A Nonconvex Variational Approach for Robust Graphical Lasso.

Alessandro Benfenati, Emilie Chouzenoux, Jean-Christophe Pesquet

To cite this version:

HAL Id: hal-01803424
https://hal.archives-ouvertes.fr/hal-01803424
Submitted on 30 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A NONCONVEX VARIATIONAL APPROACH FOR ROBUST GRAPHICAL LASSO

Alessandro Benfenati¹, Emilie Chouzenoux¹², Jean–Christophe Pesquet²

¹LIGM, UMR CNRS 8049, Université Paris Est Marne la Vallée, Champs-sur-Marne, France
²CVN, 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 cost function, 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 algorithm, where the underlying cost functions reads as the sum of (i) a minus log–determinant function, (ii) the component–wise ℓ₁ norm (of the matrix entries) and (iii) a linear trace term. 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 Re–

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

2. PROPOSED MAJORIZE–MINIMIZE APPROACH

2.1. Problem Statement

Let us consider the following signal model [25]:

\[ (\forall i \in \{1, \ldots, N\} \quad x^{(i)} = A s^{(i)} + e^{(i)} \quad (1) \]

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

$$\Sigma = A^T E A + \sigma^2 I_d.$$  \hspace{1cm} (2)\hfill

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 $x^{(i)}$'s in (1) is

$$S = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}(x^{(i)})^\top,$$  \hspace{1cm} (3)\hfill

which can be viewed as a rough approximation of (2), especially when $N$ is small. Assuming that the noise level $\sigma$ is known, we propose here to estimate more tightly the precision matrix, i.e., the inverse of the covariance matrix (2) from the input data $(x^{(i)})_{1 \leq i \leq N}$, by a penalized maximum likelihood approach leading to the following minimization problem:

$$\min_{C \in S_+^{n+}} (F(C) \triangleq f(C) + T_S(C) + g_0(C) + g_1(C))$$  \hspace{1cm} (4)\hfill

where

$$(\forall C \in S_+^{n+}) \quad f(C) \triangleq \log \det (C^{-1} + \sigma^2 I_d)$$  \hspace{1cm} (5)\hfill

$$(\forall C \in S_+^{n+}) \quad T_S(C) \triangleq \text{tr} \left( (I_d + \sigma^2 C)^{-1} CS \right),$$  \hspace{1cm} (6)\hfill

the first (resp. second) function value being taken equal to $+\infty$ outside $S_+^{n+}$ (resp. $S_n^+$). A hybrid regularization term is considered, which consists of $g_0 + g_1$. Function $g_0$ is a spectral symmetric function, i.e. $(\forall C \in S_n) \quad g_0(C) = \psi(Pd)$, where $P$ is any permutation matrix, $d = [d_1, \ldots, d_n]^\top$ is the vector of eigenvalues of $C$, and $\psi \in \Gamma_0(\mathbb{R}^n)$. Moreover, $g_1 \in \Gamma_0(\mathbb{R}^{n \times n})$ acts on the whole matrix $C$. It is worthwhile noticing that, when $\sigma = 0$, $g_0 = 0$ and $g_1 = \| \cdot \|_1$, (4) 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 $F$ in (4) is a difference of concave functions, which will be a key property in the optimization approach we propose.

**Lemma 1.** Consider (4) 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 $T_S$ is concave on $S_+^{n+}$.

**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 (6) is a positive semi-definite linear operator on $S_n$.

According to Lemma 1, the cost function $F$ is not convex, thus we propose to adopt a Majorize-Minimize (MM) strategy [29, 25, 30, 31] to solve (4). At each iteration $\ell \in \mathbb{N}$, we upper bound $F$ 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:

$$C^{(\ell+1)} = \arg\min_{C \in S_+^{n+}} \left( f(C) + \text{tr} \left( \nabla T_S(C^{(\ell)}) C \right) + g_0(C) + g_1(C) \right).$$  \hspace{1cm} (7)\hfill

In (7), $\nabla T_S(C^{(\ell)})$ is the gradient of $T_S$ at $C^{(\ell)}$, which is given by $\nabla T_S(C^{(\ell)}) = (I_d + \sigma^2 C^{(\ell)})^{-1} S \cdot (I_d + \sigma^2 C^{(\ell)})^{-1}$. 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 (7), the DR method minimizes the sum of $f + \text{tr} \left( \nabla T_S(C^{(\ell)}) \cdot \cdot \right) + g_0$ and $g_1$ by alternatingly 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 $\gamma h$ at $C$ will be denoted by $\text{prox}_{\gamma h}(C)$ and corresponds to the unique minimizer of $\gamma h + \| \cdot - C \|^2/2$, where $\| \cdot \|_F$ is the Frobenius norm. Most practical choices for $g_1$ allow to compute its proximity operator via an explicit formula, while the subsequent lemma reveals a practical way to compute the proximity operator of the other term appearing in (7).

**Lemma 2.** Let $\gamma \in [0, +\infty]$ and $\overline{C} \in S_n$. For every $d = [d_1, \ldots, d_n]^\top \in \mathbb{R}^n$, let $\varphi(d) = \sum_{j=1}^{n} \log \left( (1 + \sigma^2 d_j)/d_j \right)$ if $d \in [0, +\infty]$ 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 $\text{dom} \varphi \cap \text{dom} \psi \neq \emptyset$. Let $\lambda \in \mathbb{R}^n$ and $U$ be an orthogonal matrix such that $\overline{C} = \gamma \nabla T_S(C^{(\ell)}) = U \text{Diag}(\lambda) U^\top$. Then

$$\text{prox}_{\gamma \varphi(f \pm \text{tr} \left( \nabla T_S(C^{(\ell)}) \cdot \cdot \right) + g_0)}(\overline{C}) = U \text{Diag} \left( \text{prox}_{\gamma \varphi(\varphi + \psi)}(\lambda) \right) U^\top.$$  \hspace{1cm} (8)\hfill

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.

**Theorem 1.** Let $(C^{(\ell)})_{\ell \geq 0}$ be a sequence generated by (7). Assume that $\text{dom} f \cap \text{dom} g_0 \cap \text{dom} g_1 \neq \emptyset$, $f + g_0 + g_1$ is coercive, and $\{C \in S_n | F(C) \leq F(C^{(0)})\}$ is included in the relative interior of $\text{dom} g_0 \cap \text{dom} g_1$. Then, the following properties hold:

1. $(F(C^{(\ell)}))_{\ell \geq 0}$ is a decaying sequence converging to $F \in \mathbb{R}$.\hfill\Box
Algorithm 1 MMDR: Majorization–Minimization algorithm with Douglas–Rachford inner steps

1: Set $\gamma > 0$, $C^{(0)} \in S_n$, let $S \in S_n^+$ be the given data
2: for $l = 0, 1, \ldots$ do
3:  Set $C^{(l)} = C^{(l)}$
4:  for $k = 0, 1, \ldots$ do
5:  Compute $U^{(k)}, \Lambda^{(k)} = \text{Diag}((\Lambda^{(k)})$ such that
6:  \[ C^{(\ell,k+i)} = \text{prox}_{\gamma T_S}(C^{(\ell,k)}) \]
7:  \[ Y^{(\ell,k)} = \text{prox}_{\gamma\phi}(2C^{(\ell,k+i)} - C^{(\ell,k)}) \]
8:  Choose $\alpha_k \in [0,2]$ 
9:  \[ C^{(\ell,k+1)} = C^{(\ell,k)} + \alpha_k (Y^{(\ell,k)} - C^{(\ell,k+i)}) \]
10: \[ C^{(\ell+1)} = C^{(\ell,k+\frac{1}{2})} \]
11: end for 
12: end for 

2. $(C^{(\ell)})_{\ell\geq0}$ has a cluster point.

3. Every cluster point $\hat{C}$ of $(C^{(\ell)})_{\ell\geq0}$ is such that $F(\hat{C}) = \hat{F}$ and it is a critical point of $F$, i.e. $-\nabla f(\hat{C}) = \nabla T_S(\hat{C}) \in \partial(g_0 + g_1)(\hat{C})$.

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:\footnote{http://stanford.edu/~boyd/papers/admm/covsel/covsel_example.html} a sparse precision matrix $C_0$ of dimension $n \times n$ is randomly created, then its inverse $\Sigma_0$ is used to generate $N$ realizations $(x^{(i)})_{1 \leq i \leq N}$ of a Gaussian random variable with zero mean and covariance $\Sigma_0$. Gaussian noise with variance $\sigma^2$ is added to these realizations, hence the final covariance matrix $\Sigma$ fulfills exactly (2), with $\Sigma_0 = \Lambda^T \Sigma A$.

The settings of our experiments are $n = 100$, $N = 10000$, the number of nonzero elements in $C_0$ is $10$, $g_1 = \mu_1 \cdot ||1||_1$, $\mu_1 > 0$ and $g_0(C) = \mu_0 R_1(C^{-1})$, where $R_1$ is the Schatten–1–norm (also called nuclear norm) and $\mu_0 > 0$. The parameters $\mu_0$ and $\mu_1$ allow us to adjust the incidence of the regularization functionals on the final 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 $F$ 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 $\text{rmse}$ on the covariance matrix, i.e. $\text{rmse} = ||\Sigma_0 - \Sigma_{\text{rec}}||_F/||\Sigma_0||_F$, where $\Sigma_{\text{rec}}$ is the reconstructed covariance matrix. Panels (a-b) in Figure 1 are useful to visually compare the true covariance matrix $\Sigma_0$ and the recovered one $\Sigma_{\text{rec}}$: 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 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 (4), 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. $\sigma$ is set to 0 in (4): applying the DR algorithm to the result-
ing 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 $\sigma$.

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


