Pixel-level signal modelling with spatial correlation for two-colour microarrays

Claus Thorn Ekstrøm and Søren Bak and Mats Rudemo

11th November 2004

Abstract

Statistical models for spot shapes and signal intensities are used in image analysis of laser scans of microarrays. Most models have essentially been based on the assumption of independent pixel intensity values, but models that allow for spatial correlation among neighbouring pixels can accommodate errors in the microarray slide and should improve the model fit. Five spatial correlation structures, exponential, Gaussian, linear, rational quadratic and spherical, are compared for a dataset with 50-mer two-colour oligonucleotide microarrays and 452 probes for selected Arabidopsis genes. Substantial improvement in model fit is obtained for all five correlation structures compared to the model with independent pixel values, and the Gaussian and the spherical models seem to be slightly better than the other three models. We also conclude that for the data set analysed the correlation seems negligible for non-neighbouring pixels.
Source code for R is available at
`http://www.matfys.kvl.dk/~ekstrom/spotshapes/`

Introduction

In two-colour microarray experiments, gene expression intensities are extracted for each spot from the red and green channel image files after it has been determined which pixels constitute the gene (the foreground) and which pixels that are part of the background.

Improved models for spot shapes and signal intensities help identify and define spots more clearly on the image file and will refine the classification of pixels as foreground or background. Thus, improved spot shape models provide more precise information on spot intensity level and on local background necessary for (local) normalization. Statistical models for spot intensities will enable us to extract (unbiased) gene expression intensities from highly expressed genes, when some of the pixels values are censored at the upper limit of the scanner, typically at $2^{16} - 1 = 65535$ with 16-bit precision, and will also provide us with more detailed information about the nature of the gene expression intensities.

Spatial models that allow for correlation between neighbouring pixel intensities can capture imperfections in the microarray slide, correct for "overspill" caused by excess DNA and they can compensate for optical point spread functions extended over more than one pixel. Spatial models can obviously be used to estimate the correlation structure between pixels and are likely to improve the fit to the image data. In addition, accurate pixel-level spot intensity models can be used to simulate more realistic microarray data in order to validate the efficiency of methods to extract information from microarray images.

In Ekstrøm et al. (2004) we presented a polynomial-hyperbolic spot shape model with the following three desirable properties: (i) isotropic, i.e., that the average intensity at a pixel x only depends on the distance from x to the spot centre and not on the direction from the centre, (ii) should allow for spot-shapes resembling both "plateaus" and "volcanos/craters/donuts" as spot intensities are often highest near the edge of the spot and smaller near the spot centre making the resulting spot shape resemble a volcano, and (iii) allow for spatial correlation, i.e., intensities at pixels close together should be more correlated than intensities at pixels further apart. Although the polynomial-hyperbolic spot shape model developed in Ekstrøm et al. (2004) did allow for spatial correlation, we only considered independent pixel intensities. In the present paper we allow for spatial correlation structure in order to improve the fit of the polynomial-hyperbolic spot shape model and the reconstruction of the individual spot signals. Different spatial correlation structures for the polynomial-hyperbolic spot shape model are studied and we estimate the corresponding extension of the correlation. The correlation should result in a substantial increase in the fit of the model and may provide improved estimates for saturated pixel values.

The models are applied to a dataset obtained with a specially designed spotted 50-mer oligonucleotide microarray. Here the expression of 452 selected genes in transgenic Arabidopsis plants are compared to the corresponding genes in wild-type plants (Kristensen et al., 2005). Data include scans with different photometric gains ranging from no saturation to heavy saturation.

Materials

The data used for shape modelling and data transformation are based on a transcriptome analysis (Kristensen et al., 2005) of metabolically altered *Arabidopsis thaliana* plants (Tattersall et al., 2001; Bak et al., 1999, 2000). The array is a custom designed 50-mer oligonucleotide array, 9×18 mm, 350 μm dot spacing, spotted by MWG Biotech using a single pin on epoxy coated glass slides. The array contains probes for 452 selected *A. thaliana* genes spotted in duplicate. The 50-mer oligonucleotides were designed by MWG Biotech, essentially as described by Kane et al. (2000). mRNA was isolated from 30 days old *A. thaliana* rosette leaves using MicroPoly(A)Pure^{TM} small scale mRNA purification kit (Ambion). 3-3.5 μg mRNA was used for direct incorporation of cy3- and cy5-fluorescent dyes (Amersham Pharmacia Biotech) using Superscript II kit (Invitrogen). Hybridizations and washings were performed essentially according to the manufacturers instructions and subsequently scanned using a GMS 418 Array Scanner (Affymetrix) using four different photomultiplier gains: 30, 40, 50, 60 while keeping the laser power at 30.

The resulting 16 bit gray scale tif-images are available for wildtype (wt) and the transgenic line 3x.8 (Tattersall et al., 2001), four photomultiplier gains: 30, 40, 50, 60, and two dye swap experiments: cy3, cy5, for a total of 16 images.

Spatial spot shape models

The polynomial-hyperbolic spot shape model

We start this section with an outline of the polynomial-hyperbolic spot shape model. Let S denote the set of spots and for each spot $s, s \in S$, we associate a set A_s of pixels such that no pixel can belong to more than one such set. $Y = Y(x)$ denotes the suitably transformed intensity at a pixel, x, with pixel coordinates $x = (x_1, x_2)$. We prefer a Box-Cox transformation (Box and Cox, 1964) of the measured pixel values from the laser scanner (Ekstrøm et al., 2004).

Consider a spot s and pixels $x \in A_s$. Let $c_s = (c_{s1}, c_{s2})$ be the spot centre of spot s, and let $r_s(x) = ||x - c_s||$ be the distance from pixel x to the spot centre. The

polynomial-hyperbolic spot shape model is

$$Y(x) = B_s h_s(r_s(x)) + b_s + \varepsilon(x), \quad x \in A_s \tag{1}$$

where B_s measures the intensity of spot s, b_s is a constant representing the background, $\varepsilon(x)$ corresponds to zero-mean Gaussian noise at x. The spot shape function is

$$h_s(r) = \begin{cases} \frac{K_s}{\sigma_s^2} \exp(g_s(r/\sigma_s)) & \text{if} \quad 0 \le r < \gamma_s \sigma_s \\ 0 & \text{if} \quad r \ge \gamma_s \sigma_s, \end{cases} \tag{2}$$

with

$$g_s(r) = \sum_{i=1}^{2} b_{si} r^i - \frac{a_s}{\gamma_s - r}, \quad 0 \le r < \gamma_s,$$

where $a_s > 0$ and $\gamma_s > 1$, σ_s represents the radius of the spot, K_s is a normalizing constant and

$$\begin{aligned} b_{s1} &= a_s/\gamma_s^2 \\ b_{s2} &= \frac{a_s}{2}\left\{\frac{1}{(\gamma_s-1)^2} - \frac{1}{\gamma_s^2}\right\}. \end{aligned}$$

Some spot shape parameters may be common for all spots and some may be spot-specific.

We assume that $(Y(x), x \in A_s)$ has a multivariate normal distribution with mean vector μ_s, where $\mu_s = B_s h_s(r_s(x)) + b_s$ and covariance matrix C_s. In Ekstrøm et al. (2004) we only considered the simplest possible covariance model where C_s is proportional to the identity matrix such that all pixel values are independent. Here, we examine more complex correlation structures.

Spatial correlation structures

We will consider five different one-parameter isotropic spatial correlation structures such that

$$C_s = \mathrm{cov}(Y,Y) = \sigma_e^2 R, \tag{3}$$

where R is the correlation matrix with elements

$$r_{x,x'} = \psi(d(x,x'),\rho), x \in A_s, x' \in A_s. \tag{4}$$

ρ is a non-negative parameter, ψ is a correlation function, and $d(x,x')$ is a distance function between pixels x and x', and for our purpose here, we use the Euclidian distance between the pixel coordinates $d(x,x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}$.

The five different correlation functions we consider are

Exponential

$$\psi(d,\rho) = \exp(-d/\rho) \tag{5}$$

Gaussian

$$\psi(d,\rho) = \exp(-(d/\rho)^2) \tag{6}$$

Linear

$$\psi(d,\rho) = (1-d/\rho)\,\mathbf{1}(d<\rho) \tag{7}$$

Rational quadratic

$$\psi(d,\rho) = \frac{1}{1+(d/\rho)^2} \tag{8}$$

Spherical

$$\psi(d,\rho) = (1-\frac{3}{2}(d/\rho)+\frac{1}{2}(d/\rho)^3)\,\mathbf{1}(d<\rho) \tag{9}$$

All five correlation function depend on a single parameter, ρ, and Cressie (1993) and Pinheiro and Bates (2000) provide additional details about these correlation structures.

Estimation of parameters and saturated pixel values

Spot shape parameters are estimated by maximizing the log likelihood function related to the above assumed multivariate normal distribution,

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|C_s|) - \frac{1}{2}(y-\mu_s)^{\intercal}C_s^{-1}(y-\mu_s), \tag{10}$$

where μ_s contains the parameters for the polynomial-hyperbolic spot shape model and where C_s contains the residual variance and the correlation parameter.

However, some pixels may be censored at an upper limit, ℓ_c, in which case we have a missing data problem. We use the expectation-maximization (EM) algorithm (Dempster et al., 1977) to maximize the likelihood in the presence of saturated pixels. Let $A_s' = \{x \in A_s : Y(x) < \ell_c\}$ and $A_s'' = \{x \in A_s : Y(x) \geq \ell_c\}$ denote the set of pixels in A_s that are uncensored and censored, respectively, at the level ℓ_c. Furthermore, let $Y_{\text{obs}} = (Y_1,\ldots,Y_m)$ denote the observed intensities of the m pixels from A_s' and $Y_{\text{cens}} = (Y_{m+1},\ldots,Y_n)$ the $n-m$ censored observations from A_s''. The log likelihood function can be written as

$$\log L = \log L(Y_{\text{cens}}|Y_{\text{obs}}) + \log L(Y_{\text{obs}}) \tag{11}$$

where $L(Y_{\text{obs}})$ is the marginal likelihood of the uncensored data and $L(Y_{\text{cens}}|Y_{\text{obs}})$ is the conditional likelihood of the censored data given the observed data. Y_{obs} follows a regular multivariate Gaussian distribution and is easily evaluated. If the observations in Y_{cens} had not been censored, the conditional distribution of $Y_{\text{cens}}|Y_{\text{obs}}$ would also have been multivariate normal and we could maximize the log likelihood (11) using standard iterative maximization techniques. However, since Y_{cens} are all censored at ℓ_c, we need to evaluate the multiple integral

$$L(Y_{\text{cens}}|Y_{\text{obs}}) = \int_{\ell_c}^{\infty} \cdots \int_{\ell_c}^{\infty} dP(y_{m+1}, \ldots, y_n) \tag{12}$$

where P is the (underlying) multivariate Gaussian distribution of the censored pixels given the observed pixel intensities. The observations in Y_{cens} are not independent because of the spatial correlation structure and evaluation of (12) becomes costly for more than a few censored observations. Monte Carlo methods can be used to evaluate (12) resulting in a doubly iterative algorithm to maximize (11).

Instead we propose to use an expectation-maximization-like (EM) approach where we impute values for the missing (censored) observations and then maximize the log likelihood (11) by standard iterative maximization techniques as if all pixels had been observed. Imputation is done by inserting the conditional marginal expectation, i.e., for iteration $k+1$ we replace each $Y_i, i = m+1, \ldots, n$, with

$$\hat{Y}_i^{k+1} = E(Y_i|Y_{\text{obs}}, Y_i \geq \ell_c, \theta^k) \tag{13}$$

where θ^k is the vector of current estimates of the parameters in the model. We stop the iterations when there is no change in the complete data log-likelihood and in the parameters. Once the complete data likelihood and parameter values have stabilized, we stop the algorithm when the average change in predicted censored pixels is less than a certain threshold ε, i.e. when

$$\frac{\sum_{i=m+1}^{n} (\hat{Y}_i^{k+1} - \hat{Y}_i^k)^2}{n-m} < \varepsilon \tag{14}$$

This additional option may be necessary to end the algorithm when a large percentage of the pixels are censored, see discussion. Note that the predicted values in (13) can also be used to predict saturated (censored) pixel values (see below).

Results

Parameter estimation

In this paper we have two major objectives: 1) to compare the model fit of the polynomial-hyperbolic spot shape model with spatial correlation to the model with

no correlation; and 2) identify realistic correlation structures between pixels. If the introduction of the spatial correlation structure results in a substantial increase in model fit we also want to assess how well the polynomial-hyperbolic spot shape model with spatial correlation predicts saturated (censored) pixel values relative to the model, where no correlation structure is applied.

We use log likelihoods to compare the relative fit of the polynomial-hyperbolic spot shape model with one of the five different isotropic correlation structures (5)–(9) to the spot shape model with independent pixel values. The results are shown in Table 1 and are based on analysis of 25 spots scanned at four different photometric gains for a total of 100 datasets. These spots are the same spots that were used for the analyses in Ekstrøm et al. (2004) and have no saturated pixels such that the log likelihood values for the different models are indeed comparable. A Box-Cox transformation Ekstrøm et al. (2004) was applied before analysis.

Table 1 shows that all five different correlation structures result in a substantial improvement in log likelihood and that there — except for the correlation parameter ρ — is virtually no difference in estimated parameter values for the five correlation structures. There are small variations in the median log likelihood improvements with the Gaussian correlation structure having the largest overall log likelihood improvement. Twice the log likelihood difference is approximately χ^2 distributed with one degree of freedom so a comparison of the spatial correlation models to the uncorrelated model will generally yield highly significant likelihood ratio tests.

The precise value of the correlation parameter ρ may not be very important for the fit of the correlation models as long as the correlation structure is there to capture some of the correlation between neighbouring pixels. Therefore, we fitted the five different correlation structures when the correlation parameter ρ was held fixed at the (rounded) median value found previously (i.e., ρ was fixed at 2, 1, 1, 1, or 1.4 for the spherical, exponential, Gaussian, rational and linear model, respectively). The results from these analyses are also listed in Table 1 and show that apart from the linear correlation model (7) there is no significant reduction in the median log likelihood when we compare a model where we maximize the correlation parameter ρ to the model where ρ is fixed.

Figure 1 shows the median estimated correlation as a function of Euclidian distance for the five different correlation structures. Vertical lines indicate the only possible observable distances in the data. The five different correlation structures split into two groups: the exponential and rational quadratic in one group which shows correlation at longer distances and the remaining three correlations structures in the other group. The two correlation structures corresponding to the largest improvement in median log likelihood values (the Gaussian and the spherical, see Table 1) are almost identical — especially at the distances possible with the spot data. Interestingly, the correlation function for both the Gaussian and the spheri-

Correlation structure	Parameter estimate							log like. improv.
	b_s	B_s	σ_s	γ_s	α_s	σ_e	ρ	
Independence	0.226	15.00	5.11	1.75	0.596	0.0408	–	–
Spherical	0.226	16.35	5.11	1.75	0.763	0.0401	1.970	78
	0.226	17.05	5.11	1.77	0.766	0.0373	2†	77.5
Exponential	0.227	16.45	5.19	1.73	0.757	0.0416	0.995	69
	0.226	16.15	5.16	1.75	0.771	0.0388	1†	69
Gaussian	0.227	16.25	5.11	1.76	0.746	0.0401	0.956	82
	0.227	16.30	5.11	1.77	0.820	0.0397	1†	80.5
Rational quadratic	0.227	16.50	5.18	1.74	0.767	0.0439	0.875	75
	0.227	16.90	5.18	1.74	0.816	0.0432	1†	72.5
Linear	0.227	16.20	5.12	1.76	0.736	0.0391	1.395	73.5
	0.227	16.05	5.10	1.77	0.760	0.0387	1.4†	47.5

Table 1: Comparison of correlation structures: Median estimated parameter values from five correlation structures for 25 spots and four gains. The log likelihood improvement is the median increase of the log likelihood for a given correlation structure relative to the log likelihood of the uncorrelated model. The rows where ρ-values are marked with † correspond to fixed values of ρ.

cal correlation structure are essentially 0 from a distance of 2 or greater suggesting that there is virtually no correlation between non-neighbouring pixels. Figure 1 also shows the median empirical correlation coefficients for the observable distances. The empirical correlation coefficients coincide nicely with the estimated Gaussian/spherical correlation function for pixel distance less than 2 and for pixel distances equal to or above 2 the empirical correlations are somewhat larger than the estimated Gaussian and spherical correlation functions.

Reconstruction of saturated values

Table 1 shows a substantial increase in log-likelihood so it is clear that the introduction of a spatial correlation structure increases the fit of the polynomial-hyperbolic spot shape model. We proceed to investigate if the correlated model provides improved estimates of the saturated (censored) pixel values.

In Figure 2 we show the estimated spot shape profiles when the pixels for spots 242, 352 and 787 (scanned with photometric laser gain 60) are artificially censored at different intensities ($\ell_c = 0.9, 0.8, 0.7$ and 0.6). These three spots were chosen as those with the highest intensity level not exceeding the upper limit. The spots were artificially censored so we knew the actual intensities of all pixels and therefore could compare the reconstructed pixel values with the true pixel values. We

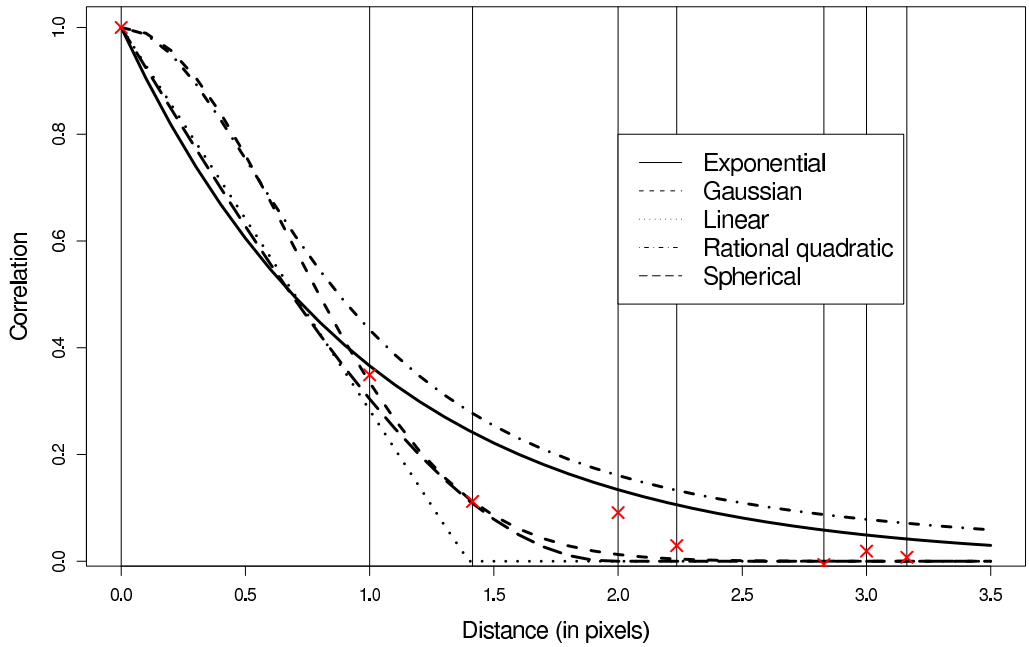

Figure 1: Median estimated correlation functions for the five examined correlation structures. Vertical lines indicate the only possible observable distances between pixels and the $\times$'s show the median empirical correlation coefficients.

estimated the spot shape model (1) both with and without a Gaussian spatial correlation structure (6) since that correlation structure provided the largest log likelihood improvement in Table 1. The leftmost plots in Figure 2 show the estimated spot profiles for each of these spots when there is no censoring, while the other diagrams show reconstruction for varying degrees of censoring.

With a small degree of censoring corresponding to the second or third column in Figure 2 the reconstruction is satisfactory and the model with Gaussian correlation structure provides slightly improved estimates of the saturated pixels relative to the model with no correlation structure. For higher degrees of censoring (the two rightmost columns in Figure 2) the Gaussian correlation structure does not appear to provide better predictions for the saturated pixels than the polynomial hyperbolic model with independent errors, but both the model with independent and the model with correlated pixel values are clearly better than just using the censored values.

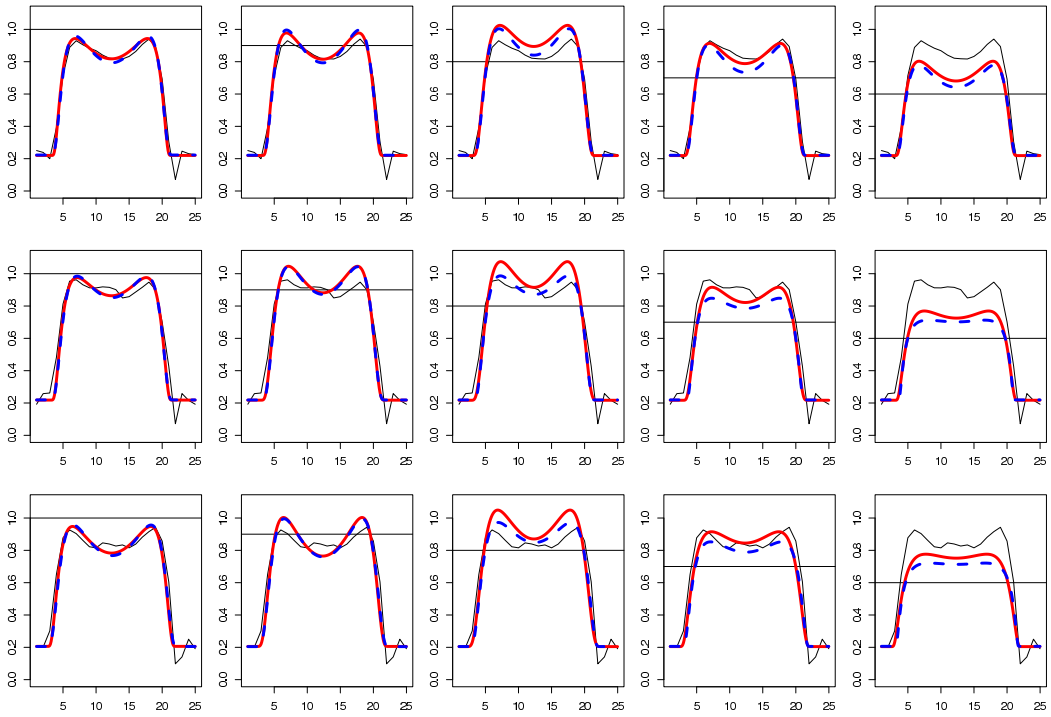

Figure 2: Horizontal normalized intensity profiles through the centres of spots 242, 352 and 787 (each spot represented by a row) at photometric gain 60 with different levels of (artificial) censoring as indicated by the horizontal lines. For each profile both data (thin lines) and estimated spot profile for non-correlated (thick solid line) and spatially correlated (Gaussian) polynomial-hyperbolic model (thick dashed line) are shown. The average fraction of pixels that were censored among the 25×25 pixels regarded for each spot were (from the left) in the five columns: 0%, 17%, 29%, 30% and 32%, respectively.

Discussion

The polynomial-hyperbolic spot shape model examined previously (Ekstrøm et al., 2004) provides a good fit to spot intensities measured on the pixel level, and the model can capture the volcano/donut shapes caused by surface tension when the spot dries. Initially, however, the polynomial-hyperbolic spot shape model assumed independent errors but correlated pixel values may occur for example because of imperfections in the array, dust, ineffective washing or from the point spread function in the laser scanning. The correlation structure might also compensate for minor local systematic deviations between the true spot shape and the spot shape

model, regard for instance the central part of the spot shown in Figure 3 in Ekstrøm et al. (2004).

In this paper we consider spatial correlation extensions for the polynomial-hyperbolic spot shape model for spot intensities measured on the pixel scale.

The results seen in Table 1 show that introduction of spatial correlation provides a substantial increase in log likelihood. This suggests that positive correlation between neighbouring pixels values are indeed present and should be addressed. If we fix the correlation parameter at a value close to the estimated median value, we see that the median increase in log likelihoods is hardly changed relative to the situation where the correlation parameter can vary freely. The only exception is the linear correlation structure, where a substantial decrease in log likelihood is observed when the parameter is fixed. The reason for the decrease with the linear correlation structure is its inflexibility; the correlation is exactly zero from a certain point and when the fixed value we used in the analyses is too small we will get a large decrease in log likelihood. The computations required for maximizing the model are significantly reduced (and the computation time significantly decreased) when the correlation parameters are fixed and the results from Table 1 suggest that we only get a minor decrease in log likelihood if we maximize the polynomial-hyperbolic spot shape model with a fixed spatial correlation structure without having to estimate the correlation parameter.

Table 1 also indicates that the Gaussian correlation structure may provide the best fit but since the difference between some of the correlation structures is small and we are looking at median increase in log likelihood it is difficult to give conclusive evidence that the Gaussian correlation structure is consistently best. That the specific choice of correlation structure may have little influence on the actual estimated correlation is also seen in Figure 1. The two best-fitting correlation functions (the Gaussian and the spherical) are almost identical at the Euclidian distances observable in the dataset and these two functions have estimated correlation functions that are essentially 0 from a distance of 2 pixels or greater. While this suggests that virtually no correlation exists between non-neighbouring pixels we can see from the empirical correlation that there is still some correlation remaining at distance 2. However, the correlation at distance 2 (median empirical correlation coefficient of 0.09) is not particularly large.

As shown in Ekstrøm et al. (2004), the polynomial-hyperbolic spot shape model provides a considerable improvement in log-likelihood compared to models earlier described in literature such as the top hat model, the Gaussian model and the Gaussian difference model suggested in Wierling et al. (2002). If we compare the spatial model fit, we note that the median increase in log-likelihood after a spatial correlation structure is introduced exceeds the median log-likelihood improvement obtained by using a Box-Cox transformation of the pixel values instead of a simple log

transform, see Table 1 in Ekstrøm et al. (2004). Also, the log-likelihood increase of the spatial model is more than half the increase when going from a cylindrical spot shape model to the more flexible polynomial-hyperbolic spot shape model. Thus, the improvements in log-likelihood after of the spatial polynomial-hyperbolic spot shape model in considerable.

Figure 2 suggests that with minor censoring (less than 30%, say) the polynomial-hyperbolic model with Gaussian correlation predicts censored pixels satisfactory and slightly better than the model with independent errors.

The polynomial-hyperbolic model with independent errors had a tendency to "overshoot" the estimated spot profiles (see three middle graphs in Figure 2) in some situations. Generally, the predicted profile for models with Gaussian correlation lies slightly lower than the profiles for models with independent errors, so this overshot is lessened after introducing spatial correlation. However, the introduction of spatial correlation only goes so far. When censoring is large and only the edge and the background pixels of the spot is observed (i.e., when $\ell_c = 0.6$ or 0.7) the polynomial-hyperbolic model with Gaussian correlation fares no better than to model with independent errors. In general, observed pixels from the spot centre are often needed to estimate the parameters in the model well and thereby provide good predictions of the censored pixel values.

The biggest problem with the proposed spatial extension to the polynomial-hyperbolic spot shape model is the difficulties in evaluating the multiple integral (12) required for the incomplete data likelihood. It is possible to evaluate it numerically, but the additional time needed to evaluate the integral would render the method virtually useless.

Using the approximation (13) can make it difficult to determine when the algorithm has converged since we can only monitor the complete data likelihood and the data used for the calculations change from iteration to iteration. This normally poses no problems in the situations where some of pixels in the spot centre are observed so we can find stable estimates of the parameters. However, if there is no information about the pixels in the spot centre the likelihood becomes flat and the parameters change a little from iteration to iteration with virtually no change in log likelihood. The resulting spot profiles are indistinguishable from iteration to iteration so this really gives no problems with respect to prediction of saturated pixels but it does make it more difficult to determine when to stop the algorithm. We found that the criteria described above work well in our setting.

The results from Figure 1 suggest that a pixel value in essence is only correlated with the eight neighbouring pixels. Thus, we may avoid the approximation (13) and use the multiple integral (12) if we assume that each censored pixel is independent of all non-neighbouring pixels. This approach requires the evaluation of a multiple Gaussian integral in nine dimensions (if all eight neighbouring pixels are also cen-

sored) and in less dimensions if not all pixels are censored. Numerical evaluation of multiple integrals with more than a few dimensions is, however, so computational expensive that this approach is useless in practice (a single iteration for spot 352 when $\ell_c = 0.9$ takes several days to evaluate using the C/Fortran-routines implemented in R). It may be argued based on Figure 1 that it is only necessary to condition on the four non-diagonal neighbouring pixels and while that results in a substantial reduction in computation time it is still too costly to be useful for any real array with current computational power (a single iteration takes 5 hours to complete).

In conclusion, the polynomial-hyperbolic spot shape model with spatially correlated errors results in a substantially better fit than the model with independent errors and may prove useful for both extraction of more precise gene expression intensities and as a model for simulating realistic microarray image data. The results suggest that the range of the correlation does not extend several pixels but that the existing local correlations are highly significant. Finally, the spatially correlated polynomial-hyperbolic spot shape model will lessen the information loss for spots with a low or moderate number of saturated pixels intensities due to censoring. This improves the usefulness of the polynomial-hyperbolic model as a tool to predict censored pixel values.

Source code for R is available at `www.matfys.kvl.dk/~ekstrom/spotshapes`

References

Bak, S., Olsen, C. E., Halkier, B. A., and Møller, B. L. (2000). Transgenic tobacco and arabidopsis plants expressing the two multifunctional sorghum cytochromes P450, CYP79A1 and CYP71E1, are cyanogenic and accumulate metabolites derived from intermediates in dhurrin biosynthesis. *Plant Physiology*, 123:1437–1448.

Bak, S., Olsen, C. E., Petersen, B. L., Møller, B. L., and Halkier, B. A. (1999). Metabolic engineering of p-hydroxybenzylglucosinolate in Arabidopsis by expression of the cyanogenic CYP79A1 from Sorghum bicolor. *Plant Journal*, 20:663–672.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–252.

Cressie, N. (1993). *Statistics for Spatial Data, revised edition*. Wiley.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statistical Society, Series B*, 39:1–38.

Ekstrøm, C. T., Kristensen, C., Bak, S., and Rudemo, M. (2004). Spot shape modelling and data transformations for microarrays. *Bioinformatics*, 20:2270–2278.

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research*, 28:4552–4557.

Kristensen, C., Morant, M., Olsen, C. E., Ekstrøm, C. T., Galbraith, D. W., Møller, B. L., and Bak, S. (2005). Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome. *PNAS*, 102:1779–1784.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York, USA.

Tattersall, D. B., Bak, S., Jones, P. R., Olsen, C. E., Nielsen, J. K., Hansen, M. K., Høj, P. B., and Møller, B. L. (2001). Resistance to an herbivore through engineered cyanogenic glucoside synthesis. *Science*, 293:1826–1828.

Wierling, C. K., Steinfach, M., Elge, T., Schulze-Kremer, S., Aanstad, P., Clark, M., Lehrach, H., and Herwig, R. (2002). Simulation of dna array hybridization

experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics*, 3:29.