Fundamentos da Estimação Paramétrica

GABRIELA W GABRIEL

IEE-S / ITA – Sala 195 – Ramal 5991

ggabriel@ita.br / gabriela.gabriel@gp.ita.br

São José dos Campos, 11 de Maio de 2021

Conteúdo

Definição do Problema

- Técnicas de estimação de **parâmetros ESTÁTICOS**, posteriormente estendido para estimação de **parâmetros DINÂMICOS** no tempo.

- Vamos considerar a forma geral na forma de **Regressão Pseudolinear**

$$y[k] = \varphi(k,\theta)'\theta + w[k]$$

em que, θ é o vetor de parâmetros desconhecidos
φ é o vetor dos regressores referente a k medidas conhecidas
$w[k]$ é um ruído que corrompe a leitura de $\varphi'\theta$

Definição do Problema

- Consideremos um conjunto de k medidas $y[j], u[j],\ j = 1, 2, \cdots, k$ do parâmetro θ na presença do ruído de medida $w[\cdot]$, representado por $\varphi^k \triangleq \{y[j], u[j]\}_{j=1}^{k}$ (**conhecido**)

- Nosso objetivo é

  Encontrar uma função destas k observações do parâmetro θ do tipo

  $$\hat{\theta}_k = \hat{\theta}(k, \varphi^k)$$

  e que **ESTIME** o valor de θ segundo algum critério preestabelecido.

- Para isso, definiremos o **erro da estimação** realizada como

$$\tilde{\theta} \triangleq \theta - \hat{\theta}$$

Definição do Problema

- Os **MÉTODOS DE ESTIMAÇÃO PARAMÉTRICA** corresponde ao mapeamento $\varphi^k \to \hat{\theta}$

- Como escolhemos o **melhor** estimador? Dentre **quais** métodos?

- Para isso, precisamos entender duas possíveis abordagens:
  - θ é **determinístico**. Há UM ÚNICO valor verdadeiro
  - θ é uma **variável aleatória**. Sua p.d.f é CONHECIDA E CONSTANTE durante as medições

Parâmetros Determinísticos

- Conhecida como abordagem não bayesiana (abordagem de Fisher)
- Nesta abordagem definimos uma **função de verossimilhança** $\Lambda_y(\theta)$.

DEFINIÇÃO V.1

A **função de verossimilhança** *corresponde a p.d.f das medidas de* θ *condicionadas a* θ*.*

$$\Lambda_y(\theta) \triangleq f_{y|\theta}(\beta|\alpha)$$

- Observe que, neste caso, não há conhecimento prévio da estatística do parâmetro

Parâmetro Estocástico

- Conhecida como abordagem bayesiana

- Temos **conhecimento a priori** de $f_\theta(\alpha)$, que permanece o mesmo durante todo o processo de medição.

- Dado um conjunto de k (fixo) medidas com $f_y(\beta) = c$, seguindo a fórmula de Bayes, dado $(f_{y|\theta}(\beta|\alpha))$, determinamos a **densidade a posteriori** conforme

$$f_{\theta|y}(\alpha|\beta) = \frac{f_{y|\theta}(\beta|\alpha) f_\theta(\alpha)}{c}$$

c também é dita constante de normalização.

Métodos de Estimação

Escolhida uma **FAMÍLIA DE MODELOS**

$\Downarrow$

Objetivo:

Encontrar uma estimativa coerente para o vetor θ.

- Diferentes modelos geram diferentes erros de predição

$\Downarrow$

- Escolha do modelo baseada no **ERRO DE PREDIÇÃO**

$$\varepsilon(k, x) = y[k] - \hat{y}(k, x)$$

segundo algum critério estocástico, dado $\varphi^k = \begin{bmatrix} y[1] & \cdots & y[k] & \cdots & u[k] & \cdots \end{bmatrix}$.

Métodos de Estimação

- Vamos considerar a família de modelos genérica

$$y[k] = G(q,\theta)u[k] + H(q,\theta)e[k]$$

- Seu preditor associado corresponde a

$$\hat{y}(k|\theta) = \underbrace{W_y(q,\theta)}_{[1-H^{-1}(q,\theta)]} y[k] + \underbrace{W_u(q,\theta)}_{H^{-1}(q,\theta)G(q,\theta)} u[k]$$

Exercício: Reescrever o sistema

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k + Du_k + \nu_k \end{aligned}$$

na representação padrão e indicar a qual família de modelos pertence.

Métodos de Estimação

NOTAÇÃO DE ESTIMADORES

- *Aqui, $\hat{x}_k$ denota a estimativa de x dadas k observações*
- *Para um conjunto de k observações de x, denotamos por*

$$\hat{x}(Y) \triangleq \hat{x}(k, Y^k) \triangleq \hat{x}(k, \varphi^k)$$

- *Faremos ainda a simplificação*

$$f_{x|y}(\alpha|\beta) = f_{x|y}(x|y) \text{ denotada apenas por } f_{x|y}$$

Analogamente para as p.d.f. marginais.

NOTA: *x é o parâmetro a ser estimado, y_j, a medida realizada no instante j e w_j, o ruído de medida que incide sobre a j-ésima leitura.*

Máxima Verossimilhança

- Estimador de máxima verossimilhança (ML – *Maximum Likelihood*)

- Método **não bayesiano** com estimador

$$\hat{x}^{\mathrm{ML}}(Y) = \arg\max_{x} \Lambda_Y(x) = \arg\max_{x} f_{y|x}$$

- Há **um único** valor verdadeiro $x = x_0$

- $\hat{x}^{\mathrm{ML}}(Y)$ é um vetor de medidas (**aleatório**) dada a presença de $w[\cdot]$, obtido segundo

$$\frac{d\Lambda_Y(x)}{dx} = \frac{df_{y|x}}{dx} = 0$$

Máximo a Posteriori

- Estimador de máximo a posteriori (MAP – *Maximum A Posteriori*)

- Método **bayesinao**, pressupõe que x é uma v.a. e sua estimativa é obtida a partir da p.d.f. a posteriori

$$\hat{x}^{\mathrm{MAP}}(Y) = \arg\max_{x} f_{x|y} = \arg\max_{x} \left[\frac{f_{y|x} f_x}{f_y} \right]$$

onde $f_y \neq 0$.

- Tanto o parâmetro x quanto sua estimativa $\hat{x}$ são v.a.

Exercícios

1. Obtenha o melhor estimador $\hat{x}^{\mathrm{ML}}$ para um parâmetro x, dado que o ruído que entra no sistema é $w \sim \mathcal{N}(0, \sigma^2)$, e sua medida descrita por

$$y = x + w$$

2. Vamos considerar agora que temos k medidas realizadas do parâmetro x, e portanto

$$y[j] = x + w[j], \quad j = 1, 2, \cdots, k$$

sendo os ruídos conhecidos com $w[j] \sim \mathcal{N}(0, \sigma^2)$, para todo j. Calcule o melhor estimador $\hat{x}^{\mathrm{ML}}$ nesta nova situação.

3. Por fim, obtenha o melhor estimador $\hat{x}^{\mathrm{MAP}}$ de $x \sim \mathcal{N}(\bar{x}, \sigma_x^2)$, para um experimento descrito conforme

$$y = x + w, \quad w \sim \mathcal{N}(0, \sigma^2)$$

4. Estimar (utilizando o estimador ML, o valor de um parâmetro x a partir de um conjunto de k medidas y_i, $i = 1, 2, \cdots, k$, sendo sua média nula e variância também desconhecida.

Mínimos Quadrados

- Estimador dos mínimos quadrados (LS – *Least Squares*) é outro método **não bayesiano**
- Dado um conjunto de medidas $y_j = h(j, x) + w_j$ com $j = 1, 2, \cdots, k$, a estimativa de x é tal que

$$\hat{x}^{\mathrm{LS}}(Y) = \arg\min_{x} \left\{ \sum_{j=1}^{k} [y_j - h(j, x)]^2 \right\}$$

- y é uma função não linear de x através de $h(j, x)$
- Não há hipótese sobre o ruído w. Para w gaussiano, o estimador LS corresponde ao estimador ML.

Mínimo Erro Quadrático Médio

- Estimador do mínimo erro quadrático médio (MMSE – *Minimum Mean Square Error*)

- Contrapartida aleatória do estimador dos mínimos quadrados, ou seja, x é uma v.a. com p.d.f. conhecida a priori

- Estimador calculado conforme

$$\hat{x}^{\mathrm{MMSE}}(Y) = \mathcal{E}\left(x|Y\right) \triangleq \int_{-\infty}^{\infty} x f_{x|y} dx$$

que é a solução de

$$\hat{x}^{\mathrm{MMSE}}(Y) = \arg\min_{\hat{x}} \mathcal{E}\left((\hat{x}-x)^2|Y\right)$$

- A estimativa MMSE será uma v.a. em decorrência de x e w serem v.a.

Polarização de Estimadores

- ***BIAS*** é um valor persistente que distancia a estimativa do valor verdadeiro de um parâmetro.

> **_DEFINIÇÃO V.1_**
>
> *Um estimador é dito* **Não Polarizado** ***(unbiased)*** *se o erro de estimativa possui média nula, ou seja,*
>
> $$\mathcal{E}\left(\tilde{x}\right) = 0, \qquad \tilde{x} \triangleq x - \hat{x}$$

- Interpretações:
  - Abordagem **não bayesianas**

    $$\mathcal{E}\left(\hat{x}(Y)\right) = x_0$$

    onde x_0 é o valor verdadeiro do parâmetro x.
  - Abordagem **bayesianas**.

    $$\mathcal{E}\left(\hat{x}(Y)\right) = \mathcal{E}\left(x\right)$$

    sendo f_x conhecida a priori e $f_{x|y}$ calculada.

Polarização de Estimadores

DEFINIÇÃO V.2

Um estimador é dito (simplesmente) **NÃO POLARIZADO** *se a Definição V.1 é válida para todo k.*

DEFINIÇÃO V.3

Um estimador é dito **ASSINTOTICAMENTE NÃO POLARIZADO** *se a Definição V.1 é válida para $k \to \infty$.*

Variância e Erro Quadrático Médio de um Estimador

DEFINIÇÃO V.4

A **variância (covariância) de um estimador** *é definida como*

$$\operatorname{var}(\hat{x}) \triangleq \mathcal{E}\left([\hat{x}(Y) - \mathcal{E}(\hat{x}(Y))]^2\right)$$
$$\operatorname{cov}(\hat{x}) \triangleq \mathcal{E}\left([\hat{x}(Y) - \mathcal{E}(\hat{x}(Y))]\,[\hat{x}(Y) - \mathcal{E}(\hat{x}(Y))]'\right)$$

DEFINIÇÃO V.5

O **erro quadrático médio** *(MSE) definido, para o caso escalar, como*

$$\operatorname{MSE}(\hat{x}(Y)) = \mathcal{E}\left(\tilde{x}^2\right)$$

onde $\tilde{x} \triangleq x - \hat{x}$*, ou para o caso vetorial*

$$\operatorname{MSE}(\hat{x}(Y)) = \mathcal{E}\left(\tilde{x}\tilde{x}'\right)$$

Variância e Erro Quadrático Médio de um Estimador

- O $\mathrm{MSE}(\hat{x})$ indica a dispersão do estimador em torno do valor verdadeiro do parâmetro x enquanto a $\mathrm{cov}(\hat{x})$ é utilizada para avaliar sua convergência para o valor médio.
- Para o caso de estimador **unbiased**, temos que $\mathcal{E}\left(\hat{x}(Y)\right) = x$.

  - Se x for ainda <u>NÃO BAYESIANO</u>

$$\begin{aligned}\mathrm{MSE}\left(\hat{x}(Y)\right) &= \mathcal{E}\left(\tilde{x}^2\right) = \mathcal{E}\left([\hat{x}(Y) - x_0]^2\right) = \mathcal{E}\left([\hat{x}(Y) - \mathcal{E}\left(\hat{x}(Y)\right)]^2\right)\\ &= \mathrm{var}(\hat{x}(Y))\end{aligned}$$

  - Se x for <u>BAYESIANO</u>

$$\begin{aligned}\mathrm{MSE}\left(\hat{x}(Y)\right) &\triangleq \mathcal{E}\left([\hat{x}(Y) - x]^2\right) = \mathcal{E}\left(\mathcal{E}\left([\hat{x}(Y) - x]^2 \,\middle|\, Y\right)\right)\\ &= \mathcal{E}\left(\mathrm{MSE}\left(\hat{x}(Y)\middle| Y\right)\right)\end{aligned}$$

$\mathcal{E}\left(\mathrm{MSE}\left(\hat{x}(Y)\middle| Y\right)\right)$ é o **erro médio quadrático condicional**.

Desvio Padrão de um Estimador

DEFINIÇÃO V.6

O **desvio padrão** *de um estimador é dado por*

$$\sigma_{\hat{x}(y)} = \sqrt{\mathrm{var}\left(\hat{x}(y)\right)}$$

Consistência de um Estimador

DEFINIÇÃO V.6

Um estimador é dito **consistente** *se a estimativa convergir para o valor verdadeiro em algum sentido estocástico.*

- Equivale a exigir que o erro de estimação convirja para zero, sob algum sentido estocástico
- Calculamos
  - Para o caso não bayesiano

$$\lim_{k\to\infty} \mathcal{E}\left((\hat{x}(k, Y^k) - x_0)^2\right)$$

  - Para o caso bayesiano,

$$\lim_{k\to\infty} \mathcal{E}\left((\hat{x}(k, Y^k) - x)^2\right)$$

- Nos dois casos acima, se $\lim_{k\to\infty} \mathcal{E}\left(\tilde{x}^2\right) = 0$ o estimador será consistente.

Limitante Inferior de Cramer-Rao e Informação de Fisher

- $\Lambda(x) = f_{y|x}$, por construção, pode ser entendida como a quantidade de informação que temos a priori a respeito de um determinado parâmetro a ser estimado
- O funcional J, associado a $\Lambda(x)$, quantifica esta informação, a Informação de Fisher.

DEFINIÇÃO V.7

A **Informação de FISHER**, *para o caso escalar, é*

$$J \triangleq -\mathcal{E}\left(\frac{\partial^2 \ln f_{y|x}}{\partial x^2}\right)_{x=x_0} = \mathcal{E}\left(\left[\frac{\partial \ln f_{y|x}}{\partial x}\right]^2\right)_{x=x_0}$$

em que x_0 é o valor verdadeiro do parâmetro x. Para o caso vetorial,

$$J \triangleq -\mathcal{E}\left(\bigtriangledown_x \bigtriangledown_x' \ \ln f_{y|x}\right)_{x=x_0}$$

conhecida como **Matriz de Informação de Fisher** *(FIM).*

Desigualdade de Cramer-Rao

- O valor verdadeiro do parâmetro x_0 não é conhecido, portanto, o valor exato de J, também não será conhecido
- A quantidade de informação disponível corresponde a quantidade de informação extraída pelo Estimador de Máxima Verossimilhança (ML)

TEOREMA V.1

A estimativa de um parâmetro escalar x dado um estimador unbiased $\hat{x}(Y)$ tem sua variância limitada de acordo com

$$\mathcal{E}\left([\hat{x}(Y) - x_0]^2\right) \geq J^{-1}$$

em que x_0 é o valor verdadeiro do parâmetro x, a função densidade de probabilidade de Y^k é $f_y(x_0, Y^k)$ e suas derivadas de primeira e segunda ordem existem. Esta desigualdade é conhecida como **Desigualdade de Cramer-Rao** *ou (CRLB).*

Eficiência de um Estimador

- A eficiência dos estimadores é uma medida da quantidade de informação que pode ser extraída de um parâmetro

DEFINIÇÃO V.8

Um estimador é dito **eficiente** *se sua* VARIÂNCIA *é igual ao limitante inferior de Cramer-Rao (J^{-1}).*

- um estimador é então eficiente se a quantidade de informação que ele pode extrair é igual a quantidade de informação existente.

Melhor Estimador

- Buscamos o **MELHOR ESTIMADOR** segundo algum critério pré-definido.
  - Simplicidade
  - Requeira a menor quantidade de informação a priori

⬇

- △ Funções lineares
- △ Estimadores não polarizados (*unbiased*)
- △ Erro de estimação descorrelacionado da medida/observação
- △ **Princípio da ortogonalidade**

Melhor Estimador

- Critério que minimiza o erro (quadrático) de estimação **MMSE**
- Problema

$$\min_{\hat{x}} \mathcal{E}\left([x-\hat{x}]^2 \,\middle|\, Y\right)$$

com $\hat{x}$ obtido a partir do conjunto de medidas Y^k

Estimadores Lineares

- Supondo relação **linear** de $\hat{x}$ com as medidas Y^k

$$\hat{x} = \sum_{j=1}^{k} \alpha_j y_j$$

onde $\alpha_j > 0$ são coeficientes normalizados tais que $\sum_{j=1}^{k} \alpha_j = 1$

- Substituindo o $\hat{x}$ na expressão anterior
  - Problema linear de fácil solução
  - Solução independe das funções densidade de probabilidades

Estimadores Lineares

- y_i, $i = 1, 2, \cdots, k$, escalares (ou vetores), v.a. com médias nulas e reais, y_i definido no espaço vetorial de Hilbert

- Espaço fechado em relação a soma e a multiplicação, com norma $\langle y_i, y_i \rangle = \mathcal{E}\left(y_i^2\right) = \|y_i\|^2$, sendo

$$\langle y_i, y_k \rangle = \mathcal{E}\left(y_i y_k\right)$$

Da álgebra linear, dois vetores são ortogonais ($y_i \perp y_k$) se e somente se $\langle y_i, y_k \rangle = 0$

$$\Downarrow$$

$$\langle y_i, y_k \rangle = \mathcal{E}\left(y_i y_k\right) = 0$$

$$\Downarrow$$

y_i e y_k são **descorrelacionadas**, uma vez que possuem média nula

Estimadores Lineares

- Considere um **estimador linear** de x

  Encontrar $\hat{x}$ a partir da minimização de

$$\|\tilde{x}\|^2 = \mathcal{E}\left([x - \hat{x}]^2\right)$$

$$\Downarrow$$

$$\mathcal{E}\left(-2xy_i + 2y_i \sum_{j=1}^{k} \alpha_j y_j\right) = 0 \Longrightarrow \mathcal{E}\left(\left(x - \sum_{j=1}^{k} \alpha_j y_j\right) y_i\right) = 0$$

$$\Downarrow$$

$$\langle \tilde{x}, y_i \rangle = 0, \quad \forall i = 1, 2, \cdots, k$$

Princípio da Ortogonalidade

- De $\langle \tilde{x}, y_i \rangle = 0$ temos $\tilde{x} \perp y_i$, $i = 1, 2, \cdots, k$.

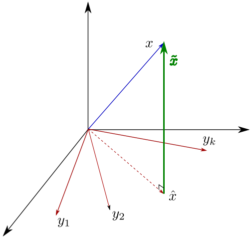
> O melhor (MMSE) estimador linear de um parâmetro x é aquele cujo erro é ortogonal ao conjunto de medidas y_i, $i = 1, 2, \cdots, k$.

Princípio da Ortogonalidade

PRINCÍPIO DA ORTOGONALIDADE

Para que um estimador tenha **NORMA MÍNIMA** *seu erro deve ser* **ORTOGONAL** *às observações.*

OBSERVAÇÃO : O estimador ótimo $\hat{x}$, dado um conjunto de medidas Y^k de x, é a projeção ortogonal de x no espaço gerado pelo conjunto de k medidas y



Estimadores Lineares – Considerações

- Para uma v.a. x, o melhor estimador linear, no caso geral, será

$$\hat{x} = \beta_0 + \sum_{j=1}^{k} \beta_i y_i$$

- O erro médio quadrático (MSE) será

$$\mathcal{E}\left(\tilde{x}^2\right) = \mathrm{var}\left(\tilde{x}\right) - \mathcal{E}\left(\tilde{x}\right)^2$$

- Condições para determinar $\hat{x}$:
  - Estimador não polarizado $\mathcal{E}\left(\tilde{x}\right) = \mathcal{E}\left(x - \hat{x}\right) = 0$
  - Estimador MMSE sendo, $\mathrm{MSE} = \mathcal{E}\left(\tilde{x}^2\right)$