arXiv:1207.4255v1 [cs.LG] 18 Jul 2012

Simultaneous and Group-Sparse Multi-Task Learning of Gaussian Graphical Models

Jean Honorio, Dimitris Samaras
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794
{jhonorio,samaras}@cs.sunysb.edu

Abstract

In this paper, we present $\ell_{1,p}$ *multi-task structure learning* for Gaussian graphical models. We discuss the 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_p regularized quadratic minimization subproblems. We further show that this subproblem leads to the *continuous quadratic knapsack problem* for $p = \infty$ and to a separable version of the well-known *quadratic trust-region problem* for $p = 2$, for which very efficient methods exist. Finally, we show promising results in synthetic experiments as well as in two real-world datasets.

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 Meinshausen and 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 and Lin [2007].

Structure learning techniques are very useful for analyzing datasets for which probabilistic dependencies are not known apriori. For instance, these techniques allow for modeling interactions between brain regions, based on measured activation levels through imaging. Suppose that we want to learn the structure of brain region interactions for one person.

We can expect that the interaction patterns in the brains of two persons are not exactly the same. On the other hand, when learning the structure for one person, we would like to use evidence from other persons as side information in our learning process. This becomes more important in settings with limited amount of data and high variability, 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 $\ell_{1,p}$ 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 $\ell_{1,p}$-norm, also known as the simultaneous prior [Turlach et al., 2005, Tropp, 2006] for $p = \infty$ or the group-sparse prior [Yuan and Lin, 2006, Meier et al., 2008] for $p = 2$.

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 $\ell_{1,p}$ multi-task structure learning problem and the continuous quadratic knapsack problem for $p = \infty$, which allows us to use existing efficient methods [Helgason et al., 1980, Brucker, 1984, Kiwiel, 2007]. We also show the connection between our multi-task structure learning problem and the quadratic trust-region problem for $p = 2$, which can be efficiently solved by one-dimensional optimization. Third, we discuss penalization of the diagonals of the precision matrices and experimentally show that penalizing the diagonals does not lead to better generalization performance, when compared to not penalizing the diagonals.

Compared to our related shorter conference paper [Honorio and Samaras, 2010], we present a more general framework which assumes $p > 1$, while [Honorio and Samaras, 2010] assumes $p = \infty$. We present a new algorithm for $p = 2$ and experimentally show that our method recovers the ground truth edges and the probability distribution always better than the $\ell_{1,2}$ method of Varoquaux et al. [2010] for every regularization level. We discuss penalization of the diagonals of the precision matrices which leads to additional optimization problems, namely the *continuous logarithmic knapsack problem* for $p = \infty$ and the *separable logarithmic trust-region problem* for $p = 2$. We show that our method outperforms others in recovering the topology of the ground truth model through synthetic experiments. In addition to the small fMRI dataset used in [Honorio and Samaras, 2010], we include validation in a considerably larger fMRI dataset. We experimentally show that the cross-validated log-likelihood of our method is higher than competing methods in both real-world datasets.

Section 2 introduces Gaussian graphical models as well as techniques for learning such structures from data. Section 3 sets up the $\ell_{1,p}$ multi-task structure learning 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. Section 6 shows the connection to the quadratic trust-region problem. Section 1 presents our algorithm in detail. Section 8 discusses penalization of the diagonals of the precision matrices. Experimental results are shown and explained in Section 9. Main contributions and results are summarized in Section 10.

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. $\sqrt{\sum_n c_n^2}$
$\mathbf{diag}(\mathbf{c}) \in \mathbb{R}^{N \times N}$	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
$\|\mathbf{A}\|_1$	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. $\sum_{mn} a_{mn} b_{mn}$

Table 1: Notation used in this paper.

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:

$$\max_{\mathbf{\Omega} \succ \mathbf{0}} \left(\ell_{\widehat{\mathbf{\Sigma}}}(\mathbf{\Omega}) - \rho \|\mathbf{\Omega}\|_1 \right) \tag{1}$$

for regularization parameter $\rho > 0$. The term $\|\mathbf{\Omega}\|_1$ 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 and Lin [2007]. Instead of solving eq.(1), the *Meinshausen-Bühlmann approximation* [Meinshausen and 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, Schmidt et al., 2009] and unknown block-variable assignments [Marlin and K.Murphy, 2009, Marlin et al., 2009], spatial coherence [Honorio et al., 2009], sparse changes in controlled experiments [Zhang and Wang, 2010], power law regularization in scale free networks [Liu and Ihler, 2011], or variable selection [Honorio et al., 2012].

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

Structure learning through ℓ_1-regularization has been also proposed for different types of graphical models: Markov random fields (MRFs) by a clique selection heuristic and approximate inference [Lee et al., 2006]; Bayesian networks on binary variables by logistic regression [Schmidt et al., 2007]; Conditional random fields by pseudo-likelihood and block regularization in order to penalize all parameters of an edge simultaneously [Schmidt et al., 2008]; and Ising models, i.e. MRFs on binary variables with pairwise interactions, by logistic regression [Wainwright et al., 2006] which is similar in spirit to Meinshausen and Bühlmann [2006].

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 goals are to learn one structure for each task that best explains the observed data, and to promote a common sparseness pattern of edges for all 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. $\omega^{(1)}_{n_1 n_2}, \ldots, \omega^{(K)}_{n_1 n_2}$). 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 $\ell_{1,p}$ *multi-task structure learning problem* is defined as:

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

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

$$\|\mathbf{\Omega}\|_{1,p} = \sum_{n_1 n_2} \|(\omega^{(1)}_{n_1 n_2}, \ldots, \omega^{(K)}_{n_1 n_2})\|_p \tag{4}$$

We assume that $p > 1$, since for $p = 1$, the multi-task problem in eq.(3) reduces to K *single-task* problems as in eq.(1), and for $p < 1$, eq.(3) is not convex. The number of samples $T^{(k)}$ 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.

The $\ell_{1,2}$ multi-task structure learning problem was originally proposed in [Varoquaux et al., 2010], where the authors minimize the original non-smooth objective function by using a sequence of smooth quadratic upper bounds. Varoquaux et al. [2010] do not provide any guarantee of positive definiteness, eigenvalue bounds or convergence. The $\ell_{1,\infty}$ multi-task problem was originally proposed in Honorio and Samaras [2010]. In this paper, we analyze the computational aspects of the more general $\ell_{1,p}$ multi-task problem for $p > 1$. While the $\ell_{1,\infty}$ multi-task problem of Honorio and Samaras [2010] leads to the *continuous quadratic knapsack problem*, we show that the $\ell_{1,2}$ multi-task problem leads to the *quadratic trust-region problem*. Another multi-task penalty has been proposed in Guo et al. [2010], however this penalty is non-convex.

3.2 Bounds

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

Lemma 1. *For $\rho > 0$ and $p > 1$, the $\ell_{1,p}$ multi-task structure learning problem in eq.(3) is a maximization problem with a 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, $\boldsymbol{\Omega}^{(k)} \succ \mathbf{0}$ is a convex constraint. □

Theorem 2. *For $\rho > 0$ and $p > 1$, the optimal solution to the $\ell_{1,p}$ multi-task structure learning problem in eq.(3) is unique and bounded as follows:*

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

Proof. Let the $\ell_{p'}$-norm be the dual of the ℓ_p-norm, i.e. $\frac{1}{p} + \frac{1}{p'} = 1$. By using the identity for dual norms $\rho\|\mathbf{c}\|_p = \max_{\|\mathbf{a}\|_{p'}\leq\rho} \mathbf{a}^{\mathrm{T}}\mathbf{c}$ in eq.(3), we get:

$$\max_{(\forall k)\ \boldsymbol{\Omega}^{(k)}\succ\mathbf{0}}\ \min_{(\forall n_1 n_2)\ \|\mathbf{a}_{n_1n_2}\|_{p'}\leq\rho} \sum_k T^{(k)}\left(\log\det\boldsymbol{\Omega}^{(k)} - \langle\widehat{\boldsymbol{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}, \boldsymbol{\Omega}^{(k)}\rangle\right) \tag{6}$$

where $\mathbf{a}_{n_1n_2} = (a^{(1)}_{n_1n_2}, \ldots, a^{(K)}_{n_1n_2})^{\mathrm{T}}$ 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 $\boldsymbol{\Omega}^{(k)} = (\widehat{\boldsymbol{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}})^{-1}$. By replacing this solution in eq.(6), we get the dual problem of eq.(3):

$$\min_{(\forall n_1 n_2)\ \|\mathbf{a}_{n_1n_2}\|_{p'}\leq\rho} -\sum_k T^{(k)}\log\det\left(\widehat{\boldsymbol{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}\right) - NK \tag{7}$$

In order to find a lower bound for the minimum eigenvalue of $\mathbf{\Omega}^{(k)^*}$, note that $\|\mathbf{\Omega}^{(k)^{*-1}}\|_2 = \|\widehat{\mathbf{\Sigma}}^{(k)} + \frac{\mathbf{A}^{(k)}}{T^{(k)}}\|_2 \leq \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \|\frac{\mathbf{A}^{(k)}}{T^{(k)}}\|_2 = \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{1}{T^{(k)}}\|\mathbf{A}^{(k)}\|_2 \leq \|\widehat{\mathbf{\Sigma}}^{(k)}\|_2 + \frac{1}{T^{(k)}}\|\mathbf{A}^{(k)}\|_{\mathfrak{F}}$. Since $\|\mathbf{a}_{n_1 n_2}\|_{p'} \leq \rho$, it follows that $|a^{(k)}_{n_1 n_2}| \leq \rho$ and therefore $\|\mathbf{A}^{(k)}\|_{\mathfrak{F}} \leq N\rho$.

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

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

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

4 Block Coordinate Descent Method

In this section, we develop a block coordinate descent method for our $\ell_{1,p}$ multi-task structure learning problem, and discuss some of its properties.

Since the objective function in eq.(3) 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 [Duchi and Singer, 2009]. 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 [Honorio et al., 2009, Honorio and Samaras, 2010, Honorio et al., 2012], unlike covariance selection [Banerjee et al., 2006] and graphical lasso [Friedman et al., 2007] which optimize the dual. We choose to optimize the primal because 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 will first assume that the diagonals of $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$ are not penalized by our multi-task regularizer defined in eq.(4). In Section 8, we will discuss penalization of the diagonals, for which an additional *continuous logarithmic knapsack problem* for $p = \infty$ or *separable logarithmic trust-region problem* for $p = 2$ needs to be solved. We point out that all the following theorems and lemmas still hold in that case.

Lemma 3. *The solution sequence generated by the block coordinate descent method is bounded and every cluster point is a solution of the $\ell_{1,p}$ multi-task structure learning problem in eq.*(3)*.*

Proof. The non-smooth regularizer $\|\mathbf{\Omega}\|_{1,p}$ is separable into a sum of $\mathcal{O}(N^2)$ individual functions of the form $\|(\omega^{(1)}_{n_1 n_2}, \ldots, \omega^{(K)}_{n_1 n_2})\|_p$. These functions are defined over blocks of K variables, i.e. $\omega^{(1)}_{n_1 n_2}, \ldots, \omega^{(K)}_{n_1 n_2}$. 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 $\ell_{1,p}$ 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:

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

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:

$$\max_{(\forall k)\ \mathbf{\Omega}^{(k)} \succ \mathbf{0}} \begin{pmatrix} \sum_k T^{(k)} \left(\log(z^{(k)} - \mathbf{y}^{(k)\mathrm{T}} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)}) - 2\mathbf{u}^{(k)\mathrm{T}} \mathbf{y}^{(k)} - v^{(k)} z^{(k)} \right) \\ -2\rho \sum_n \| (y_n^{(1)}, \ldots, y_n^{(K)}) \|_p \end{pmatrix} \tag{10}$$

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:

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

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. □

Remark 5. *Note that eq.*(11) *defines the "diagonal update step" of the block coordinate descent method. For each k we set* $z^{(k)}$ *to its optimal value, i.e.* $z^{(k)^*} = \frac{1}{v^{(k)}} + \mathbf{y}^{(k)\mathrm{T}} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)}$.

Theorem 6. *The "off-diagonal update step" of the block coordinate descent method for the* $\ell_{1,p}$ *multi-task structure learning problem in eq.*(3) *is equivalent to solving a sequence of strictly convex* $\ell_{1,p}$ *regularized quadratic subproblems:*

$$\min_{(\forall k)\ \mathbf{y}^{(k)} \in \mathbb{R}^{N-1}} \begin{pmatrix} \sum_k T^{(k)} \left(\frac{1}{2} \mathbf{y}^{(k)\mathrm{T}} v^{(k)} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)} + \mathbf{u}^{(k)\mathrm{T}} \mathbf{y}^{(k)} \right) \\ +\rho \sum_n \| (y_n^{(1)}, \ldots, y_n^{(K)}) \|_p \end{pmatrix} \tag{12}$$

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. □

As we will show in Section 8, the Schur complement is still positive when we penalize the diagonals, i.e. $z^{(k)} - \mathbf{y}^{(k)\mathrm{T}} \mathbf{W}^{(k)^{-1}} \mathbf{y}^{(k)} = \xi > 0$. Note that in such case, $\xi \neq \frac{1}{v^{(k)}}$ in contrast to eq.(11) but we can still perform the replacement in eq.(10), and therefore Theorem 6 still holds when penalizing the diagonals.

Lemma 7. *Let the* $\ell_{p'}$*-norm be the dual of the* ℓ_p*-norm, i.e.* $\frac{1}{p} + \frac{1}{p'} = 1$. *If the* $\ell_{\infty,p'}$ *norm* $\max_n \| (T^{(1)} u_n^{(1)}, \ldots, T^{(K)} u_n^{(K)}) \|_{p'} \leq \rho$, *the* $\ell_{1,p}$ *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. $(\exists \mathbf{A} \in \mathbb{R}^{N-1\times K})\ (T^{(1)}\mathbf{u}^{(1)}, \ldots, T^{(K)}\mathbf{u}^{(K)}) + \mathbf{A} = \mathbf{0} \wedge \max_n \|(a_{n1}, \ldots, a_{nK})\|_{p'} \leq \rho$. This condition is true for $\max_n \|(T^{(1)}u_n^{(1)}, \ldots, T^{(K)}u_n^{(K)})\|_{p'} \leq \rho$. □

Remark 8. *By using Lemma 7, 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 9. *The coordinate descent method for the $\ell_{1,p}$ regularized quadratic problem in eq.*(12) *is equivalent to solving a sequence of strictly convex ℓ_p regularized separable quadratic subproblems:*

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

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:

$$\mathbf{W}^{(k)^{-1}} = \begin{bmatrix} \mathbf{H}_{11}^{(k)} & \mathbf{h}_{12}^{(k)} \\ \mathbf{h}_{12}^{(k)^{\mathrm{T}}} & h_{22}^{(k)} \end{bmatrix},\ \mathbf{y}^{(k)} = \begin{bmatrix} \mathbf{y}_1^{(k)} \\ x_k \end{bmatrix},\ \mathbf{u}^{(k)} = \begin{bmatrix} \mathbf{u}_1^{(k)} \\ u_2^{(k)} \end{bmatrix} \tag{14}$$

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

In terms of the variable $\mathbf{x}$ and the constants $q_k = T^{(k)}v^{(k)}h_{22}^{(k)}$, $c_k = -T^{(k)}(v^{(k)}\mathbf{h}_{12}^{(k)^{\mathrm{T}}}\mathbf{y}_1^{(k)} + u_2^{(k)})$, the $\ell_{1,p}$ regularized quadratic problem in eq.(12) can be reformulated as in eq.(13). Moreover, since $(\forall k)\ T^{(k)} > 0 \wedge v^{(k)} > 0 \wedge h_{22}^{(k)} > 0 \Rightarrow \mathbf{q} > \mathbf{0}$, 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 $\mathcal{O}(K \log K)$ 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., 2009a] provides an $\mathcal{O}(K \log K)$ algorithm. We point out that [Duchi et al., 2008b, Liu et al., 2009a] 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. $(\forall k)\ q_k > 0$.

Theorem 10. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, $p = \infty$, the ℓ_∞ regularized separable quadratic problem in eq.*(13) *is equivalent to the separable quadratic problem with one ℓ_1 constraint:*

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

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

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

Remark 12. *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* $\mathbf{r}^* = \mathbf{c}$ *is inside the feasible region given that* $\|\mathbf{r}^*\|_1 \leq \rho$*.*

Lemma 13. *For* $\mathbf{q} > \mathbf{0}$*,* $(\forall k)\ c_k \neq 0$*,* $\|\mathbf{c}\|_1 > \rho$*, the optimal solution* $\mathbf{r}^*$ *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.* $(\forall k)\ r_k^* c_k \geq 0$*.*

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

Theorem 14. *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:*

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

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

Proof. By invoking Lemma 13, we can replace $(\forall k)\ r_k = \mathrm{sgn}(c_k)g_k,\ g_k \geq 0$ in eq.(15). Finally, we change the inequality constraint $\mathbf{1}^{\mathrm{T}}\mathbf{g} \leq \rho$ to an equality constraint since $\|\mathbf{c}\|_1 > \rho$ and therefore, the optimal solution must be on the boundary of the constraint set. □

Lemma 15. *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:*

$$g_k(\nu) = \max(0, |c_k| - \nu q_k) \tag{17}$$

for some ν*, and furthermore, the optimal solution fulfills the condition:*

$$\mathbf{g}^* = \mathbf{g}(\nu) \Leftrightarrow \mathbf{1}^{\mathrm{T}}\mathbf{g}(\nu) = \rho \tag{18}$$

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

$$\min_{\mathbf{g} \geq \mathbf{0}} \left(\sum_k \frac{1}{2q_k}(g_k - |c_k|)^2 + \nu(\mathbf{1}^{\mathrm{T}}\mathbf{g} - \rho) \right) \tag{19}$$

Both results can be obtained by invoking the Karush-Kuhn-Tucker optimality conditions on eq.(19). □

Remark 16. *Note that $g_k(\nu)$ in eq.*(17) *is a decreasing piecewise linear function with breakpoint $\nu = \frac{|c_k|}{q_k} > 0$. By Lemma 15, finding the optimal $\mathbf{g}^*$ is equivalent to finding ν in a piecewise linear function $\mathbf{1}^{\mathrm{T}}\mathbf{g}(\nu)$ that produces ρ.*

Lemma 17. *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 $g_k^* = \max(0, |c_k| - \nu^* q_k)$ for:*

$$\frac{|c_{\pi_{k^*}}|}{q_{\pi_{k^*}}} \geq \nu^* = \frac{\sum_{k=1}^{k^*} |c_{\pi_k}| - \rho}{\sum_{k=1}^{k^*} q_{\pi_k}} \geq \frac{|c_{\pi_{k^*+1}}|}{q_{\pi_{k^*+1}}} \tag{20}$$

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^*, we search for the range in which $\mathbf{1}^{\mathrm{T}}\mathbf{g}\left(\frac{|c_{\pi_{k^*}}|}{q_{\pi_{k^*}}}\right) \leq \rho \leq \mathbf{1}^{\mathrm{T}}\mathbf{g}\left(\frac{|c_{\pi_{k^*+1}}|}{q_{\pi_{k^*+1}}}\right)$. □

Theorem 18. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, $p = \infty$, the ℓ_∞ regularized separable quadratic problem in eq.*(13) *has the optimal solution:*

$$\begin{aligned}
&\|\mathbf{c}\|_1 \leq \rho \Rightarrow \mathbf{x}^* = \mathbf{0} \\
&\|\mathbf{c}\|_1 > \rho \wedge k > k^* \Rightarrow x_{\pi_k}^* = \frac{c_{\pi_k}}{q_{\pi_k}} \\
&\|\mathbf{c}\|_1 > \rho \wedge k \leq k^* \Rightarrow x_{\pi_k}^* = \operatorname{sgn}(c_{\pi_k}) \frac{\sum_{k=1}^{k^*} |c_{\pi_k}| - \rho}{\sum_{k=1}^{k^*} q_{\pi_k}}
\end{aligned} \tag{21}$$

Proof. For $\|\mathbf{c}\|_1 \leq \rho$, from Remark 12 we know that $\mathbf{r}^* = \mathbf{c}$. By Theorem 10, the optimal solution of eq.(13) is $\mathbf{x}^* = \mathbf{diag}(\mathbf{q})^{-1}(\mathbf{c} - \mathbf{r}^*) = \mathbf{0}$, and we prove the first claim.

For $\|\mathbf{c}\|_1 > \rho$, by Theorem 10, the optimal solution of eq.(13) $x_{\pi_k}^* = \frac{1}{q_{\pi_k}}(c_{\pi_k} - r_{\pi_k}^*)$. By Theorem 14, $x_{\pi_k}^* = \frac{1}{q_{\pi_k}}(c_{\pi_k} - \operatorname{sgn}(c_{\pi_k}) g_{\pi_k}^*)$. By Lemma 17, $x_{\pi_k}^* = \frac{c_{\pi_k}}{q_{\pi_k}} - \operatorname{sgn}(c_{\pi_k}) \max(0, \frac{|c_{\pi_k}|}{q_{\pi_k}} - \nu^*)$.

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

6 Separable Quadratic Trust-Region Problem

In this section, we show the connection between the $\ell_{1,2}$ multi-task structure learning problem and the separable quadratic trust-region problem, which can be efficiently solved by one-dimensional optimization.

The trust-region problem has been extensively studied by the mathematical optimization community [Forsythe and Golub, 1965, Moré and Sorensen, 1983, Boyd and Vandenberghe, 2006]. Trust-region methods arise in the optimization of general convex functions. In that context, the strategy behind trust-region methods is to perform a local second-order approximation to the original objective function. The quadratic model for local optimization is "trusted" to be correct inside a circular region (i.e. the trust region). Separability is usually not assumed, i.e. a symmetric matrix $\mathbf{Q}$ is used instead of $\mathbf{diag}(\mathbf{q})$ in eq.(13), and

therefore the general algorithms are more involved than ours. In the context of machine learning, [Duchi and Singer, 2009] provides a closed form solution for the separable version of the problem when 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. $(\forall k)\ q_k > 0$. A closed form solution is not possible in this general case, but the efficient one-dimensional *Newton-Raphson method* can be applied.

Theorem 19. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, $p = 2$, the ℓ_2 regularized separable quadratic problem in eq.(13) is equivalent to the separable quadratic trust-region problem:*

$$\min_{\|\mathbf{r}\|_2 \leq \rho} \left(\frac{1}{2} (\mathbf{r} - \mathbf{c})^{\mathrm{T}} \mathbf{diag}(\mathbf{q})^{-1} (\mathbf{r} - \mathbf{c}) \right) \tag{22}$$

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

Proof. By Lagrangian duality, the problem in eq.(22) is the dual of the problem in eq.(13). Furthermore, strong duality holds in this case. □

Remark 20. *In eq.(22), we can assume that $(\forall k)\ c_k \neq 0$. If $(\exists k)\ c_k = 0$, the partial optimal solution is $r_k^* = 0$, and since this assignment does not affect the constraint, we can safely remove r_k from the optimization problem.*

Remark 21. *In what follows, we assume that $\|\mathbf{c}\|_2 > \rho$. If $\|\mathbf{c}\|_2 \leq \rho$, the unconstrained optimal solution of eq.(22) is also its optimal solution, since $\mathbf{r}^* = \mathbf{c}$ is inside the feasible region given that $\|\mathbf{r}^*\|_2 \leq \rho$.*

Lemma 22. *For $\mathbf{q} > \mathbf{0}$, $(\forall k)\ c_k \neq 0$, $\|\mathbf{c}\|_2 > \rho$, the separable quadratic trust-region problem in eq.(22) is equivalent to the problem:*

$$\min_{\lambda \geq 0} \left(\sum_n \frac{c_n^2}{q_n + \lambda q_n^2} + \rho^2 \lambda \right) \tag{23}$$

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

Proof. By Lagrangian duality, the problem in eq.(23) is the dual of the problem in eq.(22). Furthermore, strong duality holds in this case. □

Corollary 23. *For the special case $\mathbf{q} = \mathbf{1}$ of Duchi and Singer [2009], the trust-region dual problem in eq.(23) has the closed form solution $\lambda^* = \max\left(0, \frac{\|\mathbf{c}\|_2}{\rho} - 1\right)$.*

Proof. For $\mathbf{q} = \mathbf{1}$, the problem in eq.(23) becomes $\min_{\lambda \geq 0} \left(\frac{\|\mathbf{c}\|_2^2}{1+\lambda} + \rho^2 \lambda \right)$. By minimizing with respect to λ and by noting that $\lambda \geq 0$, we prove our claim. □

Theorem 24. *For $\mathbf{q} > \mathbf{0}$, $\rho > 0$, $p = 2$, the ℓ_2 regularized separable quadratic problem in eq.(13) has the optimal solution:*

$$\begin{aligned} &\|\mathbf{c}\|_2 \leq \rho \Rightarrow \mathbf{x}^* = \mathbf{0} \\ &\|\mathbf{c}\|_2 > \rho \Rightarrow \mathbf{x}^* = \lambda^* \mathbf{diag}(\mathbf{1} + \lambda^* \mathbf{q})^{-1} \mathbf{c} \end{aligned} \tag{24}$$

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)
    For $p = \infty$, solve the ℓ_∞ regularized separable quadratic problem by eq.(21), either by sorting the breakpoints or using medians of breakpoint subsets. For $p = 2$, solve the ℓ_2 regularized separable quadratic problem by eq.(24) by using the Newton-Raphson method for solving the trust-region dual problem in eq.(23)
  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, $\mathbf{\Omega}^{(k)} \succ \mathbf{0}$

Proof. For $\|\mathbf{c}\|_2 \leq \rho$, from Remark 21 we know that $\mathbf{r}^* = \mathbf{c}$. By Theorem 19, the optimal solution of eq.(13) is $\mathbf{x}^* = \mathbf{diag}(\mathbf{q})^{-1}(\mathbf{c} - \mathbf{r}^*) = \mathbf{0}$, and we prove the first claim.

For $\|\mathbf{c}\|_2 > \rho$, by Theorem 19, the optimal solution of eq.(13) is $(\forall k)\ x_k^* = \frac{1}{q_k}(c_k - r_k^*)$. By Lemma 22, $x_k^* = \frac{1}{q_k}(c_k - \frac{1}{1+\lambda^* q_k} c_k) = \frac{\lambda^*}{1+\lambda^* q_k} c_k$, and we prove the second claim. □

7 Algorithm

Algorithm 1 shows the block coordinate descent method in detail. A careful implementation of the algorithm allows obtaining a time complexity of $\mathcal{O}(LN^3K)$ for L iterations, N variables and K tasks. In our experiments, the algorithm converges quickly in usually $L = 10$ iterations. The polynomial dependence $\mathcal{O}(N^3)$ 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. For $p = \infty$, the linear-time dependence $\mathcal{O}(K)$ 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 $\mathcal{O}(K \log K)$ algorithm is obtained by initially sorting the breakpoints and searching the range for which Lemma 17 holds. For $p = 2$, the linear-time dependence $\mathcal{O}(K)$ on the number of tasks can be accomplished by using the one-dimensional Newton-Raphson method for solving the trust-region dual problem in eq.(23). In our implementation, we initialize $\lambda = 0$ and perform 10 iterations of the Newton-Raphson method.

8 Penalizing the Diagonals

In this section, we discuss penalization of the diagonals of the precision matrices. It is unclear whether diagonal penalization leads to better models with respect to structure as well as generalization performance. For the *single-task* problem, covariance selection [Banerjee et al., 2006] and graphical lasso [Friedman et al., 2007] penalize the weights of

the diagonal elements. In contrast, the analysis of consistency in structure recovery of Ravikumar et al. [2008] assumed that diagonals are not penalized.

Note that, when the diagonals are not penalized, the "diagonal update step" (Remark 5) reduces to setting for each k, ${z^{(k)}}^* = \frac{1}{v^{(k)}} + {\mathbf{y}^{(k)}}^{\mathrm{T}} {\mathbf{W}^{(k)}}^{-1} \mathbf{y}^{(k)}$. Penalization of the diagonals of the precision matrices is more involved, since it requires the solution of additional optimization problems, namely the *continuous logarithmic knapsack problem* for $p = \infty$ and the *separable logarithmic trust-region problem* for $p = 2$. First, we discuss the general problem for arbitrary $p > 1$.

Lemma 25. *When penalizing the diagonals of the precision matrices, the "diagonal update step" of the block coordinate descent method for the $\ell_{1,p}$ multi-task structure learning problem in eq.(3) is equivalent to solving a sequence of strictly convex ℓ_p regularized separable logarithmic subproblems:*

$$\max_{(\forall k)\ z^{(k)} > b_k} \left(\sum_k q_k \log(z^{(k)} - b_k) - \mathbf{c}^{\mathrm{T}} \mathbf{z} - \rho \|\mathbf{z}\|_p \right) \tag{25}$$

where $\mathbf{z} = (z^{(1)}, \ldots, z^{(K)})^{\mathrm{T}}$ and $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$. Moreover, the block coordinate descent method generates a sequence of positive definite solutions.

Proof. When we choose to penalize the diagonals of the precision matrices $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$, eq.(10) contains an additional ℓ_p penalty, i.e. $\rho \|\mathbf{z}\|_p$. In terms of the variables $\mathbf{y}^{(k)}, z^{(k)}$ and the constant matrix $\mathbf{W}^{(k)}$ introduced in eq.(9), the multi-task structure learning problem in eq.(3) can be reformulated as:

$$\max_{(\forall k)\ \mathbf{\Omega}^{(k)} \succ \mathbf{0}} \begin{pmatrix} \sum_k T^{(k)} \left(\log(z^{(k)} - {\mathbf{y}^{(k)}}^{\mathrm{T}} {\mathbf{W}^{(k)}}^{-1} \mathbf{y}^{(k)}) - 2 {\mathbf{u}^{(k)}}^{\mathrm{T}} \mathbf{y}^{(k)} - v^{(k)} z^{(k)} \right) \\ -2\rho \sum_n \|(y_n^{(1)}, \ldots, y_n^{(K)})\|_p - \rho \|\mathbf{z}\|_p \end{pmatrix} \tag{26}$$

Let $q_k = T^{(k)} > 0$, $c_k = T^{(k)} v^{(k)} > 0$ and $b_k = {\mathbf{y}^{(k)}}^{\mathrm{T}} {\mathbf{W}^{(k)}}^{-1} \mathbf{y}^{(k)} \geq 0$ since $\mathbf{W}^{(k)} \succ \mathbf{0}$ and $\mathbf{y}^{(k)}$ is an arbitrary vector (including the case $\mathbf{y}^{(k)} = 0$). We obtain eq.(25) by noting that we are maximizing with respect to $\mathbf{z}$ and by enforcing $(\forall k)\ z^{(k)} > b_k$ since $\log(z^{(k)} - b_k)$ is undefined for $z^{(k)} \leq b_k$.

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 $\mathbf{W}^{(k)} \succ \mathbf{0}$. Note that the Schur complement $z^{(k)} - {\mathbf{y}^{(k)}}^{\mathrm{T}} {\mathbf{W}^{(k)}}^{-1} \mathbf{y}^{(k)} = z^{(k)} - b_k$ and therefore it is strictly positive for every feasible solution given the constraints $(\forall k)\ z^{(k)} > b_k$.

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. □

Lemma 26. *For $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$, $\rho > 0$, $p > 1$, the ℓ_p regularized separable logarithmic problem in eq.(25) is equivalent to the separable logarithmic problem with one $\ell_{p'}$ constraint:*

$$\min_{\substack{\mathbf{r} \geq \mathbf{0} \\ \|\mathbf{r}\|_{p'} = \rho}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}} \mathbf{r} \right) \tag{27}$$

Furthermore, their optimal solutions are related by ${z^{(k)}}^* = b_k + \frac{q_k}{c_k + r_k^*}$.

Proof. By Lagrangian duality, the problem in eq.(27) is the dual of the problem in eq.(25). Furthermore, strong duality holds in this case.

The constraint $\mathbf{r} \geq \mathbf{0}$ comes from the fact that $\mathbf{z} > \mathbf{0}$ since it is the diagonal of positive definite matrices. Note that for a general $\mathbf{z} \in \mathbb{R}^K$, we have $\rho\|\mathbf{z}\|_p = \max_{\|\mathbf{r}\|_{p'} \leq \rho} \mathbf{r}^{\mathrm{T}}\mathbf{z}$. In order to maximize this expression, $\mathbf{r}$ will take values on the non-negative orthant since $\mathbf{z} > \mathbf{0}$.

We changed the inequality constraint $\|\mathbf{r}\|_{p'} \leq \rho$ to an equality constraint since the objective is separable and decreasing with respect to each r_k and therefore, the optimal solution must be on the boundary of the constraint set. □

In what follows, we focus on the case $p = \infty$ and show that this problem can be solved by a combination of sorting and the Newton-Raphson method.

Lemma 27. *For $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$, $\rho > 0$, $p = \infty$, the separable logarithmic problem with one $\ell_{p'}$ in eq.(27) is the continuous logarithmic knapsack problem:*

$$\min_{\substack{\mathbf{r} \geq \mathbf{0} \\ \mathbf{1}^{\mathrm{T}}\mathbf{r} = \rho}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}}\mathbf{r} \right) \tag{28}$$

which has the solution:

$$r_k(\nu) = \begin{cases} +\infty, & \nu \leq b_k \\ \frac{q_k}{\nu - b_k} - c_k, & b_k < \nu < \frac{q_k}{c_k} + b_k \\ 0, & \nu \geq \frac{q_k}{c_k} + b_k \end{cases} \tag{29}$$

for some ν, and furthermore, the optimal solution fulfills the condition:

$$\mathbf{r}^* = \mathbf{r}(\nu) \Leftrightarrow \mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu) = \rho \tag{30}$$

Proof. The Lagrangian of eq.(28) is:

$$\min_{\mathbf{r} \geq \mathbf{0}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}}\mathbf{r} + \nu(\mathbf{1}^{\mathrm{T}}\mathbf{r} - \rho) \right) \tag{31}$$

Both results can be obtained by invoking the Karush-Kuhn-Tucker optimality conditions on eq.(31). □

Remark 28. *Note that for $\nu \leq b_k$, we have $r_k(\nu) = +\infty$ in eq.(29), therefore $\mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu)$ is finite if and only if $\nu > \max_k b_k = \|b\|_\infty$. Additionally, for $\nu > \|b\|_\infty$ we have that $r_k(\nu)$ in eq.(29) is a decreasing piecewise inverse function with breakpoint $\nu = \frac{q_k}{c_k} + b_k > 0$. By Lemma 27, finding the optimal $\mathbf{r}^*$ is equivalent to finding ν in a piecewise inverse function $\mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu)$ that produces ρ.*

Similarly as in Lemma 17, we sort the breakpoints in decreasing order, i.e. we find a permutation π of the indices $1, 2, \ldots, K$ such that $\frac{q_{\pi_1}}{c_{\pi_1}} + b_{\pi_1} \geq \frac{q_{\pi_2}}{c_{\pi_2}} + b_{\pi_2} \geq \cdots \geq \frac{q_{\pi_K}}{c_{\pi_K}} + b_{\pi_K} \geq \frac{q_{\pi_{K+1}}}{c_{\pi_{K+1}}} + b_{\pi_{K+1}} \equiv 0$. Then, we search for the optimal breakpoint k^* or equivalently we search for the range in which $\mathbf{1}^{\mathrm{T}}\mathbf{r}\left(\frac{q_{\pi_{k^*}}}{c_{\pi_{k^*}}} + b_{\pi_{k^*}}\right) \leq \rho \leq \mathbf{1}^{\mathrm{T}}\mathbf{r}\left(\frac{q_{\pi_{k^*+1}}}{c_{\pi_{k^*+1}}} + b_{\pi_{k^*+1}}\right)$. After finding k^*, ν^*

can be found by the Newton-Raphson method in order to fulfill the condition $\mathbf{1}^{\mathrm{T}}\mathbf{r}(\nu^*) = \rho$ in Lemma 27. In our implementation, we initialize ν at one of the extremes of the optimal range, i.e. $\nu = \max(\|b\|_\infty + \varepsilon, \frac{q_{\pi_{k^*}}}{c_{\pi_{k^*}}} + b_{\pi_{k^*}})$ for some small $\varepsilon > 0$. We then perform 10 iterations of the Newton-Raphson method.

Next, we focus on the case $p = 2$ and show that this problem can be solved by the Newton-Raphson method.

Lemma 29. *For $\mathbf{q}, \mathbf{c}, \mathbf{b} > \mathbf{0}$, $\rho > 0$, $p = 2$, the separable logarithmic problem with one $\ell_{p'}$ in eq.(27) is the separable logarithmic trust-region problem:*

$$\min_{\substack{\mathbf{r}\geq\mathbf{0} \\ \|\mathbf{r}\|_2\leq\rho}} \left(-\sum_k q_k \log(r_k + c_k) - \mathbf{b}^{\mathrm{T}}\mathbf{r} \right) \tag{32}$$

which can be solved by one-dimensional optimization of:

$$\max_{\lambda\geq 0} \left(-\sum_k q_k \log(r_k(\lambda) + c_k) - \mathbf{b}^{\mathrm{T}}\mathbf{r}(\lambda) + \frac{\lambda}{2}\left(\mathbf{r}(\lambda)^{\mathrm{T}}\mathbf{r}(\lambda) - \rho^2\right) \right) \tag{33}$$

where $r_k(\lambda) = \frac{b_k - \lambda c_k + \sqrt{(b_k+\lambda c_k)^2 + 4\lambda q_k}}{2\lambda}$.

Proof. By Lagrangian duality, the problem in eq.(33) is the dual of the problem in eq.(32). Furthermore, strong duality holds in this case. □

In our implementation, we initialize $\lambda = \frac{1}{K}\sum_k \frac{q_k + b_k(c_k+\rho)}{\rho(c_k+\rho)}$ and perform 10 iterations of the Newton-Raphson method. Our initialization rule follows from using the average of the k independently optimal values of λ. That is, we consider only one task k at a time from eq.(33) which leads to $\max_{\lambda\geq 0}\left(-q_k\log(r_k(\lambda)+c_k) - b_k r_k(\lambda) + \frac{\lambda}{2}(r_k^2(\lambda) - \rho^2)\right)$. Then, we compute the optimal value of λ under this setting, which is $\frac{q_k + b_k(c_k+\rho)}{\rho(c_k+\rho)}$. Finally, we average these optimal values for all k which leads to our initialization rule for λ.

9 Experimental Results

We begin with a synthetic example to test the ability of the method to recover the ground truth structure from data. The model contains $N = 50$ variables and $K = 5$ tasks. For each of 50 repetitions, we generate a topology (undirected graph) $\mathbf{\Upsilon}_g \in \{0,1\}^{N\times N}$ with a required edge density (either 0.1,0.3,0.5). For each task k, we first generate a Gaussian graphical model $\mathbf{\Omega}_g^{(k)}$ with topology $\mathbf{\Upsilon}_g$ where each edge weight is generated uniformly at random from $[-1; +1]$. We ensure positive definiteness of $\mathbf{\Omega}_g^{(k)}$ by verifying that its minimum eigenvalue is at least 0.1. We then generate a dataset of $T^{(k)} = 50$ samples.

In order to measure the closeness of the recovered models to the ground truth, we measured the Kullback-Leibler divergence, sensitivity (one minus the fraction of falsely excluded edges) and specificity (one minus the fraction of falsely included edges). For comparison purposes, we used the following *single-task* methods: covariance selection [Banerjee et al.,

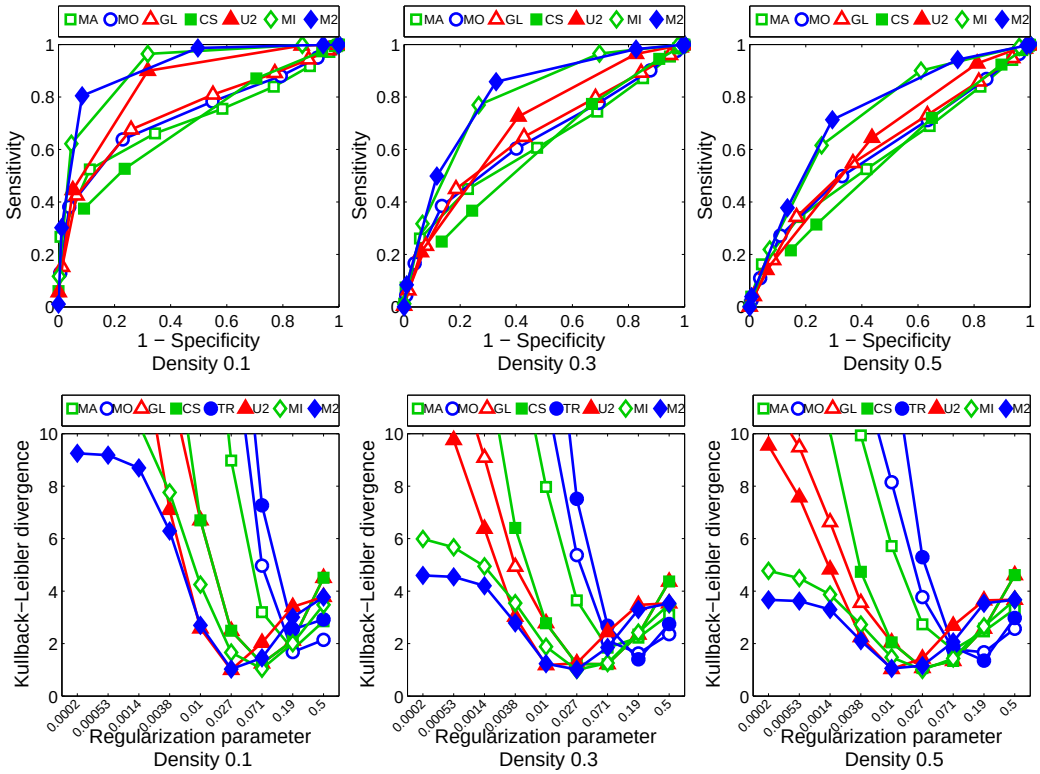


Figure 1: ROC curves (top) and cross-validated Kullback-Leibler divergence (bottom) between the recovered models and the ground truth for low (left), moderate (center) and high (right) edge density. Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods recover the ground truth edges remarkably better than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). The Kullback-Leibler divergence of our $\ell_{1,2}$ method is always lower than the $\ell_{1,2}$ upper bound method for all the regularization values.

2006], graphical lasso [Friedman et al., 2007], Meinshausen-Bühlmann approximation [Meinshausen and Bühlmann, 2006] and Tikhonov regularization. We also compared our method to the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010].

Figure 1 shows the ROC curves and Kullback-Leibler divergence between the recovered models and the ground truth. Note that both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods recover the ground truth edges remarkably better (higher ROC) than the comparison methods, including the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010]. Our $\ell_{1,2}$ method always produces better probability distributions (lower Kullback-Leibler divergence) than the $\ell_{1,2}$ upper bound method for all the regularization values. We did not observe a significant difference in Kullback-Leibler divergence between penalizing the weights in the diagonals versus not penalizing the diagonals for most regularization levels $\rho < 0.19$. For $\rho \geq 0.19$, diagonal penalization leads to a slightly worse Kullback-Leibler divergence. Furthermore, the ROC curves for our methods with and without diagonal penalization were the same. Therefore, we chose to report only the results without diagonal penalization.

For experimental validation on a real-world dataset, we first use 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 16 cocaine addicted

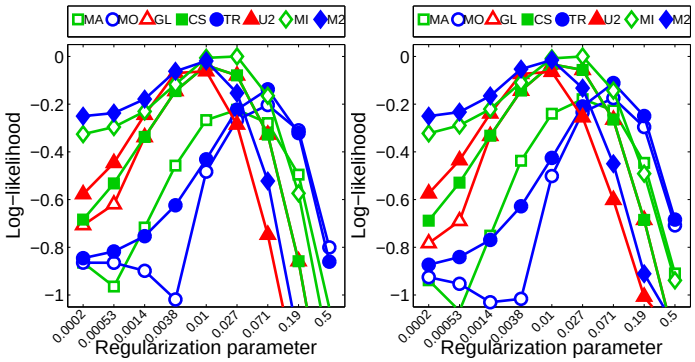


Figure 2: Cross-validated log-likelihood of structures learnt for each of the six sessions on cocaine addicted subjects (a) and control subjects (b). Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods have higher log-likelihood than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). Our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method for all the regularization values.

subjects and 12 control subjects. 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 (`http://www.fil.ion.ucl.ac.uk/spm/`). We extracted voxels from the gray matter only, and grouped them into 157 regions by using standard labels (Please, see Appendix A), given by the Talairach Daemon (`http://www.talairach.org/`). 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 2, we can observe that the log-likelihood of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods is better than the comparison methods. Moreover, our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010] for all the regularization values. We did not observe a significant difference in log-likelihood between penalizing the weights in the diagonals versus not penalizing the diagonals for most regularization levels $\rho < 0.19$. For $\rho \geq 0.19$, diagonal penalization leads to a slightly worse log-likelihood. Therefore, we chose to report only the results without diagonal penalization.

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 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 3, we can observe that the log-likelihood of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods is better than the comparison methods. Moreover, our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010] for all the regularization values. Finally, both our $\ell_{1,\infty}$ and $\ell_{1,2}$ methods are

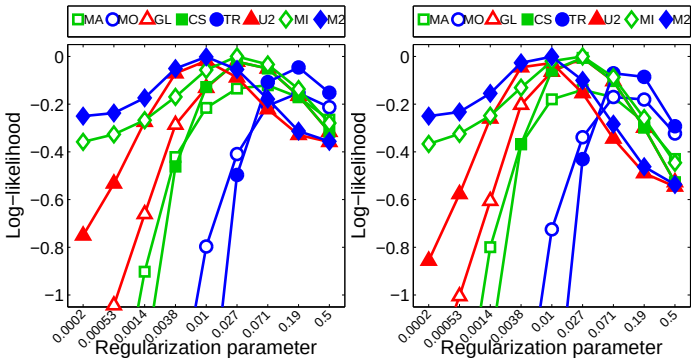


Figure 3: Cross-validated log-likelihood of structures learnt for each subject on cocaine addicted subjects (a) and control subjects (b). Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods have higher log-likelihood than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). Our $\ell_{1,2}$ method is always better than the $\ell_{1,2}$ multi-task upper bound method for all the regularization values. For low regularization levels, our methods are more stable than the comparison methods.

Method	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**
MA	27.4	14.7	9.1	12.0	18.9	10.4	9.0	19.6	9.6	8.1	8.5	17.2	23.2	19.2	19.5	10.3
MO	25.6	17.0	10.4	13.7	19.4	10.4	10.7	20.4	13.5	11.2	10.1	18.5	22.5	17.6	21.1	13.9
GL	2.0	2.7	1.9	1.5	0.7*	1.8	2.7	2.2	4.3	2.9	1.8	1.8	1.8	2.2	4.7	2.9
CS	2.0	2.6	1.9	1.5	0.7*	1.8	2.7	2.1	4.3	2.9	1.8	1.8	1.8	2.2	4.7	2.9
TR	15.4	5.1	3.6	6.3	10.3	6.7	3.5	12.0	3.2	2.2	3.7	8.8	8.8	11.4	8.0	5.0
U2	5.5	2.3	1.7	2.0	4.0	1.9	1.1*	4.3	1.3	0.7*	1.3	3.9	3.3	3.9	2.9	1.2*
M2	0.8*	-0.3*	0.1*	0.3*	0.7*	0.5*	-0.1*	1.0*	-0.1*	-0.2*	-0.0*	0.8*	0.5*	1.1*	0.3*	-0.3*

Table 2: Z-statistic for the difference of log-likelihoods between our $\ell_{1,\infty}$ technique and each other method, for 16 cocaine addicted subjects. Expect for few cases (marked with an asterisk), our method is statistically significantly better (90%, $Z > 1.28$) than the $\ell_{1,2}$ upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). The $\ell_{1,\infty}$ and $\ell_{1,2}$ (M2) methods are not statistically significantly different.

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 Tables 2 and 3, we report the two sample Z-statistic for the difference of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ techniques minus each competing method. Except for few subjects, the cross-validated log-likelihood of both our $\ell_{1,\infty}$ and $\ell_{1,2}$ methods is statistically significantly higher (90%, $Z > 1.28$) than the comparison methods, including the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010]. Our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods are not statistically significantly different.

We show a subgraph of learnt structures for three randomly selected cocaine addicted subjects in Figure 4. We can observe that the sparseness pattern of the structures produced by our multi-task method is consistent across subjects.

Next, we present experimental results on a considerably larger real-world dataset. The *1000 functional connectomes* dataset contains resting-state fMRI of over 1128 subjects collected on several sites around the world. The dataset is publicly available at `http://www.`

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
MA	27.1	15.2	9.2	11.9	18.6	10.2	9.3	19.0	9.9	8.4	8.7	16.7	23.1	18.5	19.5	10.9
MO	25.3	17.5	10.5	13.7	19.1	10.2	10.9	19.8	13.8	11.5	10.3	18.1	22.4	16.9	21.1	14.4
GL	1.3	3.0	1.8	1.2*	0.0*	1.4	2.9	1.3*	4.5	3.1	1.9	1.1*	1.4	1.2*	4.5	3.3
CS	1.3	2.9	1.8	1.2*	0.0*	1.4	2.9	1.2*	4.5	3.1	1.9	1.0*	1.4	1.3*	4.5	3.2
TR	14.9	5.5	3.6	6.2	9.7	6.4	3.7	11.1	3.4	2.5	3.8	8.1	8.5	10.3	7.9	5.4
U2	4.8	2.6	1.7	1.8	3.3	1.5	1.2*	3.4	1.5	0.9*	1.4	3.2	2.8	2.9	2.6	1.5
MI	-0.8*	0.3*	-0.1*	-0.3*	-0.7*	-0.5*	0.1*	-1.0*	0.1*	0.2*	0.0*	-0.8*	-0.5*	-1.1*	-0.3*	0.3*

Table 3: Z-statistic for the difference of log-likelihoods between our $\ell_{1,2}$ technique and each other method, for 16 cocaine addicted subjects. Expect for few cases (marked with an asterisk), our method is statistically significantly better (90%, $Z > 1.28$) than the $\ell_{1,2}$ upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). The $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ methods are not statistically significantly different.

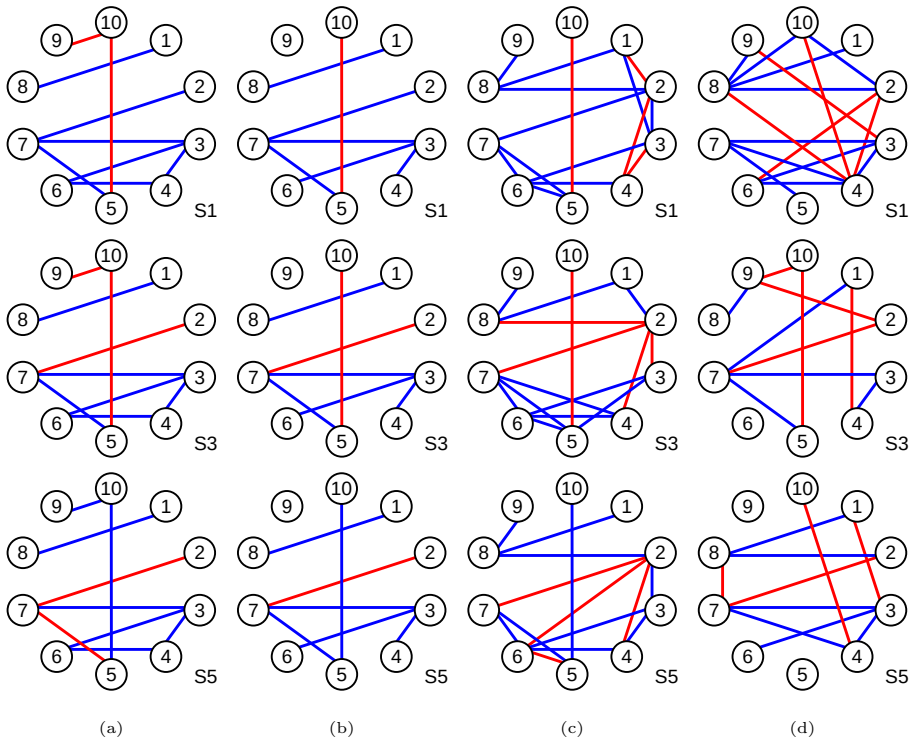


Figure 4: Subgraph of ten randomly selected brain regions from learnt structures for three randomly selected cocaine addicted subjects, for (a) our $\ell_{1,\infty}$ multi-task method, (b) our $\ell_{1,2}$ multi-task method, (c) the $\ell_{1,2}$ upper bound method and (d) graphical lasso. Regularization parameter $\rho = 0.01$. Positive interactions are shown in blue, negative interactions are shown in red. Notice that sparseness of our structures (a,b) are consistent across subjects, while the remaining methods (c,d) fail to obtain a consistent sparseness pattern.

`nitrc.org/projects/fcon_1000/`. Resting-state fMRI is a procedure that captures brain function of an individual that is not focused on the outside world, while his brain is at wakeful rest. Registration of the dataset to the same spatial reference template (Talairach space) and spatial smoothing was performed in SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`). We extracted voxels from the gray matter only, and grouped them into 157 regions by using standard labels (Please, see Appendix A), given by the Talairach Daemon (`http://www.talairach.org/`). These regions span the entire brain (cerebellum, cerebrum and

Site	Subjects	Scans	Site	Subjects	Scans	Site	Subjects	Scans
AnnArbor_a	23	295	Cleveland1	17	125	NewYorkA2	24	192
Baltimore	46	120	Cleveland2	14	125	NewYorkB	20	168
Bangor	20	256	Dallas	24	114	Newark	19	135
Beijing1	40	225	ICBM	42	128	Ontario	11	100
Beijing2	42	225	Leiden1	12	210	Orangeburg	20	162
Beijing3	41	225	Leiden2	19	210	Oulu1	57	243
Beijing4	30	225	Leipzig	37	192	Oulu2	47	243
Beijing5	45	225	NYU_TRT1A	13	192	Oxford	22	175
Berlin	26	192	NYU_TRT1B	12	192	PaloAlto	17	234
Cambridge1	48	117	NYU_TRT2A	13	192	Queensland	18	189
Cambridge2	46	117	NYU_TRT2B	12	192	SaintLouis	31	125
Cambridge3	49	117	NYU_TRT3A	13	192	Taipei_a	13	256
Cambridge4	55	117	NYU_TRT3B	12	192	Taipei_b	8	160
CambridgeWG	35	144	NewYorkA1	35	192			

Table 4: Number of subjects per collection site and number of scans per subject in the *1000 functional connectomes* dataset

brainstem). In order to capture laterality effects, we have regions for the left and right side of the brain. Table 4 shows the number of subjects per collection site as well as the number of scans per subject.

We learn one Gaussian graphical model for each of the 41 collection sites, i.e. each site is a task. For each site, we used one third of the subjects for training, one third for validation and the remaining third for testing. We performed six repetitions by making each third of the subjects take turns as training, validation and testing sets. We report the negative log-likelihood on the testing set in Fig.5 (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 both our $\ell_{1,\infty}$ and $\ell_{1,2}$ multi-task methods is better than the comparison methods. Moreover, our $\ell_{1,2}$ method is better than the $\ell_{1,2}$ multi-task upper bound method of Varoquaux et al. [2010]. Our results also suggest that diagonal penalization does not produce better generalization performance.

We show a subgraph of learnt structures for three randomly selected collection sites in Figure 6. We can observe that the sparseness pattern of the structures produced by our multi-task method is consistent across collection sites.

10 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 $\ell_{1,p}$-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 $\ell_{1,\infty}$ multi-task structure learning problem and the continuous quadratic knapsack problem, as well as the connection between our $\ell_{1,2}$ multi-task structure learning problem and the quadratic trust-region problem. In synthetic experiments, we showed that our method outperforms others in recovering the topology of the ground truth model. The cross-validated log-likelihood of our method is higher than competing methods in two real-world brain fMRI datasets. For the $\ell_{1,2}$ problem, our block coordinate descent method leads to better ground-truth recovery and

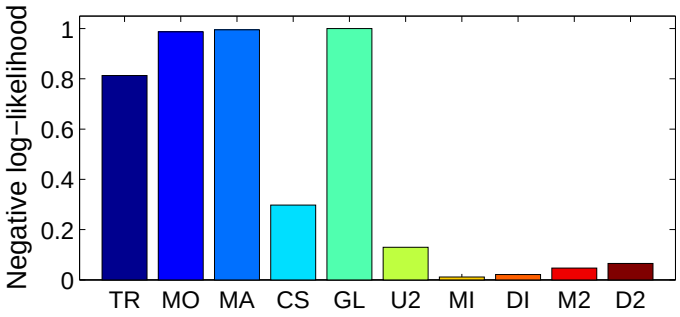


Figure 5: Test negative log-likelihood of structures learnt for the *1000 functional connectomes* dataset. Differences between our multi-task methods and the rest are statistically significant (99%, $Z > 2.33$). Both our $\ell_{1,\infty}$ (MI) and $\ell_{1,2}$ (M2) multi-task methods (without diagonal penalization) as well as our $\ell_{1,\infty}$ (DI) and $\ell_{1,2}$ (D2) multi-task methods (with diagonal penalization), have better log-likelihood than the $\ell_{1,2}$ multi-task upper bound method (U2), Meinshausen-Bühlmann with AND-rule (MA), OR-rule (MO), graphical lasso (GL), covariance selection (CS) and Tikhonov regularization (TR). Our $\ell_{1,2}$ method is better than the $\ell_{1,2}$ multi-task upper bound method. Our results also suggest that diagonal penalization does not produce better generalization performance.

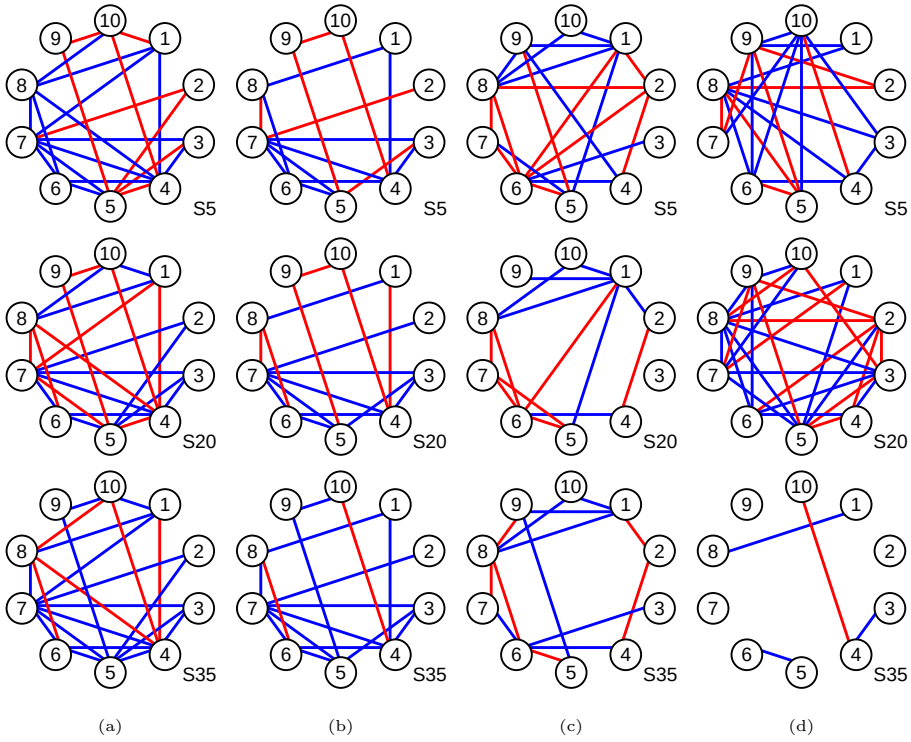


Figure 6: Subgraph of ten randomly selected brain regions from learnt structures for three randomly selected collection sites, for (a) our $\ell_{1,\infty}$ multi-task method, (b) our $\ell_{1,2}$ multi-task method, (c) the $\ell_{1,2}$ upper bound method and (d) covariance selection. Regularization parameter $\rho = 0.0002$. Positive interactions are shown in blue, negative interactions are shown in red. Notice that sparseness of our structures (a,b) are consistent across collection sites, while the remaining methods (c,d) fail to obtain a consistent sparseness pattern.

generalization when compared to the upper bound method of Varoquaux et al. [2010]. We experimentally found that diagonal penalization does not lead to better generalization

performance, when compared to not penalizing the diagonals. Our methods with and without diagonal penalization recover the ground truth edges similarly well. Therefore, we believe the negative impact of diagonal penalization is not on structure recovery but on parameter learning.

There are several ways of extending this research. 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. Model selection consistency when the number of samples grows to infinity needs to be proved. Finally, we hope the connection to the quadratic knapsack and trust-region problems will be useful for other multi-task problems, e.g. regression.

Acknowledgments

We thank Rita Goldstein for providing us the cocaine addiction fMRI dataset, Dardo Tomasi for providing us the *1000 functional connectomes* dataset, and Luis Ortiz for his guidance and helpful comments. This work was supported in part by NIDA Grants 1 R01 DA020949, 1 R01 DA023579 and NIBIB Grant 1 R01 EB007530.

References

O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex Optimization Techniques for Fitting Sparse Gaussian Graphical Models. *ICML*, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2006.

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

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

J. Duchi, S. Gould, and D. Koller. Projected Subgradient Methods for Learning Sparse Gaussians. *UAI*, 2008a.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions. *ICML*, 2008b.

J. Duchi and Y. Singer. Efficient Learning using Forward-Backward Splitting. *NIPS*, 2009.

G. Forsythe and G. Golub. On the Stationary Values of a Second-Degree Polynomial on the Unit Sphere. *SIAM Journal of the Society for Industrial and Applied Mathematics*, 1965.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 2007.

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.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint Estimation of Multiple Graphical Models. *Biometrika*, 2010.

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

J. Honorio, L. Ortiz, D. Samaras, N. Paragios, and R. Goldstein. Sparse and Locally Constant Gaussian Graphical Models. *NIPS*, 2009.

J. Honorio and D. Samaras. Multi-Task Learning of Gaussian Graphical Models. *ICML*, 2010.

J. Honorio, D. Samaras, I. Rish, and G. Cecchi. Variable Selection for Gaussian Graphical Models. *AISTATS*, 2012.

T. Jebara. Multi-Task Feature and Kernel Selection for SVMs. *ICML*, 2004.

K. Kiwiel. On Linear-Time Algorithms for the Continuous Quadratic Knapsack Problem. *Journal of Optimization Theory and Applications*, 2007.

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

S. Lee, V. Ganapathi, and D. Koller. Efficient Structure Learning of Markov Networks Using ℓ_1-Regularization. *NIPS*, 2006.

E. Levina, A. Rothman, and J. Zhu. Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty. *The Annals of Applied Statistics*, 2008.

H. Liu, M. Palatucci, and J. Zhang. Blockwise Coordinate Descent Procedures for the Multi-task Lasso, with Applications to Neural Semantic Basis Discovery. *ICML*, 2009a.

J. Liu, S. Ji, and J. Ye. Multi-Task Feature Learning Via Efficient $\ell_{2,1}$-Norm Minimization. *UAI*, 2009b.

Q. Liu and A. Ihler. Learning Scale Free Networks by Reweighted ℓ_1 regularization. *AISTATS*, 2011.

B. Marlin and K.Murphy. Sparse Gaussian Graphical Models with Unknown Block Structure. *ICML*, 2009.

B. Marlin, M. Schmidt, and K. Murphy. Group Sparse Priors for Covariance Estimation. *UAI*, 2009.

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

N. Meinshausen and P. Bühlmann. High Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 2006.

J. Moré and D. Sorensen. Computing a Trust Region Step. *SIAM Journal on Scientific and Statistical Computing*, 1983.

A. Niculescu-Mizil and R. Caruana. Inductive Transfer for Bayesian Network Structure Learning. *AISTATS*, 2007.

Y. Qi, D. Liu, L. Carin, and D. Dunson. Multi-Task Compressive Sensing with Dirichlet Process Priors. *ICML*, 2008.

P. Ravikumar, G. Raskutti, M. Wainwright, and B. Yu. Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1-regularized MLE. *NIPS*, 2008.

M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure Learning in Random Fields for Heart Motion Abnormality Detection. *CVPR*, 2008.

M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning Graphical Model Structure Using ℓ_1-Regularization Paths. *AAAI*, 2007.

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.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 1996.

J. Tropp. Algorithms for simultaneous sparse approximation, Part II: convex relaxation. *Signal Processing*, 2006.

P. Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*, 2001.

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

G. Varoquaux, A. Gramfort, J. Poline, and B. Thirion. Brain Covariance Selection: Better Individual Functional Connectivity Models Using Population Prior. *NIPS*, 2010.

M. Wainwright, P. Ravikumar, and J. Lafferty. High dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression. *NIPS*, 2006.

A. Wilson, A. Fern, S. Ray, and P. Tadepalli. Multi-Task Reinforcement Learning: A Hierarchical Bayesian Approach. *ICML*, 2007.

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

M. Yuan and Y. Lin. Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 2007.

B. Zhang and Y. Wang. Learning Structural Changes of Gaussian Graphical Models in Controlled Experiments. *UAI*, 2010.

A List of Brain Regions

We present the list of the 157 regions used in Section 9. Parenthesis, e.g. "(Left, Right) Amygdala", indicate that we used two regions: Left Amygdala and Right Amygdala.

- Cerebellum: Cerebellar Lingual
- Cerebellum: (Culmen, Declive, Pyramis, Tuber, Uvula) of Vermis
- Cerebellum: (Left, Right) (Cerebellar Tonsil, Culmen, Declive, Dentate, Fastigium, Inferior Semi-Lunar Lobule, Nodule, Pyramis, Tuber, Uvula)
- Cerebrum: Hypothalamus
- Cerebrum: (Left, Right) (Amygdala, Claustrum, Hippocampus, Pulvinar, Putamen)
- Cerebrum: (Left, Right) (Anterior, Lateral Dorsal, Lateral Posterior, Medial Dorsal, Midline, Ventral Anterior, Ventral Lateral, Ventral Posterior Lateral, Ventral Posterior Medial) Nucleus
- Cerebrum: (Left, Right) Brodmann area (1,2,. . . ,47)
- Cerebrum: (Left, Right) Caudate (Body, Head, Tail)
- Cerebrum: (Left, Right) (Lateral, Medial) Globus Pallidus
- Brainstem: (Left, Right) (Mammillary Body, Red Nucleus, Substantia Nigra, Subthalamic Nucleus)