Sparse Gaussian Graphical Models with Unknown Block Structure

Benjamin M. Marlin and Kevin P. Murphy
Department of Computer Science
University of British Columbia

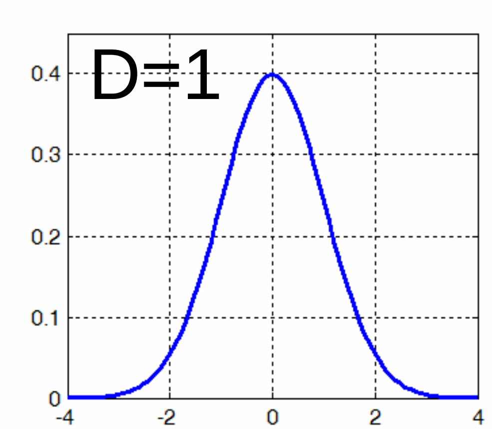
Outline

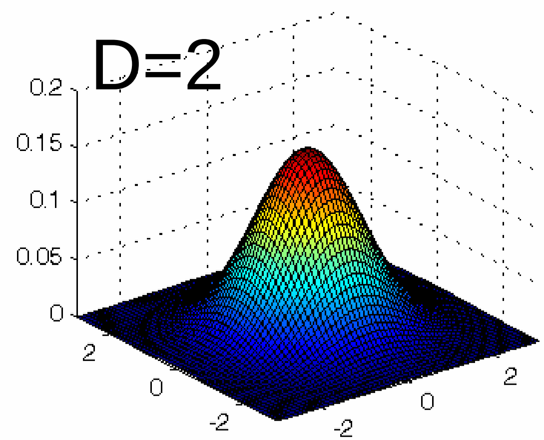
- **Introduction**
- **Related Work**
  - **Graphical Lasso**
  - **Group L1 Penalized Maximum Likelihood**
  - **Sparse Dependency Networks**
- **Unknown Block Structure**
  - **Model**
  - **Variational Inference**
- **Experiments and Results**
- **Conclusions**

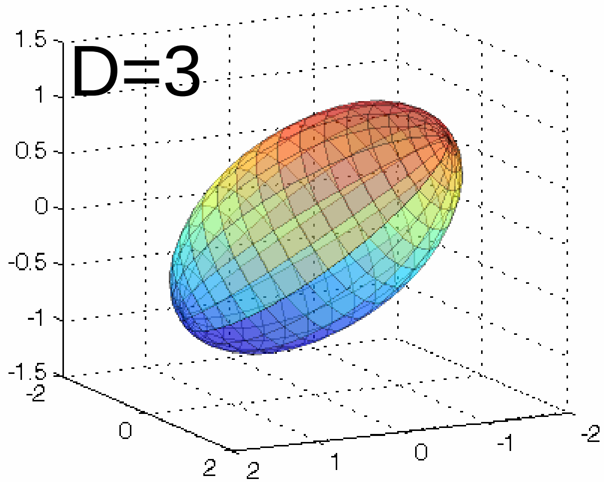
Introduction: Covariance Estimation

- Estimating the covariance matrix Σ of a Gaussian distribution is known to be difficult when the number of data cases N is low relative to the number of data dimensions D.

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$





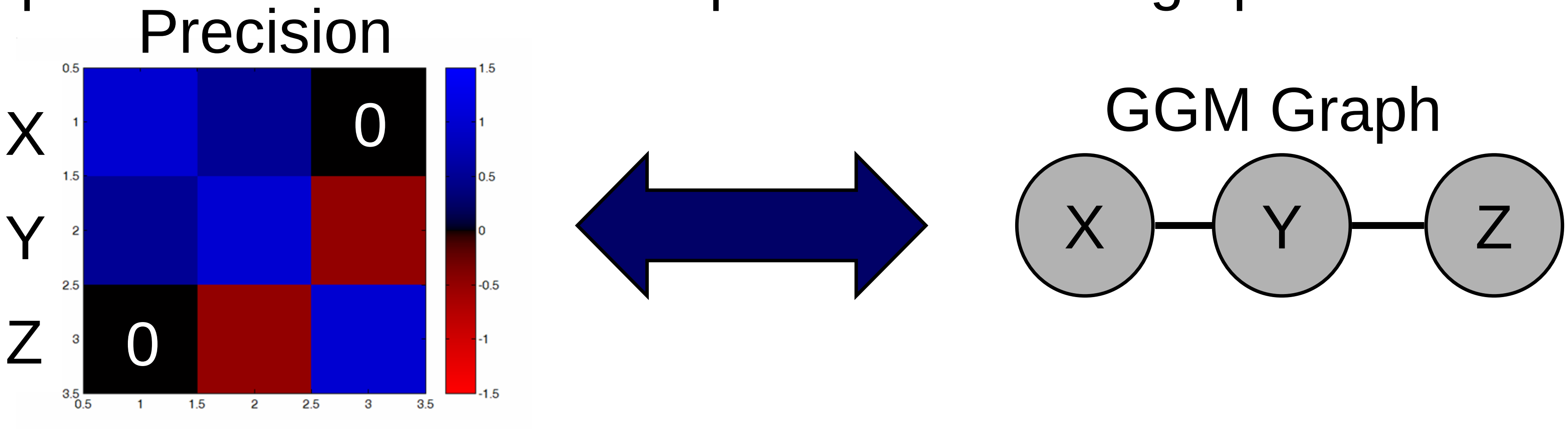


......

Introduction: Covariance Selection

- In 1972, Dempster proposed clamping some of the elements of the precision matrix $\Omega = \Sigma^{-1}$ to zero as a way of controlling complexity and deriving better covariance estimates.

- Zeros in the precision matrix correspond to absent edges in the Gaussian Graphical Model (GGM). Favoring sparse precision matrices corresponds to favoring sparse GGMs.

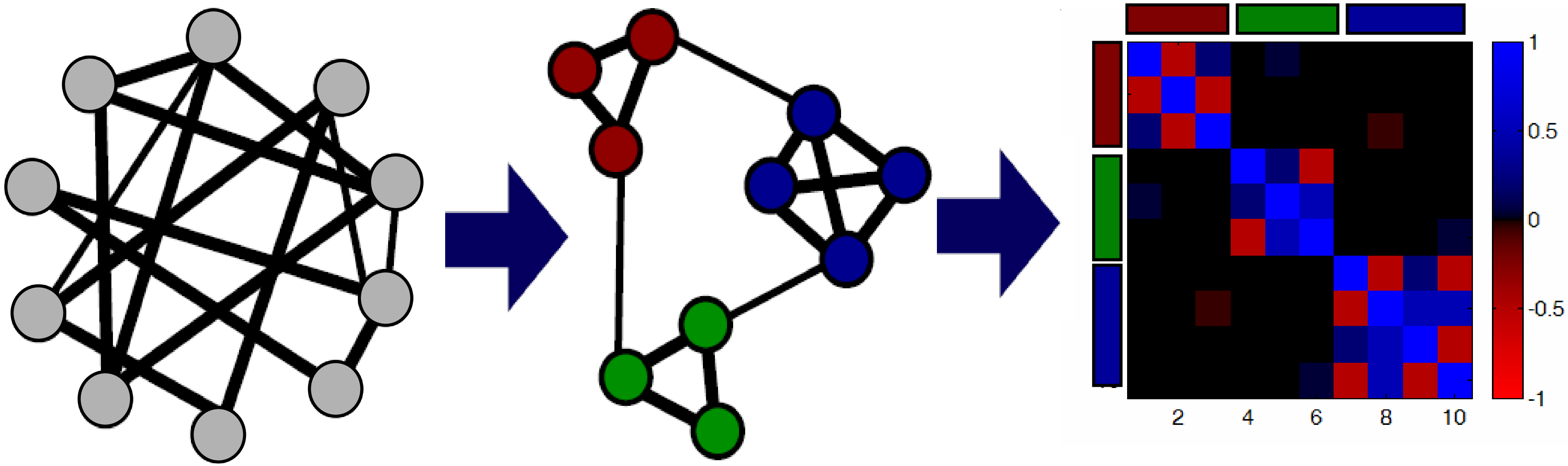


Introduction: Group Sparsity

- For some kinds of data, the variables can be clustered or grouped into types that share similar connectivity or correlation patterns.
- If we can infer these groups, we can use them to regularize precision matrix estimation in the $N \approx D$ and $N < D$ regimes.

Introduction: Problem Statement

• The problem we address in this work is how to estimate sparse, block-structured Gaussian precision matrices when the blocks are not known *a priori*.



Outline

- Introduction
- **Related Work**
    - Graphical Lasso
    - Group L1 Penalized Maximum Likelihood
    - Sparse Dependency Networks
- Unknown Block Structure
    - Model
    - Variational Inference
- Experiments and Results
- Conclusions

Related Work: Graphical Lasso

- The Graphical Lasso is a technique for sparse precision estimation based on independently penalizing the L1 norm of each precision matrix entry [Banerjee et al, Yuan & Lin].

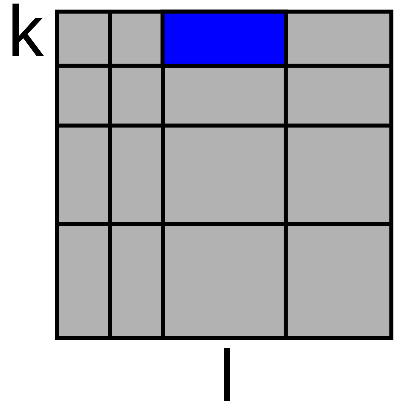
$$\hat{\Omega} = \arg\max_{\Omega \in S^{++}} \Big(\log\det(\Omega) - \mathrm{tr}(S\Omega) - \lambda \sum_{i=1}^{D} \sum_{j \neq i}^{D} |\Omega_{ij}| - \nu \sum_{i=1}^{D} |\Omega_{ii}| \Big)$$

- **S:** Empirical covariance matrix.
- ν**:** Diagonal regularization parameter.
- λ**:** Off-diagonal regularization parameter.

Related Work: Group Graphical Lasso

- The graphical lasso has been extended to group sparsity by penalizing the norm of each block of the precision matrix given a known grouping of the variables [Duchi et al, Schmidt et al].

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



- **G_k :** Set of variables in group k.
- λ_{kl} **:** Penalty parameter for entries between groups k and l.
- **p_{kl} :** Norm on entries between groups k and l.
- Schmidt et al. use **p_{kl}** = 1 within groups and **p_{kl}** = 2 between.

Related Work: Sparse Dependency Nets

- In a sparse dependency net we penalize the L1 norm of the linear regression weights for each node j regressed on every other node i≠j [Meinshausen and Buhlmann]. We can extract a graph and fit GGM using IPF/gradient-based optimization

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

- **w**$_{ji}$: Linear regression weight for node j given node i.
- **x**$_{nj}$: Value of data dimension j for data case n.
- **x**$_{n-j}$: Value of all data dimensions but j for data case n.
- λ : Penalty parameter.

Outline

- Introduction
- Related Work
    - Graphical Lasso
    - Group L1 Penalized Maximum Likelihood
    - Sparse Dependency Networks
- **Unknown Block Structure**
    - Model
    - Variational Inference
- Experiments and Results
- Conclusions

Unknown Block Structure: Overview

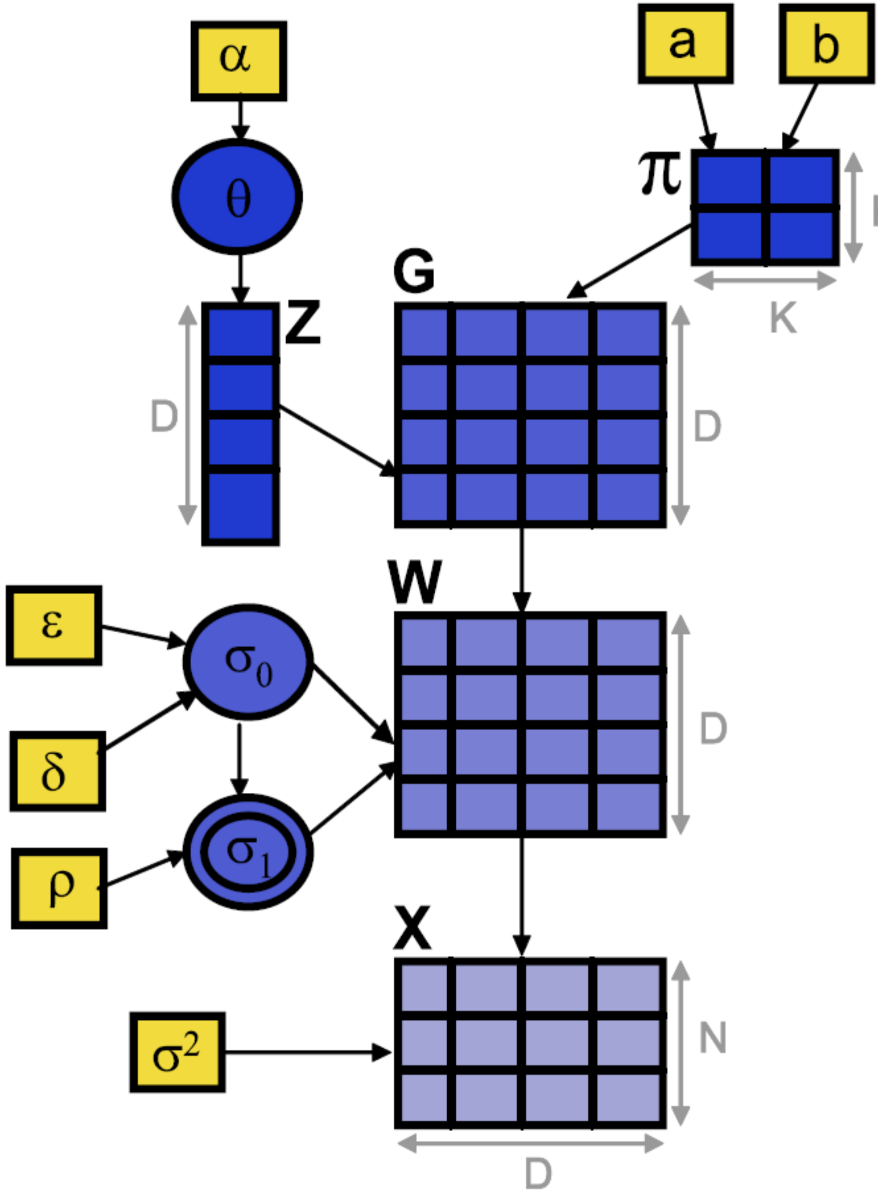
A Two-Stage Approach to Precision Estimation:

1. Use a hierarchical dependency network-based model to infer a grouping of the variables.
2. Fix the grouping and estimate the precision matrix using the Group L1/L2 method of Schmidt et al.

- Using group graphical lasso to estimate the precision matrix gives us **block sparsity** when it is well supported by the data, and **block shrinkage** in general.

Unknown Block Structure: Model

- **Stochastic Block Model**
- **Dependency Network**
- **Spike and Slab style prior**

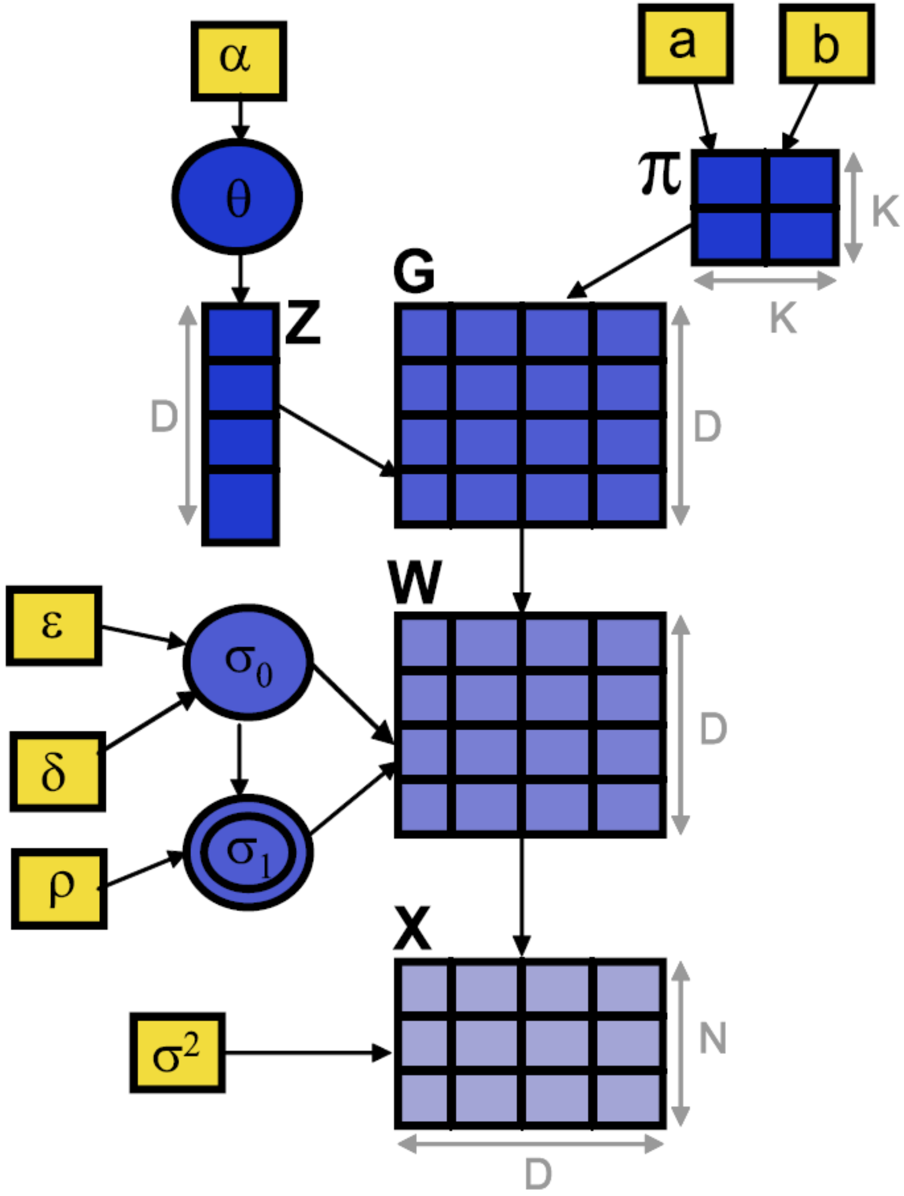


Unknown Block Structure: Model

- $\theta \sim \text{Dir}(\frac{\alpha}{K})$
- $z_d \sim \text{Multi}(\theta, 1)$
- $\pi_{k,k'} \sim \text{Beta}(a_{k,k'}, b_{k,k'})$
- $G_{d,d'} \sim \text{Ber}(\pi_{z_d, z_{d'}})$

- $w_{d,d'} \sim \mathcal{N}(0, \sigma_1^2)^{G_{d,d'}} \mathcal{N}(0, \sigma_0^2)^{1-G_{d,d'}}$
- $x_{dn} \sim \mathcal{N}(\boldsymbol{w}_d^T \boldsymbol{x}_{-d,n}, \sigma^2)$

- $\sigma_0^2 \sim \text{Ga}(\epsilon, \delta)$ and $\sigma_1^2 = \rho\sigma_0^2$.



Unknown Block Structure: Inference

Variational Bayes Approximation: We use a fully factorized variational Bayes approximation for learning.

$$Q(Z, \theta, \pi, G, W, \sigma_0) = Q(Z)Q(\theta)Q(\pi)Q(G)Q(W)Q(\sigma_0)$$

$$\begin{aligned}
Q(Z_d) &= \mathrm{Multi}(\phi_d, 1) \\
Q(\theta) &= \mathrm{Dir}(\alpha^*) \\
Q(\pi_{k,k'}) &= \mathrm{Beta}(a^*_{k,k'}, b^*_{k,k'}) \\
Q(G_{d,d'}) &= \mathrm{Ber}(\gamma_{d,d'}) \\
Q(1/\sigma_0^2) &= \mathrm{Ga}(\epsilon^*, \delta^*) \\
Q(W) &= \delta(W - \hat{w})
\end{aligned}$$

Unknown Block Structure: Inference

Variational Bayes Learning Algorithm:

$$a^*_{k,k'} \leftarrow a_{k,k'} + \sum_{d=1}^{D}\sum_{d'=1}^{D} \phi_{d,k}\phi_{d',k'}\gamma_{d,d'} \qquad \epsilon^* \leftarrow \epsilon + \frac{D(D-1)}{2}$$

$$b^*_{k,k'} \leftarrow b_{k,k'} + \sum_{d=1}^{D}\sum_{d'=1}^{D} \phi_{d,k}\phi_{d',k'}(1-\gamma_{d,d'}) \qquad \delta^* \leftarrow \delta + \sum_{d=1}^{D}\sum_{d'\neq d} \frac{w^2_{d,d'}}{2}\left(\frac{\gamma_{d,d'}}{\rho} + (1-\gamma_{d,d'})\right)$$

$$\overline{\pi}_{k,k',1} \leftarrow \Psi(a^*_{k,k'}) - \Psi(a^*_{k,k'} + b^*_{k,k'}) \qquad \overline{\pi}_{k,k',0} \leftarrow \Psi(b^*_{k,k'}) - \Psi(a^*_{k,k'} + b^*_{k,k'})$$

$$\phi_{dk} \leftarrow \mathrm{softmax}\left(\sum_{d'\neq d}\sum_{k'=1}^{K} \phi_{d',k'}(\gamma_{d,d'}\overline{\pi}_{k,k',1} + (1-\gamma_{d,d'})\overline{\pi}_{k,k',0}) + \Psi(\alpha^*_k) - \Psi(\sum_{k=1}^{K}\alpha^*_k)\right)$$

$$\gamma_{d,d'} \leftarrow \mathrm{logistic}\left(\sum_{k\leftarrow 1}^{K}\sum_{k'\leftarrow 1}^{K} \phi_{d,k}\phi_{d',k'}(\overline{\pi}_{k,k',1} - \overline{\pi}_{k,k',0}) + \tfrac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)(w^2_{d,d'} + w^2_{d',d}) - \tfrac{1}{2}\log\rho\right)$$

$$\alpha^*_k \leftarrow \alpha_k + \sum_{d=1}^{D}\sum_{k=1}^{K} \phi_{dk}$$

$$\Lambda_d \leftarrow \mathrm{diag}\left(\frac{\epsilon^*}{\delta^*}\left(\frac{\gamma_{d,d'}}{\rho} + (1-\gamma_{d,d'})\right)\right)$$

$$\widehat{w}_d \leftarrow (\sigma^2\Lambda_d + \sum_{n=1}^{N} X^T_{-dn}X_{-dn})^{-1}(\sum_{n=1}^{N} X_{dn}X_{-dn})$$

Unknown Block Structure: *Inference*

Extensions to Basic Variational Inference:

- The variational updates for the cluster indicators are tightly coupled together. To help get around this problem we introduce explicit cluster splitting steps based on graph cuts.

- For large problems, the dependency network weight updates are very costly at $O(d^4)$ per iteration. We use a fast adaptive variational update schedule to help with this problem.

Outline

- Introduction
- Related Work
  - Graphical Lasso
  - Group L1 Penalized Maximum Likelihood
  - Sparse Dependency Networks
- Unknown Block Structure
  - Model
  - Variational Inference
- **Experiments and Results**
- Conclusions

Experiments: Methods

- **T:** Tikhonov Regularization $\hat{\Sigma} = S + \nu I$
- **IL1:** Independent L1 penalized maximum likelihood (aka graphical lasso)
- **KGL1:** Group L1/L2 penalized maximum likelihood with known groups.
- **UGL1:** Group L1/L2 penalized maximum likelihood with groups inferred by our hierarchical dependency network.
- **UGL1F:** Group L1/L2 penalized maximum likelihood with groups inferred by our hierarchical dependency network. Uses fast update schedule.

Experiments: Empirical Protocol

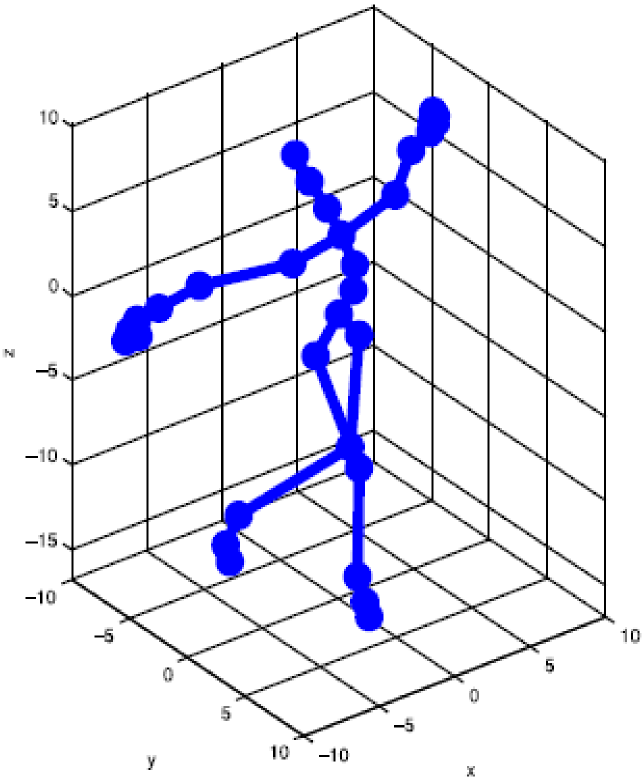
- We used fixed hyper-parameters for the hierarchical dependency network to infer the groups for UGL1 and UGL1F.

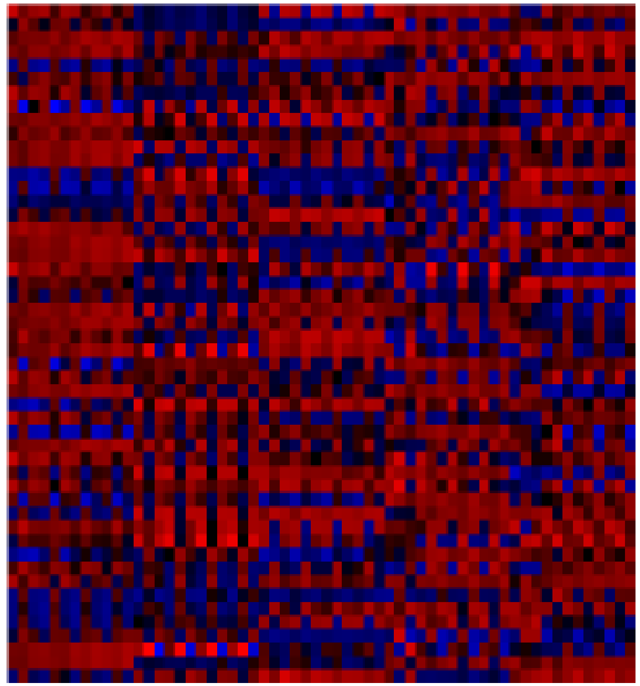
- We report five-fold cross validation test log likelihood estimates (relative to the Tikhonov baseline) as a function of the regularization parameter λ.

- We present results on two data sets.

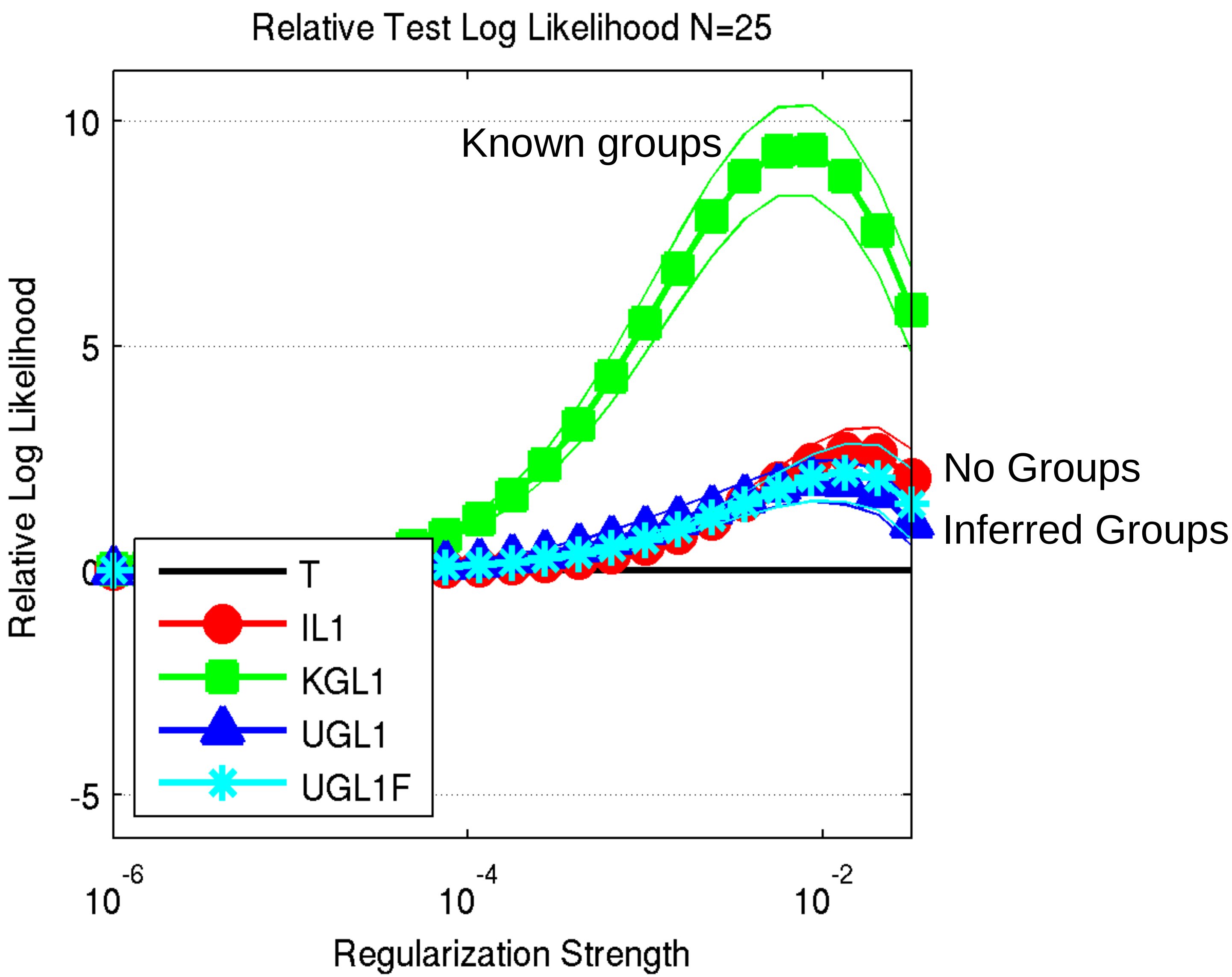
Results: *CMU Data Set*

CMU Motion Capture Data Set (N={25,50,75,100}, D=60):

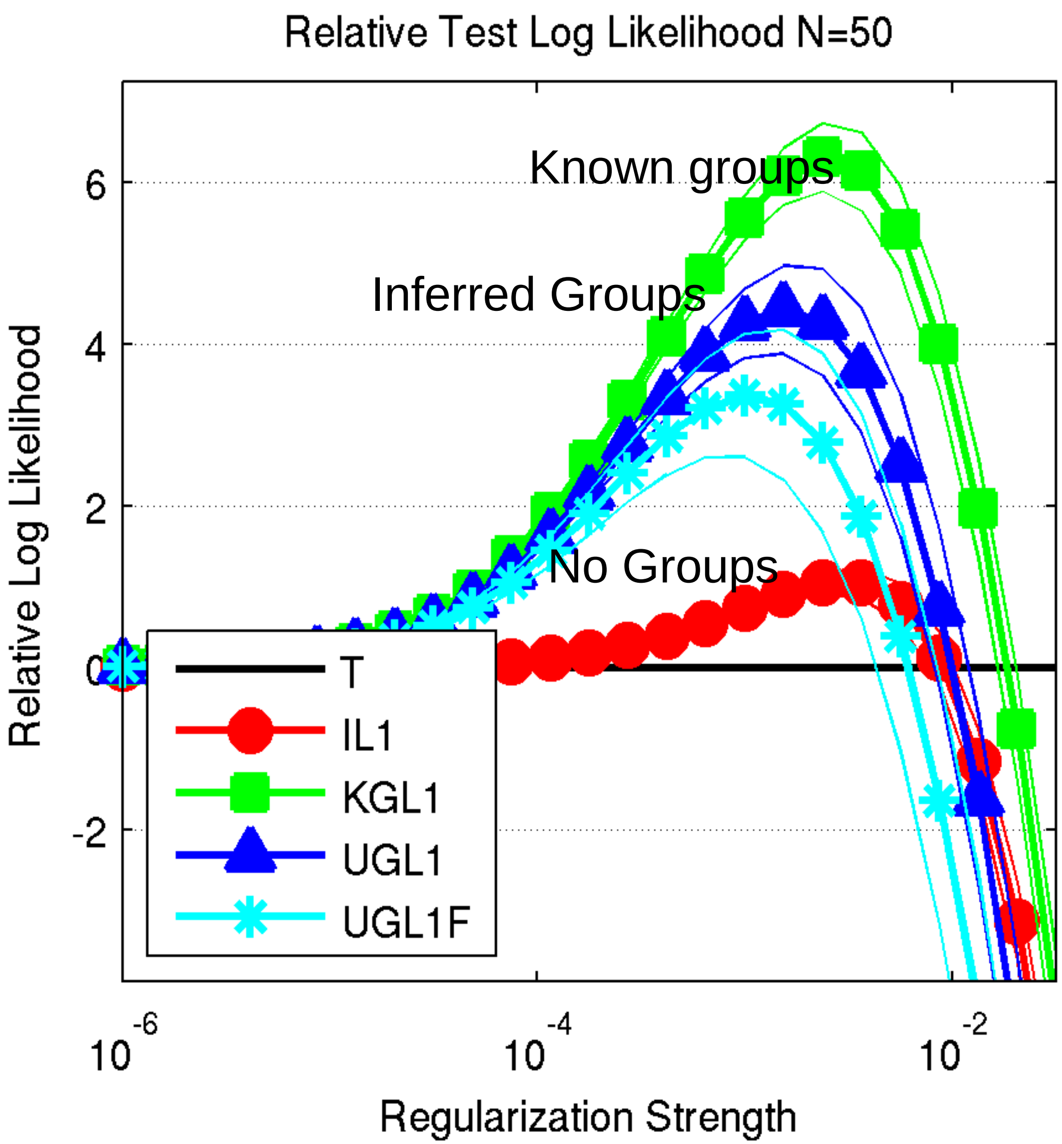




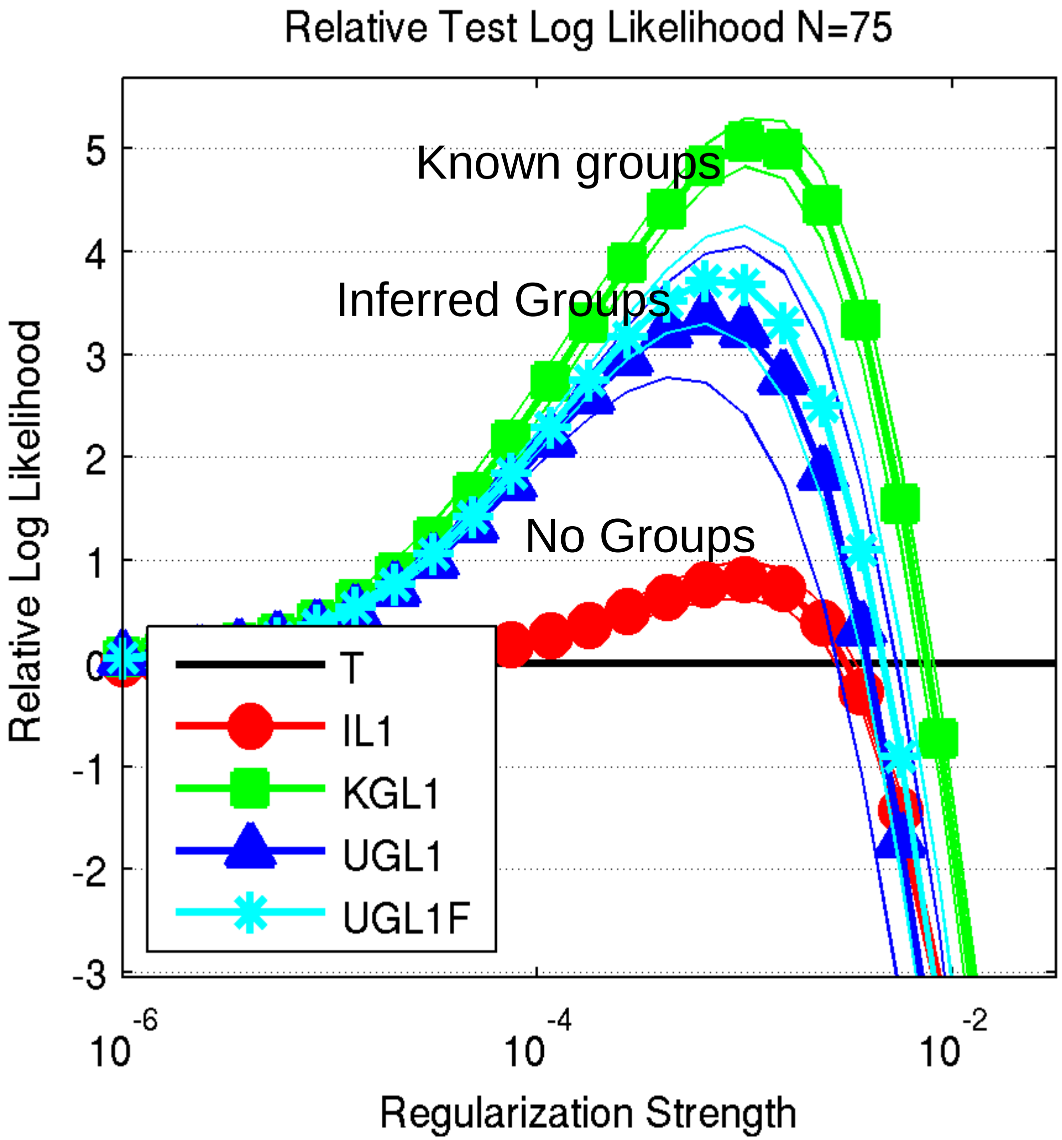
Results: CMU Test Log Likelihood N=25



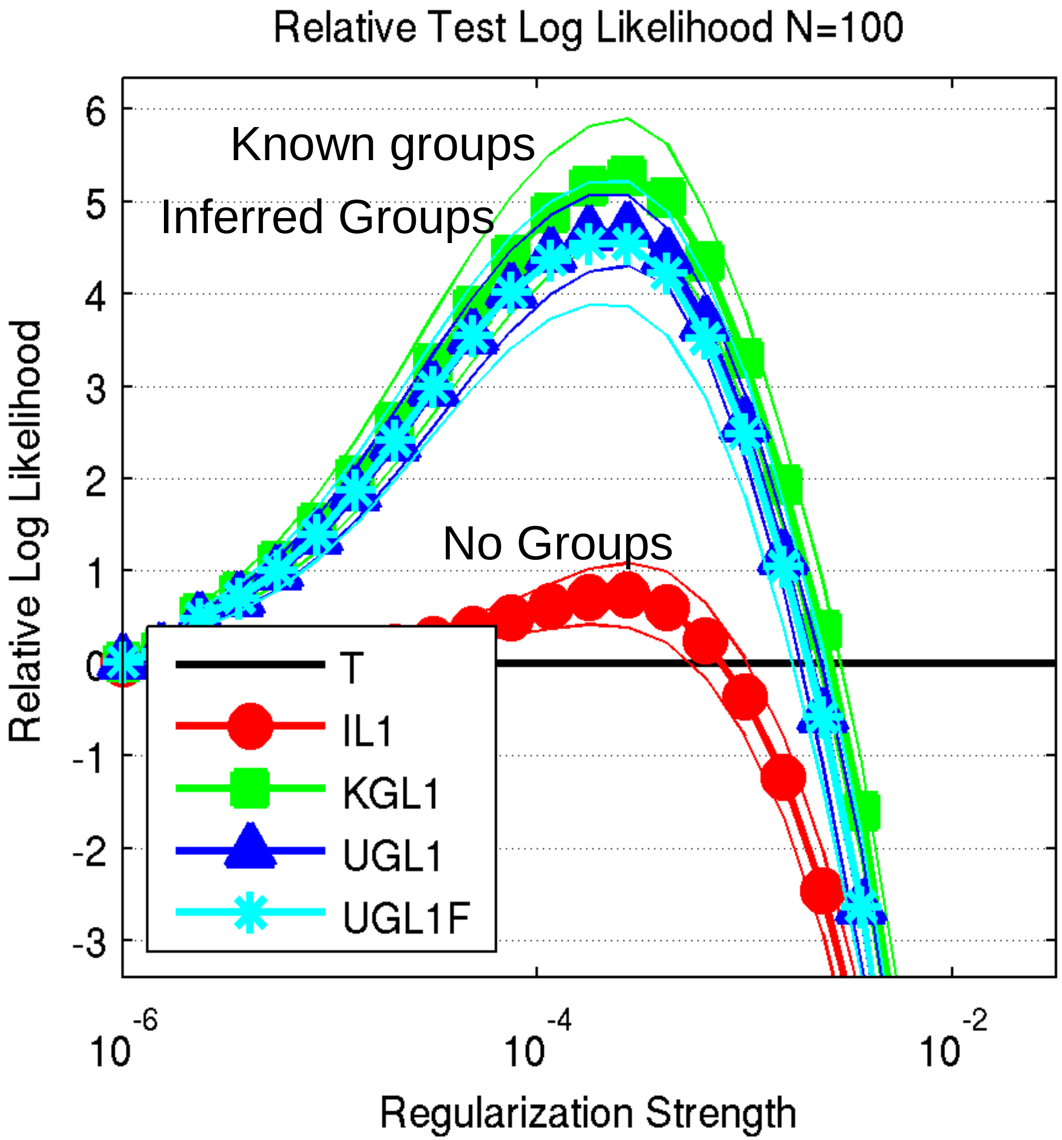
Results: CMU Test Log Likelihood N=50



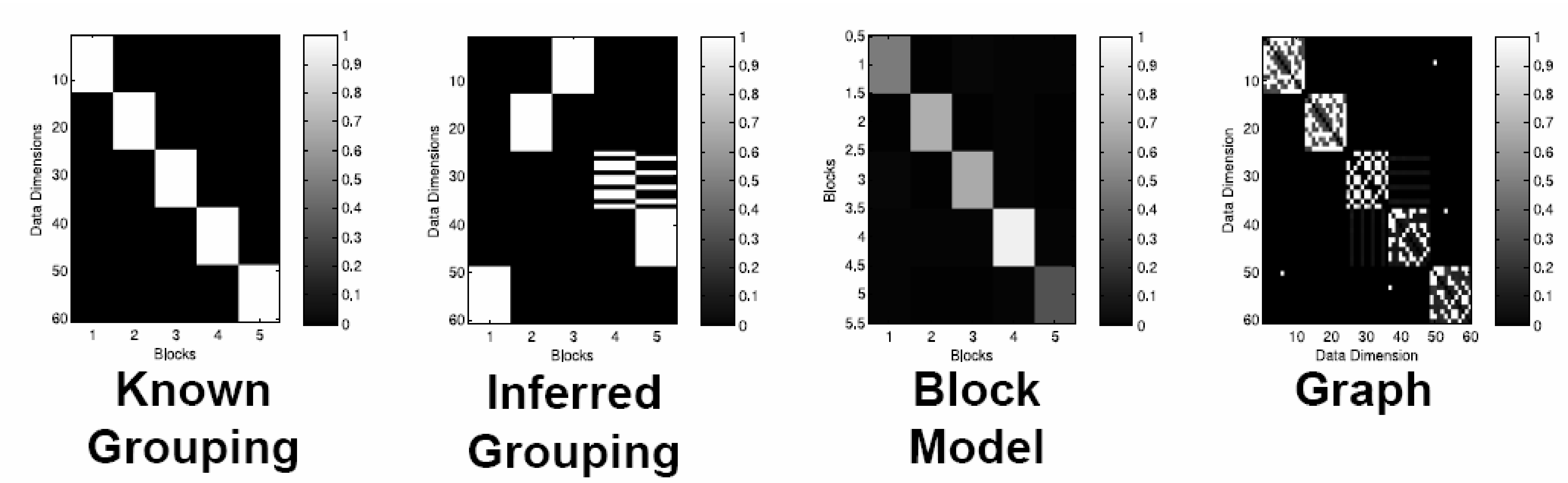
Results: *CMU Test Log Likelihood N=75*



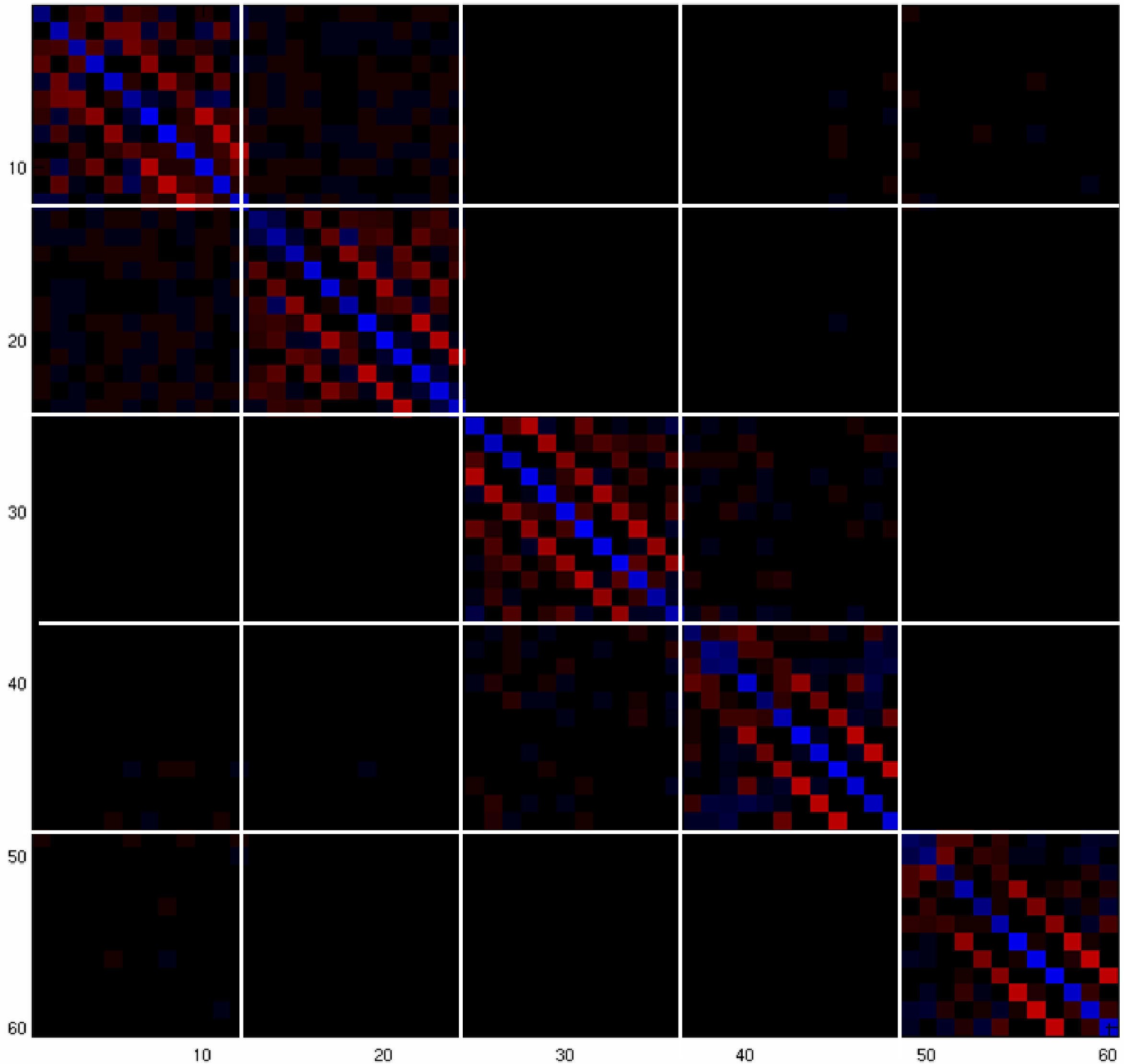
Results: *CMU Test Log Likelihood N=100*



Results: *CMU Inferred Structures N=50*



Results: CMU Estimated Precision Matrix



Results: *Gasch Genes Data Set*

Gasch Genes Data Set (N=174,D=667):

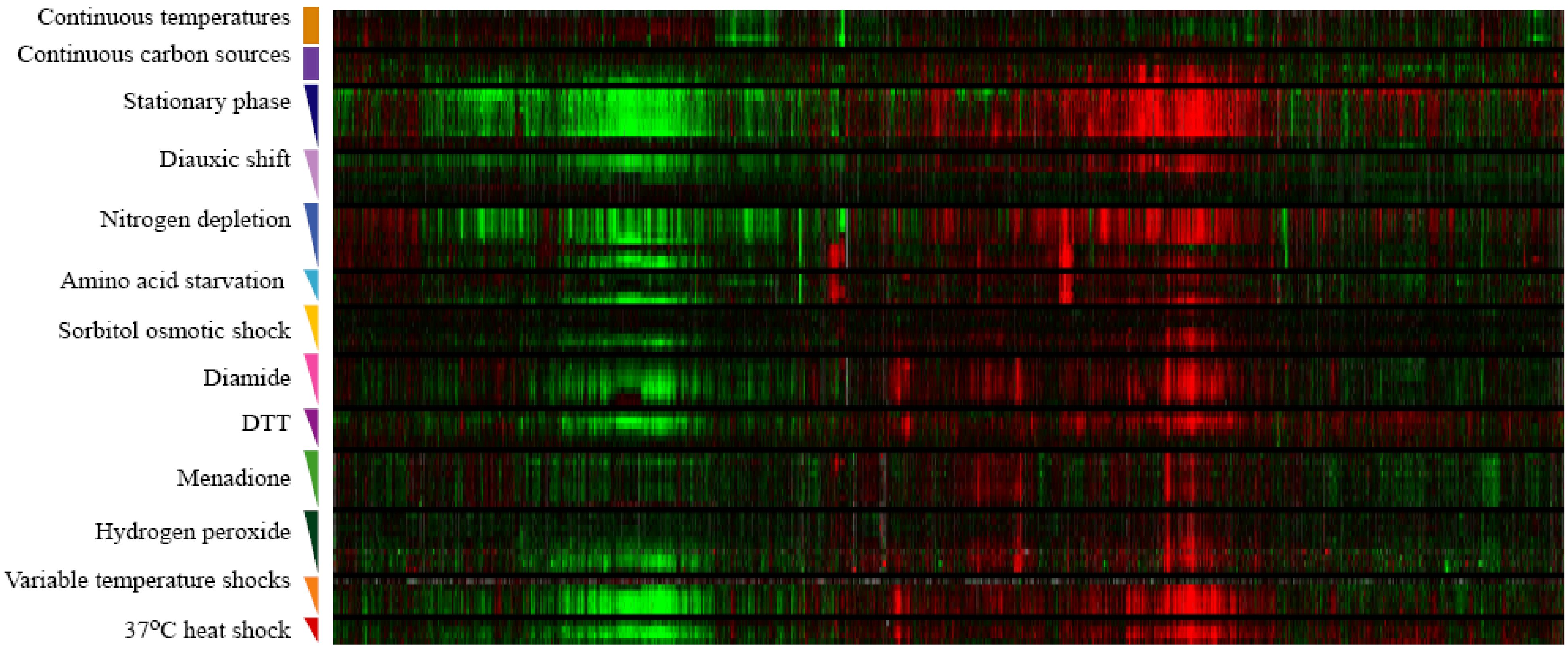
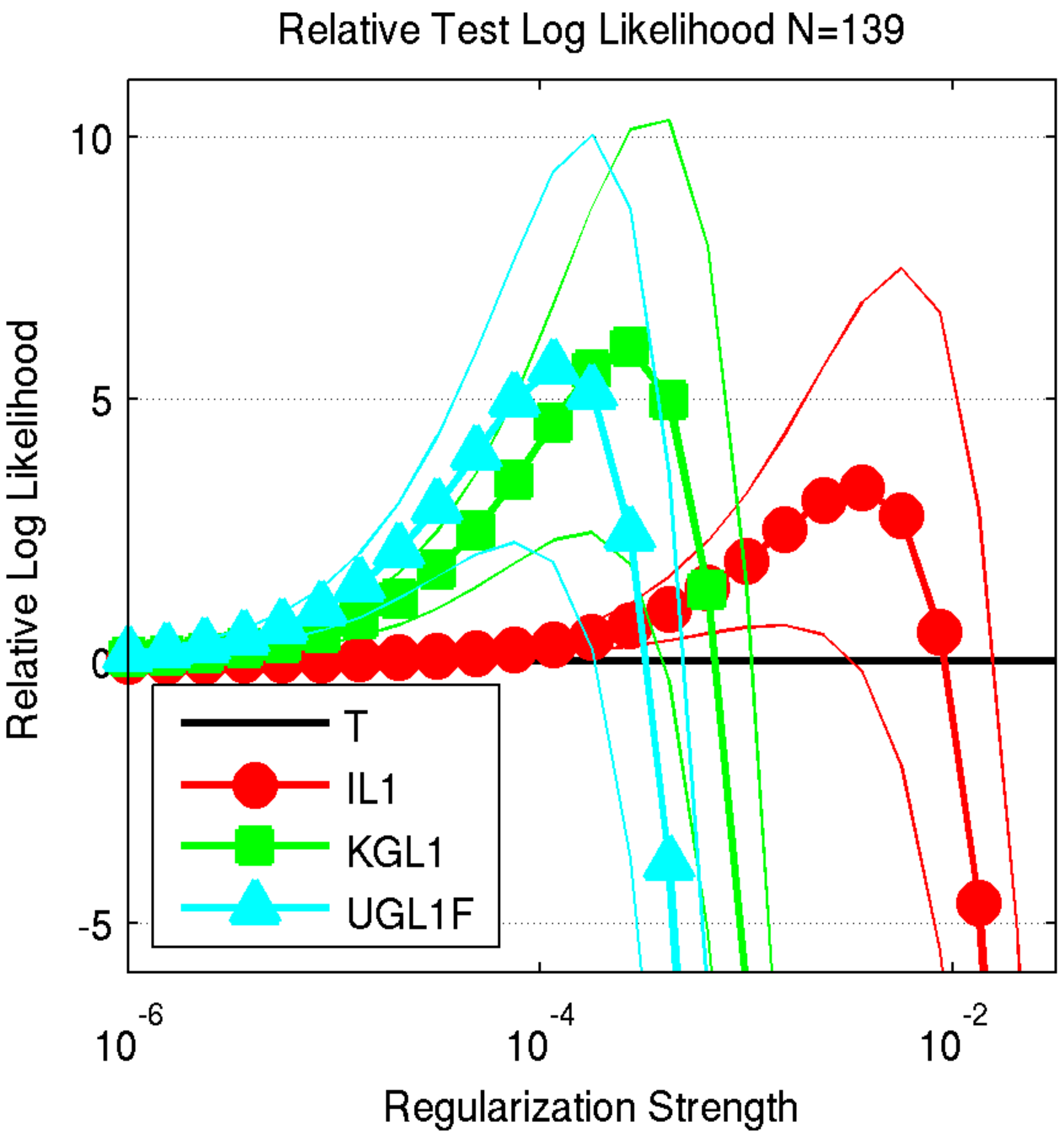
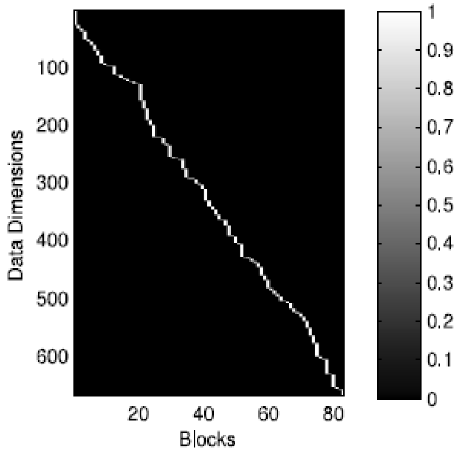


Image adapted from Gasch et al. (2000) supplemental materials.

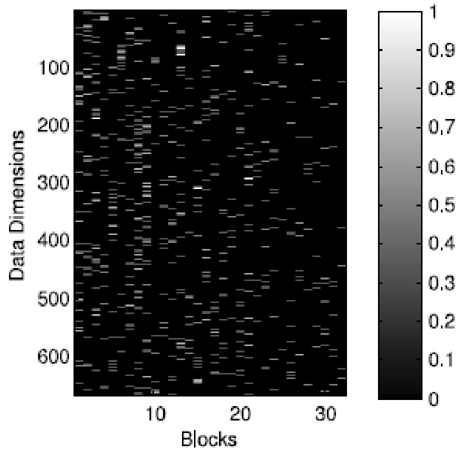
Results: *Genes Test Set Log Likelihood*



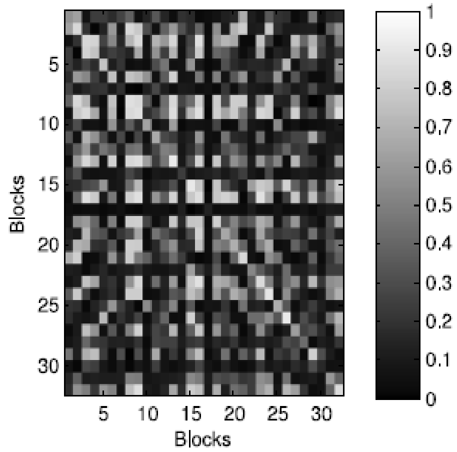
Results: *Genes Inferred Structures*



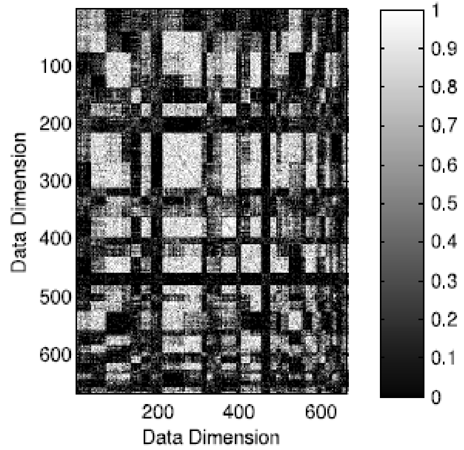
Known Grouping



Inferred Grouping

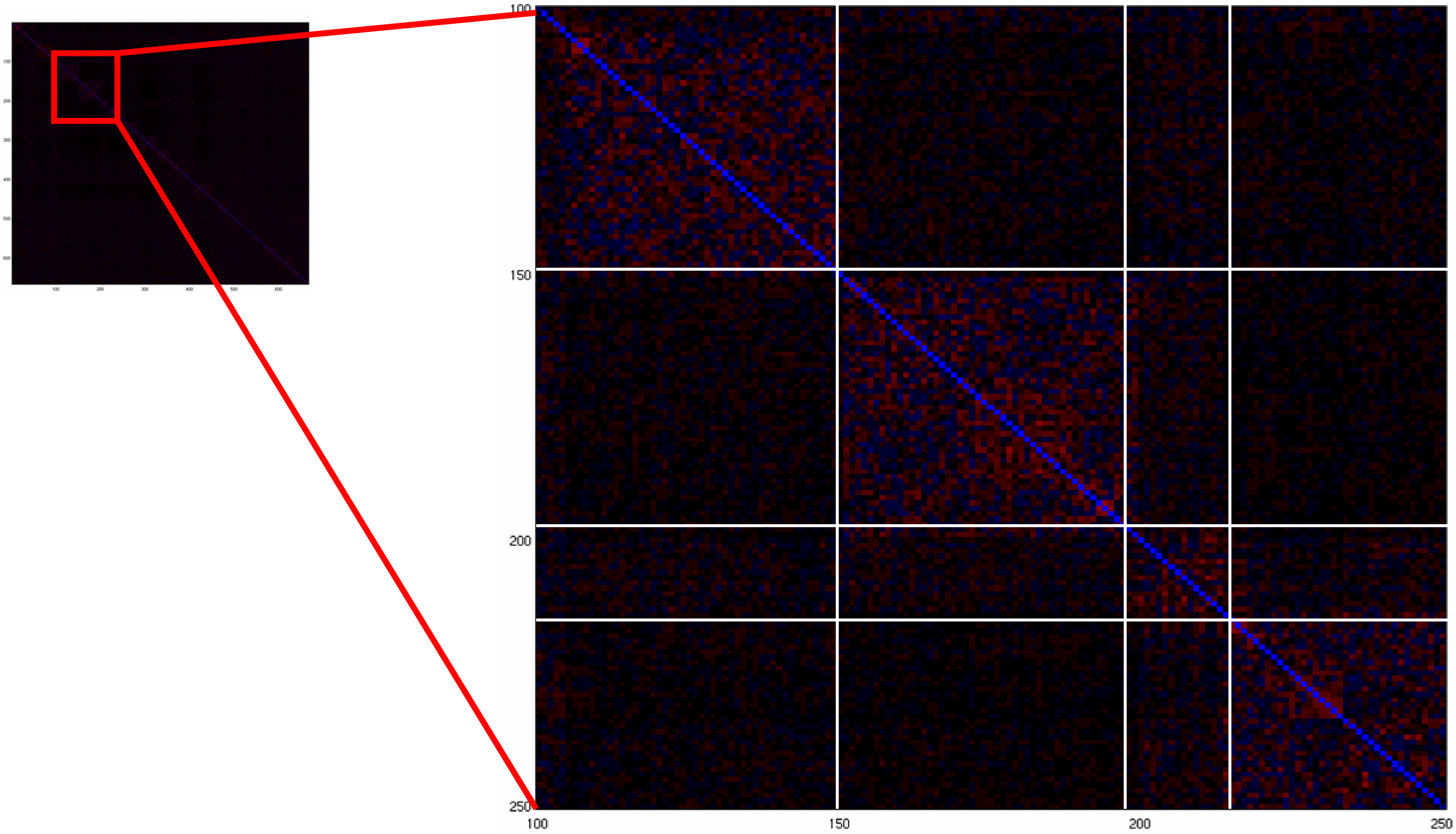


Block Model



Graph

Results: *Genes Estimated Precision*



Outline

- Introduction
- Related Work
    - Graphical Lasso
    - Group L1 Penalized Maximum Likelihood
    - Sparse Dependency Networks
- Unknown Block Structure
    - Model
    - Variational Inference
- Experiments and Results
- **Conclusions**

Conclusions and Future Work

- We have demonstrated a method for estimating sparse block-structured precision matrices when the blocks are not known *a priori*.
- The method is based on using variational inference in a hierarchical dependency network model to estimate the blocks, combined with convex optimization to estimate the precision matrix given the blocks.
- In work appearing at UAI'09, we present an alternative approach based on converting the graphical lasso and group L1/L2 penalty functions into distributions on positive definite matrices.

The End