Análise de Correlação Canônica

Caroline Vasconcelos

Glicose Sanguínea

Marcas de Saquê

A análise de correlação canônica busca identificar e quantificar as associações entre dois conjuntos de variáveis, é útil quando se deseja entender as relações complexas entre múltiplas variáveis independentes e dependentes simultaneamente. As funções canônicas resultantes da análise podem ser interpretadas como padrões específicos de associação entre os conjuntos de variáveis. Permite também avaliar a validade e generalização das relações identificadas para novos grupos de dados.

A técnica se concentra na correlação entre uma combinação linear de variáveis em um conjunto e uma combinação linear de variáveis em outro conjunto, a ideia é primeiro determinar o par de combinações lineares que possui a maior correlação. Depois, determinar o par de combinações lineares que possui a maior correlação entre todos os pares não correlacionados com o par inicialmente selecionado, e assim por diante. Os pares de combinações lineares são chamados de variáveis canônicas e suas correlações são chamadas de correlações canônicas, elas medem a força da associação entre os dois conjuntos de variáveis. O aspecto de maximização da técnica representa uma tentativa de concentrar uma relação de alta dimensão entre dois conjuntos de variáveis em alguns pares de variáveis canônicas.

- Glicose Sanguínea

O conjunto de dados contém medições dos níveis de glicose no sangue em três ocasiõesem 50 mulheres. Os y's representam medições de glicose em jejum em três ocasiões e os x's são medições de glicose 1 hora após a ingestão de açúcar. A figura 1 mostra os gráficos de dispersão bivariadas, densidade e as correlações entre as variáveis. Importante notar que todas as correlações são baixas, nenhuma é superior a 0,5. Os gráficos de dispersão bivariadas mostram vários pontos fora da nuvem de pontos em cada um dos gráficos. Usando o pacote MVN, o p-valor igual a 0,9e-4 do teste de Royston, nos dá evidências de que este conjunto de dados não apresenta normalidade multivariada. No teste univariado de Anderson-Darling 4 variáveis apresentaram normalidade, foram elas: y_1, x_1, x_2, x_3.

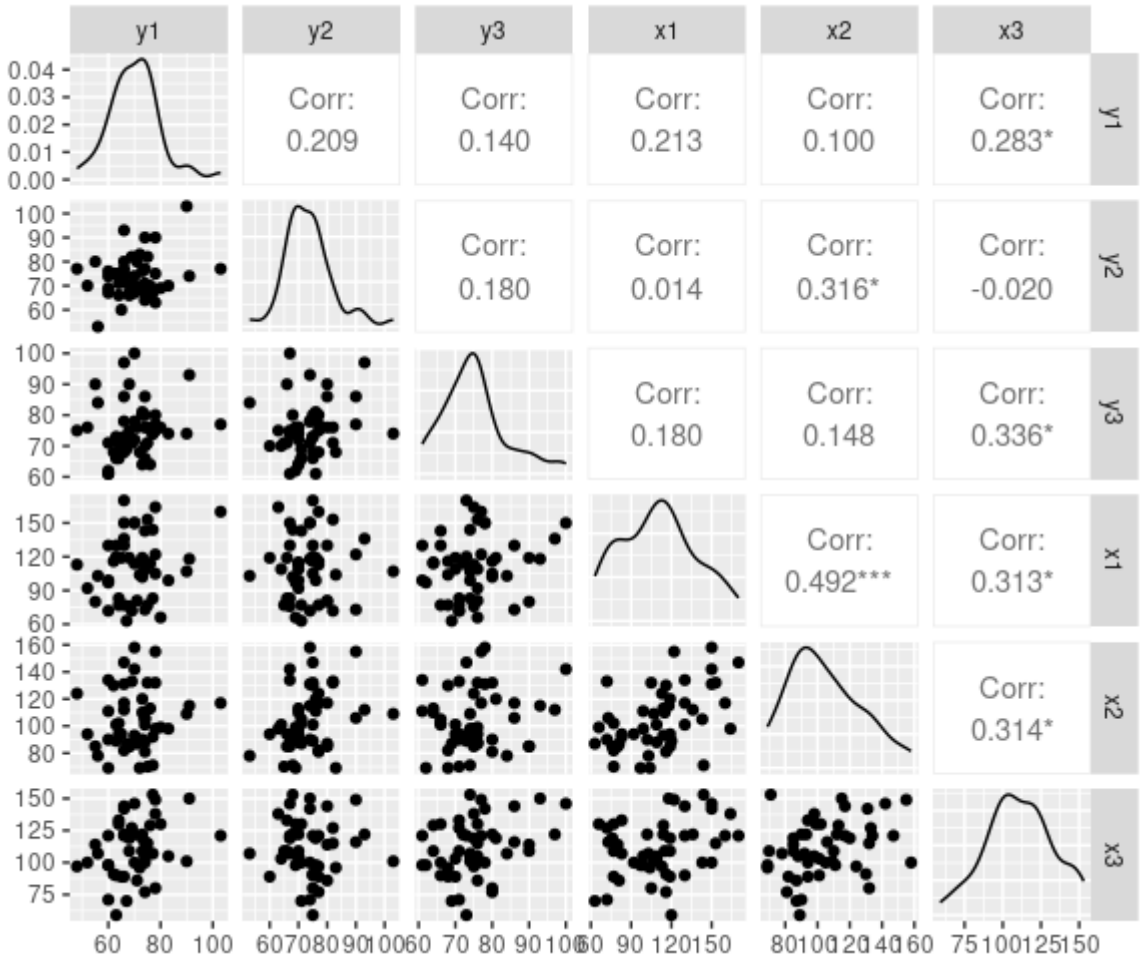

Figura 1: Correlações bivariadas das variáveis glicêmicas.

Para obter a correlação canônica entre as variáveis y_1, y_2 e y_3, que representam as quantidades glicêmicas das 50 pacientes em jejum em momentos diferentes assim como também três outros momentos, x_1, x_2 e x_3, após ingestão de açúcar, utilizou-se o pacote candisc. A tabela 1 mostra as variâncias explicadas e acumuladas das três variáveis canônicas que representam os pares de variáveis x_1 e y_1, x_2 e y_2, e x_3 e y_3. O par canônico CR1 retém 70,94% da variância explicada, juntamente com o par canônico CR2, eles representam 99,36% da variância acumulada.

Table 1: Variâncias explicadas e acumuladas.

	CR1	CR2	CR3
Variância explicada	70,94%	28,41%	0,63%
Variância acumulada	70,94%	99,36%	100,00%

A tabela 2 mostra as correlações entre os pares canônicos, o primeiro par representa a combinação linear entre x_1 e y_1, e assim por diante. Novamente vemos que todas as correlações entre as variáveis canônicas e os pares x e y são baixas e ficaram abaixo de 0,5, no entanto, foram todas positivas. A correlação do primeiro par é a maior correlação entre as combinações lineares entre os três pares.

Table 2: Correlações entre as variáveis canônicas e as variáveis originais.

Variáveis	ρ
CR1	0,493
CR2	0,338
CR3	0,053

Os coeficientes nas variáveis canônicas refletem as diferenças de dimensão entre as variáveis, bem como diferenças na contribuição das variáveis para a correlação canônica. Para remover o efeito de dimensão, eles podem ser padronizados multiplicando pelos desvios padrão das variáveis correspondentes. Os coeficientes padronizados mostram a contribuição das variáveis na presença uma da outra. Assim, se algumas das variáveis forem excluídas e outras adicionadas, os coeficientes mudarão. As variáveis que mais contribuem para a correlação entre y e x são y_2, y_3, x_1, x_2 e x_3.

Table 3: Coeficientes canônicos padronizados de x.

Variáveis originais	Xcan1	Xcan2	Xcan3
y_1	-6.2607	-1.6150	-7.8184
y_2	5.8115	-6.7081	-1.0777
y_3	-5.2936	-3.7185	6.2417

Table 4: Coeficientes canônicos padronizados de y.

Variáveis originais	Ycan1	Ycan2	Ycan3
x_1	-14.7643	5.7129	-28.5758
x_2	14.5473	-22.0427	1.2759
x_3	-18.4587	-5.9124	13.5368

Existem vários testes para avaliar a significância das correlações canônicas, ou seja, testa-se se as n primeiras correlações canônicas são as correlações significativas, e portanto, as variáveis canônicas correspondentes seriam as mais importantes para a caracterização da informação (Mingoti, 2007). O teste de significância de Wilks (1935) testa a independência entre grupos de

variáveis. Existem outros testes de significância, como o apresentado por Morrisson (1976), que afirma que a distribuição do maior autovalor segue a distribuição da maior raiz característica de Roy, esse teste foi generalizado por Wilks (Ferreira, 2012). Conforme mostrado na tabela 5, não rejeitamos a hipótese de que a correlação do primeiro par canônico seja igual a 0, e para os demais pares, o contrário. Pelo teste de Roy (tabela 6), ao nível de significância de 5%, temos que o primeiro par canônico é o mais significativo e o que mais retém informação.

Table 5: Teste de significância de Wilks de cada correlação canônica.

Variáveis canônicas	Aproximação F	p- valor
CR1	2,15	0.0311
CR2	1,44	0.2261
CR3	0,13	0.7167

Table 6: Teste de significância de Roy.

Variável canônica	Aproximação F	p-valor
CR1	4,94	0,0046

- **Marcas de saquê**

Trinta marcas de saquê foram avaliadas com relação a gosto, odor, pH, acidez 1, acidez 2, densidade, açúcar redutor, quantidade total de açúcar, quantidade de álcool e nitrogênio, nesta ordem respectivamente. Na figura 2, observam-se, em geral, correlações pequenas, sendo as correlações mais altas entre as variáveis y_1 e y_2, gosto e odor, x_2 e x_3 acidez 1 e acidez 2, x_3 e x_8, acidez 2 e nitrogênio, x_5 e x_6, açũcar redutor e quantidade total de açúcar. Os gráficos de dispersão não nos dão evidências de normalide, muito menos os de densidade. Usando o pacote **MVN** e o teste de Royston, o p-valor igual a 3,07e-11 nos dá evidências de que os dados não apresentam normalidade multivariada. No teste univariado de Anderson-Darling, 6 variáveis apresentaram evidências de normalidade (y_1, y_2, x_2, x_4, x_5 e x_6).

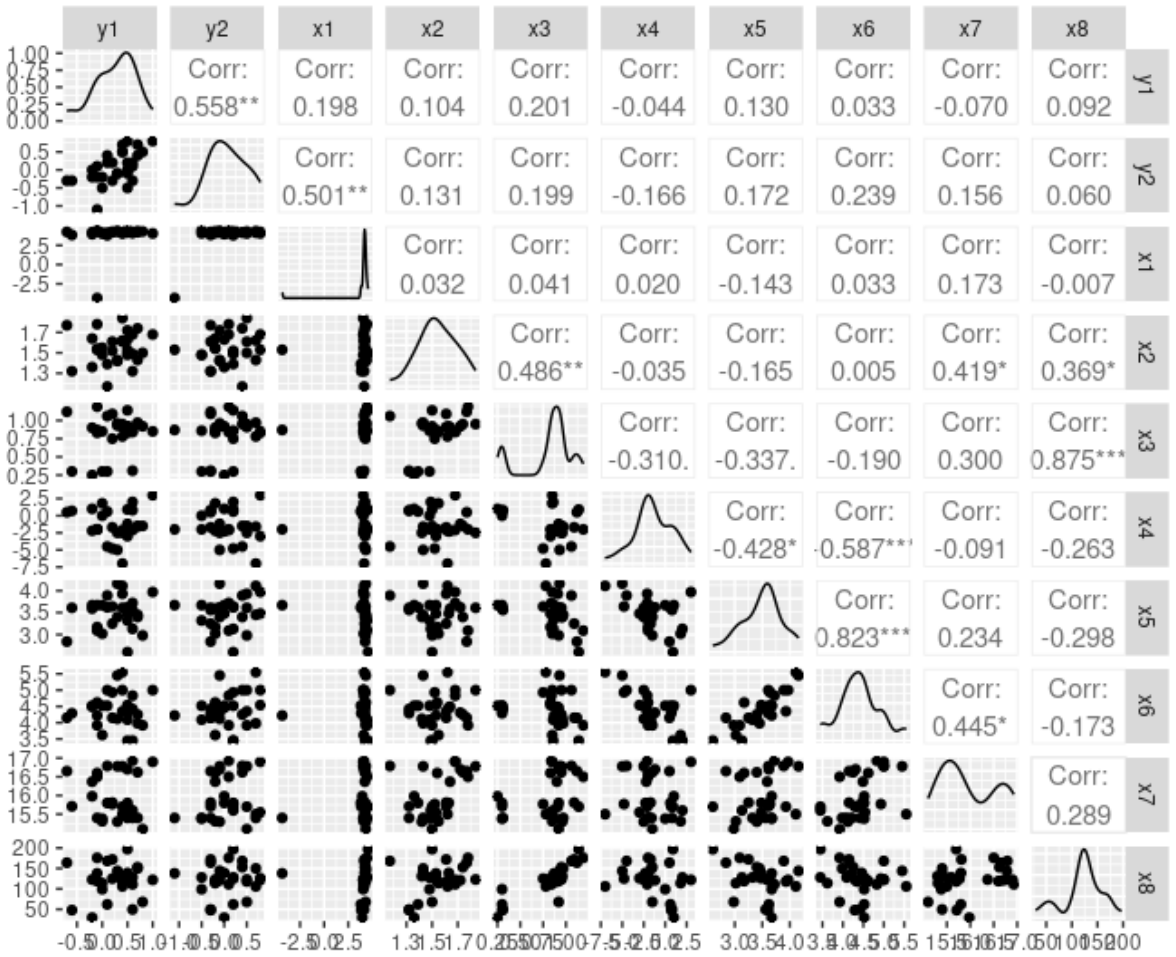

Figura 2: Correlações, gráficos de dispersão e densidade do banco de dados marcas de saquê.

A análise de correlação canônica foi feita utilizando os pacotes **candisc**, **CCA** e **CCP**. Para a análise, as variáveis foram separadas em 2 grupos, grupo 1: y_1 e y_2 (gosto e odor), grupo 2: x_1, x_2, x_3, x_4, x_5, x_6, x_7 e x_8 (pH, acidez 1 e acidez 2, densidade, açúcar redutor, açúcar total, quantidade de álcool e nitrogênio). A tabela 7 mostra as variâncias explicadas e acumuladas dos pares canônicos, observa-se que todos o primeiro pares retém a maior variância explicada, portanto, a maior quantidade de informações. Na tabela 8, temos as correlações entre os pares canônicos e as variáveis originais, todas foram positivas, e o primeiro par apresentou correlação ligeiramente maior em relação ao segundo.

Table 7: Variâncias explicadas e variâncias acumuladas.

Pares canônicos	Variância Explicada	Variância Acumulada
CR1	82,67	82,67
CR2	17,33	100,00

Pares canônicos	Variância Explicada	Variância Acumulada

Table 8: Correlações entre os pares canônicos e as variáveis originais.

Pares canônicos	ρ
CR1	0,70
CR2	0,40

As tabelas 9 e 10 apresentam os coeficientes canônicos padronizados. Como visto anteriormente, os coeficientes canônicos padronizados mostram a contribuição das variáveis na presença das outras variáveis. As variáveis que mais contribuiram para a correlação entre y e x foram: y_2, x_1, x_7 e x_8, referentes a odor, pH, quantidade de álcool e nitrogênio.

Table 9: Coeficientes canônicos padronizados de x.

Variáveis originais	Xcan1	Xcan2
y_1	-0.1012	-0.4687
y_2	-0.3664	0.3814

Table 10: Coeficientes canônicos padronizados de y.

Variáveis originais	Ycan1	Ycan2
x_1	-1.2156	0.3672
x_2	-0.0095	-0.0294
x_3	-0.3585	-0.1917
x_4	-1.3676	-1.4522
x_5	-0.2330	-0.3833
x_6	-0.2804	0.1829
x_7	0.3598	0.5017
x_8	19.0102	-3.6605

Com relação aos testes de significância, como o teste de Wilks (1935) testa a independência entre grupos de variáveis, o que observamos na tabela 11 é que existem evidências para que se rejeite a hipótese de que existe independência entre as variáveis canônicas. O teste de Roy, na tabela 12, mostra que o primeiro par canônico é o mais representativo.

Table 11: Teste de significância de Wilks de cada correlação canônica.

Variáveis canônicas	Aproximação F	p- valor
CR1	1,17	0.3281
CR2	0,80	0.5296

Table 12: Teste de significância de Roy.

Variável canônica	Aproximação F	p-valor
CR1	2,79	0,0601

Referências

FERREIRA, D. F. Estatística Multivariada. 1a. edição. Lavras: Editora UFLA, 2008.

JOHNSON, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2007.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.