2.15 Estimação de parâmetros

O problema da estimação de parâmetros pode ser definido como: dadas medições y_k relacionadas com variáveis descritivas x_k tais que

$$y_k = f(x_k, \theta)$$

com $f(\cdot, \theta)$ sendo uma função de parâmetros θ (escalar ou vetor), determinar uma estimativa $\hat{\theta}$ de θ de modo que $\hat{\theta}$ seja o mais próximo possível de θ. As soluções a este problema serão classificadas em três abordagens distintas: quadrados mínimos, máximo de verossimilhança e estimação bayesiana.

Quanto ao problema, esse pode ser classificado em linear ou não-linear, dependendo da forma de $f(\cdot, \theta)$ com relação a θ.

Exemplo: estimação linear ou não-linear?

Se $f(x_k, \theta)$ puder ser escrito na forma

$$f(x_k, \theta) = \varphi(x_k)^T \cdot \theta$$

com $\varphi(\cdot)^T : \mathbb{R}^n \to \mathbb{R}^m$ sendo um vetor coluna que é função de $x_k \in \mathbb{R}^n$ e $\theta \in \mathbb{R}^m$ sendo o vetor de parâmetros, então o problema de estimação é dito linear. Se $f(x_k, \theta) = a \cdot \sin(x_k) + b \cdot \cos(x_k)$, então

$$\varphi(x_k)^T = [\sin(x_k) \quad \cos(x_k)]$$

$$\theta^T = [a \quad b]$$

Mas se $f(x_k, \theta) = a.\sin(x_k) + b.\cos(x_k) + 1$, que é uma função afim não pode ser escrita como $\varphi(x_k)^T \cdot \theta$. Mas, com a seguinte mudança

$$y_k - 1 = a.\sin(x_k) + b.\cos(x_k)$$

$$= \underbrace{[\sin(x_k) \quad \cos(x_k)]}_{\varphi(x_k)^T} \cdot \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\theta}$$

o problema se torna linear.

Com $f(x_k) = \cos(x_k + \theta)$, temos então um problema não-linear.

Algumas propriedades desejadas para estimadores são:

i) Estimativa não-tendenciosa:

$$E\{\hat{\theta}\} = \theta$$

ii) Estimativa não-tendenciosa de variância mínima:

Para toda estimativa θ^* que satisfaz

$$E\{\theta^*\} = \theta$$

$\hat{\theta}$ é uma estimativa não-tendenciosa de variância mínima se $E\{\hat{\theta}\} = \theta$ e

$$VAR\{\hat{\theta}\} \leq VAR\{\theta^*\}.$$

iii) Estimativa coerente

$$\lim_{K \to \infty} Pr\{|\hat{\theta} - \theta| > \varepsilon\} = 0$$

para todo $\varepsilon > 0$. Assim, uma estimativa é

coerente se $\lim_{n\to\infty} E\{\hat{\theta}\} = 0$ e $\lim_{n\to\infty} Var\{\hat{\theta}\} = 0$

As definições acima se aplicam a problemas de estimação determinística, em que θ é considerado uma constante (i.e., $Var\{\theta\} = 0$). Em estimação estocástica temos ainda

iv) Estimativa consistente

$$E\{\hat{\theta} - \theta\} = 0 \quad e \quad Var\{\hat{\theta}\} = Var\{\theta\}$$

para uma estimativa inconsistente, basta $E\{\hat{\theta} - \theta\} \neq 0$

v) Estimativa consistente pessimista

$$E\{\hat{\theta} - \theta\} = 0 \quad e \quad Var\{\hat{\theta}\} > Var\{\theta\}$$

também chamada de estimativa conservativa.

vi) Estimativa consistente otimista

$$E\{\hat{\theta} - \theta\} = 0 \quad e \quad Var\{\hat{\theta}\} < Var\{\theta\}$$

Em muitas aplicações, estimativas otimistas são perigosas quando são usadas em processo de decisão.

2.15.1 Estimação por mínimos quadrados

Nos métodos de mínimos quadrados, o problema se escreve da forma seguinte:

$$\hat{\theta} = \arg\min_{\theta} J(r, \theta)$$

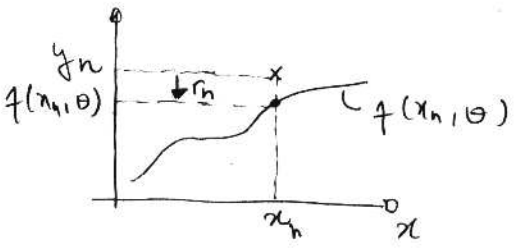
com r sendo os resíduos (ou erro) de estimação,

$J(r,\theta)$ é uma função do tipo

$$J(r,\theta) = \sum_{n=1}^{N} r_n(x_n, y_n, \theta)^2$$

Assim, $\hat{\theta}$ é a estimativa que minimiza a soma dos quadrados dos resíduos. Em função da forma como os resíduos são definidos, temos os mínimos quadrados (MQ) ou ainda os mínimos quadrados totais (MQT). Em MQ, os resíduos são simplesmente

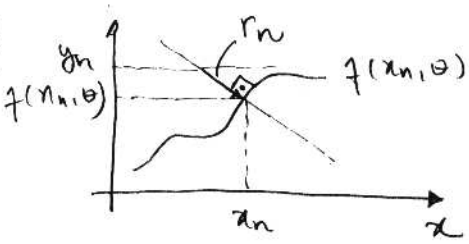
$$r_n = y_n - f(x_n, \theta) \Rightarrow$$



enquanto que, no MQT,

$r_n \Rightarrow$ distância ortogonal

$\Rightarrow$



<u>Exemplo</u>: Estimação dos parâmetros de uma reta

Seja o modelo

$$y = a.x + b$$

Este modelo pode ser escrito como

$$y_n = \varphi(x_n)^T.\theta, \text{ com } \varphi(x_n)^T = [x_n \; 1]$$

e $\theta = [a \; b]^T$. Pela abordagem MQ, temos

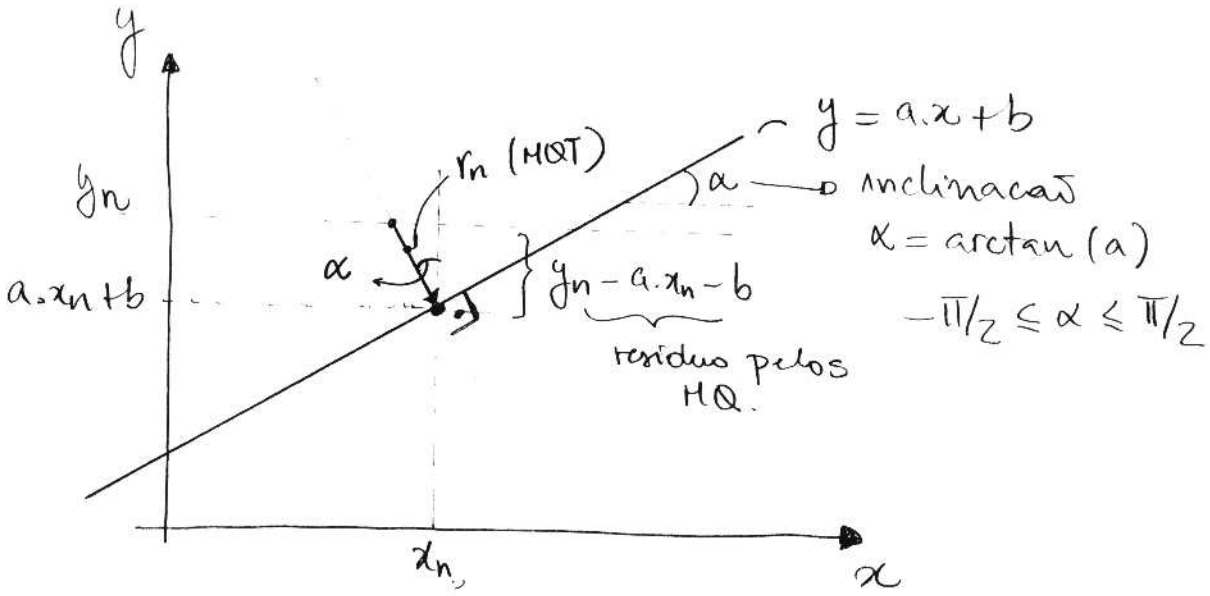
$$r_n = y_n - a.x_n - b$$

enquanto que, com a abordagem MQT, os resíduos

são definidos como

$$r_n = \left(y_n - (a \cdot x_n + b)\right) / \cos(\alpha)$$

com $\alpha = \arctan(a)$. O esboço de r_n é mostrado abaixo.



Pelo método dos mínimos quadrados, a solução deste problema é dada pelo mínimo de $J(r, \theta)$ com relação a θ, que satisfaz a

$$\left.\frac{\partial J(r,\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

$$J(r,\theta) = \sum_{n=1}^{N} (y_n - a \cdot x_n - b)^2 = \sum_{n=1}^{N} r_n^2$$

$$\frac{\partial J}{\partial \theta} = \begin{bmatrix} \frac{\partial J}{\partial a} \\ \frac{\partial J}{\partial b} \end{bmatrix} = \sum_{n=1}^{N} 2 \cdot \frac{\partial r_n}{\partial \theta} \cdot r_n$$

$$\frac{\partial J}{\partial a} = \sum_{n=1}^{N} 2 \cdot (y_n - a \cdot x_n - b) \cdot (-x_n)$$

$$\frac{\partial J}{\partial b} = \sum_{n=1}^{N} 2 \cdot (y_n - a \cdot x_n - b) \cdot (-1)$$

$$\left.\frac{\partial J}{\partial a}\right|_{\substack{a=\hat{a}\\ b=\hat{b}}} = 0 \Rightarrow \sum_{n=1}^{N} \left(\hat{a}.x_n^2 + \hat{b}.x_n - x_n y_n\right) = 0$$

$$\Rightarrow \hat{a}.\sum_{n=1}^{N} x_n^2 + \hat{b}.\sum_{n=1}^{N} x_n = \sum_{n=1}^{N} x_n.y_n$$

$$\Rightarrow \hat{a}.\frac{1}{N}.\sum_{n=1}^{N} x_n^2 + \hat{b}.\frac{1}{N}.\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{i=1}^{N} x_n y_n$$

$$\left.\frac{\partial J}{\partial b}\right|_{\substack{a=\hat{a}\\ b=\hat{b}}} = 0 \Rightarrow \sum_{n=1}^{N} \left(\hat{a}.x_n + \hat{b} - y_n\right) = 0$$

$$\Rightarrow \hat{a}.\sum_{n=1}^{N} x_n + \hat{b}.\sum_{n=1}^{N} 1 = \sum_{n=1}^{N} y_n$$

$$\Rightarrow \hat{a}.\frac{1}{N}.\sum_{n=1}^{N} x_n + \hat{b} = \frac{1}{N}.\sum_{n=1}^{N} y_n$$

Definindo:

$$\bar{x} = \frac{1}{N}.\sum_{n=1}^{N} x_n \quad ; \quad \bar{y} = \frac{1}{N}.\sum_{n=1}^{N} y_n$$

$$C_{xx} = \frac{1}{N}.\sum_{n=1}^{N} x_n^2 \quad ; \quad C_{xy} = \frac{1}{N}.\sum_{n=1}^{N} x_n.y_n$$

Temos que resolver o sistema abaixo para encontrar as estimativas $\hat{a}$ e $\hat{b}$:

$$\hat{a}.C_{xx} + \hat{b}.\bar{x} = C_{xy}$$

$$\hat{a}.\bar{x} + \hat{b} = \bar{y}$$

fazendo $\hat{b} = \bar{y} - \hat{a}.\bar{x}$, temos

$$\hat{a}.C_{xx} + \bar{x}.\bar{y} - \hat{a}.\bar{x}^2 = C_{xy}$$

e assim

$$\hat{a} = \frac{C_{xy} - \bar{x}.\bar{y}}{C_{xx} - \bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{b} = \bar{y} - \frac{S_{xy}}{S_{xx}}.\bar{x}$$

com $S_{xy} = \frac{1}{N} \cdot \sum_{n=1}^{N} (x_n - \bar{x}) \cdot (y_n - \bar{y}) = C_{xy} - \bar{x}\bar{y}$

e $S_{xx} = \frac{1}{N} \cdot \sum_{n=1}^{N} (x_n - \bar{x})^2 = C_{xx} - \bar{x}^2$

Pode-se perceber que se $S_{xx} = 0$ e $S_{xy} \neq 0$, então $\hat{a} \to \infty$ (reta vertical com medições perfeitas). Neste caso, a parametrização $y = ax + b$ é limitada. Melhor seria usar o modelo

$$a.y + b.x + c = 0$$

com resíduo $r_n = a.y_n + b.x_n + c$.

No caso dos MQT, a solução de

$$\hat{\theta} = \arg\min_{\theta} \sum_{n=1}^{N} \{(y_n - (a.x_n + b)) / \cos(\arctan(a))\}^2$$

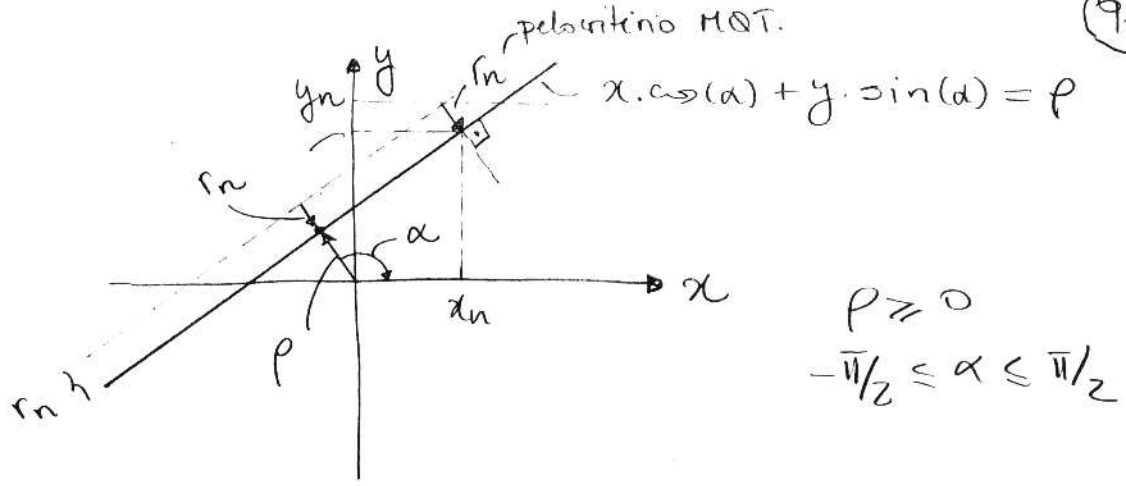
parece um pouco complicado, mas é resolvível. Ainda assim, a escolha do modelo $y = ax + b$ não é apropriada com retas verticais. O modelo

$$ay + bx + c = 0$$

é mais apropriado e se aplica a qualquer tipo de reta. No entanto, o resíduo pelo método MQT parece complexo. Para reduzir a complexidade, o modelo de reta dado por

$$\rho = x.\cos(\alpha) + y.\sin(\alpha)$$

é mais interessante pois envolve apenas dois parâmetros (ρ e α).



Temos que ρ e α são as coordenadas polares do ponto da reta que está mais próximo da origem

Pode-se verificar que, para o MQT, temos

$$r_n = \rho - x_n \cdot \cos(\alpha) - y_n \cdot \sin(\alpha)$$

E a minimização de

$$J = \sum_{n=1}^{N} \left(\rho - x_n \cdot \cos(\alpha) - y_n \cdot \sin(\alpha)\right)^2$$

é obtida com parâmetros $\hat{\rho}$, $\hat{\alpha}$ tais que

$$\left.\frac{\partial J}{\partial \rho}\right|_{\substack{\rho=\hat{\rho} \\ \alpha=\hat{\alpha}}} = \left.\frac{\partial J}{\partial \alpha}\right|_{\substack{\rho=\hat{\rho} \\ \alpha=\hat{\alpha}}} = 0$$

Que resulta em:

$$\hat{\rho} = \bar{x} \cdot \cos(\hat{\alpha}) + \bar{y} \cdot \sin(\hat{\alpha})$$

$$\hat{\alpha} = \frac{1}{2} \cdot \arctan\left(\frac{-2 \cdot S_{xy}}{S_{yy} - S_{xx}}\right)$$

com $\bar{x}$, $\bar{y}$, S_{xy} e S_{xx} já definidos acima, e

$$S_{yy} = \frac{1}{N} \cdot \sum_{n=1}^{N} (y - \bar{y})^2$$

A generalização dos modelos apresentados a pouco pode ser dada pelo modelo linear abaixo:

$$y_n = \varphi(x_n)^T \cdot \theta + \varepsilon_n$$

com $y_k \in \mathbb{R}$ e ε_k sendo o ruído de medição. Então

$$J = \sum_{n=1}^{N} \left(y_n - \varphi(x_n)^T \cdot \theta \right)^2$$

Cujo mínimo é obtido fazendo

$$\left.\frac{\partial J}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

$$= \sum_{n=1}^{N} 2 \cdot (-\varphi(x_n)) \cdot \left(y_n - \varphi(x_n)^T \cdot \hat{\theta} \right)$$

Então

$$\sum_{n=1}^{N} y_n \cdot \varphi(x_n) = \sum_{n=1}^{N} \varphi(x_n) \cdot \varphi(x_n)^T \cdot \hat{\theta}$$

Assim, a estimativa $\hat{\theta}$ é obtida por:

$$\underbrace{\left(\sum_{n=1}^{N} \varphi(x_n) \cdot \varphi(x_n)^T \right)}_{m \times m \text{ simétrica}} \cdot \hat{\theta} = \underbrace{\sum_{n=1}^{N} \varphi(x_n) \cdot y_n}_{1 \times m} \quad \text{Lo escalar}$$

$m = \dim \hat{\theta}$

$$\hat{\theta} = \underbrace{\left(\sum_{n=1}^{N} \varphi(x_n) \cdot \varphi(x_n)^T \right)^{-1}}_{m \times m} \cdot \underbrace{\left(\sum_{n=1}^{N} \varphi(x_n) \cdot y_n \right)}_{m \times 1}$$

Exemplo: Aproximação polinomial

Dado pares de medições $\{x_n, y_n\}$, encontrar os parâmetros $\theta = [a_0 \; a_1 \; \cdots \; a_m]^T$ para o seguinte modelo polinomial

$$y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_m \cdot x^m$$
$$= \varphi^T(x) \cdot \theta$$

com $\varphi^T(x) = [1 \; x \; x^2 \cdots x^m]$.

Exemplo: Identificação de sistemas

Considere o seguinte modelo ARX:

$$y(k) = b_1 \cdot u(k-1) + \ldots + b_m \cdot u(k-m) - a_1 \cdot y(k-1) - \ldots - a_m \cdot y(k-1)$$

Este modelo relaciona amostras da entrada e saída de um sistema cuja função de transferência discreta é dada por

$$\frac{y(k)}{u(k)} = \frac{b_1 \cdot z^{-1} + \ldots + b_m \cdot z^{-m}}{1 + a_1 \cdot z^{-1} + \ldots + a_m \cdot z^{-m}}$$

Considerando $\theta = [b_1, \ldots, b_m, a_1, \ldots, a_m]^T$ e $\varphi^T(u(k)) = [u(k-1) \ldots u(k-m) \; -y(k-1) \ldots -y(k-m)]$, $\hat{\theta}$ pode ser obtido pelo método dos mínimos quadrados.

Observa-se que, sendo

$$\varepsilon_n = y_n - \varphi(x_n)\cdot\theta$$

se considerarmos que $\varepsilon_n \sim N(0, \sigma_\varepsilon^2)$, então

$$\sigma_\varepsilon^2 = \frac{1}{N-1}\cdot\sum_{n=1}^{N}\varepsilon_n^2 = \frac{1}{N-1}\cdot J(\theta)$$

(valor da função ou custo p/ $\hat{\theta}$.)

A estimativa $\hat{\theta}$ pode ser escrita

$$\hat{\theta} = (\phi^T\cdot\phi)^{-1}\cdot\phi^T\cdot Y$$

com $\phi = \underbrace{\begin{bmatrix}\varphi(x_1)^T\\ \vdots \\ \varphi(x_N)^T\end{bmatrix}}_{N\times m}$ e $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N\end{bmatrix}$

ϵ o modelo de erro

$$Y = \phi\cdot\theta + \varepsilon$$

com $\varepsilon = [\varepsilon_1 \dots \varepsilon_N]^T$, θ sendo o verdadeiro vetor de parâmetros e $\varepsilon \sim N(0, P_\varepsilon)$ com

$$P_\varepsilon = E\{\varepsilon\cdot\varepsilon^T\} = \sigma_\varepsilon^2\cdot I_N$$

considerando ε_i e ε_j, $i\neq j$, independentes, e I_N é a matriz identidade de dimensão N. Então ε_n é iid.

O estimador $\hat{\theta} = (\phi^T\cdot\phi)^{-1}\cdot\phi^T\cdot Y$ tem as seguintes propriedades:

$$E\{\hat{\theta} - \theta\} = E\{(\phi^T\phi)^{-1}\cdot\phi^T\cdot(\underbrace{\phi\cdot\theta+\varepsilon}_{Y}) - \theta\}$$

$$= E\{\underbrace{(\phi^T\cdot\phi)^{-1}\cdot\phi^T\cdot\phi}_{I_m \text{ pois } A\cdot A^{-1}=I}\cdot\theta + (\phi^T\cdot\phi)^{-1}\cdot\phi^T\cdot\varepsilon - \theta\}$$

$$= E\{\theta - \theta + (\phi^T\phi)^{-1}\cdot\phi^T\varepsilon\}$$

$$= (\phi^T\cdot\phi)^{-1}\cdot\phi^T\, E\{\varepsilon\} = 0$$

ou seja, $\hat{\theta}$ é não tendencioso (observe que θ é tratado como uma constante). Assim,

$$E\{\hat{\theta}-\theta\} = E\{\hat{\theta}\} - E\{\theta\} = E\{\hat{\theta}\} - \theta = 0$$

significando que $E\{\hat{\theta}\} = \theta$. Então

$$P_{\hat{\theta}} = E\{(\hat{\theta} - E\{\hat{\theta}\})\cdot(\hat{\theta} - E\{\hat{\theta}\})^T\}$$

$$= E\{(\hat{\theta}-\theta)\cdot(\hat{\theta}-\theta)^T\}, \text{ pois } \hat{\theta}-\theta = (\phi^T\phi)^{-1}\cdot\phi^T\varepsilon$$

$$= E\{(\phi^T\cdot\phi)^{-1}\cdot\phi^T\cdot\varepsilon\cdot\varepsilon^T\cdot\phi\cdot(\phi^T\cdot\phi)^{-1}\}$$

$$= (\phi^T\cdot\phi)^{-1}\cdot\phi^T\cdot P_\varepsilon\cdot\phi\cdot(\phi^T\phi)^{-1}$$

($P_\varepsilon \hookrightarrow \sigma_\varepsilon^2\cdot I_N$)

$$= \underbrace{(\phi^T\cdot\phi)^{-1}\,\phi^T\cdot\phi}_{I_m}\cdot(\phi^T\cdot\phi)^{-1}\cdot\sigma_\varepsilon^2$$

$$P_{\hat{\theta}} = (\phi^T\cdot\phi)^{-1}\cdot\sigma_\varepsilon^2 = \left(\sum_{n=1}^{N}\varphi(x_n)\,\varphi^T(x_n)\right)^{-1}\cdot\sigma_\varepsilon^2$$

com σ_ε^2 podendo ser estimado a partir de $J(\hat{\theta})$

$$\sigma_\varepsilon^2 = \frac{1}{N-1}\cdot J(\hat{\theta})$$

como $J(\hat{\theta}) \leq J(\theta^*)$ pois $\hat{\theta}$ minimiza J, então $P_{\hat{\theta}} \leq P_{\theta^*}$, e $\hat{\theta}$ é também conhecido como estimador de mínima variância.

pois $\hat{\theta}$ é uma estimativa não tendenciosa de θ.

Conforme foi exposto, o método dos mínimos quadrados considera que todas as medições têm pesos idênticos (i.e., mesma importância). No entanto

um y_i mais confiável do que um y_j pode ter um peso $w_i > w_j$. Isto resulta nos mínimos quadrados ponderados (MQP):

$$J = \sum_{n=1}^{N} w_n \cdot (y_n - \varphi^T(x_n) \cdot \theta)^2$$

$$= (Y - \Phi \cdot \theta)^T \cdot W \cdot (Y - \Phi \cdot \theta)$$

com $W = \text{diag}(w_1, \ldots, w_N)$, então pode-se verificar que

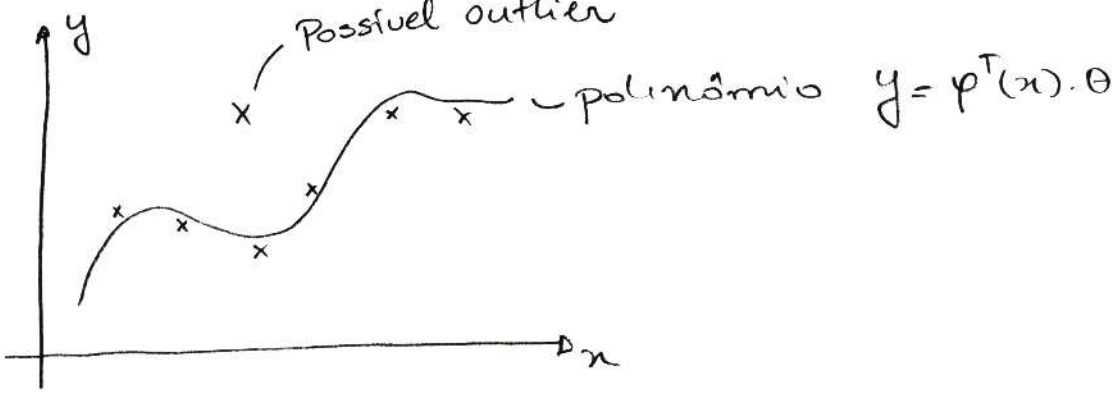
$$\left.\frac{\partial J}{\partial \theta}\right|_{\theta=\hat{\theta}} = \sum_{n=1}^{N} -w_n \cdot (y_n - \varphi^T(x_n) \cdot \hat{\theta}) \cdot \varphi^T(x_n) = 0$$

tem como solução

$$\hat{\theta} = \left(\sum_{n=1}^{N} \varphi(x_n) \cdot \varphi(x_n)^T \cdot w_n \right)^{-1} \cdot \left(\sum_{n=1}^{N} \varphi(x_n) \cdot w_n \cdot y_n \right)$$

$$= (\Phi^T \cdot W \cdot \Phi)^{-1} \cdot \Phi^T \cdot W \cdot Y$$

Se w_n for escolhido apropriadamente, então a n-ésima medição pode ser descartada da estimação se ela for um "outlier", ou seja, uma medida incompatível com o modelo:



Pode-se perceber que o mínimos quadrados é um caso particular do MQP, com $w_n = 1$, ou seja não fazendo distinção entre bons e mais medidas. Entas, o MQ é dito um estimador não-robusto pois basta um outlier para polarizar a estimação. Uma abordagem robusta consiste em usar o seguinte algoritmo:

i) Iniciar todos os pesos w_n iguais a 1 (MQ)
ii) Contados de iteração $K=0$.
iii) Calcular $\hat{\theta}_K$ usando MQP
iv) Re-calcular $w_n = \mu(r_n)$
v) incrementar K e retornar ao passo (iii) se o critério de parada não for verificado

O algoritmo acima calcula pesos w_n de forma iterativa usando uma funcional $\mu(r_n)$, com r_n sendo o resíduo de estimação usando a estimativa da K-ésima iteração:

$$r_n = y_n - \varphi(x_n)^T \cdot \hat{\theta}_K$$

Uma destas funcionais é a função de Huber

$$w_n = \min\left(1, \frac{c \cdot MAD}{abs(r_n)}\right)$$

com $c = 1{,}345$ e $MAD = MED\left(|r_n - MED(r_n)|\right)$. O estimador obtido com a função de Huber pertence à classe dos M-estimadores robustos.

2.15.2 Máximo da Verossimilhança

Dadas medições

$$y = f(x, \theta),$$

estimador do máximo de verossimilhança (MV) é dado por

$$\hat{\theta} = \arg\max_{\theta} p(y|\theta)$$

com $p(y|\theta)$ sendo a função de dens. condicional de y dado θ, também chamada de função de verossimilhança. (likelihood function).

Maximizar $p(y|\theta)$ é equivalente a maximizar $\ln(p(y|\theta))$, uma vez que $\ln(\cdot)$ é uma função monotônica crescente para $p(y|\theta) \geq 0$:

$$\hat{\theta} = \arg\max_{\theta} \ln(p(y|\theta))$$

Esta versão do MV é interessante quando $p(y|\theta)$ envolve exponenciais.

Exemplo: Seja $y = \varphi^T(x) \cdot \theta$ um modelo linear em θ do qual pares $\{y_n, x_n\}$ são obtidos (estimação linear) se as medições y_n são consideradas $iid \sim N\{E\{y_n\}, \sigma_y^2\}$ com $E\{y_n\} = \varphi^T(x_n) \cdot \theta$, uma vez que θ não é uma va, mas sim uma constante. Então

$$p(y_n|\theta) = \frac{1}{\sqrt{2\pi} \cdot \sigma_y} \cdot \exp\left\{-\frac{1}{2} \frac{(y_n - \varphi^T(x_n) \cdot \theta)^2}{\sigma_y^2}\right\}$$

Sendo $p(y|\theta) = p(y_1, y_2, \dots, y_N|\theta)$ e y_n iid, então

$$p(y|\theta) = \prod_{n=1}^{N} p(y_n|\theta)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\cdot\sigma_y} \cdot \exp\left(\frac{-1}{2}\cdot\frac{(y_n - \varphi(x_n)^T\theta)^2}{\sigma_y^2}\right)$$

$$= \frac{1}{(\sqrt{2\pi}\cdot\sigma_y)^N} \cdot \exp\left(\sum_{n=1}^{N} \frac{-1}{2}\frac{(y_n - \varphi(x_n)^T\cdot\theta)^2}{\sigma_y^2}\right)$$

Como $p(y|\theta)$ é uma função exponencial, é melhor maximizar $\ln(p(y|\theta))$:

$$\ln(p(y|\theta)) = \underbrace{-N\cdot\ln(\sqrt{2\pi}\,\sigma_y)}_{\text{Termo constante}} - \underbrace{\frac{1}{2\sigma_y^2}\sum_{n=1}^{N}(y_n - \varphi(x_n)^T\cdot\theta)^2}_{\text{único termo que depende de } \theta}$$

O máximo de $\ln(p(y|\theta))$ satisfaz a

$$\left.\frac{\partial \ln(p(y|\theta))}{\partial\theta}\right|_{\theta=\hat{\theta}} = \frac{1}{\sigma_y^2}\cdot\sum_{n=1}^{N}(y_n - \varphi(x_n)^T\cdot\hat{\theta})\cdot\varphi(x_n)^T$$

$$= 0$$

Ou seja, deve-se resolver

$$\sum_{n=1}^{N}(y_n - \varphi(x_n)^T\cdot\hat{\theta})\cdot\varphi(x_n)^T = 0$$

resultando em

$$\hat{\theta} = \left(\sum_{n=1}^{N}\varphi(x_n)\cdot\varphi(x_n)^T\right)^{-1}\cdot\left(\sum_{n=1}^{N}\varphi(x_n)^T\cdot y_n\right)$$

que é o mesmo resultado obtido pelos mínimos quadrados.

Isto significa que, no caso de estimação linear com erro de medição Gaussiano de média nula e iid, HV e MQ possuem resultados idênticos

Exemplo: Considere o seguinte modelo não-linear

$$y = \theta^2 + \varepsilon$$

com $\theta \in \mathbb{R}$ sendo uma constante a ser estimada e ε representando o erro de medição com densidade $p_\varepsilon(\varepsilon)$. usando as técnicas de propagação de densidade, temos que a função de verossimilhança é dada por

$$p(y|\theta) = p_\varepsilon(y - \theta^2)$$

se $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ então

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\cdot\sigma_\varepsilon} \cdot \exp\left\{ -\frac{1}{2} \cdot \frac{(y-\theta^2)^2}{\sigma_\varepsilon^2} \right\}$$

se dispomos de N medidas, elas são iid, e

$$p(y|\theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\cdot\sigma_\varepsilon} \cdot \exp\left\{ -\frac{1}{2} \cdot \frac{(y_n-\theta^2)^2}{\sigma_\varepsilon^2} \right\}$$

e $$\ln(p(y|\theta)) = -N \cdot \ln(\sqrt{2\pi}\cdot\sigma_\varepsilon) - \frac{1}{2\cdot\sigma_\varepsilon^2} \cdot \sum_{n=1}^{N} (y_n - \theta^2)^2$$

O máximo de $\ln(p(y|\theta))$ satisfaz a

$$\left.\frac{\partial \ln(p(y|\theta))}{\partial \theta}\right|_{\theta=\hat{\theta}} = \frac{1}{2\cdot\sigma_\varepsilon^2} \cdot \sum_{n=1}^{N} 2\cdot(y_n - \hat{\theta}^2)\cdot 2\cdot\hat{\theta}$$

$$= \frac{2}{\sigma_\varepsilon^2} \cdot \sum_{n=1}^{N} (y_n - \hat{\theta}^2)\cdot\hat{\theta}$$

b) $\hat{\theta}_2 = \sqrt{\bar{y}} \Rightarrow \dfrac{\partial^2 \ln(p(y|\theta))}{\partial\theta^2} = -\dfrac{4N}{\sigma_\varepsilon^2} \cdot \bar{y}$

$\hat{\theta}_2$ é máximo se $\bar{y} > 0$

c) $\hat{\theta}_3 = -\sqrt{\bar{y}} \Rightarrow \dfrac{\partial^2 \ln(p(y|\theta))}{\partial\theta^2} = -\dfrac{4N}{\sigma_\varepsilon^2} \cdot \bar{y}$

$\hat{\theta}_3$ é máximo se $\bar{y} > 0$.

Portanto, sendo $\bar{y}$ um número real, temos como estimativas:

$$\hat{\theta} = \begin{cases} 0, & \text{se } \bar{y} < 0 \\ \pm\sqrt{\bar{y}}, & \text{se } \bar{y} \geq 0 \end{cases}$$

Observa-se que para $y \geq 0$, $p(y|\hat{\theta}_1) = p(y|\hat{\theta}_2)$, ou seja, o problema de estimação MV tem duas soluções que correspondem aos modos de $p(y|\theta)$.

Deve ser observado que, embora o parâmetro θ a ser estimado tenha sido até então considerado uma constante, a estimativa $\hat{\theta}$ é uma variável aleatória. Por exemplo, no caso anterior, $\hat{\theta} = \sqrt{\bar{y}}$ é uma va uma vez que $\hat{\theta}$ é uma função da va $\bar{y}$, que por sua vez é função das vas $y_1, \ldots, y_N$. Sendo assim $\hat{\theta}$ possui uma função de densidade associada.

É ainda, $\hat{\theta}$ possui uma variância (medida de incerteza). Como $\hat{\theta} = \sqrt{\bar{y}}$, sua variância pode ser obtida da variância de $\bar{y}$

$$V_{AR}\{\bar{y}\} = V_{AR}\left\{\frac{1}{N} \cdot \sum_{n=1}^{N} y_n\right\}$$

Sendo $y_n = \theta^2 + \varepsilon_n$, com $\varepsilon_n \sim N(0, \sigma_\varepsilon^2)$, então $V_{AR}\{y_n\} = \sigma_\varepsilon^2$. Sendo $\bar{y}$ uma função linear de $y_1, y_2, \ldots, y_n$, então

$$V_{AR}\{\bar{y}\} = \frac{1}{N^2} \cdot V_{AR}\{y_1\} + \ldots + \frac{1}{N^2} \cdot V_{AR}\{y_N\}$$

$$= \frac{1}{N^2} \sum_{n=1}^{N} V_{AR}\{y_n\} = \frac{1}{N^2} \cdot \sum_{n=1}^{N} \sigma_\varepsilon^2 = \frac{1}{N} \sigma_\varepsilon^2$$

$$\sigma_{\bar{y}}^2 = \frac{\sigma_\varepsilon^2}{N}$$

Então, $V_{AR}\{\hat{\theta}\}$ pode ser obtido através da variância de $\bar{y}$ por meio da propagação de covariâncias.

Considerando agora a teoria em geral, se θ^* é uma estimativa de θ (θ^* não é obrigatoriamente de variância mínima), a relação

$$V_{AR}\{\theta^*\} \geq \left\{ E\left\{ \left(\frac{\partial \ln(p(Y|\theta))}{\partial \theta} \right)^2 \right\} \right\}^{-1}$$

com o termo à direita podendo ser substituído por

$$\left\{ E\left\{ \left(\frac{\partial \ln(p(y|\theta))}{\partial \theta} \right)^2 \right\} \right\}^{-1} = - \left\{ E\left\{ \frac{\partial^2 \ln(p(y|\theta))}{\partial \theta^2} \right\} \right\}^{-1}$$

é conhecida como limite inferior de Cramer-Rao. Esta relação se verifica se θ^* for uma estimativa não tendenciosa de θ. E sendo $\hat{\theta}$ uma estimativa não tendenciosa e de variância mínima,

$$\frac{\partial \ln(p(y|\theta))}{\partial \theta} = (\hat{\theta}(y) - \theta).C(\theta) = 0,$$

com $C(\theta)$ sendo uma função independente de y (dados coletados). O limite inferior de Cramer-Rao é alcançado por $\theta^* = \hat{\theta}$. Observa-se que a relação acima tem como soluções $C(\theta) = 0$ e $\hat{\theta}(y) = \theta$. $C(\theta) = 0$ é independente das medições, sendo assim desinteressante para nós. A única solução interessante é obtida por $\hat{\theta}(y) = \theta$.

Exemplo: Estimação não-linear (pg 105) revisitada.

Temos que

$$\frac{\partial \ln(p(y|\theta))}{\partial \theta} = \frac{2}{\sigma_\varepsilon^2} \sum_{n=1}^{N} (y_n - \hat{\theta}^2).\hat{\theta}$$

$$= \frac{2}{\sigma_\varepsilon^2} \left\{ N.\bar{y} - N.\hat{\theta}^2 \right\}.\hat{\theta}$$

$$= \underbrace{\frac{2N}{\sigma_\varepsilon^2}.\hat{\theta}}_{C(\theta)|_{\theta=\hat{\theta}}} . \underbrace{\left\{ \bar{y} - \hat{\theta}^2 \right\}}_{(\hat{\theta}(y) - \theta)}$$

Então, a solução $\hat{\theta} = 0$ não é interessante pois não depende de $y_1, \ldots, y_N$

2.15.3 Estimação Bayesiana

No paradigma Bayesiano, θ é tratado como uma variável cuja distribuição a priori dada por $p_\theta(\theta)$. Com a realização das medidas y (i.e., vetor de medições), então, pela regra de Bayes,

$$p(\theta|y) = \frac{p(y|\theta)\cdot p_\theta(\theta)}{p_Y(y)} = \frac{p(y|\theta)\cdot p_\theta(\theta)}{\int_{-\infty}^{\infty} p(y|\theta)\, p_\theta(\theta)\, d\theta}$$

Na relação acima temos $p(y|\theta)$ que é a função de verossimilhança. $p(\theta|y)$ é chamada de densidade a posteriori. Sendo $p_Y(y)$ independente de θ, $p(\theta|y)$ é também escrita como

$$p(\theta|y) = \beta \cdot p(y|\theta)\cdot p_\theta(\theta) \propto p(y|\theta)\cdot p_\theta(\theta) \qquad \beta = \frac{1}{p_Y(y)}$$

$\propto$ símbolo "proporcional a"

Ao assumir independência entre as componentes y_n do vetor y, usa-se $p(y|\theta) = \prod_{n=1}^{N} p(y_n|\theta)$, o que facilita enormemente a determinação de $p(\theta|y)$.

No caso discreto, a regra de Bayes se escreve como

$$F(\theta_i|y_j) = \frac{F(y_j|\theta_i)\cdot F_\theta(\theta_i)}{F_Y(y_j)} = \frac{F(y_j|\theta_i)\cdot F_\theta(\theta_i)}{\sum_{K=1}^{N_\theta} F(y_j|\theta_K)\cdot F_\theta(\theta_K)}$$

Na estimação bayesiana, dois tipos de estimativas são comumente empregadas: a média condicional $E\{\theta|y\}$ e o máximo a posteriori (MAP) $\arg\max_\theta p(\theta|y)$.

2.15.3.1 Média como estimativa Bayesiana

Nesta abordagem,

$$\hat{\theta} = E\{\theta|y\} = \int_{-\infty}^{\infty} \theta \cdot p(\theta|y) \cdot d\theta$$

que pode ser determinado de forma analítica ou através de simulações $\theta_1, \ldots, \theta_m$ usando Monte Carlo:

$$\hat{\theta} \approx \frac{1}{m} \cdot \sum_{k=1}^{m} \theta_k$$

com as amostras θ_k geradas a partir de $p(\theta|y)$.

Considerando θ o verdadeiro parâmetro, temos

$$Var\{\hat{\theta}|y\} = E\{(\hat{\theta} - E\{\hat{\theta}\})^2 | y\}$$

como variância da estimativa $\hat{\theta}$. Se $E\{\hat{\theta}\} = \theta$, então

$$Var\{\hat{\theta}|y\} = E\{(\hat{\theta} - \theta)^2 | y\}$$

$$= \int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 \cdot p(\theta|y)\, d\theta$$

Sendo $Var\{\hat{\theta}|y\} \geq 0$, então esta é uma função cujo mínimo satisfaz a

$$\frac{\partial Var\{\hat{\theta}|y\}}{\partial \hat{\theta}} = 2 \cdot \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \cdot p(\theta|y)\, d\theta = 0$$

Então,

$$\hat{\theta} \cdot \underbrace{\int_{-\infty}^{\infty} p(\theta|y)\, d\theta}_{1} = \int_{-\infty}^{\infty} \theta \cdot p(\theta|y)\, d\theta$$

Assim

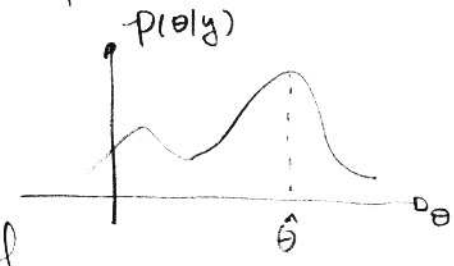
$$\hat{\theta} = \int_{-\infty}^{\infty} \theta \cdot p(\theta|y)\, d\theta = E\{\theta|y\}$$

Portanto, $\hat{\theta} = E\{\theta|y\}$ é a estimativa de variância mínima de θ. Observe que $\hat{\theta}$ é uma va pois depende de y, que é uma va.

2.15.3.2 MAP como estimativa Bayesiana

A estimativa MAP é dada por

$$\hat{\theta} = \arg\max_{\theta} p(\theta|y) \Rightarrow$$



Ou seja, $\hat{\theta}$ é o modo principal de $p(\theta|y)$ e satisfaz a

$$\left.\frac{\partial p(\theta|y)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \quad \text{e} \quad \left.\frac{\partial^2 p(\theta|y)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$$

Sendo $p(\theta|y) = \dfrac{p(y|\theta)\, p_\theta(\theta)}{p_Y(y)} = \beta \cdot p(y|\theta)\, p_\theta(\theta)$

então $\dfrac{\partial p(\theta|y)}{\partial \theta} = \beta \cdot \dfrac{\partial\{p(y|\theta)\cdot p_\theta(\theta)\}}{\partial \theta}$, com $\beta = p_Y(y)$

o que significa que $p_Y(y)$ é dispensável para calcular a estimativa MAP pois esta satisfaz a

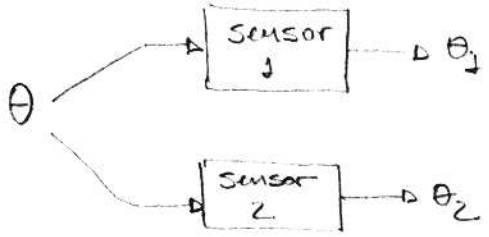
$$\frac{\partial p(\theta|y)}{\partial \theta} = 0 \Rightarrow \frac{\partial\{p(y|\theta)\cdot p_\theta(\theta)\}}{\partial \theta} = 0$$

sendo $\beta \neq 0$.

Exemplo: Fusão de dados multissensoriais

Considere dois sensores fornecendo medidas θ_1 e θ_2 da grandeza θ de acordo com os modelos

$$\left.\begin{aligned}\theta_1 &= \theta + \varepsilon_1 \\ \theta_2 &= \theta + \varepsilon_2\end{aligned}\right\}$$



com $\varepsilon_1 \sim N(0, \sigma^2_{\varepsilon_1})$ e $\varepsilon_2 \sim N(0, \sigma^2_{\varepsilon_2})$. Temos que $E\{\theta_1\} = E\{\theta_2\} = \theta$ e $\theta_1 \sim N(\theta, \sigma^2_{\varepsilon_1})$ e $\theta_2 \sim N(\theta, \sigma^2_{\varepsilon_2})$.

Se nenhum conhecimento a priori de θ é dado, então usar $\hat{\theta} = \theta_1$ como estimativa inicial permite construir esta informação a priori. Assim sendo, $\hat{\theta} \sim N(\theta_1, \sigma^2_{\varepsilon_1})$ e

$$p_{\theta_1}(\theta) = \frac{1}{(2\pi\sigma^2_{\varepsilon_1})^{1/2}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(\theta - \theta_1)^2}{\sigma^2_{\varepsilon_1}}\right\}$$

Então, $p_\theta(\theta) = p_{\theta_1}(\theta)$ vira como informação a priori.

E $$p(\theta_2|\theta) = \frac{1}{(2\pi\sigma^2_{\varepsilon_2})^{1/2}} \cdot \exp\left\{-\frac{1}{2} \frac{(\theta_2 - \theta)^2}{\sigma^2_{\varepsilon_2}}\right\}$$

Assim,

$$p(\theta_2|\theta)\, p_\theta(\theta) = \frac{1}{(4\pi^2 \sigma^2_{\varepsilon_1} \cdot \sigma^2_{\varepsilon_2})^{1/2}} \cdot \exp\left\{-\frac{1}{2} \cdot \left\{\frac{(\theta_2 - \theta)^2}{\sigma^2_{\varepsilon_2}} + \frac{(\theta - \theta_1)^2}{\sigma^2_{\varepsilon_1}}\right\}\right\}$$

como

$$\hat{\theta} = \arg\max_\theta p(\theta_2|\theta) \cdot p_\theta(\theta) = \arg\max_\theta \ln\left(p(\theta_2|\theta) \cdot p_\theta(\theta)\right)$$

$$\ln\left(p(\theta_2|\theta)\, p_\theta(\theta)\right) = -\ln\left(2\pi\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}\right) - \frac{1}{2}\cdot\left\{\frac{(\theta_2-\theta)^2}{\sigma^2_{\varepsilon_2}} + \frac{(\theta-\theta_1)^2}{\sigma^2_{\varepsilon_1}}\right\}$$

e então,

$$\left.\frac{\partial \ln\left(p(\theta_2|\theta)\cdot p_\theta(\theta)\right)}{\partial\theta}\right|_{\theta=\hat\theta} = -\frac{1}{2}\cdot\left\{\frac{2\cdot(\theta_2-\hat\theta)(-1)}{\sigma^2_{\varepsilon_2}} + \frac{2\cdot(\hat\theta-\theta_1)}{\sigma^2_{\varepsilon_1}}\right\}$$

$$= 0$$

Portanto,

$$\frac{\theta_2-\hat\theta}{\sigma^2_{\varepsilon_2}} = \frac{\hat\theta-\theta_1}{\sigma^2_{\varepsilon_1}}$$

$$\theta_2\cdot\sigma^2_{\varepsilon_1} - \hat\theta\cdot\sigma^2_{\varepsilon_1} = \hat\theta\cdot\sigma^2_{\varepsilon_2} - \theta_1\cdot\sigma^2_{\varepsilon_2}$$

$$\hat\theta\cdot\left(\sigma^2_{\varepsilon_1}+\sigma^2_{\varepsilon_2}\right) = \theta_1\cdot\sigma^2_{\varepsilon_2} + \theta_2\cdot\sigma^2_{\varepsilon_1}$$

$$\hat\theta = \frac{\sigma^2_{\varepsilon_2}}{\sigma^2_{\varepsilon_1}+\sigma^2_{\varepsilon_2}}\cdot\theta_1 + \frac{\sigma^2_{\varepsilon_1}}{\sigma^2_{\varepsilon_1}+\sigma^2_{\varepsilon_2}}\cdot\theta_2$$

Portanto, $\hat\theta$ é uma média ponderada de θ_1 e θ_2 com pesos $\alpha = \sigma^2_{\varepsilon_2}/(\sigma^2_{\varepsilon_1}+\sigma^2_{\varepsilon_2})$ e $1-\alpha$ tais que

$$\hat\theta = \alpha\cdot\theta_1 + (1-\alpha)\cdot\theta_2$$

A variância de $\hat\theta$ é dada por

$$\sigma^2_{\hat\theta} = \alpha^2\cdot\sigma^2_{\varepsilon_1} + (1-\alpha)^2\cdot\sigma^2_{\varepsilon_2}$$

Pode ser verificado que, dentre a classe de estimadores $\hat\theta = \alpha\cdot\theta_1 + (1-\alpha)\theta_2$, com $0\le\alpha\le 1$,

a variância mínima é alcançada com

$$\alpha_{MV} = \sigma^2_{\varepsilon_2} / (\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2})$$

que satisfaz a

$$\left.\frac{\partial \sigma^2_{\hat{\theta}}}{\partial \alpha}\right|_{\alpha = \alpha_{MV}} = 0.$$

E ainda, $\hat{\theta}$ pode ser escrito como

$$\hat{\theta} = \theta_1 + \underbrace{(1 - \alpha_{mv})}_{g_{mv}} \cdot (\theta_2 - \theta_1)$$

com g_{mv} sendo o ganho que resulta na variância mínima de $\sigma^2_{\hat{\theta}}$ obtido de α_{mv} e que multiplica a diferença entre θ_2 e θ_1. $(\theta_2 - \theta_1)$ é chamado de inovação, θ_2 é a medição e θ_1 é a estimativa a priori de θ. Observa-se que

$$g_{mv} = \frac{\sigma^2_{\varepsilon_1}}{\sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_1}} = \begin{cases} 0, & \text{se } \sigma^2_{\varepsilon_1} = 0 \text{ e } \sigma^2_{\varepsilon_2} \neq 0 \\ 1, & \text{se } \sigma^2_{\varepsilon_1} \neq 0 \text{ e } \sigma^2_{\varepsilon_2} = 0 \end{cases}$$

Assim, se a informação a priori é precisa ($\sigma^2_{\varepsilon_1} = 0$), temos $g_{mv} = 0$ e $\hat{\theta} = \theta_1$. Ou seja, o estimador "confia" apenas na informação a priori. Sendo a medição precisa ($\sigma^2_{\varepsilon_2} = 0$), $g_{mv} = 1$ e $\hat{\theta} = \theta_2$, prevalecendo como estimativa θ_2.

Neste exemplo específico, g_{MV} é chamado de Ganho de Kalman, pois o problema é linear, Gaussiano, e o filtro de Kalman, se aplicado, dará o mesmo resultado. Assim, nestas condições, o filtro de Kalman é um caso especial do estimador MAP Bayesiano (p/ sistemas lineares Gaussianos)
→ continua em 116a.

Exemplo: Problema não-linear (pg 105) em versão Bayesiana:

$$y = \theta^2 + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

com $\theta \in \mathbb{R}$. Dadas medições $y_1, \ldots, y_N$, foi obtido

$$p(y|\theta) = \prod_{n=1}^{N} \frac{1}{(2\pi\sigma_\varepsilon^2)^{1/2}} \cdot \exp\left\{ -\frac{1}{2} \cdot \frac{(y_n - \theta^2)^2}{\sigma_\varepsilon^2} \right\}$$

$$= (2\pi\sigma_\varepsilon^2)^{-N/2} \cdot \exp\left\{ \sum_{n=1}^{N} -\frac{1}{2} \frac{(y_n - \theta^2)^2}{\sigma_\varepsilon^2} \right\}$$

Se tivermos como conhecimento a priori $\theta \sim N(\theta_0, \sigma_{\theta_0}^2)$ então $p_\theta(\theta) = \frac{1}{(2\pi\sigma_{\theta_0}^2)^{1/2}} \cdot \exp\left\{ -\frac{1}{2} \frac{(\theta - \theta_0)^2}{\sigma_{\theta_0}^2} \right\}$

Portanto,

$$p(y|\theta) \cdot p_\theta(\theta) = (2\pi\sigma_\varepsilon^2)^{-N/2} \cdot (2\pi\sigma_{\theta_0}^2)^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \left\{ \frac{(\theta - \theta_0)^2}{\sigma_{\theta_0}^2} + \sum_{n=1}^{N} \frac{(y_n - \theta^2)^2}{\sigma_\varepsilon^2} \right\} \right\}$$

e

$$\ln\left(p(y|\theta) \cdot p_\theta(\theta)\right) = -\frac{N}{2} \cdot \ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2} \cdot \ln(2\pi\sigma_{\theta_0}^2)$$
$$- \frac{1}{2} \cdot \left\{ \frac{(\theta - \theta_0)^2}{\sigma_{\theta_0}^2} + \sum_{n=1}^{N} \frac{(y_n - \theta^2)^2}{\sigma_\varepsilon^2} \right\}$$

→ continua em 117.

A variância de $\hat{\theta}$ considerando α_{mv} é dada por

$$\sigma^2_{\hat{\theta}} = \alpha^2_{mv} \cdot \sigma^2_{\varepsilon_1} + (1-\alpha_{mv})^2 \cdot \sigma^2_{\varepsilon_2}$$

com $\alpha_{mv} = \dfrac{\sigma^2_{\varepsilon_2}}{\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2}}$, temos

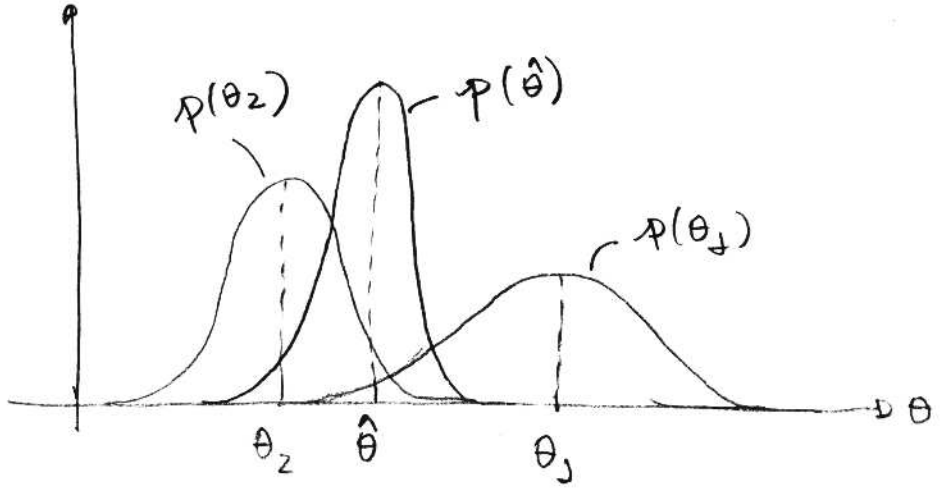
$$\sigma^2_{\hat{\theta}} = \frac{\sigma^4_{\varepsilon_2} \cdot \sigma^2_{\varepsilon_1}}{(\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2})^2} + \frac{\sigma^4_{\varepsilon_1} \cdot \sigma^2_{\varepsilon_2}}{(\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2})^2}$$

$$\sigma^2_{\hat{\theta}} = \frac{\sigma^2_{\varepsilon_1} \cdot \sigma^2_{\varepsilon_2} \cdot (\sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_1})}{(\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2})^2} = \frac{\sigma^2_{\varepsilon_1} \cdot \sigma^2_{\varepsilon_2}}{(\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2})}$$

Pode ser verificado que

$$\sigma^2_{\hat{\theta}} \leq \sigma^2_{\varepsilon_1} \quad \text{e} \quad \sigma^2_{\hat{\theta}} \leq \sigma^2_{\varepsilon_2}$$

ou seja, a incerteza associada a $\hat{\theta}$ é menor ou igual à menor das incertezas de θ_1 e θ_2 (medidas). Isso justifica o uso de técnicas estatísticas em fusão de dados sensoriais. Percebe-se que



$\in \hat{\theta}$ satisfaz a

$$\frac{2.(\hat{\theta}-\theta_0)}{\sigma^2_{\theta_0}} + \sum_{n=1}^{N} \frac{2.(y_n-\hat{\theta}^2).(-2).\hat{\theta}}{\sigma^2_\varepsilon} = 0$$

$$\frac{(\hat{\theta}-\theta_0)}{\sigma^2_{\theta_0}} - \frac{2}{\sigma^2_\varepsilon}.\left\{ \underbrace{\sum_{n=1}^{N} y_n.\hat{\theta}}_{\hat{\theta}.N\bar{y}} - \underbrace{\sum_{n=1}^{N} \hat{\theta}^3}_{\hat{\theta}^3.N} \right\} = 0$$

$$\frac{\hat{\theta}-\theta_0}{\sigma^2_{\theta_0}} - \frac{2.\hat{\theta}.N\bar{y}}{\sigma^2_\varepsilon} + \frac{2.\hat{\theta}^3.N}{\sigma^2_\varepsilon} = 0$$

$$\frac{\hat{\theta}}{\sigma^2_{\theta_0}} - \frac{2.\hat{\theta}.N.\bar{y}}{\sigma^2_\varepsilon} + \frac{2.\hat{\theta}^3 N}{\sigma^2_\varepsilon} = \frac{\theta_0}{\sigma^2_{\theta_0}}$$

$$\boxed{\hat{\theta}^3.\frac{N.2}{\sigma^2_\varepsilon} + \hat{\theta}.\left(\frac{1}{\sigma^2_{\theta_0}} - \frac{N.\bar{y}.2}{\sigma^2_\varepsilon}\right) - \frac{\theta_0}{\sigma^2_{\theta_0}} = 0}$$

Deve ser resolvida esta equação do terceiro grau.

Neste caso, apesar da complexidade, a solução para a estimação MAP é ainda tratável por ser analítica. Entretanto, por fatores diversos (e.g., multiplos modelos ou casos, multidimensionalidade de θ, etc.), a solução analítica é difícil de se obter. Nestes casos, a abordagem computacional é bem-vinda.

(ver simulação estimacao_bayesiana)

2.15.3.3 Aspectos computacionais

Dada a equação de Bayes

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$$

foram apresentados dois estimadores que consideram informação a priori: A média condicional $E\{\theta|y\}$ e o máximo a posteriori (MAP). Apesar da forma simples que $p(y|\theta)$ e $p(\theta)$ podem ter, determinar estimativas usando $p(\theta|y)$ pode ser bastante complicado. E ainda, com o aumento do número de dimensões de θ, o uso de técnicas numéricas tradicionais fica proibitivo para determinar

$$p(y) = \int_{-\infty}^{\infty} p(y|\theta) \cdot p(\theta) \cdot d\theta$$

$$E\{\theta|y\} = \int_{-\infty}^{\infty} \theta \cdot p(\theta|y)\, d\theta$$

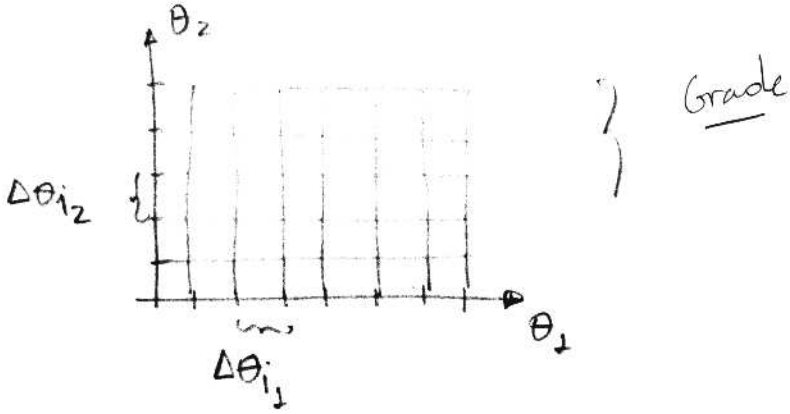
Neste sentido, algumas aproximações numéricas se fazem necessárias.

a) Discretização do espaço de parâmetro (Grid method ou lattice)

Uma abordagem clássica na determinação de $p(\theta|y)$ consiste em discretizar o espaço do parâmetro θ em $\theta_i = [\theta_{i1}, \ldots, \theta_{in}]$, com

$$\theta_{ij} - \theta_{ij-1} = \Delta\theta_{i1}, \ldots, \theta_{in} - \theta_{in-1} = \Delta\theta_{in}.$$

Por exemplo, com $n=2$:



Esta discretização somente faz sentido se os parâmetros $\theta_1, \ldots, \theta_n$ tiverem valores limitados a intervalos. Assim, $p(\theta|y)$ é estimada para cada θ na grade. Deve-se então avaliar

$$p(y, \theta_i) = p(y|\theta_i) \cdot p(\theta_i)$$

e calcular

$$p(y) = \sum_{i_1} \cdots \sum_{i_n} \overbrace{p(y|\theta_i) \cdot p(\theta_i)}^{p(y,\theta_i)} \cdot \Delta i_1 \cdots \Delta i_n$$

De posse de $p(y)$ e $p(y, \theta_i)$, temos então

$$p(\theta_i|y) = \frac{p(y, \theta_i)}{p(y)}$$

b) Amostragem de importância (Importance Sampling) (IS)

A idéia principal do IS é aproximar integrais do tipo $\int f(x)dx$ usando uma função de densidade auxiliar $g(x)$ de onde amostras são geradas:

$$\int f(x)dx = \int \left(\frac{f(x)}{g(x)}\right) \cdot g(x) \cdot dx$$

$$= \frac{1}{M} \cdot \sum_{m=1}^{M} \frac{f(x_m)}{g(x_m)}$$

com $x_m \sim g(x)$.

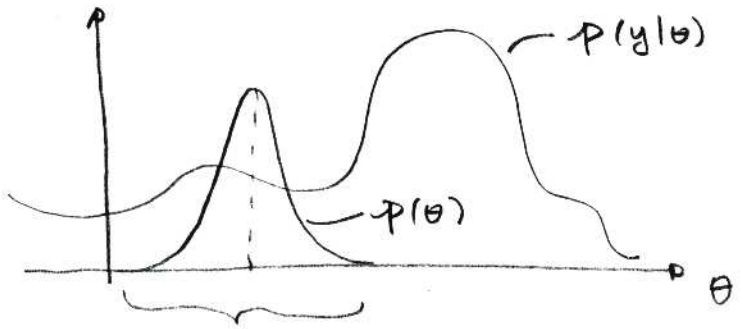
$g(x)$ é chamada de densidade de importância. De forma similar ao método aceitação/rejeição, $g(x)$ deve ter a forma mais próxima de $f(x)$, de forma que a estimativa da integral seja melhor possível.

O emprego deste método em estimação bayesiana está na avaliação das integrais

$$p(y) = \int_{-\infty}^{\infty} \underbrace{p(y|\theta) \cdot p(\theta)}_{f(\theta)} \cdot d\theta$$

$$E\{\theta|y\} = \int_{-\infty}^{\infty} \underbrace{\theta \cdot p(\theta|y)}_{f(\theta)} \cdot d\theta$$

Observa-se que para o caso de $p(y)$ for escolhida $g(\theta) = p(\theta)$, então caímos no caso da simulação de monte-carlo para aproximação de $p(y)$. No entanto, esta escolha para $g(\theta)$ pode não ser eficiente (e geralmente não é) se $p(\theta)$ não for da mesma forma de $p(y|\theta)$:



amostras ficam concentradas nesta zona.

c) Sampling-importance-resampling (SIR)

Em vez de aproximar integrais como no método IS, o método SIR gera amostras diretamente de $p(\theta|y)$. A ideia consiste em gerar amostras seguindo $p(\theta)$ e modificá-las de modo a obter amostras de $p(\theta|y)$. Sendo

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{\int_{-\infty}^{\infty} p(y|\theta) \cdot p(\theta) \cdot d\theta} = \frac{f(\theta)}{\int_{-\infty}^{\infty} f(\theta) d\theta}$$

com $f(\theta) = p(y|\theta) \cdot p(\theta)$ sendo uma função. Se existir uma constante M que satisfaz a

$$f(\theta) \leq M \cdot p(\theta)$$

então o método aceitação/rejeição pode ser usado. Observa-se que esta relação pode ser escrita como

$$p(y|\theta) \cdot p(\theta) \leq M \cdot p(\theta)$$

o que leva a $p(y|\theta) \leq M$. Esta relação é satisfeita com $\theta = \hat{\theta}_{MV}$ (estimativa de máximo de verossimilhança) tendo então que se escolher $M = p(y|\hat{\theta}_{MV})$. Neste caso, o problema de máximo de verossimilhança deve ser resolvido primeiro. Como $\int_{-\infty}^{\infty} f(\theta) d\theta \geq 0$, então

$$\underbrace{\left| \frac{f(\theta)}{\int_{-\infty}^{\infty} f(\theta) d\theta} \right|}_{p(\theta|y)} \leq \frac{M \cdot p(\theta)}{\int_{-\infty}^{\infty} f(\theta) d\theta}$$

ou seja, $f(\theta) \leq M \cdot p(\theta)$ é sempre válida para qualquer θ que satisfaça a $p(\theta|y) \leq \frac{M \cdot f(\theta)}{\int_{-\infty}^{\infty} f(\theta) d\theta}$. Como f foi definida, o método aceitação/rejeição deveria gerar uma amostra θ_i de $p(\theta)$ e aceitá-la com probabilidade

$$\frac{p(\theta|y)}{\frac{M \cdot p(\theta)}{\int_{-\infty}^{\infty} f(\theta) d\theta}}$$

que, como $p(\theta_i|y) \cdot \int_{-\infty}^{\infty} f(\theta) d\theta = f(\theta_i)$, esta probabilidade é igual a

$$\frac{f(\theta)}{M \cdot p(\theta_i)} = \frac{p(y|\theta_i) \cdot p(\theta_i)}{p(y|\hat{\theta}_{MV}) \cdot p(\theta_i)} = \frac{p(y|\theta_i)}{p(y|\hat{\theta}_{MV})}$$

Assim, amostras $\theta_i \sim p(\theta|y)$ são guardadas. Se M não for conhecido, estas amostras de $p(\theta|y)$ podem ser geradas da seguinte forma: a partir de N amostras $\theta_1, \ldots, \theta_N$ de $p(\theta)$, calcula-se probabilidades

$$q_n = \frac{w_n}{\sum_{n=1}^{N} w_n}, \text{ com } w_n = \frac{f(\theta_n)}{p(\theta_n)} = \underbrace{p(y|\theta_n)}_{\text{verossimilhança}}$$

e escolhe-se como amostra θ^* tirado das amostras $\theta_1, \ldots, \theta_N$ com probabilidade q_n. Ou seja, cada θ_n de amostra tem probabilidade q_n de ser sorteado.

Pode-se verificar que θ^* é uma amostra de $p(\theta|y)$ através da seguinte relação:

$$P_r\{\theta^* \le a\} = \sum_{n=1}^{N} f_n \cdot u(\theta_n - a) \Rightarrow \text{Distribuição discreta de } \theta_1, \dots, \theta_N$$

com $u(\theta_n - a) = \begin{cases} 1, & \text{se } \theta_n \le a \\ 0, & \text{caso contrário} \end{cases}$. Então,

$$P_r\{\theta^* \le a\} = \frac{\sum_{n=1}^{N} \omega_n \cdot u(\theta_n - a)}{\sum_{i=1}^{N} \omega_i}$$

então, como

$$\lim_{N\to\infty} \sum_{n=1}^{N} \omega_n \cdot u(\theta_n - a) = \lim_{N\to\infty} \sum_{n=1}^{N} p(y|\theta_n) \cdot \underbrace{u(\theta_n - a)}_{0 \text{ se } \theta_n > a}$$

$$= N \cdot \int_{-\infty}^{a} p(y|\theta) \cdot p(\theta) \cdot d\theta$$

e

$$\lim_{N\to\infty} \sum_{n=1}^{N} \omega_n = \lim_{N\to\infty} \sum_{n=1}^{N} p(y|\theta_n)$$

$$= N \cdot \int_{-\infty}^{\infty} p(y|\theta) \cdot p(\theta) \cdot d\theta$$

(pois θ foi gerado de $p(\theta)$)

Assim,

$$\lim_{N\to\infty} P_r\{\theta^* \le a\} = \frac{\int_{-\infty}^{a} p(y|\theta) \cdot p(\theta) \cdot d\theta}{\int_{-\infty}^{\infty} p(y|\theta) \cdot p(\theta) \cdot d\theta} = P_r\{\theta \le a | y\}$$

Então, como a densidade de $P_r\{\theta \le a|y\}$ é $p(\theta|y)$, então, com $N \to \infty$, θ^* é uma amostra de $p(\theta|y)$.

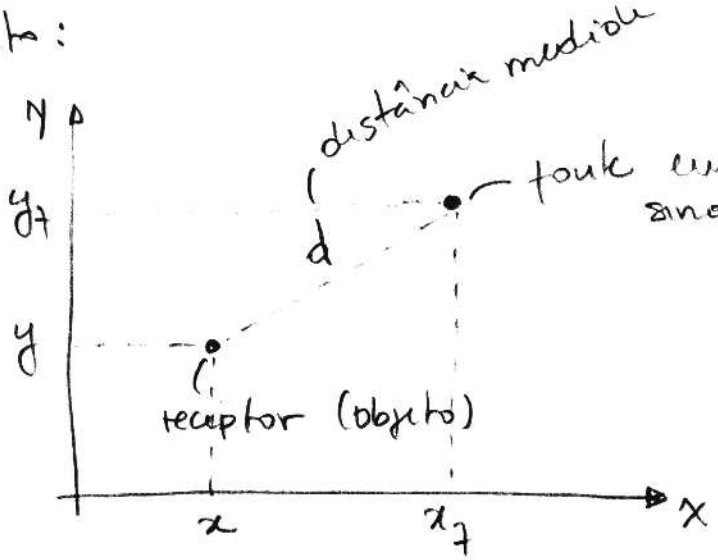
Observa-se que

$$q_n = \frac{p(y|\theta_n)}{\sum_{i=1}^{N} p(y|\theta_i)}$$

A principal vantagem do algoritmo SIR sobre o método em Grade é que este último se aplica a parâmetros θ ilimitados (e.g., intervalo de reta) e sua complexidade computacional aumenta com a redução da discretização $(\Delta\theta_{1j}, \ldots, \Delta\theta_{in})$. No SIR, o número de amostras a serem geradas de $p(\theta|y)$ é constante, sendo que este número deve ser suficientemente grande de forma que as amostras passam ser usadas de forma satisfatória para calcular $\hat{\theta}$. Além do mais, $p(\theta)$ deve descrever uma distribuição no mínimo pessimista da informação a priori. Caso contrário $p(\theta)$ não faria parte do suporte de $p(\theta|y)$, e poucas amostras utilizáveis seriam geradas.

Exemplo: localização usando RF (posição apenas)

Considere o seguinte caso: uma fonte emissora de sinais de RF é usada como "landmark" para a estimação da posição 2D (no plano) de um objeto (robô, automóvel, pessoa) usando apenas uma medida d da distância do emissor com relação ao objeto:



$$d^2 = (x-x_7)^2 + (y-y_7)^2$$

(x_7, y_7) : coordenadas cartesianas da fonte

(x, y) : " " do objeto

Define-se:

$p(\theta)$: densidade a priori da posição $\theta = (x,y)^T$ do objeto, ou ainda $p(x,y)$.

$p(d|\theta)$: densidade condicional de d dado θ ou ainda $p(d|x,y)$

Se modelarmos a distância medida d como

$$d = \left((x-x_7)^2 + (y-y_7)^2\right)^{1/2} + \varepsilon_d$$

com $\varepsilon_d \sim N(0, \sigma^2_{\varepsilon_d})$

Temos que

$$p(d|\theta) = p_{\varepsilon_d}\left(d - \left[(x-x_7)^2 + (y-y_7)^2\right]^{1/2}\right)$$

$$= \frac{1}{(2\pi\sigma^2_{\varepsilon_d})^{1/2}} \cdot \exp\left\{\frac{-1}{2\cdot\sigma^2_{\varepsilon_d}} \cdot \left(d - \sqrt{(x-x_7)^2 + (y-y_7)^2}\right)^2\right\}$$

Assim, dado $p(\overbrace{x,y}^{\theta})$ como informação a priori, então

$$p(x,y|d) \propto p(d|x,y) \cdot p(x,y)$$

(Ver simulação estimacao_bayesiana_computacional)