

PERSPECTIVAS NA ANÁLISE ESTATÍSTICA DE DADOS DE MICROARRAYS



Fernando Henrique Ferraz Pereira da Rosa
Instituto de Matemática e Estatística
Universidade de São Paulo
feferraz@ime.usp.br

Júlia Maria Pavan Soler
Instituto de Matemática e Estatística
Universidade de São Paulo
pavan@ime.usp.br

Resumo

A abordagem genética no estudo de problemas fisiológicos, como o da hipertensão no ser humano, oferece uma perspectiva desafiadora para pesquisas nessa área. A multiplicidade de genes envolvidos e a interação desses genes com eles próprios, com o meio ambiente e com o próprio organismo do indivíduo, a heterogeneidade inerente às populações humanas e as imprecisões das técnicas atuais são fatores que contribuem com esse prospecto.

A tecnologia de microarrays de cDNA, em particular, permite o estudo dos níveis de expressão de dezenas de milhares de genes simultaneamente. Em geral, o objetivo desse tipo de experimento é a identificação de genes associados à regulação de determinadas características de interesse. De forma a atingir esse objetivo é necessário conduzir um delineamento experimental adequado [3] e utilizar as técnicas de análise mais apropriadas [4], de acordo com as questões de interesse do pesquisador e o delineamento adotado, visando eliminar as fontes de variação de confundimento e evidenciar a variabilidade biológica.

No presente trabalho, descrevemos duas classes de técnicas diferentes: técnicas de normalização para conjuntos de lâminas de microarrays, e técnicas de identificação de grupos de genes diferentemente expressos, através da abordagem da Teoria da Resposta ao Item. A motivação e conjuntos de dados vem de experimentos realizados no Laboratório de Cardiologia Molecular do Instituto do Coração (InCor-USP).

Introdução

Estudos em genética envolvendo experimentos com microarrays permitem a quantificação e a comparação simultânea dos níveis de expressão de dezenas de milhares de genes. Em geral, o objetivo desse tipo de experimento é a identificação de genes associados à regulação de determinadas características de interesse. De forma a atingir esse objetivo é necessário conduzir um delineamento experimental adequado [3] e utilizar as técnicas de análise mais apropriadas [4], de acordo com as questões de interesse do pesquisador e o delineamento adotado, visando eliminar as fontes de variação de confundimento e evidenciar a variabilidade biológica.

A elaboração dos delineamentos experimentais e a análise dos dados nos diferentes estágios do processo, envolve o uso de diversas ferramentas estatísticas e computacionais, colocando sempre novos desafios para os pesquisadores que se deparam com problemas nessa área. A escolha da abordagem de análise mais adequada é essencial para garantir o maior poder de detecção de genes e deve ser feita levando em conta diversos fatores. Nas seções que se seguem, descrevemos um experimento em andamento no Laboratório de Cardiologia Molecular do Instituto do Coração (InCor-USP) e algumas das diferentes abordagens que utilizamos para lidar com os problemas encontrados na análise dos dados provenientes desse experimento.

Descrição do experimento *ex vivo*

Em um experimento realizado no Laboratório de Cardiologia Molecular do Instituto do Coração (InCor-USP), foram obtidas amostras de mRNA dos tecidos de veia safena de pacientes que foram operados em razão de alguma doença cardíaca. Esses tecidos foram mantidos em uma cultura *ex vivo* e submetidos a duas condições experimentais: regime arterial e regime venoso. Cada amostra de tecido de veia safena, de cada paciente, foi submetida a cada uma das condições experimentais. Para cada amostra e condição experimental foi hibridizada uma lâmina de microarray da plataforma Codelink (cDNA de uma cor). Foram feitas lâminas para amostras de mRNA de quatro indivíduos diferentes. Na Tabela 1 temos um resumo das lâminas disponíveis.

Indivíduo	Tratamento	
	arterial	venoso
105	VS105A	VS105V
128	VS128A	VS128V
130	VS130A	VS130V
131	VS131A	VS131V

Tabela 1: Descrição esquemática das lâminas do experimento *ex vivo*.

Técnicas de normalização

As técnicas de normalização objetivam minimizar parte das variações sistemáticas entre os níveis de intensidade de expressão gênica de um conjunto de lâminas de microarrays, com a intenção de fazer com que a variabilidade da técnica não se sobressaia à variabilidade biológica e, principalmente, àquela devida às diferentes condições experimentais, tornando comparáveis as medidas de lâminas diferentes [7].

Esses métodos de normalização podem ser divididos em duas classes: os que buscam tornar comparáveis as intensidades dentro das lâminas e os que buscam tornar comparáveis as lâminas entre si. Dentro de cada classe há uma série de técnicas disponíveis, como por exemplo a normalização por intensidade global [2] e as técnicas que utilizam modelos de regressão não paramétricos, como a utilizada por Soler et al. [6].

No caso específico do experimento *ex vivo* consideramos as possibilidades de abordagem para normalização entre lâminas. Entretanto, antes de entrar nas técnicas de normalização, precisamos introduzir o gráfico MA, que é, em geral, utilizado para representar conjuntos de lâminas de microarrays. Define-se

$$M = \log_2 R/G \quad \text{e} \quad A = \log_2 \sqrt{RG},$$

onde R são os níveis de intensidade de expressão gênica para um dado corante (no caso desse experimento, uma dada lâmina) e G os níveis de intensidade do outro corante (aqui, outra lâmina que vamos comparar com a anterior), e faz-se o diagrama de dispersão de M por A . Na Figura 1 temos os gráficos MA para todos os pareamentos entre lâminas na condição arterial. A linha horizontal passando pela origem representa o eixo X, enquanto a outra curva representa um ajuste pelo método de Kernel ao padrão de associação entre os dados. Quanto mais próximas as duas curvas se localizam, mais concordante é o par de lâminas sob estudo.

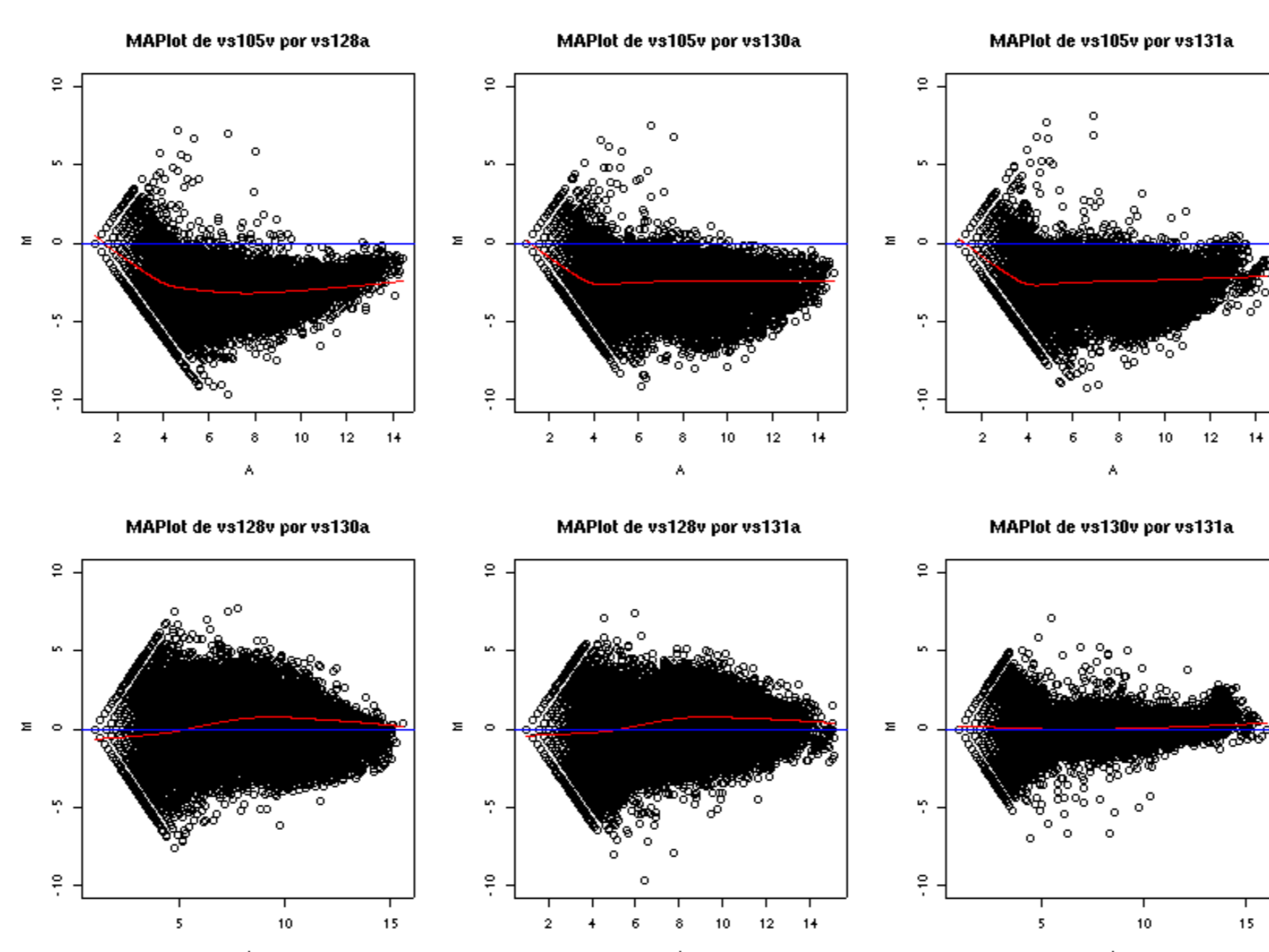


Figura 1: Pareamentos dentro da condição arterial.

Na Figura 2, consideramos as mesmas lâminas, mas após a normalização quantílica, proposta em Bolstad et al. [1], que ajusta simultaneamente todas as intensidades entre lâminas, a partir dos pareamentos das combinações possíveis de lâminas dentro de um determinado tratamento.

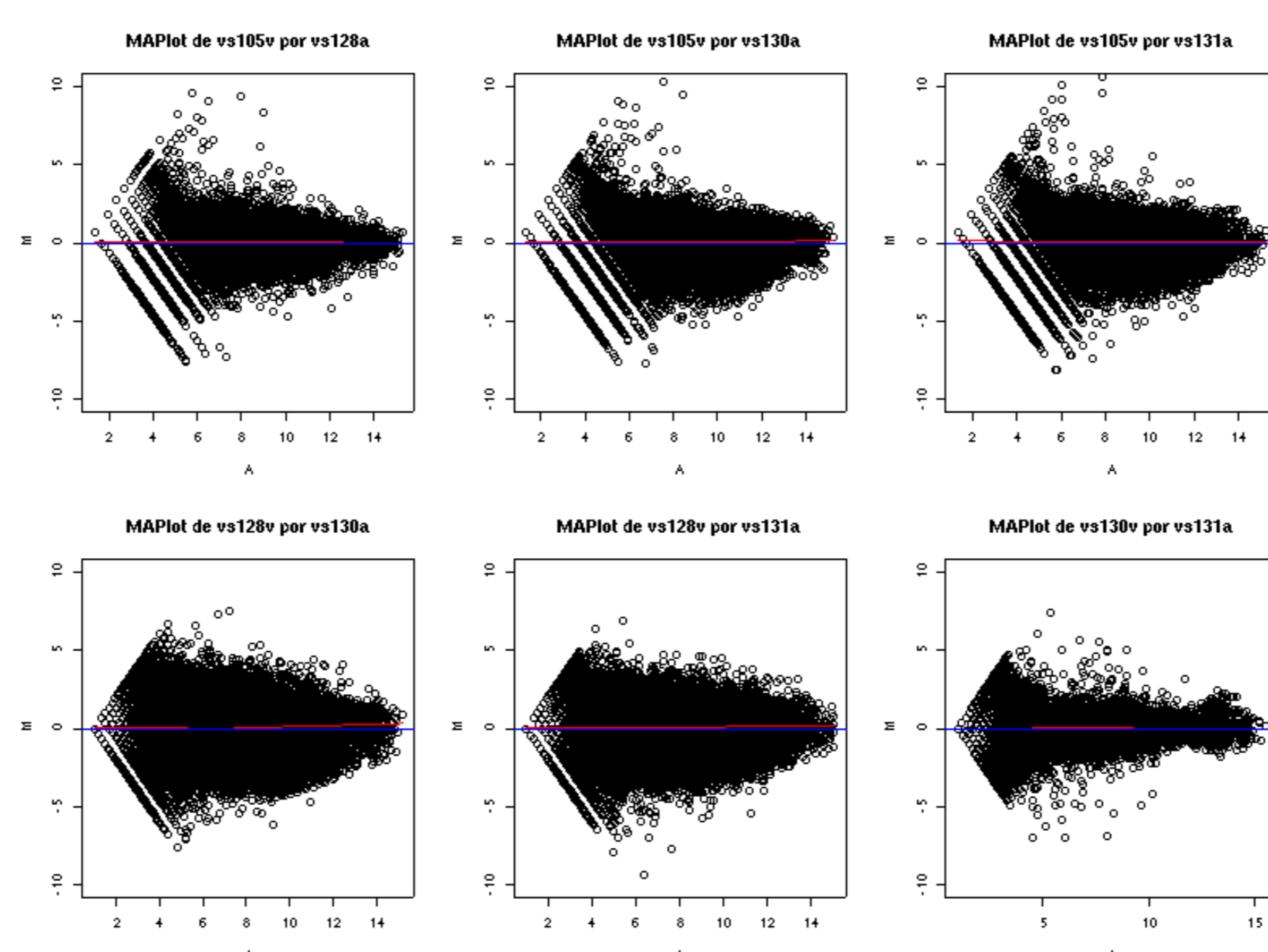


Figura 2: Pareamentos dentro da condição arterial, após normalização quantílica.

Teoria da Resposta ao Item

A Teoria da Resposta ao Item (TRI) é uma metodologia estatística que faz parte de uma classe de modelos chamados de Modelos de Traços Latentes, originalmente aplicada na área de avaliações educacionais e psicologia. A grande vantagem dessa teoria é que o uso de modelos probabilísticos bem definidos para cada situação permite um poder de generalização grande para outras áreas de aplicação. Consideramos dois dos modelos mais comuns utilizados para modelar problemas em TRI

O modelo de um parâmetro para itens dicotômicos

Seja um teste composto por I questões, respondido por J indivíduos, onde cada questão tem correção dicotômica: ou o indivíduo acerta ou erra a questão. Nesse contexto, seja U_{ij} uma variável indicadora, definida para o indivíduo j e o item i , da seguinte forma:

$$U_{ij} = \begin{cases} 1, & \text{se o indivíduo } j \text{ acerta o item } i \\ 0, & \text{se ele erra} \end{cases} \quad \begin{matrix} i = 1, \dots, I \\ j = 1, \dots, J \end{matrix}$$

O modelo logístico de um parâmetro (ML1), modela a probabilidade de acerto do item i pelo indivíduo j , pela expressão:

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}, \quad (1)$$

onde θ_j representa a habilidade (traço latente) do indivíduo j , e b_i é um parâmetro de dificuldade para o item i .

O modelo de Rasch para respostas ordinais

Consideremos agora a modelagem de itens com respostas do tipo ordinal. Seja um teste com I itens aplicado a J indivíduos. Denotemos por U_{ij} a resposta do j -ésimo indivíduo ao i -ésimo item, onde a resposta pode ser uma dentre $m + 1$ possíveis categorias: $0, \dots, m$. O modelo de crédito parcial de Rasch, modela a probabilidade de uma resposta h por um indivíduo com traço latente θ_j por:

$$P(U_{ij} = h|\theta_j) = \frac{\exp[b_{ih} + h\theta_j]}{\sum_{l=0}^m \exp[b_{il} + l\theta_j]} \quad (2)$$

Em [5], duas abordagens diferentes são propostas para modelagem do problema, ambas tratadas através de um modelo de Rasch para respostas ordinais.

A segunda formulação considera genes como 'pessoas' e cada condição experimental como um 'item'. A intenção é identificar traços latentes associados com cada gene baseando-se nas medidas de expressão desses genes em diversas condições experimentais.

Abordagem um

Nessa abordagem considera-se os genes como 'itens' e as condições experimentais como 'pessoas'. A ideia é identificar diversos traços latentes associados com uma dada condição experimental a partir das medidas de expressão de diversos genes nessa condição.

Notemos que as expressões dos genes para cada condição experimental são valores contínuos, e o modelo de Rasch proposto trabalha com respostas graduais de m categorias. Em primeiro lugar, para podermos ajustar o modelo, fazemos uma discretização dos dados, na qual substituímos o nível de expressão x_{ij} por um valor de 0 a m , através por exemplo do seu quantil observado.

Abordagem dois

Nessa abordagem considera-se genes como 'pessoas' e condições experimentais como 'itens'. Agora o interesse está na identificação de genes diferentemente expressos entre as condições experimentais. A intenção é identificar traços latentes associados com cada gene baseando-se nas medidas de expressão desses genes em diversas condições experimentais.

Essa formulação permite a comparação da diferença de expressão de genes dentro de uma mesma condição experimental e a identificação de genes mais ou menos expressos nessa dada condição. Além disso, a comparação de traços latentes do mesmo gene entre diversas condições experimentais, permite identificar genes que estejam diferentemente expressos entre essas diferentes condições.

Exemplo de uma curva característica nesse contexto

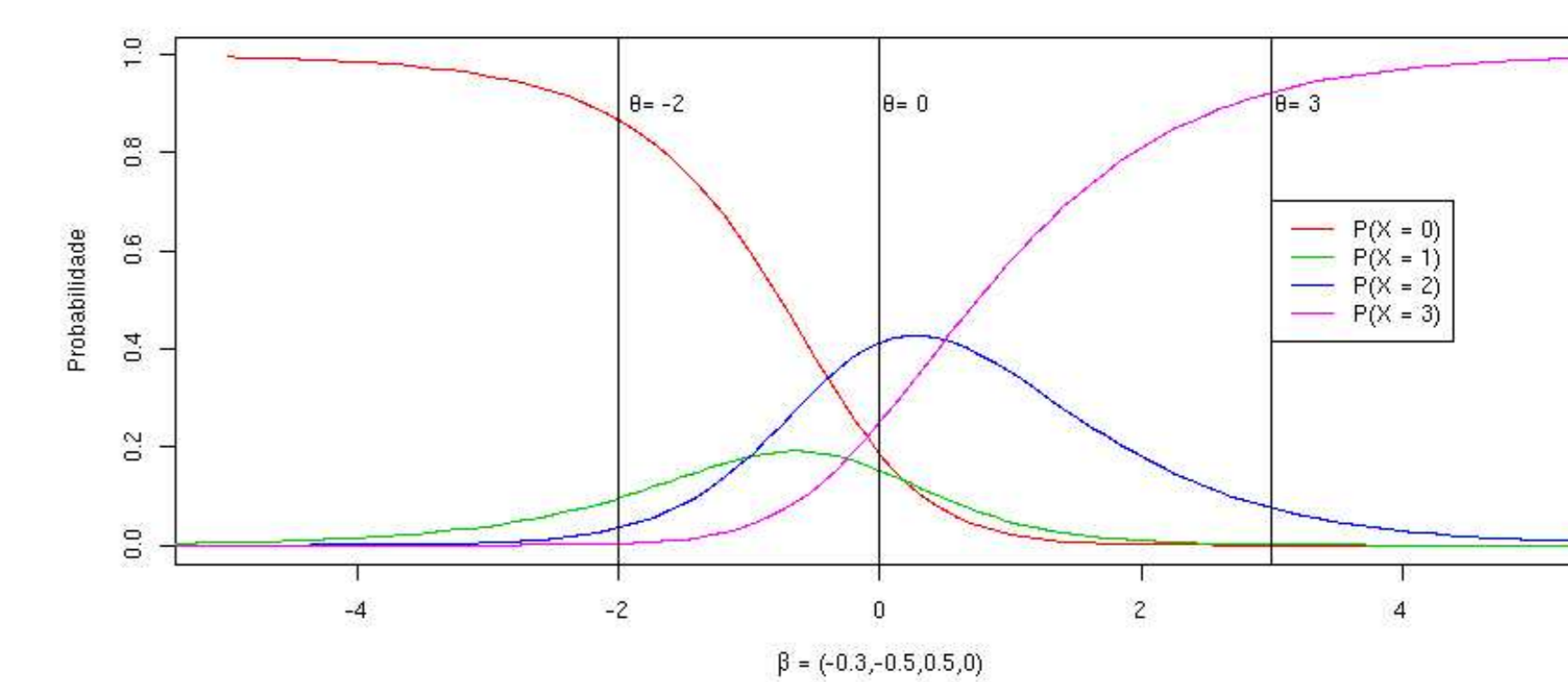


Figura 3: Curva característica de um modelo de Rasch tentativo.

Referências

- [1] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.
- [2] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2(4):364–374, 1997.
- [3] G. A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nat. Gen. Sup.*, 32:490–495, 2002.
- [4] W. Huber, A. von Heydebreck, and M. Vingron. *Handbook of Statistical Genetics*, chapter Analysis of Microarray Gene Expression Data, pages 162–187. John Wiley & Sons, Ltd., 2003.
- [5] H. Li and F. Hong. Cluster-Rasch models for microarray gene expression data. *Genome Biology*, 2(8):research0031.1–0031.13, 2001.
- [6] 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.
- [7] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4), 2002.