POLITECNICO DI MILANO

Introduction to Bayesian estimation

Marco Lovera

Dipartimento di Scienze e Tecnologie Aerospaziali, Politecnico di Milano

Conditional probability

Consider a random experiment defined by {Ω, **C**, P} and study the probabilities of two events *A* and *C*.

The conditional probability of *A* given *C* is defined as

$$P(A \backslash C) = \frac{P(A \cap C)}{P(C)}$$

Conditional probability

$$P(A\backslash C) = \frac{P(A \cap C)}{P(C)}$$

Example: rolling a dice.

Ω={1,2,3,4,5,6}

C= all subsets of Ω

Consider

A={1,2,3,5} and C={2,4,6}.

Clearly P(A)=4/6=2/3 and P(C)=3/6=1/2.

Conditional probability

$$P(A\backslash C) = \frac{P(A \cap C)}{P(C)}$$

A∩C= {2}, so P(A∩C)=1/6.

Therefore

P(A\C)=1/3.

If now we *fix* C and consider the function

$$P(\cdot \backslash C)$$

defined in **C** and taking values in [0,1], we have defined the probability of any event in **C** given event C.

It has to be checked that this function is a well-defined probability function, *i.e.*, it satisfies the properties defined earlier on.

P is a function mapping C to the [0, 1] interval, satisfying:

- P(Ω) = 1: $P(\Omega \backslash C) = \frac{P(\Omega \cap C)}{P(C)} = \frac{P(C)}{P(C)} = 1$
- If for N < ∞ events $A_1, A_2, \ldots, A_N \in$ **C**, and

$$A_i \bigcap A_j = 0, \quad \forall i, j$$

then

$$P(\bigcup_i A_i) = \sum_i P(A_i)$$

The second property holds as we have the following.

$$P(\bigcup_i A_i \backslash C) = \frac{P(\bigcup_i A_i \cap C)}{P(C)} = \frac{P(\bigcup_i (A_i \cap C))}{P(C)} =$$

$$= \frac{\sum_i P((A_i \cap C))}{P(C)} = \sum_i P(A_i \backslash C)$$

Constrained random experiment

We can now consider a *constrained* random experiment defined by

$\{\Omega, \mathbf{C}, P(\cdot \backslash C)\}$

as a random experiment constrained to the event C.

Partition of Ω

A partition of Ω is defined as a set

$$\Pi = \{C_1, C_2, \ldots, C_n\}, \quad C_i \subseteq \Omega$$

with the following properties:

- The sets C_i are all disjoint
- $\cup_i C_i = \Omega$.

Total probability theorem

Given a random experiment and a partition Π such that

$$\Pi \subseteq \mathbf{C}$$

and

$$P(C_i) \neq 0$$

then we have

$$P(A) = \sum_i P(A \backslash C_i) P(C_i) \quad \forall A \in \mathbf{C}$$

Proof: A can be written as

$$A = A \cap \Omega = A \cap (\cup_i C_i) = \cup_i (A \cap C_i)$$

so in terms of probabilities

$$P(A) = P(\cup_i (A \cap C_i)) = \sum_i P(A \cap C_i) = \sum_i P(A \backslash C_i) P(C_i)$$

Bayes Theorem

For two events A and $B \in \mathbf{C}$ with P(A), P(B) $\neq$ 0 it holds that

$$P(A \backslash B) = \frac{P(B \backslash A) P(A)}{P(B)}$$

Proof: multiply both sides by P(B) to get P(A $\cap$ B) on both sides of the equation.

Extension of Bayes Theorem

Let

$$\Pi = \{A_1, A_2, \ldots, A_n\}, \quad A_i \subseteq \mathbf{C}$$

a partition of Ω and consider an event B $\in$ **C**.

Then

$$P(A_i \backslash B) = \frac{P(B \backslash A_i) P(A_i)}{\sum_i P(B \backslash A_i) P(A_i)}.$$

Usual nomenclature:

- $P(A_i)$: *a priori* probability
- $P(A_i \backslash B)$: *a posteriori* probability

with respect to the conditioning to *B*.

Independent events

Two events A and B $\in$ **C** are called *independent* if and only if

$$P(A \cap B) = P(A)P(B)$$

Clearly for independent events we have, in terms of conditional probabilities

$$P(A \backslash B) = P(A)$$

$$P(B \backslash A) = P(B)$$

The above ideas can lead to the definition of conditional distributions and conditional densities, as follows.

Consider a random experiment and a random variable *v* defined on it.

Then pick an event $C \in \mathbf{C}$: $P(C) \neq 0$.

Then the distribution function for *v* conditional to C is defined as the distribution function for the constrained experiment.

Conditional distribution function

Consider the random experiment {Ω, **C**, P(·\C)} and random variable *v*, then the conditional distribution is

$$F(q \backslash C) = \frac{P(v \leq q, s \in C)}{P(C)}, \qquad \forall q \in \bar{\mathbb{R}}$$

where we can write equivalently

$$P(v \leq q, s \in C) = P(\phi^{-1}([-\infty, q]) \cap C)$$

Conditional probability density function

A conditional probability density function for a given conditional distribution can be defined as

$$f(q \backslash C) = \frac{dF(q \backslash C)}{dq}$$

Total probability theorem (for distributions)

Consider a partition

$$\Pi = \{C_1, C_2, \ldots, C_n\}, \quad C_i \subseteq \mathbf{C}$$

such that $P(C_i) \neq 0 \ \forall \ i$.

Then

$$F(q) = \sum_i F(q \backslash C_i) P(C_i), \quad \forall q \in \bar{\mathbb{R}}$$

If the conditioning event is given by

$$C = \phi^{-1}([-\infty, r]), \quad r \in \bar{\mathbb{R}}$$

then by definition

$$F(q \backslash C) = \frac{P(v \leq q, v \leq r)}{P(v \leq r)} = \frac{P(v \leq q, v \leq r)}{F(r)}$$

But clearly $P(v \leq q, v \leq r) = P(v \leq \min(q, r))$ so

$$F(q \backslash C) = \begin{cases} \frac{F(q)}{F(r)} & q \leq r \\ 1 & q > r \end{cases}$$

As a consequence, if

$$F(q \backslash C) = \begin{cases} \frac{F(q)}{F(r)} & q \leq r \\ 1 & q > r \end{cases}$$

then in terms of densities we have

$$f(q \backslash C) = \frac{dF(q \backslash C)}{dq} = \begin{cases} \frac{f(q)}{F(r)} & q \leq r \\ 0 & q > r \end{cases}$$

or equivalently

$$f(q \backslash C) = \frac{dF(q \backslash C)}{dq} = \begin{cases} \frac{f(q)}{\int_{-\infty}^{r} f(w)dw} & q \leq r \\ 0 & q > r \end{cases}$$

More on conditional distributions and densities

For a generic conditioning event *E* we have the conditional density

$$f(q \backslash E) = \begin{cases} \dfrac{f(q)}{\int_E f(w)dw} & q \notin E \\ 0 & q \in E \end{cases}$$

and the corresponding distribution

$$F(q \backslash v \in E) = \int_{-\infty}^{q} f(r \backslash v \in E)dr.$$

Conditional expectation

Given a real random variable *v* and the conditional density function f(q\C) the conditional expectation of *v* given C is defined as

$$E[v\backslash C] = \int_{-\infty}^{+\infty} q f(q\backslash C) dq.$$

Furthermore, if C is defined on *v*, we have

$$E[v\backslash v \in E] = \int_{-\infty}^{+\infty} q f(q\backslash v \in E) dq = \int_E q f(q\backslash v \in E) dq = \\ = \frac{\int_E q f(q) dq}{\int_E f(q) dq}.$$

Consider the random experiment $\{\Omega$, **C**, $P(\cdot\backslash C)\}$ and a vector random variable *v*, then the conditional distribution is

$$F(q\backslash C) = \frac{P(v_1 \leq q_1, \dots, v_n \leq q_n, s \in C)}{P(C)}, \qquad \forall q \in \bar{\mathbb{R}}^n$$

where we can write equivalently

$$P(v_1 \leq q_1, \dots, v_n \leq q_n, s \in C) =$$
$$= P(\phi^{-1}(v_1 \leq q_1, \dots, v_n \leq q_n) \cap C)$$

Vector conditional density

Similarly, for the conditional density function we get

$$f(q_1,\ldots,q_n\backslash C) = \frac{\partial F(q_1,\ldots,q_n\backslash C)}{\partial q_1 \ldots \partial q_n}.$$

and if the event C is defined on v as v ∈ E we get

$$f(q_1,\ldots,q_n\backslash C) = \begin{cases} \frac{f(q_1,\ldots,q_n)}{\int_E f(q_1,\ldots,q_n)dq_1,\ldots,dq_n} & q \notin E \\ 0 & q \in E \end{cases}$$

Vector conditional density: a particular case

What if the conditioning event corresponds to a line?

We get a conditional density given by (*n*=2 case)

$$f(q_1 \backslash v_2 = q_2) = \frac{f(q_1, q_2)}{f(q_2)}$$

At the level of vector conditional densities they can be stated as

$$f(q_1) = \int_{-\infty}^{+\infty} f(q_1 \backslash q_2) f(q_2) dq_2$$

$$f(q_1 \backslash q_2) = \frac{f(q_2 \backslash q_1) f(q_1)}{f(q_2)}$$

Bayesian estimation

The basic estimation problem can be formulated as follows.

- We have two random variables θ and d:
  - d is the observed variable
  - θ is the unknown we want to estimate.
- The value of the two variables is defined by a *joint* random experiment,
- We want to estimate the value of θ given a sample x *of d.*
- To solve the problem we need prior knowledge about the joint probability density function of the two variables, which will be defined in the following.

We define an estimator for *θ* as a function *h*(*d*).

Our problem is to find the estimator $h^o(d)$ such that

$$E[(\theta - h^o(d))^2] \leq E[(\theta - h(d))^2], \quad \forall h(\cdot).$$

The solution to the problem is given by the following

Theorem: function $h^o(\cdot)$ is given by

$$h^o(d) = E[\theta \backslash d = x].$$

Proof.

Let $E[(\theta - h(d))^2] = E[g(d,\theta)]$ and denote the joint probability density function of *θ* and *d* as

$$f(q_1, q_2).$$

Then $E[g(d,\theta)]$ can be written explicitly as

$$E[g(d,\theta)] = \int\int g(q_1, q_2) f(q_1, q_2) dq_1 dq_2$$

where

- q_1 is the running variable for *d*
- q_2 is the running variable for *θ*.

Recall now that

$$f(q_1, q_2) = f(q_2 \backslash q_1) f(q_1)$$

therefore substituting we have

$$E[g(d,\theta)] = \int\int g(q_1, q_2) f(q_1, q_2) dq_1 dq_2 =$$
$$= \int\int [g(q_1, q_2) f(q_2 \backslash q_1) dq_2] f(q_1) dq_1 .$$

The inner integral is the conditional expectation of $g(d,\theta)$ given d, so

$$\int g(q_1, q_2) f(q_2 \backslash q_1) dq_2 = E[g(d,\theta) \backslash d = q_1]$$

which in turn can be computed explicitly.

Recall now that

$$E[g(d,\theta)\backslash d = q_1] = E[(\theta - h(d))^2 \backslash d = q_1]$$

Expanding the square, $E[(\theta - h(d))^2 \backslash d = q_1]$ becomes

$$E[\theta^2 \backslash d = q_1] + E[-2\theta h(d) \backslash d = q_1] + E[h(d)^2 \backslash d = q_1]$$

and recalling that

$$E[f(w)\backslash w = z] = f(z)$$

we get

$$E[(\theta - h(d))^2 \backslash d = q_1] = E[\theta^2 \backslash d = q_1] - 2h(q_1)E[\theta \backslash d = q_1] + h(q_1)^2.$$

Finally, completing the square we get

$$E[\theta^2 \backslash d = q_1] - 2h(q_1)E[\theta \backslash d = q_1] + h(q_1)^2 \pm E[\theta \backslash d = q_1]^2$$

and

$$E[(\theta - h(d))^2 \backslash d = q_1] = (E[\theta \backslash d = q_1] - h(q_1))^2 + \text{terms indep. from h.}$$

So our performance criterion becomes

$$E[(\theta - h(d))^2] = \int (E[\theta \backslash d = q_1] - h(q_1))^2 f(q_1) dq_1 + \ldots$$

which is clearly minimised by

$$h^o(d) = E[\theta \backslash d = x].$$

Consider now the particular case in which *θ* and *d* are scalar and jointly Gaussian:

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_{dd} & \lambda_{d\theta} \\ \lambda_{\theta d} & \lambda_{\theta\theta} \end{bmatrix}\right).$$

The conditional density of *θ* given *d* is given by

$$f(\theta \backslash d) = \frac{f(d,\theta)}{f(d)}$$

where

$$f(d) = c_1 e^{-\frac{1}{2}\frac{d^2}{\lambda_{dd}}}$$

$$f(d,\theta) = c_2 e^{-\frac{1}{2}\begin{bmatrix} d & \theta \end{bmatrix}\begin{bmatrix} \lambda_{dd} & \lambda_{d\theta} \\ \lambda_{\theta d} & \lambda_{\theta\theta} \end{bmatrix}^{-1}\begin{bmatrix} d \\ \theta \end{bmatrix}}.$$

Bayesian estimation: the scalar Gaussian case

The inverse of the covariance is given by

$$\begin{bmatrix} \lambda_{dd} & \lambda_{d\theta} \\ \lambda_{\theta d} & \lambda_{\theta\theta} \end{bmatrix}^{-1} = \frac{1}{\lambda_{dd}\lambda_{\theta\theta} - \lambda_{\theta d}^2} \begin{bmatrix} \lambda_{\theta\theta} & -\lambda_{d\theta} \\ -\lambda_{\theta d} & \lambda_{dd} \end{bmatrix}$$

or equivalently

$$\begin{bmatrix} \lambda_{dd} & \lambda_{d\theta} \\ \lambda_{\theta d} & \lambda_{\theta\theta} \end{bmatrix}^{-1} = \frac{1}{\lambda^2} \begin{bmatrix} \frac{\lambda_{\theta\theta}}{\lambda_{dd}} & -\frac{\lambda_{d\theta}}{\lambda_{dd}} \\ -\frac{\lambda_{\theta d}}{\lambda_{dd}} & 1 \end{bmatrix}$$

where ($\lambda_{d\theta} = \lambda_{\theta d}$) $\lambda^2 = \lambda_{\theta\theta} - \frac{\lambda_{\theta d}^2}{\lambda_{dd}}.$

We also have

$$f(d) = c_1 e^{-\frac{1}{2}\frac{d^2}{\lambda_{dd}}} = c_1 e^{-\frac{1}{2\lambda^2}\frac{d^2\lambda^2}{\lambda_{dd}}} = c_1 e^{-\frac{1}{2\lambda^2}\left(\frac{\lambda_{\theta\theta}}{\lambda_{dd}} - \frac{\lambda_{\theta d}^2}{\lambda_{dd}^2}\right)d^2}$$

Substituting in the joint density we get

$$f(d,\theta) = c_2 e^{-\frac{1}{2}\begin{bmatrix} d & \theta \end{bmatrix} \begin{bmatrix} \lambda_{dd} & \lambda_{d\theta} \\ \lambda_{\theta d} & \lambda_{\theta\theta} \end{bmatrix}^{-1} \begin{bmatrix} d \\ \theta \end{bmatrix}} = c_2 e^{-\frac{1}{2\lambda^2}(\frac{\lambda_{\theta\theta}}{\lambda_{dd}}d^2 - 2\frac{\lambda_{\theta d}}{\lambda_{dd}}\theta d + \theta^2)}$$

The conditional density of *θ* given *d* is given by

$$f(\theta \backslash d) = \frac{f(d,\theta)}{f(d)} = c_3 e^{-\frac{1}{2\lambda^2}(\cancel{\frac{\lambda_{\theta\theta}}{\lambda_{dd}}}d^2 - 2\frac{\lambda_{\theta d}}{\lambda_{dd}}\theta d + \theta^2 - \cancel{\frac{\lambda_{\theta\theta}}{\lambda_{dd}}}d^2 + \frac{\lambda_{\theta d}^2}{\lambda_{dd}^2}d^2)}$$

$$f(\theta \backslash d) = c_3 e^{-\frac{1}{2\lambda^2}(\theta - \frac{\lambda_{\theta d}}{\lambda_{dd}}d)^2}$$

which is a Gaussian:

$$f(\theta \backslash d) \approx (\frac{\lambda_{\theta d}}{\lambda_{dd}}d, \lambda^2).$$

Therefore the Bayesian estimator for θ is given by

$$\widehat{\theta} = E[\theta \backslash d] = \frac{\lambda_{\theta d}}{\lambda_{dd}} d.$$

It is easy to verify that $\mathrm{Var}[\widehat{\theta} - \theta] = \lambda^2$:

$$\begin{aligned}
\mathrm{Var}[\widehat{\theta} - \theta] = \mathrm{Var}[\frac{\lambda_{\theta d}}{\lambda_{dd}} d - \theta] = \\
= \frac{\lambda_{\theta d}^2}{\lambda_{dd}^2} \mathrm{Var}[d] + \mathrm{Var}[\theta] - 2\frac{\lambda_{\theta d}}{\lambda_{dd}} \mathrm{E}[\theta d] = \\
= \frac{\lambda_{\theta d}^2}{\lambda_{dd}^2} \lambda_{dd} + \lambda_{\theta\theta} - 2\frac{\lambda_{\theta d}}{\lambda_{dd}} \lambda_{\theta d} = \lambda^2.
\end{aligned}$$

In Bayesian estimation we use *a priori* knowledge to model the unknown and the measured variable, so we can distiguish between

- The *a priori* estimate, which we could make based on the prior knowledge alone. In our case:

$$\widehat{\theta} = E[\theta] = 0, \quad \mathrm{Var}[\widehat{\theta} - \theta] = \lambda_{\theta\theta}$$

- The *a posteriori* estimate, which we can make exploting also the measurement of *d*. In our case:

$$\widehat{\theta} = E[\theta \backslash d] = \frac{\lambda_{\theta d}}{\lambda_{dd}} d, \quad \mathrm{Var}[\widehat{\theta} - \theta] = \lambda^2$$

- Note that by construction $\lambda^2 = \lambda_{\theta\theta} - \frac{\lambda_{\theta d}^2}{\lambda_{dd}} \leq \lambda_{\theta\theta}.$

Note further that

$$\lim_{\lambda_{dd}\to\infty} \lambda^2 = \lambda_{\theta\theta}$$

so if the measurement is poorly informative then the *a posteriori* estimate converges to the *a priori* one.

Finally, the variance can be written in terms of the correlation coefficient between *θ* and *d*:

$$\lambda^2 = \lambda_{\theta\theta} - \frac{\lambda_{\theta d}^2}{\lambda_{dd}} = \lambda_{\theta\theta}(1-\rho^2), \quad \rho = \frac{\lambda_{\theta d}}{\sqrt{\lambda_{\theta\theta}\lambda_{dd}}}.$$

Bayesian estimation: the scalar Gaussian case

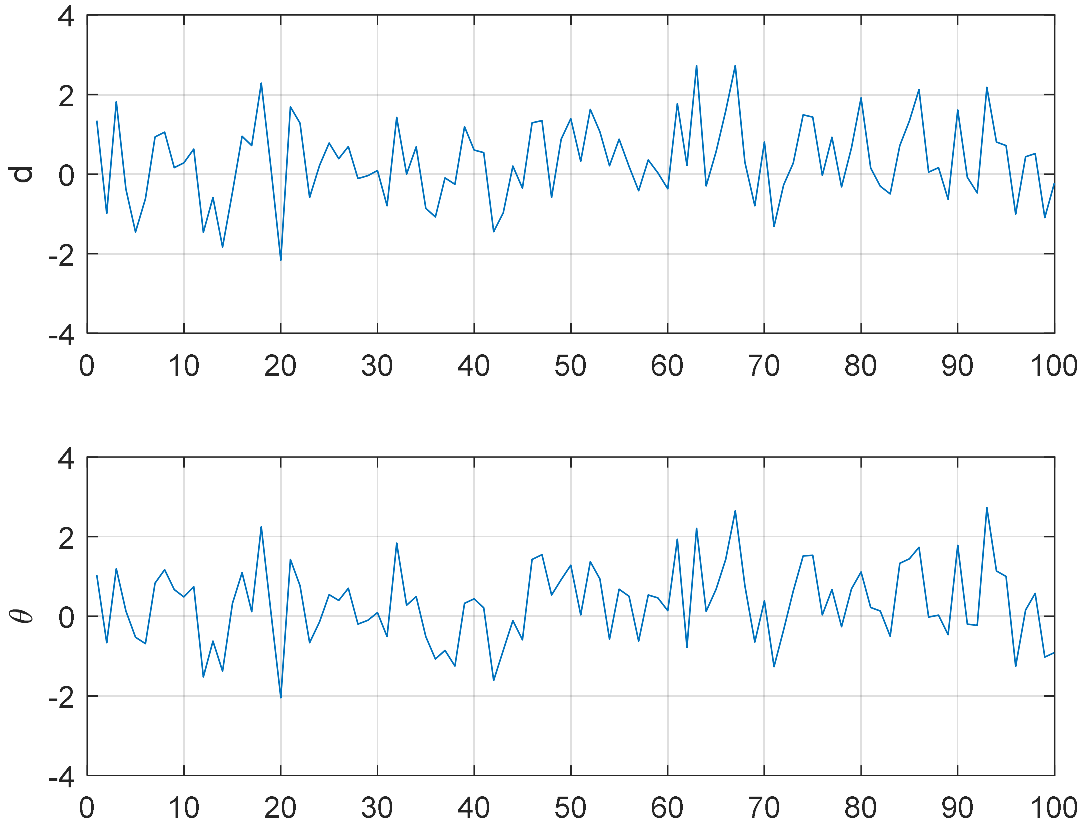
It is interesting to look at the extreme cases:

If $\rho = 0$ then $\lambda^2 = \lambda_{\theta\theta}(1-\rho^2) = \lambda_{\theta\theta}$ so *d* does not provide any information on *θ*.

If $\rho = \pm 1$ then $\lambda^2 = \lambda_{\theta\theta}(1-\rho^2) = 0$ so measuring *d* is equivalent to measuring *θ*.

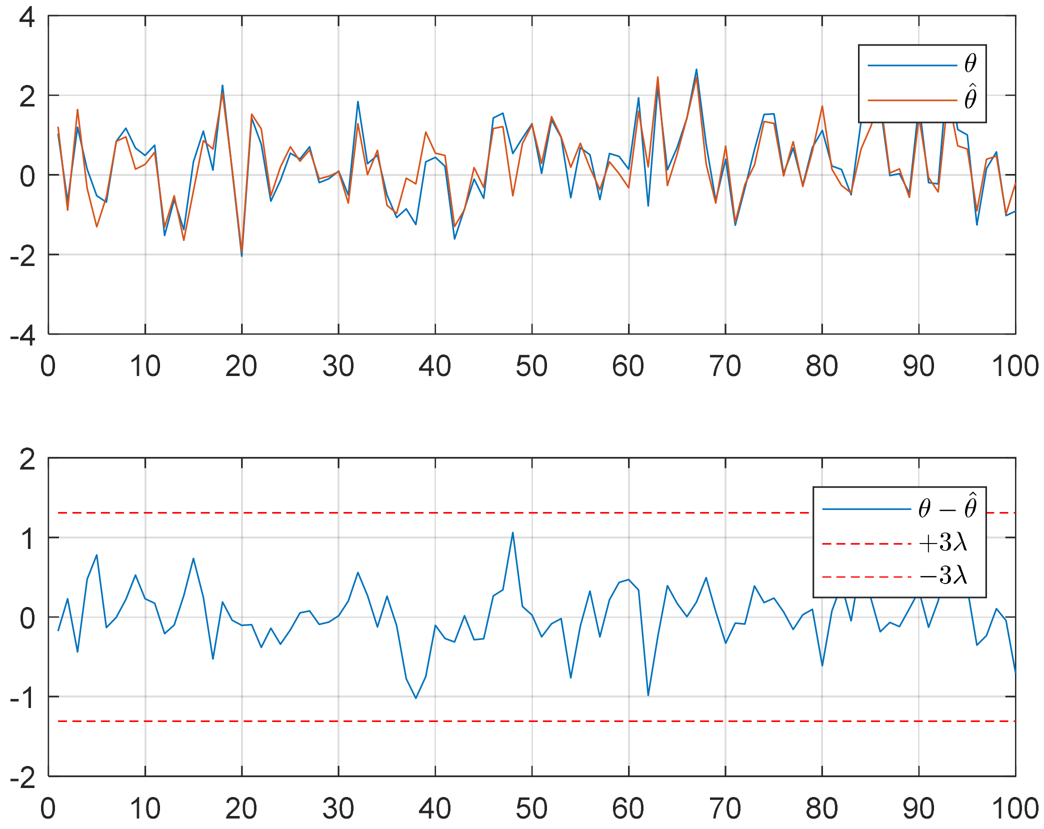
Example: high positive correlation

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$



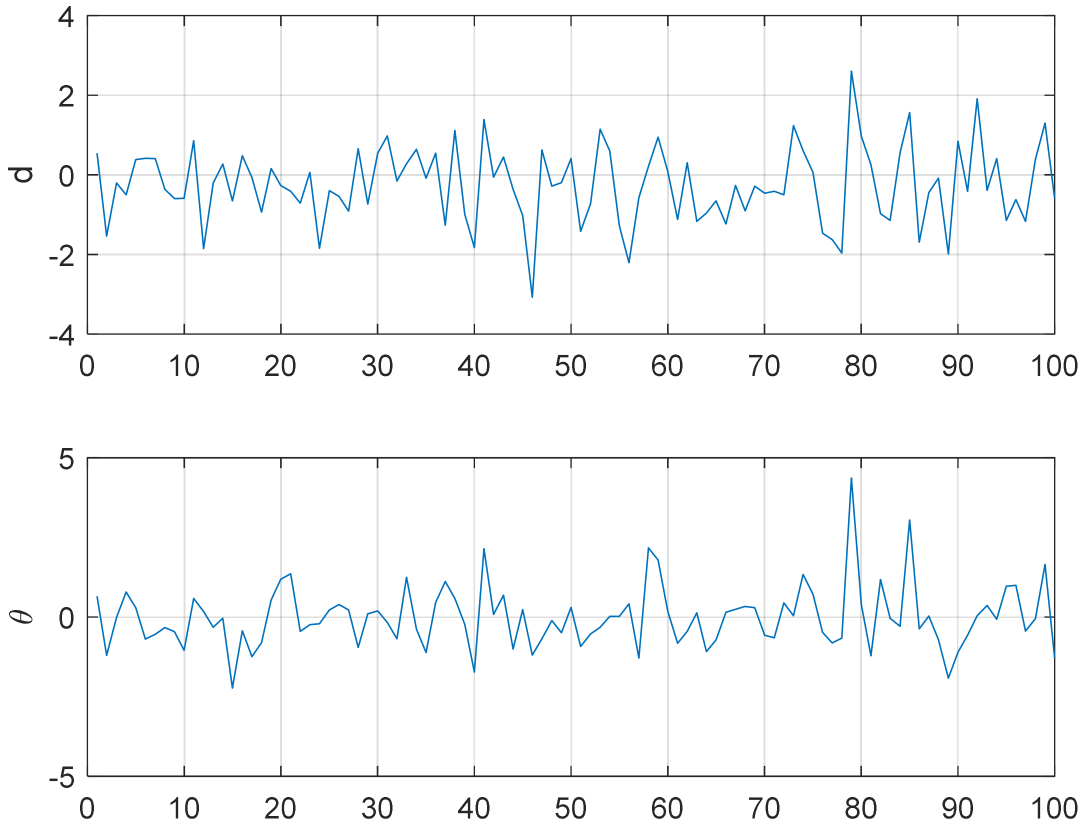
Example: high positive correlation

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$



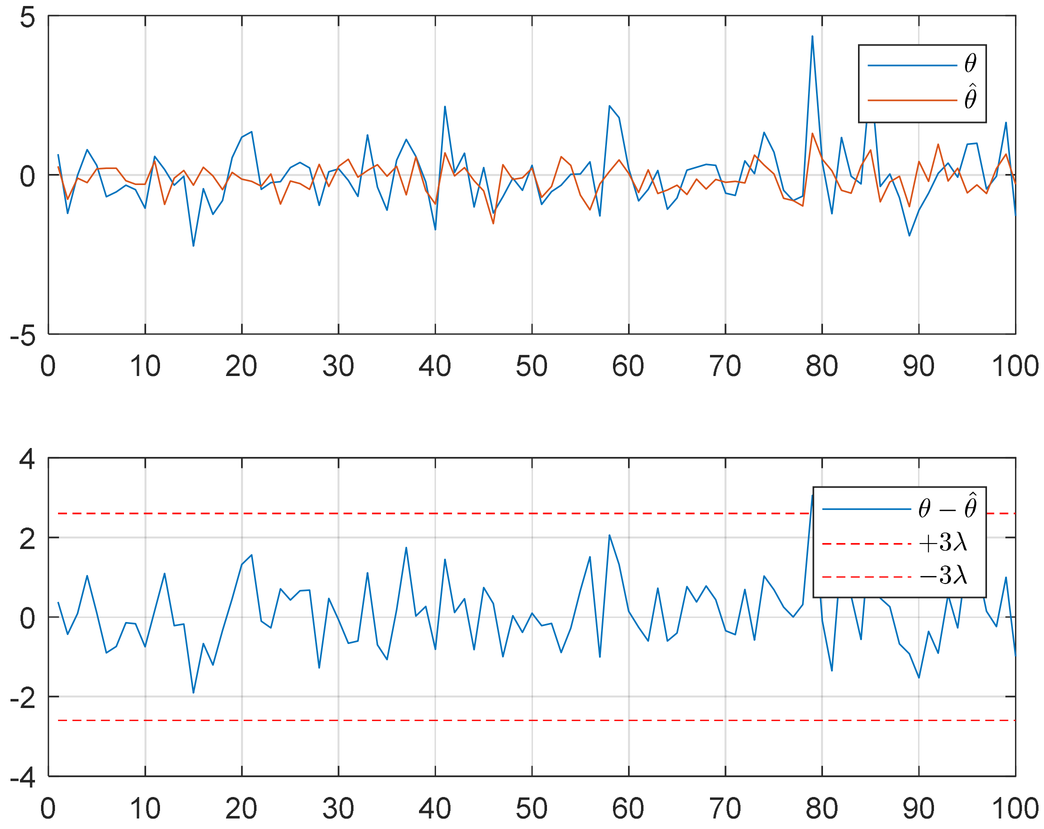
Example: moderate positive correlation

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$



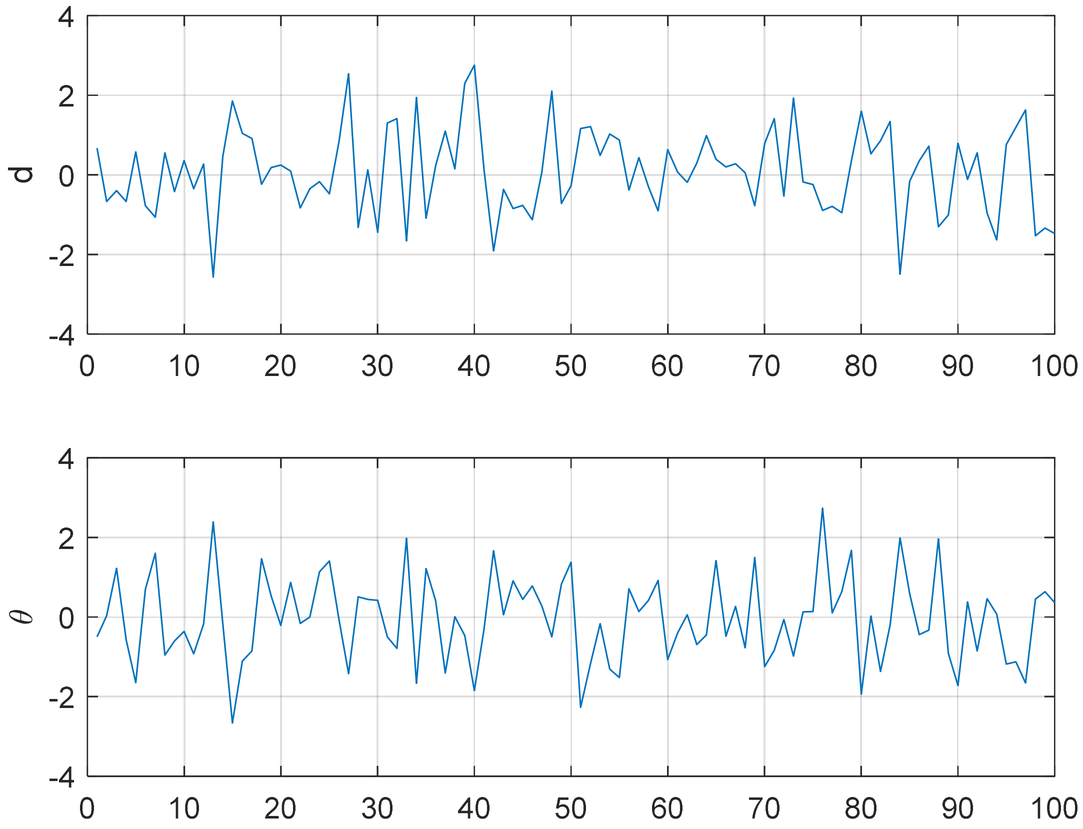
Example: moderate positive correlation

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$



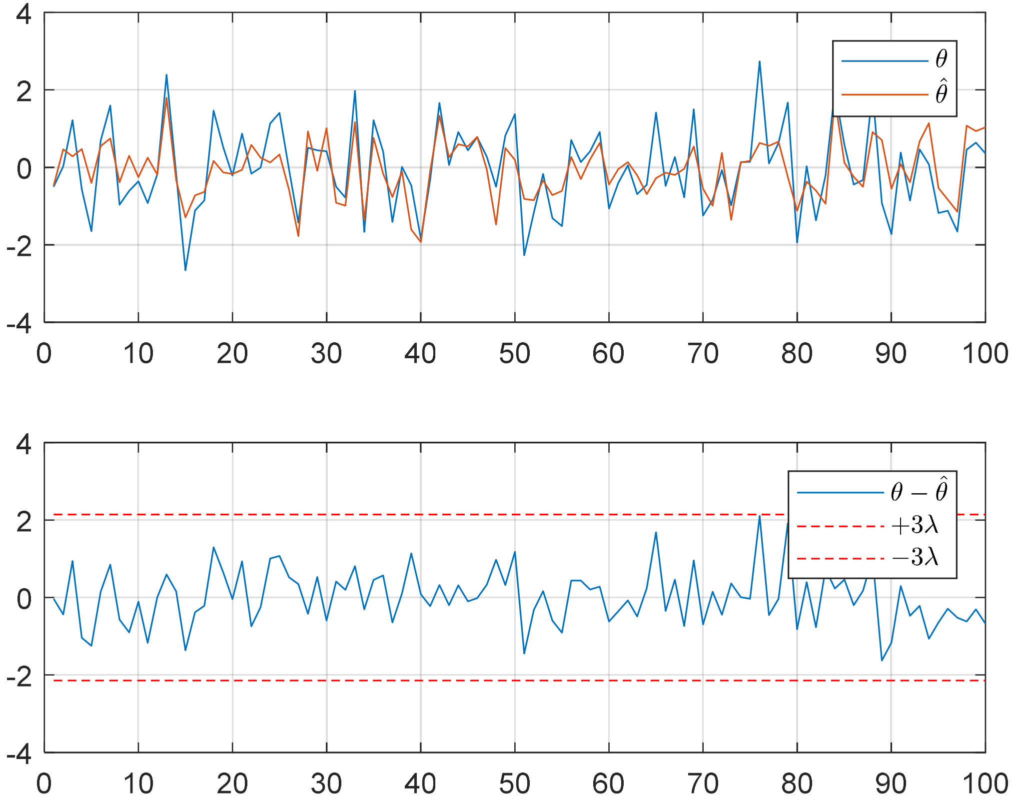
Example: negative correlation

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}\right)$$



Example: negative correlation

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}\right)$$



If θ and d have known non-zero mean, then it is sufficient to define

$$\tilde{\theta} = \theta - \theta_m, \qquad \tilde{d} = d - d_m$$

and apply Bayes rule to the new variables

$$\widehat{\tilde{\theta}} = E[\tilde{\theta} \backslash \tilde{d}] = \frac{\lambda_{\theta d}}{\lambda_{dd}} \tilde{d}, \qquad \text{Var}[\hat{\theta} - \theta] = \lambda^2$$

to finally get

$$\hat{\theta} = \theta_m + \frac{\lambda_{\theta d}}{\lambda_{dd}} (d - d_m).$$

Consider now the more general case in which θ and d are vectors and jointly Gaussian:

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \approx G\left(\begin{bmatrix} d_m \\ \theta_m \end{bmatrix}, \begin{bmatrix} \Lambda_{dd} & \Lambda_{d\theta} \\ \Lambda_{\theta d} & \Lambda_{\theta\theta} \end{bmatrix}\right).$$

One can follow the same derivation to get

$$\hat{\theta} = \theta_m + \Lambda_{\theta d}\Lambda_{dd}^{-1}(d - d_m)$$

$$\mathrm{Var}[\hat{\theta} - \theta] = \Lambda_{\theta\theta} - \Lambda_{\theta d}\Lambda_{dd}^{-1}\Lambda_{d\theta}$$

and

$$\mathrm{Var}[\hat{\theta} - \theta] = \Lambda_{\theta\theta} - \Lambda_{\theta d}\Lambda_{dd}^{-1}\Lambda_{d\theta} \leq \Lambda_{\theta\theta}.$$

In view of the application to real-time prediction and filtering we have to study the *recursive* problem, *i.e.*, how to update the estimate when new measurements of *d* arrive.

Consider the setting

$$\begin{bmatrix} \theta \\ d(1) \\ d(2) \end{bmatrix} \approx G(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_{\theta\theta} & \lambda_{\theta d(1)} & \lambda_{\theta d(2)} \\ \lambda_{d(1)\theta} & \lambda_{d(1)d(1)} & \lambda_{d(1)d(2)} \\ \lambda_{d(2)\theta} & \lambda_{d(2)d(1)} & \lambda_{d(2)d(2)} \end{bmatrix})$$

and:

- Compute a first estimate of θ given only $d(1)$
- Update it using the information provided by $d(2)$.

At time 1 we get

$$E[\theta \backslash d(1)] = \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}} d(1), \quad \text{Var}[\widehat{\theta} - \theta] = \lambda^2$$

While at time 2, having two samples of *d* we have

$$E[\theta \backslash d(1), d(2)] = \Lambda_{\theta d} \Lambda_{dd}^{-1} \begin{bmatrix} d(1) \\ d(2) \end{bmatrix} =$$

$$= \begin{bmatrix} \lambda_{\theta d(1)} & \lambda_{\theta d(2)} \end{bmatrix} \begin{bmatrix} \lambda_{d(1)d(1)} & \lambda_{d(1)d(2)} \\ \lambda_{d(2)d(1)} & \lambda_{d(2)d(2)} \end{bmatrix}^{-1} \begin{bmatrix} d(1) \\ d(2) \end{bmatrix}.$$

We can now expand this expression to relate the two estimates.

Computing the inverse we get

$$E[\theta \backslash d(1), d(2)] = \frac{1}{\lambda_{d(1)d(1)}\lambda^2} \begin{bmatrix} \lambda_{\theta d(1)} & \lambda_{\theta d(2)} \end{bmatrix} \begin{bmatrix} \lambda_{d(2)d(2)} & -\lambda_{d(2)d(1)} \\ -\lambda_{d(1)d(2)} & \lambda_{d(1)d(1)} \end{bmatrix} \begin{bmatrix} d(1) \\ d(2) \end{bmatrix}$$

where

$$\lambda^2 = \lambda_{d(2)d(2)} - \frac{\lambda^2_{d(1)d(2)}}{\lambda_{d(1)d(1)}}.$$

Expanding the products we get

$$E[\theta \backslash d(1), d(2)] = \frac{1}{\lambda_{d(1)d(1)}\lambda^2} \begin{bmatrix} \lambda_{\theta d(1)} & \lambda_{\theta d(2)} \end{bmatrix} \begin{bmatrix} \lambda_{d(2)d(2)}d(1) - \lambda_{d(2)d(1)}d(2) \\ -\lambda_{d(1)d(2)}d(1) + \lambda_{d(1)d(1)}d(2) \end{bmatrix}$$

$$E[\theta\backslash d(1),d(2)] = \frac{1}{\lambda_{d(1)d(1)}\lambda^2}(-\lambda_{\theta d(1)}\lambda_{d(2)d(1)} + \lambda_{\theta d(2)}\lambda_{d(1)d(1)})d(2) + \\ + \frac{1}{\lambda_{d(1)d(1)}\lambda^2}(\lambda_{\theta d(1)}\lambda_{d(2)d(2)} - \lambda_{\theta d(2)}\lambda_{d(1)d(2)})d(1)$$

$$E[\theta\backslash d(1),d(2)] = \frac{1}{\lambda^2}(\lambda_{\theta d(2)} - \lambda_{\theta d(1)}\frac{\lambda_{d(2)d(1)}}{\lambda_{d(1)d(1)}})d(2) + \\ + \frac{1}{\lambda_{d(1)d(1)}\lambda^2}(\lambda_{\theta d(1)}\lambda_{d(2)d(2)} - \lambda_{\theta d(2)}\lambda_{d(1)d(2)})d(1) \pm \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}}d(1)$$

$$E[\theta\backslash d(1),d(2)] = \frac{1}{\lambda^2}(\lambda_{\theta d(2)} - \lambda_{\theta d(1)}\frac{\lambda_{d(2)d(1)}}{\lambda_{d(1)d(1)}})d(2) + \\ + \frac{1}{\lambda_{d(1)d(1)}\lambda^2}(\lambda_{\theta d(1)}\lambda_{d(2)d(2)} - \lambda_{\theta d(2)}\lambda_{d(1)d(2)} - \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}}\lambda^2)d(1) + \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}}d(1)$$

Recursive Bayesian estimation

$$E[\theta \backslash d(1), d(2)] = \frac{1}{\lambda^2}(\lambda_{\theta d(2)} - \lambda_{\theta d(1)}\frac{\lambda_{d(2)d(1)}}{\lambda_{d(1)d(1)}})d(2) +$$

$$+ \frac{1}{\lambda_{d(1)d(1)}\lambda^2}(\lambda_{\theta d(1)}\lambda_{d(2)d(2)} - \lambda_{\theta d(2)}\lambda_{d(1)d(2)} - \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}}\lambda^2)d(1) + \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}}d(1)$$

$$E[\theta \backslash d(1), d(2)] = \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}}d(1) + \frac{1}{\lambda^2}(\lambda_{\theta d(2)} - \lambda_{\theta d(1)}\frac{\lambda_{d(2)d(1)}}{\lambda_{d(1)d(1)}})d(2) +$$

$$+ \frac{1}{\lambda^2}\frac{\lambda_{d(1)d(2)}}{\lambda_{d(1)d(1)}}(-\lambda_{\theta d(2)} + \lambda_{\theta d(1)}\frac{\lambda_{d(1)d(2)}}{\lambda_{d(1)d(1)}})d(1)$$

$$E[\theta \backslash d(1), d(2)] = \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}}d(1) + \frac{1}{\lambda^2}(\lambda_{\theta d(2)} - \lambda_{\theta d(1)}\frac{\lambda_{d(2)d(1)}}{\lambda_{d(1)d(1)}})(d(2) - \frac{\lambda_{d(1)d(2)}}{\lambda_{d(1)d(1)}}d(1)).$$

The quantity

$$e = d(2) - \frac{\lambda_{d(1)d(2)}}{\lambda_{d(1)d(1)}} d(1) = d(2) - E[d(2) \backslash d(1)]$$

is called the *innovation* of *d*(2) with respect to *d*(1).

It is defined as the difference between *d*(2) and its estimate based on *d*(1).

In terms of the innovation

$$E[\theta \backslash d(1), d(2)] = \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}} d(1) + \frac{1}{\lambda^2} (\lambda_{\theta d(2)} - \lambda_{\theta d(1)} \frac{\lambda_{d(2)d(1)}}{\lambda_{d(1)d(1)}}) e.$$

Properties of the innovation:

- Expected value: $E[e] = 0$
- Variance: $\lambda_{ee} = \text{Var}[e] = E[e^2] = E[(d(2) - \frac{\lambda_{d(1)d(2)}}{\lambda_{d(1)d(1)}} d(1))^2] = \ldots = \lambda^2$
- $\lambda_{\theta e} = E[\theta e] = E[\theta d(2)] - \frac{\lambda_{d(1)d(2)}}{\lambda_{d(1)d(1)}} E[\theta d(1)] = \lambda_{\theta d(2)} - \frac{\lambda_{d(1)d(2)}}{\lambda_{d(1)d(1)}} \lambda_{\theta d(1)}$

Reformulate the problem considering *d*(1) and *e* as data:

$$E[\theta \backslash d(1), e] = \begin{bmatrix} \lambda_{\theta d(1)} & \lambda_{\theta e} \end{bmatrix} \begin{bmatrix} \lambda_{d(1)d(1)} & 0 \\ 0 & \lambda_{ee} \end{bmatrix} \begin{bmatrix} d(1) \\ e \end{bmatrix} =$$

$$= \frac{\lambda_{\theta d(1)}}{\lambda_{d(1)d(1)}} d(1) + \frac{\lambda_{\theta e}}{\lambda_{ee}} e =$$

$$= E[\theta \backslash d(1)] + E[\theta \backslash e].$$

This conclusion is not surprising, as from the definition of *e* we get

$$d(2) = E[d(2) \backslash d(1)] + e.$$

Recursive Bayesian estimation: vector case

Consider the setting

$$\begin{bmatrix} \theta \\ d(1) \\ d(2) \end{bmatrix} \approx G(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda_{\theta\theta} & \Lambda_{\theta d(1)} & \Lambda_{\theta d(2)} \\ \Lambda_{d(1)\theta} & \Lambda_{d(1)d(1)} & \Lambda_{d(1)d(2)} \\ \Lambda_{d(2)\theta} & \Lambda_{d(2)d(1)} & \Lambda_{d(2)d(2)} \end{bmatrix})$$

then the estimate of θ is given by

$$\widehat{\theta} = E[\theta \backslash d(1), d(2)] = \Lambda_{\theta d(1)} \Lambda_{d(1)d(1)}^{-1} d(1) + \Lambda_{\theta e} \Lambda_{ee}^{-1} e.$$