Technical Report

TR-2013-012

Data Assimilation in Cardiovascular Fluid-Structure Interaction Problems: an introduction

by

Luca Bertagna, M. D'Elia, M. Perego, A. Veneziani

MATHEMATICS AND COMPUTER SCIENCE

EMORY UNIVERSITY

Data Assimilation in Cardiovascular Fluid-Structure Interaction Problems: an introduction

L. Bertagna, M. D'Elia, M. Perego, A. Veneziani

December 9, 2013

Abstract

Numerical methods for incompressible fluid dynamics have recently received a strong impulse from the applications to the cardiovascular system. In particular, fluid-structure interaction methods have been extensively investigated in view of an accurate and possibly fast simulation of blood flow in arteries and veins. This interest has been strongly motivated by the progressive interest in using numerical tools not only for understanding the general physiology and pathology of the vascular system. The opportunity offered by medical images properly preprocessed and elaborated to simulate blood flow in real patients highlighted the potential impact of scientific computing on the clinical practice. Therefore, *in silico* experiments are currently extensively used in bioengineering for completing (and sometimes driving) more traditional in vivo and in vitro investigations. Parallel to the development of numerical models, the need for quantitative analysis for diagnostic purposes has strongly stimulated the design of new methods and instruments for measurements and imaging. Thanks to these developments, a huge amount of data is nowadays available. *Data Assimilation* is the accurate merging of measures (including images) and numerical simulations for a mathematically sound integration of different sources of information. The outcome of this process includes both the patient-specific measures and the general principles underlying the development of mathematical models. In this way, simulations are adapted to the availability of individual data and are therefore supposed to more reliable; measures are correspondingly filtered by the mathematical models assumed to describe the underlying phenomena, resulting in a (hopefully) significant reduction of the noise.

This note provides an introduction to methods for data assimilation, mostly developed in fields like meteorology, applied to computational hemodynamics. We focus mainly on two of them: methods based on stochastic arguments (Kalman filtering) and variational methods. We also address some examples that have been approached with different techniques, in particular the estimation of vascular compliance from displacement measures.

Keywords: Data assimilation, Kalman Filter, Variational parameter identification, Fluid-structure interaction, Biological flows
MSC: 49N45,35Q35,74F10,76Z05,62M20, 49M05, 49M15, 49M29 60G25, 60G35 90C46 93E10,93E11,93E12

1 Preliminaries

Numerical methods for incompressible fluid dynamics have recently received a strong impulse from the applications to the cardiovascular system (see e.g. [24, 29]). In particular, fluid-structure interaction (FSI) methods have been extensively investigated in view of an accurate and possibly fast simulation of blood flow in arteries and veins (see e.g. the recent works by Y. Maday and by J.F. Gerbeau and M. Fernandez, Chapters 8 and 9 of [24] respectively or the Chapter of the present book by C. Grandmont, M. Lukáčová and Š. Nečasová). Beyond the intrinsic mathematical interest, the development of reliable tools for the numerical simulations of cardiovascular problems - and FSI in particular - has an impact on bioengineering and medical research. Thanks to the opportunity offered by improvements in imaging and measurement devices and the subsequent elaboration (see Chapter 4 of [24] authored by L. Antiga, D. Steinman and J. Peiró), nowadays scientific computing is not only a tool for investigating the physiopathology of the cardiovascular system at a general level, but also a way for analyzing in detail the single patient. Mathematical models, properly numerically approximated complete the patient-specific information provided by traditional (yet progressively improved) diagnostic tools. The complete patient-specific picture provided by numerical models may (and most likely will) have a diagnostic and prognostic impact and, more in general, provide a decision-making support in clinical practice. However, this fascinating perspective raises some important challenges. The general problem of quantifying and reducing uncertainty in mathematical models and to certify the quality of numerical simulations - common to any computer aided activity - is even more important when supporting the clinical practice, for its individual and social impact. This is related to the problem of *validating* numerical models. “Validation is the process of determining the extent to which the computer implementation corresponds to the real world. If solution verification has already been demonstrated, then validation asks whether the mathematical model is effective in simulating those aspects of the real world system under study” (from [22]).

Parallel to the development of numerical models, the need for quantitative analysis for diagnostic purposes has strongly stimulated the design of new methods and instruments for measurements and imaging. Thanks to these developments, a huge amount of data is nowadays available to bioengineers and medical doctors. Also, the reliability of these data and their significance in clinical practice needs to undergo a strict analysis and assessment, since they are typically affected by noise and errors.

Data Assimilation (DA) is a process for integrating the knowledge provided

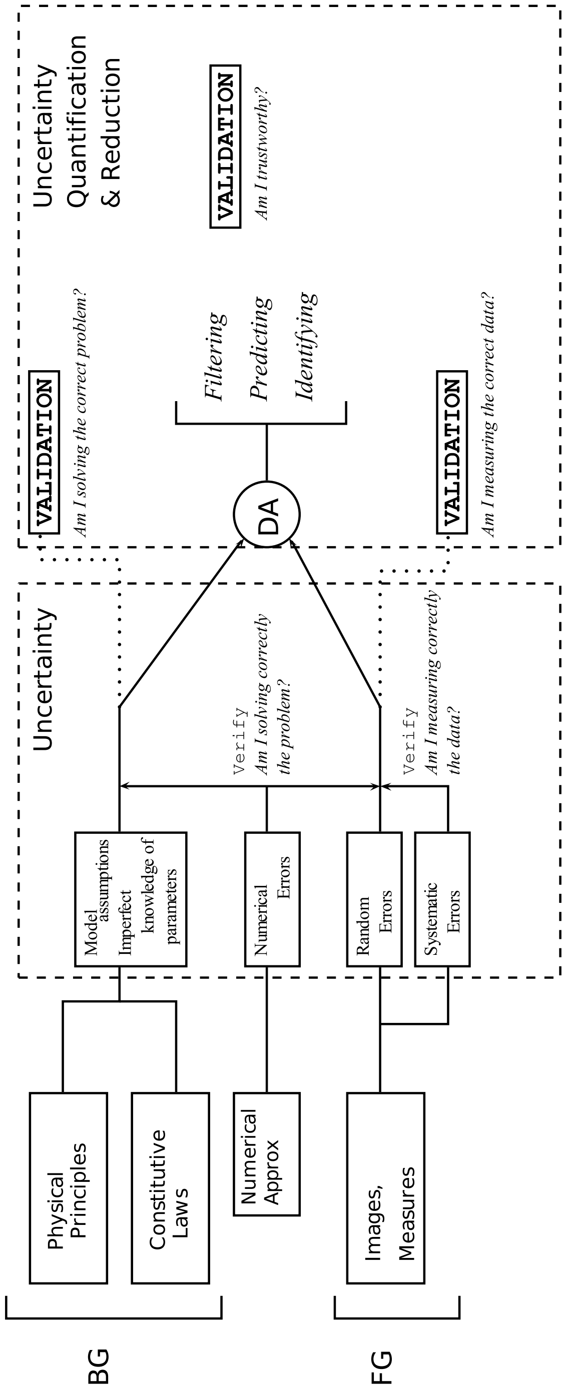


Figure 1: The general framework of Data Assimilation as a process for improving the reliability of quantitative analyses (BG=Background, FG= Foreground).

by numerical models and measurements with the purpose of improving the reliability of quantitative analysis. This approach has been developed since the mid of the 20^{th} Century having as preferential application the weather forecasting. The rationale is that the predictions provided by numerical models, that we may call a *background knowledge*, being partially based on universal physical and constitutive laws, are affected by uncertainties in real world problems. These are the consequence of simplifying assumptions as well as of an incomplete knowledge of parameters usually needed by the constitutive laws forming a mathematical model. For instance, referring to biomedical applications, blood viscosity (that in a Newton constitutive law is supposed to be constant) or compliance of an artery (that in a Hookean material is supposed to be represented by a parameter, the Young modulus) are available as estimated on samples, but when dealing with a specific patient they are in general not known, being impossible or inconvenient to measure. The integration with available measures, that we may call a *foreground knowledge*, since they are specific of the case, is expected to be beneficial to the quantitative analysis, reducing the uncertainty in the mathematical models. On the other hand, background models improve the knowledge extracted from the data, providing a way for filtering noise. In particular, this is important for at least three purposes,

1. *estimate* the state of a dynamical system (e.g. the velocity, the pressure)or its derivated quantities for which noisy data and mathematical models are available,

2. *predict* the state of a dynamical system for which data are available in the past,

3. *identify* one or more parameters involved in the mathematical model, adjusting their values on the basis of available data.

In the global picture - that we have represented in Fig. 1 - DA reduces possibilities of failure in estimating, predicting, and identifying by merging background and foreground in a unique quantitative analysis. The necessity of this process in the traditional development of numerical models in cardiovascular mathematics becomes progressively more urgent with the increment of available data and, more importantly, of patients that may benefit from quantitative analysis.

In this Chapter we want to provide an introduction to some topics brought in by DA in Cardiovascular Mathematics, with a particular emphasis to FSI problems. It is important to stress that, as such, this introduction cannot be complete. First, there are several ways for approaching DA and it is basically impossible to provide an exhaustive global picture of the possible methodologies. We refer to [8] as a more general introduction. Second, DA in cardiovascular modeling is a relatively recent topic and many questions and challenges are still open, so it is hard to draw conclusive statements about the adequacy of a methodology for a specific problem. Our perspective is to provide some examples that have been recently considered in the literature and a self-contained introduction to the methods used there. In particular, we have selected examples tackled with different approaches, providing different perspectives for solving the same problem. This is intended to give not only the idea of the complexity of the problems but also of the variety of approaches, the differences and the complementary nature of the methods. In particular, we will consider two classes of methods,

1. stochastic approaches, when some probabilistic knowledge of the uncertainty affecting the model and the noise affecting the measures are available; in particular we refer to methods related to Kalman filtering and its extension to nonlinear problems; these methods will be addressed in Section 2;

2. deterministic approaches, when no clue on the statistical features of uncertainty is available; in particular, we will see variational methods based on the minimization of the mismatch between the data and numerical results, constrained by the background model; these methods are introduced in Section 3.

The above distinction is not strict. In fact, available statistical information can be included in variational models.

The FSI problem and more in general the problems involved in cardiovascular mathematics - usually represented by a system of partial differential equations - are complex and *per se* challenging. In the framework of DA, the issue of computational costs becomes even more important, as DA typically involves the solution of *inverse problems*. In practice, these problems can be solved by iterative approaches, where the solution of the FSI system (or more in general of the "forward" problem) needs to be performed at each iteration. It is promptly realized that this requires specific techniques to make the computational costs affordable. We address this issue in Section 3, in particular referring to methods for reducing the costs of each iteration by representing the solution on a "smart" low-dimensional basis functions set that strongly reduces the number of degrees of freedom required by a traditional numerical method (like finite element or spectral methods).

Detailed examples are provided in Section 4. In particular, we consider the assimilation of velocity measures with the numerical simulation of the Navier-Stokes equations for improving the estimate of blood velocity on an artery. We address two different deterministic approaches and how they can be reinterpreted or improved by a stochastic Bayesian perspective. Finally, we present in detail the problem of estimating vascular compliance by solving an inverse FSI problem. Again, we present both a stochastic approach based on Kalman filtering and a deterministic constrained minimization approach. In the latter case, we present a technique for reducing the computational costs by representing the solution on a low-dimensional basis obtained with a Proper Orthogonal Decomposition approach.

As we have pointed out, the methodological picture in the filed of DA is pretty articulated, encompassing statistical as well as numerical issues for inverse problems [17, 71]. Here, we mention some references for the reader interested in having more details on the topics covered only partially in this introduction. The importance of uncertainty quantification in any field of scientific computing has been recently underlined in [22]. An excellent introduction to statistical methods for computational inverse problems is given in the books [47, 13, 27, 68]. A recent collection of contributions in the numerical solution of inverse problems and computational costs reduction is [7].

A classical book on methods for solving constrained minimization for flow problems is [37]. Fundamental contributions can be found in [32, 33], recently collected in [34]. Parameter estimation for partial differential equations has been extensively covered in [2] (see also the recent [3]).

An introductory example To illustrate some basic concepts in DA, we refer to an oversimplified example. Let us assume to have a pipe where an incompressible fluid flows. We also suppose that N measures of velocity are available in N points P_i $(i = 1, \dots, N)$. Our goal is to compute the shear stress at the wall of the pipe, which is defined as[1]

$$\mu \left(\nabla \mathbf{u} + \nabla \mathbf{u}^T \right) \mathbf{n} - \mu \left(\mathbf{n} \left(\nabla \mathbf{u} + \nabla \mathbf{u}^T \right) \mathbf{n} \right) \mathbf{n} \tag{1}$$

where μ is the blood viscosity, $\mathbf{u}$ is the velocity, $\mathbf{n}$ the unit vector normal to the surface. In particular, if we are interested in the estimate of the stress at the times for which velocity measures are available, this is an *estimate* or *filtering problem*. If we want to quantify the wall shear stress for time instants *after* time of measures, we have a *prediction problem*.

There are different approaches for this.

1. *Data fitting* approach: a functional form for the velocity is decided (for instance polynomial) and fitted with the data. Successively the wall shear stress is obtained by applying (1) with the fitted velocity. The quality of the computation depends on the number and location of the measures, and the amount of noise. Fitting can be performed with either interpolation or least square approximation depending on how trustworthy the measures are considered. In this approach, we do not assume any background knowledge of fluid mechanics.

2. *Model based computation*: We may assume that blood flow is an incompressible Newtonian fluid. Under these assumptions, for a cylindrical pipe described by the coordinates z, r, θ, we can derive the Poiseuille solution (outlined in color in Fig. 2),

$$u_z = \frac{G_P}{4\mu}(R^2 - r^2), \quad u_r = u_\theta = 0, \quad p = G_P z + C \tag{2}$$

   where G_P is the pressure gradient, R is the radius of the pipe and C is an arbitrary constant. Should G_P and μ be available, we compute the wall shear stress; such parameters are needed to get the final estimate of

[1] We remind that the wall shear stress is a quantity of great relevance in biomedical applications for its correlation with pathologies such as atherosclerosis - see e.g. [14].

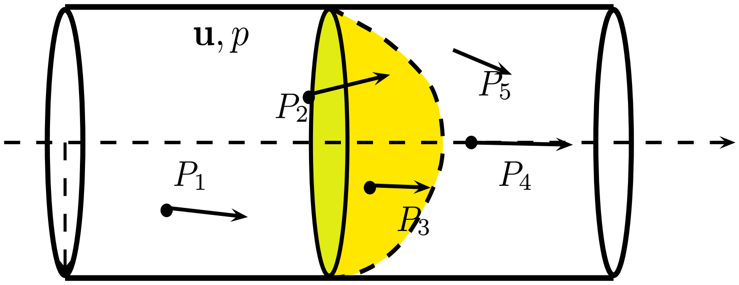


Figure 2: Pipe where an incompressible fluid flows and velocity measures are available in the points P_i: how is it possible to reliably compute the wall shear stress at the wall?

the stress, while measures are not needed. The quality of the estimate depends on the reliability of the model assumption. In general, the analytical solution may be replaced by a numerical solution. In such case, the accuracy of the estimated stress will also depend on the numerical approximation.

3. *Data assimilation* procedure: suppose that the assumptions behind the Poiseuille solution are acceptable but our knowledge is incomplete, for instance the viscosity μ and the pressure gradient G_P are unknown; in this case, we may take advantage of the velocity measures to fill the gap and eventually to compute the wall shear stress by formulating the following problem. Find μ and G_P such that $\mathbf{u}$ minimizes the mismatch

$$\mathcal{J} = \sum_{i=1}^{N} \left(\mathbf{u}_m(P_i) - \mathbf{u}_p(P_i)\right)^2$$

where $\mathbf{u}_m(P_i)$ is the measured velocity and $\mathbf{u}_p$ is the Poiseuille solution (2). In this way, we are fitting the physical parameters μ and G_P so that the background model is matching the foreground knowledge. Once μ and G_P are computed, the wall shear stress (both as an estimate or as a prediction) is quantified. Contextually, the noise affecting the data is filtered by our background knowledge of fluid mechanics in the physically-driven least squares procedure. Notice that when quantifying the viscosity we are solving an *identification problem*.

In the more realistic case that the Poiseuille solution cannot be applied, we replace $\mathbf{u}_p$ with the (numerical) solution of the Navier-Stokes equations. In this case, the minimization procedure requiring to find the stationary points of $\mathcal{J}$ regarded as a function of μ and G_P is clearly non-trivial (as we will see in the next Sections).

This simple example (that will be developed in Section 4), shows the relevance of DA in biomedical applications, especially related to the clinical practice. As a matter of fact, patient-specific knowledge of parameters that form a mathematical/numerical model is always incomplete. As for the boundary conditions, this has led to extensive investigation of the so-called "defective boundary data problems" (see for instance [25, 26]). Concerning parameter identification, we mention *elastography* as a method for detecting the rigidity of soft tissues by solving inverse elasticity problems [5, 4]. In this case, parameter identification is not only functional to the computation of a specific variable of interest, but it is by itself an important procedure for diagnostic purposes (e.g. breast cancer).

In the previous example, the DA procedure is performed without any real assumptions on the quantity we want to estimate and on the noise affecting the measures. However, in many cases some *a priori* knowledge is available and may be used to drive the assimilation process and eventually reduce the uncertainty affecting the final results. For instance, we speculate that fluid viscosity is a Gaussian variable whose average and variance are known. Similarly, we may assume that the noise features a probabilistic density function whose moments are available. Availability of trustworthy probabilistic information on quantities of interest and noise may be a discriminant for the choice of the DA methods. In the next Section 2 we address probabilistic approaches, while Section 3 is devoted to deterministic methods. It is important to stress that the two classes of approaches are somehow contiguous. As we will see in Section 4, solution obtained with one approach can be reinterpreted in the other framework, when the reliability of *a priori knowledge* tends to 0.

2 Probabilistic Approach

In this Section, we consider the estimation/prediction/identification of quantities when we assume stochastic *a priori* information to be available. We may take therefore advantage of this knowledge to integrate models and measures. The latter are in turn considered to be the realization of a stochastic process, whose features are known.

The ingredients of the problem (see Fig. 3) are: (i) a variable $\mathbf{w}$ - for generality we assume it is a n-dimensional random vector, whose probability density function (see below) is known - and (ii) a set of observations $\mathbf{z}$ - we assume to have p observations organized in a $n \times p$ matrix Z, regarded as p realizations drawn from a known probability density function. Our goal is to find an estimate $\widehat{\mathbf{w}}$ of $\mathbf{w}$ based on both the *a priori* and the *a posteriori* knowledge we have. To introduce fundamental concepts, we start considering $\mathbf{w}$ as an instantaneous (or time independent) variable. Assume for instance that the distribution of the variable of interest is a Gaussian *probability density function* (p.d.f.) then the solution of the estimate problem is given by the average and the variance (and generally the statistical moments) of this distribution[2]. We will see several ways for obtaining these quantities, namely the *minimum variance*

[2] Precise definitions of average and variance of a Gaussian variable will be given later on.

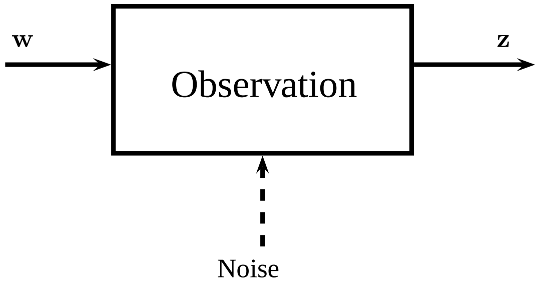


Figure 3: In/Out system: $\mathbf{z}$ is the quantity measured, $\mathbf{w}$ is the quantity to be estimated.

(MV), the *maximum a posteriori probability* (MAP), the *maximum likelihood* (ML).

Then, we consider the case when the variables of interest are part of a dynamical system. As a matter of fact, in the applications we are interested in there is a dynamics and we have a mathematical model describing how a variable of interest, that we call $\mathbf{u}$ (the *state* of the system), evolves according to a sequential parameter that will be in general the time. This may be either the fluid or the fluid coupled with the arterial wall, etc. In general, this is the knowledge given by mathematical modeling. In most of the cases, this is a system of partial differential equations and a numerical discretization procedure is necessary for its quantitative solution. The numerical model (i.e. the discretized mathematical model) reads

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k+1)}(\text{Past}, \text{Input}, \text{Noise}).$$

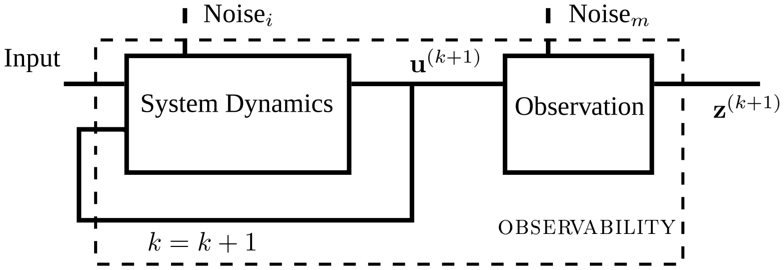


Figure 4: Possible approaches for the estimate with a system dynamics. Here we rely upon the observability of the system and the knowledge of the statistical features of the stochastic process.

Assume that we measure $\mathbf{z}^{(k)}$ and we want to estimate $\mathbf{u}^{(k)}$ (*filtering* problem) or $\mathbf{u}^{(k+p)}$ ($p-$step prediction problem). If the variable $\mathbf{u}$ that we want to estimate is a parameter of the original model (as it was the viscosity on the previous example), this is an *identification* problem. In the simplest case, assume that the system dynamics is linear and that noise affects both the input of the system and the measures. The system dynamics is regarded as the mechanism converting the p.d.f. of the input to the p.d.f. of the output. From the latter, we want to infer the state variable. If we apply the criterion of finding the estimate by minimizing the variance of the estimate, we will get a sequential estimate/prediction procedure called *Kalman filter*. Although the method strongly relies on the linearity of the system, that is not changing the nature of the p.d.f. , we will see that the method can be properly extended to non-linear cases. The effectiveness of the procedure relies upon the "observability" of the system, i.e. on how the output information is actually representative of the state of the system.

As pointed out, our goal here is to give a short introduction to probabilistic estimation theory. For a more complete and extensive presentation we refer to [68], Chapters 4,5,6.

2.1 Basic notation and concepts

We summarize some fundamental concepts and notation that are useful in the remainder of this Section. For a complete and rigorous introduction and explanation of these concepts, we refer e.g. to [55].

Random variables For a random variable w, i.e. a variable whose value depends on a random experiment ω, we introduce the *distribution function*

$$F_W(w) \equiv P(\omega : \mathrm{w}(\omega) \leq w)$$

where the notation on the right hand side represents the probability that the realization of w associated to ω is $\leq w$. Elementary properties of probability imply that $\lim_{w \to -\infty} F_W = 0$ and $\lim_{w \to +\infty} F_W = 1$ and that the function is non-decreasing. The corresponding p.d.f. is defined as

$$p_W(w) \equiv \frac{dF_W}{dw}.$$

For the properties of distribution, $p_W(w) \geq 0$ and $\int\limits_{\mathbb{R}} p_W dw = 1$.

The Gaussian p.d.f. for instance reads

$$g_W(w) = \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{(w-\mu)^2}{2\lambda^2}\right). \tag{3}$$

A p.d.f. can be characterized by its *moments*. In particular, we define the *expectation* operator $\mathcal{E}(\cdot)$ as

$$\mathcal{E}(\mathrm{w}) \equiv \int\limits_{\mathbb{R}} w\, p_W(w) dw,$$

that associates the random variable to a number called *mean.* Similarly, we may consider the moments and the central moments of order m defined respectively as

$$\mathcal{E}\left(\mathrm{w}^m\right) \equiv \int\limits_{\mathbb{R}} w^m p_W(w) dw, \qquad \mathcal{E}\left(\mathrm{w}^m\right) \equiv \int\limits_{\mathbb{R}} \left(w - \mathcal{E}\left(\mathrm{w}\right)\right)^m p_W(w) dw.$$

The central moment of order 2 is called *variance.* For the Gaussian p.d.f. g_W, the mean is μ and variance is λ^2.

Jointly distributed random variables We may consider the case of multiple random variables depending on ω according to a joint distribution

$$F_{W_1 W_2 \ldots W_n} \equiv P(\omega : \mathrm{w}_1 \leq w_1, \ldots, \mathrm{w}_n \leq w_n).$$

In this case, the joint p.d.f. reads

$$p_{W_1 W_2 \ldots W_n} \equiv \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F_{W_1 W_2 \ldots W_n}.$$

First order moments read

$$\mathcal{E}\left(w_j\right) = \int\limits_{\mathbb{R}^n} w_j \, p_{W_1 W_2 \ldots W_n} dw_1 dw_2 \ldots dw_n.$$

Second order central moments form the symmetric *covariance* matrix

$$\begin{aligned} &\lambda_{jk} \equiv \left[\mathcal{E}\left((w_j - \mathcal{E}\left(w_j\right))(w_k - \mathcal{E}\left(w_k\right))\right)\right] = \\ &\int\limits_{\mathbb{R}^n} (w_j - \mathcal{E}\left(w_j\right))(w_k - \mathcal{E}\left(w_k\right)) p_{W_1 W_2 \ldots W_n} dw_1 dw_2 \ldots dw_n \end{aligned}$$

where clearly $\lambda_{jj} = \lambda_j^2$, the variance of w_j. The *correlation coefficient* between w_j and w_k is defined as

$$\rho_{jk} \equiv \frac{\lambda_{jk}}{\lambda_j \lambda_k}. \tag{4}$$

For instance, two jointly distributed Gaussian variables have the distribution

$$p_{W_1 W_2}(w_1, w_2) = \frac{1}{2\pi\sqrt{|\Lambda|}} \exp\left(-\frac{1}{2}\boldsymbol{\delta}^T \Lambda^{-1} \boldsymbol{\delta}\right),$$

where $\boldsymbol{\delta} = \begin{bmatrix} w_1 - \mu_1 \\ w_2 - \mu_2 \end{bmatrix}$, $\Lambda = \begin{bmatrix} \lambda_1^2 & \lambda_{12} \\ \lambda_{12} & \lambda_2^2 \end{bmatrix}$ is the covariance matrix, $|\Lambda|$ stands for its determinant and μ_1 and μ_2 are the means of the two variables.

In the sequel, this distribution is denoted by $\mathcal{G}(\boldsymbol{\mu}, \Lambda)$. In particular a distribution with $\boldsymbol{\mu} = \mathbf{0}$ and Λ diagonal (that means that the components of the vector are not correlated) is considered as a model for random disturbances or *white noise*[3].

The *marginal density function* of one of the random variables w_j in the pool may be obtained by the joint one after integration over the range of the other variables,

$$p_{W_j} = \int\limits_{\mathbb{R}^{n-1}} p_{W_1 W_2 \dots W_n} dw_1 \dots dw_{j-1} dw_{j+1} \dots dw_n$$

Conditional probability The *conditional* p.d.f. of random vector $\mathbf{w}$ given the occurrence of the random vector $\mathbf{y}$ is defined as

$$p_{W|Y}(\mathbf{w}|\mathbf{y}) \equiv \frac{p_{W,Y}(\mathbf{w},\mathbf{y})}{p_Y(\mathbf{y})}.$$

Similarly, we have the definition

$$p_{Y|W}(\mathbf{y}|\mathbf{w}) \equiv \frac{p_{W,Y}(\mathbf{w},\mathbf{y})}{p_W(\mathbf{w})},$$

from which we obtain the *Bayes law*

$$p_{W|Y}(\mathbf{w}|\mathbf{y}) = \frac{p_{Y|W}(\mathbf{y}|\mathbf{w}) p_W(\mathbf{w})}{p_Y(\mathbf{y})}. \tag{5}$$

The conditional expectation is defined consequently as

$$\mathcal{E}\left(\mathbf{w}|\mathbf{y}\right) = \int\limits_{\mathbb{R}^n} \mathbf{w} p_{W|Y}(\mathbf{w}|\mathbf{y}) d\mathbf{w}.$$

From the previous relations, it follows that

$$\mathcal{E}\left(\mathbf{w}\right) = \int\limits_{\mathbb{R}^n} \mathbf{w} p_W(\mathbf{w}) d\mathbf{w} = \int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} \mathbf{w} p_{W,Y}(\mathbf{w},\mathbf{y}) d\mathbf{y} d\mathbf{w}$$
$$\int\limits_{\mathbb{R}^n} \left(\int\limits_{\mathbb{R}^n} \mathbf{w} p_{W|Y} d\mathbf{w} \right) p_Y d\mathbf{y} = \mathcal{E}\left(\mathcal{E}\left(\mathbf{w}|\mathbf{y}\right)\right).$$

2.2 Minimum Variance and other In-Out estimators

Let us consider a first example of estimator $\widehat{\mathbf{w}}$ of the random vector $\mathbf{w}$ upon data $\mathbf{z}$ in the "steady" case - Fig. 3. Let $\mathbf{e} \equiv \widehat{\mathbf{w}} - \mathbf{w}$ be the estimate error, and define

$$J(\mathbf{e}) = \mathbf{e}^T \mathrm{E} \mathbf{e}, \tag{6}$$

[3] The choice of Gaussian distribution for white noise is reasonable, but arbitrary. We could have considered other distributions for zero-mean, uncorrelated components.

where E is a symmetric positive definite (s.p.d.) matrix. We assume J to be the measure of the estimate error or “risk”. Here, $\widehat{\mathbf{w}}$ depends on $\mathbf{z}$ and $\mathbf{w}$, and it is regarded as a stochastic process. With our definition of risk we may introduce the functional $\mathcal{J}\left(\widehat{\mathbf{w}}\right)\equiv \mathcal{E}\left(J(\mathbf{e})\right) = \int\limits_{\mathbb{R}^n} J(\mathbf{e})\, p_W d\mathbf{w}$ and in order to minimize the risk we refer to the estimator

$$\widehat{\mathbf{w}} = \arg\min \mathcal{J}\left(\widehat{\mathbf{w}}\right) \equiv \int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} J(\mathbf{e})\, p_{W,Z}(\mathbf{w},\mathbf{z}) d\mathbf{z} d\mathbf{w}.$$

By exploiting the properties of p.d.f. recalled above, we rewrite the risk to minimize as

$$\mathcal{J}\left(\widehat{\mathbf{w}}\right) = \int\limits_{\mathbb{R}^n} \left(\int\limits_{\mathbb{R}^n} \mathbf{e}^T \mathrm{E} \mathbf{e}\, p_{W|Z} d\mathbf{w} \right) p_Z d\mathbf{z} = \int\limits_{\mathbb{R}^n} \mathcal{J}\left(\widehat{\mathbf{w}}|\mathbf{z}\right) p_Z d\mathbf{z},$$

for

$$\mathcal{J}\left(\widehat{\mathbf{w}}|\mathbf{z}\right) \equiv \int\limits_{\mathbb{R}^n} (\mathbf{w}-\widehat{\mathbf{w}})^T \mathrm{E} (\mathbf{w}-\widehat{\mathbf{w}}) p_{W|Z} d\mathbf{w}.$$

Since the outer integral in the definition of the cost $\mathcal{J}\left(\widehat{\mathbf{w}}\right)$ does not involve $\widehat{\mathbf{w}}$ and $p_Z \geq 0$, we minimize the risk by minimizing $\mathcal{J}\left(\widehat{\mathbf{w}}|\mathbf{z}\right)$.

Recall that for a generic vector $\mathbf{x}$ and a symmetric matrix A of proper size [60], we have $\dfrac{\partial \mathbf{x}^T \mathrm{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathrm{A}\mathbf{x}$. Then the minimization of $\mathcal{J}\left(\widehat{\mathbf{w}}|\mathbf{z}\right)$ leads to

$$0 = \frac{\partial \mathcal{J}\left(\widehat{\mathbf{w}}|\mathbf{z}\right)}{\partial \widehat{\mathbf{w}}} = -2\mathrm{E}\int\limits_{\mathbb{R}^n} (\mathbf{w}-\widehat{\mathbf{w}}) p_{W|Z} d\mathbf{w}.$$

Independently of E, we have the equation

$$\widehat{\mathbf{w}} \int\limits_{\mathbb{R}^n} p_{W|Z} d\mathbf{w} = \int\limits_{\mathbb{R}^n} \mathbf{w}\, p_{W|Z} d\mathbf{w} = \mathcal{E}\left(\mathbf{w}|\mathbf{z}\right).$$

Since $\int\limits_{\mathbb{R}^n} p_{W|Z}(\mathbf{w}|\mathbf{z}) d\mathbf{w} = 1$, we have

$$\widehat{\mathbf{w}} = \mathcal{E}\left(\mathbf{w}|\mathbf{z}\right). \tag{7}$$

This is called *minimum variance estimator*, hereafter denoted by $\widehat{\mathbf{w}}_{MV}$. An important property of this estimator is that it is *unbiased*, i.e. $\mathcal{E}\left(\mathbf{e}\right) = \mathcal{E}\left(\widehat{\mathbf{w}}_{MV} - \mathbf{w}\right) = 0$. In fact, we have

$$\mathcal{E}\left(\widehat{\mathbf{w}}_{MV}\right) = \int\limits_{\mathbb{R}^n} \widehat{\mathbf{w}}_{MV}\, p_Z d\mathbf{z} = \int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} \mathbf{w} p_{W|Z}\, d\mathbf{w}\, p_{\mathbf{z}} d\mathbf{z} =$$

$$\int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} \mathbf{w}\, p(\mathbf{w},\mathbf{z})\, d\mathbf{w} d\mathbf{z} = \mathcal{E}\left(\mathbf{w}\right).$$

It is also possible to verify that $\dfrac{\partial^2 \mathcal{J}\left(\widehat{\mathbf{w}}_{MV}\right)}{\partial \widehat{\mathbf{w}}^2} = 2\mathrm{E} > 0$, so $\widehat{\mathbf{w}}_{MV}$ is indeed a minimum [68].

2.2.1 Maximum *a posteriori* estimator

Other estimators may be computed with a similar approach but referring to a different risk $\mathcal{J}\left(\widehat{\mathbf{w}}\right)$,

$$\mathcal{J}\left(\widehat{\mathbf{w}}\right) = \mathcal{E}\left(J(\mathbf{e})\right) = \int\limits_{\mathbb{R}^n} J(\mathbf{e}) p_W d\mathbf{w} = \int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} J(\mathbf{e}) p_{W,Z} d\mathbf{w} d\mathbf{z}$$

for different choices for $J(\cdot)$. For instance, another possible choice is the "uniform" (thresholding) cost:

$$J(\mathbf{e}) = \begin{cases} 0 & \text{for} \quad \|\mathbf{e}\|_\infty < \varepsilon \\ \dfrac{1}{2\varepsilon} & \text{for} \quad \|\mathbf{e}\|_\infty \geq \varepsilon \end{cases} \cdot ,$$

where $\|\cdot\|_\infty$ is the maximum norm. With this cost function, we obtain

$$\mathcal{J}\left(\widehat{\mathbf{w}}\right) = \int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} J(\mathbf{e})\, p_{W|Z}\, d\mathbf{w} p_Z d\mathbf{z} = \int\limits_{\mathbb{R}^n} \frac{1}{2\varepsilon} \left(\int\limits_{\mathbb{R}^n \setminus \{\widehat{\mathbf{w}} + [-\varepsilon,\varepsilon]^n\}} p_{W|Z} d\mathbf{w} \right) p_Z d\mathbf{z}.$$

By definition

$$\int\limits_{\mathbb{R}^n \setminus \{\widehat{\mathbf{w}} + [-\varepsilon,\varepsilon]^n\}} p_{W|Z} d\mathbf{w} = 1 - \int\limits_{\widehat{\mathbf{w}} + [-\varepsilon,\varepsilon]^n} p_{W|Z} d\mathbf{w}$$

so that

$$\mathcal{J}\left(\widehat{\mathbf{w}}\right) = \frac{1}{2\varepsilon} \int\limits_{\mathbb{R}^n} p_Z d\mathbf{z} - \frac{1}{2\varepsilon} \int\limits_{\mathbb{R}^n} \int\limits_{\widehat{\mathbf{w}} + [-\varepsilon,\varepsilon]^n} p_{W|Z} d\mathbf{w} p_Z d\mathbf{z}.$$

The first term on the right hand side is constant and does not affect the minimization process. Let us focus on the second term. The mean value theorem states that there exists a $\xi \in x + [-\varepsilon, \varepsilon]$ such that

$$\frac{1}{2\varepsilon} \int\limits_{x-\varepsilon}^{x+\varepsilon} f(\chi) d\chi = \frac{1}{2\varepsilon} 2\varepsilon f(\xi) = f(\xi).$$

For $\varepsilon \to 0$ we have $\dfrac{1}{2\varepsilon} \int\limits_{x-\varepsilon}^{x+\varepsilon} f(\chi) d\chi = f(x)$. In particular, in our case, we obtain

$$\lim_{\varepsilon\to 0}\left(-\frac{1}{2\varepsilon}\int\limits_{\widehat{\mathbf{w}}+[-\varepsilon,\varepsilon]^n} p_{W|Z}d\mathbf{w}\right) = -p_{W|Z}(\widehat{\mathbf{w}}|\mathbf{z}).$$

In other terms, the cost function to quantify the risk selected here leads to the maximization of the *a posteriori* p.d.f. $p_{W|Z}$. This estimator, such that $\dfrac{\partial p_{W|Z}}{\partial \mathbf{w}} = 0$, is denoted $\widehat{\mathbf{v}}_{MAP}$ and it is not necessarily unbiased.

Example

In this example, we assume that the scalar variables w and z are statistically related by having a joint Gaussian distribution $\mathcal{G}([0,0],\Lambda)$. The two variables features a Gaussian marginal p.d.f. with mean and variance $0,\lambda_w^2$ and $0,\lambda_z^2$ respectively.

As for the conditional p.d.f. we have

$$p_{W|Z} = \frac{p_{WZ}(w,z)}{p_Z} = \frac{\sqrt{\lambda_z^2}}{\sqrt{2\pi|\Lambda|}}\exp\left(-\frac{1}{2}[w \quad z]^T\Lambda^{-1}[w \quad z] + \frac{z^2}{2\lambda_z^2}\right).$$

Define $\lambda^2 \equiv \lambda_w^2 - \dfrac{\lambda_{wz}^2}{\lambda_z^2}$. Then, by direct inspection, one verifies that $(\mathbf{w}|\mathbf{z})$ is a random vector with Gaussian distribution $\mathcal{G}(\dfrac{\lambda_{wz}}{\lambda_z^2}z, \lambda^2)$.

Because for Gaussian distributions, the value where the maximum is achieved corresponds to the mean, we have that

$$\hat{w}_{MV} = \hat{w}_{MAP} = \frac{\lambda_{wz}}{\lambda_z^2}z.$$

Using the definition of correlation coefficient given in (4), we have

$$\hat{w}_{MV} = \hat{w}_{MAP} = \lambda_w^2\rho_{wz}z.$$

From here, we can check the consistency of our estimate with intuitive situations: if w is not correlated to z, the estimate is 0, which is the mean value of the marginal p.d.f. of w. In this case, the knowledge of z does not bring any advantage and the best estimate remains the *a priori* expected value of w.

2.2.2 Maximum Likelihood estimate

Another reasonable approach is to compute the estimator $\widehat{\mathbf{w}}$ as the value that maximizes the probability of measuring $\mathbf{z}$. In other terms, we get $\widehat{\mathbf{w}}_{ML} = \arg\max p_{Z|W}(\mathbf{z}|\mathbf{w})$ or

$$\widehat{\mathbf{w}}_{ML} : \frac{\partial p_{Z|W}}{\partial \mathbf{w}}|_{\widehat{\mathbf{w}}_{ML}} = 0.$$

The p.d.f. $p_{Z|W}$ is a measure of the *likelihood* that $\mathbf{z}$ is measured, so this estimate is called *maximum likelihood*.

It is interesting to establish a relation between this estimator and the previous ones. We do this for the case of Gaussian variables, even though the same conclusion holds in the general case.

We know that estimator $\widehat{\mathbf{w}}_{MAP}$ is such that $\dfrac{\partial p_{W|Z}}{\partial \mathbf{w}}|_{\widehat{\mathbf{w}}_{MAP}} = 0$.

Then, thanks to the Bayes Theorem, we have

$$p_{W|Z} = \frac{p_{WZ}}{p_Z} = \frac{p_{Z|W} p_W}{p_Z}.$$

Maximizing $p_{W|Z}$ is equivalent to the maximization of its logarithm, so we have

$$\frac{\partial p_{W|Z}}{\partial \mathbf{w}} = 0 \Rightarrow \frac{\partial \ln p_{W|Z}}{\partial \mathbf{w}} = \frac{\partial \ln \frac{p_{Z|W} p_W}{p_Z}}{\partial \mathbf{w}} = 0 \Rightarrow$$

$$\frac{\partial \ln p_{Z|W}}{\partial \mathbf{w}} + \frac{\partial \ln p_W}{\partial \mathbf{w}} - \frac{\partial \ln p_Z}{\partial \mathbf{w}} = \frac{\partial \ln p_{Z|W}}{\partial \mathbf{w}} + \frac{\partial \ln p_W}{\partial \mathbf{w}} = 0$$

since p_Z is independent of $\mathbf{w}$.

Now, for a Gaussian variable, such that

$$p_W = \frac{1}{\sqrt{(2\pi)^n |\Lambda|}} \exp\left(-\frac{1}{2} (\mathbf{w} - \mathcal{E}(\mathbf{w}))^T \Lambda^{-1} (\mathbf{w} - \mathcal{E}(\mathbf{w})) \right),$$

we have

$$\frac{\partial \ln p_W}{\partial \mathbf{w}} = -\frac{1}{2} \frac{\partial (\mathbf{w} - \mathcal{E}(\mathbf{w}))^T \Lambda^{-1} (\mathbf{w} - \mathcal{E}(\mathbf{w}))}{\partial \mathbf{w}} = -\Lambda^{-1}(\mathbf{w} - \mathcal{E}(\mathbf{w})).$$

When the variance of a random variable is large, this means that our *a priori* knowledge is not trustworthy. The limit case of $\Lambda^{-1} \to \mathbf{0}$ corresponds to a total lack of *a priori* information on $\mathbf{w}$. Notice that in this condition

$$\frac{\partial \ln p_{W|Z}}{\partial \mathbf{w}} = \frac{\partial \ln p_{Z|W}}{\partial \mathbf{w}} + \frac{\partial \ln p_W}{\partial \mathbf{w}} \overset{(\Lambda^{-1} = \mathbf{0})}{=} \frac{\partial \ln p_{Z|W}}{\partial \mathbf{w}}$$

so that the maximization of likelihood leads to the MAP estimator. We conclude therefore that $\widehat{\mathbf{v}}_{ML}$ can be considered the estimator in the "limit" case, when we do not have an *a priori* distribution for $\mathbf{w}$, i.e. when we assume that the variance of $\mathbf{w}$ is approaching ∞.

Example

Let us consider two scalar variables w and z with

$$z = Hw + \nu \tag{8}$$

where $w \sim \mathcal{G}(0, \lambda_w^2)$, and the noise $\nu \sim \mathcal{G}(0, r^2)$ is assumed to be uncorrelated with w. Then it is possible to verify that $z \sim \mathcal{G}(0, H^2\lambda_w^2 + r^2)$ and that w and z have a joint Gaussian distribution with $\lambda_{wz} = H\lambda_w^2$. In fact

$$\begin{aligned}
&\mathcal{E}(z) = H\mathcal{E}(w) + \mathcal{E}(\nu) = 0 \\
&\lambda_z^2 = \mathcal{E}\left(z^2\right) = \mathcal{E}\left(H^2 w^2 + 2Hw\nu + \nu^2\right) = H^2\lambda_w^2 + 0 + r^2 \\
&\lambda_{wz} = \mathcal{E}(wz) = \mathcal{E}\left(Hw^2 + \nu w\right) = H\lambda_w^2.
\end{aligned}$$

Using the results of the previous example, in this case we can compute

$$\widehat{\mathbf{w}}_{MV} = \widehat{\mathbf{w}}_{MAP} = \frac{H\lambda_w^2}{H^2\lambda_w^2 + r^2} z = \frac{z}{H}\frac{H^2\lambda_w^2}{H^2\lambda_w^2 + r^2} = \frac{z}{H}\left(1 - \frac{r^2}{\lambda_z^2}\right).$$

Using again the result of the previous example, we find that $p_{Z|W} = \dfrac{p_{WZ}}{p_Z}$ is a Gaussian distribution with mean $\dfrac{\lambda_{wz}}{\lambda_w^2} w$. The maximum of $p_{Z|W}$ is obtain in correspondence of its mean, i.e. for $z = \dfrac{\lambda_{wz}}{\lambda_w^2} w$ or, equivalently, for $w = \dfrac{\lambda_w^2}{\lambda_{wz}} z$. Therefore the ML estimator reads

$$\widehat{\mathbf{w}}_{ML} = \frac{\lambda_w^2}{\lambda_{wz}} z = \frac{z}{H}.$$

As expected, $\lim\limits_{\lambda_w \to \infty} \widehat{\mathbf{w}}_{MAP} = \widehat{\mathbf{w}}_{ML}$. The estimators coincide also when $r^2 = 0$, and return z/H. In fact, in this case the noise is 0, so the estimators gives the deterministic relation (obtained by (8) for $\nu = 0$) $w = z/H$.

Example

Assume now that $\mathbf{w}$ and $\mathbf{z}$ are n-dimensional vectors, $\mathbf{w} \sim \mathcal{G}(\boldsymbol{\mu}, \Lambda)$ and

$$\mathbf{z} = \mathrm{H}\mathbf{w} + \boldsymbol{\nu}$$

where $\boldsymbol{\nu} \sim \mathcal{G}(0, \mathrm{R})$ is the noise independent of $\mathbf{w}$. H is called *observation matrix.*

It is possible to prove that $\mathbf{z} \sim \mathcal{G}(\boldsymbol{\mu}_\mathbf{z}, \Lambda_\mathbf{z},)$ with

$$\begin{aligned}\boldsymbol{\mu}_\mathbf{z} &= \mathcal{E}\left(\mathrm{H}\mathbf{w} + \boldsymbol{\nu}\right) = \mathrm{H}\boldsymbol{\mu}\\ \Lambda_\mathbf{z} &= \mathcal{E}\left((\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})^T(\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})\right) = \mathrm{H}\Lambda\mathrm{H}^T + \mathrm{R}.\end{aligned}$$

For the conditional probabilities, we find that

$$p_{\mathbf{w}|\mathbf{z}} = \frac{\sqrt{|\Lambda_\mathbf{z}|}}{\sqrt{(2\pi)^n|\Lambda||\mathrm{R}|}} \exp(-\frac{1}{2}\mathrm{J})$$

where $\mathrm{J} = (\mathbf{w} - \widehat{\mathbf{w}})^T\Lambda_\mathbf{e}^{-1}(\mathbf{w} - \widehat{\mathbf{w}})$ and $\Lambda_\mathbf{e}^{-1} = \Lambda^{-1} + \mathrm{H}^T\mathrm{R}^{-1}\mathrm{H}$ and

$$\widehat{\mathbf{w}}_{MV} \equiv \mathcal{E}\left(p_{\mathbf{w}|\mathbf{z}}\right) = \Lambda_\mathbf{e}\left(\mathrm{H}^T\mathrm{R}^{-1}\mathbf{z} + \Lambda^{-1}\boldsymbol{\mu}\right) = \widehat{\mathbf{w}}_{MAP}.$$

Moreover, we find that $p_{\mathbf{z}|\mathbf{w}}$ has average $\mathrm{H}\mathbf{w}$ and $\Lambda_{\mathbf{z}|\mathbf{w}} = \mathrm{R}$. If we maximize the likelihood, we obtain

$$\widehat{\mathbf{w}}_{ML} = \mathrm{H}^{-1}\mathbf{z}.$$

Again, it is possible to verify that the MV/MAP estimator is unbiased and the ML estimator is obtained by the MAP, when $\Lambda^{-1} \to 0$.

Remark 2.1 *Contrarily to what previous examples may suggest, the coincidence of MV and MAP is not true in general.*

2.3 The Kalman Filter for Linear problems

Kalman filter [48] is one of the most important algorithms of the 20th century, with an exceptional number of applications, ranging from robotics to mathematical finance. It is concerned with the case of a dynamical system, when the variable to be estimated is supposed to be the solution of a time-dependent linear system. Since for the biomedical applications of interest here, dynamics is in general given by the time discretization of a PDE (as we will see later on), here we consider a time discrete case, even though the time-continuous case can be investigated as well. The time index will be denoted by k, and we represent the dynamics of interest (indexed by k) of the system as

$$\mathbf{u}^{(k)} = \mathrm{A}_{k-1}\mathbf{u}^{(k-1)} + \mathbf{b}^{(k-1)} \tag{9}$$

where $\mathbf{b}^{(k-1)}$ is a *Gaussian white noise* in time representing the model error, i.e. $\mathbf{b}^{(k)} \sim \mathcal{G}(\mathbf{0}, \mathrm{Q}_k)$, and the errors are not correlated in time, i.e.

$$\mathcal{E}\left(\mathbf{b}^{(k)}\mathbf{b}^{(l),T}\right) = \mathrm{Q}_k\delta_{kl}.$$

Here δ_{kl} is the Kronecker delta ($= 1$ if $k = l$, 0 elsewhere).

The measurement process is denoted by

$$\mathbf{z}^{(k)} = \mathrm{H}_k\mathbf{u}^{(k)} + \boldsymbol{\nu}^{(k)}, \tag{10}$$

where $\boldsymbol{\nu}^{(\cdot)}$ is a Gaussian white noise with variance matrix R_k and assumed uncorrelated with $\mathbf{b}^{(\cdot)}$.

In absence of observations of $\mathbf{z}^{(k)}$, a natural prediction is simply the one obtained by dropping the noise in the system. In other terms a first reasonable prediction would be

$$\mathbf{u}_p^{(k)} = \mathrm{A}_{k-1}\mathbf{u}_*^{(k-1)}. \tag{11}$$

For the moment being, we assume that $\mathbf{u}_*^{(k-1)}$ is the "true" state $\mathbf{u}^{(k-1)}$.

This is a deterministic forecast that we take as starting point of our probabilistic measure. The fundamental questions now are: *how the measure* $\mathbf{z}^{(k)}$ *can improve this estimate? How can we reduce the error between the state and its prediction?*

As an arbitrary but reasonable choice, we postulate a *correction step* which is a linear combination between the prediction $\mathbf{u}_p^{(k)}$ and the data at the same instant $\mathbf{z}^{(k)}$,

$$\mathbf{u}_c^{(k)} = \mathrm{L}_k\mathbf{u}_p^{(k)} + \mathrm{K}_k\mathbf{z}^{(k)}.$$

The first term on the right hand side measures how much the deterministic model is trustworthy, the latter defines the gain due to the observation. The weighting matrices L_k, K_k need to be determined. Let us introduce the estimate errors

$$\mathbf{e}_p^{(k)} = \mathbf{u}_p^{(k)} - \mathbf{u}^{(k)}, \quad \mathbf{e}_c^{(k)} = \mathbf{u}_c^{(k)} - \mathbf{u}^{(k)}.$$

Notice that $\mathbf{e}_p^{(k)} = -\mathbf{b}^{(k-1)}$ by construction (for $\mathbf{u}_*^{(k-1)} = \mathbf{u}^{(k-1)}$).

We have then

$$\begin{aligned} \mathbf{e}_c^{(k)} = \mathrm{L}_k \mathbf{u}_p^{(k)} + \mathrm{K}_k \mathbf{z}^{(k)} - \mathrm{L}_k \mathbf{u}^{(k)} + \mathrm{L}_k \mathbf{u}^{(k)} - \mathbf{u}^{(k)} = \\ \mathrm{L}_k \mathbf{e}_p^{(k)} + \mathrm{K}_k \boldsymbol{\nu}^{(k)} + (\mathrm{L}_k + \mathrm{K}_k \mathrm{H}_k - \mathrm{I})\, \mathbf{u}^{(k)}. \end{aligned} \tag{12}$$

In order to have an unbiased correction, i.e. $\mathcal{E}\left(\mathbf{e}_c^{(k)}\right) = 0$, we write

$$\begin{aligned} \mathcal{E}\left(\mathbf{e}_c^{(k)}\right) = \mathrm{L}_k \mathcal{E}\left(\mathbf{e}_p^{(k)}\right) + \mathrm{K}_k \mathcal{E}\left(\boldsymbol{\nu}^{(k)}\right) + (\mathrm{L}_k + \mathrm{K}_k \mathrm{H}_k - \mathrm{I})\, \mathcal{E}\left(\mathbf{u}^{(k)}\right) = \\ -\mathrm{L}_k \mathcal{E}\left(\mathbf{b}^{(k-1)}\right) + \mathrm{K}_k \mathcal{E}\left(\boldsymbol{\nu}^{(k)}\right) + (\mathrm{L}_k + \mathrm{K}_k \mathrm{H}_k - \mathrm{I})\, \mathcal{E}\left(\mathbf{u}^{(k)}\right) = 0. \end{aligned} \tag{13}$$

Because we assume that the noise has null mean, the first two terms are zero. To have an unbiased estimate we need to force

$$\mathrm{L}_k + \mathrm{K}_k \mathrm{H}_k - \mathrm{I} = 0 \Rightarrow \mathrm{L}_k = \mathrm{I} - \mathrm{K}_k \mathrm{H}_k.$$

With this position, we have

$$\mathbf{u}_c^{(k)} = \mathbf{u}_p^{(k)} + \mathrm{K}_k (\mathbf{z}^{(k)} - \mathrm{H}_k \mathbf{u}_p^{(k)}).$$

This representation is extremely expressive:

1. the first term on the right hand side $\mathbf{u}_p^{(k)}$ is the deterministic estimate purely based on the *model*, with no observations;
2. $\mathbf{z}^{(k)} - \mathrm{H}_k \mathbf{u}_p^{(k)}$ is the *innovation*, i.e. what is new in $\mathbf{z}^{(k)}$ and that is not predictable by $\mathbf{u}_p^{(k)}$;
3. K_k, to be determined, is called the *gain matrix*, since it weighs the improvement brought to the deterministic estimate by the measures.

The two estimation errors are then related by the following equation

$$\begin{aligned} \mathbf{e}_c^{(k)} = \quad & \mathbf{u}_c^{(k)} - \mathbf{u}^{(k)} = \mathbf{u}_p^{(k)} - \mathbf{u}^{(k)} + \mathrm{K}_k (\mathrm{H}_k \mathbf{u}^{(k)} + \boldsymbol{\nu}^{(k)} - \mathrm{H}_k \mathbf{u}_p^{(k)}) = \\ & (\mathrm{I} - \mathrm{K}_k \mathrm{H}_k)\, \mathbf{e}_p^{(k)} + \mathrm{K}_k \boldsymbol{\nu}^{(k)}. \end{aligned} \tag{14}$$

However, in the sequential process we do not know the true state $\mathbf{u}^{(k-1)}$ in (11). We replace $\mathbf{u}_*^{(k-1)}$ with what we consider sequentially our best estimation of the state, which is $\mathbf{u}_c^{(k-1)}$. The recursive *prediction* step reads therefore

$$\mathbf{u}_p^{(k)} = \mathrm{A}_{k-1} \mathbf{u}_c^{(k-1)}. \tag{15}$$

From this equation, we derive another relation for the errors

$$\begin{aligned} \mathbf{e}_p^{(k)} = \quad & \mathbf{u}_p^{(k)} - \mathbf{u}^{(k)} = \mathrm{A}_{k-1} \mathbf{u}_c^{(k-1)} - \mathbf{u}^{(k)} = \\ & \mathrm{A}_{k-1} \left(\mathbf{u}_c^{(k-1)} - \mathbf{u}^{(k-1)}\right) - \mathbf{b}^{(k-1)} = \mathrm{A}_{k-1} \mathbf{e}_c^{(k-1)} - \mathbf{b}^{(k-1)} \end{aligned} \tag{16}$$

giving an evolution equation for the deterministic forecast error.

Remark 2.2 *In this analysis, we are considering one-step prediction estimates, where the deterministic estimate is obtained by the previous step* (15)*. We may consider also q-step predictions, obtained by dropping the noise at each step,*

$$\mathbf{u}_p^{(k)} = \prod_{j=1}^{q} \mathrm{A}_{k-j}\, \mathbf{u}_c^{(k-q)}.$$

For the sake of simplicity, here we develop the case $q = 1$ and refer the interested reader to [68].

Let us compute the variance matrix of $\mathbf{e}_p^{(k)}$ and $\mathbf{e}_c^{(k)}$, i.e.

$$\Lambda_p^{(k)} \equiv \mathcal{E}\left(\mathbf{e}_p^{(k)}\mathbf{e}_p^{(k,T)}\right), \qquad \Lambda_c^{(k)} \equiv \mathcal{E}\left(\mathbf{e}_c^{(k)}\mathbf{e}_c^{(k,T)}\right). \tag{17}$$

From (16), we have

$$\mathbf{e}_p^{(k)}\mathbf{e}_p^{(k),T} = \mathrm{A}_{k-1}\mathbf{e}_c^{(k-1)}\mathbf{e}_c^{(k-1),T}\mathrm{A}_{k-1}^T + \mathbf{b}^{(k-1)}\mathbf{b}^{(k-1),T} +$$

$$\mathrm{A}_{k-1}\mathbf{e}_c^{(k-1)}\mathbf{b}^{(k-1),T} + \mathbf{b}^{(k-1)}\mathbf{e}_c^{(k-1),T}\mathrm{A}_{k-1}^T,$$

leading to

$$\Lambda_p^{(k)} = \mathrm{A}_{k-1}\Lambda_c^{(k-1)}\mathrm{A}_{k-1}^T + \mathrm{Q}_{k-1} \tag{18}$$

because $\mathbf{b}^{(k-1)}$ has no correlation with $\mathbf{e}_c^{(k-1)}$.

Similarly, from (14) we obtain

$$\Lambda_c^{(k)} = (\mathrm{I} - \mathrm{K}_k\mathrm{H}_k)\Lambda_p^{(k)}(\mathrm{I} - \mathrm{K}_k\mathrm{H}_k)^T + \mathrm{K}_k\mathrm{R}_k\mathrm{K}_k^T. \tag{19}$$

usually called *Joseph formula.*

2.3.1 The Kalman gain matrix

Finally we determine the gain matrix. We will follow an optimality criterion. According to the (weighted) minimal variance approach, we could minimize $\mathcal{E}\left(\mathbf{e}_c^{(k,T)}\mathrm{E}_k\mathbf{e}_c^{(k)}\right)$, where E_k is a s.p.d. weight matrix. Note that

$$\mathbf{e}_c^{(k),T}\mathrm{E}_k\mathbf{e}_c^{(k)} = \mathrm{Tr}(\mathrm{E}_k\mathbf{e}_c^{(k)}\mathbf{e}_c^{(k),T}) \Rightarrow$$

$$\mathcal{E}\left(\mathbf{e}_c^{(k),T}\mathrm{E}_k\mathbf{e}_c^{(k)}\right) = \mathcal{E}\left(\mathrm{Tr}(\mathrm{E}_k\mathbf{e}_c^{(k)}\mathbf{e}_c^{(k),T})\right) = \mathrm{Tr}(\mathrm{E}_k\Lambda_c^{(k)}).$$

Using the following properties of the trace:

$$\mathrm{Tr}(\mathrm{A}+\mathrm{B}) = \mathrm{Tr}(\mathrm{A}) + \mathrm{Tr}(\mathrm{B}), \qquad \mathrm{Tr}(\mathrm{AB}) = \mathrm{Tr}(\mathrm{A}^T\mathrm{B}^T),$$

we get that

$$\mathrm{Tr}(\mathrm{E}_k\Lambda_c^{(k)}) = \mathrm{Tr}(\mathrm{E}_k\Lambda_p^{(k)}) - 2\mathrm{Tr}(\mathrm{E}_k\Lambda_p^{(k)}\mathrm{H}_k^T\mathrm{K}_k^T) + \mathrm{Tr}(\mathrm{E}_k\mathrm{K}_k(\mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T + \mathrm{R}_k)\mathrm{K}_k^T).$$

Moreover, we recall that [60] for a generic matrix A and symmetric matrices B and C we have

$$\frac{\partial \mathrm{Tr}(\mathrm{AX}^T)}{\partial \mathrm{X}} = \mathrm{A}, \qquad \frac{\partial \mathrm{Tr}(\mathrm{BXCX}^T)}{\partial \mathrm{X}} = 2\mathrm{BCX}.$$

From these relations we obtain that the gain matrix K_k that minimizes the variance is such that .

$$\begin{aligned}\frac{\partial \mathcal{E}\left(\mathbf{e}_c^{(k,T)}\mathrm{E}_k\mathbf{e}_c^{(k)}\right)}{\partial \mathrm{K}_k} &= -2\mathrm{E}_k\Lambda_p^{(k)}\mathrm{H}_k^T + 2\mathrm{E}_k\mathrm{K}_k(\mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T + \mathrm{R}_k) = 0 \\ &\Rightarrow \mathrm{K}_k^* = \Lambda_p^{(k)}\mathrm{H}_k^T\left(\mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T + \mathrm{R}_k\right)^{-1}.\end{aligned}$$

From the Joseph formula we have

$$\Lambda_c^{(k)} = (\mathrm{I} - \mathrm{K}_k\mathrm{H}_k)\Lambda_p^{(k)} - \Lambda_p^{(k)}\mathrm{H}_k^T\mathrm{K}_k^T + \mathrm{K}_k(\mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T + \mathrm{R}_k)\mathrm{K}_k^T.$$

By using K_k^* in this formula, the last two terms cancel out and we are left with

$$\Lambda_c^{(k)} = (\mathrm{I} - \mathrm{K}_k\mathrm{H}_k)\,\Lambda_p^{(k)}.$$

The matrix K_k^* is the so-called *Kalman gain matrix.*

The entire estimate process may be summarized as follows.

1. PREDICTION
   (a) $\mathbf{u}_p^{(k)} = \mathrm{A}_{k-1}\mathbf{u}_c^{(k-1)}$
   (b) $\Lambda_p^{(k)} = \mathrm{A}_{k-1}\Lambda_c^{(k-1)}\mathrm{A}_{k-1}^T + \mathrm{Q}_{k-1}$.
2. CORRECTION
   Kalman gain: $\mathrm{K}_k^* = \Lambda_p^{(k)}\mathrm{H}_k^T\left(\mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T + \mathrm{R}_k\right)^{-1}$.
   (a) State estimate:
   $$\mathbf{u}_c^{(k)} = \mathbf{u}_p^{(k)} + \mathrm{K}_k^*(\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)}). \tag{20}$$
   (b) Covariance estimate:
   $$\Lambda_c^{(k)} = (\mathrm{I} - \mathrm{K}_k^*\mathrm{H}_k)\,\Lambda_p^{(k)}. \tag{21}$$

This yields the MV *Kalman filter* estimation for a time-discrete system. When we want to *estimate* the state of the system by merging the mathematical model and the available measure, we refer to $\mathbf{u}_c^{(k)}$. When we want to *predict* the evolution of the state, using all the information we have at time t^k, we refer to the prediction[4] step $\mathbf{u}_p^{(k+1)} = \mathrm{A}_k\mathbf{u}_c^{(k)}$.

[4] The third problem addressed in the Introduction, the *identification* of the system will be addressed later on.

Remark 2.3 *In many cases, the dynamical system features a deterministic input* $\mathbf{f}^{(k)}$*, so that* (9) *modifies in*

$$\mathbf{u}^{(k)} = \mathrm{A}_{k-1}\mathbf{u}^{(k-1)} + \mathrm{C}_{k-1}\mathbf{f}^{(k-1)} + \mathbf{b}^{(k-1)}. \tag{22}$$

This reflects in a change of the prediction step, that reads

$$\mathbf{u}_p^{(k)} = \mathrm{A}_{k-1}\mathbf{u}_c^{(k-1)} + \mathrm{C}_{k-1}\mathbf{f}^{(k-1)}.$$

All the other steps drawn above remain unchanged.

2.3.2 Properties of the Kalman filter

Orthogonality of the estimate/prediction and the estimate/prediction error Following an induction argument, it is possible to prove that when we select the Kalman gain matrix to compute the estimate, then

$$\mathcal{E}\left(\mathbf{u}_c^{(k)}\mathbf{e}_c^{(k),T}\right) = 0. \tag{23}$$

With this relation, it is possible to prove a similar relation between the prediction and the prediction error

$$\mathcal{E}\left(\mathbf{u}_p^{(k+1)}\mathbf{e}_p^{(k+1),T}\right) = \mathcal{E}\left(A_k\mathbf{u}_c^{(k)}\left(A_k\mathbf{e}_c^{(k)}\right)^T\right) = 0. \tag{24}$$

These relations have an interesting geometrical interpretation that provide a justification to the "optimal" nature of Kalman estimate/prediction. The correlation operator $\mathcal{E}\left((\cdot)(\cdot)^T\right)$ is a *scalar product.* For this reason, the two equations (23) and (24) state that the estimate and the prediction are orthogonal to their respective errors. This is the feature of optimal approximations by projection. As a matter of fact, we can conclude that *among all the possible estimates generated by any possible gain matrices, Kalman provides the optimal one in the metric defined by the correlation scalar product* (see Fig. 2.3.2).

Innovation As we pointed out, the *innovation*

$$\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)} = \mathbf{z}^{(k)} - \mathrm{H}_k\mathrm{A}_k\mathbf{u}_c^{(k-1)}$$

plays an important role in understanding how the Kalman estimate works. When we do not have other *a priori* information, from A_k and $\mathbf{u}_c^{(k-1)}$ the best we can do is

- to predict the state at k as $\mathbf{u}_p^{(k)} = \mathrm{A}_k\mathbf{u}_c^{(k-1)}$;
- to guess accordingly an "expected measure" $\mathrm{H}_k\mathrm{A}_k\mathbf{u}_c^{(k-1)}$.

This is the part of knowledge in the measure we could extract from the state at the previous time step, or *from the past.* We do expect that $\mathbf{z}^{(k)}$ is adding new information. The novel part of the information added by the measure is exactly the innovation $\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)}$.

Notice that

$$\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)} = \mathrm{H}_k\mathbf{u}^{(k)} + \boldsymbol{\nu}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)} = \boldsymbol{\nu}^{(k)} - \mathrm{H}_k\mathbf{e}_p^{(k)},$$

consequently

$$\mathcal{E}\left(\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)}\right) = \mathcal{E}\left(\boldsymbol{\nu}^{(k)}\right) - \mathrm{H}_k\mathcal{E}\left(\mathbf{e}_p^{(k)}\right) = 0.$$

In addition, we compute the variance of the innovation.

$$\begin{aligned}&\mathcal{E}\left((\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)})(\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)})^T\right) = \\ &\mathcal{E}\left((\boldsymbol{\nu}^{(k)} - \mathrm{H}_k\mathbf{e}_p^{(k)})(\boldsymbol{\nu}^{(k)} - \mathrm{H}_k\mathbf{e}_p^{(k)})^T\right) = \mathrm{R}_k + \mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T\end{aligned}$$

because the noise at k is not correlated to $\mathbf{e}_p^{(k)} = \mathrm{A}_k(\mathbf{u}_c^{(k)} - \mathbf{u}^{(k-1)}) - \mathbf{b}^{(k-1)}$.

It is also possible to prove [68] that for $j \geq 1$

$$\mathcal{E}\left((\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}_p^{(k)})(\mathbf{z}^{(k-j)} - \mathrm{H}_k\mathbf{u}_p^{(k-j)})^T\right) = 0.$$

This means that the innovation at time k has no correlation with the innovation at the previous time steps, so that we can conclude that the innovation is a *white process*.

Also in this case, we may give a geometrical interpretation to this equation, concluding that the splitting $\mathbf{z}^{(k)} = \text{Predicted measure} + \text{Innovation}$ is actually an orthogonal decomposition. Since the predicted measure $\mathrm{H}_k\mathbf{u}_p^{(k)}$ depends entirely on the past, we say that *innovation is orthogonal to the past* (see Fig. 2.3.2).

Variance reduction Let us establish a relation between the variance of $\mathbf{u}_p^{(k)}$ and of $\mathbf{u}_c^{(k)}$. We show that the Kalman correction in fact reduces the variance of the estimate. Let us introduce an auxiliary variable that we call *pseudo-observation*, i.e. an observation based on the prediction of the measure added by noise,

$$\mathbf{z}_{po}^{(k)} = \mathrm{H}_k\mathbf{u}_p^{(k)} + \boldsymbol{\nu}^{(k)}.$$

It is possible to verify that

$$\Lambda_{po}^{(k)} = \mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T + \mathrm{R}_k, \quad \Lambda_{p,po}^{(k)} := \mathcal{E}\left(\mathbf{u}_p^{(k)}\,\mathbf{z}_{po}^{(k),T}\right) = \Lambda_p^{(k)}\mathrm{H}_k^T.$$

With this notation, we may rewrite the correction step of the Kalman filter as follows,

$$\mathrm{K}_k^* = \Lambda_{p,po}^{(k)}(\Lambda_{po}^{(k)})^{-1}, \quad \Lambda_c^{(k)} = \Lambda_p^{(k)} - \Lambda_{p,po}^{(k)}(\Lambda_{po}^{(k)})^{-1}\Lambda_{p,po}^{(k),T}.$$

Since $\Lambda_{po}^{(k)}$ is s.p.d., we have that $\Lambda_p^{(k)} - \Lambda_c^{(k)}$ is positive. This relation outlines the reduction of the variance induced by the correction step with respect to the variance of the prediction.

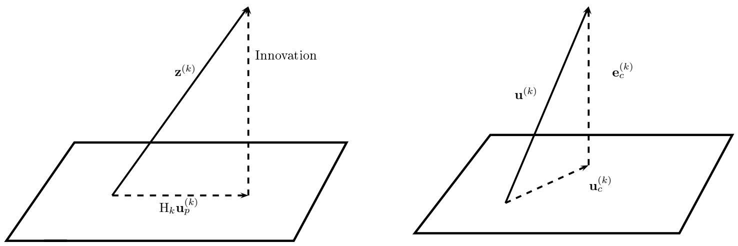


Figure 5: The innovation is orthogonal to the past (left) according to the correlation scalar product. The estimate error is orthogonal to the estimate itself (right) according to the correlation scalar product, that qualifies the Kalman correction as the orthogonal projection of the state to be estimated on the subspace of possible estimations obtained for different gain matrices.

Recursive formula for the variance equations, Riccati equations Let us eliminate $\Lambda_c^{(k)}$ from the equations of the Kalman filter, in particular we compute the variance $\Lambda_p^{(k+1)}$ as function of $\Lambda_c^{(k)}$ and $\Lambda_p^{(k)}$. We get

$$\begin{aligned}\Lambda_p^{(k+1)} = \quad & \mathrm{A}_k\Lambda_c^{(k)}\mathrm{A}_k^T + \mathrm{Q}_k = \\ & \mathrm{A}_k(\Lambda_p^{(k)} - \mathrm{K}_k\mathrm{H}_k\Lambda_p^{(k)})\mathrm{A}_k^T + \mathrm{Q}_k = \\ & \mathrm{A}_k\Lambda_p^{(k)}\mathrm{A}_k^T + \mathrm{Q}_k - \mathrm{A}_k\Lambda_p^{(k)}\mathrm{H}_k^T(\mathrm{H}_k\Lambda_p^{(k)}\mathrm{H}_k^T + \mathrm{R}_k)^{-1}\mathrm{H}_k\Lambda_p^{(k)}\mathrm{A}_k^T.\end{aligned} \tag{25}$$

In addition, by using the well know Sherman-Morrison-Woodbury formula [35], we can write also the recursive variance matrix equation of the Kalman estimate

$$\begin{aligned}\Lambda_c^{(k)} = (\mathrm{I} - \mathrm{K}_k^*\mathrm{H}_k)\,\Lambda_p^{(k)} = \left((\Lambda_p^{(k)})^{-1} + \mathrm{H}_k^T\mathrm{R}_k^{-1}\mathrm{H}_k\right)^{-1} = \\ \left(\left(\mathrm{A}_{k-1}\Lambda_c^{(k-1)}\mathrm{A}_{k-1}^T + \mathrm{Q}_{k-1}\right)^{-1} + \mathrm{H}_k^T\mathrm{R}_k^{-1}\mathrm{H}_k\right)^{-1}.\end{aligned} \tag{26}$$

Let us assume that the matrices $\mathrm{A}, \mathrm{H}, \mathrm{R}, \mathrm{Q}$ do not depend on k. The latter equation in (25) reads then

$$\Lambda_p^{(k+1)} = \mathrm{A}\Lambda_p^{(k)}\mathrm{A}^T + \mathrm{Q} - \mathrm{A}\Lambda_p^{(k)}\mathrm{H}^T(\mathrm{H}\Lambda_p^{(k)}\mathrm{H}^T + \mathrm{R})^{-1}\mathrm{H}\Lambda_p^{(k)}\mathrm{A}^T.$$

This is called *Difference Riccati Equation* (DRE). A reasonable question related to this equation for time-independent dynamics refers to the existence of a stationary variance matrix. This is a variance matrix such that

$$\Lambda_p^{(k+1)} = \Lambda_p^{(k)} = \Lambda_p.$$

The latter can be clearly obtained as a fixed point of the DRE. This leads to solve the so called *Algebraic Riccati Equation* (ARE)

$$\Lambda_p = \mathrm{A}\Lambda_p\mathrm{A}^T + \mathrm{Q} - \mathrm{A}\Lambda_p\mathrm{H}^T(\mathrm{H}\Lambda_p\mathrm{H}^T + \mathrm{R})^{-1}\mathrm{H}\Lambda_p\mathrm{A}^T.$$

This equation has been largely investigated by several authors [1, 46, 51], to determine under which conditions the solution Λ_{ARE} exists and it can be computed as the asymptotic limit of the corresponding DRE. In particular, let us assume that the dynamical system is asymptotically stable and converges to a steady solution. Clearly a good predictor is expected to follow the system dynamics, converging to the asymptotic estimate. Correspondingly, in this case we expect $\Lambda_p^{(k)}$ to converge to the asymptotic matrix Λ_{ARE}. Otherwise, our predictor would be unable to follow the system dynamics converging to the stationary solution. As a matter of fact, it is possible to prove that *if the system is stable, then the predictor is stable and its variance gets closer to the solution of the associated ARE* (see e.g. [46] for a precise statement of the Theorem).

In addition, we point out that this solution can be interpreted as an "approximate“ Kalman filter, where the matrix $\Lambda_p^{(k)}$ is replaced by Λ_{ARE} to save the computational costs of computing $\Lambda_p^{(k)}$ at each step. This provides a stationary filter which is clearly sub-optimal, since the associated error is not orthogonal to the estimate. However, it may be computationally convenient.

Another possible use of Λ_{ARE} is to provide a bound to the variance of the "optimal case" when we apply the Kalman filter with no approximations.

Example

Let us consider the scalar case, with

$$\begin{array}{ll} u^{(k)} = u^{(k-1)} & (A = 1, b = 0) \\ z^{(k)} = u^{(k)} + \nu^{(k)} & (H = 1, \nu \sim \mathcal{G}(0,1)). \end{array}$$

Assume also that the initial data $u^{(1)} \sim \mathcal{G}(\mu, 1)$. Set $u_p^{(1)} = \mu$. Then, the Kalman filter formulas read

$$\begin{aligned} & u_p^{(k)} = u_c^{(k-1)}, \quad \lambda_p^{(k),2} = \lambda_c^{(k-1),2} \\ & K_k = \frac{\lambda_p^{(k),2}}{\lambda_p^{(k),2} + 1} \\ & u_c^{(k)} = u_p^{(k)} + \frac{\lambda_p^{(k),2}}{\lambda_p^{(k),2} + 1}(z^{(k)} - u_p^{(k)}) = \frac{1}{\lambda_p^{(k),2} + 1} u_p^{(k)} + \frac{\lambda_p^{(k),2}}{\lambda_p^{(k),2} + 1} z^{(k)} \\ & \lambda_c^{(k),2} = \frac{\lambda_p^{(k),2}}{\lambda_p^{(k),2} + 1} = \frac{\lambda_c^{(k-1),2}}{\lambda_c^{(k-1),2} + 1}. \end{aligned}$$

We have therefore

$$u_p^{(1)} = \mu, \ \lambda_p^{(1)} = 1, \ K_1 = \frac{1}{2}, \ u_c^{(1)} = \frac{\mu + z^{(1)}}{2} = u_p^{(2)}.$$

Notice that the prediction at $k = 2$ is just the sample average of the "past" and the new data. Similarly we obtain at a generic step k

$$u_p^{k+1} = u_c^k = \frac{\mu + \sum_{j=1}^k z^{(j)}}{k+1}.$$

Actually, we have the arithmetic average of the available data at t^k, that is somehow intuitively expected. Moreover, we have the recursive formula

$$\lambda_p^{(k+1),2} = \frac{\lambda_p^{(k),2}}{\lambda_p^{(k),2}+1}, \quad \text{with } \lambda_p^{(1)} = 1.$$

By induction one can check that $\lambda_p^{(k),2} = \dfrac{1}{k}$. Consequently we have that

1. $\lim\limits_{k\to\infty} \lambda_p^{(k)} = 0$, i.e. the prediction is asymptotically exact; similarly, the estimate is asymptotically exact;
2. the ARE $\lambda^2 = \lambda^2/(1+\lambda^2)$ has only one solution, that is 0;
3. the Kalman filter is asymptotically stable, whereas the dynamic system is not asymptotically stable.

This example provides the case of an asymptotically stable estimator even when the dynamical system is not stable. As we have pointed out, the "reverse" situation (system is stable, estimator is unstable) is not possible: when the system is stable, the predictor is automatically stable.

An alternative look at the Kalman filter The Kalman filter can be obtained in different ways. Among the others, in particular here we mention a recent approach presented in [42], where the algorithm is the result of an application of the Newton root finding method with an appropriate initial guess. Beyond its intrinsic interest, this approach is actually oriented to extension to nonlinear systems in the form of an application the Gauss-Newton method.

More precisely, assuming to have the exact initial state $\mathbf{u}^{(0)}$, let us consider the prediction-mismatch functional

$$\mathcal{J}_{k,p} = \frac{1}{2}\sum_{j=1}^{k} \|\mathbf{u}^{(j)} - \mathrm{A}_{j-1}\mathbf{u}^{(j-1)}\|^2_{\mathrm{Q}_j^{-1}} + \frac{1}{2}\sum_{j=1}^{k-1} \|\mathbf{z}^{(j)} - \mathrm{H}_j\mathbf{u}^{(i)}\|^2_{\mathrm{R}_j^{-1}}.$$

and the corresponding one for the estimate

$$\mathcal{J}_{k,c} = \mathcal{J}_{k,p} + \frac{1}{2}\|\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}^{(k)}\|^2_{\mathrm{R}_k^{-1}}.$$

The latter has the "natural" recursive formulation

$$\mathcal{J}_{k,c} = \mathcal{J}_{k-1,c} + \frac{1}{2}\|\mathbf{u}^{(k)} - \mathrm{A}_{k-1}\mathbf{u}^{(k-1)}\|^2_{\mathrm{Q}_k^{-1}} + \frac{1}{2}\|\mathbf{z}^{(k)} - \mathrm{H}_k\mathbf{u}^{(k)}\|^2_{\mathrm{R}_k^{-1}}.$$

We estimate $\mathbf{u}^{(k)}$ as the arg min of $\mathcal{J}_{k,c}$. When solving $\nabla\mathcal{J}_{k,c} = 0$, we apply the Newton method, that reads

$$\mathcal{H}\left(\mathbf{u}_{new} - \mathbf{u}_{old}\right) = -\nabla\mathcal{J}_{k,c}(\mathbf{u}_{old}) \tag{27}$$

where $\mathcal{H}$ is the Hessian matrix associated to $\mathcal{J}_{k,c}$. By selecting $\mathbf{u}_{old} = \mathbf{u}_p^{(k)} = \mathrm{A}_{k-1}\mathbf{u}_c^{(k-1)}$, it is possible to prove [42] that the Kalman estimate $\mathbf{u}_c^{(k)}$ is the solution $\mathbf{u}_{new}$ of (27). In the case of a linear system, $\mathbf{u}_{new} = \mathbf{u}_c^{(k)}$ minimizes $\mathcal{J}_{k,c}$. In fact, Newton method converges in one iteration when applied to linear equations.

2.3.3 Computational issues associated with the Kalman Filter

There are several issues associated with practical computation of the Kalman filter. Here we mention just a few.

From the numerical view point the implementation of the filter following closely the formulas given above has a cost of $\mathcal{O}(n^3)$ operations at each time iteration, where n is the dimension of the matrix A. This cost is basically driven by the computation of the variance and gain matrices. For systems coming from the discretization of partial differential equations, n may be a fairly large number. However, the matrix is usually sparse and - as it is well known - this may reduce significantly the storage requirements and the number of operations. In addition, computational cost may still be an issue and specific methods for reducing the costs are mandatory. Among the others, we mention the replacement of the estimate covariance matrix with the asymptotic one (when the system dynamics is time independent) obtained by solving the ARE, as pointed out above.

Another computational issue is *numerical stability*. In particular, when computing the estimate variance matrix, equation (21) depends linearly on the computation error associated with the Kalman gain matrix K. In this respect, using Joseph formula is beneficial, since numerical errors are propagated quadratically. More in general, numerical errors may lead to computing non-positive covariance matrices. This problem can be faced by resorting to appropriate Cholesky or LDL^T factorizations of the covariance matrices that guarantee their numerical positiveness, leading to the so-called *square root form* of the filter.

2.4 Extension of the Kalman Filter to nonlinear problems

The most relevant limitation of the Kalman filter theory presented is that it relies upon linearity of the dynamical system, and that Gaussian densities remain Gaussian after linear transformations. However, in most of practical applications, the problem to solve is nonlinear. We show an important example hereafter (and many others later on, for the applications relevant to the contents of the present book).

We need therefore to find a way for extending the method to nonlinear cases, by properly approximating the procedures. We see methods based on both linearization as well as sampling.

2.4.1 Parameter identification via Kalman filter

Consider a problem represented by a dynamical system with some parameters that we want to identify. We defined this as an *identification* problem. A possible approach to the problem is to add the parameter to the list of state variables and then to perform an estimation procedure. In general, this leads to a nonlinear dynamics. This approach is called *state augmentation technique.* We illustrate this in a case with a linear dynamics for the state variable $\mathbf{u}$

$$\begin{aligned} \mathbf{u}^{(k)} &= \mathrm{A}(\vartheta)\mathbf{u}^{(k-1)} + \mathbf{b}^{(k)} \\ \mathbf{z}^{(k)} &= \mathrm{H}(\vartheta)\mathbf{u}^{(k)} + \boldsymbol{\nu}^{(k)}. \end{aligned} \tag{28}$$

We assume for simplicity that ϑ is a time independent stochastic variable, so we have

$$\vartheta^{(k)} = \vartheta^{(k-1)} + \varepsilon^{(k)}$$

with $\varepsilon^{(k)} \sim \mathcal{G}(0, s_k)$, uncorrelated with other sources of noise. We augment the list of state variables of the parameter, so we have

$$\mathbf{v}^{(k)} = \begin{bmatrix} \mathbf{u}^{(k)} \\ \vartheta^{(k)} \end{bmatrix} \Rightarrow \mathbf{v}^{(k)} = \begin{bmatrix} \mathrm{A}(\vartheta^{(k-1)})\mathbf{u}^{(k-1)} \\ \vartheta^{(k-1)} \end{bmatrix} + \begin{bmatrix} \mathbf{b}^{(k)} \\ \varepsilon^{(k)} \end{bmatrix}, \tag{29}$$

with

$$\mathbf{z}^{(k)} = \mathrm{H}(\vartheta^{(k)})\mathbf{u}^{(k)} + \boldsymbol{\nu}^{(k)}.$$

In general, this is now a nonlinear augmented system so that the Kalman filter method presented above cannot be applied.

2.4.2 The Extended Kalman Filter

The Extended Kalman Filter (EKF) is the most immediate approach to extend the filter based on the *linearization* of both the system dynamics and of the observation process. Let us consider the nonlinear dynamic system

$$\begin{cases} \mathbf{u}^{(k)} = \mathcal{A}(\mathbf{u}^{(k-1)}) + \mathbf{b}^{(k)} \\ \mathbf{z}^{(k)} = \mathcal{H}(\mathbf{u}^{(k)}) + \boldsymbol{\nu}^{(k)}. \end{cases}$$

We still follow the minimal variance approach and introduce the *tangent operators*, i.e. the Jacobian matrices

$$\mathcal{A}'(\cdot) = \frac{\partial \mathrm{A}(\cdot)}{\partial \mathbf{u}}, \quad \mathcal{H}'(\cdot) = \frac{\partial \mathrm{H}(\cdot)}{\partial \mathbf{u}}.$$

After linearization, we get an extension of the Kalman filter. This reads

1. Prediction

   (a) $\mathbf{u}_p^{(k)} = \mathcal{A}(\mathbf{u}_c^{(k-1)})$,

   (b) $\Lambda_p^{(k)} = \mathcal{A}'(\mathbf{u}_c^{(k-1)})\Lambda_c^{(k-1)}(\mathcal{A}'(\mathbf{u}_c^{(k-1)}))^T + \mathrm{Q}_{k-1}$.

2. Correction

   Kalman gain:

   $$\mathrm{K}_k = \Lambda_p^{(k)} \mathcal{H}'(\mathbf{u}_p^{(k)})^T \left(\mathcal{H}'(\mathbf{u}_p^{(k)}) \Lambda_p^{(k)} (\mathcal{H}'(\mathbf{u}_p^{(k)}))^T + \mathrm{R}_k \right)^{-1}.$$

   (a) State estimate:

   $$\mathbf{u}_c^{(k)} = \mathbf{u}_p^{(k)} - \mathrm{K}_k \, \left(\mathbf{z}^{(k)} - \mathcal{H}(\mathbf{u}_p^{(k)}) \right).$$

   (b) Covariance estimate:

   $$\Lambda_c^{(k)} = \Lambda_p^{(k)} - \mathrm{K}_k \, \mathcal{H}'(\mathbf{u}_p^{(k)}) \Lambda_p^{(k)}.$$

As to be expected, most of the analysis holding for the linear case cannot be trivially extended to this case, since the covariance matrices associated with the errors depend on the linearization procedure. In particular, they depend on the set of observations so they are a random process. In addition, they are only an approximation of the error covariance and this leads to biased state estimates ($\mathcal{E}\left(\mathbf{e}_c^{(k)}\right) \neq 0$). Another drawback is the computational cost associated with the tangent operators, that for problems coming from the discretization of partial differential equations may be fairly expensive.

Nevertheless, we address the case of parameter estimation with the EKF.

EKF and parameter estimation Let us apply EKF to (29), with

$$\mathcal{A}'(\mathbf{v}) = \begin{bmatrix} \mathrm{A}(\vartheta) & \dfrac{\partial \mathrm{A}}{\partial \vartheta}\mathbf{u} \\ O & \mathrm{I} \end{bmatrix}, \mathcal{H}'(\mathbf{v}) = \begin{bmatrix} \mathrm{H}(\vartheta) & \dfrac{\partial \mathrm{H}}{\partial \vartheta}\mathbf{u} \end{bmatrix}.$$

In many cases, H is independent of ϑ so that the last entry in $\mathcal{H}'(\mathbf{v})$ is zero. We assume that this is the case hereafter. In this form, the parameter estimation is performed following the EKF steps. . The covariance matrix of the augmented status will be

$$\Lambda_{augm,\cdot} \equiv \left[\begin{array}{cc} \Lambda_{u,\cdot} & \Lambda_{u\vartheta,\cdot} \\ \Lambda_{u\vartheta,\cdot} & \Lambda_{\vartheta,\cdot} \end{array} \right]$$

where the dot can be either p in the prediction or c for the correction (or estimate). Then, we have the following steps.

1. Prediction

   (a) $\mathbf{u}_p^{(k)} = \mathcal{A}(\vartheta_c^{(k-1)})\mathbf{u}_c^{(k-1)}, \quad \vartheta_p^{(k)} = \vartheta_c^{(k-1)}$

   (b) $$\Lambda_{augm,p}^{(k)} = \left[\begin{array}{cc} \mathrm{A}\Lambda_{u,c}^{(k-1)}\mathrm{A}^T + \mathrm{B} + \mathrm{B}^T + \dfrac{\partial \mathrm{A}}{\partial \vartheta}\mathbf{u}_c^{(k-1)}\Lambda_{\vartheta,c}^{(k-1)}\mathbf{u}_c^{(k-1),T}\dfrac{\partial \mathrm{A}^T}{\partial \vartheta} & \mathrm{C} \\ \mathrm{C}^T & \Lambda_{\vartheta,c}^{(k-1)} \end{array} \right]$$

   where $\mathrm{B} = \mathrm{A}\Lambda_{u\vartheta,c}^{(k-1)}\mathbf{u}_c^{(k-1),T}\dfrac{\partial \mathrm{A}^T}{\partial \vartheta}$, $\mathrm{C} = \mathrm{A}\Lambda_{u\vartheta,c}^{(k-1)} + \dfrac{\partial \mathrm{A}}{\partial \vartheta}\mathbf{u}_c^{(k-1)}\Lambda_{\vartheta,c}^{(k-1)}$ and all the occurrences of A and its derivative are computed in $\vartheta_c^{(k-1)}$.

2. Correction

Kalman gain:

$$\mathrm{K}_k = \begin{bmatrix} \Lambda^{(k)}_{p,u}\mathrm{H}^T \\ \Lambda^{(k)}_{p,u\vartheta}\mathrm{H}^T \end{bmatrix} \left(\mathrm{H}\Lambda^{(k)}_{p,u}\mathrm{H}^T + \mathrm{R}_k\right)^{-1} = \begin{bmatrix} \mathrm{K}_{k,1} \\ \mathrm{K}_{k,2} \end{bmatrix}.$$

(a) State and Parameter:

$$\mathbf{u}^{(k)}_c = \mathbf{u}^{(k)}_p - \mathrm{K}_{k,1}\left(\mathbf{z}^{(k)} - \mathrm{H}(\vartheta^{(k)}_p)\mathbf{u}^{(k)}\right),$$
$$\vartheta^{(k)}_c = \vartheta^{(k)}_p - \mathrm{K}_{k,2}\left(\mathbf{z}^{(k)} - \mathrm{H}(\vartheta^{(k)}_p)\mathbf{u}^{(k)}\right).$$

(b) Covariance estimate:

$$\Lambda^{(k)}_{augm,c} = \Lambda^{(k)}_{augm,p} - \begin{bmatrix} \mathrm{K}_{k,1}\mathrm{H}(\vartheta^{(k)}_p)\Lambda^{(k)}_{p,u} & \mathrm{K}_{k,1}\mathrm{H}(\vartheta^{(k)}_p)\Lambda^{(k)}_{p,u\vartheta} \\ \mathrm{K}_{k,2}\mathrm{H}(\vartheta^{(k)}_p)\Lambda^{(k)}_{p,u} & \mathrm{K}_{k,2}\mathrm{H}(\vartheta^{(k)}_p)\Lambda^{(k)}_{p,u\vartheta} \end{bmatrix}.$$

It is worth noting that in this way we have a sort of adaptive filtering, since the improvement of the knowledge of the parameter affects the quality of the state estimate in a self-learning process.

As we have pointed out in the Introduction, there are several ways to perform parameter estimation (see e.g. [2, 3]), this one is just an example. In Section 4.2 we present an example relevant to fluid-structure interaction. Since EKF suffers from the computation of the tangent operators, this can be avoided by resorting to a different extension of the Kalman Filter, that we introduce in the next Section.

Remark 2.4 *EKF can be regarded as the result of the application of one iteration of the Gauss-Newton method for the minimization of a suitable mismatch functional, as we have seen for the linear case. For more details, see [42]*

2.5 The Unscented Kalman Filter (UKF)

As pointed out above, errors associated with the linearization of EKF lead in general to sub-optimal performances. In the UKF [45], the basic idea is to approximate the evolution of the nonlinear dynamic system not by linearization but by deterministic sampling, following the so-called *unscented transformation* (UT). The basic idea of UT is that "it is easier to approximate a Gaussian distribution than it is to approximate an arbitrary nonlinear function or transformation" [44]. For this reason, the nonlinear dynamics in UKF is statistically approximated by mean and covariance of samples suitably selected for the state variable to be estimated.

For instance, suppose to have a scalar Gaussian random variable $u^{(k)}$ with mean μ and variance λ^2. At the first step we determine two samples of $u^{(k)}$, as $s_{1,2} = \mu \pm \lambda$. If we need to approximate a nonlinear evolution $u^{(k+1)} = f(u^{(k)})$, we compute the samples $f_i \equiv f(s_i)$ and take

$$\mathcal{E}\left(u^{(k+1)}\right) \approx w_1 f_1 + w_2 f_2 \equiv \overline{f},$$
$$\mathcal{E}\left(\left(f(u^{(k)}) - \mathcal{E}\left(f(u^{(k)})\right)\right)^2\right) \approx w_1(f_1 - \overline{f})^2 + w_2(f_2 - \overline{f})^2 +,$$

where w_i are suitable weighting coefficients.

The selection of the sampling points (called σ-points) is clearly of paramount importance and can be done in different ways. In general [73, 45, 44], a canonical choice for a state variable $\mathbf{u}^{(k)}$ of size n, with Gaussian distribution with mean $\mathcal{E}\left(\mathbf{u}^{(k)}\right)$ and covariance matrix Λ reads

$$\begin{array}{ll} \mathbf{s}_0 = \mathcal{E}\left(\mathbf{u}^{(k)}\right), \mathbf{s}_i = \mathcal{E}\left(\mathbf{u}^{(k)}\right) + \left(\sqrt{(n+\kappa)\Lambda}\right)_i, \mathbf{s}_{i+n} = \mathcal{E}\left(\mathbf{u}^{(k)}\right) - \left(\sqrt{(n+\kappa)\Lambda}\right)_i, & \\ w_0 = \dfrac{\kappa}{\kappa+n}, \quad w_i = w_{i+n} = \dfrac{\kappa+n}{2}, & i = 1, 2, \ldots n, \end{array}$$

where κ is a real scaling factor and $\left(\sqrt{(n+\kappa)\Lambda}\right)_i$ is the i-th row of the matrix $\sqrt{(n+\kappa)\Lambda}$. This can be computed by a Cholesky factorization of the s.p.d. matrix. Other criteria for sampling can be however pursued [6].

The UKF will eventually consist of a sampling step, followed by the "Kalman-like" prediction and correction steps.

1. Sampling – Let $\mathsf{C}(\cdot)$ denote the Cholesky decomposition of a s.p.d. matrix. We take

$$\begin{aligned} \mathrm{C}_{(k-1)} &= \sqrt{n+\kappa}\,\mathsf{C}(\Lambda_p^{(k-1)}) \\ \mathbf{u}_0^{(k-1)} &= \mathbf{u}_c^{(k-1)}, \\ \mathbf{u}_i^{(k-1)} &= \mathbf{u}_c^{(k-1)} + \mathrm{C}_{(k-1),i}, \quad i = 1, 2, \ldots n, \\ \mathbf{u}_{i+n}^{(k-1)} &= \mathbf{u}_c^{(k-1)} + \mathrm{C}_{(k-1),i}, \quad i = 1, 2, \ldots n. \end{aligned}$$

2. Prediction – Let w_i be the weight coefficients.

$$\begin{aligned} \mathbf{u}_{p,i}^{(k)} &= \mathrm{A}(\mathbf{u}_{p,i}^{(k-1)}) \quad \text{sample evolution} \\ \mathbf{u}_p^{(k)} &= \sum_i w_i \mathbf{u}_{p,i}^{(k)}, \quad \Lambda_p^{(k)} = \sum_i w_i \left(\mathbf{u}_{p,i}^{(k)} - \mathbf{u}_p^{(k)}\right)\left(\mathbf{u}_{p,i}^{(k)} - \mathbf{u}_p^{(k)}\right)^T \end{aligned}$$

3. Correction

$$\begin{aligned} \mathbf{z}_i^{(k)} &= \mathrm{H}(\mathbf{u}_{p,i}^{(k)}) \\ \Lambda_{po}^{(k)} &= \sum_i w_i \left(\mathbf{z}_i^{(k)} - \mathrm{H}(\mathbf{u}_p^{(k)})\right)\left(\mathbf{z}_i^{(k)} - \mathrm{H}(\mathbf{u}_p^{(k)})\right)^T \\ \Lambda_{p,po}^{(k)} &= \sum_i w_i \left(\mathbf{z}_i^{(k)} - \mathrm{H}(\mathbf{u}_p^{(k)})\right)\left(\mathbf{u}_{p,i}^{(k)} - \mathbf{u}_p^{(k)}\right)^T \\ \mathrm{K}_k &= \Lambda_{p,po}^{(k)}\left(\Lambda_{po}^{(k)}\right)^{-1} \\ \mathbf{u}_c^{(k)} &= \mathbf{u}_p^{(k)} + \mathrm{K}_k\left(\mathbf{z}^{(k)} - \mathrm{H}(\mathbf{u}_p^{(k)})\right) \\ \Lambda_c^k &= \Lambda_p^k - \Lambda_{p,po}^{(k)}\left(\Lambda_{po}^{(k)}\right)^{-1}\Lambda_{p,po}^{(k),T}. \end{aligned}$$

Examples of this method can be found in [73]. *UKF dual estimation* is in particular the identification of parameters of the model simultaneous to the state estimation, similarly to what we have illustrated for the EKF. In this respect, we will see an example in Section 4.2. A smart implementation of the methods may be necessary for problems coming from the discretization of partial differential equations, using for example the so-called *Factorized UKF*. In particular, for parameter estimation, computational cost can be reduced by assuming that uncertainty affects only the parameter of interest and not the entire state. For more details, see [6, 54].

3 Deterministic variational assimilation methods

In this Section, we consider a different approach for data assimilation, based on a deterministic approach. We do not necessarily rely upon *a priori* statistical knowledge of the process and we formulate the problem as a minimization procedure, where the mathematical paradigm acts as a constraint. For instance, referring to Fig. 3, with an educated guess we can decide a functional form for **w** and then fit this form with the measures. In other words we find **w** belonging to some class of functions V such that

$$dist(\mathbf{z}, \mathsf{Observation}(\mathbf{w})) \leq dist(\mathbf{z}, \mathsf{Observation}(\mathbf{v}))$$

for all $\mathbf{v} \in V$, where $\mathsf{Observation}(\mathbf{v})$ is the application of the mathematical representation of the measure process to **v**, to be compared with the "real" measure **z**. If the measures are trustworthy, we could derive a model for capturing exactly the data pursuing an interpolation approach. In general measures are noisy and we resort to a Least Squares (LS) procedure, so that the model fits the data in a "weaker" sense. Notice that the definition of the distance is somehow arbitrary. For instance it could include an *a priori* knowledge of the location of data more trustworthy than others by means of proper coefficients that give more relevance to these data.

In the general case of interest for our applications when we have a dynamical system evolving, we can recast the assimilation procedure as a *control problem*. In a very abstract setting, we may list the ingredients of this approach as follows (see Fig. 6).

A mathematical model describing the dynamics of interest for the variables or physical quantities describing the state of the system we are interested in. In our problems, this model or paradigm is given by a system of partial differential equations and, more precisely, a model describing fluid-structure interaction. In this case, the state variables are represented by the velocity, the pressure of the fluid and the displacement of the structure.

A set of observations or measurements of the state variables or, more in general, of a function of the state.

A functional $\mathcal{J}$ to be minimized. In general, this is the discrepancy between the results obtained by the mathematical (numerical) model and the available data.

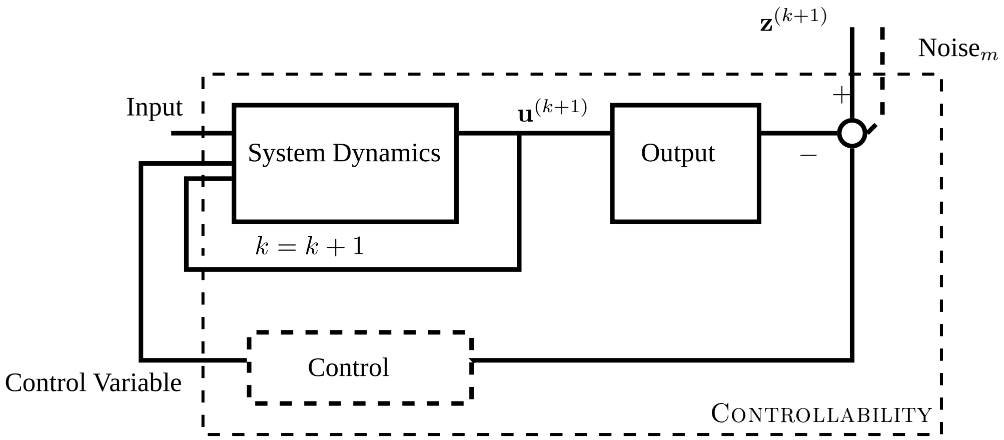


Figure 6: Possible approaches for the estimate with a system dynamics: here we do not use stochastic knowledge and we refer to the concept of CONTROLLABILITY: the estimate is reformulated as a control process.

A control variable (CV), which is the variable that we tune to get the minimization done. Its choice strongly depends on the purpose of the assimilation. For instance, in identification (parameter estimation) problems, the parameter(s) to be identified will be the control variable(s) to drive the minimization.

Solution of constrained minimization problems with distributed models (partial differential equations) acting as constraint has been considered by several authors [34, 69, 66, 2]. With no claim to be exhaustive, in the present Section we provide some general solution methods with simple examples, that have been used in applications of interest for biomedical fluid-structure interaction problems. Since minimization procedures resort typically to iterative methods, the solution of the system of partial differential equation representing the model needs typically to be solved several times. This rapidly leads to high computational costs, in particular when working on unsteady problems, as the ones we are interested in. We need to address therefore the problem of reducing the computational costs.

The key concept in this case is *controllability* - which is the dual concept of observability advocated in the previous Section - in other terms the effectiveness of the control strongly depends on the *sensitivity* of the functional to be minimized to the control variable. More the functional is sensitive to the control and most likely the minimization will be successful. This is somehow a change in the usual perspective of solving problems in engineering. As a matter of fact, *high sensitivity* comes from a lack of stability or robustness and the control action is intended to recover these properties. For instance, in the case of fluids we should expect a control to be more effective when the Reynolds number is high, because in this case, in general, the variable of interests are more sensible to perturbations and for this reasons they may be controlled.

The entire Section is largely based on [37], Chapters 2 and 5.

3.1 Least squares estimators

As we have done in the previous Section, we start with some considerations on the “steady” case, when we perform an In-Out constrained minimization (Fig. 3).

Suppose that we have a sequence of measures of the same variable $\mathbf{w} \in \mathbb{R}^n$ affected by noise,

$$\mathbf{z}_i = \mathrm{H}_i \mathbf{w} + \text{noise}_i, \quad i = 1, 2, \dots m.$$

We do not postulate any *a priori* probabilistic knowledge of the noise. The problem of estimating $\mathbf{w}$ from these measures has a classical deterministic formulation given by the Least-Squares (LS) approach. More precisely, the problem is formulated as: find the optimal $\mathbf{w}$ such that

$$\mathbf{w} = \arg\min \mathcal{J},$$

where

$$\mathcal{J} = \frac{1}{2}\sum_{i=1}^{m} (\mathbf{z}_i - \mathrm{H}_i \mathbf{w})^T \Omega_m^{-1} (\mathbf{z}_i - \mathrm{H}_i \mathbf{w})$$

and Ω_m^{-1} is a $n \times n$ weight matrix, which is assumed to be s.p.d. Let

$$\widehat{\mathrm{H}}_m = \begin{bmatrix} \mathrm{H}_1 \\ \mathrm{H}_2 \\ \dots \\ \mathrm{H}_m \end{bmatrix} \in \mathbb{R}^{nm,n}, \quad \widehat{\mathbf{z}} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \dots \\ \mathbf{z}_m \end{bmatrix} \in \mathbb{R}^{nm}, \quad \widehat{\Omega}_m = \begin{bmatrix} \Omega_m & & \mathrm{O} \\ & \ddots & \\ \mathrm{O} & & \Omega_m \end{bmatrix}$$

then,

$$\mathcal{J} = \frac{1}{2}(\widehat{\mathbf{z}} - \widehat{\mathrm{H}}_m \mathbf{w})^T \widehat{\Omega}_m^{-1} (\widehat{\mathbf{z}} - \widehat{\mathrm{H}}_m \mathbf{w})$$

Solving

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}} = 0$$

we find

$$\widehat{\mathrm{H}}_m^T \widehat{\Omega}_m^{-1} \left(\widehat{\mathbf{z}} - \widehat{\mathrm{H}}_m \mathbf{w}_{LS} \right) = 0.$$

Thus,

$$\mathbf{w}_{LS} = \left(\widehat{\mathrm{H}}_m^T \widehat{\Omega}_m^{-1} \widehat{\mathrm{H}}_m \right)^{-1} \widehat{\mathrm{H}}_m^T \widehat{\Omega}_m^{-1} \widehat{\mathbf{z}} = \Lambda_m \widehat{\mathrm{H}}_m^T \widehat{\Omega}_m^{-1} \widehat{\mathbf{z}}.$$

where $\Lambda_m = (\widehat{\mathrm{H}}_m^T \Omega_m^{-1} \widehat{\mathrm{H}}_m)^{-1}$.

Remark 3.1 *Let us consider a recursive formulation of this problem, obtained adjusting an available estimate (based on previous observations) when a new observation becomes available. Let us write recursively for $k > 1$*

$$\widehat{\mathrm{H}}_k = \begin{bmatrix} \widehat{\mathrm{H}}_{k-1} \\ \mathrm{H}_k \end{bmatrix}, \qquad \widehat{\mathbf{z}}_k = \begin{bmatrix} \widehat{\mathbf{z}}_{k-1} \\ \mathbf{z}_k \end{bmatrix}.$$

We also write

$$\widehat{\Omega_k}^{-1} = \begin{bmatrix} \widehat{\Omega}_{k-1}^{-1} & 0 \\ 0 & \Omega_k^{-1} \end{bmatrix},$$

where Ω_k^{-1} is a s.p.d. matrix. With this notation, we may write

$$\widehat{\mathrm{H}}_k^T \widehat{\Omega}_k^{-1} \widehat{\mathrm{H}}_k = \widehat{\mathrm{H}}_{k-1}^T \widehat{\Omega}_{k-1}^{-1} \widehat{\mathrm{H}}_{k-1} + \mathrm{H}_k^T \Omega_k^{-1} \mathrm{H}_k.$$

or, with the (suggestive) notation introduced above

$$\Lambda_k^{-1} = \Lambda_{k-1}^{-1} + \mathrm{H}_k^T \Omega_k^{-1} \mathrm{H}_k.$$

By the Sherman-Morrison-Woodbury formula we obtain

$$\Lambda_k = \left(\Lambda_{k-1}^{-1} + \mathrm{H}_k^T \Omega_k^{-1} \mathrm{H}_k\right)^{-1} = \Lambda_{k-1} - \Lambda_{k-1} \mathrm{H}_k^T \left(\Omega_k + \mathrm{H}_k \Lambda_{k-1} \mathrm{H}_k^T\right)^{-1} \mathrm{H}_k \Lambda_{k-1}.$$

Let us introduce the matrix

$$\mathrm{G}_k = \Lambda_{k-1} \mathrm{H}_k \left(\Omega_k + \mathrm{H}_k \Lambda_{k-1} \mathrm{H}_k^T\right)^{-1},$$

so we have

$$\Lambda_k = \left(\mathrm{I} - \mathrm{G}_k \mathrm{H}_k\right) \Lambda_{k-1}.$$

From here, we can obtain the recursive formula (see [68], Section 4.3)

$$\widehat{\mathbf{w}}_k^{LS} = \widehat{\mathbf{w}}_{k-1}^{LS} + \mathrm{G}_k \left(\mathbf{z}_k - \mathrm{H}_k \widehat{\mathbf{w}}_{k-1}^{LS}\right),$$

that has a formal analogy with Kalman filter formulas, even though in this case the dynamics is not related to the state (no dynamics occurs on $\mathbf{w}$*) but just to the addition of new measures. In this respect, $\mathbf{z}_k - \mathrm{H}_k \widehat{\mathbf{w}}_{k-1}^{LS}$ represents the net content of new information brought by the new measure.*

When we assume that the state evolves, the mathematical equation describing the dynamics may be used as a constraint to the minimization process. Algebraic aspects of constrained LS problems, with both equality and inequality constraints, have been addressed in [35] Chapter 12. For instance, we may consider the problem: find $\mathbf{x}$ such that

$$\mathbf{x} = \arg\min \|\mathrm{A}\mathbf{x} - \mathbf{b}\|_2, \quad \mathrm{B}\mathbf{x} = \mathbf{d}$$

where A is a $m \times n$ matrix, B is $p \times n$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{d} \in \mathbb{R}^b$. We assume that the matrices are both full-rank. The problem can be solved by an application of the Generalized Singular Value Decomposition (GSVD),

$$\mathrm{U}^T \mathrm{A} \mathrm{X} = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n) = \mathrm{D}_A, \quad \mathrm{V}^T \mathrm{B} \mathrm{X} = \mathrm{diag}(\beta_1, \beta_2, \ldots, \beta_p) = \mathrm{D}_B,$$

with U and V orthogonal matrices and $\mathbf{x}_i$ are the columns of X for $i = 1, 2, \ldots, n$. The solution to this problem then reads [35]

$$\mathbf{x} = \sum_{i=1}^{p} \frac{\mathbf{v}_i^T \mathbf{d}}{\beta_i} \mathbf{x}_i + \sum_{i=p+1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\alpha_i} \mathbf{x}_i.$$

We can consider the associated unconstrained LS problem

$$\mathbf{x} = \arg\min \| \begin{bmatrix} \mathrm{A} \\ \lambda \mathrm{B} \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \lambda \mathbf{d} \end{bmatrix} \|_2$$

This can be solved with an ordinary LS procedure. Using the GSVD decomposition, it is possible to find the solution

$$\mathbf{x}(\lambda) = \sum_{i=1}^{p} \frac{\alpha_i \mathbf{u}_i^T \mathbf{b} + \lambda^2 \beta_i^2 \mathbf{v}_i^T \mathbf{d}}{\alpha_i^2 + \lambda^2 \beta_i^2} \mathbf{x}_i + \sum_{i=p+1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\alpha_i} \mathbf{x}_i,$$

from which it is promptly realized that the solution to the constrained minimization problem $\mathbf{x} = \lim_{\lambda \to \infty} \mathbf{x}(\lambda)$. In the next Section, we see a similar approach for solving unconstrained minimization when the constraint is represented by a partial differential equation.

3.2 Constrained minimization problems with PDEs: a simple working example

To be concrete, we illustrate techniques of constrained minimization with partial differential equations on the following problem. Let $\Omega \subset \mathbb{R}^n$ $(n = 2, 3)$ and $\partial\Omega$ be denoted by Γ. We assume that $u(\mathbf{x})$ is the state variable that obeys the following equation

$$-\Delta u + \mathbf{b} \cdot \nabla u + \gamma u + u^3 = \sum_{i=1}^{K} \alpha_i f_i, \quad \text{in } \Omega \tag{30}$$

γ and α_i $(i = 1, \ldots K)$ are real coefficient. $\boldsymbol{\alpha} \in \mathbb{R}^K$ is the vector with entries α_i. $\mathbf{b}$ is a divergence free vector function. In our applications, $\mathbf{b}$ may represent the blood flow, and u the concentration of some solute in the blood. We assume to have a *reference* or desired function $d(\mathbf{x})$ we would like to be approximated by $u(\mathbf{x})$. We assume moreover

$$u = 0 \quad \text{on } \partial\Omega. \tag{31}$$

We assume that γ and α_i are unknown parameters. They need to be computed so to drive the state variable to the reference behavior. We formulate therefore the problem[5]: find $\gamma, \boldsymbol{\alpha}$ to minimize

$$\mathcal{J}(u) = \frac{1}{2} \int_{\Omega} (u(\mathbf{x}) - d(\mathbf{x}))^2 d\mathbf{x},$$

[5] A similar problem has been investigated as a simplified model of superconductivity in [69]

where $u(\mathbf{x})$ solves (30), (31).

Quite often $\mathcal{J}$ is added with a term depending explicitly on the parameters to be estimated in the form

$$\mathcal{J}_R(u, \boldsymbol{\alpha}, \gamma) = \mathcal{J}(u) + \frac{\sigma_1}{2}\|\boldsymbol{\alpha}\|^2 + \frac{\sigma_2}{2}\|\gamma - \gamma_{ref}\|^2, \tag{32}$$

where σ_1 and σ_2 are constants and $\|\cdot\|$ denotes a generic (convenient) norm; in the remainder of the Section we assume $\|\cdot\| = \|\cdot\|_2$. This modification may have both practical and theoretical reasons.

1. First practical motivation: when the control variable corresponds to a physical control, like the coefficients $\boldsymbol{\alpha}$, it implies a practical cost (intended in a broad sense as the energy required to apply it). For this reason, the "size" of the control cannot be too large. The correction of $\mathcal{J}$ with $\frac{\sigma_1}{2}\|\boldsymbol{\alpha}\|^2$ is a "penalization" that includes the cost of the control[6].
2. Second practical motivation: in some cases, in particular when identifying a parameter, a "nominal" reference guess is available, based for instance on averaging available measures or samples. This is denoted here by γ_{ref} and the real value is supposed to be "not too far" from this value. This leads to the term $\frac{\sigma_2}{2}\|\gamma - \gamma_{ref}\|^2$ that penalizes the difference respect to the nominal value.
3. Mathematical motivation: If we hypothetically consider only the terms $\mathcal{J}_R(u, d) = \frac{\sigma_1}{2}\|\boldsymbol{\alpha}\|^2 + \frac{\sigma_2}{2}\|\gamma - \gamma_{ref}\|^2$, the function to be minimized has excellent mathematical properties. It is actually quadratic and the minimization leads clearly to the solution $(\boldsymbol{\alpha}, \gamma) = (\mathbf{0}, \gamma_{ref})$. We infer therefore that the term $\mathcal{J}_R$ has a regularizing effect on the minimization properties, balancing the bad (or not so good) properties of the original constrained minimization. As a matter of fact, in general the original problem may be ill-posed, featuring multiple local minima or none. The term $\mathcal{J}_R$ with a proper selection of the weights σ_1 and σ_2 allows us in general to have a well-posed problem. For this reason, when solving this kind of *inverse problems*, this term is often called *regularization* (*Tikhonov regularization* in the form in (32)). Other forms of regularization may be considered in practice, but they will not be addressed here (see [23, 38]).

The appropriate selection of wieghts $\sigma_{1,2}$ is not trivial. It is actually a trade-off between the minimization of the mismatch (that requires these weights to be small) and the regularization of the problem (that in general is improved for large positive values). Different strategies are possible. A general approach is to identify values such that the impact of additional terms on the non-regularized functional is bounded by the numerical errors, so to reduce the effects on mismatch minimization within the range acceptable after approximations, while improving the conditioning properties of the problem. See e.g. [67, 72].

[6] This could be done also with unilateral constraints $\|\boldsymbol{\alpha}\| \leq$ max-cost-allowed.

3.2.1 Gâteaux and Fréchet derivatives

For the solution of a PDE constrained optimization problem, we need to be able to differentiate operators acting between functional spaces. In particular, let $\mathcal{F} : X \to Y$, X and Y being appropriate functional spaces and let $u,\, v \in X$. The derivative of $\mathcal{F}$, in the direction v can be computed as

$$D\,\mathcal{F}(u;\, v) := \lim_{\varepsilon \to 0} \frac{\mathcal{F}(u+\varepsilon v) - \mathcal{F}(u)}{\varepsilon}.$$

Such derivative is called *Gâteaux* derivative. As an example, take $\mathcal{G}(u) = (u - f)^2$, then

$$D\,\mathcal{G}(u;\, v) = \lim_{\varepsilon \to 0} \frac{(u+\varepsilon v - f)^2 - (u-f)^2}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{2\varepsilon (u-f)v + \varepsilon^2 v^2}{\varepsilon} = 2\,(u{-}f)\,v.$$

It is often possible to write the Gâteaux derivative of $\mathcal{F}$ in any direction v, as the application of a bounded linear operator $\left.\dfrac{D\mathcal{F}}{Du}\right|_u$ to v. Such operator is called *Fréchet* derivative. In the following we assume that the Fréchet derivative exists and we write

$$D\,\mathcal{F}(u;\, v) = \left.\frac{D\mathcal{F}}{Du}\right|_u (v).$$

In our example, $\left.\dfrac{D\mathcal{F}}{Du}\right|_u = 2(u-f)$. It is possible to show that the Gâteaux derivative of $\mathcal{G}(\boldsymbol{u}) = \boldsymbol{u}^T A\,\boldsymbol{u}$, in the direction $\boldsymbol{v}$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are vector functions and A is a constant square matrix of compatible dimensions, reads

$$D\,\mathcal{G}(\boldsymbol{u},\, \boldsymbol{v}) = \boldsymbol{u}^T A\,\boldsymbol{v} + \boldsymbol{v}^T A\,\boldsymbol{u} = \boldsymbol{u}^T (A + A^T)\boldsymbol{v},$$

while its Fréchet derivative reads

$$\left.\frac{D\mathcal{G}}{D\boldsymbol{u}}\right|_{\boldsymbol{u}} = \boldsymbol{u}^T (A + A^T),$$

with the understanding, in this case, that the application of the operator $\left.\dfrac{D\mathcal{G}}{D\boldsymbol{u}}\right|_{\boldsymbol{u}}$ to $\boldsymbol{v}$ is the usual matrix vector product of the one-row matrix $\left.\dfrac{D\mathcal{G}}{D\boldsymbol{u}}\right|_{\boldsymbol{u}}$ and the vector $\boldsymbol{v}$. As another example, consider $\mathcal{G}(u) = -\Delta u$, then $D\mathcal{G}(u,\, v) = -\Delta v$ and $\left.\dfrac{D\,\mathcal{G}}{Du}\right|_u = -\Delta$. In general, the derivative of a linear operator (the Laplacian operator in this case) is the linear operator itself.

The usual chain rule holds for the differentiation of composite functions

$$D\,(\mathcal{G} \circ \mathcal{F})(u,\, v) = D\,\mathcal{G}\,(\mathcal{F}(u),\, D\mathcal{F}(u, v)),$$

or

$$\left.\frac{D\,(\mathcal{G} \circ \mathcal{F})}{Du}\right|_u = \left.\frac{D\,(\mathcal{G} \circ \mathcal{F})}{D\mathcal{F}}\right|_{\mathcal{F}(u)} \left(\left.\frac{D\,\mathcal{F}}{Du}\right|_u \right).$$

3.2.2 Gradient-based optimization approaches

A common and effective approach to deal with optimization constrained by partial differential equations, is to include directly the constraint in the functional to be minimized. In this way, the minimization procedure is recast in an unconstrained case and the solution is obtained with classical arguments. In particular, the first order necessary conditions are obtained by setting to 0 the gradient of the functional.

In our case, this means that the solution u is computed as a function of the control variables $\boldsymbol{\alpha}$ and γ and the total derivative of $\mathcal{J}_R$, regarded as function of these variables, is set to 0. This procedure admits an iterative implementation. Let us denote the state problem (30), (31) with the abstract notation $\mathcal{F}(u, \boldsymbol{\alpha}, \gamma) = 0$.

Assume that an initial guess $\boldsymbol{\alpha}^{(0)}$ and $\gamma^{(0)}$ is given. Typically, we take $\gamma^{(0)} = \gamma_{ref}$. Then, we perform the following steps for $j = 0, 1, 2, \ldots$:

- find the state variable $u^{(j)}$ solution to $\mathcal{F}(u, \boldsymbol{\alpha}^{(j)}, \gamma^{(j)}) = 0$;
- compute $D\mathcal{J}_R(u^{(j)}, \boldsymbol{\alpha}^{(j)}, \gamma^{(j)})/D\boldsymbol{\alpha}\big|_{\boldsymbol{\alpha}^{(j)}}$ and $D\mathcal{J}_R u^{(j)}, \boldsymbol{\alpha}^{(j)}, \gamma^{(j)}/D\gamma\big|_{\gamma^{(j)}}$;
- **if** $\|D\mathcal{J}_R(u^{(j)}, \boldsymbol{\alpha}^{(j)}, \gamma^{(j)})/D[\boldsymbol{\alpha}, \gamma]\|$ is sufficiently small, solution is reached; **else** compute a new guess $\boldsymbol{\alpha}^{(j+1)}$, $\gamma^{(j+1)}$, for instance by setting

$$\begin{aligned} \alpha_i^{(j+1)} &= \alpha_i^{(j)} - \omega_i^{(j)} \frac{D\mathcal{J}_R(u^{(j)}, \boldsymbol{\alpha}^{(j)}, \gamma^{(j)})}{D\alpha_i}, \quad i = 1, 2, \ldots K \\ \gamma^{(j+1)} &= \gamma^{(j)} - \omega_{K+1}^{(j)} \frac{D\mathcal{J}_R(u^{(j)}, \boldsymbol{\alpha}^{(j)}, \gamma^{(j)})}{D\gamma}, \end{aligned} \tag{33}$$

  where $\omega_i^{(j)}$, $i = 1, 2, \ldots, K+1$ are numerical coefficients that drive the convergence of the procedure.

This approach, based on (33), belongs to the family of *steepest descent methods* and the parameters ω_i define the step performed in updating the solution along the line identified by the gradient. These coefficients, in general, may be dynamically determined at each iteration. Other iterative methods may be considered for the sake of effectiveness. Among the others, a method that usually outperforms the steepest descent approach is the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) method (see e. g. [58]); another common choice is the Gauss-Newton method. The latter finds the roots of $D\mathcal{J}_R/D[\boldsymbol{\alpha}, \gamma] = 0$ using the Newton method, that means that, at each iteration j, the minimization of the paraboloid tangent to $\mathcal{J}_R$ in $\boldsymbol{\alpha}^{(j)}, \gamma^{(j)}$ is performed. The method is potentially second order, but it has the drawback that the Hessian of the functional $\mathcal{J}_R$ is needed.

The most troublesome step in the previous algorithm is the computation of the gradients $D\mathcal{J}_R(u^{(j)}, \boldsymbol{\alpha}^{(j)}, \gamma^{(j)})/D[\boldsymbol{\alpha}, \gamma]$. Let us address two possible methods.

3.2.3 Gradient computation through sensitivities

A possible way for computing the gradients relies upon the chain rule

$$\left.\frac{D\mathcal{J}_R}{D\alpha_i}\right|_{\alpha^{(j)}} = \left.\frac{\partial \mathcal{J}_R}{\partial u}\right|_{u^{(j)}} \left(\left.\frac{\partial u}{\partial \alpha_i}\right|_{\alpha_i^{(j)}} \right) + \left.\frac{\partial \mathcal{J}_R}{\partial \alpha_i}\right|_{\alpha_i^{(j)}}, \quad i = 1, 2, \dots, K+1$$

where for easiness of notation we set $\alpha_{K+1} = \gamma$. We call *sensitivities* the derivatives

$$\phi_i \equiv \frac{\partial u}{\partial \alpha_i}, \quad \forall i = 1, 2, \dots K+1$$

as they quantify the sensitivity of the solution to each control variable. From

$$\mathcal{F}(u^{(j)}, \boldsymbol{\alpha}^{(j)}) = 0 \Rightarrow \left.\frac{D\mathcal{F}}{D\alpha_i}\right|^{(j)} = \left.\frac{\partial \mathcal{F}}{\partial u}\right|_{u^{(j)}} \left(\phi_i^{(j)}\right) + \left.\frac{\partial \mathcal{F}}{\partial \alpha_i}\right|^{(j)} = 0,$$

we have

$$\left.\frac{\partial \mathcal{F}}{\partial u}\right|_{u^{(j)}} \left(\phi_i^{(j)}\right) = - \left.\frac{\partial \mathcal{F}}{\partial \alpha_i}\right|^{(j)}. \tag{34}$$

Sensitivities can be retrieved by solving this set of equations for $i = 1, 2, \dots, K+1$.

In particular, for our working example we have

$$\left.\frac{D\mathcal{J}_R}{D\alpha_i}\right|_{\alpha_i^{(j)}} = \left.\frac{\partial \mathcal{J}_R}{\partial u}\right|_{u^{(j)}} \left(\phi_i^{(j)}\right) + \frac{\partial \mathcal{J}_R}{\partial \alpha_i}^{(j)} = \int_\Omega (u-d)\phi_i + \sigma_1 \alpha_i, \quad i = 1, 2, \dots K$$

$$\frac{D\mathcal{J}_R}{D\gamma}^{(j)} = \left.\frac{\partial \mathcal{J}_R}{\partial u}\right|_{u^{(j)}} \left(\phi_{K+1}^{(j)}\right) + \left.\frac{\partial \mathcal{J}_R}{\partial \alpha_i}\right|_{\alpha_{K+1}^{(j)}} = \int_\Omega (u-d)\left(\phi_{K+1}\right) + \sigma_2(\gamma - \gamma_{ref}).$$

Notice that from the state equations (30), (31), we have for $i = 1, 2, \dots K+1$

$$\left.\frac{\partial \mathcal{F}}{\partial u}\right|_{u^{(j)}} (\phi_i) = -\Delta\phi_i + \mathbf{b} \cdot \nabla\phi_i + \gamma\phi_i + 3(u^{(j)})^2\phi_i,$$

and

$$\left.\frac{\partial \mathcal{F}}{\partial \alpha_i}\right|^{(j)} = -f_i, \;\; i = 1, 2, \dots K$$

$$\left.\frac{\partial \mathcal{F}}{\partial \gamma}\right|^{(j)} = u^{(j)}.$$

Then, the *sensitivities equations* read

$$\begin{cases} -\Delta\phi_i + \mathbf{b} \cdot \nabla\phi_i + \gamma\phi_i + 3(u^{(j)})^2\phi_i = f_i & \text{in } \Omega \\ -\Delta\phi_{K+1} + \mathbf{b} \cdot \nabla\phi_{K+1} + \gamma\phi_{K+1} + 3(u^{(j)})^2\phi_{K+1} = -u^{(j)} & \text{in } \Omega \\ \phi_i = 0, \;\; i = 1, 2, \dots K+1 & \text{on } \partial\Omega. \end{cases} \tag{35}$$

Notice that these equations are linear in the sensitivities. Finally, we have

$$\frac{D\mathcal{J}_R}{D\alpha_i}\bigg|^{(j)} = \int\limits_\Omega (u^{(j)} - d)\phi_i + \sigma_1\alpha_i^{(j)}, \quad \frac{D\mathcal{J}_R}{D\gamma}\bigg|^{(j)} = \int\limits_\Omega (u^{(j)} - d)\phi_{K+1} + \sigma_2\gamma^{(j)}.$$

Gradients of the functional with respect to the control variables following this approach requires therefore the solution of the $K + 1$ sensitivity equations.

3.2.4 Gradient computation through adjoint equations

In the following, we omit the iteration index j for simplicity. In the previous Section we computed the operator $\left.\frac{\partial\mathcal{F}}{\partial u}\right|_u$ applied to the sensitivities ϕ_i. Let us consider the *adjoint* of this operator, which is the operator $\left.\left(\frac{\partial\mathcal{F}}{\partial u}\right)^*\right|_u$ such that

$$< \left.\left(\frac{\partial\mathcal{F}}{\partial u}\right)^*\right|_u (\rho)\, ,\, v > = < \rho\, ,\, \left.\frac{\partial\mathcal{F}}{\partial u}\right|_u (v) >, \tag{36}$$

for any v belonging to an appropriate functional space. Here $< \cdot, \cdot >$ indicates a duality pairing. In particular, in a finite dimensional setting, $< \cdot, \cdot >$ typically denotes the usual Euclidean dot product, while in the continuous setting, it denotes one of the integrals

$$\begin{cases} < u, v > \equiv \int\limits_\Omega u\, v, & \text{for scalar functions,} \\ < \mathbf{u}, \mathbf{v} > \equiv \int\limits_\Omega \mathbf{u} \cdot \mathbf{v}, & \text{for vector functions,} \\ < U, V > \equiv \int\limits_\Omega U : V & \text{for tensor functions,} \end{cases} .$$

In our example, we have

$$< \rho, \left.\frac{\partial\mathcal{F}}{\partial u}\right|_u (v) > = \int\limits_\Omega \rho \left(-\Delta v + \mathbf{b} \cdot \nabla v + \gamma v + 3u^2 v\right).$$

Integrating by parts, and choosing ρ to vanish on Γ, we get[7]

$$< \rho, \left.\frac{\partial\mathcal{F}}{\partial u}\right|_u (v) > = \int\limits_\Omega \left(-\Delta\rho - \mathbf{b} \cdot \nabla\rho + \gamma\rho + 3u^2\rho\right) v.$$

Therefore, the adjoint operator reads

$$\left.\left(\frac{\partial\mathcal{F}}{\partial u}\right)^*\right|_u = -\Delta\rho - \mathbf{b} \cdot \nabla\rho + \gamma\rho + 3u^2\rho.$$

[7] We remind that we assumed $\mathbf{b}$ to be divergence free.

We consider the following adjoint problem, whose solution, as we will see later, is crucial to find the derivatives of $\mathcal{J}$ with respect to the parameters.

$$< \left(\frac{\partial \mathcal{F}}{\partial u}\right)^*\bigg|_u (\rho)\,,\, v >= \frac{\partial \mathcal{J}_R}{\partial u}\bigg|_u (v), \tag{37}$$

for any v belonging to appropriate functional spaces. In our specific example, this problem reads

$$\int\limits_\Omega \left(-\Delta\rho - \mathbf{b}\cdot\nabla\rho + \gamma\rho + 3u^2\rho\right) v = \int\limits_\Omega (u-d)\, v,$$

for any v, with ρ vanishing on Γ. Since such equation must hold for any v, we get the strong form of the adjoint equation

$$\begin{cases} -\Delta\rho - \mathbf{b}\cdot\nabla\rho + \gamma\rho + 3u^2\rho = u - d & \text{in } \Omega \\ \rho = 0 & \text{on } \Gamma \end{cases}.$$

Notice that once ρ is computed by solving this equation, we may write for $i = 1, 2, \dots K+1$

$$\begin{aligned} \frac{D\mathcal{J}_R}{D\alpha_i} =& \frac{\partial \mathcal{J}_R}{\partial u}\left(\frac{\partial u}{\partial \alpha_i}\right) + \frac{\partial \mathcal{J}_R}{\partial \alpha_i} =< \left(\frac{\partial \mathcal{F}}{\partial u}\right)^*\bigg|_u (\rho)\,,\, \frac{\partial u}{\partial \alpha_i} > + \frac{\partial \mathcal{J}_R}{\partial \alpha_i} = \\ & < \rho\,,\, \frac{\partial \mathcal{F}}{\partial u}\bigg|_u \left(\frac{\partial u}{\partial \alpha_i}\right) > + \frac{\partial \mathcal{J}_R}{\partial \alpha_i} = - < \rho\,,\, \frac{\partial \mathcal{F}}{\partial \alpha_i} > + \frac{\partial \mathcal{J}_R}{\partial \alpha_i}, \end{aligned} \tag{38}$$

where we exploit (37) (36) and (34). In other words, all the gradients needed by the iterative procedure are promptly computed after ρ is calculated. In the example, this reads for $i = 1, 2 \dots K$

$$\begin{aligned} \frac{D\mathcal{J}_R}{D\alpha_i} &= \sigma_1\alpha_i + \int\limits_\Omega (u-d)\frac{\partial u}{\partial \alpha_i} = \sigma_1\alpha_i + \int\limits_\Omega (u-d)\phi_i \\ &= \sigma_1\alpha_i + \int\limits_\Omega \left(-\Delta\rho - \mathbf{b}\cdot\nabla\rho + \gamma\rho + 3u^2\rho\right)\phi_i \\ &\underset{\text{(by parts)}}{=} \sigma_1\alpha_i + \int\limits_\Omega \left(-\Delta\phi_i + \mathbf{b}\cdot\nabla\phi_i + \gamma\phi_i + 3u^2\phi_i\right)\rho \\ &= \sigma_1\alpha_i + \int\limits_\Omega f_i\rho, \end{aligned}$$

and similarly we obtain $\dfrac{D\mathcal{J}_R}{D\gamma} = \sigma_1(\gamma - \gamma_{ref}) - \displaystyle\int\limits_\Omega u\rho.$

According to this procedure, it is enough to solve a differential problem in the adjoint operator to compute all the gradients needed by the iterative procedure. This approach is therefore more efficient, when it is possible (and doable) the computation of the adjoint operator.

3.2.5 The Lagrange Multiplier approach and the KKT conditions

Let us consider a different approach for reformulating the constrained minimization problem into an unconstrained one. It is a classical argument in which a companion functional is introduced to include the constraints. We stick to our simple working example to introduce the idea, referring to the mentioned references for a more complete presentation. Let us consider the functional

$$\mathcal{L}(u, \boldsymbol{\alpha}, \gamma, \chi) = \mathcal{J}_R(u, \boldsymbol{\alpha}, \gamma) - < \chi, \, \mathcal{F}(u, \boldsymbol{\alpha}, \gamma) >,$$

where χ is the adjoint (or co-state) function , the so-called *Lagrange multiplier*. The idea behind this approach is that *solutions of the constrained minimization problem are stationary points of* $\mathcal{L}$. As such they solve the following system of equations, representing the *first order necessary conditions* of optimality

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \chi} = 0 & \text{State equations} \\ \dfrac{\partial \mathcal{L}}{\partial u} = 0 & \text{Adjoint/Co} - \text{state equations} \\ \dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = 0 & \text{Optimality conditions} \\ \dfrac{\partial \mathcal{L}}{\partial \gamma} = 0 & \text{Optimality condition.} \end{cases} \tag{39}$$

Here, each variable is independent of the others since no constraint holds. In our specific example, we have[8]

$$\mathcal{L}(u, \boldsymbol{\alpha}, \gamma, \chi_1, \chi_2) = \mathcal{J}_R(u, \boldsymbol{\alpha}, \gamma) - \int_\Omega \chi_1 \left(-\Delta u + \mathbf{b} \cdot \nabla u + \gamma u + u^3 - \sum_{i=1}^{K} \alpha_i f_i \right) - \int_\Gamma \chi_2 u.$$

Here, we considered the integral formulation of (30,31), where χ_1 and χ_2 are the functions enforcing the constraint given by the state equation. To obtain the stationary points, we need to perform the Gateaux differentiation

$$\left.\frac{\partial \mathcal{L}}{\partial \chi_1}\right|_{\chi_1} (\delta_{\chi_1}) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\mathcal{L}(u, \boldsymbol{\alpha}, \gamma, \chi_1 + \varepsilon \delta_{\chi_1}, \chi_2) - \mathcal{L}(u, \boldsymbol{\alpha}, \gamma, \chi_1, \chi_2) \right) \tag{40}$$

where δ_{χ_1} is an admissible variation. We find

$$\int_\Omega \delta_{\chi_1} \left(-\Delta u + \mathbf{b} \cdot \nabla u + \gamma u + u^3 - \sum_{i=1}^{K} \alpha_i f_i \right) = 0.$$

[8] Here we used the Lagrange multiplier χ_2 to prescribe the Dirichlet homogeneous boundary condition. Often, such condition is prescribed without using Lagrange multipliers but requiring directly that u and χ_1 vanish on the boundary.

Since δ_{χ_1} is arbitrary, from this equation we promptly obtain the state problem (30). Similarly,

$$\left.\frac{\partial \mathcal{L}}{\partial \chi_2}\right|_{\chi_2}(\delta_{\chi_2}) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(\mathcal{L}(u, \boldsymbol{\alpha}, \gamma, \chi_1, \chi_2 + \varepsilon\delta_{\chi_2}) - \mathcal{L}(u, \boldsymbol{\alpha}, \gamma, \chi_1, \chi_2)\right) = \int\limits_{\Gamma} \delta_{\chi_2} u = 0 \tag{41}$$

leading to (31).

Let us write explicitly now the adjoint equation

$$\begin{gathered}\left.\frac{\partial \mathcal{L}}{\partial u}\right|_{u}(\delta_u) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(\mathcal{L}(u + \varepsilon\delta_u, \boldsymbol{\alpha}, \gamma, \chi_1, \chi_2) - \mathcal{L}(u, \boldsymbol{\alpha}, \gamma, \chi_1, \chi_2)\right) = \\ \int\limits_{\Omega} (u - d)\, \delta_u - \int\limits_{\Omega} \chi_1 \left(-\Delta\delta_u + \mathbf{b} \cdot \nabla\delta_u + \gamma\delta_u + 3u^2\delta_u\right) - \int\limits_{\Gamma} \chi_2\delta_u = 0.\end{gathered} \tag{42}$$

Let us factor out the arbitrary variation δ_u. If we integrate by parts the second and first order terms, we get

$$\begin{gathered}\int\limits_{\Omega} \delta_u \left(u - d + \Delta\chi_1 + \mathbf{b} \cdot \nabla\chi_1 - \gamma\chi_1 - 3u^2\chi_1\right) + \\ \int\limits_{\partial\Omega} \chi_1 \nabla\delta_u \cdot \mathbf{n} - \int\limits_{\partial\Omega} (\nabla\chi_1 \cdot \mathbf{n} + \chi_1 \mathbf{b} \cdot \mathbf{n} + \chi_2)\delta_u = 0.\end{gathered}$$

Because δ_u is arbitrary, this equation is equivalent to

$$\begin{aligned} &-\Delta\chi_1 - \mathbf{b} \cdot \nabla\chi_1 + \gamma\chi_1 + 3u^2\chi_1 = u - d && \text{in } \Omega \\ &\chi_1 = 0 && \text{on } \partial\Omega \\ &\chi_2 = -\mathbf{b} \cdot \nabla\chi_1 \cdot \mathbf{n} - \chi_1 \mathbf{n} && \text{on } \partial\Omega. \end{aligned}$$

Notice that χ_2 does not affect the solution of the problem, therefore in the following we drop the last equation because we are not interested in the particular value assumed by χ_2. Finally, we compute the derivative with respect to the control variables.

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \int\limits_{\Omega} \chi_1 f_i + \sigma_1 \alpha_i, \quad \frac{\partial \mathcal{L}}{\partial \gamma} = \int\limits_{\Omega} u\chi_1 + \sigma_2(\gamma - \gamma_{ref}). \tag{43}$$

Summarizing, the optimality system to be solved reads

$$\begin{cases} -\Delta u + \mathbf{b} \cdot \nabla u + \gamma u + u^3 = \sum\limits_{i=1}^{K} \alpha_i f_i \quad \text{in } \Omega \\ u = 0 \quad \text{on } \partial\Omega \end{cases} \qquad \text{State equations}$$

$$\begin{cases} -\Delta \chi_1 - \mathbf{b} \cdot \nabla \chi_1 + \gamma \chi_1 + 3u^2 \chi_1 = u - d \text{ in } \Omega \\ \chi_1 = 0 \text{ on } \partial\Omega \end{cases} \qquad \text{Adjoint equations}$$

$$\begin{cases} \alpha_i = -\dfrac{1}{\sigma_1} \displaystyle\int\limits_{\Omega} \chi_1 f_i \quad i = 1, \dots, K \\ \gamma = \gamma_{ref} - \dfrac{1}{\sigma_2} \displaystyle\int\limits_{\Omega} \chi_1 u \end{cases} \qquad \text{Optimality conditions} \tag{44}$$

This set of equations represents the so-called *Karush-Khun-Tucker (KKT) conditions* [69].

In principle, this system provides the solution to the optimization problem in a monolithic or "one-shot" fashion. In practice, the cases of interest when the system can be solved directly are rare - in particular for nonlinear state problems, and we need again to resort to iterative procedures.

Let a guess for $\boldsymbol{\alpha}$ and γ be given at the iteration j. Again, typically, we take $\gamma^{(0)} = \gamma_{ref}$. A reasonable iterative procedure reads as follows.

1. Solve the state equations to compute $u^{(j+1)}$;
2. Solve the adjoint problem to compute $\chi_1^{(j+1)}$ and $\chi_2^{(j+1)}$.
3. Update the control variables using the optimality conditions. In this example it is natural to choose

$$\alpha_i^{(j+1)} = -\frac{1}{\sigma_1} \int\limits_{\Omega} \chi_1^{(j+1)} f_i \text{ and } \gamma^{(j+1)} = \gamma_{ref} - \frac{1}{\sigma_2} \int\limits_{\Omega} \chi_1^{(j+1)} u^{(j+1)},$$

until a convergence criterion is satisfied.

This procedure corresponds in fact to a fixed-step steepest descent method for $\mathcal{J}_R$ regarded as a function of the control variables. In fact, notice that the Lagrange multiplier χ introduced here corresponds to ρ introduced in the previous Section. With this perspective, equation (38) reads

$$\frac{D\mathcal{J}_R}{D\alpha} = 0$$

that is exactly what we want to obtain when we are looking for a minimum of $\mathcal{J}_R$. As a matter of fact, the iterative algorithm introduced in the previous Section to minimize $\mathcal{J}_{\mathcal{R}}$, is an iterative algorithm to solve the KKT conditions.

Sequential Quadratic Programming Algorithm In contrast with what is done in the unconstrained approach considered so far, constrained algorithms try to compute the solution to the minimization problem thorough the convergence of the state and parameters variables $(u^{(j)}, \alpha^{(j)})$ simultaneously. This approach can be very effective in presence of nonlinear constraints, as the constraints need not to be solved at each iteration. In this Section we consider one of these methods, the sequential quadratic programming (SQP) method [12] which consists of iteratively approximating the original problem with a quadratic problem subject to linear constraints. Such quadratic problem is then solved using quadratic programming (QP) algorithms. Assume that the problem is already discretized, and let the vector $\mathbf{x}^{(j)}$ include both the state $(\mathbf{u}^{(j)})$ and the parameter $(\boldsymbol{\alpha}^{(j)})$ vectors

$$\mathbf{x}^{(j)} = \begin{bmatrix} \mathbf{u}^{(j)} \\ \boldsymbol{\alpha}^{(j)} \end{bmatrix}.$$

The Lagrangian functional of the problem $\mathcal{L}(\mathbf{x}, \boldsymbol{\chi}) = \mathcal{J}_R(\mathbf{x}) - < \boldsymbol{\chi}, \mathcal{F}(\mathbf{x}) >$ is approximated at iteration j with the paraboloid tangent to the Lagrangian in $\mathbf{x}^{(j)}$, i.e.

$$\mathcal{L}\left(\mathbf{x}, \boldsymbol{\chi}^{(j)}\right) \approx \mathcal{L}(\mathbf{x}^{(j)}, \boldsymbol{\chi}^{(j)}) + L_{\mathbf{x}}^{(j),T} \boldsymbol{\delta}_x^{(j)} + \frac{1}{2} \boldsymbol{\delta}_x^{(j),T} \mathrm{H}^{(j)} \boldsymbol{\delta}_x^{(j)},$$

where $L_{\mathbf{x}}^{(j)} = \left.\dfrac{\partial \mathcal{L}}{\partial \mathbf{x}}\right|_{\mathbf{x}^{(j)}}$, $\boldsymbol{\delta}_x^{(j)} = \mathbf{x} - \mathbf{x}^{(j)}$, and $\mathrm{H}^{(j)} = \left.\dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2}\right|_{\mathbf{x}^{(j)}}$ is the *Hessian* matrix. Such approximation of the Lagrangian is minimized w.r.t. $\boldsymbol{\delta}_x^{(j)}$, subject to the linearization of the constraint $\mathcal{F}(\mathbf{x}) = 0$

$$\mathcal{F}\left(\mathbf{x}^{(j)}\right) + \mathrm{F}_{\mathbf{x}}^{(j),T} \boldsymbol{\delta}_x^{(j)} = 0. \tag{45}$$

where the matrix $\mathrm{F}_{\mathbf{x}}^{(j)} = \left.\dfrac{\partial \mathcal{F}}{\partial \mathbf{x}}\right|_{\mathbf{x}^{(j)}}$. Exploiting the fact that $\mathrm{F}_{\mathbf{x}}^{(j),T} \boldsymbol{\delta}_x^{(j)}$ is constant w.r.t. $\boldsymbol{\delta}_x^{(j)}$ because of (45), one can reformulate the quadratic programming problem as

$$\begin{aligned} \boldsymbol{\delta}_x^{(j)} &= \operatorname{argmin} \quad J_{\mathbf{x}}^{(j),T} \boldsymbol{\delta}_x^{(j)} + \frac{1}{2} \boldsymbol{\delta}_x^{(j),T} \mathrm{H}^{(j)} \boldsymbol{\delta}_x^{(j)} \\ \text{s. t. } \mathrm{F}_{\mathbf{x}}^{(j),T} \boldsymbol{\delta}_x^{(j)} &= -\mathcal{F}\left(\mathbf{x}^{(j)}\right), \end{aligned} \tag{46}$$

where the column vector $J_{\mathbf{x}}^{(j)} = \left.\dfrac{\partial \mathcal{J}_R}{\partial \mathbf{x}}\right|_{\mathbf{x}^{(j)}}$. The value $\mathbf{x}^{(j+1)}$ is obtained as $\mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} + \zeta \boldsymbol{\delta}_x^{(j)}$, where the step length $\zeta \in (0, 1]$ is chosen using a line search method.

The Lagrangian multiplier $\boldsymbol{\chi}^{(j+1)}$ can be computed as $\boldsymbol{\chi}^{(j+1)} = \boldsymbol{\chi}^{(j)} + \gamma(\boldsymbol{\chi}^{opt} - \boldsymbol{\chi}^{(j)})$, where $\boldsymbol{\chi}^{opt}$ is the optimal Lagrangian multiplier associated to problem (46). In order to avoid the computational costs associated with the evaluation of the Hessian, the matrix H can be replaced by a suitable approximation. A common approach is to use instead the matrix generated by the BFGS method. In general the effectiveness of the SQP method relies on the method used to solve the QP problem. Inequality constraints (e.g. the constraint that some parameters must be non negative) can be easily handled using SQP approach. In addition, we point out that when the exact Hessian is used, $\gamma = 1$, and only equality constraints are considered, the method is equivalent to solve the KKT conditions with the Newton method.

Notice that from the formulation of the SQP problem that the solution at iteration j does not need to be feasible, i.e. to satisfy the constraints. This approach allows to save a lot of time because we do not have to enforce the feasibility of the solution at each iteration. However, this lack of feasibility might affect the robustness of the method.

Unsteady problems The procedure illustrated above can be extended to unsteady problems, that are of major interest in fluid-structure interaction applications. However, in this case, it is important to notice that the adjoint problem (in any of the formulations we encountered) is a *final*-boundary value problem. This means that it is *backward* in time. This feature introduces high computational costs either when we solve the problem via the KKT system or we follow a gradient-based procedure based on the adjoint problem. In fact, the state problem (which is forward in time) and the adjoint problem need to be solved all together in space-time. The computational costs for this approach are therefore in many cases not affordable, not to mention the storage cost of the solutions at each time step. For this reason, different workarounds have been considered. For instance [37], the solution may be stored only on a predefined set of instants T_k (subset of the time discretization steps) called *checkpoints* and the state required by the optimization for computing the adjoint solution is locally recomputed or approximated.

Following a different approach, time discretization may be performed before the optimization, leading to a sequence of pseudo-steady optimization problems at each time step. An example of this approach will be provided in the next Section for estimating the compliance of an artery.

Interplay between numerical discretization and solution of the control problem In the numerical solution of control problems there is an usual dilemma, concerning the order of the steps for the optimization and the numerical approximation. We may summarize this as "Discretize then Optimize" (DO) vs. "Optimize then Discretize" (OD). The two operations are in general non-commutative and the solutions obtained with the two approaches are in general different. It is difficult to draw general indications about the most appropriate approach, which is basically a trade-off between accuracy and computational costs. The issue is extensively discussed in [37]. There are clearly pros and cons in both the sequences. With DO we may say that

- we avoid inconsistencies induced by the numerical differentiation of the KKT conditions; in other terms, the numerical approximation of the KKT conditions introduces a discrepancy with the real optimization condition;

- we can even use automatic differentiation software;
- we can split an unsteady problem into a sequence of pseudo-steady optimization problems.

On the other hand with OD:

- we do not deal with the differentiation of numerical artificial terms (like stabilization of advection terms for high Reynolds numbers);
- managing moving boundary problems as in shape optimization is easier, since we do not need the derivative of the grid with respect to the optimization parameters.

In the examples that follow we stick to a DO approach. A parameter estimation procedure based on OD can be found in [74] for the estimate of cardiac conductivities.

3.3 Reducing the costs via solution reduction

As we have pointed out several times, the optimization methods presented above suffer from high computational costs for different reasons. The state equations and possibly the adjoint problem need to be solved at each iteration, not to mention the additional costs in terms of computations and storage for unsteady problems, that need to be truly tackled in 4D (space and time).

In order to reduce the computing time we need to reduce either the number of iterations or the cost of each iteration (or both). The number of iterations may be reduced by using effective optimization algorithms as the BFGS method for updating the current solution. The cost of each iteration can be reduced by treating the constraints in a "flexible" way. This means that the fulfillment of the constraints may be relaxed in particular in the first iterations when this does not prevent the convergence to the admissible solution. This can be done by accepting a solution to the state equations featuring relatively large residuals or by replacing the state equations themselves with a simplified model. These approaches are mostly problem-dependent, being based on the possible simplifications offered by the problem at hand. For instance in electrocardiology, the Bidomain equations that describe the dynamics of the extra and intra-cellular potentials may be replaced by the simplified Monodomain system (see e.g. [16]). When solving fluid-structure interaction problems in hemodynamics a fully 3D coupling may be downscaled to a 3D Fluid/2D Structure problem [57], as we see in the next Section.

Here we address another (somehow complementary) way for reducing the computational costs, which is based on reducing more specifically the number of degrees of freedom required to give an accurate representation of the solution. As a matter of fact, a function in a (separable) Hilbert space (like for instance L^2 or H^1) admits the representation

$$u = \sum_{i=1}^{\infty} U_i \psi_i,$$

for a proper selection of the basis functions ψ_i. In the Galerkin approach for approximating the solution, we generally find a basis function set to represent the approximate solution $u_N(x)$ as

$$u_N = \sum_{i=1}^{N} U_{N,i}\varphi_i.$$

The basis functions may be piecewise polynomials as in the finite element method, or global polynomials as in spectral methods. In general, those basis functions can be defined to be *general purpose*, in the sense that they do not specifically rely on the feature of the problem to be solved and can be applied to a vast class of problems. This versatility has the drawback that, in general, to achieve accurate solutions the number N of degrees of freedom is high. This clearly implies high computational costs as the associated linear(ized) systems are large.

A somehow opposite approach would be to give up pursuing a general basis, using an "educated" basis that incorporates specific information of the problem. For instance, in *modal analysis* the solution is represented on the basis given by the eigenfunctions of the problem. The basis is therefore problem-dependent, bringing intrinsically information on the problem to be solved. The gain is that it is generally possible to achieve a good accuracy when truncating to a low number of degrees of freedom. However, this is not for free, as the basis needs to be specifically computed. In particular, computation of eigenfunctions is in general not trivial and quite costly [15].

Following the same idea of constructing an informed basis, we may consider *snapshot-based* approaches. In this case, the basis is the result of the elaboration of the solutions of the problem for particular configurations useful for the solution of the state problem in the optimization process. For instance, when the control variable is a parameter to be identified (as γ was in our working example), snapshots may be the solution of the state problem for a particular set of values of the parameter. The proper identification of this set is clearly crucial for the effectiveness of the entire procedure. This can be realized by considering that if the optimal value of the control variable falls within the range considered in the snapshots, the entire procedure configures as a sophisticated "interpolation", for which several convergence results are available. On the contrary, if the range of the snapshots computation is not well defined, we are actually performing an "extrapolation" and the convergence is not necessarily guaranteed. Again, the final goal is to keep the size N of the finite dimensional approximation of the solution as small as possible, thanks to the information contained in the basis.

From the computational standpoint, this snapshot-based approach relies on the *off-line/on-line* paradigm, namely

1. computation of the basis is "off-line", and it is intended to be an accurate (and therefore expensive) numerical approximation of the solution for different configurations that are considered to be relevant to the basis;

2. solution of the optimization problem, and in particular the computation of the coefficients $U_{N,i}$ along the iterations of the minimization procedure is "on-line", and contributes to the actual cost of the control procedure.

In this way, the computational costs are factorized, the major contribution being carried out in a step preliminary to the optimization. This paradigm clearly makes sense whenever the "off-line" part can be recycled for the solution of several optimization problems[9].

Among the different snapshot-based strategies, we mention the *reduced basis method* and the *Proper Orthogonal Decomposition* (POD). In the former, the snapshots are computed for values of the parameters that are evaluated to perform the best control of the error on the basis of rigorous error estimates (see [64, 70]). In particular, we mention [52] for an application of the reduced basis method to fluid-structure interaction problems. The latter is known also as Karhunen-Loève decomposition or *principal component analysis* and it is illustrated more in detail in the next paragraph.

POD basis selection We start assuming that a set of size M of solutions is available for instance by computing snapshots for M different values of the parameter of interest after a uniform sampling of an appropriate range. We assume that M is still large for the purpose of reducing the computational costs and that a reduction of the size of the basis is required, by properly filtering redundancy in the snapshots set. Denote by $\boldsymbol{\rho}_j \in \mathbb{R}^N$ the M snapshots of the (approximate) solution, with $j = 1, \dots M$. Then, we perform the following steps.

1. *Sample average*: $\overline{\boldsymbol{\rho}} = \sum_{j=1}^{M} \boldsymbol{\rho}_j$.

2. *Sample Covariance*: Compute $\mathrm{C} \in \mathbb{R}^{M \times M}$, whose elements are defined as $c_{ij} := \dfrac{1}{M}(\boldsymbol{\rho}_i - \overline{\boldsymbol{\rho}})^T(\boldsymbol{\rho}_j - \overline{\boldsymbol{\rho}})$. Matrix C is positive semidefinite and symmetric so the eigenvalues are all real and the eigenvectors $\{\mathbf{x}_j\}$ form an orthonormal basis in $\mathbb{R}^M$. We order the eigenvalues as

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq 0.$$

3. *Thresholding*: Select a tolerance $\tau \in [0, 1]$ and pick the minimum $\widetilde{M}$ such that

$$\frac{\sum_{i=1}^{\widetilde{M}} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \geq \tau.$$

[9] This can be problematic in a clinical context, where patient-specific geometries differ one from the other and the snapshot computation is not trivially recycled. Anatomical atlas mapping ideal to real geometries are required.

Here γ is a threshold that identifies the "essential" information. Hopefully, this happens for $\widetilde{M} \ll M$.

4. *New basis.* Let us select a new basis $\{\mathbf{y}_i\}$ consistent with the eigenvalues threshold. We take for $i = 1, \dots \widetilde{M}$

$$\mathbf{y}_i = \sum_{j=1}^{\widetilde{M}} (\mathbf{x}_i)_j (\boldsymbol{\rho}_j - \overline{\boldsymbol{\rho}}),$$

where $(\mathbf{x}_i)_j$ is the $j-$th entry of the $i-$th eigenvector. Then, we normalize $\mathbf{y}_i^* = \dfrac{1}{\|\mathbf{y}_i\|}\mathbf{y}_i$.

This is by construction an orthonormal basis . In addition and more importantly, this basis fulfils an optimal property. As a matter of fact [37, 15], the space spanned by the POD basis is the best $\tilde{M}$-dimensional subspace approximation of the space spanned by the snapshots (in the 2-norm sense). A vector $\mathbf{x}$ in $\mathbb{R}^M$ can then be approximated in terms of the POD basis as

$$\mathbf{x} = \bar{\boldsymbol{\rho}} + \sum_{i=1}^{\tilde{M}} c_i \mathbf{y}_i^*$$

For particular problems, such as progressive waves, reduction of the size for the solution and eventually of the computational costs after this procedure may be not enough. In this case, other reduced solution techniques may be considered [30]. Nevertheless, an example of POD for the solution of an inverse fluid-structure interaction problem is illustrated in the next Section.

Remark 3.2 *Here we have presented a particular application of POD for reducing the dimension of the solution forward problem. However, POD can be used for reducing the dimensionality in different contexts. For instance, in [7], Chapter 7, POD is advocated also for reducing the dimensionality of the size of the parameter space, which is crucial when the parameter is a function represented by a large number of degrees of freedom.*

Remark 3.3 *Here we introduced the POD using the eigenvalue decomposition of the sample covariance matrix. Alternatively, one can perform the Singular Value Decomposition (SVD) of the snapshots matrix $X = [\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_M]$. This can be efficiently done by first performing a QR factorization of X and then by computing the SVD of the triangular factor. In other words, $X = QR = QU\Sigma V^T = \tilde{U}\Sigma V^T$. The POD basis is then made of the first $\tilde{M}$ left singular vectors (the columns of $\tilde{U}$), where $\tilde{M}$ is chosen with the same procedure as before, using the singular values of the snapshots matrix rather than the eigenvalues of the covariance matrix.*

4 Some applications of Data Assimilation in Hemodynamics problems

In this Section we consider some applications of DA and Parameter Estimation in computational hemodynamics.

First, we present the problem of reconstructing the blood flow in a vessel assimilating sparse noisy measures of the velocity with the numerical results obtained by solving the incompressible Navier-Stokes equations. Successively, we consider the problem of estimating the compliance of a vessel based on measures of the displacement retrieved from medical images. The solution to this problem leads to an *inverse fluid-structure interaction* (IFSI) problem. These are not the only examples of data assimilation in biomedical applications. We mention for instance the work in electrocardiology for the set up of patient-specific models in [43], and for estimating cardiac conductivities [36, 74, 31]. Other applications can be found e.g. in [17, 28]. In particular, in [25, 26] DA methods are advocated for filling the gap between available boundary data and mathematical conditions required to solve the problem.

We have selected these examples because they offer the opportunity to see in action different methodologies based on the techniques illustrated in the previous Sections.

4.1 Assimilation of velocity measures in blood flow simulations

We consider the problem of merging velocity measures and the numerical simulation of blood flow. The DA problem can be addressed in several and diverse ways, as described in the previous Sections. More precisely, we present two approaches introduced in two recent papers; in the first one [18] the problem is faced with a variational (control) method, where the control variable is the normal component of the stress at the inflow section of the vessel. In the second paper [39] the authors exploit a Least Squares Finite Element (LSFE) approximation treating internal layers, where measures are available, as artificial boundaries. This approach can be reinterpreted as a MAP Bayesian estimate, as pointed out in [21].

We introduce the formal statement of the problem. Let us denote by Ω a domain in $\mathbb{R}^d$ ($d = 2, 3$; in real applications $d = 3$). We assume (see Figure 7) that Ω features an inflow boundary Γ_{in}, an outflow boundary Γ_{out} and the physical wall of the vessel Γ_{wall}. Γ_{in} and Γ_{out} can possibly consist of several sections. The variables of interest are the velocity $\mathbf{u}(\mathbf{x}) \in [H^1(\Omega)]^d$ and the pressure $p(\mathbf{x}) \in L^2_0(\Omega)$. Also, we assume to have some velocity measures as in, e.g., Figure 7 or sparse in the domain.

Velocity and pressure are assumed to obey the incompressible Navier-Stokes

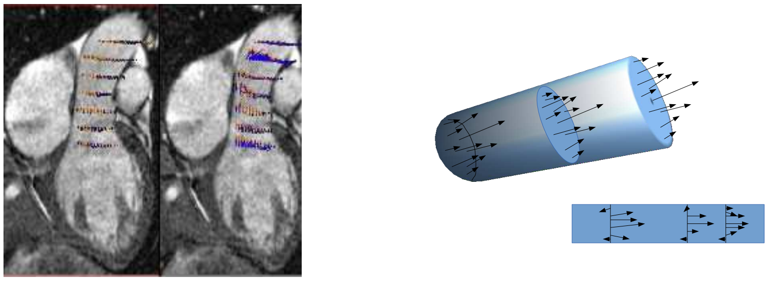


Figure 7: On the left, view of blood measured velocities in an MRI of the ascending aorta. On the right, examples of a three-dimensional and two-dimensional domain for which data are collected on internal layers transversal to the flow.

equations (NSE).

$$
\begin{aligned}
&\frac{\partial \mathbf{u}}{\partial t} - \mu\, \nabla \cdot (\nabla \mathbf{u} + \nabla \mathbf{u}^{\mathrm{T}}) + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} && \text{in } \Omega, \\
&\nabla \cdot \mathbf{u} = 0 && \text{in } \Omega, \\
&\mathbf{u} = \mathbf{0} && \text{on } \Gamma_{wall}, \\
&-\mu\,(\nabla \mathbf{u} + \nabla \mathbf{u}^{\mathrm{T}}) \cdot \mathbf{n} + p \cdot \mathbf{n} = \mathbf{h} && \text{on } \Gamma_{in}, \\
&-\mu\,(\nabla \mathbf{u} + \nabla \mathbf{u}^{\mathrm{T}}) \cdot \mathbf{n} + p \cdot \mathbf{n} = \mathbf{g} && \text{on } \Gamma_{out}.
\end{aligned}
\tag{47}
$$

A Newtonian rheology is supposed to hold, which is a common assumption in large and medium vessels [24], and μ is the constant kinematic viscosity. The choice of homogeneous Dirichlet boundary conditions on Γ_{wall} reflects the fact that we consider fixed geometries.

In this Section we consider the steady case $\dfrac{\partial \mathbf{u}}{\partial t} = \mathbf{0}$.

4.1.1 Variational approach

In [18] a variational DA procedure for the inclusion of velocity measures in the simulation of the NSE in hemodynamics is proposed. Sparse noisy velocity measures $d_1, \dots d_{N_s}$ are assumed to be available in the domain and possibly on the boundary at some sites[10] $\mathbf{x}_i^m \in \Omega$, $i = 1, \dots N_s$, that do not necessarily belong to a plane or a layer inside of Ω.

The assimilation technique in [18] is formulated as a control problem where the misfit between computed data (the NSE solution) and observed data is minimized. The equations of incompressible fluid dynamics are the constraint to the

[10] Notice that we use the word "sites" for the location of measurements, as opposed to the word "nodes" for points where velocities are computed. In general sites and nodes are different, but we do not exclude that the intersection of sites set and nodes set in non-empty.

minimization procedure. The control variable is selected to be the inflow normal (or natural) stress $\mathbf{h}$; knowledge of this quantity is quite often not available in practice. The variational problem is formulated as

$$\begin{aligned} &\min_{\mathbf{h}} \mathcal{J}(\mathbf{u},\, \mathbf{h}) = \|f(\mathbf{u}) - \mathbf{d}\|_{l^2} + \mathcal{R}(\mathbf{h}) \\ &\text{s.t. Steady version of (47)} \end{aligned} \tag{48}$$

Here, f is a filtering vector function that returns the value of the velocity field evaluated on the measurement sites; $\mathcal{R}$ is a regularization term added to prevent potential ill-posedness and ill-conditioning of the problem due to the location of data and the presence of noise.

For the numerical solution of the problem (48) we first consider the linearized NSE; then, we discuss the nonlinear case; in the linearized formulation the term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ is substituted by $\boldsymbol{\beta} \cdot \nabla \mathbf{u}$, where $\boldsymbol{\beta}$ is a known advection field. We follow a DO approach (see Section 3.2), thus, after the discretization (via e.g. the finite element method) of the functional and the linearized state equations we resort to the following algebraic optimization problem

$$\begin{aligned} &\min_{\mathbf{H}} J(\mathbf{V},\, \mathbf{H}) = \frac{1}{2}\|\mathrm{D}\mathbf{V} - \mathbf{d}\|_2^2 + \frac{\alpha}{2}\|\mathrm{L}\mathbf{H}\|_2^2 \\ &\text{s.t. } \mathrm{S}\mathbf{V} = \mathrm{R}_{in}^{\mathrm{T}}\mathrm{M}_{in}\mathbf{H} + \mathbf{F}. \end{aligned} \tag{49}$$

Here, $\mathbf{V} = [\mathbf{U}\ \mathbf{P}] \in \mathbb{R}^{N_u+N_p}$ is the vector of discretized velocity $\mathbf{U} \in \mathbb{R}^{N_u}$ and pressure $\mathbf{P} \in \mathbb{R}^{N_p}$; $\mathbf{H} \in \mathbb{R}^{N_{in}}$ is the discretization of the control variable $\mathbf{h}$; N_{in} is the number of degrees of freedom of the velocity on Γ_{in}; $\mathbf{d} = [d_1\ \ldots\ d_{N_s}] \in \mathbb{R}^{N_s}$ is the vector of the available measures. For $\alpha > 0$, $\frac{\alpha}{2}\|\mathrm{L}\mathbf{H}\|_2^2$ is a Tikhonov regularization term, see Sect. 3.2. The matrix S is defined as follows

$$\mathrm{S} = \begin{bmatrix} \mathrm{C} + \mathrm{A} & \mathrm{B}^{\mathrm{T}} \\ \mathrm{B} & \mathrm{O} \end{bmatrix}, \tag{50}$$

where, C, A $\in \mathbb{R}^{N_u,N_u}$ and B $\in \mathbb{R}^{N_p,N_u}$ are the discretization of the diffusion, advection and divergence operators. D is the selection or observation matrix and it is defined as D $= [\mathrm{Q}\quad \mathrm{O}]$, where Q $\in \mathbb{R}^{dN_s,N_u}$ is such that $[\mathrm{Q}\mathbf{U}]_i$ is the numerical solution evaluated at the data sites. $\mathrm{R}_{in} \in \mathbb{R}^{N_{in},N_u+N_p}$ is a restriction matrix which selects the degrees of freedom of the velocity on Γ_{in}. $\mathrm{M}_{in} \in \mathbb{R}^{N_{in},N_{in}}$ is the discretization of the mass operator restricted to inlet boundary nodes.

For the solution of problem (49) we use the Lagrange multiplier approach, so we consider the Lagrange functional

$$L(\mathbf{V},\mathbf{H},\mathbf{X}) = \frac{1}{2}\|\mathrm{D}\mathbf{V} - \mathbf{d}\|_2^2 + \frac{\alpha}{2}\ \|\mathrm{L}\mathbf{H}\|_2^2 + \mathbf{X}^{\mathrm{T}}(\mathrm{S}\mathbf{V} - \mathrm{R}_{in}^{\mathrm{T}}\mathrm{M}_{in}\mathbf{H} - \mathbf{F}), \tag{51}$$

where $\mathbf{X} \in \mathbb{R}^{N_u+N_p}$ is the discrete Lagrange multiplier. The set of necessary conditions for optimality is given by the KKT system

$$\left\{\begin{array}{l}\dfrac{\partial L}{\partial \mathbf{V}} = \mathrm{D}^{\mathrm{T}}(\mathrm{D}\mathbf{V} - \mathbf{d}) + \mathrm{S}^{\mathrm{T}}\mathbf{X} = \mathbf{0} \\ \dfrac{\partial L}{\partial \mathbf{H}} = \alpha \mathrm{L}^{\mathrm{T}}\mathrm{L}\,\mathbf{H} - \mathrm{M}_{in}^{\mathrm{T}}\mathrm{R}_{in}\mathbf{X} = \mathbf{0} \\ \dfrac{\partial L}{\partial \mathbf{X}} = \mathrm{S}\mathbf{V} - \mathrm{R}_{in}^{\mathrm{T}}\mathrm{M}_{in}\mathbf{H} - \mathbf{F} = \mathbf{0}.\end{array}\right. \tag{52}$$

By defining $\mathrm{Z} = \mathrm{DS}^{-1}\mathrm{R}_{in}^{\mathrm{T}}\mathrm{M}_{in}$ and $\mathrm{W} = \mathrm{Z}^{\mathrm{T}}\mathrm{Z} + \alpha\,\mathrm{L}^{\mathrm{T}}\mathrm{L}$ the reduced system, obtained by block elimination, reads $\mathrm{W}\mathbf{H} = \mathrm{Z}^{\mathrm{T}}(\mathbf{d} - \mathrm{DS}^{-1}\mathbf{F})$, where W is the so-called *reduced Hessian* matrix.

The following theorem states the necessary and sufficient conditions for the well-posedness of problem (49).

Theorem 4.1 *For $\alpha = 0$, W is non-singular, i.e. (49) is well-posed, $\Leftrightarrow$*

$$\mathit{Null}(\mathrm{D}) \cap \mathit{Range}(\mathrm{S}^{-1}\mathrm{R}_{in}^{\mathrm{T}}\mathrm{M}_{in}) = \{\mathbf{0}\}. \tag{53}$$

For the proof see [18]. Condition (53) is satisfied when “enough” data are available on the inflow section (number and location of the measures that guarantee the well-posedness depend on the discretization method used).

In order to consider the nonlinear advection term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ and to solve the nonlinear PDE constrained optimization problem (48), we combine the DA procedure for the linearized case and classical fixed point schemes for the solution of the NSE. In particular, we refer to the Picard and Newton methods [63]. The assimilation problem is solved iteratively as follows.

Given a guess for the velocity field at iteration $k+1$, $\mathbf{U}_k$,

$$\textit{solve} \quad \left\{\begin{array}{l}\displaystyle\min_{\mathbf{H}_k} \frac{1}{2}\|\mathrm{D}\mathbf{V}_{k+1} - \mathbf{d}\|_2^2 + \frac{\alpha}{2}\|\mathrm{L}\mathbf{H}_{k+1}\|_2^2 \\ \text{s.t.} \quad \mathrm{S}_k\mathbf{V}_{k+1} = \mathrm{R}_{in}^{\mathrm{T}}\mathrm{M}_{in}\mathbf{H}_{k+1} + \mathbf{F}_k\end{array}\right. \tag{54}$$

until $\|\mathbf{V}_k - \mathbf{V}_{k+1}\| \leq \delta$, being δ a user defined tolerance. Here,

$$\mathrm{S}_k = \left[\begin{array}{cc}\mathrm{C} + \mathrm{A}_k & \mathrm{B}^{\mathrm{T}} \\ \mathrm{B} & \mathrm{O}\end{array}\right], \qquad \text{and} \qquad \mathbf{F}_k = \mathbf{F} + w\mathbf{Y}_k. \tag{55}$$

A_k comes from the discretization of $(\overline{\mathbf{u}}_k \cdot \nabla)\mathbf{u}_{k+1} + w(\mathbf{u}_{k+1} \cdot \nabla)\overline{\mathbf{u}}_k$ ($w = 0$ for Picard method, 1 for Newton); $\mathbf{Y}_k$ is the discretization of $(\overline{\mathbf{u}}_k \cdot \nabla)\overline{\mathbf{u}}_k$. Here $\overline{\mathbf{u}}_k$ is defined as $\vartheta\mathbf{u}_{k-1} + (1-\vartheta)\mathbf{u}_k$, being $\vartheta \in [0, 1]$, w is a relaxation parameter, chosen empirically.

Numerical tests. In Figure 8 we report the numerical results obtained on two geometries approximating blood vessels. In Figure 8 (left) the computational grid is an approximation of a carotid artery; the colored vector field consists in the actual data used in the assimilation, these are generated adding Gaussian noise to a reference solution; to appreciate the presence of the noise the noise-free data are also reported in black. In the center, the magnitude of the assimilated velocity is displayed; a comparison with a reference solution (conducted in [18]) shows that the noise is filtered and that the assimilated solution is close to the reference one.

On the right, a three-dimensional cylindrical domain is reported, this case is treated with an axisymmetric formulation. On selected internal surfaces the assimilated field and its magnitude are reported; it is important to note that the noise affecting the components of the velocity parallel to the flow is significantly filtered.

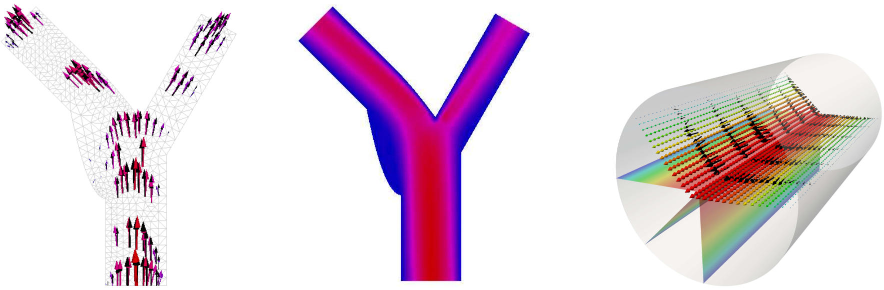


Figure 8: On the left, the colored vector field consists in the available measures whereas the black one corresponds to the noise-free data. In the center, the magnitude of the assimilated vector field is reported. On the right, the colored vector field corresponds to the assimilated velocity, the black one to the noisy data and the colored field in the background corresponds to the magnitude of the velocity.

Next, we consider the problem of estimating the wall shear stress (WSS) already described in the introductory example of Section 1. An accurate approximation of the WSS is fundamental in the investigation of cardiovascular pathologies since it is an index of the possibility of rupture of the vessel wall and formation of stenosis [24]. Approximations of the WSS retrieved from indirect measurements are in general unreliable because of the post-processing numerical errors and the noise affecting the measures. Including measurements in simulations is a way for improving the reliability of the computed solutions and, on the other hand, the introduction of the mathematical (numerical) model results in noise filtering. For the geometry of Fig. 8 (left), we compute the WSS computed on a selected internal wall. In order to quantify the accuracy of this solution we compare the assimilated WSS with the one associated with a reference solution, we introduce the index of accuracy $E_{\mathrm{WSS}} = \|\mathrm{WSS} - \mathrm{WSS}_{\mathrm{FE}}\|_2 / \|\mathrm{WSS}_{\mathrm{FE}}\|_2$ where $\mathrm{WSS}_{\mathrm{FE}}$ is the value retrieved from the reference solution. In correspondence of decreasing values of signal to noise ratio[11] (SNR), the WSS errors

[11] We define the signal to noise ratio as the ratio between the maximum of the absolute

SNR	$E_{\text{WSS,DA}}$	$E_{\text{WSS,FW}}$
100	0.2536	0.2667
20	0.2591	0.3030
10	0.2738	0.3861
5	0.3149	0.6114

Table 1: Comparison of relative errors for the WSS computed with DA and forward solution.

obtained with the assimilated velocity field, $E_{\text{WSS,DA}}$, compared with those obtained from a forward simulation on the same grid with the same noisy measures (used for DA) as boundary data on the inflow section are reported in Tab. 1. With high SNR the gain obtained with DA is not significant, however as we decrease SNR we can obtain up to the 50% of gain with respect to the forward simulation.

4.1.2 Bayesian approach

A *Variational Bayesian* approach to the assimilation problem is possible. This formulation features an overlap between statistical and variational techniques; both point estimators and confidence regions for the velocity are considered. Here, we recall the method for the computation of the MAP and ML estimators (see Section 2.2) and present some numerical results that illustrate how the knowledge of the nature of the measurement noise can significantly improve the quality of the estimation with respect to the deterministic estimates.

We assume to deal with discretized variables, all treated as random; in the remainder of this Section the bold variables denote random vectors while the capital plain variables a specific realization. With an abuse of notation we introduce the random variable $\mathbf{H}$ which describes the normal stress of the fluid at the inflow section; $\mathbf{M}$ is the random variable that describes the measures and $\boldsymbol{\nu}$ the noise perturbing the measurements. We let p_H be the p.d.f. of $\mathbf{H}$, or its a priori distribution, and p_ν the one of $\boldsymbol{\nu}$; these distributions are assumed to be known. As described in Section 2.2 the purpose of the Bayesian procedure is to estimate the posterior distribution $p_{H|M}$ exploiting the Bayes formula (5) in the form

$$p_{H|M} = \frac{p_{M|H}\, p_H}{p_M}; \tag{56}$$

where p_M is the p.d.f. of the measures.

First, we assume that the relation between the random vectors $\mathbf{H}$ and $\mathbf{U}$, the random variable that describes the velocity, is linear (i.e. we consider the linearized NSE), then we treat the nonlinear case.

In the linearized formulation $\mathbf{H}$ and $\mathbf{M}$ are related by the following *additive noise* relation

$$\mathbf{U} + \boldsymbol{\nu} = \mathbf{M} \quad \Rightarrow \quad \mathbf{Z}\mathbf{H} + \boldsymbol{\nu} = \mathbf{M}. \tag{57}$$

value of the signal and the standard deviation of the noise

Here, the *observation operator* between $\mathbf{H}$ and $\mathbf{M}$ (see the example in Section 2.2 for Gaussian vectors), is actually the inverse of a (discrete) differential operator; $\mathrm{Z} = \mathrm{DS}^{-1}\mathrm{R}_{in}\mathrm{M}_{in}$ has been introduced after (52) and it describes the deterministic relation between the velocity and the normal stress. The random variable $\boldsymbol{\nu}$ accounts for the measurement noise. We make the assumption of mutual independence of $\mathbf{U}$ and $\boldsymbol{\nu}$. Since $\mathbf{H}$ and $\mathbf{U}$ are related by a linear relation this implies the independence of $\mathbf{H}$ and $\boldsymbol{\nu}$. As a consequence, the p.d.f. of $\boldsymbol{\nu}$ is independent of any realization of $\mathbf{H}$ and the likelihood function, $p_{M|H}$, can be expressed as

$$p_{M|H}(M) = p_{M|H}(\nu + \mathrm{Z}H) = p_{\nu}(M - \mathrm{Z}H). \tag{58}$$

Next, we consider the realization $M = \mathbf{d}$ (the vector of available velocity measures introduced previously), we have

$$p_{H|M}(H) \propto p_{M|H}(\mathbf{d})\, p_H(H) = p_{\nu}(\mathbf{d} - \mathrm{Z}H)\, p_H(H). \tag{59}$$

Now we make the assumption that all variables are Gaussian and we define the a priori distribution and the noise distribution as follows

$$\begin{aligned} p_H = g_H \propto exp\left\{-\frac{1}{2}(H - H_0)^{\mathrm{T}}\Lambda_H^{-1}(H - H_0)\right\}, \\ p_{\nu} = g_{\nu} \propto exp\left\{-\frac{1}{2}(\nu - \nu_0)^{\mathrm{T}}\Lambda_{\nu}^{-1}(\nu - \nu_0)\right\}; \end{aligned} \tag{60}$$

where H_0 and ν_0 are the expectation values and Λ_H and Λ_ν are the correlation matrices for $\mathbf{H}$ and $\boldsymbol{\nu}$ respectively. The analysis of Section 2.2 shows that the posterior distribution $p_{H|M}$ is a Gaussian distribution itself with covariance and mean given by

$$\begin{aligned} \Lambda_{H|M} &= (\Lambda_H^{-1} + \mathrm{Z}^{\mathrm{T}}\Lambda_{\nu}^{-1}\mathrm{Z})^{-1}, \\ \mathcal{E}(\mathbf{H}) &= \Lambda_{H|M}^{-1}(\mathrm{Z}^{\mathrm{T}}\Lambda_{\nu}^{-1}(\mathbf{d} - \nu_0) + \Lambda_H^{-1}H_0). \end{aligned} \tag{61}$$

We recall that the mean value of the posterior distribution is the value that maximizes $p_{H|M}$, and then, by definition, it is the MAP estimator of $\mathbf{H}$, say $\widehat{\mathbf{H}}_{MAP}$. On the other hand, the value that maximizes the likelihood function, with respect to H, corresponds to the ML estimator for $\mathbf{H}$ and has the following expression

$$\widehat{\mathbf{H}}_{ML} = (\mathrm{Z}^{\mathrm{T}}\Lambda_{\nu}^{-1}\mathrm{Z})^{-1}(\mathrm{Z}^{\mathrm{T}}\Lambda_{\nu}^{-1}(\mathbf{d} - \nu_0)). \tag{62}$$

In treating the nonlinearity we consider an iterative approach similar to the deterministic one described in the previous Section; in fact, also in this case, we rely on the Newton method for the NSE. The distribution $p_{H|M}$ for the nonlinear model is still Gaussian, the following algorithm is used to determine its mean and covariance.

Given a guess for the random vector $\mathbf{U}_k = Z_k \widehat{\mathbf{H}}_{MAP,k}$ *at iteration* $k+1$,

$$\begin{aligned} &(1)\ \textit{compute} \quad \Lambda_{H|M,k+1} = \Lambda_H^{-1} + \mathrm{Z}_k^{\mathrm{T}} \Lambda_\nu^{-1} \mathrm{Z}_k \\ &(2)\ \textit{solve} \qquad \Lambda_{H|M,k+1}(\widehat{\mathbf{H}}_{MAP,k+1}) = \mathrm{Z}^{\mathrm{T}} \Lambda_\nu^{-1}(\mathbf{d} - \varepsilon_0) + \Lambda_H^{-1} h_0, \end{aligned} \tag{63}$$

until a convergence criterion is satisfied.

Here, for $\mathrm{Z}_k = \mathrm{D}\mathrm{S}_k^{-1}\mathrm{R}_{in}^{\mathrm{T}}\mathrm{M}_{in}$ we define

$$\mathrm{S}_k = \left[\begin{array}{ll} \mathrm{C} + \mathrm{A}_k & \mathrm{B}^{\mathrm{T}} \\ \mathrm{B} & \mathrm{O} \end{array} \right]. \tag{64}$$

A_k is the discretization of the advection operator with advection field $\mathbf{U}_k$, the velocity vector associated with the normal stress $\mathcal{E}(H)_k$. Note that with this formulation $\mathbf{H}$ and $\mathbf{U}$, at each iteration, are related by a linear model and, for this reason, $\mathbf{U}$ can still be considered normally distributed.

Numerical tests. We assume to have an exact, analytic, solution of the NSE and we compare the accuracy of the MAP and ML estimators vs. the "deterministic estimator" introduced in the previous Section, i.e. the solution of the variational formulation. The index of accuracy is related to the velocity fields, $\widehat{\mathbf{U}}$, retrieved from $\widehat{\mathbf{H}}_{MAP}$, $\widehat{\mathbf{H}}_{ML}$ and $\widehat{\mathbf{H}}_{det}$ (the deterministic estimate); it is defined as $E(\widehat{\mathbf{U}}) = \frac{\|\widehat{\mathbf{U}} - \mathbf{U}_{anl}\|_2}{\|\mathbf{U}_{anl}\|_2}$, where $\mathbf{U}_{anl}$ is the discretized analytic solution. We also define an average error over a set of noise realizations $\{\boldsymbol{\nu}\}_{i=1}^n$, $\overline{E}(\widehat{\mathbf{U}}) = \frac{1}{n}\sum_{i=1}^n E(\widehat{\mathbf{U}}, i)$ where $E(\widehat{\mathbf{U}}, i)$ is associated with the i-th realization of noise $\boldsymbol{\nu}_i$. In addition, we consider a measure of the gain, γ, in using statistical estimators as opposed to deterministic ones: $\gamma = 1 - \frac{\overline{E}(\widehat{\mathbf{U}}_{stat})}{\overline{E}(\widehat{\mathbf{U}}_{det}}$ where *stat* stands for either MAP or ML.

The details of the numerical tests are fully reported in [19].

In a square domain we consider data on Γ_{in} and internal data located on 10 internal slices. In Table 2 we report results obtained in correspondence of SNR of 20 and 10. In the computation of $\widehat{\mathbf{H}}_{MAP}$ and $\widehat{\mathbf{H}}_{det}$ the regularization parameter $\alpha = 0.5$ is chosen empirically (left table in Tab. 2). In the computation of $\widehat{\mathbf{H}}_{ML}$ and $\widehat{\mathbf{H}}_{det}$ on the right table the regularization parameter α is set to 0. From the results we infer the following facts. (1.) Compared to the deterministic estimator, the statistical estimators are always more accurate since they take into account additional information brought by statistical properties of the data. (2.) The computational time required in solving the statistical formulations is, in average, 1.3 times bigger than the one required by the deterministic one. (3.) The poor gain in correspondence of SNR = 20 means that statistical information associated with a low amount of noise is not significant enough to make a considerable difference with respect to deterministic estimates in terms of accuracy.

As a second example we consider the same problem setting of the previous Section for the flow in a cylinder, see Figure 8 (right); we consider measures

SNR	$\overline{E}_{U,det}$	$\overline{E}_{U,MAP}$	γ
20	0.0822	0.07371	10%
10	0.1394	0.1041	25%

SNR	$\overline{E}_{U,det}$	$\overline{E}_{U,ML}$	γ
20	0.0855	0.0579	6%
10	0.1675	0.1363	18%

Table 2: Accuracy results for statistical and deterministic solutions for the NSE.

SNR	$\overline{E}_{U,det}$	$\overline{E}_{U,MAP}$	γ
20	0.0396	0.0308	22%
10	0.1423	0.0978	31%

Table 3: Accuracy results for statistical and deterministic solutions for the axisymmetric case.

on the inflow boundary and internal data located on 5 internal slices. We only compute the MAP estimator (the problem for the computation of $\widehat{\mathbf{H}}_{ML}$ is ill-posed). In this experiment $\alpha =$ 1e-7; results in Table 3 show that with the MAP estimator we have a significant gain in accuracy. Moreover, the computational time required by the statistical estimators is the same as for the deterministic one.

4.1.3 Weighted least squares finite element method

Another approach to the assimilation of measured velocities has been proposed in [39]. This work is mainly inspired by the development of a new experimental technique, the particle imaging velocimetry [40], that can be used to determine two components of the blood velocity along a single plane within the ventricle of the heart. The proposed method relies therefore on the hypothesis that the measures are collected inside a three-dimensional region on a two-dimensional plane (as in Figure 7); the latter is basically treated as a (artificial) boundary.

This variational technique exploits a weighted least squares finite element method (WLSFEM), based on the LSFEM [9, 10, 11]; the latter has been utilized in general for the solution of PDEs. It features great flexibility in the enforcement of various types of boundary conditions. However, the LSFE method has been also applied to inverse problems since the 90's for the numerical solution of PDE constrained control problems; main contributors are Bochev and Gunzburger [9, 10, 11].

If we consider the problem of solving the following generic boundary value problem

$$\begin{aligned} Lu &= f \qquad \text{in } \Omega \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned} \tag{65}$$

where L is a first order linear differential operator and $J(u)$ is a cost functional defined as

$$J(u) = \|Lu - f\|^2_{L^2(\Omega)} + \|u - g\|^2_{H^{1/2}(\partial\Omega)}. \tag{66}$$

Then, the LSFE solution u is obtained as the minimal of $J(u)$.

Assume we have N_s measures $d_i(\mathbf{x})$ of the variable u on some layers, $\Gamma_1, \ldots, \Gamma_{N_s}$, internal to Ω. We want to perform DA for problem (65), i.e. we want to merge $\{d_i\}$ and the numerical solution of (65). The idea of the WLSFEM is to add penalization terms to the functional J. The internal layers are considered part of the boundary and the corresponding measures are treated as boundary data; these terms are then properly weighted according to the level of confidence of the measure. Thus, the cost functional is defined as

$$\widehat{J}(u) = J(u) + w_1 \|u - d_1\|^2_{H^{1/2}(\Gamma_1)} + \ldots w_{N_s} \|u - d_{N_s}\|^2_{H^{1/2}(\Gamma_{N_s})}, \tag{67}$$

where $w_1, \ldots, w_{N_s}$ are the weights. Note that (1.) with the introduction of these additional terms that penalize the difference between the observed data and the solution, the assimilation is weakly enforced; (2.) DA introduces an additional exra cost to LSFE calculation.

When applying the WLSFE method to the NSE one has to keep in mind that it is designed for first order linear differential operators; thus, we must recast the fluid dynamic equations into a linearized first order differential system. To this end, we consider a non-primitive variable set: we introduce the variable $\omega = -\nabla \times \mathbf{u}$, the negative vorticity, and the variable

$$\mathbf{r} = \nabla p + \frac{\sqrt{Re}}{2} \nabla |\mathbf{u}|^2,$$

commonly referred to as the "gradient of pressure", where Re is the Reynolds number. Then, we apply the WLSFE method to the equivalent problem in Ω (see [39] for details on how to derive the following system)

$$\begin{aligned}
&\nabla \times \mathbf{u} + \omega = 0\\
&\nabla \cdot \mathbf{u} = 0\\
&\frac{1}{\sqrt{Re}} \nabla \times \omega - \mathbf{r} - \sqrt{Re}(\mathbf{u} \times \omega) = 0\\
&\nabla \cdot \omega = 0\\
&\nabla \times \mathbf{r} = 0\\
&\nabla \cdot \mathbf{r} - \sqrt{Re}(\omega \cdot \omega) - Re(\mathbf{u} \cdot \mathbf{r}) = 0.
\end{aligned} \tag{68}$$

The optimization problem, formulated as in (66) with $Lu = f$ given by equations (68), is then solved with standard techniques from the calculus of variations. Again, we stress the fact that, being L a linear operator, the cost of the WLSFE formulation is of the same order of the solution of the NSE. However, this approach, as opposed to primitive variables formulations might be less conducive than the straightforward inclusion of available measures.

Numerical tests. Consider a cylindrical geometry and assume the measures to be located on internal layers parallel to the flow; in Figure 9 (left) the noisy data on the layer crossing the axis of symmetry of the cylinder are reported.

It is also assumed that the noise affects the boundary data (whereas in [18] they are considered exact), which is always the case in real applications. For the numerical solution with the WLSFE method the boundary and internal data are properly weighted according to the noise level (assumed known in this particular experiment); the assimilated solution, on an internal layer close to the measurement one, is reported in Figure 9 (right). The filtering action of the DA on the noise is evident. Quantitative analysis, not reported here, reveals a good level of accuracy [39].

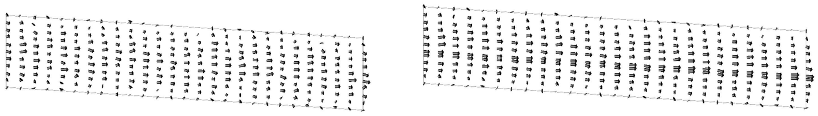


Figure 9: On the left the synthetic measures generated adding Gaussian noise to an analytical solution are reported on a layer crossing the axis of symmetry. On the right, on a layer close to the one where the measures are collected, the assimilated velocity is reported. Adapted from [39].

4.1.4 WLSFEM as a Bayesian approach to DA

In [21] a reinterpretation of the WLSFEM in terms of Bayesian approach to DA is proposed; in fact, in [39] the method is not presented in an inverse problem framework. Here we show that the WLSFE solution can be interpreted as the *maximum a posteriori* (MAP) estimator in a variational Bayesian approach to DA, for a certain choice of a priori distribution and likelihood function. A statistical interpretation of the weights is also provided.

In describing the method we refer to the general boundary value problem (65). We recall that in a Bayesian approach to DA all variables are treated as random, the goal is to determine the p.d.f. of u conditioned on realizations of the measures $d_1, \dots d_{N_s}$ available on the internal layers $\Gamma_1, \dots \Gamma_{N_s}$. We assume that the measures are affected by the measurement noise $\nu_1, \dots \nu_{N_s}$ such that $d_i(\mathbf{x}) = u(\mathbf{x})|_{\Gamma_i} + \nu(\mathbf{x})_i$, for $i = 1, \dots N_s$. To apply the Bayes theorem we need to define an a priori distribution for u, p_u, based on our prior belief on u and a likelihood function for the measurement noise ν_i, $p_{\nu,i}$. In order to show the equivalence between the WLSFE deterministic solution, or WLSFE estimator, and the MAP estimator in the Bayesian setting we make the following choices.

We define a prior distribution which is large when u satisfies the governing equations (65) "well" and small otherwise; in this way the prior describes to what extent the equations are a good model for the observations. Formally

$$p_u(u) \propto \exp\left\{-J(u)\right\},$$

where J is defined as in (66).

Next, in defining the likelihood functions for ν_i, we assume that the measurement errors ν_i are independent and normally distributed with null mean and variance $\frac{1}{2w_i}$, being w_i the weights introduced in the previous Section. Applying the Bayes theorem we have

$$p_{u|d_1 \dots d_{N_s}} \propto \exp\left\{-J(u) - w_1 \|u - d_1\|^2_{H^{1/2}(\Gamma_1)} + \dots w_{N_s} \|u - d_{N_s}\|^2_{H^{1/2}(\Gamma_{N_s})}\right\} \Rightarrow$$

$$p_{u|d_1 \dots d_{N_s}} \propto \exp\left\{-\widehat{J}(u)\right\},$$

for $\widehat{J}$ as in (67). The MAP estimator is then the value of u that maximizes the posterior distribution $p_{u|d_1 \dots d_{N_s}}$, thus

$$\widehat{u}_{MAP} = \arg\max p_{u|d_1 \dots d_{N_s}} = \arg\min \widehat{J}(u) = u_{\text{WLSFE}},$$

This leads to the conclusion that the WLSFE solution, u_{WLSFE}, is actually a Bayesian estimator for the variable u; thus, we have the following statistical interpretations

- the mathematical model encodes our prior belief on u;
- the data is a correcting likelihood;
- the weights reflect the variance of the measurement noise, i.e. are an index of the reliability of the measures.

This procedure, and the associated considerations, naturally apply to the first order form of the NSE so that the velocity estimated via WLSFEM is a Bayesian estimator. The latter differs from the one introduced in [19] in the choice of the prior distribution and likelihood function. The first approach is certainly more general as does not require the measures to be on a plane and more straightforward because formulated for the primitive variable, on the other hand, the second is computationally cheaper as it deals with the recast (linear) form of the NSE. As for the accuracy, an extensive comparison is still missing.

4.2 Estimation of the arterial compliance from measurements of displacement: an inverse fluid-structure interaction problem

As a second example, we consider the estimation of the compliance of an artery. The problem consists of estimating the compliance of an artery wall, based on (noisy) data of the displacement of the wall, obtained using medical devices such as Magnetic Resonance Imaging (MRI) during an heart beat. We focus on two approaches that have been recently adopted in the literature. In [59] a variational approach is pursued: the compliance is used as control variable for minimizing the misfits between the results of a fluid-structure interaction problem and the displacement of the vessel (possibly retrieved from images). In [6] a Reduced Order Unscented Kalman Filter is advocated to solve the same problem. In the following we summarize these two approaches and present some results of these works. For details we refer the reader to the corresponding works.

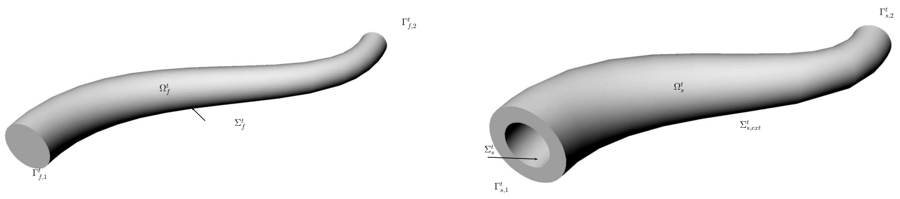


Figure 10: Representation of the domain of the FSI problem: fluid domain on the left, structure domain on the right.

4.2.1 Problem formulation

We consider a domain of a vessel (structure domain) perfused by blood (fluid domain) as depicted in Figure 10. We make the simplistic assumption that the vessel is (linearly) elastic, with the stress tensor $\boldsymbol{\sigma}_s$ depending on the vessel displacement $\boldsymbol{\eta}$ as

$$\boldsymbol{\sigma}_s(\boldsymbol{\eta}) \equiv \gamma_1(\nabla\boldsymbol{\eta} + (\nabla\boldsymbol{\eta})^T) + \gamma_2(\nabla\cdot\boldsymbol{\eta})\boldsymbol{I},$$

where

$$\gamma_1 := \frac{E}{2(1+\nu)}, \qquad \gamma_2 := \frac{E\nu}{(1+\nu)(1-2\nu)},$$

are the Lamé constants, $\boldsymbol{I}$ is the identity tensor, E is the Young's modulus and ν is the Poisson's ratio. For the sake of notation, we factor the Young's modulus E out of the stress tensor, and so that we can write

$$\boldsymbol{\sigma}_s = E\,\tilde{\boldsymbol{\sigma}}_s, \qquad \tilde{\boldsymbol{\sigma}}_s := \frac{1}{2(1+\nu)}(\nabla\boldsymbol{\eta} + (\nabla\boldsymbol{\eta})^T) + \frac{\nu}{(1+\nu)(1-2\nu)}(\nabla\cdot\boldsymbol{\eta})\boldsymbol{I}.$$

The vessel deforms under the stress coming from the blood, and in turn, the elastic structure of the vessel affects the blood flow. This problem has been largely investigated in other Chapters of this book (see also e.g [24]). For the sake of numerical approximation of the problem, the problem is formulated on a frame of reference moving with the physical wall of the artery and fixed on the artificial boundaries (inflow/outflow). This approach is known as the Arbitrary Lagrangian Eulerian (ALE) formulation, see, e.g [41, 20]. We write the problem according to the ALE frame of reference. At time t the blood velocity $\mathbf{u}$ and pressure p live in the fluid domain Ω^t_f, whereas the vessel displacement $\boldsymbol{\eta}$ lives in the structure vessel domain Ω^t_s. We denote the interface between the fluid and the solid domains with Σ^t (see Figure 10). It is more convenient to model the structure displacement $\boldsymbol{\eta}$ in the reference configuration $\hat{\Omega}_s$; we denote a variable in the reference configuration with a $\hat{}$, e.g. $\hat{\boldsymbol{\eta}}$.

1. *Fluid-Structure problem.* Find fluid velocity $\mathbf{u}$, pressure p and structure displacement $\boldsymbol{\eta}$ such that

$$\begin{cases} \rho_f \dfrac{D^A\mathbf{u}}{Dt} + \rho_f((\mathbf{u}-\mathbf{w})\cdot\nabla)\mathbf{u} - \nabla\cdot\boldsymbol{\sigma}_f = \mathbf{f}_f & \text{in } \Omega_f^t \times (0,T), \\ \nabla\cdot\mathbf{u} = 0 & \text{in } \Omega_f^t \times (0,T), \\ & \\ \rho_s \dfrac{\partial^2 \widehat{\boldsymbol{\eta}}}{\partial t^2} - \nabla\cdot(E\,\widehat{\boldsymbol{\sigma}}_s) = \widehat{\mathbf{f}}_s & \text{in } \Omega_s \times (0,T), \\ \mathbf{u} = \dfrac{\partial \boldsymbol{\eta}}{\partial t} & \text{on } \Sigma^t \times (0,T), \\ \boldsymbol{\sigma}_s\,\mathbf{n} - \boldsymbol{\sigma}_f\,\mathbf{n} = \mathbf{0} & \text{on } \Sigma^t \times (0,T), \end{cases} \tag{69}$$

where $\boldsymbol{\sigma}_f(\mathbf{u},p) = -p\,\boldsymbol{I} + \mu(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$, ρ_f and ρ_s are the fluid and structure density, μ is the constant blood viscosity, $\mathbf{f}_f$ and $\mathbf{f}_s$ are the forcing terms. Here, $\dfrac{D^A}{Dt}$ is the so called ALE derivative and $\mathbf{w}$ is a lifting of the velocity at Σ^t in Ω_f^t. Typically (but not necessarily) this lifting is obtained by solving a Poisson problem (harmonic lifting). At the inlet and outlet sections, proper boundary conditions have to be prescribed. In particular, it is important to use *absorbing* boundary conditions at the outlet to avoid unphysical solutions. The two matching conditions enforced at the interface are $(69)_4$ (*continuity of fluid and structure velocities*) and $(69)_5$ (*continuity of stresses*).

Before moving to the problem of the estimation of the parameters, to reduce computational costs a simplified set of equations for the Fluid-Structure Interaction (FSI) problem can be used. In fact, for large arteries, the wall thickness is in general significantly smaller than the dimension of the lumen, so that the arterial wall can be described as a 2D surface rather than a 3D structure. If we also assume that the displacement occurs only in the normal direction[12] it can be shown [57] that the structure equations reduce to

$$\rho_s h_s \frac{\partial^2 \eta}{\partial t^2} + E\beta\eta = f_s$$

where η refers to the normal displacement on the boundary of the vessel, β is a parameter embedding both geometrical and physical properties of the membrane, whose expression is given by $\beta = \frac{h_s}{1-\nu^2}(4k_m^2 - 2(1-\nu)k_g)$. Here k_m and k_g are respectively the mean and gaussian curvature of the membrane and h_s is the wall thickness. Discretizing the problem in time (using for instance backward Euler) and imposing the conservation of the normal stresses on the

[12] This assumption may be questionable for arteries close to the heart (like the aortic arch), however it is in general quite acceptable.

interface, the Fluid-Membrane Interaction (FMI) problem can be written as

$$\begin{cases} \dfrac{\rho_f}{\Delta t}(\mathbf{u}^n - \mathbf{u}^{n-1}) + \rho_f((\mathbf{u}^* - \mathbf{w}^*)\cdot\nabla)\mathbf{u}^n + \nabla\cdot\boldsymbol{\sigma}_f = \mathbf{f}_f & \text{in } \Omega^* \\ \nabla\cdot\mathbf{u}^n = 0 & \text{in } \Omega^* \\ \mathbf{u}^n \ \cdot\boldsymbol{\tau} = 0 & \text{on } \Sigma^* \\ \mathbf{n}\cdot\boldsymbol{\sigma}_f\mathbf{n} + \left(\dfrac{\rho_s h_s}{\Delta t} + E^n\beta\Delta t\right)\mathbf{u}^n\cdot\mathbf{n} = \left(\dfrac{\rho_s h_s}{\Delta t}\mathbf{u}^{n-1}\cdot\mathbf{n} - E^n\beta\eta^{n-1}\right) & \text{on } \Sigma^* \\ \eta^n = \eta^{n-1} + \Delta t\mathbf{u}^n\cdot\mathbf{n} & \text{on } \Sigma^* \end{cases} \tag{70}$$

where $\boldsymbol{\tau}$ is any unit versor in the tangent space to Σ^*. Notice how the effect of the structure is now expressed as a Robin-type boundary condition for the fluid equations. The superscript * denotes a suitable extrapolation of the quantity at the time t^n. Notice that if we use a semi-implicit scheme to deal with convective and geometric non-linearities, so that $\Omega^* = \Omega^{n-1}$ and the fluid and structure equations are then decoupled (within the time step). In particular, the equation for η^n can be promptly solved once the fluid equations have been solved. We will make use of this model in the following Sections.

4.2.2 Parameter estimation problem

The displacement of a vessel can be retrieved from images properly segmented and registered in time. This means that at each available snapshots, the arterial wall is reconstructed as a triangulated surface; then, a map is properly computed to identify the image of each point at the subsequent snapshots (see e.g. [53, 61]). The map is obtained by minimizing the mismatch between the image of its application to a snapshot and the successive one. In particular we denote by τ_k the instants when images are available and by $\Delta\tau$ the length of each time interval. Once the map is computed, the displacement is promptly available. In [61] for instance the *Iterative Closest Point* criterion is used to quantify the mismatch and to compute the map. Our goal now is to estimate the compliance of the vessel such that the mismatch between the retrieved and the computed displacement is minimized. To this end, we introduce the following cost functional

$$\mathcal{J}_1 = \frac{1}{2}\sum_{k=1}^{N}\int\limits_{\Sigma}\left(\boldsymbol{\eta}_{meas}(\boldsymbol{x},\tau_k) - \boldsymbol{\eta}(\mathbf{x},\tau_k)\right)^2 d\sigma. \tag{71}$$

where $\boldsymbol{\eta}(\mathbf{x},\tau_k)$ solves equations (69) at instants τ_k and $\boldsymbol{\eta}_{meas}$ is the (noisy) observed displacement. Here, we are assuming to have a continuous displacement field $\boldsymbol{\eta}_{meas}$ defined on Ω_s. In case we only have sparse measurements of the displacement, it is reasonable to use the following cost functional

$$\mathcal{J}_2 = \frac{1}{2}\sum_{k=1}^{N}\sum_{j=1}^{M}\|\Delta\eta_{j,\,k}\|^2_{\mathrm{R}_k^{-1}}, \tag{72}$$

where $\Delta\eta_{j,\,k} = \boldsymbol{\eta}_{meas}(\mathbf{x}_j, \tau_k) - \boldsymbol{\eta}(\mathbf{x}_j, \tau_k)$, R_k^{-1} is a weight s.p.d. matrix. Should probabilistic information on the displacement be available, R_k^{-1} is the covariance matrix of the noise of the displacement retrieval process.

As anticipated, we consider two approaches to solve this problem: a deterministic variational approach and a Kalman-based approach.

Remark 4.2 *Typically, the time step Δt of the numerical scheme is smaller than the time sample $\Delta\tau$, requiring more observations than those available. A common practice is to recover the observation at needed time steps by interpolation. In the following we will use this approach.*

Variational approach In order to minimize $\mathcal{J}_1$ we can use a gradient based optimization approach as discussed in Section 3.2. However, as outlined there, the solution of an unsteady minimization problem, such as the FSI problem, would be very expensive beacuse all the steps are coupled toghether, and it would also require the evaluation of shape derivatives since the geometry is evolving in time. To reduce the computational costs and the algorithm complexity, we exploit the fact that the parameter E does not change in time and solve the following suboptimal problem. First, we discretize the system in time. Then, at each time instant t^n we solve a steady suboptimal optimization problem, finding the value E^n which minimizes the functional

$$\mathcal{J}_3^n = \frac{1}{2}\int\limits_{\Sigma} \left(\boldsymbol{\eta}_{meas}(\boldsymbol{x}, \tau_n) - \boldsymbol{\eta}(\mathbf{x}, \tau_n)\right)^2 d\sigma, \tag{73}$$

constrained by the time-discrete fluid-structure interaction problem at time t^n. Finally, we compute E as the average of E^n: $E = \frac{1}{N}\sum_{n=1}^{N} E^n$.

Numerical solution For the sake of clarity, we focus on the simplified membrane model (70), already discretized in time. However, note that the optimization strategy described in the following, has been applied to the original FSI problem (69) in [59]. When considering the membrane approximation, the cost functional J_3^n becomes

$$\mathcal{J}_m^n = \frac{1}{2}\int\limits_{\Sigma} \left(\eta_{meas}^n - \eta^n\right)^2 d\sigma,$$

and the adjoint of problem (70) reads [13]

$$\begin{cases} \dfrac{\rho_f}{\Delta t}(\boldsymbol{\chi}^n - \mathbf{u}^{n-1}) - \rho_f((\mathbf{u}^* - \mathbf{w}^*)\cdot\nabla)\boldsymbol{\chi}^n + \nabla\cdot\boldsymbol{\sigma}_f(\boldsymbol{\chi}^n) = \Delta t(\eta_{meas}^n - \eta^n)\mathbf{n} & \text{in } \Omega^* \\ \nabla\cdot\boldsymbol{\chi}^n = 0 & \text{in } \Omega^* \\ \boldsymbol{\chi}^n \ \cdot\boldsymbol{\tau} = 0 & \text{on } \Sigma^* \\ \mathbf{n}\cdot\boldsymbol{\sigma}_f(\boldsymbol{\chi})\mathbf{n} + \left(\dfrac{\rho_s h_s}{\Delta t} + E^n\beta\Delta t\right)\boldsymbol{\chi}^n\cdot\mathbf{n} = 0 & \text{on } \Sigma^* \\ \eta^n = \eta^{n-1} + \Delta t\boldsymbol{\chi}^n\cdot\mathbf{n} & \text{on } \Sigma^*. \end{cases}$$

[13] see (37), and note that here the adjoint variable is denoted with $\boldsymbol{\chi}$ as in this context ρ is used for the density.

The gradient of the cost functional with respect to the parameter E^n is obtained using the adjoint variable $\boldsymbol{\chi}$ and relation (38), which, for the problem at hand reads

$$\left.\frac{D\mathcal{J}_m^n}{DE}\right|_{E^n} = -\int\limits_{\Sigma^*} \beta\left(\Delta t\mathbf{u}^n\cdot\mathbf{n} + \eta^{n-1}\right)(\boldsymbol{\chi}^n\cdot\mathbf{n})d\sigma.$$

The optimization is performed using the BFGS method. In particular, at each time step, for a given initial guess of the parameter $E^{n,(0)}$, the BFGS method iteratively provides parameter guesses $E^{n,(j)}$, based on the values of the cost functional $\mathcal{J}_m^n\left(E^{n,(j-1)}\right)$ and its derivative $\left.\dfrac{D\mathcal{J}_m}{DE}\right|_{E^{n,(j-1)}}$. The iterative procedure stops when the norm of $\left.\dfrac{D\mathcal{J}_m}{DE}\right|_{E^{n,(j-1)}}$ is less than a given tolerance.

Remark 4.3 *BFGS is a method devised for unconstrained optimization, while the problem at hand features the constraint $E > 0$. Unilateral constraints can be managed as indicated in [58]. Here, we include this, with a simple change of variable, by using as a control variable $\psi = \log(E)$, so that $E = \exp(\psi) > 0$ for every ψ.*

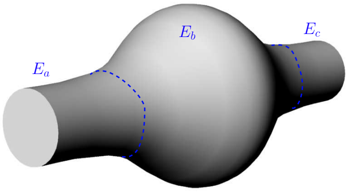
Numerical results on a simplified geometry representing an abdominal aneurysm In the numerical results presented in this Section, we will use the simplified membrane model (70). The optimization strategy depicted above, however, can be equally applied to this simplified problem.

We consider a 2D axisymmetric geometry which represents an abdominal aneurysm (see Figure 11, top-left). The radius of the vessel varies from 1 cm to 2.5 cm and the vessel length is 6 cm. We perform a synthetic simulation in which we prescribe the piecewise linear Young's modulus shown in Figure 11 (bottom-left). For the forward simulation, we take $E_a = 4\cdot 10^6\, dyne/cm^2, E_b = 10^7\, dyne/cm^2, E_c = 5\cdot 10^6\, dyne/cm^2$. We prescribe at the inlet a parabolic profile for the velocity, whose maximum u_{max} lies on the axis of symmetry and it is given by

$$u_{max} = u_{max}^0 + A\max\left\{\sin\left(\frac{2\pi t}{T}\right); 0\right\},$$

where $u_{max}^0 = 5\ cm/s$, $A = 55\ cm/s$ and $T = 0.6\ s$. At the outlet we prescribe the absorbing boundary conditions proposed in [57]. We run the simulation for two heart beats, i. e. for $0 < t \leq 2T$. We add a uniform noise ν_P to the forward displacement η_{fwd} and we use the result as data for the control problem. In Figure 11 (bottom-right) we report a comparison between the displacement obtained with the forward simulation, the noisy data and the computed displacement at time $t = 0.96\ s$. The agreement is very good.

In Table 4, we report the average, over the 10 realizations, of the estimated values of E_a, E_b and E_c and the number of times the state and the adjoint problem have needed to be solved. Different noise percentage P are considered. The initial guess is $E_{a,0} = E_{b,0} = E_{c,0} = 2\cdot 10^7 dyne/cm^2$. The estimated



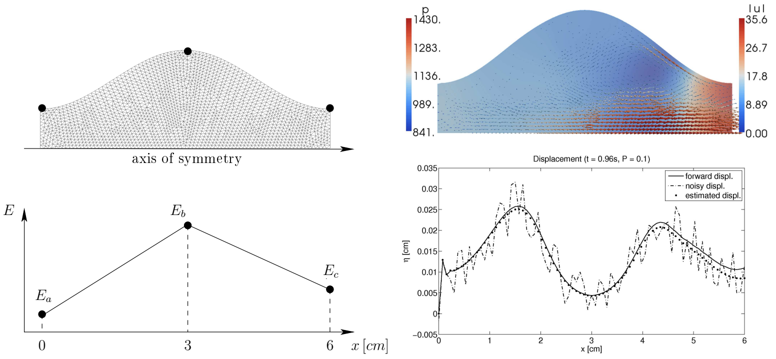


Figure 11: Aneurysm simulation. Top-left: mesh used for the simulation. Bottom-left: piecewise linear approximation of the Young's modulus E in the forward simulation. Top-right: Velocity vectors and pressure at time $t = 0.96$ s. Bottom-right: Comparison between the displacement obtained with the forward simulation, the noisy data and the computed displacement, at time $t = 0.96$ s and for $P = 0.1$.

SNR	E_a	E_b	E_c	$iter.(state\|adjoint)$
34	4.047 ± 0.118 (1.2%)	10.19 ± 0.295 (1.9%)	5.194 ± 0.240 (3.9%)	12.9|3.5
17	4.034 ± 0.281 (0.9%)	10.40 ± 0.505 (4%)	5.507 ± 0.584 (10%)	14.8|3.8
8	4.200 ± 0.550 (5%)	10.89 ± 0.850 (8.9%)	–	16.0|4.2

Table 4: Noisy case. Mean and standard deviation (to be multiplied by 10^6) of the ten estimates for E_a, E_b, E_c and number of state and adjoint iterations (bottom) for different values of the noise percentage P. The initial guess is $E_{a,0} = E_{b,0} = E_{c,0} = 2 \cdot 10^7 dyne/cm^2$.

values for $P = 0.1$ and $P = 0.2$ are quite accurate. For $P = 0.3$ we do not find a converged value for E_c. To overcome this problem, we add a regularization term to the cost functional $\mathcal{J}_3^n$, penalizing values of E far from the initial guess. Table 5 shows that the regularization term is effective. The estimates for E_c are still the more sensible to the noise, but now the estimated values are acceptable. In the first time steps of the simulation, the displacements computed by the FSI solver are very small for $x > 3\ cm$, hence the data is dominated by the noise in that region. This fact can be an explanation of the high sensibility to the noise of the estimated value for E_c.

SNR	E_a	E_b	E_c	$iter.(state\|adjoint)$
34	4.032 ± 0.119 (0.8%)	10.15 ± 0.320 (1.5%)	5.123 ± 0.129 (2.5%)	13.1|3.7
17	4.222 ± 0.238 (5.5%)	10.17 ± 0.510 (1.7%)	5.349 ± 0.368 (7.0%)	14.2|3.6
11	4.446 ± 0.426 (11%)	10.57 ± 0.780 (5.7%)	7.036 ± 3.90 (41%)	15.5|4.1
8.3	4.386 ± 0.570 (9.6%)	11.09 ± 1.519 (11%)	7.802 ± 4.12 (56%)	16.9|4.1

Table 5: Noisy case with regularization term. Mean and standard deviation (to be multiplied by 10^6) of the ten estimates for E_a, E_b, E_c and number of state and adjoint iterations (bottom) for different values of the noise percentage P. The initial guess is $E_{a,0} = E_{b,0} = E_{c,0} = 10^7 dyne/cm^2$.

Reduction of the Computational Costs via POD In this Section we show an example of how the POD procedure explained in Section 3.3 can be used for reducing the computational costs of the problem of the estimation of the Young's modulus explained in the previous paragraph. We will assume that the structure is approximated as a 2D membrane, which allows us to use the simplified model (70). Furthermore, we divide the structure in a predetermined number k of regions along the axial direction of the vessel, and we consider the case where the Young's modulus is globally piecewise constant, with a constant value in each (predetermined) region. This choice is driven by both practical and theoretical reasons. On one hand, it can allow us to model the scenario where, due for instance to the presence of some pathology, the local properties of the tissue are altered. On the other hand, this choice guarantees the existence of a solution for the inverse problem, as shown in [59].

To show why a POD approach is reasonable, let us consider a flow in a cylinder, where the membrane has been divided in three regions in which the Young's modulus is constant, and let us consider for the inflow a sinusoidal pressure wave of the type

$$p(t) = 500 \sin(50\pi t).$$

We solve the forward problem for different values of each of the Young's moduli in the three regions and we compute the correlation matrix of the Finite

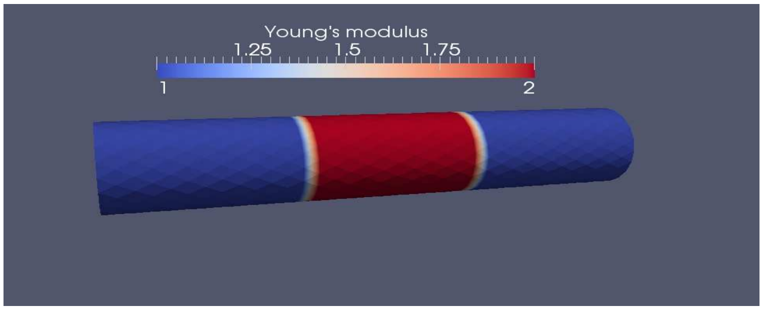


Figure 12: Example of piecewise constant (along the axial direction) Young's modulus.

Element snapshots for fluid velocity, denoted $\mathbf{u}_h$, and membrane displacement, denoted by η_h. The number of degrees of freedom is 9186 for the fluid velocity and 3540 for the membrane displacement.

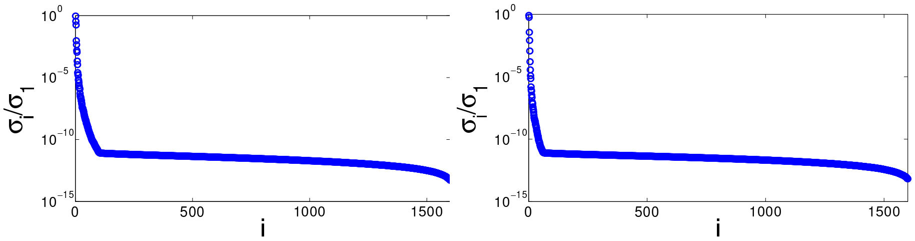


Figure 13: Singular values of the velocity (left) and membrane displacement (right) snapshot matrix. Here we chose $E = (E_1, E_2, E_3)$, with $E_i \in \{1, 1.5, 2\} \cdot 10^6$ dyn/cm^2, and performed 60 time step.

The figures suggest that the unknowns can be well approximated by vectors belonging to spaces of dimensions much lower than the corresponding Finite Element ones. Therefore, it is natural to think to POD, as a possible strategy to reduce the computational costs.

At time t^n the fully discrete Inverse Fluid-Membrane Interaction (IFMI) problem reads: find E^n_{opt} such that

$$\begin{aligned} E^n_{opt,h} &= \arg\min_{E\in\mathbb{R}^k_+} \mathcal{J}^n(E) \\ s.t. \quad & \begin{bmatrix} \mathrm{C}(E) & \mathrm{B}^T & \mathrm{O} \\ \mathrm{B} & \mathrm{O} & \mathrm{O} \\ -\Delta t\mathrm{P} & \mathrm{O} & \mathrm{I} \end{bmatrix} \begin{bmatrix} \mathbf{u}^n_h \\ \delta p^n_h \\ \eta^n_h \end{bmatrix} = \begin{bmatrix} f^n_h(E_h) \\ 0 \\ \eta^{n-1}_h \end{bmatrix} \end{aligned} \tag{74}$$

where P is a projection matrix, that extracts the normal component of the computed velocity on Σ. Here the dependence of the velocity matrix C and the velocity right hand side f_h on E_h comes from the Robin boundary conditions in (70). We also point out that here we switched to an incremental formulation for the pressure, and we included the term $\mathrm{B}^T p_h^{n-1}$ in $f_h^n(E_h)$ for the sake of brevity.

To generate the POD basis, we solve the forward problem (70) for different values of the Young's modulus and we store the snapshots at each timestep. By this we mean that $\mathbf{u}_h^n$ and $\mathbf{u}_h^{n+1}$ are two different snapshots, even if they corresponds to the same Young's modulus. However, we expect the solution to change smoothly in time, and therefore the singular values of the snapshots matrix should decay fast. This is confirmed by the numerical experiments, as we showed in Figure 13.

When building our reduced order model for the FMI problem, we want to exploit the divergence-free velocity snapshots. Assume we have stored the velocity vectors in the matrix W_u. When we project the momentum equation onto the range of W_u we obtain

$$\mathrm{W}_u^T \mathrm{C}\mathbf{u}^n + \mathrm{W}_u^T \mathrm{B}^T \delta p_h^n = \mathrm{W}_u^T f_h^n \tag{75}$$

Should the geometry be constant in time, then the product $\mathrm{W}_u^T \mathrm{B}^T = (\mathrm{B}\mathrm{W}_u)^T$ would be identically zero, being the discrete space divergence-free, and the pressure increment term would disappear. When the geometry is moving, this is not true, since each snapshot is strictly speaking divergence free only in the geometry in which it was computed. However, for a small time step (and for small displacements) we do expect the increment δp^n to be small. For the sake of the computational costs, we drop the pressure correction term in the reduced problem. This can be regarded as an explicit treatment of the pressure in the time advancing scheme. Once the reduced momentum equation has been solved, the pressure can be recovered by solving the least square problem in the full Finite Element space, that is,

$$p_h^n = \min_{q_h \in Q_h} \; ||f_h^n - \mathrm{C}u_h^n - \mathrm{B}^T q||^2 \tag{76}$$

The solution to this problem exists and is unique, provided that the velocity and pressure FE spaces satisfy the *inf-sup* condition, which guarantees that B^T has full column rank. In order to have a representation of the pressure in the reduced space, one has to make sure that the reduced saddle point problem is non-singular. In literature this issue has been tackled by enriching the velocity reduced space [65].

We therefore construct the reduced basis only for the fluid velocity and membrane displacement fields. To this end, we solve the forward problem for a given set of Young moduli $E_1, \dots E_M$ and store the corresponding solutions (snapshots) $\mathbf{u}_{h,i}, \eta_{h,i}$. In order to deal with non-homogeneous boundary conditions at the inflow/outflow sections, we modify the velocity snapshots in the following way

$$\hat{\mathbf{u}}_{h,i} = \mathbf{u}_{h,i} - \mathbf{u}_\ell \tag{77}$$

where $\mathbf{u}_\ell$ is the solution of a *steady* rigid-wall Stokes problem used as a *lifting function* for the non-homogeneous boundary conditions. This choice allows us to preserve the divergence-free nature of the snapshots which are then collected (amended by the lifting) in the snapshots matrices X_u and X_η. We compute the SVD of these matrices and let W_α be the matrices containing the first k_α left singular vectors of X_α ($\alpha = u, \eta$), with k_α such that

$$\sum_{i=1}^{k_\alpha} \sigma_i \geq \tau \sum_{i=1}^{\mathcal{N}_\alpha} \sigma_i \tag{78}$$

where σ_i are the singular values of X_α, τ is the fraction of data variability that we want to capture (typically we take $\tau = 0.9, 0.95$ or 0.99) and $\mathcal{N}_\alpha$ is the dimension of the FE space. The columns of W_u and W_η form the reduced basis for the fluid velocity and membrane displacement spaces.

If we project the IFMI problem (74) onto the reduced space, we then obtain

$$\begin{aligned} E_h^n = \quad & \arg\min_{E_h \in \mathbb{R}^k} \mathcal{J}_r^n(E_h) = \frac{1}{2}||\eta_{r,h}^n - d_r^n||_\Sigma^2 + \mathcal{R}(E_h) \\ & \text{s.t.} \begin{bmatrix} \mathrm{C}_r & \mathrm{O} \\ -\Delta t \mathrm{P}_r & \mathrm{I} \end{bmatrix} \begin{bmatrix} \mathbf{u}_r^n \\ \eta_r^n \end{bmatrix} = \begin{bmatrix} f_{r,h}^n \\ \eta_{r,h}^{n-1} \end{bmatrix} \end{aligned} \tag{79}$$

where $\mathrm{C}_r = \mathrm{W}_u^T \mathrm{C} \mathrm{W}_u$, $\mathrm{M}_r = \mathrm{W}_\eta^T \mathrm{M}_\Sigma \mathrm{W}_\eta$, $\mathrm{P}_r = \mathrm{W}_\eta^T \mathrm{P} \mathrm{W}_u$, $f_{r,h}^n = \mathrm{W}_u^T (f_h^n - \mathrm{B}^T p_h^*)$, $d_{r,h}^n = \mathrm{W}^T \eta_h \eta_{meas}^n$, and the dependence of C and f_h^n on E is understood for brevity.

The minimization problem is then solved as in the previous Section, using the BFGS method. In particular, in order to evaluate the functional and its gradient, we solve the state and adjoint problems respectively, which are given by

$$\begin{cases} \begin{bmatrix} \mathrm{C}_r^T & -\Delta t \mathrm{P}_r^T \\ \mathrm{O} & \mathrm{I} \end{bmatrix} \begin{bmatrix} \lambda_u \\ \lambda_\eta \end{bmatrix} = \begin{bmatrix} 0 \\ -\mathrm{M}_r(\eta_r^n - d_r^n) \end{bmatrix} & \text{(State)} \\ \\ \begin{bmatrix} \mathrm{C}_r & \mathrm{O} \\ -\Delta t \mathrm{P}_r & \mathrm{I} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{r,h}^n \\ \eta_{r,h}^n \end{bmatrix} = \begin{bmatrix} f_{r,h}^n \\ \eta_{r,h}^{n-1} \end{bmatrix} & \text{(Adjoint)} \end{cases} \tag{80}$$

Numerical results on an idealized aortic arch In this Section we study the flow in a curved pipe resembling the shape of an idealized aortic arch. In particular, the geometry consists of a half torus joint with a cylinder. We chose the major and minor radii of the torus (i.e. the distance between the center of the torus and the centerline of the pipe and the radius of the pipe respectively) to be $R = 1.5cm$ and $r = 0.5cm$, while the length of the cylindrical part is $L = 5cm$. At the inflow/outflow sections we prescribe the Neumann conditions

$$p\boldsymbol{n} - \nu \left(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T \right) = g\boldsymbol{n}$$

with $g = 0$ at the outflow and $g(t) = 500 \sin(100\pi t)$ at the inflow.

As in the previous Section, we solve the forward problem for a given Young's modulus and we store the corresponding membrane displacement. This will

τ	0.9	0.95	0.99
N_u	5	8	22
N_η	5	7	12

Table 6: Dimension of the fluid velocity and membrane displacement POD basis for different values of the POD threshold for the idealized aortic arch test case.

provide the synthetic measures to be used in the DA procedure. In order to not commit an "inverse crime" we solve the forward problem on a finer mesh, then we add some noise to the computed membrane displacement and we project it on the (coarser) mesh used for the solution of the inverse problem. In other words, we use the measures given by

$$\eta_m = \Pi\, \eta_{f,h} + ||\eta_{f,h}||_\infty \xi e$$

where Π is a projection from the fine to the coarse mesh, $\eta_{f,h}$ is the displacement computed on the fine mesh, $e \sim \mathcal{U}(-1,1)$ is a random vector and ξ is the noise level, reciprocal of the (SNR).

The Young modulus used to generate the measures is $E = [1.3, 1.8, 1.3] \times 10^6$ dyn/cm^2, assuming a piecewise constant profile along the axial direction. In particular, E_1 is the value of the Young modulus for the first quarter of the torus, E_2 is the value for the second quarter, and E_3 is the value in the cylindrical part. For the generation of the POD basis, we use the sample $S = \{E \in \mathbb{R}^3 : E_i \in \{1,2\} \times 10^6 \text{ dyn/cm}^2\}$. In Table 6 we report the dimension N_u, N_η of the velocity and displacement POD basis for different choices of the POD threshold τ.

In this test we compared the reduced space approach with the full space approach (i.e., the minimization in the full Finite Element space). The history of the estimates at each time step for the case SNR=10 and $\tau = 0.95$ is shown in Figure 14, while in Table 7 we report their performance. The optimal estimate for the Young's modulus is computed by averaging all but the first 10 time steps estimates, which are clearly significantly affected by the initial guess.

We can see in addition that the reduced space approach estimates are as good as the full space approach. Moreover, the error on the estimates is remarkably smaller than the intensity of the noise in the measures, for both the approaches, dropping from 10% to about 2%, that shows also how DA filters the noise in the measures. Regarding the behavior of the estimates with respect to the POD threshold, in Table 8 we report the time averages (excluding the first 10 time steps) and the corresponding relative error for three different POD thresholds.

Finally, in Figure 15 we show the history of the Young modulus estimates for different choices of the POD threshold in the case of SNR=5. It is interesting to notice that, despite the fact that the level of the noise is as large as 20% of the intensity of the signal, the average estimates are still close to the correct values. In particular, even when using a low dimensional size for the reduced model, the optimization procedure clearly detects that the Young modulus in the second region is larger than in the other two regions.

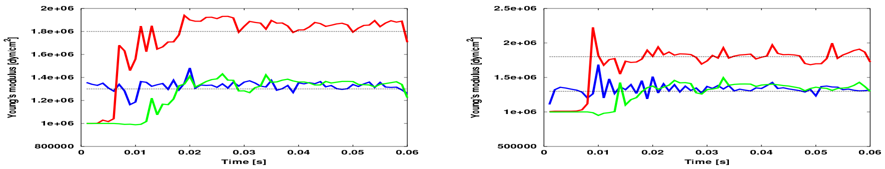


Figure 14: History of the Young modulus estimates for SNR=10. On the left, the estimates obtained by solving the inverse problem on the full finite element space. On the right, the estimates obtained by first projecting the problem on the reduced space. The colors refer to the different components of the vector E (E_1 (blue), E_2 (red), E_3 (green)).

	FS	RS
E	$[1.33,\ 1.84,\ 1.31]\times 10^6$	$[1.34,\ 1.80,\ 1.33]\times 10^6$
rel. error	1.91%	2.01%
exec. time	3176s	277s
NS solves	492s	480s

Table 7: Comparison between Full Space (FS) and Reduced Space (RS) performance for the idealized aortic arch test case.

	$\tau = 0.9$	$\tau = 0.95$	$\tau = 0.99$
E	$[1.37,\ 1.81,\ 1.32]\times 10^6$	$[1.34,\ 1.80,\ 1.32]\times 10^6$	$[1.30,\ 1.78,\ 1.29]\times 10^6$
rel. error	2.83%	1.97%	0.87%

Table 8: Time average of the estimates and relative error for different values of the POD threshold for the idealized arch test case.

4.2.3 A Kalman-based Parameter Estimation Approach

Let us consider the FSI system after time-space discretization and linearization that we write as

$$\boldsymbol{U}^k = A^{k-1}\boldsymbol{U}^{k-1} + \boldsymbol{F}^{k-1},$$

where $\boldsymbol{U}^k \in \mathbb{R}^N$ is the vector of velocity and pressure degrees of freedom. In order to estimate the parameter $\boldsymbol{E} \in \mathbb{R}^p$ the augmented state approach is used. Define $\boldsymbol{X}^{(k)} := [\boldsymbol{U}^{(k)}, \boldsymbol{E}^{(k)}]$, then the system becomes

$$\boldsymbol{X}^{(k)} = A_X^{(k-1)}\boldsymbol{X}^{(k-1)} + \boldsymbol{F}_X^{(k-1)}, \qquad A_X^{(k)} = \begin{bmatrix} A^{(k)} & 0 \\ 0 & I \end{bmatrix}, \qquad \boldsymbol{F}_X^{(k)} = \begin{bmatrix} \boldsymbol{F}^{(k)} \\ 0 \end{bmatrix}. \tag{81}$$

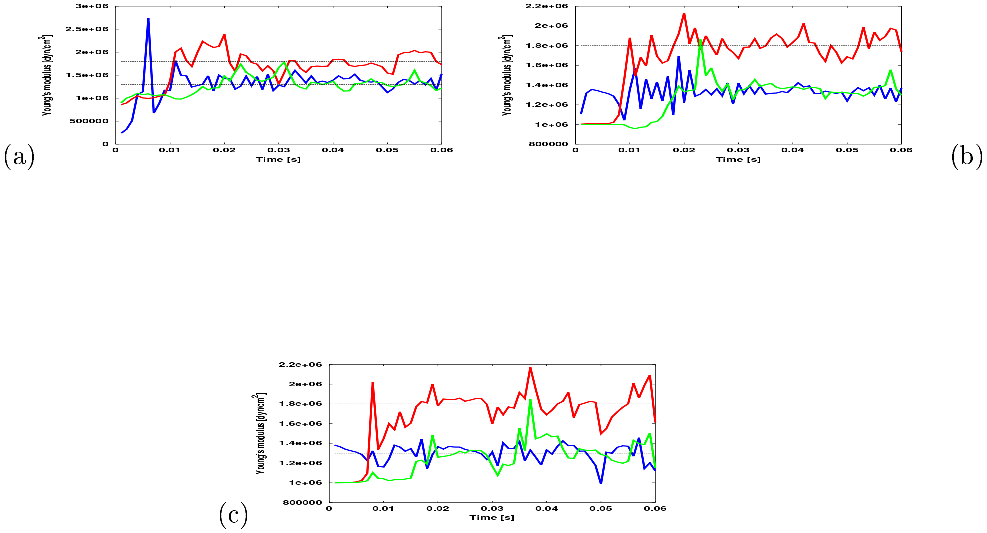


Figure 15: History of the Young modulus estimates for SNR=5 for different values of the POD threshold for the aortic arch test case. (a) $\tau = 0.9$ ($N_u = 5,\ N_\eta = 5$), (b) $\tau = 0.95$ ($N_u = 8,\ N_\eta = 7$), (c) $\tau = 0.99$ ($N_u = 22,\ N_\eta = 12$). The colors refer to the different components of the vector E (E_1 (blue), E_2 (red), E_3 (green)).

The initial state is assumed to be $\boldsymbol{X}^{(0)} = [\boldsymbol{U}^{(0)}, \boldsymbol{E}_{ref} + \boldsymbol{\theta}^{(0)}]$, and $\mathbf{b}^{(k)} = \mathbf{0}$, i.e. the initial velocity and displacement are assumed to be known without uncertainty and the model is considered exact. The variables to be estimated are denoted by $\boldsymbol{\theta}$. The measures of the displacement are affected by a white noise $\boldsymbol{\nu}^{(k)}$, i.e.

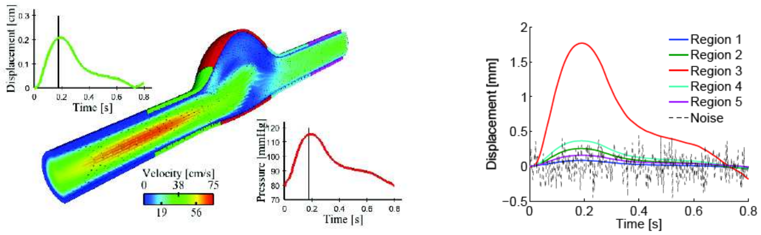
$$\boldsymbol{\eta}_{meas}^{(k)} = H_k \boldsymbol{X}^{(k)} + \boldsymbol{\nu}^{(k)}.$$

Since the problem is nonlinear, an unscented Kalman filter approach (see Section 2.5) is used where $(N + p + 1)$ sample points (for details see [6]) are needed to approximate the average and the covariance of the evolving state. As explained in Section 2.5, the predictor phase consists in evaluating $\boldsymbol{X}_i^{(k)}$ for each sample $\boldsymbol{X}_i^{(k-1)}$, which requires the solution of the FSI problem $(N+p+1)$ times at each time step. This is computationally prohibitive, therefore a model reduction is performed. The idea is to exploit the fact that the initial covariance is given by

$$\Lambda^{(0)} = \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{Cov}\left(\boldsymbol{\theta}^{(0)}\right) \end{bmatrix},$$

and to use a factorized formulation of the unscented Kalman filter. In this way [6] it is possible to use only $p + 1$ sample points, which significantly reduce the computational cost of the method when $p \ll N$, i.e. when the number of parameters is much smaller then the dimension of the state.

Consider the idealized 3D geometry of an abdominal aortic aneurysm showed in Figure 16, left. The structure is divided, a priori, in five regions featuring different values of the Young modulus E, corresponding to different colors in Figure 16.The typical displacements and noise recorded in the five regions are shown in Figure 16, right.



Region	SNR
1	0.43
2	1.25
3	8.8
4	1.8
5	0.75

Figure 16: Left: idealized abdominal aortic aneurysm geometry with subregions and fluid velocity field; displacement and pressure fields at the outlet as a function of time. Rigth: Noise compared to the typical wall displacements in the five regions and signal to noise ratios. Adapted from [6].

Numerical Results Consider the idealized 3D geometry of an abdominal aortic aneurysm showed in Figure 16, left. The length of the geometry, the minimum and maximum diameters are $23cm$, $1.7cm$ and $5cm$ respectively. The Poisson ratio, density and viscosity of the structure are 0.46, 1.2 g/cm^3 and 10^{-3} s, respectively. The fluid density and viscosity are 1 g/cm^3 and 0.035 Po, respectively. A Windkessel boundary condition is used at the outflow (see 16 for details). We assume to have displacement measures at each grid point of the mesh and that $\boldsymbol{\nu}^{(k)} \sim \mathcal{N}(0, \sigma^2 I)$. In analogy with the variational approach, comparing the FE discretization of the cost functional (71) and the cost functional (72), we assume the covariance matrix R_k to be inversely proportional to the M_k^Σ, the mass matrix on Σ. In particular we take

$$R_k^{-1} = \beta\sigma^{-2}\frac{\tau_m}{T_{ref}}\frac{M_k^\Sigma}{|\Sigma|},$$

where τ_m is the time sampling of the measurements, T_{ref} is a reference time, and β is a positive scalar used to weigth the importance of the measurements. Also we assume that $\theta^{(0)} \sim N(0, \alpha I)$. Figure 17, top, shows the reconstructed Young modulus in the different regions, as a function of α and β. The coefficient β represent the level of confidence attributed to the displacement measures, whereas α is the *a priori* covariance. As expected, the sensitivity with respect to β is higher when α is larger and the sensitivity with respect to α is higher in regions with smaller SNR. Together with the estimated parameters, the Kalman filter provides also their covariances, which is an important index to evaluate the confidence we should have in the results. The results are in fact more (less) reliable when the covariances are small (large). The estimated Young moduli and the corresponding standard deviations are showed in Figure 17, bottom.

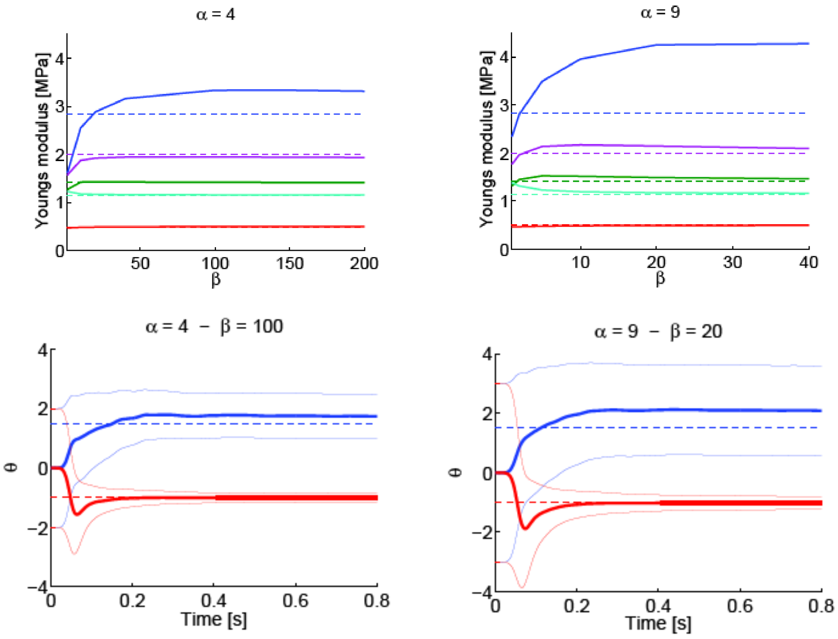


Figure 17: Top: Estimated Young modulus in the five regions as a function of β, for $\alpha = 4$ (left) and $\alpha = 9$ (right). The dashed lines represent the reference values. Bottom: mean values (thick solid lines) and plus/minus standard deviations (thin solid lines) of the logarithm of estimated Young modulus, for $\alpha = 4$, $\beta = 100$ (left) and $\alpha = 9, \beta = 20$ (right). Adapted from [6].

With respect to the variational method, the filtering approach has the advantages that only the solution of the forward problem is needed and that it provides an estimate of the covariance of the parameters. Also, it is computationally cheaper when the parameter space is much smaller than the state space. However, the nonlinearities are not solved accurately and this can lead to a suboptimal estimate of the parameters. Also, when the space of the parameters is large (e.g. $\boldsymbol{E}$ is a finite element field with as many DOFs as the number of grid points), the Kalman approach may become expensive.

5 Conclusions

Cardiovascular Mathematics is nowadays a mature discipline not only for understanding and improving basic knowledge of diseases, but also for supporting the clinical practice, with an accurate quantitative estimate, prediction, identification of optimal therapies. In particular, the common denominator of this exciting perspective is the presence of *inverse problems*, where problems related to blood flow and FSI, traditionally per se challenging, need to be solved several times, assimilated to available measures, analyzed with probabilistic tools. This is true not only for DA, but also for the identification of the optimal realization of a therapy or, more specifically, of a surgical intervention. For instance, in [56, 50, 49] the identification of the optimal placement of leads for optimizing pace making action in the heart is addressed; the computation of a personalized patient-specific peritoneal dialysis is addressed in [62], Chapter 7. In Fig. 18 we report the aortic blood flow simulated with different options of a Left Ventricular Assisted Device (LVAD) implant; in particular, the emphasis is on the location of the cannula from the pump. The identification of the optimal location is still an open problem whose solution certainly depends on the patient-specific morphology.

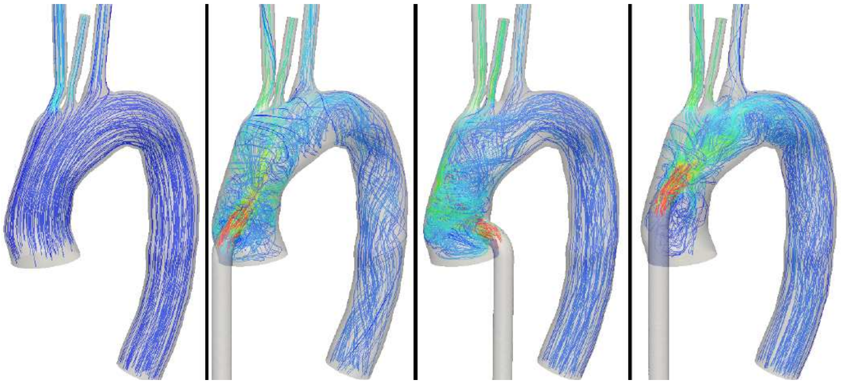


Figure 18: Simulation of different locations of the cannula of a LVAD in a real aorta. Leftmost: pre-op fluid dynamics (images provided by D. Gupta, Emory University Hospital, image processing and simulations in collaboration with M. Piccinelli (Radiology, Emory University) and T. Passerini (Math & CS, Emory University).

This process bringing complex quantitative analyses from the computer to the bedside requires a strong integration with available data, shifting the goal of performing a patient-specific computation to the patient-specific “assimilation” [71]. This is a crucial step for improving reliability of numerical elaborations, reducing uncertainty and eventually the risks of failure.

Several methods can be pursued to this goal and extensive investigation is required to establish the most appropriate approach for the different problems. A genuinely numerical-statistical research is necessary for understanding how to reduce the computational costs to be able to tackle challenges presented by clinical problems that typically feature short timelines and large number of patients.

This Chapter intended to offer a short introduction with a special emphasis on FSI problems to some possible methods and to their interplay. Far to be a conclusive and exhaustive presentation, we aimed at turning on interest for the emerging topic of Inverse Cardiovascular Mathematics, with the final - ambitious but possible - goal of introducing mathematically advanced methods in the clinical practice to improve doctors activity and - more importantly - patients healthcare.

References

[1] Hisham Abou-Kandil, Gerhard Freiling, Vlad Ionescu, and Gerhard Jank. *Matrix Riccati Equations: In Control and Systems Theory*. Springer, 2003.

[2] H. T. Banks and K. Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Birkhauser, 1989.

[3] H.T. Banks. *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*. Taylor & Francis, 2012.

[4] P.E. Barbone and A.A. Oberai. Elastic modulus imaging: some exact solutions of the compressible elastography inverse problem. *Physics in Medicine and Biology*, 52:1577, 2007.

[5] P.E. Barbone, C.E. Rivas, I. Harari, U. Albocher, A.A. Oberai, and Y. Zhang. Adjoint-weighted variational formulation for the direct solution of inverse problems of general linear elasticity with full interior data. *International Journal for Numerical Methods in Engineering*, 81(13):1713–1736, 2010.

[6] Cristóbal Bertoglio, Philippe Moireau, and Jean-Frederic Gerbeau. Sequential parameter estimation for fluid–structure problems: Application to hemodynamics. *International Journal for Numerical Methods in Biomedical Engineering*, 28(4):434–455, 2012.

[7] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. Waanders, K. Willcox, and Y. Marzouk. *Large-Scale Inverse Problems and Quantification of Uncertainty*. Wiley Series in Computational Statistics. Wiley, 2011.

[8] Jacques Blum, François-Xavier Le Dimet, and I. Michael Navon. Data assimilation for geophysical fluids. In P.G. Ciarlet, editor, *Handbook of Numerical Analysis*, volume 14 of *Handbook of Numerical Analysis*, pages 385–441. Elsevier, 2009.

[9] P.B. Bochev. Analysis of least-squares finite element methods for the navier-stokes equations. *SIAM Journal on Numerical Analysis*, 34:1817–1844, 1997.

[10] P.B. Bochev and M.D. Gunzburger. Accuracy of least-squares methods for the navier-stokes equations. *Computers & Fluids*, 22:549–563, 1993.

[11] P.B. Bochev and M.D. Gunzburger. *Least-Squares Finite Element Methods*. Springer, 2009.

[12] Paul T. Boggs and Jon W. Tolle. Sequential quadratic programming. *Acta Numerica*, 4:1–51, 1 1995.

[13] D. Calvetti and E. Somersalo. *An Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing.* Surveys and Tutorials in the Applied Mathematical Sciences. Springer Science+Business Media, 2007.

[14] Ian Campbell and W. Robert Taylor. Flow and atherosclerosis. In *Hemodynamics and Mechanobiology of Endothelium.* World Scientific, 2010.

[15] Dominique Chapelle, Asven Gariah, and Jacques Sainte-Marie. Galerkin approximation with proper orthogonal decomposition : new error estimates and illustrative examples. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46:731–757, 7 2012.

[16] R H Clayton, O Bernus, E M Cherry, H Dierckx, F H Fenton, L Mirabella, A V Panfilov, F B Sachse, G Seemann, and H Zhang. Models of cardiac tissue electrophysiology: progress, challenges and open questions. *Prog. Biophys. Mol. Biol.*, 104(1-3):22–48, 2011.

[17] M. D'Elia, L. Mirabella, T. Passerini, M. Perego, M. Piccinelli, C. Vergara, and A. Veneziani. *Some applications of variational data assimilation in computational hemodynamics*, volume D. Ambrosi, A. Quarteroni and G. Rozza Eds, Modelling of Physiological Flows of *MS&A series*, pages 363–394. Springer Verlag, 2011.

[18] M. D'Elia, M. Perego, and A. Veneziani. A variational data assimilation procedure for the incompressible navier stokes equations in hemodynamics. *J Sci Comp*, 52(2):340–359, 2012.

[19] Marta D'Elia and Alessandro Veneziani. Uncertainty quantification for data assimilation in a steady incompressible navier-stokes problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47:1037–1057, 7 2013.

[20] J. Donea, S. Giuliani, and J.P. Halleux. An arbitrary lagrangian-eulerian finite element method for transient dynamic fluid-structure interactions. *Computer Methods in Applied Mechanics and Engineering*, 33(1–3):689–723, 1982.

[21] R.P. Dwight. Bayesian inference for data assimilation using least-squares finite element methods. In *IOP Conference Series: Materials Science and Engineering*, volume 10, page 012224. IOP Publishing, 2010.

[22] B. Einarsson. *Accuracy and reliability in scientific computing*, volume 18. Society for Industrial Mathematics, 2005.

[23] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Mathematics and Its Applications. Springer, 1996.

[24] L. Formaggia, A. Quarteroni, and A. Veneziani. *Cardiovascular Mathematics: Modeling and simulation of the circulatory system*, volume 1. Springer Verlag, 2009.

[25] L. Formaggia, A. Veneziani, and C. Vergara. A new approach to numerical solution of defective boundary value problems in incompressible fluid dynamics. *SIAM Journal on Numerical Analysis*, 46(6):2769–2794, 2008.

[26] L. Formaggia, A. Veneziani, and C. Vergara. Flow rate boundary problems for an incompressible fluid in deformable domains: Formulations and solution methods. *Comp Meth Appl Mech Eng*, 9(12):677–688, 2010.

[27] B. Fristedt, N. Jain, and N.V. Krylov. *Filtering and Prediction: A Primer*. Number v. 10 in Filtering and prediction: a primer. American Mathematical Society, 2007.

[28] Kenichi Funamoto and Toshiyuki Hayase. Reproduction of pressure field in ultrasonic-measurement-integrated simulation of blood flow. *International Journal for Numerical Methods in Biomedical Engineering*, 29(7):726–740, 2013.

[29] GP Galdi, AM Robertson, R. Rannacher, and S. Turek. Hemodynamical flows: Modeling, analysis and simulation. oberwolfach seminar series vol. 35, 2007.

[30] J.F. Gerbeau and D. Lombardi. Reduced-order modeling based on approximated lax pairs. Technical Report RR 8137, INRIA, arXiv:1211.4153v1, November 2012.

[31] E. Gilboa, P.S. La Rosa, and Arye Nehorai. Estimating electrical conductivity tensors of biological tissues using microelectrode arrays. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, page 1040–1044, 2012.

[32] R. Glowinski and J. L. Lions. Exact and approximate controllability for distributed parameter systems. *Acta Numerica*, 3:269–378, 1994.

[33] R. Glowinski and J. L. Lions. Exact and approximate controllability for distributed parameter systems. *Acta Numerica*, 4:159–328, 1995.

[34] Roland Glowinski, Jacques-Louis Lions, and Jiwen He. *Exact and Approximate Controllability for Distributed Parameter Systems: A Numerical Approach (Encyclopedia of Mathematics and Its Applications)*. Cambridge University Press, New York, NY, USA, 1 edition, 2008.

[35] G.H. Golub and C.F. Van Loan. *Matrix computations*, volume 3. Johns Hopkins Univ Pr, 1996.

[36] L.S. Graham and D. Kilpatrick. Estimation of the bidomain conductivity parameters of cardiac tissue from extracellular potential distributions initiated by point stimulation. *Ann Biomed Eng*, 38(12):3630–3648, 2010.

[37] M.D. Gunzburger. *Perspectives in flow control and optimization*, volume 5. Society for Industrial Mathematics, 2003.

[38] Per Christian Hansen. *Rank-deficient and discrete ill-posed problems.* SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[39] JJ Heys, TA Manteuffel, SF McCormick, M. Milano, J. Westerdale, and M. Belohlavek. Weighted least-squares finite elements based on particle imaging velocimetry data. *Journal of Computational Physics*, 229(1):107–118, 2010.

[40] K. Hinsch. 3-dimensional particle velocimetry. *Measurement Sci. Tech.*, 6:742–753, 1995.

[41] Thomas J.R. Hughes, Wing Kam Liu, and Thomas K. Zimmermann. Lagrangian-eulerian finite element formulation for incompressible viscous flows. *Computer Methods in Applied Mechanics and Engineering*, 29(3):329–349, 1981.

[42] J. Humpherys, P. Redd, and J. West. A fresh look at the kalman filter. *SIAM Review*, 54(4):801–823, 2012.

[43] H. Delingette M. Sermesant R. Cabrera-Lozoya C. Tobon-Gomez P. Moireau R.M. Figueras i Ventura K. Lekadir A. Hernandez M. Garreau E. Donal C. Leclercq S.G. Duckett K. Rhode C.A. Rinaldi A.F. Frangi R. Razavi D. Chapelle N. Ayache S. Marchesseau. Personalization of a cardiac electromechanical model using reduced order unscented kalman filtering from regional volumes. *Medical Image Analysis*, 17:816–829, 2013.

[44] Simon J. Julier and Jeffrey K. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *Proc. SPIE 3068, Signal Processing, Sensor Fusion, and Target Recognition VI, 182*, pages 182–193, 1997.

[45] S.J. Julier and J.K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[46] T. Kailath. *Lectures Notes on Wiener and Kalman Filtering.* Springer-Verlag, 1981.

[47] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems (Applied Mathematical Sciences) (v. 160).* Springer, 1 edition, December 2004.

[48] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME J.Basic Engrg*, 82:35–45, 1960.

[49] K. Kunisch and M. Wagner. Optimal control of the bidomain system (iii): Existence of minimizers and first-order optimality conditions. Technical Report SFB-Report No. 2011/031, TU Graz, http://math.uni-graz.at/mobis/publications/SFB-Report-2011-031.pdf, 2011.

[50] Karl Kunisch and Marcus Wagner. Optimal control of the bidomain system (ii): uniqueness and regularity theorems for weak solutions. *Annali di Matematica Pura ed Applicata*, pages 1–36, 2012.

[51] Peter Lancaster and Leiba Rodman. *Algebraic Riccati Equations*. Oxford Science Publications, 1995.

[52] Toni Lassila, Andrea Manzoni, Alfio Quarteroni, and Gianluigi Rozza. A reduced computational and geometrical framework for inverse problems in hemodynamics. *International Journal for Numerical Methods in Biomedical Engineering*, 29(7):741–776, 2013.

[53] J. Modersitzki. *FAIR: Flexible Algorithms for Image Registration*. Fundamentals of Algorithms. Society for Industrial and Applied Mathematics, 2009.

[54] P. Moireau and D. Chapelle. Reduced-order unscented kalman filtering with application to parameter identification in large-dimensional systems. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(02):380–405, 2011.

[55] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, 1974.

[56] Chamakuri Nagaiah, Karl Kunisch, and Gernot Plank. Numerical solutions for optimal control of monodomain equations. *PAMM*, 9(1):609–610, 2009.

[57] F. Nobile and C. Vergara. An effective fluid-structure interaction formulation for vascular dynamics by generalized Robin conditions. *SIAM J Sc Comp*, 30(2):731–763, 2008.

[58] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, April 2000.

[59] M. Perego, A. Veneziani, and C. Vergara. A variational approach for estimating the compliance of the cardiovascular tissue: An inverse fluid-structure interaction problem. *SIAM Journal on Scientific Computing*, 33(3):1181–1211, 2011.

[60] Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. Technical report, http://matrixcookbook.com, 2008.

[61] Marina Piccinelli, Lucia Mirabella, Tiziano Passerini, Eldad Haber, and Alessandro Veneziani. 4d image-based cfd simulation of a compliant blood vessel. Technical report, Technical Report TR-2010-27, Department of Mathematics & CS, Emory University, www. mathcs. emory. edu, 2010.

[62] A. Quarteroni, L. Formaggia, and A. Veneziani. *Complex Systems in Biomedicine*. Springer, 2007.

[63] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. Texts in Applied Mathematics Series. Springer-Verlag GmbH, 2000.

[64] G. Rozza, D.B.P. Huynh, and A.T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):229–275, 2008.

[65] G. Rozza and K. Veroy. On the stability of the reduced basis method for stokes equations in parametrized domains. *Computational Methods in Applied Mechanics and Engineering*, 196(7):1244–1260, 2007.

[66] S. Salsa. *Partial differential equations in action: from modelling to theory.* Springer Verlag, 2008.

[67] O. Scherzer. The use of morozov's discrepancy principle for tikhonov regularization for solving nonlinear ill-posed problems. *Computing*, 51(1):45–60, 1993.

[68] R. Todling. Estimation theory and foundations of atmospheric data assimilation. *DAO Office Note*, 1:1999, 1999.

[69] F. Tröltzsch. *Optimal control of partial differential equations: theory, methods, and applications*, volume 112. Amer Mathematical Society, 2010.

[70] Karsten Urban and Anthony T. Patera. A new error bound for reduced basis approximation of parabolic partial differential equations. *Comptes Rendus Mathematique*, 350(3–4):203–207, 2012.

[71] A. Veneziani and C. Vergara. Inverse problems in cardiovascular mathematics: toward patient-specific data assimilation and optimization. *International Journal for Numerical Methods in Biomedical Engineering*, 29(7):723/725, 2013. Editorial of the special issue "Inverse Problems in Cardiovascular Mathematics".

[72] C.R. Vogel. *Computational Methods for Inverse Problems.* Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, 2002.

[73] E.A. Wan and R. Van der Merwe. The unscented kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 153–158, 2000.

[74] H. Yang and A. Veneziani. Variational estimation of cardiac conductivities by a data assimilation procedure. Technical Report TR-2013-007, Math&CS, Emory University, July 2013.