PSYCHOMETRIC MODELING OF STRUCTURE IN FLUORESCENCE INTENSITIES OF SNP ARRAYS

Ralph C.A. Rippe

data theory group fsw, leiden university, p.o. box 9555,
2300 rb leiden, the netherlands, rrippe@fsw.leidenuniv.nl

Paul H.C. Eilers

data theory group fsw, leiden university, p.o. box 9555,
2300 rb leiden, the netherlands, p.eilers@fsw.leidenuniv.nl

Jacqueline J. Meulman

mathematical institute, leiden university, p.o. box 9512,
2300 ra leiden, the netherlands,
jmeulman@math.leidenuniv.nl

December 19, 2008

Corresponding author: Jacqueline J. Meulman (jmeulman@math.leidenuniv.nl)

PSYCHOMETRIC MODELING OF STRUCTURE IN FLUORESCENCE INTENSITIES OF SNP ARRAYS

Abstract

In this paper we present applications of models from the psychometric toolbox in a case study of data from genomics. The paper is meant to serve two purposes. On the one hand, we wish to show that an "individual differences" model, where we use quantification (optimal scaling) of categorical variables, can be useful in modeling data from a domain other than psychometrics. On the other hand, we wish to arouse the interest of the psychometric community for challenging data that are currently generated in fields such as genomics, proteomics and metabolomics, as well as biological psychology. In the present case study, we analyze genetic measurements of allelic mutations, called Single Nucleotide Polymorphisms (SNPs), obtained for a number of persons. The data for the persons are captured in arrays. Each SNP contains a combination of two alleles, which are variant forms of genes, and each allele is associated with a categorical variable, with the three genotypes as categories. Our task is to quantify the categories of this categorical variable for each allele, and estimate an array effect that is shared by all variables (SNPs), as well a SNP effect that is shared by the persons (arrays). We use an additive model with restrictions on the SNP and person parameters, where the restrictions follow from both statistical as well as biological considerations. To label the different categories in the genotype variable, we used the results from a preliminary classification obtained from a mixture model. Our data analytic results show systematic variation between arrays, equally in both alleles, as well as systematic variations between the

SNPs (the variables). In one of the alleles, the genotype quantifications that we obtained in the optimal scaling process systematically deviate from theoretical difference factors. We show that these quantifications can be visualized as the centers of gravity of three clusters of SNPs, each denoting a particular genotype. We propose that the estimates for the parameters can be used to improve the genotype classification.

Key words: high-throughput data; genomics; SNP arrays; fluorescence intensities; genotype classification; optimal scaling; category quantification; alternating least squares; block relaxation.

1. Introduction

The psychometrics toolbox contains an impressive array of methods for analyzing complex data. True to their origin, most of them are being used on data from the behavioral sciences. In our view, however, the research area subsumed under the name *systems biology* (genomics, proteomics, metabolomics and other omics), as well as biological psychology, generates data that are pre-eminently suited for the application of psychometric methods. These high-dimensional data are also called high-throughput data, where the number of variables is much larger than the number of observations (usually denoted as $p >> n$, where p denotes the number of variables and n the number of observations). In this paper, we present applications of an "individual differences" model in the genomics domain; specifically, we analyze genetic measurements of allelic mutations, called Single Nucleotide Polymorphisms (SNPs), obtained for a number of persons, captured in arrays. (In the current text, we use the term *array* in a biological sense, as opposed to the traditional mathematical meaning. We will refer to a mathematical array as a *data array*. This choice has a simple reason: the array in biological sense will be much more often used in the sequel.) Although the fact that $p >> n$ creates challenging problems in prediction tasks such as multiple regression, it does not amount to special problems in the present case.

We will elaborate on the necessary biological context and technical details in section 2, but here it suffices to know the following. Usually, we have a fixed combination of two alleles - two variant forms of genes - in a known position on the chromosome. However, in some positions, mutations with (possibly strong) biological implications can occur, which is the motivation of the analysis of SNPs. Each allele is associated with a categorical variable - the three possible genotypes. Each allele forms a data source, where each source constitutes a three-way data array (in the present case of dimensions $p = 1624$ by $n = 24$ by 3 (the number of genotypes)), in which 2/3 of the cells cannot be observed, because each slide of the three-way data array is associated with one particular genotype. Moreover, we have four separate linkage panels, where each of the linkage panels constitutes a number of chromosomes.

In this paper, we propose and study a set of models with optimal scaling features, where the latter are used to quantify the categories of the genotype variable. The complexity of the models depends on the amount of interaction we allow between the three dimensions in the data array, and on whether the models for the two separate three-way data arrays (one for each allele) share sets of parameters. The paper is structured as follows: first we present some necessary background information on the biology and the measurement technique(s). Second, we fully describe the data, after which we introduce the statistical models that we propose. The data analytic results of the application of our models will then be evaluated and discussed.

2. Biological background and technical background

It is commonly known that our DNA is contained in chromosomes and that chromosomes usually come in pairs. Each chromosome is a string of nucleotides which come from an "alphabet" of four letters: A, C, G and T that represent the chemical building blocks adenosine, cytosine, guanine and thymine. We can define a position on a chromosome by counting the number of nucleotides from a unique end. At most positions we will always find the same nucleotide on both chromosomes. But there are millions of places (about 1 in every 1000 nucleotides) where we see small mutations. These are called Single Nucleotide Polymorphisms (SNPs, pronounced as "snips"). Each SNP has two variants, called alleles which we indicate here by A and B, where both A and B are one of the four possible nucleotides. SNPs are being studied extensively in biology and medicine (Adorjan et al., 2002; Altshuler et al., 2005). The driving force behind their popularity is the recent availability of cost-effective methods for measuring (hundreds of) thousands of SNPs at the same time.

We give a simplified description of the Illumina BeadArray system (Fan et al., 2006). At each SNP position two different nucleotides can be observed. Their exact type is immaterial, so we will simply call them A and B. A single BeadArray consists of about 50000 tiny beads. Each bead is covered with one type of molecule, specific for a known SNP (Eckhardt et al., 2006; Fan et al., 2003). The specificity means that, in a

watery solution of DNA, only fragments containing that SNP attaches (hybridizes) to that bead.

Consider one SNP in the DNA of one person. Because chromosomes are paired, the possible pairs of alleles are AA, AB, BA and BB. However only the unordered pairs AA, AB and BB are observable, since it is technically not possible to discern the two chromosomes. These are called genotypes.

Concentrations are being measured by fluorescence. So-called fluorophores are attached to the DNA fragments. These are molecules that emit light of a certain color when illuminated by laser light of the right color. The fluorophores are chosen in such a way that red light corresponds to the A allele and green light to the B allele. The strength of the fluorescence is proportional to the concentration of the DNA fragments on a bead.

When A (B) is present, green (red) light is observed. So, when the genotype is AA (BB), all fragments will be of type A (B) and no red (green) light should be observed, but only green (red) light of double strength. When the genotype is AB, the corresponding beads will attract A and B alleles in equal amounts. We should observe both red and green light, having equal strengths. Hence the relative concentrations of hybridized fragments on the beads reflect the genotypes of the corresponding SNPs.

Two (one for Red, one for Green) high resolution digital images of the full array allow the quantification of the strength of the fluorescence of each bead. For each SNP on average about 30 beads are on the array. Their red and green signal are averaged to reduce noise. The final result is a list of 1624 pairs of red and green intensities.

If the technology we just described worked perfectly, our story would end here. In practice we observe a number of interesting and relevant patterns:

- The strength of the fluorescence signals varies systematically between SNPs. If a SNP signal is strong, it is strong in all arrays. A weak SNP signal is weak in all arrays.
- The strength of the signals varies between biological samples (which is unavoidable; it is caused by differences in the quality of the biological material and the efficiency of

DNA extraction).

- Systematic deviations from the theoretical genotype factors 1 and 2 occur (1 or 2 times the allele). These deviations could be due to copy number variations (1 or 3 alleles, instead of 2), especially in humans.
- Noise and background signals are present.
- Nonlinearities occur in hybridization and fluorescence.

We have several reasons to investigate statistical models for the BeadArray fluorescence data. In the first place we like to know how well the theoretical model (intensities of red and green fluorescence proportional to the number of allele) holds in practice. The deviations that we find might point to flaws in the technology. A second goal is look for interesting, quantifiable, biological phenomena. Our final, and most important, goal is to use estimates for the model parameters to correct the data in such a way that genotyping will be improved.

3. The data

The type of BeadArray we consider here contains beads to measure 1624 different SNPs. The biological samples arrays are collected on a plate with 96 with small wells (also called microtiters), arranged in an 8 by 12-grid. The optical fiber (bead) bundles are held together in a metal frame, organized in the same 8 by 12 grid, above the wellplate. The arrays are lowered into the wells to analyze 96 samples at the same time. After averaging of the individual beads per SNP, two 1624 by 96 matrices are obtained, one for Red (the red fluorescence signal, allele A), the other for Green (the green fluorescence signal, allele B). The 96 arrays are assembled in four groups of 24. The groups represent four so-called linkage panels, each covering a subset of the chromosomes.

At each combination of SNP and array, one of the three genotypes AA, AB or BB can occur. The genotypes have to be determined from the same fluorescence signals. We use our own method (see Appendix A), which resembles some approaches in the

literature (Giannoulatou, Yau, Colella, Ragoussis, & Holmes, 2008; Xiao, Segal, Yang, & Yeh, 2007). Given the genotypes, we can construct a three-mode array with 1624 rows, 96 columns and three layers for each of the colors. Because only one genotype can occur at a given combination of SNP and array, only one of the three positions is filled. Hence, only one third of the arry is filled and the data are fundamentally incomplete. Figure 1 presents a cartoon of the arrays and Figure 2 shows the three genotype layers, where the dots in the cells denote empty cells. The filled cells show different saturations, indicating varying intensities. An alternative representation uses three complete matrices: one for the genotypes, and the two others for the Red and Green intensities.

Figure 3 shows a small subset of the raw data in linkage panel 1. The two graphs at the left show Red and Green for subsets of SNPs (rows) and arrays (columns). Careful inspection shows that generally intensity levels are similar in a row. High (low) intensities go together, although changes caused by different genotypes blur the pattern. To emphasize the pattern, the two graphs at the right show the same subsets, but now the Red signal is used to sort rows and columns of both Red and Green matrices. The signal has been sorted in ascending order for both the SNPs and the arrays. The resulting ordering is also applied to the Green signal. Now, we can see a similar pattern in both the Red and Green signal.

We conclude that some arrays show stronger average intensities over all SNPs than others. This is well known; it is caused by differences in DNA quality and quantity in different biological samples. We also conclude that some SNPs show consistently higher intensities than others. This not common knowledge and it is the motivation for our modeling effort.

Note that the data matrix is transposed compared to the psychometric/statistical tradition that uses rows for individuals and columns for variables. In our data matrices, the columns correspond to biological samples. This is standard in modern high-throughput microbiology (Bibikova et al., 2006). The reason for this is prosaic: spreadsheet programs, to which automatic instruments deliver their data used to allow no more than 256 (2^8) columns but up to 65536 (2^{16}) rows.

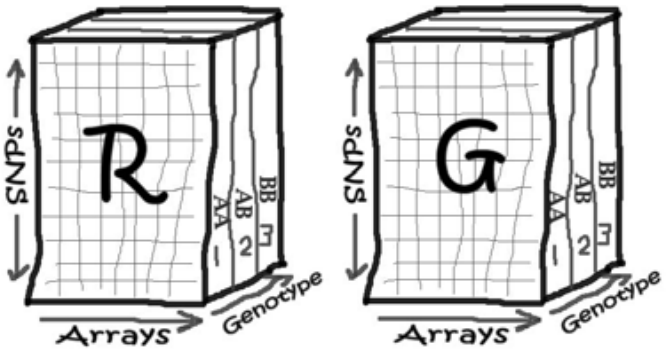

FIGURE 1.

Red and Green data with three modes: SNPs in the rows, samples in the columns and genotypes in the layers. In a single linkage panel, we have 24 arrays. There are four linkage panels.

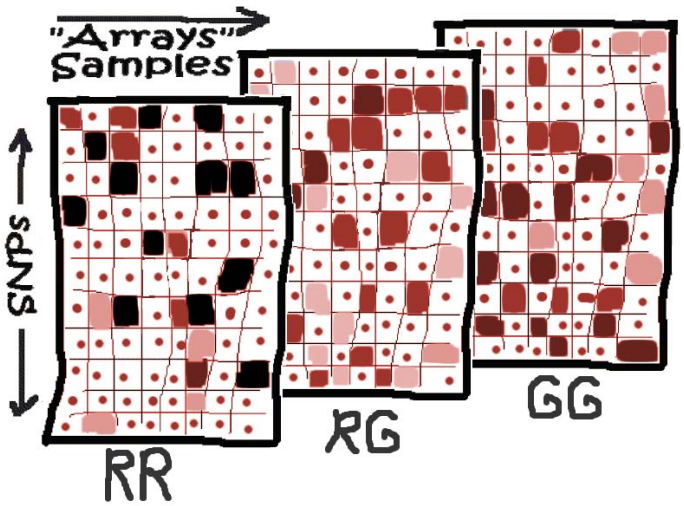

FIGURE 2.

Exploded view of the 3-dimensional data. The dotted cells denote empty cells. The filled cells show signals with varying intensity. In every combination of row and column, only one cell is filled in one of the three layers. Aggregation of the layers results in a single, but complete intensity matrix without genotype information.

4. Models and parameters

We expect the data to be proportional: the strength of each SNP signal is increased by the array strength and by the genotype (the number of alleles). This is a so-called

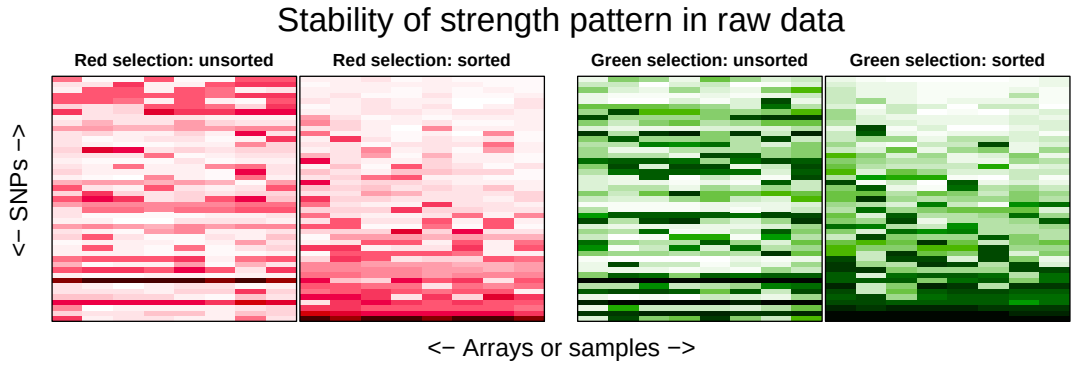

FIGURE 3.

Selection of raw data. Left: signals unsorted. Right: both signals sorted for rows and columns of red intensities.

'dimmer'-effect: if you turn down the light in a room, the whole room gets darker, but the different objects can still be distinguished. This phenomenon works in a similar way for the arrays (if the array is dimmed, so are all SNPs in that array) and the SNPs (if a SNP is dim, it will be in all arrays in which it occurs). Since this theoretically holds for both the SNPs and array, we suggest a multiplicative model. Also, in principle, dependent on the genotypes AA, AB and BB, the intensities of red will be proportional to 2, 1, 0, and those of green to 0, 1, 2. We expect the intensities to be proportional to a product that symbolically can be written as $SNP * array * genotype$.

The assumed proportionality suggests a three-way analysis (Kroonenberg, 2008; Smilde, Geladi, & Bro, 2004), but there are complications. First, it is not known how available three-mode techniques behave in a situation with two-third of the data missing. Second, equal variances cannot be assumed. To a first approximation, the variance increases with the square of the signal mean. However, on a (base 10) logarithmic scale we can write, again symbolically, $SNP + array + genotype$, leading to a linear model with approximately constant error variance.

Suppose X_c represents the data for color c, and $Y_c = \log(X_c) = \{y_{ijc}\}]$, with $i = 1 \ldots m$ and $j = 1 \ldots n$. The linear model we propose for the data for color c can

then be written as:

$$y_{ijc} = \mu_c + \alpha_{ic} + \beta_{jc} + \sum_{k=1}^{3} \gamma_{kc} h_{ijk} + e_{ijc}, \tag{1}$$

where μ_c is the grand mean, α_{ic} describes the overall level of SNP i, and β_{jc} describes the overall level of sample (array) j. The index k denotes the genotype, (1 = RR, 2 = RG, 3 = GG), where we now will use R for Red (B) and G for Green (A). Thus, $k = 1$ represents the RR matrix in Figure 2, $k = 2$ represents the RG matrix, and $k = 3$ represents the GG matrix. Genotypes are coded by the indicator matrix $H = \{h_{ijk}\}$, and γ_{kc} denotes the parameter for genotype k. To make the model identifiable, we introduce without loss of generality the constraints $\sum_i \alpha_{ic} = 0$ and $\sum_j \beta_{jc} = 0$.

4.1. Model estimation

To estimate the parameters of the model in (1), we minimize the usual sum of squares of residuals,

$$S = \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \mu - \alpha_i - \beta_j - \sum_{k=1}^{3} \gamma_k h_{ijk})^2,$$

where m is the number of SNPs and n the number of arrays. To simplify the notation, we dropped the index c for color for all parameters, since at this point the two colors are handled separately. We set $\hat{\mu}$ equal to the average of all observations:

$$\hat{\mu} = \sum_{i}^{m} \sum_{j}^{n} y_{ij}/mn.$$

In addition, we have the constraints $\sum_i^n \alpha_i = 0$ and $\sum_j^n \beta_j = 0$.

Given $\hat{\mu}$, estimation of the other parameters could be achieved by linear regression, after vectorizing $Y = \{y_{ij}\}$ and translating H into the proper design matrix. Instead, we use block relaxation, because the system of normal equations would be very large (of the order of 5000). In our application, block relaxation is very easy, because it boils down to repeated computation of averages (alternating least squares). Linear regression would give us standard errors of estimated parameters, but we have no need for them (yet). Constraints on parameters (like those on α and β) are easily handled within block relaxation.

Suppose we have the approximations $\tilde{\alpha} = \{\tilde{\alpha}_i\}$ and $\tilde{\gamma} = \{\tilde{\gamma}_k\}$. Given these, we compute $r_{ij} = y_{ij} - \hat{\mu} - \tilde{\alpha}_i - \sum_k h_{ijk}\tilde{\gamma}_k$ and minimize $\sum_i \sum_j (r_{ij} - \beta_j)^2$ to obtain estimates $\tilde{\beta}_j$. It is easy to see that $\tilde{\beta}_j = \sum_i r_{ij}/m$ gives the solution. So we have to average the columns of the matrix R. To enforce $\sum_j \beta_j = 0$, we subtract the mean of the newly computed $\tilde{\beta} = \{\tilde{\beta}_j\}$. By analogy, for $\tilde{\alpha}_i$ we find $\tilde{\alpha}_i = \sum_j r_{ij}/n$, where now $r_{ij} = y_{ij} - \hat{\mu} - \tilde{\beta}_j \sum_k h_{ijk}\tilde{\gamma}_k$.

In a similar way, given approximations $\tilde{\alpha}$ and $\tilde{\beta}$, we compute $r_{ij} = y_{ij} - \hat{\mu} - \tilde{\alpha}_i - \tilde{\beta}_j$ and minimize $\sum_i \sum_j (r_{ij} - \sum_k h_{ijk}\gamma_k)^2$. The solution is $\tilde{\gamma}_k = \sum_i \sum_j h_{ijk} r_{ij} / \sum_i \sum_j h_{ijk}^2$. Because H is an indicator matrix, containing only zeros and ones, this is equivalent to computing the averages of r_{ij} selectively for the three genotypes.

From this point, we repeat the procedure, improving α, β and γ in turn. We monitor the changes in β from iteration to iteration and stop when these are small enough. Our experience is that after five or six iterations relative changes are of the order of 10^{-6} or less.

4.2. Alternative models

Up to now we estimated the parameters for the two colors separately. In view of the similar patterns in Figure 3, it is reasonable to also consider models in which the vectors β or both vectors α and β are the same for each color. Re-introducing the index c to indicate either green or red, this leads to either

$$y_{ijc} = \mu_c + \alpha_{ic} + \beta_j + \sum_{k=1}^{3} \gamma_{kc} h_{ijk} + e_{ijc}, \tag{2}$$

or

$$y_{ijc} = \mu_c + \alpha_i + \beta_j + \sum_{k=1}^{3} \gamma_{kc} h_{ijk} + e_{ijc}. \tag{3}$$

In the next section we will study the correlations between the α and β vectors for the two colors and will use the standard errors of the residuals to decide which sets of parameters can be shared.

5. Results

5.1. Statistics of the raw data

Before we go into details, we give an idea on the order of magnitude of the signals we are working with. The mean of the log transformed intensities lies around 3.5, corresponding to a median of the intensities around 3000. As Table 1 shows, the standard deviations lie around 0.6, corresponding to a factor 4 on the intensity scale. This illustrates that we are dealing with signals that vary over a large range.

TABLE 1.

Standard deviations of the residuals. Results for the null-model (centered for $\hat{\mu}$, no estimates for α, β and γ).

	Red	Green
1	0.595	0.573
2	0.622	0.591
3	0.632	0.599
4	0.635	0.579

5.2. Model fit

We judge the fit of models by the standard deviation of the residuals; these are collected, for the four linkage panels, in Table 2. Except for model (3) with equal α and β, we obtain an approximately five-fold reduction. To put these numbers in perspective, we note that $10^{0.12} \approx 1.3$, so the variation on the scale of the intensities has been reduced from a factor 4 to about 30%.

The model fit in Table 2 shows that, compared to the null-model, the standard deviations σ of the three models' residuals are all very small. We can see that models (1) and (2) fit similarly well. Sharing β between colors, model (2), hardly increases the residuals. This makes sense, as we expect that the intensities of the two colors are

influenced by variations in biological samples in the same way. In contrast, sharing α as well, model (3), increases the standard deviation of the residuals by about one third. We conclude that the model (2) with shared β is to be preferred, and therefore all results that follow will be based on this model.

Linkage panels 1 to 3 show very similar results. For linkage panel 4 the residuals are somewhat higher. A possible explanation might be that this linkage panel covers the X chromosome, which is special (women have two X chromosomes and males have X and Y). In the current study, we have not investigated this further.

TABLE 2.

Standard deviations of the residuals. Results for the models from sections 4 and 4.2.

Panel	Model	Red	Green
1	1) α and β per color	0.120	0.115
1	2) β equal	0.121	0.116
1	3) α and β equal	0.161	0.158
2	1) α and β per color	0.113	0.109
2	2) β equal	0.114	0.109
2	3) α and β equal	0.161	0.158
3	1) α and β per color	0.131	0.116
3	2) β equal	0.131	0.116
3	3) α and β equal	0.176	0.166
4	1) α and β per color	0.159	0.139
4	2) β equal	0.159	0.140
4	3) α and β equal	0.196	0.181

Figure 4 illustrates the fit graphically; the results are for a typical sample: array 11 in linkage panel 1. The scatter plot indicates a very good fit for a linear relationship. This is also expressed by the Pearson correlation coefficients: 0.892 for Red and 0.877 for Green. However, overall correlations can be misleading. If the centers of the clouds

for the three genotypes are far apart but near a straight line, we will always find a high overall correlation, even if it is low within each separate cloud. Figure 4 shows that this is not the case for the present results, as is also shown in Table 3.

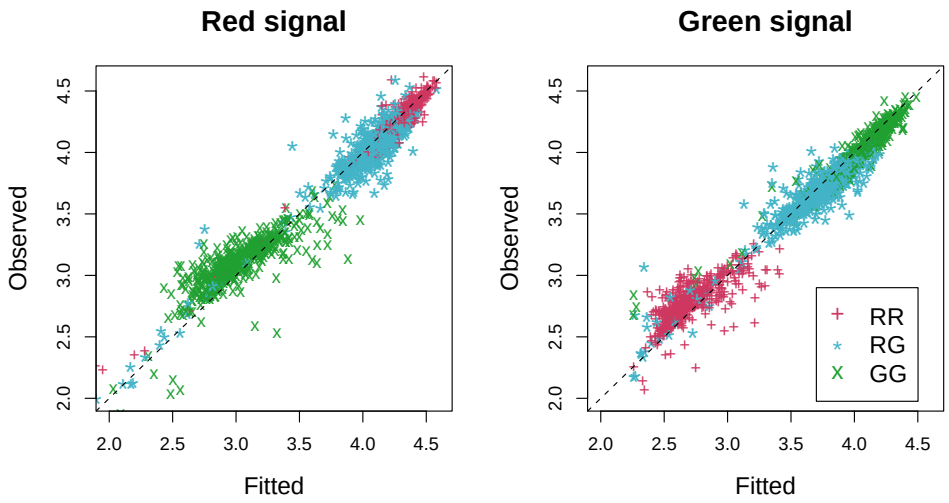

FIGURE 4.

The observed data versus the fitted values from model 2 for a typical array 11 in Linkage Panel 1: strong overall and within-cluster correlations.

For both the red and the green signal, the within-cluster correlations are somewhat lower than the overall correlation. For the red signal, the largest correlation is always for the RR cluster, while for the green signal, the largest correlation is always for the GG cluster. In fact, the order of the correlations in the red and green column are always reversed. Although the overall correlation is higher for the red signal compared to the green signal, for the within-cluster correlations, those for the green signal are consistently higher than those for the red signal.

5.3. Genotype parameters

According to the biochemical background of the BeadArray technology, we expect to obtain red (green) fluorescence signal strengths in the ratio 2, 1, 0 (0, 1, 2) for the RR, RG and GG genotypes. We never really get 0, because of background, small

TABLE 3.

The two columns represent the correlations in the Red and Green signal, from model 2, in array 11 in Linkage Panel 1.

Panel		R_{red}	R_{green}	Panel		R_{red}	R_{green}
1	Overall	0.892	0.877	3	Overall	0.872	0.870
1	RR	0.793	0.697	3	RR	0.811	0.677
1	RG	0.771	0.771	3	RG	0.731	0.728
1	GG	0.629	0.897	3	GG	0.627	0.829
2	Overall	0.891	0.852	4	Overall	0.845	0.841
2	RR	0.794	0.583	4	RR	0.776	0.571
2	RG	0.754	0.765	4	RG	0.744	0.750
2	GG	0.695	0.837	4	GG	0.585	0.821

(positive) noise, and so-called crosstalk between the two colors. Hence we will get a small number instead of 0. But nevertheless the ratio 2:1 when comparing RR to RG for red (GG to RG for green) should be obtained, because noise, background and crosstalk are too small to have an influence there.

Whether these assumptions are reflected by the experimental data can be checked by computing $10^{\gamma_1-\gamma_2}$ for red and $10^{\gamma_3-\gamma_2}$ for green. The numbers are collected in Table 4. For the green signal, the numbers are near their expected values, but surprisingly those for red are consistently too low. We were unable to find an explanation for this result.

6. Improved genotyping

In our modeling efforts we have assumed that the genotypes are known, and reliable, but in fact they are estimated from the same fluorescence intensity, and with limited accuracy. The procedure that we have used, genotypes all SNPs on an array in one go. In this respect it is different from the commercial Illumina software, but similar

TABLE 4.

Genotype patterns from model 2 are used to compute the DSR (Double-Single Ratio) between the homozygotic genotype RR (GG) and the heterozygotic genotype RG for the red (green) signal, for the four linkage panels.

	Panel 1	Panel 2	Panel 3	Panel 4
DSR Red	1.517	1.514	1.521	1.514
DSR Green	2.244	2.234	2.193	2.099

to recent proposals in the literature (Giannoulatou et al., 2008; Xiao et al., 2007). Figure 5 shows two graphs (the first for array 1 and the second for array 11) with the values of log(Red/Green) versus log(Red + Green). In this standard graph, the RR, RG and GG genotypes show up as elongated clouds. The intensity data are modeled as a mixture of three straight lines with normally distributed noise around them. Each cluster is described by three parameters (intercept, slope, and noise standard deviation). Given the cluster parameters, one can compute for each observation, cluster membership probabilities. Given these probabilities, one can estimate cluster parameters by weighted regression. We alternate between these steps, which is known as the EM algorithm. In our experience, the convergence is quick. Good starting values are easily found by initial separation of the data in three groups, based on log(Red/Green). As can be seen from the graph, there is some overlap between the clusters and hence the estimated genotypes of the corresponding SNPs are less reliable.

Figure 6 shows the fluorescence intensity data again, for the same two arrays, 1 and 11. In the two graphs at the left, log(Green) is plotted versus log(Red); the same three elongated genotype clusters are visible as in Figure 5, but here they are rotated. For the two graphs in the middle of Figure 6, the data have first been corrected by using the estimates obtained for the parameters of our model (2); thus we use the values of $log(y_{ijc}) - \hat{\mu}_c - \hat{\alpha}_{ic} - \hat{\beta}_j$. In the two graphs at the right, we use the corrected data $y_{ijc} - \hat{\mu}_c - \hat{\alpha}_{ic} - \hat{\beta}_j$ to compute the Ratio and Sum of Green and Red to produce the RS-plot. As we compare these scatterplots to those in Figure 5, it is clear that the

clusters have become more compact, allowing improved genotyping in a second round. Elaboration of this idea goes beyond the scope of the present paper, but one can easily imagine a procedure thats iterates between cluster estimation, modeling and correction until convergence. The presentation of the corrected (logs of) fluorescence intensities also allows us to visualize the γ's for the two colors, the quantification of the categories of the genotype variable by our optimal scaling procedure. The values for the γ's now give the centers of the three clusters in the two plots in the middle. The anomaly for Red is clearly visible by the small gap between the RG and RR cluster on the Red scale (the horizontal axis).

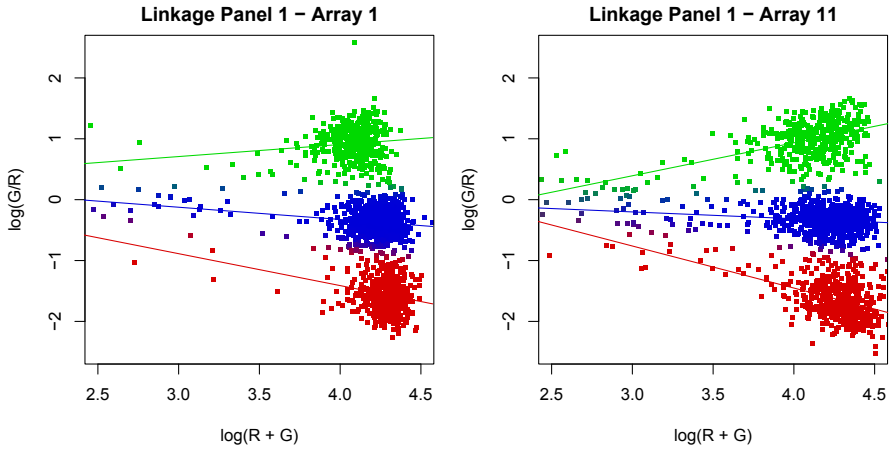

FIGURE 5.

RSPlot (Ratio versus Sum of Red and Green) with model lines for a representative sample.

7. Conclusion and Discussion

We have analyzed a relatively large data set from modern molecular biology using tools that are familiar in psychometrics. From an optimal scaling perspective, the task has been to replace the categories for the genotype variable (labeled AA, AB and BB) in a matrix G by quantifications γ_1, γ_2 and γ_3, to obtain a matrix C such that the fit of the additive model $\alpha_i + \beta_j + c_{ij}$ to a matrix $Y = \{y_{ij}\}$ of (logs of) observed fluorescence intensities is optimized. Because of the additive structure, a large linear regression

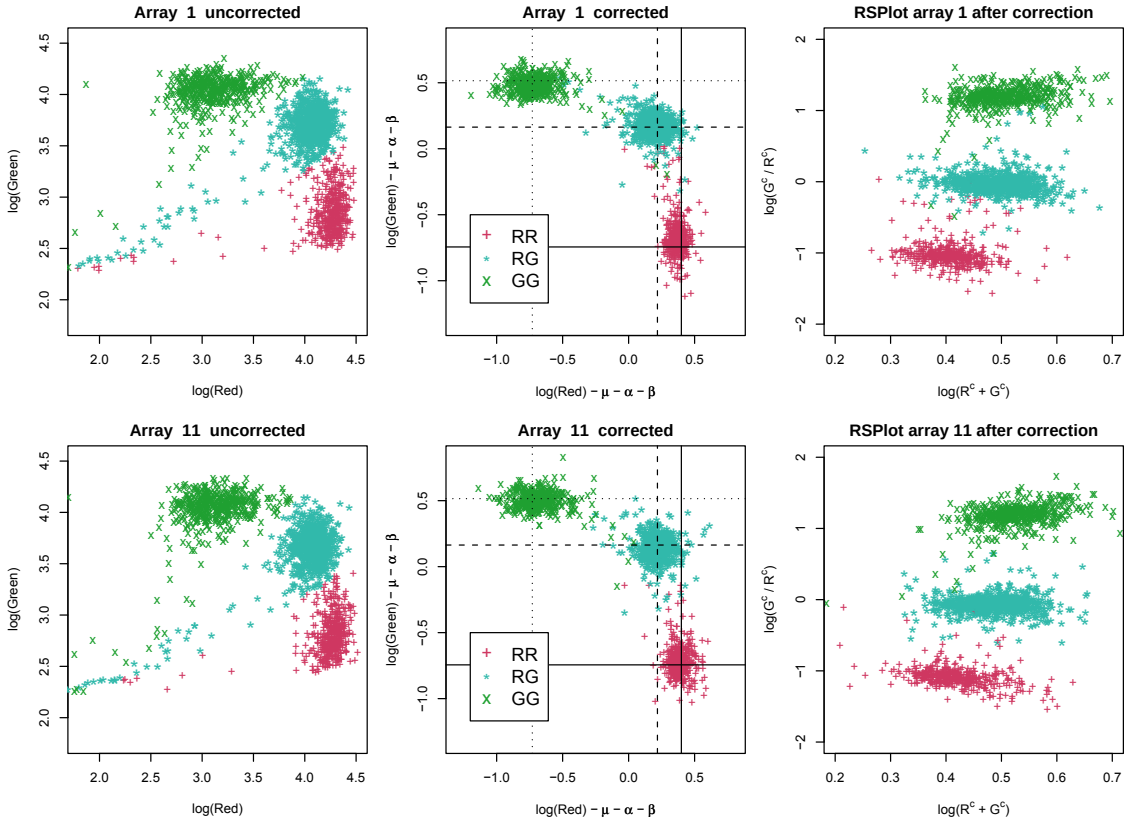

FIGURE 6.

Two selected arrays from the first linkage panel. In the left panels, we present the uncorrected data on the log scales. In the middle panels, the data on both axes are corrected for the relating αs, βs and μs. The lines represent the red (x-axis) and green (y-axis) genotype quantifications. In the right panels, we see the RSPlots for the same corrected data. In the axis labels, $Red^c = log(Red) - \mu - \alpha - \beta$ and $Green^c = log(Green) - \mu - \alpha - \beta$.

model follows, allowing for straightforward estimation of all parameters using block relaxation (alternating least squares).

We found excellent fit of the model to the data of the two fluorescence colors, red and green. We investigated whether parameters could be shared between colors (equality constraints), without degrading the fit (the standard deviation of the residuals) too much. It turned out that only the β's, the parameters for the normalization of the biological samples can be shared.

The biochemical background prescribes certain theoretical relationships between

the optimal scaling quantifications (in γ) for the genotypes. These seem to hold for the green fluorescence, but certainly not for red. This is interesting and possibly important information for biochemists, but we will not elaborate on this result in the present paper.

Much more interesting and useful, is the fact that the parameter estimates from the model can be used to improve the genotype classification. The latter is based on a model-based clustering procedure. We showed that clusters become more compact and hence better defined after correction of the intensities with the estimates of the parameters in the model. Presently we are working on an iterative algorithm that repeatedly performs clustering, modeling and correction, to optimize genotyping.

The data set we analyzed here may look large, but it is small by modern standards in molecular biology. Nowadays arrays that can analyze over half a million SNPs in one go are being used routinely, increasing the size of the data by at least two orders of magnitude. A computation that now takes 10 seconds will take an hour then. Computational efficiency will become an additional goal.

Extensions of the model are also required. We analyzed a rather well-behaved data set, stemming from non-pathological human samples. In tumor samples, many aberrations can be found of which copy number variations (CNV) and loss of heterozygosity (LOH) are most common. Copy number variations (CNV) are deviations from the AA, AB, BA or BB pattern, resulting in patterns like AAB or ABB ("gain") or only A or B ("loss"). LOH occurs when (part of) one original chromosome got lost, but the remaining one was copied to form a pair again. Only the genotypes AA and BB can occur then. Finally, recent reports indicate that an un-typed third allele may occur in some SNPs (in normal as well as in tumor samples), allowing six genotypes: 00, 0A, 0B, AA, AB and BB (Franke et al., 2008). The latter situation extends the optimal scaling task to quantification of six genotype categories.

Acknowledgements

We would like to thank the Pathology Department of the Leiden University Medical Center for providing the data (Lips et al., 2005).

References

Adorjan, P., Distler, J., Lipscher, E., Muller, F., Muller, J., Pelet, C., et al. (2002). Tumour class prediction and discovery by microarray-based dna methylation analysis. *Nucleic Acids Research*, *30*, e21.

Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., Donnelly, P., et al. (2005). A haplotype map of the human genome. *Nature*, *437*, 1299-1320.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Wickham-Garcia, E., Wu, B., et al. (2006). High-throughput dna methylation profiling using universal bead arrays. *Genome Research*, *16(3)*, 383-93.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., et al. (2006). Dna methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, *38(12)*, 1378-85.

Fan, J. B., Gunderson, K., Bibikova, M., Yeakley, J. M., Chen, J., Wickham-Garcia, E., et al. (2006). Illumina universal bead arrays. *Methods in Enzymology*, *410*, 57-73.

Fan, J. B., Oliphant, A., Shen, R., Kermani, B., Garcia, F., Gunderson, K., et al. (2003). Highly parallel snp genotyping. *Cold Spring Harbor Symposia on Quantitative Biology*, *68*, 69-78.

Franke, L., Kovel, C. de, Aulchenko, Y., Trynka, G., Zhernakova, A., Hunt, K., et al. (2008). Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *American Journal of Human Genetics*, *82(6)*, 1316-1333.

Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J., & Holmes, C. (2008). Genosnp: a variational bayes within-sample snp genotyping algorithm that does not require a reference population. *Bioinformatics*, *24(19)*, 2209-2214.

Kroonenberg, P. (2008). *Applied multiway data analysis*. Hoboken, NJ: Wiley.

Lips, E., Dierssen, J., Eijk, R. van, Oosting, J., Eilers, P., Tollenaar, R., et al. (2005). Reliable high-throughput genotyping and loss-of-heterozygosity detection in formalin-fixed, paraffin-embedded tumors using single nucleotide polymorphism arrays. *Cancer Research*, *65*, 10188–10191.

Smilde, A., Geladi, P., & Bro, R. (2004). *Multi-way analysis in chemistry*. Chichester, UK: Wiley.

Xiao, Y., Segal, M. R., Yang, Y. H., & Yeh, R. F. (2007). A multi-array multi-snp genotyping algorithm for affymetrix snp microarrays. *Bioinformatics*, *23-12*, 14591467.