Estimation of parameters in dynamical models

Morten Hagdrup & John Bagterp Jørgensen

DTU Compute
Technical University of Denmark
November 1, 2016

1 Nonlinear Data Fitting

1.1 Introduction

We shall in this section consider the problem of fitting a model to observed data. As the headline indicates we take on the general question of fitting a model structure nonlinear in its parameters to data. By a model structure we shall mean a parametrized family of functions on $\mathbb{R}$. The vector of parameters we shall agree to denote by θ and assume that θ belongs to some set $\Theta \subseteq \mathbb{R}^p$. The function corresponding to the parameter vector θ we shall denote $\phi(\cdot, \theta)$ and thus its value at time τ by $\phi(\tau, \theta)$.

Given a set of observations $y_1, y_2, \ldots, y_n$ collected at times $t_1, t_2, \ldots, t_n$ the problem is therefore that of finding the value of $\theta \in \Theta$ that ensures the best fit of function values $\phi(t_i, \theta)$ to the y_i with respect to some chosen criterion. Different criteria will in general give rise to different optimal parameter vectors for a given set of observations. In this section we shall consider the least squares criterion and this gives rise to a class of optimization problems known as *Nonlinear Least Squares* problems (NLS). The use of the least squares criterion means that we seek to minimize w.r.t. $\theta \in \Theta$ the cost function

$$\sum_{i=1}^{n} |y_i - \phi(t_i, \theta)|^2 \tag{1}$$

Since all measurements are at least to some extent subject to measurement noise the question presents itself how much this noise influences the parameter estimate. In the following our concern will therefore be the statistical properties of estimators of NLS type. The analysis of these statistical properties will be based on the scenario where the data generating system belongs to the model structure $\phi(\cdot, \theta), \theta \in \Theta$, with the measurements corrupted by additive noise. In other words we shall assume that

$$y_i = \phi(t_i, \theta_*) + \varepsilon_i, \qquad i \in \{1, 2, \ldots, n\} \tag{2}$$

for some value θ_* and some sequence $(\varepsilon_i)_{i=1}^n$ of random variables. Often one assumes that $\varepsilon_i \sim N_{iid}(0, \sigma^2)$, i.e. that the noise samples are described by a sequence of independent and identically distributed normal random variables. This is the assumption we shall make here in our study of non-linear least squares estimation. Since the focus is here on estimation of parameters we shall (by abuse of notation) occasionally also write

$$\phi_{t_i}(\theta) = \phi(t_i, \theta), \qquad i \in \{1, 2, \ldots, n\} \tag{3}$$

1.2 Notation

In the calculations to follow we shall find it convenient to employ matrix notation. Specifically we gather the n observations and the function values at the corresponding times in column vectors as follows:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ \ldots \\ y_n \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \ldots \\ \varepsilon_n \end{bmatrix} \qquad \text{and} \qquad \phi_n(\theta) = \begin{bmatrix} \phi(t_1, \theta) \\ \phi(t_2, \theta) \\ \ldots \\ \ldots \\ \phi(t_n, \theta) \end{bmatrix} \tag{4}$$

For the Jacobian of ϕ_n we employ the notation

$$J_n(\theta) = \begin{bmatrix} \left(\frac{\partial \phi}{\partial \theta_1}\right)_{t=t_1} & \cdots & \left(\frac{\partial \phi}{\partial \theta_p}\right)_{t=t_1} \\ \left(\frac{\partial \phi}{\partial \theta_1}\right)_{t=t_2} & \cdots & \left(\frac{\partial \phi}{\partial \theta_p}\right)_{t=t_2} \\ \vdots & & \vdots \\ \left(\frac{\partial \phi}{\partial \theta_1}\right)_{t=t_n} & \cdots & \left(\frac{\partial \phi}{\partial \theta_p}\right)_{t=t_n} \end{bmatrix} \in \mathbb{R}^{n \times p} \tag{5}$$

In order to keep the notation relatively light we have chosen not to include an index specifying the size of y and ε. Since these quantities will always appear in connections where it should be obvious from the context what their size is it is deemed unnecessary to have a special index denoting this. This is strictly speaking abuse of notation but we hope the reader will agree that this is a price worth paying.

1.3 Consistency of Least Squares Estimators

1.3.1 Review of Linear Least Squares Estimation

Let us first recall the familar case of linear least squares data fitting where we assume that $\phi(\cdot, \theta)$ is linear in the parameters θ. In this case the Jacobian J_n of ϕ_n is independent of θ so we shall refer to it merely by J_n. That is we assume that

$$\phi_n(\theta) = J_n \, \theta \tag{6}$$

The linear least squares data fitting problem then becomes that of determining the value $\hat{\theta}_n$ of θ which minimizes the objective function

$$\theta \to \|y - J_n \, \theta\|_2^2 \tag{7}$$

It is well-known that $\hat{\theta}_n$ is a solution to this linear least squares problem if and only if

$$({J_n}' J_n)\hat{\theta}_n = {J_n}' y \tag{8}$$

and if ${J_n}' J_n$ is invertible then

$$\hat{\theta}_n = ({J_n}' J_n)^{-1} {J_n}' y \tag{9}$$

becomes the unique solution of the linear least squares problem.

Assuming that the data y is generated by a linear system of the considered structure (6) with added noise ε

$$y = J_n \theta_* + \varepsilon \tag{10}$$

then the solution (9) specializes to

$$\hat{\theta}_n = \theta_* + ({J_n}' J_n)^{-1} {J_n}' \, \varepsilon \tag{11}$$

From the assumption that $\varepsilon_i \sim N_{iid}(0, \sigma^2)$ follows immediately by considering the ensemble average over the noise process $\boldsymbol{\varepsilon}$ that

$$E\{\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}\} = \theta_* \tag{12}$$

While it is surely comforting to know that the linear least squares estimate $\hat{\theta}_n$ is not systematically offset what is of greater importance to study is in fact how the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ behaves asymptotically. That is when the number of samples tends to infinity. After all, if the estimates resulting from a particular choice of estimator were systematically offset by a certain amount one would at least in principle be able to correct for this. Keeping the variance of the estimator low is another matter though. An attractive property for $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ to have would therefore be if its covariance matrix could be made arbitrarily small just by choosing the number of observations large enough. This property of a family of estimators $\{\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} : n \in \mathbb{N}\}$ is known as *consistency* or *asymptotic consistency*.

From the expression (11) for the linear least squares estimate as a function of the noise realization ε it immediately follows using (12) that the variance of the Linear Least Squares Estimator equals

$$E\{(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*)(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*)'\} = (J_n{}' J_n)^{-1} J_n{}' \, E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\} \, J_n \, (J_n{}' J_n)^{-1} \tag{13}$$

This somewhat cumbersome expression which yields the covariance of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ in terms of the covariance $E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\}$ of the noise sequence may be significantly simplified in the case that $\boldsymbol{\varepsilon_i} \sim N_{iid}(0, \sigma^2)$. For this particular case it is clear that

$$E\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right\} = \sigma^2 I_{n\times n} \tag{14}$$

where by $I_{n\times n}$ we denote the $(n \times n)$ identity matrix. The resulting covariance expression now becomes the much simpler

$$E\left\{\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*\right)\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*\right)'\right\} = \sigma^2 \left(J_n{}' J_n\right)^{-1} \tag{15}$$

For fixed noise variance σ^2 the question of asymptotic consistency of the family of estimators $\{\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} : n \in \mathbb{N}\}$ therefore becomes that of whether the smallest eigenvalue of the matrix $(J_n{}' J_n)^{-1}$ tends to zero as more and more samples are considered. It is maybe tempting to think that this will always be the case as one considers ever more data points, but as the following examples show this is not necessarily so.

1.3.2 Two examples

Consider the case of the simple 1-dimensional linear model structure

$$\phi(t;\theta) = \frac{\theta}{t} \tag{16}$$

and the generating system is that given by

$$y_i = \phi(t_i, \theta_*) + \varepsilon_i \tag{17}$$

with ε_i denoting Additive White Gaussian Noise as above. For this set-up we obtain for the observations taken at times $t_i = i$, $i \in \{1, \ldots, n\}$, that $J_n{}' J_n$ becomes the scalar

$$J_n{}' J_n = \sum_{i=1}^{n} \frac{1}{i^2} \to \frac{\pi^2}{6} < +\infty \tag{18}$$

In this case we do *not* achieve that $\sigma^2({J_n}' J_n)^{-1}$ tends to zero as the number of observations increases beyond all bounds. This is obviously due to the fact that we have chosen samples from the system output for which the signal-to-noise ratio becomes so poor with increasing i that it prevents ${J_n}' J_n$ from tending to $+\infty$. In fact one may in general consider the expression

$$\frac{{J_n}' J_n}{\sigma^2} \tag{19}$$

to embody a matrix-valued representation of a signal-to-noise ratio for the identification experiment. It is in fact a special case of the matrix known in Mathematical Statistics as *Fisher's Information Matrix*. It is important to note though that the individual observations do not themselves have to be collected with ever increasing signal-to-noise ratio in order for the estimator covariance to tend to zero. As long as there is sufficient signal power the independence of the noise contributions will allow for the averaging to take effect. A borderline case would result if we took $t_i = \sqrt{i}$ instead of $t_i = i$. In that case one would get:

$${J_n}' J_n = \sum_{i=1}^{n} \frac{1}{i} \to +\infty \quad \text{for } n \to +\infty \tag{20}$$

Here the ensemble average

$$\frac{\sigma^2}{\phi(t_i;\theta_*)^2} = \frac{i \cdot \sigma^2}{\theta_*^2} \tag{21}$$

of the noise-to-signal ratio at the system output at t_i also tends to infinity with increasing index i but now sufficiently slowly for the averaging of the independent noise contributions to become effective. So in this case the variance of the parameter estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ tends to zero.

In addition to discussing the notion of consistency of estimators we have in this subsection provided a glimpse of the branch of mathematics called *Optimal Design of Experiments*. The goal of that field is to provide guidelines as to how to select most appropriately the conditions for an experiment to be used for parameter estimation. The chief objective being of course to obtain parameter estimates of the highest possible quality given the circumstances.

1.3.3 Nonlinear Least Squares Estimation

As is clear from the previous subsections there are already within the framework of Linear Least Squares Estimation certain conditions to check in order to ensure that the estimator under consideration is consistent. One might therefore think that additional difficulties would arise when considering the more general case of Nonlinear Least Squares Estimation. Apart from certain reasonable regularity assumptions, though, it turns out that the results known from the linear case carry over to the nonlinear one.

We present here the results of an analysis carried out by Jennrich in [2]. In the subsequent sections we attempt to provide the reader with a derivation of the results stated. We do skip some of the more technical details though. So for a rigorous derivation the reader is referred to [2]. In order to state the results we need to introduce a bit of notation.

Let $x = (x_i)$ and $y = (y_i)$ be two sequences of real numbers and let

$$(x, y)_n = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \tag{22}$$

If $(x,y)_n$ converges to a real number as $n \to +\infty$ then its limit (x,y) will be called the *tail product* of x and y. Assume now that instead of being real numbers x_i and y_i are continuous functions on some Euclidean space and the convergence $(x,y)_n$ is uniform on some subset of this space. In that case the limit function (x,y) defined pointwise by (22) is a continuous function which we shall refer to as the *tail cross product*.

Assumption 1. *A sequence $y = (y_i)$ of real-valued responses has the structure*

$$y_i = \phi_{t_i}(\theta_*) + \varepsilon_i \tag{23}$$

where the ϕ_{t_i} are continuous functions on a compact subset Θ of a Euclidean space and the ε_i are independent Gaussian random variables with zero mean and finite variance $\sigma^2 > 0$.

Any vector $\hat{\theta} \in \Theta$ which minimizes

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi_{t_i}(\theta))^2 \tag{24}$$

will be called a least squares estimate of θ_* based on the first n values of y_i.

Assumption 2. *The tail cross product of $\phi = (\phi_{t_i})$ with itself exists and*

$$Q(\theta) = |\phi(\theta) - \phi(\theta_*)|^2 \text{ has a unique minimum at } \theta = \theta_* \tag{25}$$

One should note that Assumption 2 is in fact a so-called *identifiability condition* for the model structure $\theta \to \phi(\cdot,\theta)$. What it states is that in the absence of noise it is possible to identify the true value θ_* of the parameter perfectly from the recorded observations (t_i, y_i).

Under assumptions 1 & 2 the function Q is continuous on Θ and the following theorem holds:

Theorem 1. *Let $(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}})$ be a sequence of least squares estimators. Under assumptions 1 and 2 $(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}})$ and $(Q_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}))$ are strongly consistent estimators of θ_* and σ^2 respectively.*

What this means is that for almost every realization ε of the noise process $\boldsymbol{\varepsilon}$ the sequences of estimates $\hat{\theta}_n$ and $Q_n(\hat{\theta}_n)$ converge to θ_* and σ^2 respectively.

To establish the asymptotic normality of a sequence of least squares estimators one needs the existence of the following derivatives:

$$\phi'_{t,i}(\theta) = \frac{\partial}{\partial\theta_i}\phi_t(\theta) \quad \text{and} \quad \phi''_{t,ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\phi_t(\theta) \tag{26}$$

for $i,j = 1,\ldots,p$.

Assumption 3. *For all $i,j = 1,\ldots,p$ the derivatives $\phi_{t,i}$ and $\phi_{t,ij}$ exist and are continuous on Θ. Furthermore all tail cross products of the form $[g,h]$ exist, where g and h are of the form ϕ, ϕ'_i or ϕ''_{ij}.*

For each $\theta \in \Theta$ we define for $i, j = 1, \ldots, p$ and $n \in \mathbb{N}$

$$\Phi_{n,ij}(\theta) = (\phi_i'(\theta), \phi_j'(\theta))_n, \quad \Phi_n(\theta) = (\Phi_{n,ij}(\theta)) \tag{27}$$

and for the cross tail products correspondingly

$$\Phi_{ij}(\theta) = (\phi_i'(\theta), \phi_j'(\theta)), \quad \Phi(\theta) = (\Phi_{ij}(\theta)) \tag{28}$$

We further make the following

Assumption 4. *The true parameter θ_* is an interior point of the compact set $\Theta \subset \mathbb{R}^p$ and the matrix $\Phi(\theta_*)$ is non-singular.*

With those preparations in place we may formulate the following result

Theorem 2. *Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ be a sequence of least squares estimators of θ_*. Under assumptions 1 - 4 it holds that*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*) \xrightarrow{d} N(0, \sigma^2 \Phi^{-1}(\theta_*)) \text{ as } n \to +\infty \tag{29}$$

where the notion of convergence at play is that of convergence in distribution. Furthermore $(\boldsymbol{\Phi}_{\boldsymbol{n}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}))$ is a strongly consistent sequence of estimators of $\Phi(\theta_*)$

It should be noted here that in addition to the fact that $(\boldsymbol{\Phi}_{\boldsymbol{n}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}))$ is a strongly consistent sequence of estimators of $\Phi(\theta_*)$ it is a trivial consequence of Assumption 3 that

$$\frac{1}{n} J_n(\theta_*)' J_n(\theta_*) \to \Phi(\theta_*) \text{ as } n \to +\infty \tag{30}$$

in whichever norm one considers on the vector space of the space of $(p \times p)$-matrices with real entries. It therefore follows from Assumption 4 that the matrix $J_n(\theta_*)' J_n(\theta_*)$ is invertible for all sufficiently large n. From the asymptotic result of (29) we therefore have for sufficiently large values of n the following approximate distribution of $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*$:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_* \sim N(0, \sigma^2 (J_n(\theta_*)' J_n(\theta_*))^{-1}) \tag{31}$$

Here the value θ_* would in practice be replaced by the available estimate $\hat{\theta}_n$.

1.3.4 NLS versus ML

Gauss-Newton [3]

1.4 Derivation of the results on NLS estimators

In this subsection we present a non-rigorous but hopefully illustrative derivation of the statistical result (31) stated in subsection 1.3.3. The analysis we present relies on replacing in our calculations the model structure $\phi(t, \theta)$ by its 1st order Taylor expansion at θ_* with respect to θ. The success of this approach hinges upon the combined effects of the higher order derivatives of the model structure and measurement noise not being too strong. Fortunately from the theorems in subsection 1.3.3 it follows that with increasing n one can be increasingly certain of the estimates $\hat{\theta}_n$ being close to θ_*. This is the reason why things go well in the derivation below.

1.4.1 Mean and variance of $\hat{\boldsymbol{\theta}}$

The problem is that of minimizing $\|y - \phi_n(\theta)\|_2^2$ w.r.t. θ which upon insertion of (2) is seen to be equivalent to that of minimizing the expression

$$\|\varepsilon - (\phi_n(\theta) - \phi_n(\theta_*))\|_2^2 \tag{32}$$

As mentioned above we now assume that the combined effect of the higher order derivatives of the estimator and measurement noise are sufficiently weak for the approximation

$$\phi_n(\theta) \approx \phi_n(\theta_*) + J_n(\theta_*)(\theta - \theta_*) \tag{33}$$

to be good enough for our purposes.

Accepting the approximation (33) we are faced with the task of minimizing the expression

$$\|\varepsilon - J_n(\theta_*)(\theta - \theta_*)\|_2^2 \tag{34}$$

with respect to θ.

From the theory of linear least squares problems we know that $\hat{\theta}_n$ is a minimizer of (34) if and only if the following *normal equations* are satisfied

$$J_n(\theta_*)' J_n(\theta_*)(\hat{\theta}_n - \theta_*) = J_n(\theta_*)' \varepsilon \tag{35}$$

In a data fitting scenario as considered here the number of observations n is always assumed to be larger than p, the dimension of the parameter vector θ. For a suitably chosen model structure $\phi(\cdot, \theta)$ we will therefore have that the matrix product $J_n(\theta_*)' J_n(\theta_*)$ is invertible. To the extent that the approximations we have made are valid we have the following closed-form expression for the deviation $\left(\hat{\theta}_n - \theta_*\right)$ of our estimate from the true parameter value as a function of the concrete realization ε of the noise process:

$$\hat{\theta}_n - \theta_* = \left(J_n(\theta_*)' J_n(\theta_*)\right)^{-1} J_n(\theta_*)' \varepsilon \tag{36}$$

Since the ensemble mean of the noise process $\boldsymbol{\varepsilon}$ equals 0 we obtain as a result of the approximations made that $E\{\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}\} = \theta_*$. For the variance of the estimator it now follows that

$$E\left\{\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*\right)\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \theta_*\right)'\right\} = \left(J_n(\theta_*)' J_n(\theta_*)\right)^{-1} J_n(\theta_*)' E\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right\} J_n(\theta_*)\left(J_n(\theta_*)' J_n(\theta_*)\right)^{-1} \tag{37}$$

Summing up we have established using the approximations spelled out above that the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ follows a Gaussian distribution with the following parameters

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} \sim N\left(\theta_*, \sigma^2\left(J_n(\theta_*)' J_n(\theta_*)\right)^{-1}\right) \tag{38}$$

1.4.2 Estimating σ^2

Equation (2) gives an expression for the distribution that the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ follows. However, we do not know the value of σ^2 in advance. It therefore remains to provide an estimate $\hat{\sigma}^2$ of the variance σ^2 of the residual.

For a given realization ε of the measurement noise process we have within the range of validity of our approximations established that the deviation $\hat{\theta}_n - \theta_*$ is given by (36). By insertion of (36) into (33) we now obtain the following approximation for the residual:

$$y - \phi_n(\hat{\theta}_n) \approx \left(I_{n\times n} - J_n\left(\theta_*\right)\left(J_n\left(\theta_*\right)' J_n\left(\theta_*\right)\right)^{-1} J_n\left(\theta_*\right)'\right)\varepsilon \tag{39}$$

It will now be helpful to take a geometric view of the linear map $P : \mathbb{R}^n \to \mathbb{R}^n$ where

$$P\varepsilon = J_n\left(\theta_*\right)\left(J_n\left(\theta_*\right)_n' J\left(\theta_*\right)\right)^{-1} J_n\left(\theta_*\right)'\,\varepsilon \tag{40}$$

A straightforward calculation will reveal that it is *idempotent*, that is $P^2 = P$. Secondly by denoting the range of P by U

$$U = \{u \in \mathbb{R}^n |\, \exists v \in \mathbb{R}^n : u = Pv\} \tag{41}$$

we obtain that the restriction of P to the subspace U is the identity. In other words

$$\forall u \in U : Pu = u \tag{42}$$

Denoting by $I_{\mathbb{R}^n}$ the identity map on $\mathbb{R}^n$ then the map $I_{\mathbb{R}^n} - P$ may likewise be shown to be idempotent. It furthermore equals the identity map on the subspace W given by

$$W = \{w \in \mathbb{R}^n |\, \exists v \in \mathbb{R}^n : w = (I_{\mathbb{R}^n} - P)v\} \tag{43}$$

We have now written $\mathbb{R}^n$ as a *direct sum* $\mathbb{R}^n = U \oplus W$ in the sense that $U \cap W = \{0\}$ and each $x \in \mathbb{R}^n$ may be decomposed as a sum $x = Px + (I_{\mathbb{R}^n} - P)x$ of its *projections* onto U and W respectively. The dimension of the subspace U equals the number of parameters p. We now pick an orthonormal basis $(u_1, \ldots, u_p)$ for U. This basis we augment to an orthonormal basis of $\mathbb{R}^n$ by also selecting an orthonormal basis $(u_{p+1}, \ldots, u_n)$ for W. From our geometric analysis now follows that each of the basis vectors $u_1, \ldots, u_p$ is an eigenvector of P corresponding to the eigenvalue $\lambda = 1$ and each of the basis vectors $u_{p+1}, \ldots, u_n$ is an eigenvector of P corresponding to the eigenvalue $\lambda = 0$. Define now the matrix Ξ whose columns are made up of the coordinates of the basis vectors $(u_1, \ldots, u_n)$ with respect to the usual basis of $\mathbb{R}^n$. That is let Ξ denote the orthogonal matrix

$$\Xi = \left[u_1 | u_2 | \ldots | u_n\right] \tag{44}$$

Then we have for the linear map P the following decomposition

$$\Xi^{-1} P\, \Xi = \begin{bmatrix} I_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & 0_{(n-p)\times(n-p)} \end{bmatrix} \equiv \Lambda \tag{45}$$

We now obtain for the residual $y - \phi_n(t, \hat{\theta}) \approx (I - P)\varepsilon$ that

$$\begin{aligned} (y - \phi_n(t,\hat{\theta}))'(y - \phi_n(t,\hat{\theta})) &\approx \varepsilon'(I_{n\times n} - \Xi^{-1}\Lambda\Xi)(I_{n\times n} - \Xi^{-1}\Lambda\Xi)\varepsilon \\ &= \varepsilon'(I_{n\times n} - \Xi^{-1}\Lambda\Xi)\varepsilon \\ &= (\Xi\varepsilon)'(I_{n\times n} - \Lambda)(\Xi\varepsilon) \end{aligned} \tag{46}$$

Let us now recall yet another result from multivariate statistics. It states that a random vector $\boldsymbol{x}$ follows an n-dimensional normal distribution $\boldsymbol{x} \sim N(\mu, \Sigma)$ if and only if $\Omega\boldsymbol{x} \sim N(\mu, \Sigma)$ for any orthogonal matrix $\Omega \in \mathbb{R}^{n\times n}$. If we now assume that $\boldsymbol{\varepsilon_i} \sim N_{iid}(0, \sigma^2)$ and hence in particular that (14) holds it follows that also $\boldsymbol{\xi} = \Xi\boldsymbol{\varepsilon}$ satisfies

$$E\left\{\boldsymbol{\xi}\boldsymbol{\xi'}\right\} = \sigma^2 I_{n\times n} \tag{47}$$

One may therefore approximate the residual power as follows

$$(y - \phi_n(\hat{\theta}))'(y - \phi_n(\hat{\theta})) = \xi' \begin{bmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{(n-p)\times(n-p)} \end{bmatrix} \xi \tag{48}$$

Normalizing by the noise variance σ^2 now yields

$$\frac{(\boldsymbol{y} - \phi_n(\hat{\theta}))'(\boldsymbol{y} - \phi_n(\hat{\theta}))}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=p+1}^{n} \boldsymbol{\xi_i^2} \tag{49}$$

This writes the normalized residual power as the sum of squares of $(n-p)$ independent standard normal random variables (that is zero-mean and with unit variance). The normalized residual power therefore follows a χ^2-distribution with $(n-p)$ degrees of freedom:

$$\frac{(\boldsymbol{y} - \phi_n(\hat{\theta}))'(\boldsymbol{y} - \phi_n(\hat{\theta}))}{\sigma^2} \sim \chi^2(n-p) \tag{50}$$

Now it is well-known (or easily derived) that the mean value of the χ^2-distribution with $n-p$ degrees of freedom equals $n-p$. This implies that the ensemble average of

$$\frac{(\boldsymbol{y} - \phi_n(\hat{\theta}))'(\boldsymbol{y} - \phi_n(\hat{\theta}))}{n-p} \tag{51}$$

equals the noise variance σ^2. We do not have access to the ensemble average in question. In fact we only have access to (51) for a single realization so with that we have to make do. We thus have the following estimate $\hat{\sigma}^2$ for the measurement noise variance σ^2:

$$\hat{\sigma}^2 = \frac{\|y - \phi_n(\theta)\|_2^2}{n-p} \tag{52}$$

or in terms of the individual vector components:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} |y_i - \phi(t_i, \hat{\theta})|^2 \tag{53}$$

Our analysis shows that hereby defined estimator $\boldsymbol{\hat{\sigma}^2}$ is a central estimator, i.e. that it is *unbiased*. It should be noted here that this does *not* mean that it is a constant estimator. What it means is that its ensemble average over all realizations of the measurement noise process $\boldsymbol{\varepsilon}$ equals the true value σ^2. So the estimator $\boldsymbol{\hat{\sigma}^2}$ still has a non-zero variance depending on σ^2 and on the size n of the data set.

In summary we also note that the analysis above has shown that the estimator $\boldsymbol{\hat{\sigma}^2}$ is given by

$$\boldsymbol{\hat{\sigma}^2} \sim \frac{\sigma^2}{n-p}\boldsymbol{\chi^2(n-p)} \tag{54}$$

where by $\boldsymbol{\chi^2(n-p)}$ we denote a random variable which follows a χ^2-distribution with $n-p$ degrees of freedom.

1.5 Parameter confidence intervals

In subsection 1.4.1 we saw that the distribution of the parameter estimator $\hat{\boldsymbol{\theta}}$ for *known* noise variance σ^2 is given by the expression (38). In order to extract the variances on the individual elements $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ of $\hat{\boldsymbol{\theta}}$ we recall a fundamental result of multivariate statistics. This states that a random variable $\boldsymbol{x}$ follows a multivariate Gaussian distribution of dimension p if and only if $a'\boldsymbol{x}$ follows a univariate Gaussian distribution for any vector $a \in \mathbb{R}^p$ and that

$$a'\boldsymbol{x} \sim N\left(a'\mu, a'\Sigma a\right) \tag{55}$$

if $\boldsymbol{x} \sim N(\mu, \Sigma)$. From (38) it therefore follows that

$$a'\hat{\boldsymbol{\theta}} \sim N\left(a'\theta_*, \sigma^2 a'\left(J\left(\theta_*\right)' J\left(\theta_*\right)\right)^{-1} a\right) \tag{56}$$

and in particular that

$$\frac{a'\hat{\boldsymbol{\theta}} - a'\theta_*}{\sigma\sqrt{a'\left(J\left(\theta_*\right)' J\left(\theta_*\right)\right)^{-1} a}} \sim N(0,1) \tag{57}$$

1.5.1 Enter Student's t-distribution

However, the value of σ^2 in (57) is *not* known in advance. It has to be estimated from data using the formula (53). Replacing σ^2 by $\boldsymbol{\sigma^2}$ in turn alters the distribution of the left hand side of (57). In order to arrive at the modified distribution based on an estimated noise variance we need to recall a result from statistics. It states that if $\boldsymbol{x}$ and $\boldsymbol{y}$ are random variables which follow a standard normal and a χ^2-distrbution with ν degrees of freedom respectively

$$\boldsymbol{x} \sim N(0,1) \quad \text{and} \quad \boldsymbol{y} \sim \chi^2(\nu) \tag{58}$$

then the quotient

$$\boldsymbol{z} = \frac{\boldsymbol{x}}{\sqrt{\frac{\boldsymbol{y}}{\nu}}} \sim t(\nu) \tag{59}$$

follows *Student's t-distribution* with ν degrees of freedom. In a course on Probability Theory one would derive the probability density function the t-distribution with ν degrees of freedom and arrive at the following result

$$p(x;\nu) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + x^2\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R} \tag{60}$$

One notes in particular that the probability density function is symmetric with respect to $x = 0$. The corresponding cumulative density function (cdf)

$$P(x;\nu) = \int_{-\infty}^{x} p(\xi;\nu)\, d\xi \tag{61}$$

is easily evaluated using the MATLAB® function `tinv` from the Statistical Toolbox as we shall now show. If one wants to determine the number $x \geq 0$ with the property that only the fraction $0 < \alpha \leq 1$ of the total probability mass of $t(\nu)$ resides outside the interval $[-x; x]$, that is such that

$$\int_{-x}^{x} p(\xi;\nu)\, d\xi = 1 - \alpha \tag{62}$$

then it is clear that we need to find the x which satisfies the equation

$$P(-x;\nu) = \int_{-\infty}^{-x} p(\xi;\nu)\, d\xi = \frac{\alpha}{2} \tag{63}$$

Using `tinv` we have that

$$x = -\texttt{tinv}(\alpha/2, \nu) \tag{64}$$

and we shall henceforth use the notation

$$x = -t_\nu(\alpha/2) \tag{65}$$

for this value.

Finally it should be noted that as the number of degrees of freedom tends to infinity Student's t-distribution tends to the standard normal distribution. This is in fact only reasonable since the precision of the variance estimate also becomes better and better with an increasing number of independent observations.

1.5.2 Confidence intervals expressed in terms of the t-distribution

Using the properties of Student's t-distribution introduced above we may now calculate approximate confidence intervals for the parameter estimates $\hat{\theta}_i$. From (57) we have an example of a random variable which follows a standard normal distribution while the estimator

$$\frac{n-p}{\sigma^2}\hat{\boldsymbol{\sigma}}^{\mathbf{2}} \tag{66}$$

follows a χ^2-distribution with $n-p$ degrees of freedom. From (59) we now see that

$$T = \frac{a'\hat{\boldsymbol{\theta}} - a'\theta_*}{\sqrt{\hat{\boldsymbol{\sigma}}^{\mathbf{2}} a' \left(J\left(\theta_*\right)' J\left(\theta_*\right)\right)^{-1} a}} \sim t(n-p) \tag{67}$$

follows a t-distribution with $(n-p)$ degrees of freedom. The approximate $100(1-\alpha)\%$ confidence interval for the estimate $a'\hat{\theta}$ therefore becomes

$$a'\hat{\theta} \pm t_{n-p}(\alpha/2)\left(\hat{\sigma}^2 a' F^{-1} a\right)^{1/2} \tag{68}$$

where

$$F = J(\hat{\theta})' J(\hat{\theta}) \tag{69}$$

We have thus approximated the unknown correct value θ_* by the estimate $\hat{\theta}$.

For the individual parameter estimates θ_i we now obtain the respective confidence interval by choosing as a the relevant canonical basis vector of $\mathbb{R}$. That is in the case of $\hat{\theta}_i$ we pick $a = e_i$ which only differs from the zero column vector at entry i which equals 1. The resulting approximative confidence intervals become

$$\hat{\theta}_i \pm t_{n-p}(\alpha)\hat{\sigma}\sqrt{C_{ii}} \quad \text{where} \quad C_{ii} = [(J(\hat{\theta})' J(\hat{\theta}))^{-1}]_{ii} \tag{70}$$

for $i = 1, \ldots, p$.

1.5.3 A word of caution

It may be beneficial to pause here for a moment and contemplate the expression (67) used to approximate the distribution of $a'\hat{\boldsymbol{\theta}}$. Instead of the correct value θ_* we have used the estimate $\hat{\theta}$ resulting from a single realization y of the output process $\boldsymbol{y}$. With the number of observations fixed and increasing noise level σ^2 the quality of the used estimates becomes poorer and thus eventually renders (67) a poor approximation of the distribution of $a'\hat{\boldsymbol{\theta}}$. On the other hand with increasing noise levels our whole approach to studying the statistics of $\hat{\boldsymbol{\theta}}$ by linearizing as above becomes questionable. To sum up we can therefore only expect (67) to reflect the statistical properties of $a'\hat{\boldsymbol{\theta}}$ for sufficiently large number of observations relative to the value of σ^2.

2 Extended Kalman Filter

In this section we shall show how one may attempt to apply the filtering theory developed for linear systems also in the case of nonlinear dynamics. The extension made will rely on linearization using Taylor's formula and will as such assume only small deviations around some nominal trajectory. The presentation given below will largely follow that of Jazwinski [1].

We shall consider the continuous-time SDE model with observations performed at discrete times t_k, $k = 0, 1, \ldots$

$$\begin{aligned} d\boldsymbol{x}(t) &= f(\boldsymbol{x}(t), u(t), \theta)dt + \sigma(\theta)d\boldsymbol{w}(t) \\ \boldsymbol{y}(t_k) &= g(\boldsymbol{x}(t_k), \theta) + \boldsymbol{v}(t_k) \end{aligned} \tag{71}$$

The initial state and noise variables are assumed to have the following distributions:

$$\begin{aligned} \boldsymbol{x}(t_0) &\sim N(\hat{x}_{0|-1}(\theta), P_{0|-1}(\theta)) \\ d\boldsymbol{w}(t) &\sim N_{iid}(0, Idt) \\ \boldsymbol{v}(t_k) &\sim N_{iid}(0, R(\theta)) \end{aligned} \tag{72}$$

Here *iid* signifies that the random variables in question are independent and identically distributed.

2.1 The idea of a nominal trajectory

We introduce the idea of a *nominal trajectory* $\bar{\boldsymbol{x}}(t)$ as the solution to the *deterministic* initial-value problem

$$\frac{d\bar{\boldsymbol{x}}(t)}{dt} = f(\bar{\boldsymbol{x}}(t), u(t), \theta), \qquad \bar{\boldsymbol{x}}(t_0) = \hat{x}_{0|0} \tag{73}$$

and the corresponding nominal measurement equation

$$\bar{\boldsymbol{y}}(t_k) = g(\bar{\boldsymbol{x}}(t_k), \theta) \tag{74}$$

With the initial condition $\bar{\boldsymbol{x}}(t_0) = \hat{x}_{0|0}$ in (73) we indicate that we assume that the *a priori* estimate $\hat{x}_{0|-1}$ has been adjusted using an incoming measurement $\boldsymbol{y}(t_0)$ resulting in the *a posteriori* estimate $\hat{x}_{0|0}$.

The idea behind the Extended Kalman Filter is that of calculating a nominal solution as a first approximation to the state estimate of the system in the hope that the stochastic disturbances are not too large. Then one linearizes around the resulting nominal solution in order to obtain a linear system with the task of tracking the stochastic disturbances. When a measurement $\boldsymbol{y}(t_k)$ arrives our knowledge of the system state $\boldsymbol{x}$ improves and this should give rise to a re-evaluation of the nominal trajectory around which we linearize and obtain our problem in the deviation variables. To facilitate this analysis we now establish the connection between the reference trajectory $\bar{\boldsymbol{x}}$ and the conditional expectation $\hat{x}_k$ of $\boldsymbol{x}$ given observations $Y_k = \{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$.

By writing the SDE

$$d\boldsymbol{x}(t) = f(\boldsymbol{x}(t), u(t), \theta)dt + \sigma(\theta)d\boldsymbol{w}(t) \tag{75}$$

in integral form we obtain that for $t \geq t_k$

$$\boldsymbol{x}(t) = \boldsymbol{x}(t_k) + \int_{t_k}^{t} f(\boldsymbol{x}(\tau), u(\tau), \theta)\, d\tau + \int_{t_k}^{t} \sigma(\theta)\, d\boldsymbol{\omega}(\tau) \tag{76}$$

By taking the ensemble average conditioned on Y_k and in the process using the definition of a stochastic integral in the sense of Itô it follows that

$$\hat{\boldsymbol{x}}_k(t) = \hat{\boldsymbol{x}}_k(t_k) + \int_{t_k}^{t} E\{f(\boldsymbol{x}(\tau), u(\tau), \theta)|Y_k\}\, d\tau \tag{77}$$

where

$$\hat{\boldsymbol{x}}_k(t) = E\{\boldsymbol{x}(t)|Y_k\} \tag{78}$$

and in particular

$$\hat{\boldsymbol{x}}_k(t_k) = E\{\boldsymbol{x}(t_k)|Y_k\} = \hat{x}_{k|k} \tag{79}$$

The function f being nonlinear we now expand it in a 1st order Taylor series around $\boldsymbol{x} = \hat{\boldsymbol{x}}_k$:

$$f(\boldsymbol{x}(\tau), u(\tau), \theta) \approx f(\hat{\boldsymbol{x}}_k(\tau), u(\tau), \theta) + \left(\frac{\partial f}{\partial \boldsymbol{x}}\right)_{\substack{\boldsymbol{x}=\hat{\boldsymbol{x}}_k(\tau) \\ u=u(\tau)}} (\boldsymbol{x}(\tau) - \hat{\boldsymbol{x}}_k(\tau)) \tag{80}$$

The fundamental assumption upon which the derivation of the Extended Kalman Filter rests is that despite the presence of non-linearities these are supposed to be sufficiently mild to allow us to assume that $\boldsymbol{x}(\tau)$ follows an approximately Gaussian distribution. In particular it will be assumed that the distribution of $\boldsymbol{x}(\tau)$ is symmetric w.r.t. the mean. Up to first order one thus obtains the approximation

$$E\{f(\boldsymbol{x}(\tau), u(\tau), \theta)|Y_k\} \approx f(\hat{\boldsymbol{x}}_k(\tau), u(\tau), \theta) \tag{81}$$

and using this the equation (77) takes the form

$$\hat{\boldsymbol{x}}_k(t) = \hat{\boldsymbol{x}}_k(t_k) + \int_{t_k}^{t} f(\hat{\boldsymbol{x}}_k(\tau), u(\tau), \theta)\, d\tau \tag{82}$$

In other words: with the approximations that we have employed by Taylor expanding to first order, $\hat{\boldsymbol{x}}_k(t)$ solves the initial-value problem

$$\begin{aligned} \frac{d\hat{\boldsymbol{x}}_k(t)}{dt} &= f(\hat{\boldsymbol{x}}_k(t), u(t), \theta), \qquad t \in [t_k; t_{k+1}] \\ \hat{\boldsymbol{x}}_k(t_k) &= \hat{x}_{k|k} \end{aligned} \tag{83}$$

Apart from the choice of initial-value we observe that this is equivalent to the initial-value problem (73).

2.2 Linearizing the system

With a view to linearizing the nonlinear stochastic system (71) we introduce on the interval $t_k \leq t < t_{k+1}$ the stochastic process $\boldsymbol{\xi}_k$ as the difference

$$\boldsymbol{\xi}_k(t) \equiv \boldsymbol{x}(t) - \hat{\boldsymbol{x}}_k(t) \tag{84}$$

conditioned on the measurements Y_k recorded up to and including time $t = t_k$. From (71) and (83) it is now clear that $\boldsymbol{\xi}_k(t)$ is a stochastic process satisfying the SDE

$$d\boldsymbol{\xi}_k(t) = (f(\boldsymbol{x}(t), u(t), \theta) - f(\hat{\boldsymbol{x}}_k(t), u(t), \theta))\, dt + \sigma(\theta) d\boldsymbol{w}(t) \tag{85}$$

with initial condition

$$\boldsymbol{\xi}_k(t_k) \sim N(0, P_{t_k}) \tag{86}$$

Proceeding heuristically it is now tempting to make the approximation

$$\begin{aligned} f(\boldsymbol{x}(t), u(t), \theta) - f(\hat{\boldsymbol{x}}_k(t), u(t), \theta) &\approx A_k(t, \theta)\,(\boldsymbol{x}(t) - \hat{\boldsymbol{x}}_k(t)) \\ &= A_k(t, \theta)\, \boldsymbol{\xi}_k(t) \end{aligned} \tag{87}$$

where the matrix A_k is given by

$$A_k(t, \theta) \equiv \left[\frac{\partial f(\boldsymbol{x}, u, \theta)}{\partial \boldsymbol{x}} \right]_{\substack{\boldsymbol{x}=\hat{\boldsymbol{x}}_k(t) \\ u=u(t)}} \tag{88}$$

We emphasize the dependence on $\hat{\boldsymbol{x}}_k(t)$, the nominal trajectory along which the partial derivatves are to be calculated.

For the *deviation variable* $\boldsymbol{\xi}_k$ we therefore consider following approximate *linear* stochastic initial-value problem:

$$\begin{aligned} d\boldsymbol{\xi}_k(t) &= A_k\left(t,\theta\right)\boldsymbol{\xi}_k(t)dt + \sigma(\theta)d\boldsymbol{w}(t) \\ \boldsymbol{\xi}_k(t_k)) &\sim N(0, P_{t_k}) \end{aligned} \tag{89}$$

We now linearize the measurement equation around the nominal measurement trajectory. The nominal measurement $\bar{\boldsymbol{y}}(t_k)$ by the time a measurement arrives at t_k is given in terms of the nominal trajectory calculated for the interval $t_k \leq t \leq t_{k+1}$:

$$\bar{\boldsymbol{y}}(t_k) = g(\hat{\boldsymbol{x}}_{k-1}(t_k)) \tag{90}$$

Introducing the deviation variable

$$\boldsymbol{\eta}_k \equiv \boldsymbol{y}(t_k) - g(\hat{\boldsymbol{x}}_{k-1}(t_k)) \tag{91}$$

we are led to the linearized measurement equation

$$\boldsymbol{\eta}_k = C_{k|k-1}\left(\theta\right)\boldsymbol{\xi}_{k-1}(t_k) + \boldsymbol{v}(t_k) \tag{92}$$

where

$$C_{k|k-1}\left(\theta\right) \equiv \left[\frac{\partial g(\boldsymbol{x},\theta)}{\partial \boldsymbol{x}}\right]_{\boldsymbol{x}=\hat{\boldsymbol{x}}_{k-1}(t_k)} \tag{93}$$

The deviation variables $\boldsymbol{\xi}_k$ and $\boldsymbol{\eta}_k$ therefore satisfy the following *linear* stochastic system on $t_k \leq t \leq t_{k+1}$

$$\begin{aligned} d\boldsymbol{\xi}_k(t) &= A_k(t,\theta)\boldsymbol{\xi}_k(t)dt + \sigma(\theta)d\boldsymbol{w}(t) \\ \boldsymbol{\eta}_k &= C_{k|k-1}(\theta)\boldsymbol{\xi}(t_k) + \boldsymbol{v}(t_k) \end{aligned} \tag{94}$$

with initial conditions and noise distributions given by

$$\begin{aligned} \boldsymbol{\xi}_k(t_k)) &\sim N(0, P_{t_k}) \\ d\boldsymbol{w}(t) &\sim N_{iid}(0, Idt) \\ \boldsymbol{v}(t_k) &\sim N_{iid}(0, R(\theta)) \end{aligned} \tag{95}$$

2.3 The measurement update

The linearized stochastic system (94), (95) in the *deviation variables* $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ is one to which the linear theory of Kalman Filtering applies. In particular it allows for measurement update of $\hat{\boldsymbol{\xi}}_k$ based on an incoming measurement $\boldsymbol{y}(t_{k+1})$ (through $\boldsymbol{\eta}_{k+1} = \boldsymbol{y}(t_{k+1}) - g(\hat{\boldsymbol{x}}_{k+1|k})$).

At $t = t_{k+1}$ immediately prior to the arrival of a new measurement we would thus have the following estimate

$$\hat{\boldsymbol{x}}_{k+1|k} = \hat{\boldsymbol{x}}_k(t_{k+1}) \tag{96}$$

For an estimate of the variance of the estimate (96) we have to rely on the linearized expression (94) for the evolution of the stochastic part. Denoting the *a posteriori* state error covariance at $t = t_k$ by $P_{k|k}$ we have from (94) that the *a priori* covariance $P_{k+1|k}$ is given by $P_k(t_{k+1})$ where P_k is found by solving the following set of coupled differential equations:

$$\begin{aligned} \frac{d\hat{\boldsymbol{x}}_k(t)}{dt} &= f(\hat{\boldsymbol{x}}_k(t), u(t), \theta), && t \in [t_k; t_{k+1}] \\ A_k(t) &= \frac{\partial f}{\partial x}(\hat{\boldsymbol{x}}_k(t), u(t), \theta) && t \in [t_k; t_{k+1}] \\ \frac{dP_k}{dt} &= A_k(t)P_k(t) + P_k(t)A_k(t)^{'} + \sigma(\theta)\sigma(\theta)^{'} && t \in [t_k; t_{k+1}] \end{aligned} \tag{97}$$

whose solution in turn must be coupled with that of (83) since $A_k(t)$ depends on $\hat{\boldsymbol{x}}_k(t)$.

As already mentioned above the standing assumption underlying our derivation of the Extended Kalman Filter is that despite the presence of nonlinearities these are mild enough for the resulting distributions still to be approximately Gaussian. This is reflected in the linearity of the system (94) governing the evolution of the deviation variables.

From (91) we see that an incoming measurement $\boldsymbol{y}(t_k)$ amounts to a measurement of the deviation variable $\boldsymbol{\eta}_k$.

Assuming only mild nonlinearities in the measurement function g we use Taylor expansion around $\boldsymbol{x} = \hat{\boldsymbol{x}}_{k|k-1}$ to make the approximation

$$g(\boldsymbol{x}(t_k)) \approx g(\hat{\boldsymbol{x}}_{k|k-1}) + \left(\frac{\partial g}{\partial x}\right)_{\boldsymbol{x}=\hat{\boldsymbol{x}}_{k|k-1}} (\boldsymbol{x}(t_k) - \hat{\boldsymbol{x}}_{k|k-1}) \tag{98}$$

Since $\hat{\boldsymbol{x}}_{k|k-1} = E\{\boldsymbol{x}(t_k)|Y_{k-1}\}$ we therefore infer that

$$\hat{\boldsymbol{y}}_{k|k-1} = E\{g(\boldsymbol{x}(t_k))|Y_{k-1}\} \approx g(\hat{\boldsymbol{x}}_{k|k-1}) \tag{99}$$

Equation (91) thus has an equivalent formulation as

$$\boldsymbol{\eta}(t_k) = \boldsymbol{y}(t_k) - \hat{\boldsymbol{y}}_{k|k-1} \tag{100}$$

In keeping with the jargon known from the conventional theory of linear Kalman Filtering we shall therefore refer to $\boldsymbol{\eta}(t_k)$ as the *innovation* and denote it by $\boldsymbol{e}_k$. From (94) we note that its covariance matrix is given by

$$R_{e,k}(\theta) = C_{k|k-1}(\theta)P_{k|k-1}C_{k|k-1}(\theta)' + R(\theta) \tag{101}$$

To perform the measurement update of $\boldsymbol{x}$ at $t = t_k$ we note that

$$\begin{aligned} E\{\boldsymbol{x}(t_k)|Y_k\} &\approx E\{\hat{\boldsymbol{x}}_{k-1}(t_k) + \boldsymbol{\xi}_{k-1}(t_k)|Y_k\} \\ &= \hat{\boldsymbol{x}}_{k-1}(t_k) + E\{\boldsymbol{\xi}_{k-1}(t_k)|Y_k\} \\ &= \hat{\boldsymbol{x}}_{k|k-1} + E\{\boldsymbol{\xi}_{k-1}(t_k)|Y_k\} \end{aligned} \tag{102}$$

Updating $\boldsymbol{x}$ thus boils down to performing the standard measurement update of $\boldsymbol{\xi}$ for the linearized measurement equation

$$\boldsymbol{\eta}_k = C_{k|k-1}(\theta)\boldsymbol{\xi}_{k-1}(t_k) + \boldsymbol{v}(t_k) \tag{103}$$

using that the *a priori* deviation state error covariance at $t = t_k$ equals $P_{k|k-1}$ and $\boldsymbol{v}(t_k) \sim N(0, R(\theta))$ is independent of $\boldsymbol{x}(t_k)$.

Denoting the *Kalman gain* for the system (94)-(95) by

$$K_{f,k} = P_{k|k-1}(\theta)C_{k|k-1}(\theta)'R_{e,k}(\theta)^{-1} \tag{104}$$

one obtains that

$$\begin{aligned} E\{\boldsymbol{\xi}_{k-1}(t_k)|Y_k\} &= E\{\boldsymbol{\xi}_{k-1}(t_k)|Y_{k-1}\} + K_{f,k}(\boldsymbol{\eta}_k - \hat{\boldsymbol{\eta}}_{k|k-1}) \\ &= \hat{\boldsymbol{\xi}}_{k|k-1} + K_{f,k}(\boldsymbol{\eta}_k - C\left[\hat{\boldsymbol{x}}_{k-1}(t_k), \theta\right]\hat{\boldsymbol{\xi}}_{k|k-1}) \\ &= K_{f,k}\boldsymbol{\eta}_k \\ &= K_{f,k}\boldsymbol{e}_k \end{aligned} \tag{105}$$

since $\hat{\boldsymbol{\xi}}_{k|k-1} = 0$. This is due to the fact that we realign the initial conditions at every measurement update so that $\hat{\boldsymbol{\xi}}_{k-1|k-1} = 0$. The linearity of the SDE governing the evolution of $\boldsymbol{\xi}_{k-1}$ implies that also $\hat{\boldsymbol{\xi}}_{k|k-1} = 0$.

2.4 Summary of the EKF algorithm

In the two frames below we summarize in compact form the filtering algorithm derived in the previous three subsections. We separate into Measurement Updates and Time Updates. First we list the Measurement Update.

Continuous-Discrete EKF algorithm - Measurement Update

1. One step prediction of the measurement

$$\hat{\boldsymbol{y}}_{k|k-1} = g(\hat{\boldsymbol{x}}_{k|k-1})$$
$$C_{k|k-1}(\theta) = \left[\frac{\partial g(\boldsymbol{x},\theta)}{\partial \boldsymbol{x}}\right]_{\boldsymbol{x}=\hat{\boldsymbol{x}}_{k-1}(t_k)}$$

2. Innovation

$$\boldsymbol{e}_k = \boldsymbol{y}(t_k) - \hat{\boldsymbol{y}}_{k|k-1}$$
$$R_{e,k}(\theta) = C_{k|k-1}(\theta)P_{k|k-1}C_{k|k-1}(\theta)' + R(\theta)$$

3. Filter Constant

$$K_{f,k} = P_{k|k-1}(\theta)C_{k|k-1}(\theta)'R_{e,k}(\theta)^{-1}$$

4. Filtered State

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + K_{f,k}\boldsymbol{e}_n$$
$$P_{k|k} = P_{k|k-1} - K_{f,k}R_{e,k}K'_{f,k}$$

And we continue with the Time Update, the projection ahead in between samples.

Continuous-Discrete EKF algorithm - Time Update

1. Initial conditions

$$\hat{\boldsymbol{x}}_k(t_k) = \hat{\boldsymbol{x}}_{k|k}(\theta)$$
$$P_k(t_k) = P_{k|k}(\theta)$$

2. Solve the coupled differential equations

$$\frac{d\hat{\boldsymbol{x}}_k(t)}{dt} = f(\hat{\boldsymbol{x}}_k(t), u(t), \theta), \qquad t \in [t_k; t_{k+1}]$$
$$A_k(t) = \frac{\partial f}{\partial x}(\hat{\boldsymbol{x}}_k(t), u(t), \theta) \qquad t \in [t_k; t_{k+1}]$$
$$\frac{dP_k}{dt} = A_k(t)P_k(t) + P_k(t)A_k(t)' + \sigma(\theta)\sigma(\theta)' \qquad t \in [t_k; t_{k+1}]$$

3. One-step prediction of the state and the associated covariance

$$\hat{\boldsymbol{x}}_{k+1|k}(\theta) = \hat{\boldsymbol{x}}_k(t_{k+1})$$
$$P_{k+1|k} = P_k(t_{k+1})$$

3 A variant of the standard SDE model

Sometimes the modelling process gives rise to models which may not be of the standard form (71). Indeed it is conceivable that it is for one reason or another more convenient to formulate the model as

$$dg(\boldsymbol{x}(t)) = f(\boldsymbol{x}(t), u(t), \theta)\, dt + \sigma(\theta)\, d\boldsymbol{\omega}(t) \tag{106}$$

or in equivalent integral form

$$g(\boldsymbol{x}(t)) - g(\boldsymbol{x}(t_0)) = \int_{t_0}^{t} f(\boldsymbol{x}(\tau), u(\tau), \theta)\, d\tau + \int_{t_0}^{t} \sigma(\theta)\, d\boldsymbol{\omega}(\tau) \tag{107}$$

In the further development of this example we shall assume that $g : \mathbb{R}^n \to \mathbb{R}^n$ is a mapping between Euclidean spaces of the same dimension as that of the state vector $\boldsymbol{x}$. We shall also assume that the state vector $\boldsymbol{x}$ assumes values in an open subset $\Omega \subset \mathbb{R}^n$ to which the restriction of g is one-to-one. We introduce the notation $\boldsymbol{y}(t) = g(\boldsymbol{x}(t))$ and see that $\boldsymbol{y}$ must satisfy the following SDE in standard form:

$$d\boldsymbol{y}(t) = f(g^{-1}(\boldsymbol{y}(t)), u(t), \theta)\, dt + \sigma(\theta)\, d\boldsymbol{\omega}(t) \tag{108}$$

Let us now impose the zero-order-hold (ZOH) assumption, i.e. that $u(t) = u_k$ for $t_k \leq t < t_{k+1}$. We consider a given set of sampling times $t_0, t_1, \ldots, t_N$ and assume that the components of the state vector $\boldsymbol{x}$ are all measurable and neglect the issue of measurement noise. Then we have full knowledge of $\boldsymbol{y}(t_i)$ at any sampling time t_i. Under these assumption we now apply the material developed in the preceding section on the Extended Kalman Filter (EKF). This involves approximating the development of the conditional mean and covariance of $\hat{\boldsymbol{y}}$ by solving the following ordinary differential equations for $t_i \leq t \leq t_{i+1}$:

$$\begin{aligned} \dot{\hat{\boldsymbol{y}}}_{t|t_i} &= f(g^{-1}(\hat{\boldsymbol{y}}_{t|t_i}), u(t), \theta), & \hat{\boldsymbol{y}}_{t_i|t_i} &= \boldsymbol{y}(t_i) \\ \dot{P}_{t|t_i} &= A_i(t)P_{t|t_i} + P_{t|t_i}A_i(t)^T + \sigma(\theta)\sigma(\theta)^T, & P_{t_i|t_i} &= 0 \end{aligned} \tag{109}$$

where

$$A_i(t) = \left(\frac{\partial (f g^{-1})}{\partial \boldsymbol{y}} \right)_{\boldsymbol{y} = \hat{\boldsymbol{y}}_{t|t_i}} \tag{110}$$

We now employ this machinery to perform ML-estimation of parameters in the model. That is, given observations $\{\boldsymbol{x}(t_0), \boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_N)\}$ or equivalently $\{\boldsymbol{y}(t_0), \boldsymbol{y}(t_1), \ldots, \boldsymbol{y}(t_N)\}$ we shall attempt to maximize the conditional probability

$$P\left(\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\} | \boldsymbol{y}_0, \theta\right) \tag{111}$$

with respect to the parameter vector θ.

In our special case where we have assumed all components of the state vector $\boldsymbol{y}$ perfectly measurable we may factor (111) as follows

$$P\left(\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n\}|\boldsymbol{y}_0,\theta\right)=\prod_{k=1}^{N}P\left(\boldsymbol{y}_k|\boldsymbol{y}_{k-1},\theta\right) \tag{112}$$

We make the assumption that the conditional probability $P\left(\boldsymbol{y}_k|\boldsymbol{y}_{k-1},\theta\right)$ may be well approximated by the Gaussian distribution with mean value μ_k and variance P_k given by

$$\mu_k=\mu_k(\theta)=\hat{\boldsymbol{y}}_{t_k|t_{k-1}}\qquad P_k=P_k(\theta)=P_{t_k|t_{k-1}} \tag{113}$$

So

$$P\left(\boldsymbol{y}_k|\boldsymbol{y}_{k-1},\theta\right)\approx\frac{1}{\sqrt{(2\pi)^n|P_k|}}\exp\left(-\frac{1}{2}(\boldsymbol{y}_k-\mu_k)^TP_k^{-1}(\boldsymbol{y}_k-\mu_k)\right) \tag{114}$$

where by $|P_k|$ we have denoted the determinant of the matrix P_k. The task of maximizing (111) therefore is equivalent to that of maximizing the expression

$$\prod_{k=1}^{N}\frac{1}{\sqrt{|P_k|}}\exp\left(-\frac{1}{2}(\boldsymbol{y}_k-\mu_k)^TP_k^{-1}(\boldsymbol{y}_k-\mu_k)\right) \tag{115}$$

In turn since the logarithm is a monotonically increasing function this is equivalent to that of mimimizing the function

$$\Phi(\theta)=\sum_{k=1}^{N}\log|P_k(\theta)|+(\boldsymbol{y}_k-\mu_k(\theta))^TP_k^{-1}(\theta)(\boldsymbol{y}_k-\mu_k(\theta)) \tag{116}$$

with respect to θ.

References

[1] A. H. Jazwinski. *Stochastic Processes and Filtering Theory.* Academic Press, San Diego, CA, 1970.

[2] Robert I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643, 04 1969.

[3] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization, 2nd Edition.* Springer, New York, USA, 2006.