

# Statistical genomics and spatial statistics: Incorporating biological knowledge of genes into analysis of genomic data

Wei Pan

(joint work with Peng Wei)

Division of Biostatistics, SPH

University of Minnesota

Dec 2007

## Outline

- Problem
- Standard mixture model
- Stratified mixture model
- Spatially correlated mixture model
- Numerical Results: real and simulated data
- Discussion

## Introduction

- Problem: genomic discoveries  
which of the  $G$  genes satisfy a specified condition?
- Problem 1: detecting differentially expressed (DE) genes based on microarray expression data
- Problem 2: detecting binding targets of a TF based on ChIP-chip data
- A related problem: prostate cancer (PC) screening by PSA
- Features:
  - Unsupervised learning/discovery: no or few known cases/controls; e.g. cannot apply logistic regression; use mixture model/clustering.
  - Many genes/subjects: somewhat similar; borrow info.
  - Data: high noise level.

- Statistical problem: testing  $H_{0,i}$  vs  $H_{1,i}$  for each gene  $i$ .
  - $H_{0,i}$ : gene  $i$  is equally expressed for Problem 1;
  - $H_{0,i}$ : gene  $i$  is not a target of the TF for Problem 2;
  - $H_{0,i}$ : subject  $i$  does not have PC for Problem PC;
  - $H_{1,i}$ : opposite of  $H_{0,i}$  (i.e. gene  $i$  is DE for Problem 1, is a target for Problem 2, subject  $i$  has PC for Problem PC).
- Given microarray data  $\implies Z_i$ 's
 

$Z_i$ : a summary statistic against  $H_{0,i}$  for gene  $i$ ;  
e.g. a fold change, t-type statistic, or even p-value.
- We transform  $Z_i$  such that the null distribution of  $Z_i$ 's (i.e. for those genes satisfying  $H_{0,i}$ ) is  $N(0, 1)$ .  
e.g. If  $Z_i = P_i$  is a p-value,  $z_i = \Phi^{-1}(1 - P_i)$ .
- The null distribution may not be exactly  $N(0, 1)$ , called theoretical null, and hence may need to be estimated as  $N(\mu_0, \sigma_0)$ , called empirical null (Efron 2004, JASA)

- From now on, we work with  $z_i$ 's (i.e. transformed  $Z_i$ 's).

## Standard mixture model

- Many references: Efron et al (2001, JASA); Newton et al (2001, JCB);...
- A hierarchical model:
- Prior probability:  $\pi_0 = \text{Prob}(H_{0,i})$  for any  $i$ .  
a constant! common across the genes!
- Null distr:  $f(z_i|H_{0,i}) = f_0(z_i)$ ;
- Non-null distr:  $f(z_i|H_{1,i}) = f_1(z_i)$ ;
- Marginally,  $z_i$ 's are iid from  
 $f(z_i) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i)$ ,  
a standard mixture model.
- Key:  
all the genes are treated equally and independently *a priori*;  
reasonable?

- Inference:

$$Pr(H_{1,i}|z_i) = \frac{(1-\pi_0)f_1(z_i)}{f(z_i)} = 1 - \frac{\pi_0 f_0(z_i)}{f(z_i)} \propto \frac{f_1(z_i)}{f_0(z_i)} = LRT.$$

Rank the genes based on their  $Pr(H_{1,i}|z_i)$  or LRT.

- False discovery rate (FDR) estimation (Newton 2004, *Biostatistics*)

Decision rule: for any given cut-off value  $c$ , rejects  $H_{0i}$  if and only if  $Pr(H_{1,i}|z_i) > 1 - c$ , then

$$\widehat{FDR}(c) = \frac{\sum_i [1 - Pr(H_{1,i}|z_i)] 1[Pr(H_{1,i}|z_i) > 1 - c]}{\sum_i 1[Pr(H_{1,i}|z_i) > 1 - c]}.$$

$$FDR = E \left( \frac{\# \text{false positives}}{\# \text{claimed positives}} \right).$$

## Stratified mixture model

- Reference: Pan (2005, *Statistical Applications in Genetics and Molecular Biology*)
- Known: the genes are annotated in  $K > 1$  GO categories or pathways,  $G_1, \dots, G_K$ .  
known: the genes in the same group should be *more similar* to each other than those from different groups!
- How to take advantage?  
treat the genes in different groups **differently** *a priori*.
- Prior probability:  $\pi_0^{(k)} = \text{Prob}(H_{0,i} | i \in G_k)$ .  
NOT a constant; group-dependent!
- Null distr: same as before;  $f(z_i | H_{0,i}) = f_1(z_i)$ .
- Non-null distr: group-specific;  $f(z_i | H_{1,i}, i \in G_k) = f_1^{(k)}(z_i)$ .
- Marginally,  $z_i$ 's for those in  $G_k$  are iid as



$$f(z_i|i \in G_k) = \pi_0^{(k)} f_0(z_i) + (1 - \pi_0^{(k)}) f_1^{(k)}(z_i),$$

but the marginal distribution depends on  $k$ : genes from different  $G_k$  have different distributions!

$\implies$  treat genes differently *a priori*

- Inference: same as before except working on each  $G_k$  one by one—stratified analysis!
- Efron (in press, AoAS): a general problem; theory.
- Problem PC: use age-adjusted PSA!
  - According to NCI website:
  - A cut-off point for each 10-year age group;
  - For example, claimed normal (NCI website):
  - 1) if  $PSA < 2.4$  ng/mL for men  $< 50$  year-old;
  - 2) if  $PSA < 6.5$  ng/mL for men  $\geq 70$  year-old.
  - Why? PSA level increases with age.

- But still controversial.
- A practical problem: depends on the choice of  $G_k$ 's  
GO: thousands of the groups;  
GO: DAG; hierarchical: higher level categories are more general, while lower ones more specific  
 $\implies$  trade-off: group homogeneity vs group size!
- Hierarchical mixture model (Pan 2006, *Applied Statistics*)
- Main ideas:
  - 1) each GO category is a stratum;
  - 2) borrowing information: parameters from a category are related to that of its parents; shrinking its sample estimate towards that of its parent!

## Spatially correlated mixture model

- A problem with the stratified mixture model: choice of  $G_k$ 's.
- Some argue that gene functions should be characterized by some categories, rather, by their inter-relationships (Marcotte)  
 $\implies$  gene networks
- gene networks: many types; can be general here.  
undirected graph: genes are nodes; an edge indicates “direct relationship” between the two genes.  
**basic assumption:** any two connected genes in a network are more similar (i.e. more likely to satisfy or not satisfy  $H_0$  together) than two random picks.
- Prior probability:  $\pi_{i,0} = \text{Prob}(H_{0,i})$  for gene  $i$ .  
Key: gene-specific!
- Null distr: same;  $f(z_i|H_{0,i}) = f_0(z_i)$ .

- Non-null distr: same;  $f(z_i|H_{1,i}) = f_1(z_i)$ .
- Marginally  $z_i$  is distributed as  

$$f_i(z_i) = \pi_{i,0}f_0(z_i) + \pi_{i,1}f_1(z_i),$$
- Too many parameters  $\pi$ 's  $\implies$  borrowing information!  
 have not used information in network yet!
- Assume two latent Markov random fields  
 $\mathbf{x}_j = \{x_{i,j}; i = 1, \dots, G\},$   
 $\pi_{i,j} = \exp(x_{i,j}) / [\exp(x_{i,0}) + \exp(x_{i,1})]$  for  $j = 0, 1$ .
- $\mathbf{x}_j$ : intrinsic Gaussian conditional autoregression (CAR) model  
 (Besag and Kooperberg 1995, B'ka)  

$$x_{i,j} | \mathbf{x}_{(-i),j} \sim N \left( \frac{1}{m_i} \sum_{l \in \delta_i} x_{l,j}, \frac{\sigma_{C_j}^2}{m_i} \right),$$
 where  $\delta_i$ : indices for the neighbors of gene  $i$ ;  $m_i = |\delta_i|$ .  
 neighborhoods: determined by a gene network!
- A Bayesian implementation ... see Wei and Pan (RR

#2007-032)

used MCMC; inference is based on posterior probabilities, e.g.  $\widehat{Pr}(H_{0,i}|data)$ .

a standard mixture model can be similarly implemented.

- Originally proposed by Fernandez and Green (2002, JRSS-B) for spatial statistics; to avoid over-smoothing near “edges”. applied to CGH data by Broet and Richardson (2006, *Bioinfo.*): 1-dim smoothing over a chromosome to “change point” detection.

## An example

- Data: 3 replicates of ChIP-chip experiments for yeast *S. cerevisiae* by Lee et al (2002, Science);  $G \approx 6000$   
TF: GCN4; involved in response to amino acid starvation;  
Used their  $p$ -values.
- Positive (negative) control set: genes believed to be (not to be) the transcriptional targets of GCN4;  $n = 80$  (900).  
compiled by Pokholok et al (2005, Cell); based on 3 sources of data: a newer generation of ChIP-chip; gene expression; DNA motif analysis).
- Gene network: *computationally* constructed by Lee et al (2004, Science).  
two connected genes: functional linkage;  
based on multiple data sources: gene expression, protein-protein interaction, gene co-citation, gene fusion and

phylogentic profiles;

- Used their ‘ConfidentNet’: 4681 nodes, 34000 edges.  
summary of # direct neighbors: min=1, 25%=2, 50%=6,  
75%=13, max=188.
- Merged the data and network.  
 $G = 4616$  genes/nodes, 33432 edges;  
positive control set: 66 genes;  
negative control set: 770 genes;
- Subnetwork with only control genes: Fig 1  
clustering?
- Evaluation: used only the two control sets to estimate  
sensitivity and specificity  $\implies$  ROC curve.
- Model fitting: Fig 2.

Standard:

$$\hat{f}(z_i) = 0.91\phi(z_i; 0, .80^2) + 0.037\phi(z_i; -1.98, .40^2) + 0.058\phi(z_i; 1.67, 1.94^2),$$

Spatial:

$$\hat{f}(z_i) = \hat{\pi}_{i,0,1}\phi(z_i; 0, .63^2) + \hat{\pi}_{i,0,2}\phi(z_i; -0.38, 1.02^2) + \hat{\pi}_{i,1,1}\phi(z_i; 0.75, 1.53^2)$$

averages of  $\hat{\pi}_{i,0,1}$ ,  $\hat{\pi}_{i,0,2}$ ,  $\hat{\pi}_{i,1,1}$ : 0.500, 0.314 and 0.186.

- Statistical power: ROC curves in Fig 3



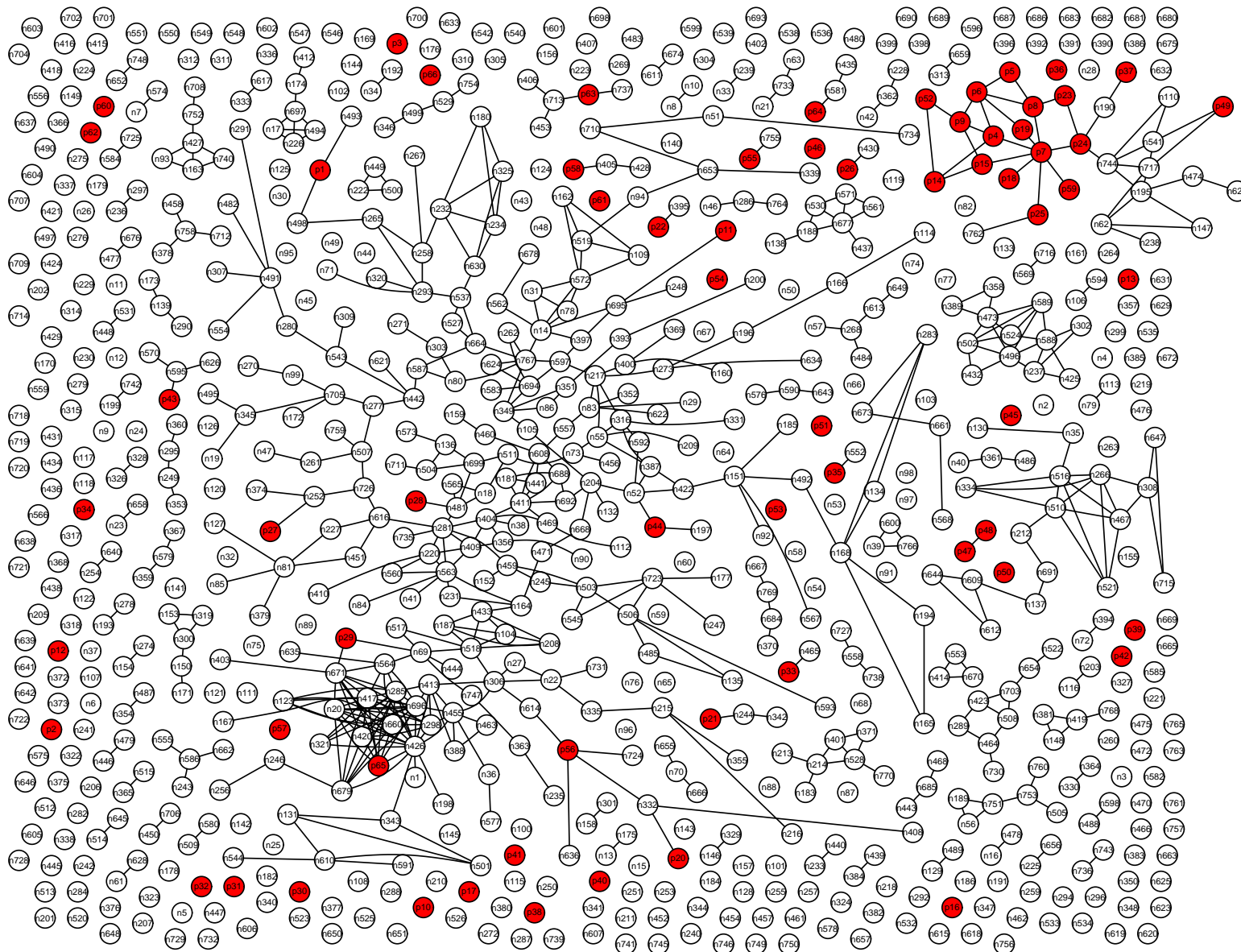


Figure 1: Subnetwork consisting of positive control genes (dark ones)

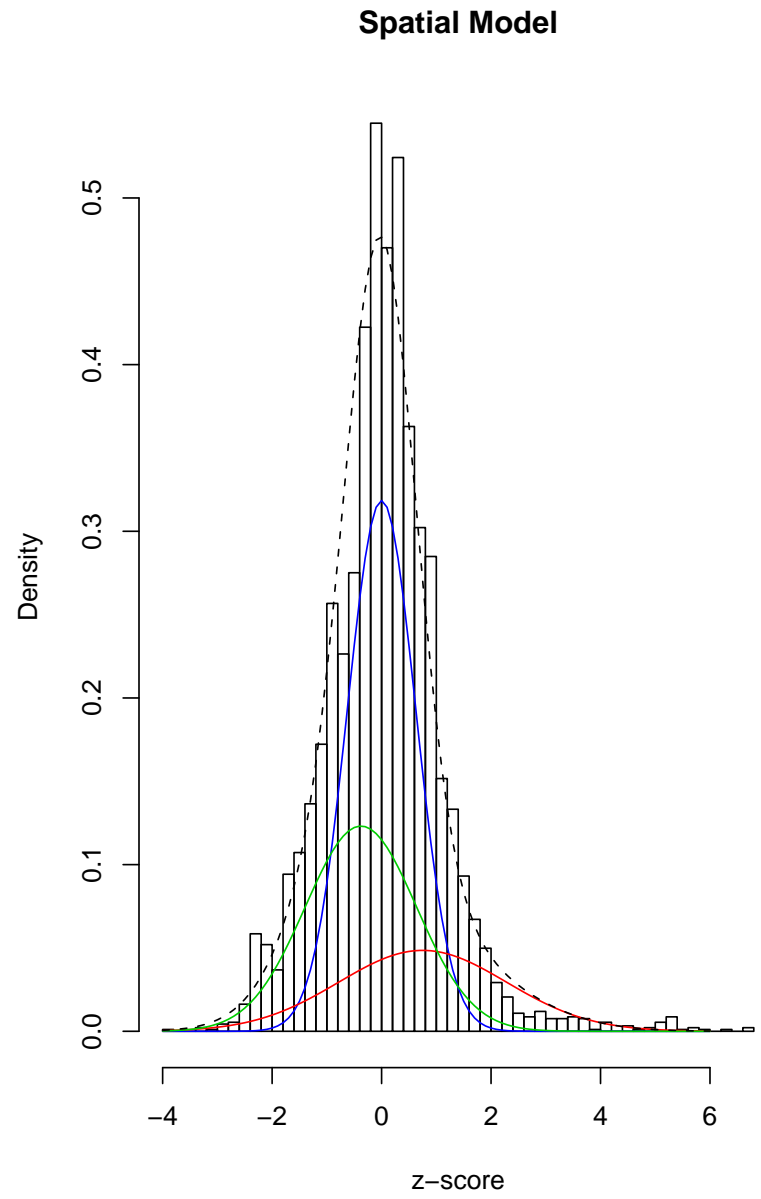
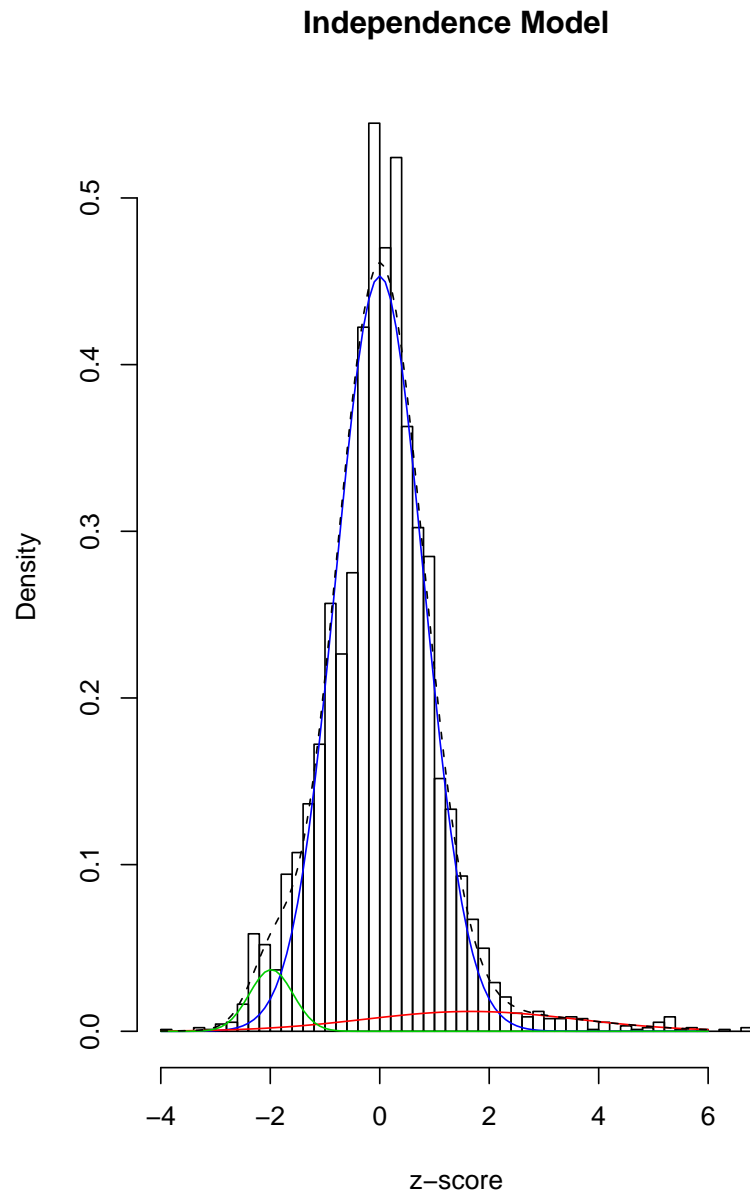


Figure 2: Fitted mixture models.

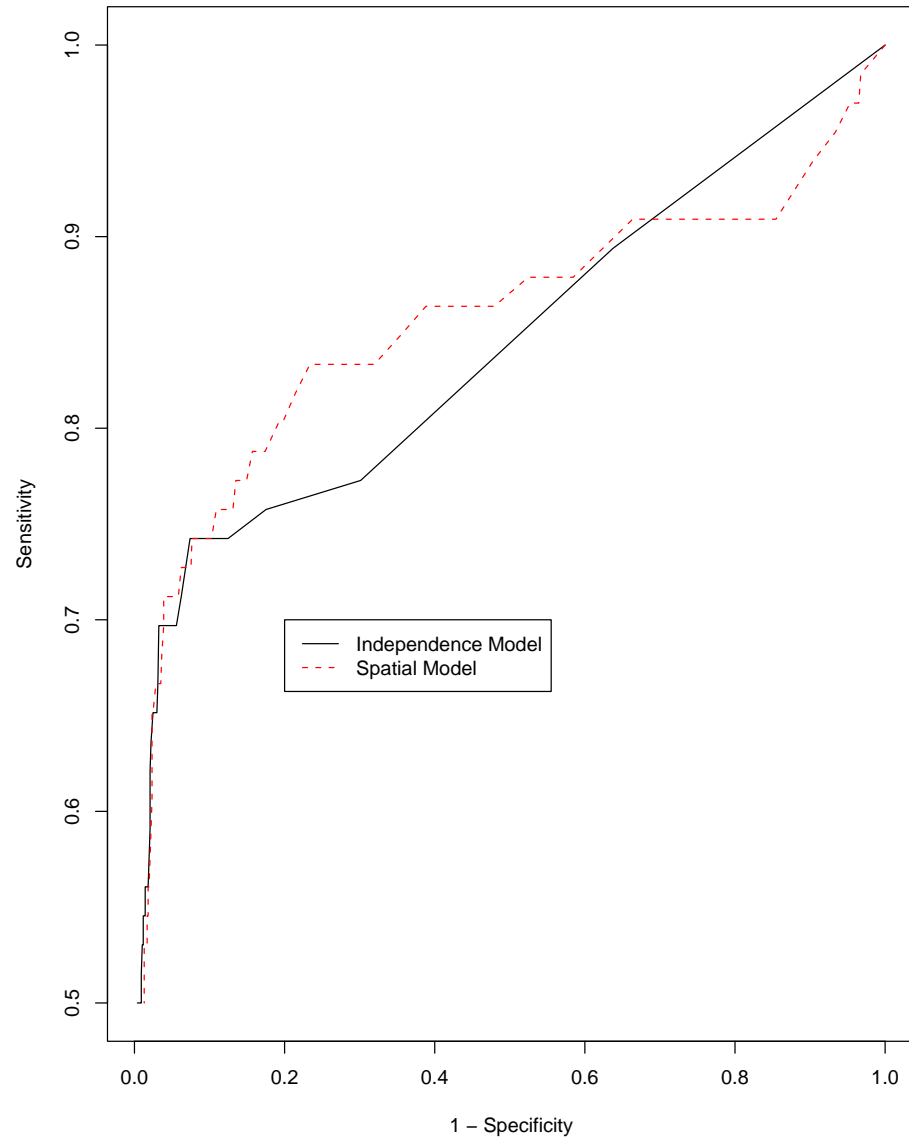


Figure 3: ROC curves for the two methods applied to the real data.

## Example genes

- ARG8: in the positive control set.
  - posterior prob: =0.728 by the spatial model; =0.023 by the standard model.
  - data in Lee et al (rich medium): binding ratio=1.02; used here.
  - new data by Harbison et al (2004, Nature) (plus other conditions: amino acid starvation and nutrition deprivation): binding ratio=5.0; p-value= $10^{-11}$ .
  - ARG8: annotated in GO *BP: amino acid biosynthetic process*, while GCN4 is a transcriptional activator of amino acid biosynthetic genes in response to amino acid starvation. –a reasonable target.
  - How detected by the spatial model? ARG8 is the direct neighbor of 4 positive control genes but of *none* negative

control genes. –borrowing information: its prior prob was estimated to be 0.733 by the spatial model, in contrast to 0.058 by the standard mdoel.

- TRP5: not in either control set.
  - Prior prob: 0.716 by the spatial model vs 0.058 by the standard model;
  - Posterior prob: 0.723 vs 0.032;
  - binding ratio: =1.15 in Lee et al; =1.21 in Harbison et al;
  - Beyer et al (2006, PLoS Comp Biol): predicted to be a target of GCN4;
  - Annotated in GO ‘BP: amino acid biosynthetic process’; likely a target!
- ICY2: a positive gene; has 6 nighbors: 2 negative and none positive.
  - Prior prob: 0.668 by the spatial model vs 0.058 by the

standard model;

– Posterior prob: 0.836 vs 0.548. –detected!

– its two negative control genes: ADY2 and CRS5,

– 1) ADY2:

Prior prob: 0.08 by the spatial model vs 0.058 by the standard model;

Posterior prob: 0.06 vs 0.02;

– 2) CRS5:

Prior prob: 0.12 by the spatial model vs 0.058 by the standard model;

Posterior prob: 0.09 vs 0.02;

—both negative neighbors are not false positives!

## Simulation

- Starting from the same network as in the real data, simulated a binary MRF for the *latent states* (i.e. whether  $H_{0,i}$  holds or not).
  - Note: MRF not for  $\mathbf{x}_j$  as used in our model; we have a mis-specified spatial model!
  - updated according to the conditional distribution; stopped after 10 iterations, nearly stable;
  - 4609 nodes, 33432 edge; 183 positive genes, and others negatives.
  - accordingly simulated  $z_i$  from the fitted model:  $\phi(0, 0.63^2)$  for the null distr,  $\phi(0.75, 1.53^2)$  for non-null.
- Simulated 5 datasets: ROC curves, Fig 4
- Sensitivity to mis-specified network structures: Fig 5 randomly removed 5% edges;

randomly added 5% edges;

randomly removed 5% and then added 5% edges.

- Sensitivity to hyperparameters: Fig 6  
prior for the precision of the mixture model; tried to use non-informative priors when possible.



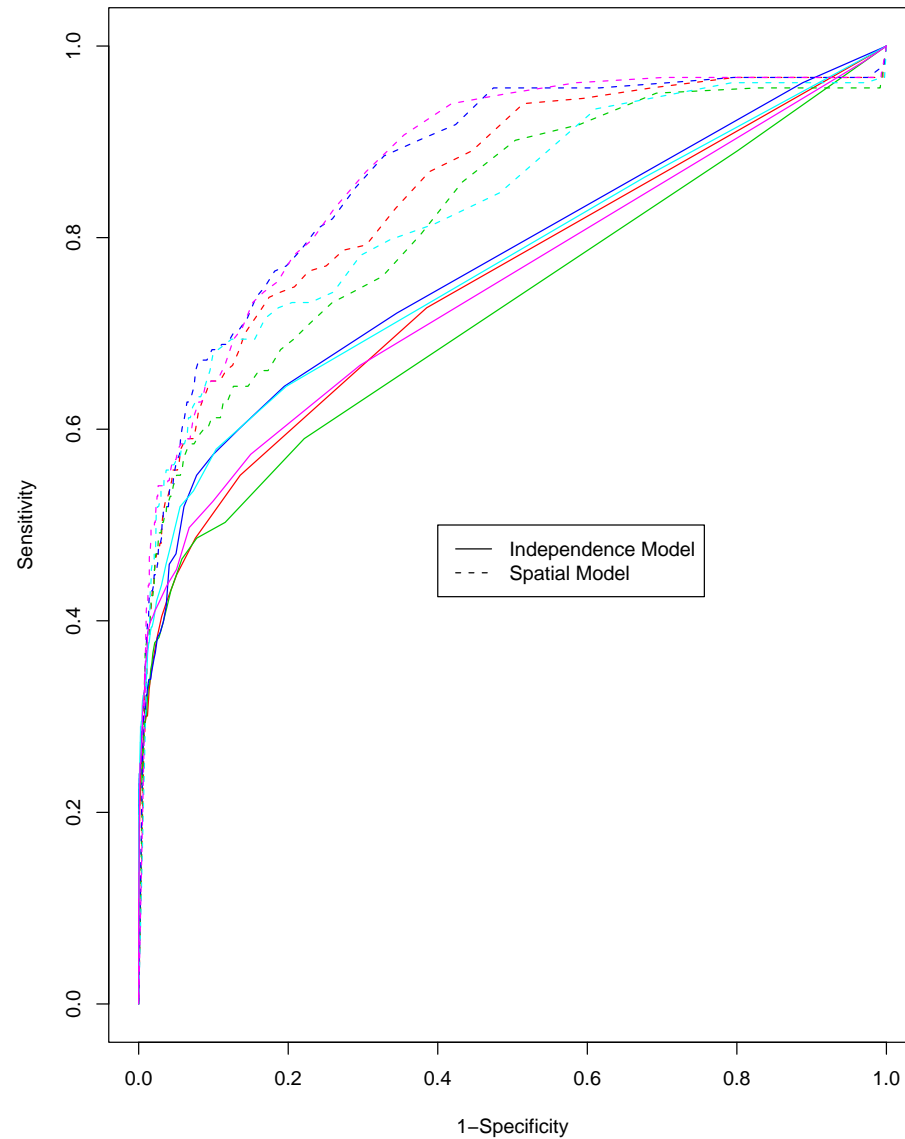


Figure 4: ROC curves for the two methods applied to five simulated data sets. Dashed lines are for the spatial model; solid lines are for the independence model.

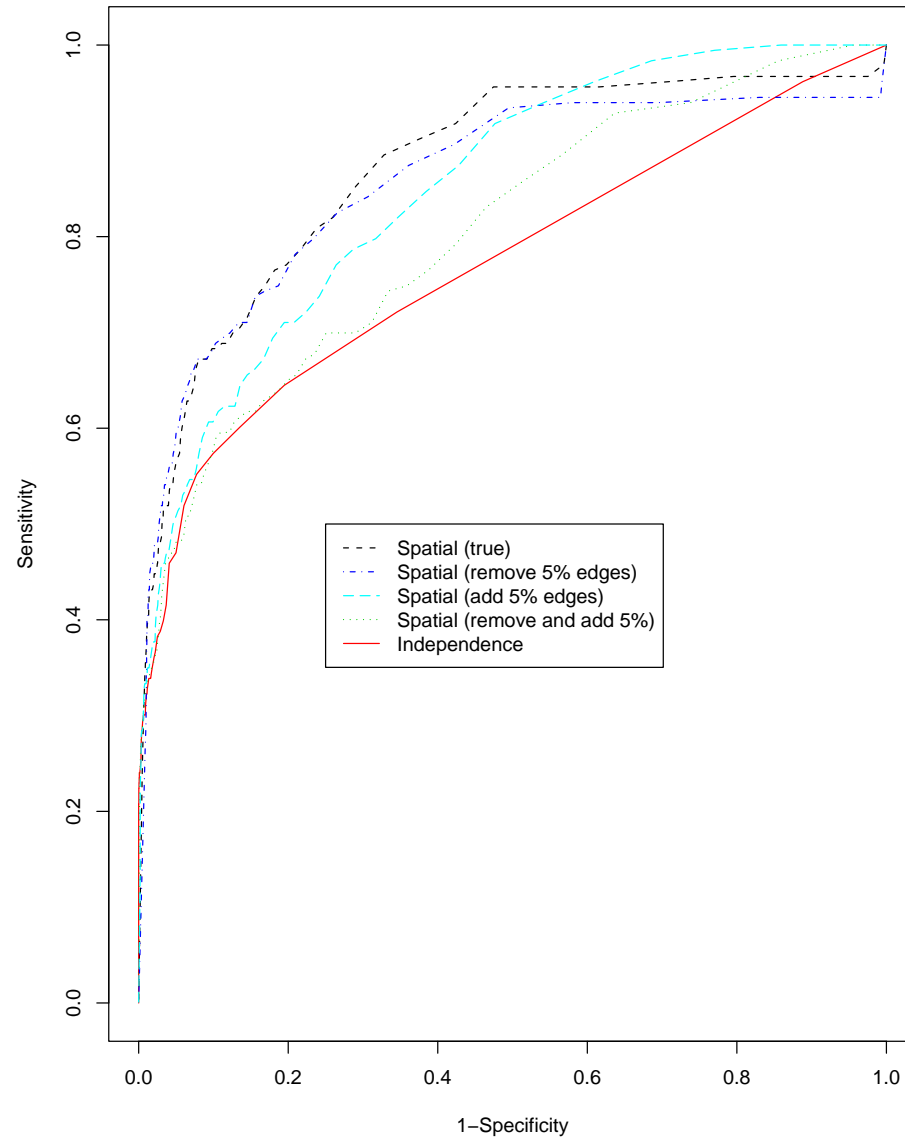


Figure 5: ROC curves for misspecified network structures.

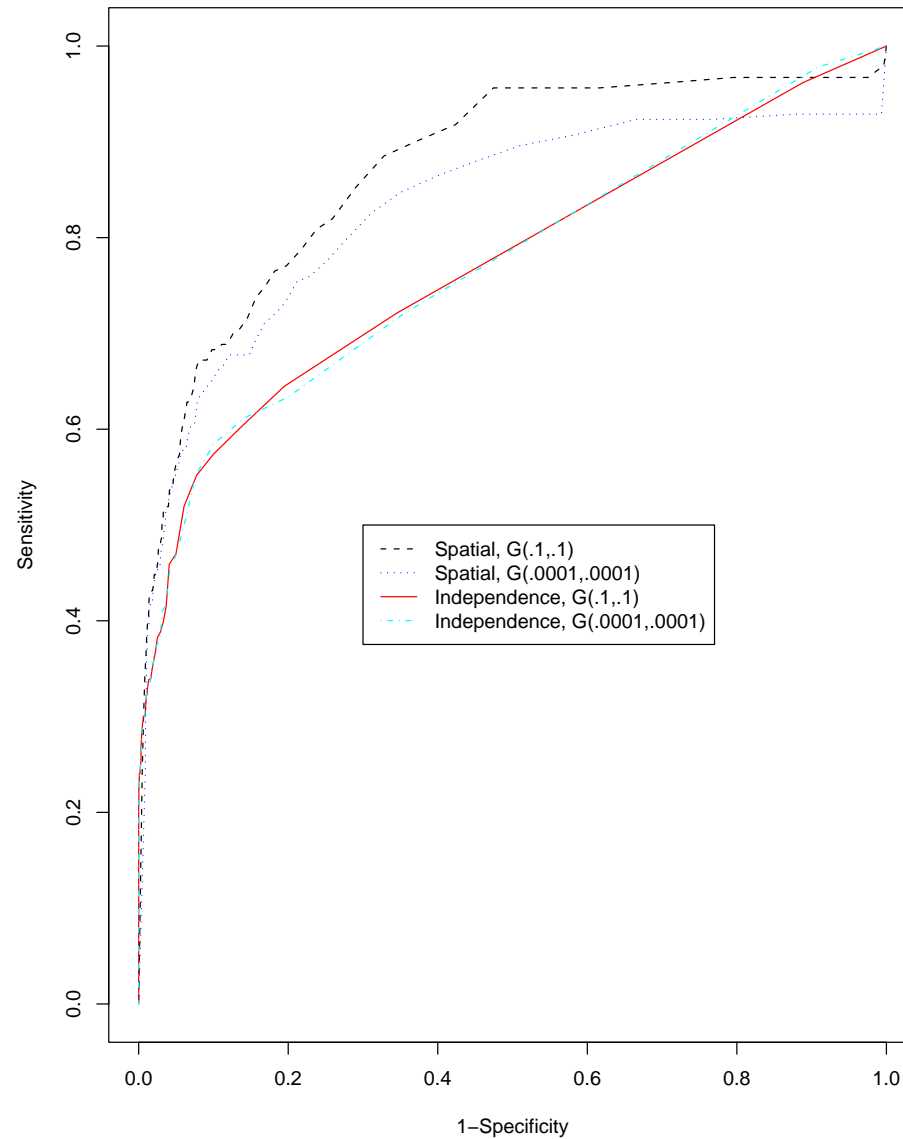


Figure 6: ROC curves for sensitivity analysis (two different priors for the precision parameters of the normal mixture components).

## Discussion

- A (happy or productive?) marriage of statistical genomics and spatial statistics.
- More comparisons, applications (e.g. to expression data) and extensions.
  - Wei and Li (2007, *Bioinformatics*): modeling the states of  $H_{0,i}$  as a binary MRF; use ICM (Besag, 1986, JRSS-B).  
give only point estimates; sensitivity to mis-specified network?  
alternative: fully Bayesian.
  - Integrating multiple sources of data (Pan et al in press, *Statistica Sinica*; Pan et al in press, PSB'08; Xie 2006 PhD Thesis).
  - Peng Wei's thesis?
- Applicable: clustering genes with expression profiles for gene

function discovery.

stratified model: Pan (2006, *Bioinformatics*).

challenge here: computationally too demanding?

penalized methods: connected to Bayesian

- Extensions:
  - variable/gene selection in sample classifications.  
Feng Tai's thesis?
  - variable/gene selection in sample clustering.  
Benhuai Xie's thesis?
- My longer-term plan: apply to genome-wide association studies with SNP data.
  - a high-dim problem;
  - are stat genomics and stat genetics converging?
  - E.g., using gene chromosome location, functional groups/pathways or porotein-protein interaction networks...

- Using linkage analysis as prior for association study (Roeder et al 2006, AJHG) using weighted p-values.  
Extending to incorporating network?
- Yanni Zhu's thesis?

Acknowledgement: This research was supported by NIH and a UM AHC Faculty Research Development grant.

You can download our papers from  
<http://www.biostat.umn.edu/rrs.php>

**Thank you!**