2^T$ is the reduced SVD of $Q_1^T Q_2$, and W is an arbitrary orthogonal matrix satisfying $W^T W = WW^T = I$. In [13,57], the authors proposed algorithm SCCA which fixes $W = I$ and solves

$$\min_{W_x}\{\|W_x\|_1 \mid U_1^T W_x = \Sigma_1^{-1} P_1\} \quad \text{and} \quad \min_{W_y}\{\|W_y\|_1 \mid V_1^T W_y = \Sigma_2^{-1} P_2\} \tag{30}$$

by using the accelerated linearized Bregman iteration method. Notice that the feasible set of problem (30) is just a subset of (29), to find optimal W from the set of orthogonal matrices, we consider

$$\min_{W_x,W_y} \|W_x\|_1 + \lambda\|W_y\|_1 \quad \text{s.t.} \quad U_1^T W_x = \Sigma_1^{-1} P_1 W,\ V_1^T W_y = \Sigma_2^{-1} P_2 W,\ W^T W = I \tag{31}$$

directly. When fixing $W = I$, problem (31) reduces to two independent problems in (30)

We name algorithms obtained by applying Algorithm 2 to problems (28) and (31) as SCCAALN and WSCCAAL, and outline details in Algorithm 6 and Algorithm 7, respectively, in Appendix D. In the implementation of SCCAALN, we choose $\gamma_1 = \gamma_2 = 0.51$, $\alpha^k = \frac{1.01}{\min\{1,\sigma_{min}(X^TX)\}}$, $\beta^k = \frac{1.01}{\min\{1,\sigma_{min}(Y^TY)\}}$, where $\sigma_{min}(X^TX)$ denotes the smallest positive singular value of X^TX. In the implementation of WSCCAAL, we choose $\gamma_1 = \gamma_2 = 0.51$, $\gamma_3 = 0.51(\frac{1}{\sigma_{min}(X^TX)} + \frac{1}{\sigma_{min}(Y^TY)})$. Parameter ρ_0 was set to be $1.01 \times \max\{\sigma_{min}(XX^T), \sigma_{min}(YY^T)\}$. We compare SCCAALN and WSCCAAL with SCCA [13], whose implementation is publicly available [6]. For SCCA, we terminate iteration when

$$\max\{\|U_1^T(W_x)^k - \Sigma_1^{-1}P_1\|_F/\|\Sigma_1^{-1}P_1\|_F,\ \|V_1^T(W_y)^k - \Sigma_2^{-1}P_2\|_F/\|\Sigma_2^{-1}P_2\|_F\} < 1E-5.$$

For SCCAALN and WSCCAAL, we terminate iteration when

$$\max\left\{\frac{\|W_x^k - W_x^{k-1}\|_F}{\max\{1,\|W_x^{k-1}\|_F\}}, \frac{\|W_y^k - W_y^{k-1}\|_F}{\max\{1,\|W_y^{k-1}\|_F\}}\right\} < 1E-3 \text{ and } \max\{\|R_1^k\|_\infty,\ \|R_2^k\|_\infty\} < 1E-5,$$

where R_1^k and R_2^k are defined in Algorithm 6 and Algorithm 7 for each algorithm.

5.3.1 Gene Expression Data Classification

It has been shown in [13] that ULDA is a special case of CCA when X is the data matrix and Y is an indicator matrix constructed from class information. Thus, SCCA can also be used for data classification. We test sparse CCA algorithms SCCA, SCCAALN and WSCCAAL on four gene expression data sets: Lymphoma, Srbct, Brain, and Carcinom, described in Table 4. For algorithms SCCAALN and WSCCAAL, we set regularization parameters $\lambda_1 = \lambda_2 = 0.8$ and $\lambda = 2$, respectively. Results are shown in Table 6, where 'Accuracy' denotes the classification accuracy using nearest neighbor classifier, 'Correlation' computed by $\frac{tr(W_x^TXY^TW_y)}{\sqrt{tr(W_x^TXX^TW_x)tr(W_y^TYY^TW_y)}}$ denotes canonical correlation between XW_x and YW_y on training (first component) and testing (second component) data, 'Orthogonality' records values $(\|W_x^TXX^TW_x - I_l\|_F/\sqrt{l},\ \|W_y^TYY^TW_y - I_l\|_F/\sqrt{l})$ on training data, 'Sparsity' and 'Selec. Fea.' record percentage of zeros and number of selected variables in W_x, respectively. It is observed that SCCAALN and WSCCAAL require much less time than SCCA while at the same time achieve comparable results in 'Accuracy', 'Correlation', 'Sparsity' and 'Selec. Fea.' as SCCA.

5.3.2 Cross-Language Document Retrieval

CCA is also capable of detecting semantic similarities in content between documents written in different languages. Thus, CCA has also been deployed to address the problem of cross-language document retrieval. Specifically, cross-language document retrieval involves a collection of documents with each being represented in different languages (e.g. English and French), and for a query document in one language (e.g. French) one is required to retrieve the most related document in another language (e.g. English). Details about this topic can be found in [13,45,46].

[6] The MATLAB implementation of SCCA with acceleration is available at `http://web.bii.a-star.edu.sg/~zhangxw/Publications.html`.

Table 6 Average results over 10 training-testing splittings: CPU time (CPU), classification accuracy in percentage (Accuracy), canonical correlation on training and testing data (Correlation), orthogonality constraints for W_x and W_y (Orthogonality), percentage or zero entries (Sparsity), and the number of selected features (Selec. Fea.) in W_x.

Data set	Algarithm	CPU	Accuracy	Correlation	Orthogonality	Sparsity	Selec. Fea.
Lymphoma	SCCA	342.18	98.67	(1.00, 0.90)	(5.48e-07, 5.36e-07)	99.22	58.70
	SCCAALN	2.11	98.33	(0.98, 0.90)	(1.50E-03, 5.25E-06)	97.74	164.40
	WSCCAAL	2.98	99.33	(1.00, 0.90)	(3.66E-04, 2.56E-06)	98.72	94.00
Srbct	SCCA	303.48	98.06	(1.00, 0.91)	(4.82e-07, 7.48e-07)	98.64	76.60
	SCCAALN	8.90	96.77	(0.96,0.90)	(4.76e-04, 5.83e-06)	98.88	60
	WSCCAAL	11.62	99.35	(1.00, 0.92)	(7.71E-05, 2.97E-06)	98.26	98.60
Brain	SCCA	519.66	73.33	(1.00, 0.67)	(6.37E-07, 3.25E-07)	99.63	79.40
	SCCAALN	3.52	79.05	(0.99,0.74)	(6.42E-04, 3.18E-06)	98.87	234.30
	WSCCAAL	5.34	75.24	(1.00, 0.70)	(2.41E-04, 4.05E-06)	99.25	155.60
Carcinom	SCCA	6889.27	92.94	(1.00, 0.86)	(6.72E-07, 1.07E-06)	99.04	621.10
	SCCAALN	89.90	92.94	(0.94, 0.85)	(3.74E-04, 1.88E-06)	99.36	382.00
	WSCCAAL	58.99	91.65	(1.00, 0.87)	(7.76E-05, 1.01E-06)	98.49	923.30

In this experiment, we test our approach on two data sets: Hansards [7] and Europarl [8]. The Hansards data set was obtained from the Aligned hansands of the 36th parliament of Canada, consisting of million pairs of text chunks aligned into English and French translations. Following the same pre-processing as in [13,57], we get a 5383×818 English term-by-document matrix and a 8015×818 French term-by-document matrix. The Europarl data set was obtained from the Europarl parallel corpus data set [27], where we get a 23308×202 English term-by-document matrix and a 33986×202 French term-by-document matrix. For each data set, we randomly split each term-by-document matrix into two parts, one is used as training data to obtain sparse CCA transformations and the other is used as testing data to test the retrieval ability of different algorithms. In this way, we obtain 400 and 100 training documents for Hansards and Europarl, respectively.

We compare our algorithms SCCAALN and WSCCAAL with SCCA, and use $\rho_0 = 1$ as the initial value of the Lagrangian penalty parameter for our algorithms . To evaluate the precision of document retrieval, we adopted the average area under the ROC (AROC) [10, 13, 19, 45] as the evaluation metric where a larger AROC value means more accurate. Performance statistics are listed in Table 7, from which we observe that our algorihtms take much less time than SCCA, and WSCCAAL is the most efficient algorithm. In the line of accuracy, SCCA and WSCCAAL achieve similar performance, since both algorithms attempt to find the sparsest solution from the set of minimum dimensional solutions. On the other hand, SCCAALN achieves higher accuracy than SCCA and WSCCAAL by a large margin on the Europarl data set. This may be attributed to the usage of regularization in the objective function. However, one drawback of SCCAALN is that the resulting solution is only an approximate of CCA solution, instead of exact solution as in SCCA and WSCCAAL.

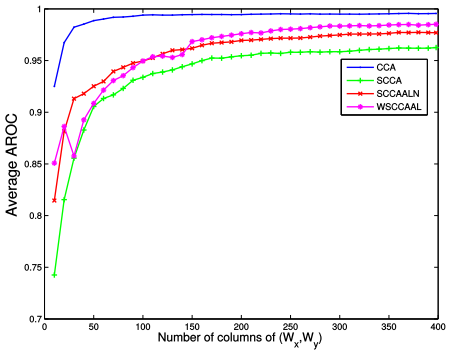
We also study the performance of the comparison algorithms when the number of columns in W_x and W_y changes, and plot the AROC and canonical correlation on the testing data in Fig. 3, where we also include the results of (non-sparse) CCA as a baseline method. It is observed that for all algorithms the retrieval performance improves as more columns of W_x and W_y are used. This is reasonable as more information is used to retrieve documents. It is worth noting that canonical correlation on the testing data correlation obtained by WSCCAALN is higher than that of SCCA and SCCAALN.

[7] `http://www.isi.edu/natural-language/download/hansard/`

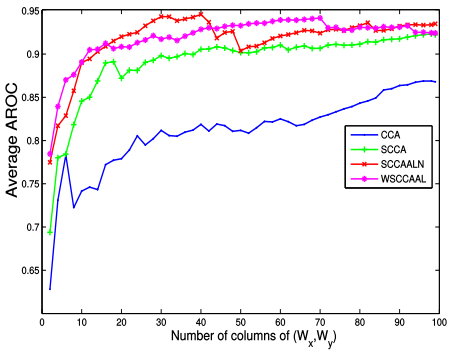
[8] `http://www.statmt.org/europarl/`

Table 7 Results of cross-language document retrieval on Hansands and Europarl: 'AROC' stands for Area under ROC, 'Correlation' records canonical correlation of (training data, testing data), 'Orthogonality' measures the orthogonality constraints of (W_x, W_y), 'Sparsity' measures the percentage of zeros in (W_x, W_y), and 'Selec.Fea.' measures the number of selected variables in (W_x, W_y).

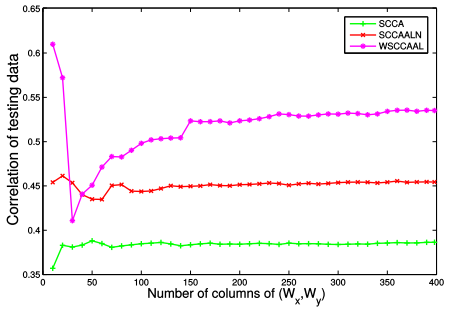
Data set	Algarithm	CPU	AROC	Correlation	Orthogonality	Sparsity	Selec. Fea.
Hansands	SCCA	9.43E+05	96.54	(1.00, 0.38)	(1.81E-05, 1.59E-05)	(92.09, 94.65)	(2604, 3223)
	SCCAALN	8.82E+04	97.81	(1.00, 0.45)	(1.32E-06, 1.26E-06)	(87.98, 91.83)	(2931, 3797)
	WSCCAAL	8.21E+02	98.51	(1.00, 0.54)	(5.42E-07, 6.44E-07)	(89.99, 92.94)	(2724, 3483)
Europarl	SCCA	2.25E+06	92.81	(1.00, 0.60)	(1.50E-05, 1.53E-05)	(99.51, 99.66)	(603, 597)
	SCCAALN	4.59E+04	97.06	(1.00, 0.74)	(2.75E-06, 2.69E-06)	(98.73, 99.12)	(1322, 1331)
	WSCCAAL	1.03E+02	92.43	(1.00, 0.75)	(6.60E-07, 1.11E-07)	(99.27, 99.11)	(723, 1143)



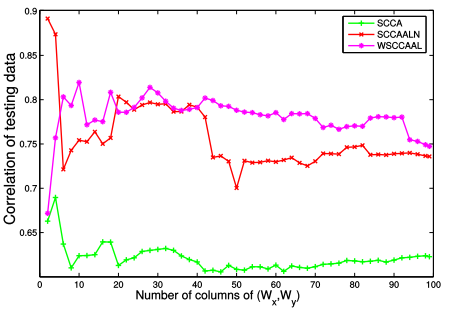
(a) Average AROC on Hansards data



(b) Average AROC on Europarl data



(c) Canonical correlation on Hansards testing data



(d) Canonical correlation on Europarl testing data

Fig. 3 Average AROC and canonical correlation of testing data as functions of the number of columns in (W_x, W_y). Left: Hansands data set; Right: Europarl data set.

6 Conclusions

In this paper, we propose an approximate augmented Lagrangian method, based on the augmented Lagrangian scheme and the extended proximal alternating linearized minimization method, to solve a class of nonconvex and nonsmooth problems with generalized orthogonality constraints. We also provide global convergence results for the proposed method in the way that if the sequence generated by our method is bounded, then every limit point is a KKT point. Numerical experiments indicate that the proposed algorithms are efficient and

achieve comparable performance with existing methods. In the future, we plan to study the acceleration of our methods.

7 Acknowledgement

The authors would like to thank the two anonymous referees for their valuable comments and suggestions. The work of L.-Z. Liao was supported in part by grants from Hong Kong Baptist University (FRG) and General Research Fund (GRF) of Hong Kong.

Appendix A Proof of Lemma 2

To prove Lemma 2, we need a few preliminary results regarding the limiting subdifferential of indicator functions. For any closed set $\mathscr{X}$, it is well-known [37] that the limiting subdifferential of indicator function $\delta_{\mathscr{X}}$ at x is given by the normal cone to $\mathscr{X}$ with respect to x denoted by $N_{\mathscr{X}}(x)$, that is,

$$\partial \delta_{\mathscr{X}}(x) = N_{\mathscr{X}}(x),\ \forall x \in \mathscr{X}.$$

Next, we consider the normal cone of two specific closed sets. Let $\ell(X) = \mathscr{A}X : \mathbb{R}^{m\times n} \to \mathbb{R}^{\gamma\times n}$ be a linear mapping with $\mathscr{A} \in \mathbb{R}^{\gamma\times m}$, and define the closed set $\mathscr{X} := \{X \in \mathbb{R}^{m\times n} \mid \mathscr{A}X = 0\}$. It is easy to show that

$$N_{\mathscr{X}}(X) = \{\mathscr{A}^T\Gamma \mid \Gamma \in \mathbb{R}^{\gamma\times n}\}. \tag{32}$$

Let $\mathscr{M} = \{X \in \mathbb{R}^{n\times q} \mid X^TMX = I_q, M \in S_+^n\}$ be the set of matrices satisfying the generalized orthogonality constraints, it follows that for any curve $Y(t) \in \mathscr{M}$ with $Y(0) = X$, $(Y'(t))^TMY(t) + Y(t)^TMY'(t) = 0$. Let $t = 0$, notice that $Y'(0) \in T_{\mathscr{M}}(X)$, we have

$$\eta^TMX + X^TM\eta = 0 \quad \forall\ \eta \in T_{\mathscr{M}}(X).$$

Let $MX = \bar{U}\bar{\Sigma}\bar{V}^T$ be the reduced SVD of MX, and $(MX)_\perp$ be a column orthogonal matrix such that $[\bar{U}\ (MX)_\perp]$ is an orthogonal matrix, then η can be written as

$$\eta = \bar{U}\eta_1 + (MX)_\perp\eta_2,$$

where $\eta_1^T\bar{\Sigma}\bar{V}^T + \bar{V}\bar{\Sigma}\eta_1 = 0$, which means $\bar{V}\bar{\Sigma}\eta_1$ is skew-symmetric. Moreover,

$$\begin{aligned}\eta &= \bar{U}\eta_1 + (MX)_\perp\eta_2\\ &= \bar{U}\bar{\Sigma}^{-1}\bar{V}^T(\bar{V}\bar{\Sigma}\eta_1) + (MX)_\perp\eta_2\\ &= (X^TM)^\dagger(\bar{V}\bar{\Sigma}\eta_1) + (MX)_\perp\eta_2,\end{aligned}$$

where $(X^TM)^\dagger$ is the pseudo-inverse of X^TM. Therefore, the tangent space of $\mathscr{M}$ is given by

$$T_{\mathscr{M}}(X) = \{(X^TM)^\dagger\Omega + (MX)_\perp K \mid \Omega^T + \Omega = 0\}.$$

In addition, any $\zeta \in N_{\mathscr{M}}(X)$ can be written as

$$\zeta = \bar{U}\zeta_1 + (MX)_\perp\zeta_2 = (MX)\bar{V}\bar{\Sigma}^{-1}\zeta_1 + (MX)_\perp\zeta_2.$$

Since $\langle \zeta, \eta \rangle = 0$ for any $\eta \in T_{\mathscr{M}}(X)$, we have

$$\langle \eta_1, \zeta_1 \rangle = 0 \quad \text{and} \quad \zeta_2 = 0.$$

The first equality implies that

$$0 = \langle \eta_1, \zeta_1 \rangle = \langle \bar{V}\bar{\Sigma}\eta_1, \bar{V}\bar{\Sigma}^{-1}\zeta_1 \rangle,$$

and thus $\bar{V}\bar{\Sigma}^{-1}\zeta_1$ must be symmetric. Hence, the normal cone of $\mathscr{M}$ at X is given by

$$N_{\mathscr{M}}(X) = \{(MX)S \mid S \in S^q\}. \tag{33}$$

Now, we are ready to prove Lemma 2.

Proof Equalities

$$AX^* + BY^* - C = 0,\ X^* - G^* = 0 \text{ and } (G^*)^T M G^* = I_q$$

hold since (X^*, Y^*, G^*) is feasible as a local minimizer. For the convenience of analysis, we denote $W := \begin{pmatrix} X \\ Y \\ G \end{pmatrix} \in \mathbb{R}^{(2n+m)\times q}$ and define $g_1 : \mathbb{R}^{(2n+m)\times q} \to \mathbb{R}^{(l+n)\times q}$ as

$$g_1(W) = \begin{pmatrix} A & B & 0 \\ I & 0 & -I \end{pmatrix} \begin{pmatrix} X \\ Y \\ G \end{pmatrix}.$$

Let $\Omega = \{W \in \mathbb{R}^{(2n+m)\times q} \mid g_1(W) = 0\}$, then problem (9) is equivalent to

$$\min_W\ f(X) + g(Y) + h(X,Y) + \delta_{\mathscr{M}}(G) + \delta_{\Omega}(W).$$

Since (X^*, Y^*, G^*) is a local minimizer, by the generalized Fermat's rule and subdifferentiability property [15,43], we have

$$0 \in \begin{pmatrix} \partial f(X^*) + \partial_X h(X^*, Y^*) \\ \partial g(Y^*) + \partial_Y h(X^*, Y^*) \\ \partial \delta_{\mathscr{M}}(G^*) \end{pmatrix} + \partial \delta_{\Omega}(W^*).$$

As described at the beginning of this Appendix A, subdifferential of indicator functions $\delta_{\mathscr{M}}$ and δ_{Ω} are given by the normal cones (33) and (32), respectively. In particular,

$$\partial \delta_{\mathscr{M}}(G^*) = N_{\mathscr{M}}(G^*) = \{MG^*S \mid S \in S^q\},$$

and

$$\partial \delta_{\Omega}(W^*) = N_{\Omega}(W^*) = \left\{ \begin{pmatrix} A^T & I \\ B^T & 0 \\ 0 & -I \end{pmatrix} \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix} \mid \Lambda_1 \in \mathbb{R}^{l\times q},\ \Lambda_2 \in \mathbb{R}^{n\times q} \right\}.$$

Therefore, there exist $v^* \in \partial f(X^*),\ w^* \in \partial g(Y^*),\ \Lambda_1^* \in \mathbb{R}^{l\times q},\ \Lambda_2^* \in \mathbb{R}^{n\times q},\ \Lambda_3^* \in S^q$ such that

$$0 = \begin{pmatrix} v^* + \partial_X h(X^*, Y^*) \\ w^* + \partial_Y h(X^*, Y^*) \\ 2MG^*\Lambda_3^* \end{pmatrix} + \begin{pmatrix} A^T & I \\ B^T & 0 \\ 0 & -I \end{pmatrix} \begin{pmatrix} \Lambda_1^* \\ \Lambda_2^* \end{pmatrix} = \begin{pmatrix} v^* + \partial_X h(X^*, Y^*) \\ w^* + \partial_Y h(X^*, Y^*) \\ 0 \end{pmatrix} + \begin{pmatrix} A^T & I & 0 \\ B^T & 0 & 0 \\ 0 & -I & 2MG^* \end{pmatrix} \begin{pmatrix} \Lambda_1^* \\ \Lambda_2^* \\ \Lambda_3^* \end{pmatrix},$$

which proves equality (12). Moreover, it yields that $\Lambda_2^* = 2MG^*\Lambda_3^*$. Substitute this into (12) and eliminate G^*, we get equality (13). This completes the proof. ■

Appendix B Proof of Theorem 1

Proof For any limit point (X^*,Y^*,G^*) of the bounded sequence $\{(X^k,Y^k,G^k)\}_{k\in\mathbb{N}}$, there exists an index set $\mathscr{K}\subset\mathbb{N}$ such that $\{(X^k,Y^k,G^k)\}_{k\in\mathscr{K}}$ converges to (X^*,Y^*,G^*). To prove that (X^*,Y^*,G^*) is a KKT point, we first show that it is a feasible point. The equality $(G^*)^TMG^*=I_q$ is trivial to check since $(G^k)^TMG^k=I_q$ holds for any $k\in\mathbb{N}$. If $\{\rho_k\}$ is bounded, then by the updating rule of ρ_k in Algorithm 2, there exists an $k_0\in\mathbb{N}$ such that

$$\|R_j^k\|_\infty\le\tau\|R_j^{k-1}\|_\infty\quad\forall k\ge k_0,\ j=1,2.$$

By the definition of R_j^k, $j=1,2$, it holds that

$$\begin{cases}\|AX^k+BY^k-C\|_\infty\le\tau\|AX^{k-1}+BY^{k-1}-C\|_\infty,\\ \|X^k-G^k\|_\infty\le\tau\|X^{k-1}-G^{k-1}\|_\infty,\end{cases}$$

for any $k\ge k_0$. Thus

$$\begin{cases}AX^*+BY^*-C=0,\\ X^*-G^*=0.\end{cases}$$

If $\{\rho_k\}$ is unbounded, by the generalized Fermat rule, finding a solution satisfying the constraint (11) is equivalent of calculating a point (X^k,Y^k,G^k) such that

$$\left\|\begin{pmatrix}v^k+\partial_X h(X^k,Y^k)\\ w^k+\partial_Y h(X^k,Y^k)\\ 0\end{pmatrix}/\rho_{k-1}+\begin{pmatrix}A^T & I & 0\\ B^T & 0 & 0\\ 0 & -I & 2MG^k\end{pmatrix}\begin{pmatrix}\frac{\bar\Lambda_1^{k-1}}{\rho_{k-1}}+(AX^k+BY^k-C)\\ \frac{\bar\Lambda_2^{k-1}}{\rho_{k-1}}+(X^k-G^k)\\ \Lambda_3^k/\rho_{k-1}\end{pmatrix}\right\|_\infty\le\frac{\varepsilon_{k-1}}{\rho_{k-1}},\tag{34}$$

for some $v^k\in\partial f(X^k)$, $w^k\in\partial g(Y^k)$, $\Lambda_3^k\in\mathscr{S}^q$, and $\varepsilon_k\downarrow 0$ as $k\to\infty$. Notice that $\{\bar\Lambda_1^k\}$ and $\{\bar\Lambda_2^k\}$ are bounded, $\{v^k\}_{k\in\mathscr{K}}$, $\{w^k\}_{k\in\mathscr{K}}$, $\{\partial_X h(X^k,Y^k)\}$ and $\{\partial_Y h(X^k,Y^k)\}$ are bounded under Assumption 1 (iii)-(iv). Let $k\in\mathscr{K}$ go to infinity, equation (34) implies that

$$\begin{pmatrix}A^T & I\\ B^T & 0\end{pmatrix}\begin{pmatrix}AX^*+BY^*-C\\ X^*-G^*\end{pmatrix}=0.$$

Recall that B has full row rank, we get $AX^*+BY^*-C=0$ and $X^*-G^*=0$. Therefore, in both cases, we show that (X^*,Y^*,G^*) is a feasible point.

Next, we show that there exist $\Lambda_1^*\in\mathbb{R}^{k\times q}$, $\Lambda_2^*\in\mathbb{R}^{n\times q}$ and $\Lambda_3^*\in\mathscr{S}^q$ such that $(X^*,Y^*,G^*;\Lambda_1^*,\Lambda_2^*,\Lambda_3^*)$ satisfies (12). If $\{X^k,Y^k,G^k\}_{k\in\mathbb{N}}$ is bounded, there exists an index set $\mathscr{K}\subseteq\mathbb{N}$, such that $\lim_{k\in\mathscr{K}}(X^k,Y^k,G^k)=(X^*,Y^*,G^*)$. Since $\{v^k\}_{k\in\mathscr{K}}$ is bounded, there exists a subsequence $\mathscr{K}_2\subseteq\mathscr{K}$ such that $\lim_{k\in\mathscr{K}_2}v^k=v^*$. Moreover, by the closedness property of the limiting subdifferential, we get

$$v^*\in\partial f(X^*).$$

Similarly, there exists a subsequence $\mathscr{K}_3\subseteq\mathscr{K}_2$ such that $\lim_{k\in\mathscr{K}_3}w^k=w^*$, and

$$w^*\in\partial g(Y^*).$$

Combining with the updating formula of $\bar\Lambda_1^k$ and $\bar\Lambda_2^k$ in **Step 2** of Algorithm 2, (34) implies that there exists a ξ^k with $\|\xi^k\|_\infty\le\frac{\varepsilon_{k-1}}{\rho_{k-1}}$ such that

$$\begin{pmatrix} A^T & I & 0 \\ B^T & 0 & 0 \\ 0 & -I & 2MG^k \end{pmatrix} \begin{pmatrix} \Lambda_1^k \\ \Lambda_2^k \\ \Lambda_3^k \end{pmatrix} / \rho_{k-1} = \xi^k - \begin{pmatrix} v^k + \partial_X h(X^k, Y^k) \\ w^k + \partial_Y h(X^k, Y^k) \\ 0 \end{pmatrix} / \rho_{k-1}. \tag{35}$$

Define

$$\Xi^k = \begin{pmatrix} A^T & I & 0 \\ B^T & 0 & 0 \\ 0 & -I & 2MG^k \end{pmatrix} \in R^{(2n+m)\times(2k+q)}, \quad \Upsilon^k = \begin{pmatrix} \Lambda_1^k \\ \Lambda_2^k \\ \Lambda_3^k \end{pmatrix},$$

we can rewrite (35) as

$$\Xi^k \Upsilon^k = \rho_{k-1}\xi^k - \begin{pmatrix} v^k + \partial_X h(X^k, Y^k) \\ w^k + \partial_Y h(X^k, Y^k) \\ 0 \end{pmatrix}.$$

Since the columns of Ξ^k are linearly independent, $(\Xi^k)^T(\Xi^k)$ is nonsingular. Thus

$$\Upsilon^k = ((\Xi^k)^T \Xi^k)^{-1}(\Xi^k)^T \left[\rho_{k-1}\xi^k - \begin{pmatrix} v^k + \partial_X h(X^k, Y^k) \\ w^k + \partial_Y h(X^k, Y^k) \\ 0 \end{pmatrix} \right]. \tag{36}$$

Taking limit on (36) as $k \in \mathscr{K}_3$ goes to infinity, and noticing that $\|\xi^k\|_\infty \leq \frac{\varepsilon_{k-1}}{\rho_{k-1}}$ with $\varepsilon_k \downarrow 0$ as $k \to \infty$, we have

$$\lim_{k\in\mathscr{K}_3, k\to\infty} \Upsilon^k = \Upsilon^* := -((\Xi^*)^T \Xi^*)^{-1}(\Xi^*)^T \begin{pmatrix} v^* + \partial_X h(X^*, Y^*) \\ w^* + \partial_Y h(X^*, Y^*) \\ 0 \end{pmatrix},$$

where

$$\Xi^* = \begin{pmatrix} A^T & I & 0 \\ B^T & 0 & 0 \\ 0 & -I & 2MG^* \end{pmatrix}$$

has full column rank. From the definition of Υ^k, taking limit $k \in \mathscr{K}_3$ goes to infinity on both sides of (35) yields that

$$\begin{pmatrix} v^* + \partial_X h(X^*, Y^*) \\ w^* + \partial_Y h(X^*, Y^*) \\ 0 \end{pmatrix} + \begin{pmatrix} A^T & I & 0 \\ B^T & 0 & 0 \\ 0 & -I & 2MG^* \end{pmatrix} \begin{pmatrix} \Lambda_1^* \\ \Lambda_2^* \\ \Lambda_3^* \end{pmatrix} = 0,$$

where $\Lambda_3^* \in \mathscr{S}^q$ since $\Lambda_3^k \in \mathscr{S}^q$ for any $k \in \mathbb{N}$. According to Lemma 2, (X^*, Y^*, G^*) is a KKT point of problem (9). Moreover, (X^*, Y^*) is a KKT point of problem (1). ■

Appendix C Proof of Proposition 1

Proof In this proof, we assume k is fixed and for the simplicity of notations, we let $W := \begin{pmatrix} X \\ Y \\ G \end{pmatrix}$ and $L_k(W) := L_{\rho_{k-1}}(X,Y,G;\bar{\Lambda}^{k-1})$.

1) By the first-order optimality conditions of three subproblems in (15), there exist $v^{k,j} \in \partial f_1^k(X^{k,j})$, $w^{k,j} \in \partial f_2^k(Y^{k,j})$, $\nu^{k,j} \in \partial f_3^k(G^{k,j})$ such that

$$\begin{cases} 0 = v^{k,j} + \nabla_X H_k(X^{k,j-1}, Y^{k,j-1}, G^{k,j-1}) + B_1^{k,j-1}(X^{k,j} - X^{k,j-1}), \\ 0 = w^{k,j} + \nabla_Y H_k(X^{k,j}, Y^{k,j-1}, G^{k,j-1}) + B_2^{k,j-1}(Y^{k,j} - Y^{k,j-1}), \\ 0 = \nu^{k,j} - \bar{\Lambda}_1^{k-1}/\rho_{k-1} + (G^{k,j} - X^{k,j}) + B_3^k(G^{k,j} - G^{k,j-1}). \end{cases}$$

Combined with the above relationships, by simple calculations, $A^{k,j}$ defined in (19) has the following properties,

$$\begin{aligned} A_X^{k,j} &= \nabla_X H_k(W^{k,j}) - \nabla_X H_k(W^{k,j-1}) + B_1^{k,j-1}(X^{k,j-1} - X^{k,j}) \\ &= v^{k,j} + \nabla_X H_k(W^{k,j}) \in \partial_X L_k(W^{k,j}), \\ A_Y^{k,j} &= \nabla_Y H_k(W^{k,j}) - \nabla_Y H_k(X^{k,j}, Y^{k,j-1}, Z^{k,j-1}) + B_2^{k,j-1}(Y^{k,j-1} - Y^{k,j}), \\ &= w^{k,j} + \nabla_Y H_k(W^{k,j}) \in \partial_Y L_k(W^{k,j}), \\ A_G^{k,j} &= B_3^k(G^{k,j-1} - G^{k,j}) = \nu^{k,j} - \bar{\Lambda}_1^{k-1}/\rho_{k-1} + (G^{k,j} - X^{k,j}) \\ &= \nu^{k,j} + \nabla_G H_k(W^{k,j}) \in \partial_G L_k(W^{k,j}). \end{aligned}$$

Since $H_k(W)$ is continuously differentiable, by subdifferentiability property [5,43],

$$\partial L_k(W) = \partial_X L_k(W) \times \partial_Y L_k(W) \times \partial_G L_k(W),$$

which implies

$$A^{k,j} \in \partial L_k(W^{k,j}) = \partial L_{\rho_{k-1}}(X^{k,j}, Y^{k,j}, G^{k,j}; \bar{\Lambda}^{k-1}) \quad \forall j \in \mathbb{N}.$$

To show $\|A^{k,j}\|_\infty \to 0$, it suffices to show that $L_k(W)$ satisfies the conditions in Assumption 2 and apply Lemma 1 c).

We first show that $\{W^{k,j}\}_{j\in\mathbb{N}}$ generated by Algorithm 3 is bounded. This can be accomplished via proof by contradiction. Notice that $\tilde{L}_k(W) = \rho_{k-1}L_k(W)$ is a coercive function under the assumption that $\phi(X,Y) + \frac{\rho_0}{2}\|AX + BY - C\|_F^2$ is coercive function and the fact that $\{\rho_k\}_{k\in\mathbb{N}}$ is non-decreasing, $\delta_{\mathscr{M}}(G)$ is a coercive function and $\frac{\rho_{k-1}}{2}\|X - G + \bar{\Lambda}_2^{k-1}\|_F^2 - \frac{1}{2\rho_{k-1}}\|\bar{\Lambda}^{k-1}\|_F^2$ is bounded from below. Suppose $\lim_{j\to\infty}\|W^{k,j}\|_\infty = +\infty$, then there must hold

$$\lim_{j\to\infty} \tilde{L}_k(W^{k,j}) = +\infty.$$

On the other hand, we know from Lemma 1 a) that $\{L_k(W^{k,j})\}_{j\in\mathbb{N}}$ is a decreasing sequence, thus $\{\tilde{L}_k(W^{k,j})\}_{j\in\mathbb{N}}$ is non-increasing, which implies

$$\lim_{j\to\infty} \tilde{L}_k(W^{k,j}) \le \tilde{L}_k(W^{k,0}) < +\infty.$$

Hence, a contradiction, and $\{W^{k,j}\}_{j\in\mathbb{N}}$ is bounded.

Now, we verify that $L_k(W)$ satisfies the conditions in Assumption 2. By the definitions of $L_k(W)$, $H_k(W)$, f_i^k, $i = 1,2,3$ and Assumption 1 (i), it is easy to see that for any given $\bar{\Lambda}^{k-1}$ and ρ_{k-1}, the following results hold:

(a) f_i^k, $i=1,2,3$, are proper and lower semicontinuous functions satisfying $\inf f_i^k > -\infty$, H_k is a C^1 function, and

$$\begin{aligned} H_k(W) =& \frac{1}{\rho_{k-1}} h(X,Y) + \frac{1}{2}\|AX+BY-C+\bar{\Lambda}_1^{k-1}/\rho_{k-1}\|_F^2 \\ &+ \frac{1}{2}\|X-G+\bar{\Lambda}_2^{k-1}/\rho_{k-1}\|_F^2 - \frac{1}{2\rho_{k-1}^2}\|\bar{\Lambda}^{k-1}\|_F^2. \end{aligned} \tag{37}$$

Thus $\inf_W H_k(W) \geq \frac{1}{\rho_{k-1}} \min_{X,Y} h(X,Y) - \frac{1}{2(\rho_{k-1})^2}\|\bar{\Lambda}^{k-1}\|_F^2 > -\infty$.

(b) Since H_k is a quadratic function with respect to G, $\nabla_G H_k$ is obviously Lipschitz continuous. Regarding the Lipschitz continuity of partial derivatives $\nabla_X h(X,Y)$ and $\nabla_Y h(X,Y)$, we have

$$\|\nabla_X H_k(X,Y^{k,j-1},G^{k,j-1}) - \nabla_X H_k(\tilde{X},Y^{k,j-1},G^{k,j-1})\| \leq L_1^{k,j-1}\|X-\tilde{X}\|,$$

where $L_1^{k,j-1} = \frac{L_1(Y^{k,j-1})}{\rho_{k-1}} + \|A^TA\| + 1$, and

$$\|\nabla_Y H_k(X^{k,j},Y,G^{k,j-1}) - \nabla_Y H_k(X^{k,j},\tilde{Y},G^{k,j-1})\| \leq L_2^{k,j-1}\|Y-\tilde{Y}\|,$$

where $L_2^{k,j-1} = \frac{L_2(X^{k,j})}{\rho_{k-1}} + \|B^TB\| + 1$. In addition, let $\bar{L}_1 = \sup_Y L_1(Y)$ and $\bar{L}_2 = \sup_X L_2(X)$, then the boundedness of $\{W^{k,j}\}_{j\in\mathbb{N}}$ and Assumption 1 (iii) imply that $\bar{L}_1 < \infty$ and $\bar{L}_2 < \infty$, and we have

$$\|A^TA\|+1 \leq L_1^{k,j-1} \leq \bar{L}_1/\rho_0 + \|A^TA\| + 1, \quad \|B^TB\|+1 \leq L_2^{k,j-1} \leq \bar{L}_2/\rho_0 + \|B^TB\| + 1,$$

which imply

$$\inf_j\{L_1^{k,j-1}\} \geq \|A^TA\|+1 > -\infty, \quad \inf_j\{L_2^{k,j-1}\} \geq \|B^TB\|+1 > -\infty,$$

and

$$\sup_j\{L_1^{k,j-1}\} \leq \bar{L}_1/\rho_0 + \|A^TA\|+1 < +\infty, \quad \sup_j\{L_2^{k,j-1}\} \leq \bar{L}_2/\rho_0 + \|B^TB\|+1 < +\infty.$$

Moreover, Assumption 2 (iii) holds by the definition of $B_i^{k,j-1}$, $i=1,2$ and B_3^k. Thus, Assumption 2 (i)-(iii) hold.

(c) Assumption 2 (iv) holds since $h(X,Y)$ satisfies Assumption 1 (iii).

2) From the proof of 1), we know that $\{W^{k,j}\}_{j\in\mathbb{N}}$ is bounded. Then the proof of 2) remains to show that $L_k(W)$ is a K-L function and apply Lemma 1 d). Notice that

$$\begin{aligned} L_k(W) =& \frac{1}{\rho_{k-1}}\phi(X,Y) + \frac{1}{\rho_{k-1}}\delta_{\mathscr{M}}(G) + \frac{1}{2}\|AX+BY-C+\bar{\Lambda}_1^{k-1}/\rho_{k-1}\|_F^2 \\ &+ \frac{1}{2}\|X-G+\bar{\Lambda}_2^{k-1}/\rho_{k-1}\|_F^2 - \frac{1}{2\rho_{k-1}^2}\|\bar{\Lambda}^{k-1}\|_F^2, \end{aligned} \tag{38}$$

i.e., $L_k(W)$ satisfies the K-L properties as a finite sum of functions satisfying the K-L properties, then the result holds directly. ■

Appendix D Outline of Algorithms for CMs, Sparse ULDA and Sparse CCA

In this section, we describe the detailed derivations of algorithms for CMs, sparse ULDA and sparse CCA in Section 5.

D.1 Compressed Modes

The scaled augmented Lagrangian function associated with (21) is

$$L_{\rho_{k-1}}(\Psi,X;\bar{\Lambda}^{k-1}) = \frac{1}{\kappa\rho_{k-1}}\|\Psi\|_1 + \frac{1}{\rho_{k-1}}\delta_{\mathscr{X}}(X) + H_k(\Psi,X),$$

where $\mathscr{X} = St(n,N)$ and

$$H_k(\Psi,X) = \frac{1}{\rho_{k-1}}tr(\Psi^T H\Psi) + \langle \bar{\Lambda}^{k-1}/\rho_{k-1}, \Psi - X\rangle + \frac{1}{2}\|\Psi - X\|_F^2.$$

Applying Algorithm 3 with $B_1^{k,j-1} = \gamma_1^k I$ and $B_2^{k,j-1} = \gamma_2 I$, we get the following updating of $(\Psi^{k,j}, X^{k,j})$ for any fixed $k \in \mathbb{N}$

$$\Psi^{k,j} = \mathbf{shrink}\left(\Psi^{k,j-1} - \frac{1}{\gamma_1}\left(\frac{2}{\rho_{k-1}}H\Psi^{k,j-1} + \frac{\bar{\Lambda}^{k-1}}{\rho_{k-1}} + \Psi^{k,j-1} - X^{k,j-1}\right),\ \frac{1}{\gamma_1^k\kappa\rho_{k-1}}\right), \tag{39}$$

$$X^{k,j} = \arg\max_{X^TX=I}\left\langle X,\ \Psi^{k,j} + \frac{\bar{\Lambda}^{k-1}}{\rho_{k-1}} + (\gamma_2 - 1)X^{k,j-1}\right\rangle = \tilde{U}\tilde{V}^T, \tag{40}$$

where $\mathbf{shrink}(x,\eta) = \mathrm{sign}(x)\odot\max\{|x| - \eta, 0\}$ is the soft-shrinkage operator and $\odot$ denotes component-wise product, and $\tilde{U}\tilde{\Sigma}\tilde{V}^T = \Psi^{k,j} + \frac{\bar{\Lambda}^{k-1}}{\rho_{k-1}} + (\gamma_2 - 1)X^{k,j-1}$ is the reduced SVD. Outline of the algorithm is given in Algorithm 4.

D.2 Sparse ULDA

By introducing auxiliary variable $Z = X$, the scaled augmented Lagrangian function associated with (25) is

$$L_{\rho_{k-1}}(X,G,Z;\bar{\Lambda}^{k-1}) = \frac{1}{\rho_{k-1}}\|G\|_1 + \frac{1}{\rho_{k-1}}\delta_{\mathscr{O}}(Z) + H_k(X,G,Z),$$

where $\mathscr{O}$ denotes the set of orthogonal matrices and

$$H_k(X,G,Z) = \left\langle \bar{\Lambda}_1^{k-1}/\rho_{k-1},\ \mathscr{A}X + \mathscr{B}G\right\rangle + \frac{1}{2}\|\mathscr{A}X + \mathscr{B}G\|_F^2 + \left\langle \bar{\Lambda}_2^{k-1}/\rho_{k-1}, X - Z\right\rangle + \frac{1}{2}\|X - Z\|_F^2.$$

Algorithm 4 (CMsAL: Algorithm 2 for CMs problem (21))

Input: Data matrix H, parameters $\{\varepsilon_k\}_{k\in\mathbb{N}}\downarrow 0$, $\{\bar{\Lambda}_{min}\le\bar{\Lambda}_{max}\}$, $\tau\in[0,1)$, $\mu>1$, $\rho_0>0$, and $\{\gamma_i\}_{1\le i\le 2}$.
Output: (Ψ^k,X^k).

Choose $\{\Psi^0,X^0\}$ randomly such that $(X^0)^TX^0=I$. Let $k=1$.
while stopping criterion is not satisfied **do**
  Let $(\Psi^{k,0},X^{k,0})=(\Psi^{k-1},X^{k-1})$, $j=1$ and compute $A^{k,j}$.
  while $\|A^{k,j}\|_\infty>\frac{\varepsilon_{k-1}}{\rho_{k-1}}$ **do**
    Compute $(\Psi^{k,j},X^{k,j})$ using (39)-(40) for CMsAL.
    Compute $A^{k,j}$ as in (19).
  end while
  Let $\Psi^k=\Psi^{k,j}$ and $X^k=X^{k,j}$.
  Update the Lagrangian multiplier

$$\Lambda^k=\bar{\Lambda}^{k-1}+\rho_{k-1}(\Psi^k-X^k),$$

  where $\bar{\Lambda}^k$ is the projection of Λ^k on $\{\Lambda:\bar{\Lambda}_{min}\le\Lambda\le\bar{\Lambda}_{max}\}$.
  Update the penalty parameter

$$\rho^k=\begin{cases}\rho_{k-1}, & \text{if } \|\Psi^k-X^k\|_\infty\le\tau\|\Psi^{k-1}-X^{k-1}\|_\infty,\\ \mu\rho_{k-1}, & \text{otherwise.}\end{cases}$$

end while
return Ψ^k and X^k.

Algorithm 5 (SULDAAL: Algorithm 2 for sparse ULDA problem (25)) / (GSULDAAL: Algorithm 2 for group sparse ULDA problem (26))

Input: Data matrix A with class labels, parameters $\{\varepsilon_k\}_{k\in\mathbb{N}}\downarrow 0$, $\{\bar{\Lambda}_{i,min}\le\bar{\Lambda}_{i,max}\}_{i=1,2}$, $\tau\in[0,1)$, $\mu>1$, $\rho_0>0$, and $\{\gamma_i\}_{1\le i\le 3}$.
Output: (X^k,G^k,Z^k).

Compute $\mathscr{A}$ and $\mathscr{B}$ from data matrix and choose $\{X^0,G^0,Z^0\}$ randomly such that $(Z^0)^TZ^0=I$. Let $k=1$.
while stopping criterion is not satisfied **do**
  Let $(X^{k,0},G^{k,0},Z^{k,0})=(X^{k-1},G^{k-1},Z^{k-1})$, $j=1$ and compute $A^{k,j}$.
  while $\|A^{k,j}\|_\infty>\frac{\varepsilon_{k-1}}{\rho_{k-1}}$ **do**
    Compute $(X^{k,j},G^{k,j},Z^{k,j})$ using (41)-(43) for SULDAAL (using (41), (44), (43) for GSULDAAL).
    Compute $A^{k,j}$ as in (19).
  end while
  Let $X^k=X^{k,j}$, $G^k=Y^{k,j}$ and $Z^k=Z^{k,j}$.
  Update the Lagrangian multipliers

$$\begin{cases}\Lambda_1^k=\bar{\Lambda}_1^{k-1}+\rho_{k-1}(\mathscr{A}X^k+\mathscr{B}G^k),\\ \Lambda_2^k=\bar{\Lambda}_2^{k-1}+\rho_{k-1}(X^k-Z^k),\end{cases}$$

  where $\bar{\Lambda}_i^k$ is the projection of Λ_i^k on $\{\Lambda_i:\bar{\Lambda}_{i,min}\le\Lambda_i\le\bar{\Lambda}_{i,max}\}_{i=1,2}$.
  Update the penalty parameter

$$\rho^k=\begin{cases}\rho_{k-1}, & \text{if } \|R_i^k\|_\infty\le\tau\|R_i^{k-1}\|_\infty,\ i=1,2,\\ \mu\rho_{k-1}, & \text{otherwise,}\end{cases}$$

  where $R_1^k:=\mathscr{A}X^k+\mathscr{B}G^k$, $R_2^k:=X^k-Z^k$.
end while
return X^k, G^k and Z^k.

Applying Algorithm 3 with $B_1^{k,j-1}=\gamma_1 I$, $B_2^{k,j-1}=\gamma_2 I$ and $B_3^k=\gamma_3 I$, we get the following updating of $(G^{k,j},X^{k,j},Z^{k,j})$ for any fixed $k\in\mathbb{N}$

$$X^{k,j}=X^{k,j-1}-\frac{1}{\gamma_1}\left[\mathscr{A}^T\left(\mathscr{A}X^{k,j-1}+\mathscr{B}G^{k,j-1}+\frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}}\right)+X^{k,j-1}-Z^{k,j-1}+\frac{\bar{\Lambda}_2^{k-1}}{\rho_{k-1}}\right], \tag{41}$$

$$G^{k,j}=\mathbf{shrink}\left(G^{k,j-1}-\frac{1}{\gamma_2}\mathscr{B}^T(\mathscr{A}X^{k,j}+\mathscr{B}G^{k,j-1}+\frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}}),\ \frac{1}{\gamma_2\rho_{k-1}}\right), \tag{42}$$

$$Z^{k,j}=\arg\max_{Z^TZ=I}\left\langle Z,\ X^{k,j}+\gamma_3 Z^{k,j}+\frac{\bar{\Lambda}_2^{k-1}}{\rho_{k-1}}\right\rangle=\tilde{U}\tilde{V}^T, \tag{43}$$

Algorithm 6 (SCCAALN: Algorithm 2 for sparse CCA problem (28))

Input: Data matrix X, Y, parameters $\{\varepsilon_k\}_{k\in\mathbb{N}} \downarrow 0$, $\{\bar{\Lambda}_{i,min} \le \bar{\Lambda}_{i,max}\}_{i=1,2}$, $\tau \in [0,1)$, $\lambda_1 = \lambda_2 = 0.01$, $\rho_0 > 0$, $\mu > 1$, and $\{\gamma_i\}_{1\le i\le 2}$.

Output: (W_x^k, W_y^k, P^k, Q^k).

1: Choose $\{W_x^0, W_y^0, P^0, Q^0\}$ randomly such that $P^0 \in \mathscr{X}$, $Q^0 \in \mathscr{Y}$. Compute $T_1^{k,0} = \Sigma_1^{-1}(U_1^T P^{k,0})$, $(U_1)_\perp Z_1^{k,0} = P^{k,0} - U_1(U_1^T P^{k,0})$, $T_2^{k,0} = \Sigma_2^{-1}(V_1^T Q^{k,0})$, $(V_1)_\perp Z_2^{k,0} = Q^{k,0} - V_1(V_1^T Q^{k,0})$. Let $k = 1$.

2: **while** stopping criterion is not satisfied **do**

3: Let $(W_x^{k,0}, W_y^{k,0}, P^{k,0}, Q^{k,0}) = (W_x^{k-1}, W_y^{k-1}, P^{k-1}, Q^{k-1})$, $j = 1$ and compute $A^{k,j}$.

4: **while** $\|A^{k,j}\|_\infty > \frac{\varepsilon_{k-1}}{\rho_{k-1}}$ **do**

5: Compute $(W_x^{k,j}, W_y^{k,j})$ using (45)-(46) for SCCAAL.

6: Compute $\Delta_1 = U_1^T(W_x^{k,j} - \bar{\Lambda}_1^{k-1}/\rho_{k-1})$ and $\Delta_1' = (W_x^{k,j} - \bar{\Lambda}_1^{k-1}/\rho_{k-1}) - U_1\Delta_1$.
Compute $\Delta_2 = V_1^T(W_y^{k,j} - \bar{\Lambda}_2^{k-1}/\rho_{k-1})$ and $\Delta_2' = (W_y^{k,j} - \bar{\Lambda}_2^{k-1}/\rho_{k-1}) - V_1\Delta_2$.

7: Compute $\Delta_x = \Sigma_1^{-1}\Delta_1' + (\alpha^k I - \Sigma_1^{-2})T_1^{k,j-1}$ and its reduced SVD $\Delta_x = \tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1^T$.
Compute $\Delta_y = \Sigma_2^{-1}\Delta_2' + (\beta^k I - \Sigma_2^{-2})T_2^{k,j-1}$ and its reduced SVD $\Delta_y = \tilde{U}_2\tilde{\Sigma}_2\tilde{V}_2^T$.

8: Compute $T_1^{k,j} = \tilde{U}_1\tilde{V}_1^T$ and $(U_1)_\perp Z_1^{k,j} = \frac{1}{\alpha^k}\Delta_x + (1 - \frac{1}{\alpha^k})((U_1)_\perp Z_1^{k,j})$.
Compute $T_2^{k,j} = \tilde{U}_2\tilde{V}_2^T$ and $(V_1)_\perp Z_2^{k,j} = \frac{1}{\beta^k}\Delta_y + (1 - \frac{1}{\beta^k})((V_1)_\perp Z_2^{k,j})$.

9: Compute $A^{k,j}$ as in (19).

10: **end while**

11: Let $W_x^k = X^{k,j}$, $W_y^k = Y^{k,j}$, $P^k = U_1\Sigma_1^{-1}T_1^{k,j} + (U_1)_\perp Z_1^{k,j}$, and $Q^k = V_1\Sigma_2^{-1}T_2^{k,j} + (V_1)_\perp Z_2^{k,j}$.

12: Update the Lagrangian multiplier

$$\begin{cases} \Lambda_1^k = \bar{\Lambda}_1^{k-1} + \rho_{k-1}(W_x^k - P^k), \\ \Lambda_2^k = \bar{\Lambda}_2^{k-1} + \rho_{k-1}(W_y^k - Q^k), \end{cases}$$

where $\bar{\Lambda}_i^k$ is the projection of Λ_i^k on $\{\Lambda_i : \bar{\Lambda}_{i,min} \le \Lambda \le \bar{\Lambda}_{i,max}\}$, $i = 1,2$.

13: Update the penalty parameter

$$\rho^k = \begin{cases} \rho_{k-1}, & \text{if } \|R_i^k\|_\infty \le \tau\|R_i^{k-1}\|_\infty, i = 1,2, \\ \mu\rho_{k-1}, & \text{otherwise,} \end{cases}$$

where $R_1^k = W_x^k - P^k$, $R_2^k = W_y^k - Q^k$.

14: **end while**

15: **return** W_x^k, W_y^k, P^k, and Q^k.

where $\tilde{U}\tilde{\Sigma}\tilde{V}^T = X^{k,j} + \gamma_3 Z^{k,j} + \bar{\Lambda}_2^{k-1}/\rho_{k-1}$ is the reduced SVD.

Applying Algorithm 2 to problems (25) and (26) are the same except that in the latter case we compute $G^{k,j}$ as follows: Let $\Delta_G^{k,j-1} := G^{k,j-1} - \frac{1}{\gamma_2}\mathscr{B}^T(\mathscr{A}X^{k,j} + \mathscr{B}G^{k,j-1} + \frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}})$, then

$$G_{i,:}^{k,j} = \max\left\{0,\ 1 - 1/\left(\gamma_2\rho_{k-1}\|(\Delta_G^{k,j-1})_{i,:}\|_2\right)\right\}(\Delta_G^{k,j-1})_{i,:}, \tag{44}$$

where subscript $i,:$ means the i-th row.

Summarizing the above procedure leads to sparse ULDA algorithm SULDAAL and group sparse ULDA algorithm SULDAAL outlined in Algorithm 5.

D.3 Sparse CCA

By introducing auxiliary variables $P = W_x$, $Q = W_y$, and denoting $\mathscr{X} = \{P \mid P^T X^T X P = I\}$, $\mathscr{Y} = \{Q \mid Q^T Y^T Y Q = I\}$, we can reformulate (28) as

$$\min_{W_x,\, W_y,\, P,\, Q} \{-tr(W_x^T X^T Y W_y) + \lambda_1 \|W_x\|_1 + \lambda_2 \|W_y\|_1 + \delta_{\mathscr{X}}(P) + \delta_{\mathscr{Y}}(Q) \mid W_x - P = 0,\, W_y - Q = 0\}.$$

The scaled augmented Lagrangian function associated with (28) is

$$\begin{aligned} L_{\rho_{k-1}}(W_x, W_y, P, Q; \bar{\Lambda}^{k-1}) = &\frac{\lambda_1}{\rho_{k-1}} \|W_x\|_1 + \frac{\lambda_2}{\rho_{k-1}} \|W_y\|_1 + \frac{1}{\rho_{k-1}} \delta_{\mathscr{X}}(P) + \frac{1}{\rho_{k-1}} \delta_{\mathscr{Y}}(Q) \\ &+ H_k(W_x, W_y, P, Q), \end{aligned}$$

where

$$\begin{aligned} H_k(W_x, W_y, P, Q) = &-\frac{1}{\rho_{k-1}} tr(W_x^T X^T Y W_y) + \langle \frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}}, W_x - P \rangle + \langle \frac{\bar{\Lambda}_2^{k-1}}{\rho_{k-1}}, W_y - Q \rangle \\ &+ \frac{1}{2} \|W_x - P\|_F^2 + \frac{1}{2} \|W_y - Q\|_F^2. \end{aligned}$$

Applying Algorithm 3 with $B_1^{k,j-1} = \gamma_1 I$, $B_2^{k,j-1} = \gamma_2 I$, $B_3^{k,j} = \alpha^k X^T X - I_{d_1} + \alpha^k (I_{d_1} - U_1 U_1^T)$ and $B_4^{k,j} = \beta^k Y^T Y - I_{d_2} + \beta^k (I_{d_2} - V_1 V_1^T)$, where U_1 and V_1 are obtained from the reduced SVD of X and Y, respectively, as in equation (29), we get the following updating of $(W_x^{k,j}, W_y^{k,j})$ for any fixed $k \in \mathbb{N}$

$$W_x^{k,j} = \mathbf{shrink}\left(W_x^{k,j-1} - \frac{1}{\gamma_1} (-\frac{X^T Y W_y^{k,j-1}}{\rho_{k-1}} + \frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}} + W_x^{k,j-1} - P^{k,j-1}),\ \frac{\lambda_1}{\gamma_1 \rho_{k-1}} \right), \quad (45)$$

$$W_y^{k,j} = \mathbf{shrink}\left(W_y^{k,j-1} - \frac{1}{\gamma_2} (-\frac{Y^T X W_x^{k,j-1}}{\rho_{k-1}} + \frac{\bar{\Lambda}_2^{k-1}}{\rho_{k-1}} + W_y^{k,j-1} - Q^{k,j-1}),\ \frac{\lambda_2}{\gamma_2 \rho_{k-1}} \right). \quad (46)$$

The resulting algorithm SCCAALN is outlined in Algorithm 6.

For problem (31), the associated scaled augmented Lagrangian function is

$$L_{\rho_{k-1}}(W_x, W_y, W; \bar{\Lambda}^{k-1}) = \frac{1}{\rho_{k-1}} \|W_x\|_1 + \frac{\lambda}{\rho_{k-1}} \|W_y\|_1 + \frac{1}{\rho_{k-1}} \delta_{\mathscr{O}}(W) + H_k(W_x, W_y, W),$$

where

$$\begin{aligned} H_k(W_x, W_y, W) = &\langle \frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}}, U_1^T W_x - \Sigma_1^{-1} P_1 W \rangle + \langle \frac{\bar{\Lambda}_2^{k-1}}{\rho_{k-1}}, V_1^T W_y - \Sigma_2^{-1} P_2 W \rangle \\ &+ \frac{1}{2} \|U_1^T W_x - \Sigma_1^{-1} P_1 W\|_F^2 + \frac{1}{2} \|V_1^T W_y - \Sigma_2^{-1} P_2 W\|_F^2. \end{aligned}$$

Applying Algorithm 3 with $B_1^{k,j-1} = \gamma_1 I$, $B_2^{k,j-1} = \gamma_2 I$ and $B_3^{k,j} = \gamma_3 I$, we get the following updating of $(W_x^{k,j}, W_y^{k,j}, W^{k,j})$ for any fixed $k \in \mathbb{N}$

$$W_x^{k,j} = \mathbf{shrink}\left(W_x^{k,j-1} - \frac{1}{\gamma_1} U_1 (\frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}} + U_1^T W_x^{k,j-1} - \Sigma_1^{-1} P_1 W^{k,j-1}),\ \frac{1}{\gamma_1 \rho_{k-1}} \right), \quad (47)$$

$$W_y^{k,j} = \mathbf{shrink}\left(W_y^{k,j-1} - \frac{1}{\gamma_1} V_1 (\frac{\bar{\Lambda}_2^{k-1}}{\rho_{k-1}} + V_1^T W_y^{k,j-1} - \Sigma_2^{-1} P_2 W^{k,j-1}),\ \frac{\lambda}{\gamma_1 \rho_{k-1}} \right), \quad (48)$$

$$W^{k,j} = \arg \max_{W^T W = I} \left\langle W,\ W^{k,j-1} + \frac{1}{\gamma_3} (P_1^T \Sigma_1^{-1} \Delta_x^{k,j} + P_2^T \Sigma_2^{-1} \Delta_y^{k,j}) \right\rangle = \tilde{U} \tilde{V}^T, \quad (49)$$

Algorithm 7 (WSCCAAL: Algorithm 2 for sparse CCA problem (31))

Input: Data matrix Σ_1, U_1, P_1, Σ_2, U_2, P_2, parameters $\{\varepsilon_k\}_{k\in\mathbb{N}}\downarrow 0$, $\{\bar{\Lambda}_{i,min}\le\bar{\Lambda}_{i,max}\}_{i=1,2}$, $\tau\in[0,1)$, $\lambda>0$, $\rho_0>0$, $\mu>1$, and $\{\gamma_i\}_{1\le i\le 2}$.
Output: (W_x^k, W_y^k, W^k).

1: Choose $\{W_x^0, W_y^0, W^0\}$ randomly such that $(W^0)^T(W^0)=I$. Let $k=1$.
2: **while** stopping criterion is not satisfied **do**
3: Let $(W_x^{k,0}, W_y^{k,0}, W^{k,0})=(W_x^{k-1}, W_y^{k-1}, W^{k-1})$, $j=1$ and compute $A^{k,j}$.
4: **while** $\|A^{k,j}\|_\infty > \frac{\varepsilon_{k-1}}{\rho_{k-1}}$ **do**
5: Compute $(W_x^{k,j}, W_y^{k,j}, W^{k,j})$ using (47)-(49) for WSCCAAL.
6: Compute $A^{k,j}$ as in (19).
7: **end while**
8: Let $W_x^k=W_x^{k,j}$, $W_y^k=W_y^{k,j}$ and $W^k=W^{k,j}$.
9: Update the Lagrangian multiplier

$$\begin{cases}\Lambda_1^k=\bar{\Lambda}_1^{k-1}+\rho_{k-1}(U_1^TW_x^k-\Sigma_1^{-1}P_1W^k),\\ \Lambda_2^k=\bar{\Lambda}_2^{k-1}+\rho_{k-1}(V_1^TW_y^k-\Sigma_2^{-1}P_2W^k),\end{cases}$$

where $\bar{\Lambda}_i^k$ is the projection of Λ_i^k on $\{\Lambda_i:\bar{\Lambda}_{i,min}\le\Lambda_i\le\bar{\Lambda}_{i,max}\}$, $i=1,2$.
10: Update the penalty parameter

$$\rho^k=\begin{cases}\rho_{k-1}, & \text{if } \|R_i^k\|_\infty\le\tau\|R_i^{k-1}\|_\infty,\ i=1,2,\\ \mu\rho_{k-1}, & \text{otherwise,}\end{cases}$$

where $R_1^k=U_1^TW_x^k-\Sigma_1^{-1}P_1W^k$, $R_2^k=V_1^TW_y^k-\Sigma_2^{-1}P_2W^k$.
11: **end while**
12: **return** W_x^k, W_y^k, and W^k.

where $\Delta_x^{k,j}=\frac{\bar{\Lambda}_1^{k-1}}{\rho_{k-1}}+U_1^TW_x^{k,j}$, $\Delta_y^{k,j}=\frac{\bar{\Lambda}_2^{k-1}}{\rho_{k-1}}+V_1^TW_y^{k,j}$ and $\tilde{U}\tilde{\Sigma}\tilde{V}^T=W^{k,j-1}+\frac{1}{\gamma_3}(P_1^T\Sigma_1^{-1}\Delta_x^{k,j}+P_2^T\Sigma_2^{-1}\Delta_y^{k,j})$ is the reduced SVD. The resulting algorithm WSCCAAL is outlined in Algorithm 7.

References

1. Abrudan, T., Eriksson, J., Koivunen, V.: Steepest descent algorithms for optimization under unitary matrix constraint. Signal Processing, IEEE Transactions on **56**(3), 1134–1147 (2008)
2. Abrudan, T., Eriksson, J., Koivunen, V.: Conjugate gradient algorithm for optimization under unitary matrix constraint. Signal Processing **89**(9), 1704–1714 (2009)
3. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press (2009)
4. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: On augmented lagrangian methods with general lower-level constraints. SIAM Journal on Optimization **18**(4), 1286–1309 (2007)
5. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. Mathematics of Operations Research **35**(2), 438–457 (2010)
6. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. Mathematical Programming **137**(1-2), 91–129 (2013)
7. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic press (1982)
8. Bertsekas, D.P.: Nonlinear programming. Athena scientific (1999)
9. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming **146**(1-2), 459–494 (2014)
10. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition **30**(7), 1145–1159 (1997)

11. Chen, W.: Wavelet frames on the sphere, high angular resolution diffusion imagining and l_1-regularized optimization on stiefel manifolds. Ph.D. thesis, The National University of Singapore (2015)
12. Chen, W., Ji, H., You, Y.: An augmented lagrangian method for ℓ_1-regularized optimization problems with orthogonality constraints. SIAM journal on Scientific Computing **38**(4), B570–B592 (2016)
13. Chu, D., Liao, L.Z., Ng, M.K., Zhang, X.: Sparse canonical correlation analysis: new formulation and algorithm. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(12), 3050–3065 (2013)
14. Chu, M.T., Trendafilov, N.T.: The orthogonally constrained regression revisited. Journal of Computational and Graphical Statistics **10**(4), 746–771 (2001)
15. Clarke, F.H., Ledyaev, Y.S., Stern, R.J., Wolenski, P.R.: Nonsmooth analysis and control theory, vol. 178. Springer Science & Business Media (2008)
16. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics **53**(4) (2011)
17. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley (2000)
18. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM journal on Matrix Analysis and Applications **20**(2), 303–353 (1998)
19. Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)
20. Francisco, J.B., Martínez, J.M., Martínez, L., Pisnitchenko, F.: Inexact restoration method for minimization problems arising in electronic structure calculations. Computational Optimization and Applications **50**(3), 555–590 (2011)
21. Grubišić, I., Pietersz, R.: Efficient rank reduction of correlation matrices. Linear algebra and its applications **422**(2), 629–653 (2007)
22. Hardoon, D.R., Shawe-Taylor, J.: Sparse canonical correlation analysis. Machine Learning **83**(3), 331–353 (2011)
23. Hestenes, M.R.: Multiplier and gradient methods. Journal of optimization theory and applications **4**(5), 303–320 (1969)
24. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
25. Howland, P., Jeon, M., Park, H.: Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. SIAM Journal on Matrix Analysis and Applications **25**, 165–179 (2003)
26. Jiang, B., Dai, Y.H.: A framework of constraint preserving update schemes for optimization on stiefel manifold. Mathematical Programming **153**(2), 535–575 (2015)
27. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit, vol. 5, pp. 79–86. Citeseer (2005)
28. Kokiopoulou, E., Chen, J., Saad, Y.: Trace optimization and eigenproblems in dimension reduction methods. Numerical Linear Algebra with Applications **18**(3), 565–602 (2011)
29. Kovnatsky, A., Glashoff, K., Bronstein, M.M.: Madmm: a generic algorithm for non-smooth optimization on manifolds. arXiv preprint arXiv:1505.07676 (2015)
30. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l'institut Fourier **48**(3), 769–783 (1998)
31. Lai, R., Osher, S.: A splitting method for orthogonality constrained problems. Journal of Scientific Computing **58**(2), 431–449 (2014)
32. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM Journal on Optimization **25**(4), 2434–2460 (2015)
33. Lojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. Les équations aux dérivées partielles pp. 87–89 (1963)
34. Lu, Z., Zhang, Y.: An augmented lagrangian approach for sparse principal component analysis. Mathematical Programming **135**(1-2), 149–193 (2012)
35. Merchante, L., Grandvalet, Y., Govaert, G.: An efficient approach to sparse linear discriminant analysis. In: Preceedings of the 29th International Conference on Machine Learning (2012)
36. Mordukhovich, B.S.: Variational analysis and generalized differentiation I: Basic theory, vol. 330. Springer Science & Business Media (2006)
37. Mordukhovich, B.S., Shao, Y.: On nonconvex subdifferential calculus in banach spaces. Journal of Convex Analysis **2**(1/2), 211–227 (1995)
38. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France **93**, 273–299 (1965)
39. Nishimori, Y., Akaho, S.: Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. Neurocomputing **67**, 106–135 (2005)
40. Ozoliņš, V., Lai, R., Caflisch, R., Osher, S.: Compressed modes for variational problems in mathematics and physics. Proceedings of the National Academy of Sciences **110**(46), 18,368–18,373 (2013)
41. Powell, M.J.: A method for non-linear constraints in minimization problems. UKAEA (1967)

42. Rockafellar, R.T.: Augmented lagrange multiplier functions and duality in nonconvex programming. SIAM Journal on Control **12**(2), 268–285 (1974)
43. Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2009)
44. Savas, B., Lim, L.H.: Quasi-newton methods on grassmannians and multilinear approximations of tensors. SIAM Journal on Scientific Computing **32**(6), 3352–3393 (2010)
45. Sriperumbudur, B.K., Torres, D.A., Lanckriet, G.R.: A majorization-minimization approach to the sparse generalized eigenvalue problem. Machine learning **85**(1-2), 3–39 (2011)
46. Vinokourov, A., Cristianini, N., Shawe-Taylor, J.S.: Inferring a semantic representation of text via cross-language correlation analysis. In: Advances in neural information processing systems, pp. 1473–1480 (2002)
47. Voorhees, E.M.: The sixth text retrieval conference (trec-6). Information Processing & Management **36**(1), 1–2 (2000)
48. Wang, Y., Yin, W., Zeng, J.: Global convergence of admm in nonconvex nonsmooth optimization. arXiv preprint arXiv:1511.06324 (2015)
49. Wen, Z., Yang, C., Liu, X., Zhang, Y.: Trace-penalty minimization for large-scale eigenspace computation. Journal of Scientific Computing **66**, 1175–1203 (2016)
50. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Mathematical Programming **142**(1-2), 397–434 (2013)
51. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3), 515–534 (2009)
52. Yang, C., Meza, J.C., Wang, L.W.: A trust region direct constrained minimization algorithm for the kohn-sham equation. SIAM Journal on Scientific Computing **29**(5), 1854–1875 (2007)
53. Yang, K., Cai, Z., Li, J., Lin, G.: A stable gene selection in microarray data analysis. BMC bioinformatics **7**(1), 228 (2006)
54. Ye, J.: Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. Journal of Machine Learning Research **6**(4), 483–502 (2005)
55. Zhang, L., Li, R.: Maximization of the sum of the trace ratio on the stiefel manifold, i: Theory. Science China Mathematics **57**(12), 2495–2508 (2014)
56. Zhang, L., Li, R.: Maximization of the sum of the trace ratio on the stiefel manifold, ii: Computation. Science China Mathematics **58**(7), 1549–1566 (2015)
57. Zhang, X.: Sparse dimensionality reduction methods: Algorithms and applications. Ph.D. thesis, The National University of Singapore (2013)
58. Zhang, X., Chu, D.: Sparse uncorrelated linear discriminant analysis. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13), pp. 45–52 (2013)
59. Zhang, X., Chu, D., Tan, R.C.: Sparse uncorrelated linear discriminant analysis for undersampled problems. Neural Networks and Learning Systems, IEEE Transactions on **27**(7), 1469–1485 (2015)
60. Zhao, Y., Karypis, G.: Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning **55**(3), 311–331 (2004)