Variable Selection for Gaussian Graphical Models

Jean Honorio[1] jhonorio@cs.sunysb.edu
Dimitris Samaras[1] samaras@cs.sunysb.edu
Irina Rish[2] rish@us.ibm.com
Guillermo Cecchi[2] gcecchi@us.ibm.com

[1]Stony Brook University, Stony Brook, NY 11794, USA
[2]IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

Abstract

We present a *variable-selection structure learning* approach for Gaussian graphical models. Unlike standard sparseness promoting techniques, our method aims at selecting the most-important variables besides simply sparsifying the set of edges. Through simulations, we show that our method outperforms the state-of-the-art in recovering the ground truth model. Our method also exhibits better generalization performance in a wide range of complex real-world datasets: brain fMRI, gene expression, NASDAQ stock prices and world weather. We also show that our resulting networks are more interpretable in the context of brain fMRI analysis, while retaining discriminability. From an optimization perspective, we show that a block coordinate descent method generates a sequence of positive definite solutions. Thus, we reduce the original problem into a sequence of strictly convex (ℓ_1,ℓ_p) regularized quadratic minimization subproblems for $p \in \{2, \infty\}$. Our algorithm is well founded since the optimal solution of the maximization problem is unique and bounded.

1 Introduction

Structure learning aims to discover the topology of a probabilistic graphical model such that this model represents accurately a given dataset. Accuracy of representation is measured by the likelihood that the model explains the observed data. From an algorithmic point of view, one challenge faced by structure learning is that the number of possible structures is super-exponential in the number of variables. From a statistical perspective, it is very important to find good regularization techniques in order to avoid over-fitting and to achieve better generalization performance. Such regularization techniques will aim to reduce the complexity of the graphical model, which is measured by its number of parameters.

For Gaussian graphical models, the number of parameters, the number of edges in the structure and the number of non-zero elements in the inverse covariance or precision matrix are equivalent measures of complexity. Therefore, several techniques focus on enforcing sparseness of the precision matrix. An approximation method proposed in [1] relied on a sequence of sparse regressions. Maximum likelihood estimation with an ℓ_1-norm penalty for encouraging sparseness is proposed in [2, 3, 4].

In this paper, we enforce a particular form of sparseness: that only a small number of nodes in the graphical model interact with each other. Intuitively, we want to select these "important" nodes. However, the above methods for sparsifying network structure do not directly promote variable selection, i.e. group-wise elimination of all edges adjacent to an "unimportant" node. Variable selection in graphical models present several advantages. From a computational point of view, reducing the number of variables can significantly reduce the number of precision-matrix parameters. Moreover, group-wise edge elimination may serve as a more aggressive regularization, removing all "noisy" edges associated with nuisance variables at once, and potentially leading to better generalization performance, especially if, indeed, the underlying problem structure involves only a limited number of "important" variables. Finally, variable selection improves interpretability of the graphical model: for example, when learning a graphical model of brain area connectivity, variable selection may help to localize brain areas most relevant to particular mental states.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

Table 1: Notation used in this paper.

Notation	Description
$\|\mathbf{c}\|_1$	ℓ_1-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\sum_n \|c_n\|$
$\|\mathbf{c}\|_\infty$	ℓ_∞-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\max_n \|c_n\|$
$\|\mathbf{c}\|_2$	Euclidean norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\sqrt{\sum_n c_n^2}$
$\mathbf{A} \succeq \mathbf{0}$	$\mathbf{A} \in \mathbb{R}^{N\times N}$ is symmetric and positive semidefinite
$\mathbf{A} \succ \mathbf{0}$	$\mathbf{A} \in \mathbb{R}^{N\times N}$ is symmetric and positive definite
$\|\mathbf{A}\|_1$	ℓ_1-norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\sum_{mn} \|a_{mn}\|$
$\|\mathbf{A}\|_\infty$	ℓ_∞-norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\max_{mn} \|a_{mn}\|$
$\|\mathbf{A}\|_2$	spectral norm of $\mathbf{A} \in \mathbb{R}^{N\times N}$, i.e. the maximum eigenvalue of $\mathbf{A} \succ \mathbf{0}$
$\|\mathbf{A}\|_{\mathfrak{F}}$	Frobenius norm of $\mathbf{A} \in \mathbb{R}^{M\times N}$, i.e. $\sqrt{\sum_{mn} a_{mn}^2}$
$\langle \mathbf{A}, \mathbf{B} \rangle$	scalar product of $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{M\times N}$, i.e. $\sum_{mn} a_{mn} b_{mn}$

Our contribution is to develop variable-selection in the context of learning sparse Gaussian graphical models. To achieve this, we add an $\ell_{1,p}$-norm regularization term to the maximum likelihood estimation problem, for $p \in \{2, \infty\}$. We optimize this problem through a block coordinate descent method which yields sparse and positive definite estimates. We show that our method outperforms the state-of-the-art in recovering the ground truth model through synthetic experiments. We also show that our structures have higher test log-likelihood than competing methods, in a wide range of complex real-world datasets: brain fMRI, gene expression, NASDAQ stock prices and world weather. In particular, in the context of brain fMRI analysis, we show that our method produces more interpretable models that involve few brain areas, unlike standard sparseness promoting techniques which produce hard-to-interpret networks involving most of the brain. Moreover, our structures are as good as standard sparseness promoting techniques, when used for classification purposes.

Sec.2 introduces Gaussian graphical models and techniques for learning them from data. Sec.3 sets up the $\ell_{1,p}$-regularized maximum likelihood problem and discusses its properties. Sec.4 describes our block coordinate descent method. Experimental results are in Sec.5.

2 Background

In this paper, we use the notation in Table 1.

A *Gaussian graphical model* is a graph in which all random variables are continuous and jointly Gaussian. This model corresponds to the multivariate normal distribution for N variables with covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$. Conditional independence in a Gaussian graphical model is simply reflected in the zero entries of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ [5]. Let $\mathbf{\Omega} = \{\omega_{n_1 n_2}\}$, two variables n_1 and n_2 are conditionally independent if and only if $\omega_{n_1 n_2} = 0$.

The estimation of sparse precision matrices was first introduced in [6]. It is well known that finding the most sparse precision matrix which fits a dataset is a NP-hard problem [2]. Therefore, several ℓ_1-regularization methods have been proposed.

Given a dense sample covariance matrix $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, the problem of finding a sparse precision matrix $\mathbf{\Omega}$ by regularized maximum likelihood estimation is given by:

$$\max_{\mathbf{\Omega}\succ\mathbf{0}} \Big(\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle - \rho \|\mathbf{\Omega}\|_1 \Big) \qquad (1)$$

for $\rho > 0$. The term $\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle$ is the Gaussian log-likelihood. The term $\|\mathbf{\Omega}\|_1$ encourages sparseness of the precision matrix or conditional independence among variables.

Several algorithms have been proposed for solving eq.(1): *covariance selection* [2], *graphical lasso* [3] and the *Meinshausen-Bühlmann approximation* [1].

Besides sparseness, several regularizers have been proposed for Gaussian graphical models, for enforcing diagonal structure [7], spatial coherence [8], common structure among multiple tasks [9], or sparse changes in controlled experiments [10]. In particular, different group sparse priors have been proposed for enforcing block structure for known block-variable assignments [11, 12] and unknown block-variable assignments [13, 14], or power law regularization in scale free networks [15].

Variable selection has been applied to very diverse problems, such as linear regression [16], classification [17, 18, 19] and reinforcement learning [20].

Structure learning through ℓ_1-regularization has been also proposed for different types of graphical models: Markov random fields [21]; Bayesian networks on binary variables [22]; Conditional random fields [23]; and Ising models [24].

3 Preliminaries

In this section, we set up the problem and discuss some of its properties.

3.1 Problem Setup

We propose priors that are motivated from the variable selection literature from regression and classification, such as group lasso [25, 26, 27] which imposes an $\ell_{1,2}$-norm penalty, and simultaneous lasso [28, 29] which imposes an $\ell_{1,\infty}$-norm penalty.

Recall that an edge in a Gaussian graphical model corresponds to a non-zero entry in the precision matrix. We promote variable selection by learning a structure with a small number of nodes that interact with each other, or equivalently a large number of nodes that are disconnected from the rest of the graph. For each disconnected node, its corresponding row in the precision matrix (or column given that it is symmetric) contains only zeros (except for the diagonal). Therefore, the use of row-level regularizers such as the $\ell_{1,p}$-norm are natural in our context. Note that our goal differs from sparse Gaussian graphical models, in which sparseness is imposed at the edge level only. We additionally impose sparseness at the node level, which promotes conditional independence of variables with respect to all other variables.

Given a dense sample covariance matrix $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, we learn a precision matrix $\mathbf{\Omega} \in \mathbb{R}^{N \times N}$ for N variables. The *variable-selection structure learning problem* is defined as:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \Big(\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle - \rho \|\mathbf{\Omega}\|_1 - \tau \|\mathbf{\Omega}\|_{1,p} \Big) \quad (2)$$

for $\rho > 0$, $\tau > 0$ and $p \in \{2, \infty\}$. The term $\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle$ is the Gaussian log-likelihood. $\|\mathbf{\Omega}\|_1$ encourages sparseness of the precision matrix or conditional independence among variables. The last term $\|\mathbf{\Omega}\|_{1,p}$ is our variable selection regularizer, and it is defined as:

$$\|\mathbf{\Omega}\|_{1,p} = \sum_n \|(\omega_{n,1}, \ldots, \omega_{n,n-1}, \omega_{n,n+1}, \ldots, \omega_{n,N})\|_p \quad (3)$$

In a technical report, [30] proposed an optimization problem that is similar to eq.(2). The main differences are that their model does not promote sparseness, and that they do not solve the original maximum likelihood problem, but instead build upon an approximation (pseudo-likelihood) approach of Meinshausen-Bühlmann [1] based on independent linear regression problems. Finally, note that regression based methods such as [1] have been already shown in [3] to have worse performance than solving the original maximum likelihood problem. In this paper, we solve the original maximum likelihood problem.

3.2 Bounds

In what follows, we discuss uniqueness and boundedness of the optimal solution of our problem.

Lemma 1. *For $\rho > 0$, $\tau > 0$, the variable-selection structure learning problem in eq.(2) is a maximization problem with concave (but not strictly concave) objective function and convex constraints.*

Proof. The Gaussian log-likelihood is concave, since $\log\det$ is concave on the space of symmetric positive definite matrices, and since the linear operator $\langle \cdot, \cdot \rangle$ is also concave. Both regularization terms, the negative ℓ_1-norm as well as the negative $\ell_{1,p}$-norm defined in eq.(3) are non-smooth concave functions. Finally, $\mathbf{\Omega} \succ \mathbf{0}$ is a convex constraint. □

For clarity of exposition, we assume that the diagonals of $\mathbf{\Omega}$ are penalized by our variable selection regularizer defined in eq.(3).

Theorem 2. *For $\rho > 0$, $\tau > 0$, the optimal solution to the variable-selection structure learning problem in eq.(2) is unique and bounded as follows:*

$$\left(\frac{1}{\|\widehat{\mathbf{\Sigma}}\|_2 + N\rho + N^{1/p'}\tau} \right) \mathbf{I} \preceq \mathbf{\Omega}^* \preceq \left(\frac{N}{\max(\rho, \tau)} \right) \mathbf{I} \quad (4)$$

where $\ell_{p'}$-norm is the dual of the ℓ_p-norm, i.e. $(p = 2, p' = 2)$ or $(p = \infty, p' = 1)$.

Proof. By using the identity for dual norms $\kappa \|\mathbf{c}\|_p = \max_{\|\mathbf{d}\|_{p'} \leq \kappa} \mathbf{d}^{\mathrm{T}} \mathbf{c}$ in eq.(2), we get:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \min_{\substack{\|\mathbf{A}\|_\infty \leq \rho \\ \|\mathbf{B}\|_{\infty,p'} \leq \tau}} \Big(\log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}} + \mathbf{A} + \mathbf{B}, \mathbf{\Omega} \rangle \Big) \quad (5)$$

where $\|B\|_{\infty,p'} = \max_n \|(b_{n,1}, \ldots, b_{n,N})\|_{p'}$. By virtue of Sion's minimax theorem, we can swap the order of max and min. Furthermore, note that the optimal solution of the inner equation is given by $\mathbf{\Omega} = (\widehat{\mathbf{\Sigma}} + \mathbf{A} + \mathbf{B})^{-1}$. By replacing this solution in eq.(5), we get the dual problem of eq.(2):

$$\min_{\substack{\|\mathbf{A}\|_\infty \leq \rho \\ \|\mathbf{B}\|_{\infty,p'} \leq \tau}} \Big(-\log\det(\widehat{\mathbf{\Sigma}} + \mathbf{A} + \mathbf{B}) - N \Big) \quad (6)$$

In order to find a lower bound for the minimum eigenvalue of $\mathbf{\Omega}^*$, note that $\|\mathbf{\Omega}^{*-1}\|_2 = \|\widehat{\mathbf{\Sigma}} + \mathbf{A} + \mathbf{B}\|_2 \leq \|\widehat{\mathbf{\Sigma}}\|_2 + \|\mathbf{A}\|_2 + \|\mathbf{B}\|_2 \leq \|\widehat{\mathbf{\Sigma}}\|_2 + N\|\mathbf{A}\|_\infty + N^{1/p'}\|\mathbf{B}\|_{\infty,p'} \leq \|\widehat{\mathbf{\Sigma}}\|_2 + N\rho + N^{1/p'}\tau$. (Here we used $\|\mathbf{B}\|_2 \leq N^{1/p'}\|\mathbf{B}\|_{\infty,p'}$ as shown in Appendix A)

In order to find an upper bound for the maximum eigenvalue of $\mathbf{\Omega}^*$, note that, at optimum, the primal-dual gap is zero:

$$-N + \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^* \rangle + \rho \|\mathbf{\Omega}^*\|_1 + \tau \|\mathbf{\Omega}^*\|_{1,p} = 0 \quad (7)$$

The upper bound is found as follows: $\|\mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}^*\|_{\mathfrak{F}} \leq \|\mathbf{\Omega}^*\|_1 = (N - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^* \rangle - \tau \|\mathbf{\Omega}^*\|_{1,p})/\rho$. Note that $\tau \|\mathbf{\Omega}^*\|_{1,p} \geq 0$, and since $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$ and $\mathbf{\Omega}^* \succ \mathbf{0}$, it follows that $\langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^* \rangle \geq 0$. Therefore, $\|\mathbf{\Omega}^*\|_2 \leq \frac{N}{\rho}$. In

a similar fashion, $\|\mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}^*\|_{1,p} = (N - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^* \rangle - \rho\|\mathbf{\Omega}^*\|_1)/\tau$. (Here we used $\|\mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}^*\|_{1,p}$ as shown in Appendix A). Note that $\rho\|\mathbf{\Omega}^*\|_1 \geq 0$ and $\langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega}^* \rangle \geq 0$. Therefore, $\|\mathbf{\Omega}^*\|_2 \leq \frac{N}{\tau}$. □

4 Block Coordinate Descent Method

Since the objective function in eq.(2) contains a non-smooth regularizer, methods such as gradient descent cannot be applied. On the other hand, subgradient descent methods very rarely converge to non-smooth points [31]. In our problem, these non-smooth points correspond to zeros in the precision matrix, are often the true minima of the objective function, and are very desirable in the solution because they convey information of conditional independence among variables.

We apply block coordinate descent method on the primal problem [8, 9], unlike covariance selection [2] and graphical lasso [3] which optimize the dual. Optimization of the dual problem in eq.(6) by a block coordinate descent method can be done with quadratic programming for $p = \infty$ but not for $p = 2$ (i.e. the objective function is quadratic for $p \in \{2, \infty\}$, the constraints are linear for $p = \infty$ and quadratic for $p = 2$). Optimization of the primal problem provides the same efficient framework for $p \in \{2, \infty\}$. We point out that a projected subgradient method as in [11] cannot be applied since our regularizer does not decompose into disjoint subsets. Our problem contains a positive definiteness constraint and therefore it does not fall in the general framework of [25, 26, 27, 32, 28, 29] which consider unconstrained problems only. Finally, more recent work of [33, 34] consider subsets with overlap, but it does still consider unconstrained problems only.

Theorem 3. *The block coordinate descent method for the variable-selection structure learning problem in eq.(2) generates a sequence of positive definite solutions.*

Proof. Maximization can be performed with respect to one row and column of all precision matrices $\mathbf{\Omega}$ at a time. Without loss of generality, we use the last row and column in our derivation. Let:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{W} & \mathbf{y} \\ \mathbf{y}^{\mathrm{T}} & z \end{bmatrix} \quad , \quad \widehat{\mathbf{\Sigma}} = \begin{bmatrix} \mathbf{S} & \mathbf{u} \\ \mathbf{u}^{\mathrm{T}} & v \end{bmatrix} \tag{8}$$

where $\mathbf{W}, \mathbf{S} \in \mathbb{R}^{N-1 \times N-1}$, $\mathbf{y}, \mathbf{u} \in \mathbb{R}^{N-1}$.

In terms of the variables $\mathbf{y}, z$ and the constant matrix $\mathbf{W}$, the variable-selection structure learning problem in eq.(2) can be reformulated as:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \begin{pmatrix} \log(z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}) - 2\mathbf{u}^{\mathrm{T}}\mathbf{y} - (v+\rho)z \\ -2\rho\|\mathbf{y}\|_1 - \tau\|\mathbf{y}\|_p - \tau\sum_n \|(y_n, t_n)\|_p \end{pmatrix} \tag{9}$$

where $t_n = \|(w_{n,1}, \ldots, w_{n,n-1}, w_{n,n+1}, \ldots, w_{n,N})\|_p$.

If $\mathbf{\Omega}$ is a symmetric matrix, according to the Haynsworth inertia formula, $\mathbf{\Omega} \succ \mathbf{0}$ if and only if its Schur complement $z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y} > 0$ and $\mathbf{W} \succ \mathbf{0}$. By maximizing eq.(9) with respect to z, we get:

$$z - \mathbf{y}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y} = \frac{1}{v+\rho} \tag{10}$$

and since $v > 0$ and $\rho > 0$, this implies that the Schur complement in eq.(10) is positive. Finally, in our iterative optimization, it suffices to initialize $\mathbf{\Omega}$ to a matrix known to be positive definite, e.g. a diagonal matrix with positive elements. □

Theorem 4. *The block coordinate descent method for the variable-selection structure learning problem in eq.(2) is equivalent to solving a sequence of strictly convex (ℓ_1,$\ell_{1,p}$) regularized quadratic subproblems for $p \in \{2, \infty\}$:*

$$\min_{\mathbf{y} \in \mathbb{R}^{N-1}} \begin{pmatrix} \frac{1}{2}\mathbf{y}^{\mathrm{T}}(v+\rho)\mathbf{W}^{-1}\mathbf{y} + \mathbf{u}^{\mathrm{T}}\mathbf{y} \\ +\rho\|\mathbf{y}\|_1 + \frac{\tau}{2}\|\mathbf{y}\|_p + \frac{\tau}{2}\sum_n \|(y_n, t_n)\|_p \end{pmatrix} \tag{11}$$

Proof. By replacing the optimal z given by eq.(10) into the objective function in eq.(9), we get eq.(11). Since $\mathbf{W} \succ \mathbf{0} \Rightarrow \mathbf{W}^{-1} \succ \mathbf{0}$, hence eq.(11) is strictly convex. □

Lemma 5. *If $\|\mathbf{u}\|_\infty \leq \rho + \tau/(2(N-1)^{1/p'})$ or $\|\mathbf{u}\|_{p'} \leq \rho + \tau/2$, the ($\ell_1, \ell_{1,p}$) regularized quadratic problem in eq.(11) has the minimizer $\mathbf{y}^* = \mathbf{0}$.*

Proof. Note that since $\mathbf{W} \succ \mathbf{0} \Rightarrow \mathbf{W}^{-1} \succ \mathbf{0}$, $\mathbf{y}^* = \mathbf{0}$ is the minimizer of the quadratic part of eq.(11). It suffices to prove that the remaining part is also minimized for $\mathbf{y}^* = \mathbf{0}$, i.e. $\mathbf{u}^{\mathrm{T}}\mathbf{y} + \rho\|\mathbf{y}\|_1 + \frac{\tau}{2}\|\mathbf{y}\|_p + \frac{\tau}{2}\sum_n \|(y_n, t_n)\|_p \geq \frac{\tau}{2}\sum_n t_n$ for an arbitrary $\mathbf{y}$. The lower bound comes from setting $\mathbf{y}^* = \mathbf{0}$ in eq.(11) and by noting that $(\forall n)\ t_n > 0$.

By using lower bounds $\sum_n \|(y_n, t_n)\|_p \geq \sum_n t_n$ and either $\|\mathbf{y}\|_p \geq \|\mathbf{y}\|_1/(N-1)^{1/p'}$ or $\|\mathbf{y}\|_1 \geq \|\mathbf{y}\|_p$, we modify the original claim into a stronger one, i.e. $\mathbf{u}^{\mathrm{T}}\mathbf{y} + (\rho + \tau/(2(N-1)^{1/p'}))\|\mathbf{y}\|_1 \geq 0$ or $\mathbf{u}^{\mathrm{T}}\mathbf{y} + (\rho + \tau/2)\|y\|_p \geq 0$. Finally, by using the identity for dual norms $\kappa\|\mathbf{y}\|_p = \max_{\|\mathbf{d}\|_{p'} \leq \kappa} \mathbf{d}^{\mathrm{T}}\mathbf{y}$, we have $\max_{\|\mathbf{d}\|_\infty \leq \rho + \tau/(2(N-1)^{1/p'})} (\mathbf{u}+\mathbf{d})^{\mathrm{T}}\mathbf{y} \geq 0$ or $\max_{\|\mathbf{d}\|_{p'} \leq \rho + \tau/2} (\mathbf{u}+\mathbf{d})^{\mathrm{T}}\mathbf{y} \geq 0$, which proves our claim. □

Remark 6. *By using Lemma 5, we can reduce the size of the original problem by removing variables in which this condition holds, since it only depends on the dense sample covariance matrix.*

Theorem 7. *The coordinate descent method for the* $(\ell_1, \ell_{1,p})$ *regularized quadratic problem in eq.*(11) *is equivalent to solving a sequence of strictly convex* (ℓ_1, ℓ_p) *regularized quadratic subproblems:*

$$\min_x \left(\frac{1}{2} q x^2 - cx + \rho |x| + \frac{\tau}{2} \|(x,a)\|_p + \frac{\tau}{2} \|(x,b)\|_p \right) \tag{12}$$

Proof. Without loss of generality, we use the last row and column in our derivation, since permutation of rows and columns is always possible. Let:

$$\mathbf{W}^{-1} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{h}_{12} \\ \mathbf{h}_{12}{}^{\mathrm{T}} & h_{22} \end{bmatrix} \quad , \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ x \end{bmatrix} \quad , \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ u_2 \end{bmatrix} \tag{13}$$

where $\mathbf{H}_{11} \in \mathbb{R}^{N-2 \times N-2}$, $\mathbf{h}_{12}, \mathbf{y}_1, \mathbf{u}_1 \in \mathbb{R}^{N-2}$.

In terms of the variable x and the constants $q = (v+\rho)h_{22}$, $c = -((v+\rho)\mathbf{h}_{12}{}^{\mathrm{T}}\mathbf{y}_1 + u_2)$, $a = \|\mathbf{y}_1\|_p$, $b = t_n$, the $(\ell_1, \ell_{1,p})$ regularized quadratic problem in eq.(11) can be reformulated as in eq.(12). Moreover, since $v > 0 \wedge \rho > 0 \wedge h_{22} > 0 \Rightarrow q > 0$, and therefore eq.(12) is strictly convex. □

For $p = \infty$, eq.(12) has five points in which the objective function is non-smooth, i.e. $x \in \{-\max(a,b), -\min(a,b), 0, \min(a,b), \max(a,b)\}$. Furthermore, since the objective function is quadratic on each interval, it admits a closed form solution.

For $p = 2$, eq.(12) has only one non-smooth point, i.e. $x = 0$. Given the objective function $f(x)$, we first compute the left derivative $\partial_- f(0) = -c - \rho$ and the right derivative $\partial_+ f(0) = -c + \rho$. If $\partial_- f(0) \leq 0 \wedge \partial_+ f(0) \geq 0 \Rightarrow x^* = 0$. If $\partial_- f(0) > 0 \Rightarrow x^* < 0$ and we use the one-dimensional *Newton-Raphson method* for finding x^*. If $\partial_+ f(0) < 0 \Rightarrow x^* > 0$. For numerical stability, we add a small $\varepsilon > 0$ to the ℓ_2-norms by using $\sqrt{x^2 + a^2 + \varepsilon}$ instead of $\|(x,a)\|_2$.

Algorithm 1 shows the block coordinate descent method in detail. A careful implementation leads to a time complexity of $\mathcal{O}(KN^3)$ for K iterations and N variables. In our experiments, the algorithm converges quickly in usually $K = 10$ iterations. Polynomial dependence $\mathcal{O}(N^3)$ on the number of variables is expected since no algorithm can be faster than computing the inverse of the sample covariance in the case of an infinite sample.

Algorithm 1 Block Coordinate Descent

Input: $\widehat{\mathbf{\Sigma}} \succeq \mathbf{0}$, $\rho > 0$, $\tau > 0$, $p \in \{2, \infty\}$
Initialize $\mathbf{\Omega} = \mathbf{diag}(\widehat{\mathbf{\Sigma}})^{-1}$
for each iteration $1, \ldots, K$ and each variable $1, \ldots, N$ **do**
  Split $\mathbf{\Omega}$ into $\mathbf{W}, \mathbf{y}, z$ and $\widehat{\mathbf{\Sigma}}$ into $\mathbf{S}, \mathbf{u}, v$ as described in eq.(8)
  Update $\mathbf{W}^{-1}$ by using the Sherman-Woodbury-Morrison formula (Note that when iterating from one variable to the next one, only one row and column change on matrix $\mathbf{W}$)
  for each variable $1, \ldots, N-1$ **do**
    Split $\mathbf{W}^{-1}, \mathbf{y}, \mathbf{u}$ as in eq.(13)
    Solve the (ℓ_1, ℓ_p) regularized quadratic problem in closed form ($p = \infty$) or by using the Newton-Raphson method ($p = 2$)
  end for
  Update $z \leftarrow \frac{1}{v+\rho} + \mathbf{y}^{\mathrm{T}} \mathbf{W}^{-1} \mathbf{y}$
end for
Output: $\mathbf{\Omega} \succ \mathbf{0}$

5 Experimental Results

We test with a synthetic example the ability of the method to recover ground truth structure from data. The model contains $N \in \{50, 100, 200\}$ variables. For each of 50 repetitions, we first select a proportion of "connected" nodes (either 0.2,0.5,0.8) from the N variables. The unselected (i.e. "disconnected") nodes do not participate in any edge of the ground truth model. We then generate edges among the connected nodes with a required density (either 0.2,0.5,0.8), where each edge weight is generated uniformly at random from $\{-1, +1\}$. We ensure positive definiteness of $\mathbf{\Omega}_g$ by verifying that its minimum eigenvalue is at least 0.1. We then generate a dataset of 50 samples. We model the ratio $\bar{\sigma}_c / \bar{\sigma}_d$ between the standard deviation of connected versus disconnected nodes. In the "high variance confounders" regime, $\bar{\sigma}_c / \bar{\sigma}_d = 1$ which means that on average connected and disconnected variables have the same standard deviation. In the "low variance confounders" regime, $\bar{\sigma}_c / \bar{\sigma}_d = 10$ which means that on average the standard deviation of a connected variable is 10 times the one of a disconnected variable. Variables with low variance produce higher values in the precision matrix than variables with high variance. We analyze both regimes in order to evaluate the impact of this effect in structure recovery.

In order to measure the closeness of the recovered models to the ground truth, we measured the Kullback-Leibler (KL) divergence, sensitivity (one minus the fraction of falsely excluded edges) and specificity (one minus the fraction of falsely included edges). We compare to the following methods: covariance selection [2], graphical lasso [3], Meinshausen-Bühlmann approximation [1] and Tikhonov regularization. For our method, we found that the variable selection parameter $\tau = 50\rho$ provides reasonable results, in both synthetic and real-world experiments. Therefore, we report results only with respect to the sparseness parameter ρ.

First, we test the performance of our methods for in-

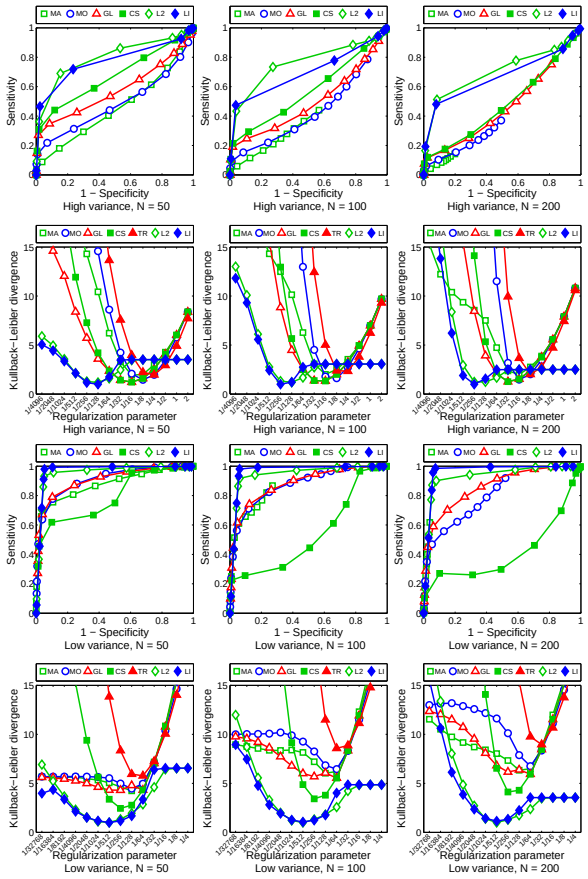

Figure 1: ROC curves (first row) and KL divergence (second row) for the "high variance confounders" regime. ROC curves (third row) and KL divergence (fourth row) for the "low variance confounders" regime. Left: $N = 50$ variables, center: $N = 100$ variables, right: $N = 200$ variables (connectedness 0.8, edge density 0.5). Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) recover edges better and produce better probability distributions than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). Our methods degrade less in recovering the ground truth edges when the number of variables grows.

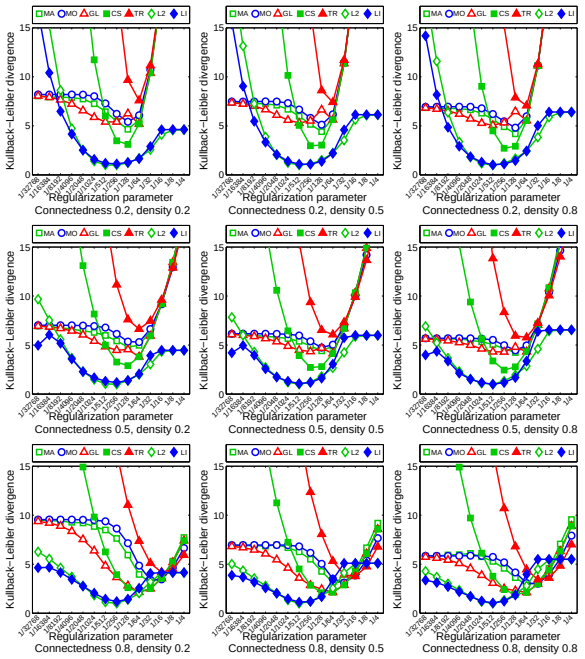

Figure 2: Cross-validated KL divergence for structures learnt for the "low variance confounders" regime ($N = 50$ variables, different connectedness and density levels). Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) produce better probability distributions than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR).

creasing number of variables, moderate edge density (0.5) and high proportion of connected nodes (0.8). Fig.1 shows the ROC curves and KL divergence between the recovered models and the ground truth. In both "low" and "high variance confounders" regimes, our $\ell_{1,2}$ and $\ell_{1,\infty}$ methods recover ground truth edges better than competing methods (higher ROC) and produce better probability distributions (lower KL divergence) than the other methods. Our methods degrade less than competing methods in recovering the ground truth edges when the number of variables grows, while the KL divergence behavior remains similar.

Second, we test the performance of our methods with respect to edge density and the proportion of connected nodes. Fig.2 shows the KL divergence between the recovered models and the ground truth for the "low variance confounders" regime. Our $\ell_{1,2}$ and $\ell_{1,\infty}$ methods produce better probability distributions (lower KL divergence) than the remaining techniques. (Please, see Appendix B for results on ROC and the "high variance confounders" regime.)

Our $\ell_{1,2}$ method takes 0.07s for $N = 100$, 0.12s for $N = 200$ variables. Our $\ell_{1,\infty}$ method takes 0.13s for $N = 100$, 0.63s for $N = 200$. Graphical lasso [3], the fastest and most accurate competing method in our evaluation, takes 0.11s for $N = 100$, 0.49s for $N = 200$. Our $\ell_{1,\infty}$ method is slightly slower than graphical lasso, while our $\ell_{1,2}$ method is the fastest. One reason for this is that Lemma 5 eliminates more variables in the $\ell_{1,2}$ setting.

For experimental validation on real-world datasets, we use datasets with a diverse nature of probabilistic relationships: brain fMRI, gene expression, NASDAQ stock prices and world weather. The *brain fMRI* dataset collected by [35] captures brain function of 15 cocaine addicted and 11 control subjects under conditions of monetary reward. Each subject contains 87 scans of $53 \times 63 \times 46$ voxels each, taken ev-

ery 3.5 seconds. Registration to a common spatial template and spatial smoothing was done in SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`). After sampling each $4 \times 4 \times 4$ voxels, we obtained 869 variables. The *gene expression* dataset contains 8,565 variables and 587 samples. The dataset was collected by [36] from drug treated rat livers, by treating rats with a variety of fibrate, statin, or estrogen receptor agonist compounds. The dataset is publicly available at `http://www.ebi.ac.uk/`. In order to consider the full set of genes, we had to impute a very small percentage (0.90%) of missing values by randomly generating values with the same mean and standard deviation. The *NASDAQ stocks* dataset contains daily opening and closing prices for 2,749 stocks from Apr 19, 2010 to Apr 18, 2011 (257 days). The dataset was downloaded from `http://www.google.com/finance`. For our experiments, we computed the percentage of change between the closing and opening prices. The *world weather* dataset contains monthly measurements of temperature, precipitation, vapor, cloud cover, wet days and frost days from Jan 1990 to Dec 2002 (156 months) on a 2.5×2.5 degree grid that covers the entire world. The dataset is publicly available at `http://www.cru.uea.ac.uk/`. After sampling each 5×5 degrees, we obtained 4,146 variables. For our experiments, we computed the change between each month and the month in the previous year.

For all the datasets, we used one third of the data for training, one third for validation and the remaining third for testing. Since the brain fMRI dataset has a very small number of subjects, we performed six repetitions by making each third of the data take turns as training, validation and testing sets. In our evaluation, we included scale free networks [15]. We did not include the covariance selection method [2] since we found it is extremely slow for these high-dimensional datasets. We report the negative log-likelihood on the testing set in Fig.3 (we subtracted the entropy measured on the testing set and then scaled the results for visualization purposes). We can observe that the log-likelihood of our method is remarkably better than the other techniques for all the datasets.

Regarding comparison to group sparse methods, in our previous experiments we did not include block structure for known block-variable assignments [11, 12] since our synthetic and real-world datasets lack such assignments. We did not include block structure for unknown assignments [13, 14] given their time complexity ([14] has a $\mathcal{O}(N^5)$-time Gibbs sampler step for N variables and it is applied for $N = 60$ only, while [13] has a $\mathcal{O}(N^4)$-time ridge regression step). Instead, we evaluated our method in the *baker's yeast* gene expression dataset in [11] which contains 677 variables

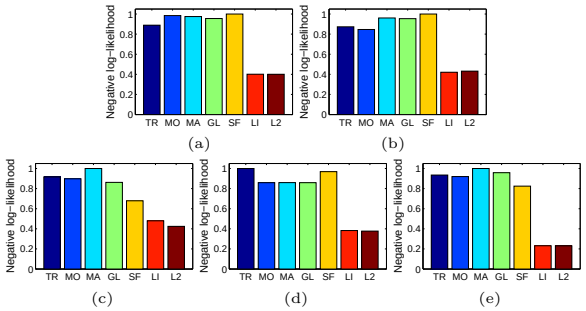

Figure 3: Test negative log-likelihood of structures learnt for (a) addicted subjects and (b) control subjects in the brain fMRI dataset, (c) gene expression, (d) NASDAQ stocks and (e) world weather. Our proposed methods $\ell_{1,2}$ (L2) and $\ell_{1,\infty}$ (LI) outperforms the Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), Tikhonov regularization (TR) and scale free networks (SF).

and 173 samples. We used the experimental settings of Fig.3 in [13]. For learning one structure, [13] took 5 hours while our $\ell_{1,2}$ method took only 50 seconds. Our method outperforms block structures for known and unknown assignments. The log-likelihood is 0 for Tikhonov regularization, 6 for [11, 13], 8 for [12], and 22 for our $\ell_{1,2}$ method.

We show the structures learnt for cocaine addicted and control subjects in Fig.4, for our $\ell_{1,2}$ method and graphical lasso [3]. The disconnected variables are not shown. Note that our structures involve remarkably fewer connected variables but yield a higher log-likelihood than graphical lasso (Fig.3), which suggests that the discarded edges from the disconnected nodes are not important for accurate modeling of this dataset. Moreover, removal of a large number of nuisance variables (voxels) results into a more interpretable model, clearly demonstrating brain areas involved in structural model differences that discriminate cocaine addicted from control subjects. Note that graphical lasso (bottom of Fig.4) connects most of the brain voxels in both populations, making them impossible to compare. Our approach produces more "localized" networks (top of the Fig.4) involving a relatively small number of brain areas: cocaine addicted subjects show increased interactions between the visual cortex (back of the brain, on the left in the image) and the prefrontal cortex (front of the brain, on the right in the image), while at the same time decreased density of interactions between the visual cortex with other brain areas (more clearly present in control subjects). The alteration in this pathway in the addict group is highly significant from a neuroscientific perspective. First, the trigger for reward was a visual stimulus. Abnormalities in the visual cortex was reported in [37] when comparing cocaine abusers to control subjects. Second, the prefrontal cortex is involved in higher-order

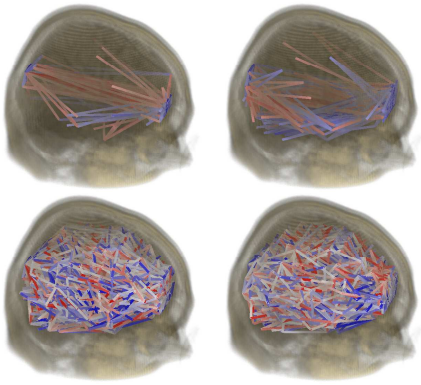

Figure 4: Structures learnt for cocaine addicted (left) and control subjects (right), for our $\ell_{1,2}$ method (top) and graphical lasso (bottom). Regularization parameter $\rho = 1/16$. Positive interactions in blue, negative interactions in red. Our structures are sparser (density 0.0016) than graphical lasso (density 0.023) where the number of edges in a complete graph is ≈378000.

cognitive functions such as decision making and reward processing. Abnormal monetary processing in the prefrontal cortex was reported in [38] when comparing cocaine addicted individuals to controls. Although a more careful interpretation of the observed results remains to be done in the near future, these results are encouraging and lend themselves to specific neuroscientific hypothesis testing.

In a different evaluation, we used generatively learnt structures for a classification task. We performed a five-fold cross-validation on the subjects. From the subjects in the training set, we learned one structure for cocaine addicted and one structure for control subjects. Then, we assigned a test subject to the structure that gave highest probability for his data. All methods in our evaluation except Tikhonov regularization obtained 84.6% accuracy. Tikhonov regularization obtained 65.4% accuracy. Therefore, our method produces structures that retain discriminability with respect to standard sparseness promoting techniques.

6 Conclusions and Future Work

In this paper, we presented variable selection in the context of learning sparse Gaussian graphical models by adding an $\ell_{1,p}$-norm regularization term, for $p \in \{2, \infty\}$. We presented a block coordinate descent method which yields sparse and positive definite estimates. We solved the original problem by efficiently solving a sequence of strictly convex (ℓ_1,ℓ_p) regularized quadratic minimization subproblems.

The motivation behind this work was to incorporate variable selection into structure learning of sparse Markov networks, and specifically Gaussian graphical models. Besides providing a better regularizer (as observed on several real-world datasets: brain fMRI, gene expression, NASDAQ stock prices and world weather), key advantages of our approach include a more accurate structure recovery in the presence of multiple noisy variables (as demonstrated by simulations), significantly better interpretability and same discriminability of the resulting network in practical applications (as shown for brain fMRI analysis).

There are several ways to extend this research. In practice, our technique converges in a small number of iterations, but an analysis of convergence rate needs to be performed. Consistency when the number of samples grows to infinity needs to be proved.

Acknowledgments

We thank Rita Goldstein for providing us the fMRI dataset. This work was supported in part by NIH Grants 1 R01 DA020949 and 1 R01 EB007530.

References

[1] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 2006.

[2] O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. *ICML*, 2006.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2007.

[4] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 2007.

[5] S. Lauritzen. *Graphical Models*. Oxford Press, 1996.

[6] A. Dempster. Covariance selection. *Biometrics*, 1972.

[7] E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2008.

[8] J. Honorio, L. Ortiz, D. Samaras, N. Paragios, and R. Goldstein. Sparse and locally constant Gaussian graphical models. *NIPS*, 2009.

[9] J. Honorio and D. Samaras. Multi-task learning of Gaussian graphical models. *ICML*, 2010.

[10] B. Zhang and Y. Wang. Learning structural changes of Gaussian graphical models in controlled experiments. *UAI*, 2010.

[11] J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse Gaussians. *UAI*, 2008.

[12] M. Schmidt, E. van den Berg, M. Friedlander, and K. Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. *AISTATS*, 2009.

[13] B. Marlin and K.Murphy. Sparse Gaussian graphical models with unknown block structure. *ICML*, 2009.

[14] B. Marlin, M. Schmidt, and K. Murphy. Group sparse priors for covariance estimation. *UAI*, 2009.

[15] Q. Liu and A. Ihler. Learning scale free networks by reweighted ℓ_1 regularization. *AISTATS*, 2011.

[16] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 1996.

[17] A. Chan, N. Vasconcelos, and G. Lanckriet. Direct convex relaxations of sparse SVM. *ICML*, 2007.

[18] S. Lee, H. Lee, P. Abbeel, and A. Ng. Efficient ℓ_1 regularized logistic regression. *AAAI*, 2006.

[19] J. Duchi and Y. Singer. Boosting with structural sparsity. *ICML*, 2009.

[20] R. Parr, L. Li, G. Taylor, C. Painter-Wakefield, and M. Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. *ICML*, 2008.

[21] S. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using ℓ_1-regularization. *NIPS*, 2006.

[22] M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning graphical model structure using ℓ_1-regularization paths. *AAAI*, 2007.

[23] M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. *CVPR*, 2008.

[24] M. Wainwright, P. Ravikumar, and J. Lafferty. High dimensional graphical model selection using ℓ_1-regularized logistic regression. *NIPS*, 2006.

[25] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 2006.

[26] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society*, 2008.

[27] G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 2010.

[28] B. Turlach, W. Venables, and S. Wright. Simultaneous variable selection. *Technometrics*, 2005.

[29] J. Tropp. Algorithms for simultaneous sparse approximation, part II: convex relaxation. *Signal Processing*, 2006.

[30] J. Friedman, T. Hastie, and R. Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. *Technical Report, Stanford University*, 2010.

[31] J. Duchi and Y. Singer. Efficient learning using forward-backward splitting. *NIPS*, 2009.

[32] A. Quattoni, X. Carreras, M. Collins, and T. Darrell. An efficient projection for $\ell_{1,\infty}$ regularization. *ICML*, 2009.

[33] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing. Smoothing proximal gradient method for general structured sparse learning. *UAI*, 2011.

[34] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. *NIPS*, 2010.

[35] R. Goldstein, D. Tomasi, N. Alia-Klein, L. Zhang, F. Telang, and N. Volkow. The effect of practice on a sustained attention task in cocaine abusers. *NeuroImage*, 2007.

[36] G. Natsoulis, L. El Ghaoui, G. Lanckriet, A. Tolley, F. Leroy, S. Dunlea, B. Eynon, C. Pearson, S. Tugendreich, and K. Jarnagin. Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Research*.

[37] J. Lee, F. Telang, C. Springer, and N. Volkow. Abnormal brain activation to visual stimulation in cocaine abusers. *Life Sciences*, 2003.

[38] R. Goldstein, N. Alia-Klein, D. Tomasi, J. Honorio, T. Maloney, P. Woicik, R. Wang, F. Telang, and N. Volkow. Anterior cingulate cortex hypoactivations to an emotionally salient task in cocaine addiction. *Proceedings of the National Academy of Sciences, USA*, 2009.