MAE5776 – Análise Multivariada 1º Sem/2022 Questões - Consolidando o Aprendizado

Aluno:
Thiago Ferreira Miranda - nº USP: 11925711

1. Entender a estrutura dos dados sob análise é importante em todas as análises estatísticas? Particularmente, em análise multivariada de dados, discuta por que isso é importante.

R: A estrutura dos dados é importante em análises estatísticas, em análises multivariadas de dados não é diferente, uma vez que é de suma importância entender características como o tamanho amostral (para saber por exemplo se iremos trabalhar com análises classicas, ou se estamos no big n ou big p), se as variáveis são independentes, se são quantitativas, se há variáveis categóricas, e se estas efetuam o papel de definição de grupo ou não, se o conjunto de variáveis pode ou não ser formulado como dois ou mais subconjuntos de variáveis. Tais exemplos de características na estrutura dos dados nos dá informações que podem nortear quais técnicas podem ser mais adequadas para a análise de dados multivariados.

2. Como a teoria de Espaços Duais pode ser útil na redução de dimensionalidade de dados?

R: A representação dos dados de formas quadráticas dos indivíduos ou das variáveis são úteis pois fornecem entradas para a decomposição espectral ou decomposição em valores sigulares para a geração de autovalores e autovetores, o que auxilia a criação de escores em menores dimensões. Num problema multivariado podemos utilizar os dados disponíveis na forma retangular da matriz de dados $Y_{n\times p}$, bem como na matriz de covariância $S_{p\times p}$ ou matriz de distâncias $D_{n\times n}$. Assim quando falamos em espaços duais vemos que muita informação que está contida na matriz $Y_{n\times p}$ também está contida na matriz de covariância $S_{p\times p}$ e matriz de distâncias $D_{n\times n}$, assim com a decomposição em valores singulares da matriz $Y_{n\times p}$, que resulta na matriz de autovetores U, matriz diagonal de autovalores e matriz de autovetores V, ou com a decomposição espectral da matriz de covariância $S_{p\times p}$ ou matriz de distâncias $D_{n\times n}$, as quais apresentam os mesmos autovalores, e neste sentido podemos por esta equivalência obter Componentes Principais ou Coordenadas Principais, para o caso métrico, por exemplo.

3. Justifique a afirmação: "As técnicas matriciais de Decomposição em Valores Singulares, bem como de Decomposição Espectral, são a base de muitas das análises de redução de dimensionalidade". R: Bloco

A decomposição em valores singulares da matriz retangular de dados $Y_{n\times p}$ ou a decomposição espectral das matrizes quadradas de covariancia $S_{p\times p}$ ou matriz de distâncias $D_{n\times n}$, tem os mesmos autovalores, com isto podemos obter Componentes Principais, por meio dos autovetores V relacionado às colunas (obtidos via decomposição em valores singulares), neste sentido, uma equivalência (que se dá para a solução métrica) pode também ser obtida com Coordenadas Principais,

por meio dos autovetores U relacionado às linhas (obtidos via decomposição em valores singulares), sendo que CP e CoP estão baseadas na decomposição espectral e,…….

Para justificar tal afirmação podemos citar as técnicas de Análise de Componentes Principais e Análise Coordenadas Principais que estão totalmente baseadas na Decomposição Espectral das formas quadráticas das matrizes de covariância $S_{p\times p}$ e de distâncias $D_{n\times n}$, respectivamente. E a decomposição espectral de uma formas quadráticas está também associada à Decomposição em Valores Singulares da matriz $Y_{n\times p}$, uma vez que por meio desta há uma partição que resulta na matriz de autovetores U (relativa às linhas), matriz diagonal de autovalores e matriz de autovetores V (relativa às colunas) e, tais resultados serão utéis para as análises de CP e CoP supracitadas. Além disso, vemos em CP que podemos efetuar a redução de dimensionalidade com a penalização de alguns autovetores

4. Considere a matriz de dados $Y_{n\times p}$ e as correspondentes formas quadráticas YY' e $Y'Y$.

(a) Se λ é um autovalor de $Y'Y$ com autovetor v. Mostre que λ é um autovalor de YY' com autovetor Yv (equivalentemente, com autovetor padronizado $Yv\lambda^{-1/2}$).

R: Podemos escrever a decomposição em valores singulares da seguinte forma (considerando dados padronizados):

$Y_{n\times p} = U_{n\times n}\Lambda^{1/2}V'_{p\times p}$, em que temos a matriz de autovetores U (relativa às linhas), Λ matriz diagonal de autovalores e matriz de autovetores V (relativa às colunas).

Se observarmos a matriz de similaridade entre indíviduos (representada na forma quadrática YY') ao realizar sua decomposição espectral teremos como resultado $YY' = U\Lambda U'$, aqui um ponto importante é a obtenção do mesmo autovalor Λ.

Já se observarmos a matriz de similaridade entre as variáveis (representada na forma quadrática $Y'Y$) ao realizar sua decomposição espectral teremos como resultado $Y'Y = V\Lambda^{1/2}V'$, e novamente aqui temos a obtenção do mesmo autovalor Λ. Este relacionamento entre as decomposições, o fato da obtermos o mesmo autovalor torna possível a equivalência entre as análises de Componentes Principais e Coordenadas Principais

(b) Como esse resultado pode ser usado para relacionar Componentes Principais e Coordenadas Principais? R: Resumindo temos que $Y_{n\times p} = U_{n\times n}\Lambda^{1/2}V'_{p\times p} \Longrightarrow YV = U\Lambda^{1/2}$, sendo $YV \longrightarrow$ Componentes Principais e $U\Lambda^{1/2} \longrightarrow$ Coordenadas Principais. Esta equivalência é garantida para o caso métrico, com distâncias Euclidianas. Para dados originais se faz necessária a utilização de uma matriz H pré multiplicando Y, com o intuito de fornecer linhas centradas de Y. Portanto, para o caso de dados originais, a decomposição em valores singulares é expressa por $(HY)_{n\times p} = U_{n\times n}\Lambda^{1/2}V'_{p\times p}$.

5. Considere a matriz de dados $Y_{n\times p}$ e as correspondentes formas quadráticas YY' e $Y'Y$.

(a) Estabeleça relações entre os autovalores e autovetores da decomposição em valores singulares da matriz retangular Y e das correspondentes decomposições espectrais das formas quadráticas. R: Podemos escrever a decomposição em valores singulares da seguinte forma (considerando dados padronizados):

$Y_{n\times p} = U_{n\times n}\Lambda^{1/2}V'_{p\times p}$, em que temos a matriz de autovetores U (relativa às linhas), Λ matriz diagonal de autovalores e matriz de autovetores V (relativa às colunas).

Se observarmos a matriz de similaridade entre indíviduos (representada na forma quadrática YY') ao realizar sua decomposição espectral teremos como resultado $YY' = U\Lambda U'$, aqui um ponto importante é a obtenção do mesmo autovalor Λ.

Já se observarmos a matriz de similaridade entre as variáveis (representada na forma quadrática $Y'Y$) ao realizar sua decomposição espectral teremos como resultado $Y'Y = V\Lambda^{1/2}V'$, e novamente aqui temos a obtenção do mesmo autovalor Λ. Este relacionamento entre as decomposições, o fato da obtermos o mesmo autovalor torna possível a equivalência entre as análises de Componentes Principais e Coordenadas Principais, resumindo temos que $Y_{n\times p} = U_{n\times n}\Lambda^{1/2}V'_{p\times p} \Longrightarrow YV = U\Lambda^{1/2}$, sendo $YV \longrightarrow$ Componentes Principais e $U\Lambda^{1/2} \longrightarrow$ Coordenadas Principais.

Esta equivalência é garantida para o caso métrico, com distâncias Euclidianas. Para dados originais se faz necessária a utilização de uma matriz H pré multiplicando Y, com o intuito de fornecer linhas centradas de Y. Portanto, para o caso de dados originais, a decomposição em valores singulares é expressa por $(HY)_{n\times p} = U_{n\times n}\Lambda^{1/2}V'_{p\times p}$.

(b) Como esse resultado pode ser usado em big-data? R: No caso do big-data, se pensarmos no big-p por exemplo, podemos utilizar esta equivalência para trabalhar com Coordenadas Principais por meio da decomposição espectral da matriz de distâncias $D_{n\times n}$, dado que no big-p a matriz de covariâncias se torna muito grande, o que pode acarretar problemas computacionais, dificuldade na leitura e visualização dos dados, instabilidade númerica para obtenção de autovalores e autovetores, fazendo com que a análise no espaço $n \times n$ seja uma melhor alternativa.

6. O Biplot é uma ferramenta útil na visualização de dados multivariados. Explique por que e como esse gráfico é construído.\ R: O gráfico biplot representa cada observação dos dados, sendo construído pela dispersão de pares de escores gerados por alguma análise de redução de dimensionalidade e pelo sentido das cargas que geram tais escores. O que proporciona uma representação gráfica simultânea dOs n escores e p componentes/fatores em R^2.

7. Como a Análise de Correspondência pode ser formulada a partir da Análise de Coordenadas Principais? R: Por meio da tabela de contigência utilizada na Análise de Correspondência, pode-se construir uma matriz de distâncias entre as linhas e entre as colunas, dadas as frequências relativas das linhas a seu total (matriz de distancia das linhas $r \times r$), do mesmo modo das frequências relativas das colunas comparados ao seu valor total (matriz de distancia das colunas $c \times c$). Com estas duas matrizes de distância podemos aplicar o mesmo procedimento da Análise de Coordenadas Principais, o que possibilita obter os escores das linhas e os escores das colunas, sendo que estes posteriormente poderão ser visulaizados num mesmo gráfico, por exemplo.

8. Na Análise de Fatores (Análise Fatorial), como estão definidos os fatores comuns e os específicos? Você pode usar a solução via Componentes Principais para responder. R: Com a Análise de Fatorial, em especial a Análise Fatorial Exploratória, podemos modelar as variáveis observadas Y em função de um conjunto de variáveis latentes F (não observáveis, constructos), esta modelagem pode ser expressar em sua notação matricial por $Y_i - \mu = \Phi_{p\times m}\mathbf{f}_{i\ m\times 1} + e_{i\ p\times 1}$, em que $\mathbf{f} = (F_1, ..., F_m)'$ são fatores comuns (variáveis latentes), $e = (e_1, ..., e_p)'$ são fatores específicos e $\Phi = (\phi_{ij})$ são as cargas fatoriais.

Com base nesta formulação, a matriz de covariâncias $\Sigma_{p\times p}$ é expressa por $\Sigma_{p\times p} = \Phi_{p\times m}\Phi'_{m\times p} + \Psi_{p\times p}$, em que $\Phi_{p\times m}$ e $\Psi_{p\times p}$ representam, respectivamente, a comunalidade e especificidade na variabilidade dos dados.

Podemos realizar uma solução para Análise Fatorial via Componentes Principais, em CP temos a matriz de covariâncias $\Sigma = V\Lambda V' = (V\Lambda^{1/2})(V\Lambda^{1/2})'$ e os escores dados por $Z_{n\times p} = YV$.

Portanto, ao observar este problema numa perspectiva em que buscamos predizer Y dado que conhecemos os escores Z, podemos escrever a equação como $Y_i = VZ_i \Longrightarrow Y_i = (V\Lambda^{1/2})(\Lambda^{-1/2}Z_i) + e_i$, em que $(V\Lambda^{1/2})$ equivalentes às cargas fatoriais e $(\Lambda^{-1/2}Z_i)$ os fatores comuns, sendo que estes escores fatoriais são os Componentes Principais padronizados.

9. Na análise de redução de dimensionalidade de uma matriz de dados $Y_{n\times p}$, os Componentes Principais satisfazem quais propriedades? O que garante que dois componentes reduzem bem os dados?

R: A principal propriedade é a maximização da variância total dos dados em componêntes não-correlacionados, dois componentes reduzem bem os dados quando a soma do percentual de variância explicada por ambos é superior a 80%.

10. Na análise de redução de dimensionalidade de uma matriz de dados $Y_{n\times p}$, os Eixos Discriminantes da Solução Linear de Fisher satisfazem quais propriedades? O que garante que dois eixos discriminantes reduzem bem os dados?

R: A principal propriedade é a maximização da variabilidade entre os grupos, simultaneamente, a minimização da variância dentro dos grupos.

%%% precisa dar uma conferida no material %%%% Calculados os autovalres da matriz $S_w^{-1} \times S_b$ temos que os primeiros autovalores proporcionam uma maior separação, discriminando bem os dados

11. Na análise de redução de dimensionalidade de uma matriz de dados $Y_{n\times p}$, com $p = p_1 + p_2$, os Eixos Canônicos da Correlação Canônica satisfazem quais propriedades? O que garante que o primeiro par desses eixos reduzem bem os dados? R: A principal propriedade é a máxima correlação entre os escores construídos / obtidos nos subconjuntos. O autovalor relacionado ao primeiro par tem por característica trazer uma maior explicação que os demais (com relação ao demais pares)

12. Considere a redução de dimensionalidade de uma matriz de dados $Y_{n\times p}$. Se $n << p$, quais são os problemas na realização da Análise de Componentes Principais "Clássica"? Que alternativas de análise podem ser usadas?

R: Como dificuldade inicial temos o tamanho da matriz de covariâncias $S_{p\times p}$, pois esta fica muito grande, assim pode-se ter problemas computacionais, instabilidade númerica para obtenção de autovalores e autovetores, dificuldade na leitura e visualização dos dados. Como alternativa podemos realizar a análise por Coordenadas Principais, uma vez que n é pequeno. Contudo, há situações em que a solução por CoP, pode não atender a algumas expectativas, como em soluções esparsas com a eliminação de algumas variáveis, neste caso soluções com penalização, como CP Penalizada são requeridas.

13. Na MANOVA qual é a importância da equação envolvendo o seguinte determinante: $|S_B - \lambda S_W| = 0$, em que S_B e S_W são matrizes quadradas (pxp), conhecidas, de soma de quadrados e produtos cruzados dos efeitos Entre e Dentro de grupos, respectivamente, e $\lambda\mathbb{R}^+$ é tal que $(S_W^{-1}S_B)V = \lambda V$, com $V \in \mathbb{R}^p$?

R: A equação define a decomposição espectral que retém a informação sobre a discriminação entre os grupos avaliados na MANOVA, onde a avaliação dos coeficientes dos autovetores, associados aos maiores autovalores (λ), define a importância de cada variável no efeito de tratamento.

14. Na análise de uma matriz de dados multivariados $Y_{n\times p}$, como a distância de Mahalanobis pode ser usada para definir Regiões de Concentração dos dados e Regiões de Confiança para o centroide? Como pode ser proposto um critério de diagnóstico de observações atípicas (outliers)?

R: A distância de Mahalanobis leva em consideração o fluxo de tendência, de correlação dos dados, o formato e direção da nuvem formada pelos dados. A distância de Mahalanobis mede a distância das observações ao centróide, padronizando estas pelas suas variâncias e covariâncias. Temos seu uso na construção da elipse de concentração dos dados. No BoxPlot Bivariado (Everitt, 2007), temos na elipse central a concentração de 50% dos dados. Similarmente, tendo como premissas observações independentes, dados aproximadamente normais, podemos utilizar o critério da distância de Mahalanobis para definir Regiões de Confiança, por meio de critérios de corte adotados associados à um valor de significância, que definirá, por exemplo, que a Região de Confiança para a verdadeira média, verdadeiro centróide dos dados na elipse agregam 90%, 95% dos dados (atrelada a um \$ = 5%, 10% \$, por exemplo), considerando que d_M^2 segue uma distribuição χ^2. Deste modo, este critério de corte c^2 associado à um valor de significância (α) na distribuição χ^2, indicará observações aberrantes, as quais representam pontos que ocorrem fora das caldas multivariadas.
Diagnóstico valor atípico multidimensional no R^p: $d_M^2 \overset{n\longrightarrow\infty}{\sim} \chi_p^2 \implies P(d_M^2 \leq c^2) \leq (1-\alpha)$

15. Na análise de uma matriz de dados multivariados \$Y_{n ×p} (,) \$ com $p = 2$, ilustre, em um gráfico de dispersão, possíveis diferenças entre os Intervalos de Confiança Univariados Clássicos, Intervalos de Confiança Univariados Simultâneos, Intervalos de Confiança Univariados com Correção de Bonferroni e Regiões de Confiança para o vetor μ. Compare essas quatro abordagens no contexto de correções para múltiplos testes.

R: Na figura abaixo temos a dispersão dos intervalos de confiança solicitados. É possível observar que os retângulso formado intervalos de confiança univariados com correção de bonferroni (ICBonf-;ICBonf+; ICSBonf-;ICSBonf+) flexibilizou os ICs quando comparado aos intervalos clássicos (IC-;IC+) e simultâneos. É importante notar também que estes ICs, à depender dos vetores X e Y avaliados, terão resultados distintos.

Figura 1: Intervalos de Confiança e Região de Confiança dos dados Simulados.

\

Figura 2: Intervalos de Confiança e Região de Confiança dos dados Simulados.

\

16. Em uma análise estatística, considere que na normalização de dados foi usada a seguinte transformação: $y_{ij} - v_{ij}z_{1i}$

em que, y_{ij} é a resposta do i-ésimo paciente na j-ésima variável, z_{1i} é o escore desse paciente no primeiro componente principal e v_{ij} é o j-ésima coordenada do correspondente autovetor.\

(a) Apresente situações práticas do por que adotar tal transformação dos dados em uma análise estatística.

R:

(b) Apresente críticas ao uso dessa estratégia de normalização de dados.

R:

(c) Em problemas gerais da análise de dados, cite alternativas que podem ser usadas para normalização de dados.

R:

17. Apresente uma situação prática para ilustrar a utilidade da Análise de Componentes Principais e de Coordenadas Principais.

R: Uma possível aplicação da análise da análise de componentes principais é na análise de desempenho de atletas no futebol, onde é investigado qual combinação de características explica melhor a variabilidade do desempenho dos jogadores.

Na análise de desempenho também é possível analisar as similaridades ou dissimilaridades entre scouts de atletas, buscando identificar jogadores com características mais próximas ou distintas dos demais.}

18. Apresente uma situação prática para ilustrar a utilidade da Análise de Correspondência.

R: Uma possível aplicação da análise da análise de correspondência é na avaliação de preferências de clientes que visitam uma concessionária por tipos de carros relacionando o modelo(SUV, Sedan, Hatch e etc...), com o tipo de combustível(Elétrico, Disel, Gasolina e etc...), onde o intuito é escolher os melhores carros para testes drive.

19. Apresente uma situação prática para ilustrar a utilidade da Análise Fatorial (fatores Comuns e específicos).

R: Uma possível aplicação da análise da análise fatorial é na avaliação de pacientes de um consultório psicológico, onde através de um questionário com questões relacionadas ao sono, autoestima, tristeza, alimentação; com o intuito de avaliar traços latentes relacionado à depressão, ansiedade e estresse.

20. Apresente uma situação prática para ilustrar a utilidade da Análise Discriminante e da MANOVA.

R: Uma possível aplicação da análise da análise de discriminante seria a classificação de perfis de clientes em uma financeira em bons e maus pagadores de cartão de crédito, com o objetivo de prever se um novo cliente é um bom pagador em potencial, realizada a partir de informações prévias de outros clientes. Uma possível aplicação da análise da MANOVA seria na avaliação das notas do Exame Nacional do Ensino Médio - ENEM de três escolas nas quatro área de conhecimentos: ciências humanas, ciências da natureza, linguagem e matemática, com o intuito de verificar se há alguma escola com desempenho difente nas áreas avaliadas.

21. Apresente uma situação prática para ilustrar a utilidade da Análise de Correlação Canônica.
R: Uma possível aplicação da análise de correlação canônica seria entre as notas de provas de disciplinas de probabilidade e inferência estatística dos alunos de graduação e mestrado do IME-USP, onde a ideia é quantificar a associação entre estas duas etapas de formação acadêmica, resumindo essa informação em pares de variáveis canônicas.

22. Com o apoio da Aula em que geramos dados e visualizamos direções dos vetores reducionistas de dados multivariados, compare as soluções de Componentes Principais, Análise Discriminante, Correlação Canônica e Regressão Linear.

R: Ao comparar os exemplos 1 a 6 podemos ver que para o mesmo conjunto de dados obtemos diferentes soluções de acordo com cada uma das técnicas aplicadas. No exemplo 1 podemos perceber que a direção do eixo de AD e CP são moderadamente próximos, contudo não coincidentes, pois cada um deles busca atender à objetivos e propriedades diferentes, conforme suas particularidades, neste sentido a solução via Regressão Linear apresenta um eixo ainda mais próximo do eixo de CP.

Figura 3: Exemplos

23. Qual é o problema analítico e possíveis soluções da análise de dados em Big-n?

R: Muitas análises se utilizam do espaço $n \times n$ e tais análises, frequentemente, sofrem problemas computacionais, como: alocação de memória e alto tempo computacional em leitura e armazenamento. As principais soluções se baseiam na divisão dos dados em conjuntos menores, como Split-Apply-Combine ou Divide-Recombine, que após a divisão calcula a estatística de interesse e depois gera uma medida agregada dessas sub-amostras. Há também o problema de visualização dos dados, onde as disperções geram borrões, como solução recomenda-se a construção de do gráfico em coordenadas paralelas, ou até mesmo a representação somentes dos dados mais frequentes.