ON l_q ESTIMATION OF SPARSE INVERSE COVARIANCE

Goran Marjanovic, Member, IEEE and Alfred O. Hero, Fellow, IEEE

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Recently, major attention has been given to penalized log-likelihood estimators for sparse precision (inverse covariance) matrices. The penalty is responsible for inducing sparsity, and a very common choice is the convex l_1 norm. However, it is not always the case that the best estimator is achieved with this penalty. So, to improve sparsity and reduce biases associated with the l_1 norm, one must move to non-convex penalties such as the l_q ($0 \leq q < 1$). In this paper we introduce the resulting non-concave l_q penalized log-likelihood problem, and derive the corresponding optimality conditions. A novel cyclic descent algorithm is presented for penalized log-likelihood optimization, and we show how the derived conditions can be used to reduce algorithm computation. We illustrate by comparing reconstruction quality over the range $0 \leq q \leq 1$ for several experiments.

Index Terms— sparsity, l_q penalty, non-convex, precision matrix, optimality conditions.

1. INTRODUCTION

Graphical models have a long history [1–3] and provide a systematic way of reducing large dimensional data. The structure of the graph identifies meaningful interactions among the data variables. Assuming the data is Gaussian with mean $\boldsymbol{\mu} = \mathbf{0}_{p\times 1}$ and covariance $\boldsymbol{\Sigma}_{p\times p}$, the graphical model is an undirected graph specified by the precision matrix $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$. Specifically, no edge between nodes i and j ($i \neq j$) in the graph denotes the conditional independence of variables i and j given the other variables, which in turn corresponds to having $\boldsymbol{\Omega}(i,j) = 0$, [1–3]. Thus, obtaining an undirected graph is equivalent to obtaining $\boldsymbol{\Omega}$.

Following the parsimony principle, the objective is to choose the simplest model, i.e., the sparsest graph that adequately explains the data. The sparsity requirement improves the interpretability of the model and prevents over-fitting. Equivalently, in order to estimate a sparse $\boldsymbol{\Omega}$, attention has been given to maximizing a sparsity **P**enalized **L**og-**L**ikelihood (**PLL**) objective function. The log-likelihood measures the goodness-of-fit of the estimator while the penalty forces many of its entries to become zero. The most common sparsity penalties can be characterised using the l_q “norm”[1], which for $0 < q \leq 1$ is defined by:

$$\|\boldsymbol{\Omega}\|_q := \left(\sum_{i=1}^{p}\sum_{j=1}^{p}|\boldsymbol{\Omega}(i,j)|^q\right)^{\frac{1}{q}} \tag{1}$$

The function $\|\boldsymbol{\Omega}\|_q^q$ approaches the total number of non-zero entries in $\boldsymbol{\Omega}$ as $q \to 0^+$. Thus, for $q = 0$ the l_q penalty, known as the l_0 “norm”, is defined as this limit, and denoted by: $\|\boldsymbol{\Omega}\|_0^0$. So, for $0 \leq q \leq 1$ the l_q PLL problem is:

$$\max_{\boldsymbol{\Omega}\succ 0} \mathcal{L}_{\lambda,q}(\boldsymbol{\Omega}) := \log\det(\boldsymbol{\Omega}) - \mathrm{tr}(\boldsymbol{\Omega}\mathbf{S}) - \lambda\|\boldsymbol{\Omega}^-\|_q^q \tag{2}$$

where $\boldsymbol{\Omega}$ is symmetric ($\boldsymbol{\Omega}^T = \boldsymbol{\Omega}$), $\mathbf{S} \succeq 0$ is the sample covariance matrix, $\lambda > 0$ is a constant penalty parameter, and $\boldsymbol{\Omega}^-$ is $\boldsymbol{\Omega}$ with diagonal entries set to zero.

Even though l_0 ($q = 0$) is the most natural penalty in (2), the l_1 norm ($q = 1$) has become the dominant sparsity promoting penalty, see [4–11]. This is motivated by the convexity of $\|\boldsymbol{\Omega}\|_1$ and the use of the l_1 norm in sparse linear regression [12]. Its convexity makes (2) a desirable concave problem, allowing global solutions to be obtained.

It has recently been noted that using non-convex penalties can alleviate the biases of the l_1 norm for estimation of sparse precision matrices and similar problems, see [13–24]. It is expected that the l_q “norm” with $q < 1$ can achieve this and, at the same time, result in more aggressive shrinking/hard-thresholding, which produces relatively sparser solutions. The advantages of l_q penalties for $0 \leq q < 1$ have been shown for related estimation problems [15, 16, 20–25]. This paper establishes that similar improvements can be achieved for the PLL problem (2).

The remainder of the paper is organised as follows: Section 2 relates to prior work. Section 3 gives the optimality conditions for problem (2) when $0 \leq q < 1$, and Section 4 states a novel algorithm for its optimization. Section 5 shows how algorithm computation can be reduced by considering the optimality conditions, and provides estimator comparisons for $0 \leq q \leq 1$. Section 6 has concluding remarks.

Notation: For $\boldsymbol{\Omega}_{p\times p}$ matrix, $\boldsymbol{\Omega}_{/i/j}$ is a $(p-1)\times(p-1)$ sub-matrix produced by removing row i and column j in $\boldsymbol{\Omega}$. $\boldsymbol{\beta}_{-i}$ denotes $\boldsymbol{\beta}$ with the i-th entry set to 0. sgn(β) is the sign of $\beta \neq 0$ and 0 otherwise. $\mathbf{e}_i$ denotes a vector with 1 in the i-th entry and 0 in the rest. $1_{\{\cdot\}}$ is the indicator function, equaling 1 if the statement in $\{\cdot\}$ holds, and zero otherwise.

2. RELATION TO PRIOR WORK

Optimality Conditions: The existing literature on the PLL problem has predominantly focused on the concave l_1 PLL formulation, i.e., (2) with $q = 1$. Some references include [4–11], bearing in mind that even though [7] mentions the use of the l_q penalty, the work is limited to $q \geq 1$. The optimality conditions for $q = 1$ are well known, for example see [4]. They can easily be derived and follow directly from the convex l_1 linear regression setting studied, for example, in [26, 27].

This research was partially supported by AFSOR grant FA9550-13-1-0043.

[1] The l_q function is not a norm for $0 \leq q < 1$.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Other non-concave PLL formulations are given in [13,14,17,19]. The primary focus of [14, 17, 19] is not the l_q penalty ($q < 1$) and the work does not derive optimality conditions. The same is true for [13], even though it focused solely on the l_0 PLL problem. So, as far as we are aware, except for [13], the non-concave l_q PLL formulation with $q < 1$ has never been studied before, and hence, the derivation of the corresponding optimality conditions has never been attempted. We lastly note that, in order to derive these conditions one cannot apply the convexity flavoured ideas and arguments used when $q = 1$. Consequently, we take an alternative approach that takes advantage of the features unique to the objective function in (2).

Cyclic Descent (CD): The general CD algorithm is given in [28, 29], while the CD algorithms that were specifically designed for the PLL problem are given in [4, 6, 9, 11]. None of these, however, can handle (2) when $q < 1$, i.e., they do not provide a way to construct CD updates for $0 \leq q < 1$. Our novel algorithm is a block type CD method shown in the first author's recent PhD thesis [22]. Other block CD methods for the PLL problem are given in [4, 6], but these can only handle (2) with $q = 1$. More specifically, they are derived using duality arguments for convex/concave objective functions, which are not applicable when $0 \leq q < 1$. As a result, our method is fundamentally different and is derived by direct arguments.

3. OPTIMALITY CONDITIONS

In this section we derive the necessary optimality conditions for problem (2) with $0 \leq q < 1$. To do this, the following two theorems are needed:

Theorem 1. *For $0 \leq q < 1$ consider the (scalar) problem:*

$$\min_{\beta} \frac{1}{2}(z-\beta)^2 + \lambda|\beta|^q \tag{3}$$

where $|\beta|^q := 1_{\{\beta\neq 0\}}$ for $q = 0$. Then, all its solutions are:

$$\mathcal{T}_\lambda(z) = \begin{cases} 0 & \text{if} \quad |z| < h_\lambda \\ 0 & \text{if} \quad |z| = h_\lambda \\ \text{sgn}(z)\beta_\lambda & \text{if} \quad |z| = h_\lambda \\ \text{sgn}(z)\hat{\beta} & \text{if} \quad |z| > h_\lambda \end{cases} \tag{4}$$

where

$$\beta_\lambda := [2\lambda(1-q)]^{\frac{1}{2-q}} \text{ and } h_\lambda := \frac{1}{2}\left(\frac{2-q}{1-q}\right)\beta_\lambda \tag{5}$$

and

$$\hat{\beta} = \begin{cases} |z| & \text{if} \quad q = 0 \\ |z| - \lambda q \hat{\beta}^{q-1} \in (\beta_\lambda, |z|) & \text{if} \quad 0 < q < 1 \end{cases} \tag{6}$$

For $0 < q < 1$, $\hat{\beta}$ is found by iterating: $\beta_{k+1} = |z| - \lambda q \beta_k^{q-1}$ using $\beta_\lambda \leq \beta_0 \leq |z|$.

Proof: See first author's [15, Theorem 1 and Remark 3].

Theorem 2. *Consider the following block partitions of $p \times p$ symmetric matrices $\mathbf{\Omega}_\pi \succ 0$ and $\mathbf{S}_\pi \succeq 0$:*

$$\mathbf{\Omega}_\pi = \begin{bmatrix} \mathbf{V} & \mathbf{u} \\ \mathbf{u}^{\mathrm{T}} & u_0 \end{bmatrix}, \quad \mathbf{S}_\pi = \begin{bmatrix} \mathbf{\Gamma} & \boldsymbol{\gamma} \\ \boldsymbol{\gamma}^{\mathrm{T}} & \gamma_0 \end{bmatrix} \tag{7}$$

where $\mathbf{V} \succ 0$ and $\mathbf{\Gamma} \succeq 0$ are $(p-1)\times(p-1)$ symmetric, $\mathbf{u}, \boldsymbol{\gamma}$ are $(p-1)\times 1$, and $u_0, \gamma_0 > 0$ are scalars. Denote the i-th column of $\mathbf{V}^{-1}$ by $\mathbf{v}_i^-$, and its i-th entry by $v_{ii}^- > 0$. For any $i \in \{1, \ldots, p-1\}$, define:

$$\hat{\mathbf{u}} := \mathbf{u}_{-i} + \mathcal{T}_{\lambda_i}(z_i)\mathbf{e}_i, \quad \hat{u}_0 := \hat{\mathbf{u}}^{\mathrm{T}}\mathbf{V}^{-1}\hat{\mathbf{u}} + \gamma_0^{-1} \tag{8}$$

where $\mathcal{T}_\lambda(\cdot)$ is given by (4), and:

$$z_i := -\frac{\gamma_0 \mathbf{u}_{-i}^{\mathrm{T}}\mathbf{v}_i^- + \boldsymbol{\gamma}(i)}{\gamma_0 v_{ii}^-}, \quad \lambda_i := \frac{\lambda}{\gamma_0 v_{ii}^-} \tag{9}$$

Then, the following inequality holds for any $0 \leq q < 1$:

$$\mathcal{L}_{\lambda,q}\left(\begin{bmatrix} \mathbf{V} & \mathbf{u} \\ \mathbf{u}^{\mathrm{T}} & u_0 \end{bmatrix}\right) \leq \mathcal{L}_{\lambda,q}\left(\begin{bmatrix} \mathbf{V} & \hat{\mathbf{u}} \\ \hat{\mathbf{u}}^{\mathrm{T}} & \hat{u}_0 \end{bmatrix}\right) \tag{10}$$

For the proof see the Appendix. Theorem 2 strongly motivates the necessary optimality conditions for (2) with $0 \leq q < 1$:

Theorem 3. **(l_q PLL Optimality Conditions)** *Given a symmetric $\mathbf{\Omega}_{p\times p} \succ 0$, define the index set of off-diagonal zero entries $\mathcal{Z}(\Omega) := \{(i,j) : \mathbf{\Omega}(i,j) = 0\}$, the off-diagonal non-zero entries $\mathcal{Z}^{\mathrm{C}}(\Omega) := \{(i,j) : \mathbf{\Omega}(i,j) \neq 0\}$, and the diagonal entries $\mathcal{D}(\Omega)$. If $\mathbf{\Omega}$ is a (global) solution of (2) with $0 \leq q < 1$, then:*

$\mathbf{C}_1$: *For $(i,j) \in \mathcal{Z}(\Omega)$,*

$$\left|\mathbf{\Omega}^{-1}(i,j) - \mathbf{S}(i,j)\right| \leq \left\{\mathbf{S}(j,j)\mathbf{\Omega}_{/j/j}{}^{-1}(i,i)\right\}^{\frac{1-q}{2-q}} h_\lambda$$

$\mathbf{C}_2$: *For $(i,j) \in \mathcal{Z}^{\mathrm{C}}(\Omega)$,*

$$|\mathbf{\Omega}(i,j)| \geq \left\{\mathbf{S}(j,j)\mathbf{\Omega}_{/j/j}{}^{-1}(i,i)\right\}^{-\frac{1}{2-q}} \beta_\lambda$$

$\mathbf{C}_3$: *For $(i,j) \in \mathcal{Z}^{\mathrm{C}}(\Omega)$,*

$$\mathbf{\Omega}^{-1}(i,j) - \mathbf{S}(i,j) - \lambda q|\mathbf{\Omega}(i,j)|^{q-1}\text{sgn}\left(\mathbf{\Omega}(i,j)\right) = 0$$

$\mathbf{C}_4$: *For $(j,j) \in \mathcal{D}(\Omega)$,*

$$\mathbf{\Omega}^{-1}(j,j) = \mathbf{S}(j,j)$$

where β_λ and h_λ are from (5) in Theorem 1

For the proof see the Appendix.

Remark 1. *Letting $\frac{0}{0} := 1$, $\mathbf{C}_{1-4}$ reduce to the necessary optimality conditions for (2) when $q = 1$. In general, having optimality conditions is useful when considering algorithm development, initialization, stopping criteria and convergence analysis, for example see [4, 26, 27, 30]. In Section 5 we exploit some of these conditions for initialization and reducing algorithm computation.*

4. THE ALGORITHM

Inspired by Theorem 2, we state the block CD algorithm:

The l_qCOV Algorithm

Initialization: Choose a diagonal $\mathbf{\Omega} \succ 0$ and compute $\mathbf{\Omega}^{-1}$. Then, for $k = 1, 2, \ldots, p, 1, 2, \ldots, p, \ldots$, repeat:

(1) Let $\mathbf{\Omega}$ be the current iterate with the k-th column denoted by the vector $[\mathbf{u}^T u_0]^T$, where $u_0 = \mathbf{\Omega}(k,k)$. Similarly, denote the k-th column of $\mathbf{S}$ by $[\boldsymbol{\gamma}^T \gamma_0]^T$. Both $\mathbf{u}$ and $\boldsymbol{\gamma}$ are $(p-1)\times 1$. Let $\mathbf{V} := \mathbf{\Omega}_{/k/k}$, and $\mathbf{\Gamma} := \mathbf{S}_{/k/k}$.

(2) Calculate $\mathbf{V}^{-1}$, and update each $\mathbf{u}(i) = \mathbf{\Omega}(i,k)$, cyclically using (8). With the updated $\mathbf{u}$, lastly update u_0, using (8).

(3) Using $\mathbf{\Omega}$ with the updated k-th row/column, update $\mathbf{\Omega}^{-1}$ and $\mathbf{V}^{-1}$ using the standard formula for block matrix inversion.

Theorem 4. (l_qCOV): if $\mathbf{\Omega} \succ 0$ and $\mathbf{\Omega}^+$ is the current and the next iterate respectively, $\mathbf{\Omega}^+ \succ 0$ and $\mathcal{L}_{\lambda,q}(\mathbf{\Omega}) \leq \mathcal{L}_{\lambda,q}(\mathbf{\Omega}^+)$.

Proof: Follows easily from Theorem 2.

5. SIMULATIONS

Here we compare the quality of estimators: $\hat{\mathbf{\Omega}}$ obtained by optimizing (2). For $0 \leq q < 1$, l_qCOV is used, while for $q = 1$ we consider the weighted GLASSO method in [4], where the diagonal weights in the penalty are set to 0 and the rest to 1.

Initialization and Stopping using $\mathbf{C}_{1-4}$: All algorithms are initialized with a diagonal matrix, and so, we make the diagonal entries satisfy $\mathbf{C}_4$. Thus, they are given by $1/\mathbf{S}(i,i)$.

All algorithms are terminated when $\mathbf{C}_{1-4}$ are satisfied. To measure the progress, for a particular iterate $\mathbf{\Omega}$, we let the "distance to optimality" $d_{\mathbf{\Omega}}(i,j)$ be 1 if the optimality conditions for (i,j) in Theorem 3 are not satisfied, and 0 otherwise. Then, the total distance is $d(\mathbf{\Omega}) := \sum_{i,j} d_{\mathbf{\Omega}}(i,j)$.

Reducing Computation with $\mathbf{C}_1$: Inspired by the computation reducing method for CD methods with $q = 1$ in [9, 31], we suggest a corresponding method when $0 \leq q < 1$.

The method in [9, 31] involves updating a subset of the entries in $\mathbf{\Omega}$, called "Active Entries", in each iteration with the intent of reducing algorithm computation. In [31], the "Active Entries" are the non-zero entries, while in [9] they are additionally the zero entries that do not satisfy $|\mathbf{\Omega}^{-1}(i,j) - \mathbf{S}(i,j)| \leq \lambda$. This is precisely $\mathbf{C}_1$ when $q \to 1$, see Remark 1. As a result, we extend this idea to $0 \leq q < 1$: In Step (2) of l_qCOV, the "Active Entries" are those for which $\mathbf{\Omega}(i,j) \neq 0$, or $\mathbf{\Omega}(i,j) = 0$ and $\mathbf{C}_1$ does not hold. $\mathbf{C}_1$ involves calculating ${\mathbf{\Omega}_{/j/j}}^{-1}$ but this is already given in Step (2). Due to lack of space, the theoretical aspects of this method will be presented elsewhere. Fig.1 presents experimental results.

Simulations for Random & Star Graph Configurations: Two popular types of graph configurations of size $p = 50$ are considered for recovery: random and star, with corresponding precision matrices respectively denoted by $\mathbf{\Omega}_R$ and $\mathbf{\Omega}_S$. The former is constructed randomly using the matlab function `sprandsym` and the latter by using the procedure in [22, p.138]. The non-zero elements in both $\mathbf{\Omega}_R$ and $\mathbf{\Omega}_S$ are drawn from the Gaussian distribution.

In all simulations, the corresponding $\mathbf{S}$ are constructed with $n = 30$ data instances drawn from the Gaussian distribution with mean $\boldsymbol{\mu} = \mathbf{0}$ and respective covariances $\mathbf{\Omega}_R^{-1}$ and $\mathbf{\Omega}_S^{-1}$. We constrained $||\mathbf{\Omega}_R||_0^0 = ||\mathbf{\Omega}_S||_0^0 = 0.05 \times p^2$. Fig.2 presents experimental results.

We also consider recovering $\mathbf{\Omega}$ constructed by stochastically combining $\mathbf{\Omega}_R$ and $\mathbf{\Omega}_S$. In this case, $\mathbf{\Omega}(i,j) \neq 0$ if $z_{ij} \neq 0$, where z_{ij} is a Bernoulli random variable with probability parameter:

$$p_{ij} = (1-\alpha)1_{\{\mathbf{\Omega}_R(i,j)\neq 0\}} + \alpha 1_{\{\mathbf{\Omega}_S(i,j)\neq 0\}} \qquad (11)$$

and $\alpha \in (0,1)$. When $z_{ij} \neq 0$ we let:

$$\mathbf{\Omega}(i,j) = (1-\alpha)\mathbf{\Omega}_R(i,j) + \alpha\mathbf{\Omega}_S(i,j) \qquad (12)$$

Thus, $||\mathbf{\Omega}(\alpha)||_0^0 = 0.05 \times p^2$ is consistent for any $\alpha \in [0,1]$.

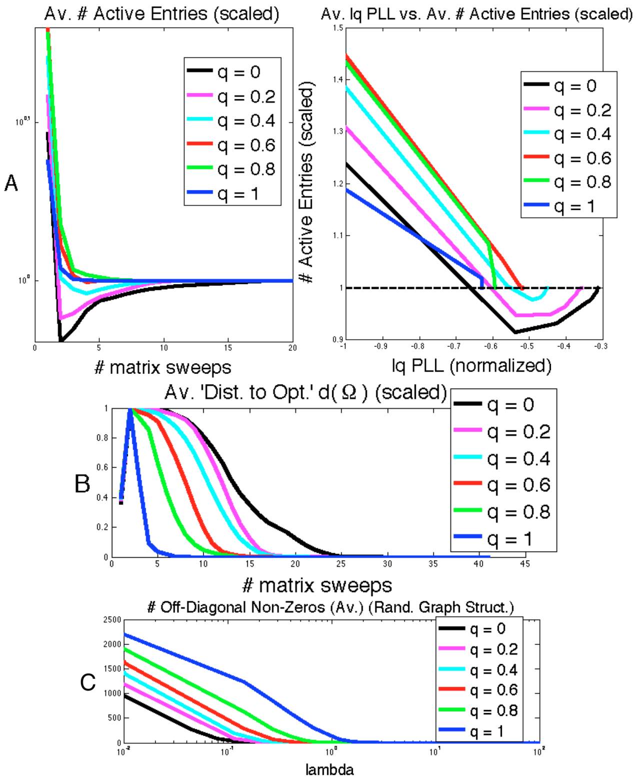

Fig. 1: Examples of algorithm runs (# variables is $p = 50$ and # samples is $n = 30$): $\mathbf{\Omega}_{p\times p}$ has random configuration and entries, and is generated using the matlab function `sprandsym` s.t. $||\mathbf{\Omega}||_0^0 = 0.2 \times p^2$. The tuning parameter λ was chosen s.t. $||\hat{\mathbf{\Omega}}||_0^0 = ||\mathbf{\Omega}||_0^0$. The (av.) # "Active Entries" in panel A are scaled by $||\mathbf{\Omega}||_0^0$. $\mathcal{L}_{\lambda,q}(\cdot)$ is normalized with its initial value. In panel B, $d(\cdot)$ is scaled by the number of trials. We see from panel A that the number of matrix entries needed to be updated in l_qCOV becomes much less than p^2 for all $0 \leq q \leq 1$. As it can be seen from panel B, all $\hat{\mathbf{\Omega}}$ end up satisfying $\mathbf{C}_{1-4}$. We found this to hold in general. Panel C shows (av.) $||\hat{\mathbf{\Omega}}||_0^0 - p$ as λ changes. For a given λ we see that $\hat{\mathbf{\Omega}}$ is sparser as $q \to 0$.

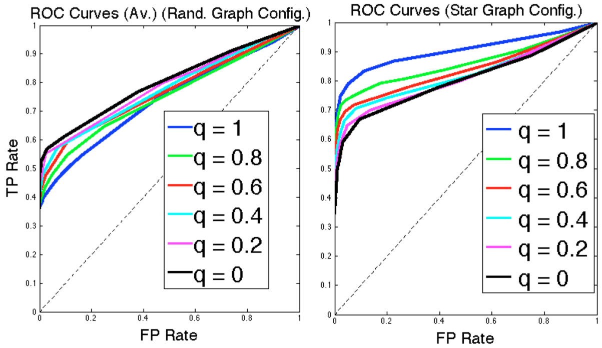

Fig. 2: (av.) ROC (**R**eceiver **O**perating **C**haracteristic) curves for (left) $\hat{\mathbf{\Omega}}_R$ and, (right) $\hat{\mathbf{\Omega}}_S$. The ROC measures structure, and we see that the best structure is convincingly achieved when (left) $q = 0$, and (right) $q = 1$. The (av.) $\min_\lambda$ **K**ullback-**L**eibler (KL(λ, q)) loss is also achieved when (left) $q = 0$, and (right) $q = 1$.

The ROC results were similar to those in Fig.2, i.e., we also found that the curves favoured $q = 0$ and $q = 1$ when $\alpha < 0.5$ and $\alpha > 0.5$ respectively. However, they approached each other as $\alpha \to 0.5$. For $\alpha < 0.5$, the (av.) $\min_\lambda$ **K**ullback-**L**eibler (KL(λ, q)) loss is achieved when $q = 0$, while for $\alpha > 0.5$, it is achieved when $q = 1$. Fig.3 summarizes this.

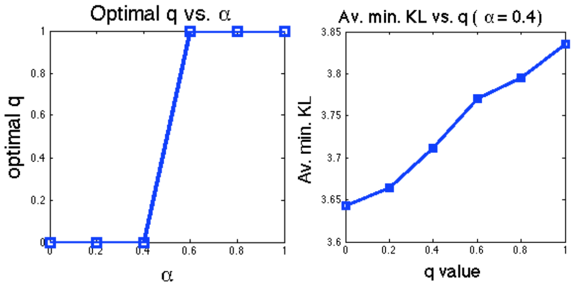

Fig. 3: (left) Structurally (ROC) and in terms of KL, as it can be seen, the best $\hat{\mathbf{\Omega}}(\alpha)$'s are obtained only with binary values of q. (right) q vs. (av.) $\min_\lambda$ KL((λ, q)). When $\alpha = 0.4$, we see that the smallest KL is achieved when $q = 0$.

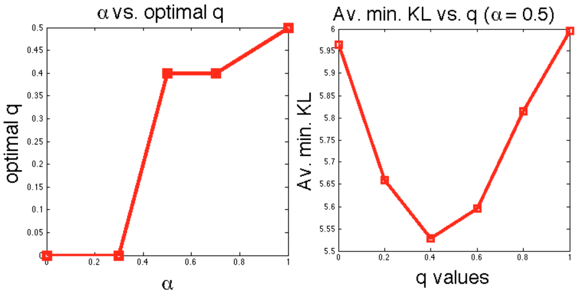

Fig. 4: We noticed that for each α considered, the ROC curves for $\hat{\mathbf{\Omega}}(\alpha)$ were approximately equal for all $0 \leq q \leq 1$. As a result, this implies equal quality in terms of estimator structure (since ROC measures only structure), and so, we can not reject the hypothesis that the optimal q is again either 0 or 1. Thus, to distinguish the estimators, we further looked at the KL measure, which takes into account not only structure but estimator values. In this case, we noticed that the best quality $\hat{\mathbf{\Omega}}(\alpha)$ is achieved with $0 \leq q < 1$. (left) The optimal q based on the KL loss as α varies. We see that as sparsity decreases ($\alpha \to 1$) the optimal q increases. (right) q vs. (av.) $\min_\lambda$ KL(λ, q) for $\alpha = 0.5$, corresponding to $\|\mathbf{\Omega}(\alpha)\|_0^0 = 0.15 \times p^2$. For this example, the optimal $q = 0.4$.

Simulations for Random Graph Configurations: Here we replace $\mathbf{\Omega}_S$ with a less sparse $\mathbf{\Omega}_R$, denoted by $\mathbf{\Omega}'_R$, where $\|\mathbf{\Omega}'_R\|_0^0 = 0.2 \times p^2$. With this change we use $\mathbf{\Omega}_R$ and $\mathbf{\Omega}'_R$ in (11) and (12) to stochastically construct $\mathbf{\Omega}(\alpha)$, except this time α varies the sparsity in $\mathbf{\Omega}(\alpha)$ between 5% and 20%. Fig.'s 4 and 5 present the experimental results.

6. CONCLUSION

We introduced the non-concave l_q penalized log-likelihood problem ($0 \leq q < 1$) for Gaussian graphical models, and derived the corresponding optimality conditions. A novel coordinate descent algorithm was given, and we showed how the derived conditions can be used to reduce computation. Simulations showed that $0 \leq q < 1$ can be used to improve on $q = 1$.

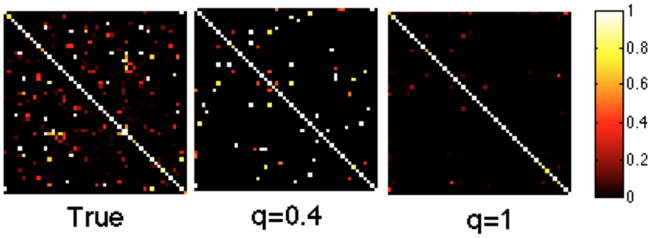

Fig. 5: For $\alpha = 0.5$, we show the size of entries in (left) $\mathbf{\Omega}(\alpha)$, and (centre) $\hat{\mathbf{\Omega}}(\alpha)$ with $q = 0.4$, and (right) $\hat{\mathbf{\Omega}}(\alpha)$ with $q = 1$. The penalty parameter λ is tuned s.t. both $\hat{\mathbf{\Omega}}(\alpha)$'s have an FP rate of 0.02. Both $\hat{\mathbf{\Omega}}(\alpha)$'s have an approximately equal number of non-zeros too, but despite this, we see that LASSO ($q = 1$) produces non-zero entries that are severely shrunken towards zero, unlike the l_q-penalized reconstruction with $q = 0.4$. The same phenomenon was observed for a wide range of the FP rate, and confirms the negative bias of the l_1 norm. This difference between l_1 and l_q is due to the magnitude over-penalization by the linear l_1 penalty as compared to the sub-linear l_1 penalty for $q = 0.4$.

7. APPENDIX

Proof of Theorem 2: By the properties of the determinant and trace for block matrices, $\mathcal{L}_{\lambda,q}(\mathbf{\Omega}_\pi) = \mathcal{L}_{\lambda,q}(\mathbf{V}) + \log\left(u_0 - \mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}\right) - 2\boldsymbol{\gamma}^T\mathbf{u} - \gamma_0 u_0 - 2\lambda\|\mathbf{u}\|_q^q$, which is differentiable w.r.t. u_0. So, setting its derivative to zero and solving for u_0 we obtain that $\hat{u}_0$ is the maximizer. Substituting $\hat{u}_0$ in for u_0, it can easily be shown that: $\mathcal{L}_{\lambda,q}(\mathbf{\Omega}_\pi) \leq \mathcal{L}_{\lambda,q}(\mathbf{V}) - c - 2J(\mathbf{u})$, where $c = \log(\gamma_0) + 1$ and $J(\mathbf{u}) = \frac{1}{2}\gamma_0\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u} + \boldsymbol{\gamma}^T\mathbf{u} + \lambda\|\mathbf{u}\|_q^q$. To obtain an additional inequality for $\mathcal{L}_{\lambda,q}(\cdot)$ we derive an inequality for $J(\cdot)$.

Substituting $\mathbf{u} = \mathbf{u}_{-i} + \mathbf{u}(i)\mathbf{e}_i$ in $J(\mathbf{u})$, and using Theorem 1 it is easy to obtain $\mathcal{T}_{\lambda_i}(z_i) = \arg\min_{\mathbf{u}(i)} J(\mathbf{u})$. So, $J(\hat{\mathbf{u}}) \leq J(\mathbf{u})$, which finally gives (10). By Sylvester's criterion, $\hat{u}_0 > 0$ implies the matrix with $\mathbf{V}$, $\hat{\mathbf{u}}$ and $\hat{u}_0$ is positive definite.

Proof of Theorem 3: Let $\mathbf{\Omega}_\pi$ and $\mathbf{S}_\pi$ denote the respective π-permutations of $\mathbf{\Omega}$ and $\mathbf{S}$, i.e., having the j-th row and column of $\mathbf{\Omega}$ and $\mathbf{S}$ placed last, see [4, 6, 13, 22]. Then, in Theorem 2, we have: $\mathbf{V} = \mathbf{\Omega}_{/j/j}$ and $\mathbf{\Gamma} = \mathbf{S}_{/j/j}$, and $\mathbf{u}$ and $\boldsymbol{\gamma}$ become the respective j-th columns of $\mathbf{\Omega}$ and $\mathbf{S}$ truncated not to include the j-th entry. Also, $u_0 = \mathbf{\Omega}(j,j)$ and $\gamma_0 = \mathbf{S}(j,j)$. Since the log-likelihood and the l_q penalty are invariant under this π-permutation, we have $\mathcal{L}_{\lambda,q}(\mathbf{\Omega}) = \mathcal{L}_{\lambda,q}(\mathbf{\Omega}_\pi)$. Thus, we can use (10): since $\mathbf{\Omega}_\pi$ maximizes $\mathcal{L}_{\lambda,q}(\cdot)$, we must have **(i)** $u_0 = \hat{u}_0$, and $\mathbf{u} = \mathbf{u}_{-i} + \mathcal{T}_{\lambda_i}(z_i)\,\mathbf{e}_i$ that reduces to **(ii)** $\mathbf{u}(i) = \mathcal{T}_{\lambda_i}(z_i)$ for each $i \neq j$. So, using **(i)** in the block matrix inversion formula to invert $\mathbf{\Omega}_\pi$, we obtain that:

$$(\mathbf{\Omega}_\pi)^{-1} = \begin{bmatrix} \cdots & -\gamma_0\mathbf{V}^{-1}\mathbf{u} \\ -\gamma_0\mathbf{u}^T\mathbf{V}^{-1} & \gamma_0 \end{bmatrix} \qquad (13)$$

Since $(\mathbf{\Omega}_\pi)^{-1} = (\mathbf{\Omega}^{-1})_\pi$, in (13) we firstly see that $\mathbf{\Omega}^{-1}(j,j) = \gamma_0$, giving $\mathbf{C}_4$, and secondly, that $-\gamma_0\mathbf{V}^{-1}\mathbf{u}$ is the j-th column of $\mathbf{\Omega}^{-1}$ truncated not to include $\mathbf{\Omega}^{-1}(j,j)$. Its i-th entry is $-\gamma_0\mathbf{u}^T\mathbf{v}_i^-$, which is thus $\mathbf{\Omega}^{-1}(i,j)$. Since $\mathbf{u}(i) = \mathbf{\Omega}(i,j)$, this allows us to re-write the numerator in z_i as $\mathbf{\Omega}^{-1}(i,j) - \mathbf{S}(i,j) + \gamma_0 v_{ii}^-\mathbf{\Omega}(i,j)$.

So, to derive $\mathbf{C}_{1-3}$, we use **(ii)**, Theorem 1 and the definition of z_i. When $\mathbf{u}(i) = 0$ we have: $|z_i| \leq h_{\lambda_i}$, which simplifies to $\mathbf{C}_1$. When $\mathbf{u}(i) \neq 0$, we have $|z_i| \geq h_{\lambda_i}$ implying $|\mathbf{u}(i)| \geq \beta_{\lambda_i}$, which simplifies to $\mathbf{C}_2$. Using (6) we obtain $\mathbf{C}_3$ after re-arranging and simplifying.

8. REFERENCES

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

[2] J. Whittaker, *Graphical Models in Applied Mathematical Analysis*. New York: Wiley, 1990.

[3] S. Lauritzen, *Graphical Models*. Oxford: Oxford University Press, 1996.

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

[5] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[6] O. Banerjee, L. Ghaoui, and A. dAspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, 2008.

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

[8] K. Scheinberg, S. Ma, and D. Goldfarb, "Sparse inverse covariance selection via alternating linearization methods," 2010, http://books.nips.cc/papers/files/nips23/NIPS2010_0109.pdf.

[9] C. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "Sparse inverse covariance matrix estimation using quadratic approximation," 2013, arXiv: http://arxiv.org/abs/1306.3212.

[10] C. Hsieh, I. S. Dhillon, P. Ravikumar, and A. Banerjee, "A divide-and-conquer method for sparse inverse covariance estimation," *NIPS*, vol. 24, 2012.

[11] K. Scheinberg and I. Rish, "Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach," *Lect. Notes Comput. Sc.*, vol. 6323, pp. 196–212, 2010.

[12] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[13] G. Marjanovic and V. Solo, "l_0 sparse graphical modeling," *IEEE ICASSP*, pp. 2084–2087, 2011.

[14] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, pp. 1348–1360, 2001.

[15] G. Marjanovic and V. Solo, "On l_q optimization and matrix completion," *IEEE T. Signal Proces.*, vol. 60, no. 11, pp. 5714–5724, 2012.

[16] ——, "l_q matrix completion," *IEEE ICASSP*, pp. 3885–3888, 2012.

[17] J. Fan, Y. Feng, and Y. Wu, "Network exploration via the adaptive LASSO and SCAD penalties," *Ann. Appl. Stat.*, vol. 3, no. 2, pp. 521–541, 2009.

[18] G. Marjanovic and V. Solo, "On exact l_q denoising," *IEEE ICASSP*, pp. 6068–6072, 2013.

[19] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Ann. Appl. Stat.*, vol. 37, no. 6, pp. 4254–4278, 2009.

[20] A. Seneviratne and V. Solo, "On vector l_0 penalized multivariate regression," *IEEE ICASSP*, pp. 3613–3616, 2012.

[21] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, 2010.

[22] G. Marjanovic, "l_q sparse signal estimation with applications," PhD Thesis, 2013, http://www.unsworks.unsw.edu.au.

[23] B. D. Rao and K. K. Delgado, "An affine scaling methodology for best basis selection," *IEEE T. Signal Proces.*, vol. 47, no. 1, pp. 187–200, 1999.

[24] G. Marjanovic and V. Solo, "l_q sparsity penalized linear regression with cyclic descent," *IEEE T. Signal Proces.*, vol. 62, no. 6, pp. 1464–1475, 2014.

[25] X. Chen, F. Xu, and Y. Ye, "Lower bound theory of nonzero entries in solutions of l_2-l_p minimization," *SIAM J. Sci. Comput.*, vol. 32, pp. 2832–2852, 2010.

[26] J. J. Fuchs, "Convergence of a sparse representations algorithm applicable to real or complex data," *IEEE J. Sel. Top. Signa.*, vol. 1, no. 4, pp. 598–605, 2007.

[27] S. Alliney and S. Ruzinsky, "An algorithm for the minimisation of mixed l_1 and l_2 norms with application to Bayesian estimation," *IEEE T. Signal Proces.*, vol. 42, pp. 618–627, 1994.

[28] D. Luenberger and Y. Ye, *Linear and Nonlinear Programming*. Springer Science, 2008.

[29] P. Huard, *Mathematical Programming Study 10*. North-Holland Publishing Company, 1979.

[30] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for l_1 regularization: A comparative study and two new approaches," *Lect. Notes Comput. Sc.*, pp. 286–297, 2007.

[31] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Soft.*, vol. 33, no. 1, pp. 1–22, 2010.