
Parameter Identification Based on Conditional Expectation

Elmar Zander,^{1,*} Noémi Friedman² and Hermann G. Matthies¹

Parameter identification is an important issue in many scientific and technical disciplines. We present here a technique that updates our knowledge of the system parameters based on the so-called conditional expectation. It will be shown that for linear systems with normally distributed uncertainties this reproduces exactly the Bayes' posterior and is thus a non-linear extension of the Kalman filter.

4.1 Introduction

Estimation of model parameters is a very common problem in the natural and technical sciences. To eliminate problem specific details we can cast it into the following abstract form: Let q be a vector of parameters and u the complete state of the system; then there is a relation

$$A(u; q) = 0 \tag{4.1}$$

where A is a model of the system, given, for example, by a set of algebraic or differential equations. If the model is well-posed in the sense of Hadamard [84], then there is a unique solution for u given the parameters q and this dependence is continuous in q – that is, small changes in q invoke only small changes in u .

In many cases of interest, the full state of the system \mathbf{u} is not directly observable. Rather, we can observe some measurements performed on the system that give us data

$$y = M(u; q) + \epsilon, \tag{4.2}$$

which depends on the state u and possibly also on the parameters q and is usually contaminated by measurement noise ϵ . Here, M signifies a mathematical model of the measurement process.

Example 7 *A simple example for demonstration is the flow of groundwater where parameters like permeability or boundary conditions shall be inferred from measurements inside of the domain. A standard linear model for groundwater flow is given by Darcy's equation*

$$-\nabla \cdot (\kappa \nabla u) = f$$

¹ Technische Universität Braunschweig, Germany.

² Institute for Computer Science and Control (SZTAKI), Budapest, Hungary.

* Corresponding author: elmar@zandere.de

on the domain \mathcal{D} with Neumann conditions

$$\kappa \nabla u = g$$

on the boundary $\partial\mathcal{D}$. Here, u is the hydraulic head, f describes the sinks and sources inside the domain, and g denotes the inflow and outflow via the boundary.

Let the hydraulic head be measured at N locations x_1, \dots, x_N in \mathcal{D} and the parameters to be inferred $q = (\kappa, f, g)$, which are assumed to be constant on the domain for simplicity. The system model is then given by

$$A(u; q) = \begin{pmatrix} -\nabla \cdot (\kappa \nabla u) - f \\ (\kappa \nabla u - g)|_{\partial\mathcal{D}} \end{pmatrix}$$

and the measurement operator by¹

$$M(u; q) = (u(x_1), \dots, u(x_n))$$

Using the system model A , the measurement model M , and values of the parameters q we can make predictions about the expected outcome of the measurements. We call this prediction the system response or response surface $G(q)$. However, given some actually observed measurements y_m , most likely they will not be identical to the predicted measurements $y = G(q)$. This disagreement between measurement and prediction can have multiple causes of which the most important ones are the aforementioned measurement error ϵ , wrong values for the parameters q not consistent with the true parameters q_{true} , numerical errors in computing u or y , and finally, an inadequate or inaccurate model for A or M , maybe omitting important physical effects.

In this chapter, we will only deal with errors of the first and second kind, that is, we assume the models are adequate for the process under investigation and the computational errors are negligible compared to the other types of errors. For the treatment and incorporation of modelling errors we refer the interested reader to the literature; see, for example [179, 170].

The disagreement between y and y_m can be seen as nuisance, but also as an opportunity to infer better knowledge about the true values of the parameters. In classical parameter estimation procedures we can use this information to update our belief about the value of the parameters q . A common approach is to choose a new set of parameters q' by minimising the distance between the predicted value y and y_m with respect to some loss or cost function. However, this is very often an ill-posed problem as the minimiser is usually not uniquely determined. A typical way to go is to use some regularisation scheme, for example, restricting q' in some norm or the difference between q and q' in another.

Though those schemes can turn the problem into a well-posed one, they suffer from being somewhat arbitrary and there is usually no good reason to choose one regularisation scheme over another – at least from a modelling point of view, maybe from a computational viewpoint there is.

In this chapter, we choose the Bayesian point of view to which this book is also devoted, acknowledging our lack of knowledge by modelling the parameter not as a single, deterministic value, but as a random variable (see Section 4.1.1). Our uncertainty about the true value is then determined by the variance of the random variable. Point estimates for q can then be given by the mean or median of the random variable.

In order to distinguish deterministic values from random variables we will use small letters for the former and capital letters for the latter. So, Q would denote the random variable corresponding

¹For a more realistic and elaborate example see Chapter 8.

to the parameters and Y the one for the measurements. The aim of the present chapter is therefore to present a method to compute a new random variable Q' , such that actual measurements y_m of Y are taken into account and the probability distribution of Q' is adjusted accordingly.

4.1.1 Preliminaries—basics of probability and information

Note, that a thorough treatment of the topics discussed here would need some familiarity with measure theory, which we do not assume here and only give an intuitive (mathematicians would call it sloppy) introduction to the definitions and ideas, that are needed in later sections. For more formal definitions and exact theorems see, for example, [80].

The mathematical definition of a probability space consists of three ingredients: a set of elementary events Ω , a set of events \mathcal{F} – mathematically this is a structure called a σ -algebra – and a probability measure \mathbb{P} . The elementary events $\omega \in \Omega$ are the singular outcomes that will happen by chance or are the possible underlying events that characterise our insufficient knowledge, depending on whether we model with the probability space truly random events (aleatory uncertainty) or lack of knowledge (epistemic uncertainty).

For practical purposes it is often not necessary – and often also not possible – to know which elementary event exactly happened, but rather whether it lay in some specific subset of Ω or not. Such a set is commonly called an *event*. As the notion of an “event” is not really suitable for the case that the set describes uncertain knowledge, we call it here a “hypothesis”, meaning the proposition whose true but unknown value lies inside the given set.

This also simplifies the definition of the probability measure, as it has to be defined only for those events in \mathcal{F} . Mathematically, the probability measure $\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}^+$ is thus a function from the set of events \mathcal{F} to the positive real numbers. Since from all possible events in Ω , one must have happened (or equivalently, one value of ω must be the true value), we have the condition $\mathbb{P}(\Omega) = 1$, which is also called the *normalisation condition*.

The notion of hypotheses or events as subsets of the sample space Ω is also connected with that of *information*. Practically, we are usually not interested in which elementary event exactly will happen, but rather in the answer to questions like “Will it rain tomorrow?” or “Will the structure withstand the load?” If we put all outcomes ω for which the answer is “yes” into a set, say A , we can only answer questions about the probability of those hypotheses, when this set A is an element of \mathcal{F} . From this it becomes clear, that the more refined the subsets in \mathcal{F} are, the more detailed questions we can answer within our probability model.

4.1.1.1 Random variables

In most elementary texts on probability, a random variable is a variable that “somehow” takes on different values according to some probability distribution. Though for basic results in probability, such a casual notion will suffice, we need a more formal approach here.

In the formal, axiomatic treatment of probability after Kolmogorov [112] a *random variable*, say X , is a so-called measurable function from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into the real numbers. That means that every “sensible” subset of the real numbers² U has a preimage in \mathcal{F} , that is, there is a set $A \in \mathcal{F}$ such that $X(A) = U$, or equivalently $X^{-1}(U) \in \mathcal{F}$.

By collecting all preimages of a random variable, we can thus also define a sigma algebra. This is often denoted by $\sigma(X)$, the sigma algebra generated by the random variable X . The sets in

²Mathematically, those sets are called the Borel sets, but we don’t need that kind of mathematical rigour here.

$\sigma(X)$ define thus what kind of information can be differentiated by observing values of the random variable X .

In some contexts, random variables that take values in a finite dimensional vector space are called *random vectors* or if the target space is a more general topological vector space (like a Banach or Hilbert space) they are called *random elements*. We will generally not make this distinction here and call all of them random variables.

4.1.2 Bayes' theorem

Although a detailed introduction to Bayes' theorem and Bayesian inverse problems was provided in Chapter 1, a summary is provided here for the sake of consistency and unified notation.

As discussed in Chapter 1, epistemic uncertainties due to a lack of knowledge are not intrinsic to the natural or technical system under investigation itself, but rather characterise our state of knowledge about that system. We encode that knowledge into specifying probabilities for different hypotheses over that state or over the parameters describing the system. Gaining knowledge by observing or performing measurements using the system will consequently increase our knowledge and thus modify the probability we will assign to these hypotheses $H \in \mathcal{F}$. We will call the probability measure before we have the new information of the *prior* probability and after incorporation of the information of the *posterior* probability.

Let us say that the information that we have – maybe from gathering data by performing measurements on the system – is codified in some set $E \in \mathcal{F}$, also called the *evidence*. We are interested now in the probability of some hypothesis H , given that we have already learned the evidence E – this probability will be denoted by $\mathbb{P}(H|E)$. Now, all events for which H is true, given that we have the additional information E , have to be in the intersection $H \cap E$, and thus the probability of the hypothesis H given knowledge of the evidence E must be proportional to that of $H \cap E$. Because the probability of E given that E is known must trivially be 1, it becomes obvious that the constant of proportionality must be $1/\mathbb{P}(E)$. Thus, we can define the conditional probability of H given E by

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(H \cap E)}{\mathbb{P}(E)}. \quad (4.3)$$

As this relation must also be true when the roles of H and E are interchanged, that is,

$$\mathbb{P}(E|H) = \frac{\mathbb{P}(E \cap H)}{\mathbb{P}(H)}, \quad (4.4)$$

it follows easily that

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(E)}, \quad (4.5)$$

which is the well-known *Bayes' theorem*. In the Bayesian framework, the terms appearing in Equation (4.5) have special meanings: $\mathbb{P}(H)$ is called the *prior probability* or just the *prior*, which is the state of knowledge before any measurements are available, while $\mathbb{P}(H|E)$, the *posterior probability* represents the state of knowledge after having learned the evidence. $\mathbb{P}(E|H)$ is called the *likelihood* and denotes the probability of measuring the evidence given that the hypothesis H is true.

The significance of the Bayes' theorem stems from the fact that it relates quantities that are easily calculable with quantities that are of interest. For example, it is generally easy to calculate

the likelihood of getting the evidence E given that the hypothesis is true, since this is a direct consequence of the modelling assumptions. But what is really of interest (and this is in contrast to classical statistical methods such as maximum likelihood) is the probability of the hypothesis being true after acquiring the evidence, and this link can be established by Bayes' theorem. One possible drawback is that for the calculation of $\mathbb{P}(H|E)$, we need to make assumptions about the prior probability of the hypothesis $\mathbb{P}(H)$. The choice of the prior is often a subjective issue, which caused some controversy about the objectivity of Bayesian methods. However, in classical statistical methods there are also hidden assumptions which have to be made explicit in the Bayesian framework. Furthermore, there have been many advances to make the prior assumptions less subjective by using, for example, uninformative priors or empirical Bayes methods. For detailed accounts see [36, 74].

Example 8 Referring back to the context given in the introduction where we had parameters $q \in \mathcal{Q}$ and measurements $y \in \mathcal{Y}$, the hypothesis H could correspond to the event that the parameters q are in some specific subset of the parameters space $\mathcal{Q}_H \subset \mathcal{Q}$, that is, $H = \{\omega | Q(\omega) \in \mathcal{Q}_H\} = Q^{-1}(\mathcal{Q}_H)$. The evidence E could correspond to the event that the measurements y lie in a specific subset of the space $\mathcal{Y}_E \subset \mathcal{Y}$, that is, $E = \{\omega | \mathcal{Y}_E(\omega) \in \mathcal{Y}_E\} = Y^{-1}(\mathcal{Y}_E)$.

In many practical problems we need to condition on single measurements $Y_E = \{y_m\}$. Unfortunately, if the range of the measurements is continuous, the probability of such an event is zero in general. Since then the evidence $\mathbb{P}(E)$ and the likelihood $\mathbb{P}(E|H)$ is zero, the right hand side of Equation (4.5) is undefined. We can avoid this by conditioning on a finite set $\mathcal{Y}_E = \{y | d(y, y_m) \leq \Delta y\}$, where d is some distance function, for example, the Euclidean distance. Letting Δy go to zero and doing the equivalent thing for the hypothesis leads to

$$\pi(q|y) = \frac{\pi(y|q)\pi(q)}{\pi(y)}, \quad (4.6)$$

where π is the probability density of the given random variable. This is called the Bayes' theorem for densities.

Note that this limiting process can be somewhat problematic, as the density is not invariant under non-linear transformations of the measurement variable. For an example, the so-called Borel-Kolmogorov "paradox", see, for example, [102]. However, if the limit process is consistent with the measurement process, that is, the measured values are taken as is or only linearly transformed, the problems can be circumvented.

4.1.3 Conditional expectation

The paradoxes involved in using conditional probabilities can also be resolved by switching to conditional expectations instead. The classical way to define the conditional expectation is with respect to an event.

In the case that Q is a continuous random variable and $E \subset \mathcal{F}$ is an event, the conditional expectation of Q given E is given by

$$\mathbb{E}[Q|E] = \frac{1}{\mathbb{P}(E)} \int_E Q(\omega) \mathbb{P}(d\omega). \quad (4.7)$$

Here, $\mathbb{P}(d\omega)$ means the probability of the infinitesimally small set of size $d\omega$ located at ω , which can also be written as $\pi(\omega)d\omega$ if the \mathbb{P} has continuous density π on Ω .

In case the event has zero probability, which again happens for events E of the form $Y^{-1}(\{y_m\})$, Equation (4.7) also becomes indefinite. The key in generalising this is to first rearrange the formula into

$$\int_Q \mathbb{E}[Q|E] \mathbb{P}(d\omega) = \int_E Q(\omega) \mathbb{P}(d\omega) \quad (4.8)$$

which is trivially true, if $\mathbb{P}(E) = 0$. Now, we define the conditional expectation as a random variable Q_Y by requiring that

$$\int_E Q_Y(\omega) \mathbb{P}(d\omega) = \int_E Q(\omega) \mathbb{P}(d\omega) \quad (4.9)$$

holds for all $E \in \sigma(Y)$. The conditional expectation Q_Y is very often denoted by $\mathbb{E}[Q|Y]$, but we prefer the notation Q_Y , because it makes it more evident, that this is a random variable and not a deterministic value.

It is apparent from the last equation that the conditional expectation $\mathbb{E}[Q|Y]$ depends only on the permitted sets $Q \in \sigma(Y)$ and is thus independent of transformations of the measurements Y . Counter-intuitive results such as the Borel-Kolmogorov paradox are thus impossible in this context and we will thus base our updating strategy rather on the conditional expectation than on conditional probabilities.

Note, that the definition of the conditional expectation as given above, only implicitly defines it, but does not say how to compute it. However, there is the closely related notion of the *minimum mean square error* estimator which allows efficient numerical approximations as will be shown in the next section.

4.2 The Mean Square Error Estimator

We call an *estimator* any possible function from the space of measurements \mathcal{Y} to the space of unknowns \mathcal{Q} . From all of these functions – from which most would not even deserve the name estimator in the usual sense – we can select one by defining a measure on how close the estimates $\varphi(Y)$ come to the true value Q .

A common measure of closeness, which also has nice analytical properties, is the *mean squared error*, defined in the present setting by

$$e_{\text{MSE}}^2 = \mathbb{E}[\|Q - \varphi(Y)\|^2] \quad (4.10)$$

$$= \int_{\Omega} \|Q(\omega) - \varphi(Y(\omega))\|^2 \mathbb{P}(d\omega) \quad (4.11)$$

This assigns to each estimator φ the mean value of the squared error that is made, when trying to predict the parameter Q from any possible realisation of the measurement random variable Y . The minimum mean square error estimator $\hat{\varphi}$ is the one that minimises the mean square error e_{MSE} , and can be written as

$$\hat{\varphi} = \arg \min_{\varphi \in \mathcal{L}(\mathcal{Q}, \mathcal{Y})} \mathbb{E}[\|Q - \varphi(Y)\|^2]. \quad (4.12)$$

Here, $\mathcal{L}(\mathcal{Q}, \mathcal{Y})$ denotes the space of measurable functions, which includes a large amount of functions, for example, all the continuous functions from \mathcal{Q} to \mathcal{Y} .

It can be shown that the MMSE estimator $\hat{\varphi}$ is the conditional expectation of Q given the measurements Y , that is,

$$\hat{\varphi}(Y) = Q_Y. \quad (4.13)$$

REMARK 1.1: A heuristic argument that helps to see the equality in (4.13) is, that for some random variable X , the real number u that minimises $\mathbb{E}[(X - u)^2]$ is given by $u = \mathbb{E}[X]$. For simple random variables, which take on only finitely many different values, this leads directly to the proof taking $X = Q \cdot \mathbb{1}_{Y=y}$ for some $y \in \mathcal{Y}$ and $\mathbb{1}_A$ being the indicator function for the set A . For a more rigorous treatment see, for example, [147] or [191]

4.2.1 Numerical approximation of the MMSE

In Equation (4.12) the minimisation is done over the whole space of measurable functions $\mathcal{L}(\mathcal{Q}, \mathcal{Y})$. Because this is in general an infinite dimensional space, we have to restrict it to a finite dimensional subspace $\mathcal{V}_\varphi \subset \mathcal{L}(\mathcal{Q}, \mathcal{Y})$ to make the problem computationally feasible. Let this space be defined by basis functions Ψ_γ , where $\gamma \in \mathcal{J}$ is an index and \mathcal{J} the set of indices. The functions Ψ_γ can be, for example, some sort of multivariate polynomials and the γ corresponding multiindices, but other function systems are also possible (e.g. tensor products of sines and cosines). An element φ of this function space has a representation as a linear combination

$$\varphi(y) = \sum_{\gamma \in \mathcal{J}} \varphi_\gamma \Psi_\gamma(y) \quad (4.14)$$

of these basis functions. Let us suppose for the moment that Q is a scalar-valued random variable and the coefficients φ_γ are therefore scalar-valued, too. Minimising expression (4.11) for φ then becomes equivalent to solving

$$\frac{\partial}{\partial \varphi_\delta} \mathbb{E}[(Q - \sum_{\gamma} \varphi_\gamma \Psi_\gamma(Y))^2] = 0 \quad (4.15)$$

for all $\delta \in \mathcal{J}$. Using the linearity of the derivative operator and of the expectation leads to

$$\sum_{\gamma} \varphi_\gamma \mathbb{E}[\Psi_\gamma(Y) \Psi_\delta(Y)] = \mathbb{E}[Q \Psi_\delta(Y)]. \quad (4.16)$$

As this is a linear system of equations we can rewrite it in the compact form

$$\mathbf{A}\boldsymbol{\varphi} = \mathbf{b} \quad (4.17)$$

with

$$[\mathbf{A}]_{\gamma\delta} = \mathbb{E}[\Psi_\gamma(Y) \Psi_\delta(Y)], \quad (4.18)$$

$$[\mathbf{b}]_\delta = \mathbb{E}[Q \Psi_\delta(Y)] \quad (4.19)$$

coefficients φ_γ collected in the vector $\boldsymbol{\varphi}$. Note, that for the actual computation some linear ordering needs to be imposed on the indices $\gamma \in \mathcal{J}$, but this is not essential here and can be left to the implementation.

If the unknown Q and the measurements Y are given by polynomials – like, for example, a Wiener polynomial chaos expansion (PCE) or a generalized polynomial chaos expansion (GPC) [75, 192] – and the function space \mathcal{V}_φ consists of polynomials, the expectations could in principle be computed exactly using the polynomial algebra. However, this is computationally very expensive and non-trivial to implement. More efficient is to approximate \mathbf{A} and \mathbf{b} by numerical integration via

$$\mathbb{E}[\Psi_\gamma(Y) \Psi_\delta(Y)] \approx \sum_k w_k \Psi_\gamma(Y(\xi_k)) \Psi_\delta(Y(\xi_k)) \quad (4.20)$$

and

$$\mathbb{E}[Q \Psi_\delta(Y)] \approx \sum_k w_k Q(\xi_k) \Psi_\delta(Y(\xi_k)). \quad (4.21)$$

Choosing an integration rule of sufficient polynomial exactness these relations can also be made exact.

REMARK 1.2: Suppose Y and Q are polynomials of total degree p_Y and p_Q , respectively, and φ has total degree p_φ . Then the maximum degree in the expression for \mathbf{A} will be $2p_Y p_\varphi$ and a Gauss integration rule of order $p_Y p_\varphi + 1$ will suffice. For the computation of \mathbf{b} a rule of order $\lceil (p_Q + p_Y p_\varphi + 1)/2 \rceil$ will suffice for exactness of the integration. In practice, usually integration rules of lower degree of exactness have shown to work well.

In the case that Q is a vector-valued random variable, which it usually is, the component functions φ_i of φ in expression (4.11) approximating the components Q_i for $i \in [1 \dots n]$ are completely independent. So, the problem of computing the minimiser in Equation (4.15) essentially factors into n independent problems and can be done component-wise. In order to compute the estimator $\hat{\varphi}$ now for a vector valued Q the vectors φ_i and \mathbf{b}_i (defined by Equation (4.19) for each Q_i) can be collected into matrices and the whole system

$$\mathbf{A}[\varphi_1, \dots, \varphi_n] = [\mathbf{b}_1, \dots, \mathbf{b}_n] \quad (4.22)$$

solved at once, which makes the process often more efficient, especially if factorisation of the matrix \mathbf{A} is involved.

4.2.2 Numerical examples

In this section, we present two examples for the numerical approximation of the conditional expectation via the MMSE. In the first we take two random variables for Q and Y – created arbitrarily via some multivariate polynomials with random coefficients – and approximate the MMSE estimator from \mathcal{Y} to \mathcal{Q} . It will be shown that the approximations $\varphi_p(Y)$ converge to Q if the measurements Y give enough information about the underlying probability space. If the measurements are not sufficiently informative it can be seen that only the mean square error is better minimised with increasing approximation order p . In the second example we have Y given by a non-linear measurement function plus some additive noise, for which the analytical computation of the conditional expectation and Bayes' posterior is possible. This allows better comparison and exact computation of the errors made in the MMSE approximation.

Example 9 *In the following examples, the sample space Ω is \mathbb{R}^d with a standard Gaussian product measure, which can be thought of simply as a collection of d independent standard Gaussian random variables. If the number of parameters is denoted n , then the random variable Q is a function from Ω to \mathbb{R}^n , which we create artificially as a vector of n polynomials of total degree p_q with randomly generated coefficients in the d standard Gaussian random variables each. The number of measurements is m and the random variable $Y : \Omega \rightarrow \mathbb{R}^m$ is generated likewise as m multivariate polynomials in d variables up to total degree p_y . The non-linear MMSE was then used to approximate the “unknown” random vector Q by the “measurements” Y .*

Figure 4.1 shows the non-linear MMSE for $d = m = n = 2$ and different values of p_φ , the polynomial degree of the estimator $\hat{\varphi}$. Since $d = m$ the estimator can be expected to converge for large values of p_φ , that is, $\lim_{p_\varphi \rightarrow \infty} \|Q - \hat{\varphi}(Y; p_\varphi)\| = 0$. This can be seen in the figures by noting

that the crosses (x), denoting the approximated values $\hat{Q} = \hat{\varphi}(Y; p_\varphi)$, are increasingly better centered in the circles (o), denoting the true values of Q , in the sequence of increasing p_φ .

In the two leftmost graphs in Figure 4.2, one can see the MMSE estimation with two-dimensional sample and parameter space, like in the previous example, but only one measurement, that is, $m = 1$. As expected, the estimate can be only a one-dimensional subset, which is in the left figure for $p_\varphi = 1$, the linear estimator, a straight line. For the cubic estimator with $p_\varphi = 3$ in the middle figure, the estimate is non-linear, and matches better the shape of the original distribution. However, convergence cannot be achieved as there is just not enough information available in the measurements.

In the rightmost figure the parameters are $m = n = 3$, $d = 5$ and $p_\varphi = 4$. Even though, the number of measurements is the same as the number of parameters to estimate and the polynomial degree is relatively high, there is no apparent convergence, since the dimension of the sample space d is higher than m . In this setting, a measurement y is not sufficient to determine the exact event ω but only some subset of Ω that can have led to it, and thus the determination of the corresponding parameter q has still some remaining uncertainty. So, even for high value of p_φ no convergent approximation can be expected.

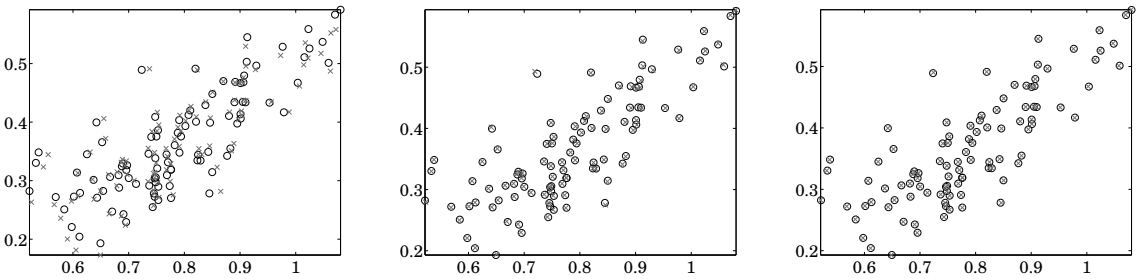


Figure 4.1: MMSE estimation with increasing polynomial degrees $p_\varphi = 1, 2$, and 3 from left to right (for $m = n = d = 2$). Shown are the true parameter values q_i in the Q -plane (marked by o's) and the MMSE estimates $\hat{q}_i = \varphi(y_i)$ (marked by x's).

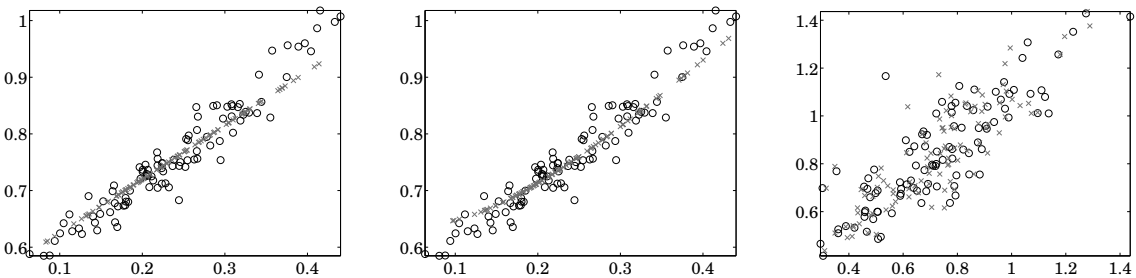


Figure 4.2: MMSE estimation with $m = 1, p_\varphi = 1$ (left), $m = 1, p_\varphi = 3$ (middle), and $m = 3, d = 5, p_\varphi = 4$ (right). True values Q are marked by o, and estimated values $\hat{Q} = \varphi(Y)$ are marked by x.

Example 10 In this example, the measurement function is non-linear, but still conditional probabilities and conditional expectation can be computed analytically. Let Q have a uniform prior $Q \sim U[-1, 1]$. Let E have a uniform distribution $E \sim U[-\delta, \delta]$ with $\delta = 0.5$. Let the system response G be given by $y = M(q) = \sin(q)$, such that $Y = \sin(Q) + E$. Then the MMSE is given by

$$q = \hat{\varphi}(y) = \frac{1}{2} (\arcsin(\min(\sin(1), y + \delta)) + \arcsin(\max(-\sin(1), y - \delta))) \quad (4.23)$$

The MMSE and polynomial approximations of degree 1, 3, and 5 (even degree terms are zero since $\hat{\varphi}$ is an odd function) are shown in Figure 4.3 (left). The discontinuities in the prior and the error probability density functions introduce kinks in the MMSE at $\pm(\sin(1) - \delta)$, where the polynomial approximation is not very good. This can also be seen in Figure 4.3 (right), where the error is displayed, by reduced convergence at those points. This behaviour is mitigated for smooth prior and error distributions with much faster convergence of the polynomial approximations.

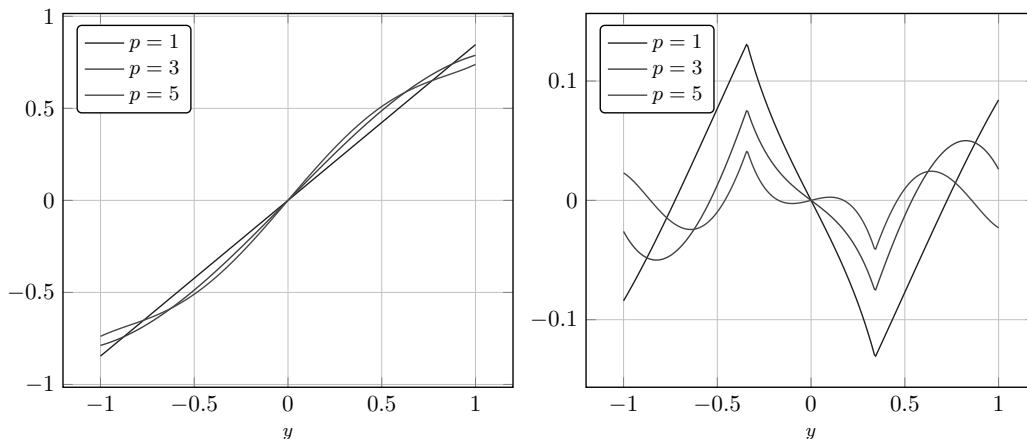


Figure 4.3: Approximation of the conditional expectation for different polynomial degrees (left) and difference to the true conditional expectation (right).

4.3 Parameter identification using the MMSE

The MMSE as derived in the last sections can be used directly for point estimates of the parameters q , that is, the posterior mean $q_m = \hat{\varphi}(y_m)$. However, the mean is often not informative enough, because it does not tell us anything about how certain this value is or how much trust we can put into it. In a Bayesian framework, we thus want to have a distribution which characterises the posterior density.

4.3.1 The MMSE filter

Suppose we have a parameter $q \in \mathcal{Q}$ and a corresponding measurement value $y = M(u; q) \in \mathcal{Y}$. From the section about the MMSE we know that $\hat{\varphi}(y)$ is the best estimate for q in the mean square

sense. In terms of the random variables, we can therefore restate that as: for each ω , the conditional expectation $E(Q|Y)(\omega)$, or equivalently $\hat{\varphi}(Y(\omega))$, is the best estimate for Q given $Y(\omega)$. We can thus decompose Q into two components such that

$$\begin{aligned} Q &= \underbrace{(Q - Q_Y)} + Q_Y \\ &= Q_Y^\top + Q_Y \end{aligned}$$

where Q_Y is the conditional expectation and Q_Y^\top is the residual part of Q . Q_Y and Q_Y^\top are orthogonal, that is, uncorrelated, random variables. This can be easily seen, because Q_Y follows from the minimisation of $\mathbb{E}\|Q - \varphi(Y)\|^2$, and thus $Q_Y^\top = Q - Q_Y$ must be orthogonal to $\varphi(Y)$ for every φ including $Q_Y = \hat{\varphi}(Y)$.³

Two uncorrelated random variables that are known to be (jointly) Gaussian are also independent. So, if Q_Y and Q_Y^\top are Gaussian, then they are independent and thus Q_Y^\top is also independent of Y and hence does not contain any information that can be inferred from the measurements Y . In this case Q_Y can be called the “predictable part” of the parameters Q by Y , and Q_Y^\top the “unpredictable part”, as there is no information in Y that can reduce our uncertainty in Q_Y . If the random variables are not Gaussian, then this is not strictly true, that is, Q_Y^\top does contain information from Y . However, in many cases where the random variables are not too far from Gaussianity this assumption is a good approximation. Quantitative measures or estimates in this case are yet missing but currently under investigation.

The MMSE filter is now based on the following idea: the “predictable part” of Q is the component of the parameter uncertainty that can be removed by knowledge of an outcome of measurement Y . That means, if we have a concrete measurement y_m we can replace the Q_Y in Equation (4.24) below by the concrete prediction $q_m = \hat{\varphi}(y_m)$, the best estimate for the parameters. The new model Q' with reduced uncertainty is given by best estimate q_m plus remaining uncertainty Q_Y^\top , that is,

$$Q' = q_m + Q_Y^\top \tag{4.24}$$

This can be written as an update equation for Q in the form

$$Q' = Q + \hat{\varphi}(y_m) - \hat{\varphi}(Y) \tag{4.25}$$

or using conditional expectations

$$Q' = Q + \mathbb{E}[Q|Y = y_m] - \mathbb{E}[Q|Y], \tag{4.26}$$

which constitutes the so-called *MMSE filter*.

The implementation is straightforward given that the MMSE has been calculated beforehand. Depending on how the random variables Q and Y are represented in the code, the representation of Q' should be chosen accordingly. For example, if Q and Y are given as functions, then Q' could be defined as a new function that forwards its arguments to Q and Y respectively, like

$$Q' = \omega \mapsto \left(Q(\omega) + q_m - \sum_{\gamma \in \mathcal{J}} \varphi_\gamma \Psi_\gamma(Y(\omega)) \right) \tag{4.27}$$

³This can be visualised in the following way: Suppose in 3D space we have a point and a plane and we are looking for the point in the plane that minimises the distance to the original point. Then the residual, that is, the vector from the point to the minimiser is orthogonal to every vector in the plane. In our stochastic setting, the point corresponds to Q , the plane corresponds to the subspace in \mathcal{Q} made up by all $\varphi(Q)$ and the Euclidean distance to the mean squared error. For a rigorous treatment in Hilbert spaces see, for example [123, Classical Projection Theorem].

If Q and Y are given via GPC expansions, then the GPC for Q' can be attained from the above, for example, by projection. Before we show numerical examples for the MMSE filter, we will first show that under certain conditions the filter is equivalent to the well-known *Kalman filter*, and can thus be regarded as a non-linear extension of the latter.

4.3.2 The Kalman filter

The Kalman filter is a numerical procedure for state estimation in dynamical systems [107]. It consists of a prediction step that describes how the distribution of the state estimate develops over one time step and a data assimilation step that incorporates new, but uncertain sensor data into the current state estimate. In the setting of parameter estimation, we are only interested in the data assimilation step, where the current state of the Kalman filter now corresponds to the parameter vector we want to estimate.

The model underlying the assimilation part of the Kalman filter can be summarised as follows: Let the current best estimate for the uncertain parameters be described by the random variable Q having a Gaussian distribution with mean q and variance C_Q , that is, $Q \sim N(q, C_Q)$. The observations shall be given by a measurement model $Y = HQ + E$ where H is the *observation matrix* and E is a mean-free Gaussian noise term with covariance C_E , i.e. $E \sim N(0, C_E)$. Then, after observing y_m , the best new estimate for the mean is given by

$$q' = q + K(y_m - Hq) \quad (4.28)$$

where $K = C_Q H^\top (C_E + H C_Q H^\top)^{-1}$ is the *Kalman gain*, and the variance of this updated estimate is given by

$$C_{Q'} = (I - KH)C_Q(I - KH)^\top + KC_E K^\top, \quad (4.29)$$

that is, then $Q' \sim N(q', C_{Q'})$. For a more extensive treatment we refer the interested reader to [123] and [130]. We show now that under the same conditions, that is, a linear observation operator and Gaussian uncertainties, the MMSE filter leads to exactly the same equations.

Since the Kalman filter is a linear filter, the only functions in the basis will be the constant and linear polynomials that we assign to the basis Ψ_i for $(i = 0 \dots m)$, that is,

$$\Psi_0(Y) = 1, \Psi_1(Y) = Y_1, \dots, \Psi_m(Y) = Y_m. \quad (4.30)$$

Since we assume Q to be vector valued, we have trial functions of the form

$$\varphi_i(Y) = \alpha_i + \beta_{i1}Y_1, \dots, \beta_{im}Y_m \quad (4.31)$$

with $i = 0 \dots n$. Collecting the α_i in a vector $\alpha = (\alpha_1, \dots, \alpha_n)^\top$ and the β_{ij} in a matrix K (the naming will become clear later) such that $(K)_{ij} = \beta_{ij}$, we can write this as

$$\varphi(Y) = \alpha + KY \quad (4.32)$$

In this setting, Equation (4.22) becomes

$$\begin{bmatrix} \mathbb{E}[1] & \mathbb{E}[Y^\top] \\ \mathbb{E}[Y] & \mathbb{E}[YY^\top] \end{bmatrix} \begin{bmatrix} \alpha^\top \\ K^\top \end{bmatrix} = \begin{bmatrix} \mathbb{E}[Q^\top] \\ \mathbb{E}[YQ^\top] \end{bmatrix}. \quad (4.33)$$

Using the expressions

$$\mathbb{E}[YY^\top] = C_Y + C_E + \mu_Y\mu_Y^\top \quad (4.34)$$

$$\mathbb{E}[YQ^\top] = C_{YQ} + \mu_Y\mu_Q^\top, \quad (4.35)$$

where $C_Y = \mathbb{E}[(Y - \mu_Y)(Y - \mu_Y)^\top]$ is the covariance matrix of Y and $C_{YQ} = \mathbb{E}[(Y - \mu_Y)(Q - \mu_Q)^\top]$ is the cross covariance matrix between Y and Q , we get

$$\begin{bmatrix} 1 & \mu_Y^\top \\ \mu_Y & C_Y + C_E + \mu_Y\mu_Y^\top \end{bmatrix} \begin{bmatrix} \alpha^\top \\ K^\top \end{bmatrix} = \begin{bmatrix} \mu_Q^\top \\ C_{YQ} + \mu_Y\mu_Q^\top \end{bmatrix}. \quad (4.36)$$

From the first row of Equation (4.36), we can express α as

$$\alpha = \mu_Q - K\mu_Y \quad (4.37)$$

and inserting this into the second row we attain

$$K = C_{QY}(C_Y + C_E)^{-1}, \quad (4.38)$$

whose expression corresponds to the K Kalman gain. The estimator then reads:

$$\hat{\varphi}(Y) = \mu_Q + K(Y - \mu_Y) \quad (4.39)$$

and the MMSE filter with the orthogonal decomposition becomes

$$Q' = Q - \hat{\varphi}(Y) + \hat{\varphi}(y_m) = Q + K(y_m - Y). \quad (4.40)$$

If we compute the mean and the variance of both sides of the last equation we get the usual update equations for the Kalman filter

$$\mu'_Q = \mu_Q + K(y_m - \mu_Y) \quad (4.41)$$

$$C'_Q = C_Q + KC_YK^\top. \quad (4.42)$$

This means, in the linear case, the MMSE filter reduces to the Kalman filter, and the former can thus be seen as a non-linear generalization of the latter. In Section 4.3.3, a comparison of the performance between the MMSE and the Kalman filter for a simple non-linear example will be shown.

4.3.3 Numerical examples

In the following, we show two numerical examples for the MMSE filter with quadratic non-linearities. The examples were chosen such that the action of the filter can well be studied. For a realistic case study, the reader is referred to Chapter 8.

Example 11 *The one-dimensional case is well-suited for comparison to the Kalman filter. Let the system model be given by the quadratic relation*

$$A(u; q) = u - \alpha(q - q_0)^2 = 0 \quad (4.43)$$

with $q_0 = -3$ and $\alpha = 0.03$ and a measurement operator given by the identity

$$M(u; q) = u. \tag{4.44}$$

Thus, the relation between the parameters and the measurements is given by the response surface

$$G(q) = \alpha(q - q_0)^2. \tag{4.45}$$

As a prior, we assume a normal distribution with standard deviation 2, that is, $Q \sim N(0, 2^2)$. The true value of the parameter is $q_{\text{true}} = 3$, of course taken to be unknown, and the measured value $y_m = G(q_{\text{true}}) = 1.08$, where no measurement error has been added.

As the Bayes posterior is a combination of our prior belief and the information given by the data, it should be somewhere between the maximum of the prior (wider curve) and the black vertical line (the true parameter value) in Figure 4.4. The MMSE posterior for $p_\varphi = 1, 2, 3$, and 4 is shown as the narrower curve in Figure 4.4. It can be observed that in the linear case that corresponds to the Kalman filter, the posterior overshoots and lies even further away from the truth than the

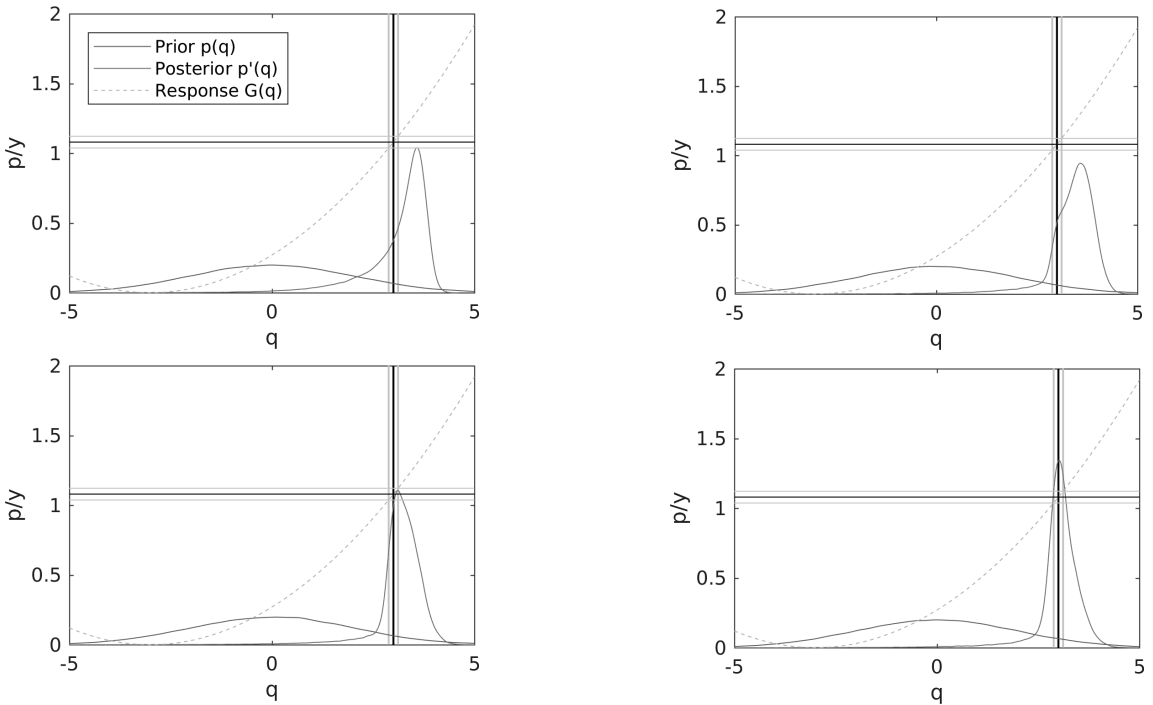


Figure 4.4: MMSE filter with different polynomial orders. The system response $G(q)$ is given by the dashed line, the prior by the wider curve and the MMSE filter posterior for different polynomials orders, that is, $p_\varphi = 1$, $p_\varphi = 2$, $p_\varphi = 3$, and $p_\varphi = 4$ from left to right and top to bottom, by the narrower curve. The measurement y_m is indicated by the horizontal black line, and plus and minus one standard deviation by the parallel grey lines. The corresponding parameter value is indicated by the vertical black line; again, the horizontal lines indicate one standard deviation.

prior would indicate. The reason is that the slope for the linear estimator is determined by sampling of the response surface where the prior is large. Since the response is relatively flat there, the inverse mapping has a steep slope and so the estimator will overestimate the true MMSE estimates. The higher-order estimators, taking non-linear terms into account, produce apparently much better results, especially for $p_\varphi = 3$ and $p_\varphi = 4$.

Example 12 In the second example, we use a two-dimensional response surface and a Gaussian prior, such that depending on their parameters, the posterior can take very different shapes.

The prior on Q is given by normal distribution with mean at the center $(0, 0)^\top$ and variance one, depicted by the coloured plane at $z = 0$ in Figure 4.5. The system model is given by

$$A(u; q) = u - \alpha \|q - q_0\|^2 \quad (4.46)$$

with $q_0 = (1, 3)^\top$ and $\alpha = 1.1$ and the measurement operator again, like in the last example, by the identity $M(u; q) = u$. Thus the relation between the parameters and the response is

$$G(q) = \alpha \|q - q_0\|^2. \quad (4.47)$$

The measurement was assumed to be $y_m = 12$ here. Depending on the location of q_0 , the value of α and the measurement y_m , very different shapes of the posterior distribution can be achieved: everything from circular shapes, over nearly Gaussian-like bumps, to very long, thin and straight shapes are possible. For the parameters described here, the resulting posterior is a slightly bent, “banana”-shaped bump, which can be seen in the left plot of Figure 4.5. The center plot shows sample points generated by a Metropolis-Hastings MCMC method, which follows the Bayes posterior very well.

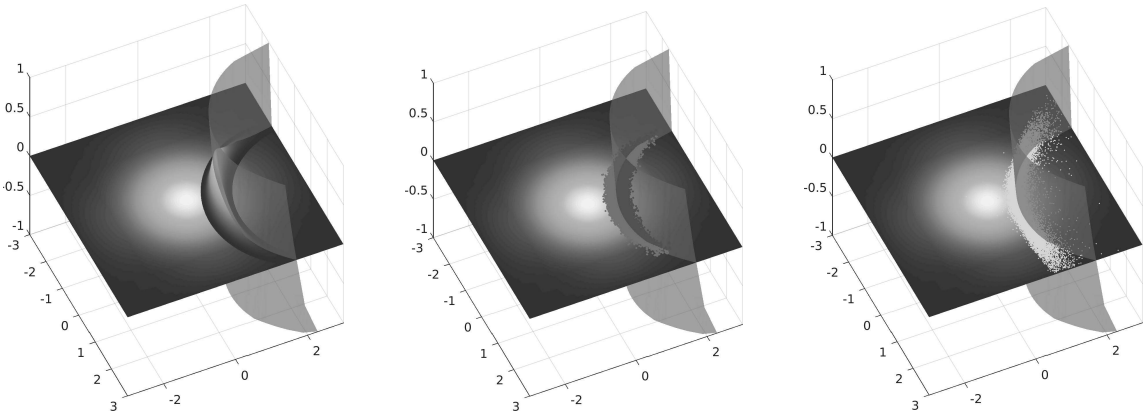


Figure 4.5: Comparison of the MMSE filter with $p_\varphi = 3$ (right) with a true Bayes posterior (left) and an MCMC simulation (center) for a two-dimensional example. The prior density is a standard normal indicated on the $z = 0$ plane. The response surface is given by a paraboloid shown by the light-blue transparent surface, shifted such a way that the measurement coincides with the zero plane. That way the set of parameters that are in agreement with the measurement are given by the intersection of those two surfaces, and the posterior should be close to this set (a circle) and to the maximum of prior density.

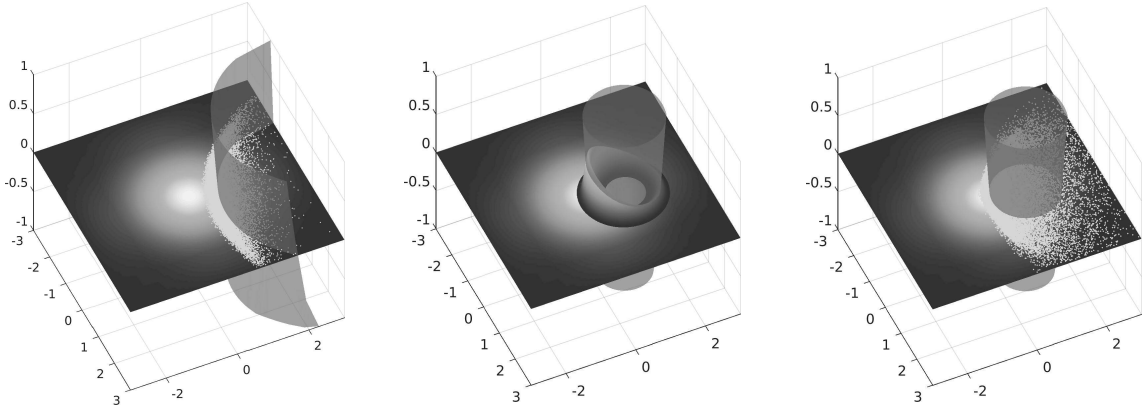


Figure 4.6: Continuation of Fig. 4.5. On the left the MMSE filter with $p_\varphi = 1$. In the center, the true posterior density for a Bayes posterior that is more difficult to approximate and on the right samples from the MMSE filter with $p_\varphi = 3$.

The plot on the right show samples generated from the cubic MMSE filter posterior, which also captures the center region well, however, it deviates at the bent edges of posterior. In Figure 4.6, on the left, samples from the linear MMSE filter are displayed. Here, the part of the posterior close to the center of the prior is also matched quite well; however, there are too many samples in the center of the response paraboloid than there should be.

The example in the center and on the right of Figure 4.6 shows results for the MMSE filter for a set of parameters ($\alpha = 3.3, q_0 = (0.3, 0.9)^\top, y_m = 3$) which gives rise to a more complicated, circular posterior. Here, the filter even with higher orders cannot capture the structure of the posterior density well. This behaviour that is, that the MMSE filter captures the Bayes posterior around the conditional mean very well, but deviates at the tails if the distribution is strongly curved or multi-modal, could be observed in other examples as well.

4.4 Conclusion

The MMSE filter has shown good performance in a variety of problems. Its advantage being that it is a deterministic method with calculable runtime, in contrast to, for example, Monte Carlo methods, with a non-predictable number of iterations for burn-in and convergence. The performance is generally superior to Kalman filters that use only linearisations (e.g. the EKF) as the MMSE filter can also take non-linearities into account. However, the performance can suffer during strongly non-linear mappings. Here, either different basis functions (like e.g. trigonometric functions or rational basis functions) should be used or approaches like the ensemble Kalman filter or MCMC methods could be considered. If only the conditional mean or the conditional variance is of interest, there are now also more efficient methods available (see e.g. [191]), that do not construct the MMSE as a function but directly compute its value for a given measurement.