BIOINFORMATICS

Vol. 20 no. 14 2004, pages 2270–2278
doi:10.1093/bioinformatics/bth237

Spot shape modelling and data transformations for microarrays

Claus Thorn Ekstrøm[1,], Søren Bak[2], Charlotte Kristensen[2,†] and Mats Rudemo[1]*

[1]Department of Mathematics and Physics and [2]Plant Biochemistry Laboratory, Department of Plant Biology, Center of Molecular Plant Physiology (PlaCe), Royal Veterinary and Agricultural University, Thorvaldsensvej 40, DK-1871 Frederiksberg C., Denmark

Received on October 30, 2003; revised on March 4, 2004; accepted on March 25, 2004
Advance Access publication April 1, 2004

ABSTRACT

Motivation: To study lowly expressed genes in microarray experiments, it is useful to increase the photometric gain in the scanning. However, a large gain may cause some pixels for highly expressed genes to become saturated. Spatial statistical models that model spot shapes on the pixel level may be used to infer information about the saturated pixel intensities. Other possible applications for spot shape models include data quality control and accurate determination of spot centres and spot diameters.

Results: Spatial statistical models for spotted microarrays are studied including pixel level transformations and spot shape models. The models are applied to a dataset from 50mer oligonucleotide microarrays with 452 selected *Arabidopsis* genes. Logarithmic, Box–Cox and inverse hyperbolic sine transformations are compared in combination with four spot shape models: a cylindric plateau shape, an isotropic Gaussian distribution and a difference of two-scaled Gaussian distribution suggested in the literature, as well as a proposed new polynomial-hyperbolic spot shape model. A substantial improvement is obtained for the dataset studied by the polynomial-hyperbolic spot shape model in combination with the Box–Cox transformation. The spatial statistical models are used to correct spot measurements with saturation by extrapolating the censored data.

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

Contact: ekstrom@dina.kvl.dk

INTRODUCTION

In order to study lowly expressed genes in microarray experiments, it is useful to increase the photometric gain in the scanning. However, a large gain may cause some pixels for highly expressed genes to become saturated, i.e. the registered pixel values become censored at the upper limit, which with 16-bit precision is $2^{16} - 1 = 65535$. Techniques for adjustment of highly expressed signal intensities are given in Wit and McClure (2003) based on a small set of available spot summaries, such as spot mean, spot median and spot variance. As mentioned in Wit and McClure (2003), it should be possible to get more accurate adjustments when all pixel values are available. In the present paper, we study spatial statistical models for pixel values that should enable such adjustments.

A convenient type of modelling is to transform data to become approximately Gaussian distributed with a mean value function determined by gene intensities and spot shapes and a corresponding covariance function. For such models, censored pixel values can be estimated optimally. We investigate several types of transformations on the pixel level such as the logarithmic transformation, the Box–Cox family (Box and Cox, 1964) and the inverse hyperbolic sine transformation (Huber *et al*., 2002; Durbin *et al*., 2002), also called the generalized logarithm (Rocke and Durbin, 2003). The inverse hyperbolic sine transformation has been proven useful for analyzing microarray spot intensities, but here we apply it at the pixel level. The Box–Cox transformation with exponent 0.5, i.e. a square root transformation optimal for Poisson distributed counts, has been used at pixel level analysis of microarray data by Glasbey and Ghazal (2003).

The spot shapes studied include three types suggested by Wierling *et al*. (2002): (i) a cylindric plateau spot distribution, (ii) an isotropic two-dimensional (2D) Gaussian distribution and (iii) a crater spot distribution consisting of a difference between two scaled isotropic 2D Gaussian distributions. These models does not seem to provide a satisfactory description for the dataset considered, and we introduce a new class of models with polynomial-hyperbolic spot shape. With a second degree polynomial we get a considerably improved performance. This spot shape may be regarded as a generalization of the cylindric plateau spot shape.

*To whom correspondence should be addressed.

†Present address: Poalis A/S, Bülowsvej 25, 1870 Frederiksberg C, Denmark

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

DATA, TRANSFORMATIONS AND EXPLORATORY ANALYSIS

Materials

The data used for shape modelling and data transformation are based on a transcriptome analysis (Kristensen and Bak, personal communication) of metabolically altered *Arabidopsis* plants (Tattersall *et al*., 2000). The array is a custom designed 50mer 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 *Arabidopsis* genes designed to cover the cytochrome P450 (Paquette *et al*., 2000; Werck-Reichhart *et al*., 2002) (see http://www.biobase.dk/P450/) and glycosyltransferase (UGT) (Paquette *et al*., 2003) multigene families as well as genes that relate to aromatic amino acid biosynthesis, secondary metabolism and stress. The 50mer oligonucleotides were designed by MWG Biotech, essentially as described by Kane *et al*. (2000). mRNA was isolated from 30 days old *Arabidopsis* rosette leaves using MicroPoly(A)Pure™ small-scale mRNA purification kit (Ambion). About 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 manufacturer's 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 grey scale tiff-images are available for two varieties: wild-type wt, transgenic line 3x.8, four photomultiplier gains: 30, 40, 50, 60 and two dye swap experiments: cy3, cy5, for a total of 16 images.

Transformations

Let $Z = Z(x)$ denote the intensity of a pixel x. Here, Z is a 16-bit integer, i.e. $0 \leq Z \leq 2^{16} - 1 = 65535$. Let $Y(x)$ denote a transformation of $Z(x)$,

$$Y(x) = f(Z(x), \boldsymbol{\lambda}), \tag{1}$$

where $f(\cdot, \boldsymbol{\lambda})$ is a family of transformation depending on the parameter vector $\boldsymbol{\lambda}$.

In the following, we shall consider three transformations: A logarithmic transformation

$$Y = k \log(Z + \lambda_1), \tag{2}$$

where λ_1 is a positive offset parameter. A Box–Cox transformation

$$Y = \begin{cases} k((Z + \lambda_1)^{\lambda_2} - 1)/\lambda_2 & \text{if } \lambda_2 \neq 0 \\ k \log(Z + \lambda_1) & \text{if } \lambda_2 = 0, \end{cases} \tag{3}$$

where $\lambda_1 > 0$, and an inverse hyperbolic sine transformation

$$Y = k \operatorname{arsinh}\left(\frac{Z + \lambda_1}{\lambda_2}\right), \quad \lambda_2 > 0. \tag{4}$$

The constant k is used in all three transformations to scale the transformed pixel intensities such that a saturated pixel (i.e. a pixel with intensity $Z = 2^{16} - 1 = 65535$) corresponds to a value of $Y = 1$.

Note that $\operatorname{arsinh}(z) = \log(z + \sqrt{z^2 + 1})$ for $z > 0$, so for large z we have $\operatorname{arsinh}(z) \approx \log(2z)$. As a result, the logarithmic transformation is essentially (at least for large values of z) a special case of both the Box–Cox transformation (with $\lambda_2 = 0$) and the inverse hyperbolic sine transformation (with $\lambda_2 = 2$).

Figure 1 shows the inverted grey scale image of the Y-intensities for the logarithmic transformation (2) with $\lambda_1 = 20$ and $k = 1/\log(2^{16} + 20 - 1)$ for the wild-type green (cy3) channel with photometric gain 60. Pixels with Y-values close to 1 are shown in black and pixels with Y-values close to 0 are light grey. The middle panel of Figure 1 shows the corresponding saturated pixels where $Y = 1$. In the right panel, we show the intensities for three spots with photometric gains 30, 40, 50 and 60, respectively, along horizontal lines through the spot centres. Each of the intensity curve is given for 25 pixels, the centre pixel and 12 pixels on each side. The spots are the 6th, 10th and 12th spots in the 9th row in the left panel; these three spots have numbers 102, 106 and 108, and show medium, high and low spot intensity, respectively, and the intensity curves are overlayed for all four gains. Spots 102, 106 and 108 are marked with boxes in the left panel of Figure 1.

SPOT SHAPE MODELS

Based on empirical observations of spot intensity profiles as seen in Figure 1 as well as in Duggan *et al*. (1999) (Fig. 2) and Glasbey and Ghazal (2003) (Fig. 1), we desire a spatial spot shape model to have the following three 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 'volcanos/craters/donuts' and 'plateaus'. 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 (middle panel of Fig. 1); and (iii) allow for spatial correlation, i.e. pixels close together and with the same distance from the spot centre should be more correlated than pixels further apart.

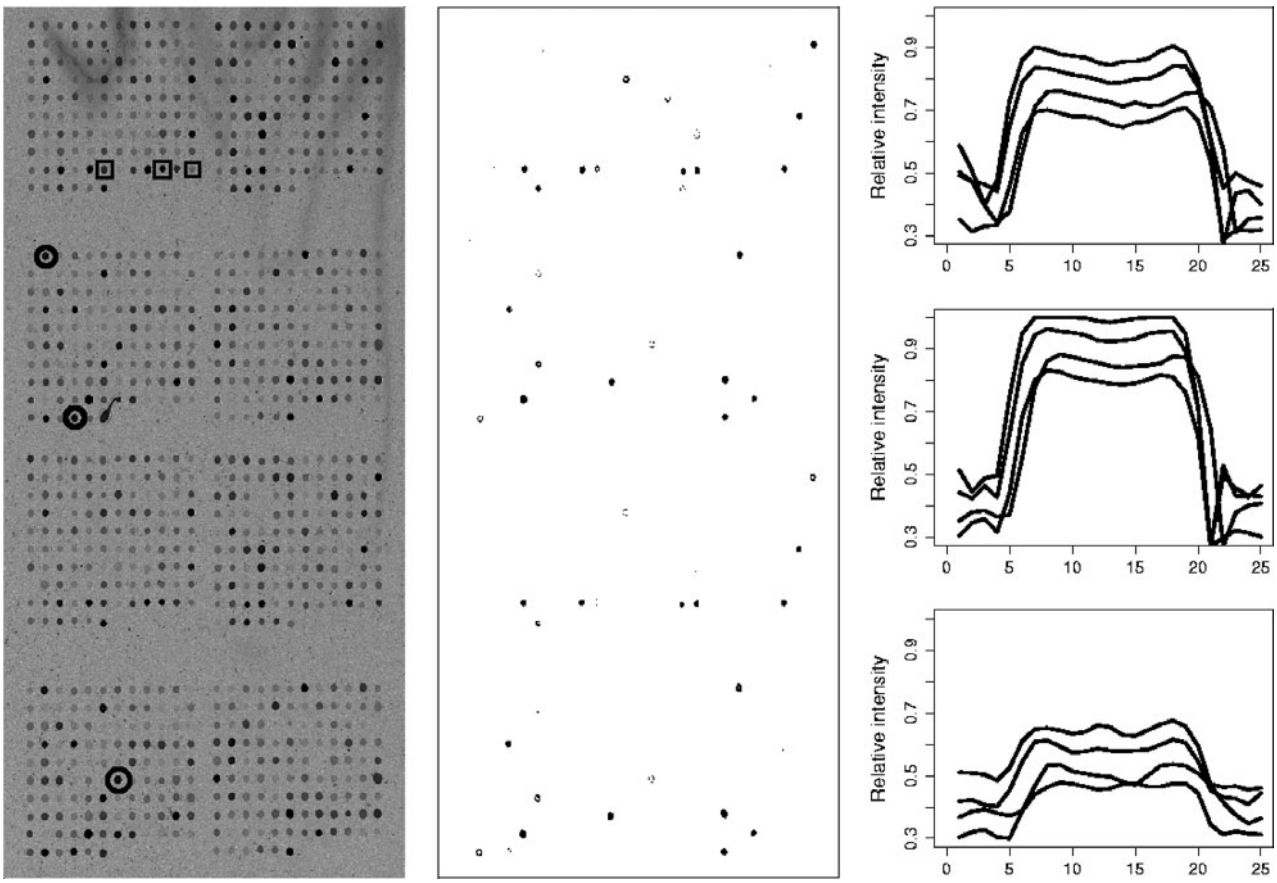

Fig. 1. Inverted grey scale image of cy3 wild-type with photometric gain 60 (left panel) and the corresponding saturated pixels (middle panel). Horizontal intensity profiles through spot centres for wild-type cy3 images with gains 30, 40, 50 and 60 are shown for three spots (right panel). The three spots depicted in the right panel are the ones marked with boxes on the left panel. The three spots marked with circles are used in Figure 5.

Let S denote the set of spots. With each spot $s, s \in S$, we associate a set A_s of pixels. We assume that no pixel belongs to more than one such set, and some pixels may not be associated with any spot. Let $Y = Y(x)$ denote the (possibly transformed) intensity at a pixel, x, with pixel centre coordinates $x = (x_1, x_2)$. We assume that $Y(x)$ and $Y(x')$ are independent if x and x' are associated with different spots.

Consider now 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. Assume that

$$Y(x) = B_s h_s(r_s(x)) + b_s + \epsilon(x), \quad x \in A_s, \tag{5}$$

where B_s measures the intensity of spot s, b_s is a constant representing the background, $h_s(r)$ is a spot shape function and $\epsilon(x)$ corresponds to zero-mean noise at x. We assume that $[Y(x), x \in A_s]$ has a multivariate normal distribution with mean vector μ_s and covariance matrix C_s. Thus

$$\mu_s(x) = B_s h_s(r_s(x)) + b_s,$$

and the spot shape function $h_s(r)$ may depend on parameters. Some spot shape parameters may be common for all spots but some may be spot-specific.

In the present paper, we only consider the simplest covariance model where each pixel intensity is assumed independent, i.e. $\epsilon(x) \sim N(0, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. More complicated spatial correlation structures will be investigated further in a later publication.

We consider the following four spot shape models:

The cylindrical shape model. Let

$$h_s(r) = \frac{1}{\pi \sigma_s^2} 1(r \le \sigma_s), \tag{6}$$

where $1(P) = 1$ if P is true and $1(P) = 0$ otherwise. The parameter $\sigma_s > 0$ is the radius of the spot.

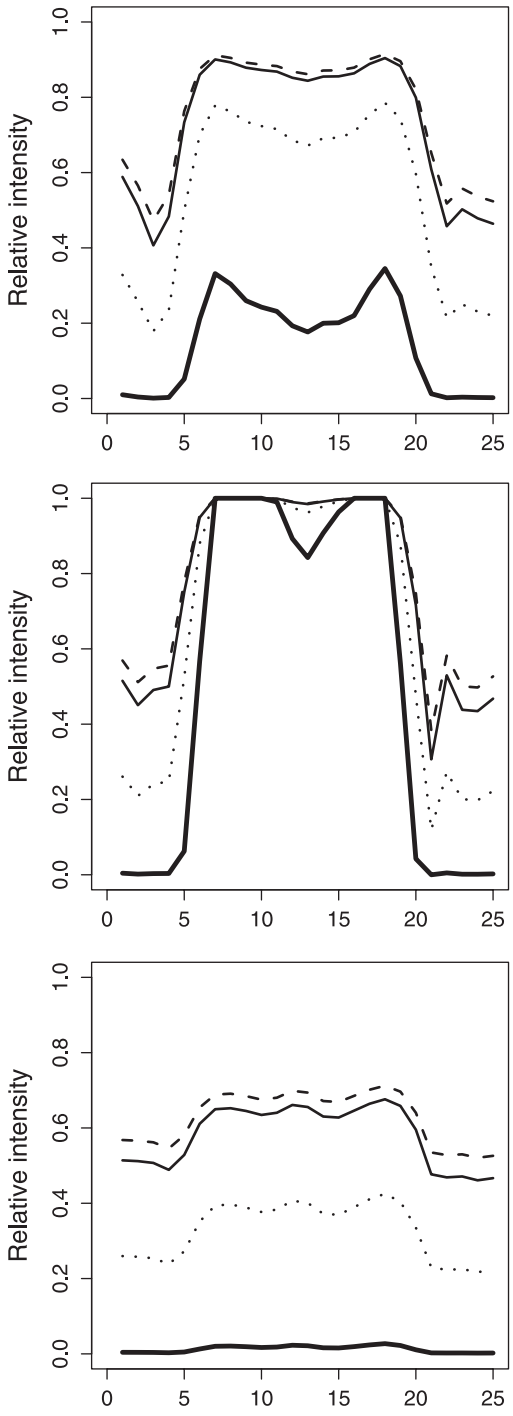

Fig. 2. Examples of transformations of spots 102, 106 and 108 at gain 60. Original data (▬), logarithmic transformation (—), inverse hyperbolic sine (· · ·) and Box–Cox (– - –).

The Gaussian shape model. Put

$$h_s(r) = \frac{1}{\sqrt{2\pi}\sigma_s^2}\phi\left(\frac{r}{\sigma_s}\right), \tag{7}$$

where $\sigma_s > 0$ and ϕ is the standardized 1D normal density

$$\phi(r) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}r^2\right).$$

The Gaussian difference shape model. Let

$$h_s(r) = \frac{1+\alpha_s}{\sqrt{2\pi}\sigma_s^2}\phi\left(\frac{r}{\sigma_s}\right) - \frac{\alpha_s}{\sqrt{2\pi}(\beta_s\sigma_s)^2}\phi\left(\frac{r}{\beta_s\sigma_s}\right), \tag{8}$$

where $\sigma_s > 0$, $\alpha_s > 0$ and $0 < \beta_s < 1$.

A polynomial-hyperbolic spot shape family. Put

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

where $I \ge 2$, $a_s > 0$ and $\gamma_s > 1$. Put further

$$h_s(r) = \begin{cases} \frac{K_s}{\sigma_s^2}\exp(g_s(r/\sigma_s)) & \text{if } 0 \le r < \gamma_s\sigma_s \\ 0 & \text{if } r \ge \gamma_s\sigma_s, \end{cases} \tag{9}$$

where σ_s represents the radius of the spot and $\sigma_s\gamma_s$ is the distance from the spot centre where there are no more signal from the spot. The constant K_s is a function of the parameters $b_{s1}, \ldots, b_{sI}, a_s, \gamma_s$ such that

$$\int_0^\infty \int_0^{2\pi} h_s(r) r dr d\theta = 1,$$

a condition that is also satisfied by the spot shapes (6), (7) and (8). The parameters a_s and γ_s determine the steepness of the spot edge. It may be noted that the spot function (9) is zero outside a circle around the centre for $r > \gamma_s\sigma_s$, similar to the cylindric spot function (6), which is zero for $r > \sigma_s$. While the cylindric spot shape function is discontinuous, the function (9) is continuous and infinitely differentiable. However, the cylindric spot shape may be obtained as a limiting case of the polynomial-hyperbolic spot shape, see below.

We require the boundary condition

$$g_s'(0) = 0, \tag{10}$$

i.e. that the spot intensities are flat near the centre of the spot. Most often, we would also require that

$$g_s'(1) = 0, \tag{11}$$

such that the spot intensity starts to decrease at value 1 (i.e. when the pixel is at distance σ_s away from the spot centre).

For $I = 2$, the boundary conditions (10) and (11) result in the following constraints on the parameters in the polynomial:

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

If we here let a_s tend to zero and γ_s tend to one we get the the cylindric spot shape as a limiting case.

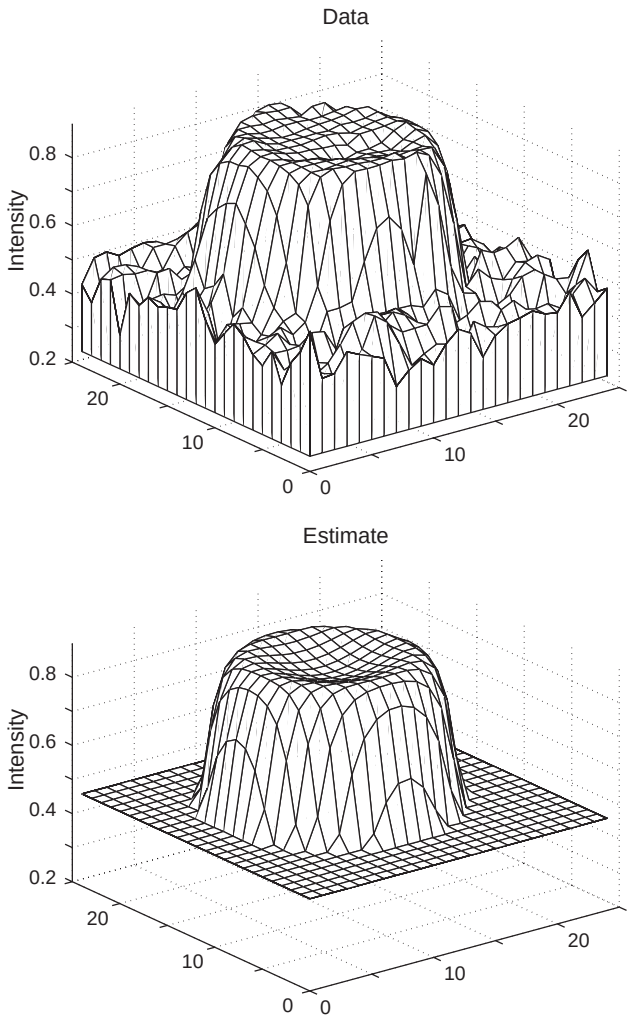

Fig. 3. Three-dimensional plot of observed intensities for spot 102 at gain 60 (top surface) and estimated spot shape from the polynomial-hyperbolic shape model (bottom surface).

We will use the condition (11) in the sequel but it may be noted that if we specify $g_s'(1)$ as a negative constant we may obtain spot shapes with a dome shape, and if we specify $g_s'(1)$ as a positive constant we may obtain more pronounced crater shapes.

Data and estimated spot shape for the polynomial-hyperbolic spot shape model are shown in 3D in Figure 3 for spot 102 at gain 60, corresponding to the upper curve in the right panel of Figure 1.

Fitted cylindrical, Gaussian, Gaussian difference and polynomial-hyperbolic shape functions corresponding to the spots in the right panel of Figure 1 are shown in Figure 4. It is clearly seen that the polynomial-hyperbolic spot shape model fits the data better than the three other models.

ESTIMATION OF PARAMETERS AND SATURATED (CENSORED) VALUES

Parameter estimation

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. Spot shape parameters may be estimated by maximizing the log-likelihood function

$$L_Y = L_1 + L_2, \tag{12}$$

where

$$L_1 = \sum_{x \in A_s'} \log \left\{ \frac{1}{\sigma_e} \phi \left(\frac{Y(x) - B_s h_s[r_s(x)] - b_s}{\sigma_e} \right) \right\}$$

and

$$L_2 = \sum_{x \in A_s''} \log \left\{ 1 - \Phi \left(\frac{\ell_c - B_s h_s[r_s(x)] - b_s}{\sigma_e} \right) \right\},$$

where ϕ and Φ are the standardized normal density function and distribution function, respectively. The log-likelihood (12) can be maximized by standard iterative maximization techniques, e.g. quasi-Newton or Nelder-Mead.

We note that if the spot shape parameters are varied individually for spots we get six parameters for the spot shape models (6) and (7): B_s, c_{s1}, c_{s2}, b_s, σ_e and σ_s, and eight parameters for the models (8) and (9). The additional parameters are α_s and β_s for model (8) and a_s and γ_s for model (9).

To estimate also parameters in the transformation (1), we maximize

$$L_Z = L_Y + \sum_x \log \left(\frac{\partial Y(x)}{\partial Z} \right), \tag{13}$$

Prediction of saturated (censored) values

For $x \in A_s''$, we denote the transformed estimated (predicted) intensity by

$$\hat{Y}(x) = \hat{B}_s \hat{h}_s(r_s(x)) + \hat{b}_s,$$

where $\hat{B}_s$ and $\hat{b}_s$ denote estimated parameters and $\hat{h}_s$ denotes the spot shape function with estimated parameters. If f is the transformation employed, e.g. (2), (3) or (4), then the corresponding estimated intensity is

$$\hat{Z}(x) = f^{-1}(\hat{Y}(x)).$$

Once the predicted values for the saturated pixels are obtained, we can plug in these values and analyze the spot as if all pixels were completely observed.

RESULTS

Choice of transformation and spot shape model

The Box–Cox (3) and the inverse hyperbolic sine transformation (4) both contain the logarithmic transformation (2) as

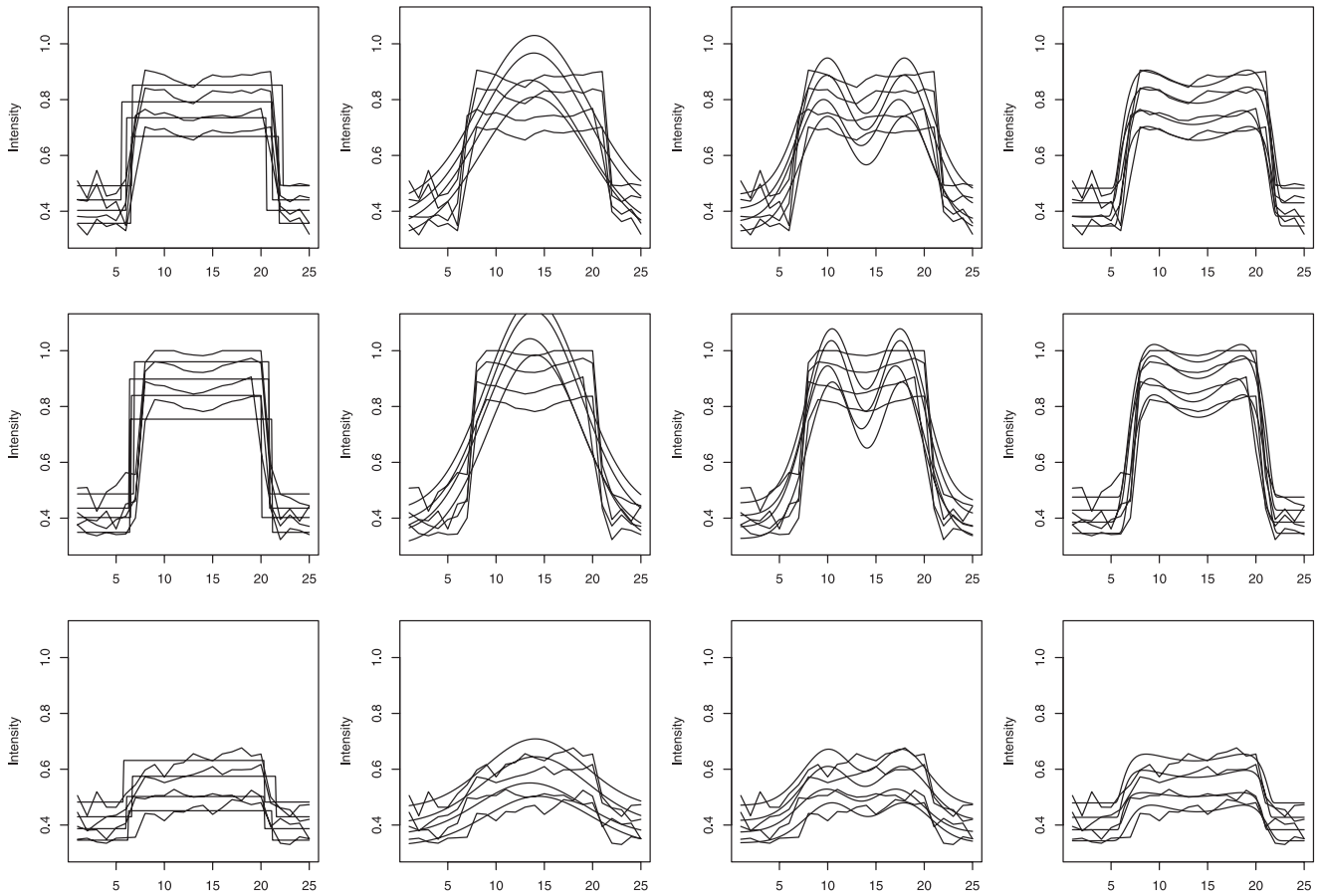

Fig. 4. Horizontal intensity profiles through the centres of spots 102, 106 and 108 (each spot represented by a row) together with estimated spot shapes for each of the four spot shape models: cylindrical, Gaussian, Gaussian difference, polynomial-hyperbolic (corresponding to each column) for gains 30, 40, 50 and 60. Estimates are, for each spot, based on pixels in 25×25 area, but are here (as well as in Fig. 5) displayed as profiles through the spot centre.

Table 1. Comparison of transformations and spot shape models

Transformation	Spot shape model			
	Cylindrical	Gaussian	Gaussian difference	Polynomial-hyperbolic
Logarithm[a]	171.57	330.45	191.98	57.79
Logarithm	136.30	329.60	185.41	17.00
Arsinh	127.19	258.70	144.39	13.86
Box–Cox	134.32	320.30	178.23	0.00

Median increase in log-likelihood (13) for 25 spots and four gains relative to polynomial-hyperbolic spot shape model with Box–Cox transformation.
[a]Fixed λ_1 to 1.

a special case. Thus we can use log-likelihoods to test if either of them gives a significant improvement relative to the logarithmic transformation for a given spot shape model.

The results shown in Table 1 are based on the analysis of 25 spots and for each of them four gains, which gives 100 datasets. The choice of spots was made so that both low, median and high-intensity levels were represented but with a slight over-representation of high intensities as one of our main objective was to study reconstruction of spots with saturated pixels.

Table 1 shows that the polynomial-hyperbolic spot shape model clearly turned out superior to the other spot shape models studied with the log-likelihood as criterion. This was also suggested by Figure 4 where the polynomial-hyperbolic shape gives a considerably better fit than the other three spot shape models. The analyses also show, that the

Box–Cox transformation provides the best transformation for the polynomial-hyperbolic spot shape model, while the inverse hyperbolic sine transformation yields better fits for the cylindrical, Gaussian and Gaussian difference models.

Interestingly, the second best fit is provided by the simple cylindrical model while the two Gaussian models give the worst fit. This was also suggested by Figure 4 where the Gaussian models—in contrast to the cylindrical and the polynomial hyperbolic shape models—fit equally bad on the spot boundary and at the spot centres.

For comparison, we fitted the polynomial-hyperbolic spot shape model using the Box–Cox transformation with fixed values of $\lambda_1 = 1$ and $\lambda_2 = 0.2$. This model provides a better median fit (median log-likelihood increase of 7.03) than does the logarithmic and inverse hyperbolic transformations with variable λ values. The logarithmic transformation with variable offset parameter λ_1 turned out to be considerably better than the standard logarithmic transformation with fixed $\lambda_1 = 1$ (we use $\lambda_1 = 1$ as fixed value rather than $\lambda_1 = 0$ as some pixel values were zero).

A priori, we can not formally test the Box–Cox and the inverse hyperbolic sine transformations against each other as the statistical models are not nested. However, for the polynomial-hyperbolic spot shape model it turned out that in most of the 100 datasets the logarithmic and inverse hyperbolic sines were close, while the Box–Cox transformation gave a considerable improvement relative to the logarithmic transformation. Therefore, we conclude that the Box–Cox transformation also was superior to the inverse hyperbolic sine in the present study.

Selected median parameter estimates from the polynomial-hyperbolic spot shape model with Box–Cox transformation were $\hat{a}_s = 0.68$, $\hat{\gamma}_s = 1.75$, $\hat{\lambda}_1 = 0.99$ and $\hat{\lambda}_2 = 0.185$.

Reconstruction of saturated values

Figure 5 shows the estimated spot profiles for the polynomial-hyperbolic spot shape model when the pixels for spots 242, 352 and 787 are artificially censored at different intensities. These three spots were chosen as those with the highest level not exceeding the upper limit. Thus the the leftmost diagrams show for each of these spots the estimate without censoring, while the other diagrams show reconstruction for varying degrees of censoring. In these diagrams, the Box–Cox transformation was used with fixed λ-values, $\lambda_1 = 1$ and $\lambda_2 = 0.2$. The parameters a_s and γ_s were also fixed and chosen empirically to mimic the results from the previous section, $a_s = 0.65$ and $\gamma_s = 1.75$.

The conclusion from Figure 5 is that with a small degree of censoring corresponding to the second column in Figure 5 the reconstruction is satisfactory. For a higher degree of censoring corresponding to the third column in Figure 5 we get some overshoot. With increasing degree of censoring an improvement is in fact seen in the fourth column, while the rightmost column corresponding to censoring at level 0.6 gives a clear undershot. This undershot is even more pronounced for censoring at level 0.5 (data not shown).

DISCUSSION

In this paper, we consider models for spot intensities on the pixel scale and different transformations to approximate normality and variance constance.

An empirical observation is that a logarithmic transformation with no offset is found to result in non-homogeneous variation: low-range pixel intensities show larger variation than mid-range or high-range pixels. The results from the analyses show that inclusion of an offset λ_1 improves the logarithmic transformation and that a further improvement is obtained with either the inverse hyperbolic sine or the Box–Cox transformation. The proposed polynomial-hyperbolic spot shape model (9) is more flexible than both the cylindrical, Gaussian and Gaussian difference models and is found to provide by far the best fit (Table 1). The results from Table 1 also indicate, that a value of λ_2 near 0.2 (i.e. the 5th root) is optimal for the Box–Cox transformation in combination with the polynomial-hyperbolic spot shape model.

The results seen in Figure 5 indicate that with a small percentage of censoring (<30%, say) it should be possible to estimate parameters and predict pixel intensities for the censored pixels in a satisfactory way. An obvious consequence of this is, that the photometric laser gain in some situations may be increased such that some pixels are saturated in order to improve the pixel intensities of the low intensity spots without any serious loss of information for the spot with highest pixels intensities.

Figure 5 also suggests that some pixels from the spot centre need to be observed in order to estimate censored pixel values well. When only the edge and the background pixels of the spot are observed (corresponding to the last column with artificial censoring at level 0.6 and even more pronounced at level 0.5, data not shown), the polynomial-hyperbolic spot shape models has difficulties in reconstructing the non-observed saturated pixel values.

It should also be possible to combine several runs with varying gains, compare the right panel in Figure 1. For spots with saturated pixels, pixel values may be reconstructed as shown in this paper. But if the censoring is too hard the corresponding estimate should be down-weighted when combined with signal intensities for runs with lower gains. To find optimal weights further studies are necessary.

The proposed spot shape model may be improved by considering several spots and/or combining data from both channels simultaneously. It can be argued that all spots should share the same transformation parameters, λ_1 and λ_2, such that their estimates should be based on the joint analysis of all spots. With measurements in two channels additional information may be gained by estimating parameters common for both channels simultaneously. In particular, the spot centre

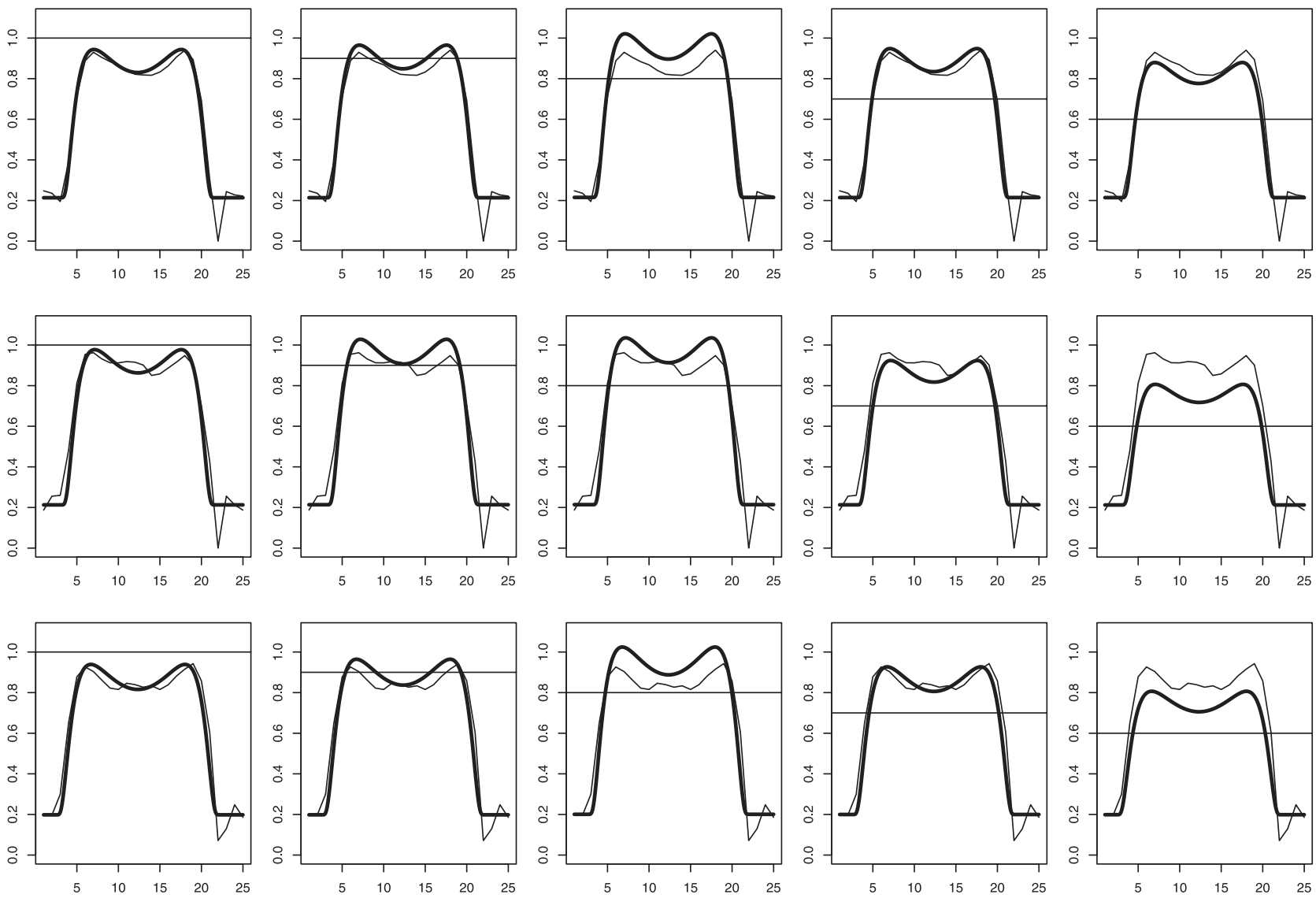

Fig. 5. Horizontal intensity profiles through the centres of spots 242, 352 and 787 (each spot represented by a row) at photometric gain 60 for the polynomial-hyperbolic spot shape model for different levels of (artificial) censoring as indicated by a horizontal line. For each profile both data (thin lines) and the reconstruction 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.

parameters, $c_s = (c_{s1}, c_{s2})$ and the spot size, σ_s are obvious choices, since they should be identical for both channels.

We conclude by listing some additional items that may be studied by use of a good spot shape model:

- To find accurate estimates of the local background level. We note that the model (5) contains such a parameter b_s for the local background at spot s.
- To make a quality control by finding spots that deviate in some way as may be seen in left panel in Figure 1 (e.g. the second spot to the right of the middle circled spot and several spots in the upper part of the chip).
- To find improved estimates of spot centres and spot diameters. It is also possible that the estimate of the parameter B_s in (5) could be used to estimate the total intensity of a spot, but we rather think that an average within an accurately determined circular disk would give a more robust intensity estimate for spots with all pixels uncensored.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for valuable comments that lead to improvement of the manuscript. This work was supported by Danish National Research Foundation (C.K. and S.B.) and The Danish Veterinary & Agricultural Research Council (C.K. and S.B.).

REFERENCES

Box,G.E.P. and Cox,D.R. (1964) An analysis of transformations. *J. R. Stat. Soc., Ser. B,* **26**, 211–252.

Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nat. Genet.*, **21**(suppl.), 10–14.

Durbin,B.P., Hardin,J.S., Hawkins,D.M. and Rocke,D.M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**(Suppl. 1), S105–S110.

Glasbey,C. and Ghazal,P. (2003) Combinatorial image analysis of DNA microarray features. *Bioinformatics*, **19**, 194–203.

Huber,W., von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl. 1), S96–S104.

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 Res.*, **28**, 4552–4557.

Paquette,S.M., Bak,S. and Feyereisen,R. (2000) Intron–exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.*, **19**, 307–317.

Paquette,S.M., Møller,B.L. and Bak,S. (2003) On the origin of family 1 plant glycosyltransferases. *Phytochemistry*, **62**, 399–413.

Rocke,D.M. and Durbin,B. (2003) Approximate variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **19**, 966–972.

Tattersall,D.B., Bak,S., Jones,P.R., Olsen,C.E., Nielsen,J.K., Hansen,M.K., Høj,P.B. and Møller,B.L. (2000) Resistance to an herbivore through engineered cyanogenic glucoside synthesis. *Science*, **293**, 1826–1828.

Werck-Reichhart,D., Bak,S. and Paquette,S.M. (2002) Cytochromes P450. In Somercille,C.R. and Meyerowitz,E.M. (eds) *The Arabidopsis Book*. American Society of Plant Biologists, Rockville, MD.

Wierling,C.K., Steinfath,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.

Wit,E. and McClure,J. (2003) Statistical adjustment of signal censoring in gene expression experiments. *Bioinformatics*, **19**, 1055–1060.