Multi-Task Learning of Gaussian Graphical Models

Jean Honorio JHONORIO@CS.SUNYSB.EDU
Dimitris Samaras SAMARAS@CS.SUNYSB.EDU
Stony Brook University, Stony Brook, NY 11794, USA

Abstract

We present *multi-task structure learning* for Gaussian graphical models. We discuss uniqueness and boundedness of the optimal solution of the maximization problem. A block coordinate descent method leads to a provably convergent algorithm that generates a sequence of positive definite solutions. Thus, we reduce the original problem into a sequence of strictly convex ℓ_∞ regularized quadratic minimization subproblems. We further show that this subproblem leads to the *continuous quadratic knapsack problem*, for which very efficient methods exist. Finally, we show promising results in a dataset that captures brain function of cocaine addicted and control subjects under conditions of monetary reward.

1. Introduction

Structure learning aims to discover the topology of a probabilistic network of variables such that this network represents accurately a given dataset while maintaining low complexity. Accuracy of representation is measured by the likelihood that the model explains the observed data, while complexity of a graphical model is measured by its number of parameters. Structure learning faces several challenges: the number of possible structures is super-exponential in the number of variables while the required sample size might be even exponential. Therefore, finding good regularization techniques is very important in order to avoid overfitting and to achieve a better generalization performance.

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 (Meinshausen & Bühlmann, 2006) relied on a sequence of sparse regressions. Maximum likelihood estimation with an ℓ_1-norm penalty for encouraging sparseness is proposed in (Banerjee et al., 2006; Friedman et al., 2007; Yuan & Lin, 2007).

Suppose that we want to learn the structure of brain region interactions for one person. We can expect that the interaction patterns of two persons are not same. On the other hand, when learning the structure for one person, we would like to use evidence from other persons as a side information in our learning process. This becomes more important in settings with limited amount of data, such as in functional magnetic resonance image (fMRI) studies. Multi-task learning allows for a more efficient use of training data which is available for multiple related tasks.

In this paper, we consider the computational aspect of multi-task structure learning, which generalizes the learning of sparse Gaussian graphical models to the multi-task setting by replacing the ℓ_1-norm regularization with an $\ell_{1,\infty}$-norm.

Our contribution in this paper is three-fold. First, we present a block coordinate descent method which is provably convergent and yields sparse and positive definite estimates. Second, we show the connection between our multi-task structure learning problem and the continuous quadratic knapsack problem, which allows us to use existing efficient methods (Helgason et al., 1980; Brucker, 1984; Kiwiel, 2007). Finally, we experimentally show that the cross-validated log-likelihood of our method is more stable and statistically significantly higher than the competing methods in a fMRI dataset that captures brain function of cocaine addicted and control subjects under conditions of monetary reward.

Section 2 introduces Gaussian graphical models as well as techniques for learning such structures from data.

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

Table 1. Notation used in this paper.

Notation	Description
$\lVert\mathbf{c}\rVert_1$	ℓ_1-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\sum_n \lvert c_n \rvert$
$\lVert\mathbf{c}\rVert_\infty$	ℓ_∞-norm of $\mathbf{c} \in \mathbb{R}^N$, i.e. $\max_n \lvert c_n \rvert$
$\mathbf{diag}(\mathbf{c}) \in \mathbb{R}^{N \times N}$	matrix with elements of $\mathbf{c} \in \mathbb{R}^N$ on its diagonal
$\mathbf{A} \succeq \mathbf{0}$	$\mathbf{A} \in \mathbb{R}^{N \times N}$ is symmetric and positive semidefinite
$\mathbf{A} \succ \mathbf{0}$	$\mathbf{A} \in \mathbb{R}^{N \times N}$ is symmetric and positive definite
$\lVert\mathbf{A}\rVert_1$	ℓ_1-norm of $\mathbf{A} \in \mathbb{R}^{M \times N}$, i.e. $\sum_{mn} \lvert a_{mn} \rvert$
$\lVert\mathbf{A}\rVert_\infty$	ℓ_∞-norm of $\mathbf{A} \in \mathbb{R}^{M \times N}$, i.e. $\max_{mn} \lvert a_{mn} \rvert$
$\lVert\mathbf{A}\rVert_2$	spectral norm of $\mathbf{A} \in \mathbb{R}^{N \times N}$, i.e. the maximum eigenvalue of $\mathbf{A} \succ \mathbf{0}$
$\lVert\mathbf{A}\rVert_{\mathfrak{F}}$	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. $\sum_{mn} a_{mn} b_{mn}$

Section 3 sets up the problem and discusses some of its properties. Section 4 describes our block coordinate descent method. Section 5 shows the connection to the continuous quadratic knapsack problem. Experimental results are shown and explained in Section 6. Main contributions and results are summarized in Section 7.

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 $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ (Lauritzen, 1996). 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 concept of robust estimation by performing covariance selection was first introduced in (Dempster, 1972) where the number of parameters to be estimated is reduced by setting some elements of the precision matrix $\mathbf{\Omega}$ to zero. Since finding the most sparse precision matrix which fits a dataset is a NP-hard problem (Banerjee et al., 2006), in order to overcome it, several ℓ_1-regularization methods have been proposed for learning Gaussian graphical models from data.

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:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \left(\ell_{\widehat{\mathbf{\Sigma}}}(\mathbf{\Omega}) - \rho \lVert \mathbf{\Omega} \rVert_1 \right) \tag{1}$$

for $\rho > 0$. The term $\lVert \mathbf{\Omega} \rVert_1$ encourages sparseness of the precision matrix or conditional independence among variables, while the term $\ell_{\widehat{\mathbf{\Sigma}}}(\mathbf{\Omega})$ is the Gaussian log-likelihood, and it is defined as:

$$\ell_{\widehat{\mathbf{\Sigma}}}(\mathbf{\Omega}) = \log\det \mathbf{\Omega} - \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Omega} \rangle \tag{2}$$

Several optimization techniques have been proposed for eq.(1): a sequence of box-constrained quadratic programs in the *covariance selection* (Banerjee et al., 2006), solution of the dual problem by sparse regression in the *graphical lasso* (Friedman et al., 2007) or an approximation via standard determinant maximization with linear inequality constraints in (Yuan & Lin, 2007). Instead of solving eq.(1), the *Meinshausen-Bühlmann approximation* (Meinshausen & Bühlmann, 2006) obtains the conditional dependencies by performing a sparse linear regression for each variable, by using *lasso* regression (Tibshirani, 1996).

Besides sparseness, several regularizers have been proposed for Gaussian graphical models for *single-task* learning, for enforcing diagonal structure (Levina et al., 2008), block structure for known block-variable assignments (Duchi et al., 2008a) and unknown block-variable assignments (Marlin & K.Murphy, 2009; Marlin et al., 2009), or spatial coherence (Honorio et al., 2009).

Multi-task learning has been applied to very diverse problems, such as linear regression (Liu et al., 2009), classification (Jebara, 2004), compressive sensing (Qi et al., 2008), reinforcement learning (Wilson et al., 2007) and structure learning of Bayesian networks (Niculescu-Mizil & Caruana, 2007).

3. Preliminaries

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

3.1. Problem Setup

We propose a prior that is motivated from the multi-task learning literature. Given K arbitrary tasks, our goal is to learn one structure for each task, and to promote a consistent sparseness pattern across tasks.

For a given task k, we learn a precision matrix $\mathbf{\Omega}^{(k)} \in \mathbb{R}^{N \times N}$ for N variables. Our multi-task regularizer penalizes corresponding edges across tasks (i.e. $\omega_{n_1 n_2}^{(1)}, \dots, \omega_{n_1 n_2}^{(K)}$) with the same strength whether it appears in $1, 2, \dots$ or K tasks. As a result, only edges that help to explain the observed data for almost every task, will appear in the learnt structures.

Let $\widehat{\mathbf{\Sigma}}^{(k)} \succeq \mathbf{0}$ be the dense sample covariance matrix for task k, and $T^{(k)} > 0$ be the number of samples in task k. The *multi-task structure learning problem* is defined as:

$$\max_{(\forall k)\mathbf{\Omega}^{(k)} \succ \mathbf{0}} \left(\sum_k T^{(k)} \ell_{\widehat{\mathbf{\Sigma}}^{(k)}}(\mathbf{\Omega}^{(k)}) - \rho \|\mathbf{\Omega}\|_{1,\infty} \right) \qquad (3)$$

for $\rho > 0$. The term $\ell_{\widehat{\mathbf{\Sigma}}^{(k)}}(\mathbf{\Omega}^{(k)})$ is the Gaussian log-likelihood defined in eq.(2), while the term $\|\mathbf{\Omega}\|_{1,\infty}$ is our multi-task regularizer, and it is defined as:

$$\|\mathbf{\Omega}\|_{1,\infty} = \sum_{n_1 n_2} \max_k |\omega^{(k)}_{n_1 n_2}| \qquad (4)$$

The number of samples is a term that is usually dropped for covariance selection and graphical lasso as in eq.(1). For the multi-task structure learning problem, it is important to keep this term when adding the log-likelihood of several tasks into a single objective function.

3.2. Bounds

In what follows, we discuss uniqueness and boundedness of the optimal solution of the multi-task structure learning problem.

Lemma 1. *For $\rho > 0$, the multi-task structure learning problem in eq.(3) is a maximization problem with concave (but not strictly concave) objective function and convex constraints.*

Proof. The Gaussian log-likelihood defined in eq.(2) is concave, since $\log\det$ is concave on the space of symmetric positive definite matrices and $\langle\cdot,\cdot\rangle$ is a linear operator. The multi-task regularizer defined in eq.(4) is a non-smooth convex function. Finally, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$ is a convex constraint. □

Theorem 2. *For $\rho > 0$, the optimal solution to the multi-task structure learning problem in eq.(3) is unique and bounded as follows:*

$$\left(\frac{1}{\|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{N\rho}{T^{(k)}}} \right) \mathbf{I} \preceq {\mathbf{\Omega}^{(k)}}^* \preceq \left(\frac{NK}{\rho} \right) \mathbf{I} \qquad (5)$$

Proof. By using the identity $\rho\|\mathbf{c}\|_\infty = \max_{\|\mathbf{a}\|_1 \le \rho} \mathbf{a}^{\mathrm{T}}\mathbf{c}$ in eq.(3), we get:

$$\max_{(\forall k)\mathbf{\Omega}^{(k)} \succ \mathbf{0}} \; \min_{\substack{(\forall n_1 n_2) \\ \|\mathbf{a}_{n_1 n_2}\|_1 \le \rho}} \sum_k T^{(k)} \begin{pmatrix} \log\det \mathbf{\Omega}^{(k)} \\ -\langle \widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}, \mathbf{\Omega}^{(k)} \rangle \end{pmatrix} \qquad (6)$$

where $\mathbf{a}_{n_1 n_2} = (a^{(1)}_{n_1 n_2}, \dots, a^{(K)}_{n_1 n_2})^{\mathrm{T}}$ and $\mathbf{A}^{(k)} \in \mathbb{R}^{N \times N}$. 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 independent for each k and is given by $\mathbf{\Omega}^{(k)} = (\widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}})^{-1}$. By replacing this solution in eq.(6), we get the dual problem of eq.(3):

$$\min_{\substack{(\forall n_1 n_2) \\ \|\mathbf{a}_{n_1 n_2}\|_1 \le \rho}} -\sum_k T^{(k)} \log\det \left(\widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}} \right) - NK \qquad (7)$$

In order to find a lowerbound for the minimum eigenvalue of ${\mathbf{\Omega}^{(k)}}^*$, note that $\|{\mathbf{\Omega}^{(k)}}^{*-1}\|_2 = \|\widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}\|_2 \le \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \|\frac{\mathbf{A}^{(k)}}{T^{(k)}}\|_2 = \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{1}{T^{(k)}}\|\mathbf{A}^{(k)}\|_2 \le \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{N}{T^{(k)}}\|\mathbf{A}^{(k)}\|_\infty$. Since $\|\mathbf{a}_{n_1 n_2}\|_1 \le \rho$, it follows that $|a^{(k)}_{n_1 n_2}| = \rho$ in the extreme case in which $(\forall k_1 \ne k) a^{(k_1)}_{n_1 n_2} = 0$, and therefore $\|\mathbf{A}^{(k)}\|_\infty \le \rho$.

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

$$-NK + \sum_k T^{(k)} \langle \widehat{\mathbf{\Sigma}}^{(k)}, {\mathbf{\Omega}^{(k)}}^* \rangle + \rho \|\mathbf{\Omega}^*\|_{1,\infty} = 0 \qquad (8)$$

The upperbound is found as follows: $\|{\mathbf{\Omega}^{(k)}}^*\|_2 \le \|{\mathbf{\Omega}^{(k)}}^*\|_{\mathfrak{F}} \le \|{\mathbf{\Omega}^{(k)}}^*\|_1 \le \|\mathbf{\Omega}^*\|_{1,\infty} = (NK - \sum_k T^{(k)} \langle \widehat{\mathbf{\Sigma}}^{(k)}, {\mathbf{\Omega}^{(k)}}^* \rangle)/\rho$ and since $\mathbf{\Sigma}^{(k)} \succeq \mathbf{0}$ and ${\mathbf{\Omega}^{(k)}}^* \succ \mathbf{0}$, it follows that $\langle \widehat{\mathbf{\Sigma}}^{(k)}, {\mathbf{\Omega}^{(k)}}^* \rangle \ge 0$. □

4. Block Coordinate Descent Method

In this section, we develop a block coordinate descent method for our multi-task structure learning problem, and discuss some of its properties.

We apply block coordinate descent method on the primal problem, unlike covariance selection (Banerjee et al., 2006) and graphical lasso (Friedman et al., 2007) which optimize the dual. Our choice of optimizing the primal follows from the fact that the dual formulation in eq.(7) leads to a sum of K terms ($\log\det$ functions) which cannot be simplified to a quadratic problem unless $K = 1$.

For clarity of exposition, we assume that the diagonals of $\mathbf{\Omega}^{(1)}, \dots, \mathbf{\Omega}^{(K)}$ are not penalized by our multi-task regularizer defined in eq.(4). In case they are penalized, an additional *continuous logarithmic knapsack problem* needs to be solved. We point out that all the following theorems and lemmas still hold in this case.

Lemma 3. *The solution sequence generated by the block coordinate descent method is bounded and every* *cluster point is a solution of the multi-task structure learning problem in eq.*(3)*.*

Proof. The non-smooth regularizer $\|\mathbf{\Omega}\|_{1,\infty}$ is separable into a sum of $O(N^2)$ individual functions of the form $\max_k |\omega^{(k)}_{n_1 n_2}|$. These functions are defined over blocks of K variables, i.e. $\omega^{(1)}_{n_1 n_2}, \dots, \omega^{(K)}_{n_1 n_2}$. The objective function in eq.(3) is continuous on a compact level set. By virtue of Theorem 4.1 in (Tseng, 2001), we prove our claim. □

Theorem 4. *The block coordinate descent method for the multi-task structure learning problem in eq.*(3) *generates a sequence of positive definite solutions.*

Proof. Maximization can be performed with respect to one row and column of all precision matrices $\mathbf{\Omega}^{(k)}$ at a time. Without loss of generality, we use the last row and column in our derivation, since permutation of rows and columns is always possible. Let:

$$\mathbf{\Omega}^{(k)} = \begin{bmatrix} \mathbf{W}^{(k)} & \mathbf{y}^{(k)} \\ \mathbf{y}^{(k)\mathrm{T}} & z^{(k)} \end{bmatrix}, \widehat{\mathbf{\Sigma}}^{(k)} = \begin{bmatrix} \mathbf{S}^{(k)} & \mathbf{u}^{(k)} \\ \mathbf{u}^{(k)\mathrm{T}} & v^{(k)} \end{bmatrix} \quad (9)$$

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

In terms of the variables $\mathbf{y}^{(k)}, z^{(k)}$ and the constant matrix $\mathbf{W}^{(k)}$, the multi-task structure learning problem in eq.(3) can be reformulated as:

$$\max_{(\forall k)\mathbf{\Omega}^{(k)}\succ\mathbf{0}} \begin{pmatrix} \sum_k T^{(k)} \begin{pmatrix} \log(z^{(k)} - \mathbf{y}^{(k)^\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)}) \\ -2\mathbf{u}^{(k)^\mathrm{T}}\mathbf{y}^{(k)} - v^{(k)}z^{(k)} \end{pmatrix} \\ -2\rho\sum_n \max_k |y^{(k)}_n| \end{pmatrix} \quad (10)$$

If $\mathbf{\Omega}^{(k)}$ is a symmetric matrix, according to the Haynsworth inertia formula, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$ if and only if its Schur complement $z^{(k)} - \mathbf{y}^{(k)^\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)} > 0$ and $\mathbf{W}^{(k)} \succ \mathbf{0}$. By maximizing eq.(10) with respect to $z^{(k)}$, we get:

$$z^{(k)} - \mathbf{y}^{(k)^\mathrm{T}}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)} = \frac{1}{v^{(k)}} \quad (11)$$

and since $v^{(k)} > 0$, this implies that the Schur complement in eq.(11) is positive.

Finally, in an iterative optimization algorithm, it suffices to initialize $\mathbf{\Omega}^{(k)}$ to a matrix that is known to be positive definite, e.g. a diagonal matrix with positive elements. □

Theorem 5. *The block coordinate descent method for the multi-task structure learning problem in eq.*(3) *is equivalent to solving a sequence of strictly convex* $\ell_{1,\infty}$ *regularized quadratic subproblems:*

$$\min_{(\forall k)\mathbf{y}^{(k)}\in\mathbb{R}^{N-1}} \begin{pmatrix} \sum_k T^{(k)} \begin{pmatrix} \frac{1}{2}\mathbf{y}^{(k)^\mathrm{T}} v^{(k)}\mathbf{W}^{(k)^{-1}}\mathbf{y}^{(k)} \\ +\mathbf{u}^{(k)^\mathrm{T}}\mathbf{y}^{(k)} \end{pmatrix} \\ +\rho\sum_n \max_k |y^{(k)}_n| \end{pmatrix} \quad (12)$$

Proof. By replacing the optimal $z^{(k)}$ given by eq.(11) into the objective function in eq.(10), we get eq.(12). Since $\mathbf{W}^{(k)} \succ \mathbf{0} \Rightarrow \mathbf{W}^{(k)^{-1}} \succ \mathbf{0}$, hence eq.(12) is strictly convex. □

Lemma 6. *If* $\max_n \sum_k T^{(k)}|u^{(k)}_n| \le \rho$*, the* $\ell_{1,\infty}$ *regularized quadratic problem in eq.*(12) *has the minimizer* $(\forall k)\mathbf{y}^{(k)^*} = \mathbf{0}$.

Proof. The problem in eq.(12) has the minimizer $(\forall k)\mathbf{y}^{(k)^*} = \mathbf{0}$ if and only if $\mathbf{0}$ belongs to the subdifferential set of the non-smooth objective function at $(\forall k)\mathbf{y}^{(k)} = \mathbf{0}$, i.e. $(\exists \mathbf{A} \in \mathbb{R}^{N-1\times K})(T^{(1)}\mathbf{u}^{(1)}, \dots, T^{(K)}\mathbf{u}^{(K)}) + \mathbf{A} = \mathbf{0} \wedge \max_n \sum_k |a_{nk}| \le \rho$. This condition is true for $\max_n \sum_k |T^{(k)}u^{(k)}_n| \le \rho$ and since $(\forall k)T^{(k)} > 0$, we prove our claim. □

Remark 7. *By using Lemma 6, 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 8. *The coordinate descent method for the* $\ell_{1,\infty}$ *regularized quadratic problem in eq.*(12) *is equivalent to solving a sequence of strictly convex* ℓ_∞ *regularized separable quadratic subproblems:*

$$\min_{\mathbf{x}\in\mathbb{R}^K} \left(\frac{1}{2}\mathbf{x}^\mathrm{T}\mathbf{diag}(\mathbf{q})\mathbf{x} - \mathbf{c}^\mathrm{T}\mathbf{x} + \rho\,\|\mathbf{x}\|_\infty\right) \quad (13)$$

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:

$$\mathbf{W}^{(k)^{-1}} = \begin{bmatrix} \mathbf{H}^{(k)}_{11} & \mathbf{h}^{(k)}_{12} \\ \mathbf{h}^{(k)^\mathrm{T}}_{12} & h^{(k)}_{22} \end{bmatrix}, \mathbf{y}^{(k)} = \begin{bmatrix} \mathbf{y}^{(k)}_1 \\ x_k \end{bmatrix}, \mathbf{u}^{(k)} = \begin{bmatrix} \mathbf{u}^{(k)}_1 \\ u^{(k)}_2 \end{bmatrix} \quad (14)$$

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

In terms of the variable $\mathbf{x}$ and the constants $q_k = T^{(k)}v^{(k)}h^{(k)}_{22}$, $c_k = -T^{(k)}(v^{(k)}\mathbf{h}^{(k)^\mathrm{T}}_{12}\mathbf{y}^{(k)}_1 + u^{(k)}_2)$, the $\ell_{1,\infty}$ regularized quadratic problem in eq.(12) can be reformulated as in eq.(13). Moreover, since $(\forall k)T^{(k)} > 0 \wedge v^{(k)} > 0 \wedge h^{(k)}_{22} > 0 \Rightarrow \mathbf{q} > \mathbf{0}$, and therefore eq.(13) is strictly convex. □

5. Continuous Quadratic Knapsack Problem

In this section, we show the connection between the multi-task structure learning problem and the continuous quadratic knapsack problem, for which very efficient methods exist.

The continuous quadratic knapsack problem has been solved in several areas. (Helgason et al., 1980) provides an $O(K \log K)$ algorithm which initially sort the breakpoints. (Brucker, 1984) and later (Kiwiel, 2007) provide deterministic linear-time algorithms by using medians of breakpoint subsets. In the context of machine learning, (Duchi et al., 2008b) provides a randomized linear-time algorithm, while (Liu et al., 2009) provides an $O(K \log K)$ algorithm. We point out that (Duchi et al., 2008b; Liu et al., 2009) assume that the weights of the quadratic term are all equal, i.e. $(\forall k) q_k = 1$. In this paper, we assume arbitrary positive weights, i.e. $(\forall k) q_k > 0$.

We point out to the reader, that the variables $\mathbf{y}, \mathbf{z}$ used in this section have a different meaning with respect to the previous sections. We prefer to use them, since those are variables regularly used as unknowns.

Theorem 9. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, the ℓ_∞ regularized separable quadratic problem in eq.(13) is equivalent to the separable quadratic problem with one ℓ_1 constraint:*

$$\min_{\|\mathbf{y}\|_1 \leq \rho} \left(\frac{1}{2} (\mathbf{y} - \mathbf{c})^{\mathrm{T}} \mathbf{diag}(\mathbf{q})^{-1} (\mathbf{y} - \mathbf{c}) \right) \tag{15}$$

Furthermore, their optimal solutions are related by $\mathbf{x}^* = \mathbf{diag}(\mathbf{q})^{-1}(\mathbf{c} - \mathbf{y}^*)$.

Proof. By Lagrangian duality, the problem in eq.(15) is the dual of the problem in eq.(13). Furthermore, strong duality holds in this case. □

Remark 10. *In eq.(15), we can assume that $(\forall k) c_k \neq 0$. If $(\exists k) c_k = 0$, the partial optimal solution is $y_k^* = 0$, and since this assignment does not affect the constraint, we can safely remove y_k from the optimization problem.*

Remark 11. *In what follows, we assume that $\|\mathbf{c}\|_1 > \rho$. If $\|\mathbf{c}\|_1 \leq \rho$, the unconstrained optimal solution of eq.(15) is also its optimal solution, since $\mathbf{y}^* = \mathbf{c}$ is inside the feasible region given that $\|\mathbf{y}^*\|_1 \leq \rho$.*

Lemma 12. *For $\mathbf{q} > \mathbf{0}$, $(\forall k) c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the optimal solution $\mathbf{y}^*$ of the separable quadratic problem with one ℓ_1 constraint in eq.(15) belongs to the same orthant as the unconstrained optimal solution $\mathbf{c}$, i.e. $(\forall k) y_k^* c_k \geq 0$.*

Proof. We prove this by contradiction. Assume $(\exists k_1) y_{k_1}^* c_{k_1} < 0$. Let $\mathbf{y}$ be a vector such that $y_{k_1} = 0$ and $(\forall k_2 \neq k_1) y_{k_2} = y_{k_2}^*$. The solution $\mathbf{y}$ is feasible, since $\|\mathbf{y}^*\|_1 \leq \rho$ and $\|\mathbf{y}\|_1 = \|\mathbf{y}^*\|_1 - |y_{k_1}^*| \leq \rho$. The difference in the objective function between $\mathbf{y}^*$ and $\mathbf{y}$ is $\frac{1}{2}(\mathbf{y}^* - \mathbf{c})^{\mathrm{T}} \mathbf{diag}(\mathbf{q})^{-1} (\mathbf{y}^* - \mathbf{c}) - \frac{1}{2}(\mathbf{y} - \mathbf{c})^{\mathrm{T}} \mathbf{diag}(\mathbf{q})^{-1} (\mathbf{y} - \mathbf{c}) = \frac{1}{2q_{k_1}} ({y_{k_1}^*}^2 - 2c_{k_1} y_{k_1}^*) > \frac{{y_{k_1}^*}^2}{2q_{k_1}} > 0$. Thus, the objective function for $\mathbf{y}$ is smaller than for $\mathbf{y}^*$ (the assumed optimal solution), which is a contradiction. □

Theorem 13. *For $\mathbf{q} > \mathbf{0}$, $(\forall k) c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the separable quadratic problem with one ℓ_1 constraint in eq.(15) is equivalent to the continuous quadratic knapsack problem:*

$$\min_{\substack{\mathbf{z} \geq \mathbf{0} \\ \mathbf{1}^{\mathrm{T}}\mathbf{z} = \rho}} \sum_k \frac{1}{2q_k} (z_k - |c_k|)^2 \tag{16}$$

Furthermore, their optimal solutions are related by $(\forall k) y_k^* = \mathrm{sgn}(c_k) z_k^*$.

Proof. By invoking Lemma 12, we can replace $(\forall k) y_k = \mathrm{sgn}(c_k) z_k$, $z_k \geq 0$ in eq.(15). Finally, we change the inequality constraint $\mathbf{1}^{\mathrm{T}}\mathbf{z} \leq \rho$ to an equality constraint since $\|\mathbf{c}\|_1 > \rho$ and therefore, the optimal solution must be on the boundary of the constraint set. □

Lemma 14. *For $\mathbf{q} > \mathbf{0}$, $(\forall k) c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the continuous quadratic knapsack problem in eq.(16) has the solution $z_k(\nu) = \max(0, |c_k| - \nu q_k)$ for some ν, and furthermore:*

$$\mathbf{z}^* = \mathbf{z}(\nu) \Leftrightarrow \mathbf{1}^{\mathrm{T}}\mathbf{z}(\nu) = \rho \tag{17}$$

Proof. The Lagrangian of eq.(16) is:

$$\min_{\mathbf{z} \geq \mathbf{0}} \sum_k \frac{1}{2q_k} (z_k - |c_k|)^2 + \nu(\mathbf{1}^{\mathrm{T}}\mathbf{z} - \rho) \tag{18}$$

Both results can be obtained by invoking the Karush-Kuhn-Tucker optimality conditions on eq.(18). □

Remark 15. *Note that $z_k(\nu) = \max(0, |c_k| - \nu q_k)$ is a decreasing piecewise linear function with breakpoint $\nu = \frac{|c_k|}{q_k} > 0$. By Lemma 14, finding the optimal $\mathbf{z}^*$ is equivalent to finding ν in a piecewise linear function $\mathbf{1}^{\mathrm{T}}\mathbf{z}(\nu)$ that produces ρ.*

Lemma 16. *For $\mathbf{q} > \mathbf{0}$, $(\forall k) c_k \neq 0$, $\|\mathbf{c}\|_1 > \rho$, the continuous quadratic knapsack problem in eq.(16) has the optimal solution $z_k^* = \max(0, |c_k| - \nu^* q_k)$ for:*

$$\frac{|c_{\pi_{k^*}}|}{q_{\pi_{k^*}}} \geq \nu^* = \frac{\sum_{k=1}^{k^*} |c_{\pi_k}| - \rho}{\sum_{k=1}^{k^*} q_{\pi_k}} \geq \frac{|c_{\pi_{k^*+1}}|}{q_{\pi_{k^*+1}}} \tag{19}$$

where the breakpoints are sorted in decreasing order by a permutation π of the indices $1, 2, \ldots, K$, i.e. $\frac{|c_{\pi_1}|}{q_{\pi_1}} \geq \frac{|c_{\pi_2}|}{q_{\pi_2}} \geq \cdots \geq \frac{|c_{\pi_K}|}{q_{\pi_K}} \geq \frac{|c_{\pi_{K+1}}|}{q_{\pi_{K+1}}} \equiv 0$.

Proof. Given k^*, ν^* can be found straightforwardly by using the equation of the line. In order to find k^*, note that we want to find the range in which $\mathbf{1}^{\mathrm{T}}\mathbf{z}\left(\frac{|c_{\pi_{k^*}}|}{q_{\pi_{k^*}}}\right) \leq \rho \leq \mathbf{1}^{\mathrm{T}}\mathbf{z}\left(\frac{|c_{\pi_{k^*+1}}|}{q_{\pi_{k^*+1}}}\right)$. □

Theorem 17. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, the ℓ_∞ regularized separable quadratic problem in eq.(13) has the optimal solution:*

$$\begin{array}{l} \|\mathbf{c}\|_1 \leq \rho \Rightarrow \mathbf{x}^* = \mathbf{0} \\ \|\mathbf{c}\|_1 > \rho \wedge k > k^* \Rightarrow x^*_{\pi_k} = \frac{c_{\pi_k}}{q_{\pi_k}} \\ \|\mathbf{c}\|_1 > \rho \wedge k \leq k^* \Rightarrow x^*_{\pi_k} = \mathrm{sgn}(c_{\pi_k}) \frac{\sum_{k=1}^{k^*} |c_{\pi_k}| - \rho}{\sum_{k=1}^{k^*} q_{\pi_k}} \end{array} \quad (20)$$

Proof. For $\|\mathbf{c}\|_1 \leq \rho$, from Remark 11 we know that $\mathbf{y}^* = \mathbf{c}$. By Theorem 9, the optimal solution of eq.(13) is $\mathbf{x}^* = \mathbf{diag}(\mathbf{q})^{-1}(\mathbf{c} - \mathbf{y}^*) = \mathbf{0}$, and we prove the first claim.

For $\|\mathbf{c}\|_1 > \rho$, by Theorem 9, the optimal solution of eq.(13) $x^*_{\pi_k} = \frac{1}{q_{\pi_k}}(c_{\pi_k} - y^*_{\pi_k})$. By Theorem 13, $x^*_{\pi_k} = \frac{1}{q_{\pi_k}}(c_{\pi_k} - \mathrm{sgn}(c_{\pi_k}) z^*_{\pi_k})$. By Lemma 16, $x^*_{\pi_k} = \frac{c_{\pi_k}}{q_{\pi_k}} - \mathrm{sgn}(c_{\pi_k}) \max(0, \frac{|c_{\pi_k}|}{q_{\pi_k}} - \nu^*)$.

If $k > k^* \Rightarrow \frac{|c_{\pi_k}|}{q_{\pi_k}} < \nu^* \Rightarrow x^*_{\pi_k} = \frac{c_{\pi_k}}{q_{\pi_k}}$, and we prove the second claim.

If $k \leq k^* \Rightarrow \frac{|c_{\pi_k}|}{q_{\pi_k}} \geq \nu^* \Rightarrow x^*_{\pi_k} = \mathrm{sgn}(c_{\pi_k}) \nu^*$, and we prove the third claim. □

Algorithm 1 shows the block coordinate descent method in detail. A careful implementation of the algorithm allows obtaining a time complexity of $O(LN^3K)$ for L iterations, N variables and K tasks. In our experiments, the algorithm converges quickly in usually $L = 10$ iterations. The polynomial dependence $O(N^3)$ on the number of variables is expected since we cannot produce an algorithm faster than computing the inverse of the sample covariance in the case of an infinite sample. The linear-time dependence $O(K)$ on the number of tasks can be accomplished by using a deterministic linear-time method for solving the continuous quadratic knapsack problem, based on medians of breakpoint subsets (Kiwiel, 2007). A very easy-to-implement $O(K \log K)$ algorithm is obtained by initially sorting the breakpoints and searching the range for which Lemma 16 holds.

Algorithm 1 Block Coordinate Descent

```
Input: $\rho > 0$, for each $k$, $\widehat{\mathbf{\Sigma}}^{(k)} \succeq \mathbf{0}$, $T^{(k)} > 0$
Initialize for each $k$, $\mathbf{\Omega}^{(k)} = \mathbf{diag}(\widehat{\mathbf{\Sigma}}^{(k)})^{-1}$
for each iteration $1, \ldots, L$ and each variable $1, \ldots, N$ do
  Split for each $k$, $\mathbf{\Omega}^{(k)}$ into $\mathbf{W}^{(k)}, \mathbf{y}^{(k)}, z^{(k)}$ and $\widehat{\mathbf{\Sigma}}^{(k)}$ into $\mathbf{S}^{(k)}, \mathbf{u}^{(k)}, v^{(k)}$ as described in eq.(9)
  Update for each $k$, ${\mathbf{W}^{(k)}}^{-1}$ 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 $\mathbf{W}^{(k)}$)
  for each variable $1, \ldots, N-1$ do
    Split for each $k$, ${\mathbf{W}^{(k)}}^{-1}, \mathbf{y}^{(k)}, \mathbf{u}^{(k)}$ as in eq.(14)
    Solve the $\ell_\infty$ regularized separable quadratic problem by eq.(20), either by sorting the breakpoints or using medians of breakpoint subsets
  end for
  Update for each $k$, $z^{(k)} \leftarrow \frac{1}{v^{(k)}} + {\mathbf{y}^{(k)}}^{\mathrm{T}} {\mathbf{W}^{(k)}}^{-1} \mathbf{y}^{(k)}$
end for
Output: for each $k$, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$
```

6. Experimental Results

For experimental validation, we used a fMRI dataset that captures brain function of cocaine addicted and control subjects under conditions of monetary reward. The dataset collected by (Goldstein et al., 2007) contains 28 subjects: 16 cocaine addicted and 12 control. Six sessions were acquired for each subject. Each session contains 87 scans taken every 3.5 seconds.

Registration of the dataset to the same spatial reference template (Talairach space) and spatial smoothing was performed in SPM2[1]. We extracted voxels from the gray matter only, and grouped them into 157 regions by using standard labels, given by the Talairach Daemon[2]. These regions span the entire brain (cerebellum, cerebrum and brainstem). In order to capture laterality effects, we have regions for the left and right side of the brain.

First, we test the idea of learning one Gaussian graphical model for each of the six sessions, i.e. each session is a task. We performed five-fold cross-validation on the subjects, and report the log-likelihood on the testing set (scaled for visualization purposes). In Figure 1, we can observe that the log-likelihood of our method is higher than the competing methods.

Second, we test the idea of learning one Gaussian graphical model for each subject, i.e. each subject is a task. It is well known that fMRI datasets have more

[1] `http://www.fil.ion.ucl.ac.uk/spm/`

[2] `http://www.talairach.org/`

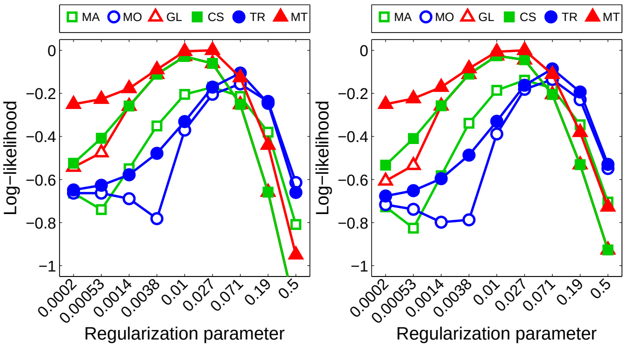

Figure 1. Cross-validated log-likelihood of structures learnt for each of the six sessions on cocaine addicted subjects (left) and control subjects (right). Our multi-task method (MT) outperforms Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR).

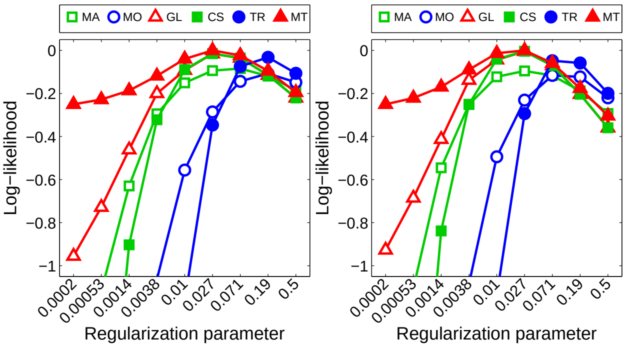

Figure 2. Cross-validated log-likelihood of structures learnt for each subject on cocaine addicted subjects (left) and control subjects (right). Our multi-task method (MT) is more stable for low regularization levels and outperforms Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR).

variability across subjects than across sessions of the same subject. Therefore, our cross-validation setting works as follows: we use one session as training set, and the remaining five sessions as testing set. We repeat this procedure for all the six sessions and report the log-likelihood (scaled for visualization purposes). In Figure 2, similar to the previous results, we can observe that the log-likelihood of our method is higher than the competing methods. Moreover, our method is more stable for low regularization levels than the other methods in our evaluation, which perform very poorly.

In order to measure the statistical significance of our previously reported log-likelihoods, we further compared the best parameter setting for each of the techniques. In Table 2, we report the two sample Z-statistic for the difference of our technique minus each competing method. Except for few subjects, the cross-validated log-likelihood of our method is statistically significantly higher (95%, $Z > 1.65$).

Table 2. Z-statistic for the difference of log-likelihoods between our technique and each competing method, for 16 cocaine addicted subjects. Except for few cases (marked with an asterisk), our method is statistically significantly better (95%, $Z > 1.65$) than Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR).

Method	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
MA	27.4	14.7	9.1	12.0	18.9	10.4	9.0	19.6
MO	25.6	17.0	10.4	13.7	19.4	10.4	10.7	20.4
GL	2.0	2.7	1.9	1.5*	0.7*	1.8	2.7	2.2
CS	2.0	2.6	1.9	1.5*	0.7*	1.8	2.7	2.1
TR	15.4	5.1	3.6	6.3	10.3	6.7	3.5	12.0
Method	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**
MA	9.6	8.1	8.5	17.2	23.2	19.2	19.5	10.3
MO	13.5	11.2	10.1	18.5	22.5	17.6	21.1	13.9
GL	4.3	2.9	1.8	1.8	1.8	2.2	4.7	2.9
CS	4.3	2.9	1.8	1.8	1.8	2.2	4.7	2.9
TR	3.2	2.2	3.7	8.8	8.8	11.4	8.0	5.0

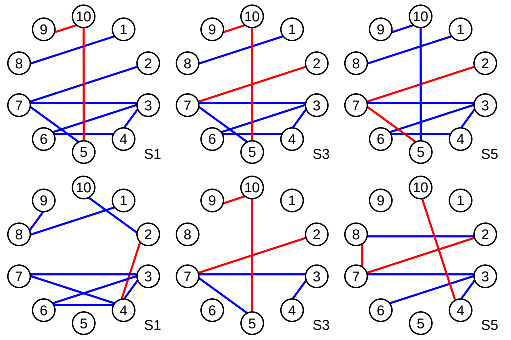

Figure 3. Subgraph of ten brain regions from learnt structures for three randomly selected cocaine addicted subjects, for our multi-task method (top) and graphical lasso (bottom). Regularization parameter $\rho = 0.027$. Positive interactions are shown in blue, negative interactions are shown in red. Notice that sparseness of our structures is consistent across subjects.

We show a subgraph of learnt structures for three randomly selected cocaine addicted subjects in Figure 3. We can observe that the sparseness pattern of the structures produced by our multi-task method is consistent across subjects.

7. Conclusions and Future Work

In this paper, we generalized the learning of sparse Gaussian graphical models to the multi-task setting by replacing the ℓ_1-norm regularization with an $\ell_{1,\infty}$-norm. We presented a block coordinate descent method which is provably convergent and yields sparse and positive definite estimates. We showed the connection between our multi-task structure learning problem and the continuous quadratic knapsack problem. Finally, we experimentally showed that the cross-validated log-likelihood of our method is more stable and statistically significantly higher than the competing methods in a brain fMRI dataset.

There are several ways of extending this research. Methods for selecting the regularization parameter need to be further investigated. In practice, our technique converges in a small number of iterations, but a more precise analysis of the rate of convergence needs to be performed. Finally, model selection consistency when the number of samples grows to infinity needs to be proved.

Acknowledgments

We thank Rita Goldstein for providing us the fMRI dataset, and Luis Ortiz for his guidance and helpful comments. This work was supported in part by NIDA Grants 1 R01 DA020949 and 1 R01 DA023579.

References

Banerjee, O., El Ghaoui, L., d'Aspremont, A., and Natsoulis, G. Convex optimization techniques for fitting sparse Gaussian graphical models. *International Conference on Machine Learning*, 2006.

Brucker, P. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 1984.

Dempster, A. Covariance selection. *Biometrics*, 1972.

Duchi, J., Gould, S., and Koller, D. Projected subgradient methods for learning sparse Gaussians. *Uncertainty in Artificial Intelligence*, 2008a.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. *International Conference on Machine Learning*, 2008b.

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

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

Helgason, K., Kennington, J., and Lall, H. A polynomially bounded algorithm for a singly constrained quadratic program. *Mathematical Programming*, 1980.

Honorio, J., Ortiz, L., Samaras, D., Paragios, N., and Goldstein, R. Sparse and locally constant Gaussian graphical models. *Neural Information Processing Systems*, 2009.

Jebara, T. Multi-task feature and kernel selection for SVMs. *International Conference on Machine Learning*, 2004.

Kiwiel, K. On linear-time algorithms for the continuous quadratic knapsack problem. *Journal of Optimization Theory and Applications*, 2007.

Lauritzen, S. *Graphical Models.* Oxford Press, 1996.

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

Liu, H., Palatucci, M., and Zhang, J. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. *International Conference on Machine Learning*, 2009.

Marlin, B. and K.Murphy. Sparse Gaussian graphical models with unknown block structure. *International Conference on Machine Learning*, 2009.

Marlin, B., Schmidt, M., and Murphy, K. Group sparse priors for covariance estimation. *Uncertainty in Artificial Intelligence*, 2009.

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

Niculescu-Mizil, A. and Caruana, R. Inductive transfer for Bayesian network structure learning. *Artificial Intelligence and Statistics*, 2007.

Qi, Y., Liu, D., Carin, L., and Dunson, D. Multi-task compressive sensing with Dirichlet process priors. *International Conference on Machine Learning*, 2008.

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

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 2001.

Wilson, A., Fern, A., Ray, S., and Tadepalli, P. Multi-task reinforcement learning: A hierarchical Bayesian approach. *International Conference on Machine Learning*, 2007.

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