29

Bayesian Mixture Models for Gene Expression and Protein Profiles

Michele Guindani, Kim-Anh Do, Peter Müller and Jeff S. Morris
M.D. Anderson Cancer Center

Abstract

We review the use of semi-parametric mixture models for Bayesian inference in high throughput genomic data. We discuss three specific approaches for microarray data, for protein mass spectrometry experiments, and for SAGE data. For the microarray data and the protein mass spectrometry we assume group comparison experiments, i.e., experiments that seek to identify genes and proteins that are differentially expressed across two biologic conditions of interest. For the SAGE data example we consider inference for a single biologic sample. For all three applications we use flexible mixture models to implement inference. For the microarray data we define a Dirichlet process mixture of normal model. For the mass spectrometry data we introduce a mixture of Beta model. And the proposed inference for SAGE data is based on a semi-parametric mixture of Poisson distributions.

29.1 Introduction

We discuss semi-parametric Bayesian data analysis for high throughput genomic data. We introduce suitable semi-parametric mixture models to implement inference for microarray data, mass spectrometry data and SAGE data. The proposed models include a Dirichlet process mixture of normals for microarray data, a mixture of Beta distributions with a random number of terms for mass spectrometry data, and a Dirichlet process mixture of Poisson model for SAGE data. For the microarray data and the protein mass spectrometry data we consider experiments that compare two biologic conditions of interest. We assume that the aim of the experiment is to find genes and proteins, respectively, that

are differentially expressed under the two conditions. For the SAGE example, we propose data analysis for a single biologic sample.

Several aspects of data analysis for microarray and other high throughput gene and protein expression experiments give rise to mixture models. One important application of mixture models is for flexible modeling of sampling distributions. This is attractive, for example, when the number of genes on a microarray is the relevant sample size, thus allowing flexible semi-parametric representations. Such approaches are discussed, among others, in Broet et al. (2002), Dahl (2003) or Tadesse et al. (2005). The latter exploit the clustering implicitely defined by the mixture model to identify biologically interesting subclasses. Also, see Dahl (2006); Tadesse et al. (2006) in this volume. In this chapter we review three approaches that are typical examples of this literature. In Section 29.2 we discuss the use of Dirichlet process mixtures for model based inference about differential gene expression. In Section 29.3 we describe a mixture of Beta model for the mass/charge spectrum in MALDI-TOF mass spectrometry experiments. In Section 29.4 we introduce a semiparametric mixture of Poisson model for SAGE data.

Another important class of applications for mixture models in data analysis for high throughput gene expression data are finite mixtures, with each term in the mixture corresponding to a different condition of interest. A typical example is the model used in Parmigiani et al. (2002) who construct a sampling model for observed gene expression in microarray experiments as a mixture of three terms corresponding to normal, under- and over-expression. Newton et al. (2001) define a Gamma/Gamma hierarchical model with a mixture induced by an indicator for ties between two biologic conditions of interest. Kendziorsky et al. (2005) use mixtures for expression QTL mapping. See also Chen and Kendziorski (2006) in this volume. Kendziorsky et al. (2003) use finite mixtures to identify patterns of differential expression across multiple biologic conditions.

Naturally, the distinction between the two types of mixtures, i.e., flexible mixtures for an unknown sampling model versus mixtures of submodels with a biologically meaningful interpretation, is not strict. A typical example is the use of semi-parametric mixtures to define a probability model for clustering of genes or samples. Inference about clusters can often be interpreted as inference on biologically meaningful groups of genes or subpopulations corresponding to biologically distinct sub-types of a disease. From a modeling perspective, the intention of our distinc-

tion is to focus on semi-parametric mixture models with a random and, at least in spirit, unconstrained size mixture.

Also, approaches that use hierarchical models to define flexible sampling models could alternatively be considered as mixture models. Collapsing the hierarchical model by marginalizing with respect to some intermediate level parameters one can often rewrite the hierarchical model as a mixture. See, for example, Hein et al. (2005) or Hein et al. (2006) in this volume.

In this chapter we only focus on the use of semi-parametric mixtures to represent an unknown sampling model, i.e., applications of infinite size mixtures, and will not discuss the other type of mixture models.

29.2 A Non-parametric Bayesian model for Differential Gene Expression

We consider inference for microarray group comparsion experiments. Assume that the data has been summarized as a set of difference scores, z_i, $i = 1, \ldots, n$, for n genes. The difference score z_i could be, for example, a two-sample t-statistic for observed fluorescence intensities for gene i in samples under two biologic conditions of interest. See Efron et al. (2001) for a discussion of appropriate data pr-eproscesing and Baggerly et al. (2006), in this volume, for an explaination of the experimental setup and important issues in data analysis for such experiments. We assume that the set $i = 1, \ldots, n$, of genes is partitioned into a subset of differentially expressed genes and non-differentially expressed genes. Inference proceeds by assuming that for differentially expressed genes, the difference scores z_i arise by independent sampling from some unknown distribution f_1; for non-differentially expressed genes, z_i are independent samples from an unknown distribution f_0. For a reasonable choice of difference scores, the distribution f_0 should be a unimodal distribution centered at zero. The distribution f_1 should be a bimodal distribution with symmetric modes to the left and right of zero corresponding to over- and underexpressed genes. Figure 29.1 show possible histograms for observed difference scores generated from f_0 and f_1. Of course, the partition into differentially and non-differentially expressed genes is unknown. Thus, instead of samples from f_0 and f_1, we can only work with the sample z_i, $i = 1, \ldots, n$, generated from a mixture of f_0 and f_1. Let p_0 denote the unknown proporition of non-differentially genes. We

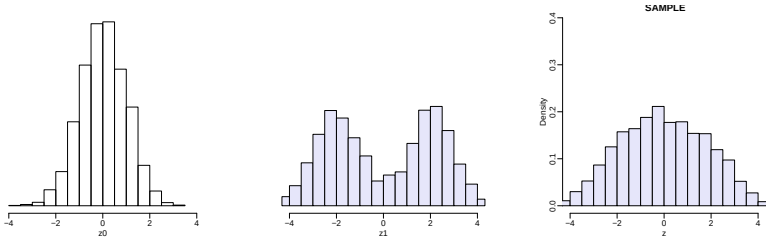


Fig. 29.1. Hypothetical distribution of difference scores for non-differentially expressed (left, f_0) and differentially expessed genes (center, f_1), and the observed mixture (right, f).

assume

$$z_i \overset{iid}{\sim} f(z) = p_0\, f_0(z) + (1-p_0)\, f_1(z), \quad i = 1, \ldots, n. \tag{29.1}$$

The main goal of inference in the two group comparison microarray experiment can be formally described as the deconvolution of (29.1). We introduce a latent indicator variables $r_i \in \{0, 1\}$ to rewrite (29.1) equivalently as a hierarchical model

$$\begin{aligned} p(z_i \mid r_i = j) &= f_j(z_i) \\ Pr(r_i = 0) &= p_0. \end{aligned} \tag{29.2}$$

The latent variable r_i can be interpreted as indicator for gene i being differentially expressed. Efron et al. (2001) propose cleverly chosen point estimates for p_0, f_0 and f_1 and report the implied inference for r_i. To develop the point estimate they introduce an additional set of difference scores, z_i, $i = n+1, \ldots, 2n$. The additional difference scores are generated using the same original data, but deliberately computing difference scores for samples under the same biologic conditions. Thus,

$$z_i \sim f_0(z_i), \quad i = n+1, \ldots, 2n,$$

for this additional null sample.

In Do et al. (2005) we propose a model-based semiparametric Bayesian approach to inference in this problem. We recognize f_0, f_1 and p_0 as unknown quantities and proceed by defining a suitable prior probability model. Probability models for unknown functions, including distributions such as f_0 and f_1 in this problem, are known as non-parametric Bayesian models. See, for example Müller and Quintana (2004) for a

recent review of non-parametric Bayesian inference. The term "non-parametric" is a misnomer, as the random functions are infinite dimensional parameters. However, the name is traditionally used because implied posterior inference closely resembles inference under classical non-parametric methods.

In chosing a prior probability model for f_0 and f_1 we face two competing aims. On one hand we wish to generalize traditional parametric models, like a normal sampling model. On the other hand we want to retain as much computational simplicity as possible. This leads us to use a mixture of normal model, with a non-parametric prior on the mixing measure. Inference under this model is almost as straightforward as under a simple normal model, yet, subject to some technical constraints, the mixture of normal model can approximate arbitrary sampling distributions. As probabilty model for the mixing measure we use a Dirichlet process (DP) prior (Ferguson, 1973; Antoniak, 1974). For reasons of computational simplicity and ease of interpretation, the DP prior is one of the most widely used non-parametric Bayes models. The DP model has two parameters, a base measure and a total mass parameter. We write $G \sim DP(G^\star, M)$ to indicate that G has a DP prior with a base measure $G^\star$ and total mass M. The base measure has the interpretation as mean measure, in fact $E(G) = G^\star$. The total mass parameter can be interpreted as a precision parameter. The larger M, the closer the random G will be to $G^\star$. Another important implication of the total mass parameter is mentioned below.

In summary, we assume the following model. Let $N(z;\ m, s)$ denote a normal distribution for the random variable z, with moments (m, s). We define a probability model for the random distributions f_0 and f_1 as:

$$\begin{aligned} f_j(z) &= \int N(z;\ \mu, \sigma)\ dG_j(\mu) \\ G_j &\sim DP(G_j^\star, M). \end{aligned} \tag{29.3}$$

One of the critical properties of the DP prior is that a DP generated random measure is almost surely discrete. Thus the integral in (29.3) is simply a sum over all point masses in G_j. The total mass parameter M determines the distribution of the weights attached to these point masses. Mixture models with respect to a mixting measure with DP prior, such as (29.3), are known as mixture of DP (MDP) models and are widely used in non-parametric Bayesian inference. See, for example, MacEachern and Müller (2000) for a review of such models.

We complete the model given by the likelihood (29.2) and prior (29.3) with a hyperprior on the base measures $G_j^\star$. We assume $G_0^\star = N(0, \tau^2)$ with a conjugate inverse Gamma hyperprior on τ^2, and $G_1^\star = \frac{1}{2}N(-b, \tau^2) + \frac{1}{2}N(b, \tau^2)$ with a conjugate normal hyperprior on b. Finally, we assume a Beta prior for p_0, $p_0 \sim Be(\alpha, \beta)$. The hyperparameters α, β, M are fixed.

Inference in the proposed model is implemented by Markov chain Monte Carlo (MCMC) simulation. See Do et al. (2005) for a detailed description of the posterior MCMC algorithm. A direct implication of the model (29.2) and (29.3) is that the marginal posterior probability of differential expression, $Pr(r_i = 1 \mid data)$, is the same for all genes with equal difference score z_i. Thus posterior inference can be summarized as a function $Pr(r_i = 1 \mid z_i = z, data)$. Starting with model (29.2), a straightforward use of Bayes theorem shows

$$Pr(r_i = 0 \mid z_i = z, f_0, f_1, p_0) = p_0\, f_0(z) / [\underbrace{p_0\, f_0(z) + (1 - p_0)\, f_1(z)}_{f(z)}].$$

Let $P_1 = p_0\, f_0 / f$. Then the posterior expectation $\overline{P}_1 = E(P_1 \mid data)$ is exactly the desired marginal posterior probability of differential expression, $\overline{P}_1 = Pr(r_i = 1 \mid z_i = z, data)$. Figure 29.2 shows posterior inference for a simulation experiment. The figure shows the simulation truth, the reported posterior mean curve $\overline{P}_1(z)$, and pointwise posterior credible intervals for $P_1(z)$. The curve $\overline{P}_1(z)$ allows one to readily read off the marginal posterior probability of differential expression for each gene. In contrast to reasonable but ad-hoc point estimates, the reported probabilities are interpreted as marginal probabilities in one coherent encompassing probability model. This leads to a straightforward definition, evaluation and control of false discovery rates. See Newton et al. (2004) or Do et al. (2005) for a discussion.

29.3 A Mixture of Beta Model for MALDI-TOF Data

Matrix assisted laser disorption – time of flight (MALDI-TOF) experiments allow the investigator to simultaneously measure abundance for a large number of proteins. Details of the experimental setup are described, for example, in Baggerly et al. (2003) or Baggerly et al. (2006), in this volume. Briefly, the biological sample for which we wish to determine protein abundance is fixed in a matrix. A laser beam is used to break free and ionize individual protein molecules. The experiment is

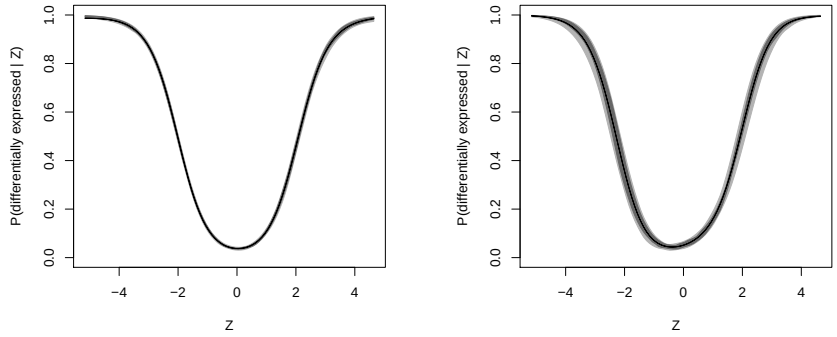


Fig. 29.2. $\overline{P}_1(z)$: posterior mean probability of differential expression as a function of the observed difference score z (solid black line). The left figure conditions on the full data, $z_i, i = 1, \ldots, 2n$, including the null data. The right figure does not make use of the null data, conditioning only on z_i, $i = 1, \ldots, n$. The dark grey shaded band shows the central 50% posterior density interval. Light grey shows a 75% posterior interval The dark and light grey shaded areas are very narrow and can hardly be distinguished from the posterior mean curve.

arranged such that ionized proteins are exposed to an electric field that accelerates molecules along a flight tube. On the other end of the flight tube molecules hit a detector that records a histogram of number of molecules that hit over time. Assumig that all ionized molecules carry a unit charge, the time of flight is deterministically related to the molecule mass. The histogram of detector events over time can therefore be be changed to a histogram of detector events over protein masses. Allowing for multiple charges, the mass scale is replaced by a scale of mass/charge ratios. The histogram of detector events is known as mass/charge spectrum. Figure 29.3 shows typical spectra.

Ideally, each protein that is present in the original probe should correspond to a peak in the spectrum. Because of the random initial velocities when proteins are ionized by the laser impact we would expect to see peaks rather than sharp lines even in an idealized experiment. Many additional artifacts of the experiment add to the idealized description, leading to an additional baseline that adds to the protein peaks. See the data shown in Figure 29.3.

Assume we observe spectra for experiments $k = 1, \ldots, K$. Let $y_k(m_i)$ denote the recorded count for sample k at mass/charge grid point m_i,

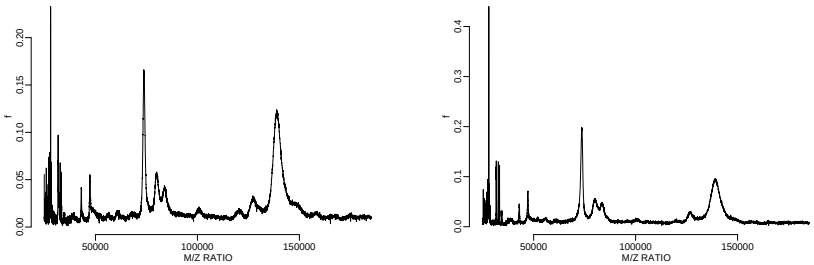


Fig. 29.3. Spectra for a normal samples (left) and a tumor samples (right), on grid of size $I = 60,000$.

and let $f_k(m_i)$ denote the assumed underlying cleaned spectra corresponding to detected proteins only. The desired inference about the unknown protein abundance in the original probes can be formalized as (i) removing noise and baseline from the observed spectra y_{ki} to impute f_k; (ii) finding peaks in f_k; and (iii) reporting the relative sizes of these peaks. The relative size of the peaks corresponds to the relative abundance of the corresponding protein in the probe. If samples are collected under different biologic conditions we need additional inference about different versus equal abundance of different proteins.

In Müller et al. (2006) we develop a non-parametric Bayes model to allow such inference. Based on the above stylized description of the experiment we consider y_k as the empirical histogram of detector events. We represent it as a mixture of a baseline B_k corresponding to detector noise, protein fragments, etc., and a cleaned spectrum f_k:

$$p_k(m) = p_{0k}\, B_k(m) + (1 - p_{0k})\, f_k(m).$$

The spectrum f_k is a sum of peaks, with each detected protein contributing a peak centered at its mass/charge value. The experimental arrangement implies a finite support for f_k. Motivated by nonparametric models for random distributions on a finite support developed in Petrone (1999) and Robert and Rousseau (2002) we use a mixture of Beta distributions to define the random distribution f_k. The location for each Beta kernel is interpreted as the mass/charge ratio of the protein giving rise to this peak. To facilitate later interpretation, we use a non-standard parametriztion of the Beta distribution. We write $Be(x;\, \epsilon, \alpha)$ for a Beta kernel for the random variable x, with mean and standard deviation ϵ and α (with appropriate constraints on α).

Let x denote the biologic condition of sample k. We assume a two group comparison, i.e., $x \in \{0, 1\}$. Then

$$f_k(m) = \sum_{j=1}^{J} w_{xj} \text{ Beta}(m;\ \epsilon_j, \alpha_j). \tag{29.4}$$

In words, the k-th spectrum is a mixture of Beta kernels, corresponding to J distinct proteins with mass/charge values ϵ_j. The relative weight w_{xj}, i.e., relative abundance of protein j, is assumed the same for all samples under the same biologic condition. For reasons of technical convenience we chose a similar mixture of Beta prior for the baseline B_k. Different hyperparameters reflect the fact that the baseline is much smoother than f_k and we expect fewer terms in the mixture. $B_k(m) = \sum_{j=1}^{J_k} v_{kj} \text{ Beta}(m;\ \eta_{kj}, \beta_{kj})$. The sizes of the mixtures are random. We use truncated Poisson priors for J and J_k, $k = 1, \ldots, K$. Baseline B_k and mean spectrum f_k are combined to define the distribution of mass/charge ratios $p_k = p_{0k}\, B_k + (1 - p_{0k})\, f_k$. Following the idealized description of the experimental setup, the sampling model is random sampling from p_k. Let $y_k = (y_{ki},\ i = 1, \ldots, I)$ denote the empirical spectrum for the k-th sample over the grid of mass/charge values. Typically I is large, say 60,000, defining a very fine grid. Let $\theta = (J, J_k, w_{xj}, v_{kj}, \epsilon_j, \alpha_j, \eta_i, \beta_i,\ x = 0, 1,\ j = 1, \ldots, J,\ k = 1, \ldots, K,\ i = 1, \ldots, J_k)$ denote the parameter vector. The likelihood is

$$\log p(y_k \mid \theta) = \sum_{i=1}^{I} y_{ki} \log p_k(m_i). \tag{29.5}$$

Instead of the random sampling model (29.5) many authors use a regression likelihood, assuming normal residuals, $y_{ki} \sim N(p_k(m_i), \sigma^2)$. Little changes in the following discussion if we were to replace (29.5) by this regression likelihood.

The model is completed with a prior for the Beta parameters and the weights. For the weights w_{xj} we use a hierarchical prior with indicators λ_j for ties

$$\lambda_j = I(w_{0j} = w_{1j}).$$

Posterior inference on the λ_j and the locations ϵ_j summarizes the desired inference on proteins that are differentially expressed across the two groups of samples.

Implementation of posterior inference requires MCMC over a varying dimension parameter space, as the dimension of the parameter space

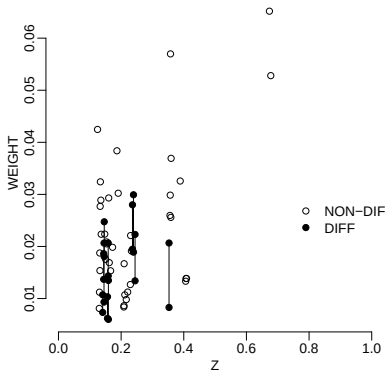


Fig. 29.4. Posterior mean abundance of detected proteins. All peaks with posterior probability of differential expression greater 50% are marked as solid dots, with a line combining $\overline{w}_{0j} = E(w_{0j} \mid data)$ and $\overline{w}_{1j}$. Mass/charge ratios on the horizontal axis are rescaled to the unit interval.

depends on the sizes J and J_k of the mixtures. We use reversible jump MCMC (RJMCMC) as proposed in Green (1995) and, specifically for mixture models, in Richardson and Green (1997). See Müller et al. (2005) for a detailed description of the MCMC algorithm.

A minor complication arises in reporting and summarizing posterior inference about distinct proteins and their mass/charge ratios. The mixture f_k only includes exchangeable indices j, leading to the complication that the Beta kernel corresponding to a specific protein might have different indices at different iterations of the posterior MCMC simulation. In other words, the protein identity is not part of the probability model. To report posterior inference on the mean abundance of a given protein requires additional post-processing to match Beta kernels that correspond to the same protein across iterations. We use a reasonable ad-hoc rule. Any two peaks j and h with a difference in masses below a certain threshold are counted as arising from the same protein. Specifically, we use the condition $|\epsilon_j - \epsilon_h| < 0.5\alpha_j$ to match peaks. Here j indexes the peak that was imputed in an earlier MCMC iteration than the peak h. The problem of reporting inference related to the terms in a mixture is known as the label switching problem (C. C. Holmes and Stephens, 2005).

Figure 29.4 summarizes estimated masses and abundance of detected proteins. Assuming that the main inference goal is to identify proteins

with differential expression across the two biologic conditions, we focus on inference about the indicator for differential expression, $1 - \lambda_j$. The figure indicates all protein masses with $Pr(\lambda_j = 0 \mid data, \ldots) > 50\%$, i.e., with posterior probability greater than 50% for differential expression. The probability is evaluated conditional on the protein being present in the probe (therefore the "..." in the conditioning set). Also, only proteins are reported in the figure that have posterior probability greater than 5% of being present, i.e., a peak being identified at the corresponding mass. In a data analysis, the list of reported protein masses would now be compared against a list of known protein masses to match the discovered peaks with specific proteins.

29.4 A Semi-Parametric Mixture Model for SAGE Data

Consider data from a SAGE (Serial Analysis of Gene Expression) experiment. See Baggerly et al. (2006) for a description of the experimental setup, and the nature of the data. We consider inference for data from one biologic sample. Let y_i, $i = 1, \ldots, k$, denote observed tag frequencies for k distinct transcripts. Let $n = \sum y_i$ denote the total number of recorded transcripts, and let π_i denote the unknown true abundance of the i-th transcript in the probe. For large y_i, the empirical frequency $\widehat{\pi}_i = y_i/n$ is an appropriate point estimate for π_i. The associated uncertainty, formalized as variance of the maximum likelihood estimator or as posterior standard deviation in a suitable model, is negligible. However, for scarce tags with small π_i, more elaborate estimates are required. The empirical frequency for scarce tages includes considerable sampling variability. Also, when the data includes samples across different biologic conditions, the inference goal might not be restricted to estimating the transcript frequencies. For discrimination and classification additional inference about differences in transcript frequencies, and related probability statements are required. In addition to inference on π_i for a specific tag i, one might be interested in the distribution of tag frequencies across different transcripts. This can be achieved by model-based posterior inference.

Morris et al. (2003) introduce an approach that is based on a hierarchical model with a mixture of two Dirichlet distributions as population distribution prior for the π_i. See Morris et al. (2006) in this volume for a review of this approach. Building on this model, we introduce a semi-parametric Bayesian mixture model, replacing the two-component

mixture of Dirichlet distributions by an unknown random measure, with a nonparametric Bayesian prior model.

For the following model construction it is convenient not to condition on n. In other words, instead of assuming that the set of observed counts arise as a multinomial sample with cell frequencies π_i, we assume that, conditional on hyperparameters, the counts y_i arise as independent samples from some distribution. Specifically, we assume that the counts y_i are sampled from a mixture of Poisson model. Let $\text{Poi}(x;\ \lambda)$ denote a Poisson distribution for the random variable x with parameter λ. We assume

$$y_i \sim \int \text{Poi}(y_i;\ \lambda)\ \text{d}G(\lambda),$$

$i = 1, \ldots, n$, independently conditional on G. We specify a prior distribution for the mixture model by assuming a nonparametric prior on the mixing measure, chosing a DP prior as in (29.3),

$$G \sim DP(G^\star, M). \tag{29.6}$$

The mixture model can alternatively be written as a hierarchical model

$$y_i \mid \lambda_i \sim \text{Poi}(\lambda_i) \text{ with } \lambda_i \sim G. \tag{29.7}$$

The discrete nature of the DP random measure G implies a positive probability for ties among the λ_i. We denote with L the number of distinct values.

A minor complication arises from the fact that $y_i = 0$ is not observed; it is censored. Let k_0 denote the number of tags with non-zero count, i.e., the number of tags recorded in a SAGE library as shown in Baggerly et al. (2006). One could augment the model to include inference on k, $k \geq k_0$. Alternatively, we follow Stollberg et al. (2000), and fix k by imputing a point estimate for the uknown number of unobserved tags, i.e., tags with $y_i = 0$.

Model (29.6) and (29.7) defines a DP mixture of Poisson distributions. Such models are popular choices for non-parametric Bayesian data analysis. See, for example, MacEachern and Müller (2000) for a review of such models, including implementation of posterior inference by MCMC simulation. Choosing the base measure $G^\star$ to be conjugate with the Poisson distribution we define a conjugate DP mixture, greatly facilitating the MCMC implementation. Let $\text{Ga}(x;\ \alpha, \beta)$ denote a Gamma distribution with mean α/β. We use

$$G^\star(\lambda) = Ga(x;\ \alpha, \beta),$$

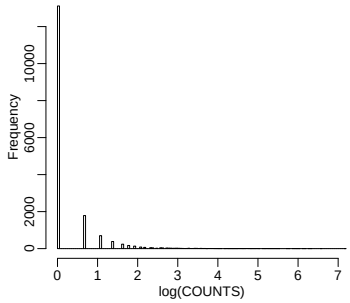


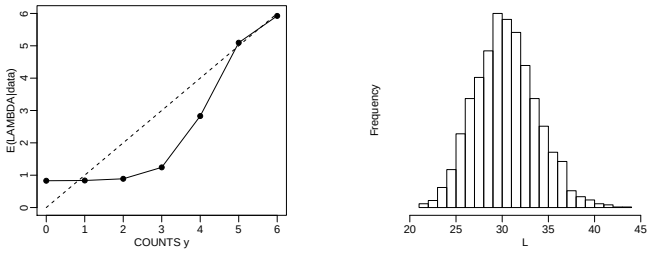
Fig. 29.5. Observed tag counts y_i. The highly skewed nature is typial for SAGE data.

with fixed hyperparameters α and β.

To illustrate the model we implemented posterior inference for a SAGE library reported in Zhang et al. (1997). The same data was used in Morris et al. (2003), and is available at
http://www.sagenet.org/SAGEData/NC1.htm. It records counts for $k_0 = 17703$ distinct transcripts, with a total number of $n = \sum y_i = 49610$ recorded tags. We use the estimte from Stollberg et al. (2000), and set $k = 25336$, with $y_i = 0$ for $i = k_0 + 1, \ldots, k$, i.e., we estimate the number of tags with censored counts $y_i = 0$, as $\sum I(y_i = 0) = 8072$. Figure 29.5a shows a histogram of observed counts y_i in the data. Figure 29.6 summarizes posterior inference for the transcript abundances. The figure plots posterior mean estimates $E(\lambda_i \mid data)$ versus observed counts y_i. The nature of the shrinkage follows patterns reported in Morris et al. (2003). For censored tags, with $y_i = 0$, the posterior mean estimate inflates the m.l.e. and reports $E(\lambda_i \mid data) \approx 0.9$. For rare tags with non-zero counts, posterior inference shrinks the maximum likelihood estimate. For abundant tags, posterior inference is driven only by the observed count, and $E(\lambda_i \mid data) \approx y_i$. Figure 29.7 shows the estimated distribution of tag abundances λ_i.

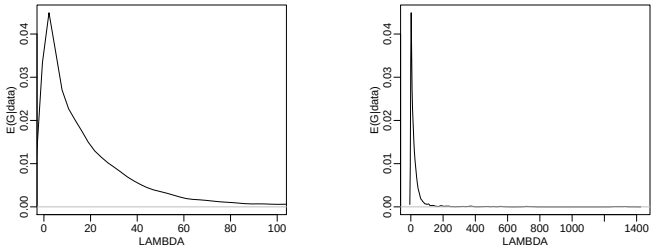
29.5 Summary

We have illustrated the use of mixture models for Bayesian inference with gene expression and proteomics data. We focused on the use of mixtures as a flexible class of distributions to parametrize random distributions.



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

Fig. 29.6. Posterior means $\overline{\lambda}_i = E(\lambda_i \mid data)$ versus observed counts y_i (panel a). Note the strong shrinkage for small counts. For large counts, $y_i > 6$, posterior shrinkage quickly becomes negligible. Posterior distribution for the number of clusters L (panel b).



(a) $E(G \mid data)$, $\lambda < 100$ (b) $E(G \mid data)$

Fig. 29.7. Estimated mixing measure G. The left panel zooms in on the lower end, $\lambda < 100$. The highly skewed nature of $G^\star = E(G \mid data)$ reflects the same feature in the recorded data y_i.

Another important use of mixtures arises in models where the submodels in the mixture correspond to different biologic conditions. Such models are extensively reviewed in other chapters in this volume.

We introduced DP mixtures of normals models to model microarray gene expression data, DP mixtures of Poissons to model tag counts in SAGE data, and location/scale mixtures of Beta kernels to represent mass/charge spectra in protein mass spectrometry experiments. The underlying theme in all three applications is the use of model-

based inference, with a probability model on the random distribution (or mass/charge spectrum). This is in contrast to traditional, and very resonable, multi-step methods. The power of the model-based methods lies in the full probabilistic description of all related uncertainties. Many important inference problems go beyond point estimates. For example, consider the decision problem of flagging genes for differential expression, or the problem of identifying a set of proteins that can serve as biomarker panel, or sample size choice for a microarray experiment. A decision theoretic answer to these question relies on a description of all uncertainties in one coherent probability model.

Acknowledgments

Jeff Morris was supported by NCI grant CA-107304. Michele Guindani and Peter Müller were supported by NCI grant CA75981.

Bibliography

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, 2, 1152–1174.

Baggerly, K. A., Coombes, K. R., and Morris, J. S. (2006), "Bayesian Inference for Gene Expression and Proteomics," in Do et al. (2006), chap. An Introduction to High-Throughput Bioinformatics Data, pp. xxx–xxx.

Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L. C., and Coombes, K. R. (2003), "A comprehensive approach to analysis of MALDI-TOF proteomics spectra from serum samples," *Proteomics*, 3, 1667–1672.

Broet, P., Richardson, S., and Radvanyi, F. (2002), "Bayesian hierarchical model for identifying changes in gene expression from microarray experiments," *J Comput Biol.*, 9, 671–83.

C. C. Holmes, A. J. and Stephens, D. A. (2005), "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling," *Statistical Science*, 20, 50–67.

Chen, M. and Kendziorski, C. (2006), "Bayesian Inference for Gene Expression and Proteomics," in Do et al. (2006), chap. Interval mapping for expression quantitative trait loci, pp. xxx–xxx.

Dahl, D. (2003), "Modeling differential gene expression using a Dirichlet Process mixture model," in *2003 Proceedings of the American Statistical Association, Bayesian Statistical Sciences Section*, Alexandria, VA: American Statistical Association.

— (2006), "Bayesian Inference for Gene Expression and Proteomics," in Do et al. (2006), chap. Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model, pp. xxx–xxx.

Do, K., Müller, P., and Tang, F. (2005), "A Bayesian mixture model for

differential gene expression," *Journal of the Royal Statistical Society C*, 54, 627–644.

Do, K.-A., Müller, P., and Vannucci, M. (eds.) (2006), *Bayesian Inference for Gene Expression and Proteomics*, Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Hein, A., Richardson, S., Causton, H., Ambler, G., and Green, P. (2005), "BGX: A fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data," *Biostatistics*, 6, 349–373.

Hein, A.-M. K., Lewin, A., and Richardson, S. (2006), "Bayesian Inference for Gene Expression and Proteomics," in Do et al. (2006), chap. Bayesian hierarchical models for inference in microarray data, pp. xxx–xxx.

Kendziorsky, C. M., Chen, M., Yuan, M., Lan, H., and Attie, A. (2005), "Statistical methods for expression trait loci (eQTL) mapping," *Biometrics*, to appear.

Kendziorsky, C. M., Newton, M., Lan, H., and Gould, M. (2003), "On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles," *Statistics in Medicine*, 22, 3899–3914.

MacEachern, S. N. and Müller, P. (2000), "Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models," in *Robust Bayesian Analysis*, eds. Ruggeri, F. and Ríos-Insua, D., New York:Springer-Verlag, pp. 295–316.

Morris, J. S., Baggerly, K. A., and Coombes, K. R. (2003), "Bayesian Shrinkage Estimation of the Relative Abundance of MRNA Transcripts Using SAGE," *Biometrics*, 59, 476–486.

— (2006), "Bayesian Inference for Gene Expression and Proteomics," in Do et al. (2006), chap. Shrinkage Estimation for SAGE Data using a Mixture Dirichlet Prior, pp. 1–37.

Müller, P., Do, K.-A., Bandyopadhyay, R., and Baggerly, K. (2006), "A Bayesian Mixture Model for Protein Biomarker Discovery," Tech. rep., M.D. Anderson Cancer Center.

Müller, P. and Quintana, F. A. (2004), "Nonparametric Bayesian Data Analysis," *Statistical Science*, 19, 95–110.

Newton, M., Noueriry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting differential gene expression with a semiparametric heirarchical mixture model," *Biostatistics*, 5, 155–176.

Newton, M. A., Kendziorsky, C. M., Richmond, C. S., R., B. F., and Tsui, K. W. (2001), "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data," *Journal Computational Biology*, 8, 37–52.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002), "A statistical framework for expression-based molecular classification in cancer," *Journal of the Royal Statistical Society B*, 64, 717–736.

Petrone, S. (1999), "Bayesian density estimation using Bernstein polynomials," *Canadian Journal of Statistics*, 27, 105–126.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society B*, 59, 731–792.

Robert, C. and Rousseau, J. (2002), "A mixture approach to Bayesian goodness of fit," Tech. rep., CEREMADE.

Stollberg, J. Urschitz, J., Urban, Z., and Boyd, C. (2000), "A quantitative evaluation of Sage," *Genome Research*, 10, 1241–1248.

Tadesse, M., Sha, N., Kim, S., and Vannucci, M. (2006), "Bayesian Inference for Gene Expression and Proteomics," in Do et al. (2006), chap. Identification of Biomarkers in Classification and Clustering of High-Throughput Data, pp. 1–91.

Tadesse, M., Sha, N., and Vannucci, M. (2005), "Bayesian variable selection in clustering high-dimensional data," *Journal of the American Statistical Association*, 100, 602–617.

Zhang, L., Zhou, W., Velculescu, V., Kern, S., Hruban, R., Hamilton, S., Vogelstein, B., and Kinzler, K. (1997), "Gene Expression Profiles in Normal and Cancer Cells," *Science*, 276, 1268–1272.