

Extração multilinear de informações discriminantes em imagens de ressonância magnética do cérebro humano

Rafael D. Leão¹, João R. Sato², Carlos E. Thomaz¹

¹ Departamento de Engenharia Elétrica – FEI – São Bernardo do Campo – SP – Brasil

² Centro de Matemática, Computação e Cognição – UFABC – Santo André – SP – Brasil

rafaeldoroleao@gmail.com

Abstract. *This paper presents a study about multilinear extraction of discriminant information using a set of magnetic resonance (MR) images of the human brain. The images set investigated is composed of 169 healthy subjects that have been analyzed using 6 distinct two-group characteristics (height, ever smoked, gender, hypertension, age and obesity). Univariate and multivariate statistical methods were used to extract discriminant information. A geometric analysis highlights the main principles and limitations of each statistical method, and the differences between the results are presented and discussed visually and quantitatively.*

Resumo. *Este artigo apresenta um estudo sobre extração multilinear de informações discriminantes de um conjunto de imagens RM estrutural do cérebro humano. O conjunto analisado é formado por 169 indivíduos saudáveis e são analisados através de 6 características distintas entre grupos de amostras (altura, hábito de fumar, gênero, hipertensão, idade e obesidade). Métodos estatísticos univariado e multivariado foram utilizados para a extração de informações discriminantes. Uma análise geométrica enfatiza as características principais e limitações de cada método estatístico, e são apresentadas e discutidas as diferenças entre os resultados de forma visual e quantitativa.*

1. Introdução

O avanço da tecnologia na área de neuroimagem tem possibilitado a geração de imagens de ressonância magnética (RM) com resoluções cada vez maiores, permitindo extrair diferenças estruturais sutis entre características distintas de indivíduos saudáveis. Atualmente, pode-se trabalhar com imagens cujo tamanho de voxel seja igual ou inferior a 1mm^3 , gerando imagens com resolução de milhares de voxels. Contudo, já existem técnicas que permitem alcançar uma resolução com tamanho de voxel inferior a 10 nanômetros [Degen et al. 2009]. Esta alta dimensionalidade dos dados torna cada vez mais difícil uma avaliação visual precisa da imagem por especialistas, a medida que a quantidade de informação aumenta para uma mesma região analisada. Assim, fica evidente a necessidade de uma análise computacional que permita destacar as diferenças entre conjuntos de imagens para auxiliar no trabalho do especialista (CAD - Computer Aid Diagnosis).

As primeiras análises computacionais de imagens RM, realizadas em meados da década de 1980, utilizavam regiões de interesse ou *Regions of Interest* (ROI), que não

permitted identifying differences beyond ROIs between groups of samples. Thus, in the final decade of 1980, images were initially segmented to be analyzed statistically voxel-by-voxel. As ROIs were manually delimited, there was in the era the need to overcome problems inherent to a global analysis, such as: the need for non-brain tissues from images, the spatial alignment of the same, the aggravation of the problem of multiple comparisons in massively univariate analyses, and the need to propose computationally efficient multivariate analysis methods that would allow investigating all voxels of the image simultaneously. In the beginning of the 1990s, the GLM (*General Linear Model*) was introduced in neuroimaging providing the basis for the SPM (*Statistical Parametric Mapping*) [Friston et al. 1995] which performs a global analysis, but still massively univariate, of the images. This method was intensively used in many works of the 1990s and is still quite used today. In researches using global images, spatial normalization methods were being automated through linear and non-linear transformations, which initially used reference images with markers and later without markers. Since the beginning of the 2000s, with the application of feature selection methods and recognition of patterns, the global analysis of images in a multivariate way has been processed and evaluated successfully in problems of information extraction in RM images [Golland et al. 2005, Kawasaki et al. 2007, Thomaz et al. 2007, Sato et al. 2008, Sato et al. 2009].

In this work, a comparative study is presented between univariate and multivariate analyses of discriminant information in RM structural brain images, analyzed based on the intensity of the voxels. A set of images from healthy individuals was evaluated through 6 different characteristics: height, smoking habit, gender, hypertension, age and obesity. As in the acquisition of these images, the brain volumes of the individuals do not end up exactly in the same spatial region of the generated images and these volumes can present different sizes or shapes making a direct comparison between the images, a pre-processing step was used to normalize spatially these images. In problems where the number of variables (or voxels) is very large, as in this case, an analysis as univariate faces a known problem of multiple comparisons, while a multivariate analysis faces a problem of few samples. Thus, in this work a comparison is also made between these methods with the aim of facilitating the understanding of the characteristics and limitations of each statistical approach. The results of these analyses were compared both visually and quantitatively, highlighting the differences and evidencing the importance of an analysis that coherently highlights the differences between the groups.

2. Análises estatísticas

This section aims to present the differences between a univariate statistical analysis, which does not take into account the size of the image, since it performs a voxel-by-voxel analysis, and a multivariate statistical analysis, which considers the image as a whole, where each image represents a point in an n -dimensional space, where n is the total number of voxels in the image.

2.1. Estatística univariada

O conceito envolvido por trás do modelo univariado baseia-se em um teste de hipóteses, onde uma probabilidade está associada a uma certa diferença encontrada ser ou não estatisticamente significativa. Neste trabalho, foi utilizado o teste t como teste de hipóteses. O t-valor de cada voxel da imagem é dado pela diferença entre as médias de cada grupo, ponderado pelo desvio padrão, conforme a seguinte equação:

$$t_k = \frac{\bar{x}_{1,k} - \bar{x}_{2,k}}{\sigma_k \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}, \quad (1)$$

onde $\bar{x}_{1,k}$ e $\bar{x}_{2,k}$ são respectivamente as médias do voxel k para os grupos 1 e 2, σ_k é o desvio padrão ponderado de todas as amostras para o voxel k , N_1 e N_2 são respectivamente o total de amostras do grupo 1 e do grupo 2, e t_k é o t-valor correspondente à variação de cada voxel. O desvio padrão ponderado do conjunto de amostras é definido como:

$$\sigma_k = \sqrt{\frac{(N_1 - 1)(\sigma_{1,k})^2 + (N_2 - 1)(\sigma_{2,k})^2}{N_1 + N_2 - 2}}, \quad (2)$$

onde $\sigma_{1,k}$ e $\sigma_{2,k}$ são respectivamente os desvios padrões da variável k para os grupos 1 e 2.

Entretanto, como uma taxa de erro é admitida no teste de hipóteses, e estas probabilidades estão associadas a testes independentes, quando estas são analisadas massivamente, um problema conhecido como problema de múltiplas comparações reduz a probabilidade de uma diferença apontada como estatisticamente significativa ser verdadeira, devido ao erro admitido ser ampliado por comparações sucessivas.

2.2. Estatística multivariada

O conceito envolvido por trás da análise multivariada está relacionado com a construção de um classificador que separe as amostras de dois grupos distintos no espaço n -dimensional.

Um dos problemas enfrentados neste tipo de análise é calcular o espalhamento dos dados quando estamos analisando amostras com milhares de variáveis. Por este motivo, para viabilizar esta análise, a técnica de PCA (*Principal Component Analysis*) [Fukunaga 1990] foi utilizada com a finalidade de reduzir a dimensionalidade das amostras sem perda de informação. A técnica de PCA realiza uma mudança de base vetorial, onde os eixos desta nova base descrevem as direções de maior espalhamento destes dados. Assim, para conjuntos de amostras onde o número de variáveis é muito maior que o número de amostras, estes dados podem ser representados, sem perda de informação, pelo número de amostras (N) menos 1 [Thomaz et al. 2007]. O cálculo da matriz de transformação para projeção das amostras para o espaço do PCA está descrito com maiores detalhes na seção 4 de Experimentos.

Para a construção do hiperplano de separação, descrito pela direção de maior espalhamento das amostras, foi utilizado o método LDA (*Linear Discriminant Analysis*). Neste método, Fisher [Fisher 1936] propõe a maximização da matriz de espalhamento inter-classes e a minimização da matriz de espalhamento intra-classes. A maximização da razão proposta por Fisher é dada por:

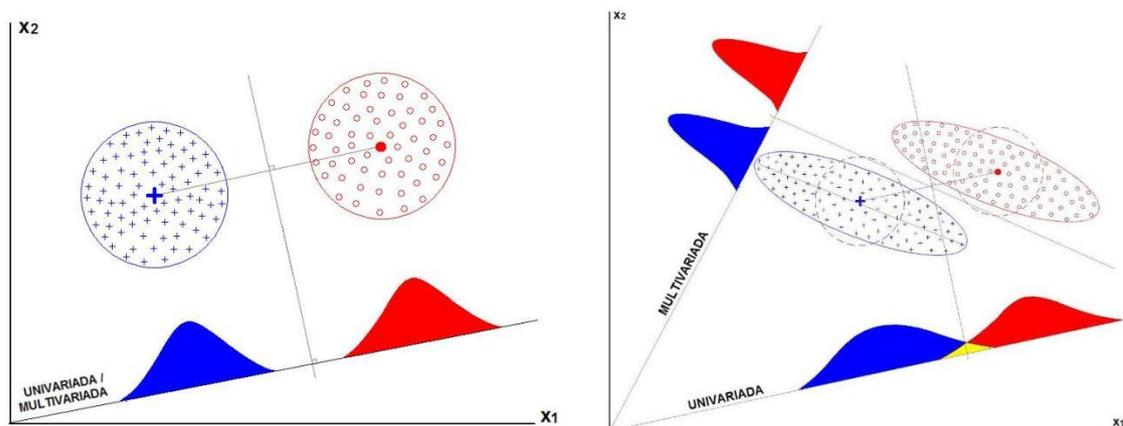


Figura 1. A figura ilustra geometricamente a separação dos grupos através de análises univariada e multivariada, para uma distribuição descrita por hiperesferas (imagem à esquerda) e hiperelipses (imagem à direita).

$$P_{lda} = \operatorname{argmax} \frac{|P^T S_b P|}{|P^T S_w P|}, \quad (3)$$

onde S_w é a matriz de espalhamento intra-classes e S_b a matriz de espalhamento inter-classes.. No entanto, existe uma instabilidade na inversão de S_w , quando o número de amostras é inferior ao número de variáveis. Para superar este problema, foi utilizado o método MLDA (*Maximum uncertainty Linear Discriminant Analysis*) que substitui a matriz S_w por uma matriz regularizada, conforme descrito em [Thomaz et al. 2006].

2.3. Comparação geométrica

Com a finalidade de enfatizar as diferenças entre os resultados obtidos pelas análises univariada e multivariada, a Figura 1 ilustra do lado esquerdo grupos de amostras com espalhamentos descritos por hiperesferas e do lado direito descritos por hiperelipses. Por esta análise geométrica, nota-se que os resultados obtidos na extração de informações discriminantes para grupos de amostras diferentes é influenciado fundamentalmente pelo tipo de espalhamento dessas amostras. Nota-se que quando o espalhamento das amostras é descrito por uma hiperesfera as diferenças discriminantes encontradas entre os grupos é igual para as análises univariada e multivariada e não há restrição com relação a utilização de um método em comparação ao outro, com exceção do problema de múltiplas comparações inerente à análise massivamente univariada. No entanto, quando o espalhamento das amostras não descreve uma hiperesfera, somente uma análise como a multivariada consegue extrair linearmente informações que descrevem as diferenças relevantes entre os grupos. Em outras palavras, na análise univariada pode-se detectar diferenças que na verdade não são discriminantes e estão relacionadas fundamentalmente à hipótese (inválida) de ambos espalhamentos poderem ser descritos por hiperesferas.

3. Material utilizado

Foram utilizadas 169 imagens de RM do cérebro humano de indivíduos saudáveis. Estas imagens são do tipo T1 e foram corrigidas espacialmente e adquiridas previamente

pelo Departamento de Computação do *Imperial College London*, sendo alinhadas para o padrão MNI (*Montreal Neurological Institute*) ICBM152 (*International Consortium for Brain Mapping*) pelo software SPM5 [Friston et al. 1995]. Estas imagens foram redimensionadas para 182x218x182.

Os indivíduos foram analisados com relação às características de altura, hábito de fumar, gênero, hipertensão, idade e obesidade. Para a análise de cada característica, os indivíduos foram classificados respectivamente como: baixos e altos, não fumantes e fumantes, homens e mulheres, não hipertensos e hipertensos, jovens e idosos, magros e obesos. Foram considerados baixos os indivíduos com tamanho inferior à 1,70m, jovens os indivíduos com menos de 46 anos e magros aqueles com IMC (Índice de Massa Corporal) [ABESO 2008] inferior à 24. Estes valores limiares foram definidos de forma a deixar os grupos com uma quantidade mais homogênea de indivíduos.

4. Experimentos

Inicialmente foi realizada uma etapa de pré-processamento com a finalidade de remover os tecidos não cerebrais das imagens e normalizar espacialmente a região cerebral. Para isso, foi utilizado o software SPM5 [Friston et al. 1995] para fazer uma segmentação das imagens em substância branca, cinza e líquido cefalorraquidiano, e posteriormente realizar uma suavização através de um filtro Gaussiano com largura de 8mm, das imagens segmentadas. Neste pré-processamento as imagens foram redimensionadas para 91x108x91. Como a finalidade deste trabalho é extrair as diferenças estruturais do cérebro, as imagens referentes ao líquido cefalorraquidiano não foram utilizadas na análise.

Na análise estatística univariada, foram calculados os t-valores de cada voxel das imagens, formando um mapa de t-valores para cada característica investigada e para cada substância cerebral. Para a identificação das regiões com maior probabilidade de diferenças estatisticamente significantes, foi calculado o t-valor de corte através do software “R” obtido em <http://www.r-project.org/>.

Na análise estatística multivariada, com a finalidade de reduzir ainda mais a dimensionalidade dos dados, foram removidos os fundos das imagens através de uma máscara criada para cada substância cerebral. Essa máscara foi gerada definindo um valor de limiar para a imagem da média das amostras, de modo que abrangesse toda região cerebral. Essa remoção do fundo das imagens reduziu em mais de 3 vezes a dimensionalidade das imagens. Estas imagens sem fundo foram, então, projetadas no espaço do PCA e em seguida no espaço do MLDA. Assim, no espaço do MLDA, foram calculadas as médias e desvios padrões de cada grupo investigado e estas informações puderam ser projetadas de volta para o espaço original das imagens, de forma que fosse possível calcular o mapa de t-valores e, assim, visualizar as regiões com diferenças mais significativas em uma análise multivariada. Para cada substância cerebral de cada característica, foram calculados os respectivos mapas de t-valores. A Figura 2 ilustra os processos de treinamento, classificação e reconstrução das imagens pela análise multivariada.

Para estimar a taxa de classificação dos hiperplanos calculados, foi utilizado a técnica de *cross-validation*. Nesta estimativa, o conjunto de amostras foi dividido em 10 partes, mantendo suas respectivas proporções para cada característica, e separando uma das partes para ser testada e as demais para o cálculo do hiperplano classificador.

A comparação visual e quantitativa dos resultados foi obtida determinando uma

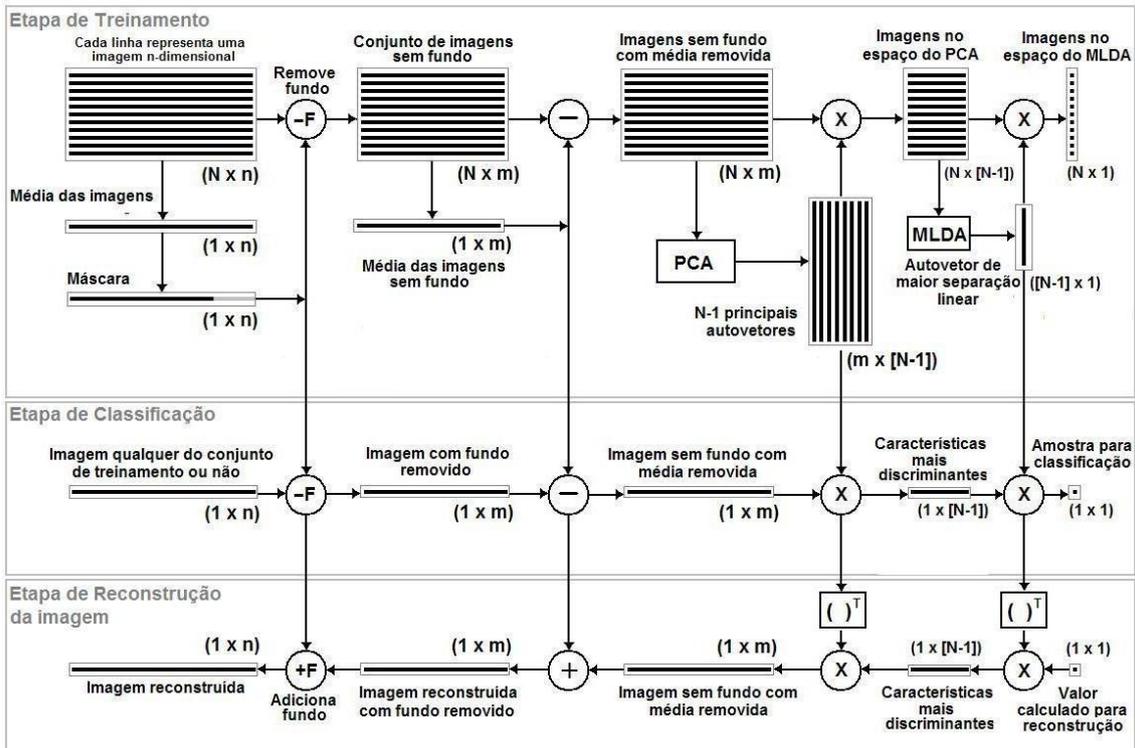


Figura 2. Representação gráfica da análise multivariada das imagens.

quantidade fixa de voxels mais significativos para os mapas de t-valores da análise multivariada (Ex.: 1% ou 5% dos maiores t-valores em módulo) e estes resultados foram comparados com os resultados do teste t da análise univariada, utilizando, por exemplo, 1% ou 5% de nível de significância. Assim, as regiões destacadas puderam ser comparadas visualmente e o valor de limiar destas regiões comparados quantitativamente. A análise das regiões exclusivas de cada característica permitiu comparar visualmente o espalhamento das regiões identificadas por cada modelo estatístico.

5. Resultados

A característica de idade conferiu maior confiabilidade para as regiões selecionadas linearmente como mais discriminantes pelos classificadores, apresentando uma taxa de classificação superior a 85%. Observou-se para a característica de obesidade que a taxa de classificação geral superestimou a classificação para o grupo de indivíduos magros. A Figura 3 apresenta as taxas de classificação para cada uma das características analisadas e cada substância cerebral.

A Figura 4 apresenta alguns *slices* da comparação visual entre os resultados das análises univariada e multivariada, representando por cores diferentes e destacando também as regiões coincidentes. Nota-se por esta figura que a diferença conceitual dos dois métodos estatísticos reflete diferenças consideráveis em seus resultados. Por exemplo, para a característica de idade, foram encontradas evidências de diferenças significativas para a substância cinzenta nas regiões do giro parietal superior, giro temporal transversal, tálamo lateral, amígdala e hipocampo, conforme também observado pela análise massivamente univariada realizada em [Good et al. 2001].

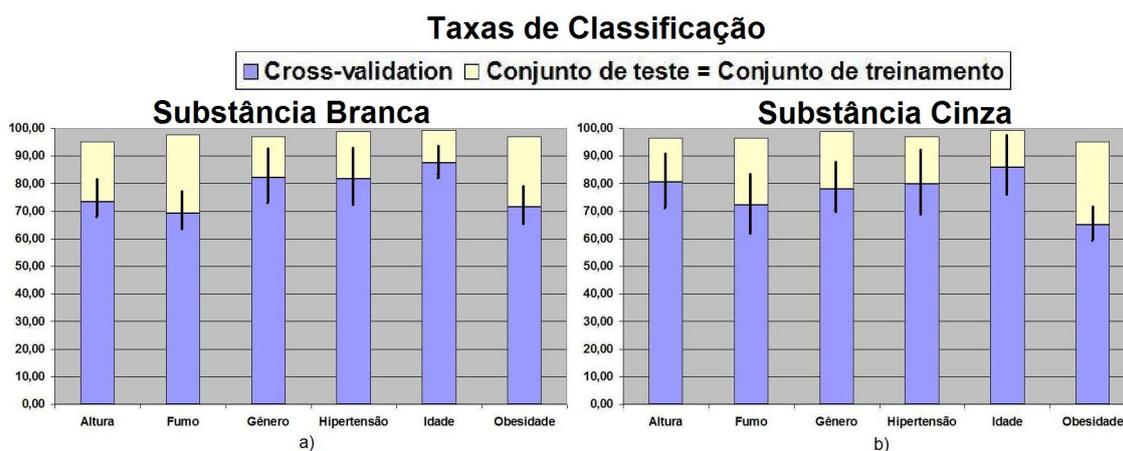


Figura 3. Taxa de classificação dos testes de cross-validation, com média e desvio padrão, e taxa de classificação com o conjunto de teste igual ao conjunto de treinamento.

A tabela da Figura 5 apresenta os t-valores de limiar das regiões selecionadas, permitindo realizar uma comparação quantitativa dos resultados. Observa-se que a análise univariada apresenta valores próximos ou acima do t-valor correspondente ao nível de significância. Nota-se também que, para a análise multivariada, os t-valores de limiar foram sempre inferiores aos da análise univariada. Isso indica que uma análise univariada destaca uma quantidade muito maior de diferenças significantes e, conseqüentemente, a possibilidade de se destacarem falsos positivos como diferenças relevantes entre grupos é maior.

Adicionalmente, para as 6 características estudadas, outra comparação realizada se referiu à identificação das regiões exclusivas de cada característica, conforme Figura 6. Nota-se para as regiões exclusivas da análise univariada que houve maior dispersão das regiões. Para a análise multivariada, a única característica que pode ser considerada como apresentando regiões exclusivas mais evidentes é a característica de idade. Essa diferença da quantidade de regiões exclusivas para a característica de idade evidencia o fato desta característica ter apresentado uma taxa de classificação maior.

6. Conclusão

A análise multivariada, em todos os casos, apresentou valores de limiar inferiores aos obtidos pela análise univariada, indicando que para um mesmo t-valor uma quantidade menor de voxels seriam destacados como diferenças estatisticamente significantes. Por isso, deve-se tomar cuidado ao se utilizar o modelo univariado, pois como o teste de hipótese admite uma certa taxa de erro, para um conjunto de variáveis muito grande, como neste caso, a quantidade de variáveis identificadas erroneamente pode-se tornar significativa para a interpretação dos resultados. Os resultados deste trabalho corroboram as afirmações de Davatzikos [Davatzikos 2004].

Observou-se através da análise das regiões exclusivas de cada característica, que os espalhamentos das amostras investigados pela análise multivariada permitiu definir regiões menos dispersas do que para a univariada e destacar as características que apresentam regiões que podem discriminar linearmente imagens RM de grupos distintos, como no caso da idade. Portanto, acredita-se que a análise multivariada é mais adequada que

a univariada para extração de informações discriminantes em imagens RM do cérebro humano, pois analisa o cérebro como um todo e está baseada na construção de um classificador que permite julgar se os grupos de amostras a serem avaliados podem ser separados linearmente com altas taxas de acerto. Como continuação deste trabalho, pretende-se realizar uma análise baseada em modelos deformáveis, para permitir avaliar o tipo de deformação (compressão, expansão) de cada região.

Agradecimentos

Os autores deste trabalho gostariam de agradecer aos colaboradores do Departamento de Computação do *Imperial College London*, especialmente ao Professor Dr. Daniel Rueckert, pelas imagens disponibilizadas para a realização deste trabalho.

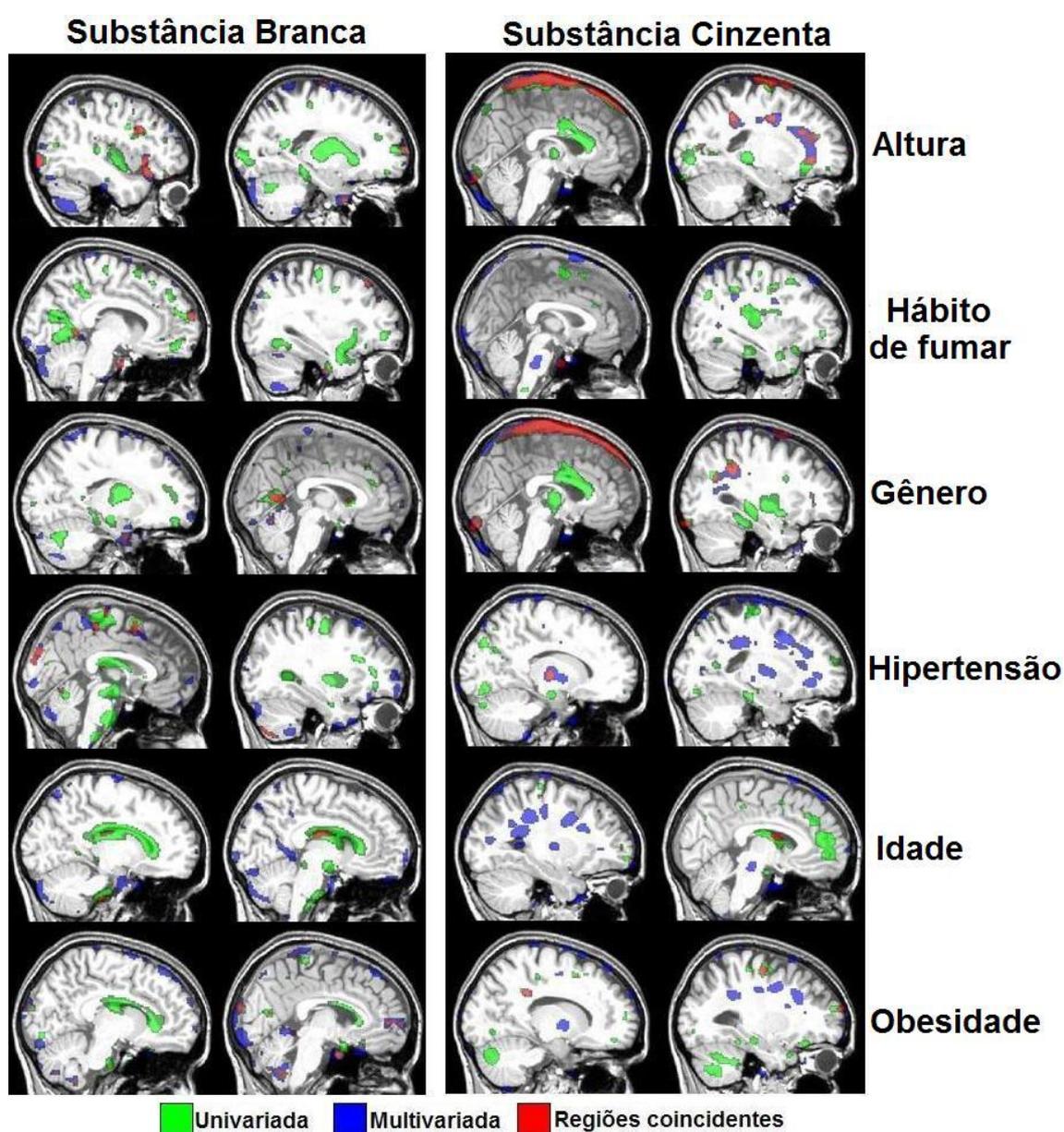


Figura 4. Regiões destacadas em cada tipo de análise, com 1% dos voxels com maiores coeficientes Studentizados em módulo selecionados.

Maiores voxels ativos Valor-t correspondente	Substância Cinza				Substância Branca			
	5% (45132 voxels) t ≥ 1.9743		1% (9027 voxels) t ≥ 2.6056		5% (45132 voxels) t ≥ 1.9743		1% (9027 voxels) t ≥ 2.6056	
	Univariada	Multivariada	Univariada	Multivariada	Univariada	Multivariada	Univariada	Multivariada
Altura	1,7606	0,4477	2,8425	0,9940	1,6257	0,5465	2,5089	1,1535
Fumo	1,5116	0,3459	2,2345	0,7007	1,6254	0,4921	2,4756	1,0450
Gênero	1,9082	0,4798	3,0578	1,0558	1,7386	0,5938	2,7856	1,2430
Hipertensão	3,2304	0,4081	4,0652	0,8248	1,7881	0,5748	2,8306	1,1974
Idade	6,8520	0,7520	8,9637	1,5173	2,5169	0,7694	4,8495	1,5280
Obesidade	2,7249	0,3701	3,4782	0,7597	1,6850	0,5323	2,4472	1,1055

Figura 5. Análise quantitativa com t-valores de limiar.

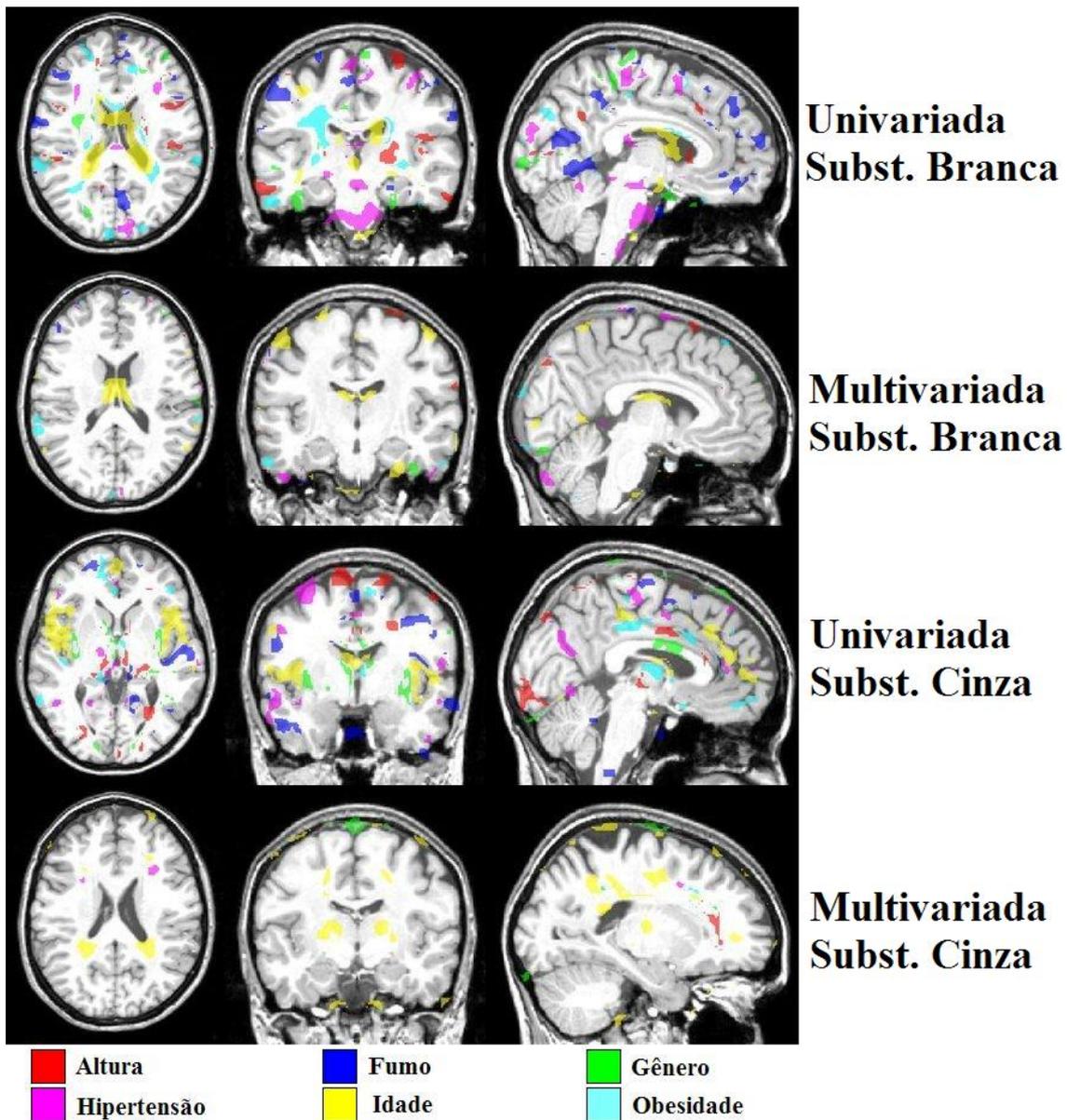


Figura 6. Regiões exclusivas de cada característica investigada.

Referências

ABESO (2008). Abeso - associação brasileira para o estudo da obesidade e da síndrome metabólica. <http://www.abeso.org.br/index.htm>. acessado em 15/07/2008.

- Davatzikos, C. (2004). Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage*, 23.
- Degen, C. L., Poggio, M., Mamin, H. J., Rettner, C. T., and Rugar, D. (2009). Nanoscale magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the USA*, pages 1–5.
- Fisher, R. A. (1936). The use of multiple measures in taxonomic problems. *Annals of Eugenics*, 7:179–188.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C., and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210.
- Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, 2 edition.
- Golland, P., Grimson, W. E. L., Shenton, M. E., and Kikinis, R. (2005). Detection and analysis of statistical differences in anatomical shape. *Medical Image Analysis*, 9(1):69–86.
- Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N. A., Friston, K. J., and Frackowiak, R. S. J. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, (14):21–36.
- Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S., Nakamura, K., Matsui, M., Sumiyoshi, T., Seto, H., and Kurachi, M. (2007). Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *NeuroImage*, 34(1):235–242.
- Sato, J. R., Fujita, A., Thomaz, C. E., Morais-Martin, M. G., Mourao-Miranda, J., Brammer, M. J., and Junior, E. A. (2009). Evaluating svm and mlda in the extraction of discriminant regions for mental state prediction. *NeuroImage*, 46(1):105–114.
- Sato, J. R., Thomaz, C. E., Cardoso, E. F., Fujita, A., Morais-Martin, M. G., and Junior, E. A. (2008). Hyperplane navigation: a method to set individual scores in fmri group datasets. *NeuroImage*, 42(4):1473–1480.
- Thomaz, C. E., Boardman, J., Counsell, S., Hill, D. L. G., Hajnal, J. V., Edwards, D., Rutherford, M. A., Gillies, D. F., and Rueckert, D. (2007). A multivariate statistical analysis of the developing human brain in preterm infants. *Image and Vision Computing*, 25(6):981–994.
- Thomaz, C. E., Kitani, E. C., and Gillies, D. F. (2006). A maximum uncertainty lda-based approach for limited sample size problems - with application to face recognition. *Journal of the Brazilian Computer Society*, 12(2):7–18.