

# Semi-parametric Bayesian Inference for High-Throughput Gene Expression Data

PETER MÜLLER

Dpt of Biostat., M.D. Anderson Cancer Center

# Outline

## Intro

- Random functions = nonparametric Bayes
- High-throughput arrays for gene and protein expression

# Outline

## Intro

- Random functions = nonparametric Bayes
  - High-throughput arrays for gene and protein expression
1. **Microarrays:** Differential gene expression
  2. **Mass spectrometry:** Mass/charge spectra
  3. **SAGE:** Poisson/Gamma DP mixture

# Nonparametric Bayesian Inference

- Probability model on infinite dimensional space, i.e., infinite dimensional parameter vector;

# Nonparametric Bayesian Inference

- Probability model on infinite dimensional space, i.e., infinite dimensional parameter vector;
- Prob models on random functions (and densities);

# Nonparametric Bayesian Inference

- Probability model on infinite dimensional space, i.e., infinite dimensional parameter vector;
- Prob models on random functions (and densities);
- Avoids critical dependence on parametric assumptions;

# Nonparametric Bayesian Inference

- Probability model on infinite dimensional space, i.e., infinite dimensional parameter vector;
- Prob models on random functions (and densities);
- Avoids critical dependence on parametric assumptions;
- Robustifies parametric models (non-parametric model centered at parametric model);

# Nonparametric Bayesian Inference

- Probability model on infinite dimensional space, i.e., infinite dimensional parameter vector;
- Prob models on random functions (and densities);
- Avoids critical dependence on parametric assumptions;
- Robustifies parametric models (non-parametric model centered at parametric model);
- Model diagnostic and sensitivity analysis.



# High-Throughput Assays

DNA  $\rightarrow$  mRNA  $\rightarrow$  proteins  $\rightarrow$  us ...

# High-Throughput Assays

DNA  $\rightarrow$  mRNA  $\rightarrow$  proteins  $\rightarrow$  us ...

## Microarrays:

- Measure mRNA for a (large) number of selected genes,  $g = 1, \dots, G$ .

# High-Throughput Assays

DNA  $\rightarrow$  mRNA  $\rightarrow$  proteins  $\rightarrow$  us ...

## Microarrays:

- Measure mRNA for a (large) number of selected genes,  $g = 1, \dots, G$ .
- Usually multiple arrays (samples):  $t = 1, \dots, N$ .

# High-Throughput Assays

DNA  $\rightarrow$  mRNA  $\rightarrow$  proteins  $\rightarrow$  us ...

## Microarrays:

- Measure mRNA for a (large) number of selected genes,  $g = 1, \dots, G$ .
- Usually multiple arrays (samples):  $t = 1, \dots, N$ .
- **Data:**  $(G \times N)$  matrix  $x_{gt}$  of gene expression for gene  $g$ , sample  $t$ .

# High-Throughput Assays

DNA  $\rightarrow$  mRNA  $\rightarrow$  proteins  $\rightarrow$  us ...

## Microarrays:

- Measure mRNA for a (large) number of selected genes,  $g = 1, \dots, G$ .
- Usually multiple arrays (samples):  $t = 1, \dots, N$ .
- **Data:**  $(G \times N)$  matrix  $x_{gt}$  of gene expression for gene  $g$ , sample  $t$ .

## Protein Mass spectrometry:

- Record proteins (mass, time-of-flight) in a probe.

# High-Throughput Assays

DNA  $\rightarrow$  mRNA  $\rightarrow$  proteins  $\rightarrow$  us ...

## Microarrays:

- Measure mRNA for a (large) number of selected genes,  $g = 1, \dots, G$ .
- Usually multiple arrays (samples):  $t = 1, \dots, N$ .
- **Data:** ( $G \times N$ ) matrix  $x_{gt}$  of gene expression for gene  $g$ , sample  $t$ .

## Protein Mass spectrometry:

- Record proteins (mass, time-of-flight) in a probe.
- **Data:** histogram (“spectrum”) with peaks corresponding to detected proteins.

## SAGE: Serial Analysis of Gene Expression

- Measure mRNA (tags of 10 base pairs) present in probe.

## SAGE: Serial Analysis of Gene Expression

- Measure mRNA (tags of 10 base pairs) present in probe.
- **Data:** tag counts.



## SAGE: Serial Analysis of Gene Expression

- Measure mRNA (tags of 10 base pairs) present in probe.
- **Data:** tag counts.

**Pre-processing:** Critically important, but not usually np-bayes.

# Microarrays: Differential Gene Expression

**Two-group Comparisons:** considers comparison of samples under two biologic conditions,  $x_t \in \{0, 1\}$

# Microarrays: Differential Gene Expression

**Two-group Comparisons:** considers comparison of samples under two biologic conditions,  $x_t \in \{0, 1\}$

**Mixtures:**

- Efron et al. (2001 JASA), empirical Bayes

# Microarrays: Differential Gene Expression

**Two-group Comparisons:** considers comparison of samples under two biologic conditions,  $x_t \in \{0, 1\}$

**Mixtures:**

- Efron et al. (2001 JASA), empirical Bayes
- Parmigiani et al. (2002 JRSSB), mixture of uniform (under-expression), normal (typical) and uniform (over)

# Microarrays: Differential Gene Expression

**Two-group Comparisons:** considers comparison of samples under two biologic conditions,  $x_t \in \{0, 1\}$

**Mixtures:**

- Efron et al. (2001 JASA), empirical Bayes
- Parmigiani et al. (2002 JRSSB), mixture of uniform (under-expression), normal (typical) and uniform (over)
- Ibrahim et al. (2002 JASA), mixture with point mass for non-expressed genes

## Hierarchical models:

- Newton et al. (2001 J Comp Bio), Gamma/Gamma hierarchical model with indicator for non-differential expression

## Hierarchical models:

- Newton et al. (2001 J Comp Bio), Gamma/Gamma hierarchical model with indicator for non-differential expression
- Hein et al. (2005 Biostat) and Lewin et al. (2005 Biometrics): hierarchical models.
- and many many others!

## Hierarchical models:

- Newton et al. (2001 J Comp Bio), Gamma/Gamma hierarchical model with indicator for non-differential expression
- Hein et al. (2005 Biostat) and Lewin et al. (2005 Biometrics): hierarchical models.
- and many many others!

**Dependence:** Network models (e.g., Dobra et al. 2004 J MvAnal), CART (Pittman et al, 2004 PNAS), factor models, PCA



## Hierarchical models:

- Newton et al. (2001 J Comp Bio), Gamma/Gamma hierarchical model with indicator for non-differential expression
- Hein et al. (2005 Biostat) and Lewin et al. (2005 Biometrics): hierarchical models.
- and many many others!

**Dependence:** Network models (e.g., Dobra et al. 2004 J MvAnal), CART (Pittman et al, 2004 PNAS), factor models, PCA

**Sample size:** Power, ROC curve, parametrized learning curve, decision theoretic

# A Semiparametric Mixture of Normal Model

with K.-A. DO and F. TANG (M.D. Anderson Cancer Center)

- Microarray experiments: Measure gene expression for many ( $G = 6,500$ ) genes simultaneously;

# A Semiparametric Mixture of Normal Model

with K.-A. DO and F. TANG (M.D. Anderson Cancer Center)

- Microarray experiments: Measure gene expression for many ( $G = 6,500$ ) genes simultaneously;
- Under different conditions: e.g., normal vs. tumor tissue

# A Semiparametric Mixture of Normal Model

with K.-A. DO and F. TANG (M.D. Anderson Cancer Center)

- Microarray experiments: Measure gene expression for many ( $G = 6,500$ ) genes simultaneously;
- Under different conditions: e.g., normal vs. tumor tissue
- **Data:** difference scores  $x_g$  for each gene,  $g = 1, \dots, G$ ,

# A Semiparametric Mixture of Normal Model

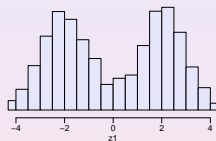
with K.-A. DO and F. TANG (M.D. Anderson Cancer Center)

- Microarray experiments: Measure gene expression for many ( $G = 6,500$ ) genes simultaneously;
- Under different conditions: e.g., normal vs. tumor tissue
- **Data:** difference scores  $x_g$  for each gene,  $g = 1, \dots, G$ , e.g., t-statistic for each gene.

# Differences Scores

Affected Genes:  
differentially expressed genes,  
difference score  $x_g$  for differ-  
ence of normal vs. tumor tis-  
sue

$$f_1(x)$$

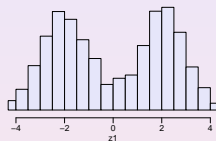


# Differences Scores

Affected Genes:

differentially expressed genes,  
difference score  $x_g$  for differ-  
ence of normal vs. tumor tis-  
sue

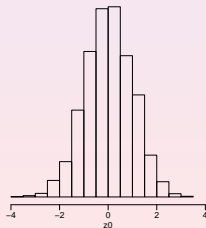
$$f_1(x)$$



Non affected genes:

non differentially expressed  
genes, differences normal vs.  
tumor

$$f_0(x)$$

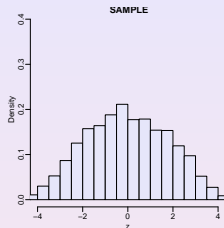


## Data:

mixture of  $f_0$  and  $f_1$   
need deconvolution



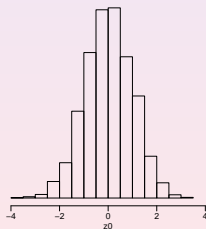
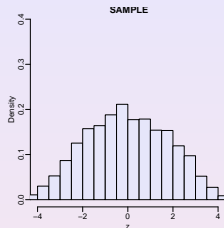
Data:  
mixture of  $f_0$  and  $f_1$   
need deconvolution



**Data:**

mixture of  $f_0$  and  $f_1$   
need deconvolution

“Null sample”  
(Fake) differences between  
equal conditions:  
 $x \sim f_0(x)$



## Likelihood:

$$p(x_g) = p_0 f_0(x_g) + (1 - p_0) f_1(x_g): \text{ for } g = 1, \dots, G$$

# Likelihood:

$$p(x_g) = p_0 f_0(x_g) + (1 - p_0) f_1(x_g): \text{ for } g = 1, \dots, G$$
$$p(x_g) = f_0(x_g) \quad \quad \quad : \text{ for } g = G + 1, \dots, 2G$$

“null sample”

# Likelihood:

$$p(x_g) = p_0 f_0(x_g) + (1 - p_0) f_1(x_g): \text{ for } g = 1, \dots, G$$
$$p(x_g) = f_0(x_g) \quad \quad \quad : \text{ for } g = G + 1, \dots, 2G$$

“null sample”

Parameters  $p_0$

# Likelihood:

$$\begin{aligned} p(x_g) &= p_0 f_0(x_g) + (1 - p_0) f_1(x_g) && \text{for } g = 1, \dots, G \\ p(x_g) &= f_0(x_g) && \text{for } g = G + 1, \dots, 2G \\ &\text{"null sample"} \end{aligned}$$

Parameters  $p_0$  and (!! ) unknown distributions  $f_0, f_1$

# Likelihood:

$$\begin{aligned} p(x_g) &= p_0 f_0(x_g) + (1 - p_0) f_1(x_g): \text{ for } g = 1, \dots, G \\ p(x_g) &= f_0(x_g) : \text{ for } g = G + 1, \dots, 2G \\ &\text{"null sample"} \end{aligned}$$

Parameters  $p_0$  and (!! ) unknown distributions  $f_0, f_1$

Prior:  $p(p_0)$ ,  $p(f_0)$  and  $p(f_1)$

# Likelihood:

$$\begin{aligned} p(x_g) &= p_0 f_0(x_g) + (1 - p_0) f_1(x_g) && \text{for } g = 1, \dots, G \\ p(x_g) &= f_0(x_g) && \text{for } g = G + 1, \dots, 2G \\ &\text{"null sample"} \end{aligned}$$

Parameters  $p_0$  and (!! ) unknown distributions  $f_0, f_1$

Prior:  $p(p_0), p(f_0)$  and  $p(f_1)$

Posterior inference:  $p(p_0, f_0, f_1 \mid x)$

... and inference for any function of  $p_0, f_0, f_1$ .



# DP Mixture of Normals

DP mixture of normals:

- $f_j$  : mixture of normals with random mixing measure  $F_j$

# DP Mixture of Normals

DP mixture of normals:

- $f_j$  : mixture of normals with random mixing measure  $F_j$
- DP prior for  $F_j$

$$f_j(x) = \int N(x; \mu, \sigma) dF_j(\mu)$$
$$F_j \sim DP(F^*, M).$$

# DP Mixture of Normals

DP mixture of normals:

- $f_j$  : mixture of normals with random mixing measure  $F_j$
- DP prior for  $F_j$

$$f_j(x) = \int N(x; \mu, \sigma) dF_j(\mu)$$
$$F_j \sim DP(F^*, M).$$

Base measure:

$F_0^* = N(0, 1)$  unimodal around 0;

# DP Mixture of Normals

DP mixture of normals:

- $f_j$  : mixture of normals with random mixing measure  $F_j$
- DP prior for  $F_j$

$$f_j(x) = \int N(x; \mu, \sigma) dF_j(\mu)$$
$$F_j \sim DP(F^*, M).$$

Base measure:

$F_0^* = N(0, 1)$  unimodal around 0;

$F_1^* = 0.5N(-b, 1) + 0.5N(+b, 1)$ , bimodal around 0.

# Posterior MCMC

Random partition:

- $F_0$  is a.s. discrete  $\rightarrow$  ties

# Posterior MCMC

Random partition:

- $F_0$  is a.s. discrete  $\rightarrow$  ties
- $\{\mu_1^*, \dots, \mu_L^*\}$ : unique  $\mu_g$ 's

# Posterior MCMC

Random partition:

- $F_0$  is a.s. discrete  $\rightarrow$  ties
- $\{\mu_1^*, \dots, \mu_L^*\}$ : unique  $\mu_g$ 's
- Indicators  $s_g$  with  $s_g = j$  iff  $\mu_g = \mu_j^*$

# Posterior MCMC

Random partition:

- $F_0$  is a.s. discrete  $\rightarrow$  ties
- $\{\mu_1^*, \dots, \mu_L^*\}$ : unique  $\mu_g$ 's
- Indicators  $s_g$  with  $s_g = j$  iff  $\mu_g = \mu_j^*$

Joint prior: marginalize w.r.t.  $F_0 \rightarrow p(s, \mu) = p(s) p(\mu | s)$



# Posterior MCMC

Random partition:

- $F_0$  is a.s. discrete  $\rightarrow$  ties
- $\{\mu_1^*, \dots, \mu_L^*\}$ : unique  $\mu_g$ 's
- Indicators  $s_g$  with  $s_g = j$  iff  $\mu_g = \mu_j^*$

Joint prior: marginalize w.r.t.  $F_0 \rightarrow p(s, \mu) = p(s) p(\mu | s)$

$$p(s) = \frac{M^L \Gamma(M) \prod_{j=1}^L \Gamma(n_j)}{\Gamma(M+G)} \text{ and } p(\mu_j^* | s) = F^*(\mu_j^*)$$

where  $n_j = \sum I(s_i = j)$ .

Easy to show from Polya urn scheme.

Conjugate DP mixture:

- Conjugate normal base measure  $F_0^*$

## Conjugate DP mixture:

- Conjugate normal base measure  $F_0^*$
- marginalize  $p(x | s, \mu^*)$  w.r.t.  $\mu^*$  to find  $p(x | s)$

## Conjugate DP mixture:

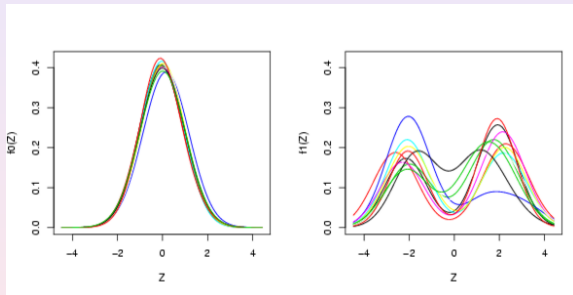
- Conjugate normal base measure  $F_0^*$
- marginalize  $p(x | s, \mu^*)$  w.r.t.  $\mu^*$  to find  $p(x | s)$
- easy MCMC

## Conjugate DP mixture:

- Conjugate normal base measure  $F_0^*$
- marginalize  $p(x | s, \mu^*)$  w.r.t.  $\mu^*$  to find  $p(x | s)$
- easy MCMC

$f_1$ : same thing ...

# Posterior inference: RPM



Posterior draws  $f_0 \sim p(f_0 \mid data)$  (left)  $f_1 \sim p(f_1 \mid data)$  (right).

## Posterior inference: Differential expression

Recall splg model:  $x_g \sim p_0 f_0(x) + (1 - p_0) f_1(x)$ .

# Posterior inference: Differential expression

Recall splg model:  $x_g \sim p_0 f_0(x) + (1 - p_0) f_1(x)$ .  
Equivalent hierarchical model:

$$p(x_g \mid r_g = j) = f_j(x_g)$$



# Posterior inference: Differential expression

Recall splg model:  $x_g \sim p_0 f_0(x) + (1 - p_0) f_1(x)$ .  
Equivalent hierarchical model:

$$\begin{aligned} p(x_g \mid r_g = j) &= f_j(x_g) \\ Pr(r_g = 0) &= p_0 \end{aligned}$$

# Posterior inference: Differential expression

Recall splg model:  $x_g \sim p_0 f_0(x) + (1 - p_0) f_1(x)$ .  
Equivalent hierarchical model:

$$\begin{aligned} p(x_g \mid r_g = j) &= f_j(x_g) \\ Pr(r_g = 0) &= p_0 \end{aligned}$$

Interpret  $r_g$ : indicator for diff expression.

# Posterior inference: Differential expression

Recall splg model:  $x_g \sim p_0 f_0(x) + (1 - p_0) f_1(x)$ .

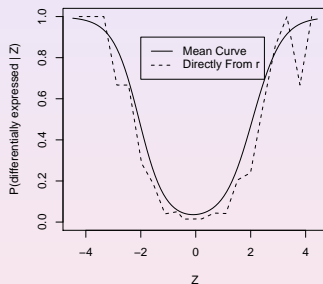
Equivalent hierarchical model:

$$\begin{aligned} p(x_g | r_g = j) &= f_j(x_g) \\ Pr(r_g = 0) &= p_0 \end{aligned}$$

Interpret  $r_g$ : indicator for diff expression.

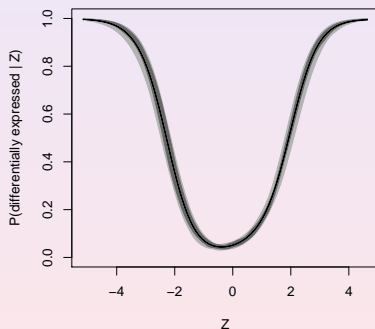
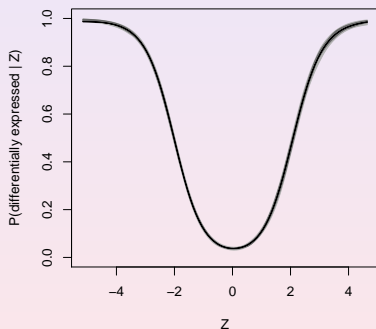
Posterior: Can show  $E(r_g | data) = E(P_1(x_g) | data)$  for

$$P_1(x_g) = \frac{(1 - p_0)f_1(x)}{p_0 f_0(x) + (1 - p_0) f_1(x)}$$



$E(P_1(x_g) | data)$  (solid curve) and truth (dashed) against  $x_g$ .

# With and Without Null Sample



# Limitations and Extensions

Difference scores: Not clear what is the right way to define  $x_g$ .

# Limitations and Extensions

**Difference scores:** Not clear what is the right way to define  $x_g$ .

**Dependence:** Gene expression is dependent across  $g$  — arrgh!

# Limitations and Extensions

**Difference scores:** Not clear what is the right way to define  $x_g$ .

**Dependence:** Gene expression is dependent across  $g$  — arrgh!

**Design:** Only considered two-group comparison. More general layouts are used.



## Limitations and Extensions

**Difference scores:** Not clear what is the right way to define  $x_g$ .

**Dependence:** Gene expression is dependent across  $g$  — arrgh!

**Design:** Only considered two-group comparison. More general layouts are used.

**Too easy!** Using *null sample* you essentially nail  $f_0$ .

# Protein Mass/Charge Spectra

MALDI-TOF: Matrix Assisted Laser Desorption Ionization

- Suspend a sample in a matrix

# Protein Mass/Charge Spectra

## MALDI-TOF: Matrix Assisted Laser Desorption Ionization

- Suspend a sample in a matrix
- Laser ionizes molecules from sample (laser-induced desorption process)

# Protein Mass/Charge Spectra

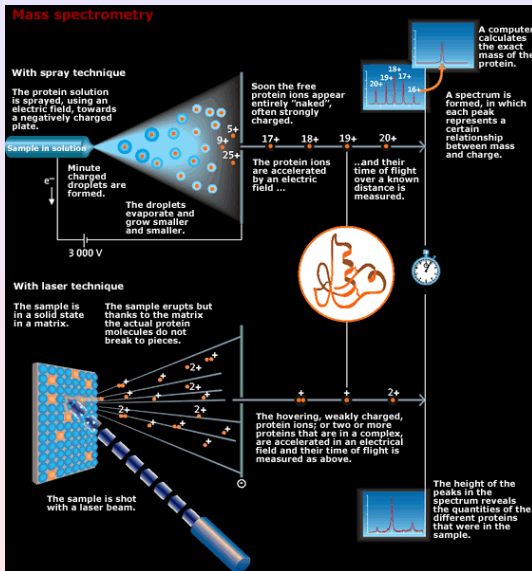
## MALDI-TOF: Matrix Assisted Laser Desorption Ionization

- Suspend a sample in a matrix
- Laser ionizes molecules from sample (laser-induced desorption process)
- Electric field accelerates particles

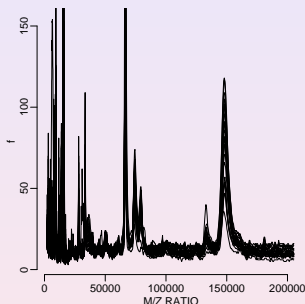
# Protein Mass/Charge Spectra

## MALDI-TOF: Matrix Assisted Laser Desorption Ionization

- Suspend a sample in a matrix
- Laser ionizes molecules from sample (laser-induced desorption process)
- Electric field accelerates particles
- Time Of Flight: separates ions by mass/charge
  - $\text{TOF} \propto (m/z)^{1/2}$
  - Measure the proportions of ions with size  $m/z$

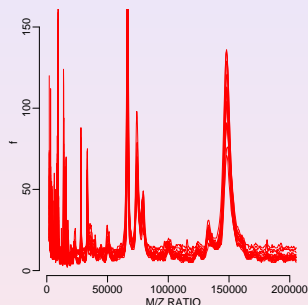
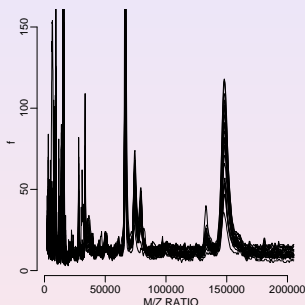


## Data



$G_0 = 17$  normal samples,  
histogram of mass/charge ratios on grid of size  $l = 60,000$ .  
First Annual Conf on Proteomics & Data Mining at Duke U.

## Data



$G_0 = 17$  normal samples,  $G_1 = 24$  tumor samples;  
histogram of mass/charge ratios on grid of size  $l = 60,000$ .  
First Annual Conf on Proteomics & Data Mining at Duke U.



Multi-step methods. Baggerly et al. (2003, Proteomics):

- Multi-step methods. Baggerly et al. (2003, Proteomics):
- baseline subtraction (with windowed local min);

Multi-step methods. Baggerly et al. (2003, Proteomics):

- baseline subtraction (with windowed local min);
- sinusoidal noise removal (! a/c current);

Multi-step methods. Baggerly et al. (2003, Proteomics):

- baseline subtraction (with windowed local min);
- sinusoidal noise removal (! a/c current);
- windowed dimension reduction to define peaks;

Multi-step methods. Baggerly et al. (2003, Proteomics):

- baseline subtraction (with windowed local min);
- sinusoidal noise removal (! a/c current);
- windowed dimension reduction to define peaks;
- genetic algorithm and exhaustive search to find subsets of peaks.

**Multi-step methods.** Baggerly et al. (2003, Proteomics):

- baseline subtraction (with windowed local min);
- sinusoidal noise removal (! a/c current);
- windowed dimension reduction to define peaks;
- genetic algorithm and exhaustive search to find subsets of peaks.

**Wavelet-based smoothing.** Morris et al. (2005 Biometrics):  
represent spectra in wavelet basis → dimension reduction  
and convenient smoothing.

# A Mixture of Beta Model for Protein Mass/charge Spectra

with KIM-ANH DO, KEITH BAGGERLY and RAJ  
BANDYOPADHYAY

**Data:** spectrum = histogram  $y_t(m)$  of observed counts,  
sample  $t$ , mass/charge  $m$

# A Mixture of Beta Model for Protein Mass/charge Spectra

with KIM-ANH DO, KEITH BAGGERLY and RAJ  
BANDYOPADHYAY

**Data:** spectrum = histogram  $y_t(m)$  of observed counts,  
sample  $t$ , mass/charge  $m$

**Parameter:**  $p_t(m_i) =$  true abundance of  $m/z$  ratio  $m_i$  in the probe.



# A Mixture of Beta Model for Protein Mass/charge Spectra

with KIM-ANH DO, KEITH BAGGERLY and RAJ  
BANDYOPADHYAY

**Data:** spectrum = histogram  $y_t(m)$  of observed counts,  
sample  $t$ , mass/charge  $m$

**Parameter:**  $p_t(m_i) = \text{true}$  abundance of  $m/z$  ratio  $m_i$  in the probe.

**Goal:** Decompose  $p_t$  into background  $B_t$  and protein peaks  $f_t$ .

# A Mixture of Beta Model for Protein Mass/charge Spectra

with KIM-ANH DO, KEITH BAGGERLY and RAJ  
BANDYOPADHYAY

**Data:** spectrum = histogram  $y_t(m)$  of observed counts,  
sample  $t$ , mass/charge  $m$

**Parameter:**  $p_t(m_i) = \text{true}$  abundance of  $m/z$  ratio  $m_i$  in the probe.

**Goal:** Decompose  $p_t$  into background  $B_t$  and protein peaks  $f_t$ .

- *Background:* detector noise, protein fragments, matrix
- ...

# A Mixture of Beta Model for Protein Mass/charge Spectra

with KIM-ANH DO, KEITH BAGGERLY and RAJ  
BANDYOPADHYAY

**Data:** spectrum = histogram  $y_t(m)$  of observed counts,  
sample  $t$ , mass/charge  $m$

**Parameter:**  $p_t(m_i)$  = true abundance of  $m/z$  ratio  $m_i$  in the probe.

**Goal:** Decompose  $p_t$  into background  $B_t$  and protein peaks  $f_t$ .

- *Background:* detector noise, protein fragments, matrix  
...
- *Protein peaks:* each protein with  $m/z$  ratio  $m$   
plus noise due to initial velocity dist & mmt error  
→ peak centered around  $m$ .

# A Mixture of Beta Model for Protein Mass/charge Spectra

with KIM-ANH DO, KEITH BAGGERLY and RAJ  
BANDYOPADHYAY

**Data:** spectrum = histogram  $y_t(m)$  of observed counts,  
sample  $t$ , mass/charge  $m$

**Parameter:**  $p_t(m_i)$  = true abundance of  $m/z$  ratio  $m_i$  in the probe.

**Goal:** Decompose  $p_t$  into background  $B_t$  and protein peaks  $f_t$ .

- *Background:* detector noise, protein fragments, matrix  
...
- *Protein peaks:* each protein with  $m/z$  ratio  $m$   
plus noise due to initial velocity dist & mmt error  
→ peak centered around  $m$ .

Prob model for  $f_t$  and  $B_t$  →

# A Mixture of Beta Model for Protein Mass/charge Spectra

with KIM-ANH DO, KEITH BAGGERLY and RAJ  
BANDYOPADHYAY

**Data:** spectrum = histogram  $y_t(m)$  of observed counts,  
sample  $t$ , mass/charge  $m$

**Parameter:**  $p_t(m_i)$  = true abundance of  $m/z$  ratio  $m_i$  in the probe.

**Goal:** Decompose  $p_t$  into background  $B_t$  and protein peaks  $f_t$ .

- *Background:* detector noise, protein fragments, matrix  
...
- *Protein peaks:* each protein with  $m/z$  ratio  $m$   
plus noise due to initial velocity dist & mmt error  
→ peak centered around  $m$ .

Prob model for  $f_t$  and  $B_t$  →

- inference on peaks,
- expression of peaks across conditions.

# Mixture of Betas

Peaks: Kernels  $Be(m, s)$ , location  $m$ , scale  $s$ .

$$f_t(m) = \sum_{g=1}^G w_{xg} \text{Be}(m; \epsilon_g, \alpha_g)$$

# Mixture of Betas

Peaks: Kernels  $Be(m, s)$ , location  $m$ , scale  $s$ .

$$f_t(m) = \sum_{g=1}^G w_{xg} \text{Be}(m; \epsilon_g, \alpha_g)$$

biologic cond  $x = x_t$

# Mixture of Betas

Peaks: Kernels  $Be(m, s)$ , location  $m$ , scale  $s$ .

$$f_t(m) = \sum_{g=1}^G w_{xg} \text{Be}(m; \epsilon_g, \alpha_g)$$

biologic cond  $x = x_t$

Baseline:  $B_t(y) = \sum_{j=1}^{J_t} v_{tj} \text{Be}(m_j | \eta_{tj}, \beta_{tj})$ .



# Mixture of Betas

Peaks: Kernels  $Be(m, s)$ , location  $m$ , scale  $s$ .

$$f_t(m) = \sum_{g=1}^G w_{xg} Be(m; \epsilon_g, \alpha_g)$$

biologic cond  $x = x_t$

Baseline:  $B_t(y) = \sum_{j=1}^{J_t} v_{tj} Be(m_j | \eta_{tj}, \beta_{tj})$ .

Spectrum:  $p_t(m) = p_{0k} B_t(m) + (1 - p_{0k}) f_t(m)$

## Likelihood:

- $y_t(m)$  count of events at mass  $m$  with  $p_t(m)$ .  
empirical distr of  $n$  samples from  $p_t$

# Likelihood:

- $y_t(m)$  count of events at mass  $m$  with  $p_t(m)$ .  
empirical distr of  $n$  samples from  $p_t$

$$\log p(y | \theta) = \sum_{t=1}^N \sum_{i=1}^I y_t(m_i) \log p_t(m_i)$$

# Likelihood:

- $y_t(m)$  count of events at mass  $m$  with  $p_t(m)$ .  
empirical distr of  $n$  samples from  $p_t$

$$\log p(y | \theta) = \sum_{t=1}^N \sum_{i=1}^I y_t(m_i) \log p_t(m_i)$$

(density estimation likelihood)

# Prior:

- Indep prior on  $(\epsilon_g, \alpha_g)$

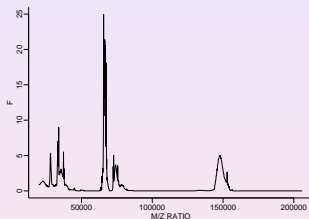
# Prior:

- Indep prior on  $(\epsilon_g, \alpha_g)$
- Hierarchical prior on  $w_{xg}$ 
  - Indicator  $\lambda_g = I(w_{0g} = w_{1g})$
  - Dirichlet prior on  $(w_{0g}; \lambda_g = 1)$  and  $(w_{xg}; \lambda_g = 0)$ ,  $x = 0, 1$ .

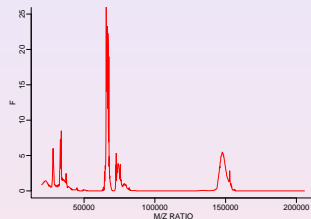
# Prior:

- Indep prior on  $(\epsilon_g, \alpha_g)$
- Hierarchical prior on  $w_{xg}$ 
  - Indicator  $\lambda_g = I(w_{0g} = w_{1g})$
  - Dirichlet prior on  $(w_{0g}; \lambda_g = 1)$  and  $(w_{xg}; \lambda_g = 0)$ ,  $x = 0, 1$ .
- Prior on number of peaks  $G$   
→ random size mixture → RJMCMC

# Estimated Spectra



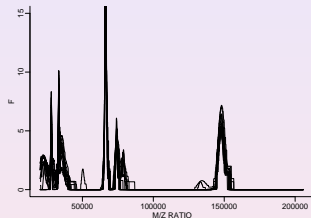
(a)  $E[f_t(\cdot) | Y]$ ,  
normal  $x_t = 0$



(b)  $E[f_t(\cdot) | Y]$ ,  
tumor  $x_t = 1$

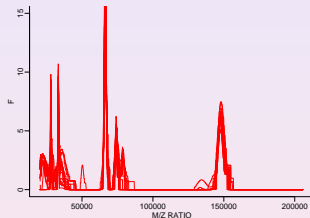
$E[f_t(m) | Y, x_t = x]$ . Estimated spectrum for normal and tumor samples.



Prob Model on  $f_t$ :

$$(a) f_t \sim p[f_t(\cdot) | Y],$$

normal  $x_t = 1$



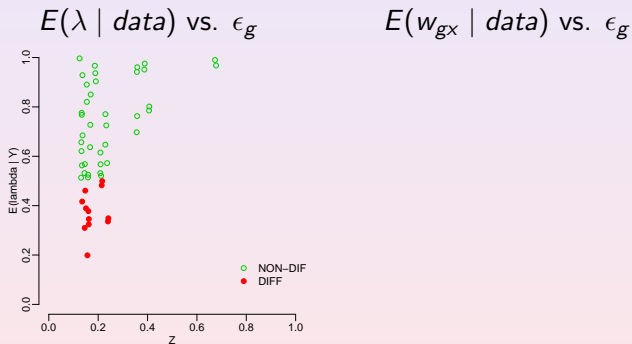
$$(b) f_t \sim p[f_t(\cdot) | Y],$$

tumor  $x_t = 1$

Random draws from the posterior on the unknown spectra.

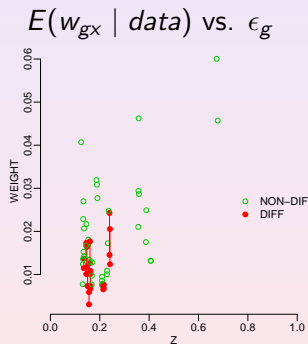
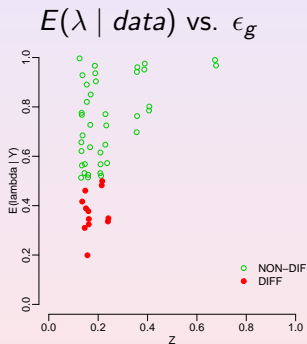
# Differential expression

Marginal posterior probabilities of differential expr.

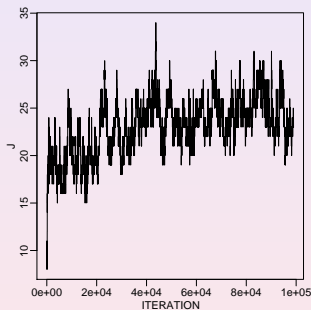
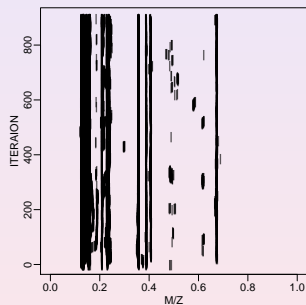


# Differential expression

Marginal posterior probabilities of differential expr.



## Results – MCMC

 $J$  vs. iteration $\epsilon_g$  vs. iteration

Some aspects of the posterior simulation

## Limitations and Extensions

**Sampling model:** Used  $w_{xg}$ , same for all samples with same biological condition  $x$ . Additional variability is reasonable.

## Limitations and Extensions

**Sampling model:** Used  $w_{xg}$ , same for all samples with same biological condition  $x$ . Additional variability is reasonable.

**Prior:** Peaks for higher mass proteins are wider. Could use this in prior.

## Limitations and Extensions

- Sampling model:** Used  $w_{xg}$ , same for all samples with same biol condition  $x$ . Additional variability is reasonable.
- Prior:** Peaks for higher mass proteins are wider. Could use this in prior.
- Protein identity:** Need to match different  $\epsilon_g$  with actual proteins (mode matching problem).

## Limitations and Extensions

**Sampling model:** Used  $w_{xg}$ , same for all samples with same biol condition  $x$ . Additional variability is reasonable.

**Prior:** Peaks for higher mass proteins are wider. Could use this in prior.

**Protein identity:** Need to match different  $\epsilon_g$  with actual proteins (mode matching problem).

**Design:** Usually more than two samples.



# Limitations and Extensions

**Sampling model:** Used  $w_{xg}$ , same for all samples with same biol condition  $x$ . Additional variability is reasonable.

**Prior:** Peaks for higher mass proteins are wider. Could use this in prior.

**Protein identity:** Need to match different  $\epsilon_g$  with actual proteins (mode matching problem).

**Design:** Usually more than two samples.

**Likelihood:** Neither is perfect:

- Density estimation:  $y_t$  as empirical distribution of a random sample from  $p_t$
- Regression:  $y_t = p_t + \text{residual}$ .

# Serial Analysis of Gene Expression (SAGE)

Data: tags counts  $y_g, g = 1, \dots, G_0$

# Serial Analysis of Gene Expression (SAGE)

Data: tags counts  $y_g$ ,  $g = 1, \dots, G_0$

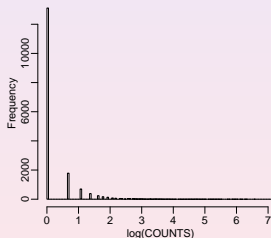
Censoring: tags with  $y_g = 0$  are not recorded

# Serial Analysis of Gene Expression (SAGE)

**Data:** tags counts  $y_g$ ,  $g = 1, \dots, G_0$

**Censoring:** tags with  $y_g = 0$  are not recorded

**Skewed data:** few tags with large count; many with small counts



Zhang et al. (1997, Science).

Mixture of two Dirichlets: Morris et al. (2003 Biometrics),

Mixture of two Dirichlets: Morris et al. (2003 Biometrics),

- Multinomial sampling  $y \sim Mn(\pi; n)$

Mixture of two Dirichlets: Morris et al. (2003 Biometrics),

- Multinomial sampling  $y \sim Mn(\pi; n)$
- (latent) split into scarce and abundant tags

Mixture of two Dirichlets: Morris et al. (2003 Biometrics),

- Multinomial sampling  $y \sim Mn(\pi; n)$
- (latent) split into scarce and abundant tags
- Dirichlet prior for scarce and abundant tag frequencies



# A DP Mixture Model for SAGE Data

Goal: generalize mix of two Dirichlet ...

# A DP Mixture Model for SAGE Data

**Goal:** generalize mix of two Dirichlet . . .

**First:** replace multinomial by Poisson sampling

# A DP Mixture Model for SAGE Data

**Goal:** generalize mix of two Dirichlet . . .

**First:** replace multinomial by Poisson sampling

**Sampling:** Indep Poisson  $y_g \sim Poi(\lambda_g)$

## A DP Mixture Model for SAGE Data

**Goal:** generalize mix of two Dirichlet . . .

**First:** replace multinomial by Poisson sampling

**Sampling:** Indep Poisson  $y_g \sim Poi(\lambda_g)$

**Prior:**  $\lambda_g \sim F$   $F$  should capture skewed nature of data

## A DP Mixture Model for SAGE Data

**Goal:** generalize mix of two Dirichlet . . .

**First:** replace multinomial by Poisson sampling

**Sampling:** Indep Poisson  $y_g \sim Poi(\lambda_g)$

**Prior:**  $\lambda_g \sim F$   $F$  should capture skewed nature of data

**Hyperprior:**  $F \sim DP(F^*, M)$

# DP Mixture Model

Model:  $y_g \sim \int \text{Poi}(y_g; \lambda_g) dF(\lambda_g)$  and  $F \sim DP(F^*, M)$

# DP Mixture Model

Model:  $y_g \sim \int Poi(y_g; \lambda_g) dF(\lambda_g)$  and  $F \sim DP(F^*, M)$

Random partition: etc., as in the normal-normal DP mixture earlier

# DP Mixture Model

Model:  $y_g \sim \int Poi(y_g; \lambda_g) dF(\lambda_g)$  and  $F \sim DP(F^*, M)$

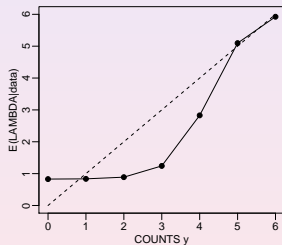
Random partition: etc., as in the normal-normal DP mixture earlier

Conjugate DP mixture:

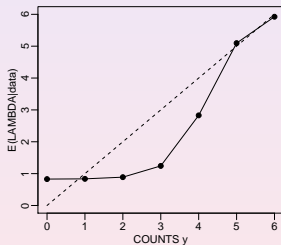
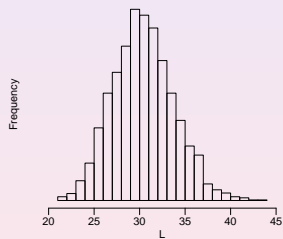
- Conjugate (Gamma) base measure.
- Marginalize w.r.t.  $\lambda^*$  to find  $p(y | s)$
- easy MCMC



## Posterior inference

 $\bar{\lambda}_i$  vs.  $y_i$  $p(L | data)$ 

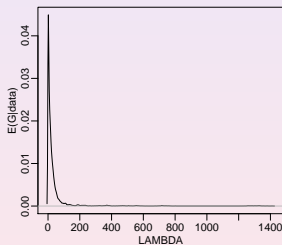
## Posterior inference

 $\bar{\lambda}_i$  vs.  $y_i$  $p(L | \text{data})$ 

## Posterior Random Measure

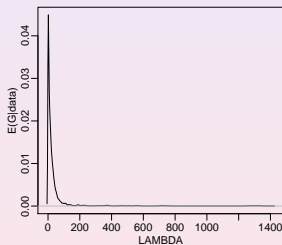
$$E(F \mid \text{data})$$

$$E(F \mid \text{data}), \lambda < 100$$

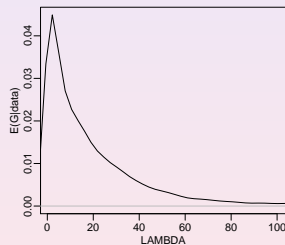


## Posterior Random Measure

$$E(F \mid \text{data})$$



$$E(F \mid \text{data}), \lambda < 100$$



# Summary

- NP Bayes to represent random distributions and functions for massive gene and protein expression data.
- If sample size = number of genes, then we have ample data.
- Joint description of all uncertainties is important to address multiplicities
- We have only discussed two-group comparisons.  
Most experiments involve more complicated designs (ANOVA etc.)