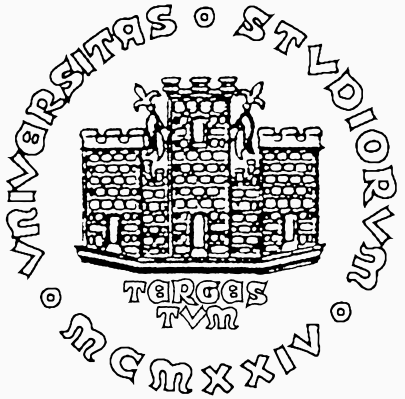
Systems Dynamics

Course ID: 267MI – Fall 2018

Thomas Parisini
Gianfranco Fenu

University of Trieste
Department of Engineering and Architecture



267MI –Fall 2018

Lecture 9
Bayes Estimation

Bayes Estimation

Considerations

- We look for an estimation method allowing to embed the possible a-priori knowledge on the unknown quantity to be estimated
- In the framework of Bayes estimation also the unknown vector is interpreted as a random vector
- The probability density function $p(\vartheta)$ in absence of observed data is the a-priori probability density function embedding the available information on ϑ before collecting the data.
- Hence, in the absence of data, the a-priori estimator could be

$$\hat{\vartheta} = E(\vartheta) = \int \vartheta \, p(\vartheta) \, d\vartheta$$

and the uncertainty $\mathrm{var}(\vartheta)$ of the estimate would be the a-priori estimate

Bayes Estimation (cont.)

- Clearly, as soon as new data are collected, the probability density function $p(\vartheta)$ changes.
- As a consequence, $E(\vartheta)$ and $\mathrm{var}(\vartheta)$ change as well.
- In particular, we expect $\mathrm{var}(\vartheta)$ to decrease
- Summing up, the basic idea is to consider a joint random experiment with respect to d and ϑ and this is the conceptual peculiarity of the Bayes estimation approach.

Bayes Estimation (cont.)

- Consider the generic estimator as function of the data

$$\hat{\vartheta} = h(d)$$

and define the cost functional

$$J[h(\cdot)] = E\left[\|\vartheta - h(d)\|^2\right]$$

- The goal is to determine an estimator $h^\circ(\cdot)$ such that $J[h(\cdot)]$ is minimised, that is we have to determine

$$h^\circ(\cdot) \; : \; E\left[\|\vartheta - h^\circ(d)\|^2\right] \leq E\left[\|\vartheta - h(d)\|^2\right] \; , \quad \forall\, h(\cdot)$$

where the expected values are computed with reference to the joint random experiment

Bayes Estimation (cont.)

- Assume for simplicity that d and ϑ are scalar:

$$E\left[\|\vartheta - h(d)\|^2\right] = E\left[\vartheta^2 - 2\vartheta\, d + h(d)^2\right]$$

and setting $f(d,\vartheta) = \vartheta^2 - 2\vartheta\, d + h(d)^2$ one gets:

$$E\left[f(d,\vartheta)\right] = \int_{x,y} f(x,y)\, p(x,y)\, dxdy$$

where x and y are the current values taken on by d and ϑ and $p(d,\vartheta)$ is the joint probability density of d and ϑ

- Recall the Bayes formula (of very general validity):

$$p(x,y) = p(y\,|x)\, p(x)$$

Bayes Estimation (cont.)

- Hence:

$$E\left[f(d,\vartheta)\right] = \int_{x,y} f(x,y)\, p(y\,|x)\, p(x)\, dxdy$$
$$= \int_x \left[\int_y f(x,y)\, p(y\,|x)\, dy \right] p(x)\, dx$$

- On the other hand, by definition one has:

$$\int_y f(x,y)\, p(y\,|x)\, dy = E\left[\, f(d,\vartheta) \,|\, d = x \,\right]$$

  and thus:

$$E\left[\, f(d,\vartheta) \,|\, d = x \,\right]$$
$$= E\left[\, \vartheta^2 \,|\, d = x \,\right] - 2\, E\left[\, \vartheta\, h(d) \,|\, d = x \,\right] + E\left[\, h(d)^2 \,|\, d = x \,\right]$$

Bayes Estimation (cont.)

- Setting $d = x$ implies that $h(d)$ becomes a deterministic quantity and hence

$$E\left[\, f(d,\vartheta) \,|\, d = x \,\right] = E\left[\, \vartheta^2 \,|\, d = x \,\right] - 2\, h(x)\, E\left[\, \vartheta \,|\, d = x \,\right] + h(x)^2$$

- Adding and subtracting $\left\{E\left[\, \vartheta \,|\, d = x \,\right]\right\}^2$ one gets (completing the squares)

$$\begin{aligned} E\left[\, f(d,\vartheta) \,|\, d = x \,\right] &= \left\{E\left[\, \vartheta \,|\, d = x \,\right]\right\}^2 - 2\, h(x)\, E\left[\, \vartheta \,|\, d = x \,\right] + h(x)^2 \\ &\quad + E\left[\, \vartheta^2 \,|\, d = x \,\right] - \left\{E\left[\, \vartheta \,|\, d = x \,\right]\right\}^2 \\ &= \left\| E\left[\, \vartheta \,|\, d = x \,\right] - h(x) \right\|^2 + E\left[\, \vartheta^2 \,|\, d = x \,\right] - \left\{E\left[\, \vartheta \,|\, d = x \,\right]\right\}^2 \end{aligned}$$

Bayes Estimation (cont.)

- Therefore:

$$
\begin{aligned}
E\left[\|\vartheta - h(d)\|^2\right] &= \int_x \left[\int_y f(x,y)\, p(y\,|x)\, dy\right] p(x)\, dx \\
&= \int_x \Big[\|E\left[\,\vartheta \,|\, d = x\,\right] - h(x)\|^2 + E\left[\,\vartheta^2 \,|\, d = x\,\right] \\
&\qquad - \left\{E\left[\,\vartheta \,|\, d = x\,\right]\right\}^2\Big]\; p(x)dx \\
&= \int_x \left[\underbrace{\|E\left[\,\vartheta \,|\, d = x\,\right] - h(x)\|^2}_{\geq 0} + \underbrace{\mathrm{var}\left[\,\vartheta \,|\, d = x\,\right]}_{\geq 0}\right] p(x)dx
\end{aligned}
$$

- Hence, one concludes that:

$$h^{\circ}(x) = E\,\left(\vartheta \,|\, d = x\right)$$

Bayes Estimation (cont.)

Optimal Bayes Estimator

The optimal Bayes estimator is the expected value conditioned to the actual observed data:

$$\hat{\vartheta} = h^{\circ}(\delta) = E\left(\vartheta \,|\, d = \delta\right)$$

where δ is the specific value taken on by d as outcome of the random experiment

Remark. The generalisation to the vector case is trivial

Bayes Estimation in the Gaussian Case

Assume that d and ϑ are marginally and jointly Gaussian random variables:

$$\begin{bmatrix} d \\ \vartheta \end{bmatrix} \sim G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_{dd} & \lambda_{d\vartheta} \\ \lambda_{\vartheta d} & \lambda_{\vartheta\vartheta} \end{bmatrix} \right)$$

and

$$p(d,\vartheta) = C \exp\left(-\frac{1}{2} \begin{bmatrix} d & \vartheta \end{bmatrix} \begin{bmatrix} \lambda_{dd} & \lambda_{d\vartheta} \\ \lambda_{\vartheta d} & \lambda_{\vartheta\vartheta} \end{bmatrix}^{-1} \begin{bmatrix} d \\ \vartheta \end{bmatrix} \right)$$

Letting $\lambda^2 = \lambda_{\vartheta\vartheta} - \lambda_{\vartheta d}^2/\lambda_{dd}$ and recalling that $\lambda_{d\vartheta} = \lambda_{\vartheta d}$ one gets:

$$\begin{aligned} \begin{bmatrix} \lambda_{dd} & \lambda_{\vartheta d} \\ \lambda_{\vartheta d} & \lambda_{\vartheta\vartheta} \end{bmatrix}^{-1} &= \frac{1}{\lambda_{dd}(\lambda_{\vartheta\vartheta} - \lambda_{\vartheta d}^2/\lambda_{dd})} \begin{bmatrix} \lambda_{\vartheta\vartheta} & -\lambda_{\vartheta d} \\ -\lambda_{\vartheta d} & \lambda_{dd} \end{bmatrix} \\ &= \frac{1}{\lambda^2} \begin{bmatrix} \lambda_{\vartheta\vartheta}/\lambda_{dd} & -\lambda_{\vartheta d}/\lambda_{dd} \\ -\lambda_{\vartheta d}/\lambda_{dd} & 1 \end{bmatrix} \end{aligned}$$

Bayes Estimation in the Gaussian Case (cont.)

Therefore:

$$\frac{1}{2}\begin{bmatrix} d & \vartheta \end{bmatrix}\begin{bmatrix} \lambda_{dd} & \lambda_{\vartheta d} \\ \lambda_{\vartheta d} & \lambda_{\vartheta\vartheta} \end{bmatrix}^{-1}\begin{bmatrix} d \\ \vartheta \end{bmatrix} = \cdots = \frac{1}{2\lambda^2}\left(\frac{\lambda_{\vartheta\vartheta}}{\lambda_{dd}}d^2 - 2\frac{\lambda_{\vartheta d}}{\lambda_{dd}}d\vartheta + \vartheta^2\right)$$

Moreover, by assumption: $p(d) = C'\exp\left(-\frac{1}{2\lambda_{dd}}\,d^2\right)$. Hence:

$$\begin{aligned} p(\vartheta \mid d) = \frac{p(d,\vartheta)}{p(d)} &= \frac{C}{C'}\exp\left[-\frac{1}{2\lambda^2}\left(\frac{\lambda_{\vartheta\vartheta}}{\lambda_{dd}}d^2 - 2\frac{\lambda_{\vartheta d}}{\lambda_{dd}}d\vartheta + \vartheta^2 - \frac{\lambda^2 d^2}{\lambda_{dd}}\right)\right] \\ &= \frac{C}{C'}\exp\left\{-\frac{1}{2\lambda^2}\left[\frac{d^2}{\lambda_{dd}}\left(\lambda_{\vartheta\vartheta} - \lambda^2\right) - 2\frac{\lambda_{\vartheta d}}{\lambda_{dd}}d\vartheta + \vartheta^2\right]\right\} \\ &= \frac{C}{C'}\exp\left[-\frac{1}{2\lambda^2}\left(\frac{\lambda_{\vartheta d}^2}{\lambda_{dd}^2}\,d^2 - 2\frac{\lambda_{\vartheta d}}{\lambda_{dd}}d\vartheta + \vartheta^2\right)\right] \\ &= \frac{C}{C'}\exp\left[-\frac{1}{2\lambda^2}\left(\vartheta - \frac{\lambda_{\vartheta d}}{\lambda_{dd}}\,d\right)^2\right] \end{aligned}$$

Bayes Estimation in the Gaussian Case (cont.)

Optimal Bayes Estimator in the Gaussian Case

$$p(\vartheta \mid d) = \frac{C}{C'} \exp\left[-\frac{1}{2\lambda^2}\left(\vartheta - \frac{\lambda_{\vartheta d}}{\lambda_{dd}}\, d\right)^2\right]$$

$p(\vartheta \mid d)$ is Gaussian with:

- Expected value: $\dfrac{\lambda_{\vartheta d}}{\lambda_{dd}}\, d$
- Variance: $\lambda^2 = \lambda_{\vartheta\vartheta} - \dfrac{\lambda_{\vartheta d}^2}{\lambda_{dd}}$

Thus, the Optimal Bayes Estimator is given by:

$$\hat{\vartheta} = h^{\circ}(x) = E\left(\vartheta \mid d = x\right) = \frac{\lambda_{\vartheta d}}{\lambda_{dd}}\, d$$

and

$$\mathrm{var}\left(\vartheta - \hat{\vartheta}\right) = E\left[(\vartheta - \hat{\vartheta})^2\right] = \lambda_{\vartheta\vartheta} - \frac{\lambda_{\vartheta d}^2}{\lambda_{dd}} = \lambda^2$$

Optimal Linear Estimator

- Let us remove the assumption that d and ϑ are marginally and jointly Gaussian random variables
- Let again $E(d^2) = \lambda_{dd}$, $E(\vartheta^2) = \lambda_{\vartheta\vartheta}$, $E(\vartheta d) = \lambda_{\vartheta d}$
- Impose that the estimator takes on a linear structure:

$$\hat{\vartheta} = \alpha d + \beta$$

  where α and β are suitable parameters to be determined.
- Introduce the cost function:

$$J = E\left[\left(\vartheta - \hat{\vartheta}\right)^2\right] = E\left[(\vartheta - \alpha\, d - \beta)^2\right]$$

Optimal Linear Estimator (cont.)

Thus, one gets:

$$\begin{aligned} J &= E\left(\vartheta^2 + \alpha^2 d^2 + \beta^2 - 2\alpha\vartheta d - 2\beta\vartheta + 2\alpha\beta d\right) \\ &= \lambda_{\vartheta\vartheta} + \alpha^2\lambda_{dd} + \beta^2 - 2\alpha\lambda_{\vartheta d} - 2\beta E(\vartheta) + 2\alpha\beta E(d) \end{aligned}$$

Hence:

$$\begin{cases} \dfrac{\partial J}{\partial \alpha} = 2\alpha\lambda_{dd} - 2\lambda_{\vartheta d} \quad \Longrightarrow \quad \alpha = \dfrac{\lambda_{\vartheta d}}{\lambda_{dd}} \\ \dfrac{\partial J}{\partial \beta} = 2\beta \quad \Longrightarrow \quad \beta = 0 \end{cases}$$

thus getting the Optimal Linear Estimator:

$$\hat{\vartheta} = \frac{\lambda_{\vartheta d}}{\lambda_{dd}}\, d$$

Its variance is given by:

$$\mathrm{var}\,(\vartheta - \hat{\vartheta}) = E\left[(\vartheta - \hat{\vartheta})^2\right] = \lambda_{\vartheta\vartheta} + \alpha^2\lambda_{dd} + \beta^2 - 2\alpha\lambda_{\vartheta d} = \cdots = \lambda^2$$

Optimal Linear Estimator (cont.)

Remarks:

- The optimal linear estimator is formally equal to the Bayes one.
- If the Gaussian assumption on the random variables holds, then the optimal linear estimator actually is the best possible in the minimum variance sense
- If the Gaussian assumption on the random variables does not hold, then the linear estimator is sub-optimal, but still it is the best estimator constrained to take on a linear structure in the case in which no further assumptions are introduced on the probabilistic characteristics of the random variables

Bayes Estimation: Generalisations

- If $E(d) = d_m\,,\ E(\vartheta) = \vartheta_m$, then:

$$\begin{cases} \hat{\vartheta} = \vartheta_m + \dfrac{\lambda_{\vartheta d}}{\lambda_{dd}}\,(d - d_m) \\ \mathrm{var}\,(\vartheta - \hat{\vartheta}) = \lambda_{\vartheta\vartheta} - \dfrac{\lambda_{\vartheta d}^2}{\lambda_{dd}} \end{cases}$$

- If d and ϑ are vectors with $E(d) = d_m\,,\ E(\vartheta) = \vartheta_m$ and

$$\mathrm{var}\left(\begin{bmatrix} d \\ \vartheta \end{bmatrix}\right) = \begin{bmatrix} \Lambda_{dd} & \Lambda_{d\vartheta} \\ \Lambda_{\vartheta d} & \Lambda_{\vartheta\vartheta} \end{bmatrix} \qquad \Lambda_{d\vartheta} = \Lambda_{\vartheta d}^{\top}$$

Then:

$$\begin{cases} \hat{\vartheta} = \vartheta_m + \Lambda_{\vartheta d}\,{\Lambda_{dd}}^{-1}\,(d - d_m) \\ \mathrm{var}\,(\vartheta - \hat{\vartheta}) = \Lambda_{\vartheta\vartheta} - \Lambda_{\vartheta d}\,{\Lambda_{dd}}^{-1}\Lambda_{d\vartheta} \end{cases}$$

Bayes Estimation: Interpretations and Remarks

- Consider for simplicity the Bayes estimator in the case:

$$\hat{\vartheta} = \vartheta_m + \frac{\lambda_{\vartheta d}}{\lambda_{dd}} (d - d_m)$$

Then:

- $\vartheta_m = E(\vartheta)$ is the a priori estimate: in case of no availability of observations , it is the "more reasonable" estimate. In this case, we have:

$$\text{var}\,(\vartheta - \hat{\vartheta}) = \lambda_{\vartheta\vartheta} = \text{var}\,(\vartheta)$$

- Instead, when observations are available, we have:

$$\hat{\vartheta} = \underbrace{\vartheta_m}_{\text{a-priori estimate}} + \underbrace{\frac{\lambda_{\vartheta d}}{\lambda_{dd}} (d - d_m)}_{\text{correction due to the observation}}$$

Bayes Estimation: Interpretations and Remarks (cont.)

- Clearly:
  - If $\lambda_{\vartheta d} = 0$ then $\hat{\vartheta} = \vartheta_m$ and this is correct: it means that the data observation d is uncorrelated with ϑ and hence it does not convey useful information for the estimate: the a-posteriori estimate coincides with the a-priori one.
  - If $\lambda_{\vartheta d} \neq 0$ then the estimate is corrected on the basis of the observed data:
    - If $\lambda_{\vartheta d} > 0$ then $\hat{\vartheta} - \vartheta_m$ and $d - d_m$ in the average keep the same sign and the correction is more likely to keep the same sign as well
    - If $\lambda_{\vartheta d} < 0$ then $\hat{\vartheta} - \vartheta_m$ and $d - d_m$ in the average have a different sign and the correction is more likely to change the same sign as well

Bayes Estimation: Interpretations and Remarks (cont.)

- It also very important to enhance the role played by the variance λ_{dd} that "quantifies" the degree of uncertainty of the observed data:

$$\hat{\vartheta} = \vartheta_m + \frac{\lambda_{\vartheta d}}{\lambda_{dd}} (d - d_m)$$

  Hence: the larger λ_{dd}, the smaller the applied correction, that is, the update is "more cautious"

- Moreover:

$$\text{var}\left(\vartheta - \hat{\vartheta}\right) = \lambda_{\vartheta\vartheta} - \frac{\lambda_{\vartheta d}^2}{\lambda_{dd}} = \lambda_{\vartheta\vartheta} \left(1 - \frac{\lambda_{\vartheta d}^2}{\lambda_{\vartheta\vartheta}\lambda_{dd}}\right)$$

  and thus $\text{var}\left(\vartheta - \hat{\vartheta}\right) \leq \text{var}\left(\vartheta\right)$ and

$$\text{var}\left(\vartheta - \hat{\vartheta}\right) < \text{var}\left(\vartheta\right) \text{ if } \lambda_{\vartheta d} \neq 0$$

  The estimate cannot but improve whenever the observed data convey useful information

Bayes Estimation: Geometric Interpretation

- Assume that d and ϑ are marginally and jointly Gaussian random variables:

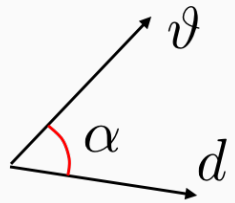
$$\begin{bmatrix} d \\ \vartheta \end{bmatrix} \sim G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_{dd} & \lambda_{d\vartheta} \\ \lambda_{\vartheta d} & \lambda_{\vartheta\vartheta} \end{bmatrix} \right)$$

  Hence d and ϑ can be interpreted as vectors in a vector space
- Define the scalar product $(\vartheta, d) = E(\vartheta \cdot d)$
- The usual properties of vector spaces equipped with scalar product hold true. In particular:

$$\|\vartheta\| = \sqrt{(\vartheta, \vartheta)}$$

$$\|d\| = \sqrt{(d, d)}$$

$$(\vartheta, d) = \|\vartheta\| \, \|d\| \, \cos\alpha$$



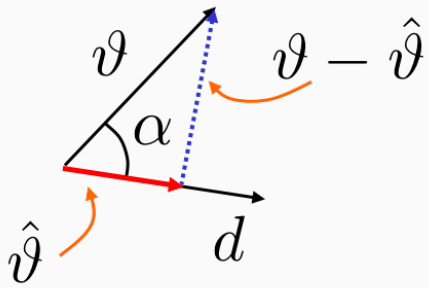
- Uncorrelated random variables: orthogonal vectors

Bayes Estimation: Geometric Interpretation

- Now:

$$\hat{\vartheta} = \frac{\lambda_{\vartheta d}}{\lambda_{dd}}\, d = \frac{E(\vartheta \cdot d)}{E(d \cdot d)}\, d = \frac{(\vartheta, d)}{\|d\|^2}\, d = \frac{(\vartheta, d)}{\|d\|^2}\, \frac{\|\vartheta\|}{\|\vartheta\|}\, d$$
$$= \frac{(\vartheta, d)}{\|\vartheta\|\|d\|}\, \|\vartheta\|\, \frac{d}{\|d\|} = \|\vartheta\|\, \cos\alpha\, \frac{d}{\|d\|}$$

The optimal estimate $\hat{\vartheta}$ is the projection of ϑ on the data vector d



- Consider the vector $\vartheta - \hat{\vartheta}$. It follows that:

$$\|\vartheta - \hat{\vartheta}\|^2 = \|\vartheta\|^2 - \|\hat{\vartheta}\|^2 = \|\vartheta\|^2 - \|\vartheta\|^2\, (\cos\alpha)^2$$
$$= \lambda_{\vartheta\vartheta} - \lambda_{\vartheta\vartheta}\, \frac{\lambda_{\vartheta d}^2}{\lambda_{dd}\lambda_{\vartheta\vartheta}} = \lambda_{\vartheta\vartheta} - \frac{\lambda_{\vartheta d}^2}{\lambda_{dd}}$$

The square of the length of vector $\vartheta - \hat{\vartheta}$ is the variance of the estimation error and is minimal.

267MI –Fall 2018

Lecture 9
Bayes Estimation

END