

A Estrutura Dual na Redução de Dimensionalidade: Aplicação em Dados Multiômicos

Julia Pavan Soler¹.

Dados multiômicos (coletados do Genoma, Transcriptoma, Proteoma, Fenoma, etc.) inauguram uma nova fase de pesquisa nas diferentes áreas factuais, como Medicina, Agropecuária e até mesmo no Marketing. Os desafios impostos se estendem desde o planejamento do estudo, coleta, armazenamento e análise dos dados, até a validação dos resultados, mediados por dificuldades decorrentes da natureza multidisciplinar envolvida. Além disso, o estrondo provocado pelo big-data também tem produzido ecos na pesquisa multiômica, sedenta por representações apropriadas, em baixa dimensão, de dados em altíssima dimensão (especificamente, baixo tamanho amostral relativamente ao número de variáveis, isto é, $n \ll p$). Nesse contexto, muitos métodos de redução de dimensionalidade têm se apoiado nas propriedades de espaços duais, sob as quais, a análise de matrizes retangulares ($n \times p$) de dados é realizada com base nas (co)variâncias (espaço $p \times p$) ou nas distâncias (espaço $n \times n$), como é o caso de Componentes Principais e Coordenadas Principais, que permitem também o uso de soluções penalizadas. Adicionalmente, levando em conta o efeito de preditores ou do desenho do estudo, a Regressão Multivariada (ou a MANOVA) também tem sua formulação clássica via (co)variâncias estendida para a análise de componentes simultâneos (ASCA) ou ainda para as correspondentes matrizes de distância, como é o caso da estatística de Mantel. Na literatura, há diferentes propostas de análises gravitando entre as representações retangulares de dados e suas formas quadráticas decorrentes, sendo imperativo levar em conta a estrutura dos dados para se alcançar os resultados pretendidos de modo apropriado e em baixa dimensão. Neste Encontro, para atender à finalidade de redução de dimensionalidade, integração de bancos de dados, bem como para o aprendizado de estruturas causais, farei considerações sobre como estender propriedades da estrutura dual em dados independentes para a modelagem de dados com dependência entre indivíduos. A motivação vem da análise de dados multiômicos avaliados em indivíduos e em seus familiares amostrados da população brasileira. Este trabalho tem a colaboração da pesquisadora Adèle Ribeiro (Pós-doutoranda na Columbia University) e é parcialmente financiado pela FAPESP (Projeto 17/05125-7)

¹IME/USP