Spatially Correlated Mixture Models with Application in Genomic Hypothesis Testing

Jie Ren

University of Minnesota Duluth

Advisor: Prof. Kang James

August 31, 2009
Abstract

An important task in genomic studies is the detection of genes satisfying certain conditions, such as being regulatory targets of a transcription factor (TF). With high-throughput data, e.g. microarray data, this is usually formulated as a simultaneous hypothesis testing problem. Gaussian mixture model (GMM) can be used in such a problem. However, standard GMM assumes that all the genes have an identical and independent distribution \textit{a priori}, which contradicts the biological fact that genes work coordinately as dictated by gene networks. Wei and Pan (2008) proposed a spatially correlated mixture model to integrate gene network information into the statistical analysis of genomic data. We applied the standard model, Wei and Pan’s model, and a similar model with modifications to a real ChIP-chip data set. By comparing the ROC curves, we found that both spatial models had better performance than did the standard model. The modified spatial model gave even higher statistical power, as well as the ability to discover more potential target genes.

\textbf{Key Words:} ChIP-chip; Conditional autoregression; Gene networks; Hypotheses testing; Markov random field; Microarray; Mixture models.
Acknowledgments

In the first place I would like to express my deepest appreciation to my advisor Dr. Kang James for her supervision, advice and encouragement. She has been providing me with persistent help during my two-year study at the University of Minnesota Duluth (UMD). She helped me pick this interesting topic, guided me through those difficult theory parts, and kept encouraging me throughout this project. Without her support, this report would not have been possible.

I gratefully thank my committee members, Dr. Barry James and Dr. Marshall Hampton, for their precious time to read this report and their valuable comments and suggestions. Dr. Barry James helped me correct errors in the earlier draft. I have also benefited by his knowledge and ideas from his two fabulous statistical courses.

Last but not least, I owe my warmest thanks to all my friends who have offered their generous help and warm encouragement during this project. I also would like to thank the Department of Mathematics and Statistics at UMD for providing me with financial support as well as a great study environment and research facilities.
Table of Contents

1 Introduction ........................................................................................................... 7

2 Biological Background ........................................................................................ 9

   2.1 ChIP-chip experiments .................................................................................. 9
   2.2 Gene networks ................................................................................................. 11

3 Methods ................................................................................................................. 13

   3.1 Problem .......................................................................................................... 13
   3.2 Standard Gaussian mixture model ................................................................. 13
   3.3 Spatially correlated mixture model ................................................................. 14
      3.3.1 Modifications .......................................................................................... 15
   3.4 Inference .......................................................................................................... 16

4 Example .................................................................................................................. 18

   4.1 Data ................................................................................................................. 18
   4.2 Parameter estimates ....................................................................................... 19
   4.3 Statistical power .............................................................................................. 23

5 Discussion .............................................................................................................. 25

A Appendix ............................................................................................................... 26

   A.1 Prior specifications ......................................................................................... 26
      A.1.1 For a three-component standard mixture model ..................................... 26
      A.1.2 For three-component spatial mixture models with zero and logit constraints .... 26
   A.2 MCMC Algorithm .......................................................................................... 26
   A.3 WinBUGS codes for implementing the three methods .................................... 27
      A.3.1 For a three-component standard mixture model ..................................... 27
      A.3.2 For a three-component spatial mixture model with zero constraint ............. 28
      A.3.3 For a three-component spatial mixture model with logit constraint and singletons..30

References ............................................................................................................... 33
List of Figures

2.1-2.7 ChIP-chip process ................................................................. 10
2.8 3D layout of the 950 strongest interactions in “ConfidentNet” .................. 12
4.1 Fitted plots against data histograms .................................................. 21
4.3 Convergence check ........................................................................ 22
4.4 ROC curves for the three models applied to the ChIP-chip data ................. 24
List of Tables

4.1 Sample ChIP-chip data ................................................................. 19
4.2 Parameter estimates for spatial models ........................................... 20
4.3 Parameter estimates for modeling Markov random fields by ICAR .......... 22
1 Introduction

A major task in genomic studies is the identification of genes involved in certain diseases or biological processes. As high-throughput microarray technologies are emerging as a powerful tool for genomic studies, the resulting large amounts of data to analyze poses new challenges to statisticians. Instead of testing each gene or subsets of genes separately, multiple hypotheses can be tested at the same time. Now 500 or 5,000 or even 50,000 tests may need to be evaluated simultaneously in genomics and image processing [1]. Many statistical methods such as Empirical Bayes (EB) [2] and Gaussian mixture models (GMM) [3] have been proposed for such a purpose. However, most of these existing approaches treat the genes equally and independently. For example, it is common to assume that all genes to be tested have an identical and independent distribution (iid) in a standard mixture model. On the other hand, genes are known to express coordinately. In recent years, more and more genome-wide gene networks have been constructed based on the rapidly growing biological knowledge. For example, Lee et al. (2004) [4] constructed an extensive, high-quality functional network for the yeast genome using a probabilistic approach. In order to take advantage of existing biological information, Wei and Li [5] proposed a discrete Markov random field (MRF) model to utilize several KEGG transcriptional pathways or other networks in discovering differently expressed genes. Wei and Pan (2008) [6] incorporated yeast gene functional linkage networks into identifying regulatory targets of a transcription factor (TF).

We generally follow Wei and Pan’s method with the same motivating example. In the example, the data were drawn from Lee et al. (2002) [7], who did ChIP-chip experiments for transcription factor GCN4 with three replicates. Our goal is to identify the regulatory target genes of TF GCN4. The rest of this report is organized as follows. We first introduce some biological background of
ChIP-chip data and gene networks. Then we review the standard GMM model and Wei and Pan’s spatially correlated mixture model. Finally we propose a similar model with modifications. We compare the three models by applying them to the real data in the motivating example. Then we end with a summary of our results and a short discussion on future work.
2 Biological background

In this chapter, we briefly introduce the biological background of ChIP-chip experiments and gene networks. The detailed description of the ChIP-chip data used in this report will be in Section 4.1.

2.1 ChIP-chip experiments

ChIP-chip is a technology for isolation and identification of genomic sites occupied by a specific DNA binding protein, for example, a transcription factor, in a living cell. “ChIP” stands for “chromatin immunoprecipitation”, which is a method for isolating DNA fragments that are bound by the DNA binding proteins. The second “chip” refers to the DNA microarray technology. A description of the ChIP-chip experiment process is as follows. We use the pictures from Zheng et al. [8] to illustrate the process.

**Step 1:** Let DNA binding proteins bind to DNA. (Figure 2.1)

**Step 2:** Chop the DNA sequences into small fragments. Therefore, among all the chopped DNA fragments, some are bound by proteins, and the rest are not. (Figure 2.2)

**Step 3:** Isolate the DNA fragments bound by proteins by chromatin immunoprecipitation (ChIP). (Figures 2.3 and 2.4)

**Step 4:** The isolated DNA fragments are released by reversing the cross-linking between DNA and protein, then amplified by LM-PCR and labeled with a fluorescent dye (Cy5). At the same time, a sample of DNA that is not enriched by the above immunoprecipitation process is also amplified by LM-PCR and labeled with another dye (Cy3). (Figures 2.5 and 2.6)
Step 5: Both enriched and unenriched pools of labeled DNA are hybridized to the same arrays (chip). The microarray is then scanned and two images corresponding to Cy5 and Cy3, respectively, are extracted. (Figure 2.7)

Figure 2.1: Proteins bind to DNA.

Figure 2.2: Chop DNA into fragments.

Figure 2.3: Antibody binds to the protein.

Figure 2.4: ChIP is used to enrich the DNA cross-linked with the complex.

Figure 2.5: Cross-linking between DNA and the complex is reversed.

Figure 2.6: ChIP-enriched DNA is amplified and dyed with Cy5 (red). A sample of normal DNA is also amplified and dyed with Cy3 (green).
A typical ChIP-chip data set contains log binding ratios, which measures the relative abundances of the proteins bound to the genes, and possibly inferred test statistics or $p$-values for rejecting the null hypothesis that each of the genes is not bound by the proteins. The log binding ratio can be calculated by $\log(C_{5}/C_{3})$. See Hughes et al. [9] for a possible way to derive $p$-values by using an error model.

## 2.2 Gene networks

A gene network is represented by an undirected or directed graph, where nodes indicate genes and edges represent gene-gene interactions. It can be extracted from existing biological databases, such as KEGG pathways, or any computationally predicted network resulting from integrated analysis of multiple sources of genomic data. In this report, we use the functional linkage network of yeast genes constructed by Lee et al. (2004). By integrating different classes of genomic data, including mRNA expression, experimental protein-protein interaction, gene context, literature mining data, and benchmark sets, Lee et al. assigned a log likelihood score to each possible gene pair via a naive Bayes approach. Two genes with a score high enough are considered linked. They constructed a “ConfidentNet” with the top ranking 34,000 log likelihood scores, which consists of 4681 nodes (genes) and 34,000 edges (gene-gene functional linkages), represented by an undirected graph. See Figure 2.8 for a 3D structure of part of the network. We will use this “ConfidentNet” in our example.
Other types of gene networks can also be used, such as chromosome locations, where neighboring genes on a chromosome are correlated.

Figure 2.8: 3D layout of the 950 strongest interactions in “ConfidentNet”. See 1099511s5.wrl in Lee et al. (2004).
3 Methods

In this chapter, we first state our motivating example, then review the standard Gaussian mixture model and the spatially correlated mixture model in Wei and Pan (2008). Then we propose a modification to the spatial mixture model.

3.1 Problem

Our goal is to identify regulatory target genes of a TF. It can be formulated as a hypothesis testing problem: for each gene $i$, $i = 1, 2, \ldots, G$, we test for a null hypothesis $H_{0,i}$: gene $i$ is not a target of the TF against an alternative $H_{1,i}$: gene $i$ is a target of the TF.

Let $z_i$ be the summarized measurement for each gene $i$. For example, $z_i$ can be a test statistics or the corresponding significance level ($p$-value) for rejecting $H_{0,i}$. In our data, we transform the $p$-values into $z$-scores by:

$$z_i = \Phi^{-1}(1 - P_i),$$

(3.1)

where $\Phi$ is the cumulative distribution function of the standard Normal distribution $N(0, 1)$, and $P_i$ is the $p$-value for gene $i$.

Define the state of gene $i$ by $T_i = I(H_{0,i} \text{ is false})$; that is, $T_i = j$ corresponds to that $H_{j,i}$ holds, where $j = 0, 1$. We denote the density functions of $z_i$ for the genes when $T_i = 0$ and $T_i = 1$ as $f_0$ and $f_1$, respectively.

3.2 Standard Gaussian mixture model

Assuming that all the genes have an identical and independent distribution, we have a marginal distribution of $z_i$ as the standard Gaussian mixture model (denoted as Standard model):

$$f(z_i) = \sum_{j=0}^{1} P_{j,i} f_j(z_i),$$

where $P_{j,i}$ is the probability that gene $i$ is a target of the TF and $f_j(z_i)$ is the density function of $z_i$ when $T_i = j$. This model can be used to identify regulatory target genes of a TF by comparing the $z$-scores with the standard Normal distribution.
\[ f(z_i) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i) , \quad (3.2) \]

where \( \pi_0 \) is the prior probability that \( H_{0,i} \) holds. Since the \( z_i \)'s are transformed from \( p \)-values, then theoretically the null distribution \( f_0 \) is exactly the standard Normal \( N(0, 1) \). In addition, the non-null distribution \( f_1 \) models the right-tail of the \( z \)-score distribution. However, \( p \)-values are often approximated and might deviate from the genuine \( p \)-values, therefore we need to estimate the null and non-null distributions by finite normal mixtures:

\[ \hat{f}_0 = \sum_{k_0=1}^{K_0} \pi_{0,k_0} \phi(\hat{\mu}_{k_0}, \hat{\sigma}_{k_0}^2) \]

\[ \hat{f}_1 = \sum_{k_1=1}^{K_1} \pi_{1,k_1} \phi(\hat{\mu}_{k_1}, \hat{\sigma}_{k_1}^2) , \]

where \( \phi(\mu, \sigma^2) \) is the probability density function of \( N(\mu, \sigma^2) \). In our example, we found that \( K_0 = 2 \) and \( K_1 = 1 \) worked well.

### 3.3 Spatially correlated mixture model

As we pointed out before, the standard model treats all the genes equally and independently, which contradicts the biological fact that genes work coordinately as dictated by gene networks. Wei and Pan (2008) proposed a spatially correlated mixture model by introducing gene-specific prior probabilities \( \pi_{i,j} = P(T_i = j) \) for \( i = 1, 2, \ldots, G \) and \( j = 0, 1 \). Then the marginal distribution of \( z_i \) is:

\[ f(z_i) = \pi_{i,0} f_0(z_i) + \pi_{i,1} f_1(z_i) . \quad (3.3) \]

The gene-specific priors are related to two latent Markov random fields \( x_j = \{x_{i,j} ; i = 1, 2, \ldots, G\} \) via a logistic transformation:

\[ \pi_{i,j} = \exp(x_{i,j}) / [\exp(x_{i,j}) + \exp(x_{i,0})] . \quad (3.4) \]

Here, each \( x_j \) is a \( G \)-dimensional latent vector, distributed according to an intrinsic Gaussian conditional autoregression model (ICAR) [10], which has been readily implemented in the
software WinBUGS [12]. For such a model, the distribution of each spatial latent variable $x_{i,j}$ conditional on $x_{(-i),j} = \{x_{k,j}; \ k \neq i\}$ depends only on the direct neighbors:

$$
    x_{i,j} \mid x_{(-i),j} \sim N\left(\frac{1}{m_i} \sum_{l \in \delta_i} x_{l,j}, \frac{\sigma_{c_j}^2}{m_i}\right),
$$

(3.5)

where $\delta_i$ is the set of indices for the neighbors of gene $i$, and $m_i$ is the corresponding number of neighbors. The parameter $\sigma_{c_j}^2$ controls the degree of dependency: smaller values of $\sigma_{c_j}^2$ will correspond to more similar prior probabilities $\pi_{i,j}$'s for those genes that are neighbors in the network.

### 3.3.1 Modifications

In the spatially correlated mixture model, to allow identifiability, the following constraint is often assumed (also as the default setting in WinBUGS):

$$
    \sum_i x_{i,j} = 0 \quad \text{for } j = 0, 1 \Rightarrow \bar{x}_{*,0} - \bar{x}_{*,1} = 0.
$$

Although adding such a constraint will not change the full conditional distribution (3.5) [11], the estimates of $\pi_{i,j}$'s tend to be large: by (3.4), we have $\logit(\pi_{i,j}) = x_{i,j} - x_{i,0}$. It follows that

$$
    \frac{1}{G} \sum_{j=1}^{G} \logit(\pi_{i,j}) = \bar{x}_{*,1} - \bar{x}_{*,0} = 0,
$$

(3.6)

that is, the $\logit(\pi_{i,j})$'s have mean 0. Thus, with this zero constraint, the estimates of $\logit(\pi_{i,j})$'s will be shrunken towards 0, or roughly, the estimates of $\pi_{i,j}$'s (prior probabilities of being a target) are shrunken towards 0.5. In fact, only a small proportion of genes could be targets of a TF (fewer than a half). In order to avoid more false positives, Wei and Pan (2009) proposed shrinking the $\logit(\pi_{i,j})$'s towards a negative constant $c$ by imposing another constraint:
\[
\frac{1}{G} \sum_{i=1}^{G} \logit(\pi_{i,j}) = \bar{x}_{i,j} - \bar{x}_{i,0} = c, \tag{3.7}
\]

which is realized by imposing \(\bar{x}_{i,j} = c\) and \(\bar{x}_{i,0} = 0\). In their report, they found that \(c = \logit(\hat{\pi}_i)\) worked well, where \(\hat{\pi}_i\) is the estimate from the standard mixture model. We will follow their approach for the constraint.

In Wei and Pan’s papers (2008, 2009), “singletons” with no neighbors, or those genes that are not in the “ConfidentNet”, are not allowed in the spatially correlated models. By merging the ChIP-chip data and the gene network, they ended up with 4,616 genes out of the total 6,270 genes in the ChIP-chip data, that is, 1,654 genes were ignored in fitting the models and will not be tested for hypotheses. We propose a modification of adding the “singletons” and replacing their prior probabilities by the estimates from the standard model. In particular, we fit a standard model with all 6,270 genes to obtain the estimates of \(\pi_i\)'s, and then instead of using (3.4), we have \(\pi_{i,j} = \hat{\pi}_j\). By this modification, we are able to make use of the entire information from the ChIP-chip experiments, but also able to identify potential target genes among the singletons.

In our example, we fit both the standard model and the spatially correlated mixture model (denoted as **Spatial model** or **Spatial model with zero constraint**) with merged data (without singletons), and apply the modified spatial mixture model (denoted as **Spatial model with logit constraint**) to the whole data (6,270 genes with singletons).

### 3.4 Inference

Although the standard model can be fitted via maximum likelihood with the EM algorithm, it is unclear how to fit the spatial models with a similar approach. For better comparison, we fit all the three models in a Bayesian hierarchical modeling framework via Markov Chain Monte Carlo (MCMC), which can be implemented in WinBUGS. MCMC is used to draw posterior samples.
for model parameters. The posterior means after a period of burn-ins are computed as the parameter estimates. See Appendix A.2 for details on the MCMC algorithm.

Statistical inference is based on the posterior probability that $H_{1,i}$ holds:

$$ P(T_i = 1 \mid z_i) = \frac{\pi_i f_i(z_i)}{f(z_i)}. $$

(3.8)

We determine whether to reject $H_{0,i}$ based on whether the estimate $\hat{P}(T_i = 1 \mid z_i)$ is smaller than a threshold $b$. By varying $b$, we are able to plot a Receiver Operating Characteristic (ROC) curve to evaluate the statistical power of each model.
4 Example

We apply the three models introduced in the previous chapter to the same motivating example in Wei and Pan (2008, 2009). We begin with a brief description of the data, and then state our results.

4.1 Data

The ChIP-chip data used in this report was downloaded from Lee et al.’s website (http://web.wi.mit.edu/young/regulator_network/). Log binding ratios and \( p \)-values of the yeast \textit{Saccharomyces cerevisiae} genes with 106 different TF’s are available. Log binding ratios were calculated as average values based on the results from three independent replicated ChIP-chip experiments. \( p \)-values were derived using an error model proposed by Hughes et al. (2000). See Table 4.1 for the structure of the ChIP-chip data. We use the \( p \)-values for 6,270 genes with TF GCN4, and transformed them into \( z \)-scores by (3.1). GCN4 (General Control Nondepressible 4) is known as an important transcriptional activator in response to amino acid starvation in yeast. Pokholok et al. [13] identified a set of 84 genes most likely to be targeted by GCN4 (positive controls) and a set of 945 genes least likely to be targets of GCN4 (negative controls). By treating the positive and negative control sets as true positives and true negatives, we are able to evaluate the performance of three models.

After merging the ChIP-chip data, positive and negative control sets, and the “ConfidentNet” discussed in Section 2.2, we end up with a 4,665-node, 33,941-edge network, 64 positive controls, 781 negative controls, and 1,605 “singletons” that are in the ChIP-chip data but not in the “ConfidentNet”. We apply the standard model and the spatial model with zero constraint to the
4,665 $z$-scores, and then fit the modified spatial model with all 6,270 $z$-scores. To facilitate comparison, we use the merged control sets to construct ROC curves for the three models.

<table>
<thead>
<tr>
<th>Index</th>
<th>Gene name</th>
<th>Binding ratio</th>
<th>$p$-value</th>
<th>$z$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YAL064C-A</td>
<td>0.29</td>
<td>0.96</td>
<td>-1.75</td>
</tr>
<tr>
<td>2</td>
<td>YAL008W</td>
<td>0.73</td>
<td>0.92</td>
<td>-1.41</td>
</tr>
<tr>
<td>3</td>
<td>YBR005W</td>
<td>0.92</td>
<td>0.69</td>
<td>-0.50</td>
</tr>
<tr>
<td>4</td>
<td>YAL001C</td>
<td>1.05</td>
<td>0.43</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>YBR258C</td>
<td>1.35</td>
<td>0.06</td>
<td>1.13</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6270</td>
<td>YAR050W</td>
<td>2.03</td>
<td>0.00041</td>
<td>3.35</td>
</tr>
</tbody>
</table>

### 4.2 Parameter estimates

The software WinBUGS is used to implement the three models. Before MCMC sampling, all unknown quantities need to be given prior distributions. Following Wei and Pan, we use vague or non-informative priors. See Appendix A.1 for full prior settings.

Convergence is checked both by visual inspection of the history plot and autocorrelation plot, and by running multiple chains from dispersed initial values. The standard model converges the fastest among the three models. Parameters of interest are estimated as the posterior mean of 6,000 MCMC samples after discarding 4,000 burn-ins for the standard model, and 9,000 burn-ins for the two spatial models.

Wei and Pan (2008) pointed out that lack-of-fit of a two-component standard mixture model was observed when applying to the same ChIP-chip data. They fixed it by adding a third Normal component. Then the model becomes a three-component standard Gaussian mixture model, in
which the first two components with zero and negative means are treated as \( \hat{f}_0 \) and the third component with positive mean as \( \hat{f}_1 \) (since departures from the null distribution are indicated by large positive \( z \)-scores). We fitted both two-component and three-component standard models. By comparing the \( \chi^2 \) test statistics (results not shown) and the fitted distributions against the data histograms (Figure 4.1a), we also came to the same conclusion that adding the third component gives a better fit. The fitted three-component standard model was:

\[
\hat{f}(z_i) = .90\phi(z_i; 0, .80^2) + .046\phi(z_i; 2.10, .49^2) + .056\phi(z_i; 1.71, 1.94^2),
\]

which fits reasonably well except at the peak of the data histogram and the “bump” around -2 (Figure 4.1b). Similarly, three-component spatial models with both zero and logit constraints were fitted:

\[
\hat{f}(z_i) = \hat{\pi}_{i,0,1}\phi(z_i; 0, \hat{\sigma}_{0,1}^2) + \hat{\pi}_{i,0,2}\phi(z_i; \hat{\mu}_{0,2}, \hat{\sigma}_{0,2}^2) + \hat{\pi}_{i,1}\phi(z_i; \hat{\mu}_1, \hat{\sigma}_1^2),
\]

where the parameter estimates are shown in Table 4.2. Table 4.3 shows the estimates of parameters in the ICAR model (3.5).

By visual examination of the fitted marginal and component-wise distributions against data histograms, we found that both the spatial models fit well (Figure 4.1c, 4.1d). History plots for the three models are shown in Figure 4.3 to check convergence of the MCMC simulations.

Table 4.2: Parameter estimates for spatial models.
Note that estimates for the gene-specific prior probabilities are average values.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \pi_{0,1} )</th>
<th>( \pi_{0,2} )</th>
<th>( \pi_1 )</th>
<th>( \mu_{0,1} )</th>
<th>( \mu_{0,2} )</th>
<th>( \mu_1 )</th>
<th>( \sigma_{0,1} )</th>
<th>( \sigma_{0,2} )</th>
<th>( \sigma_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>.90</td>
<td>.046</td>
<td>.056</td>
<td>0</td>
<td>-2.10</td>
<td>1.71</td>
<td>.80</td>
<td>.49</td>
<td>1.94</td>
</tr>
<tr>
<td>Spatial with zero constraint</td>
<td>.57</td>
<td>.33</td>
<td>.10</td>
<td>0</td>
<td>-.37</td>
<td>1.03</td>
<td>.68</td>
<td>1.14</td>
<td>1.75</td>
</tr>
<tr>
<td>Spatial with logit constraint</td>
<td>.46</td>
<td>.50</td>
<td>.038</td>
<td>0</td>
<td>-.14</td>
<td>1.79</td>
<td>.61</td>
<td>1.16</td>
<td>2.17</td>
</tr>
</tbody>
</table>
Figure 4.1: Fitted plots against data histograms: a. two-component standard model; b. (three-component) standard model; c. spatial model with zero constraints; d. spatial model with logit constraint.
On each panel, the dashed line is for the marginal distribution, while the three solid and colored lines are for the three components.
Table 4.3: Parameter estimates for modeling Markov random fields by ICAR.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_{C0,1}$</th>
<th>$\sigma_{C0,2}$</th>
<th>$\sigma_{C1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial with zero constraint</td>
<td>85.79</td>
<td>1.61</td>
<td>11.71</td>
</tr>
<tr>
<td>Spatial with logit constraint</td>
<td>43.22</td>
<td>.16</td>
<td>.27</td>
</tr>
</tbody>
</table>

Figure 4.3: Convergence check.
4.3 Statistical power

By treating the 64 positive and 781 negative control genes as true positives (targets of GCN4) and true negatives (non-targets), we constructed the ROC curves for the three models to evaluate their performance.

As shown in Figure 4.4, by borrowing information from the gene networks, both the two spatial models gave a higher statistical power than did the standard model. In particular, when the false positive rate ranged from 0.1–0.8 (i.e. with a specificity of 0.2–0.9, which is usually desired in
practice), the two spatial models had much higher true positive rate (i.e. sensitivity) in detecting the targets.

Figure 4.4: ROC curves for the three models applied to the ChIP-chip data.

For a better comparison, we also calculated the areas under the ROC curve (AUC). Results are .8364, .8780, and .8929 for standard model, spatial model with zero constraint, and spatial model with logit constraint, respectively. Again, both spatial models had larger AUCs, resulting in better performances. The modified spatial model gave an even larger AUC, because of its slightly higher sensitivity at low specificity (< 0.5).
5 Discussion

In this report, we have illustrated that by borrowing gene network information, spatially correlated mixture models have better performance in genomic discovery. In addition, the modified spatial model gained extra statistical power as well as the ability to identify more potential true positives. Even if the network data does not contain all the genes, whole-genome hypotheses test still can be conducted without ignoring any information. Wei [14] pointed out another modification of the spatial models by integrating multiple gene networks. They applied the model to another ChIP-chip data set and two different networks, and found that ranks of known target genes were boosted by the new model. Wei and Pan (2008) conducted a simulation study to evaluate the impact of mis-specified gene networks. They perturbed the network and apply the standard model and spatial model with zero constraint to a simulated ChIP-chip data set. Their results showed that the spatial model still gained more power than did the standard model regardless of the perturbations, that is, the spatial model were reasonably robust to network mis-specifications.

In our example, we were able to use positive and negative control sets in order to construct ROC curves of the three models. However, in most real problems, control gene sets like the ones we used are not available. Other evaluation criteria are needed if we apply our methods to other data sets.

Another consideration is the computation speed as well as the convergence of MCMC sampling. The standard mixture model showed quick and good convergence, while the two spatial models took much longer time to reach an acceptable convergence. In particular, the modified spatial model with logit constraint required a better specification of initial values and 30 minutes to hours to converge.
A Appendix

A.1 Prior specifications

We followed Wei and Pan (2008, 2009) for prior specifications. They used truncated Normal as prior distributions of the Normal means to ensure unique labeling. See Wei and Pan (2008) for a discussion on how different hyperparameters in the Inverse Gamma distribution might influence the results.

A.1.1 For a three-component standard mixture model

\[ \mu_{0} = 0; \mu_{0,2} \sim N(0, 10^{6})I(a, 0), a = \min z_{i}; \mu_{1} \sim N(0, 10^{6})I(0, b), b = \max z_{i}; \]

\[ \sigma_{0,1}^{2}, \sigma_{0,2}^{2}, \sigma_{1}^{2} \sim \text{Inverse Gamma}(0.1, 0.1); \]

\[ (\pi_{01}, \pi_{02}, \pi_{1}) \sim \text{Dirichlet}(1, 1, 1). \]

A.1.2 For three-component spatial mixture models with zero and logit constraints

\[ \mu_{0} = 0; \mu_{0,2} \sim N(0, 10^{6})I(a, 0), a = \min z_{i}; \mu_{1} \sim N(0, 10^{6})I(0, b), b = \max z_{i}; \]

\[ \sigma_{0,1}^{2}, \sigma_{0,2}^{2}, \sigma_{1}^{2} \sim \text{Inverse Gamma}(0.1, 0.1); \]

\[ \sigma_{c_{01}}^{2}, \sigma_{c_{02}}^{2}, \sigma_{c_{1}}^{2} \sim \text{Inverse Gamma}(0.01, 0.01). \]

A.2 MCMC Algorithm

Denote the full conditional distribution of \( \alpha \), the distribution of \( \alpha \) conditional on everything else in the model, by \( (\alpha \mid \cdots) \). For the spatial mixture models, the joint posterior distribution is

\[ p(T, \theta_{1}, x_{0}, x_{1}, x_{2}, \tau_{c_{0}}, \tau_{c_{1}}, \tau_{c_{2}} \mid z) \propto p(z \mid T, \theta_{1}) p(T \mid x_{0}, x_{1}, x_{2}) p(\theta_{1}) \prod_{j=0}^{2} p(x_{j} \mid \tau_{c_{j}}) p(\tau_{c_{j}}). \]
Note that we have changed the subscripts from 0, 1, 2 and 1 to 0, 1 and 2 for convenience. The MCMC samplers are generated by the full conditional distributions using Gibbs sampling.

For $\mu_j (j = 1, 2)$,

$$ (\mu_j | \cdots) \sim N \left( \frac{\tau_j \sum_{(i:T_i = j)} z_i}{10^{-6} + n_j \tau_j}, \frac{1}{10^{-6} + n_j \tau_j} \right) \left( I_{(0,b)}(\mu_i)I(j = 1) + I_{(0,a)}(\mu_2)I(j = 2) \right), $$

where

$$ n_j = |i : T_i = j|. $$

For $\tau_j (j = 0, 1, 2)$, $(\tau_j | \cdots) \sim Gamma \left( \tau_j \mid \frac{n_j}{2} + 0.1, \frac{1}{2} \sum_{(i:T_i = j)} (z_i - \mu_j)^2 + 0.1 \right)$.

For $T_i$, $(T_i | \cdots) \sim Multinomial \left( 1; p_{i0}, p_{i1}, p_{i2} \right)$, where

$$ p_j = \frac{\exp(x_{ij}) \phi(z_i; \mu_j, \sigma_j)}{\sum_{k=0}^{2} \exp(x_{ik}) \phi(z_i; \mu_k, \sigma_k^2)}. $$

For $x_{ij}$, $(x_{ij} | \cdots) \sim \exp \left\{ \frac{\sum_{k=0}^{2} x_{ik} I(T_i = k)}{\sum_{k=0}^{2} \exp(x_{ik})} \exp \left\{ -\frac{m_i \tau_{cj}}{2} (x_j - \frac{1}{m_i} \sum x_{ij})^2 \right\} \right\}$, which is log-concave and therefore can be updated by slice sampling implemented in WinBUGS.

For $\tau_{cj} (j = 0, 1, 2)$, $(\tau_{cj} | \cdots) \sim Gamma \left( \tau_{cj} \mid \frac{G-1}{2} + 0.01, \frac{x_j Q x_j}{2} + 0.01 \right)$.

### A.3 WinBUGS codes for implementing the three methods

#### A.3.1 For a three-component standard mixture model

```plaintext
model
{
```
for( i in 1:N ) {

Z[i] ~ dnorm(muR[i], tauR[i]) # z-scores
muR[i] <- mu[T[i]]
tauR[i] <- tau[T[i]]
T[i] ~ dcat(pi[ ]) # latent variable (zero/negative/postive components)
T1[i] <- equals(T[i],1) ; T2[i] <- equals(T[i],2); T3[i]<-equals(T[i],3);
}

# prior for mixing proportions
pi[1:3] ~ ddirch(alpha[ ])

# priors (means of normal mixture components)
mu[1] <- 0 # empirical null (zero component)
mu[2] ~ dnorm(0, 1.0E-6)I(a, 0.0) # empirical null (negative component)
mu[3] ~ dnorm(0, 1.0E-6)I(0.0, b) # non-null (positive) component

# priors (precision/variance of normal mixture components)
tau[1]~dgamma(0.1, 0.1)
tau[2]~dgamma(0.1, 0.1)
tau[3]~dgamma(0.1, 0.1)
}

A.3.2 For a three-component spatial mixture model with zero constraint

model
{

}
for( i in 1:N ) {

Z[i] ~dnorm(muR[i], tauR[i]) # z-scores
muR[i] <- mu[T[i]]
tauR[i] <- tau[T[i]]

# logistic transformation
pi[i,1] <--1/(1+exp(x2[i]-x1[i])+exp(x3[i]-x1[i]))
pi[i,2] <--1/(1+exp(x1[i]-x2[i])+exp(x3[i]-x2[i]))
pi[i,3] <--1/(1+exp(x1[i]-x3[i])+exp(x2[i]-x3[i]))

T[i] ~dcat(pi[i,1:3]) # latent variable (zero/negative/positive components)
T1[i] <-equals(T[i],1) ;T2[i] <-equals(T[i],2); T3[i] <-equals(T[i],3)
}

# Random Fields specification
x1[1:N] ~car.normal(adj[], weights[], num[], tauC[1])
x2[1:N] ~car.normal(adj[], weights[], num[], tauC[2])
x3[1:N] ~car.normal(adj[], weights[], num[], tauC[3])

# weights specification
for(k in 1:sumNumNeigh) { weights[k] <- 1 }

# priors (precision/variance for MRF)
tauC[1] ~dgamma(0.01, 0.01)I(0.0001,)
tauC[2] ~dgamma(0.01, 0.01)I(0.0001,)
tauC[3] ~dgamma(0.01, 0.01)I(0.0001,)
sigma2C[1] <- 1/tauC[1]
# priors (means of normal mixture components)
mu[1] <- 0 # empirical null (zero component)
mu[2] ~dnorm(0, 1.0E-6)I(a,0.0) # empirical null (negative component)
mu[3] ~dnorm(0, 1.0E-6)I(0.0,b) # non-null (positive) component

# priors (precision/variance of normal mixture components)
tau[1]~dgamma(0.1, 0.1)
tau[2]~dgamma(0.1, 0.1)
tau[3]~dgamma(0.1, 0.1)

A.3.3 For a three-component spatial mixture model with logit constraint and singletons

model
{
for( i in 1:N ) {
Z[i] ~dnorm(muR[i], tauR[i]) # z-scores
muR[i] <- mu[T[i]]
tauR[i] <- tau[T[i]]

# logistic transformation
pi[i,1] <-(1 - lambda[i]) *(1/(1+exp(x2[i]-x1[i])+exp(x3[i]+c-x1[i])))#lambda[i] * 0.8835
pi[i,2] <-(1 - lambda[i]) *(1/(1+exp(x1[i]-x2[i])+exp(x3[i]+c-x2[i])))#lambda[i] * 0.0489

}
pi[i,3] <- (1 - lambda[i]) *(1/(1+exp(x1[i]-x3[i]-c)+exp(x2[i]-x3[i]-c)))+lambda[i] * 0.0676

T[i] ~ dcat(pi[i,1:3]) # latent variable (zero/negative/positive components)

T1[i] <- equals(T[i],1) ; T2[i] <- equals(T[i],2); T3[i] <- equals(T[i],3)

} # Random Fields specification

x1[1:N] ~ car.normal(adj[], weights[], num[], tauC[1])
x2[1:N] ~ car.normal(adj[], weights[], num[], tauC[2])
x3[1:N] ~ car.normal(adj[], weights[], num[], tauC[3])

# weights specification

for(k in 1:sumNumNeigh) { weights[k] <- 1 }

# prior (precision/variance for MRF)

tauC[1] ~ dgamma(0.01, 0.01)I(0.0001,)
tauC[2] ~ dgamma(0.01, 0.01)I(0.0001,)
tauC[3] ~ dgamma(0.01, 0.01)I(0.0001,)

sigma2C[1] <- 1/tauC[1]

# priors (means of normal mixture components)

mu[1] <- 0 # empirical null (zero component)

mu[2] ~ dnorm(0, 1.0E-6)I(a,0.0) # empirical null (negative component)

mu[3] ~ dnorm(0, 1.0E-6)I(0.0,b) # non-null (positive) component

# priors (precision/variance of normal mixture components)

tau[1]~ dgamma(0.1, 0.1)
tau[2] ~ dgamma(0.1, 0.1)

tau[3] ~ dgamma(0.1, 0.1)


# logit constraint

c <- logit(0.0676)

}
References


