Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2007 *Paper* 18

A Bayesian hierarchical model for spot fluorescence in microarrays

Federico Mattia Stefanini*

*University of Florence, stefanini@ds.unifi.it

http://biostats.bepress.com/cobra/art18

A Bayesian hierarchical model for spot fluorescence in microarrays

Federico Mattia Stefanini

Abstract

Microarray experiments are characterized by the presence of many sources of experimental bias and a remarkably large technical variability. The assessment of differential expression for genes transcribed into a small number of mRNA copies heavily depends on the proper quantification of background fluorescence within spot. The rough model 'observed = hybridization plus background' fluorescence is at first reformulated at spot level, then it is embedded into a Bayesian hierarchical model suited for fitting control spots. The novelties of the approach include the background correction performed on the latent mean of replicated spots, and an explicit model for outlying observations at low fluorescence values in which the probability of occurrence and their magnitude depend on the background fluorescence intensity. The analysis of unpublished data from a maize ear tissues experiment confirms the feasibility of MCMC inferences as regard the computational burden.

A Bayesian hierarchical model for spot fluorescence in microarrays

Federico M. Stefanini[1]
Department of Statistics 'G.Parenti'
University of Florence, Italy

March 27, 2007

[1]viale Morgagni 59, 50134 Firenze, Italy, stefanin@ds.unifi.it

Abstract

Microarray experiments are characterized by the presence of many sources of experimental bias and a remarkably large technical variability. The assessment of differential expression for genes transcribed into a small number of mRNA copies heavily depends on the proper quantification of background fluorescence within spot. The rough model 'observed = hybridization plus background' fluorescence is at first reformulated at spot level, then it is embedded into a Bayesian hierarchical model suited for fitting control spots. The novelties of the approach include the background correction performed on the latent mean of replicated spots, and an explicit model for outlying observations at low fluorescence values in which the probability of occurrence and their magnitude depend on the background fluorescence intensity. The analysis of unpublished data from a maize ear tissues experiment confirms the feasibility of MCMC inferences as regard the computational burden.

KEYWORDS: Background fluorescence, Control Spots, MCMC

1 Introduction

A spotted microarray is a coated glass slide on which thousands of different DNA sequences (probes) are printed as spots located on a (quite) regular lattice. A spot is a small region of the slide in which a huge number of copies of just one short DNA sequence is covalently linked to the surface. The key-point of the microarray technology is in the relationship existing between the cellular concentration of a given mRNA sequence in target samples and the foreground fluorescence intensity read on the corresponding spot after hybridization. A comprehensive account of this technology is provided by Nguyen, Arpat, Wang and Carroll (2002).

The critical role played by the measurement process in the microarray assessment of differential expression has been clear since the birth of this technology. The reliability of information has been investigated by Schuchhardt, Beule, Malik, Wolski, Eickhoff, Lehrach and Herzel (2000), while the development and characterization of control spots has been performed by Eickhoff, Korn, Schick, Poustka and van der Bosch (1999) and Thellin, Zorzi, Lakaye, De Borman, Coumans, Hennen, Grisar, Igout and Heinen (1999) to improve lab protocols. Raw fluorescence values are not suited for a straightforward assessment, for example, due to the presence of background fluorescence. In a simple normalizing transformation, spot's hybridization-specific fluorescence (FHSF) is estimated by the difference between spot foreground and spot background fluorescence read around such spot (see next section). Nevertheless several authors have found that those estimates of FHSF are prone to be negative for weakly expressed genes. It is very important to provide a sound estimate of background fluorescence within spot because the plain subtraction of out-of-spot background from within spot foreground may introduce further bias in the ratio estimator, with the overall amplification of noise and increase of fake signals of differential expression.

In this work a Bayesian hierarchical model is proposed to specifically address the relation foreground-to-background within spot. Model fitting needs control spots which must be replicated on each array, such as buffer (unprinted) spots and negative (unhybridized) controls. Blank spots and the background do not contain spotted DNA sequences, therefore the fluorescence is not due to sequence-specific hybridization. The spot of a negative control contains printed DNA but no sequence-specific hybridization is possible (at least in perfect experiments). Positive controls are expected to hybridize in the same amount across target samples, thus they provide information useful in calibration steps.

The novelties of the approach include the background correction per-

formed on the latent mean of replicated spots, and an explicit model for outlying observations at low fluorescence values in which the probability of occurrence and their magnitude depend on the background fluorescence intensity. Mixtures account for over-dispersion at two levels in the model hierarchy: at spot level (through scale parameters) and at replicates level (location parameter for replicated spots). Control spots from unpublished data are processed according to the proposed model. Some model parameters are related to the quality of data and they may be useful to improve the lab protocols.

In the first section of this paper, a formal description of the background correction problem is introduced and the spot-level model is motivated. Then, a Bayesian hierarchical model is developed and applied to a case study dealing with *Zea mais L.* ear tissues. The final section include issues to be addressed in future research.

2 Methods

Let y, x be, respectively, the foreground and the out-of-spot background fluorescence intensities of a given spot for dye c (letter c omitted in this section). The *log* scale for is adopted for convenience.

It is often assumed that $y = s_f + x$, where s_f is FHSF intensity, a latent variable. Then, the natural estimate of s_f is $y - x$, where x is the realization of a random variable whose variability may be quite large, thus negative estimates may occur.

An extension of the simple model above is obtained by introducing a latent variable for the fluorescence of the background within spot's area, S_b, and a latent variable for the FHSF intensity, S_f. Within a spot we have the following relation among random variables:

$$Y = S_f + S_b \tag{1}$$

The realized spot latent background s_b may differ from the value x observed outside the spot. It is reasonable to assume that the expected value of S_b given x is a monotonic non-decreasing function of x, at least if the array coating and the related chemistry during the experiment are smooth with respect to the spot spatial scale. We assume here that:

$$S_b = \beta_0 + \beta_1\ x + \phi, \tag{2}$$

where ϕ is a symmetrically distributed zero-mean random error.

Equation (2) leads to realized values $s_b = x$ for special values of β_0, β_1, ϕ, but it may be further motivated. There is no natural definition of background area outside a spot and while a large area increases the number of pixels entering into the estimate of background fluorescence, it may suffer the spatial heterogeneity of the array. Moreover, one does not typically know how the algorithm in the scanner equipment filters the raw background value, nor how robust the filtering is with respect to the set of possible experimental protocols, hence the β_1 parameter. Furthermore, the parameter β_0, besides allowing better fit to data, may also be partially interpreted as a tuning constant, especially for $\beta_0 < 0$, under the hypothesis that DNA molecules partially mask the background fluorescence.

In equation (2), the distribution of ϕ is critically assumed to be symmetrically centered on zero. It is widely recognized that artifacts are an intrinsic feature of microarray experiments. Salt precipitation, partial dehydration of the array and simple dust, for instance, may cause extremely large noise fluctuations. Equation (2) may account for the presence of outlying errors by setting $\phi = \alpha + \epsilon$, with ϵ a symmetrically distributed error centered on zero and with α a positive quantity. Let Z be a random variable indicating the presence ($Z = 2$) or the absence ($Z = 1$) of artifacts. Equation (2) holds given $Z = 1$ with $\phi = \epsilon$, but given $Z = 2$ equation (2) becomes:

$$S_b = \beta_0 + \beta_1 \, X + \alpha + \epsilon, \tag{3}$$

with α a random effect due to spatially dependent artifacts and/or surface irregularity. In Figure (1), a graph summarizes the relationship among observable and latent random variables.

Model parameters are estimated using the information coming from control spots: housekeeping genes, negative controls and blank spots. If a spot has been printed with just a buffer solution instead of DNA probes, like in unprinted-blank spots, then $S_f = 0$ by definition. If a spot contains DNA probes known for the impossibility to hybridize, like in negative controls, still $S_f = 0$. If a spot contains DNA probes expected to hybridize in equal amount between the two target samples, like in positive controls (called housekeeping genes), then S_f has to be estimated for each dye and the two values are expected to be equal if measured without sources of bias and without experimental variability. Finally, in regular printed spots, those under testing, the DNA hybridization may or may not occur for each dye: for instance, S_f may be large for one dye and zero for the other dye.

It is worth noticing that the comparison between blank spots and negative controls partially provides evidence about the masking ability of DNA molecules which might reduce background fluorescence.

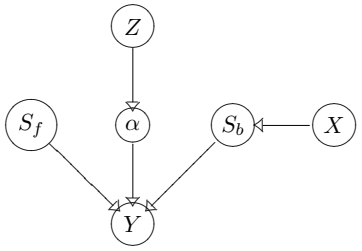

Figure 1: The graph relates the following random variable: the hybridization-specific fluorescence S_f, the observable background X, the latent S_b background, and the observable foreground Y_f. The latent allocator of outlying noise Z and its magnitude α are also included in the representation.

2.1 A Bayesian hierarchical model

The random variable $Y_{c,l,k}$, $c = 1, 2$ $k = 1, 2, 3$, refers to the *log* fluorescence intensity of the spot located in position $k = 1, 2, 3$ of the triple l and target sample marked by dye c (probe $l = 1$ in the top left corner of the array; the last triplet being in the bottom right location). The background fluorescence $x_{c,l,k}$ is quantified as the average of pixels in a ring surrounding the spot l, k. Averaged values are indicated by an overlying bar, for example $\bar{y}_{c,l} = \sum_{k=1}^{3} y_{c,l,k}/3$.

The hierarchical model is made by a likelihood function and a prior distribution which is decomposed by defining a hierarchy of hyper-prior distributions (Figure 2), thus a general comment on the definition of hyperpriors is mandatory. Prior distributions of model parameters have been specified (elicited) taking into account the role played by each parameter in the model, the known constraints and further experimental results taken from 3 other arrays printed in the same batch and hybridized in the same run. As regards constraints, the *log*-transformed fluorescence intensities are bounded due to the fixed number of bits in coding the TIFF image of each channel-dye (typically 16 bits). Random samples were drawn from prior distributions and summarized by graphical display with the aim of better calibrating the selected priors with respect to prior beliefs. The elicitation of the likelihood function has been performed according to an exploratory data analysis on location residuals performed on the other 3 arrays mentioned above. Using the Bayesian Information Criterium (BIC), a mixture

of normal distributions in which the components do not depend on dyes was selected as likelihood function.

The main features of the proposed Bayesian hierarchical model for blank spots are shown in Figure 2, and its qualitative interpretation may be summarized in plain words. As regards notation, double circles indicate observed or deterministic quantities, like theta parameters of hyper-prior distribution. Moreover the so called plates, also described by Spiegelhalter, Thomas, Best and Gilks (1996), are not represented.

In Figure 2, bottom to top, the value of the allocator variable $T_{c,l,k}$ defines the magnitude of the precision τ_0 according to hyperparameters θ_8, θ_9. Node π_0 sets the probability of observing a fluorescence intensity from the mixture component of small precision. Background fluorescence acts at a higher level in the hierarchy, i.e. on the color-location mean $\mu_{c,l}$ through regression parameters $\beta_{c,0}, \beta_{c,1}$ for the mean background $\bar{x}_{c,l}$. The subgraph defined by $\alpha_{c,l}$ and its graph ancestors explains the departure from $\mu_{c,l}$: the allocator variable $Z_{c,l}$ classifies the departure from the straight line as a regular error ($\alpha_{c,l} = 0$), or as an outlying error ($\alpha_{c,l} > 0$). The distribution of an outlying error depends on $\sigma_{c,l}$, a scale parameter which has a distribution conditional on the value taken by background and on parameters δ_1 and δ_2: the increase of $\bar{x}_{c,l}$ causes more concentration of $\mu_{c,l}$ on values close to the straight line. The probability $\pi_{c,l}$ of an outlying error also depends on the background $\bar{x}_{c,l}$: as the mean background increases, $\pi_{c,l}$ decreases for given values of γ_1 and γ_2.

At a more technical level, each component in the model hierarchy must be specified. Let $\underline{y}$ be the vector of observed foreground fluorescence intensities, $\underline{y} = \{y_{c,l,k} : \forall(c,l,k)\}$, and $\underline{\mu}$ the correspondent vector of color-location means. Then the likelihood function is:

$$p(\underline{y} \mid \underline{\mu}, \tau_0) = \prod_{c,l,k} N(y_{c,l,k} \mid \mu_{c,l}, \tau_0), \tag{4}$$

where τ_0 is the precision parameter of the normal distributions.

Given the allocator variable $T_{c,l,k}$ defined in the set $\{0, 1\}$, the conditional distribution of the precision τ_0 (Figure 2) is:

$$(\tau_0 \mid T_{c,l,k} = 0, \theta_8 = (1, 0.01)) \sim Gamma(\tau_0 \mid 1, 0.01) \tag{5}$$

$$(\tau_0 \mid T_{c,l,k} = 1, \theta_9 = (1, 0.5)) \sim Gamma(\tau_0 \mid 1, 0.5). \tag{6}$$

The allocator variable $T_{c,l,k}$ is distributed as a Bernoulli random variable:

$$(T_{c,l,k} \mid \pi_0) \sim \pi_0^{t_{c,l,k}} + (1 - \pi_0)^{1 - t_{c,l,k}}. \tag{7}$$

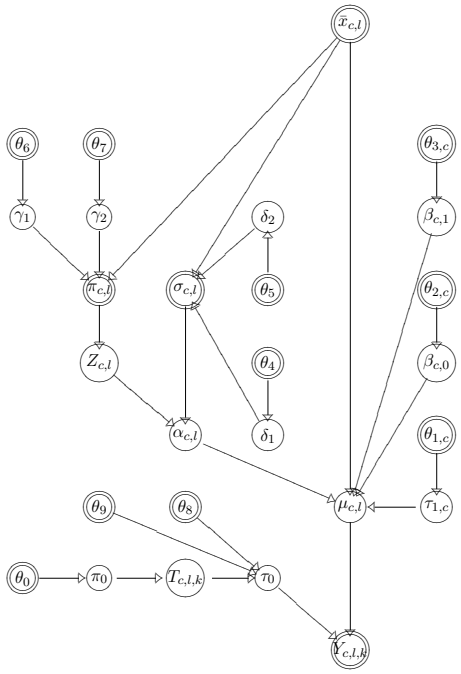

Figure 2: Graph relating the main hierarchical components of the Bayesian model. At the likelihood level, replicates are conditionally independent given their means and mixture parameters. The relation with the background acts at a higher level in the hierarchy, i.e. on the color-location mean. The double circles indicate observed or deterministic quantities.

The prior distribution of parameter π_0 has been defined as:

$$(\pi_0 \mid \theta_0 = (8, 2)) \sim Beta(8, 2). \tag{8}$$

The conditional distribution of the location mean $\mu_{c,l}$ resembles a simple linear regression with normal errors but it includes an extra term $\alpha_{c,l}$ which is interpreted like a shift in the intercept:

$$(\mu_{c,l} \mid \bar{x}_{c,l}, \beta_{c,0}, \beta_{c,1}, \alpha_{c,l}, \tau_{1,c}) \sim N(\beta_{c,0} + \alpha_{c,l} + \beta_{c,1}\bar{x}_{c,l}, \tau_{1,c}), \tag{9}$$

with α eventually equal to zero.

The hyper-parameters related to the location of $\mu_{c,l}$ are defined as:

$$(\beta_{c,0} \mid \theta_{2,c} = (0, 0.01)) \sim N(0, 0.01) \tag{10}$$

$$(\beta_{c,1} \mid \theta_{3,c} = (1, 0.01)) \sim N(1, 0.01) \tag{11}$$

$$(Z_{c,l} \mid \pi_{c,l}) \sim \pi_{c,l}^{2-z_{c,l}} + (1 - \pi_{c,l})^{z_{c,l}-1} \tag{12}$$

$$(\alpha_{c,l} \mid Z_{c,l} = 2, \sigma_{c,l}) \sim Gamma(3, \sigma_{c,l}) \tag{13}$$

$$(\alpha_{c,l} \mid Z_{c,l} = 1) = I_{\{0\}}(\alpha) \tag{14}$$

$$\sigma_{c,l} = \frac{exp(\delta_1 + \delta_2 \bar{x}_{c,l})}{1 + exp(\delta_1 + \delta_2 \bar{x}_{c,l})} \tag{15}$$

$$\pi_{c,l} = \frac{exp(\gamma_1 + \gamma_2 \bar{x}_{c,l})}{1 + exp(\gamma_1 + \gamma_2 \bar{x}_{c,l})}. \tag{16}$$

We set prior distributions of the hyper-parameters in Equations (15) and (16) on the *log* scale:

$$(log(\delta_1) \mid \theta_4 = (1, 4)) \sim N(1, 4) \tag{17}$$

$$(log(\delta_2) \mid \theta_5 = (1, 4)) \sim N(1, 4) \tag{18}$$

$$(log(\gamma_1) \mid \theta_6 = (1.751, 2.367)) \sim N(1.751, 2.367) \tag{19}$$

$$(log(\gamma_2) \mid \theta_7 = (1.751, 2.367)) \sim N(1.751, 2.367) \tag{20}$$

As regards the scale of $\mu_{c,l}$, we set the following conditional distribution:

$$(\tau_{1,c} \mid \theta_{1,c} = (5, 1.26)) \sim Gamma(5, 1.26) \tag{21}$$

Equation (9) has been modified to extend the model to housekeeping genes: S_f is the latent variable related to the FHSF of Equation (1), but averaged over replicates. The conditional expectation becomes:

$$E[\mu_{c,l} \mid \bar{x}_{c,l}, \beta_{c,0}, \beta_{c,1}, \alpha_{c,l}, s_{f,c,l}] = \beta_{c,0} + \alpha_{c,l} + \beta_{c,1}\bar{x}_{c,l} + s_{f,c,l} \tag{22}$$

with the precision $\tau_{1,c}$ of $\mu_{c,l}$ unchanged.

The prior distribution of $S_{f,c,l}$ has been defined as:

$$S_{f,c,l} \sim N(0, 0.25), \tag{23}$$

thus it is weakly informative about the FHSF intensity. Note that there is the possibility of strengthening the prior distribution. By definition of FHSF, the prior distribution of S_f should have a support defined in the set of positive numbers.

As regards negative controls, the model developed for blanks still holds if no masking ability is assumed for unhybridized DNA. The model modified through (22) is also suited to fit negative controls with masking behavior: the latent variable $S_{f,c,l}$ is substituted by $\Delta_{c,l}$ to represent the DNA masking

effect ($\Delta_{c,l} < 0$) and/or the presence of unspecific DNA hybridization ($\Delta_{c,l} > 0$). The prior distribution of $\Delta_{c,l}$ is also defined as weakly informative:

$$\Delta_{c,l} \sim N(0, 0.25). \tag{24}$$

Note that $\Delta_{c,l}$ is not introduced into the model for housekeeping genes, because of identification problems, thus it is assumed to be zero.

3 Results

3.1 A case study

The experiment deals with cereal ear tissues collected from plants of two genotypes. We consider only control spots, thus further protocol details will be omitted. The array layout is 180 rows times 90 columns. The printing head is made by 4 times 2 tips, therefore the spots are grouped in 8 subarrays (grids) of size 45×45 spots. The array is designed with 3 replicates for each probe, and they are located in adjacent column positions of the same row. Detailed explanations about the array manufacturing can be found at the URL address http://www.zmdb.iastate.edu/ on internet, array batch number 605.03.

In this study, we consider 15 of the 86 controls printed on the array (Table 1). They are located in the top two and the bottom two rows of each subarray, therefore blocks of controls are separated by blocks of biologically relevant probes. Sample sizes of blank spots are 1164 (582 spots times 2 channels) on the array, while the sample size of controls is 90 (45 spots times 2 channels).

Raw data were checked for spot quality and only well-shaped spots were considered in the analysis.

3.2 Descriptive summaries and model assumptions

Scattergrams in Figure 3 show the relationship between background fluorescence and foreground in blank spots given dye. There is a common pattern of association in the two diagrams: a straight line represents a good fit for most of the observations. A few large departures from this line are also evident, especially at small background values. Although the visual impression depends on the marginal distribution of the background, such behavior agrees with the proposed model.

In order to check the model assumptions in 4, residuals were calculated on each spot as $r_{c,l,k} = y_{c,l,k} - \bar{y}_{c,l}, l = 1, \ldots, L$, because the quantity $r_{c,l,k}$

Table 1: List of 15 controls considered in the case study on *Zea mais L.* ear tissues. The full list of controls may be found at maizedb web site (http://www.maizedb.com). Note that the number in square brackets on the left of each name is the value taken by index g. Index $g = 13$ indicates blank (unprinted) spots. EST $g = 1$ to $g = 12$ are housekeeping genes. EST $g = 14, 15$ are negative controls.

[1] Tubulin Alpha 1/2 chain	[2] Tubulin Alpha 1/2/3 chain
[3] Tubulin Alpha 3 chain	[4] GENE ubiquitin2 (skuqbgii cDNA)
[5] Histone H2A homolog	[6] Histone H1 homolog
[7] MNBB DNA binding protein	[8] Histone H4 homolog
[9] Elongation Factor 1A	[10] GENE bronze1 (bz1−ex2(5'))
[11] Rubisco Subunit subB	[12] Arath act 1/7 actin 11
[13] Blank	[14] B. thuringiensis
[15] cry1AC Myosin heavy chain	

is a source of information about the departure of spot replicates from their location mean. In Figure 4, the histogram (top left) and the boxplot (top right) of $r_{c,l,k}$ values confirms that residuals approximately follow a normal distribution, but with some extreme values on the tails, a feature recognized by the model through scale mixture at likelihood level. In Figure 4, $r_{c,l,k}$ residuals are plotted against the background (bottom left) and against location average (bottom right) without revealing unexpected patterns. No appreciable difference is obtained by conditioning the two scattergrams with respect to dye (results not shown). The pattern of variability shown in Figure 3 is therefore widely explained by the presence of noise acting on a spatial scale covering several spots.

The structure of noise found in Figure 3 does not disappear by averaging over location. In Figure 5, the scattergram of location means is shown for each dye. The pattern of dependence of $\overline{y}_{c,l}$ on $\overline{x}_{c,l}$ as well as the cloud of widely dispersed points for small background values are still apparent. The presence of a region of exclusion in which a large background value is never associated to a large spot fluorescence value is a feature captured by the hierarchical model.

Scattergrams of other control spots show patterns close to those found for blanks (results not shown): for example, the regression line of positive controls is shifted upwards due to the FHSF within the spot, and a cloud of widely dispersed values with small background is clearly appreciable.

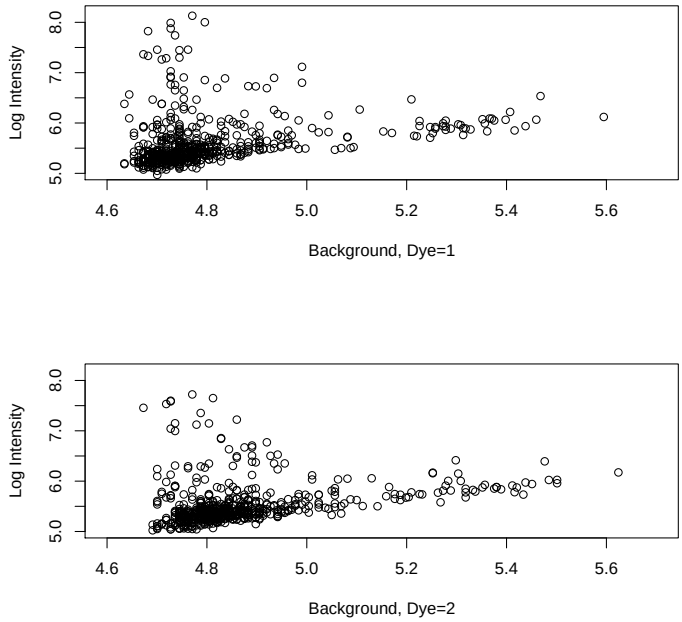

Figure 3: Scattergram of blank spots, log-fluorescence intensity against background given each dye.

3.3 Model fitting and output diagnostic for blank spots

The simulation software BUGS (Spiegelhalter et al. 1996) ran for $1.5 \cdot 10^5$ updates in 20 minutes on Pentium 2.2 GHz. The first $5 \cdot 10^4$ were discarded as burn-in. The last $1 \cdot 10^5$ steps originated $1 \cdot 10^4$ draws from the posterior distribution after thinning by 10 steps.

Initial values were selected by defining sets of plausible values for each parameter, by considering natural bounds due to the 16-bit coded fluorescence intensity and from some descriptive statistics calculated on other arrays.

The output of the MCMC simulation has been analyzed using the CODA suite of functions for output diagnostic (Plummer, Best, Cowles and Vines 2006) implemented in R (R Development Core Team 2005).

Autocorrelation of parameters listed in Table (2) is always well below 0.2 at lag = 1 and the last significant lag is not greater than 2 anyway. The inspection of the time series trace at full scale and for chunks of 300 steps

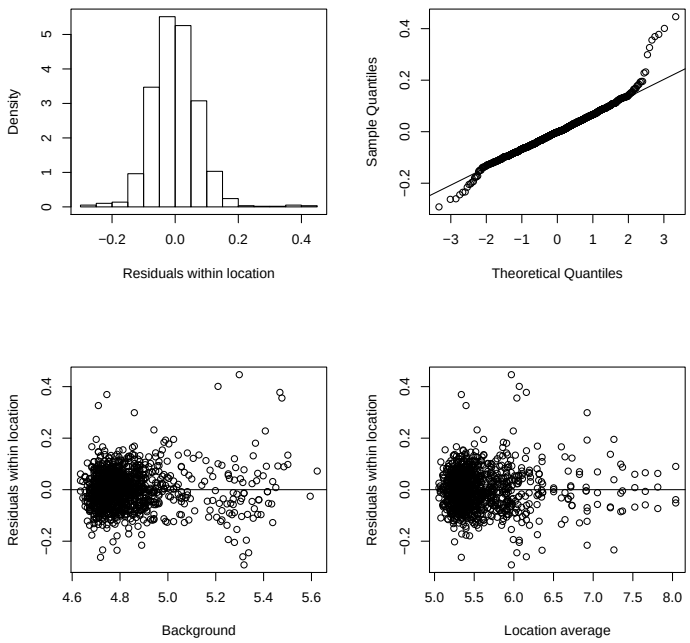

Figure 4: Histogram (top left) and quantile-quantile plot (top right) of log-fluorescence residuals within each location for blank spots. Scattergrams of residuals within location against background values (bottom left) and against location averages (bottom right) given dye suggest the absence of relevant patterns.

did not show evidences of bad mixing after thinning by 10 for all model parameters.

Statistical tests performed on simulation output included those proposed by Geweke (1996) , Heidelberg Heidelberger and Welch (1983) , and Raftery and Lewis (1992). Given the difficulty in checking all the unknowns, especially those involving latent-allocator variables, we checked out a sample of them and we found no evidence supporting the lack of convergence of the chain. We considered the chain as converged.

The approximated marginal distribution of some parameters has been summarized by the mean and some quantiles (Table 2).

In Figure 6, a density estimate of some marginal distributions is shown.

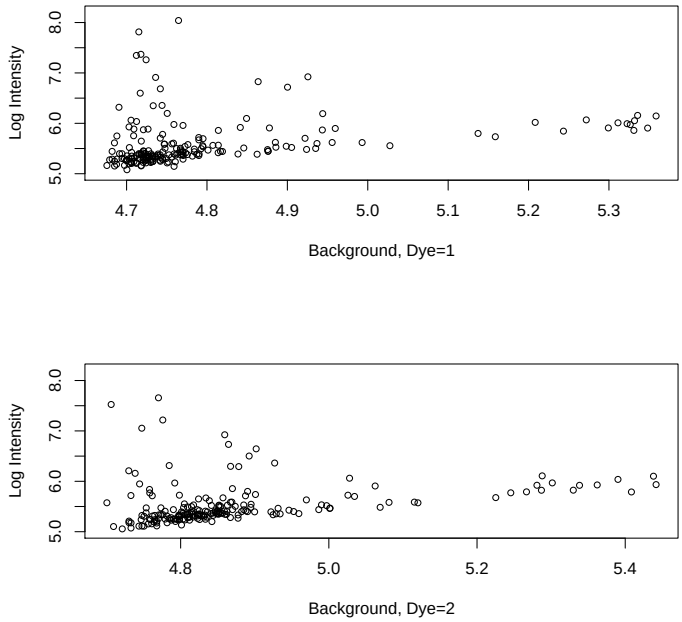

Figure 5: Scattergram of blank spots, log-fluorescence averages over location against averaged log-background values by dye.

Note that the distributions of $\beta_{c,0}$ and $\beta_{c,1}$ related to different dyes for the regression parameters are quite well separated.

Interesting indices of protocol quality are based on parameters included in our model. A point estimate of the probability of a 'large error' (namely the marginal posterior probability $1 - \widehat{\pi}_{c,l} = P[\alpha_{c,l} > 0 \mid x, data]$ when the background has value x) is obtained by mapping x through the inverse logit whose parameters γ_1 and γ_2 are set to the marginal median of the posterior distribution. For example, if $x = 4.7$ than the estimated probability is about 0.16.

3.4 Fitting a model to non-blank controls

The model for non-blank controls closely follows what described for blanks.

The MCMC run in BUGS took 64 minutes on a Pentium 2.2 Ghz with the following features: random sampling (BUGS algorithm) of initial values

Table 2: Summary of some approximated marginal posterior distributions given values of blank spots and their background (sample size $1 \cdot 10^4$ after thinning by ten). From left to right, columns are: parameter name, posterior mean, standard deviation, standard error of the mean (time series estimate), some quantiles. Here $\tau_{0,0}$ stands for $\tau_0 \mid T = 0$ and $\tau_{0,1}$ stands for $\tau_0 \mid T = 1$.

	Mean	SD	T.S. SE	2.5%	25%	50%	75%	97.5%
π_0	0.849	0.035	< 0.001	0.770	0.827	0.852	0.874	0.908
$\beta_{1,0}$	5.478	0.014	< 0.001	5.450	5.469	5.478	5.488	5.506
$\beta_{2,0}$	5.377	0.015	< 0.001	5.349	5.368	5.378	5.387	5.406
$\beta_{1,1}$	1.126	0.087	0.001	0.950	1.066	1.127	1.185	1.296
$\beta_{2,1}$	1.081	0.090	0.001	0.907	1.020	1.081	1.142	1.258
δ_1	2.995	0.260	0.003	2.515	2.817	2.987	3.168	3.516
δ_2	2.851	1.278	0.014	0.998	1.930	2.630	3.540	5.963
γ_1	2.109	0.186	0.002	1.763	1.981	2.102	2.228	2.496
γ_2	4.622	1.554	0.018	1.996	3.495	4.494	5.585	8.042
$\tau_{0,0}$	183.475	12.377	0.136	160.398	175.100	183.000	191.400	209.100
$\tau_{0,1}$	30.424	5.889	0.070	20.069	26.170	30.080	34.190	42.810
$\tau_{1,1}$	35.462	4.056	0.041	27.990	32.590	35.330	38.110	43.880
$\tau_{1,2}$	34.196	4.019	0.041	27.000	31.400	33.940	36.700	42.590

for unobserved quantities, burn-in of 1.5 10^5 steps, 1.5 10^5 values as an approximation of the final distribution.

Autocorrelations are below 0.2 for all but four parameters, and significant lags, if present, fall down to zero almost always after lag five. The Geweke, Raftery-Lewis, Heidelberger and Welch output diagnostics were overall satisfactory.

In Table 3, the marginal posterior distributions for the gene effects given dye have been summarized. The posterior marginal averages of precision parameters $\tau_{1,1}, \tau_{1,2}$ differ from those fitted to blank spots (Table 2), a difference also found in the analysis of summary statistics. The difference in marginal posterior mean between dyes suggests that some model pruning might be feasible, e.g., by using just one parameter for the two dyes. The precision parameter in the likelihood function for housekeeping spots also differs from the same parameter fitted to blanks (Table 2).

As regards negative controls, the marginal posterior averages of $\Delta_{c,l}$ are clearly greater than zero and eventually different between dyes, a result which suggests the presence of unspecific DNA hybridization.

In Table (3), marginal posterior distributions of housekeeping-related

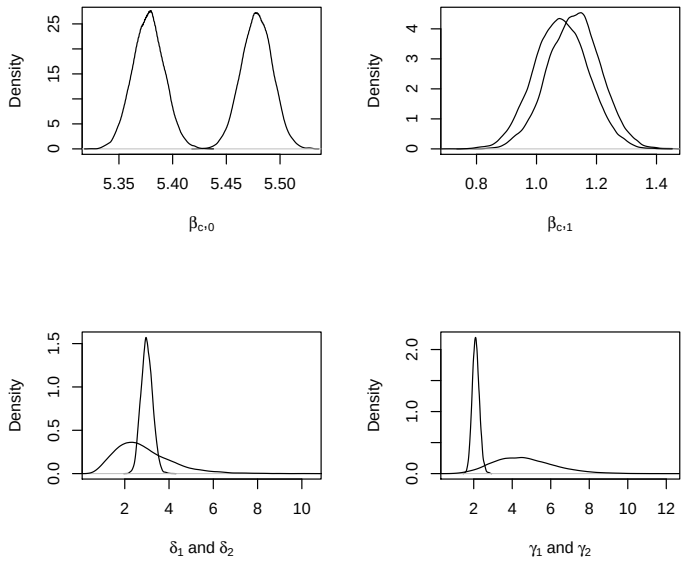

Figure 6: Approximated marginal distributions of some model parameters. Regression parameters for the location mean are shown on top, one density estimate for each dye. Regression parameters for the estimate of the probability of a large error are shown bottom right, those for the estimate of the scale of such errors are bottom left.

parameters typically differ in average between dyes. This finding is compatible with the presence of genetic modulation, that is, they are not actual housekeeping genes, otherwise the presence of an overall dye unbalancing at hybridization/scanning time should be hypothesized.

4 Discussion

Results from the analysis of the case study show that the magnitude of outlying fluorescence values in blank spots is conditioned on the fluorescence intensity, and that the deletion of an observation which is 'far' from values of the adjacent replicates might be a poor decision if the noise structure is not properly recognized.

The model developed in Section (2.1) accounts for the presence of outly-

Table 3: Summary of some parameters fitted for non-blank controls (sample size $1.5 \cdot 10^5$). Columns are labeled like in Table 2. The symbol $S_{f,c,g}$ stands for FHSF intensity, dye c and EST g (see Table 1). Here $\tau_{0,0}$ stands for $\tau_0 \mid T = 0$ and $\tau_{0,1}$ stands for $\tau_0 \mid T = 1$.

	Mean	SD	T.S. SE	2.5%	25%	50%	75%	97.5%
$\tau_{1,1}$	4.382	1.749	0.006	1.740	3.111	4.125	5.369	8.502
$\tau_{1,2}$	4.284	1.733	0.006	1.687	3.014	4.023	5.268	8.356
$\tau_{0,0}$	24.121	6.844	0.025	12.450	19.230	23.540	28.360	39.110
$\tau_{0,1}$	29.562	5.968	0.022	20.580	25.450	28.610	32.620	44.060
$S_{f,1,1}$	2.989	0.958	0.005	0.733	2.469	3.144	3.645	4.504
$S_{f,1,2}$	2.888	0.956	0.005	0.652	2.362	3.045	3.552	4.392
$S_{f,1,3}$	3.245	0.946	0.005	1.011	2.726	3.398	3.897	4.738
$S_{f,1,4}$	2.292	0.724	0.003	0.672	1.892	2.346	2.753	3.581
$S_{f,1,5}$	2.817	0.982	0.005	0.516	2.263	2.979	3.502	4.364
$S_{f,1,6}$	2.883	0.761	0.003	1.117	2.477	2.958	3.377	4.183
$S_{f,1,7}$	2.330	0.923	0.005	0.171	1.827	2.472	2.955	3.818
$S_{f,1,8}$	2.676	0.900	0.004	0.565	2.193	2.808	3.283	4.127
$S_{f,1,9}$	1.241	0.847	0.004	−0.709	0.786	1.339	1.794	2.684
$S_{f,1,10}$	2.000	0.869	0.004	−0.043	1.535	2.116	2.580	3.424
$S_{f,1,11}$	2.258	0.926	0.005	0.073	1.754	2.394	2.890	3.760
$S_{f,1,12}$	3.202	0.894	0.004	1.077	2.729	3.337	3.807	4.627
$S_{f,2,1}$	2.875	0.830	0.004	0.925	2.428	2.981	3.428	4.246
$S_{f,2,2}$	2.695	0.912	0.005	0.555	2.203	2.832	3.313	4.149
$S_{f,2,3}$	3.072	0.848	0.004	1.047	2.625	3.188	3.640	4.448
$S_{f,2,4}$	2.113	0.701	0.003	0.570	1.715	2.158	2.562	3.386
$S_{f,2,5}$	2.895	0.857	0.004	0.854	2.441	3.015	3.468	4.287
$S_{f,2,6}$	2.823	0.735	0.003	1.145	2.425	2.885	3.296	4.106
$S_{f,2,7}$	2.566	0.880	0.004	0.483	2.097	2.690	3.158	3.997
$S_{f,2,8}$	2.807	0.809	0.004	0.900	2.379	2.905	3.343	4.158
$S_{f,2,9}$	1.508	0.804	0.004	−0.348	1.074	1.589	2.030	2.901
$S_{f,2,10}$	1.918	0.813	0.004	0.011	1.489	2.014	2.446	3.302
$S_{f,2,11}$	2.548	0.923	0.005	0.365	2.054	2.691	3.174	4.029
$S_{f,2,12}$	2.996	0.861	0.004	0.946	2.537	3.113	3.571	4.400
$\Delta_{1,14}$	1.110	0.815	0.004	−0.761	0.662	1.192	1.641	2.537
$\Delta_{1,15}$	1.814	0.857	0.004	−0.184	1.352	1.920	2.381	3.255
$\Delta_{2,14}$	1.198	0.881	0.004	−0.846	0.718	1.313	1.784	2.682
$\Delta_{2,15}$	1.660	0.970	0.005	−0.575	1.107	1.815	2.330	3.229

ing observations and puts more structure in the foreground-to-background relationship than what is typically considered in the literature. It is worth noticing that, although prior distributions of model parameters have been elicited to represent subjective beliefs, some of them are objective random variables induced by the technical variability existing in the array manufacturing process and in the equipment work-flow. Anyway, prior distributions were selected by considering the information from previous experiments and from the literature.

The hierarchical model has meaningful parameters which can be used to evaluate some quality components on a quantitative basis. For example, the median of the marginal posterior distribution of the precision parameters $\tau_{1,1}$ and $\tau_{1,2}$ given observed blank spots are 35.33 and 33.94, respectively (thus variances are 0.0283 and 0.0295). Values of these two parameters depend on the quality of the experimental setup (array coating, imaging algorithms, etc.) because a larger precision implies a better estimate of background fluorescence for test spots. The quality assessment might also be based on functions of model parameters. In equation (16), the probability $\pi_{c,l}$ that the allocator variable is equal to 1 (thus $\alpha = 0$) is defined. The value of α also depends on the presence of artifacts, thus a high value of $\pi_{c,l}$ indicates a small probability of an artifact, i.e. good quality. The median of the marginal (univariate) posterior distribution of γ_1 and γ_2 are, respectively, 2.102 and 4.494 (see Table 2) and by plugging these values in (16) we obtain a probability value of 0.839 for $\bar{x} = 4.7$, a value of 0.953 for $\bar{x} = 5.0$ and a value of 0.987 for $\bar{x} = 5.3$, with 0.999 for $\bar{x} = 5.8$. These results suggest the possibility of obtaining a closed-form estimate of $\tau_{1,c}$ using only the averaged observations which have a mean background greater that 5.3. By doing this on color 1, we obtained 8 observations and a point estimate $1/(S^2)$ of $\tau_{1,1}$ based on residuals of the least squares regression on such points equal to 84.44. The estimate is about two times greater than the Bayesian estimate and it is based on just 4% of blank values. The least squares regression coefficient $\beta_{1,1}$ is equal to 1.941, quite different from the Bayesian point estimate that is equal to 1.126.

The Bayesian model of section 2.1 has been fitted by a MCMC that needed one hour on a Pentium 2.2 Ghz to run the more complex model with all the control spots. Although this is not a negligible amount of time it is still reasonable for researchers interested in the assessment of quality and if the uncertainty due to normalization must be taken into account. The most important recommendation is about the initialization of the Markov Chain, especially as regards latent allocator variables, to improve convergence. Using the output of descriptive scattergrams, like (5), we found a

good initialization by setting the value of Zs for points on the straight line of Figure 5 to one.

The normalization of raw fluorescence intensities measured on test spots, those printed with EST sequences to be investigated, may be performed with or without fitting an expanded model through MCMC. Test spots resemble housekeeping genes, but for each EST the two latent FSHF intensities may be different: this fitting requires the model for housekeeping genes already developed. It is worth noticing that we fit the model for housekeeping genes by keeping distinct parameters for each dye, although theoretically just one parameter was needed. Each pair of latent variables $S_{f,1,l}$ and $S_{f,2,l}$ carries information to be further exploited for normalization; in fact if $S_{f,1,l}$ and $S_{f,2,l}$ are very different, then the hypothesized housekeeping gene might be a false positive control, an event well documented in the literature (Eickhoff et al. 1999; Thellin et al. 1999). Otherwise, a considerable experimental dye unbalancing might be present and further model extensions would have to be introduced.

A second and simpler approach to normalization is based on the value taken by the out-of-spot background. In our case study, for values below 5.0 foreground values of test spots must be analyzed through the full Bayesian model which provides marginal posterior distributions of Z and α. If the out-of-spot background is above 5.0 then we obtain a normalized value as $\widehat{S}_f = y_f - \widehat{\beta}_0 - \widehat{\beta}_1 x.$, If model fitting on negative controls would show that $\Delta \neq 0$ then we could also subtract from r.h.s the estimated $\widehat{\Delta}$. A full statistical characterization of the simplified approach would be very useful in view of of the normalization of large scale experiments.

Besides boosting the computation, the current model might be improved upon through model calibration. Results from model checking (not shown) performed using discrepancy variables (Gelman and Meng 1996) have shown that, even on the tails of the predictive distribution the model behaves well, although some improvements are possible in a region below the straight line, for small background values.

Further model extensions might involve spike controls. Their use is almost mandatory if the interest is not focused on genome-wide expression but on a limited number of genes, like it may happen in a pharmacogenomic study targeting a specific metabolic pathway. Here the set of invariant EST required by many normalization algorithms may be empty, and due to the challenging cost of the overall study, all the information should be extracted with minimum bias.

Acknowledgments

This paper is partially supported by PRIN2005 from the Italian MIUR. Thanks are due to Rick Pè and Andy Gallavotti, University of Milan, for providing the unpublished data and for sharing several discussions on the experimental system.

References

Eickhoff, B., Korn, B., Schick, M., Poustka, A., and van der Bosch, J. (1999), "Normalization of array hybridization experiments in differential gene expression analysis," *Nucleic Acids Research*, 27, e33.

Gelman, A., and Meng, X. L. (1996), "Model checking and model improvement," in *Markov Chain Monte Carlo in practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter Chapman & Hall.

Geweke, J. (1996), "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," in *In Bayesian Statistics 4*, eds. J. Bernado, J. Berger, A. Dawid, and A. Smith Clarendon Press.

Heidelberger, P., and Welch, P. (1983), "Simulation run length control in the presence of an initial transient," *Operations Research*, 31, 1109–1144.

Nguyen, D., Arpat, A. B., Wang, N., and Carroll, R. J. (2002), "DNA microarray experiments: biological and technological aspects," *Biometrics*, 58, 701–717.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006), *CODA: output nalysis and diagnostics for MCMC*. http://www-fis.iarc.fr/coda/.

R Development Core Team (2005), *R: A language and environment for statistical computing*. http://www.R-project.org.

Raftery, A., and Lewis, S. (1992), "One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo," *Statistical Science*, 7, 493–497.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000), "Normalization strategies for cDNA microarrays," *Nucleic Acids Research*, 28, E47.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996), *BUGS: Bayesian inference using Gibbs Sampling, Version 0.5* MRC Biostatistics Unit, Cambridge.

Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. (1999), "Housekeeping genes as internal standards: use and limits," *Journal of Biotechnology*, 75, 291295.