Discover Common and Differential Enrichment: A Multivariate Bayesian Variable Selection Approach

Sierra M. Li and Giovanni Parmigiani

Division of Oncology Biostatistics
The Sidney Kimmel Cancer Center at Johns Hopkins

November 7, 2007

Model and notation

- Model
  - Hierarchical model with cross-profile shrinkage
  - Enrichment types: latent variable γ
    - no enrichment ($\gamma = 0$)
    - common enrichment: similar magnitude across profiles ($\gamma = 1$)
    - differential enrichment: different magnitude across profiles ($\gamma = 2$)
- Notation
  - Gene expression: $\boldsymbol{Y}_{n\times q}$ response matrix
  - gene-set: $\boldsymbol{X}_{n\times p}$ design matrix
  - Enrichment measure: $\boldsymbol{B}_{p\times q} = (\beta_{ij})$ regression coefficients
    - β_{ij}: enrichment for gene-set i in profile j
    - β_{ij} measures the change of average expression in a gene-set
  - $\boldsymbol{\mu}$: p-vector for average enrichment across profiles

Conjugate model and parameters

- Gene level:

$$\boldsymbol{Y} - \mathbf{1}\boldsymbol{\alpha}^{'} - \boldsymbol{X}\boldsymbol{B} = \boldsymbol{\Omega} \sim \mathcal{N}(\boldsymbol{I}_n, \sigma^2 \boldsymbol{I}_p)$$

- Gene-set level:

$$\boldsymbol{\alpha}^{'} - \boldsymbol{\alpha}_0^{'} \sim \mathcal{N}(h, \sigma^2 \boldsymbol{I}_p)$$

$$\boldsymbol{B} - \boldsymbol{\mu}_{p\times 1}\mathbf{1}_{q\times 1}^{'} \sim \mathcal{N}(\boldsymbol{H}_\gamma, \sigma^2 \boldsymbol{I}_p)$$

- Gene-set across profiles:

$$\boldsymbol{\mu}^{'} - \mathbf{0}^{'} \sim \mathcal{N}(\sigma^2, \boldsymbol{G}_\gamma)$$

- Matrices related to variable selection:

$\boldsymbol{H}_\gamma = \boldsymbol{D}_\gamma \boldsymbol{R} \boldsymbol{D}_\gamma$ and $\boldsymbol{G}_\gamma = \boldsymbol{F}_\gamma \boldsymbol{R} \boldsymbol{F}_\gamma$, $\boldsymbol{R}$: correlation $\boldsymbol{D}, \boldsymbol{F}$: diagonal matrices

$$d_i^2 = \begin{cases} \tau_{i0}^2 & \text{if } \gamma_i = 0 \text{ or } 1 \\ \tau_{i1}^2 & \text{if } \gamma_i = 2 \end{cases} \qquad f_i^2 = \begin{cases} \upsilon_{i0}^2 & \text{if } \gamma_i = 0 \\ \upsilon_{i1}^2 & \text{if } \gamma_i = 1 \text{ or } 2 \end{cases}, \text{ for } i = 1, 2, \cdots, p.$$

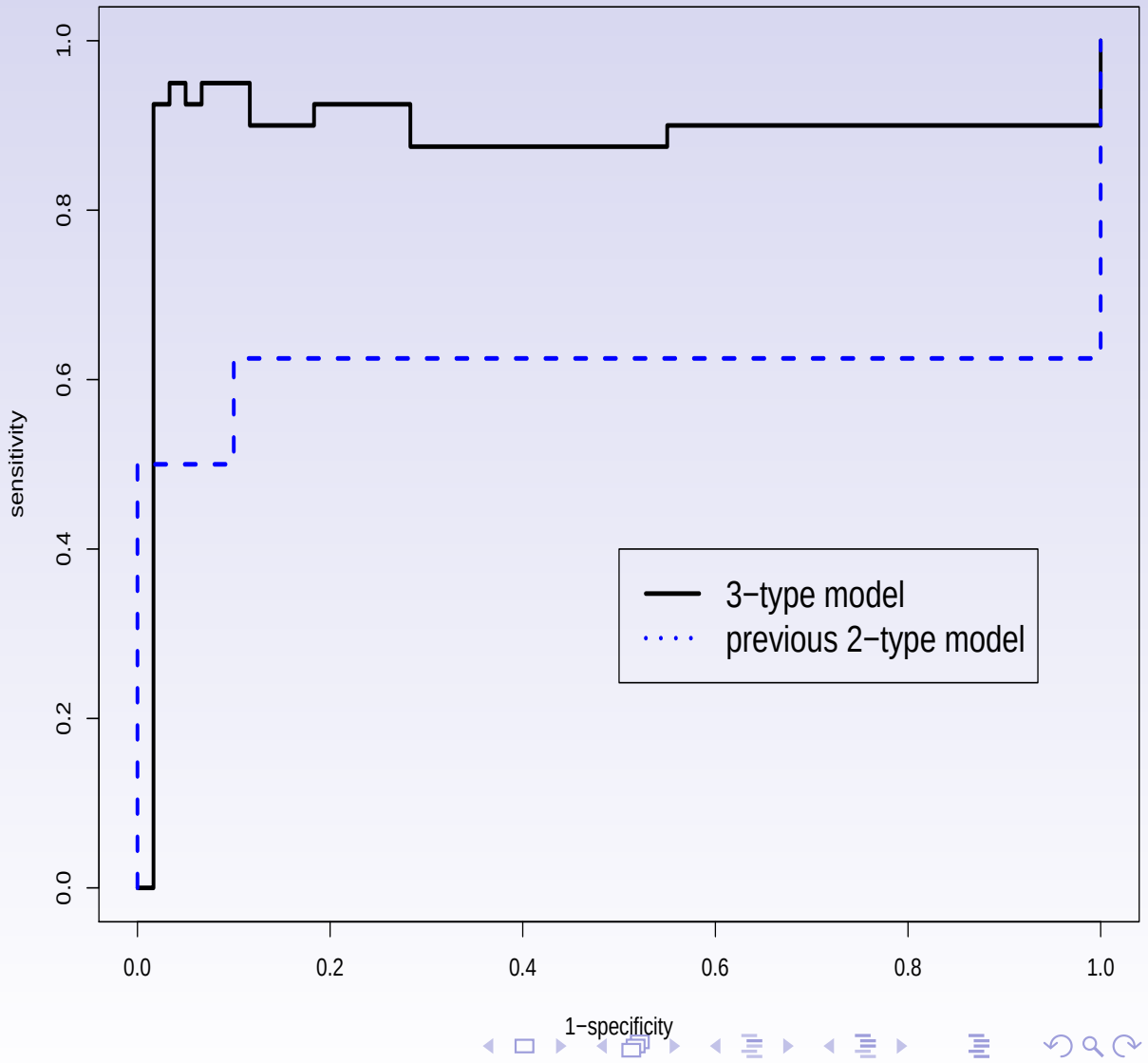
$\tau_{i0}^2 << \tau_{i1}^2$ and $\nu_{i0}^2 << \nu_{i1}^2$, trans-dimensional setting with $\tau_{i0}^2 = \nu_{i0}^2 = 0$

Prior and Posterior

- Prior:
  - $\pi(\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2) = \pi(\boldsymbol{\alpha}|\boldsymbol{\gamma}, \sigma^2)\pi(\boldsymbol{B}|\boldsymbol{\gamma}, \sigma^2)\pi(\sigma^2)\pi(\boldsymbol{\gamma})$
  - $\pi(\boldsymbol{B}|\boldsymbol{\gamma}, \sigma^2) = \pi(\boldsymbol{B}|\boldsymbol{\mu}, \boldsymbol{\gamma}, \sigma^2)\pi(\boldsymbol{\mu}|\boldsymbol{\gamma}, \sigma^2)$
  - $\sigma^2 \sim IG(a, b)$
  - $\boldsymbol{\gamma}$ product of independent multinomial
- Posterior:
  - marginal posterior $\pi(\boldsymbol{\gamma}|\boldsymbol{Y})$
  - $\pi(\boldsymbol{\gamma}|\boldsymbol{Y})$ involves the sum of the product of residual matrices on $\boldsymbol{B}$ and $\boldsymbol{\mu}$ levels with different shrinkage
- Comment : the model framework is very general, the variable selection approach can be applied to both gene-disease association and gene-set enrichment studies

Simulation : from the model with cross-profile signal

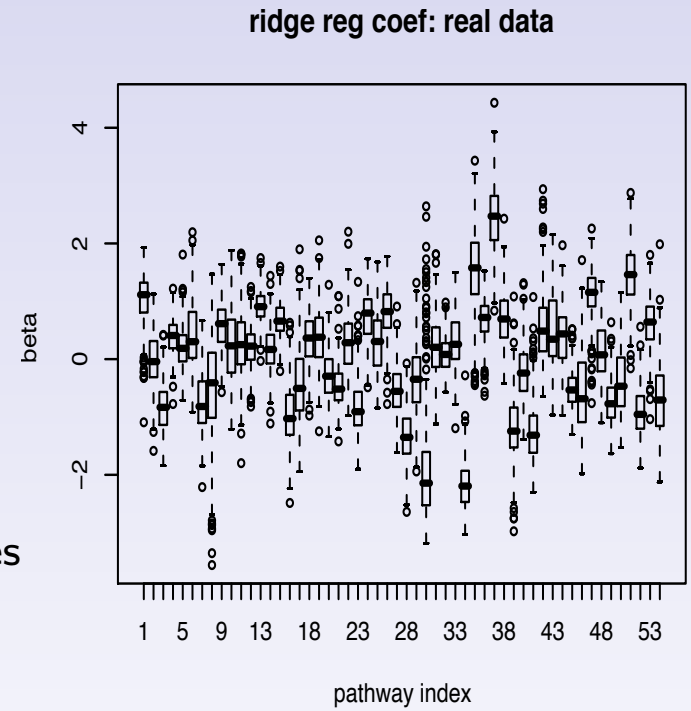
- $\boldsymbol{X}_{104\times100}$:
  - 104 yeast recombinant strains
  - randomly selection of 100 genes
  - from a real study of yeast growth under 92 different drugs
- 40 genes are in the true model:
  - 20 genes with similar association across 92 drugs
  - 20 genes with different association across 92 drugs
- generate $\boldsymbol{\mu}$, $\boldsymbol{B}|\boldsymbol{\mu}$ with specified gene types and large signal
- simulation from the model with no cross-profile signal: 3-type model and previous 2-type model have similar performance

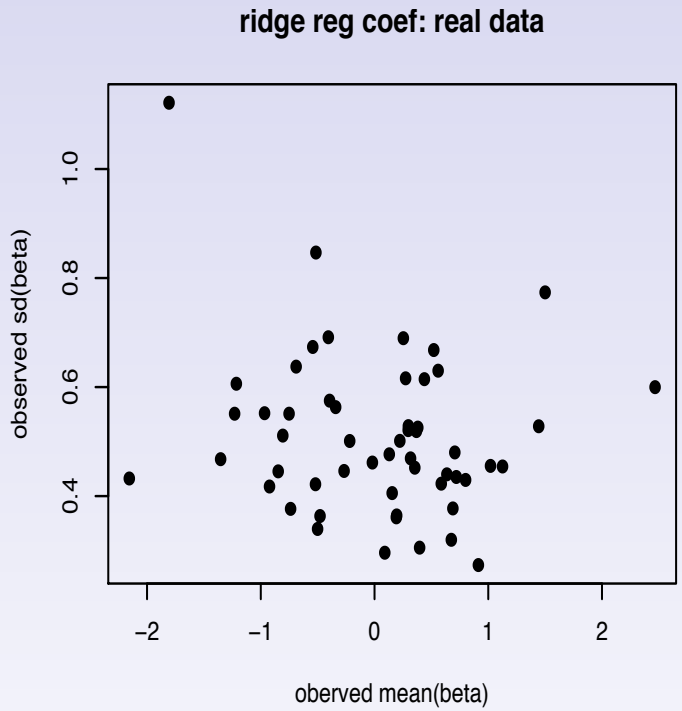


Case study

- Breast cancer by Miller et.al.
  hgu133A chip
  rma normalization
- 22283 genes
  251 samples
  190 pathways
- top 5% variable genes
  1115 genes

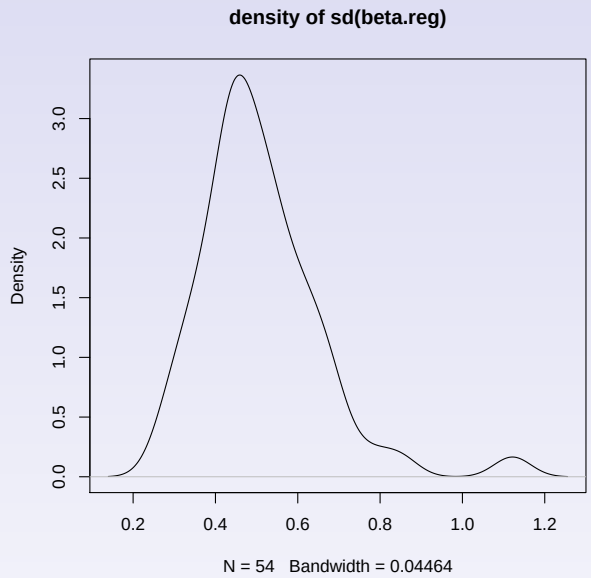
  54 sets $>$ 5 genes





- Non-conjugate model for signal and noise relationship in enrichment study
- A short Markov chain found no gene-set with high posterior probability close to 1
- Top differential enriched sets appear to be reasonable: pathway related to stage of the tumor and P53
- Top common enriched sets appear to be mostly in general biological processes
- The mean of a gene-set a conservative measure for enrichment?

[1] "Olfactory transduction"
[2] "Glioma"
[3] "Nitrogen metabolism"
[4] "Long-term potentiation"
[5] "Thyroid cancer"
[6] "Cell cycle"
[7] "Toll-like receptor signaling pathway"
[8] "Neurodegenerative Disorders"
[9] "Cell Communication"
[10] "Melanoma"
[11] "Pancreatic cancer"
[12] "Epithelial cell signaling in Helicobacter pylori infection"
[13] "Antigen processing and presentation"
[14] "Wnt signaling pathway"
[15] "Apoptosis"
[16] "Fatty acid metabolism"
[17] "Glutamate metabolism"
[18] "Regulation of actin cytoskeleton"
[19] "Gap junction"
[20] "Natural killer cell mediated cytotoxicity"
[21] "Hedgehog signaling pathway"
[22] "Adipocytokine signaling pathway"
[23] "Jak-STAT signaling pathway"



26 pathways with sd > 0.5
7 out of 17 p53 pathways

- Summary
  - The model framework is general
  - Marginal posterior of the latent variable type can be obtained
  - Conjugate or non-conjugate setting is flexible
  - Variable selection can be applied to both gene and gene-set levels

Thanks!