Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 19

Using Linear Mixed Models for Normalization of cDNA Microarrays

Philippe Haldermans* Ziv Shkedy† Suzy Van Sanden‡
Tomasz Burzykowski** Marc Aerts††

*Hasselt University, philippe.haldermans@uhasselt.be
†Hasselt University, ziv.shkedy@uhasselt.be
‡Hasselt University, suzy.vansanden@uhasselt.be
**Hasselt University, tomasz.burzykowski@uhasselt.be
††Hasselt University, marc.aerts@uhasselt.be

Using Linear Mixed Models for Normalization of cDNA Microarrays*

Philippe Haldermans, Ziv Shkedy, Suzy Van Sanden, Tomasz Burzykowski, and Marc Aerts

Abstract

Microarrays are a tool for measuring the expression levels of a large number of genes simultaneously. In the microarray experiment, however, many undesirable systematic variations are observed. Correct identification and removal of these variations is essential to allow the comparison of expression levels across experiments. We describe the use of linear mixed models for the normalization of two-color spotted microarrays for various sources of variation including printtip variation. Normalization with linear mixed models provides a parametric model of which results compare favorably to intensity dependent normalization LOWESS methods. We illustrate the use of this technique on two datasets. The first dataset contains 24 arrays, each with approximately 600 genes, replicated 3 times per array. A second dataset, coming from the apo AI experiment, was used to further illustrate the methods. Finally, a simulation study was done to compare between methods.

KEYWORDS: normalization, microarrays, linear mixed model, LOWESS

---

*The authors gratefully acknowledge the financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy) and the referees for their valuable remarks.

1 Introduction

During the last years a big expansion in the research on functional genomics has been observed. Miniaturization and automatization made it possible to determine, with the use of microarrays, which genes are differentially expressed and under which circumstances. We usually make a distinction between two types of microarrays: oligonucleotide and spotted c-DNA arrays. In this article we will focus on the second type.

In a standard cDNA microarray experiment two mRNA samples are compared by reverse transcribing them to cDNA, labeling them with green (Cy3) and red (Cy5) fluorescent dyes, and allowing them to hybridize with the DNA on the microarrays. The gene expression level of the two samples are measured to determine the ratio of the two signals. This forms a good approximation of the mRNA concentrations in the two samples.

The ratios are however subject to systematic errors, which can cause considerable biases. In order to carry out a meaningful analysis, we first have to normalize the data to remove these errors.

In order to detect systematic (non-linear) effects, Dudoit *et al.* (2002) proposed to use MA plots. These plots present the difference between the log intensity readings of the two channels ($M = \log_2 R/G$) versus their mean log-value ($A = \log_2 \sqrt{(RG)}$). If only a small number of genes is differentially expressed, it is expected to see a horizontal curve around zero. This is however often not the case. Moreover, we sometimes find that the variability of the values in the MA-plot changes according to the mean intensity level. Therefore, normalization is needed to restore the original horizontal curvature around zero.

In recent years, several methods have been proposed for normalization. Park *et al.* (2003) discussed a number of methods and suggested choosing the appropriate method according to the nature of the data. They arrange the methods in a flowchart going from a simple global normalization method to more complex models. The global normalization method assumes that the red and the green intensities are related by a constant factor (Yang *et al.*, 2002) and the normalization model can be expressed as $\mathbf{M} = \boldsymbol{\beta_0} + \boldsymbol{\varepsilon}$. Here, $\boldsymbol{\varepsilon}$ is a random error vector. Hence, in the MA plot M is independent of intensity. The linear normalization model, $\mathbf{M} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathbf{A} + \boldsymbol{\varepsilon}$, allows for linear dependence on intensity. Nonlinear intensity dependent normalization models are expressed as $\mathbf{M} = \mathbf{f}(\mathbf{A}) + \boldsymbol{\varepsilon}$. Here $\mathbf{f}(\mathbf{A})$ is assumed to be a nonlinear function of $\mathbf{A}$. Lowess smoothers (Cleveland, 1974)) are commonly used in order to estimate $\mathbf{f}(\mathbf{A})$, (Park *et al.*, 2003, Yang *et al.*, 2002, Dudoit

et al., 2002 and Fan *et al.*, 2003). Other nonlinear normalization methods such as B-splines, wavelets, kernel smoothers and support vector regression are discussed by Fujita *et al.* (2006) . Wolfinger *et al.* (2001) and Kerr *et al.* (2000) discussed the use of an ANOVA model for normalization. In the first step, data are normalized using a two-way ANOVA model (with treatment, array and treatment $\times$ array as fixed effects in the model). In the second step, the residuals obtained from the "normalization" model are used as the response variable in the "gene" model. This type of model makes it possible to account for experiment-wide systematic effects in a formal statistical way.

Dudoit *et al.* (2002) present a print-tip dependent normalization method, where they use a lowess fit for each print-tip. They show that there might be a strong print-tip or spatial effect and that it seems preferable in certain cases to normalize per print-tip and not for the whole array at once.

The focus of this paper lies on normalization using linear mixed models (LMM) as scatterplot smoothers of the MA plot. We use the methodology discussed in Ruppert *et al.* (2003) and take different systematic error possibilities into account. The presence of several replicates of one gene is also taken into consideration. The use of LMM allow us to incorporate the moralization models proposed by Dudoit *et al.* (2002) and Park *et al.* (2003) into one general framework in which all normalization models can be expressed as LMM. The normalization model which was used to normalize the data is the model with the best goodness-to-fit in the MA plot.

For illustration purposes cDNA microarray data is used from two vegetable studies designed to investigate the effect of certain vegetable diets on the gene expression in colon and lung tissue of mice (van Breda *et al.*, 2005). In both the lung and colon study, three samples of material pooled from two or three mice were available for the control group and four treatment groups. Using each of the three sets of pooled samples, the four treatment groups were compared to a control group by applying a reference design with dye-swap. Thus, in both studies 24 arrays were used in total, each containing about 600 genes that were spotted three times on every slide. The signal intensity of a spot was determined by taking the mean of the intensities of all pixels that fell within the area declared to be the spot by the scanner software.

Furthermore, we use the dataset from Yang *et al.* (2002) and Dudoit *et al.* (2002), referred to as the apo AI experiment, as a second example to illustrate the framework. In this experiment, target cDNA was obtained from eight mice for the control and treatment group respectively. These 16 microarrays consisted of 6384 cDNA probes, spotted onto the glass slides with 16 print heads. For further details about the apo AI experiment we refer to

Yang *et al.* (2002) .

The paper is organized as follows. In Section 2 we discuss the main ideas of several normalization methods and in particular, in Section 2.2 we review the linear mixed models and their application to cDNA microarray normalization. In Section 3 we illustrate the use of the proposed framework on two datasets mentioned above, while Section 4 is devoted to a simulation study in which the performance of normalization models is evaluated. All SAS and R code needed to fit the models discussed in this paper can be found in the supplemental materials for the paper and can be downloaded from the website `http://www.censtat.uhasselt.be/software/`.

2 Normalization of Microarray Data Using Linear Mixed Models

2.1 Smoothing with linear mixed models

Linear mixed models (Laird and Ware, 1982 and Verbeke and Molenberghs, 2000) are commonly used to describe the relationship between a response variable and a predictor(s) when the observations in the datasets are clustered according to a known grouping factor(s). Linear mixed models can be formulated as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is a vector of n observed random variables, $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices of dimension $n \times p$ and $n \times q$ respectively, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters representing the fixed effects, $\boldsymbol{b} \sim \mathrm{N}(0,\sigma_b^2)$ is a $q \times 1$ vector of random effects and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0,\sigma^2)$ is a $n \times 1$ vector of unobserved measurement errors.

In recent years there is an increasing interest in the statistical literature with respect to the connection between LMM and smoothing splines (Ruppert *et al.*, 2003 and Verbyla *et al.*, 1999) . The latter are used to estimate nonparametrically an unknown smooth function f when the data are assumed to follow a regression model of the form $\mathbf{Y} = \mathbf{f} + \boldsymbol{\varepsilon}$. In particular, Ruppert *et al.* (2003) considered the model

$$f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k \phi_k (x_i - t_k)_+, \; i = 1, \ldots, n, \; k = 1, \ldots, K \tag{2}$$

where x_i are the design points, $t_1, ..., t_K$ are knots chosen in advance and the basis function $\phi_k(x)$ is constructed in the following manner

$$\phi_k(x - t_k)_+ = \begin{cases} 0 & x \leq t_k, \\ x - t_k & x > t_k. \end{cases} \tag{3}$$

For a microarray with m genes, let A_i, $i = 1, \ldots, m$, be the mean intensity in the MA-plot, with $A_i = \log_2 \sqrt{R_i G_i}$ and $M_i = \log_2 \sqrt{R_i / G_i}$. We define two design matrices, an $m \times 2$ design matrix for which the ith row is $\boldsymbol{X}_i = [1, A_i]$ and an $m \times K$ matrix for which the ith row is $\mathbf{Z}_i = [(A_i - t_1)_+, \ldots, (A_i - t_K)_+]_{1 \leq k \leq K}$. Hence, the model in (2) can be rewritten as $\mathbf{M} = \mathbf{f}(\mathbf{A}) + \boldsymbol{\varepsilon}$, with $\mathbf{f}(\mathbf{A}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{b} = (b_1, \ldots, b_K)$. Thus, the model has the same form as the linear mixed model defined in (1). For the analysis presented in this paper we choose equally spaced knots. Following Ruppert *et al.* (2003) the kth knot is located at the (k+1)/(K+2) quantile. For the total number of knots, K, Ruppert *et al.* (2003) suggested K = (n/4, 35). In this paper the analysis were done using K=20, 30 and 40.

2.2 Normalization Models

Park *et al.* (2003) distinguished three normalization approaches: (1) global normalization, (2) intensity dependent linear normalization and (3) intensity dependent nonlinear normalization. Furthermore, Park *et al.* (2003) showed that all normalization methods can be expressed as regression models. As we argued above, the linear mixed model for normalization can be expressed as $\mathbf{M} = \mathbf{f}(\mathbf{A}) + \boldsymbol{\varepsilon}$. In this section we show that the three approaches can be expressed as a series of nested linear mixed models. A flowchart of the normalization procedures and their corresponding linear mixed models are shown in Figure 1.

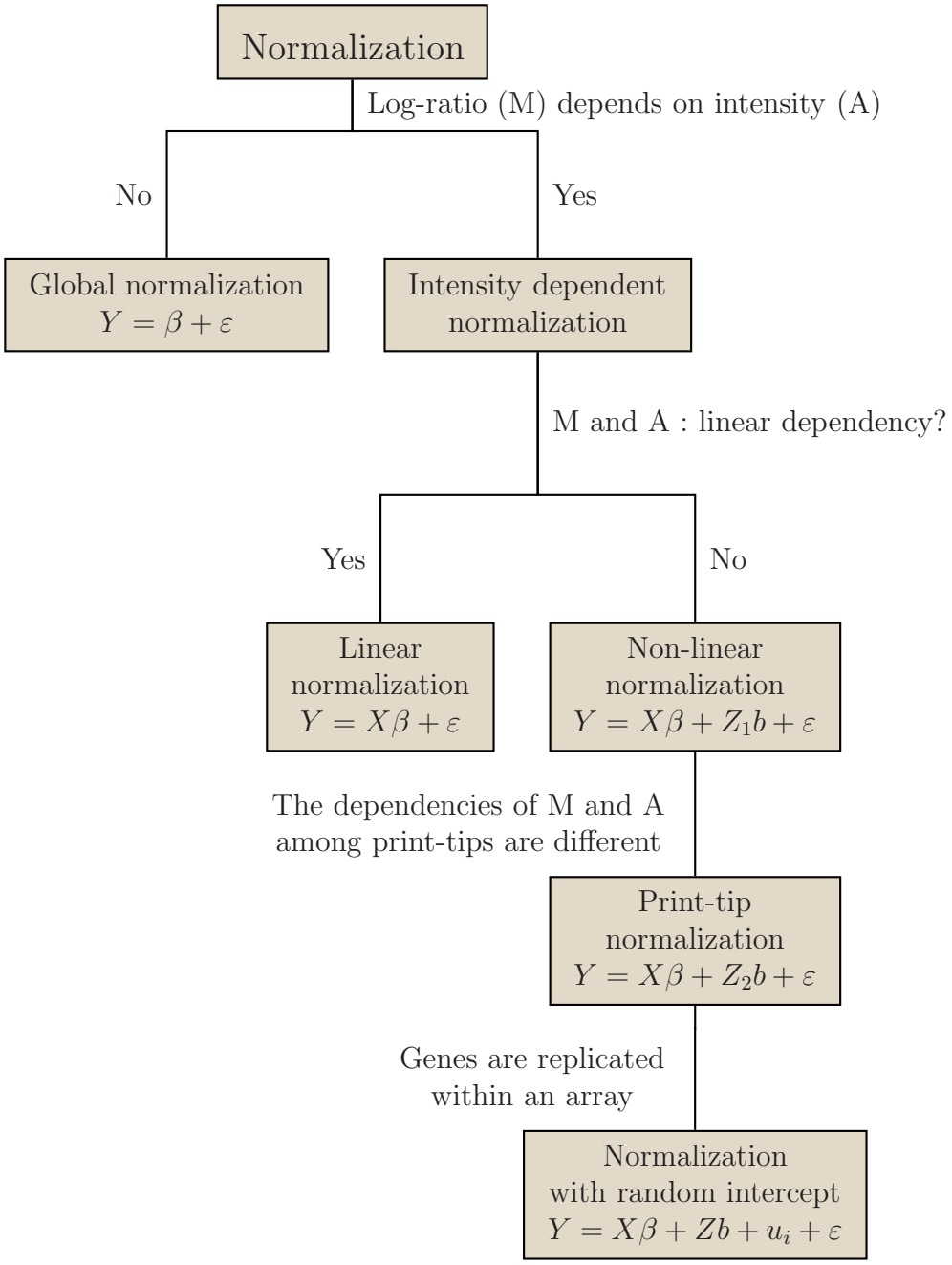


Figure 1: Normalization models and their linear (mixed) model formulation.

2.2.1 Global and linear normalization

For global and linear normalization, formulation of the normalization model as a linear model is straightforward (Park *et al.*, 2003) and the design matrices are given by

$$\text{global normalization: } \mathbf{X}_1 = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ 1 \\ 1 \end{bmatrix}, \text{ linear normalization: } \mathbf{X}_2 = \begin{bmatrix} 1 & A_1 \\ 1 & A_2 \\ . & . \\ . & . \\ 1 & A_{m-1} \\ 1 & A_m \end{bmatrix}.$$

As mentioned above, $\mathbf{f}(\mathbf{A}) = \beta_0$ and $\mathbf{f}(\mathbf{A}) = \beta_0 + \beta_1 \mathbf{A}$ for global and linear normalization models, respectively. Note that the covariance matrix for $\boldsymbol{\varepsilon}$ is, for both cases, equal to $\sigma^2\mathbf{I}$. These two models can be found in the upper part of the flowchart shown in Figure 1.

2.2.2 Nonlinear normalization

The third normalization model, the nonlinear normalization model, can be formulated using design matrix $\mathbf{X}_2$ for the fixed effects. The design matrix for the random effects $\mathbf{Z}$ is given by

$$\mathbf{Z} = \begin{bmatrix} A_1 - t_1 & A_1 - t_2 & ... & A_1 - t_K \\ A_2 - t_1 & A_2 - t_2 & ... & A_2 - t_K \\ . & . & . & . \\ . & . & . & . \\ A_m - t_1 & A_m - t_2 & ... & A_m - t_K \end{bmatrix}, \tag{4}$$

with $A_i - t_k$ as defined in the previous section. It follows that the three normalization models discussed above can be expressed as $\mathbf{M} = \mathbf{f}(\mathbf{A}) + \boldsymbol{\varepsilon}$, where

$$\mathbf{f}(\mathbf{A}) = \begin{cases} \boldsymbol{X}_1\boldsymbol{\beta}_0 & \text{global normalization,} \\ \boldsymbol{X}_2\boldsymbol{\beta} & \text{linear normalization,} \\ \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} & \text{nonlinear normalization.} \end{cases} \tag{5}$$

Alternatively, the normalization models can be expressed as $\boldsymbol{Y} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$. Within this model different parameterizations of the mean and the covariance structures lead to different normalization models:

$$\begin{array}{ll} \beta_1 = 0, \sigma_b^2 = 0 & \text{global normalization,} \\ \beta_1 \neq 0, \sigma_b^2 = 0 & \text{linear normalization,} \\ \sigma_b^2 \neq 0 & \text{nonlinear normalization,} \end{array} \tag{6}$$

with β_1 the slope and σ_b^2 the variance of the random effects in (2). The fact that different parameterizations of the mean and covariance leads to different normalization models allows us to use likelihood ratio tests or model selection criteria, such as the Akaike Information Criterion (AIC) to select the most appropriate normalization model, i.e. the model with the best goodness-of-fit for a specific array (Wager *et al.*, 2007) . Note that the normalization models in (5) are nested as one can test the null hypothesis $H_0 : \beta_1 = 0, \sigma_b^2 = 0$ against two sided alternatives in order to choose the most appropriate normalization model.

2.2.3 Print-tip Normalization

Thus far, we considered the whole array as one unit. However, sometimes it can be necessary to perform normalization on each grid separately, since each grid is printed by a different print-tip, as discussed by Dudoit *et al.* (2002) and by Park *et al.* (2003). In particular, Dudoit *et al.* (2002) use lowess normalization for each print-tip and state that it can solve spatial effects caused by the difference in print-tips.

Baird *et al.* (2004) proposed a normalization model in which the print-tip effect is included in the model as a constant (print-tip specific) fixed effect. For an array with L print-tips such a model can be implemented using a fixed term given by $[X_2|X_3][\beta|\gamma]$, where X_3 is an $m \times L$ matrix with the entry

$$[X_3]_{ij} = \begin{cases} 1 & \text{gene } i \text{ belongs to print-tip } j, \\ 0 & \text{else,} \end{cases} \tag{7}$$

and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_L)$ is the vector of print-tip specific parameters. Note that such a normalization model results in nonlinear normalization with parallel smoothers for the print-tips. As mentioned in Baird *et al.* (2004) print-tip effects can be included in the model as random effects,

$$\mathbf{M} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + a_\ell + \boldsymbol{\varepsilon}, \tag{8}$$

where a_ℓ, $\ell = 1, \ldots, L$ is a print-tip specific random intercept, $a_\ell \sim N(0, \sigma_a^2)$. L represents the number of print-tips on the array. Note that including the print-tip effects as either fixed or random lead to a normalization model in which the smoother for the print-tips is parallel lines.

Alternatively, print-tip specific nonlinear normalization models can be formulated as $\mathbf{Y} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{Z}_\ell\boldsymbol{b}_\ell + \boldsymbol{\varepsilon}$, with design matrix for the random

effects given by

$$\boldsymbol{Z}_\ell = \begin{bmatrix} \boldsymbol{Z}_1 & 0 & 0 & ... & 0 \\ 0 & \boldsymbol{Z}_2 & 0 & ... & 0 \\ . & . & . & ... & . \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 0 & 0 & 0 & ... & \boldsymbol{Z}_L \end{bmatrix}. \tag{9}$$

where $\boldsymbol{Z}_i = \boldsymbol{Z}$ $(i = 1, \ldots, L)$. Note that we assume that $\boldsymbol{b}_\ell = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_L)$ is normally distributed with mean zero and covariance matrix D, where

$$D = \sigma_b^2 \begin{bmatrix} \boldsymbol{I} & 0 & ... & 0 \\ 0 & \boldsymbol{I} & ... & 0 \\ . & . & ... & . \\ 0 & 0 & ... & \boldsymbol{I} \end{bmatrix}.$$

The fact that all random effects have the same variance σ_b^2 implies that the smoothing parameter (the ratio of the variance components $\sigma_b^2/\sigma_\varepsilon^2$) is the same for all print-tips.

2.2.4 Normalization with replicated genes

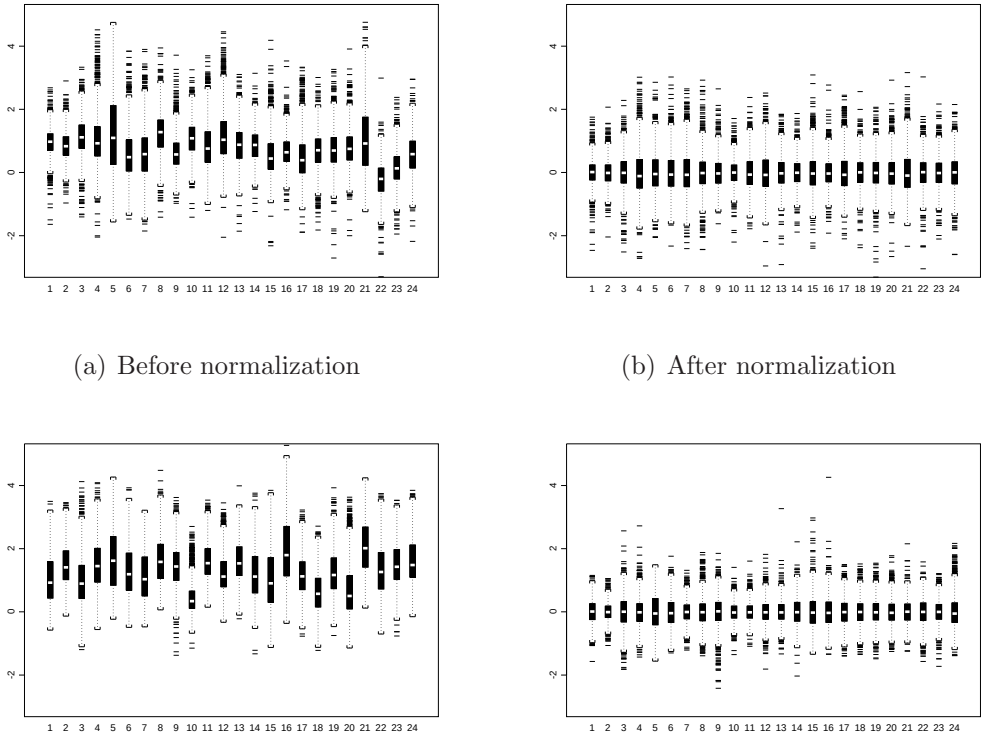
We can incorporate the fact that genes are replicated on an array by introducing a random intercept to the model. This results in the following model:

$$\mathbf{M} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{Z}_\ell\boldsymbol{b}_\ell + u_i + \boldsymbol{\varepsilon}. \tag{10}$$

Here, u_i is a random intercept for the ith gene with $u_i \sim N(0, \sigma_u^2)$. The $\mathbf{Z}_\ell$ matrix can take one of the two forms discussed in the two previous sections, depending on whether we choose to normalize by print-tip or not. This model is nested in the nonlinear model framework as can be seen in Figure 1. The covariance matrix for the random effects is adjusted in the following way:

$$D_1 = \left[\begin{array}{c|c} \sigma^2 I_{K\times K} & \\ \hline & \sigma_u I_{J\times J} \end{array}\right] \quad \text{and} \quad D_2 = \left[\begin{array}{c|c} D & \\ \hline & \sigma_u I \end{array}\right],$$

where D_1 and D_2 are the covariance matrices for nonlinear normalization and nonlinear pin by pin normalization, respectively. Note that it is assumed that the random intercept is uncorrelated with $\boldsymbol{b}$. It is important to mention that the predicted values for $\mathbf{M}$ are obtained using the empirical Bayes estimate for $\hat{\beta}$, $\hat{\mathbf{M}} = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{b}$. In that way we do not remove gene specific effect with the normalization model.



(a) Before normalization

(b) After normalization

(c) Before normalization

(d) After normalization

Figure 2: Boxplots before and after normalization. Panel a and b show the colon data, panel c and d the lung data.

3 Results

First, we applied the different normalization models discussed in the previous section to the datasets from the vegetable study mentioned in Section 1. The results are shown in Figure 2. Both boxplots represent the data separately for each array, before and after normalization. The figures reveal that the normalization centers the boxplots around zero, indicating that systematic errors were identified and minimized (Yang *et al.*, 2002). Since we applied several normalization models for each array, Akaike's Information Criterion (AIC) (Akaike, 1974) was used in order to select the best normalization model. An interesting observation is that there is not one model that is chosen all the time, indicating that there are indeed differences between arrays and that it is necessary to consider several possibilities. An objective measure for

comparison as AIC makes it possible to automate the framework to select the best normalization model for each array separately.

Table 1 presents the results for the normalization model discussed above. Model formulation for global, linear and nonlinear normalization models is given in (5). The random intercept model is a nonlinear normalization model which includes a random effect for the replicated genes, i.e. $\mathbf{f}(\mathbf{A}) = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + u_i$ with $u_i \sim N(0, \delta_u^2)$. The fixed pin model is a print-tip specific model including print-tip as a fixed effect, i.e. $\mathbf{f}(\mathbf{A}) = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \gamma_j$, with γ_j the fixed print-tip effect. Pin-by-pin refers to a print-tip specific model with a different smoother for each print-tip (see (9) for more details about the design matrix). The last model, pin-by-pin with random intercept, is a print-tip specific model with a random effect for the replicated genes as described in (10).

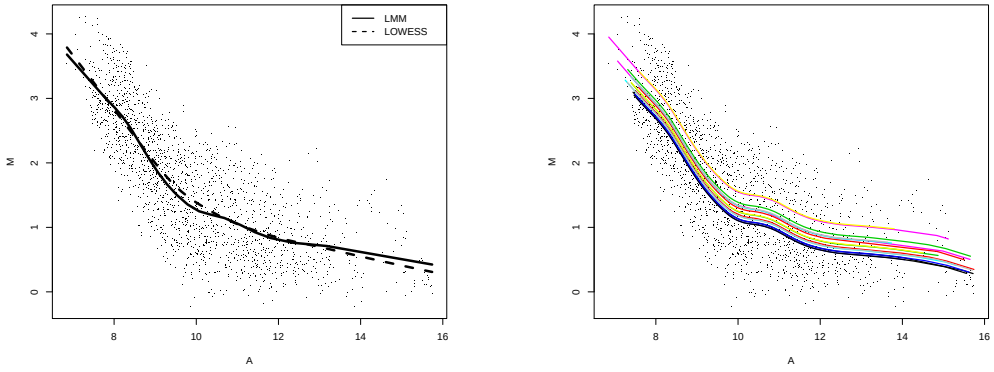
Table 1 presents the results for three arrays. The complete table is given in the supplemental material for the paper and shows that for none of the arrays the global, linear or the simple nonlinear model is preferable as the model to use for the normalization.

Table 1: AIC's of the different normalization models for the lung data

array	global	linear	non-linear	random intercept	fixed pins	pin by pin	pin by pin with random intercept
5	5018.940	3566.565	3039.857	3041.860	**3016.017**	3095.093	3097.095
15	4865.071	3429.412	2736.558	2737.403	**2656.092**	2684.612	2686.613
19	3962.601	2748.273	2344.326	2245.656	2179.008	2209.678	**2169.037**

In total, 11 out of 24 arrays have the non-linear model with random intercept coming out as best model, indicating that gene replication on an array often leads to a better model. For 9 of the 24 arrays non-linear normalization with fixed print-tip effects was considered to be the best model. In one of the cases, non-linear print-tip specific normalization is indicated as the preferred model. Finally, we find 3 arrays for which a combination of nonlinear normalization pin by pin with random intercepts is the best normalization model. Overall, we see that there does not exist a model which is the best for all arrays. Therefore we have to consider the choice of model for each array individually.

We turn now to discuss the results presented in Table 1 in more detail. Figure 3 (panel a) displays a MA plot for array 5 with the LMM smoother along with the lowess-smoother as a reference. The two smoothers reveal the same patterns, which confirms our expectations. Using the LMM smoother



(a) (b)

Figure 3: Panel a: MA plot of array 5 of the lung data with LMM and lowess-smoother. Panel b: normalization model with fixed print-tip effects.

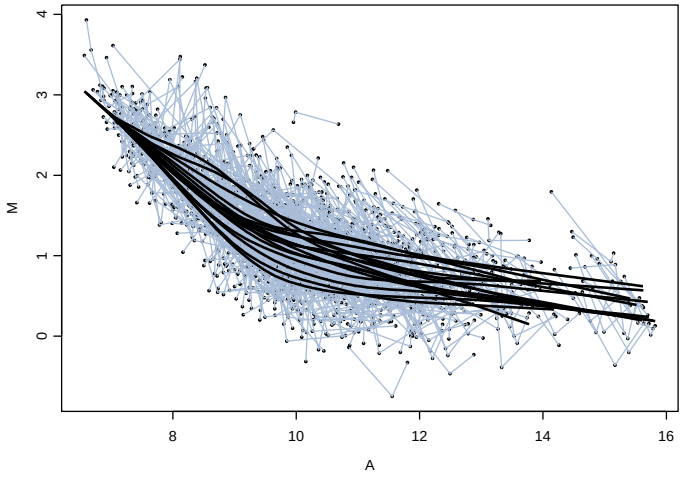


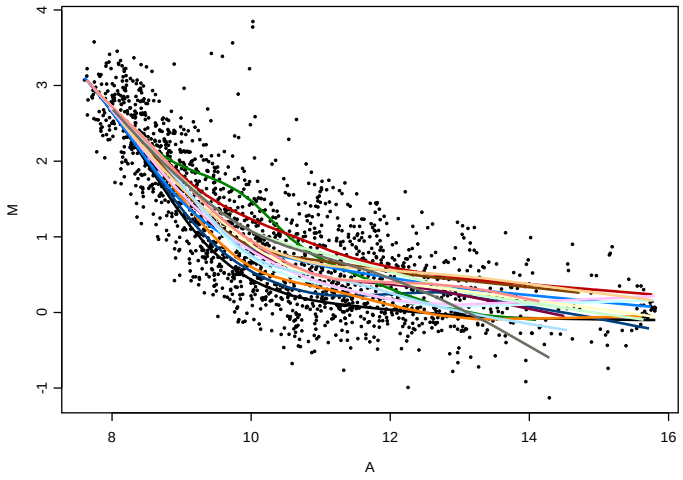
Figure 4: MA plot of array 19 of the lung data, for which the three replicates of a gene are connected by a line. A pin by pin normalization model with random effects is used to normalized the array.

allows us to choose the normalization model with the AIC as an objective model selector. For array 5, the model with the smallest AIC value is the one which includes fixed print-tip effects. This indicates that the print tip specific normalization model are non linear and parallel as shown in panel b. Note that since all normalization models are nested, a formal inference, using the likelihood ratio test, can be applied as well in order to select most appropriate normalization model. A more elaborate discussion on inference for LMM can be found in Verbeke and Molenberghs (2000).

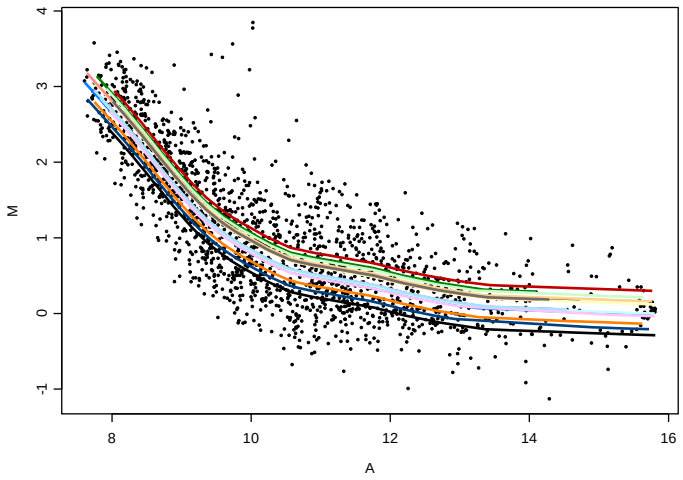
The use of a model with random intercepts is motivated by Figure 4, which shows the plot for array 19. Due to the fact that every gene has three replicates, there is a connection between the three spots.

Finally, two print-tip normalization models for microarray 15 are shown in Figure 5. Panel (a) presents a model uses the design matrix (9) for the random effect and implies nonlinear print-tip specific smoothers with the same smoothing parameter. The AIC for this normalization model is equal to 2684.612 (see Table 1). Panel (b) presents a second model which includes print-tip as fixed effects and implies that the smoothers for the print-tips are parallel lines. The AIC for this model is equal to 2656.002, indicates that second normalization model is to be preferred.

The framework was applied to the dataset from the apo AI experiment reported by Yang *et al.* (2002). The results are shown in Figure 6. Clearly, the normalization centers the boxplots around zero. This coincides with the results from Yang *et al.* (2002), although the variability in their study is bigger than in this study. Interestingly, the smallest AIC values for all arrays were obtained using nonlinear normalization with a different smoother for each print tip. This implies that there is a spatial effect in all of the arrays. This idea is confirmed by Dudoit *et al.* (2002), who state that within print-tip group dependent normalization is preferable here (AIC values are given in Table 2 in the supplemental material for the paper). Dudoit *et al.* (2002) refer to knock-out mouse number 8 to illustrate the need for print-tip specific normalization. They show that there is a difference between the print-tip groups. Specifically, they state that four groups clearly stand out from the others. From Figure 7, which shows the same array with the print-tip specific LMM smoothers, it can be seen that the same four groups have smoothers that differ significantly from the others.

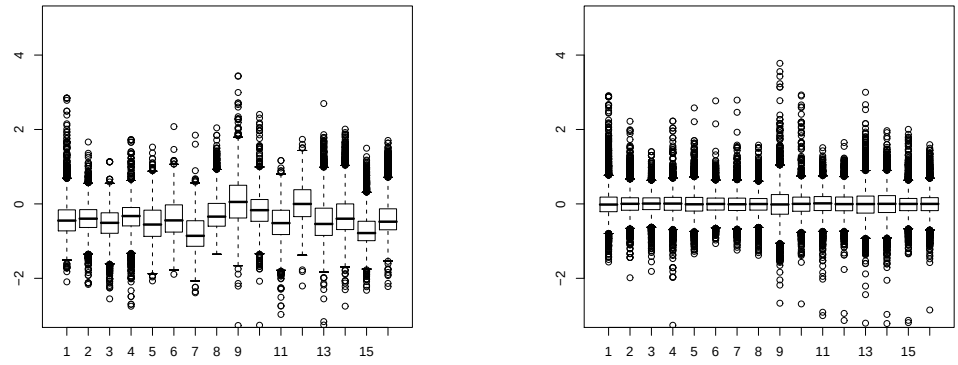


(a) Printtip specific normalization



(b) Normalization with fixed printtip effect

Figure 5: Panel a: MA plot of array 15 of the lung data, with print-tip specific nonlinear normalization. Panel b: MA plot of array 15 of the lung data, with print-tip as fixed effects which implies a nonlinear print-tip specific parallel normalization models.



(a) Before normalization (b) After normalization

Figure 6: Boxplots before and after LMM normalization for the apo AI experiment presented in Yang *et al.* (2002).

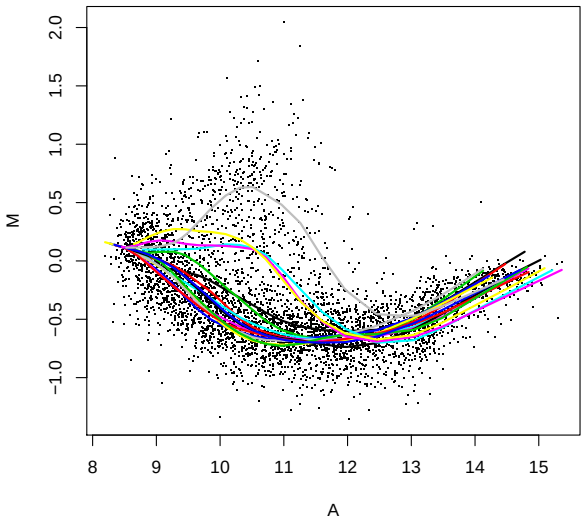


Figure 7: LMM normalization with different smoother for each printtip applied to array 8 of the apo AI experiment.

4 Simulation studies

4.1 Data generation

A simulation study was conducted in order to make an objective comparison between global, linear and non linear normalization (by lowess and linear mixed model normalization). Data were generated according to the setting discussed by Balagurunathan *et al.* (2002) and Fujita *et al.* (2006). The performance of the different normalization models were compared, similar to Fujita *et al.* (2006), by calculating the mean squared error between the estimated normalization model and the actual function from which the data were generated. In each simulation, we generate microarrays ranging from 600 to 10000 genes, 500 dataset were generated for each simulation setting. The generation of the data consists of the following steps:

1. Generate for each gene the true expression signal from an exponential distribution with $\lambda = \frac{1}{3000}$.

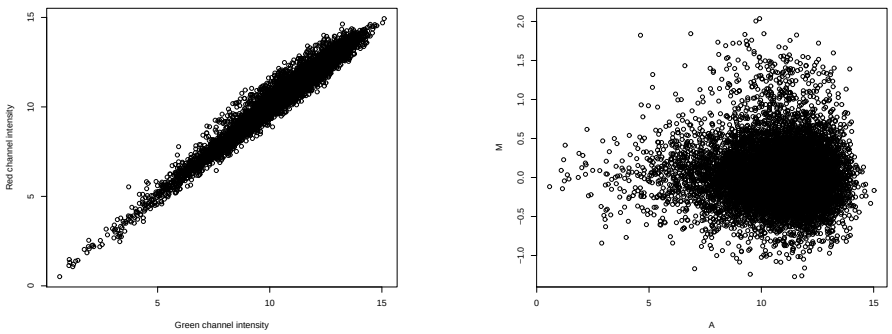
2. Simulate the red and green channel intensities for each gene from a normal distribution with mean the true expression signal from (1) and a standard deviation of 15% of the mean.

3. Include differentially expressed genes. We selected 5% of the genes to be either under- or overexpressed. The selected genes have a targeted expression ratio that is generated by $t = 10^{\pm b}$, where b follows a beta distribution, $b \sim B(1.7, 4.8)$. The expression intensity of this gene then is given by, $R' = R * \sqrt{t}$ and $G' = \frac{G}{\sqrt{t}}$ for the red and green channel respectively.

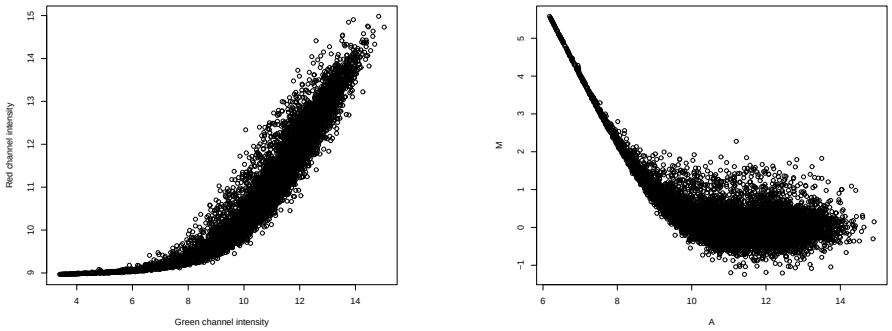
4. In order to transform these intensities to the often exhibited (non)linear patterns, we use the function family as given in Balagurunathan *et al.* (2002) by:
$$f(x) = a_3[a_0 + x(1 - e^{-x/a_1})^{a_2}] \tag{11}$$

5. Finally, the noise is added to the signal intensity of each channel.

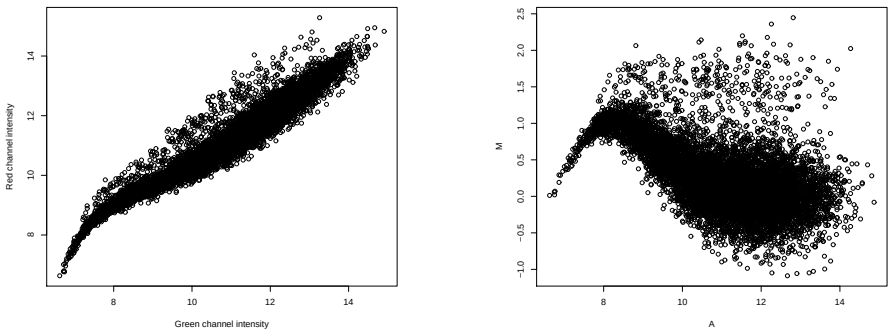
The three types of patterns as suggested by Balagurunathan *et al.* (2002) were considered. In the first setting, a pattern without alterations was considered (see Figure 8 panel a and b). Data for this pattern were generated with the parameters in (11) given by (0, 1, -1, 1) for the transformation of both red and green channel intensities . Secondly we applied a "banana shape" pattern (see Figure 8 panel c and d), where we use (0, 500, -1, 1) as parameters for



(a) (b)



(c) (d)



(e) (f)

Figure 8: Three possible patterns for the simulation study. Left panels: green versus red intensities. Right panels: the MA plots corresponding to each one of the patterns. Panel a and b: no trend, panel c and d: banana shape, panel e and f: sinusoid shape.

the transformation of the red channel intensities, and (0, 10, -1, 1) for the green channel intensities. The third pattern is a sinusoid shape (see Figure 8 panel e and f) with parameters for the red channel response function (0, $100^{1/0.7}$, -0.7, 1) and ((0, $100^{1/0.9}$, -0.9, 1) for the green channel. Each simulation setting was repeated three times,without printtip effects, with parallel printtips and with non parallel printtips.

4.2 Results

Results for the performance of the LMM normalization for the first scenario (i.e., microarray without print tip effects) are reported in Table 2. The results for the lowess method are presented for bandwidth equal to 0.4 and 0.75, corresponding respectively to the value suggested by Yang *et al.* (2002) and the default value in R. When data are generated without any trend, all normalization methods perform similarly. This confirms the results reported by Park *et al.* (2003). However, for the banana shape and sinusoid patterns the lowess and the LMM performed better than the global and linear normalization. This is in contrast with the finding reported in Park *et al.* (2003) that the linear and nonlinear normalization reveal only small difference. This can be explained by the choice of the systematic patterns in our simulation setting. When a clear systematic pattern is observed (such as in the banana shape and the sinusoid shape) the nonlinear normalization models performed better than the global and linear normalization model which are not able to capture these type of patterns in the data. Interestingly, the lowess and the LMM performed equally good, although the MSEs obtained for the lowess are slightly smaller than the MSEs obtained for the LMM for the second and the third patterns.

Table 3 reports the results for the second scenario (i.e., nonlinear, but parallel print tip effects). For this scenario the global, linear and lowess were applied for each print tip. The LMM was fitted in the usual way. Once again, for the nonlinear patterns the lowess and LMM performed much better than the global and linear normalization. Only small differences were observed between the lowess and the LMM. Although, the MSE values obtained for LMM are slightly smaller than the MSE values obtain for the lowess. Finally, similar results (presented in Table 4) were obtained for the third scenario in which data were generated with non parallel print-tip effect. We note that model selection based on the AIC criterion leads to a selection of the right model in more than 95% of the simulations (see Table 6-8 in the supplementary material).

Table 2: Simulation results without print-tip effect. All MSE values are multiplied by 10^4.

Trend	# genes	Method				
		Global	Linear	lowess (0.4)	lowess (0.75)	LMM
No trend	600	22.5	25.1	40.2	31.9	22.6
	2500	21.0	21.7	25.5	23.5	21.0
	5000	21.9	22.3	24.4	23.3	22.0
	7500	20.2	20.5	21.9	21.3	20.2
	10000	21.2	21.5	22.7	22.2	21.3
Banana	600	6030.0	8872.2	32.5	29.8	31.3
	2500	5867.2	9351.2	20.4	22.6	20.4
	5000	6454.6	10415.3	19.8	22.9	20.0
	7500	5703.0	9144.7	18.0	21.3	18.3
	10000	5963.3	9388.6	19.2	22.9	19.2
Sinusoid	600	954.7	1576.9	30.8	42.7	34.5
	2500	1056.4	1850.8	20.6	36.0	21.9
	5000	1016.7	1792.1	19.8	36.4	20.8
	7500	1002.4	1776.6	18.3	32.8	18.7
	10000	975.6	1678.6	18.5	35.5	18.8

In the simulation study discussed above, data were generated assuming 5% differentially expressed genes (DEG), from which half were assumed to be up regulated. A second simulation study was conducted in order to investigate the robustness of the normalization model for this particular assumption. Data were generated according to first scenario (with a banana shaped pattern) with 5%, 10%, 20% and 40% differentially expressed genes from which 30%, 40%, 50%, 60% or 70% were assumed to be up regulated. Only nonlinear normalization models were considered. Figure 9 presents the MSE for the lowess and the LMM (MSE values are given in Table 5 in the supplemental materials for the paper). When the proportion of DEG is relatively small (5% and 10%, respectively) the minimum MSE is obtained for the case that 50% of the genes are up regulated.

The MSEs increase slightly as the proportion of up regulated genes decrease or increase. The MSEs increase substantially, for both the lowess and the LMM, when the proportion of DEG is increase to 20% and 40%. However, it should be mentioned that it is assumed that the proportion of DEG

Table 3: Simulation results with parallel print-tips. All MSE values are multiplied by 10^4.

Trend	# genes	Method				
		Global	Linear	lowess (0.4)	lowess (0.75)	LMM
No trend	600	58.1	93.5	358.3	205.4	61.3
	2500	28.9	37.9	99.9	66.1	29.8
	5000	24.1	28.7	59.6	43.1	24.7
	7500	23.7	26.7	47.4	36.3	24.2
	10000	22.9	25.3	40.8	32.6	23.3
Banana	600	7177.4	12830.9	244.3	136.7	53.2
	2500	5927.3	9715.8	70.9	49.8	26.9
	5000	5989.6	9673.5	46.0	37.1	23.8
	7500	5796.8	9239.9	36.4	31.4	21.7
	10000	6220.4	10003.4	30.7	28.4	19.7
Sinusoid	600	1179.1	2263.3	243.5	141.3	53.7
	2500	1019.0	1788.2	70.4	58.6	27.5
	5000	986.7	1702.9	46.1	47.2	24.0
	7500	987.3	1716.9	35.9	41.0	21.9
	10000	1012.7	1762.5	30.7	40.2	20.1

is relatively small and we do not expect that any normalization method will perform well in case that the DEG is relatively high.

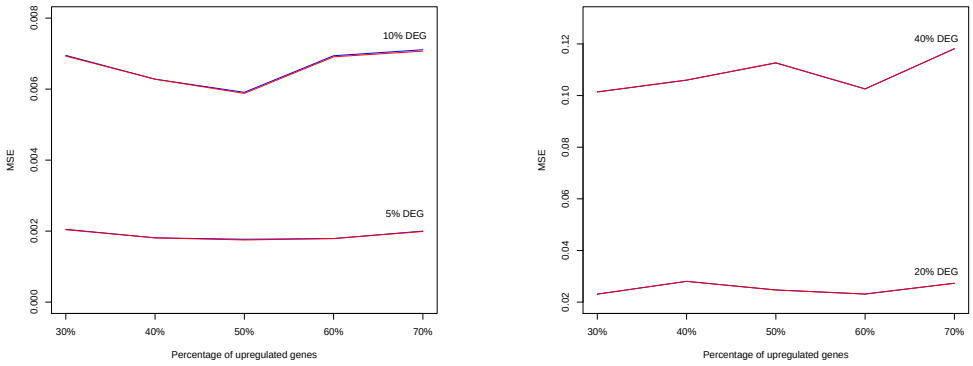
5 Conclusions

The MA-plots from the lung and colon studies clearly illustrate the need for normalization in microarray data. The plots show serious deviations from a straight line, which can cause problems in the analysis of the microarrays. In this paper we have used LMM as a smoother for the MA plot.

We have shown that different normalization models, global, linear and nonlinear, can be formulated as a LMM. The AIC criterion can be used to select the most appropriate normalization model. After the use of the LMM normalization, the curvature in the MA plots has disappeared.

Table 4: Simulation results with non-parallel print-tips. All MSE values are multiplied by 10^4.

Trend	# genes	Method				
		Global	Linear	lowess (0.4)	lowess (0.75)	LMM
Banana	600	6847.0	11128.4	263.0	143.7	55.5
	2500	5806.1	9223.3	76.0	53.7	30.4
	5000	6000.0	9551.3	45.7	37.1	23.5
	7500	5861.2	9194.0	37.9	33.4	22.7
	10000	6038.2	9707.7	32.0	29.7	20.9
Sinusoid	600	1022.9	1718.1	269.5	152.0	62.6
	2500	947.7	1616.1	75.2	64.3	30.2
	5000	974.9	1712.2	44.5	44.8	23.2
	7500	956.5	1632.4	34.7	41.8	20.5
	10000	973.7	1709.6	30.8	40.1	20.0



(a) (b)

Figure 9: MSE of normalization with LMM and lowess for different percentages of upregulated genes. Panel a: 5% and 10% of the genes are differentially expressed. Panel b: 20% of genes are differentially expressed.

We have shown that for the lung data and for the apo AI experiment, several forms of the model were needed, depending on the nature of the systematic errors. The LMM-smoother allows us to consider multiple types of normalization models within one framework, including more complex forms like the normalization per print-tip group. Different normalization models

can be defined by formulating an appropriate design matrix for the random effects in the model.

The results obtained from the simulation study have shown that when a clear pattern exists in the data nonlinear normalization models (both lowess and LMM) perform equally better than global and linear normalization models. For relative low levels of DEG the best performance of the model was achieved when half of the DEG are up regulated. The MSE increases slightly when 30% or 40% of the genes are up regulated.

For the analysis presented in this paper we follow Yang *et al.* (2002) and performed a within-slide normalization. In order to compare between slides, Yang *et al.* (2002) suggest the use of scale normalization. This might be useful in certain circumstances where changes in the settings cause scale differences between the arrays (which was not the case in our examples). The use of the LMM framework for normalization of cDNA microarrays may result in different normalization models for different arrays in the experiment. However, we have shown for both examples presented in the paper that the normalized data in the MA-plot are centered around zero. This implies that the systematic error was removed (or at least minimized), so the normalized data are comparable and the investigator can proceed to the inference step in order to detect differentially expressed genes. Note that the same approach was taken by Dudoit *et al.* (2002) who used the lowess model to perform a pin by pin nonlinear normalization. The lowess normalization presented in Dudoit *et al.* (2002) results in a different nonlinear normalization model for different print-tips (see for example Figure 3 in Dudoit *et al.* (2002)), and we obtain similar results (Figure 7). In both cases the systematic error was removed and therefore the normalized data are comparable.

Multiple array normalization can be formulated as a LMM as well by including the array as fixed or nonlinear effect. This issue is currently under investigation. The normalization models discussed in this paper assume a constant variance in the MA plot which, in some applications, might not be the case. The LMM can be reformulated in such a way that an intensity dependent variance can be incorporated in the model. This issue is a subject for future research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on automatic control*, 19:716–723.

Baird, D., Johnstone, P., and Theresa, W. (2004). Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics*, 20:3196–3205.

Balagurunathan, Y., Dougherty, E., Chen, Y., Bittner, M., and Trent, J. (2002). Simulation of cdna microarrays via a parameterized random signal model. *Journal of Biomedical Optics*, 7:507–523.

Chen, Y.-J., Ralph, K., Sistare, F., Thompson, K., Morris, S., and Chen, J. (2003). Normalization methods for analysis of microarray gene-expression data. *Journal of biopharmaceutical statistics*, 13:57–74.

Cleveland, W. (1974). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, 12(1):111–139.

Fan, J., Tam, P., Vande Woude, G., and Ren, Y. (2003). Normalization and analysis of cdna microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *PNAS*, 101:1135–1140.

Fujita, A., Sato, J. R., de Oliveira Rodrigues, L., Ferreira, C. E., and Sogayar, M. C. (2006). Evaluating different methods of microarray data normalization. *BMC Bioinformatics*, 7:469.

Kerr, M., Martin, M., and Churchill, G. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–74.

Park, T., Yi, S.-G., Kang, S.-H., Lee, S., Lee, Y.-S., and Simon, R. (2003). Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4:33.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

van Breda, S. G. J., van Agen, E., van Sanden, S., Burzykowski, T., Kienhuis, A. S., Kleinjans, J. C. S., and van Delft, J. H. M. (2005). Vegetables affect the expression of genes involved in anticarcinogenic processes in the colonic mucosa of c57bl/6 female mice. *Journal of Nutrition*.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Verbyla, A., Cullis, B., Kenward, M., and Welham, S. (1999). The analysis of designed experiments and longitudinal data using smoothing splines. *Journal of the Royal Statistics Society, Series C*, 48:269–311.

Wager, C., Vaida, F., and Kauermann, G. (2007). Model selection for p-spline smoothing using akaike information criteria. *Australian and New Zealand Journal of Statistics*. http://www.wiwi.uni-bielefeld.de/∼kauermann/research/WagerVaidaKauermann.pdf.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cdna microarray expression data via mixed models. *Journal of Computational Biology*, 8:625–637.

Yang, Y. H., Dudoit, S., Luu, P., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30:e15.