Multilevel models with applications in genomics

Brian Caffo
Johns Hopkins Bloomberg School of Public Health
Acknowledgments

- Giovanni Parmigiani
- Dongmei Liu

- Screening for Differentially Expressed Genes: Are Multilevel Models Helpful?

- Power Conjugate Multilevel Models with Applications in Genomics

- http://www.biostat.jhsph.edu/~bcaffo
Multilevel models for the analysis of gene expression microarrays

- Gene expression arrays simultaneously consider the “expression” of thousands of genes across experimental conditions.
- Expression levels share many sources of variation, both technical and biological.
- Multilevel models offer a way to model the gene to gene variability.
- They also offer a formal mechanism for shrinking gene-specific estimates.
- Early references: Newton et al. (1999, 2001), Baldi and Long (2001)
Lung cancer types

- Three studies: one using CDNA microarrays and two using Affymetrix
- We focus on studying differential expression between two cancerous tissue types: adeno and squamous cell carcinomas
- 307 genes selected as the most variable out of several thousand (see Parmigiani et al. Clinical Cancer Research 2004)
- We use two studies, Harvard and Michigan, to validate results obtained on the Stanford data set

References

  Bhattacharjee et al. 2001
  Beer et al. 2002
  Garber et al. 2001
A spiked in data set

- Experiment where expression status is known
- The same biological sample was analyzed using six Affymetrix arrays
- Several thousand genes from the null model
- Eleven genes are known to be differentially expressed between two groups of three
**Structure of the data**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Group 1</th>
<th>Group 2</th>
<th>row</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_{11}$ $y_{12}$ $\cdots$ $y_{1n_1}$ $y_{1(n_1+1)}$ $y_{1(n_1+2)}$ $\cdots$ $y_{1(n_1+n_2)}$</td>
<td>$y_1$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$y_{21}$ $y_{22}$ $\cdots$ $y_{2n_1}$ $y_{2(n_1+1)}$ $y_{2(n_1+2)}$ $\cdots$ $y_{2(n_1+n_2)}$</td>
<td>$y_2$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>G</td>
<td>$y_{G1}$ $y_{G2}$ $\cdots$ $y_{Gn_1}$ $y_{G(n_1+1)}$ $y_{G(n_1+2)}$ $\cdots$ $y_{G(n_1+n_2)}$</td>
<td>$y_G$</td>
<td></td>
</tr>
</tbody>
</table>

- Usually $n_1$ and $n_2$ are very small relative to $G$
- The true difference in the group means is the **gene-specific signal**
- The true mean expression level per gene is the **gene-specific abundance**
- The true true standard deviation per gene is the **gene-specific noise**
- Assume the gene specific noise is constant across the two groups
General approach

- Pre-process the data (see Irizarry et al. 2003)
- Choose a ranking statistic representative of the study goals
- Rank the genes according to this statistic
- Choose a threshold and select all genes with statistics above the threshold

Note:
Choice of a threshold can depend on a variety of factors such as: available funds for subsequent validation of the selected genes, having a minimally acceptable false discovery rate ...
Some analytic problems

- To shrink or not to shrink signals, noises and abundances
- How to perform the shrinkage
- If multilevel models are used, is it appropriate to assume a continuous distribution on the signals?
- Dependence between the signals, noises and abundances
- Fitting algorithms must be able to handle the large volume of data
- Choice of statistic, detecting reliably measured changes in expression or large changes in absolute expression
Top-level model

We assume that

\[ y_g \sim \text{Normal}(X\beta_g, I\lambda_g) \]

where

\[
X = \begin{pmatrix}
1 & .5 \\
\vdots & \vdots \\
1 & .5 \\
1 & - .5 \\
\vdots & \vdots \\
1 & - .5
\end{pmatrix}
\]

and

\[
\beta_g = \begin{pmatrix}
\beta_{0g} \\
\beta_{1g}
\end{pmatrix}
\]

\(X\) is constant across genes
\(\beta_{0g}\) is the gene-specific abundance
\(\beta_{1g}\) is the gene-specific signal
\(\lambda_g\) is the gene-specific noise
Power-conjugate model

For the second level model we assume that
\[
\begin{pmatrix}
\beta_{0g} \\
\beta_{1g}
\end{pmatrix}
\sim \text{Normal}
\begin{bmatrix}
\begin{pmatrix}
\mu_0 \\
\mu_1
\end{pmatrix},
\begin{pmatrix}
F_0 & F_{01} \\
F_{01} & F_1
\end{pmatrix}
\end{bmatrix}
\lambda_\delta
\]

and
\[
\lambda_g \sim \text{Inverted Gamma}(\nu, \tau)
\]

For $\delta = 1$ this is the usual conjugate model
For $\delta = 0$ independence between ($\beta_{0g}, \beta_{1g}$) and $\lambda_g$
Some notes

- When fitting a random intercept and group difference, it’s better to code the group difference as -.5 and .5 rather than 0 and 1.
- When fitting a random intercept and slope, it’s good practice to allow them to be correlated.
- When the slopes and intercepts are strongly correlated, the intercepts give information about ranking that is ignored when treating the two as independent.
- The parameter $\delta$ controls both the correlation between the signals and abundances, but also the tail decay of the marginal distribution on the $\beta_g$.
- Assuming the same delta for both the signals and abundances may be problematic.
Delta = -1

Delta = 0

Delta = 1

Delta = 2
Notes on fitting

- We use an EM algorithm to get ML estimates
  - EM is convenient but slow
  - Separate M-steps for $(\mu, F)$ and $(\nu, \tau)$
  - Closed form M-step for $(\mu, \tau)$
  - Maximizing for $(\nu, \tau)$ equivalent to maximizing a gamma likelihood (see Johnson and Kotz *Continuous Univariate Distributions*)
  - Can use some tricks to speed up computations
  - Can use method of moments approximations for starting values
  - Use a grid search to obtain $\delta$
More notes on fitting

- For $\delta \neq 1$ the relevant integrals are not tractable
  - Can reduce integrals to univariate integrals wrt a gamma kernel
  - Use Gauss-Laguerre integration as an approximation
  - Multiply and divide by the conjugate case helps the quadrature approximation
- Notice that $Z_g = HY_g \sim \text{Normal}(0, HH^t\lambda_g)$ when $H$ is comprised of vectors from the null rowspace of $X$
  - Recall that $\lambda_g \sim \text{Inverted Gamma}(\nu, \tau)$
  - Obtain estimates of $(\nu, \tau)$ using the $z_g$ eliminates the need to maximize these parameters in the M-step
Recap

- Using multilevel models to shrink estimates of the gene-specific signals, abundances and noises
- Applying a model that estimates the conjugacy relationship between (signals, abundances) and the noises
- Fit a term that models the correlation between signals and abundances
- Focus on estimating reliably measured changes in expression (signal to noise ratios)
A simulation study

- Parameters were all combinations of
  - $\sqrt{F_{11}/F_{22}} \in \{.5, 1, 2\}$
  - $F_{12}/F_{11}^{1/2} F_{22}^{1/2} \in \{0, .5\}$
  - $\delta \in \{0, .5, 1, 1.5\}$
  - CV of IG prior $\in \{.5, 1, 2\}$
  - Mean of Gamma prior $\in \{0, .5, 1, 1.5\}$.

- 1,000 genes
- Two groups of size 4
- 500 simulations
- Compared the areas under the ROC curve for each simulated data set
- Four methods: power conjugate, complete conjugacy, standard-t and SAM
ROC Curves

As the threshold is varied:

- Best: Declare everything negative
- Better: When true positive rate is high and false positive rate is low
- Worse: When the model performs worse than random guessing
- Worst: Declare everything positive
## Lung cancer datasets

Correlations within and between studies for various statistics

<table>
<thead>
<tr>
<th></th>
<th>Harvard</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC</td>
<td>T</td>
<td>SAM</td>
<td>CCI</td>
<td>III</td>
<td>PC</td>
<td>T</td>
<td>SAM</td>
<td>CCI</td>
</tr>
<tr>
<td>PC</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>0.98</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>T</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.97</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>SAM</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.95</td>
<td>0.85</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>CCI</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.97</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>III</td>
<td>0.98</td>
<td>0.97</td>
<td>0.95</td>
<td>0.97</td>
<td>0.83</td>
<td>0.82</td>
<td>0.81</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>D</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.86</td>
<td>0.83</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>T</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.82</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>SAM</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
<td>0.87</td>
<td>0.81</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>CCI</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.82</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>III</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.82</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>
The lung cancer datasets

Mean of the correlations of various statistics with the gold standard statistic

<table>
<thead>
<tr>
<th>Gold Standard</th>
<th>Power</th>
<th>$\delta = 0$</th>
<th>$\delta = 1$</th>
<th>$\delta = 0$</th>
<th>$\delta = 1$</th>
<th>T</th>
<th>SAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>0.729</td>
<td>0.727</td>
<td>0.729</td>
<td>0.698</td>
<td>0.710</td>
<td>0.693</td>
<td>0.722</td>
</tr>
<tr>
<td>II</td>
<td>0.683</td>
<td>0.683</td>
<td>0.685</td>
<td>0.645</td>
<td>0.660</td>
<td>0.639</td>
<td>0.676</td>
</tr>
<tr>
<td>SAM</td>
<td>0.715</td>
<td>0.717</td>
<td>0.720</td>
<td>0.684</td>
<td>0.701</td>
<td>0.681</td>
<td>0.718</td>
</tr>
<tr>
<td>T</td>
<td>0.728</td>
<td>0.727</td>
<td>0.729</td>
<td>0.698</td>
<td>0.711</td>
<td>0.692</td>
<td>0.724</td>
</tr>
</tbody>
</table>
The spiked-in data set

- Several thousand genes from the null distribution
- Eleven *known* genes that are differentially expressed
- Fitting a continuous random effect for the signal produces *disastrous* results
- Random effect variance for the signals is estimated to be nearly zero
<table>
<thead>
<tr>
<th>Gene name</th>
<th>Spike-in concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 1</td>
</tr>
<tr>
<td>AFFX-BioB-5-at</td>
<td>100</td>
</tr>
<tr>
<td>AFFX-BioB-3-at</td>
<td>.5</td>
</tr>
<tr>
<td>AFFX-BioB-M-at</td>
<td>1</td>
</tr>
<tr>
<td>AFFX-BioC-5-at</td>
<td>2</td>
</tr>
<tr>
<td>AFFX-CreX-5-at</td>
<td>12.5</td>
</tr>
<tr>
<td>AFFX-BioDn-3-at</td>
<td>1.5</td>
</tr>
<tr>
<td>AFFX-DapX-5-at</td>
<td>5</td>
</tr>
<tr>
<td>AFFX-DapX-3-at</td>
<td>37.5</td>
</tr>
<tr>
<td>AFFX-DapX-M-at</td>
<td>3</td>
</tr>
<tr>
<td>AFFX-BioC-3-at</td>
<td>25</td>
</tr>
<tr>
<td>AFFX-CreX-3-at</td>
<td>50</td>
</tr>
<tr>
<td>AFFX-BioDn-5-st</td>
<td>0</td>
</tr>
<tr>
<td>AFFX-BioB-5-st</td>
<td>0</td>
</tr>
<tr>
<td>AFFX-BioB-N-st</td>
<td>0</td>
</tr>
<tr>
<td>AFFX-BioB-3-st</td>
<td>0</td>
</tr>
<tr>
<td>AFFX-YEL024w/RIP1-at</td>
<td>0</td>
</tr>
<tr>
<td>AFFX-CreX-3-st</td>
<td>0</td>
</tr>
<tr>
<td>32636-f-at</td>
<td>0</td>
</tr>
<tr>
<td>AFFX-BioDn-5-st</td>
<td>0</td>
</tr>
<tr>
<td>AFFX-CreX-5-st</td>
<td>0</td>
</tr>
<tr>
<td>1764-s-at</td>
<td>0</td>
</tr>
</tbody>
</table>
Choose statistic to match experimental goals

Multilevel modeling in high throughput experiments requires careful validation of the assumptions of the model

- Are continuous random effects for signals appropriate?
- Correlation between signals and abundances
- Conjugacy may be suspect
Summary

- MLMs enjoy the benefits of a completely specified likelihood
- MLMs do not always compare favorably with non-parametric/robust statistic approaches
- Richer classes of models are needed and identified given the volume of data
- Volume of data can often reject the standard conjugacy assumptions of multilevel models
- Relieve some of the burden placed on data pre-processing
Future research

- Non-parametrically estimate the distribution of $(\beta_0^g, \beta_1^g, \lambda_g)$ (see Newton et al. 2004)
- Incorporating functional class information (Dongmei Liu)
- Incorporating gene-gene correlations (Prulak Ghosh)
- Applying similar multilevel models in fMRI research