# Group Sparse Priors for Covariance Estimation

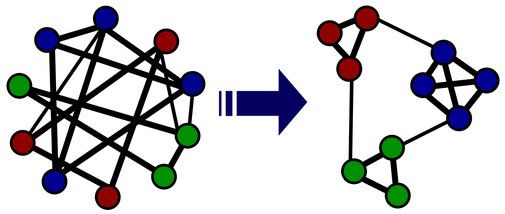
**Benjamin M. Marlin, Mark Schmidt, and Kevin P. Murphy**

**Department of Computer Science, University of British Columbia, Vancouver, Canada**

## 1.0 Introduction

**Problem:** In this work we consider the problem of sparse, block-structured Gaussian precision matrix (inverse covariance matrix) estimation when the blocks are not known *a priori*.

**Motivation:** Estimating a covariance matrix from high dimensional data using a small number of samples is known to be statistically challenging, and yet it is a problem that arises frequently in practice. For some kinds of data, it is reasonable to assume that the variables can be clustered or grouped into types that share similar connectivity or correlation patterns. For example, genes can be grouped into pathways, and connections within a pathway might be more likely than connections between pathways.



Our interest is in devising methods that simultaneously infer the block structure and a block-sparse precision matrix to provide improved regularization when there is no known block structure.

## 2.0 Related Work

**Tikhonov Regularization:** A very simple approach, which we shall call Tikhonov regularization, is to increase the diagonal of the empirical covariance matrix by adding a scalar multiple of the identity matrix.

$$\hat{\Sigma} = S + \nu I$$

**Independent L1 Regularized Precision Estimation:** Sparse precision matrix estimation can be cast as a convex optimization problem in the penalized maximum likelihood framework. An L1 penalty is imposed on the elements of the precision matrix [Yuan07, Banerjee06].

$$\hat{\Omega} = \underset{\Omega \in S^{++}}{\operatorname{argmax}} \ \log\det(\Omega) - \operatorname{tr}(S\Omega) - \lambda \sum_{i=1}^{D}\sum_{j \neq i}^{D} |\Omega_{ij}| - \nu \sum_{i=1}^{D} |\Omega_{ii}|$$

**Group L1 Regularized Precision Estimation:** If the group structure is known, one can extend the L1 penalized likelihood framework in a straightforward way, by penalizing the infinity norm [Duchi08] or the two-norm [Schmidt09] of each block separately. The resulting objective function is still convex, and encourages block-wise sparse graphs.

$$\hat{\Omega} = \arg\max_{\Omega \in S^{++}} \log\det(\Omega) - \operatorname{tr}(S\Omega) - \sum_{kl} \lambda_{kl} ||\{\Omega_{ij} : i \in G_k, j \in G_l\}||_{p_{kl}}$$

**Sparse Dependency Networks:** An alternative approach to sparse precision estimation is to learn the underlying graph by regressing each node on all the others using an L1 penalty [Meinshausen06, Marlin09].

$$\hat{\mathbf{w}}_j = \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \log p(x_{nj} | x_{n,-j}, \mathbf{w}, \sigma_j^2) + \lambda \sum_{i \neq j} |w_i|$$

## 3.0 Distributions

**Summary:** We convert the independent L1 and group L12 regularization functions into probability distributions over positive definite matrices. We embed the distributions in hierarchical models to simultaneously estimate the group structure and a corresponding group-sparse precision matrix.

**Independent L1 Distribution:** Corresponds to the independent L1 regularization function. MAP estimation under this prior is equivalent to penalized maximum likelihood estimation under L1 regularization when the penalty parameters are known and fixed.

$$P_{L1}(X|\lambda) = \ \frac{1}{\mathcal{Z}_{l1}} \mathrm{pd}(X) \prod_{i=1}^{D}\prod_{j>i}^{D} \exp(-\lambda_{ij}|X_{ij}|)$$

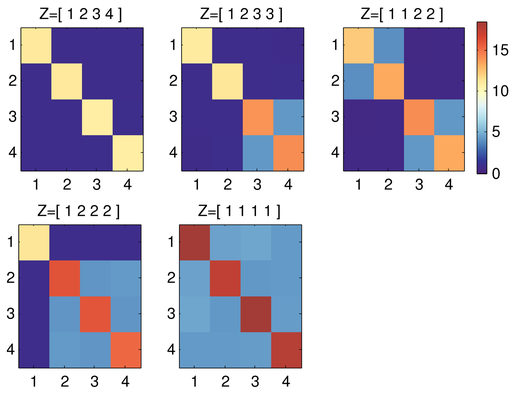
**Group L1 Distribution:** We augment the independent L1 distribution with a group indicator variable z for each dimension. We apply differential penalization that depends on the group structure giving a "soft" group sparsity effect.

$$P_{GL1}(X|\lambda, \boldsymbol{z}) = \frac{1}{\mathcal{Z}_1} \mathrm{pd}(X) \prod_{i}^{D} \exp(-\lambda_D |X_{ii}|)$$
$$\times \prod_{i=1}^{D}\prod_{j>i}^{D} \exp(-(\lambda_1 \delta_{z_i,z_j} + \lambda_0(1-\delta_{z_i,z_j}))|X_{ij}|)$$

**Group L12 Distribution:** MAP estimation under this prior is equivalent to penalized maximum likelihood estimation under L12 regularization when the penalty parameters and the grouping are known and fixed.

$$P_{GL1L2}(X|\lambda, \boldsymbol{z}) = \frac{1}{\mathcal{Z}_{12}} \mathrm{pd}(X) \prod_{i=1}^{D} \exp(-\lambda_D |X_{ii}|)$$
$$\times \prod_{i=1}^{D}\prod_{j>i}^{D} \exp(-\lambda_1 \delta_{z_i,z_j} |X_{ij}|)$$
$$\times \prod_{k=1}^{K}\prod_{l>k}^{K} \exp\left(-\lambda_0 C_{kl} \left(\sum_{i=1}^{D}\sum_{j=1}^{D} \delta_{z_i,k}\delta_{z_j,l}(X_{ij})^2\right)^{1/2}\right)$$

**Group Sparsity Property:** We illustrate the group sparsity effect in 4D by computing a Monte Carlo estimate of E[|X|] under the group L1 distribution with $\lambda_0$=1.0 and $\lambda_1$=0.1 for each unique partition. We see that between-group entries are suppressed as expected.




## 4.0 Bounds

**Summary:** The intractable normalizing terms in both distributions depend on the grouping. We adopt a strategy of upper-bounding the intractable normalization terms, which yields a tractable lower bound on the log posterior. We construct the upper bound by expanding the domain of integration from the positive definite cone to the larger space of symmetric matrices with positive diagonal.

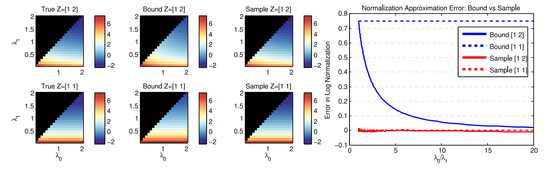
**Group L1 Bound:** Reduces to univariate exponential and Laplace integrals, which have simple closed form solutions.

$$\mathcal{Z}_1 \leq \int_{\mathbb{S}_P^D} \prod_{i=1}^{D}\prod_{j\geq i}^{D} \exp(-\lambda_{ij}|X_{ij}|)dX$$
$$= \prod_{i=1}^{D} \frac{1}{\lambda_{ii}} \cdot \prod_{i=1}^{D}\prod_{j>i}^{D} \frac{2}{\delta_{z_i,z_j}\lambda_1 + (1-\delta_{z_i,z_j})\lambda_0}$$

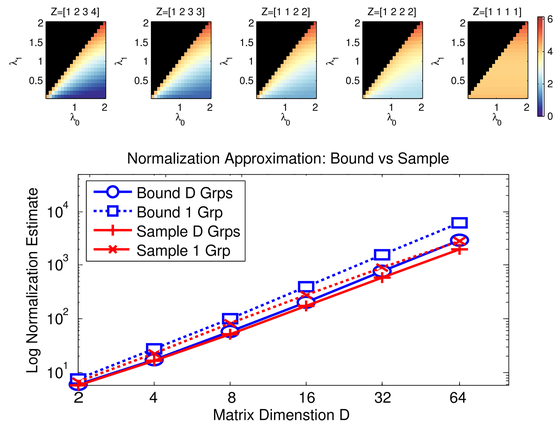
**Group L12 Bound:** Reduces to univariate exponential/Laplace and multivariate Laplace integrals, which all have closed form solutions.

$$\mathcal{Z}_{12} \leq \int_{\mathbb{S}_P^D} \prod_{i=1}^{D} \exp(-\lambda_D |X_{ii}|) \prod_{i=1}^{D}\prod_{j>i}^{D} \exp(-\lambda_1 \delta_{z_i,z_j}|X_{ij}|)$$
$$\cdot \prod_{k=1}^{K}\prod_{l>k}^{K} \exp\left(-\lambda_0 C_{kl}\left(\sum_{i=1}^{D}\sum_{j=1}^{D} \delta_{z_i,k}\delta_{z_j,l}(X_{ij})^2\right)^{1/2}\right) dX$$
$$\leq \left(\frac{1}{\lambda_D}\right)^D \left(\frac{2}{\lambda_1}\right)^{C_T} \prod_{k,l\neq k}^{K} \frac{\pi^{\frac{C_{kl}-1}{2}} \Gamma\left(\frac{C_{kl}+1}{2}\right) 2^{C_{kl}}}{(\lambda_0 C_{kl})^{C_{kl}}}$$

**Bounds in 2D:** We can evaluate the true normalization constant exactly in 2D. We compare to the bound and a Monte Carlo approximation.




**Bounds in Higher Dimension:** In higher dimensions we compare the bound to the Monte Carlo approximation. We show the error as a function of penalization strength for all partitions in 4D. We also show the bound versus the Monte Carlo estimate as a function of dimension.




## 5.0 Covariance Estimation

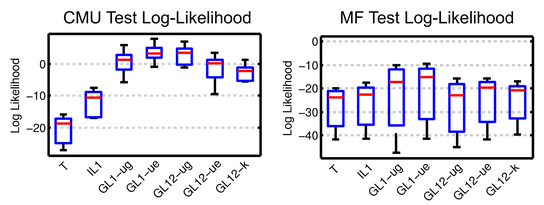
**Model:** We use our new distributions as sparsity-inducing priors on the precision matrix in a hierarchical model. The top of the hierarchy consists of a conjugate Multinomial/Dirichlet clustering model to estimate the groups.

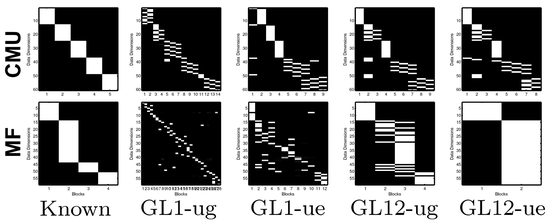
$$P(\theta|\alpha_0) = \mathcal{D}(\theta; \alpha_0) \qquad P(\Omega|\lambda, \boldsymbol{z}) = P_G(\Omega|\lambda, \boldsymbol{z})$$
$$P(z_i = k|\theta) = \theta_k \qquad P(\boldsymbol{x}_n|\mu, \Omega) = \mathcal{N}(\boldsymbol{x}_n; \mu, \Omega^{-1})$$

**Learning:** We optimize a lower bound on the log posterior based on the precision distribution bound and a Variational Bayes approximation q(z,θ|α,ϕ). We make non-local moves in partition space using explicit cluster splitting steps based on greedy or exhaustive search methods.

$$\log p(\Omega|X) = \log \int \sum_z p(X|\Omega)p(\Omega|z)p(z|\theta)p(\theta|\alpha_0)d\theta \geq \log \int \sum_z p(X|\Omega)\frac{1}{\tilde{\mathcal{Z}}_1(z)}\tilde{p}(\Omega|z)p(z|\theta)p(\theta|\alpha_0)d\theta$$
$$= \sum_z -\log \tilde{\mathcal{Z}}_1(z) + [\log p(X|\Omega) + \log \tilde{p}(\Omega|z)] + \log \int [p(z|\theta)p(\theta|\alpha_0)d\theta]$$
$$\geq \max_{\alpha,\phi} E_q\left\{-\log \tilde{\mathcal{Z}}_1(z) + [\log p(X|\Omega) + \log \tilde{p}(\Omega|z)] + \log p(z|\theta) + \log p(\theta|\alpha_0) - \log q(z,\theta|\alpha,\phi)\right\}$$
$$\geq \max_{\alpha,\phi} \frac{N}{2}(-\log(2\pi) + \log\det(\Omega)) - \frac{N}{2}\mathrm{trace}(\Omega S) + \log(pd(\Omega)) + \sum_{i=1}^{D}(-\lambda_D|\Omega_{ii}| + \log(\lambda_D))$$
$$+ \sum_{i=1}^{D}\sum_{j>i}^{D}(-\log(2) + E_q[\delta_{z_i,z_j}](-\lambda_1|\Omega_{ij}| + \log(\lambda_1)) + (1 - E_q[\delta_{z_i,z_j}])(-\lambda_0|\Omega_{ij}| + \log(\lambda_0))$$
$$+ \sum_{i=1}^{D}\sum_{k=1}^{K} E_q[\delta_{z_i,k}]E_q[\log(\theta_k)] + \log(\Gamma(\alpha_0)) - K\log(\Gamma(\alpha_0/K)) + \sum_{k=1}^{K}(\alpha_0/K - 1)E_q[\log\theta_k]$$
$$- \sum_{i=1}^{D}\sum_{k=1}^{K} E_q[\log(\phi_{ik})] - \log(\Gamma(\sum_{k=1}^{K}\alpha_k)) + \sum_{k=1}^{K}\log(\Gamma(\alpha_k)) - \sum_{k=1}^{K}(\alpha_k - 1)E_q[\log\theta_k]$$

**Experimental Results:** We test the proposed methods on two data sets. The N=100 D=60 CMU motion capture data set used by Marlin and Murphy, and the N=86, D=59 Mutual Funds (MF) data set used by Scott and Carvalho. We show test performance and inferred groups.




- **T:** Tikhonov regularization
- **IL1:** Independent L1 penalized likelihood
- **GL1-u(g/e):** Group L1 prior (unknown groups, greedy/exhaustive search)
- **GL12-u(g/e):** Group L12 prior (unknown groups, greedy/exhaustive search)
- **GL12-k:** Group L12 penalized likelihood (known groups)

## References

Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. J. of Machine Learning Research, 9, 485–516.

Dempster, A. (1972). Covariance selection. Biometrics, 28, 157-175.

Dobra, D., Hans, C., Jones, B., Nevins, J., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. J. Multivariate analysis, 90, 196-212.

Duchi, J., Gould, S., & Koller, D. (2008). Projected subgradient methods for learning sparse gaussians. Proc. of the Conf. on Uncertainty in AI.

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation the graphical lasso. Biostatistics, 432:441.

Ghahramani, Z., & Beal, M. (2000). Propagation algorithms for variational Bayesian learning. Advances in Neural Info. Proc. Systems (pp. 507.513).

Marlin, B. and Murphy, K. (2009) Sparse Gaussian Graphical Models with Unknown Block Structure. Proc. of the Intl. Conf. on Machine Learning.

Meinshausen, N., & Buhlmann, P. (2006). High dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34, 1436-1462.

Schmidt, M., van den Berg, E., Friedlander, M., & Murphy, K. (2009). Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm. AI & Statistics.

J. Scott and M. Carvalho. Feature-inclusion stochastic search for gaussian graphical models. Journal of Computational and Graphical Statistics, 17(4):790–808, 2008.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. Biometrika, 94, 19.35.