Relatório 1 : Análise Multivariada 2

Tharso Monteiro

3 de novembro de 2021

1 Resumo

Este relatório visa explicar algumas técnicas de análise multivariada e aplicá-las a uma imagem de satélite. A imagem foi obtida através do site EarthExplorer (`https://earthexplorer.usgs.gov/`), de propriedade da USGS (*United States Geological Survey*) e advém dos satélites *landsat*. Estes satélites, lançados através de uma parceria entre a USGS e a NASA, capturam imagens da superfície terrestre desde 1975.

As técnicas de análise multivariada a serem estudadas são a análise de componentes principais e a análise fatorial. A imagem de satélite utilizada como exemplo está exibida na Figura 1 e representa uma vista espacial de Pyongyang, capital da República Popular Democrática da Coreia, no dia 13/09/2021.

Todos os códigos utilizados nesse projeto foram feitos na linguagem R.

Figura 1: Imagem de satélite de Pyongyang, capital da RPDC.

2 Análise de Componentes Principais

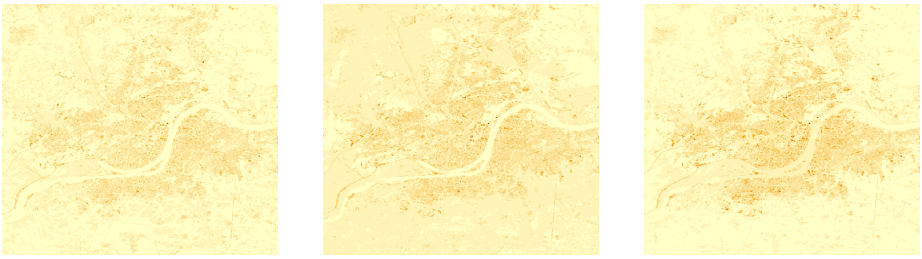
A presença de muitas variáveis num conjunto de dados implica na inviabilidade da aplicação de técnicas estatísticas convencionais. Uma das técnicas mais utilizadas na análise multivariada é a Análise de Componentes Principais (ACP), uma técnica de redução de dimensionalidade que consiste em transformar as variáveis de um conjunto de dados em novas variáveis, chamadas componentes principais.

Os componentes principais são combinações lineares não-correlacionadas das variáveis originais do conjunto de dados e ordenadas de maneira que os primeiros componentes representam a maior variação das variáveis originais, isto é, contêm mais informação que os subsequentes.

O objetivo da ACP é a criação de um pequeno conjunto de variáveis que substituam o conjunto original de dados enquanto retêm a maior quantidade de informação possível (Everitt and Hothorn, 2011). No contexto de imagens, esta técnica é especialmente útil para sua compressão. Caso a ACP seja bem sucedida, é possível representar uma imagem utilizando uma quantidade bem menor de memória após o processamento da figura original.

2.1 Aplicação

Utilizando a linguagem R, importamos a imagem exibida na Figura 1 e separamos-na em suas bandas RGB. Essas decomposições estão ilustradas na Figura 2.



(a) Decomposição da cor vermelha

(b) Decomposição da cor verde

(c) Decomposição da cor azul

Figura 2: Decomposição RGB da Figura 1

Cada banda é, então, representada por uma matriz $X_{m\times n}$ em que o elemento a_{ij} da matriz indica a intensidade da cor primária neste determinado pixel da imagem original.

Performamos a ACP em cada uma dessas matrizes e, em seguida, reconstruímos cada banda da imagem com base em p componentes principais e voltamos a juntá-las para gerar uma aproximação da imagem original.

A Figura 3 ilustra a recomposição da imagem original com $p = 5, 20, 40$.

Observa-se que com 40 componentes, é possível obter uma aproximação fiel, enquanto realiza-se uma redução de dimensionalidade significante.

(a) Recomposição da imagem original com 5 componentes principais

(b) Recomposição da imagem original com 20 componentes principais

(c) Recomposição da imagem original com 40 componentes principais

Figura 3: Recomposição da Figura 1 através de componentes principais

3 Análise Fatorial

A análise fatorial busca explicar a variância em um conjunto de variáveis mensuráveis com base em variáveis latentes, isto é, conceitos que não podem ser medidos de maneira direta, mas que por suposição, são relacionados às variáveis mensuráveis (Everitt and Hothorn, 2011).

O modelo da análise fatorial é baseado na regressão múltipla. Entretanto, diferentemente da regressão múltipla, as variáveis de interesse são modeladas com base em variáveis latentes. Por isso, a estimação direta dos coeficientes associados às mesmas não é possível. Entretanto, existem métodos de estimação indireta e de ajuste dos valores desses parâmetros de maneira que a interpretação dos fatores correspondentes seja mais intuitiva.

3.1 Aplicação

Utilizamos a mesma imagem exibida na Figura 1. Sejam $x_{red}, x_{green} e x_{blue}$ as variáveis que correspondem ao valor de um determinado pixel na decomposição RGB.

Assumimos o modelo:

$$\begin{aligned} x_{red} &= \lambda_1 f + u_1 \\ x_{green} &= \lambda_2 f + u_2 \\ x_{blue} &= \lambda_3 f + u_3 \end{aligned}$$

Em que f representa o fator de interesse e u_i representa um termo de perturbação, que engloba todas as fontes de variação em x_i que não advêm de f.

Os coeficientes $\lambda_1, \lambda_2, \lambda_3$ são chamados *factor loadings* e representam o quanto cada variável observável é afetada pelo fator f.

Utilizando a função **factanal** da linguagem R, que utiliza o método de máxima verossimilhança (Everitt and Hothorn, 2011) e rotação Varimax, obtivemos os seguintes resultados:

$$\hat{\lambda_1} = 0.978$$
$$\hat{\lambda_2} = 0.986$$
$$\hat{\lambda_3} = 0.936$$

$$\hat{\sigma^2}_{u_1} = 0.044$$
$$\hat{\sigma^2}_{u_2} = 0.028$$
$$\hat{\sigma^2}_{u_3} = 0.124$$

Este modelo explica 93.5% da variância nas variáveis de interesse.

Podemos interpretar o fator f como a captação de luz branca pelo satélite em um determinado pixel. Essa captação afeta a intensidade de cada banda de maneira quase homogênea, representado pelos valores similares de λ_1, λ_2 e λ_3.

Referências

Everitt, B. and Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.