8

Parameter Estimation

8.1 Introduction

Predictors, filters and smoothers have previously been described for state recovery under the assumption that the parameters of the generating models are correct. More often than not, the problem parameters are unknown and need to be identified. This section describes some standard statistical techniques for parameter estimation. Paradoxically, the discussed parameter estimation methods rely on having complete state information available. Although this is akin to a chicken-and-egg argument (state availability obviates the need for filters along with their attendant requirements for identified models), the task is not insurmountable.

The role of solution designers is to provide a cost benefit. That is, their objectives are to deliver improved performance at an acceptable cost. Inevitably, this requires simplifications so that the problems become sufficiently tractable and amenable to feasible solution. For example, suppose that speech emanating from a radio is too noisy and barely intelligible. In principle, high-order models could be proposed to equalise the communication channel, demodulate the baseband signal and recover the phonemes. Typically, low-order solutions tend to offer better performance because of the difficulty in identifying large numbers of parameters under low-SNR conditions. Consider also the problem of monitoring the output of a gas sensor and triggering alarms when environmental conditions become hazardous. Complex models could be constructed to take into account diurnal pressure variations, local weather influences and transients due to passing vehicles. It often turns out that low-order solutions exhibit lower false alarm rates because there are fewer assumptions susceptible to error.

Thus, the absence of complete information need not inhibit solution development. Simple schemes may suffice, such as conducting trials with candidate parameter values and assessing the consequent error performance.

In maximum-likelihood estimation [1] – [5], unknown parameters $\theta_1, \theta_2, \ldots, \theta_M$, are identified given states, x_k, by maximising a log-likelihood function, $\log f(\theta_1, \theta_2, \ldots, \theta_M \mid x_k)$. For example, the subject of noise variance estimation was studied by Mehra in [6], where maximum-likelihood estimates (MLEs) were updated using the Newton-Raphson method. Rife and Boorstyn obtained Cramér-Rao bounds for some MLEs, which "indicate the best estimation that can be made with the available data" [7]. Nayak *et al* used the pseudo-

---

"The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work" *John Von Neuman*

inverse to estimate unknown parameters in [8]. Belangér subsequently employed a least-squares approach to estimate the process noise and measurement noise variances [9]. A recursive technique for least-squares parameter estimation was developed by Strejc [10]. Dempster, Laird and Rubin [11] proved the convergence of a general purpose technique for solving joint state and parameter estimation problems, which they called the expectation-maximization (EM) algorithm. They addressed problems where complete (state) information is not available to calculate the log-likelihood and instead maximised the expectation of $\log f(\theta_1,\theta_2,...,\theta_M \mid z_k)$, given incomplete measurements, z_k. That is, by virtue of Jensen's inequality the unknowns are found by using an objective function (also called an approximate log-likelihood function), $E\{\log f(\theta_1,\theta_2,...,\theta_M \mid z_k)\}$, as a surrogate for $\log f(\theta_1, \theta_2, \ldots, \theta_M \mid x_k)$.

The system identification literature is vast and some mature techniques have evolved. It is acknowledged that subspace identification methods have been developed for general problems where a system's stochastic inputs, deterministic inputs and outputs are available. The subspace algorithms [12] – [14] consist of two steps. First, the order of the system is identified from stacked vectors of the inputs and outputs. Then the unknown parameters are determined from an extended observability matrix.

Continuous-time maximum-likelihood estimation has been mentioned previously. Here, the attention is focussed on the specific problem of joint state and parameter estimation exclusively from discrete measurements of a system's outputs. The developments proceed as follows. Section 8.2 reviews the maximum-likelihood estimation method for obtaining unknown parameters. The same estimates can be found using the method of least squares, which was pioneered by Gauss for fitting astronomical observations. Well known (filtering) EM algorithms for variance and state matrix estimation are described in Section 8.3. Improved parameter estimation accuracy can be obtained via smoothing EM algorithms, which are introduced in Section 8.4.

The filtering and smoothing EM algorithms discussed herein require caution. When perfect state information is available, the corresponding likelihood functions are exact. However, the use of imperfect state estimates leads to approximate likelihood functions, approximate Cramér-Rao bounds and biased MLEs. When the SNR is sufficiently high and the states are recovered exactly, the bias terms within the state matrix elements and process noise variances diminish to zero. Consequently, process noise variance and state matrix estimation is recommended only when the measurement noise is negligible. Conversely, measurement noise variance estimation is advocated when the SNR is sufficiently low.

"A hen is only an egg's way of making another egg." *Samuel Butler*

8.2 Maximum-Likelihood Estimation

8.2.1 General Method

Let $p(\theta \mid x_k)$ denote the probability density function of an unknown parameter θ, given samples of a discrete random variable x_k. An estimate, $\hat{\theta}$, can be obtained by finding the argument θ that maximises the probability density function, that is,

$$\hat{\theta} = \arg\max_{\theta} = p(\theta \mid x_k) . \tag{1}$$

A solution can be found by setting $\frac{\partial p(\theta \mid x_k)}{\partial \theta} = 0$ and solving for the unknown θ. Since the logarithm function is monotonic, a solution may be found equivalently by maximising

$$\hat{\theta} = \arg\max_{\theta} = \log p(\theta \mid x_k) \tag{2}$$

and setting $\frac{\partial \log p(\theta \mid x_k)}{\partial \theta} = 0$. For exponential families of distributions, the use of (2) considerably simplifies the equations to be solved.

Suppose that N mutually independent samples of x_k are available, then the joint density function of all the observations is the product of the densities

$$\begin{aligned} f(\theta \mid x_k) &= p(\theta \mid x_1) p(\theta \mid x_2) \cdots p(\theta \mid x_N) \\ &= \prod_{k=1}^{K} p(\theta \mid x_k) , \end{aligned} \tag{3}$$

which serves as a likelihood function. The MLE of θ may be found maximising the log-likelihood

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \ \log f(\theta \mid x_k) \\ &= \arg\max_{\theta} \ \sum_{k=1}^{N} p(\theta \mid x_k) \end{aligned} \tag{4}$$

by solving for a θ that satisfies $\frac{\partial \log f(\theta \mid x_k)}{\partial \theta} = \frac{\partial \log \sum_{k=1}^{N} p(\theta \mid x_k)}{\partial \theta} = 0$. The above maximum-likelihood approach is applicable to a wide range of distributions. For example, the task of estimating the intensity of a Poisson distribution from measurements is demonstrated below.

"Therefore I would not have it unknown to Your Holiness, the only thing which induced me to look for another way of reckoning the movements of the heavenly bodies was that I knew that mathematicians by no means agree in their investigation thereof." *Nicolaus Copernicus*"

Example 1. Suppose that N observations of integer x_k have a Poisson distribution $f(x_k) = \frac{e^{-\mu}\mu^{x_k}}{x_k!}$, where the intensity, μ, is unknown. The corresponding log-likelihood function is

$$\begin{aligned} \log f(\mu \mid x_k) &= \log\left(\frac{e^{-\mu}\mu^{x_1}}{x_1!}\frac{e^{-\mu}\mu^{x_2}}{x_2!}\frac{e^{-\mu}\mu^{x_3}}{x_3!}\cdots\frac{e^{-\mu}\mu^{x_N}}{x_N!} \right) \\ &= \log\left(\frac{1}{x_1!x_2!\cdots x_N!}\mu^{x_1+x_2+\ldots+x_N}e^{-N\mu} \right) \\ &= -\log(x_1!x_2!\cdots x_N!) + \log(\mu^{x_1+x_2+\ldots+x_N}) - N\mu \,. \end{aligned} \tag{5}$$

Taking $\frac{\partial \log f(\mu \mid x_k)}{\partial \mu} = \frac{1}{\mu}\sum_{k=1}^{N} x_k - N = 0$ yields

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k \,. \tag{6}$$

Since $\frac{\partial^2 \log f(\mu \mid x_k)}{\partial \mu^2} = -\frac{1}{\mu^2}\sum_{k=1}^{N} x_k$ is negative for all μ and $x_k \geq 0$, the stationary point (6) occurs at a maximum of (5). That is to say, $\hat{\mu}$ is indeed a maximum-likelihood estimate.

8.2.2 State Matrix Estimation

From the Central Limit Theorem, which was mentioned in Chapter 6, the mean of a sufficiently large sample of independent identically distributed random variables will asymptotically approach a normal distribution. Consequently, in many maximum-likelihood estimation applications it is assumed that random variables are normally distributed. Recall that the normal (or Gaussian) probability density function of a discrete random variable x_k with mean μ and covariance R_{xx} is

$$p(x_k) = \frac{1}{(2\pi)^{N/2}\left|R_{xx}\right|^{1/2}} \exp\left\{ -\frac{1}{2}(x_k - \mu)^T R_{xx}^{-1}(x_k - \mu) \right\}, \tag{7}$$

in which $\left|R_{xx}\right|$ denotes the determinant of R_{xx}. A likelihood function for a sample of N independently identically distributed random variables is

$$f(x_k) = \prod_{k=1}^{N} p(x_k) = \frac{1}{(2\pi)^{N/2}\left|R_{xx}\right|^{N/2}} \sum_{k=1}^{N} \exp\left\{ -\frac{1}{2}(x_k - \mu)^T R_{xx}^{-1}(x_k - \mu) \right\}. \tag{8}$$

In general, it is more convenient to work with the log-likelihood function

---

"How wonderful that we have met with a paradox. Now we have some hope of making progress." *Niels Henrik David Bohr*

$$\log f(x_k) = -\log\ (2\pi)^{N/2} \left|R_{xx}\right|^{N/2} - \frac{1}{2}\sum_{k=1}^{N}(x_k - \mu)^T R_{xx}^{-1}(x_k - \mu)\ . \tag{9}$$

An example of estimating a model coefficient using the Gaussian log-likelihood approach is set out below.

Example 2. Consider an autoregressive order-one process $x_{k+1} = a_0 x_k + w_k$ in which it is desired to estimate $a_0 \in \mathbb{R}$ from samples of x_k. It follows from $x_{k+1} \sim \mathcal{N}(a_0 x_k,\ \sigma_w^2)$ that

$$\log f(a_0 \mid x_{k+1}) = -\log\ (2\pi)^{N/2}\sigma_w^N - \frac{1}{2}\sum_{k=1}^{N}\sigma_w^{-2}(x_{k+1} - a_0 x_k)^2\ .$$

Setting $\dfrac{\partial \log\ f(a_0 \mid x_{k+1})}{\partial a_0}$ equal to zero gives $\sum_{k=1}^{N} x_{k+1}x_k = a_0\sum_{k=1}^{N} x_k^2$ which results in the MLE

$$\hat{a}_0 = \frac{\sum_{k=1}^{N} x_{k+1}x_k}{\sum_{k=1}^{N} x_k^2}\ .$$

Often within filtering and smoothing applications there are multiple parameters to be identified. Denote the unknown parameters by $\theta_1, \theta_2, \ldots, \theta_M$, then the MLEs may be found by solving the M equations

$$\frac{\partial \log f(\theta_1, \theta_2, \cdots \theta_M \mid x_k)}{\partial \theta_1} = 0$$

$$\frac{\partial \log f(\theta_1, \theta_2, \cdots \theta_M \mid x_k)}{\partial \theta_2} = 0$$

$$\vdots$$

$$\frac{\partial \log f(\theta_1, \theta_2, \cdots \theta_M \mid x_k)}{\partial \theta_M} = 0.$$

An vector parameter estimation example is outlined below.

Example 3. Consider the third-order autoregressive model

$$x_{k+3} + a_2 x_{k+2} + a_1 x_{k+1} + a_0 x_k = w_k \tag{10}$$

which can be written in the state-space form

"If we all worked on the assumption that what is accepted as true is really true, there would be little hope of advance." *Orville Wright*

$$\begin{bmatrix} x_{1,k+1} \\ x_{2,k+1} \\ x_{3,k+1} \end{bmatrix} = \begin{bmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{1,k} \\ x_{2,k} \\ x_{3,k} \end{bmatrix} + \begin{bmatrix} w_k \\ 0 \\ 0 \end{bmatrix}. \tag{11}$$

Assuming $x_{1,k+1} \sim \mathcal{N}(-a_2 x_{1,k} - a_1 x_{2,k} - a_0 x_{3,k}, \sigma_w^2)$ and setting to zero the partial derivatives of the corresponding log-likelihood function with respect to a_0, a_1 and a_2 yields

$$-\begin{bmatrix} \sum_{k=1}^{N} x_{3,k}^2 & \sum_{k=1}^{N} x_{2,k}x_{3,k} & \sum_{k=1}^{N} x_{1,k}x_{3,k} \\ \sum_{k=1}^{N} x_{2,k}x_{3,k} & \sum_{k=1}^{N} x_{2,k}^2 & \sum_{k=1}^{N} x_{2,k}x_{1,k} \\ \sum_{k=1}^{N} x_{1,k}x_{3,k} & \sum_{k=1}^{N} x_{2,k}x_{1,k} & \sum_{k=1}^{N} x_{1,k}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{N} x_{1,k+1}x_{3,k} \\ \sum_{k=1}^{N} x_{1,k+1}x_{2,k} \\ \sum_{k=1}^{N} x_{1,k+1}x_{1,k} \end{bmatrix}. \tag{12}$$

Hence, the MLEs are given by

$$\begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{bmatrix} = -\begin{bmatrix} \sum_{k=1}^{N} x_{3,k}^2 & \sum_{k=1}^{N} x_{2,k}x_{3,k} & \sum_{k=1}^{N} x_{1,k}x_{3,k} \\ \sum_{k=1}^{N} x_{2,k}x_{3,k} & \sum_{k=1}^{N} x_{2,k}^2 & \sum_{k=1}^{N} x_{2,k}x_{1,k} \\ \sum_{k=1}^{N} x_{1,k}x_{3,k} & \sum_{k=1}^{N} x_{2,k}x_{1,k} & \sum_{k=1}^{N} x_{1,k}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^{N} x_{1,k+1}x_{3,k} \\ \sum_{k=1}^{N} x_{1,k+1}x_{2,k} \\ \sum_{k=1}^{N} x_{1,k+1}x_{1,k} \end{bmatrix}. \tag{13}$$

8.2.3 Variance Estimation

MLEs can be similarly calculated for unknown variances, as is demonstrated by the following example.

Example 4. Consider a random variable generated by $x_k = \mu + w_k$ where $\mu \in \mathbb{R}$ is fixed and $w_k \in \mathbb{R}$ is assumed to be a zero-mean Gaussian white input sequence. Since $x_k \sim \mathcal{N}(\mu, \sigma_w^2)$, it follows that

$$\log f(\sigma_w^2 \mid x_k) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma_w^2 - \frac{1}{2}\sigma_w^{-2}\sum_{k=1}^{N}(x_k - \mu)^2$$

and

$$\frac{\partial \log f(\sigma_w^2 \mid x_k)}{\partial \sigma_w^2} = -\frac{N}{2}(\sigma_w^2)^{-1} + \frac{1}{2}(\sigma_w^2)^{-2}\sum_{k=1}^{N}(x_k - \mu)^2 .$$

From the solution of $\frac{\partial \log f(\sigma_w^2 \mid x_k)}{\partial \sigma_w^2} = 0$, the MLE is

---

"In science one tries to tell people, in such a way as to be understood by everyone, something that no-one knew before. But in poetry, it's the exact opposite." *Paul Adrien Maurice Dirac*

$$\hat{\sigma}_w^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2 \text{ , without replacement.} \tag{14}$$

If the random samples are taken from a population without replacement, the samples are not independent, the covariance between two different samples is nonzero and the MLE (14) is biased. If the sampling is done with replacement then the sample values are independent and the following correction applies

$$\hat{\sigma}_w^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu)^2 \text{ , with replacement.} \tag{15}$$

The corrected denominator within the above sample variance is only noticeable for small sample sizes, as the difference between (14) and (15) is negligible for large N. The MLE (15) is unbiased, that is, its expected value equals the variance of the population. To confirm this property, observe that

$$\begin{aligned} E\{\sigma_w^2\} &= E\left\{\frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu)^2\right\} \\ &= E\left\{\frac{1}{N-1}\sum_{k=1}^{N}x_k^2 - 2\mu x_k + \mu^2\right\} \\ &= \frac{N}{N-1}(E\{x_k^2\} - E\{\mu^2\}). \end{aligned} \tag{16}$$

Using $E\{x_k^2\} = \sigma_w^2 + \overline{x}^2$, $E\{E\{x_k^2\}\} = E\{\sigma_w^2\} + E\{\mu\}^2$, $E\{E\{x_k^2\}\} = E\{\mu^2\} = \mu^2$ and $E\{\sigma_w^2\} = \sigma_w^2/N$ within (16) yields $E\{\sigma_w^2\} = \sigma_w^2$ as required. Unless stated otherwise, it is assumed herein that the sample size is sufficiently large so that $N^{-1} \approx (N - 1)^{-1}$ and (15) may be approximated by (14). A caution about modelling error contributing bias is mentioned below.

Example 5. Suppose that the states considered in Example 4 are actually generated by $x_k = \mu + w_k + s_k$, where s_k is an independent input that accounts for the presence of modelling error. In this case, the assumption $x_k \sim \mathcal{N}(\mu,\ \sigma_w^2 + \sigma_s^2)$ leads to $\hat{\sigma}_w^2 + \hat{\sigma}_s^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2$, in which case (14) is no longer an unbiased estimate of σ_w^2.

8.2.4 Cramér-Rao Lower Bound

The Cramér-Rao Lower Bound (CRLB) establishes a limit of precision that can be achieved for any unbiased estimate of a parameter θ. It actually defines a lower bound for the variance $\sigma_{\hat{\theta}}^2$ of $\hat{\theta}$. As is pointed out in [1], since $\hat{\theta}$ is assumed to be unbiased, the variance $\sigma_{\hat{\theta}}^2$ equals the parameter error variance. Determining lower bounds for parameter error

---

"Everyone hears only what he understands." *Johann Wolfgang von Goethe*

variances is useful for model selection. Another way of selecting models involves comparing residual error variances [23]. A lucid introduction to Gaussian CRLBs is presented in [2]. An extensive survey that refers to the pioneering contributions of Fisher, Cramér and Rao appears in [4].

The bounds on the parameter variances are found from the inverse of the so-called Fisher information. A formal definition of the CRLB for scalar parameters is as follows.

Theorem 1 (Cramér-Rao Lower Bound) [2] - [5]: Assume that $f(\theta \mid x_k)$ satisfies the following regularity conditions:

(i) $\dfrac{\partial \log f(\theta \mid x_k)}{\partial \theta}$ *and* $\dfrac{\partial^2 \log f(\theta \mid x_k)}{\partial \theta^2}$ *exist for all θ, and*

(ii) $E\left\{\dfrac{\partial \log f(\theta \mid x_k)}{\partial \theta}\right\} = 0,$ *for all θ.*

Define the Fisher information by

$$F(\theta) = -E\left\{\frac{\partial^2 \log f(\theta \mid x_k)}{\partial \theta^2}\right\}, \tag{17}$$

where the derivative is evaluated at the actual value of θ. Then the variance $\sigma_{\hat{\theta}}^2$ *of an unbiased estimate* $\hat{\theta}$ *satisfies*

$$\sigma_{\hat{\theta}}^2 \geq F^{-1}(\theta) . \tag{18}$$

Proofs for the above theorem appear in [2] – [5].

Example 6. Suppose that samples of $x_k = \mu + w_k$ are available, where w_k is a zero-mean Gaussian white input sequence and $\mu \in \mathbb{R}$ is unknown. Since $w_k \sim \mathcal{N}(0, \sigma_w^2)$,

$$\log f(\mu \mid x_k) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma_w^2 - \frac{1}{2}\sigma_w^{-2}\sum_{k=1}^{N}(x_k - \mu)^2$$

and

$$\frac{\partial \log f(\mu \mid x_k)}{\partial \mu} = \sigma_w^{-2}\sum_{k=1}^{N}(x_k - \mu) .$$

Setting $\dfrac{\partial \log f(\mu \mid x_k)}{\partial \mu} = 0$ yields the MLE

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k , \tag{19}$$

"Wall street people learn nothing and forget everything." *Benjamin Graham*

which is unbiased because $E\{\hat{\mu}\} = E\left\{\frac{1}{N}\sum_{k=1}^{N} x_k\right\} = E\left\{\frac{1}{N}\sum_{k=1}^{N} \mu\right\} = \mu$. From Theorem 1, the Fisher information is

$$F(\mu) = E\left\{-\frac{\partial^2 \log f(\mu \mid x_k)}{\partial \mu^2}\right\} = E\{N\sigma_w^{-2}\} = N\sigma_w^{-2}$$

and therefore

$$\sigma_{\hat{\mu}}^2 \geq \sigma_w^2 / N \,. \tag{20}$$

The above inequality suggests that a minimum of one sample is sufficient to bound the variance of the MLE (19). It is also apparent from (20) that the error variance of $\hat{\mu}$ decreases withincreasing sample size.

The CRLB is extended for estimating a vector of parameters $\theta_1, \theta_2, \ldots, \theta_M$ by defining the $M \times M$ Fisher information matrix

$$\begin{bmatrix} \hat{x}_{k+1/k}^{(u)} \\ \hat{x}_{k/k}^{(u)} \end{bmatrix} = \begin{bmatrix} (A - K_k^{(u)}C) & K_k^{(u)} \\ (I - L_k^{(u)}C) & L_k^{(u)} \end{bmatrix} \begin{bmatrix} \hat{x}_{k/k-1}^{(u)} \\ z_k \end{bmatrix}, \tag{21}$$

for $i, j = 1 \ldots M$. The parameter error variances are then bounded by the diagonal elements of Fisher information matrix inverse

$$\sigma_{\hat{\theta}_i}^2 \geq F_{ii}^{-1}(\theta) \,. \tag{22}$$

Formal vector CRLB theorems and accompanying proofs are detailed in [2] – [5].

Example 7. Consider the problem of estimating both μ and σ_w^2 from N samples of $x_k = \mu + w_k$, with $w_k \sim \mathcal{N}(0, \sigma_w^2)$. Recall from Example 6 that

$$\log f(\mu, \sigma_w^2 \mid x_k) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma_w^2 - \frac{1}{2}\sigma_w^{-2}\sum_{k=1}^{N}(x_k - \mu)^2 \,.$$

Therefore, $\frac{\partial \log f(\mu, \sigma_w^2 \mid x_k)}{\partial \mu} = \sigma_w^{-2}\sum_{k=1}^{N}(x_k - \mu)$ and $\frac{\partial^2 \log f(\mu, \sigma_w^2 \mid x_k)}{\partial \mu^2} = -N\sigma_w^{-2}$. In Example 4 it is found that $\frac{\partial \log f(\mu, \sigma_w^2 \mid x_k)}{\partial \sigma_w^2} = -\frac{N}{2}(\sigma_w^2)^{-1} + \frac{1}{2}(\sigma_w^2)^{-2}\sum_{k=1}^{N}(x_k - \mu)^2$, which implies

$$\frac{\partial^2 \log f(\mu, \sigma_w^2 \mid x_k)}{\partial (\sigma_w^2)^2} = \frac{N}{2}(\sigma_w^2)^{-2} - (\sigma_w^2)^{-3}\sum_{k=1}^{N}(x_k - \mu)^2$$

"Laying in bed this morning contemplating how amazing it would be if somehow Oscar Wilde and Mae West could twitter from the grave." *@DitaVonTeese*

$$= \frac{N}{2}(\sigma_w^2)^{-2} - N(\sigma_w^2)^{-2}$$

$$= -\frac{N}{2}\sigma_w^{-4},$$

$$\frac{\partial^2 \log f(\mu, \sigma_w^2 \mid x_k)}{\partial \mu \partial \sigma_w^2} = -(\sigma_w^2)^{-2} \sum_{k=1}^{N} (x_k - \mu) \text{ and } E\left\{ \frac{\partial^2 \log f(\mu, \sigma_w^2 \mid x_k)}{\partial \mu \partial \sigma_w^2} \right\} = 0.$$

The Fisher information matrix and its inverse are then obtained from (21) as

$$F(u, \sigma_w^2) = \begin{bmatrix} N\sigma_w^{-2} & 0 \\ 0 & 0.5N\sigma_w^{-4} \end{bmatrix}, \; F^{-1}(u, \sigma_w^2) = \begin{bmatrix} \sigma_w^2 / N & 0 \\ 0 & 2\sigma_w^4 / N \end{bmatrix}.$$

It is found from (22) that the lower bounds for the MLE variances are $\sigma_{\hat{\mu}}^2 \geq \sigma_w^2 / N$ and $\sigma_{\hat{\sigma}_w^2}^2 \geq 2\sigma_w^4 / N$. The impact of modelling error on parameter estimation accuracy is examined below.

Example 8. Consider the problem of estimating σ_w^2 given samples of states which are generated by $x_k = \mu + w_k + s_k$, where s_k is an independent sequence that accounts for the presence of modelling error. From the assumption $x_k \sim \mathcal{N}(\mu, \; \sigma_w^2 + \sigma_s^2)$, the associated log likelihood function is

$$\frac{\partial \log f(\sigma_w^2 \mid x_k)}{\partial \sigma_w^2} = -\frac{N}{2}(\sigma_w^2 + \sigma_s^2)^{-1} + \frac{1}{2}(\sigma_w^2 + \sigma_s^2)^{-2} \sum_{k=1}^{N} (x_k - \mu)^2 ,$$

which leads to $\frac{\partial^2 \log f(\sigma_w^2 \mid x_k)}{\partial (\sigma_w^2)^2} = -\frac{N}{2}(\sigma_w^2 + \sigma_s^2)^{-2}$, that is, $\sigma_{\hat{\sigma}_w^2}^2 \geq 2(\sigma_w^2 + \sigma_s^2)^2 / N$. Thus, parameter estimation accuracy degrades as the variance of the modelling error increases. If σ_s^2 is available *a priori* then setting $\frac{\partial \log f(\sigma_w^2 \mid x_k)}{\partial \sigma_w^2} = 0$ leads to the improved estimate

$$\hat{\sigma}_w^2 = -\sigma_s^2 + \frac{1}{N} \sum_{k=1}^{K} (x_k - \mu)^2 .$$

"There are only two kinds of people who are really fascinating; people who know absolutely everything, and people who know absolutely nothing." *Oscar Fingal O'Flahertie Wills Wilde*

8.3 Filtering EM Algorithms

8.3.1 Background

The EM algorithm [3], [7], [11], [15] – [17], [19] – [22] is a general purpose technique for solving joint state and parameter estimation problems. In maximum-likelihood estimation, it is desired to estimate parameters $\theta_1, \theta_2, \ldots, \theta_M$, given states by maximising the log-likelihood $\log f(\theta_1, \theta_2, \ldots, \theta_M \mid x_k)$. When complete state information is not available to calculate the log-likelihood, the expectation of $\log f(\theta_1, \theta_2, ..., \theta_M \mid x_k)$, given incomplete measurements, z_k, is maximised instead. This basic technique was in use prior to Dempster, Laird and Rubin naming it the EM algorithm 1977 [11]. They published a general formulation of the algorithm, which consists of iterating an expectation step and a maximization step. Their expectation step involves least squares calculations on the incomplete observations using the current parameter iterations to estimate the underlying states. In the maximization step, the unknown parameters are re-estimated by maximising a joint log likelihood function using state estimates from the previous expectation step. This sequence is repeated for either a finite number of iterations or until the estimates and the log likelihood function are stable. Dempster, Laird and Rubin [11] also established parameter map conditions for the convergence of the algorithm, namely that the incomplete data log likelihood function is monotonically nonincreasing.

Wu [16] subsequently noted an equivalence between the conditions for a map to be closed and the continuity of a function. In particular, if the likelihood function satisfies certain modality, continuity and differentiability conditions, the parameter sequence converges to some stationary value. A detailed analysis of Wu's convergence results appears in [3]. Shumway and Stoffer [15] introduced a framework that is employed herein, namely, the use of a Kalman filter within the expectation step to recover the states. Feder and Weinstein [17] showed how a multiparameter estimation problem can be decoupled into separate maximum likelihood estimations within an EM algorithm. Some results on the convergence of EM algorithms for variance and state matrix estimation [19] – [20] are included within the developments below.

8.3.2 Measurement Noise Variance Estimation

8.3.2.1 EM Algorithm

The problem of estimating parameters from incomplete information has been previously studied in [11] – [16]. It is noted in [11] that the likelihood functions for variance estimation do not exist in explicit closed form. This precludes straightforward calculation of the Hessians required in [3] to assert convergence. Therefore, an alternative analysis is presented here to establish the monotonicity of variance iterations.

The expectation step described below employs the approach introduced in [15] and involves the use of a Kalman filter to obtain state estimates. The maximization step requires the calculation of decoupled MLEs similarly to [17]. Measurements of a linear time-invariant system are modelled by

---

"I'm no model lady. A model is just an imitation of the real thing." *Mary (Mae) Jane West*

$$x_{k+1} = Ax_k + Bw_k\,, \tag{23}$$

$$y_k = Cx_k + Dw_k\,, \tag{24}$$

$$z_k = y_k + v_k\,, \tag{25}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$ and w_k, v_k are stationary processes with $E\{w_k\} = 0$, $E\{w_j w_k^T\} = Q\delta_{jk}$, $E\{v_k\} = E\{w_j v_k^T\} = 0$ and $E\{v_j v_k^T\} = R\delta_{jk}$. To simplify the presentation, it is initially assumed that the direct feed-through matrix, D, is zero. A nonzero D will be considered later.

Suppose that it is desired to estimate $R = \mathrm{diag}(\sigma_{1,v}^2, \sigma_{2,v}^2, \ldots, \sigma_{p,v}^2)$ given N samples of z_k and y_k. Let $z_{i,k}$, $y_{i,k}$ and $v_{i,k}$ denote the i^{th} element of the vectors z_k, y_k and v_k, respectively. Then (25) may be written in terms of its i components, $z_{i,k} = y_{i,k} + v_{i,k}$, that is,

$$v_{i,k} = z_{i,k} - y_{i,k}\,. \tag{26}$$

From the assumption $v_{i,k} \sim \mathcal{N}(0, \sigma_{i,v}^2)$, an MLE for the unknown $\sigma_{i,v}^2$ is obtained from the sample variance formula

$$\hat{\sigma}_{i,v}^2 = \frac{1}{N}\sum_{k=1}^{N}(z_{i,k} - y_{i,k})^2\,. \tag{27}$$

An EM algorithm for updating the measurement noise variance estimates is described as follows. Assume that there exists an estimate $\hat{R}^{(u)} = \mathrm{diag}((\hat{\sigma}_{1,v}^{(u)})^2, (\hat{\sigma}_{2,v}^{(u)})^2, \ldots, (\hat{\sigma}_{p,v}^{(u)})^2)$ of R at iteration u. A Kalman filter designed with $\hat{R}^{(u)}$ may then be employed to produce corrected output estimates $\hat{y}_{k/k}^{(u)}$. The filter's design Riccati equation is given by

$$P_{k+1/k}^{(u)} = (A - K_k^{(u)}C)P_{k/k-1}^{(u)}(A - K_k^{(u)}C)^T + K_k^{(u)}\hat{R}^{(u)}(K_k^{(u)})^T + BQB^T\,, \tag{28}$$

where $K_k^{(u)} = AP_{k/k-1}^{(u)}C^T(CP_{k/k-1}^{(u)}C^T + \hat{R}^{(u)})^{-1}$ is the predictor gain. The output estimates are calculated from

$$\begin{bmatrix}\hat{x}_{k+1/k}^{(u)}\\ \hat{x}_{k/k}^{(u)}\end{bmatrix} = \begin{bmatrix}(A - K_k^{(u)}C) & K_k^{(u)}\\ (I - L_k^{(u)}C) & L_k^{(u)}\end{bmatrix}\begin{bmatrix}\hat{x}_{k+1/k}^{(u)}\\ z_k\end{bmatrix}, \tag{29}$$

$$\hat{y}_{k/k}^{(u)} = C\hat{x}_{k/k}^{(u)}\,, \tag{30}$$

where $L_k^{(u)} = P_{k/k-1}^{(u)}C^T(CP_{k/k-1}^{(u)}C^T + \hat{R}^{(u)})^{-1}$ is the filter gain.

Procedure 1 [19]. Assume that an initial estimate $\hat{R}^{(1)}$ of R is available. Subsequent estimates, $\hat{R}^{(u)}$, $u > 1$, are calculated by repeating the following two-step procedure.

"There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know." *Donald Henry Rumsfeld*

Step 1. Operate the Kalman filter (29) – (30) designed with $\hat{R}^{(u)}$ to obtain corrected output estimates $\hat{y}_{k/k}^{(u)}$.

Step 2. For $i = 1, \ldots, p$, use $\hat{y}_{k/k}^{(u)}$ instead of y_k within (27) to obtain $\hat{R}^{(u+1)} = \mathrm{diag}((\hat{\sigma}_{1,v}^{(u+1)})^2, (\hat{\sigma}_{2,v}^{(u+1)})^2, \ldots, (\hat{\sigma}_{p,v}^{(u+1)})^2)$.

8.3.2.2 Properties

The above EM algorithm involves a repetition of two steps: the states are deduced using the current variance estimates and then the variances are re-identified from the latest states. Consequently, a two-part argument is employed to establish the monotonicity of the variance sequence. For the expectation step, it is shown that monotonically non-increasing variance iterates lead to monotonically non-increasing error covariances. Then for the maximisation step, it is argued that monotonic error covariances result in a monotonic measurement noise variance sequence. The design Riccati difference equation (28) can be written as

$$P_{k+1/k}^{(u)} = (A - K_k^{(u)}C)P_{k/k-1}^{(u)}(A - K_k^{(u)}C)^T + K_k^{(u)}R(K_k^{(u)})^T + Q + S_k^{(u)}, \tag{31}$$

where $S_k^{(u)} = K_k^{(u)}(\hat{R}^{(u)} - R)(K_k^{(u)})^T$ accounts for the presence of parameter error. Subtracting x_k from $\hat{x}_{k/k}^{(u)}$ yields

$$\tilde{x}_{k/k}^{(u)} = (I - L_k^{(u)}C)\tilde{x}_{k/k-1}^{(u)} - L_k^{(u)}v_k, \tag{32}$$

where $\tilde{x}_{k/k}^{(u)} = x_k - \hat{x}_{k/k}^{(u)}$ and $\tilde{x}_{k/k-1}^{(u)} = x_k - \hat{x}_{k/k-1}^{(u)}$ are the corrected and predicted state errors, respectively. The observed corrected error covariance is defined as $\Sigma_{k/k}^{(u)} = E\{\tilde{x}_{k/k}^{(u)}(\tilde{x}_{k/k}^{(u)})^T\}$ and obtained from

$$\begin{aligned}\Sigma_{k/k}^{(u)} &= (I - L_k^{(u)}C)\Sigma_{k/k-1}^{(u)}(I - L_k^{(u)}C)^T + L_k^{(u)}R(L_k^{(u)})^T \\ &= \Sigma_{k/k-1}^{(u)} - \Sigma_{k/k-1}^{(u)}C^T(C\Sigma_{k/k-1}^{(u)}C^T + R)^{-1}C\Sigma_{k/k-1}^{(u)},\end{aligned} \tag{33}$$

where $\Sigma_{k/k-1}^{(u)} = E\{\tilde{x}_{k/k-1}^{(u)}(\tilde{x}_{k/k-1}^{(u)})^T\}$. The observed predicted state error satisfies

$$\tilde{x}_{k+1/k}^{(u)} = A\tilde{x}_{k/k}^{(u)} + Bw_k. \tag{34}$$

Hence, the observed predicted error covariance obeys the recursion

$$\Sigma_{k+1/k}^{(u)} = A\Sigma_{k/k}^{(u)}A^T + BQB^T. \tag{35}$$

Some observations concerning the above error covariances are described below. These results are used subsequently to establish the monotonicity of the above EM algorithm.

"I want minimum information given with maximum politeness." *Jacqueline (Jackie) Lee Bouvier Kennedy Onassis*

Lemma 1 [19]: *In respect of Procedure 1 for estimating R, suppose the following:*

(i) the data z_k has been generated by (23) – (25) in which A, B, C, Q are known, $|\lambda_i(A)| < 1$, i = 1, …, n, and the pair (A, C) is observable;

(ii) there exist $P_{1/0}^{(2)} \leq P_{1/0}^{(1)}$ and $R \leq \hat{R}^{(2)} \leq \hat{R}^{(1)}$ (or $P_{1/0}^{(1)} \leq P_{1/0}^{(2)}$ and $\hat{R}^{(1)} \leq \hat{R}^{(2)} \leq R$).

Then:

(i) $\Sigma_{k+1/k}^{(u)} \leq P_{k+1/k}^{(u)}$;

(ii) $\Sigma_{k/k}^{(u)} \leq P_{k/k}^{(u)}$;

(iii) $R \leq \hat{R}^{(u+1)} \leq \hat{R}^{(u)}$ *implies* $P_{k+1/k}^{(u+1)} \leq P_{k+1/k}^{(u)}$ *(or* $\hat{R}^{(u)} \leq \hat{R}^{(u+1)} \leq R$ *implies* $P_{k+1/k}^{(u)} \leq P_{k+1/k}^{(u+1)}$ *)*

for all $u \geq 1$.

Proof:

(i) *Condition (i) ensures that the problem is well-posed. Condition (ii) stipulates that* $S_k^{(1)} \geq 0$, *which is the initialisation step for an induction argument. For the inductive step, subtracting (33) from (31) yields* $P_{k+1/k}^{(u)} - \Sigma_{k+1/k}^{(u)} = (A - K_k^{(u)}C)(P_{k/k-1}^{(u)} - \Sigma_{k/k-1}^{(u)})(A - K_k^{(u)}C)^T + S_k^{(u)}$ *and thus* $\Sigma_{k/k-1}^{(u)} \leq P_{k/k-1}^{(u)}$ *implies* $\Sigma_{k+1/k}^{(u)} \leq P_{k+1/k}^{(u)}$.

(ii) *The result is immediate by considering A = I within the proof for (i).*

(iii) *The condition* $\hat{R}^{(u+1)} \leq \hat{R}^{(u)}$ *ensures that* $\begin{bmatrix} Q & A^T \\ A & -C^T(\hat{R}^{(m+1)})^{-1}C \end{bmatrix} \leq \begin{bmatrix} Q & A^T \\ A & -C^T(\hat{R}^{(m)})^{-1}C \end{bmatrix}$, *which together with* $P_{1/0}^{(u+1)} \leq P_{1/0}^{(u)}$ *within Theorem 2 of Chapter 7 results in* $P_{k+1/k}^{(u+1)} \leq P_{k+1/k}^{(u)}$. □

Thus the sequences of observed prediction and correction error covariances are bounded above by the design prediction and correction error covariances. Next, it is shown that the observed error covariances are monotonically non-increasing (or non-decreasing).

Lemma 2 [19]: *Under the conditions of Lemma 1:*

i) $\Sigma_{k+1/k}^{(u+1)} \leq \Sigma_{k+1/k}^{(u)}$ *(or* $\Sigma_{k+1/k}^{(u)} \leq \Sigma_{k+1/k}^{(u+1)}$ *) and*

ii) $\Sigma_{k/k}^{(u+1)} \leq \Sigma_{k/k}^{(u)}$ *(or* $\Sigma_{k/k}^{(u)} \leq \Sigma_{k/k}^{(u+1)}$ *).*

Proof: *To establish that the solution of (33) is monotonic non-increasing, from Theorem 2 of Chapter 7, it is required to show that*

$$\begin{bmatrix} Q + K_k^{(u+1)}R(K_k^{(u+1)})^T & (A - K_k^{(u+1)}C)^T \\ A - K_k^{(u+1)}C & 0 \end{bmatrix} \leq \begin{bmatrix} Q + K_k^{(u)}R(K_k^{(u)})^T & (A - K_k^{(u)}C)^T \\ A - K_k^{(u)}C & 0 \end{bmatrix}.$$

"Technology is so much fun but we can drown in our technology. The fog of information can drive out knowledge." *Daniel Joseph Boostin*

Since A, Q and R are time-invariant, it suffices to show that

$$\begin{bmatrix} L_k^{(u+1)}(L_k^{(u+1)})^T & (I-L_k^{(u+1)}C)^T \\ I-L_k^{(u+1)}C & 0 \end{bmatrix} \leq \begin{bmatrix} L_k^{(u)}(L_k^{(u)})^T & (I-L_k^{(u)}C)^T \\ I-L_k^{(u)}C & 0 \end{bmatrix}. \tag{36}$$

Note for an X and Y satisfying $I \geq Y \geq X \geq 0$ that $YY^T - XX^T \geq (I-X)(I-X)^T - (I-Y)(I-Y)^T$. Therefore, $\hat{R}^{(u+1)} \leq \hat{R}^{(u)}$ and $P_{k+1/k}^{(u+1)} \leq P_{k+1/k}^{(u)}$ (from Lemma 1) imply $L^{(u+1)}C \leq L^{(u)}C \leq I$ and thus (36) follows. □

It is established below that monotonic non-increasing error covariances result in a monotonic non-increasing measurement noise variance sequence.

Lemma 3 [19]: *In respect of Procedure 1 for estimating R, suppose the following:*

(i) the data z_k has been generated by (23) – (25) in which A, B, C, Q are known, $|\lambda_i(A)| < 1$, $i = 1, \ldots, n$ and the pair (A, C) is observable;

(ii) there exist $\hat{R}^{(1)} \geq R \geq 0$ and $P_{1/0}^{(u+1)} \leq P_{1/0}^{(u)}$ (or $P_{1/0}^{(u)} \leq P_{1/0}^{(u+1)}$) for all $u > 1$.

Then $\hat{R}^{(u+1)} \leq \hat{R}^{(u)}$ (or $\hat{R}^{(u)} \leq \hat{R}^{(u+1)}$) for all $u > 1$.

Proof*: Let C_i denote the i^{th} row of C. The approximate MLE within Procedure 1 is written as*

$$(\hat{\sigma}_{i,v}^{(u+1)})^2 = \frac{1}{N}\sum_{k=1}^{N}(z_{i,k} - C_i\hat{x}_{k/k}^{(u)})^2 \tag{37}$$

$$= \frac{1}{N}\sum_{k=1}^{N}(C_i\tilde{x}_{k/k}^{(u)} + v_{i,k})^2 \tag{38}$$

$$= C_i\Sigma_{k/k}^{(u)}C_i^T + \sigma_{i,v}^2 \tag{39}$$

and thus $\hat{R}^{(u+1)} = C\Sigma_{k/k}^{(u)}C^T + R$. Since $\hat{R}^{(u+1)}$ is affine to $\Sigma_{k/k}^{(u)}$, which from Lemma 2 is monotonically non-increasing, it follows that $\hat{R}^{(u+1)} \leq \hat{R}^{(u)}$. □

If the estimation problem is dominated by measurement noise, the measurement noise MLEs converge to the actual values.

Lemma 4 [19]: *Under the conditions of Lemma 3,*

$$\lim_{Q\to 0, R^{-1}\to 0, u\to\infty} \hat{R}^{(u+1)} = R\,. \tag{40}$$

"Getting information off the internet is like taking a drink from a fire hydrant." *Mitchell David Kapor*

Proof: *By inspection of* $L_k^{(u)} = P_{k/k-1}^{(u)} C^T (C P_{k/k-1}^{(u)} C^T + R^{(u)})^{-1}$, *it follows that* $\lim_{Q \to 0, R^{-1} \to 0, u \to \infty} L_k^{(u)} = 0$. *Therefore,* $\lim_{Q \to 0, R^{-1} \to 0, u \to \infty} \hat{x}_{k/k}^{(u)} = 0$ *and* $\lim_{Q \to 0, R^{-1} \to 0} z_k = v_k$, *which implies (40), since the MLE (37) is unbiased for large N.* □

Example 9. In respect of the problem (23) – (25), assume $A = 0.9$, $B = C = 1$ and $\sigma_w^2 = 0.1$ are known. Suppose that $\sigma_v^2 = 10$ but is unknown. Samples z_k and $\hat{x}_{k/k}^{(u)}$ were generated from $N = 20{,}000$ realisations of zero-mean Gaussian w_k and v_k. The sequence of MLEs obtained using Procedure 1, initialised with $(\hat{\sigma}_v^{(1)})^2 = 14$ and 12 are indicated by traces (i) and (ii) of Fig. 1, respectively. The variance sequences are monotonically decreasing, which is consistent with Lemma 3. The figure shows that the MLEs converge (to a local maximum of the approximate log-likelihood function), and are reasonably close to the actual value of $\sigma_v^2 = 10$. This illustrates the high measurement noise observation described by Lemma 4. An alternative to the EM algorithm involves calculating MLEs using the Newton-Raphson method [5], [6]. The calculated Newton-Raphson measurement noise variance iterates, initialised with $(\hat{\sigma}_v^{(1)})^2 = 14$ and 12 are indicated by traces (iii) and (iv) of Fig. 1, respectively. It can be seen that the Newton-Raphson estimates converge to those of the EM algorithm, albeit at a slower rate.

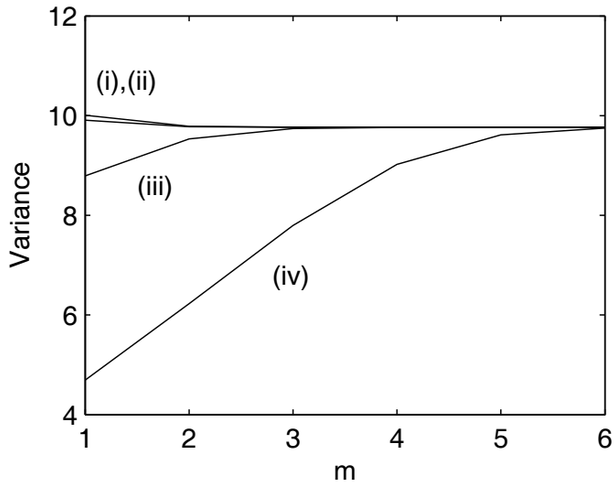


Figure 1. Variance MLEs (27) versus iteration number for Example 9: (i) EM algorithm with $(\hat{\sigma}_v^{(1)})^2 = 14$, (ii) EM algorithm with $(\hat{\sigma}_v^{(1)})^2 = 12$, (iii) Newton-Raphson with $(\hat{\sigma}_v^{(1)})^2 = 14$ and (iv) Newton-Raphson with $(\hat{\sigma}_v^{(1)})^2 = 12$.

"The Internet is the world's largest library. It's just that all the books are on the floor." *John Allen Paulos*

8.3.3 Process Noise Variance Estimation

8.3.3.1 EM Algorithm

In respect of the model (23), suppose that it is desired to estimate Q given N samples of x_{k+1}. The vector states within (23) can be written in terms of its i components, $x_{i,k+1} = A_i x_k + w_{i,k}$, that is

$$w_{i,k} = A_i x_k - x_{i,k+1}, \tag{41}$$

where $w_{i,k} = B_i w_k$, A_i and B_i refer the i^{th} row of A and B, respectively. Assume that $w_{i,k} \sim \mathcal{N}(0, \sigma_{i,w}^2)$, where $\sigma_{i,w}^2 \in \mathbb{R}$ is to be estimated. An MLE for the scalar $\sigma_{i,w}^2 = B_i Q B_i^T$ can be calculated from the sample variance formula

$$\sigma_{i,w}^2 = \frac{1}{N}\sum_{k=1}^{N} w_{i,k} w_{i,k}^T \tag{42}$$

$$= \frac{1}{N}\sum_{k=1}^{N} (x_{i,k+1} - A_i x_k)(x_{i,k+1} - A_i x_k)^T \tag{43}$$

$$= \frac{1}{N}\sum_{k=1}^{N} B_i w_k w_k^T B_i^T \tag{44}$$

$$= B_i \left(\frac{1}{N}\sum_{k=1}^{N} w_k w_k^T \right) B_i^T . \tag{45}$$

Substituting $w_k = Ax_k - x_{k+1}$ into (45) and noting that $\sigma_{i,w}^2 = B_i Q B_i^T$ yields

$$\hat{Q} = \frac{1}{N}\sum_{k=1}^{N} (Ax_k - x_{k+1})(Ax_k - x_{k+1})^T , \tag{46}$$

which can be updated as follows.

Procedure 2 [19]. Assume that an initial estimate $\hat{Q}^{(1)}$ of Q is available. Subsequent estimates can be found by repeating the following two-step algorithm.

Step 1. Operate the filter recursions (29) designed with $\hat{Q}^{(u)}$ on the measurements (25) over $k \in [1, N]$ to obtain corrected state estimates $\hat{x}_{k/k}^{(u)}$ and $\hat{x}_{k+1/k+1}^{(u)}$.

Step 2. For $i = 1, \ldots, n$, use $\hat{x}_{k/k}^{(u)}$ and $\hat{x}_{k+1/k+1}^{(u)}$ instead of x_k and x_{k+1} within (46) to obtain $\hat{Q}^{(u+1)} = \text{diag}((\hat{\sigma}_{1,w}^{(u+1)})^2, (\hat{\sigma}_{2,w}^{(u+1)})^2, \ldots, (\hat{\sigma}_{n,w}^{(u+1)})^2)$.

"Information on the Internet is subject to the same rules and regulations as conversations at a bar."
George David Lundberg

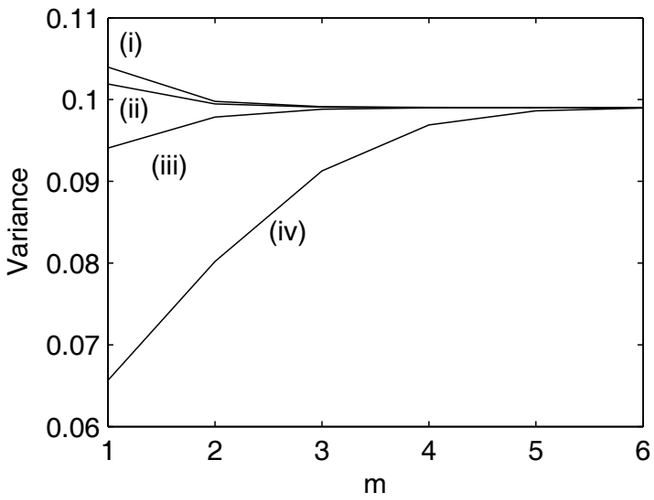


Figure 2. Variance MLEs (46) versus iteration number for Example 10: (i) EM algorithm with $(\hat{\sigma}_w^{(1)})^2$ = 0.14, (ii) EM algorithm with $(\hat{\sigma}_w^{(1)})^2$ = 0.12, (iii) Newton-Raphson with $(\hat{\sigma}_w^{(1)})^2$ = 0.14 and (iv) Newton-Raphson with $(\hat{\sigma}_w^{(1)})^2$ = 0.12.

8.3.3.2 Properties

Similarly to Lemma 1, it can be shown that a monotonically non-increasing (or non-decreasing) sequence of process noise variance estimates results in a monotonically non-increasing (or non-decreasing) sequence of design and observed error covariances, see [19]. The converse case is stated below, namely, the sequence of variance iterates is monotonically non-increasing, provided the estimates and error covariances are initialized appropriately. The accompanying proof makes use of

$$\begin{aligned}\hat{x}_{k+1/k+1}^{(u)} - A\hat{x}_{k/k}^{(u)} &= \hat{x}_{k+1/k}^{(u)} + L_{k+1}^{(u)}(z_{k+1} - C\hat{x}_{k+1/k}^{(u)}) - A\hat{x}_{k/k}^{(u)} \\ &= A\hat{x}_{k/k}^{(u)} + L_{i,k+1}(z_{k+1} - C\hat{x}_{k+1/k}^{(u)}) - A\hat{x}_{k/k}^{(u)} \\ &= L_k^{(u)}(C\tilde{x}_{k+1/k}^{(u)} + v_{k+1}) .\end{aligned} \tag{47}$$

The components of (47) are written as

$$\hat{x}_{i,k+1/k+1}^{(u)} - a_i\hat{x}_{k/k}^{(u)} = L_{i,k+1}^{(u)}(C\tilde{x}_{k+1/k}^{(u)} + v_{k+1}) , \tag{48}$$

where $L_{i,k+1}^{(u)}$ is the i^{th} row of $L_{k+1}^{(u)}$.

"I must confess that I've never trusted the Web. I've always seen it as a coward's tool. Where does it live? How do you hold it personally responsible? Can you put a distributed network of fibre-optic cable on notice? And is it male or female? In other words, can I challenge it to a fight?" *Stephen Tyrone Colbert*

Lemma 5 [19]: *In respect of Procedure 2 for estimating Q, suppose the following:*

(i) the data z_k has been generated by (23) – (25) in which A, B, C, R are known, $|\lambda_i(A)| < 1$, i = 1, ..., n and the pair (A, C) is observable;

(ii) there exist $\hat{Q}^{(1)} \geq Q \geq 0$ and $P_{1/0}^{(u+1)} \leq P_{1/0}^{(u)}$ (or $P_{1/0}^{(u)} \leq P_{1/0}^{(u+1)}$) for all u > 1.

Then $\hat{Q}^{(u+1)} \leq \hat{Q}^{(u)}$ (or $\hat{Q}^{(u)} \leq \hat{Q}^{(u+1)}$) for all u > 1.

Proof*: Using (47)within (46) gives*

$$\begin{aligned}(\hat{\sigma}_{i,w}^{(u)})^2 &= \frac{1}{N} L_{i,k+1}^{(u)} \left(\sum_{k=1}^{N} C\tilde{x}_{k+1/k} + v_{k+1} \right)^2 (L_{i,k+1}^{(u)})^T \\ &= L_{i,k+1}^{(u)} (C\Sigma_{k+1/k}^{(u)} C^T + R)(L_{i,k+1}^{(u)})^T \end{aligned} \tag{49}$$

and thus $\hat{Q}^{(u+1)} = L_{k+1}^{(u)}(C\Sigma_{k+1/k}^{(u)}C^T + R)(L_{k+1}^{(u)})^T$. Since $\hat{Q}^{(u+1)}$ varies with $L_{k+1}^{(u)}(L_{j,k+1}^{(u)})^T$ and $\Sigma_{k+1/k}^{(u)}$, which from Lemma 2 are monotonically non-increasing, it follows that $\hat{Q}^{(u+1)} \leq \hat{Q}^{(u)}$. □

It is observed that the approximate MLEs asymptotically approach the actual values when the SNR is sufficiently high.

Lemma 6 [19]: *Under the conditions of Lemma 5,*

$$\lim_{Q^{-1} \to 0, R \to 0, u \to \infty} \hat{Q}^{(u+1)} = Q \,. \tag{50}$$

Proof: *It is straight forward to show that $\lim_{Q^{-1} \to 0, R \to 0} L_{i,k} C = I$ and therefore $\lim_{Q^{-1} \to 0, R \to 0, u \to \infty} \hat{x}_{k/k}^{(u)} = x_k$, which implies (50), since the MLE (46) is unbiased for large N.* □

Example 10. For the model described in Example 8, suppose that $\sigma_v^2 = 0.01$ is known, and $(\hat{\sigma}_w^{(1)})^2 = 0.1$ but is unknown. Procedure 2 and the Newton-Raphson method [5], [6] were used to jointly estimate the states and the unknown variance. Some example variance iterations, initialised with $(\hat{\sigma}_w^{(1)})^2 = 0.14$ and 0.12, are shown in Fig. 2. The EM algorithm estimates are seen to be monotonically decreasing, which is in agreement with Lemma 5. At the final iteration, the approximate MLEs do not quite reach the actual value of $(\hat{\sigma}_w^{(1)})^2 = 0.1$, because the presence of measurement noise results in imperfect state reconstruction and introduces a small bias (see Example 5). The figure also shows that MLEs calculated via the Newton-Raphson method converge at a slower rate.

"Four years ago nobody but nuclear physicists had ever heard of the Internet. Today even my cat, Socks, has his own web page. I'm amazed at that. I meet kids all the time, been talking to my cat on the Internet." *William Jefferson (Bill) Clinton*

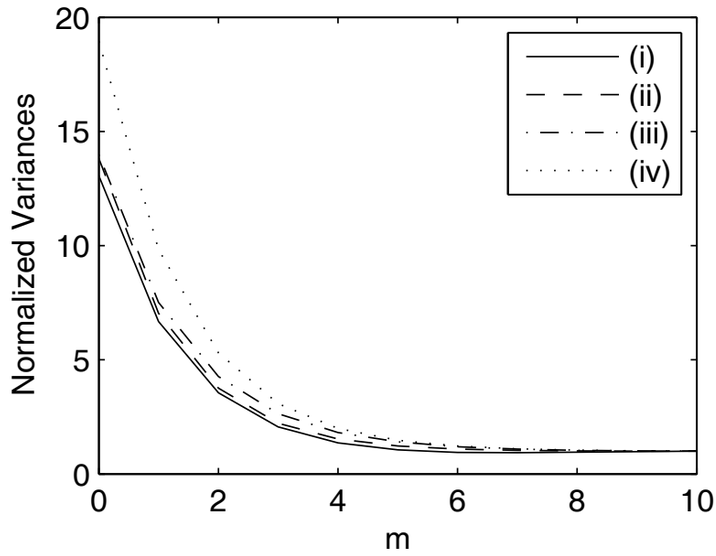


Figure 3. (i) $\hat{\sigma}^2_{1,w}$, (ii) $\hat{\sigma}^2_{2,w}$, (iii) $\hat{\sigma}^2_{3,w}$ and (iv) $\hat{\sigma}^2_{4,w}$, normalised by their steady state values, versus EM iteration number for Example 11.

Example 11. Consider the problem of calculating the initial alignment of an inertial navigation system. Alignment is the process of estimating the Earth rotation rate and rotating the attitude direction cosine matrix, so that it transforms the body-frame sensor signals to a locally-level frame, wherein certain components of accelerations and velocities approach zero when the platform is stationary. This can be achieved by a Kalman filter that uses the model (23), where $x_k \in \mathbb{R}^4$ comprises the errors in earth rotation rate, tilt, velocity and position vectors respectively, and $w_k \in \mathbb{R}^4$ is a deterministic signal which is a nonlinear function of the states (see [24]). The state matrix is calculated as $A = I + \Phi T_s + \frac{1}{2!}(\Phi T_s)^2 + \frac{1}{3!}(\Phi T_s)^3$, where T_s is the sampling interval, $\Phi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & g & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ is a continuous-time state matrix and g is the universal gravitational constant. The output mapping within (24) is $C = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$. Raw three-axis accelerometer and gyro data was recorded from a stationary Litton LN270 Inertial Navigation System at a 500 Hz data rate. In order to generate a compact plot, the initial variance estimates were selected to be 10 times the steady state values.

"On the Internet, nobody knows you're a dog." *Peter Steiner*

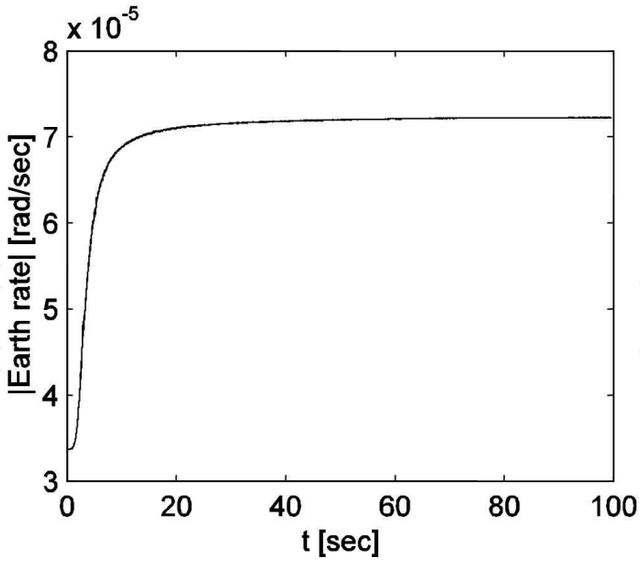


Figure 4. Estimated magnitude of Earth rotation rate for Example 11.

The estimated variances after 10 EM iterations are shown in Fig. 3. The figure demonstrates that approximate MLEs (46) approach steady state values from above, which is consistent with Lemma 5. The estimated Earth rotation rate magnitude versus time is shown in Fig. 4. At 100 seconds, the estimated magnitude of the Earth rate is 72.53 micro-radians per second, that is, one revolution every 24.06 hours. This estimated Earth rate is about 0.5% in error compared with the mean sidereal day of 23.93 hours [25]. Since the estimated Earth rate is in reasonable agreement, it is suggested that the MLEs for the unknown variances are satisfactory (see [19] for further details).

8.3.4 State Matrix Estimation

8.3.4.1 EM Algorithm

The components of the states within (23) are now written as

$$x_{i,k+1} = \sum_{j=1}^{n} a_{i,j} x_{i,k} + w_{i,k}, \tag{51}$$

where $a_{i,j}$ denotes the element in row i and column j of A. Consider the problem of estimating $a_{i,j}$ from samples of $x_{i,k}$. The assumption $x_{i,k+1} \sim \mathcal{N}(\sum_{j=1}^{n} a_{i,j} x_{i,k}, \ \sigma_{i,w}^2)$, leads to the log-likelihood

$$\log f(a_{i,j}) \mid x_{j,k+1}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma_{i,w}^2 - \frac{1}{2}\sigma_{j,w}^{-2} \sum_{k=1}^{N} \left(x_{i,k+1} - \sum_{j=1}^{n} a_{i,j} x_{i,k} \right)^2. \tag{52}$$

"It's important for us to explain to our nation that life is important. It's not only the life of babies, but it's life of children living in, you know, the dark dungeons of the internet." *George Walker Bush*

By setting $\frac{\partial \log f(a_{i,j}) \mid x_{j,k+1})}{\partial a_{i,j}} = 0$, an MLE for $a_{i,j}$ is obtained as [20]

$$\hat{a}_{i,j} = \frac{\sum_{k=1}^{N}\left(x_{i,k+1} - \sum_{j=1, j\neq i}^{n} a_{i,j} x_{i,k} \right) x_{j,k}}{\sum_{k=1}^{N} x_{j,k}^2}. \tag{53}$$

Incidentally, the above estimate can also be found using the least-squares method [2], [10] and minimising the cost function $\sum_{k=1}^{N}\left(x_{i,k+1} - \sum_{j=1}^{n} a_{i,j} x_{i,k} \right)^2$. The expectation of $\hat{a}_{i,j}$ is [20]

$$E\{\hat{a}_{i,j}\} = E\left\{ \frac{\sum_{k=1}^{N}\left(\sum_{j=1}^{n} a_{i,j} x_{i,k} + w_{i,k} - \sum_{j=1, j\neq i}^{n} a_{i,j} x_{i,k} \right) x_{j,k}}{\sum_{k=1}^{N} x_{j,k}^2} \right\}$$

$$= a_{i,j} + E\left\{ \frac{\sum_{k=1}^{N} w_{i,k} x_{j,k}}{\sum_{k=1}^{N} x_{j,k}^2} \right\}$$

$$= a_{i,j},$$

Since $w_{i,k}$ and $x_{i,k}$ are independent. Hence, the MLE (53) is unbiased.

Suppose that an estimate $\hat{A}^{(u)} = \{a_{i,j}^{(u)}\}$ of A is available at an iteration u. The predicted state estimates within (29) can be calculated from

$$\hat{x}_{k+1/k}^{(u)} = (\hat{A}^{(u)} - K_k^{(u)} C)\hat{x}_{k+1/k}^{(u)} + K_k^{(u)} z_k, \tag{54}$$

where $K_k^{(u)} = \hat{A}^{(u)} P_{k/k-1}^{(u)} C^T (C P_{k/k-1}^{(u)} C^T + R)^{-1}$, in which $P_{k/k-1}^{(u)}$ is obtained from the design Riccati equation

$$P_{k+1/k}^{(u)} = (\hat{A}^{(u)} - K_k^{(u)} C) P_{k/k-1}^{(u)} (\hat{A}^{(u)} - K_k^{(u)} C)^T + K_k^{(u)} R (K_k^{(u)})^T + Q. \tag{55}$$

An approximate MLE for $a_{i,j}$ is obtained by replacing x_k by $\hat{x}_{k/k}^{(u)}$ within (53) which results in

$$\hat{a}_{i,j}^{(u+1)} = \frac{\sum_{k=1}^{N}\left(\hat{x}_{i,k+1/k+1}^{(u)} - \sum_{j=1, j\neq i}^{n} \hat{a}_{i,j}^{(u)} \hat{x}_{i,k/k}^{(u)} \right) \hat{x}_{j,k/k}^{(u)}}{\sum_{k=1}^{N} (\hat{x}_{j,k/k}^{(u)})^2}. \tag{56}$$

"The internet is like a gold-rush; the only people making money are those who sell the pans." *Will Hobbs*

An iterative procedure for re-estimating an unknown state matrix is proposed below.

Procedure 3 [20]. Assume that there exists an initial estimate $\hat{A}^{(1)}$ satisfying $|\lambda_i(\hat{A}^{(1)})| < 1$, $i = 1, \ldots, n$. Subsequent estimates are calculated using the following two-step EM algorithm.

Step 1. Operate the Kalman filter (29) using (54) on the measurements z_k over $k \in [1, N]$ to obtain corrected state estimates $\hat{x}_{k/k}^{(u)}$ and $\hat{x}_{k+1/k+1}^{(u)}$.

Step 2. Copy $\hat{A}^{(u)}$ to $\hat{A}^{(u+1)}$. Use $\hat{x}_{k/k}^{(u)}$ within (56) to obtain candidate estimates $\hat{a}_{i,j}^{(u+1)}$, $i, j = 1, \ldots, n$. Include $\hat{a}_{i,j}^{(u+1)}$ within $\hat{A}^{(u+1)}$ if $|\lambda_i(\hat{A}^{(u+1)})| < 1$, $i = 1, \ldots, n$.

The condition $|\lambda_i(\hat{A}^{(u+1)})| < 1$ within Step 2 ensures that the estimated system is asymptotically stable.

8.3.4.2 Properties

The design Riccati difference equation (55) can be written as

$$P_{k+1/k}^{(u)} = (A - K_k^{(u)}C)P_{k/k-1}^{(u)}(A - K_k^{(u)}C)^T + K_k^{(u)}R(K_k^{(u)})^T + Q + S_k^{(u)}, \tag{57}$$

where

$$S_k^{(u)} = (\hat{A}^{(u)} - K_k^{(u)}C)P_{k/k-1}^{(u)}(\hat{A}^{(u)} - K_k^{(u)}C)^T - (A - K_k^{(u)}C)P_{k/k-1}^{(u)}(A - K_k^{(u)}C)^T \tag{58}$$

accounts for the presence of modelling error. In the following, the notation of Lemma 1 is employed to argue that a monotonically non-increasing state matrix estimate sequence results in monotonically non-increasing error covariance sequences.

Lemma 7 [20]. *In respect of Procedure 3 for estimating A, suppose the following:*

(i) the data z_k has been generated by (23) – (25) in which B, C, Q, R are known;

(ii) $|\lambda_i(\hat{A}^{(1)})| < 1$, $i = 1, \ldots, n$, the pair (A, C) is observable;

(iii) there exist $\hat{A}^{(1)} \geq A$ and $P_{1/0}^{(u+1)} \leq P_{1/0}^{(u)}$ (or $P_{1/0}^{(u)} \leq P_{1/0}^{(u+1)}$) for all $u > 1$.

Then:

(i) $\Sigma_{k+1/k}^{(u)} \leq P_{k+1/k}^{(u)}$ (or $P_{k+1/k}^{(u)} \leq \Sigma_{k+1/k}^{(u)}$);

(ii) $\Sigma_{k/k}^{(u)} \leq P_{k/k}^{(u)}$ (or $P_{k/k}^{(u)} \leq \Sigma_{k/k}^{(u)}$);

(iii) $\hat{A}^{(u+1)} \leq \hat{A}^{(u)}$ which implies $P_{k+1/k}^{(u+1)} \leq P_{k+1/k}^{(u)}$ ($\hat{A}^{(u)} \leq \hat{A}^{(u+1)}$ which implies $P_{k+1/k}^{(u)} \leq P_{k+1/k}^{(u+1)}$)

for all $u \geq 1$.

"It may not always be profitable at first for businesses to be online, but it is certainly going to be unprofitable not to be online." *Ester Dyson*

The proof follows *mutatis mutandis* from that of Lemma 1. A heuristic argument is outlined below which suggests that non-increasing error variances lead to a non-increasing state matrix estimate sequence. Suppose that there exists a residual error $s_k^{(u)} \in \mathbb{R}^n$ at iteration u such that

$$\hat{x}_{k+1/k+1}^{(u)} = \hat{A}^{(u)}\hat{x}_{k/k}^{(u)} + s_k^{(u)} . \tag{59}$$

The components of (59) are denoted by

$$\hat{x}_{i,k+1/k+1}^{(u)} = \sum_{j=1}^{n} a_{i,j}^{(u)}\hat{x}_{i,k/k}^{(u)} + s_{i,k}^{(u)} , \tag{60}$$

where $s_{i,k}^{(u)}$ is the i^{th} element of $s_k^{(u)}$. It follows from (60) and (48) that

$$s_k^{(u)} = L_k^{(u)}(C\tilde{x}_{k+1/k}^{(u)} + v_{k+1}) \tag{61}$$

and

$$s_{i,k}^{(u)} = L_{i,k}^{(u)}(C\tilde{x}_{k+1/k}^{(u)} + v_{k+1}) . \tag{62}$$

Using (61) and (63) within (57) yields

$$\begin{aligned} \hat{a}_{i,j}^{(u+1)} &= \hat{a}_{i,j}^{(u)} + \left(\sum_{k=1}^{N} s_{i,k}^{(u)}\hat{x}_{j,k/k}^{(u)}\right)\left(\sum_{k=1}^{N}(\hat{x}_{j,k/k}^{(u)})^2\right)^{-1} \\ &= \hat{a}_{i,j}^{(u)} + L_{i,k}^{(u)}C\left(\sum_{k=1}^{N}(\tilde{x}_{k+1/k}^{(u)} + C^{\#}v_{k+1})\hat{x}_{j,k/k}^{(u)}\right)\left(\sum_{k=1}^{N}(\hat{x}_{j,k/k}^{(u)})^2\right)^{-1} , \end{aligned} \tag{63}$$

where $C^{\#}$ denotes the Moore-Penrose pseudo-inverse of C. It is shown in Lemma 2 under prescribed conditions that $L^{(u+1)}C \leq L^{(u)}C \leq I$. Since the non-increasing sequence $L^{(u)}C$ is a factor of the second term on the right-hand-side of (63), the sequence $\hat{a}_{i,j}^{(u+1)}$ is expected to be non-increasing.

Lemma 8 [20]: *Under the conditions of Lemma 7, suppose that C is full rank, then*

$$\lim_{Q^{-1}\to 0, R\to 0, u\to\infty} \hat{A}^{(u+1)} = A . \tag{64}$$

Proof: *It is straight forward to show that* $\lim_{Q^{-1}\to 0, R\to 0, u\to\infty} L_{i,k}C = I$ *and therefore* $\lim_{Q^{-1}\to 0, R\to 0, u\to\infty} \hat{x}_{k/k}^{(u)} = x_k$, *which implies (64) since the MLE (53) is unbiased.* □

"New scientific ideas never spring from a communal body, however organized, but rather from the head of an individually inspired researcher who struggles with his problems in lonely thought and unites all his thought on one single point which is his whole world for the moment." *Max Karl Ernst Ludwig Planck*

An illustration is presented below.

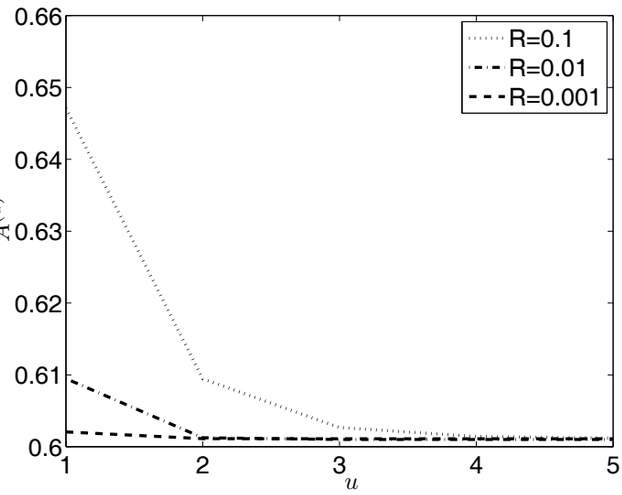


Figure 5. Sequence of $\hat{A}^{(u)}$ versus iteration number for Example 12.

Example 12. In respect of the model (23) – (25), suppose that $B = C = 1$, $\sigma_w^2 = 0.2$ are known and $A = 0.6$ is unknown. Simulations were conducted with 100 realizations of Gaussian process noise and measurement noise of length $N = 500{,}000$ for $R = 0.1$, 0.01 and 0.001. The EM algorithms were initialised with $\hat{A}^{(1)} = 0.9999$. It was observed that the resulting estimate sequences were all monotonically decreasing, however, this becomes imperceptible at $R = 0.001$, due to the limited resolution of the plot. The mean estimates are shown in Fig. 5. As expected from Lemma 8, $\hat{A}^{(u)}$ asymptotically approaches the true value of $A = 0.6$ when the measurement noise becomes negligible.

8.4 Smoothing EM Algorithms

8.4.1 Process Noise Variance Estimation

8.4.1.1 EM Algorithm

In the previous EM algorithms, the expectation step involved calculating filtered estimates. Similar EM procedures are outlined in [26] and here where smoothed estimates are used at iteration u within the expectation step. The likelihood functions described in Sections 8.2.2 and 8.2.3 are exact, provided that the underlying assumptions are correct and actual random variables are available. Under these conditions, the ensuing parameter estimates maximise the likelihood functions and their limit of precision is specified by the associated CRLBs. However, the use of filtered or smoothed quantities leads to approximate likelihood functions, MLEs and CRLBs. It turns out that the approximate MLEs approach the true parameter values under prescribed SNR conditions. It will be shown that the use of

"The best way to prepare is to write programs, and to study great programs that other people have written. In my case, I went to the garbage cans at the Computer Science Center and I fished out listings of their operating system." *William Henry (Bill) Gates III*

smoothed (as opposed to filtered) quantities results in smaller approximate CRLBs, which suggests improved parameter estimation accuracy.

Suppose that the system $\mathcal{G}$ having the realisation (23) – (24) is non-minimum phase and D is of full rank. Under these conditions $\mathcal{G}^{-1}$ exists and the minimum-variance smoother (described in Chapter 7) may be employed to produce input estimates. Assume that an estimate $\hat{Q}^{(u)} = \mathrm{diag}((\hat{\sigma}_{1,w}^{(u)})^2, (\hat{\sigma}_{2,w}^{(u)})^2, \ldots, (\hat{\sigma}_{n,w}^{(u)})^2)$ of Q is are available at iteration u. The smoothed input estimates, $\hat{w}_{k/N}^{(u)}$, are calculated from

$$\begin{bmatrix} x_{k+1/k}^{(u)} \\ \alpha_k^{(u)} \end{bmatrix} = \begin{bmatrix} A_k - K_k^{(u)} C_k & K_k^{(u)} \\ -(\Omega_k^{(u)})^{-1/2} C_k & (\Omega_k^{(u)})^{-1/2} \end{bmatrix} \begin{bmatrix} x_{k/k-1}^{(u)} \\ z_k \end{bmatrix}, \tag{65}$$

$$\begin{bmatrix} \xi_{k-1}^{(u)} \\ \gamma_{k-1}^{(u)} \\ \hat{w}_{k-1/N}^{(u)} \end{bmatrix} = \begin{bmatrix} A_k^T - C_k^T (K_k^{(u)})^T & 0 & C_k^T (\Omega_k^{(u)})^{-1/2} \\ C_k^T (K_k^{(u)})^T & A_k^T & -C_k^T (\Omega_k^{(u)})^{-1/2} \\ -\hat{Q}_k^{(u)} D_k^T (K_k^{(u)})^T & -\hat{Q}_k^{(u)} B_k^T & \hat{Q}_k^{(u)} D_k^T (\Omega_k^{(u)})^{-1/2} \end{bmatrix} \begin{bmatrix} \xi_k^{(u)} \\ \gamma_k^{(u)} \\ \alpha_k^{(u)} \end{bmatrix}, \tag{66}$$

where $K_k^{(u)} = (A_k P_{k/k-1}^{(u)} C_k^T + B_k \hat{Q}_k^{(u)} D_k^T)(\Omega_k^{(u)})^{-1}$, $\Omega_k^{(u)} = C_k P_{k/k-1}^{(u)} C_k^T + D_k \hat{Q}_k^{(u)} D_k^T + R_k$ and $P_{k/k-1}^{(u)}$ evolves from the Riccati difference equation $P_{k+1/k}^{(u)} = A_k P_{k/k-1}^{(u)} A_k^T - (A_k P_{k/k-1}^{(u)} C_k^T + B_k \hat{Q}_k^{(u)} D_k^T)(C_k P_{k/k-1}^{(u)} C_k^T + D_k \hat{Q}_k^{(u)} D_k^T + R_k)^{-1}(C_k P_{k/k-1}^{(u)} A_k^T + D_k Q_k^{(u)} B_k^T) + B_k \hat{Q}_k^{(u)} B_k^T$. A smoothing EM algorithm for iteratively re-estimating $\hat{Q}^{(u)}$ is described below.

Procedure 4. Suppose that an initial estimate $\hat{Q}^{(1)} = \mathrm{diag}((\hat{\sigma}_{1,w}^{(1)})^2, (\hat{\sigma}_{2,w}^{(1)})^2, \ldots, (\hat{\sigma}_{n,w}^{(1)})^2)$ is available. Then subsequent estimates $\hat{Q}^{(u)}$, $u > 1$, are calculated by repeating the following two steps.

Step 1. Use $\hat{Q}^{(u)} = \mathrm{diag}((\hat{\sigma}_{1,w}^{(u)})^2, (\hat{\sigma}_{2,w}^{(u)})^2, \ldots, (\hat{\sigma}_{n,w}^{(u)})^2))$ within (65) − (66) to calculate smoothed input estimates $\hat{w}_{k/N}^{(u)}$.

Step 2. Calculate the elements of $\hat{Q}^{(u+1)} = \mathrm{diag}((\hat{\sigma}_{1,w}^{(u+1)})^2, (\hat{\sigma}_{2,w}^{(u+1)})^2, \ldots, (\hat{\sigma}_{n,w}^{(u+1)})^2)$ using $\hat{w}_{k/N}^{(u)}$ from Step 1 instead of w_k within the MLE formula (46).

8.4.1.2 Properties

In the following it is shown that the variance estimates arising from the above procedure result in monotonic error covariances. The additional term within the design Riccati difference equation (57) that accounts for the presence of parameter error is now given by $S_k^{(u)} = B(\hat{Q}^{(u)} - Q)B^T$. Let $\hat{\Delta}^{(u)}$ denote an approximate spectral factor arising in the design of a

"Don't worry about people stealing your ideas. If your ideas are any good, you'll have to ram them down people's throats." *Howard Hathaway Aiken*

smoother using $P_{k/k-1}^{(u)}$ and $K_k^{(u)}$. Employing the notation and approach of Chapter 7, it is straightforward to show that

$$\hat{\Delta}^{(u)}(\hat{\Delta}^{(u)})^H = \Delta\Delta^H + C_k\mathcal{G}_0\,(P_{k/k-1}^{(u)} - P_{k+1/k}^{(u)} + S^{(u)})\mathcal{G}_0^H C_k^T\,. \tag{67}$$

Define the stacked vectors $v = [v_1^T, \ldots, v_K^T]^T$, $w = [w_1^T, \ldots, w_N^T]^T$, $\hat{w}^{(u)} = [(\hat{w}_{1/N}^{(u)})^T, \ldots, (\hat{w}_{N/N}^{(u)})^T]^T$ and $\tilde{w}^{(u)} = w - \hat{w}^{(u)} = [(\tilde{w}_{1/N}^{(u)})^T, \ldots, (\tilde{w}_{N/N}^{(u)})^T]^T$. The input estimation error is generated by $\tilde{w}^{(u)} = \mathcal{R}^{(u)}\begin{bmatrix} v \\ w \end{bmatrix}$, where $\mathcal{R}_{ei}^{(u)}(\mathcal{R}_{ei}^{(u)})^H = \mathcal{R}_{ei1}^{(u)}(\mathcal{R}_{ei1}^{(u)})^H + \mathcal{R}_{ei2}^{(u)}(\mathcal{R}_{ei2}^{(u)})^H$, in which

$$\mathcal{R}_{ei2}^{(u)} = Q\mathcal{G}^H\left((\hat{\Delta}^{(u)}(\hat{\Delta}^{(u)})^H)^{-1} - (\Delta\Delta^H)^{-1}\right)\Delta\,, \tag{68}$$

and $\mathcal{R}_{ei1}^{(u)}(\mathcal{R}_{ei1}^{(u)})^H = Q\mathcal{G}_2^H - Q\mathcal{G}^H(\Delta\Delta^H)^{-1}\mathcal{G}Q$. It is shown in the lemma below that the sequence $\left\|\tilde{w}^{(u)}(\tilde{w}^{(u)})^T\right\|_2 = \left\|\mathcal{R}_{ei}^{(u)}(\mathcal{R}_{ei}^{(u)})^H\right\|_2$ is monotonically non-increasing or monotonically non-decreasing, depending on the initial conditions.

Lemma 9: *In respect of Procedure 4 for estimating Q, suppose the following:*

(i) the system (23) – (24) is non-minimum phase, in which A, B, C, D, R are known, $|\lambda_i(\hat{A}^{(1)})| < 1$, *i = 1, …, n, the pair (A, C) is observable and D is of full rank;*

(ii) the solutions $P_{1/0}^{(1)}$, $P_{1/0}^{(2)}$ *of (57) for* $\hat{Q}^{(2)} \geq \hat{Q}^{(1)}$ *satisfy* $P_{1/0}^{(2)} \leq P_{1/0}^{(1)}$ *(or the solutions* $P_{1/0}^{(1)}$, $P_{1/0}^{(2)}$ *of (57) for* $\hat{Q}^{(1)} \geq \hat{Q}^{(2)}$ *satisfy* $P_{1/0}^{(1)} \leq P_{1/0}^{(2)}$*).*

Then:

(i) $P_{k+1/k}^{(u)} \leq P_{k/k-1}^{(u)}$ *(or* $P_{k/k-1}^{(u)} \leq P_{k+1/k}^{(u)}$*) for all k, u ≥ 1;*

(ii) $P_{k+1/k}^{(u+1)} \leq P_{k+1/k}^{(u)}$ *and* $P_{k/k-1}^{(u+1)} \leq P_{k/k-1}^{(u)}$ *(or* $P_{k+1/k}^{(u)} \leq P_{k+1/k}^{(u+1)}$ *and* $P_{k/k-1}^{(u)} \leq P_{k/k-1}^{(u+1)}$*) for all k, u ≥ 1;*

(iii) $\left\|\mathcal{R}_{ei}^{(u+1)}(\mathcal{R}_{ei}^{(u+1)})^H\right\|_2 \leq \left\|\mathcal{R}_{ei}^{(u)}(\mathcal{R}_{ei}^{(u)})^H\right\|_2$ *(or* $\left\|\mathcal{R}_{ei}^{(u)}(\mathcal{R}_{ei}^{(u)})^H\right\|_2 \leq \left\|\mathcal{R}_{ei}^{(u+1)}(\mathcal{R}_{ei}^{(u+1)})^H\right\|_2$*) for u ≥ 1.*

Proof: *(i) and (ii) This follows from* $S^{(u+1)} \leq S^{(u)}$ *within condition (iii) of Theorem 2 of Chapter 8. Since* $\mathcal{R}_{ei1}^{(u)}(\mathcal{R}_{ei1}^{(u)})^H$ *is common to* $\mathcal{R}_{ei}^{(u)}(\mathcal{R}_{ei}^{(u)})^H$ *and* $\mathcal{R}_{ei}^{(u+1)}(\mathcal{R}_{ei}^{(u+1)})^H$*, it suffices to show that*

$$\left\|\mathcal{R}_{ei2}^{(u+1)}(\mathcal{R}_{ei2}^{(u+1)})^H\right\|_2 \leq \left\|\mathcal{R}_{ei2}^{(u)}(\mathcal{R}_{ei2}^{(u)})^H\right\|_2\,. \tag{69}$$

Substituting (67) into (68) yields

$$\mathcal{R}_{ei2}^{(u)} = Q\mathcal{G}^H\left(\Delta\Delta^H + C_k\mathcal{G}_0\,(P_{k/k-1}^{(u)} - P_{k+1/k}^{(u)} + S^{(u)})\mathcal{G}_0^H C_k^T)^{-1} - (\Delta\Delta^H)^{-1}\right)\Delta\,. \tag{70}$$

"We have always been shameless about stealing great ideas." *Steven Paul Jobs*

Note for linear time-invariant systems X, $Y_1 \geq Y_2$, that

$$(XX^H)^{-1} - (XX^H + Y_1)^{-1} \geq (XX^H)^{-1} - (XX^H + Y_2)^{-1}. \tag{71}$$

Since $\left\| \mathcal{G}_0 \, (P_{k/k-1}^{(u+1)} - P_{k+1/k}^{(u+1)} + S^{(u+1)}) \mathcal{G}_0^H \right\|_2 \leq \left\| \mathcal{G}_0 \, (P_{k/k-1}^{(u)} - P_{k+1/k}^{(u)} + S^{(u)}) \mathcal{G}_0^H \right\|_2$, *(69) follows from (70) and (71).* □

As is the case for the filtering EM algorithm, the process noise variance estimates asymptotically approach the exact values when the SNR is sufficiently high.

Lemma 10: *Under the conditions of Lemma 9,*

$$\lim_{Q^{-1} \to 0, R \to 0, u \to \infty} \hat{Q}^{(u)} = Q. \tag{72}$$

Proof: *By inspection of the input estimator,* $\mathcal{H}_{IE} = Q\mathcal{G}^H(\Delta\Delta^H)^{-1} = Q\mathcal{G}^H(\mathcal{G}Q\mathcal{G}^H + R)^{-1}$, *it follows that* $\lim_{Q^{-1} \to 0, R \to 0, u \to \infty} \mathcal{H}_{IE} = \mathcal{G}^{-1}$ *and therefore* $\lim_{Q^{-1} \to 0, R \to 0, u \to \infty} \hat{w}_{k/N}^{(u)} = w_k$, *which implies (72), since the MLE (46) is unbiased for large N.* □

It is observed anecdotally that the variance estimates produced by the above smoothing EM algorithm are more accurate than those from the corresponding filtering procedure. This is consistent with the following comparison of approximate CRLBs.

Lemma 11 [26]:

$$-\left(\frac{\partial^2 \log f(\sigma_{i,w}^2 \mid \hat{x}_{k/N})}{(\partial \sigma_{i,w}^2)^2} \right)^{-1} < -\left(\frac{\partial^2 \log f(\sigma_{i,w}^2 \mid \hat{x}_{k/k})}{(\partial \sigma_{i,w}^2)^2} \right)^{-1}. \tag{73}$$

Proof: *The vector state elements within (23) can be written in terms of smoothed state estimates,* $x_{i,k+1} = A_i \hat{x}_{k/N} + w_{i,k} = A_i x_k + w_{i,k} - A_i \tilde{x}_{k/N}$, *where* $\tilde{x}_{k/N} = x_k - \hat{x}_{k/N}$. *From the approach of Example 8, the second partial derivative of the corresponding approximate log-likelihood function with respect to the process noise variance is*

$$\frac{\partial^2 \log f(\sigma_{i,w}^2 \mid \hat{x}_{k/N})}{(\partial \sigma_{i,w}^2)^2} = -\frac{N}{2} (\sigma_{i,w}^2 + A_i E\{\tilde{x}_{k/N} \tilde{x}_{k/N}^T\} A_i^T)^{-2}.$$

Similarly, the use of filtered state estimates leads to

$$\frac{\partial^2 \log f(\sigma_{i,w}^2 \mid \hat{x}_{k/k})}{(\partial \sigma_{i,w}^2)^2} = -\frac{N}{2} (\sigma_{i,w}^2 + A_i E\{\tilde{x}_{k/k} \tilde{x}_{k/k}^T\} A_i^T)^{-2}.$$

The minimum-variance smoother minimizes both the causal part and the non-causal part of the estimation error, whereas the Kalman filter only minimizes the causal part. Therefore, $E\{\tilde{x}_{k/N} \tilde{x}_{k/N}^T\} < E\{\tilde{x}_{k/k} \tilde{x}_{k/k}^T\}$. *Thus, the claim (73) follows.* □

"The power of an idea can be measured by the degree of resistance it attracts." *David Yoho*

8.4.2 State Matrix Estimation

8.4.2.1 EM Algorithm

Smoothed state estimates are obtained from the smoothed inputs via

$$\hat{x}_{k+1/N}^{(u)} = A_k \hat{x}_{k/N}^{(u)} + B_k \hat{w}_{k/N}^{(u)} . \tag{74}$$

The resulting $\hat{x}_{k/N}^{(u)}$ are used below to iteratively re-estimate state matrix elements.

Procedure 5. Assume that there exists an initial estimate $\hat{A}^{(1)}$ of A such that $|\lambda_i(\hat{A}^{(1)})| < 1$, i = 1, …, n. Subsequent estimates, $\hat{A}^{(u)}$, $u > 1$, are calculated using the following two-step EM algorithm.

Step 1. Operate the minimum-variance smoother recursions (65), (66), (74) designed with $\hat{A}^{(u)}$ to obtain $\hat{x}_{k/N}^{(u)}$.

Step 2. Copy $\hat{A}^{(u)}$ to $\hat{A}^{(u+1)}$. Use $\hat{x}_{k/N}^{(u)}$ instead of x_k within (53) to obtain candidate estimates $\hat{a}_{i,j}^{(u+1)}$, $i, j = 1, \ldots, n$. Include $\hat{a}_{i,j}^{(u+1)}$ within $\hat{A}^{(u+1)}$ if $|\lambda_i(\hat{A}^{(u+1)})| < 1$, $i = 1, \ldots, n$.

8.4.2.2 Properties

Denote $x = [x_1^T, \ldots, x_N^T]^T$, $\hat{x}^{(u)} = [(\hat{x}_{1/N}^{(u)})^T, \ldots, (\hat{x}_{N/N}^{(u)})^T]^T$ and $\tilde{x}^{(u)} = x - \hat{x}^{(u)} = [(\tilde{x}_{1/N}^{(u)})^T, \ldots, (\tilde{x}_{N/N}^{(u)})^T]^T$. Let $\boldsymbol{\mathcal{R}}^{(u)}$ be redefined as the system that maps the inputs $\begin{bmatrix} v \\ w \end{bmatrix}$ to smoother state estimation error $\tilde{x}^{(u)}$, that is, $\tilde{x}^{(u)} = \boldsymbol{\mathcal{R}}^{(u)} \begin{bmatrix} v \\ w \end{bmatrix}$. It is stated below that the estimated state matrix iterates result in a monotonic sequence of state error covariances.

Lemma 12: *In respect of Procedure 5 for estimating A and x, suppose the following:*

(i) the system (23) – (24) is non-minimum phase, in which B, C, D, Q, R are known, $|\lambda_i(\hat{A}^{(u+1)})| < 1$*, the pair (A, C) is observable and D is of full rank;*

(ii) there exist solutions $P_{1/0}^{(1)}$, $P_{1/0}^{(2)}$ *of (57) for* $AA^T \leq A^{(2)}(A^{(2)})^T \leq A^{(1)}(A^{(1)})^T$ *satisfying* $P_{1/0}^{(2)} \leq P_{1/0}^{(1)}$ *(or the solutions* $P_{1/0}^{(1)}$, $P_{1/0}^{(2)}$ *of (31) for* $A^{(1)}(A^{(1)})^T \leq A^{(2)}(A^{(2)})^T \leq AA^T$ *satisfying* $P_{1/0}^{(1)} \leq P_{1/0}^{(2)}$*).*

Then $\left\| \boldsymbol{\mathcal{R}}_{ei}^{(u+1)} (\boldsymbol{\mathcal{R}}_{ei}^{(u+1)})^H \right\|_2 \leq \left\| \boldsymbol{\mathcal{R}}_{ei}^{(u)} (\boldsymbol{\mathcal{R}}_{ei}^{(u)})^H \right\|_2$ *(or* $\left\| \boldsymbol{\mathcal{R}}_{ei}^{(u)} (\boldsymbol{\mathcal{R}}_{ei}^{(u)})^H \right\|_2 \leq \left\| \boldsymbol{\mathcal{R}}_{ei}^{(u+1)} (\boldsymbol{\mathcal{R}}_{ei}^{(u+1)})^H \right\|_2$ *) for* $u \geq 1$.

"You do not really understand something unless you can explain it to your grandmother." *Albert Einstein*

The proof is omitted since it follows *mutatis mutandis* from that of Lemma 9. Suppose that the smoother (65), (66) designed with the estimates $\hat{a}_{i,j}^{(u)}$ is employed to calculate input estimates $\hat{w}_{k/N}^{(u)}$. An approximate log-likelihood function for the unknown $a_{i,j}$ given samples of $\hat{w}_{k/N}^{(u)}$ is

$$\log f(a_{i,j} \mid \hat{w}_{i,k/K}^{(u)}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log(\sigma_{i,w}^{(u)})^2 - \frac{1}{2}(\sigma_{i,w}^{(u)})^{-2}\sum_{k=1}^{N}\hat{w}_{i,k/K}^{(u)}(\hat{w}_{i,k/N}^{(u)})^T. \quad (75)$$

Now let $\boldsymbol{\mathcal{R}}^{(u)}$ denote the map from $\begin{bmatrix} v \\ w \end{bmatrix}$ to the smoother input estimation error $\tilde{w}^{(u)} = w - \hat{w}^{(u)}$ at iteration u. It is argued below that the sequence of state matrix iterates maximises (75).

Lemma 13: *Under the conditions of Lemma 12,* $\left\|\boldsymbol{\mathcal{R}}^{(u+1)}(\boldsymbol{\mathcal{R}}^{(u+1)})^H\right\|_2 \leq \left\|\boldsymbol{\mathcal{R}}^{(u)}(\boldsymbol{\mathcal{R}}^{(u)})^H\right\|_2$ *for* $u \geq 1$.

The proof follows *mutatis mutandis* from that of Lemma 9. The above Lemma implies

$$E\{\tilde{w}^{(u+1)}(\tilde{w}^{(u+1)})^T\} \leq E\{\tilde{w}^{(u)}(\tilde{w}^{(u)})^T\}. \quad (76)$$

It follows from $\hat{w}^{(u)} = w - \tilde{w}^{(u)}$ that $E\{\tilde{w}^{(u)}(\tilde{w}^{(u)})^T\} = E\{w + \tilde{w}^{(u)})(w + (\tilde{w}^{(u)})^T\} = E\{\tilde{w}^{(u)}(\tilde{w}^{(u)})^T\} + Q$, which together with (76) implies $E\{\hat{w}^{(u+1)}(\hat{w}^{(u+1)})^T\} \leq E\{\hat{w}^{(u)}(\hat{w}^{(u)})^T\}$ and $\log f(a_{i,jw} \mid \hat{w}_{i,k/N}^{(u+1)}) \geq \log f(a_{i,jw} \mid \hat{w}_{i,k/K}^{(u)})$ for all $u \geq 1$. Therefore, it is expected that the sequence of state matrix estimates will similarly vary monotonically. Next, it is stated that the state matrix estimates asymptotically approach the exact values when the SNR is sufficiently high.

Lemma 14: *Under the conditions of Lemma 9,*

$$\lim_{Q^{-1}\to 0, R\to 0, u\to\infty} \hat{A}^{(u)} = A. \quad (77)$$

Proof: *From the proof of Lemma 10,* $\lim_{Q^{-1}\to 0, R\to 0, u\to\infty} \hat{w}_{k/N}^{(u)} = w_k$, *therefore, the states within (74) are reconstructed exactly. Thus, the claim (77) follows since the MLE (53) is unbiased.* □

It is expected that the above EM smoothing algorithm offers improved state matrix estimation accuracy.

Lemma 15:

$$-\left(\frac{\partial^2 \log f(a_{i,j} \mid \hat{x}_{k/N})}{(\partial a_{i,j})^2}\right)^{-1} \leq -\left(\frac{\partial^2 \log f(a_{i,j} \mid \hat{x}_{k/k})}{(\partial a_{i,j})^2}\right)^{-1}. \quad (78)$$

"The test of a first-rate intelligence is the ability to hold two opposed ideas in mind at the same time and still retain the ability to function." *Francis Scott Key Fitzgerald*

Proof: *Using smoothed states within (51) yields* $x_{i,k+1} = \sum_{j=1}^{n} a_{i,j}\hat{x}_{i,k/N} + w_{i,k} = \sum_{j=1}^{n} a_{i,j}x_{i,k} + w_{i,k} - \sum_{j=1}^{n} a_{i,j}\tilde{x}_{i,k/N}$, *where* $\tilde{x}_{k/N} = x_k - \hat{x}_{k/N}$. *The second partial derivative of the corresponding log-likelihood function with respect to* $a_{i,j}$ *is*

$$\frac{\partial^2 \log f(a_{i,j} \mid \hat{x}_{k/N})}{(\partial a_{i,j})^2} = -\frac{N}{2}(\sigma_{i,w}^2 + A_i E\{\tilde{x}_{k/N}\tilde{x}_{k/N}^T\}A_i^T)^{-1}\sum_{k=1}^{N} x_{j,k}^2 .$$

Similarly, the use of filtered state estimates leads to

$$\frac{\partial^2 \log f(a_{i,j} \mid \hat{x}_{k/k})}{(\partial a_{i,j})^2} = -\frac{N}{2}(\sigma_{i,w}^2 + A_i E\{\tilde{x}_{k/k}\tilde{x}_{k/k}^T\}A_i^T)^{-1}\sum_{k=1}^{N} x_{j,k}^2 .$$

The result (78) follows since $E\{\tilde{x}_{k/N}\tilde{x}_{k/N}^T\} < E\{\tilde{x}_{k/k}\tilde{x}_{k/k}^T\}$. □

Example 13.*:* Consider a system where $B = C = D = Q = 1$, $R = \{0.0001, 0.0002, 0.0003\}$ are known and $A = 0.9$ but is unknown. Simulations were conducted using 30 noise realizations with $N = 500{,}000$. The results of the above smoothing EM algorithm and the filtering EM algorithms, initialized with $\hat{A}^{(0)} = 1.03A$, are respectively shown by the dotted and dashed lines within Fig. 6. The figure shows that the estimates improve with increasing u, which is consistent with Lemma 15. The estimates also improve with increasing SNR which illustrates Lemmas 8 and 14. It is observed anecdotally that the smoother EM algorithm outperforms the filter EM algorithm for estimation of A at high signal-to-noise-ratios.

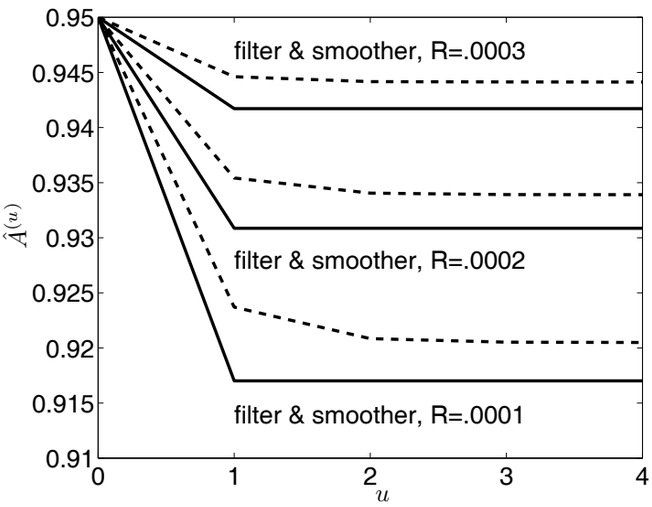


Fig. 6. State matrix estimates calculated by the smoother EM algorithm and filter EM algorithm for Example 13. It can be seen that the $\hat{A}^{(u)}$ better approach the nominal A at higher SNR.

"From the time I was seven, when I purchased my first calculator, I was fascinated by the idea of a machine that could compute things." *Michael Dell*

8.4.3 Measurement Noise Variance Estimation

The discussion of an EM procedure for measurement noise variance estimation is presented in a summary form because it follows analogously to the algorithms described previously.

Procedure 6. Assume that an initial estimate $\hat{R}^{(1)}$ of R is available. Subsequent estimates $\hat{R}^{(u)}$, $u > 1$, are calculated by repeating the following two-step procedure.

Step 1. Operate the minimum-variance smoother (7.66), (7.68), (7.69) designed with $\hat{R}^{(u)}$ to obtain corrected output estimates $\hat{y}_{k/N}^{(u)}$.

Step 2. For $i = 1, \ldots, p$, use $\hat{y}_{k/N}^{(u)}$ instead of y_k within (27) to obtain $\hat{R}^{(u+1)} = \mathrm{diag}((\hat{\sigma}_{1,v}^{(u+1)})^2, (\hat{\sigma}_{2,v}^{(u+1)})^2, \ldots, (\hat{\sigma}_{n,v}^{(u+1)})^2)$.

It can be shown using the approach of Lemma 9 that the sequence of measurement noise variance estimates are either monotonically non-increasing or non-decreasing depending on the initial conditions. When the SNR is sufficiently low, the measurement noise variance estimates converge to the actual value.

Lemma 16: *In respect of Procedure 6,*

$$\lim_{R^{-1}\to 0, Q\to 0, u\to\infty} R^{(u)} = R\,. \tag{79}$$

Proof: *By inspection of the output,* $\mathcal{H}_{OE} = \mathcal{G}Q\mathcal{G}^H(\mathcal{G}Q\mathcal{G}^H + R)^{-1}$, *it follows that* $\lim_{R^{-1}\to 0, Q\to 0, u\to\infty} \mathcal{H}_{IE} = 0$, *which together with the observation* $\lim_{R^{-1}\to 0, Q\to 0, u\to\infty} E\{zz^T\} = R$ *implies (79), since the MLE (27) is unbiased for large N.* □

Once again, the variance estimates produced by the above procedure are expected to be more accurate than those relying on filtered estimates.

Lemma 17:

$$-\left(\frac{\partial^2 \log f(\sigma_{i,v}^2 \mid \hat{y}_{k/N})}{(\partial\sigma_{i,v}^2)^2}\right)^{-1} < -\left(\frac{\partial^2 \log f(\sigma_{i,v}^2 \mid \hat{y}_{k/k})}{(\partial\sigma_{i,v}^2)^2}\right)^{-1}. \tag{80}$$

Proof: *The second partial derivative of the corresponding log-likelihood function with respect to the process noise variance is*

$$\frac{\partial^2 \log f(\sigma_{i,v}^2 \mid \hat{y}_{i,k/N})}{(\partial\sigma_{i,v}^2)^2} = -\frac{N}{2}(\sigma_{i,v}^2 + E\{\tilde{y}_{i,k/K}\tilde{y}_{i,k/K}^T\})^{-2},$$

"It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could be relegated to anyone else if machines were used." *Gottfried Wilhelm von Leibnitz*

where $\tilde{y}_{k/N}^{(u)} = y - \hat{y}_{k/N}^{(u)}$ *. Similarly, the use of filtered state estimates leads to*

$$\frac{\partial^2 \log f(\sigma_{i,v}^2 \mid \hat{y}_{i,k/k})}{(\partial \sigma_{i,v}^2)^2} = -\frac{N}{2}(\sigma_{i,v}^2 + E\{\tilde{y}_{i,k/k}\tilde{y}_{i,k/k}^T\})^{-2},$$

where $\tilde{y}_{k/k}^{(u)} = y - \hat{y}_{k/k}^{(u)}$ *. The claim (80) follows since* $E\{\tilde{y}_{k/N}\tilde{y}_{k/N}^T\} < E\{\tilde{y}_{k/k}\tilde{y}_{k/k}^T\}$ *.* □

8.5 Conclusion

From the Central Limit Theorem, the mean of a large sample of independent identically distributed random variables asymptotically approaches a normal distribution. Consequently, parameter estimates are often obtained by maximising Gaussian log-likelihood functions.

Unknown process noise variances and state matrix elements can be estimated by considering *i* single-input state evolutions of the form $x_{i,k+1} = \sum_{j=1}^{n} a_{i,j}x_{i,k} + w_{i,k}$, $a_{i,j}, x_{i,k}, w_{i,k} \in \mathbb{R}$. Similarly, unknown measurement noise variances can be estimated by considering *i* single-output observations of the form $z_{i,k} = y_{i,k} + v_{i,k}$, where $y_{i,k} + v_{i,k} \in \mathbb{R}$. The resulting MLEs are listed in Table 1 and are unbiased provided that the assumed models are correct and the number of samples is large.

The above parameter estimates rely on the availability of complete $x_{i,k}$ and $y_{i,k}$ information. Usually, both states and parameters need to be estimated from measurements. The EM algorithm is a common technique for solving joint state and parameter estimation problems. It has been shown that the estimation sequences vary monotonically and depend on the initial conditions. However, the use of imperfect states from filters or smoothers within the MLE calculations leads to biased parameter estimates. An examination of the approximate Cramér-Rao lower bounds shows that the use of smoothed states as opposed to filtered states is expected to provide improved parameter estimation accuracy.

When the SNR is sufficiently high, the states are recovered exactly and the bias terms diminish to zero, in which case $\lim_{Q^{-1}\to 0, R\to 0} \hat{\sigma}_{i,w}^2 = \sigma_{i,w}^2$ and $\lim_{Q^{-1}\to 0, R\to 0} \hat{a}_{i,j} = a_{i,j}$. Therefore, the process noise variance and state matrix estimation procedures described herein are only advocated when the measurement noise is negligible. Conversely, when the SNR is sufficiently low, that is, when the estimation problem is dominated by measurement noise, then $\lim_{Q\to 0, R^{-1}\to 0} \hat{\sigma}_{i,v}^2 = \sigma_{i,v}^2$. Thus, measurement noise estimation should only be attempted when the signal is absent. If parameter estimates are desired at intermediate SNRs then the subspace identification techniques such as [13], [14] are worthy of consideration.

"If automobiles had followed the same development cycle as the computer, a Rolls-Royce would today cost $100, get a million miles per gallon, and explode once a year, killing everyone inside." *Mark Stephens*

	ASSUMPTIONS	MAIN RESULTS
Process noise variance	$w_{i,k} \sim \mathcal{N}(0, \sigma_{i,w}^2)$	$\hat{\sigma}_{i,w}^2 = \frac{1}{N}\sum_{k=1}^{N}(x_{i,k+1} - \sum_{j=1}^{n} a_{i,j}x_{i,k})^2$
State matrix elements	$x_{i,k+1} \sim \mathcal{N}(\sum_{j=1}^{n} a_{i,j}x_{i,k},\ \sigma_{i,w}^2)$	$\hat{a}_{i,j} = \frac{\sum_{k=1}^{N}\left(x_{i,k+1} - \sum_{j=1, j\neq i}^{n} a_{i,j}x_{i,k}\right)x_{j,k}}{\sum_{k=1}^{N} x_{j,k}^2}$
Measurement noise variance	$v_{i,k} \sim \mathcal{N}(0, \sigma_{i,v}^2)$	$\hat{\sigma}_{i,v}^2 = \frac{1}{N}\sum_{k=1}^{N}(z_{i,k} - y_{i,k})^2$

Table 1. MLEs for process noise variance, state matrix element and measurement noise variance.

8.6 Problems

Problem 1.

(i) Consider the second order difference equation $x_{k+2} + a_1 x_{k+1} + a_0 x_k = w_k$. Assuming that $w_k \sim \mathcal{N}(0,\ \sigma_w^2)$, obtain an equation for the MLEs of the unknown a_1 and a_0.

(ii) Consider the n^{th} order autoregressive system $x_{k+n} + a_{n-1}x_{k+n-1} + a_{n-2}x_{k+n-2} + \ldots + a_0 x_k = w_k$, where $a_{n-1}, a_{n-2}, \ldots, a_0$ are unknown. From the assumption $w_k \sim \mathcal{N}(0,\ \sigma_w^2)$, obtain an equation for MLEs of the unknown coefficients.

Problem 2. Suppose that N samples of $x_{k+1} = Ax_k + w_k$ are available, where $w_k \sim \mathcal{N}(0,\ \sigma_w^2)$, in which σ_w^2 is an unknown parameter.

(i) Write down a Gaussian log-likelihood function for the unknown parameter, given x_k.

(ii) Derive a formula for the MLE $\hat{\sigma}_w^2$ of σ_w^2.

(iii) Show that $E\{\hat{\sigma}_w^2\} = \sigma_w^2$ provided that N is large.

(iv) Find the Cramér Rao lower bound for $\hat{\sigma}_w^2$.

(v) Replace the actual states x_k with filtered state $\hat{x}_{k/k}$ within the MLE formula. Obtain a high SNR asymptote for this approximate MLE.

"The question of whether computers can think is like the question of whether submarines can swim."
Edsger Wybe Dijkstra

Problem 3. Consider the state evolution $x_{k+1} = Ax_k + w_k$, where $A \in \mathbb{R}^{n \times n}$ is unknown and $w_k \in \mathbb{R}^n$.

(i) Write down a Gaussian log-likelihood function for the unknown components $a_{i,j}$ of A, given x_k and x_{k+1}.
(ii) Derive a formula for the MLE $\hat{a}_{i,j}$ of $a_{i,j}$.
(iii) Show that $E\{\hat{a}_{i,j}\} = a_{i,j}$. Replace the actual states x_k with the filtered state $\hat{x}_{k/k}$ within the obtained formula to yield an approximate MLE for $a_{i,j}$.
(iv) Obtain a high SNR asymptote for the approximate MLE.

Problem 4. Consider measurements of a sinusoidal signal modelled by $y_k = A\cos(2\pi fk + \varphi) + v_k$, with amplitude $A > 0$, frequency $0 < f < 0.5$, phase φ and Gaussian white measurement noise v_k.

(i) Assuming that φ and f are known, determine the Fisher information and the Cramér Rao lower bound for an unknown A.
(ii) Assuming that A and φ are known, determine the fisher information and the Cramér Rao lower bound for an unknown f_0.
(iii) Assuming that A and f are known, determine the Fisher information and the Cramér Rao lower bound .
(iv) Assuming that the vector parameter $[A^{\mathrm{T}}, f^T, \varphi^{\mathrm{T}}]^{\mathrm{T}}$ is known, determine the Fisher information matrix and the Cramér Rao lower bound. (Hint: use small angle approximations for sine and cosine, see [2].)

8.7 Glossary

SNR	Signal to noise ratio.
MLE	Maximum likelihood estimate.
CRLB	Cramér Rao Lower Bound
$F(\theta)$	The Fisher information of a parameter θ.
$x_k \sim \mathcal{N}(0, \sigma^2)$	The random variable x_k is normally distributed with mean μ and variance σ^2 .
$w_{i,k}$, $v_{i,k}$, $z_{i,k}$	i^{th} elements of vectors w_k , v_k , z_k.
$\hat{\sigma}_{i,w}^{(u)}$, $\hat{\sigma}_{i,v}^{(u)}$	Estimates of variances of $w_{i,k}$ and $v_{i,k}$ at iteration u.
$\hat{A}^{(u)}$, $\hat{R}^{(u)}$, $\hat{Q}^{(u)}$	Estimates of state matrix A, covariances R and Q at iteration u.
$\lambda_i(\hat{A}^{(u)})$	The i eigenvalues of $\hat{A}^{(u)}$.
A_i , C_i	i^{th} row of state-space matrices A and C.

"What lies at the heart of every living thing is not a fire, not warm breath, not a 'spark of life'. It is information, words, instructions." *Clinton Richard Dawkins*

$K_{i,k}$, $L_{i,k}$	i^{th} row of predictor and filter gain matrices K_k and L_k.
$S_k^{(u)}$	Additive term within the design Riccati difference equation to account for the presence of modelling error at time k and iteration u.
$a_{i,j}$	Element in row i and column j of A.
$C_k^{\#}$	Moore-Penrose pseudo-inverse of C_k.
$\boldsymbol{\mathcal{R}}^{(u)}$	A system (or map) that operates on the filtering/smoothing problem inputs to produce the input, state or output estimation error at iteration u. It is convenient to make use of the factorisation $\boldsymbol{\mathcal{R}}_{ei}^{(u)}(\boldsymbol{\mathcal{R}}_{ei}^{(u)})^H = \boldsymbol{\mathcal{R}}_{ei1}^{(u)}(\boldsymbol{\mathcal{R}}_{ei1}^{(u)})^H + \boldsymbol{\mathcal{R}}_{ei2}^{(u)}(\boldsymbol{\mathcal{R}}_{ei2}^{(u)})^H$, where $\boldsymbol{\mathcal{R}}_{ei2}^{(u)}(\boldsymbol{\mathcal{R}}_{ei2}^{(u)})^H$ includes the filter or smoother solution and $\boldsymbol{\mathcal{R}}_{ei1}^{(u)}(\boldsymbol{\mathcal{R}}_{ei1}^{(u)})^H$ is a lower performance bound.

8.8 References

[1] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison-Wesley Publishing Company Inc., Massachusetts, USA, 1990.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, Englewood Cliffs, New Jersey, ch. 7, pp. 157 – 204, 1993.

[3] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, Inc., New York, 1997.

[4] H. L. Van Trees and K. L. Bell (Editors), *Baysesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*, John Wiley & Sons, Inc., New Jersey, 2007.

[5] A. Van Den Bos, *Parameter Estimation for Scientists and Engineers*, John Wiley & Sons, New Jersey, 2007.

[6] R. K. Mehra, "On the identification of variances and adaptive Kalman filtering", *IEEE Transactions on Automatic Control*, vol. 15, pp. 175 – 184, Apr. 1970.

[7] D. C. Rife and R. R. Boorstyn, "Single-Tone Parameter Estimation from Discrete-time Observations", *IEEE Transactions on Information Theory*, vol. 20, no. 5, pp. 591 – 598, Sep. 1974.

[8] R. P. Nayak and E. C. Foundriat, "Sequential Parameter Estimation Using Pseudoinverse", *IEEE Transactions on Automatic Control*, vol. 19, no. 1, pp. 81 – 83, Feb. 1974.

[9] P. R. Bélanger, "Estimation of Noise Covariance Matrices for a Linear Time-Varying Stochastic Process", *Automatica*, vol. 10, pp. 267 – 275, 1974.

[10] V. Strejc, "Least Squares Parameter Estimation", *Automatica*, vol. 16, pp. 535 – 550, Sep. 1980.

[11] A. P. Dempster, N. M. Laid and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol 39, no. 1, pp. 1 – 38, 1977.

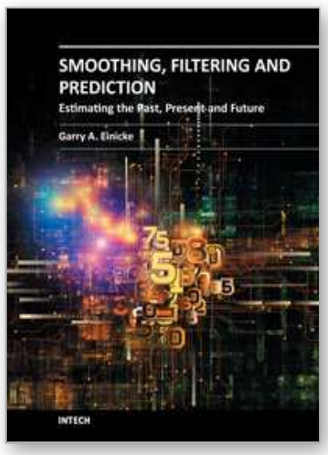
[12] P. Van Overschee and B. De Moor, "A Unifying Theorem for Three Subspace System Identification Algorithms", *Automatica*, 1995.

"In my lifetime, we've gone from Eisenhower to George W. Bush. We've gone from John F. Kennedy to Al Gore. If this is evolution, I believe that in twelve years, we'll be voting for plants." *Lewis Niles Black*

[13] T. Katayama and G. Picci, "Realization of stochastic systems with exogenous inputs and subspace identification methods", *Automatica*, vol. 35, pp. 1635 – 1652, 1999.
[14] T. Katayama, *Subspace Methods for System Identification*, Springer-Verlag London Limited, 2005.
[15] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253 – 264, 1982.
[16] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Annals of Statistics*, vol. 11,no. 1, pp. 95 – 103, Mar. 1983.
[17] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on Signal Processing*, vol. 36, no. 4, pp. 477 – 489, Apr. 198
[18] G. A. Einicke, "Optimal and Robust Noncausal Filter Formulations", *IEEE Transactions on Signal Processing,* vol. 54, no. 3, pp. 1069 - 1077, Mar. 2006.
[19] G. A. Einicke, J. T. Malos, D. C. Reid and D. W. Hainsworth, "Riccati Equation and EM Algorithm Convergence for Inertial Navigation Alignment", *IEEE Transactions on Signal Processing*, vol 57, no. 1, Jan. 2009.
[20] G. A. Einicke, G. Falco and J. T. Malos, "EM Algorithm State Matrix Estimation for Navigation", *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 437 – 440, May 2010.
[21] T. K. Moon, "The Expectation-Maximization Algorithm", *IEEE Signal Processing Magazine*, vol. 13, pp. 47 – 60, Nov. 1996.
[22] D. G. Tzikas, A. C. Likas and N. P. Galatsanos, "The Variational Approximation for Bayesian Inference: Life After the EM Algorithm", *IEEE Signal Processing Magazine*, vol. 25, Iss. 6, pp. 131 – 146, Nov. 200
[23] D. M. Titterington, A. F. M. Smith and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester and New York, 1985.
[24] R. P. Savage, *Strapdown Analytics*, Strapdown Associates, vol. 2, ch. 15, pp. 15.1 – 15.142, 2000.
[25] P. K. Seidelmann, ed., *Explanatory supplement to the Astronomical Almanac*, Mill Valley, Cal., University Science Books, pp. 52 and 698, 1992.
[26] G. A. Einicke, G. Falco, M. T. Dunn and D. C. Reid, "Iterative Smoother-Based Variance Estimation", IEEE Signal Processing letters, 2012 (to appear).

"The faithful duplication and repair exhibited by the double-stranded DNA structure would seem to be incompatible with the process of evolution. Thus, evolution has been explained by the occurrence of errors during DNA replication and repair." *Tomoyuki Shibata*

IntechOpen



Smoothing, Filtering and Prediction - Estimating The Past, Present and Future

Edited by

ISBN 978-953-307-752-9
Hard cover, 276 pages
Publisher InTech
Published online 24, February, 2012
Published in print edition February, 2012

This book describes the classical smoothing, filtering and prediction techniques together with some more recently developed embellishments for improving performance within applications. It aims to present the subject in an accessible way, so that it can serve as a practical guide for undergraduates and newcomers to the field. The material is organised as a ten-lecture course. The foundations are laid in Chapters 1 and 2, which explain minimum-mean-square-error solution construction and asymptotic behaviour. Chapters 3 and 4 introduce continuous-time and discrete-time minimum-variance filtering. Generalisations for missing data, deterministic inputs, correlated noises, direct feedthrough terms, output estimation and equalisation are described. Chapter 5 simplifies the minimum-variance filtering results for steady-state problems. Observability, Riccati equation solution convergence, asymptotic stability and Wiener filter equivalence are discussed. Chapters 6 and 7 cover the subject of continuous-time and discrete-time smoothing. The main fixed-lag, fixed-point and fixed-interval smoother results are derived. It is shown that the minimum-variance fixed-interval smoother attains the best performance. Chapter 8 attends to parameter estimation. As the above-mentioned approaches all rely on knowledge of the underlying model parameters, maximum-likelihood techniques within expectation-maximisation algorithms for joint state and parameter estimation are described. Chapter 9 is concerned with robust techniques that accommodate uncertainties within problem specifications. An extra term within Riccati equations enables designers to trade-off average error and peak error performance. Chapter 10 rounds off the course by applying the afore-mentioned linear techniques to nonlinear estimation problems. It is demonstrated that step-wise linearisations can be used within predictors, filters and smoothers, albeit by forsaking optimal performance guarantees.

Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen

IntechOpen