criticanarede.com · ISSN 1749–8457

CRÍTICA

9 de Maio de 2004 · Filosofia da ciência

Estatística multivariada

Uma visão didática-metodológica

J. M. Moita Neto

Introdução

Em qualquer decisão que tomamos em nossas vidas, sempre levamos em conta um grande número de fatores. Obviamente nem todos estes pesam da mesma maneira na hora de uma escolha. Às vezes, por tomarmos uma decisão usando a intuição, não identificamos de maneira sistemática estes fatores. Ou seja, não identificamos quais as variáveis que afetaram a nossa decisão.

Quando analisamos o mundo que nos cerca, identificamos que todos os acontecimentos, sejam eles culturais ou naturais, envolvem um grande número de variáveis. As diversas ciências têm a pretensão, de conhecer a realidade e de interpretar os acontecimentos (ciências humanas) e os fenômenos (ciências naturais), baseadas no conhecimento das variáveis intervenientes consideradas importantes nestes eventos.

Estabelecer relações, encontrar ou propor leis explicativas é o papel próprio da ciência. Para isso é necessário controlar, manipular, medir as variáveis que são consideradas relevantes ao entendimento do fenômeno analisado. Muitas são as dificuldades em traduzir as informações obtidas em conhecimento. A maior delas é de natureza epistemológica: a ciência não conhece a realidade, apenas a representa através de modelos e teorias dos diversos ramos do conhecimento.

Outra dificuldade é a aspiração de universalidade das explicações científicas. Ora, isto implica e condiciona a pesquisa a uma “padronização” metodológica. Um aspecto essencial desta padronização é a avaliação estatística das informações. A maneira própria de fazer ciência, procurando reduzir a poucas variáveis, desenvolveu muito um ramo da estatística que olha as variáveis de maneira isolada — a estatística univariada.

Somos cientificamente treinados a analisar as variáveis isoladamente e a partir desta análise fazer inferências sobre a realidade. Esta simplificação tem vantagens e desvantagens. Quando um fenômeno depende de muitas variáveis, geralmente este tipo de análise falha, pois não basta conhecer informações estatísticas isoladas, mas é necessário também conhecer a totalidade destas informações fornecida pelo conjunto das variáveis. As relações existentes entre as variáveis não são percebidas e assim efeitos antagônicos ou sinergéticos de efeito mútuo entre variáveis complicam a interpretação do fenômeno a partir das variáveis consideradas. Porém, no caso restrito de variáveis independentes entre si é possível, com razoável segurança, interpretar um fenômeno complexo usando as informações estatísticas de poucas variáveis. As informações estatísticas mais relevantes neste tipo de análise são as medidas de tendência central e de dispersão dos dados.

O desenvolvimento tecnológico oriundo das descobertas científicas tem alavancado o próprio desenvolvimento científico, ampliando em várias ordens de grandeza a capacidade de obter informações de acontecimentos e fenômenos que estão sendo analisados. Uma grande massa de informação deve ser processada antes de ser transformada em conhecimento. Portanto, cada vez mais estamos necessitando de ferramentas estatísticas que apresentem uma visão mais global do fenômeno que aquela possível numa abordagem univariada. A denominação “Análise Multivariada” corresponde a um grande número de métodos e técnicas

que utilizam simultaneamente todas as variáveis na interpretação teórica do conjunto de dados obtidos.

Para que não haja qualquer mistificação dos métodos de análise multivariada convém lembrar que estes métodos padecem dos mesmos problemas de toda a estatística. A estatística tem uma *quasi*-circularidade pouco explorada nos textos: pesquisamos para dizer algo significativo sobre o universo que elegemos, porém a pesquisa só será significativa se conhecermos suficientemente o universo para escolhermos adequadamente as variáveis e as condições de amostragem. A objetividade da pesquisa científica só começa depois da escolha das variáveis e das metodologias de análise, antes disto à atividade científica é completamente subjetiva.

Obviamente, o resultado de toda pesquisa cientifica está contaminada por este viés de nossa subjetividade. Para entender melhor, vamos exemplificar com a análise de água de um rio. O pesquisador piauiense não tem motivos para analisar mercúrio nos rios Poti ou Parnaíba pois não há atividade de garimpo nas proximidades destes rios. Não havendo registro conhecido de curtume ou de outra atividade industrial específica muito dos íons metálicos não serão pesquisados. A matéria orgânica será determinada de forma global e não se investiga substâncias específicas, a não ser que haja indícios de alguma contaminação. Considerando que aquilo que não se investiga jamais será descoberto, entende-se a subjetividade de um resultado de uma análise de água pelo que se deixou de dizer e a sua objetividade pelo que foi dito no laudo técnico de análise.

Estatística

Os diversos métodos de análise multivariados guardam entre si a necessidade de implementação computacional dos fundamentos teóricos que subjazem em suas

abordagens. A complexidade matemática, própria dos métodos multivariados, sugere, como medida de bom senso, uma descrição desmatematizada de seus conteúdos, remetendo ao uso do software estatístico o trabalho enfadonho do cálculo. Os programas estatísticos bem construídos escondem o edifício matemático atrás de uma interface amigável ao pesquisador. O professor de estatística hoje pode se dar ao luxo de transmitir o significado estatístico do tratamento de dados sem entediar os alunos com a profundidade das deduções matemáticas, fazendo uso abundante de exemplos. Deste modo, é possível trabalhar a parte mais nobre desta ciência que é a inferência estatística. Ou seja, o que posso afirmar com os dados que tenho. Ou ainda, que conhecimento científico produzi no meu trabalho.

Talvez neste momento tenhamos chegado ao paradoxo interessante, a complexidade matemática pode ser substituída por uma simplicidade didática. Através do uso de software estatístico, é possível pensar estatística sem ser estatístico. Obviamente esta seria uma grosseria com nossos colegas se não for devidamente exemplificado. Não chamo eletricista para trocar lâmpada. Não procuro médico para resfriado. Em outras palavras, as trivialidades estatísticas — incluindo a análise multivariada — estão ao alcance de todos e sem o constrangimento matemático do passado.

Esta aparente facilidade esbarra em dois problemas de ordem prática: 1) as prateleiras cheias com a diversidade de métodos estatísticos confundem o usuário que não consegue identificar a melhor solução para seu problema. Neste caso, o estatístico se transforma em psicólogo e pergunta: ”qual o seu problema?”, ou ”o que você pretende mostrar em sua pesquisa?”. Depois aponta a ferramenta adequada. 2) O usuário não conhece suficientemente o sistema de trabalho e por isso não consegue fazer uma inferência adequada. Neste caso, o estatístico não pode ajudar, pois o objeto de pesquisa em si foge de sua

especialidade. Quando não há conhecimento teórico prévio do sistema, as dificuldades começam logo na amostragem, no início do trabalho científico.

É importante ressaltar que ninguém faz ciência sem expectativa. Esta surge do conhecimento teórico e do senso comum. A pesquisa científica consiste em traduzir esta expectativa em problema, a partir do problema manifestar uma proposta de trabalho e, desta proposta, escolher um procedimento metodológico adequado. A estatística é parte constitutiva deste procedimento metodológico, estando presente no seu início (amostragem e seleção das variáveis) e no seu fim (tratamento, análise e inferência sobre os dados). Vale lembrar que, por mais avançada que esteja a estatística, ainda não se pode abrir mão da intuição e da experiência precedente do pesquisador.

Métodos multivariados

Existem vários métodos de análise multivariada com finalidades bem diversas entre si. Portanto, voltamos ao primeiro passo, que é saber que conhecimento se pretende gerar. Ou melhor, o que se pretende afirmar a respeito dos dados. Para exemplificar esta diversidade, vamos propor alguns objetivos e indicar alguns métodos possíveis. Quando o interesse é verificar como as amostras se relacionam, ou seja, o quanto estas são semelhantes segundo as variáveis utilizadas no trabalho, destaco dois métodos que podem ser utilizados: a análise por agrupamento hierárquico (HCA) e a análise por componentes principais (PCA). Quando a finalidade principal é fazer previsão, por exemplo, quando temos muitas variáveis independentes e queremos encontrar uma variável dependente, a regressão linear múltipla e redes neurais são métodos indicados para esta situação. Com uma finalidade bem diversa, existem métodos de análise multivariada que podem ser usados na etapa inicial de uma pesquisa, na própria escolha das variáveis que descreverão o sistema. Isto é muito comum nos casos

em que um processo necessita ser otimizado. Dentre os métodos que servem para otimização, citamos o simplex e o planejamento fatorial.

Os métodos estatísticos são escolhidos de acordo com os objetivos da pesquisa, por isto, mostrar, predizer ou otimizar são obtidos por diferentes métodos. Portanto, a estatística multivariada, com os seus diferentes métodos, difere de uma prateleira de supermercado abarrotada de produtos com a mesma função, pois cada método tem sua fundamentação teórica e sua faixa de aplicabilidade. Vamos apresentar aqui dois destes métodos para aprofundar melhor a teoria subjacente e explicar suas aplicações.

Análise de agrupamento Hierárquico (HCA)

A análise de agrupamento hierárquico consiste no tratamento matemático de cada amostra como um ponto no espaço multidimensional descrito pelas variáveis escolhidas (Moita Neto, J. M., Moita, Graziella Ciaramella, "Uma Introdução à Análise Exploratória de Dados Multivariados", *Química Nova*, São Paulo, SP: v. 21, n. 4, p. 467-469, 1998). Também é possível, nesta técnica, tratar cada variável como um ponto no espaço multidimensional descrito pelas amostras, ou seja, podemos ter agrupamento de amostras ou de variáveis de acordo com o interesse em cada situação. Quando uma determinada amostra é tomada como um ponto no espaço das variáveis, é possível calcular a distância deste ponto a todos os outros pontos, constituindo-se assim uma matriz que descreve a proximidade entre todas as amostras estudadas.

Existem várias maneiras de calcular a distância entre dois pontos, a mais conhecida e utilizada é a distância euclidiana, pois corresponde ao sentido trivial de distância no plano. Relembrando que, para duas variáveis, corresponde a aplicação do teorema de Pitágoras ($a^2=b^2 + c^2$): O comprimento da hipotenusa (a) é igual à raiz quadrada da soma dos quadrados dos comprimentos dos catetos

(b e c). Baseada nesta matriz de proximidade entre as amostras, se constrói um diagrama de similaridade denominado dendrograma (dendr(o) = árvore). Existem várias maneiras de aglomerar matematicamente estes pontos no espaço multidimensional para formar os agrupamentos hierárquicos. Cada um corresponde a um algoritmo específico (ou seja, o modo particular como os cálculos serão feitos pelo computador), que usa as informações da matriz de proximidade para criar um dendrograma de similaridade. A interpretação de um dendrograma de similaridade entre amostras fundamenta-se na intuição: duas amostras próximas devem ter também valores semelhantes para as variáveis medidas. Ou seja, elas devem ser próximas matematicamente no espaço multidimensional. Portanto, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas. O dendrograma hierarquiza esta similaridade de modo que podemos ter uma visão bidimensional da similaridade ou dissimilaridade de todo o conjunto de amostras utilizado no estudo. Quando o dendrograma construído é das variáveis, a similaridade entre duas variáveis aponta forte correlação entre estas variáveis do conjunto de dados estudado. Os dendrogramas de amostras são mais comuns.

A aplicação da análise de agrupamento hierárquico, quando temos variáveis de escalas diferentes, deve ser precedida por um tratamento prévio dos dados. Quando não é feito o pré-tratamento, as variáveis com valores numéricos mais altos serão mais importantes no cálculo que as variáveis com valores numéricos mais baixos. O pré-tratamento mais comumente empregado é a transformação Z, que transforma as medidas de cada variável de tal modo que o conjunto de dados tenha média zero e variância um. A finalidade deste procedimento é equalizar a importância estatística de todas as variáveis utilizadas. As dificuldades matemáticas envolvidas nestes cálculos, hoje são removidas pelos pacotes estatísticos de grande amplitude e facilidade de uso, como é o caso do SPSS

(***S**tatistical **P**ackage for the **S**ocial **S**ciences*). O SPSS fornece todas as ferramentas para a obtenção do dendrograma de similaridade incluindo as diversas opções de distância, métodos de aglomeração e modos de transformação dos dados originais.

Análise de componentes principais (PCA)

A análise de componentes principais é uma técnica estatística poderosa que pode ser utilizada para redução do número de variáveis e para fornecer uma visão estatisticamente privilegiada do conjunto de dados. A análise de componentes principais fornece as ferramentas adequadas para identificar as variáveis mais importantes no espaço das componentes principais.

Os fundamentos da análise de componentes principais serão apresentados descrevendo os passos matemáticos e estatísticos a partir das necessidades de interpretação adequada da matriz de dados. O entendimento exaustivo do assunto requer o conhecimento de operações com matrizes e por isso optamos por uma abordagem conceitual usando as noções de álgebra linear.

Um ponto no gráfico cartesiano é representado por valores das coordenadas x e y. No caso de um gráfico tridimensional, a apresentação de um ponto corresponde aos valores das coordenadas x, y e z. Traduzindo isto para o mundo das amostras e das variáveis, o ponto é uma amostra e os valores em cada uma das coordenadas correspondem aos valores das variáveis medidas. Para exemplificar isto, vamos supor que estejamos medindo duas propriedades físicas como o ponto de fusão e o ponto de ebulição de várias moléculas. A molécula de água ficaria locada nas coordenadas (0 ^{o}C , 100 ^{o}C) deste gráfico. O álcool etílico ficaria locado nas coordenadas (-114 ^{o}C, 78 ^{o}C) e assim por diante. Caso se queira transformar a escala do ponto de fusão para Kelvin e a escala do ponto de ebulição para Fahrenheit, a representação da molécula de água continua a

mesma em relação às outras moléculas, embora mudem os eixos coordenados. Ou seja, a estrutura dos dados não é alterada por uma transformação de coordenadas (Anexo).

A análise de componentes principais consiste em reescrever as variáveis originais em novas variáveis denominadas componentes principais, através de uma transformação de coordenadas. A transformação de coordenadas é um processo trivial quando feito usando matrizes. A transformação matemática das coordenadas pode ser feita de diversas maneiras conforme o interesse. A transformação das variáveis originais em componentes principais tem algumas especificidades que explicaremos agora.

Os componentes principais são as novas variáveis geradas através de uma transformação matemática especial realizada sobre as variáveis originais. Esta operação matemática está disponível em diversos softwares estatísticos especializados. Cada componente principal é uma combinação linear de todas as variáveis originais. Por exemplo, um sistema com oito variáveis, após a transformação, terá oito componentes principais. Cada uma destas componentes principais, por sua vez, será escrita como uma combinação linear das oito variáveis originais. Nestas combinações, cada variável terá uma importância ou peso diferente.

Duas são as características das componentes principais que as tornam mais efetivas que as variáveis originais para a análise do conjunto das amostras (Prado, P. I., Lewinsohn, Thomas Michael, Carmo, R. L., Hogan, D. J. “Ordenação Multivariada na Ecologia e seu Uso em Ciências Ambientais. *Ambiente e Sociedade,* Campinas, SP: v.10, p. 69-83, 2002). As variáveis podem guardar entre si correlações que são suprimidas nas componentes principais. Ou seja, as componentes principais são ortogonais entre si. Deste modo, cada

componente principal traz uma informação estatística diferente das outras. A segunda característica importante é decorrente do processo matemático-estatístico de geração de cada componente que maximiza a informação estatística para cada uma das coordenadas que estão sendo criadas. As variáveis originais têm a mesma importância estatística, enquanto que as componentes principais têm importância estatística decrescente. Ou seja, as primeiras componentes principais são tão mais importantes que podemos até desprezar as demais. Destas características podemos compreender como a análise de componentes principais: a) podem ser analisadas separadamente devido à ortogonalidade, servindo para interpretar o peso das variáveis originais na combinação das componentes principais mais importantes b) podem servir para visualizar o conjunto da amostra apenas pelo gráfico das duas primeiras componentes principais, que detêm maior parte da informação estatística.

Comparação PCA e HCA

A análise de componentes principais e a análise de agrupamento hierárquico são técnicas de análise multivariada com fundamentos teóricos bem diferentes, podendo ser aplicadas independentemente. Estas técnicas podem até ser complementares na informação sobre o conjunto de dados, dependendo do sistema analisado. Ambas fornecem a visão mais global possível das amostras dentro do conjunto de dados, conforme as variáveis usadas (Cazar, R. A. “An Exercise on Chemometrics for a Quantitative Analysis Course”. *Journal of Chemical Education*, Madison, WI: v. 80, n. 9, p. 1026-1029, 2003).

Regressão Linear múltipla de componentes principais

A regressão linear múltipla também é uma técnica multivariada cuja finalidade principal é obter uma relação matemática entre uma das variáveis (a variável dependente) e o restante das variáveis que descrevem o sistema (variáveis

independentes). Sua principal aplicação, após encontrar a relação matemática é produzir valores para a variável dependente quando se têm as variáveis independentes.Ou seja, ela pode ser usada na predição de resultados. Obviamente, a soma das contribuições de diversas variáveis para uma determinada predição pode também ser feita usando as componentes principais, pois as mesmas têm a vantagem de poder ser tratadas de modo completamente independente. Portanto, é possível também fazer regressão linear múltipla das componentes principais.

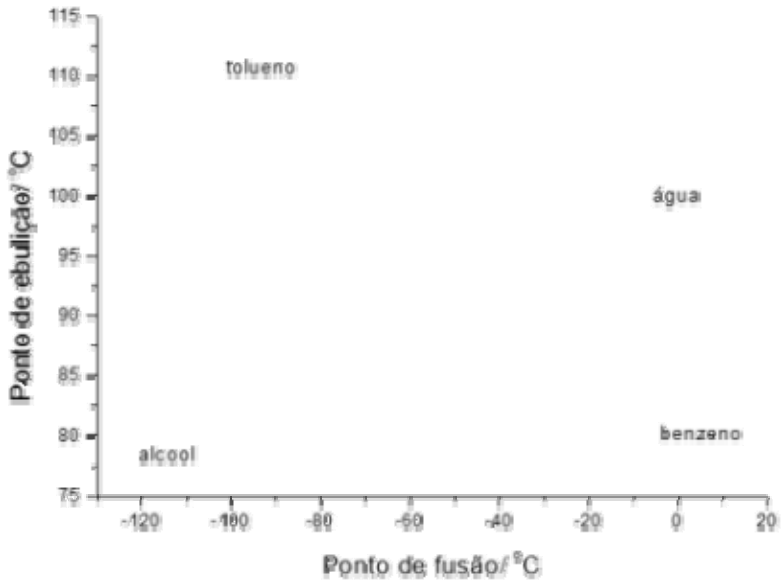
Conclusão

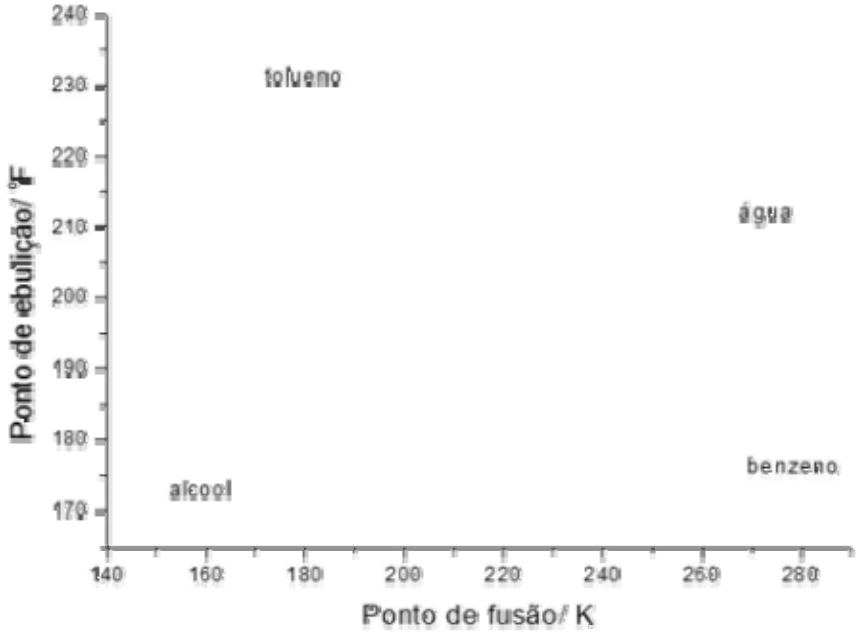
Objetivos bem precisos, desde o início da pesquisa, ajudam na consecução do trabalho e posterior tratamento estatístico. Mesmo quando o pesquisador não tem qualquer habilidade ou conhecimento de estatística, não pode deixar na mão do estatístico o seu conjunto de dados como se houvesse algum procedimento mágico para extrair informações daquele sistema. O ideal é o estabelecimento de um diálogo continuo entre pesquisador e estatístico para o primeiro apontar com clareza onde quer chegar e o que deseja dizer do sistema e o segundo informar os limites e possibilidades das técnicas estatísticas.

J. M. Moita Neto

Anexo

Transformação de coordenadas não modificam a estrutura dos dados — Mudança de escala de temperatura





Se gostou, apoie a *Crítica* fazendo uma subscrição ou clicando na publicidade. Sem o seu apoio não é possível continuar a editar a *Crítica*. Com o seu apoio, os tradutores podem ser pagos, o trabalho de formatar e editar a *Crítica* é também remunerado, e as despesas com o servidor não têm de ser suportadas pelo Director. Todas as sugestões e críticas são bem-vindas. Mais informações...