Métodos de Estimação de Matrizes de Covariância usando MATLAB®

Kenneth Brenner dos Anjos Benício - 519189

Universidade Federal do Ceará - Mestrado em Engenharia de Teleinformática

Resumo

Nesse trabalho são analisadas diferentes técnicas de obtenção das matrizes de covariância de um conjunto de dados sobre medições atmosféricas da ionosfera. Os métodos de estimação sugeridos para as matrizes de covariância são comparados diretamente com o método nativo do MATLAB®. Nessa comparação é analisado o tempo de processamento e a precisão de cada método, utilizando o conceito da norma de matriz de diferença. Por fim, é analisado a invertibilidade das matrizes de covariância global e covariância locais por meio da análise do valor de rank e do número de condicionamento das matrizes envolvidas.

1 Introdução

É abordado em detalhes nessa seção os três estimadores para as matrizes de covariância que serão utilizados nas simulações do trabalho. Além disso, é brevemente discutido os conceitos de correlação e de covariância.

1.1 Conjunto de Dados

O conjunto de dados sugerido para o exercício computacional representa uma série de medições atmosféricas sobre parâmetros da ionosfera realizado por um radar composto por um conjunto de 16 antenas [1]. O conjunto é composto por 351 observações de tal forma que cada observação é descrita por um vetor contendo 35 atributos. 0s primeiros 34 atributos desse conjunto de dados são formadas por elementos numéricos contínuos, enquanto o último atributo representa um elemento descritivo da classe ao qual aquele vetor de observações pertence. O último atributo possui apenas dois valores descritivos, "*good*"ou "*bad*", definindo assim o conjunto de dados como composto por apenas duas classes diferentes. A primeira classe, "*good*", indica quando o sistema de antenas consegue identificar a presença de certos componentes atmosféricos naquela região e a segunda classe, "*bad*", indica quando o oposto ocorre.

1.2 Matriz de Correlação

Antes de prosseguir é importante abordar brevemente o conceito de matriz de correlação. A correlação é um conceito estatístico que dita a forma como duas variáveis podem estar interrelacionadas. Nesse trabalho estamos interessados no aspecto quantitativo da correlação, então utilizaremos o conceito de coeficiente de correlação que irá nós indicar as relações presentes entre os diversos atributos do conjunto de dados. O coeficiente de correlação entre variáveis aleatórias $\boldsymbol{X_1}$ e $\boldsymbol{X_2}$ pode ser obtido pela expressão

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}, \tag{1}$$

onde σ_{12}, σ_1 e σ_2 representam a covariância conjunta entre $\boldsymbol{X_1}$ e $\boldsymbol{X_2}$, o desvio padrão de $\boldsymbol{X_1}$ e o desvio padrão de $\boldsymbol{X_2}$, respectivamente. Já a matriz de correlação representa todas as possíveis combinações entre os atributos disponíveis e, para o cada onde existem apenas 2 atributos, é possível escrevê-la como

$$\boldsymbol{R_x} = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix}. \tag{2}$$

Nesse trabalho, propõe-se obter as estimações da matriz de correlação por meio de três diferentes métodos. No primeiro, a formulação do estimador é feita utilizando apenas os vetores de informação e devido a isso esse método será chamado de **não matricial**. No segundo, o estimador é formulado considerando a estrutura matricial dos dados e por isso será chamado de método **matricial**. Por fim, o terceiro método utiliza um estimador recursivo para a matriz de correlação, onde a matriz é estimada a medida que novas observações são fornecidas para o sistema, então o método será denominado de **recursivo**. Abaixo são apresentadas as equações para os estimadores na ordem em que foram citados acima

$$\hat{\boldsymbol{R_x}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n), \tag{3}$$

$$\hat{\boldsymbol{R_x}} = \frac{1}{N} \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}, \tag{4}$$

$$\hat{\boldsymbol{R_x}}(n) = \left(\frac{n-1}{n}\right) \hat{\boldsymbol{R_x}}(n-1) + \frac{1}{n}\boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n). \tag{5}$$

1.3 Matriz de Covariância

Por fim, temos ainda a matriz de covariância. Assim como para a matriz de correlação será interessante entender a estrutura da matriz de covariância para o caso mais básico onde existem apenas dois atributos envolvidos. Dessa forma, tem-se

$$\boldsymbol{C_x} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}, \tag{6}$$

onde foi possível reescrever a primeira matriz graças a relação existente entre o coeficiente de covariância, σ_{12}, e o coeficiente de correlação, ρ_{12}. É ainda possível estimar a matriz de covariância diretamente com as seguintes expressões

$$\boldsymbol{m} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}(n), \tag{7}$$

$$\hat{\boldsymbol{C}_{\boldsymbol{x}}} = \frac{1}{N}\sum_{n=1}^{N}[\boldsymbol{x}(n)-\boldsymbol{m}][\boldsymbol{x}(n)-\boldsymbol{m}]^{\mathrm{T}}, \tag{8}$$

Como citado inicialmente existe uma relação entre o coeficiente de covariância e o coeficiente de correlação. Dessa forma, não é absurdo ter a expectativa de que possa também existir uma relação entre as matrizes de correlação e de covariância. Essa relação existe e pode ser facilmente obtida do seguinte modo

$$\begin{aligned}
&\boldsymbol{C}_{\boldsymbol{x}} = \mathbb{E}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\},\\
&\boldsymbol{C}_{\boldsymbol{x}} = \mathbb{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\},\\
&\boldsymbol{C}_{\boldsymbol{x}} =, \mathbb{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\} - \mathbb{E}\{\boldsymbol{x}\}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\mathbb{E}\{\boldsymbol{x}\}^{\mathrm{T}} + \mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\},\\
&\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{R}_{\boldsymbol{x}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}},\\
&\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{R}_{\boldsymbol{x}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}.
\end{aligned} \tag{9}$$

Por fim, ao substituirmos as Equações (3), (4) e (5) na Equação (9) obtemos os três estimadores que serão utilizados para obter as matrizes de covariância

$$\hat{\boldsymbol{C}_{\boldsymbol{x}}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n) - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}, \tag{10}$$

$$\hat{\boldsymbol{C}_{\boldsymbol{x}}} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}, \tag{11}$$

$$\hat{\boldsymbol{C}_{\boldsymbol{x}}}(n) = \hat{\boldsymbol{R}_{\boldsymbol{x}}}(n) - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}. \tag{12}$$

onde $\hat{\boldsymbol{R}_{\boldsymbol{x}}}(n)$ é dado pela Equação (5).

1.4 Software

Todos os códigos foram desenvolvidos utilizando-se o MATLAB® 2021a. Foi criada uma classe de funções com métodos correspondentes aos classificadores aqui abordados. Todos os códigos devidamente comentados foram enviados conjuntamente com esse relatório. Ademais, os resultados foram gerados por um computador com processador Intel i7-10700K (3.8GHz), 16GB de memória RAM e uma placa gráfica RX 6600.

2 Resultados

Os resultados apresentador a seguir foram obtidos em um computador com processador Intel i7-10700K (3.8GHz) e 16GB de memória RAM. Antes de tudo é necessário definir a métrica que será utilizada para analisar o erro na obtenção das matrizes de covariância entre os métodos desenvolvidos e o método nativo do MATLAB® utilizando a função COV. Assim como sugerido será utilizada a norma da matriz de diferenças que pode ser obtida diretamente a partir da expressão abaixo

$$\boldsymbol{E} = \boldsymbol{C}_{\mathrm{Proposto}} - \boldsymbol{C}_{\mathrm{Nativo}}. \tag{13}$$

Por fim, será utilizado o comando nativo NORM do MATLAB® para obter a norma da matriz de diferenças $\boldsymbol{E}$. Após essas considerações, foi possível chegar ao seguinte resultado preliminar

$$||\boldsymbol{E}_{\text{não matricial}}|| = 0.008745, \tag{14}$$

$$||\boldsymbol{E}_{\text{matricial}}|| = 0.008745, \tag{15}$$

$$||\boldsymbol{E}_{\text{recursivo}}|| = 0.008745. \tag{16}$$

Embora seja possível constatar um certo erro residual é algo que possivelmente poderá ser irrelevante a depender da aplicação ao qual esse conjunto de dados possa vir a ser destinado.

Em seguida, foi pedido para realizar um breve estudo quanto ao tempo de processamento dos métodos desenvolvidos e do método nativo do MATLAB®. Foi adotado um procedimento simples de Monte Carlo para a simulação para que assim seja possível visualizar o comportamento médio do tempo de processamento. Dessa forma, foram realizados 10 mil experimentos independentes entre si e o tempo de realização para cada experimento e cada caso foram coletados e utilizados ao final para auxiliar na obtenção de médias temporais para cada método. Sendo assim, é possível chegar à conclusão de que os estimadores descritos pelas equações (3) e (5) despendem tempo de processamento na ordem de $\mathbf{10^{-4}}$ **segundos**. Enquanto isso, o estimador descrito pela equação (4) e o nativo despendem tempo de processamento de uma ordem de grandeza menor, ou seja, $\mathbf{10^{-5}}$ **segundos**. Portanto, é interessante ver que é possível ganhar uma ordem de grandeza em termos de processamento apenas ao se utilizar o estimador da matriz de covariância em seu formato matricial.

Já com relação a análise do rank matricial das matrizes de covariância é necessário antes de tudo separar o conjunto de dados nas duas classes que o compõe. A classe *good* possui 225 observações enquanto a classe *bad* possui 126 observações. Após essa separação foi utilizado o estimador definido pela Equação (4) para a obtenção das matrizes de covariância global e locais. Após obtidas, o posto matricial das matrizes de covariância global e locais foi computado com o auxílio da função nativa RANK e os seguintes valores foram encontrados

$$rank(\boldsymbol{C}_{good}) = 32, \tag{17}$$

$$rank(\boldsymbol{C}_{bad}) = 33, \tag{18}$$

$$rank(\boldsymbol{C}_{\text{global}}) = 33. \tag{19}$$

Dessa forma, todas as matrizes de covariância envolvidas, seja global ou seja local, são de posto matricial deficiente e dessa forma é impossível realizar a operação de inversão matricial. Ademais, como já seria esperado, a função RCOND nativa retorna valor de condicionamento nulo para as matrizes de covariância obtidas indicando que de fato as matrizes são singulares.

3 Conclusão

Portanto, vimos que os estimadores sugeridos apresentam um bom desempenho comparados mesmo ao algoritmo de obtenção da matriz de covariância do próprio MATLAB®. Quando a analise é feita considerando a norma da matriz de diferenças percebe-se a presença de um erro residual em todos os métodos, mas possivelmente dentro de um intervalo aceitável. Para eliminar completamente essa questão seria necessário estudar o impacto desse erro residual em alguma aplicação desse conjunto de dados para alguma tarefa de classificação, mas foge momentaneamente ao escopo desse trabalho. Já com relação ao tempo de processamento, os métodos descritos pelas equações (3) e (5) apresentam uma ordem de grandeza em lentidão quando comparado ao método nativo e ao método descrito pela equação (4).

Por fim, com relação a invertibilidade das matrizes de covariância vimos que tanto a matriz global quanto as locais possuem posto matricial deficiente e dessa forma é impossível realizar a operação de inversão da forma que elas estão. Desse modo, seria necessário realizar alguma operação para que a invertibilidade da matriz seja possível visto que diversos métodos de classificação são baseados na inversão da matriz de covariância.

Referências

[1] Vince Sigillito. *Ionosphere Data Set*, Disponível em: <`http://archive.ics.uci.edu/ml/datasets/Ionosphere`>