Simultaneous Subset Selection via Rate-Distortion Theory

- with application to cluster and significance analysis of gene expression data

Rebecka Jörnsten

Department of Statistics, Rutgers University
rebecka@stat.rutgers.edu, http://www.stat.rutgers.edu/~rebecka

Biostatistics Day, April 25, 2008
1. **Analysis of high-dimensional data**

2. **Simultaneous Selection via Rate-distortion Theory**

3. **Cluster analysis**

4. **Significance analysis**

5. **Conclusion and Future work**
Analysis of high-dimensional data

Clustering

- Popular approach for dimension reduction
- Wide range of applications: engineering, geological data, social networks, **high-throughput biology**
  * Assign gene function via "guilt by association"
  * Suggestive of biological pathways and networks

Multiple testing

- Massive number of tests performed - how do we control the number (or proportion) of false rejections?
- Problem is encountered in e.g. clinical trials with multiple end-points, fMRI analysis, and **proteomics and genomics**.
  * Identify a set of genes whose expression levels differ between a set of experimental conditions
THINKING ABOUT THE PROBLEMS IN TERMS OF MODEL SELECTION

CLUSTERING

1. How many clusters?
2. Subset model selection: What is the most efficient description of a cluster profile?
   Example: We want to objectively be able to state that a cluster corresponds to a particular pattern across experimental conditions (e.g. static).

MULTIPLE TESTING

1. How many rejections?
2. Subset model selection: For each rejected null-hypothesis, can we identify the alternative?
   Example: We want to identify the differentially expressed genes, and the discriminatory experimental conditions.
THINKING ABOUT THE PROBLEMS IN TERMS OF MODEL SELECTION

WHY ARE THESE MODEL SELECTION TASKS SO IMPORTANT?

1. Reduce the reliance on *subjective interpretations* of the analysis outcome.
   - "This clusters *seems to* represent a static expression profile."
   - "Selected genes *appear to* primarily represent differential expression between only one of the experimental factors."

2. Waste not - want not!
   - Spend the parameter budget where it is needed.
   - If we use efficient representations of simple data structures (e.g. static cluster profiles), we may detect more subtle structures.
Challenges

Clustering

- Clustering and subset model selection are not separable.
- The search for the optimal cluster subset models is combinatorial in the number of clusters and experimental conditions.

Significance Analysis

- Double multiplicity: multiple genes, and multiple model classes for each gene. Model space is HUGE!
Proposed strategy

**Simultaneous subset selection via Rate-Distortion theory**

Challenge: clustering and subset model selection are not separable tasks

- We appeal to results in rate-distortion theory to develop a selection method that is **simultaneous** across clusters
- Generalizes to multiple testing.
We will turn the combinatorial model selection problems into a simultaneous search using results from **optimal bit allocation** in Rate-Distortion Theory.

Here: What is ”Rate”? What is ”Distortion”?
Bit-allocation

- Consider *data blocks* (block= single gene, or gene cluster).
- For each data block $k$, model $M$ results in a distortion $D_k(M)$, with rate $R_k(M)$ (e.g. $\#$ parameters $p(M)$).
- How do we allocate model complexity to each block fairly?

![Graph showing selected model and slope constraint](image)

**RD theory:** To minimize the overall distortion, (e.g. MSE), the optimal allocation is obtained at points of equal slope on the block-wise Rate-distortion curves.
The Bit-allocation/Equal Slope Principle

Why does the equal slope constraint give the optimal allocation?

For any other solution, there is at least one block-pair for which the rate-of-change of the distortion differs, and a better solution is obtained by re-allocating model complexity between these blocks.
The Bit-allocation/Equal Slope Principle

固定模型 .5bpp 与最优分配 .5bpp
Motivating Example

- Data: mRNA gene expression levels in two, divergent neural stem cell lines (one becomes neurons, the other predominantly glia).

- Timecourse; 0, 1 and 3 days after a growth factor is blocked in the media (initiates/speeds up proliferation).

- Gene cluster expression profiles appear “parallel”, “static”, “diverging”...

Motivating example

- Data: mRNA gene expression levels in two, divergent neural stem cell lines (one becomes neurons, the other predominantly glia).

- Timecourse; 0, 1 and 3 days after a growth factor is blocked in the media (initiates/speeds up proliferation).

- Gene cluster expression profiles appear “parallel”, “static”, “diverging”...

Motivating example

- Data: mRNA gene expression levels in two, divergent neural stem cell lines (one becomes neurons, the other predominantly glia).

- Timecourse; 0, 1 and 3 days after a growth factor is blocked in the media (initiates/speeds up proliferation).

- Gene cluster expression profiles appear “parallel”, “static”, “diverging”...
Model formulation

For each gene $g$ we observe a feature vector $\mathbf{x}_g$:

$$\mathbf{x}_g \mid g \text{ in cluster } k \sim \text{MVN}(\mu_k, \Sigma_k)$$

- We model each cluster profile $\mu_k = \mathcal{W}\theta_k$, where $\mathcal{W}$ is a design matrix that reflects the biological question.
- A sparse representation of $\mu_k$ is obtained if we set some of the parameters $\theta_k$ to 0.
**Model selection**

Search strategies?

- **Classification EM (CEM), a step-wise approach**
  1. Cluster the data.
  2. Perform separate model selection for each cluster.
  3. Update the clustering given the parameter constraints.

- **Combinatorial search**
  1. Cluster the data.
  2. Iterate:
     - Consider reducing a cluster specific model by one parameter.
     - Select the cluster \( k \) for which the drop does the least "damage".
  3. Stop whenever the BIC increases.

*CEM assumes clustering and cluster model selection are separable.
*The combinatorial search is greedy and computationally intensive.
Simultaneous selection

Cluster subset selection as a bit-allocation problem:

- Each cluster has its own rate-distortion curve.
- A slope constraint $\Delta$ translates to a set of subset models for the clusters, with a corresponding BIC value.
- Perform a line-search over $\Delta$ to minimize the BIC.
- This solution represents a balanced trade-off between goodness-of-fit and model complexity for all clusters.
Subset selection for cluster models

A complete search for all clusters and all models is computationally prohibitive. We compare...

- **Backward selection** = combinatorial search.
- **Simultaneous Rate-Distortion based selection** = a line search over slope constraint $\Delta$.
- **Selection via Classification EM** = separate subset selection for all clusters.
Comparison of methods

BIC as a function of the number of clusters, for the 3 search methods.

*RD is competitive with the backward search, and outperforms the CEM approach.
*We gain one cluster by using a sparse (and easy-to-interpret) representation of the cluster profiles.
The winning clustering model: many ‘static’ profiles in the neuron cell-line (e.g. clusters 1, 2, 9), cluster 4 - no cell-line/time interaction
Simulation Results

Simulating from the selected model.

Cluster selection errors

Computational cost

Number of Iterations of the clustering procedure required for model selection
Each gene is its own cluster, and has its own rate-distortion curve.

For each slope cutoff $\Delta$ we can identify the corresponding gene models.

We estimate the false discovery rate (for each $\Delta$) using bootstrap.

Alternative methods: We compare with a stepwise F-test (backward and forward).
**Selection Results**

**Gains and losses**

- With the RD method we select 909 genes at FDR=1%, compared with 808 using standard selection.
- Forward schemes are more conservative than backward schemes: RD-forward selects 890 genes.
- stepwise-F backward does not control the FDR, stepwise-F forward is overly conservative.

**Model subset classes**

- 57 sign-specific significance classes are selected (out of a possible $3^5 = 243$)...
- ... and only 10 model classes are populated by more than 20 genes.
- These 10 model classes are: main cell line effects, main time effects, interaction models where the neuron cell-line exhibits static expression.
Selection Results

Top model classes.

Class 2

Class 10

Class 11

Class 4

Class 3

Class 5

Class 8

Class 6
Simulation Results

FDR and Power across 50 simulated data sets (order RD backward/forward, F backward/forward, standard method).

- The RD method controls FDR (at 1, 5, 10%)
- ... and Power is significantly increased.
- F-backward does not control FDR, F-forward exhibits a significant loss of power.
CONCLUSIONS

SIMULTANEOUS SUBSET SELECTION VIA RATE-DISTORTION THEORY

- Cluster subset selection provides sparse and easy-to-interpret cluster profiles
- The Simultaneous Subset Selection method is fast and accurate...
- ... and can be generalized to subset selection in Multiple Testing.
- We increase Power by incorporating subset selection into Multiple Testing, and still control the FDR.

Papers are available at http://www.stat.rutgers.edu/~rebecka
Future Work

- Experiments are getting more complex - we need to automate subset selection as much as possible (e.g. consider different parameterizations for different clusters/genes simultaneously).
- Generalizations to other loss functions or coding strategies - model selection and robust estimation
- Controlling FDR at the parameter level
- Incorporating prior information into clustering (current work with Sunduz Keles).
Acknowledgements

- Professor Ron Hart and Loyal Goff, Keck Center for Collaborative Neuroscience, Rutgers University
- NSF and EPA
**Simulation Results**

Using an "incorrect" parameterization.

- RD controls FDR even for the inefficient parameterizations, and exhibits competitive Power with standard selection.

- Stepwise F does not control the FDR when an inefficient parameterization is used.
Jörnsten. Simultaneous Subset selection via Rate-Distortion Theory