

A SURVEY OF REGRESSION MODELLING TECHNIQUES FOR ANALYSING MICROARRAY DATA



Fernando Henrique Ferraz Pereira da Rosa
Department of Statistics
University of São Paulo
feferraz@ime.usp.br

Júlia Maria Pavan Soler
Department of Statistics
University of São Paulo
pavan@ime.usp.br

Abstract

Studies in genetics involving microarray experiments allow simultaneous comparison and quantification of gene expression on a large scale. Many sources of variation play an important role on the analysis of data coming from such experiments. A proper experimental design and subsequent development of suitable models for analysis are essential steps in order to assure the identification of significant genes, helping researchers distinguish between variations which are due to actual biological changes from random noise. In this work, we evaluate some of the various regression modelling strategies proposed in the literature, such as the use of mixed-effects models [5], [2] and the issue of variable selection in microarray gene expression profiling. We apply the models studied to data from a study of the Laboratory of Genetics and Molecular Cardiology, Heart Institute, University of São Paulo Medical School (InCor-USP) based on strains of rat, which aims to identify genes that have a regulatory role on the mechanism of hypertension.

Introduction

Microarrays are a widely used tool in genomic studies to assess, through the levels of mRNA in a certain tissue, how a set of thousands of genes are being expressed under different conditions. Many different strategies have been proposed to analyse data coming from such experiments, ranging from linear models to cluster analysis. In this work, we survey those methods pertaining closely to regression analysis, though a complete distinction is debatable.

In a microarray experiment, pools of differentially labelled cDNA sequences are combined and applied to a glass slide or other substrate [1], containing thousands of known complementary sequences of cDNA. Each region of the slide, containing a certain immobilised sequence is called a *spot*. For each spot, the intensity of hybridisation of the two labelled samples are measured, and through this measure we can assess the levels of expression for the gene represented by that spot. For each slide in an experiment, we have a set of $2s$ measurements, where s is the total number of spots in the slide. This number ranges usually from 20000 to 50000. We denote the measures from the sample marked with the green dye by G and those from the sample marked with the red dye by R . A generic spot i will be denoted by $(r, g)_i$.

Often these data are subject to various source of variation and a direct analysis of the raw $(r, g)_i$ values may be misleading. The first methods we address are the normalization ones, which try to control some of this variation, using nonparametric smoothing. These were the first models proposed in the literature. We then consider the approach of ANOVA and linear mixed models, and finally the use of logistic regression to identify differentially expressed genes.

Nonparametric smoothing

Nonparametric smoothing comes in the context of normalization of microarray data. Normalization techniques were devised to reduce the technical variation and to obtain comparable intensity values across different slides. These techniques are usually applied on a transformation of the original data. The pairs (R, G) are transformed into the (M, A) coordinate system, where $M = \log_2 R/G$ and $A = \log_2 \sqrt{RG}$. This new axis system, (M, A) , corresponds to a rescaling and a clockwise coordinate system rotation by 45° .

A scatter plot of these values for the spots in an array is usually called an MAPlot. This type of plot is used to display the gene expression intensities aiming a spatial identification of points. Here, the horizontal coordinate A is a measure of the average transcription level, while M is a measure of differential transcription. This graphic is akin to the variance by average scatter plots, traditionally used in exploratory data analysis for identifying heterocedasticity and dependence structures. Figure 1 displays an MAPlot for a given microarray slide.

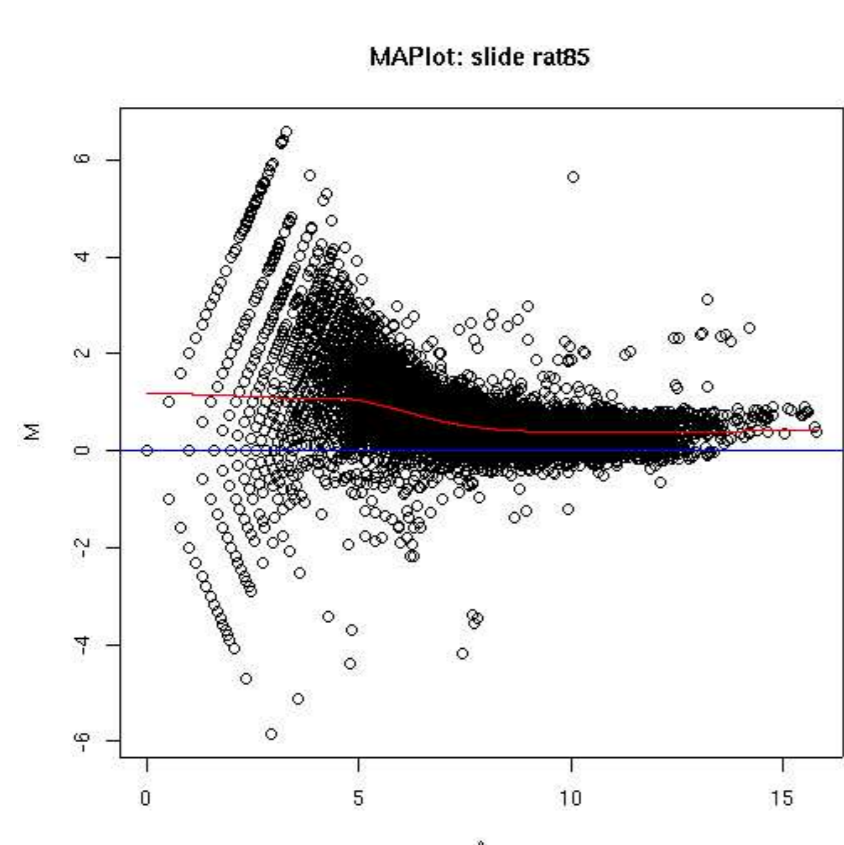


Figure 1: MAPlot for a given slide.

Normalization techniques work by transforming the M values, attempting to control systematic and random variations. We usually seek a transformation of the type:

$$M_k \rightarrow M_k - c_i(A), \quad i = 1, \dots, s,$$

where $c_i(A)$ is obtained through a smoothing function of the MAPlot for the given slide. A good choice of $c(\cdot)$ will assure that much of the variation due to technical issues is controlled, so that further analysis of the data is carried on the normalized values. The first proposed and most used smoother is the lowess [6] robust scatter plot smoother. Another possibility is the use of smoothing splines [4], as a more adaptive technique.

Other variations of these procedures, consider the physical arrangement of the spots in the slides, which is generally done by blocks or subarrays. As there is indication that subarrays have influence on the intensity values, it is common to adjust a different smoother $c_j(\cdot)$, for every subarray j , originating the normalization per subarray [6] technique. Figure 2 displays MAPlots for a set of 2 different slides, comparing two different smoothing functions, from a study on congenic rats from the Laboratory of Genetics and Molecular Cardiology, Heart Institute, University of São Paulo Medical School (InCor-USP).

Lowess and splines normalization

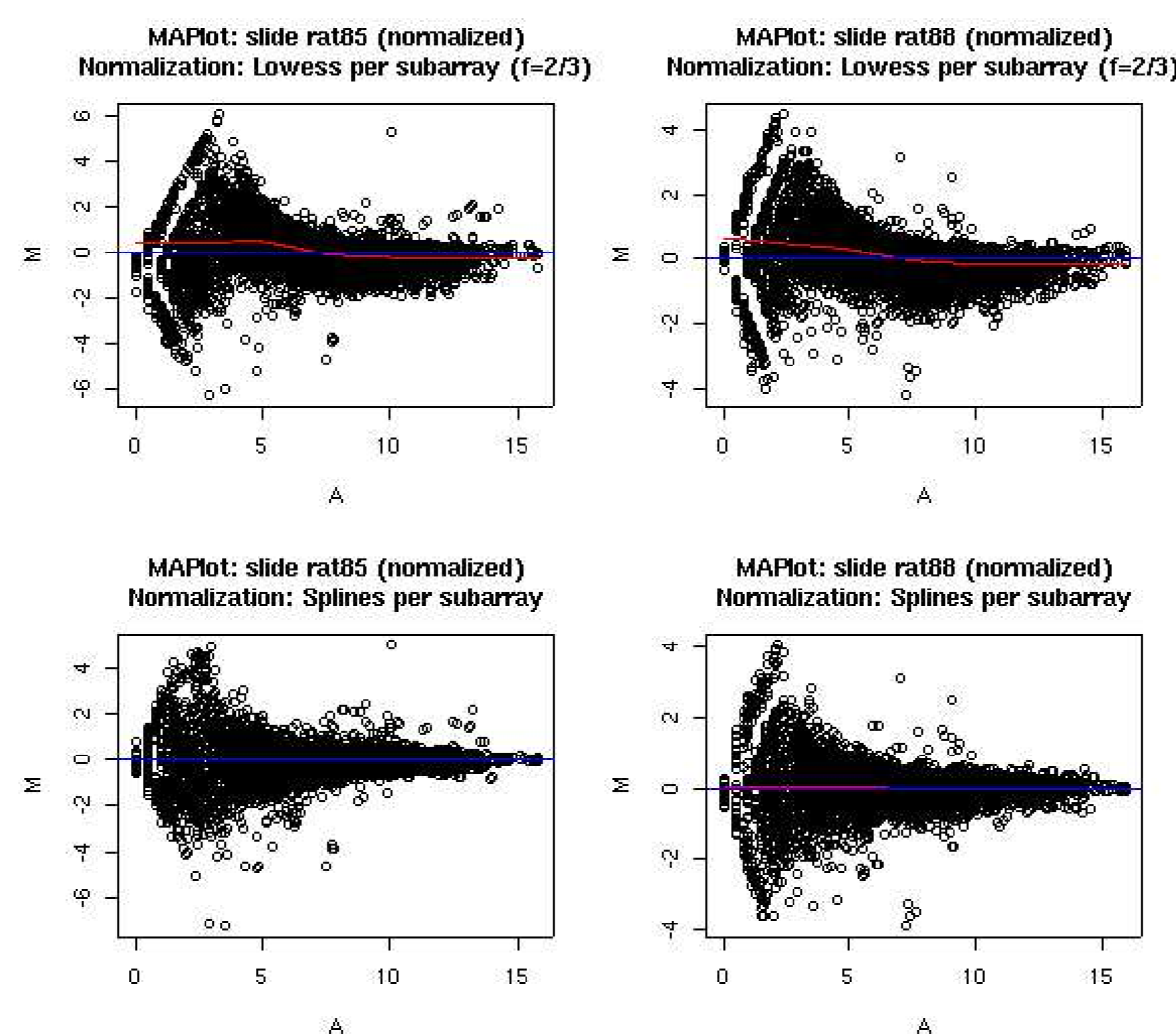


Figure 2: Lowess and splines per subarray normalization compared over a set of 2 arrays.

In general, these nonparametric smoothing techniques provide a good fit to the data, resulting in good normalization procedures. On the other hand, we know little about the structure of the model being considered, such as what effects are being accounted and what are being discarded.

ANOVA and mixed models

Another approach to analyse this type data, sometimes complementary to normalization, is the use of ANOVA techniques and linear models. This was first proposed by [2], whose model is given by:

$$\log(Y_{ijk}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijk},$$

where μ is an overall average of the signal intensities, A_i represents the effect of the i^{th} array, D_j represents the effect of the j^{th} dye, V_k the effect of the k^{th} variety, G_g the effect of the g^{th} gene and ϵ_{ijk} is a stochastic error. The interaction effect $(AG)_{ig}$ accounts for a *spot* effect, and is included in the model mainly to reduce the variability due to technical issues, not being of interest in itself. The effects of interest in this model are the interaction effects between variety and gene, given by $(VG)_{kg}$. All effects are considered fixed, and the normalization of the data is carried through the A , D and V terms, without the need to introduce preliminary manipulations.

A related method was later proposed by [5], which uses two interconnected ANOVA models, a “normalization” model and a “gene” model. The normalization model aims to control experiment-wide systematic effects while the gene model aims to identify differently expressed genes. The residuals of the normalization model are used as inputs to the gene model. Both models are mixed (with random and fixed effects), the normalization one being:

$$\log_2(Y_{gij}) = \mu + T_i + A_j + (TA)_{ij} + \epsilon_{gij},$$

where μ is an overall average, T is a variety effect, A is an array effect and TA is the interaction effect of array and variety. Effects A and TA are random, while T is fixed. Due to their experimental design, the T effect is accounting for the dyes effect, and the interaction TA is modelling the channels. This is possible as in their design the treatment was always labelled with $Cy5$, but for more general situations, a D effect for dye is usually considered. The gene model is given by:

$$r_{gij} = G_g + (GT)_{gi} + (GA)_{gj} + \gamma_{gij}$$

, where r_{gij} are the residuals from the normalization model, G is a gene effect, GT is an interaction between gene and variety and GA an interaction between gene and array. GA is a random effect, while G and GT are fixed. The GA effect is essential to this model, serving to account for spot-to-spot variability and allowing us to not form ratios.

The use of a mixed model grants us more flexibility to model the correlation structure of the data. Particularly, the use of a random effect for the array (A) allows us to model the correlations between the expression levels in a given array. Also, we may consider using a random gene effect, which would permit us to extend the inferences made to the whole population of genes, even those not present on the arrays under study.

Logistic regression

We can consider the problem of identifying genes which have a regulatory role on a certain characteristic of interest, as a problem of discriminant analysis. The characteristic of interest can be considered as a grouping factor, or binary response variable (in the case of two different groups of interest, for example normotense and hypertensive rats) and the expression levels of the genes spotted on the microarray as explanatory variables. This structure naturally leads one to a logistic regression model.

Let us consider a phenotype of interest taking two possible values: SHR (spontaneously hypertensive rat) and non SHR (normotense rat). Suppose that samples of mRNA of these two groups are obtained and cohybridized into a microarray slide, and the x_j intensities levels are measure for each channel (Cy3 and Cy5). We can then model the probability of pertaining to a certain phenotype using a logistic regression:

$$P(SHR) = \frac{1}{1 + \exp(-\beta_0 - \sum_{i=1}^s \beta_j x_j)}, \quad (1)$$

where x_j are the log-normalized gene expression levels, and β_j is a gene specific parameter. The immediate problem with directly using this approach is that $s + 1$ parameters are required to fit a model considering all genes. As the number of slides is always far smaller than that, the proposed solution [3] is to consider only a subset of best performing genes, according to an ancillary analysis, and use the logistic model to select the genes with most effect in the phenotype of interest among this subset.

Variable selection

In this context, the problem of identifying differentially expressed genes comes down to the selection of variables in the logistic model (1). A best subset of explanatory variables in this context, will be a best subset of genes which have a regulatory role on the studied treatments/phenotypes.

The procedures for carrying out the selection of the variables are well established on the literature, and are based generally on restricted likelihood criteria, like the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).

These procedures should be used with caution, though. Some algorithms tend to avoid the problem of multicollinearity by dropping one or more of the correlated variables, and this is generally not desired during the identification of genes: as much as the levels of expressions of two different genes might convey the same information regarding the classification of the experimental units in the control or treatment group, the researcher will generally be interested in all of those genes.

Another issue to be considered is that these procedures will find a subset of genes which best perform in predicting a certain phenotype when combined on a logistic regression model. The identification of individual genes, or how many genes should be stated as relevant, might be a more delicate question [3], and will depend on the significance levels adopted and assumptions the researcher will be willing to make.

It should be also observed that the approach of variable selection may also be used in the context of the ANOVA and mixed linear models, by considering a model where the predictors are genes, and selecting a subset of best performing predictors.

References

- [1] M. Kerr and G. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2:183-201, 2001.
- [2] M. K. Kerr and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, (7), 2001.
- [3] W. Li and Y. Yang. *Methods of Microarray Data Analysis*, chapter How many genes are needed for a discriminant microarray data analysis, pages 137-150. Kluwer Academic, 2002.
- [4] J. M. P. Soler, F. H. F. P. da Rosa, S. Chiavegatto, I. Aneas, and J. E. Krieger. Use of splines for normalization of microarray gene expression data. *Bioscience Journal*, Especial:101-116, 2004.
- [5] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625-637, 2001.
- [6] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. Technical Report 589, January 2001.

This work was supported by:

