

Fernando Henrique Ferraz Pereira da Rosa  
Número USP 36890349

*Métodos adaptativos em regressão não  
paramétrica e normalização de microarrays  
de cDNA*

Fernando Henrique Ferraz Pereira da Rosa  
Número USP 36890349

*Métodos adaptativos em regressão não  
paramétrica e normalização de microarrays  
de cDNA*

UNIVERSIDADE DE SÃO PAULO

# *Sumário*

## **Resumo**

<b>1</b>	<b>Introdução</b>	p. 5
1.1	Regressão via Kernel . . . . .	p. 5
1.2	Regressão via Splines . . . . .	p. 6
<b>2</b>	<b>Métodos adaptativos</b>	p. 8
2.1	Lowess . . . . .	p. 8
2.2	Smoothing splines . . . . .	p. 9
2.2.1	Método da Validação Cruzada (CV) . . . . .	p. 9
2.2.2	Método da Validação Cruzada Generalizada (GCV) . . . . .	p. 9
<b>3</b>	<b>Análise de microarrays</b>	p. 10
3.1	Introdução . . . . .	p. 10
3.2	Dados . . . . .	p. 10
3.3	Métodos . . . . .	p. 11
3.3.1	Normalização por intensidade total: normalização global . . . . .	p. 11
3.3.2	Normalização dependente da intensidade: lowess . . . . .	p. 11
3.3.3	Normalização dependente da intensidade: splines smoothing . . . . .	p. 12
3.4	Comentários . . . . .	p. 12
	<b>Apêndice A – Definições</b>	p. 13

A.1	O método do Kernel para estimação de densidades univariadas . . . . .	p. 13
A.2	A função de Validação Cruzada Generalizada . . . . .	p. 13
<b>Apêndice B – Figuras</b>		p. 14
<b>Apêndice C – Implementações</b>		p. 17
C.1	Normalização Global . . . . .	p. 17
C.2	Normalização por Lowess . . . . .	p. 17
C.3	Normalização por Smoothing Splines . . . . .	p. 18
<b>Referências Bibliográficas</b>		p. 19

# *Resumo*

Em estudos com microarrays de cDNA a utilização de métodos não paramétricos de suavização de diagramas de dispersão se consagrou na literatura ([1] e [2]), como a abordagem mais eficaz na normalização e correção das intensidades pré-análise estatística. Na maioria desses estudos entretanto, é aplicada uma classe muito restrita de técnicas não paramétricas, restringindo-se basicamente ao método de lowess (Robust Locally Weighted Regression) [3]. Há porém outras técnicas não paramétricas mais adaptativas que podem ser empregadas (regressão por kernel, splines smoothing) a esse tipo de problema. No presente trabalho introduzimos a justificativa teórica por trás desses métodos e os aplicamos à conjuntos de dados reais, provenientes de estudos com microarrays de cDNA. Os resultados obtidos sugerem que os novos métodos obtêm normalizações muito melhores que os utilizados atualmente.

# 1 Introdução

A diferença básica entre um modelo de regressão paramétrico e um outro não paramétrico é a balança entre a quantidade de suposições que o estatístico está disposto a fazer e qual peso ele quer dar aos dados por si mesmos.

Dado um conjunto de pontos  $(X_i, Y_i)_{i=1}^n$ , amostra aleatória de duas variáveis aleatórias  $X$  e  $Y$ , procuramos descobrir uma relação funcional entre as duas variáveis, na forma:

$$Y_i = g(X_i) + \epsilon_i, \quad i = 1, \dots, n \quad (1.1)$$

Um modelo paramétrico assume que  $g(X)$  é uma função desconhecida num número finito de parâmetros, e nosso trabalho é estimar os parâmetros desconhecidos, por exemplo por mínimos quadrados. Em um modelo não paramétrico a relação funcional entre as duas variáveis vive num espaço de funções muito mais amplo: assumimos apenas que  $g(X)$  está num espaço de funções seguindo algumas restrições convenientes e buscamos uma combinação linear de funções desse espaço que aproximem bem  $g(X)$ .

Descreveremos dois métodos importantes na regressão não paramétrica: o método de Kernel e o via Splines.

## 1.1 Regressão via Kernel

O método de Kernel é um método não paramétrico para estimação de curvas de densidade [4]. Ele pode ser generalizado entretanto para o caso de uma regressão, bastando para isso observarmos que no modelo descrito em (1.1), o que procuramos é uma estimativa  $\hat{g}(x)$  de  $E(Y|X = x)$ . Pela definição de esperança condicional [5], temos

$$E(Y|X = x) = \int y f_{Y|X}(y|x) dy = \frac{\int y f(x, y) dy}{f(x)} \quad (1.2)$$

Pelo método do Kernel para estimação de densidades, já temos uma estimativa para o denominador da expressão indicada em (1.2) (vide A.1). Para estimar o numerador dessa expressão, começamos notando que podemos estimar a densidade conjunta  $f(x, y)$  através

de um kernel multiplicativo [6], resultando em

$$\hat{f}_h(x, y) = n^{-1} \sum_{i=1}^n K_{h_1}(x - X_i) K_{h_2}(y - Y_i).$$

Substituindo a densidade conjunta por essa estimativa no numerador da equação (1.2) e fazendo algumas manipulações algébricas temos o estimador de Nadaraya-Watson ([7] e [8]):

$$\hat{g}_h(x) = n^{-1} \frac{\sum_{i=1}^n K_h(x - X_i) Y_i}{\sum_{j=1}^n K_h(x - X_j)}. \quad (1.3)$$

Tomando algumas suposições não muito restritivas sobre a distribuição dos erros  $\epsilon_i$  e sobre a derivabilidade e continuidade das funções de densidade de  $X$  e  $Y$ , mostra-se em [9] que o estimador proposto em (1.3) converge em probabilidade para a função  $g(x)$  do modelo descrito em (1.1), desde que  $h \rightarrow 0$  e  $nh \rightarrow \infty$ , quando  $n \rightarrow \infty$ .

Um aspecto delicado desse método de regressão é a escolha do parâmetro de suavização  $h$ . Assim como no método de Kernel para estimação de densidades, o parâmetro de suavização tem uma influência decisiva na estimativa obtida da função de regressão. Não há entretanto nenhum método ótimo para a escolha de  $h$ , sendo comum na literatura [10] a utilização de métodos de validação cruzada ou de minimização do erro de predição.

## 1.2 Regressão via Splines

Splines são uma ferramenta proveniente do cálculo numérico, que vem ganhando atenção em estatística devido ao seu forte apelo adaptativo na aproximação de funções.

Dada uma função definida em um intervalo  $[a, b]$  que queremos aproximar, a idéia por trás do conceito de splines está em dividir o intervalo original  $[a, b]$  em intervalos menores  $[x_0, x_1], \dots, [x_k, x_{k+1}]$  e ajustar polinômios de graus  $p_i$  em cada intervalo  $[x_i, x_{i+1}]$ ,  $i = 0, \dots, k$ . Esse procedimento produz um polinômio por partes  $s(x)$ , que pode ser utilizado para aproximar a função procurada.

Algumas outras restrições precisam ser impostas [10] para que possamos garantir a continuidade e suavidade de  $s(x)$ , como a colocação dos nós (*knots*), os pontos de ligação entre os polinômios. Dessas restrições e do acréscimo de *pesos* a cada polinômio, surgem as bases de splines naturais e os B-splines.

Consideremos o modelo de regressão proposto em (1.1). Suponhamos que tenhamos uma função  $f(\cdot)$  que estime a função  $g(x)$ . Um critério de bondade de ajuste poderia ser dado por

$$\sum_{i=1}^n (y_i - f(x_i))^2. \quad (1.4)$$

Como estamos tratando com modelos numéricos inicialmente criados para interpolação, somente a bondade de ajuste não é um critério bom por si só, pois no caso de uma interpolação dos dados o ajuste teria uma bondade de ajuste excelente, mas seria muito pouco suave. Acrescentamos então um critério de suavidade, a dizer

$$\int_a^b (f(x)^{(m)})^2. \quad (1.5)$$

Juntando os critérios (1.4) e (1.5) em uma única equação, temos que o nosso problema se resume a minimizar

$$A_\lambda(f) = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)})^2 \quad (1.6)$$

para  $\lambda > 0$ , onde  $f(\cdot)$  é gerada a partir de uma base de splines.

Notemos que  $\lambda$  determina o grau de suavidade da estimativa, controlando o quanto andamos na direção da interpolação dos dados ou na direção da suavização excessiva.

É bem conhecido que uma base de splines naturais cúbicos é a escolha mais apropriada para conseguir a minimização de (1.6). Uma escolha muito comum para as bases é o uso de B-splines cúbicos, pois são muito mais leves de se manipularem computacionalmente.

Fixada uma base de splines, o trabalho se resume a determinar o número e a localização dos pontos de junção (nós, ou *knots*) dos polinômios cúbicos por partes.

Ficamos então com a tarefa da escolha do melhor parâmetro de suavização e da melhor combinação de nós. O método mais bem sucedido para realizar essa tarefa é o da Validação Cruzada Generalizada [9]: a idéia consiste em tirar sucessivamente elementos da amostra e fazer uma estimativa do ponto retirado, obtendo-se um erro de predição. Procura-se então o conjunto de parâmetros que minimiza esse erro. Veremos como funcionam esses métodos em mais detalhes na sessão específica sobre smoothing splines.

## 2 Métodos adaptativos

### 2.1 Lowess

O método de lowess (Local Polynomial Regression), foi proposto por Cleveland em [3], como um método de alisamento de diagramas de dispersão robusto e resistente a outliers. Ele se baseia em regressões polinômiais locais ponderadas, de forma que pesos maiores são dados a pontos concordantes com o conjunto de dados e pesos menores são dados a outliers.

Em primeiro lugar define-se uma função de ponderação  $W(\cdot)$ , com as seguintes propriedades

1.  $W(x) > 0$  para  $|x| < 1$ ;
2.  $W(-x) = W(x)$ ;
3.  $W(x)$  é uma função decrescente de  $x$  para  $x \geq 0$ ;
4.  $W(x) = 0$  para  $|x| \geq 1$ .

Seja então  $0 < f \leq 1$  e denotemos  $r$  como o produto  $f \cdot n$  arredondado para o inteiro mais próximo. O procedimento é então o seguinte [3]. Para cada  $x_i$ , são definidos pesos  $w_k(x_i)$ , através da função de ponderação  $W(\cdot)$ . Isso é feito centrando a função  $W(\cdot)$  em  $x_i$ , de forma que a partir do  $r$ -ésimo ponto na vizinhança de  $x$ , a função  $W(\cdot)$  assumo valor zero. Uma forma de interpretar o parâmetro  $f$ , é a proporção de pontos que irão ser considerados na vizinhança de cada  $x_i$  para a estimação do ajuste polinomial local.

É realizado então um ajuste polinomial por mínimos quadrados ponderados, com pesos  $w_k(x_i)$ , obtendo valores  $\hat{y}_i$  para cada  $x_i$ . Um novo conjunto de pesos, definidos para cada par  $(x_i, y_i)$  é obtido, baseado no tamanho do resíduo  $y_i - \hat{y}_i$ . Resíduos grandes resultam em pesos pequenos, e resíduos pequenos resultam em pesos maiores, dando portanto menos peso aos outliers.

Esse processo é repetido iterativamente  $t$  vezes, reponderando os pesos dos pontos a cada passada. A escolha do parâmetro  $f$ , também conhecido como parâmetro de suavização, pode influenciar bastante a estimativa final, mas não há nenhum procedimento ótimo pra encontrá-lo. Pode-se utilizar algum conhecimento a priori sobre o conjunto dos dados para estimá-lo ou valer-se de métodos como o da minimização do erro de predição.

## 2.2 Smoothing splines

Conforme já mencionado, a escolha do parâmetros de suavização é um fator muito importante na regressão por splines. Discutiremos aqui os dois principais métodos para escolha desse parâmetro.

### 2.2.1 Método da Validação Cruzada (CV)

O método da validação cruzada propõe um procedimento automatizado [11] para estimação do parâmetro  $\lambda$ . Seja  $f_\lambda^{[k]}$  a função que minimiza

$$\frac{1}{n} \sum_{i \neq k} (y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)})^2.$$

Ou seja, procuramos uma otimização retirando cada um dos pontos  $k$ . Seguindo a notação adotada em [9], a função de validação cruzada ordinária  $V_0(\lambda)$  é definida por

$$V_0(\lambda) = \frac{1}{n} \sum_{k=1}^n (y_k - f_\lambda^{[k]}(x_k))^2.$$

Utilizando então algumas identidades ([11] e [9]) obtém-se uma forma simplificada de  $V_0$ , a qual tem um custo computacional muito mais baixo, a dizer

$$V_0(\lambda) = \frac{1}{n} \sum_{k=1}^n (y_k - f_\lambda(x_k))^2 / (1 - h_k(\lambda))^2$$

onde  $h_k(\lambda)$  é o  $k$ -ésimo elemento da matriz  $H_\lambda$ , onde  $f_\lambda = H_\lambda y$ .

### 2.2.2 Método da Validação Cruzada Generalizada (GCV)

O método da validação cruzada generalizada se baseia no método da validação cruzada mas tem algumas vantagens [9]. Ele é mais barato computacionalmente e possui algumas propriedades teóricas que não são possíveis de se obterem com o método anterior.

A função de validação cruzada generalizada pode ser no apêndice em A.2. É importante ainda notar que os métodos de validação cruzada estão bem definidos para um conjunto de pontos  $\{x_i\}_{i=1}^n$  distintos, e que portanto deve-se tomar cuidado na sua implementação para eliminar grupos de pontos não distintos antes de realizar o procedimento de otimização.

## 3 *Análise de microarrays*

### 3.1 Introdução

Um experimento com microarrays de cDNA típico [12] envolve a hibridização de duas amostras de mRNA, provenientes de um grupo controle e de um grupo experimental, que são convertidas em cDNA e marcadas com corantes fluorescentes. A partir dessas marcações, são então geradas leituras de intensidade para cada canal fluorescente, que fornecem informações sobre a expressão relativa dos genes inclusos na lâmina de microarray.

O objetivo desse tipo de estudo, em geral, consiste na identificação de genes diferentemente expressos entre os dois grupos. Essa identificação entretanto não pode ser realizada comparando-se diretamente os sinais de intensidades obtidos para cada gene. Isso ocorre pelo fato de haver uma grande variabilidade da técnica, devido a fatores como diferente peso atômico dos corantes fluorescentes utilizados (acarretando uma maior taxa de fixação para um dos corantes), variação das taxas de hibridização devido a localização espacial dos spots na lâmina, eventuais falhas na hibridização, estouro das leituras de intensidade, entre outros.

Como uma forma de contornar esses problemas, foram desenvolvidas diversas técnicas de normalização ([13], [2], [14]), que objetivam controlar as variações sistemáticas entre os níveis de intensidade medidos em uma co-hibridização, de forma que a variação *biológica* se sobressaia a variação da *técnica* e os genes diferentemente expressos possam ser identificados.

### 3.2 Dados

O conjunto de dados utilizado nas análises que se seguem é proveniente de um experimento realizado no Laboratório de Cardiologia Molecular do Instituto do Coração (InCor-USP) com ratos congênicos no qual o fenótipo de interesse é a hipertensão. Foram utilizadas lâminas de vidro com superfície de polyisina, com 26912 spots cada (cada spot pode conter uma seqüência de cDNA que pode corresponder a um gene, ESTs ou seqüências de controle).

Para cada spot de cada hibridização há leituras de intensidade para dois canais, o verde (G/green) e o vermelho (R/red). Há portanto um conjunto de 26912 pares de leituras de intensidade  $(r, g)$ .

### 3.3 Métodos

Os dados de intensidade de fluorescência são tradicionalmente analisados utilizando-se uma transformação dos dados iniciais. Toma-se  $M = \log_2 R/G$  e  $A = \log_2 \sqrt{RG}$ , dando origem ao gráfico MAPlot, que é o diagrama de dispersão de  $M$  por  $A$ . Esse gráfico nada mais é do que um tipo de gráfico de média por desvio-padrão, bem apropriado para identificação espacial de pontos observacionais, que é justamente a intenção de uma análise de microarrays. Na figura 1 temos o gráfico de dispersão de  $M$  por  $A$ , ou o *MAPlot*, para uma das hibridizações.

Sobre essa transformação dos dados, são aplicadas então as técnicas de normalização, que buscam reduzir a variação da técnica e tornar as intensidades comparáveis entre hibridizações diferentes.

#### 3.3.1 Normalização por intensidade total: normalização global

Esse foi o primeiro tipo de normalização proposto [13], e sua suposição básica é que as intensidades totais para cada canal são equivalentes, de forma que para cada spot o par  $(r, g)$  está relacionado por uma constante  $k$ , de forma que  $R = kG$ . Estima-se essa constante e aplica-se a transformação  $R \rightarrow R/k$ .

Esse tipo de normalização foi o mais utilizado no começo dos estudos microarrays, apesar das evidências [1] que ele não se comporta bem com o viés devido ao corante fluorescente nem à localização espacial.

Na figura 2 temos o MAPlot da hibridização rat85 pós-normalização global por intensidade total. A linha vermelha corresponde ao ajuste de lowess com  $f = 2/3$ . A proximidade entre o eixo x e o ajuste de lowess pode ser utilizada como critério de bondade da normalização.

#### 3.3.2 Normalização dependente da intensidade: lowess

Após a constatação das deficiências das normalizações por intensidades total, ficou consagrada na literatura [2] a utilização de métodos mais robustos, em particular o método de lowess. A idéia é fazer uma regressão robusta para  $M$  em função de  $A$  e utilizar o ajuste para corrigir as intensidades em  $M$ . Na prática, toma-se a correção

$$M_k \rightarrow M_k - C(A_k), \quad k = 1, \dots, 26912$$

onde  $C(A_k)$  é o ajuste de lowess para  $A_k$ .

Dada a natureza do problema em questão, não só podem ocorrer outliers entre as observações  $(A, M)$ , como o objetivo é justamente identificar e trabalhar com esses outliers. O método de lowess se apresenta vantajoso por não se deixar afetar tanto por outliers, conforme foi discutido na seção 2.1.

Na figura 3 temos o MAPlot dos dados normalizados pelo método dependente da intensidade por lowess.

### 3.3.3 Normalização dependente da intensidade: splines smoothing

Apesar do método de lowess ser um método robusto para regressão não paramétrica, há alguns contrapontos. Em primeiro lugar ele não é tão adaptativo quanto outros métodos, como smoothing splines. Em segundo, não há procedimentos ótimos para a escolha do seu parâmetro de suavização  $f$  ou do número de iterações  $t$ , sendo os mesmos na maioria das vezes escolhidos de forma empírica.

O método de smoothing splines, além da adaptatividade maior, possui um procedimento ótimo para escolha do seu parâmetro de suavização, o Método da Validação Cruzada Generalizada. Decidimos então verificar como se comporta uma normalização via smoothing splines, tomando a correção

$$M_k \rightarrow M_k - S(A_k), \quad k = 1, \dots, 26912$$

onde  $S(A_k)$  é o valor predito via ajuste de smoothing splines para  $A_k$ .

O resultado se encontra na figura 4, mostrando sem sombra de dúvidas a superioridade do ajuste por smoothing splines. Vale notar que para esse tamanho de amostra, um ajuste por kernel smoothing consegue resultados muito bons também, mas há o problema de decidir qual o parâmetro de suavização adequado.

## 3.4 Comentários

Conforme se observa nas seções anteriores, a normalização aqui apresentada por smoothing splines mostrou ser mais bem sucedida que a normalização até então usualmente empregada por lowess. Isso se justifica pela maior adaptatividade do método de regressão não paramétrica por splines e pelo método de Validação Cruzada Generalizada, que nos fornece um procedimento para escolha ótima dos parâmetros de suavização.

O emprego de uma normalização adequada pode influenciar razoavelmente os resultados de uma análise de microarrays, como a taxa de ocorrência de falsos positivos ou falsos negativos, de forma que é interessante dar continuidade ao estudo dessas novas técnicas, verificando o seu efeito sobre a análise final.

## *APÊNDICE A – Definições*

### **A.1 O método do Kernel para estimação de densidades univariadas**

Dada uma variável aleatória  $X$  com função de densidade  $f(x)$  e uma amostra aleatória  $X_1, \dots, X_n$ , define-se o estimador de densidade via kernel [15], como

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

Onde  $K(\cdot)$  é uma função de kernel adequada. O estimador  $\hat{f}_h(x)$  é viciado para  $f(x)$ , mas pode-se mostrar [15] que sob certas condições, a dizer  $h \rightarrow 0$  e  $nh \rightarrow \infty$  quando  $n \rightarrow \infty$ , ele é assintoticamente não viesado e seu erro quadrático médio converge para zero.

### **A.2 A função de Validação Cruzada Generalizada**

A função de Validação Cruzada Generalizada (GCV) para estimação do parâmetro de suavidade em um smoothing spline é dada por

$$V(\lambda) = \frac{1}{n} \sum_{k=1}^n (y_k - f_\lambda(x_k))^2 / (1 - \bar{h}_k(\lambda))^2 = \frac{\frac{1}{n} \|(I - H_\lambda)y\|^2}{\left(\frac{1}{n} \text{tr}(I - H_\lambda)\right)^2}$$

onde  $\bar{h}_k(\lambda) = (1/n) \text{tr}(H_\lambda)$ .

## APÊNDICE B – Figuras

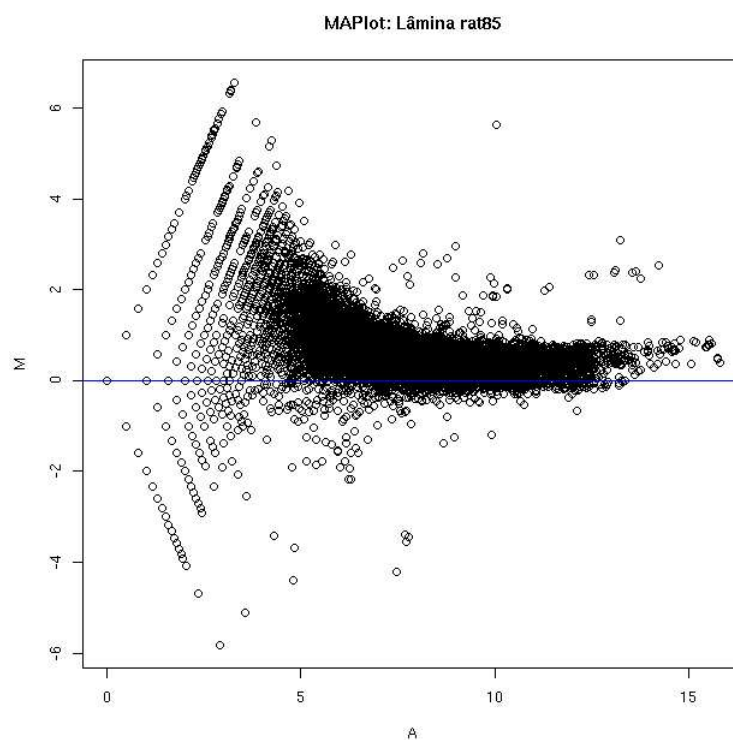


Figura 1: MAPlot da hibridização rat85. A linha azul indica o eixo x.

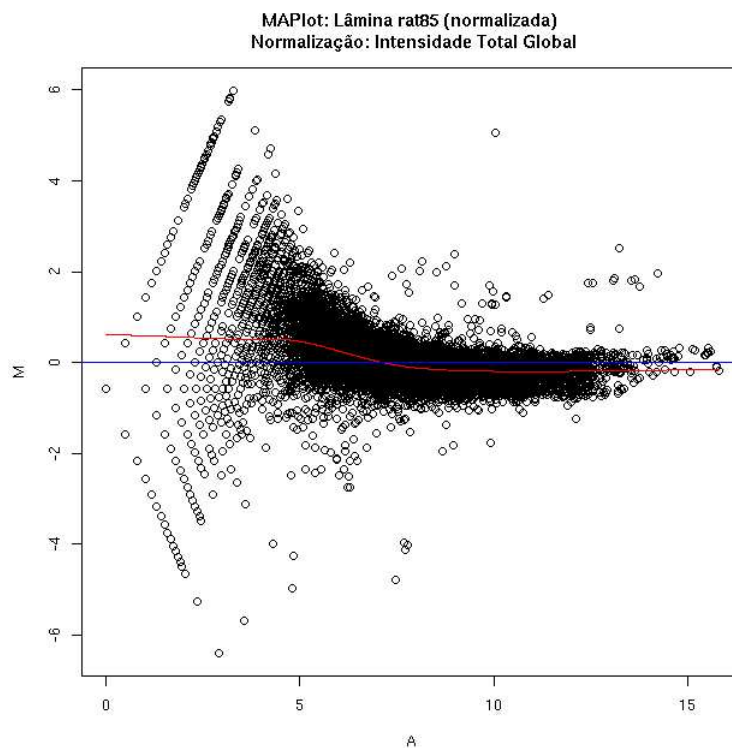


Figura 2: MAPlot da hibridização rat85 pós-normalização global.

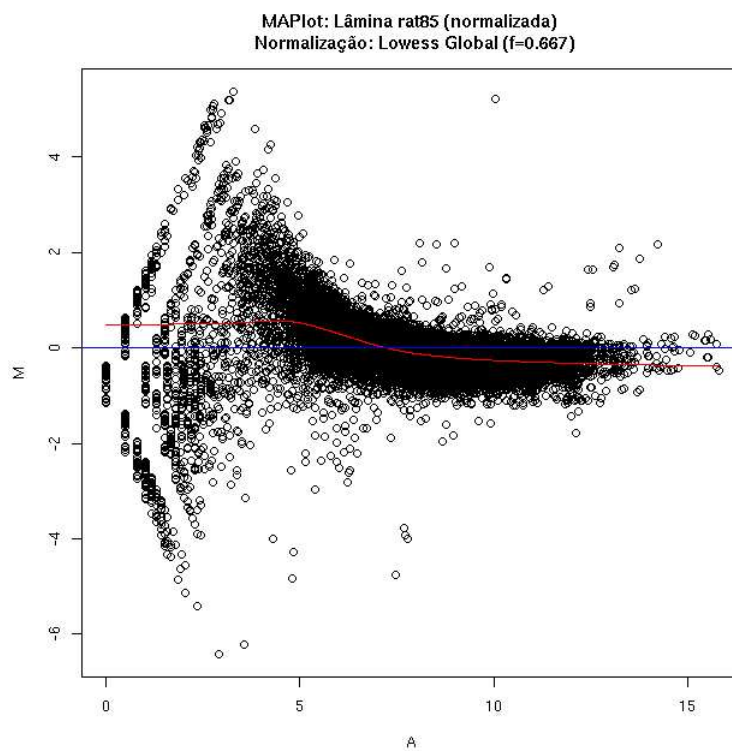


Figura 3: MAPlot da hibridização rat85 pós-normalização por lowess.

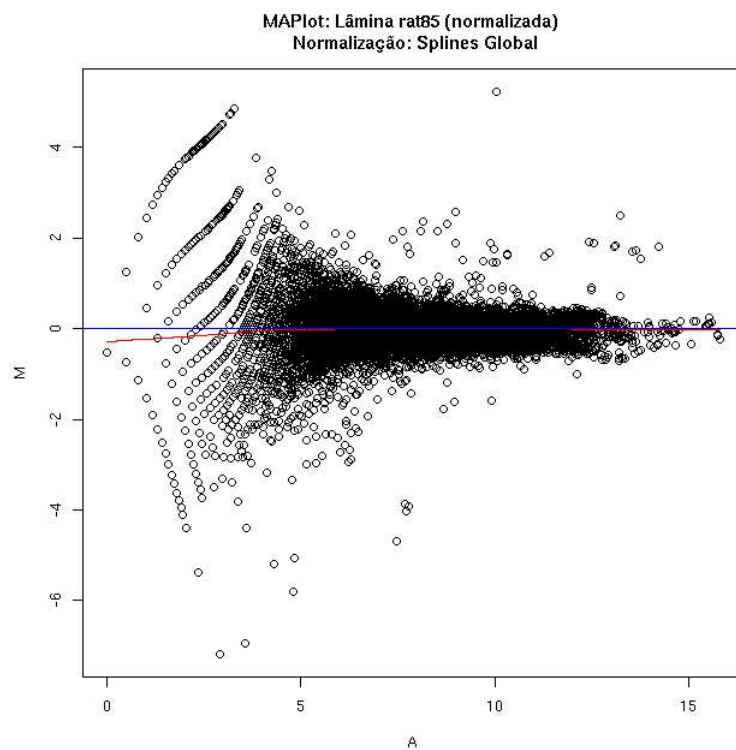


Figura 4: MAPlot da hibridização rat85 pós-normalização por smoothing splines.

## *APÊNDICE C – Implementações*

Disponibilizadas nesse apêndice estão as rotinas em R [16] utilizadas para realizar as normalizações. O pacote com as funções completas para essa análise se encontra em

<http://www.ime.usp.br/~feferraz/br/rma.html> .

### **C.1 Normalização Global**

```
global.norm <- function(madat,flaglevel=-200) {
  madat$intensities <- subset(madat$intensities,flags > flaglevel)
  norm <- madat

  c <- mean(madat$intensities$m)
  norm$intensities$m <- madat$intensities$m - c
  norm$norm <- 'Intensidade Total Global'
  norm
}
```

### **C.2 Normalização por Lowess**

```
global.lowess <- function(madat,lowess.f=2/3,flaglevel=-200) {
  madat$intensities <- subset(madat$intensities,flags > flaglevel)
  norm <- madat

  c <- lowess(madat$intensities$a,madat$intensities$m,f=lowess.f)$y
  norm$intensities$m <- norm$intensities$m - c
  norm$norm <- paste('Lowess Global (f=',format(lowess.f,digits=3),')',sep='')
  norm
}
```

## C.3 Normalização por Smoothing Splines

```
global.splines <- function(madat,lowess.f=2/3,flaglevel=-200) {  
  madat$intensities <- subset(madat$intensities,flags > flaglevel)  
  norm <- madat  
  
  smoothed <- smooth.spline(madat$intensities$a,madat$intensities$m)  
  c <- predict(smoothed,madat$intensities$a)$y  
  norm$intensities$m <- norm$intensities$m - c  
  norm$norm <- 'Splines Global'  
  norm  
}
```

## *Referências Bibliográficas*

- [1] YANG, Y. H. et al. *Normalization for cDNA microarray data*. [S.l.], January 2001. Disponível em: <www.stat.berkeley.edu/users/terry/zarray/Html/normspie.html>.
- [2] YANG, Y. H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleid Acids Research*, v. 30, n. 4, 2002.
- [3] CLEVELAND, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, v. 74, n. 368, p. 829-836, 1979.
- [4] SILVERMAN, B. W.; GREEN, P. J. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.
- [5] JAMES, B. R. *Probabilidade: Um curso em nível intermediário*. Rio de Janeiro: IMPA, 2002.
- [6] MÜLLER, H.-G. *Nonparametric Regression Analysis of Longitudinal Data*. Berlin: Springer-Verlag, 1988. (Lecture Notes in Statistics, v. 46).
- [7] WATSON, G. S. Smooth regression analysis. *Sankhya*, A26, p. 359-372, 1964.
- [8] NADARAYA, E. A. On estimating regression. *Theory of probability and its applications*, v. 10, p. 186-190, 1964.
- [9] DIAS, R. Nonparametric econometrics. *Brazilian Review of Econometrics*, v. 22, n. 1, 2002.
- [10] DIAS, R. A review of non-parametric curve estimation methods with application to econometrics. *Economia*, v. 3, n. 1, 2002.
- [11] WAHBA, G.; CRAVEN, P. Smoothing noisy data with spline functions. *Numerische Mathematik*, n. 31, p. 337-403, 1979.
- [12] YANG, H.; SPEED, T. Design issues for cDNA microarray experiments. *Nat. Gen.*, v. 3, p. 579-588, 2002.
- [13] CHEN, Y.; DOUGHERTY, E. R.; BITTNER, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, v. 2, n. 4, p. 364-374, 1997.
- [14] FINKELSTEIN, D. B.; GOLLUB, J.; CHERRY, J. M. *Normalization and systematic measurement error in cDNA microarray data*. [S.l.], 2002.

- [15] HÄRDLE, W. *Smoothing Techniques With Implementations in S*. New York: Springer-Verlang, 1990.
- [16] R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2004. ISBN 3-900051-00-3. Disponível em: <<http://www.R-project.org>>.