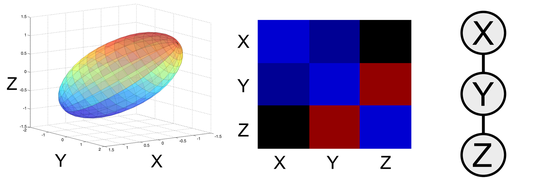
Accelerating Bayesian Structural Inference for Non-Decomposable Gaussian Graphical Models

Baback Moghaddam[1], Benjamin M. Marlin[2], Mohammad Emtiyaz Khan[2] and Kevin P. Murphy[2]

1. Jet Propulsion Laboratory, California Institute of Technology 2. Department of Computer Science, University of British Columbia

1.0 Introduction

Gaussian Graphical Models: A Gaussian graphical model is simply a multivariate Gaussian distribution where the precision matrix (the inverse of the covariance matrix) is sparse. Zeros in the precision matrix correspond to absent edges in the graphical model. Absent edges in the graphical model imply conditional independence relations.



Gaussian Distribution Precision Matrix GGM Graph

Problem: This work addresses the problem of accelerating the search for non-decomposable Gaussian Graphical Model structures.

Motivation: Estimating Gaussian Graphical Model structure can be motivated from two different perspectives:

(1) Regularization. Sparse GGM structures are sought as a means of controlling model complexity, and improving density estimation.

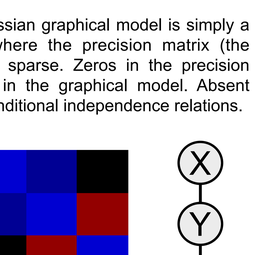
(2) Knowledge Discovery. The true graphical structure is sought for the purpose of interpreting the relationships between variables.

2.0 Graph Classes & Priors

Gaussian Distribution: We parametrize the Gaussian distribution directly in terms of the precision matrix Ω as shown below. Without loss of generality we assume the data has zero mean. S denotes the sample covariance matrix X^TX. Note that <Ω,S>=trace(Ω,S).

$$p(\mathcal{D}|\Omega) = \prod_{i=1}^{n} \mathcal{N}(x_i|\,0,\Omega^{-1}) \;\propto\; |\Omega|^{n/2}\exp(-\frac{1}{2}\langle\Omega,S\rangle) \quad [1]$$

Decomposable Graphs: Most prior work on learning the structure of GGMs has focused on the special case of decomposable graphs. Decomposable GGMs have the special property that the marginal likelihood under the conjugate Hyper-Inverse-Wishart (HIW) prior is fast to compute following local changes to the graph. However, only a small fraction of the total number of graphs on D nodes is decomposable.



Non-Decomposable Graphs: Non-Decomposable GGMs are more general than Decomposable GGMs. However, the marginal likelihood under the conjugate G-Wishart prior (given below) is intractable, making the search for general GGM structures computationally expensive.

$$W(\Omega|G,\delta_0,S_0) = \frac{I[\Omega\in\mathcal{S}_G^{++}]}{Z(G,\delta_0,S_0)}\,|\Omega|^{(\delta_0-2)/2}\,\exp(-\frac{1}{2}\langle\Omega,S_0\rangle) \quad [2]$$

$$p(\mathcal{D}|G) = \int_{\mathcal{S}_G^{++}} p(\mathcal{D}|\Omega)\,W(\Omega|G,\delta_0,S_0)\,d\Omega \;\propto\; \frac{Z(G,\delta_n,S_n)}{Z(G,\delta_0,S_0)} \quad [3]$$

3.0 Graph Scoring

Overview: Approximating the marginal likelihood for non-decomposable GGMs is one of the main contributions of this work. We compare several approximations including a Monte Carlo method, two forms of Laplace approximation, and the Bayesian Information Criterion (BIC).

Monte Carlo Approximation: Approximates the numerator and denominator of Equation 3 using sampling (Atay-Kayis and Massam, 2005). The sampling distribution is exact and consists of products of normals and chi-squareds (code available from Mike West's group).

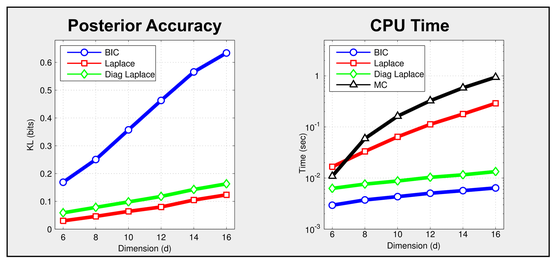
Laplace Approximation: Approximates the numerator and denominator of Equation 3 using a ratio of Laplace approximations. Requires computing the mode and the determinant of the Hessian matrix (evaluated at the mode) for both the prior and posterior G-Wishart distributions (Lenkoski and Dobra, 2008).

Diagonal Laplace Approximation: Our proposed method. Identical to the full Laplace approximation except that we approximate the determinant of the full Hessian matrix by the product of the diagonal entries.

BIC Approximation: Approximates Equation 3 using the log likelihood evaluated at the mode of the posterior plus an edge penalty term.

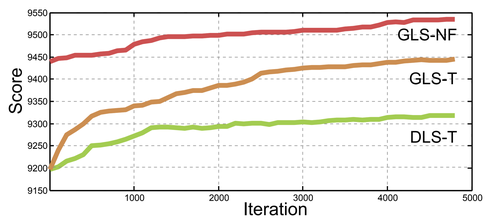
Computing G-Wishart Modes: We use a convex optimization method to compute G-Wishart prior and posterior modes. The method is similar to the G-Lasso approach of Duchi et al. (2008).

$$\hat{\Omega}_G = \arg\max_{\Omega\in\mathcal{S}_G^{++}} \log W(\Omega|G,\delta,S) = \arg\min_{\Omega\in\mathcal{S}_G^{++}} -\log|\Omega| + \left\langle\Omega,\frac{S}{\delta-2}\right\rangle \quad [4]$$



4.0 Graph Search

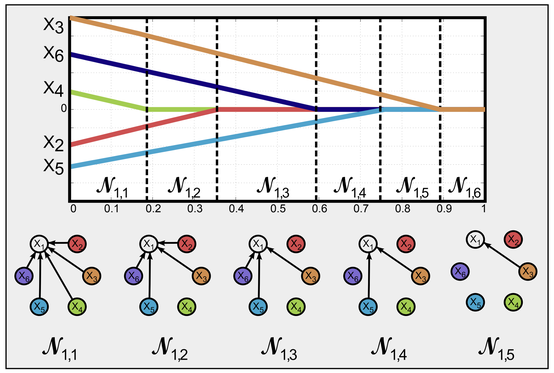
Overview: A stochastic local search algorithm tries to identify the most likely graph structure G in a given class of graphs by iterating between three steps (Scott and Carvalho, 2008): a stochastic local update to the graph based on marginal edge probabilities, scoring the graph, and updating the marginal edge probabilities. We show SLS trace plots below for decomposable graphs started from the optimal tree (DLS-T), general graphs started from the optimal tree (GLS-T), and general graphs started from Neighborhood Fusion (GLS-NF, see Section 5.0).



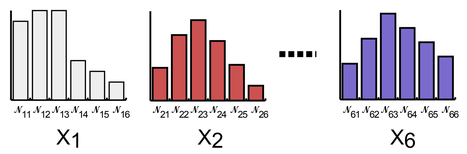
5.0 Neighborhood Fusion

Overview: Neighborhood fusion solves the problem of quickly producing large sets of high quality GGM structures. It exploits sparse linear regression techniques to compute a set of candidate neighborhood structures for each variable X_d, and specifies a mechanism for sampling and combining these neighborhoods to form undirected graphs. It is related to both the work of Meinshausen and Buhlmann (2006), and the G-Lasso (Friedman et al., 2007). We describe the algorithm below.

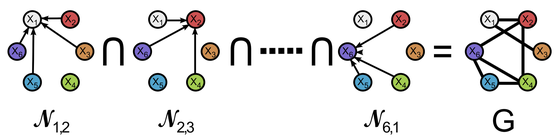
Step 1: Compute the Regularization Path & Store Neighborhoods
We compute the regularization path for each linear regression problem X_d = WX-d under a sparse prior (L_0 or L_1). We extract the distinct neighborhoods of X_d by examining the non-zero linear regression coefficients at each setting of the regularization parameter. We store the n^{th} distinct neighborhood of X_d in $\mathcal{N}_{dn}$.



Step 2: Score Neighborhoods & Compute Proposal Distribution. Next, we compute a score S_{dn} for each neighborhood $\mathcal{N}_{dn}$ of X_d expressing how well that neighborhood explains X_d. We use the linear regression Bayesian Information Criterion score. We compute a multinomial distribution Q_d over the local neighborhoods extracted for each variable X_d by exponentiating and normalizing the scores: $Q_{dn} \propto \exp(S_{dn})$.



Step 3: Sample Neighborhoods and Combine. Finally, we sample one local neighborhood for each variable X_d and intersect them to form a single symmetric graph. The graph sampling step can be repeated any number of times to quickly produce a large set of varied graphs.



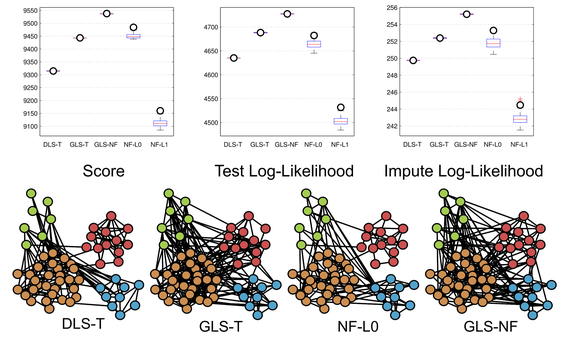
Combining NF and SLS: We can use a sample of NF graphs to accelerate SLS in two different ways. (1) We can initialize the SLS marginal edge probabilities from the NF sample. (2) We can start the SLS search from the best graph found by NF.

6.0 Experiments

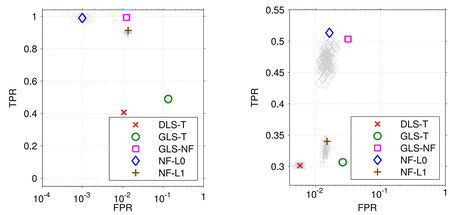
Overview: We perform experiments aimed at assessing the density estimation and structural recovery performance of GGM structure learning methods including decomposable and general SLS started from the most likely tree structure (DLS-T and GLS-T), Neighborhood Fusion using L_0 and L_1 priors (NF-L0 and NF-L1), and hybrid L_0 Neighborhood Fusion/general SLS (GLS-NF).

Data Sets: We use the D=59/N=60 mutual funds data set (previously considered by Scott and Carvalho) to assess density estimation performance and visualize MAP structures. We use D=100/N=500 and D=100/N=50 synthetic data sets to assess structural recovery.

Mutual Funds Density Estimation and Visualization: We see that the hybrid GLS-NF method is significantly better than the other methods in terms of diagonal Laplace Score, test set log-likelihood, and imputation log-likelihood. The visualizations show that the GLS-T and GLS-NF methods include more edges than the DLS and NF methods.



Synthetic Data Structural Recovery: On these higher dimensional data sets, we see that DLS-T and GLS-T perform quite poorly. NF-L0 and GLS-NF again have significantly better performance.



D=100/N=500: TPR vs FPR D=100/N=50: TPR vs FPR

Selected References

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. IEEE Trans. Info. Theory, 1968.

A. Dempster. Covariance selection. Biometrics, 28(1), 1972.

D. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao, and M.West. Sparse graphical models for exploring gene expression data. J. Multivariate analysis, 90, 2004.

J. Duchi, S. Gould, and D. Koller. Projected subgradients for learning sparse Gaussians. In UAI, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the Graphical Lasso. Biostats, 2007.

P. Giudici and P. Green. Decomposable graphical Gaussian model determination. Biometrika, 1999.

A. Lenkoski and A. Dobra. Bayesian structural learning and estimation in Gaussian graphical models. Technical Report 545, Department of Statistics, University of Washington, 2008.

N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the Lasso. The Annals of Statistics, 2006.

B. Moghaddam, A. Gruber, Y. Weiss, and S. Avidan. Sparse regression as a sparse eigenvalue problem. In Information Theory & Applications Workshop, 2008.

A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scand. J. Statistics, 29, 2002.

J. Scott and C. Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. J. of Computational and Graphical Statistics, 17(4), 2008.

T. Speed and H. Kiiveri. Gaussian Markov distributions over Finite graphs. Annals of Statistics, 1986.

M. Yuan and Yi Lin. Model selection and estimation in the GGM. Biometrika, 94(1), 2007.

T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning. In NIPS, 2008.