Slide 1

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

Peter Müller
Department of Biostat., M.D. Anderson Cancer Center

1 Intro

Slide 2

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

Slide 3

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.

Slide 4

High-Throughput Assays

DNA → mRNA → proteins → us ...

Microarrays:

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

Slide 5

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.

2 Microarrays

2.1 Intro

Slide 6

Microarrays: Differential Gene Expression

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
- Hein et al. (2005 Biostat) and Lewin et al. (2005 Biometrics): hierarchical models.
- and many many others!

Slide 7

Dependence: Nework 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

Slide 8

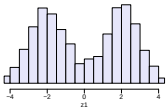
A Semiparametric Mixture of Normal Model
with K.-A. Do and F. Tang (M.D. Anderson Cancer Center)

- 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, $g = 1, \ldots, G$, e.g., t-statistic for each gene.

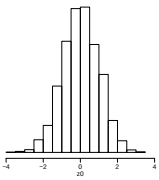
2.2 Data

Slide 9

Differences Scores
Affected Genes:
differentially expressed genes, difference score x_g for difference of normal vs. tumor tissue
$f_1(x)$

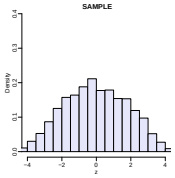


Non affected genes:
non differentially expressed genes, differences normal vs. tumor
$f_0(x)$

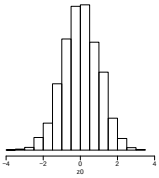


Slide 10

Data:
mixture of f_0 and f_1
need deconvolution



"Null sample"
(Fake) differences between equal conditions:
$x \sim f_0(x)$ '



2.3 Model

Slide 11

Likelihood:

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

$$p(x_g) = \quad f_0(x_g) \qquad\qquad \text{: for } g = G+1, \ldots, 2G$$

"null sample"

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.

Slide 12

DP Mixture of Normals

DP mixture of normals:

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

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

Base measure:
$F_0^\star = N(0, 1)$ unimodal around 0;
$F_1^\star = 0.5N(-b, 1) + 0.5N(+b, 1)$, bimodal around 0.

Slide 13

Posterior MCMC

Random partition:

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

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

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

Easy to show from Polya urn scheme.

Slide 14

Conjugate DP mixture:

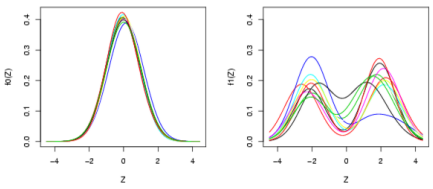
- Conjugate normal base measure $F_0^\star$
- marginalize w.r.t. μ^* to find $p(x \mid s)$
- easy MCMC

f_1**:** same thing ...

2.4 Results

Slide 15

Posterior inference: RPM



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

Slide 16

Posterior inference: Differential expression

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

Equivalent hierarchical model:

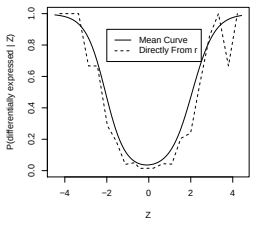
$$\begin{aligned} p(x_g \mid r_g = j) &= f_j(x_g) \\ Pr(r_g = 0) &= p_0 \end{aligned}$$

Interpret r_g as indicator for diff expression.

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

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

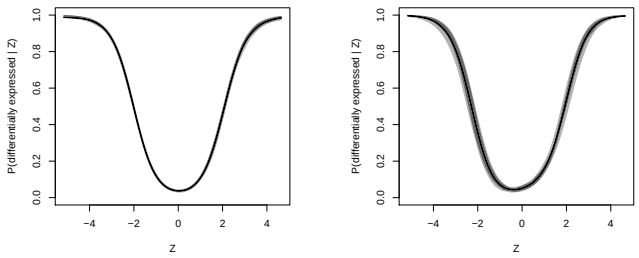
Slide 17



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

Slide 18

With and Without Null Sample



Slide 19

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.

3 Protein Mass Spectra

3.1 Intro

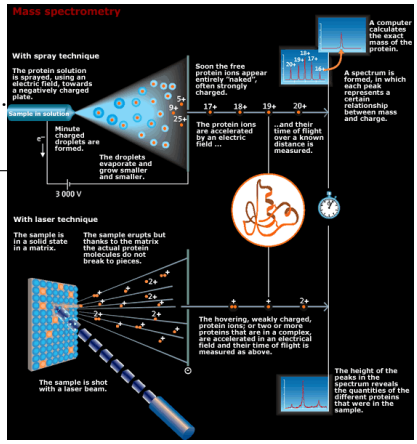
Slide 20

Protein Mass/Charge Spectra

MALDI-TOF: Matrix Assisted Laser Desorption Ionization

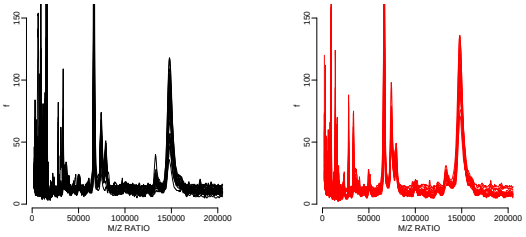
- Suspend a sample in a matrix
- Laser ionizes molecules from sample (laser-induced desorption process)
- Electric field accelerates particles
- Time Of Flight: separates ions by mass/charge
  - TOF $\propto (m/z)^{1/2}$
  - Measure the proportions of ions with size m/z

Slide 21



Slide 22

Data



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

Slide 23

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.

3.2 Model

Slide 24

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: $p_t(m_i)$ = frequency of m/z ratio m_i.

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.

Slide 25

Mixture of Betas

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

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

biologic cond $x = x_t$

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

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

Slide 26

Likelihood:

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

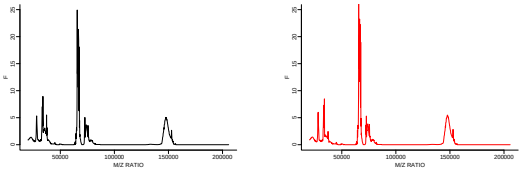
$$\log p(y \mid \theta) = \sum_{t=1}^{N} \sum_{i=1}^{I} y_t(m_i) \log p_t(m_i)$$

(density estimation likelihood)

3.3 Results

Slide 27

Estimated Spectra

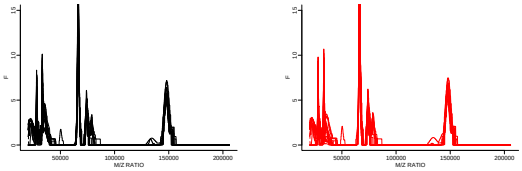


(a) $E[f_t(\cdot) \mid Y]$, normal $x_t = 0$ (b) $E[f_t(\cdot) \mid Y]$, tumor $x_t = 1$

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

Slide 28

Prob Model on f_t:



(a) $f_t \sim p[f_t(\cdot) \mid Y]$, normal $x_t = 1$ (b) $f_t \sim p[f_t(\cdot) \mid Y]$, tumor $x_t = 1$

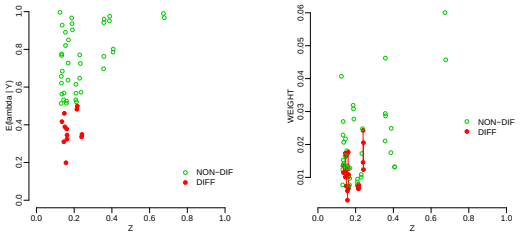
Random draws from the posterior on the unknown spectra.

Slide 29

Differential expression

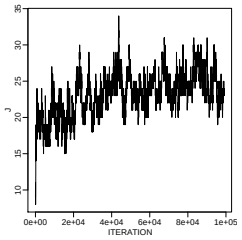
Marginal posterior probabilities of differential expr.

$E(\lambda \mid data)$ vs. ϵ_g $E(w_{gx} \mid data)$ vs. ϵ_g



Slide 30

Results – MCMC



ϵ_g vs. iteration J vs. iteration

Some aspects of the posterior simulation

3.4 Limitations ...

Slide 31

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 ϵ_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: $y_t = p_t +$ residual.

4 SAGE

4.1 Intro

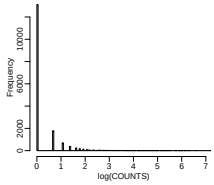
Slide 32

Serial Analysis of Gene Expression (SAGE)

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

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).

Slide 33

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 for scarce and abundant tag frequencies

4.2 Model

Slide 34

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: $\lambda_g \sim F$

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

Slide 35

DP Mixture Model

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

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

Conjugate DP mixture:

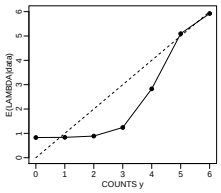
- Conjugate (Gamma) base measure.
- Marginalize w.r.t. λ^* to find $p(y \mid s)$
- easy MCMC

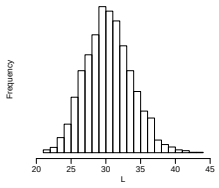
4.3 Posterior Inference

Slide 36

Posterior inference

$\overline{\lambda}_i$ vs. y_i $\quad$ $p(L \mid data)$

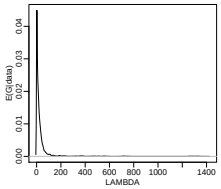


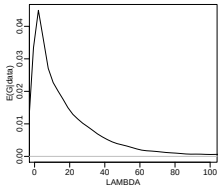


Slide 37

Posterior Random Measure

$E(F \mid data)$ $\quad$ $E(F \mid data)$, $\lambda < 100$





Slide 38

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.)