

DYNAMIC CLUSTERING OF HIGH-DIMENSIONAL BIOLOGICAL DATA

José Sánchez

IMS Meeting, Gothenburg, August 9, 2010

In collaboration with Drs. Rebecka Jornsten and Sven Nelander.





CHALMERS

Contents

- Introduction
- The problem
- The algorithm
- Examples
 - ▣ Real data
 - ▣ Simulated data

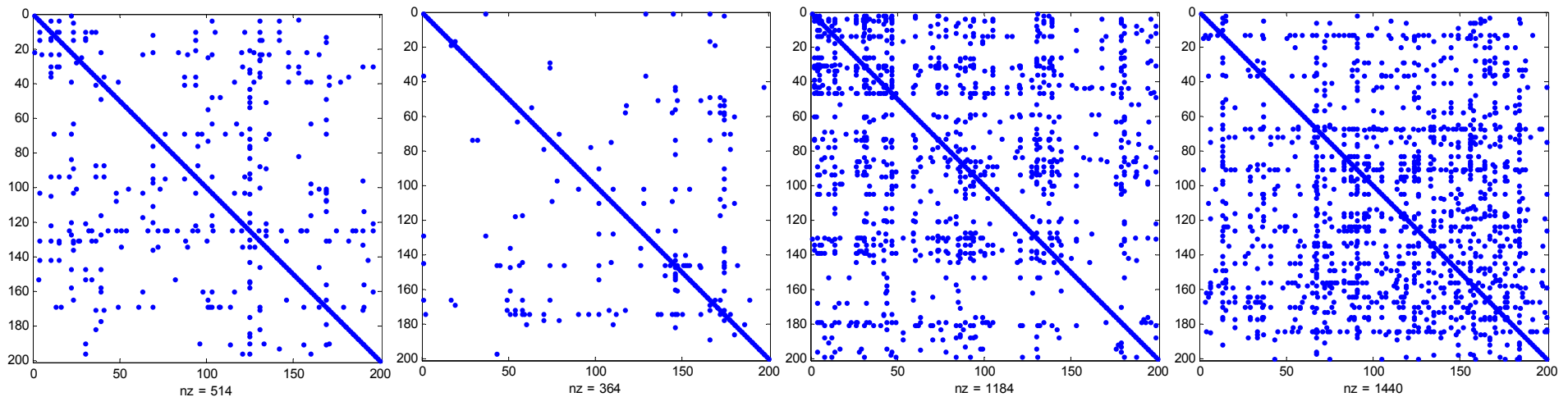
Introduction

- Analysis of gene expression data  group genes across experimental conditions. In this work we view clustering as a more dynamic problem.
- Tumors  $N(\mu, \Sigma)$ where $\Omega = \Sigma^{-1}$ will exhibit a sparse structure. Here we are interested in finding tumor clusters that reveal a change in gene-gene dependency.
- We assume a Gaussian mixture model for our data set and use a modification of the maximization step of the EM algorithm.
- We allow for specific penalization to each one of the inverse covariance matrices.

Example: 4 clusters – InvCov with non-zero entries highlighted.

* Stable gene cluster – any edges present across all tumor clusters.

* Dynamic gene cluster – any edges that are present for only a subset of tumor clusters.



The algorithm

```
for(i in 1:K)
```

```
  Par ← InitialPar
```

```
  for(j in 1:maxCV)
```

```
    Mod ← EM(data, Par)
```

```
    Mod, Par ← CV(Mod, Par)
```

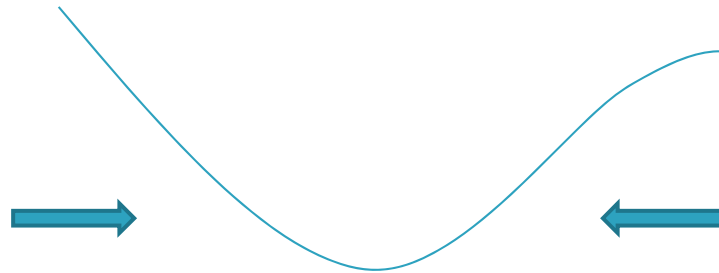
```
  endfor
```

```
endfor
```

- Data set $Y = (y_1, y_2, \dots, y_T)$ with T tumors and G genes.
- Given an initial clustering of the tumors T_1, T_2, \dots, T_K , we estimate $\Omega_1, \Omega_2, \dots, \Omega_K$ applying *glasso* to each one of the members of the initial partition.
- We then update the clustering using different penalties for *glasso* using a modified EM algorithm.
- Iterate until convergence.

Model selection

- We use successive refinement of an initial interval for the penalties.
- The used criteria to select optimal penalties and number of clusters are minimization of the BIC or maximization of the predictive likelihood.



Example: real data

- 60 patients with glioblastoma multiforme tumors from the Cancer Genome Atlas (TCGA) network.



Cancer Cell
Article

- mRNA profiles, the intersection for both platforms.
- Only 100 genes with largest variance
- Verhaak et al sub-classification of tumors available.

Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

Roel G.W. Verhaak,^{1,2,17} Katherine A. Hoadley,^{3,4,17} Elizabeth Purdom,⁷ Victoria Wang,⁸ Yuan Qi,^{4,5} Matthew D. Wilkerson,^{4,5} C. Ryan Miller,^{4,6} Li Ding,⁸ Todd Golub,^{1,10} Jill P. Mesirov,¹ Gabriele Alexe,¹ Michael Lawrence,^{1,2} Michael O'Kelly,^{1,2} Pablo Tamayo,¹ Barbara A. Weir,^{1,2} Stacey Gabriel,¹ Wendy Winckler,^{1,2} Supriya Gupta,¹ Lakshmi Jakkula,¹¹ Heidi S. Feiler,¹¹ J. Graeme Hodgson,¹² C. David James,¹² Jann N. Sarkaria,¹³ Cameron Brennan,¹⁴ Ari Kahn,¹⁵ Paul T. Spellman,¹¹ Richard K. Wilson,⁹ Terence P. Speed,^{7,16} Joe W. Gray,¹¹ Matthew Meyerson,^{1,2} Gad Getz,¹ Charles M. Perou,^{3,4,8} D. Neil Hayes,^{4,5,*} and The Cancer Genome Atlas Research Network

¹The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA

²Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA

³Department of Genetics

⁴Lineberger Comprehensive Cancer Center

⁵Department of Internal Medicine, Division of Medical Oncology

⁶Department of Pathology and Laboratory Medicine

University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

⁷Department of Statistics

⁸Group in Biostatistics

University of California, Berkeley, CA 94720, USA

⁹The Genome Center at Washington University, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA

¹⁰Department of Pediatric Oncology, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02115, USA

¹¹Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

¹²Department of Neurological Surgery, University of California, San Francisco, CA 94143, USA

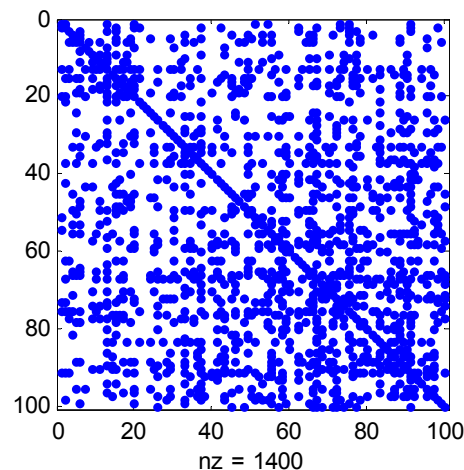
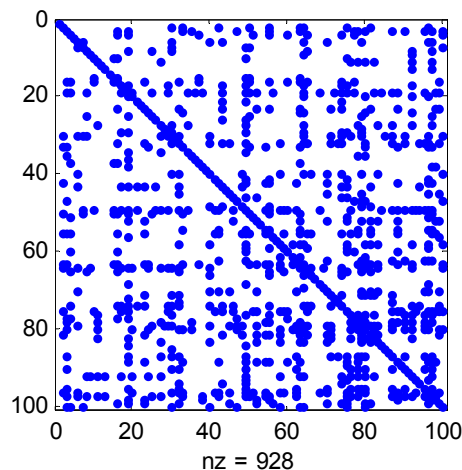
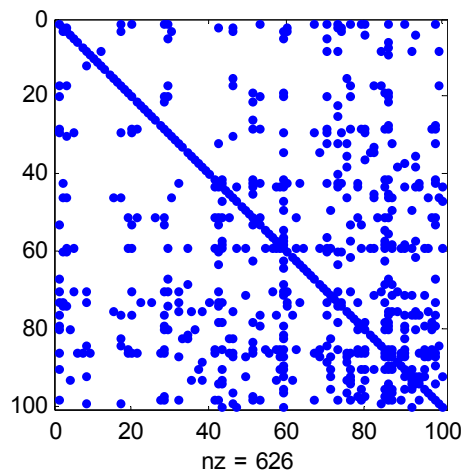
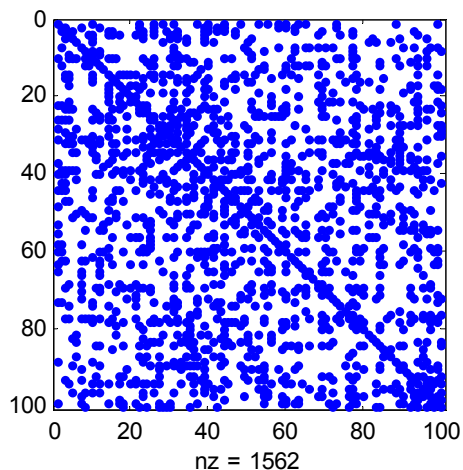
Results



- ❑ BIC tends to choose the largest penalties, stabilizing the inverse covariances matrices into almost diagonal matrices.
- ❑ Cross validation chooses a smaller penalty.
- ❑ In both cases the models with 3 or 4 clusters have the almost the same BIC or predictive likelihood.
- ❑ No large overlap with the Verhaak classification was found.

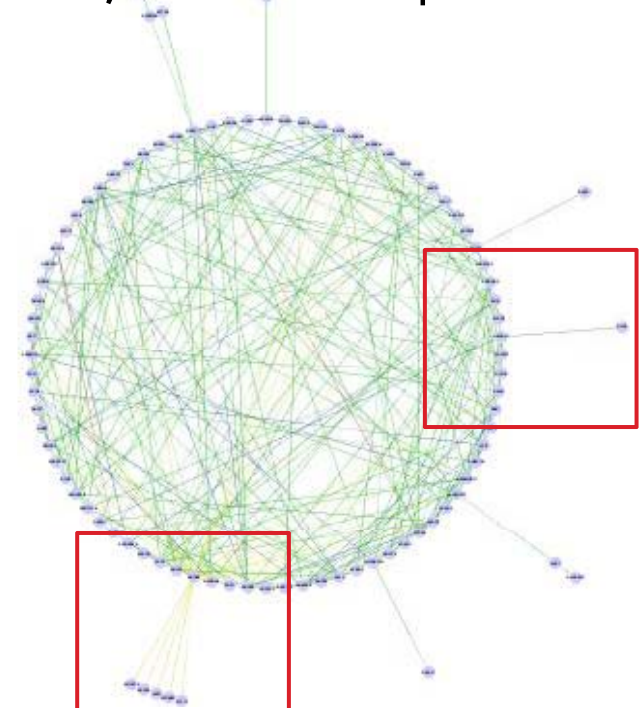
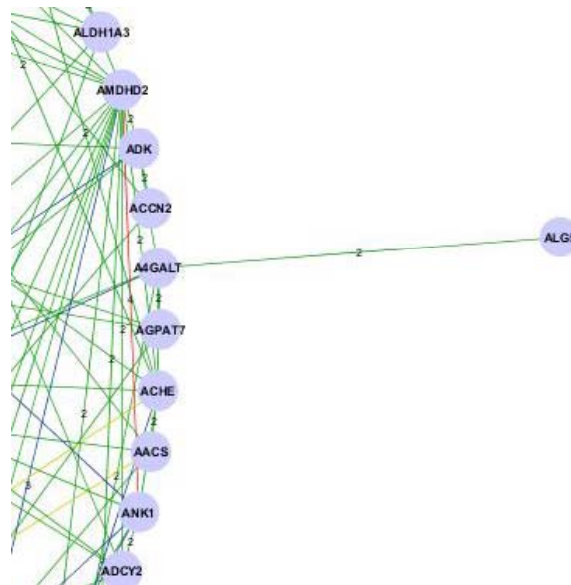
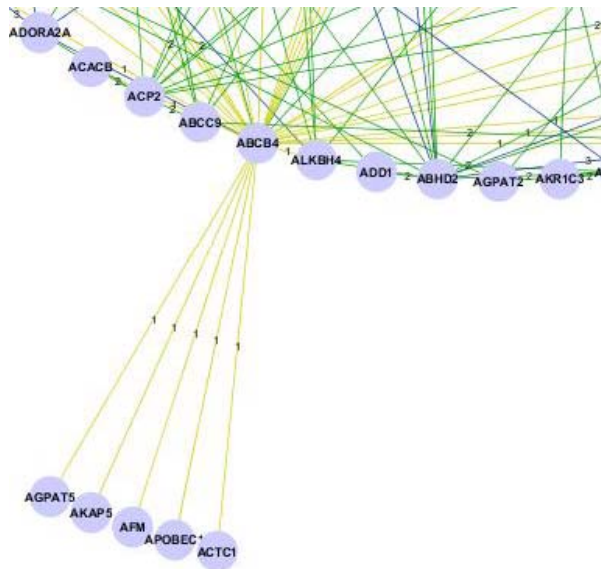
Sparsity structure of the inverse covariance matrices

- The optimal solution for a 20-fold cross validation results in different sparsity levels for the covariance matrices.
- The predictive likelihood values for 3 and 4 clusters were very close.



Gene dependencies

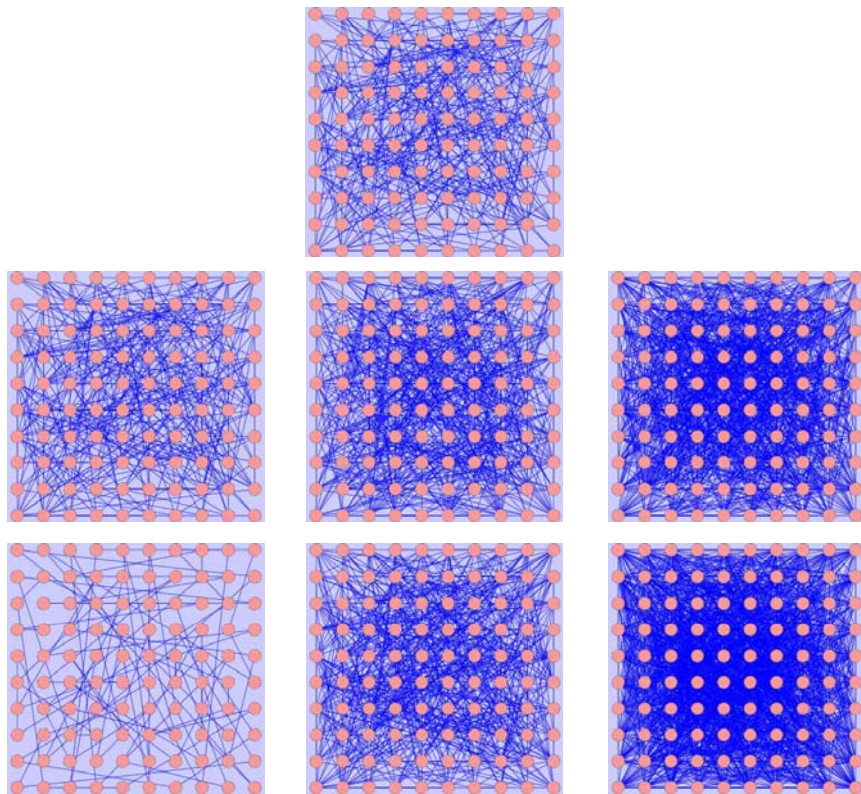
- Some gene dependencies are preserved across clusters, some are unique.
- Relevance of preserved/dynamic hubs?



Example: simulated data

- Chain networks (tridiagonal inverse covariance matrices).
- The (i,j) -th element is computed as $s_{ij} = \exp[-(s_i - s_j)/2]$ where $s_1 < s_2 < \dots < s_G$ and $s_i - s_j = U(0.5, 1)$, $i, j = 2, 3, \dots, G$.
- Heterogeneity introduced by replacing pairs of symmetrically located pairs of zeros with a value uniformly generated from the interval $[-0.1, -0.01] \cup [0.01, 0.1]$.
- We consider three settings of three clusters with different levels of sparsity.

Example: simulated data



- All three clusters with the same level of sparsity: 782 non-zero entries.
- Mildly dissimilar levels: 782, 1268 and 2238 non-zeros respectively.
- Very dissimilar levels: 298, 1268 and 3208 non-zeros respectively.

Example: simulated data



- For the same level of sparsity both criteria choose the right number of clusters.
- Again, BIC tends to choose large values for the penalties; cross validation penalizes less.
- For very dissimilar levels of sparsity cross validation performs better.

Conclusions and future work



- Resolve computational limitations to be able to process a large number of genes.
- Regularize structure of networks between the clusters as well as within.
- Investigating the results from a biological perspective.

Aknowledgements

- Rebecka Jörnsten, Chalmers University of Technology.
- Sven Nelander, Sahlgrenska Academy + members of the Nelander lab.
- The Cancer Genome Atlas consortium.



Institute of Mathematical Statistics

73rd Annual Meeting, Aug 9-13, 2010, Gothenburg, Sweden

Thank you!