Supplementary Materials for *Learning Gaussian Graphical Models with Ordered Weighted ℓ_1 Regularization*

Cody Mazza-Anthony, Bogdan Mazoure, Mark Coates

May 2019

Dual Problem Formulation for GOWL

The following formulation uses the technique from [2]. Let $\mathbf{Z} = \mathbf{\Theta}$ and its associated dual variable be $\mathbf{W} \in \mathbb{R}^{p \times p}$, which gives the Lagrangian

$$\mathcal{L}(\mathbf{\Theta}, \mathbf{Z}, \mathbf{W}) = -\log\det\,\mathbf{\Theta} + \mathrm{tr}(\boldsymbol{S}\mathbf{\Theta}) + \sum_{i=1}^{K} \lambda_i |\mathbf{vechs}(\mathbf{Z})|_{[i]} + \mathrm{tr}(\mathbf{W}(\mathbf{\Theta} - \mathbf{Z})), \tag{1}$$

Note that the above quantity is separable into terms involving $\mathbf{\Theta}$ and $\mathbf{Z}$, thus allowing for computing the infimum over auxiliary variables. We first consider the terms involving $\mathbf{Z}$ or more precisely, its vectorized form $\mathbf{z} = \mathbf{vechs}(\mathbf{Z})$:

$$g(\mathbf{z}) = \lambda_{\downarrow}^{T} \mathbf{z}_{\downarrow} - \mathrm{tr}(\mathbf{WZ})\,.$$

By the generalized rearrangement inequality [4], we know that for arbitrary vectors $\lambda, \mathbf{w}$, $\lambda_{\downarrow}^{T}\mathbf{w}_{\uparrow} \leq \lambda^{T}\mathbf{w} \leq \lambda_{\downarrow}^{T}\mathbf{w}_{\downarrow}$ and hence if $\lambda_i \geq |\mathbf{w}_i|$ for $i = 1, ..., K$ and $\mathbf{w} = \mathbf{vechs}(\mathbf{W})$, then $\inf_{\mathbf{z}} g(\mathbf{z}) = 0$. Otherwise, the problem is unbounded and the infimum is attained at $g(\mathbf{z}) = -\infty$. Therefore

$$\inf_{\mathbf{Z}} \sum_{i=1}^{K} \lambda_i |\mathbf{Z}_{[i]}| - \mathrm{tr}(\boldsymbol{W}\mathbf{Z}) = \begin{cases} 0 & \text{if } \lambda_i \geq |\mathbf{w}|_i \text{ for } i = 1, ..., K, \\ -\infty & \text{otherwise}\,. \end{cases} \tag{2}$$

The infimum over terms involving $\mathbf{\Theta}$ can be computed by taking the gradient of the log determinant under the assumption that $\mathbf{S} + W \succ 0$ and is attained at the point:

$$\inf_{\mathbf{\Theta}} [-\log\det\,\mathbf{\Theta} + \mathrm{tr}((\mathbf{S} + \mathbf{W})\mathbf{\Theta})] = \log\det\,(\mathbf{S} + \mathbf{W}) + p\,.$$

Combining (1) with the constraint (2) allows us to write the dual as:

$$\begin{aligned} \max_{\mathbf{W} \succ 0} \quad & \log\det\,(\mathbf{S} + \mathbf{W}) \\ \text{s.t.} \quad & |\mathbf{W}_i| \leq \lambda_i, \\ & (\mathbf{S} + \mathbf{W}) \succ 0\,. \end{aligned} \tag{3}$$

We proceed to solve the problem shown in (3) using G-ISTA, which requires the computation of the duality gap $\Delta = p^* - d^*$ for primal p^* and dual d^*. If we define the feasible set $B_\lambda = \{\mathbf{W} :$

$|\mathbf{W}_i| \leq \lambda_i, \forall i, \mathbf{W} \in \mathbb{S}^p_{++}\}$, then for any feasible dual point $\mathbf{W} \in B_\lambda$ and corresponding primal point $\mathbf{\Theta} = (\mathbf{S} + \mathbf{W})^{-1}$, the duality gap Δ is defined by:

$$\begin{aligned}\Delta &= p^* - d^* \qquad (4)\\ &= \operatorname{tr}(\mathbf{S}\mathbf{\Theta}) + \sum_{i=1}^{K} \lambda_i |\mathbf{\Theta}|_{[i]} - p\,.\end{aligned}$$

Proof of Theorem 1

The following proof uses the technique from [2] and builds upon Slater's condition as formulated in the following theorem:

Theorem (Slater's condition)**.** *Let a constrained optimization problem be of the form:*

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \quad & f(\mathbf{x}) \qquad (5)\\ s.t \quad & g_i(\mathbf{x}) \leq 0 \quad (i = 1, \ldots, m),\\ & h_j(\mathbf{x}) = 0 \quad (j = 1, \ldots, p),\end{aligned}$$

We say that Slater constraint qualification (SCQ) holds for (5) *if there exists* $\hat{\mathbf{x}} \in \mathcal{X}$ *such that*

$$g_i(\hat{\mathbf{x}}) < 0 \quad \textit{and} \quad h_j(\hat{\mathbf{x}}) = 0\,.$$

Furthermore, if $\hat{\mathbf{x}} \in$ *interior*$(\mathcal{X})$*, then strong duality hold at the point* $\hat{\mathbf{x}}$*.*

First, assume that $c = \max_{i,j} |\mathbf{S}|_{ij}$ is known. In the case where $\mathbf{S}$ is standardized, $c = 1$. The negative log-likelihood is convex in the precision matrix and is defined over the set of all positive-definite matrices. On the other hand, the OWL estimator is also convex in $\mathbf{\Theta}$ over the same set. Since the sum of two convex functions over the same convex set is convex, we conclude that the main objective is convex.

Using SCQ, we can say that the duality gap is zero and write the primal-dual optimal pair in the following way:

$$\mathbf{\Theta}^* = (\mathbf{S} + \mathbf{W}^*)^{-1}\,.$$

For SCQ to hold, it remains to show that there exists a point $\mathbf{W}$ in the interior of the feasible set given by $\operatorname{interior}(B_\lambda) = \{\mathbf{W} : |\mathbf{W}_i| < \lambda_i, \forall i, \mathbf{W} \in \mathbb{R}^p\}$ such that it is the solution of (2).

The goal is to choose a $\mathbf{W}$ such that $(\mathbf{S} + \mathbf{W}) \succ 0$ and ensure that the entries $|\mathbf{W}_i|$ are close to zero. First recall that $\mathbf{S}$ is a symmetric positive semi-definite matrix and since it was estimated from data, we can assume that the diagonal entries will be greater than zero with probability one. Let $\mathbf{A} = \operatorname{diag}(\mathbf{S}) \succ 0$ since the determinant of a diagonal matrix with positive entries is positive. Consequently, by Sylvester's criterion $\mathbf{A}$ is positive definite (PD). We can then write the convex combination of $\mathbf{S}$ and $\mathbf{A}$ as

$$\alpha\mathbf{S} + (1 - \alpha)\mathbf{A} \succ 0\,.$$

where $\alpha \in [0, 1)$. The above expression is itself positive definite, which we can see by taking any

$\mathbf{x} \in \mathbb{R}^p$:

$$\begin{aligned}\mathbf{x}^T\Big(\alpha\mathbf{S} + (1-\alpha)\mathbf{A}\Big)\mathbf{x} &= \mathbf{x}^T(\alpha\mathbf{S})\mathbf{x} + \mathbf{x}^T((1-\alpha)\mathbf{A})\mathbf{x} \\ &= \alpha\underbrace{(\mathbf{x}^T\mathbf{S}\mathbf{x})}_{\geq 0} + (1-\alpha)\underbrace{(\mathbf{x}^T\mathbf{A}\mathbf{x})}_{>0} > 0 \\ &> 0\,.\end{aligned}$$

Thus, we can write

$$\mathbf{S} + \mathbf{W} = \alpha\mathbf{S} + (1-\alpha)\mathbf{A} \succ 0,$$

for non-negative α strictly smaller than 1. For a given matrix of hyperparameters $\mathbf{\Lambda}$, pick $\alpha > 1 - 1/c\min_{kl}\mathbf{\Lambda}_{kl}$. In practice this can be achieved by setting $\tilde{\alpha} = 1 - 1/c\min_{kl}\mathbf{\Lambda}_{kl}$ and putting $\alpha = \tilde{\alpha} + \varepsilon$ for some $\varepsilon > 0$. Then,

$$\begin{aligned}\mathbf{W} &= \alpha\mathbf{S} + (1-\alpha)\mathbf{A} - \mathbf{S} \\ &= (1-\alpha)(\mathbf{A} - \mathbf{S})\,,\end{aligned}$$

and hence

$$\begin{aligned}|\mathbf{W}|_{ij} &= (1-\alpha)|\mathbf{S}|_{ij} \qquad (i,j = 1,\ldots,p) \\ &< 1/c\min_k \lambda_k|\mathbf{S}_{ij}| \\ &\leq \min_k \lambda_k \\ &\leq \lambda_{ij}\,.\end{aligned}$$

where we used the fact that the values in the empirical estimate of the covariance matrix $\mathbf{S}$ cannot be infinite. By the convexity of the primal objective and SCQ, we conclude that $\mathbf{W}^*$ is unique. Furthermore, since the duality gap is zero at the point $\mathbf{W}^*$, the uniqueness of $\mathbf{W}^*$ implies the uniqueness of $\mathbf{\Theta}^*$. ■

Proof of Theorem 2

The following proof uses the technique from [1]. Suppose $\hat{\beta}_k \neq \hat{\beta}_l$, then by differentiating (26) we obtain

$$-2\mathbf{x}_k^T(\mathbf{X}_{*,j} - \mathbf{X}_{*,-j}\boldsymbol{\beta}_j) - \lambda_k = 0\,,$$

and

$$-2\mathbf{x}_l^T(\mathbf{X}_{*,j} - \mathbf{X}_{*,-j}\boldsymbol{\beta}_j) - \lambda_l = 0\,,$$

By subtracting the two equations we can write:

$$-2(\mathbf{x}_l^T - \mathbf{x}_k^T)(\mathbf{X}_{*,j} - \mathbf{X}_{*,-j}\boldsymbol{\beta}_j) + (\lambda_l - \lambda_k)\,.$$

We can say that $||\mathbf{X}_{*,j} - \mathbf{X}_{*,-j}\boldsymbol{\beta}_j||_2^2 \leq ||\mathbf{X}_{*,j}||_2^2$ and $||\mathbf{x}_l^T - \mathbf{x}_k^T||_2^2 = 2(1-\rho_{kl})$ as a result of $\mathbf{X}$ being standardized. So we can write

$$|\lambda_l - \lambda_k| \leq 2||\mathbf{X}_{*,j}||_2^2\sqrt{2(1-\rho_{kl})}\,.$$

We know when the weights are constructed that $|\lambda_l - \lambda_k| \geq c = |\lambda_1 - (\lambda_1 + \lambda_2(p-1))| = |\lambda_2(p-1)|$. So we arrive at a contradiction if $c > 2||\mathbf{X}_{*,j}||_2^2\sqrt{2(1-\rho_{kl})}$. ■

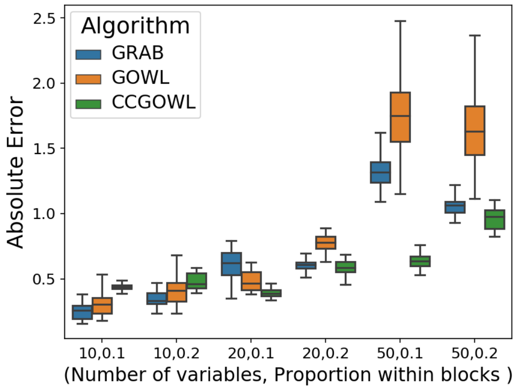

Figure 1: Absolute Error Values for GRAB, GOWL, and CCGOWL.

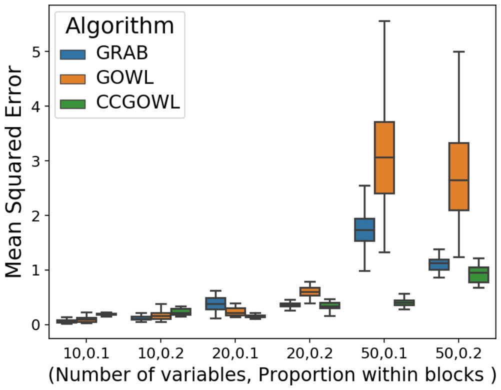

Figure 2: Mean Squared Error Values for GRAB, GOWL, and CCGOWL.

Table 1: GOWL Hyper-parameter Specification

p	Block %	λ_1	λ_2	Range Considered
10	10	0.03684211	0.01052632	(0, 0.1)
10	20	0.06842105	0.01052632	(0, 0.1)
20	10	0.06551724	0.00344828	(0, 0.1)
20	20	0.02413793	0.00344828	(0, 0.1)
50	10	0.008	0.00010	(0, 0.1)
50	20	0.006	0.00009	(0, 0.1)

Hyper-parameter Specification

Tables 1, 2, and 3 list the hyper-parameters chosen to produce the graph in Figure 2. The hyper-parameters were chosen using a 2-fold cross-validation procedure. For the gene dataset in Figure 3 the hyperparameters used were $\lambda_1 = 0.3$ and $\lambda_2 = 0.00612821$ and were chosen by using the λ_1, λ_2 provided by 2-fold cross-validation and then the sparsity hyperparameter was increased to encourage more sparsity. For the stock data set in Figure 4 the hyperparameters used were $\lambda_1 = 0.2$ and $\lambda_2 = 0.0001$. The hyperparameters were chosen arbitrarily as cross-validation required too much computational resources.

Error Fit Metrics

Absolute Squared Error and Mean Squared Error Measurements were recorded during synthetic data simulations. Figures 1 and 2 show the values recorded. Overall, the ccGOWL estimator had a lower mean squared error and absolute error for $p = 20$ and $p = 50$ and had higher errors for $p = 10$. These results are more convincing for larger p than examining weighted F_1 scores.

Table 2: ccGOWL Hyper-parameter Specification

p	Block %	λ_1	λ_2	Range Considered
10	10	0.10526316	0.05263158	(0, 0.1)
10	20	0.10526316	0.05263158	(0, 0.1)
20	10	0.23684211	0.00793103	(0, 0.1)
20	20	0.23684211	0.00793103	(0, 0.1)
50	10	0.1	0.00512821	(0, 0.1)
50	20	0.1	0.00512821	(0, 0.1)

Table 3: GRAB Hyper-parameter Specification

p	Block %	λ	Range Considered
10	10	0.1	(0, 1.0)
10	20	0.1	(0, 1.0)
20	10	0.7	(0, 1.0)
20	20	0.5	(0, 1.0)
50	10	0.5	(0, 1.0)
50	20	0.4	(0, 1.0)

Gene Expression Data

Table 4 shows which genes are highly expressed in individuals with either the ALL or AML disease as demonstrated in [3]. Figure 3 shows a plot for the GLASSO applied to gene expression data and regularization parameter was set to $\lambda = 0.4$.

Equities Data

Figure 4 shows a plot for the GLASSO applied to equities data.

References

[1] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

[2] J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. *arXiv preprint arXiv:1206.3249*, 2012.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[4] G. Hardy, J. Littlewood, and G. Pólya. Inequalities. Cambridge Mathematical Library Series, 1967.

Table 4: Gene Classification

ALL	AML
U22376	M55150
X59417	X95735
U05259	U50136
M92287	M16038
M31211	U82759
X74262	M23197
D26156	M84526
S50223	Y12670
M31523	M27891
L47738	X17042
U32944	Y00787
Z15115	M96326
X15949	U46751
X63469	M80254
M91432	L08246
U29175	M62762
Z69881	M28130
U20998	M63138
D38073	M57710
U26266	M69043
M31303	M81695
Y08612	X85116
U35451	M19045
M29696	M83652
M13792	X04085

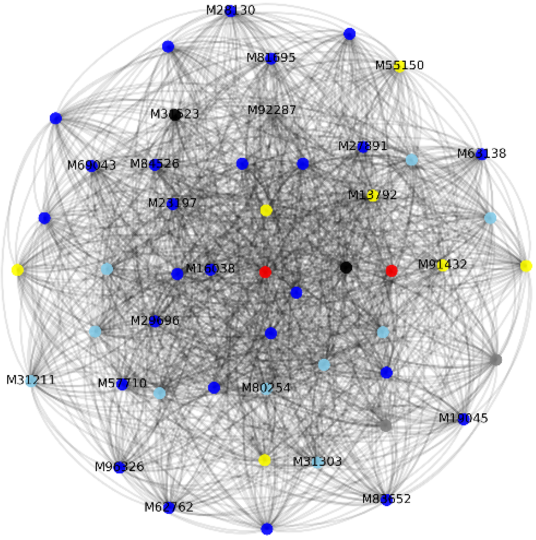

Figure 3: Network constructed by GLASSO on gene expression data. Each color represents a biological pathway: Signal Transduction (red), Immune System (blue), Cell Cycle (gray), Metabolism (yellow), Gene Expression (black), Uncategorized (skyblue).

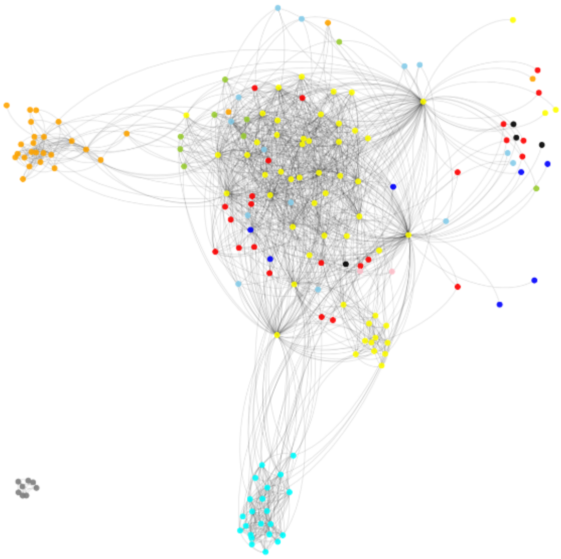

Figure 4: Network constructed by GLASSO on equities expression data. Each colour represents a GICs sector: Consumer Discretionary (red), Consumer Staples (blue), Energy (gray), Financials (yellow), Health Care (black), Industrials (skyblue), Information Technology (orange), Materials (yellow-green), Telecommunications Services (pink), Utilities (cyan).