A NONCONVEX VARIATIONAL APPROACH FOR ROBUST GRAPHICAL LASSO

Alessandro Benfenati[1]*,Emilie Chouzenoux*[1,2]*, Jean–Christophe Pesquet*[2]

[1]LIGM, UMR CNRS 8049, Université Paris Est Marne la Vallée, Champs-sur-Marne, France
[2]CVN, INRIA Saclay, CentraleSupélec, Université Paris Saclay, Gif-sur-Yvette, France

ABSTRACT

In recent years, there has been a growing interest in problems in graph estimation and model selection, which all share very similar matrix variational formulations, the most popular one being probably GLASSO. Unfortunately, the standard GLASSO formulation does not take into account noise corrupting the data: this shortcoming leads us to propose a novel criterion, where the regularization function is decoupled in two terms, one acting only on the eigenvalues of the matrix and the other on the matrix elements. Incorporating noise information into the model has the side–effect to make the cost function non–convex. To overcome this difficulty, we adopt a *majorization–minimization* approach, where at each iteration a convex approximation of the original cost function is minimized via the Douglas–Rachford procedure. The achieved results are very promising w.r.t. classical approaches.

Index Terms— Majorization–minimization, Graphical LASSO, Non–convex Optimization, Covariance Estimation, Proximal Methods.

1. INTRODUCTION

In past years, various applied areas such as shape classification [1], gene expression [2], model selection [3, 4], computer vision [5], inverse covariance estimation [6, 7, 8, 9, 10], graph estimation [11, 12, 13], social network and corporate inter-relationships analysis [14], or brain network analysis [15] have led to solving matrix optimization problems. A very popular and useful example of such problems is the graphical lasso approach, where the underlying cost function $\mathcal{G}$ reads as the sum of (*i*) a minus log-determinant function, (*ii*) the component–wise ℓ_1 norm (of the matrix entries) and (*iii*) a linear trace term, i.e.

$$\mathcal{G}(\mathbf{C}) = -\log\det(\mathbf{C}) + \mathrm{tr}(\mathbf{CS}) + \mu\|\mathbf{C}\|_1, \quad \mu \in \mathbb{R}^+, \tag{1}$$

where variable $\mathbf{C}$ is a symmetric positive–definite matrix and $\mathbf{S}$ is some given matrix. Various algorithms have been proposed to solve this problem, including the original GLASSO algorithm [6] and some of its recent variants [16]. We can also mention the dual block coordinate ascent method from [3], the SPICE algorithm [17], the gradient projection method in [1], the Refitted CLIME algorithm [18], various algorithms [9, 19, 20] based on Nesterov's smooth gradient approach [21], ADMM approaches [8, 22, 23], an inexact Newton method [10], and interior point methods [13, 24]. One of the main weaknesses of the graphical lasso model however is that it does not take into account noise perturbing the data.

In this paper, we propose a new variational approach which incorporates two key ingredients. First, we modify the data fidelity term so as to take into account the information about the noise. In addition, we consider a more versatile regularization form consisting of the sum of two different terms, one being a *symmetric spectral* function while the other acts on the whole matrix entries. Even if a convex regularization is chosen, our formulation leads to a non convex objective function. To tackle this problem, we propose to adopt a majorization–minimization approach combined with a Douglas–Rachford inner solver. We provide also a convergence result for the proposed algorithm.

The present work is organized as follows: in Section 2 we present the observation model, the proposed variational formulation, and we show that it corresponds to a non convex objective function; we describe our algorithm and discuss its convergence properties as well as some theoretical results allowing the use of the Douglas–Rachford algorithm. In Section 3 we describe the numerical experiments used to validate our approach, including comparisons with a state-of-the-art graphical lasso algorithm. Section 4 draws some conclusions on the presented work.

Notation. Bold letters and bold capital letters refer to vectors and matrices, respectively. $\mathcal{S}_n, \mathcal{S}_n^+, \mathcal{S}_n^{++}$ are the set of symmetric, symmetric positive semi–definite and symmetric positive–definite matrices. $\Gamma_0(\mathcal{H})$ designates the set of convex, proper and lower–semicontinuous functions on a Hilbert space $\mathcal{H}$.

2. PROPOSED MAJORIZE–MINIMIZE APPROACH

2.1. Problem Statement

Let us consider the following signal model [25]:

$$(\forall i \in \{1, \ldots, N\}) \quad \mathbf{x}^{(i)} = \mathbf{A}\mathbf{s}^{(i)} + \mathbf{e}^{(i)} \tag{2}$$

This work was supported by the Agence Nationale de la Recherche under grant ANR-14-CE27-0001 GRAPHSIP.

where $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $m \leq n$ and, for every $i \in \{1,\dots,N\}$, $\mathbf{s}^{(i)} \in \mathbb{R}^m$ and $\mathbf{e}^{(i)} \in \mathbb{R}^n$ are realizations of mutually independent identically distributed Gaussian multivalued random variables with zero mean and covariance matrices $\mathbf{E} \in \mathcal{S}_m^{++}$ and $\sigma^2\mathrm{I_d}$, $\sigma > 0$, respectively. Hence, the covariance matrix $\mathbf{\Sigma}$ of the observed signal in (2) is

$$\mathbf{\Sigma} = \mathbf{A}^\top \mathbf{E}\mathbf{A} + \sigma^2\mathrm{I_d}. \tag{3}$$

Such observation model is encountered in several practical applications, e.g. in the context of "Relevant Vector Machine" [26, 27]. The empirical covariance matrix of the $\mathbf{x}^{(i)}$s in (2) is

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top, \tag{4}$$

which can be viewed as a rough approximation of (3), especially when N is small. Assuming that the noise level σ is known, we propose here to estimate more tightly the precision matrix, i.e., the inverse of the covariance matrix (3) from the input data $(\mathbf{x}^{(i)})_{1\leq i\leq N}$, by a penalized maximum likelihood approach leading to the following minimization problem:

$$\underset{\mathbf{C}\in\mathcal{S}_n^{++}}{\text{minimize}} \ \big(\mathcal{F}(\mathbf{C}) \triangleq f(\mathbf{C}) + \mathcal{T}_{\mathbf{S}}(\mathbf{C}) + g_0(\mathbf{C}) + g_1(\mathbf{C})\big) \tag{5}$$

where

$$(\forall \mathbf{C} \in \mathcal{S}_n^{++}) \quad f(\mathbf{C}) \triangleq \log\det\left(\mathbf{C}^{-1} + \sigma^2\mathrm{I_d}\right) \tag{6}$$

$$(\forall \mathbf{C} \in \mathcal{S}_n^{+}) \quad \mathcal{T}_{\mathbf{S}}(\mathbf{C}) \triangleq \mathrm{tr}\left(\left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}\mathbf{C}\mathbf{S}\right), \tag{7}$$

the first (resp. second) function value being taken equal to $+\infty$ outside $\mathcal{S}_n^{++}$ (resp. $\mathcal{S}_n^{+}$). Function (6) mimics the role of $-\log\det$ in (1), including the information about the noise, while (7) directly depends on the trace operator in (1) in which $\mathbf{C}$ is substituted by $\left(\mathbf{C}^{-1} + \sigma^2\mathrm{I_d}\right)^{-1} = \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}\mathbf{C}$. A hybrid regularization term is considered, which consists of $g_0 + g_1$. Function g_0 is a *spectral symmetric* function, i.e. $(\forall \mathbf{C} \in \mathcal{S}_n) \quad g_0(\mathbf{C}) = \psi(\mathbf{P}\mathbf{d})$, where $\mathbf{P}$ is any permutation matrix, $\mathbf{d} = [d_1,\dots,d_n]^\top$ is the vector of eigenvalues of $\mathbf{C}$, and $\psi \in \Gamma_0(\mathbb{R}^n)$. Moreover, $g_1 \in \Gamma_0(\mathbb{R}^{n\times n})$ acts on the whole matrix $\mathbf{C}$. It is worthwile noticing that, when $\sigma = 0, g_0 = 0$ and $g_1 = \|\cdot\|_1$, (5) becomes equivalent to the optimization problem arising in the famous GLASSO approach [6]. Our formulation presents the advantage of accounting for a nonzero level of noise, and various choices of regularization terms.

2.2. Minimization algorithm

The subsequent lemma reveals that the functional $\mathcal{F}$ in (5) is a difference of concave functions, which will be a key property in the optimization approach we propose.

Lemma 1. *Consider* (5) *with* g_0 *and* g_1 *in* $\Gamma_0(\mathbb{R}^{n\times n})$.

(i) $f + g_0 + g_1$ *is a convex function.*

(ii) The trace term $\mathcal{T}_{\mathbf{S}}$ *is concave on* $\mathcal{S}_n^+$.

Sketch of proof. (i) is straightforwardly proved. (ii) is non trivial. Using matrix differential calculus [28], we were able to prove that the opposite of the Hessian of the trace term (7) is a positive semi–definite linear operator on $\mathcal{S}_n$. □

According to Lemma 1, the cost function $\mathcal{F}$ is not convex, thus we propose to adopt a Majorize-Minimize (MM) strategy [29, 25, 30, 31] to solve (5). At each iteration $\ell \in \mathbb{N}$, we upper bound $\mathcal{F}$ by a convex tangent approximation of it, considering simply a linear approximation of the concave term. Then, the next iterate is obtained by minimizing the approximate functional, which yields the following scheme:

$$\mathbf{C}^{(\ell+1)} = \underset{\mathbf{C}\in\mathcal{S}_n^{++}}{\mathrm{argmin}} \ \Big[f(\mathbf{C}) + \mathrm{tr}\left(\nabla\mathcal{T}_{\mathbf{S}}\big(\mathbf{C}^{(\ell)}\big)\mathbf{C}\right) + g_0(\mathbf{C}) + g_1(\mathbf{C})\Big]. \tag{8}$$

In (8), $\nabla\mathcal{T}_{\mathbf{S}}\left(\mathbf{C}^{(\ell)}\right)$ is the gradient of $\mathcal{T}_{\mathbf{S}}$ at $\mathbf{C}^{(\ell)}$, which is given by $\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}^{(\ell)}) = \left(\mathrm{I_d} + \sigma^2\mathbf{C}^{(\ell)}\right)^{-1}\mathbf{S}\left(\mathrm{I_d} + \sigma^2\mathbf{C}^{(\ell)}\right)^{-1}$. When $g_1 \neq 0$, the subproblem arising at each iteration usually has no explicit solution. Nonetheless, these subproblems are convex. We thus propose to employ the Douglas–Rachford (DR) proximal algorithm ([32, 33, 34]) to solve each inner subproblem. In the framework of (8), the DR method minimizes the sum of $f + \mathrm{tr}\left(\nabla\mathcal{T}_{\mathbf{S}}\left(\mathbf{C}^{(\ell)}\right)(\cdot)\right) + g_0$ and g_1 by alternately computing proximity operators of each of these two functions. Let us recall the definition of the proximity operator of a function $h \in \Gamma_0(\mathbb{R}^{n\times n})$: let $\gamma > 0$, the proximity operator of γh at $\overline{\mathbf{C}}$ will be denoted by $\mathrm{prox}_{\gamma h}\left(\overline{\mathbf{C}}\right)$ and corresponds to the unique minimizer of $\gamma h + \|\cdot - \overline{\mathbf{C}}\|_{\mathrm{F}}^2/2$, where $\|\cdot\|_{\mathrm{F}}$ is the Froebenius norm.
Most practical choices for g_1 allow its proximity operator to be calculated via an explicit formula, while the subsequent lemma reveals a practical way to compute the proximity operator of the other term appearing in (8).

Lemma 2. *Let* $\gamma \in]0,+\infty[$ *and* $\overline{\mathbf{C}} \in \mathcal{S}_n$. *For every* $\mathbf{d} = [d_1,\dots,d_n]^\top \in \mathbb{R}^n$, *let* $\varphi(\mathbf{d}) = \sum_{j=1}^n \log\left((1+\sigma^2 d_j)/d_j\right)$ *if* $\mathbf{d} \in]0,+\infty[^n$ *and* $+\infty$ *otherwise. Let* $g_0 \in \Gamma_0(\mathbb{R}^{n\times n})$ *be the symmetric spectral function associated with* $\psi \in \Gamma_0(\mathbb{R}^n)$ *such that* $\mathrm{dom}\,\varphi \cap \mathrm{dom}\,\psi \neq \varnothing$. *Let* $\boldsymbol{\lambda} \in \mathbb{R}^n$ *and* $\mathbf{U}$ *be an orthogonal matrix such that* $\overline{\mathbf{C}} - \gamma\nabla\mathcal{T}_{\mathbf{S}}\left(\mathbf{C}^{(\ell)}\right) = \mathbf{U}\,\mathrm{Diag}(\boldsymbol{\lambda})\mathbf{U}^\top$. *Then*

$$\mathrm{prox}_{\gamma\left(f+\mathrm{tr}\left(\nabla\mathcal{T}_{\mathbf{S}}\left(\mathbf{C}^{(\ell)}\right)\cdot\right)+g_0\right)}\left(\overline{\mathbf{C}}\right) = \mathbf{U}\,\mathrm{Diag}\left(\mathrm{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})\right)\mathbf{U}^\top.$$

The complete proposed procedure is described in Algorithm 1, whose convergence is guaranteed by the next theorem. The rather technical proof is omitted due to the lack of space.

Algorithm 1 MMDR: Majorization–Minimization algorithm with Douglas–Rachford inner steps

Set $\gamma > 0$, $\mathbf{C}^{(0)} \in \mathcal{S}_n$, let $\mathbf{S} \in \mathcal{S}_n^+$ be the given data
for $l = 0, 1, \ldots$ **do**
  Set $\mathbf{C}^{(\ell,0)} = \mathbf{C}^{(\ell)}$
  for $k = 0, 1, \ldots$ **do**
    Compute $\mathbf{U}^{(k)}, \mathbf{\Lambda}^{(k)} = \mathrm{Diag}(\boldsymbol{\lambda}^{(k)})$ such that

$$\mathbf{C}^{(\ell,k)} + \gamma \nabla \mathcal{T}_{\mathbf{S}}(\mathbf{C}^{(\ell)}) = \mathbf{U}^{(k)} \mathbf{\Lambda}^{(k)} \left(\mathbf{U}^{(k)}\right)^{\top}$$

    $\mathbf{d}^{\left(\ell,k+\frac{1}{2}\right)} = \mathrm{prox}_{\gamma(\varphi+\psi)}\left(\boldsymbol{\lambda}^{(k)}\right)$
    $\mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)} = \mathbf{U}^{(k)}\,\mathrm{Diag}\left(\mathbf{d}^{\left(\ell,k+\frac{1}{2}\right)}\right)\left(\mathbf{U}^{(k)}\right)^{\top}$
    Choose $\alpha_k \in [0, 2)$
    $\mathbf{Y}^{(\ell,k)} = \mathrm{prox}_{\gamma g_1}\left(2\mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)} - \mathbf{C}^{(\ell,k)}\right)$
    $\mathbf{C}^{(\ell,k+1)} = \mathbf{C}^{(\ell,k)} + \alpha_k\left(\mathbf{Y}^{(\ell,k)} - \mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)}\right)$
  end for
  $\mathbf{C}^{(\ell+1)} = \mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)}$
end for

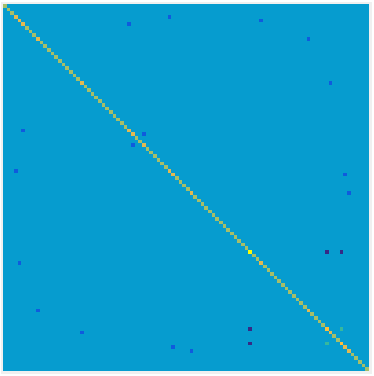
Theorem 1. *Let* $(\mathbf{C}^{(\ell)})_{\ell\geq 0}$ *be a sequence generated by* (8). *Assume that* $\mathrm{dom}\, f \cap \mathrm{dom}\, g_0 \cap \mathrm{dom}\, g_1 \neq \varnothing$, $f + g_0 + g_1$ *is coercive, and* $\{\mathbf{C} \in \mathcal{S}_n \mid \mathcal{F}(\mathbf{C}) \leq \mathcal{F}(\mathbf{C}^{(0)})\}$ *is included in the relative interior of* $\mathrm{dom}\, g_0 \cap \mathrm{dom}\, g_1$. *Then, the following properties hold:*

1. $(\mathcal{F}(\mathbf{C}^{(\ell)}))_{\ell\geq 0}$ *is a decaying sequence converging to* $\widehat{\mathcal{F}} \in \mathbb{R}$.
2. $(\mathbf{C}^{(\ell)})_{\ell\geq 0}$ *has a cluster point.*
3. *Every cluster point* $\widehat{\mathbf{C}}$ *of* $(\mathbf{C}^{(\ell)})_{\ell\geq 0}$ *is such that* $\mathcal{F}(\widehat{\mathbf{C}}) = \widehat{\mathcal{F}}$ *and it is a critical point of* $\mathcal{F}$, *i.e.* $-\nabla f(\widehat{\mathbf{C}}) - \nabla \mathcal{T}_{\mathbf{S}}(\widehat{\mathbf{C}}) \in \partial(g_0 + g_1)(\widehat{\mathbf{C}})$.

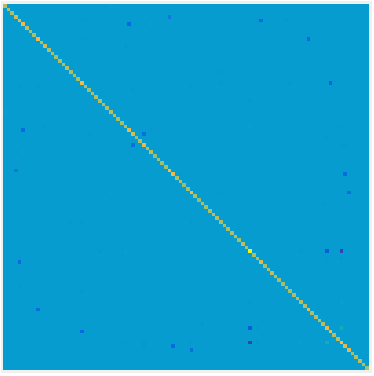
This theorem is in the spirit of asymptotic results in [35, 36]. However, unlike existing results, the differentiability of $g_0 + g_1$ is not required, which is of main importance in sparse matrix estimation problems.

3. NUMERICAL EXPERIMENTS

Let us now evaluate the performance of the proposed approach. We test our MMDR Algorithm on datasets generated with the code available on Boyd's webpage[1]: a *sparse* precision matrix $\mathbf{C}_0$ of dimension $n \times n$ is randomly created, then its inverse $\mathbf{\Sigma}_0$ is used to generate N realizations $(\mathbf{x}^{(i)})_{1\leq i\leq N}$ of a Gaussian random variable with zero mean and covariance $\mathbf{\Sigma}_0$. Gaussian noise with variance σ^2 is added to these realizations, hence the final covariance matrix $\mathbf{\Sigma}$ fulfills exactly (3), with $\mathbf{\Sigma}_0 = \mathbf{A}^{\top}\mathbf{E}\mathbf{A}$.
The settings of our experiments are $n = 100$, $N = 10000$, the number of nonzero elements in $\mathbf{C}_0$ is 10; $g_1 = \mu_1\|\cdot\|_1$, $\mu_1 > 0$ and $g_0(\mathbf{C}) = \mu_0\mathcal{R}_1\left(\mathbf{C}^{-1}\right)$, where $\mathcal{R}_1$ is the Schatten–1–norm (also called nuclear norm) and $\mu_0 > 0$. The parameters μ_0 and μ_1 allow us to adjust the incidence of the regularization functionals on the final reconstruction. The initial value $\mathbf{C}^{(0)}$ is set to $\mathbf{S} + \mathrm{I_d}$. Since the problem is not convex, the algorithm result may be sensitive to initialization. Our numerical experiments showed that the choice of the initial value affects mainly the convergence speed (in terms of total iteration number) and in a negligible way, the quality of the reconstruction. In Figure 1, the results obtained in the presence of a high noise level $\sigma = 0.5$ by setting $\mu_0 = 0.07, \mu_1 = 0.03, \gamma = 1$ are depicted. The outer cycle is stopped as soon as the relative difference of the objective function $\mathcal{F}$ reaches a tolerance of 10^{-8}, while the inner DR iterations are stopped with the same criterion on the majorant function and with a tolerance of 10^{-10}. The parameters were set so as to obtain the best relative mean square error `rmse` on the covariance matrix, i.e. $\texttt{rmse} = \|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_{\mathrm{rec}}\|_{\mathrm{F}}/\|\mathbf{\Sigma}_0\|_{\mathrm{F}}$, where $\mathbf{\Sigma}_{\mathrm{rec}}$ is the reconstructed covariance matrix. Panels



(a) $\mathbf{\Sigma}_0$



(b) $\mathbf{\Sigma}_{\mathrm{rec}}$

Fig. 1. Results for $\sigma = 0.5$. Panel (a) and (b) contain the visual inspection of the true covariance matrix and of the reconstructed one, respectively. The scale–color used is the same for both images.

(a-b) in Figure 1 are useful to visually compare the true co-

[1] `http://stanford.edu/~boyd/papers/admm/covsel/covsel_example.html`

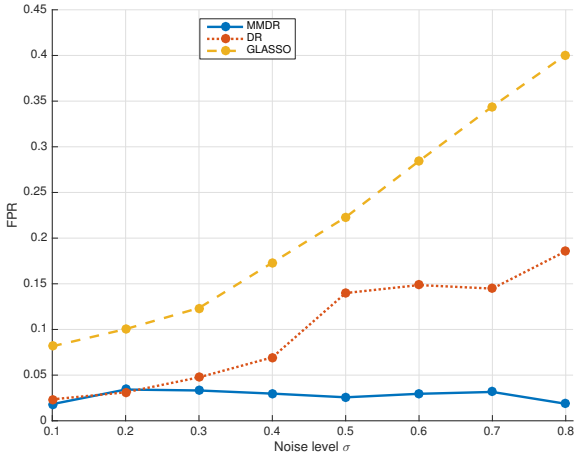


Fig. 2. Behaviour of the FPR w.r.t. noise level. Comparison for three methods: Dotted line refers to DR implementation, dashed one to GLASSO algorithm, and finally solid line to MMDR approach. The latter one allows to achieve the best estimation of the support of the precision matrix for increasing noise level.

variance matrix $\mathbf{\Sigma}_0$ and the recovered one $\mathbf{\Sigma}_{\text{rec}}$: the result appears satisfactory, and the `rmse` achieved is 0.1113, while the initial `rmse` of the empirical covariance matrix is 0.6276. We now compare the performance of MMDR with a GLASSO implementation based on an the Alternating Direction of Multipliers Method (ADMM) described in [23]. In the latter case, the information about the noise is not taken into account in the model. We generate different datasets for various noise levels ($\sigma \in \{0.1, 0.3, 0.5, 0.7\}$). Moreover, we evaluate also a simpler variational model than (5), in which we keep the splitting of the regularization functional into $g_0 + g_1$ but no information about noise is incorporated into the model, i.e. σ is set to 0 in (5): applying the DR algorithm to the resulting problem provides the solution. We are performing this further comparison to check whether including information about noise plays a critical role in terms of matrix recovery. The error measurement employed is the False Positive Rate (FPR), i.e. the percentage of nonzero entries erroneously included in the support of the recovered matrix. The True Positive Rate, i.e. the percentage of the nonzero entries correctly recognized, is 100% for each method in this example. Figure 2 compares the FPR provided by each algorithm (the non monotonic behaviour is due to the fact that the regularization parameters are set to minimize the relative mean square error in the covariance matrix estimation). The MMDR approach obviously improves the quality of the support estimation: it outperforms GLASSO which leads to poorer results as the noise corrupting the data becomes stronger. Moreover, it is clear that including the noise information into the variational model allows a more accurate reconstruction to be achieved. The improvement in terms of covariance matrix estimation is especially significant for large values of σ.

The computational cost of GLASSO and DR algorithms is mainly governed by the SVD decomposition performed at each iteration. In MMDR, the time per iteration may be higher, since it requires also the computation of the gradient of $\mathcal{T}_{\mathbf{S}}$, which consists in matrix multiplications. However, it should be pointed out that the higher complexity of the proposed model is worthy, as it brings the possibility to deal with noisy data, which is of paramount importance in practice.

4. CONCLUSIONS

In this work, we have proposed an extension of the celebrated GLASSO formulation. Our approach is designed to be robust to the presence of noise. In addition, it allows a wide range of regularization functions to be employed. An efficient MM algorithm grounded on the use of the Douglas-Rachford method has been proposed to solve the associated non convex optimization problem and its convergence properties have been investigated. The effectiveness of our method has been shown on a standard dataset.

5. REFERENCES

[1] J.C. Duchi, S. Gould, and D. Koller, "Projected Subgradient Methods for Learning Sparse Gaussians," in *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12*, 2008, pp. 145–152.

[2] S. Ma, L. Xue, and H. Zou, "Alternating direction methods for latent variable Gaussian graphical model selection," *Neural Comput.*, vol. 25, no. 8, pp. 2172–2198, Aug. 2013.

[3] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, 2008.

[4] V. Chandrasekaran, P.A. Parrilo, and A.S. Willsky, "Latent variable graphical model selection via convex optimization," *Ann. Statist.*, vol. 40, no. 4, pp. 1935–1967, 08 2012.

[5] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," *Biometrika*, vol. 98, no. 1, pp. 1, 2011.

[6] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, jul 2008.

[7] A. Dempster, "Covariance selection," *Biometrics*, vol. 28, pp. 157–175, 1972.

[8] X. Yuan, "Alternating direction method for covariance selection models," *J. Sc. Comp.*, vol. 51, no. 2, pp. 261–273, May 2012.

[9] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, "First-order methods for sparse covariance selection," *SIAM J. Matrix Anal. A.*, vol. 30, no. 1, pp. 56–66, 2008.

[10] C. Wang, D. Sun, and K.-C. Toh, "Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm," *SIAM J. Opt.*, vol. 20, no. 6, pp. 2994–3013, 2010.

[11] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, Jun 2006.

[12] P. Ravikumar, M.J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011.

[13] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19, 2007.

[14] M.S. Aslan, , X.-W. Chen, and H. Cheng, "Analyzing and learning sparse and scale-free networks using Gaussian graphical models," *Int. J. Data Science Anal.*, vol. 1, no. 2, pp. 99–109, 2016.

[15] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye, "Fused multiple graphical lasso," *SIAM J. Opt.*, vol. 25, no. 2, pp. 916–943, 2015.

[16] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electron. J. Statist.*, vol. 6, pp. 2125–2149, 2012.

[17] A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, vol. 2, pp. 494–515, 2008.

[18] T. Cai, W. Liu, and X. Luo, "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *J. Amer. Stat. Ass.*, vol. 106, no. 494, pp. 594–607, 2011.

[19] Z. Lu, "Smooth optimization approach for sparse covariance selection," *SIAM J. Opt.*, vol. 19, no. 4, pp. 1807–1827, 2009.

[20] Z. Lu, "Adaptive first-order methods for general sparse inverse covariance selection," *SIAM J. Matrix Anal. A.*, vol. 31, no. 4, pp. 2000–2016, 2010.

[21] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Progr.*, vol. 103, no. 1, pp. 127–152, 2005.

[22] K. Scheinberg, S. Ma, and D. Goldfarb, "Sparse inverse covariance selection via alternating linearization methods," in *Adv. Neural Inf. Proc. Sys. 23*, pp. 2101–2109. 2010.

[23] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[24] L. Li and K.C. Toh, "An inexact interior point method for ℓ_1–regularized sparse covariance selection," *Math. Progr. Comp.*, vol. 2, no. 3, pp. 291–315, 2010.

[25] Y. Sun, P. Babu, and D.P. Palomar, "Majorization-Minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Sig. Proc.*, vol. 65, no. 3, pp. 794–816, 2017.

[26] M.E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sept. 2001.

[27] D.P. Wipf and B.D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Sig. Proc.*, vol. 52, no. 8, pp. 2153–2164, Aug 2004.

[28] J.R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, second edition, 1999.

[29] E. Chouzenoux and J.-C. Pesquet, "Convergence Rate Analysis of the Majorize-Minimize Subspace Algorithm," *IEEE Sig. Proc. Letters*, vol. 23, no. 9, pp. 1284 – 1288, 2016.

[30] D.R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.

[31] M.W. Jacobson and J.A. Fessler, "An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms," *IEEE Trans. on Im. Proc.*, vol. 16, no. 10, pp. 2411–2422, Oct 2007.

[32] P.L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[33] P.L. Combettes and J.-P. Pesquet, "Proximal Splitting Methods in Signal Processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, 2011.

[34] P.L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Sel. Top. Sig. Proc.*, vol. 1, no. 4, pp. 564–574, Dec 2007.

[35] C.F.J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 03 1983.

[36] W.I. Zangwill, *Nonlinear programming: a unified approach*, Prentice-Hall, 1969.