Ponto 13 (Teórica) - Estatística Multivariada

Ben Dêivide

16 de Novembro de 2016

Quando em determinado problema, tem-se a necessidade de estudar simultaneamente uma série de variáveis que podem ser associados a um determinado fenômeno, diz-se que este seria um problema de "análise multivariada". Desta maneira, qualquer método estatístico que permita a análise simultânea de duas ou mais variáveis pode ser considerado como pertencente ao campo da estatística multivariada.

A importância dos métodos multivariados é devido a sua aplicação em diversas áreas do conhecimento como medicina e saúde, sociologia, economia e negócios, educação, biologia, estudos ambientais, meteorologia, geologia, psicologia, esporte, dentre outros.

Formalmente, definimos a estatística multivariada da seguinte forma

Definição 1 (Estatística Multivariada)**.** *A Estatística Multivariada consiste em um conjunto de métodos estatísticos utilizados em situações nas quais várias variáveis são medidas simultaneamente, em cada elemento amostral.*

Basicamente, a Estatística Multivariada se divide em dois grupos: um primeiro, constituindo em **técnicas exploratórias de sintetização** (ou simplificação) da estrutura de variabilidade dos dados, e um segundo, consistindo em **técnicas de inferência**. Fazem parte do primeiro grupo métodos como a análise de componentes principais, análise de correspondência, análise de correlações canônicas, análise fatorial, análise de agrupamentos e análise discriminante. **Esses métodos têm apelo prático muito interessante, pois, na sua grande maioria, independem do conhecimento da forma matemática da distribuição de probabilidade geradora dos dados amostrais**. No segundo grupo, encontram-se os métodos de estimação de parâmetros, testes de hipóteses, análise de variância, de covariância e de regressão multivariada.

Devido as restrições do tempo, nessa dissertação iremos na primeira etapa dar uma visão geral das aplicações da estatística multivariada, apresentando algumas técnicas conhecidas, e no segundo momento iremos apresentar alguns métodos de inferência sobre o vetor de médias e matriz de covariâncias.

A primeira técnica apresentada é a **Análise de Componentes Principais** (ACP), que teve origem com os trabalhos de Pearson e Hottelling. Essa técnica surge da necessidade de se conhecer a estrutura de dependência das variáveis e a *priori* não é encontrado nenhum padrão de casualidade. Seu principal objetivo é explicar a estrutura de variâncias e covariâncias de um vetor aleatório composto de p-variáveis aleatórias iniciais, podendo-se resumir sua informação.

A ACP requer que os dados das p variáveis aleatórias sejam métricos. A técnica consiste basicamente em transformar um conjunto original de variáveis $(X_1, X_2, \ldots, X_p)$ em outro conjunto de dimensão equivalente $(C_1, C_2, \ldots, C_p)$, tal que:

$$C_j = e_{1j}X_1 + e_{2j}X_2 + \ldots, e_{pj}X_p, \tag{1}$$

em que $e_{jj'}$ são os coeficientes calculados pela técnica, $j, j' = 1, 2, \ldots, p$. Ela pode ser considerada uma técnica exata, pois em sua composição não se tem a presença do erro, sendo sua estrutura basicamente matemática. O novo conjunto de variáveis possui propriedades importantes e de interesse. Essa técnica busca imprimir um tratamento estatístico a um número relativamente alto de variáveis heterogêneas, que possuam um grau considerável de aspectos comuns, isto é, com um elevado grau de correlação entre si. Desta forma, o que se busca é condensar o conjunto inicial de muitas variáveis ($X_j,\ j = 1, 2, \ldots, p$) em um número bem menor de novas variáveis ($C_k,\ k = 1, 2, \ldots, q$, sendo $q < p$) chamadas **componentes principais** e conseguir uma pequena perda de informações. A técnica de **Análise de Correspondência** (AC) pode ser considerada um caso especial de ACP, porém dirigida a dados categóricos organizados em tabelas de contingência e não a dados contínuos.

Quando almejamos estabelecer a relação entre dois vetores $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]'$ e $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_q]'$, um contendo p e outro contendo q variáveis , podemos utilizar as pq correlações obtidas a partir de cada um dos pares de variáveis. Assim, obtém-se a matriz de covariância $\boldsymbol{S}$ particionada:

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix}, \tag{2}$$

sendo $\boldsymbol{S}_{11}$ $(p \times p)$ e $\boldsymbol{S}_{22}$ $(q \times q)$ as matrizes de variâncias e covariâncias amostrais entre as variáveis do conjunto $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente. As covariâncias entre as variáveis de diferentes conjuntos, uma variável de $\boldsymbol{X}$ e outra de $\boldsymbol{Y}$, estão contidas na matriz $\boldsymbol{S}_{12}$ $(p \times q)$ e $\boldsymbol{S}_{21}$ $(q \times p)$. Analisar essas covariâncias pode ser extremamente trabalhoso, ainda mais se p e q forem grandes.

Se quisermos sumariar a informação dos pq parâmetros em um conjunto menor de coeficientes de correlação entre dois vetores, devemos escolher a técnica da **Análise de Correlação Canônica** (ACC). A ideia básica é criar um par de variáveis latentes que sejam combinações lineares das variáveis dos dois vetores, de modo que a escolha dos coeficientes dessas combinações é feita tendo-se o critério de maximização da correlação entre as duas variáveis latentes. Podemos ainda estabelecer um segundo par, não correlacionado com o primeiro, que contenha o máximo de informação remanescente que não tenha sido contemplada no primeiro par. Podemos prosseguir esse procedimento até que toda a informação de covariância ou correlação entre os dois vetores, tenham sido explicada pelo pares de variáveis latentes. Esses pares de variáveis latentes são conhecidos como **variáveis canônicas**, e a correlação entre elas é a **correlação canônica**. Nesse tipo de análise não existe distinção entre variável independente e dependente, existem somente dois conjuntos de variáveis e se busca a máxima correlação entre ambos.

Um outro tipo de técnica multivariada é a **Análise Fatorial** (AF). Essa técnica surgiu dos estudos de Charles Spearman. O modelo fatorial é motivado pelo seguinte argumento: suponha que as variáveis podem ser agrupadas por suas correlações, isto é, todas as variáveis dentro de um particular grupo sejam altamente relacionadas entre si, mas tenham correlações pequenas com variáveis em grupos diferentes. Então, é admissível que cada grupo de variáveis represente um único fator, que é responsável pelas correlações observadas. É este tipo de estrutura que a análise fatorial pretende confirmar. A AF tem como princípio básico a redução do número original de variáveis respostas a um conjunto menor de "fatores" independentes e não observados, que explicam de forma mais simples e reduzida, as variáveis originais.

Para uma situação com p variáveis, o modelo de AF ortogonal pode ser expresso da

seguinte forma:

$$(Y_j - \bar{Y}_j) = a_{ij}F_1 + \ldots + a_{jm}F_m + e_j, \quad j = 1, 2, \ldots, p, \tag{3}$$

em que: Y_j são as variáveis respostas originais; $\bar{Y}_j$ as médias das variáveis; F_i são os fatores comuns e explicam as correlações entre as variáveis, sendo $i = 1, 2, \ldots, m$; a_{ji} são as cargas fatoriais, que refletem a importância do fator i na explicação da variável j; e e_j é o erro aleatório, que capta a variação específica da variável Y_j não explicada pela combinação linear das cargas fatoriais com os fatores comum.

Diferentemente da ACP, na formação das p componentes, o erro aleatório está presente no modelo de AF, o que torna a técnica não-exata. O AF e a ACP têm como principal objetivo uma redução da dimensionalidade do espaço das variáveis. Entretanto, existem diferenças entre as duas técnicas:

a) Na ACP a ênfase é explicar a variância total, em constraste com a AF que visa explicar as covariâncias entre as variáveis respostas;

b) Na AF as variáveis são expressas como combinações lineares dos fatores, enquanto que os CP's são funções lineares das variáveis originais;

c) As CP's são únicas, enquanto que os fatores são passíveis de rotações.

Se o objetivo é encontrar e descrever alguns fatores de interesse, a AF pode ser útil se o modelo de fatores se ajusta bem aos dados. Por outro lado, se o objetivo for definir um menor número de variâncias a serem utilizados em uma outra análise, iríamos ordinariamente preferir a ACP.

Na **Análise de Agrupamento** temos a intenção de classificar objetos, itens ou indivíduos de acordo com suas semelhanças. Os objetos semelhantes são alocados em um mesmo grupo e, portanto, aqueles que pertencem a diferentes grupos são considerados dissimilares. Em geral, **objetos muito semelhantes** são ditos **similares** e aqueles com **poucas semelhanças**, são denominados **dissimilares**. A semelhança entre os objetos é quantificada por meio de uma medida de proximidade, que engloba tanto as medidas de similaridade quanto as de dissimilaridades. Nas medidas de similaridades quanto maior for o valor mensurado maior vai ser a semelhança entre os objetos e quanto mais próximo de zero forem essas medidas menor será a semelhança entre os objetos considerados. Na medidas de dissimilaridades a interpretação é o oposto, ou seja, quanto maior o valor mensurado menos semelhantes são os objetos e quanto menor, mais semelhantes são eles. Devemos diferenciar duas situações básicas quando tomamos p medidas em n objetos: na primeira queremos agrupá-los em um número desconhecido de grupos,e na segunda, temos o interesse de classificá-los em um conjunto pré-existente de grupos. O primeiro caso se refere à análise de agrupamento, e no segundo, **Análise Discriminante**.

Teríamos muitas outras abordagens de técnicas multivariadas que poderíamos comentar, contudo, pelo tempo iremos discutir sobre a parte inferencial, dando enfoque aos seguintes pontos:

- Inferências sobre vetor de médias para uma população e duas populações (Cap. 5 e 7 - FERREIRA, 2011);

- Inferências sobre matrizes de covariâncias para uma, duas e mais populações (Cap. 6 - FERREIRA, 2011);

- Inferências sobre vetores de médias com mais de duas populações (Cap. 8 - MANOVA, FERREIRA, 2011).

Antes de explanarmos sobre esses pontos, iremos apresentar algumas definições.

Definição 2 (Vetor Aleatório)**.** *Seja* $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]'$ *um vetor p-dimensional. Se cada* X_i*,* $i = 1, 2, \ldots, p$ *é uma variável aleatória, diremos que* $\boldsymbol{X}$ *é um vetor aleatório.*

Definição 3 (Vetor de médias)**.** *Seja* $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]'$ *um vetor aleatório. Se para cada variável aleatória* X_i*, temos a esperança matemática de* X_i *dada por* $E[X_i] = \mu$*, então o vetor* $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]'$ *representa o vetor de médias do vetor* $\boldsymbol{X}$*.*

Definição 4 (Variância)**.** *A variância da i-ésima variável aleatória* X_i *do vetor* $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]'$ *é dado por* $Var[X_i] = \sigma_i^2 = \sigma_{ii}$*.*

Definição 5 (Covariância)**.** *A covariância entre* X_i *e* X_j *variáveis aleatórias do vetor* $\boldsymbol{X}$ *para* $i \neq j$*, é dado por* $Cov(X_i, X_j) = \sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$*.*

Quando $i = j$, $Cov(X_i, X_i) = Var[X_i] = \sigma_{ii}$.

Definição 6 (Matriz de Covariância)**.** *A matriz de variâncias e covariâncias do vetor* $\boldsymbol{X}$ *é dado por:*

$$Cov(\boldsymbol{X}) = V(\boldsymbol{X}) = Var[\boldsymbol{X}] = \Sigma_{p\times p} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{bmatrix}.$$

Definição 7 (Correlação)**.** *O coeficiente de correlação entre as variáveis* X_i *e* X_j *do vetor* $\boldsymbol{X}$*, para* $i \neq j$*, é denotado por:*

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}},$$

sendo $-1 \leq \rho_{ij} \leq 1$*, para* $i, j = 1, 2, \ldots, p$*. Quando* $i = j$*,* $\rho_{ij} = 1$*.*

Definição 8 (Matriz de correlação)**.** *A matriz de correlação do vetor aleatório* $\boldsymbol{X}$ *é denotada por:*

$$P_{p\times p} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \ldots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \ldots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \ldots & 1 \end{bmatrix}.$$