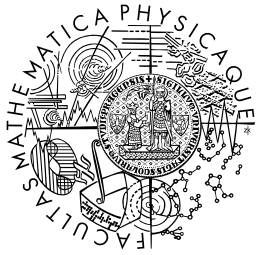

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Mgr. Ivan Kasanický

Ensemble Kalman filter on high and infinite dimensional spaces

Department of Probability and Mathematical Statistics

Supervisor of the doctoral thesis:	doc. RNDr. Daniel Hlubinka, Ph.D.
Consultant of the doctoral thesis:	prof. RNDr. Jan Mandel, CSc.
Study programme:	Mathematics
Study branch:	Probability and Mathematical Statistics

Prague 2016

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague, December, 20, 2016 Mgr. Ivan Kasanický

Title: Ensemble Kalman filter on high and infinite dimensional spaces

Author: Mgr. Ivan Kasanický

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Daniel Hlubinka, Ph.D., Department of Probability and Mathematical Statistics

Consultant: prof. RNDr. Jan Mandel, CSc., Department of Mathematical and Statistical Sciences, University of Colorado Denver

Abstract: The ensemble Kalman filter (EnKF) is a recursive filter, which is used in a data assimilation to produce sequential estimates of states of a hidden dynamical system. The evolution of the system is usually governed by a set of differential equations, so one concrete state of the system is, in fact, an element of an infinite dimensional space.

In the presented thesis we show that the EnKF is well defined on a infinite dimensional separable Hilbert space if a data noise is a weak random variable with a covariance bounded from below. We also show that this condition is sufficient for the 3DVAR and the Bayesian filtering to be well posed. Additionally, we extend the already known fact that the EnKF converges to the Kalman filter in a finite dimension, and prove that a similar statement holds even in a infinite dimension.

The EnKF suffers from a low rank approximation of a state covariance, so a covariance localization is required in real applications. The recently proposed spectral diagonal ensemble Kalman filter (SDEnKF) allows a natural localization of the state covariance, and is still computationally very efficient. We show that, under reasonable assumptions, the SDEnKF uses a much better estimate of the state covariance than the classical EnKF, and test the performance of the SDEnKF using multiple chaotic models.

Keywords: ensemble Kalman filter, data assimilation, Hilbert spaces, covariance localization

First and foremost, I would like to express my sincere gratitude to Jan Mandel for the guidance and support of my Ph.D. study, and for his generous invitation to spend nearly six months working with him at University of Colorado Denver.

My special thanks goes to Daniel Hlubinka, who supervised also my bachelor's and master's theses, and provided multiple important remarks to enhance this thesis.

I thank Martin Vejmelka for helping with the spectral diagonal ensemble Kalman filter, and Emil Pelikán, who built a great research team that I was lucky to join.

I would like to thank Kryštof Eben, Pavel Juruš, and Marie Turčičová for regular Tuesday discussions, which significantly improved the thesis.

I am grateful to the rest of the team at the Insititute of Computer Science, namely: Marek Brabec, Ondřej Konár, Pavel Krč, Marek Malý and Jaroslav Resler, who helped me to become a researcher.

Last but not the least, I would like to thank to my parents, my sister and Andrejka for supporting me throughout writing this thesis, and throughout the years spent at Charles University.

The work was supported by the Czech Science Foundation under grant No. GA13-34856S, and by the grant SVV 2016 No. 260334.

Contents

1. Introduction

1.1 Motivation and main goals

Nowadays, computers allow us to create and use sophisticated models of complex phenomenas in nature. These models are usually based on partial differential equations, which have to be discretized on appropriate meshes, and often only a numerical solution of these equations is known. Typical examples are models describing an evolution of the atmosphere (Jacobson [2005], Kalnay [2003]) or models describing an ocean current (Bennett [1992, 2002]). As an available computer performance increases, finer meshes are used for numerical computations, so the models are able to capture finer scale effects, e.g., convective rainfalls in the atmospheric models.

Nevertheless, currently used models are still far away from being perfect, and contain errors from multiple sources:

- errors caused by a necessary simplification of differential equations that govern a modeled system,
- errors caused by a numerical method used to solve the equations,
- errors due to unknown or uncertain boundary conditions,
- etc.

Therefore, the models have to be verified against available observations. Also, these observations have to be regularly used to correct forecasts produced by these models, and to augment the forecasts towards the right trajectories. The procedure of using observations to correct outputs of a numerical model is called a data assimilation.

As one can imagine, there are many obstacles with assimilation of the observations into the numerical models. The most important obstacles are summarized in the following list.

1. Available supercomputers and, already mentioned, finer meshes cause the size of the state vector to surpass any imaginable number. For example, the size of a state vector of the Aladin model, which is used for operational numerical weather forecasting in the Czech Republic, exceeds 10 million (Bénard et al. [2010]), and recently there have been done experiments with atmospheric models with dimension exceeding 100 billion (Miyoshi et al. [2014]). Therefore, many basic mathematical operations, e.g., manipulation with a full covariance matrix of the state vector, may easily become inapplicable even if one has access to the most powerful computer.

2. The observations are, similar to models, imperfect, and contain, at least, two types of errors: measurement errors and representation errors (Desroziers et al. [2005]). Hence, one never knows what the precise value of the modeled state is.

3. The observations are usually only a function of the state, and one cannot observe the modeled variables directly. For example, we may model a state of the atmosphere using air temperature, wind speed, pressure and amount of water in the atmosphere, and observe the amount of precipitation. Additionally, the observations are usually aggregated both in space and in time.

4. In many real world applications, the size of the state vector is much larger than the number of available observations. As we have already noted, a model of atmosphere can easily have a state vector of size one billion, but, without satellite images, the number of unique observations of the atmosphere at a given time interval is usually equal to a few thousands, which is obviously incomparable with the size of the state vector.

Many assimilation methods have been proposed and used over the time. One of the basic methods is the Kalman filter (Kalman [1960], Kalman and Bucy [1961]), but this method requires a manipulation with a state covariance, and that may be, as already mentioned, often impossible. Therefore, this thesis is focused on the ensemble Kalman filter, which is an assimilation method originally proposed in Evensen [1994], and later improved in Houtekamer and Mitchell [1998].

A basic idea of the ensemble Kalman filter is to use an ensemble, i.e., a set of possible trajectories of a modeled system, to represent the uncertainty in the model, and to incorporate the observations using the Kalman filter equation with a sample covariance in place of the true covariance. This approach may be implemented very efficiently, and extensive discussion of possible implementation is provided in Evensen [2009]. It is very well known (Le Gland et al. [2011], Mandel et al. [2011]) that the solution using the ensemble Kalman filter converges to the solution obtained using the Kalman filter almost surely and in $\mathcal{L}^p$ when the size of the ensemble goes to infinity and the state is finitely dimensional.

However, a state of many of the modeled phenomena, e.g., state of the atmosphere, is, in fact, a random continuous function, so it is an element of an infinite dimensional space. Hence, the interesting question is whether the ensemble Kalman filter and other well known assimilation methods may be used if we do not discretize the state. Also, some authors, e.g., Adcock and Hansen [2015], Cotter et al. [2010], Stuart [2010], argue that the discretization should be postponed to the least possible moment, so studying the ensemble Kalman filter on an infinite dimensional space is of a primary interest.

Additionally, satellites allow us to obtain high resolution images of the Earth, so one can easily assume that the observations are also elements of an infinite dimensional space. Yet, there immediately comes a question whether the ensemble Kalman filter may still be defined, and whether it possesses the same properties as in the finite dimension.

When the ensemble Kalman filter is used in a real world application, the size of the ensemble is usually much smaller than the size of a state vector. Therefore, a sample covariance is a very low rank approximation of a true covariance, and this poor covariance estimate often causes spurious correlations. Hence, many authors, e.g, Buehner [2011], Buehner and Charron [2007], Kepert [2009], Nerger et al. [2012], recommend using some kind of a covariance localization in the ensemble Kalman filter equations.

According to the previous paragraphs, we state the two main goals that we try to accomplish in this thesis.

1. To define the ensemble Kalman filter on an infinite dimensional space, and to prove results similar to Le Gland et al. [2011] but in infinite dimensions. Additionally, we will look whether the definition of the infinite dimensional ensemble Kalman filter corresponds to some other assimilation methods, such as the the 3DVAR or the Bayesian filtration.

2. To investigate properties of recently proposed FFT ensemble Kalman filter (Mandel et al. [2010b]) and DWT ensemble Kalman filter (Beezley et al. [2011]), which provide a natural localization of a forecast covariance. These methods are computationally very efficient, and are believed to require smaller ensembles than the classical ensemble Kalman filter. We will extend these methods for a usage with a general spectral transformation, and test the performance of these methods using some toy models.

1.2 Thesis outline

The thesis is divided into nine chapters. Each chapter except the first and the last one is concluded by a section that details additional notes and related references.

The first chapter contains a motivation with a few references, a thesis outline, and a notation convention.

The second chapter is devoted to a brief review of a functional analysis and a measure theory, and the main topics of this chapters are Hilbert spaces, measurable spaces and cylindrical sets. Basics definitions are recalled, but nearly all statements are given without a proof as they can be found in any textbook. A list of useful textbooks is presented in the last section of the chapter.

The third chapter introduces a probability on a separable Hilbert space. Similar to the second chapter, this chapter is a review of an already known theory, so a lot of statements are again presented without a proof. However, because the theory of random variables on an infinite dimensional space is usually not covered in basic statistical courses, we explain the problems with measurability of random variables, and review the theory of weak random variables. A large part of the chapter is devoted to Gaussian distributions, which are of our primary interest in subsequent chapters.

The fourth chapter builds a state space model, which is a basic mathematical framework for a data assimilation.

Chapters 2, 3 and 4 just review already known statements, and readers familiar with these topics may skip these chapters, but we strongly recommend to look at Definition 7 because it establishes the notation used in the all consecutive chapters.

The fifth chapter summarizes well known assimilation methods: the 3DVAR, the Kalman filter, the ensemble Kalman filter and the Bayesian filtration. Relations between solutions obtained using these methods are only briefly mentioned as they can be found in many of the references noted in the last section of the this chapter.

The sixth chapter discusses whether the methods introduced in the previous chapter may be defined when both state and observational spaces are infinite

dimensional. Sufficient and necessary conditions for some assimilation methods to be well posed are also stated.

The seventh chapter is devoted to the properties of the ensemble Kalman filter in infinite dimension. We identify the mean-field ensemble, and show that the analysis obtained using the ensemble Kalman filter converges to the mean-field ensemble in $\mathcal{L}^p$.

The eighth chapter proposes an advanced ensemble Kalman filter, which uses a diagonal sample covariance in an appropriate spectral space. We compare errors of this advanced estimate of the covariance with a classical sample covariance. Also, the performance of the proposed assimilation method is tested using multiple chaotic models.

Finally, the last chapter briefly summarizes the results obtained in this thesis.

1.3 Notation

We use calligraphic uppercase letters, e.g., $\mathcal{H}$ or $\mathcal{G}$, to denote linear spaces. Roman uppercase letters, e.g., A or T, denote linear operators between two spaces, and, because matrices are liner operators, we use the same convention to denote matrices.

If A is a complex matrix with n row and m columns, we use the notation $(\mathrm{A})_{i,j}$ to denote the element of the matrix A in the i^{th} row and the j^{th} column, and A^* denotes a conjugate transpose of A. If A is a real matrix, A^* denotes its transpose. Deterministic vectors are denoted using italic lowercase letters, e.g., u, and vectors are assumed to be columns. If $u \in \mathbb{R}^n$, then we use notation $(u)_j$ to denote its j^{th} element.

Random elements are denoted using italic uppercase letters, e.g. X, Y, regardless of the dimension of the space on which they are defined. Italic lowercase letters, e.g., x, k, denote constants or nonrandom elements of a linear space.

Angle brackets $\langle x, y \rangle$ are used to denote an inner product between x and y, and curly brackets $\{x, y, \ldots\}$ denotes a set with elements $x, y, \ldots$ A deterministic norm is denoted using single vertical bars $|\cdot|$, and double vertical bars $\|\cdot\|$ are reserved for stochastic norms.

A list of commonly used abbreviations is included at the end of the thesis.

2. Mathematical background

This chapter recalls basic mathematical definitions and theorems that we rely on in the subsequent chapters. Section 2.1 deals with measurable spaces. Section 2.2 contains a brief introduction to the functional analysis and Hilbert spaces, and Section 2.3 introduces cylindrical sets and their properties. Finally, Section 2.4 contains references to useful textbooks covering all areas as nearly all statements in this section are presented without proofs.

2.1 Measurable spaces

Given a vector space $\mathcal{W}$ over a field $\mathbb{K}$, where either $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, a function

$$\rho : \mathcal{W} \to \mathbb{R}$$

such that:

1. $\rho(ax) = |a| \rho(x)$ for all $a \in \mathbb{K}$ and all $x \in \mathcal{W}$,
2. $\rho(x + y) \leq \rho(x) + \rho(y)$ for all $x, y \in \mathcal{W}$,
3. $\rho(x) = 0$ if and only if x is the zero vector

is called a norm on $\mathcal{W}$, and a pair $(\mathcal{W}, \rho)$ is called a normed space, $\mathcal{X} = (\mathcal{W}, \rho)$. The second property of a norm is called a triangle inequality. If $\mathcal{X}$ is a normed space, then for each element $x \in \mathcal{X}$ we use single vertical bars to denote its norm

$$|x|_{\mathcal{X}} = |x| = \rho(x),$$

and reserve double vertical bars for stochastic norms introduced later. Unless it leads to confusion, we always use single vertical bars without an additional index to denote the natural norm on a given space.

A sequence $\{x_i\}_{i \in \mathbb{N}}$ of elements belonging to a normed space $\mathcal{X}$ is said to be Cauchy if for every positive ε there exist $m_\varepsilon \in \mathbb{N}$ such that

$$|x_m - x_n| \leq \varepsilon$$

for all $m, n \geq m_\varepsilon$. The space $\mathcal{X}$ is complete if every Cauchy sequence in $\mathcal{X}$ has a limit in $\mathcal{X}$.

Given a nonempty set $\mathcal{X}$ we denote $2^{\mathcal{X}}$ the collection of all subsets of $\mathcal{X}$. We say that $\mathcal{A} \subset 2^{\mathcal{X}}$ is algebra if $\mathcal{A}$ meets two conditions:

1. both $\mathcal{X}$ and $\emptyset$ are elements of $\mathcal{A}$, and
2. if $A, B \in \mathcal{A}$, than $A \setminus B \in \mathcal{A}$.

Obviously, the second condition implies that if $A, B \in \mathcal{A}$, than both $A \cap B$ and $A \cup B$ are elements of $\mathcal{A}$. Additionally, $\mathcal{A}$ is called σ-algebra if $\{A_n\}_{n=1}^{\infty} \subset \mathcal{A}$ implies that

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$

A pair $(\mathcal{X}, \mathcal{A})$ where $\mathcal{X}$ is a nonempty vector space and $\mathcal{A}$ is an σ-algebra defined on $\mathcal{X}$ is called a measurable space. If $\mathcal{F} \subset 2^{\mathcal{X}}$, then the smallest σ-algebra containing $\mathcal{F}$ is denoted $\sigma(\mathcal{F})$. In other words, let $\{\mathcal{A}_i\}_{i\in\mathcal{I}}$, where $\mathcal{I}$ is an abstract index set, be all σ-algebras in $\mathcal{X}$, then

$$\sigma(\mathcal{F}) = \bigcap_{i\in\mathcal{I}:\ \mathcal{F}\subset\mathcal{A}_i} \mathcal{A}_i.$$

A set $F \subset \mathcal{X}$ is open if for every $x \in F$ exists $\varepsilon_x > 0$ such that

$$\{y \in \mathcal{X} :\ |x - y| < \varepsilon_x\} \subset F.$$

If $\mathcal{F}$ is a collection of all open sets in $\mathcal{X}$, $\sigma(\mathcal{F})$ is called Borel σ-algebra, and we denote it as $\mathcal{B}(\mathcal{X})$.

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. A function

$$\nu :\ \mathcal{A} \to [0, \infty) \cup \{\infty\}$$

is a countable additive measure if

1. $\nu(\emptyset) = 0$, and
2. for any countable pairwise disjoint collection of sets $\{A_n\}_{n=1}^{\infty} \subset \mathcal{A}$

$$\nu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \nu(A_n).$$

According to the definition it immediately follows

$$\nu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \nu(A_n)$$

for any $\{A_n\}_{n=1}^{\infty} \subset \mathcal{A}$. If for every $x \in \mathcal{X}$ there is an open set F_x such that $x \in F_x$ and $\nu(F_x) < \infty$, we say that the measure ν is locally finite. There is only one translation invariant locally finite non zero measure on $\mathcal{B}(\mathbb{R}^n)$, $n \in \mathbb{N}$, and this measure is called the Lebesgue measure. The Lebesgue measure on $\mathbb{R}^n$ is denoted as λ^n.

Assuming that $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{D})$ are two measurable spaces, function

$$f :\ \mathcal{X} \to \mathcal{Y}$$

is called measurable if for any $D \in \mathcal{D}$ the set

$$f^{-1}(D) = \{x \in \mathcal{X} :\ \exists y \in D, y = f(x)\}$$

belongs to $\mathcal{A}$. If the function f is measurable, then $|f|$ is measurable as well, and one can easily check that the expression

$$|f|_p = \left(\int_{\mathcal{X}} |f(x)|^p \, d\nu(x)\right)^{1/p}, \tag{2.1}$$

where ν is a locally finite measure on $\mathcal{X}$, defines a norm for any $p \geq 1$. Hence, for any $p \geq 1$ we define the space

$$L^p(\mathcal{X}, \mathcal{Y}) = \left\{ f : \ \mathcal{X} \to \mathcal{Y}; \ |f|_p < \infty \right\}, \tag{2.2}$$

and we do not distinguish between functions that differ only on a set of zero measure. More precisely, the space $L^p(\mathcal{X}, \mathcal{Y})$ is a quotient space with respect to a kernel of a functional

$$f \in L^p(\mathcal{X}, \mathcal{Y}) \to |f|_p \in \mathbb{R}.$$

If $f \in L^p(\mathcal{X}, \mathcal{Y})$ and $g \in L^q(\mathcal{X}, \mathcal{Y})$ where p and q are Hölder's coefficients, i.e., $p, q \in (1, \infty)$ and $1/p + 1/q = 1$, then the Hölder's inequality states that

$$|fg|_1 \leq |f|_p |g|_q . \tag{2.3}$$

Note that the Hölder's inequality holds even when $p = 1$ with

$$|g|_\infty = \inf_{A \subset \mathcal{X}, \nu(A)=0} \left\{ \sup_{x \in \mathcal{X} \setminus A} f(x) \right\}.$$

When P is a measure on a measurable space $(\Omega, \mathcal{A})$ such that

$$P(\Omega) = 1,$$

then we say that P is a probability measure, and a triple $(\Omega, \mathcal{A}, P)$ is a probability space.

2.1.1 Radon-Nikodym theorem

Let μ and ν be two probability measures on $(\Omega, \mathcal{A})$. We say that μ is absolutely continuous with respect to ν, denoted as $\mu << \nu$, if $\nu(A) = 0$ implies that $\mu(A) = 0$ for every $A \in \mathcal{A}$. If

$$\mu << \nu \quad \text{and} \quad \nu << \mu,$$

then we say that the measures μ and ν are equivalent, $\mu \sim \nu$. If A and B are disjunct sets such that

$$\mu(A) = 1, \ \nu(A) = 0, \ \mu(B) = 0 \text{ and } \nu(B) = 1,$$

then we say that μ and ν are singular.

Theorem 1 (Radon-Nikodym theorem)**.** *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space with two countable additive locally finite measures μ and ν. If μ is absolutely continuous with respect to ν, then there exists a measurable function*

$$\frac{d\mu}{d\nu} : \ \mathcal{X} \to [0, \infty)$$

such that for each $A \in \mathcal{A}$

$$\mu(A) = \int_A \frac{d\mu}{d\nu} d\nu.$$

The function f is called a Radon-Nikodym derivative of μ with respect to ν.

2.1.2 Lipschitz continuity

Let $\mathcal{X}$ and $\mathcal{Y}$ be two normed spaces. A function $f : \mathcal{X} \to \mathcal{Y}$ is Lipschitz continuous if there exists a positive constant l such that for all $x, y \in \mathcal{X}$

$$|f(x) - f(y)| \leq l\,|x - y|,$$

and f is locally Lipschitz continuous at $x_0 \in \mathcal{X}$ if there exist $l > 0$ and $\delta_{x_0} > 0$ such that

$$|f(x_0) - f(y)| \leq l\,|x_0 - y|$$

for all $y \in \{x \in \mathcal{X} : |x_0 - x| \leq \delta_{x_0}\}$. The function f is locally Lipschitz continuous with at most polynomial growth in infinity if there exist positive constants s and l such that

$$|f(x) - f(y)| \leq l\,|x - y|\,(1 + |x|^s + |y|^s) \tag{2.4}$$

for all $x, y \in \mathcal{X}$.

Lemma 2. *If function $f : \mathcal{X} \to \mathcal{Y}$ is locally Lipschitz continuous with at most polynomial growth in infinity, then f is locally Lipschitz continuous at every $x \in \mathcal{X}$.*

Proof. By assumption there are constants $s, l > 0$ such that (2.4) holds for any $x, y \in \mathcal{X}$. Let $x_0 \in \mathcal{X}$. Then, for all $y \in \{x \in \mathcal{X} : |x_0 - x| \leq 1\}$

$$\begin{aligned} |f(x_0) - f(y)| \leq & l\,|x_0 - y|\,(1 + |x_0|^s + |y|^s) \\ \leq & l\,(1 + |x_0|^s + (|x_0| + 1)^s)\,|x_0 - y| \\ \leq & \tilde{l}\,|x_0 - y|, \end{aligned}$$

where $\tilde{l} = l\,(1 + |x_0|^s + (|x_0| + 1)^s)$, and f is locally Lipschitz continuous at x_0. □

Lemma 3. *If $f : \mathcal{X} \to \mathcal{Y}$ is locally Lipschitz continuous with at most polynomial growth in infinity, then for every $x, y \in \mathcal{X}$*

$$|f(x)| \leq l_1\left(1 + |x|^{s+1}\right), \tag{2.5}$$

$$|f(x) - f(y)| \leq l_2\,|x - y|\,(1 + |x|^s) + l_2\,|x - y|^{s+1}, \tag{2.6}$$

where l_1, l_2 are real positive constants, and s is defined by Equation (2.4).

Proof. Recalling the existence of positive constants s and l fulfilling (2.4), the proof of (2.5) relies simply on the triangle inequality,

$$\begin{aligned} |f(x)| \leq & |f(x) - f(0)| + |f(0)| \leq l\,(|x - 0|)\,(1 + |x|^s + |0|^s) + |f(0)| \\ \leq & l\left(|x| + |x|^{s+1}\right) + |f(0)| \leq l_1\left(1 + |x|^{s+1}\right), \end{aligned}$$

where the existence of the constant l_1 is obvious.

To prove (2.6), let $x, y \in \mathcal{X}$. If $|y| \leq 2\,|x - y|$, then

$$\begin{aligned} |f(x) - f(y)| \leq & l\,|x - y|\,(1 + |x|^s + |y|^s) \\ \leq & l\,|x - y|\,(1 + |x|^s) + 2^s l\,|x - y|. \end{aligned}$$

On the other hand, if $|y| > 2|x-y|$, then

$$|y| \leq |y-x| + |x| \leq \frac{1}{2}|y| + |x|,$$

which implies that $|y| \leq 2|x|$, and

$$\begin{aligned} |f(x) - f(y)| &\leq l|x-y|\left(1 + |x|^s + 2^s|x|^s\right) \\ &\leq l(1+2^s)|x-y|\left(1+|x|^s\right). \end{aligned}$$

To conclude the proof, define $l_2 = l(1+2^s)$. □

2.2 Functional analysis

When $\mathcal{X}$ is a complete normed vector space, then we say that $\mathcal{X}$ is a Banach space.

Example 1. For any real sequence $y = \{y_i\}_{i\in\mathbb{N}}$, $y \in \mathbb{R}^\infty$, and any $p \geq 1$ is ℓ^p norm defined by

$$|y|_p = \left(\sum_{i=1}^{\infty} |y_i|^p\right)^{1/p},$$

and space

$$\ell^p = \left\{y \in \mathbb{R}^\infty : \ |y|_p < \infty\right\}$$

obtained with the appropriate ℓ^p norm is Banach for any $p \geq 1$.

Example 2. The quotient space $L^p(\mathbb{R}, \mathbb{R})$ defined by (2.2) with the norm

$$|f|_p = \left(\int_{\mathbb{R}} |f|^p \, d\lambda\right)^{1/p}$$

is Banach when $p \geq 1$.

2.2.1 Inner product and Hilbert space

Given a vector space $\mathcal{H}$ defined over a scalar field $\mathbb{K}$, where again $\mathbb{K}$ stands for either $\mathbb{C}$ or $\mathbb{R}$, an inner product is a mapping from $\mathcal{H} \times \mathcal{H}$ into the scalar field $\mathbb{K}$ which associates a scalar $\langle x, y\rangle \in \mathbb{K}$ to every pair of $x, y \in \mathcal{H}$ and fulfills four conditions:

1. $\forall x, y, z \in \mathcal{H} : \ \langle x+y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$,
2. $\forall x, y \in \mathcal{H}, \forall a \in \mathbb{K} : \ \langle ax, y\rangle = a\langle x, y\rangle$,
3. $\forall x, y \in \mathcal{H} : \ \langle x, y\rangle = \overline{\langle y, x\rangle}$,
4. $\forall x \in \mathcal{H} : \ \langle x, x\rangle \geq 0$, and $\langle x, x\rangle = 0$ if and only if $x = 0$.

If $\mathbb{K} = \mathbb{R}$, then the third condition changes to

$$\forall x, y \in \mathcal{H} : \ \langle x, y \rangle = \langle y, x \rangle ,$$

and this condition also implies that $\langle x, x \rangle \in \mathbb{R}$ even if $\mathbb{K} = \mathbb{C}$. Therefore, it can easily be verified that a mapping

$$x \in \mathcal{H} \to \sqrt{\langle x, x \rangle} \in \mathbb{R}$$

defines norm on the space $\mathcal{H}$. For every $x \in \mathcal{H}$ we denote its norm generated by the inner product

$$|x|_{\mathcal{H}} = \sqrt{\langle x, x \rangle}$$

or simply $|x|$, unless this notation leads to confusion.

A pair $(\mathcal{H}, \langle ., . \rangle)$, where $\mathcal{H}$ is a complete space with respect to the norm induced by the inner product, is called a Hilbert space. It obvious that every Hilbert space is also Banach. We say that $\mathcal{H}$ is separable if it contains countable dense subset.

Example 3. The space $(\ell^2, |.|_2)$, defined in Example 1, is a separable Hilbert space with a inner product defined by

$$\langle x, y \rangle = \sum_{i=1}^{\infty} (x_i y_i)$$

for any $x, y \in \ell^2$. This space is separable as well, because the set

$$\{x \in \mathbb{Q}^{\infty} : \quad |x|_2 < \infty\}$$

is both dense and countable.

Example 4. The space $L^2(\mathbb{R}, \mathbb{R})$, defined in Example 2, is Hilbert with the norm generated by the inner product

$$\langle f, g \rangle = \int_{\mathbb{R}} f g d\lambda.$$

We say that two elements x, y of a Hilbert space $\mathcal{H}$ are orthogonal if their inner product is zero, $\langle x, y \rangle = 0$. Additionally, if their norms equal to one,

$$|x| = |y| = 1,$$

we say that x and y are orthonormal. A set $M \subset \mathcal{H}$ is called orthogonal if all elements of M are pairwise orthogonal, i.e.,

$$\forall x, y \in M, x \neq y : \ \langle x, y \rangle = 0.$$

If, additionally, all elements of M have norm one, then we call the set M orthonormal.

A set $M \subset \mathcal{H}$ is called a total orthonormal set if it is orthonormal and its span is dense in $\mathcal{H}$. Every nonempty Hilbert space contains a total orthonormal set, e.g., [Kreyszig, 1989, p. 168]. If M_1 and M_2 are two total orthonormal sets in $\mathcal{H}$, both M_1 and M_2 have the same cardinality.

Let $\mathcal{S}$ be a proper subspace of $\mathcal{H}$. The subspace

$$\mathcal{S}^{\perp} = \{x \in \mathcal{H} : \quad \langle x, y \rangle = 0 \ \forall y \in \mathcal{S}\}$$

is called a orthogonal complement of $\mathcal{S}$ in $\mathcal{H}$. For each $h \in \mathcal{H}$ there is exactly one $x_h \in \mathcal{S}$ and exactly one $y_h \in \mathcal{S}^{\perp}$ such that

$$h = x_h + y_h.$$

2.2.2 Linear operators

When $\mathcal{H}$ and $\mathcal{G}$ are two Hilbert spaces, we denote $[\mathcal{H},\mathcal{G}]$ the space of all continuous linear operators from $\mathcal{H}$ to $\mathcal{G}$. This space is equipped with the operator norm

$$|\mathrm{T}|_{[\mathcal{H},\mathcal{G}]} = \sup_{x\in\mathcal{H},x\neq 0} \frac{|\mathrm{T}x|_{\mathcal{G}}}{|x|_{\mathcal{H}}}$$

for all $\mathrm{T}\in[\mathcal{H},\mathcal{G}]$. This norm could be equivalently defined by

$$|\mathrm{T}|_{[\mathcal{H},\mathcal{G}]} = \sup_{x\in\mathcal{H},|x|_{\mathcal{H}}\leq 1} \frac{|\mathrm{T}x|_{\mathcal{G}}}{|x|_{\mathcal{H}}} = \sup_{x\in\mathcal{H},|x|_{\mathcal{H}}=1} \frac{|\mathrm{T}x|_{\mathcal{G}}}{|x|_{\mathcal{H}}},$$

and $[\mathcal{H},\mathcal{G}]$ equipped with the operator norm is a Banach space. The operator norm in not induced by an inner product, and we use convention

$$[\mathcal{H}] = [\mathcal{H},\mathcal{H}].$$

We define

$$\mathcal{H}^{\#} = \{f:\ \mathcal{H}\to\mathbb{R};\ f \text{ linear}\},$$

the space of all linear functionals on $\mathcal{H}$, and call it an algebraic dual of $\mathcal{H}$. The space of all bounded linear functionals on $\mathcal{H}$ is called a dual space of $\mathcal{H}$, and we denote it as $\mathcal{H}^*$, i.e.,

$$\mathcal{H}^* = [\mathcal{H},\mathbb{R}].$$

Now we recall one of the most important theorems in functional analysis.

Theorem 4 (Riesz's representation theorem)**.** *Let $\mathcal{H}$ be a Hilbert space. For every $f\in\mathcal{H}^*$ there exist exactly one $z\in\mathcal{H}$ such that $f(x)=\langle x,z\rangle$ for all $x\in\mathcal{H}$. The norm of z and the operator norm of f are equal,*

$$|f|_{\mathcal{H}^*} = |z|_{\mathcal{H}}.$$

The Riesz's representation theorem is crucial, and has many useful consequences. We introduce only a few of them, and the first one is the existence of an adjoint operator. For any operator $\mathrm{T}\in[\mathcal{H},\mathcal{G}]$ we define the adjoint operator T^* by equation

$$\langle y,\mathrm{T}x\rangle_{\mathcal{G}} = \langle \mathrm{T}^*y,x\rangle_{\mathcal{H}} \quad \forall y\in\mathcal{G},\forall x\in\mathcal{H}.$$

An operator $\mathrm{T}\in[\mathcal{H}]$ is selfadjoint if $\mathrm{T}=\mathrm{T}^*$, and T is positive definite if

$$\langle x,\mathrm{T}x\rangle > m|x|$$

for all $x\in\mathcal{H}$ and some positive constant m. The operator T is positive semidefinite if

$$\langle x,\mathrm{T}x\rangle \geq 0$$

for all $x\in\mathcal{H}$.

We formulate the second consequence of the Riesz's theorem as an independent corollary.

Corollary 1. Assume that $\mathcal{H}$ is a Hilbert space and α is a bilinear form defined on $\mathcal{H} \times \mathcal{H}$ such that

$$|\alpha(x, y)| \leq a\,|x|\,|y|$$

for all $x, y \in \mathcal{H}$ and some $a \in \mathbb{R}$. There exists a unique linear operator $\mathrm{A} \in [\mathcal{H}]$ such that

$$\langle x, \mathrm{A}y\rangle = \alpha(x, y)$$

for all $x, y \in \mathcal{H}$, and $|\mathrm{A}| \leq a$.

Let x and y be two elements of $\mathcal{H}$. A tensor product of x and y is a mapping from $\mathcal{H}$ to $\mathcal{H}$ defined by

$$x \otimes y :\ v \in \mathcal{H} \to x\,\langle y, v\rangle \in \mathcal{H}.$$

Example 5. When $\mathcal{H} = \mathbb{R}^n$, the tensor product of two vectors x and y is a rank one matrix xy^* since for every $u \in \mathbb{R}^n$

$$(x \otimes y)\,u = x\,\langle y, u\rangle = x\,(y^* u).$$

Given the fact that inner product is bilinear we see that the tensor product is a linear operator,

$$(x \otimes y)\,(u{+}v) = x\,\langle y, (u{+}v)\rangle = x\,\langle y, u\rangle + x\,\langle y, v\rangle = (x \otimes y)\,(u) + (x \otimes y)\,(v).$$

If $v \neq 0$, then

$$\frac{|(x \otimes y)\,v|}{|v|} = \frac{|x\,\langle y, v\rangle|}{|v|} = \frac{|x|\,\langle y, v\rangle}{|v|} \leq \frac{|x|\,|y|\,|v|}{|v|} = |x|\,|y| \tag{2.7}$$

with equality if and only if $y = v$. Therefore,

$$|x \otimes y| = |x|\,|y|.$$

Again, using Riesz's representation theorem, the tensor product $x \otimes y$ can be equivalently written as a unique element of $[\mathcal{H}]$ such that

$$\langle u, (x \otimes y)\,v\rangle = \langle x, y\rangle\,\langle u, v\rangle \quad \forall u, v \in \mathcal{H},$$

and it is obvious that the tensor product is a selfadjoint operator.

Let $M = \{e_1, \ldots, e_m\}$ be an orthonormal set and denote $\mathcal{G} = \mathrm{span}\,(M)$. The operator

$$\Pi :\ x \in \mathcal{H} \to \sum_{i=1}^{m} (e_i \otimes e_i)\,x \in \mathcal{G}$$

is called a projection of $\mathcal{H}$ onto $\mathcal{G}$. It is obvious that for every $x \in \mathcal{G}$

$$\Pi(x) = \sum_{i=1}^{m} e_i\,\langle e_i, x\rangle = x,$$

so $\Pi^2 = \Pi \circ \Pi = \Pi$. In fact, $\Pi^n = \Pi$ for any $n \in \mathbb{N}$. Since the projection is the sum of selfadjoint operators, it is also selfadjoint.

2.2.3 Spectrum of bounded linear operators

Given an operator $\mathrm{T} \in [\mathcal{H}]$ a resolvent set $\rho(\mathrm{T})$ is a set of all $\lambda \in \mathbb{C}$ such that the operator

$$\mathrm{R}_\lambda = (\mathrm{T} - \lambda \mathrm{I})^{-1}$$

satisfies 3 conditions:

1. R_λ exists,
2. R_λ is bounded, and
3. domain of R_λ is dense in $\mathcal{H}$.

The resolvent set is open in $\mathbb{C}$, and its complement

$$\sigma(\mathrm{T}) = \mathbb{C} \setminus \rho(\mathrm{T})$$

is called the spectrum of T. Clearly, if $u \in \mathcal{H}$ is such that

$$u \neq 0 \text{ and } (\mathrm{T} - \lambda \mathrm{I})\, u = 0 \tag{2.8}$$

for some $\lambda \in \mathbb{C}$, then $\lambda \in \sigma(\mathrm{T})$, and we say that λ is the eigenvalue of T and u is the corresponding eigenvector. The set of all eigenvalues $\sigma_p(\mathrm{T})$ is called the point spectrum of T.

Example 6. When $\mathcal{H} = \mathbb{R}^n$, then $[\mathcal{H}] = \mathbb{R}^{n \times n}$, and a spectrum of a matrix $\mathrm{M} \in \mathbb{R}^{n \times n}$ is a set of complex numbers λ such that the matrix $\mathrm{M} - \lambda \mathrm{I}$ is not invertible, i.e., the spectrum only consist of the eigenvalues.

On the other hand, when $\mathcal{H} = \ell^2$, defined in Example 1, the operator

$$\mathrm{L}: \ (x_1, x_2, \ldots) \in \ell^2 \to (0, x_1, x_2, \ldots) \in \ell^2$$

is obviously linear and bounded, and the operator

$$\mathrm{R}: \ (0, x_1, x_2, \ldots) \in \ell^2 \to (x_1, x_2, \ldots) \in \ell^2$$

is obviously the inverse of $\mathrm{L} - 0.\,\mathrm{I}$. But the operator R is defined only on

$$\left\{(x_1, x_2, \ldots) \in \ell^2 : \quad x_1 = 0\right\},$$

and this set is not dense in ℓ^2. Therefore, 0 belongs to the spectrum of L, but it is not an eigenvalue of L.

The previous example shows that the spectrum of a bounded linear operator may also contain other values then eigenvalues. Additionally, it is possible to construct a linear operator with no eigenvalue, even though the spectrum is always nonempty. The spectral radius $r_\sigma(\mathrm{T})$ of T is defined by

$$r_\sigma(\mathrm{T}) = \sup_{\lambda \in \sigma(\mathrm{T})} |\lambda|, \tag{2.9}$$

and it can be shown that

$$r_\sigma(\mathrm{T}) \leq |\mathrm{T}|.$$

2.2.4 Compact linear operators

An operator $\mathrm{T} \in [\mathcal{H}, \mathcal{G}]$, where $\mathcal{H}$ and $\mathcal{G}$ are Hilbert spaces, is compact if for every bounded sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$ the sequence $\{\mathrm{T}x_n\}_{n\in\mathbb{N}} \subset \mathcal{G}$ contains a convergent subsequence.

If $\mathrm{T} \in [\mathcal{H}]$ is compact and $\dim(\mathcal{H}) = \infty$, then

$$\sigma(\mathrm{T}) = \sigma_p(\mathrm{T}) \cup \{0\},$$

and zero can be the only accumulation point of the spectrum. Additionally, if T is selfadjoint, then there is a total orthonormal set $\{e_i\}_{i\in\mathbb{N}} \subset \mathcal{H}$ such that

$$\mathrm{T}e_i = \lambda_i e_i \ \forall i \in \mathbb{N},$$

with all λ_i, $i \in \mathbb{N}$, being real, and

$$\mathrm{T} = \sum_{i=1}^{\infty} \lambda_i (e_i \otimes e_i), \tag{2.10}$$

which is a spectral decomposition of T. The spectral decomposition (2.10) allows us to define

$$f(\mathrm{T}) = \sum_{i=1}^{\infty} f(\lambda_i)(e_i \otimes e_i)$$

for any continuous $f : \mathbb{R} \to \mathbb{R}$.

If $\mathrm{T} \in [\mathcal{H}]$ is a compact operator, then $\mathrm{T}^*\mathrm{T}$ is a positive semidefinite operator, so

$$\mathrm{T}^*\mathrm{T} = \sum_{i=1}^{\infty} \sigma_i^2 (e_i \otimes e_i),$$

where $\{e_i\}_{i\in\mathbb{N}}$ is a total orthonormal set and $\sigma_i^2 \geq 0$ for all $i \in \mathbb{N}$. The values σ_i are called singular values of the operator T, and it is obvious that for symmetric positive semidefinite operators the singular values and the eigenvalues coincide.

Finally, if $\mathrm{T} \in [\mathcal{H}]$ is compact, then, similarly as in a finite dimension,

$$|\mathrm{T}| = \sup_{i\in\mathbb{N}} \sigma_i,$$

where σ_i are singular values of T. And for every $\mathrm{T} \in [\mathcal{H}]$,

$$|\mathrm{T}| = r_\sigma(\mathrm{T}^*\mathrm{T}),$$

where $r_\sigma(\mathrm{T}^*\mathrm{T})$ is defined by Equation (2.9).

We say that $\mathrm{T} \in [\mathcal{H}]$ is a trace class operator if there is an orthonormal set $\{e_i\}_{i\in\mathbb{N}} \subset \mathcal{H}$ such that

$$\mathrm{Tr}\left((\mathrm{T}^*\mathrm{T})^{1/2}\right) = \sum_{i=1}^{\infty} \left\langle (\mathrm{T}^*\mathrm{T})^{1/2} e_i, e_i \right\rangle < \infty,$$

and the value $\mathrm{Tr}\left((\mathrm{T}^*\mathrm{T})^{1/2}\right)$ is called the trace of the operator T. It can be shown that the trace of T does not depend on the choice of the orthonormal set $\{e_i\}_{i\in\mathbb{N}}$,

and the set of all trace class operators on $\mathcal{H}$ is a Banach space with a norm defined by

$$|\mathrm{T}|_{\mathrm{Tr}} = \mathrm{Tr}\left((\mathrm{T}^*\mathrm{T})^{1/2}\right).$$

Additionally, every trace class operator is compact, and

$$|\mathrm{T}|_{\mathrm{Tr}} = \sum_{i=1}^{\infty} |\sigma_i|,$$

where σ_i are singular values of T.

We say that $\mathrm{T} \in [\mathcal{H}]$ is a Hilbert-Schmidt operator if there is an orthonormal set $\{e_i\}_{i\in\mathbb{N}} \subset \mathcal{H}$ such that

$$\sum_{i=1}^{\infty} \langle \mathrm{T}^*\mathrm{T}e_i, e_i\rangle < \infty.$$

The set of all Hilbert-Schmidt operators on $\mathcal{H}$ is a Hilbert space with an inner product

$$\langle \mathrm{T}, \mathrm{S}\rangle_{\mathrm{HS}} = \sum_{i=1}^{\infty} \langle \mathrm{T}e_i, \mathrm{S}e_i\rangle \tag{2.11}$$

for any Hilbert-Schmidt operators T and S, and the value of this inner product does not depend on the choice of the set $\{e_i\}_{i\in\mathbb{N}}$. Obviously, for a Hilbert-Schmidt operator T we define a Hilbert-Schmidt norm

$$|\mathrm{T}|_{\mathrm{HS}} = \langle \mathrm{T}, \mathrm{T}\rangle_{\mathrm{HS}}^{1/2} = \left(\sum_{i=1}^{\infty} |\mathrm{T}e_i|^2\right)^{1/2}, \tag{2.12}$$

and, equivalently,

$$|\mathrm{T}|_{\mathrm{HS}} = \left(\sum_{i=1}^{\infty} |\sigma_i|^2\right)^{1/2}.$$

Lemma 5. *For any $x \in \mathcal{H}$ is*

$$|x \otimes x|_{\mathrm{HS}} = |x|^2.$$

Proof. Let $\{e_i\}_{i\in\mathbb{N}}$ be a total orthonormal set. Using the definition of a tensor product and properties of an inner product,

$$\begin{aligned} |x \otimes x|_{\mathrm{HS}}^2 &= \sum_{i=1}^{\infty} |(x \otimes x)\, e_i|^2 = \sum_{i=1}^{\infty} |x\, \langle x, e_i\rangle|^2 \\ &= |x|^2 \sum_{i=1}^{\infty} |\langle x, e_i\rangle|^2 = |x|^2\, |x|^2. \end{aligned}$$

□

Generally, for $\mathrm{T} \in [\mathcal{H}]$ compact and any $p \geq 1$ the Schatten norm of T is

$$|\mathrm{T}|_p = \left(\sum_{i=1}^{\infty} |\sigma_i|^p\right)^{1/p} \tag{2.13}$$

where $\sigma_i,\ i \in \mathbb{N}$, are singular values of T. The space of all compact $\mathrm{T} \in [\mathcal{H}]$ such that $|\mathrm{T}|_p < \infty$ is Banach for all $p \geq 1$, and we can immediately see that

$$|\mathrm{T}|_1 = |\mathrm{T}|_{\mathrm{Tr}} \text{ and } |\mathrm{T}|_2 = |\mathrm{T}|_{\mathrm{HS}}\,.$$

Additionally, for any $p \leq q$ is

$$\left\{\mathrm{T} \in [\mathcal{H}] :\ |\mathrm{T}|_p < \infty\right\} \subset \left\{\mathrm{T} \in [\mathcal{H}] :\ |\mathrm{T}|_q < \infty\right\},$$

and for every $\mathrm{T} \in L_q(\mathcal{H})$ immediately yields

$$|\mathrm{T}| \leq |\mathrm{T}|_q \leq |\mathrm{T}|_p \tag{2.14}$$

for any $1 \leq p \leq q$. The following lemma is a special case of Theorem 7.8 in Weidmann [1980].

Lemma 6. *Let $p, q, r \geq 1$ be such that $1/p + 1/q = 1/r$. If $\mathrm{T}, \mathrm{S} \in [\mathcal{H}]$ are such that*

$$|\mathrm{T}|_p < \infty \quad \textit{and} \quad |\mathrm{S}|_q < \infty,$$

then

$$|\mathrm{TS}|_r \leq 2^{1/r} |\mathrm{T}|_p |\mathrm{S}|_q\,.$$

From the previous lemma follows an important corollary. If T and S are Hilbert-Schmidt operators, then TS is a trace class operator, and

$$|\mathrm{TS}|_{\mathrm{Tr}} \leq 2 |\mathrm{T}|_{\mathrm{HS}} |\mathrm{S}|_{\mathrm{HS}}\,.$$

We say that positive definite operator

$$\mathrm{R} = \sum_{i=1}^{\infty} r_i (e_i \otimes e_i)\,,$$

where $\{e_i\}$ is a total orthonormal set is bounded from bellow if

$$r = \inf_{i \in \mathbb{N}} r_i > 0.$$

If R is bounded from bellow, then $|\mathrm{R}^{-1}| < \infty$. We conclude this subsection by an estimate which can be originally found in Kwiatkowski and Mandel [2015].

Lemma 7. *Let $\mathrm{P}, \mathrm{Q} \in [\mathcal{H}]$ be positive semidefinite operators, and $\mathrm{R} \in [\mathcal{H}]$ be bounded from bellow. Then,*

$$\left|(\mathrm{P}+\mathrm{R})^{-1} - (\mathrm{Q}+\mathrm{R})^{-1}\right| \leq |\mathrm{P}-\mathrm{Q}| \left|\mathrm{R}^{-1}\right|^2.$$

Proof. The expression

$$(\mathrm{P}+\mathrm{R})^{-1} - (\mathrm{Q}+\mathrm{R})^{-1} = (\mathrm{P}+\mathrm{R})^{-1} (\mathrm{Q}-\mathrm{P}) (\mathrm{Q}+\mathrm{R})^{-1} \tag{2.15}$$

can be verified by application of $(\mathrm{P}+\mathrm{R})$ from the left side and $(\mathrm{Q}+\mathrm{R})$ from the right side. The operator R is bounded from bellow, so $|\mathrm{R}^{-1}| < \infty$, and since both operators P and Q are positive semi-definite,

$$\left|(\mathrm{P}+\mathrm{R})^{-1}\right| \leq \left|\mathrm{R}^{-1}\right| \quad \text{and} \quad \left|(\mathrm{Q}+\mathrm{R})^{-1}\right| \leq \left|\mathrm{R}^{-1}\right|. \tag{2.16}$$

The identity (2.15) together with (2.16) now gives the result

$$\begin{aligned}\left|(\mathrm{P}+\mathrm{R})^{-1} - (\mathrm{Q}+\mathrm{R})^{-1}\right| &\leq \left|(\mathrm{P}+\mathrm{R})^{-1}\right| (\mathrm{Q}-\mathrm{P}) \left|(\mathrm{Q}+\mathrm{R})^{-1}\right| \\ &\leq |\mathrm{Q}-\mathrm{P}| \left|\mathrm{R}^{-1}\right|^2.\end{aligned}$$

□

2.2.5 Commuting operators

We say that two operators P and Q from $[\mathcal{H}]$ commute if the operator

$$\mathrm{PQ} - \mathrm{QP} = 0.$$

Lemma 8. *Suppose that* P *and* Q *are two symmetric operators belonging to* $[\mathcal{H}]$, P *and* Q *commute, and* P *is compact. Then there exist a total orthonormal set* $\{e_i\}_{i\in\mathbb{N}} \subset \mathcal{H}$ *such that* $\mathrm{P}e_i = p_i e_i$ *and* $\mathrm{Q}e_i = q_i e_i$ *where* $p_i, q_i \in \mathbb{R}$, $i \in \mathbb{N}$, *are eigenvalues of* P *and* Q *respectively.*

Proof. The operator P is compact and symmetric, so it has real eigenvalues p_i, $i \in \mathbb{N}$, with eigenvectors $u_i \in \mathcal{H}$, $i \in \mathbb{N}$, where $\{u_i\}_{i\in\mathbb{N}}$ is a total orthonormal set, and the multiplicity of each eigenvalue is finite.

Assume that p_j is unique, i.e., $p_j \neq p_i$ unless $j = i$, $i \in \mathbb{N}$. Then

$$\mathrm{PQ}u_j = \mathrm{QP}u_j = p_j \mathrm{Q}u_j,$$

so $\mathrm{Q}u_j$ is also an eigenvector of P with an eigenvalue p_j. Because p_j is distinct from all other eigenvalues, there exist $q_j \in \mathbb{R}$ such that

$$\mathrm{Q}u_j = q_j u_j,$$

and u_j is an eigenvector of Q.

Now, assume that

$$p = p_1 = p_2 = \ldots = p_m$$

for some $m \in \mathbb{N}$, and define

$$\mathcal{U} = \mathrm{span}\,\{u_1, \ldots, u_m\}.$$

Since P and Q commute,

$$\mathrm{PQ}u_i = \mathrm{QP}u_i = p\mathrm{Q}u_i$$

for all $i = 1, \ldots, m$, so $v_i = \mathrm{Q}u_i \in \mathcal{U}$ because v_i is an eigenvector of P corresponding to p. Therefore, $\mathcal{U}$ is an invariant subspace of both operators P and Q. Denote by $\mathrm{P}_{\mathcal{U}}$ and $\mathrm{Q}_{\mathcal{U}}$ restrictions of the operators P and Q respectively to the subspace $\mathcal{U}$. The operators $\mathrm{P}_{\mathcal{U}}$ and $\mathrm{Q}_{\mathcal{U}}$ are symmetric, defined on a finite dimensional vector space, and commute. Therefore, they are both diagonalizable, e.g., [Hoffman and Kunze, 1971, Chapter 8, Theorem 18], and they both have a common complete orthonormal set of eigenvectors, e.g., [Hoffman and Kunze, 1971, Chapter 6, Theorem 8], which spans $\mathcal{U}$.

Hence, both operators P and Q have a common total orthonormal set of eigenvectors. □

In fact, if two symmetric linear operators commute, and one of them has a total orthonormal set of eigenvectors, then they have a common set of orthogonal eigenvectors, e.g., Levin [2002].

2.2.6 Lebesgue measure on Hilbert space

Theorem 9. *Let $\mathcal{H}$ be an infinitely dimensional separable Hilbert space. If μ is a translation invariant measure on $\mathcal{B}(\mathcal{H})$, then either*

$$\mu(B) = 0 \quad \forall B \in \mathcal{B}(\mathcal{H})$$

or

$$\mu(B) = \infty \quad \forall B \in \mathcal{B}(\mathcal{H}).$$

Proof. For any $x \in \mathcal{H}$ and $\varepsilon > 0$ we define open neighborhood $B_{x,\varepsilon}$,

$$B_{x,\varepsilon} = \{y \in \mathcal{H} : \ |x - y| < \varepsilon\}.$$

Clearly, these sets generate Borel σ-algebra $\mathcal{B}(\mathcal{H})$. The space $\mathcal{H}$ is separable, so there is an orthonormal set $\{e_i\}_{i=1}^{\infty}$ and

$$B_{e_i,1/4} \cap B_{e_j,1/4} = \emptyset \Leftrightarrow i \neq j.$$

If

$$0 < \mu\left(B_{e_i,1/4}\right) < \infty, \tag{2.17}$$

then

$$\lambda(B_{0,2}) \geq \sum_{i=1}^{\infty} \lambda\left(B_{e_i,1/4}\right) = \infty$$

because

$$B_{0,2} \supset \bigcup_{i=1}^{\infty} B_{e_i,1/4},$$

but

$$B_{0,2} = \left\{8x : \ x \in B_{0,1/4}\right\}.$$

Therefore, the measure of the set $B_{e_i,1/4}$ is either 0 or ∞, but this is a contradiction with (2.17). □

Corollary 2. There is no Lebesgue, i.e., translation invariant and locally finite, measure on an infinite dimensional Hilbert space.

2.3 Cylindrical sets

2.3.1 Definition

For any finite subset $H = \{h_1, \ldots, h_n\}$ of a Hilbert space $\mathcal{H}$ we define a mapping

$$\pi_H : \ x \in \mathcal{H} \to (\langle x, h_1\rangle, \ldots, \langle x, h_n\rangle) \in \mathbb{R}^n.$$

Definition 1. *A set $\mathcal{C} \subset \mathcal{H}$ is a cylindrical set if there is a finite set $H = \{h_1, \ldots, h_n\} \subset \mathcal{H}$ and a Borel set $B \in \mathbb{R}^n$ such that*

$$\mathcal{C} = \pi_H^{-1}(B) = \{x \in \mathcal{H} : \ \pi_H(x) \in B\}.$$

Using the notation from the previous definition, denote Π_H the projection operator from $\mathcal{H}$ to span (H). Using the properties of the projection, every $x \in \mathcal{H}$ can be written in the form $x = \Pi_H x + (\mathrm{I} - \Pi_H)\, x$, and thus for all $i = 1, \ldots, n$

$$\langle x, h_i \rangle = \langle \Pi_H x, h_i \rangle + \langle (\mathrm{I} - \Pi_H)\, x, h_i \rangle = \langle \Pi_H x, h_i \rangle .$$

Therefore, $x \in \mathcal{C}$ if and only if $\Pi_H x \in \mathcal{C}$. This observation shows that any cylindrical set $\mathcal{C}$ can be written as a direct sum

$$\mathcal{C} = B \oplus \mathcal{S}^{\perp}$$

where $\mathcal{S}$ is a finitely dimensional subspace of $\mathcal{H}$ and $B \in \mathcal{B}(\mathcal{S})$.

Let $\mathcal{C}_1 = B_1 \oplus \mathcal{S}_1^{\perp}$ and $\mathcal{C}_2 = B_2 \oplus \mathcal{S}_2^{\perp}$ be two cylindrical sets, and denote

$$\widetilde{\mathcal{S}} = \mathcal{S}_1 \cap \mathcal{S}_2.$$

Additionally, denote $\widetilde{B}_1 = B_1 \cap \widetilde{\mathcal{S}}$ and $\widetilde{B}_2 = B_2 \cap \widetilde{\mathcal{S}}$. Then, both cylindrical sets can be decomposed

$$\mathcal{C}_1 = \widetilde{B}_1 \oplus \left(B_1 \setminus \widetilde{B}_1 \right) \oplus \mathcal{S}_1^{\perp},$$
$$\mathcal{C}_2 = \widetilde{B}_2 \oplus \left(B_2 \setminus \widetilde{B}_2 \right) \oplus \mathcal{S}_2^{\perp},$$

and the union and the intersection of these sets are

$$\mathcal{C}_1 \cap \mathcal{C}_2 = \left(\widetilde{B}_1 \cap \widetilde{B}_2 \right) \oplus \left(B_1 \cap \mathcal{S}_2^{\perp} \right) \oplus \left(B_2 \cap \mathcal{S}_1^{\perp} \right) \oplus (\mathcal{S}_1 \cup \mathcal{S}_2)^{\perp},$$
$$\mathcal{C}_1 \cup \mathcal{C}_2 = \left(\widetilde{B}_1 \cup \widetilde{B}_2 \right) \oplus \left(\mathcal{S}_1^{\perp} \cap \mathcal{S}_2 \right) \oplus \left(\mathcal{S}_1 \cap \mathcal{S}_2^{\perp} \right) \oplus (\mathcal{S}_1 \cup \mathcal{S}_2)^{\perp}.$$

We may immediately observe that both union and intersection of finitely many cylindrical sets are again cylindrical sets. Hence, all cylindrical sets constitute algebra. To see that the set of all cylindrical sets is not a σ algebra, consider sets

$$\mathcal{S}_n = \left\{ x \in \mathcal{H} : \ \left(\sum_{i=1}^{n} |\langle x, h_i \rangle|^2 \right)^{1/2} < 1 \right\}.$$

Obviously for each $n \in \mathbb{N}$ $\mathcal{S}_n$ is a cylindrical set, but

$$\bigcap_{n=1}^{\infty} \mathcal{S}_n = \{ x \in \mathcal{H} : \ |x| < 1 \}$$

is not a cylindrical set. In fact the σ-algebra generated by cylindrical sets coincides with Borel σ-algebra [Balakrishnan, 1976, Lemma 6.1.1].

2.3.2 Cylindrical measure

Let $\mathcal{I}$ be an abstract index set such that

$$\{ (\mathcal{H}_i, \nu_i) \}_{i \in \mathcal{I}} \tag{2.18}$$

is a collection of all finitely dimensional subspaces $\mathcal{H}_i \subset \mathcal{H}$ and countable additive measures ν_i defined on $\mathcal{B}(\mathcal{H}_i)$. When the collection (2.18) is such that for any two $\mathcal{H}_i \subset \mathcal{H}_j$, $i, j \in \mathcal{I}$, and any cylindrical set

$$\mathcal{C} = B \oplus \mathcal{H}_i^{\perp},\ B \in \mathcal{B}(\mathcal{H}_i),$$

is

$$\nu_i(B) = \nu_j(B \oplus (\mathcal{H}_j \setminus \mathcal{H}_i)), \tag{2.19}$$

then for any cylindrical set $\mathcal{C} = B \oplus \mathcal{H}_i^{\perp},\ B \in \mathcal{B}(\mathcal{H}_i),\ i \in \mathcal{I}$, we define a cylindrical measure

$$\mu(\mathcal{C}) = \nu_i(B).$$

The condition (2.19) guarantees that $\mu(\mathcal{C})$ is the same regardless of the subspace in which a cylindrical set is measured.

We have already mentioned that set of all cylindrical measures is an algebra which is not countably additive, but it is easy to check that for any pairwise disjoint cylindrical sets $\mathcal{C}_1, \ldots, \mathcal{C}_n$ and a cylindrical measure μ is

$$\mu\left(\bigcup_{i=1}^{n} \mathcal{C}_i\right) = \sum_{i=1}^{n} \mu(\mathcal{C}_i).$$

The following example shows one possible construction of a cylindrical measure.

Example 7. Let $\{e_i\}_{i\in\mathbb{N}}$ be a total orthonormal set in a separable Hilbert space $\mathcal{H}$, and let $\mathrm{R} \in [\mathcal{H}]$ be selfadjoint and positive definite. By definition $\mathcal{C}$ is a cylindrical set if there exist $B \in \mathcal{B}(\mathbb{R}^n)$ and $\{i_1, \ldots, i_n\} \subset \mathbb{N}$ such that $\mathcal{C} = \pi^{-1}_{i_1,\ldots,i_n}(B)$ where

$$\pi_{i_1,\ldots,i_n} :\ x \in \mathcal{H} \to \left\{\left\langle x, e_{i_j}\right\rangle\right\}_{j=1}^{n} \in \mathbb{R}^n.$$

The Gaussian cylindrical measure of the set $\mathcal{C} = \pi^{-1}_{i_1,\ldots,i_n}(B)$ is

$$\mu_{\mathrm{R}}(\mathcal{C}) = \frac{1}{(2\pi)^{n/2}\left(\det(\Sigma)\right)^{1/2}} \int_B \exp\left(-\frac{1}{2}|y|^2_{\mathbf{\Sigma}^{-1}}\right) dy$$

where $\Sigma \in \mathbb{R}^{n\times n}$,

$$(\Sigma)_{i,j} = \langle \mathrm{R}e_i, e_j\rangle \quad i, j = 1, \ldots, n.$$

An important question is whether a cylindrical measure on a separable Hilbert space can be extended to be countably additive. Unfortunately, if the space is infinite dimensional, such extension is, in general, not possible.

Example 8. Denote μ Gaussian cylindrical measure with $\mathrm{R} = \mathrm{I}$ defined in Example 7. If $e_1, \ldots, e_n$ is an orthonormal set, then

$$\langle \mathrm{R}e_i, e_j\rangle = \delta_{i,j}, \quad i, j = 1, \ldots, n.$$

Denote μ_n Gaussian measure $\mathcal{N}(0, \mathrm{I})$ on $\mathbb{R}^n$,

$$\mu_n(B) = \frac{1}{(2\pi)^{n/2}} \int_B \exp\left(-\frac{1}{2}|y|^2\right) dy \quad \forall B \in \mathcal{B}(\mathbb{R}^n).$$

For $r > 0$ and $n \in \mathbb{N}$ define cylindrical set

$$\mathcal{D}_{n,r} = \bigcap_{i=1}^{n} \left\{ x \in \mathcal{H} : |\langle x, e_i \rangle|^2 \leq r^2 \right\},$$

and it follows that

$$\begin{aligned}
\mu_n(\mathcal{D}_{n,r}) =& \mu_n \left(\pi_{1,\ldots,n} \left(\bigcap_{i \leq q} \left\{ x \in \mathcal{H} : (\langle x, e_i \rangle)^2 \leq r^2 \right\} \right) \right) \\
=& \mu_n \left(\pi_{1,\ldots,n} \left(\left\{ x \in \mathcal{H} : \max_{i \leq n} (\langle x, e_i \rangle)^2 \leq r^2 \right\} \right) \right) \\
=& \mu_n \left(\left\{ (a_1, \ldots, a_q)^* \in \mathbb{R}^q : \max_{i \leq n} a_i \leq r \right\} \right) \\
=& \left(\Phi(r) - \Phi(-r) \right)^n,
\end{aligned}$$

where $\Phi(.)$ is the probability distribution function of a real standard Gaussian random variable. Therefore, $\mu(D_n) \to 0$ as n goes to ∞ for any fixed $r \in (0, \infty)$.

If S_r is a sphere with centre in 0 and diameter r,

$$S_r = \left\{ x \in \mathcal{H} : \sum_{i=1}^{\infty} (\langle x, e_i \rangle)^2 < r^2 \right\},$$

then this sphere is a subset of each $D_{n,r}$, so

$$\mu(S_r) \leq \lim_{n \to \infty} \mu(D_{n,r}) = 0$$

for any $r > 0$. On the other hand,

$$\mathcal{D}_{n,r} \subset \mathcal{H} \subset \bigcup_{r=1}^{\infty} S_r,$$

so the set S_r cannot be measurable. Hence, the measure μ cannot be extended to be a countable additive measure.

Fortunately, for a Gaussian cylindrical measure there exist a simple criteria how to determine whether the cylindrical measure can be extended to be countably additive, and the proof may be found in, e.g., Balakrishnan [1976].

Theorem 10. *Let $\mathcal{H}$ be an infinite dimensional Hilbert space, and let $\mathrm{R} \in [\mathcal{H}]$ be self adjoin and positive semidefinite. The Gaussian cylindrical measure μ_{R} can be extended to be countably additive on $\mathcal{B}(\mathcal{H})$ if and only if R is a trace class operator.*

2.4 Additional notes and references

All statements, definitions and other examples may be found in any good textbook covering these topics.

The section covering measurable spaces follows mainly Bogachev [2007] and Halmos [1950]. The different types of integral are well explained in Aliprantis and Border [1999].

Hilbert spaces, their properties and the properties of linear operators are covered by Ciarlet [2013] and Kreyszig [1989]. The extensive study of Schatten norm is provided by Reed and Simon [1980].

The construction of a cylindrical set introduced in this chapter follows Balakrishnan [1976], and additional useful informations about this topic may be found in Bogachev [1998] and Vakhania et al. [1987].

3. Probability on Hilbert spaces

This section briefly reviews the basics of the probability on an infinite dimensional Hilbert space because this topic is usually not covered by standard textbooks about probability. Similar to the previous chapter, this chapter does not cover the whole topic, and many statements are presented without proof.

The chapter is organized as follows. Section 3.1 recalls basic definitions, and shows that many useful properties, known from the theory of random vectors, hold even if random variables are defined on an infinite dimensional space. Section 3.2 introduces weak random variables, and shows their connection to classical random variables. Section 3.3 deals with Gaussian distributions, and briefly mentions the Feldman-Hayek theorem as well. Section 3.4 recalls the Markinciewicz-Zygmund inequality, which allows us to prove the law of large numbers. Section 3.5 summarizes the Bayes theorem, and, finally, Section 3.6 contains references covering the topics introduced in this chapter with greater detail.

Unless otherwise explicitly noted, through the whole section we assume that $\mathcal{H}$ is an infinite dimensional separable Hilbert space over $\mathbb{R}$.

3.1 Random variables

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $\mathcal{H}$ be a separable Hilbert space. The random variable X is a measurable mapping

$$X: \ (\Omega, \mathcal{A}, P) \to (\mathcal{H}, \mathcal{B}(\mathcal{H})),$$

and the random variable X induces a measure

$$\mu_X(B) = P\left(X^{-1}(B)\right) = P(\{\omega \in \Omega: \ X(\omega) \in B\}), \ B \in \mathcal{B}(\mathcal{H}).$$

The space of all random variables on $\mathcal{H}$ is denoted as $\mathcal{L}(\mathcal{H})$, i.e.,

$$\mathcal{L}(\mathcal{H}) = \{X: \ (\Omega, \mathcal{A}, P) \to (\mathcal{H}, \mathcal{B}(\mathcal{H})), \ X \text{ measurable}\}.$$

When $X \in \mathcal{L}(\mathbb{R})$ we say that X is a real valued random variable, and when $X \in \mathcal{L}(\mathbb{R}^n)$ we say that X is a random vector.

Random variables X and Y are equal almost surely if

$$P(X = Y) = P(\{\omega \in \Omega: \ X(\omega) = Y(\omega)\}) = 1.$$

We do not distinguish between random variables that are equal almost surely, and thus we often interchange random variable X and its induced measure μ_X. Therefore, we use both $X \in \mathcal{L}(\mathcal{H})$ and $\mu_X \in \mathcal{L}(\mathcal{H})$ to denote the fact that X is a random variable on $\mathcal{H}$. If $\mu_X \in \mathcal{L}(\mathcal{H})$, we use notation $X_1, \ldots, X_N \sim \mu_X$ to say that all has the same distribution μ_X.

Let $X \in \mathcal{L}(\mathcal{H})$ and $Y \in \mathcal{L}(\mathcal{G})$, where both $\mathcal{H}$ and $\mathcal{G}$ are separable Hilbert spaces, then a measure $\mu_{X,Y}$ is called a joint distribution of X and Y if it is induced by the measurable mapping

$$(X, Y): \omega \in (\Omega, \mathcal{A}, P) \to (X(\omega), Y(\omega)) \in (\mathcal{H} \times \mathcal{G}, \mathcal{B}(\mathcal{H} \times \mathcal{G})).$$

Random variables X and Y are independent if

$$\mu_{X,Y}\left(H \times G\right) = \mu_X\left(H\right)\mu_Y\left(G\right)$$

for all $H \in \mathcal{B}\left(\mathcal{H}\right)$ and all $G \in \mathcal{B}\left(\mathcal{G}\right)$. If random variables X and Y are independent, random vectors $\mathrm{H}X$ and $\mathrm{G}Y$ where $\mathrm{H} \in [\mathcal{H}, \mathbb{R}^n]$ and $\mathrm{G} \in [\mathcal{G}, \mathbb{R}^m]$ are independent as well. Conversely, if $\mathrm{H}X$ and $\mathrm{G}Y$ are independent for all $\mathrm{H} \in [\mathcal{H}, \mathbb{R}^n]$, $\mathrm{G} \in [\mathcal{G}, \mathbb{R}^m]$ and all $n, m \in \mathbb{N}$, then X and Y are independent.

We say that random variables $X_1, \ldots, X_N \in \mathcal{L}\left(\mathcal{H}\right)$ are exchangeable if for any measurable sets $A_1, \ldots, A_N$ in $\mathcal{H}$ and any permutation

$$\sigma : \{1, \ldots, N\} \to \{1, \ldots, N\}$$

is

$$\mu_{X_1,\ldots,X_N}\left(A_1 \times \cdots \times A_N\right) = \mu_{X_{\sigma(1)},\ldots,X_{\sigma(N)}}\left(A_1 \times \cdots \times A_N\right),$$

or, in other words, if the joint distribution of $X_1, \ldots, X_N$ is invariant under permutation of the order of the variables. It follows that if random variables $X_1, \ldots, X_N$ are exchangeable, then marginal distributions of X_i and X_j, $i, j = 1, \ldots, N$, are identical (Mandel et al. [2011]).

3.1.1 Stochastic norm

Because a continuous function of a measurable mapping is measurable, both functions

$$\begin{aligned} |X| &: \omega \in \Omega \to |X\left(\omega\right)| \in \mathbb{R}, \\ \langle h, X\rangle &: \omega \in \Omega \to \langle u, X\left(\omega\right)\rangle \in \mathbb{R} \end{aligned}$$

are measurable for any $X \in \mathcal{L}\left(\mathcal{H}\right)$ and any $u \in \mathcal{H}$, so both $|X|$ and $\langle h, X\rangle$ are real valued random variables. For any $p \geq 1$ we define a stochastic norm

$$\left\|X\right\|_p = \left(\int_\Omega |X\left(\omega\right)|^p \, dP\left(\omega\right)\right)^{1/p} = \left(\int_{\mathcal{H}} |x|^p \, d\mu_X\left(x\right)\right)^{1/p} = \left(\mathrm{E}\,|X|^p\right)^{1/p}.$$

Once again, we emphasize that we reserve double vertical bars for the stochastic norm through the whole thesis, and single bars for a deterministic norm. For any $p \geq 1$ we define a space

$$\mathcal{L}^p\left(\mathcal{H}\right) = \left\{X \in \mathcal{L}\left(\mathcal{H}\right) : \quad \left\|X\right\|_p < \infty\right\},$$

and the space $\mathcal{L}^p\left(\mathcal{H}\right)$ is a Banach space for any $p \in [1, \infty]$ with

$$\left\|X\right\|_\infty = \inf_{A \subset \Omega, P(A)=0} \left\{\sup_{\omega \in \Omega \setminus A} |X\left(\omega\right)|\right\}.$$

The space $\mathcal{L}^2\left(\mathcal{H}\right)$ is a Hilbert space with an inner product

$$\langle X, Y\rangle = \int_\Omega \langle X\left(\omega\right), Y\left(\omega\right)\rangle \, dP\left(\omega\right) = \mathrm{E}\,\langle X, Y\rangle$$

for any $X, Y \in \mathcal{L}^2(\mathcal{H})$.

If $X \in \mathcal{L}^p(\mathcal{H})$ and $Y \in \mathcal{L}^q(\mathcal{H})$ where $p, q \geq 1$ are such that $1/p + 1/q = 1$, then from (2.3) immediately follows Hölder's inequality for random variables

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q . \tag{3.1}$$

When $p = q = 2$, the inequality

$$\|XY\|_1 \leq \|X\|_2 \|Y\|_2 \tag{3.2}$$

is also known also as the Cauchy-Schwarz inequality since it can be written in the form

$$|\langle X, Y \rangle| \leq \|X\|_2 \|Y\|_2 .$$

Lemma 11. *If $X \in \mathcal{L}^{ps}(\mathcal{H})$ for some $p, s \in [1, \infty]$, then real valued random variable $|X|^s$ belongs to $\mathcal{L}^p(\mathbb{R})$.*

Proof. From the definition of stochastic norm

$$\| |X|^s \|_p = \left(\mathrm{E}\,|X|^{ps}\right)^{1/p} = \|X\|_{ps}^s < \infty.$$

□

3.1.2 Mean and covariance operator

If $X \in \mathcal{L}^1(\mathcal{H})$, then the functional

$$\mathrm{m}: \ u \in \mathcal{H} \to \mathrm{E}\,\langle u, X \rangle$$

is linear and bounded because

$$\mathrm{E}\,|\langle u, X \rangle| \leq \mathrm{E}\,(|u|\,|X|) \leq |\mathrm{u}|\,\mathrm{E}\,|X|$$

by the Cauchy-Schwarz inequality. Therefore, using the Riesz's representation theorem, Theorem 4, there is the unique $m \in H$ such that

$$\langle u, m \rangle = \mathrm{E}\,\langle u, X \rangle$$

for all $u \in \mathcal{H}$, and we define a mean value of X

$$\mathrm{E}X = m.$$

It immediately follows

$$\langle u, \mathrm{E}X \rangle = \mathrm{E}\,\langle u, X \rangle$$

which is a weak definition of a mean value, and

$$\mathrm{E}X = \int_\Omega X(\omega)\, dP(\omega)$$

is the Gelfand-Pettis integral of the random variable X (Aliprantis and Border [1999], Pettis [1938]). We also use notation

$$\mathrm{E}X = \int_{\mathcal{H}} x d\mu_X(x) = \int_{\mathcal{H}} x d\mu_X.$$

When $X \in \mathcal{L}^1(\mathcal{H})$

$$\begin{aligned}\mathrm{E}\langle u, (x \otimes X) v\rangle =& \mathrm{E}\langle u, x\langle X, v\rangle\rangle = \langle u, x\rangle \mathrm{E}\langle X, v\rangle \\ =& \langle u, x\rangle \langle \mathrm{E}X, v\rangle = \langle u, x\langle \mathrm{E}X, v\rangle\rangle = \langle u, (x \otimes \mathrm{E}X) v\rangle\end{aligned}$$

for any $x, u, v \in \mathcal{H}$, and, similarly,

$$\begin{aligned}\mathrm{E}\langle u, (X \otimes x) v\rangle =& \mathrm{E}\langle u, X\langle x, v\rangle\rangle = \langle x, v\rangle \mathrm{E}\langle u, X\rangle \\ =& \langle x, v\rangle \langle u, \mathrm{E}X\rangle = \langle u, \mathrm{E}X\langle x, v\rangle\rangle = \langle u, (\mathrm{E}X \otimes x) v\rangle.\end{aligned}$$

Hence, we immediately obtain identities

$$\mathrm{E}(x \otimes X) = x \otimes \mathrm{E}X, \tag{3.3}$$

$$\mathrm{E}(X \otimes x) = \mathrm{E}X \otimes x \tag{3.4}$$

for any $X \in \mathcal{L}^1(\mathcal{H})$ and any $x \in \mathcal{H}$.

If $X, Y \in \mathcal{L}^2(\mathcal{H})$, we define a bilinear form

$$\mathrm{B}: \ (u, v) \in \mathcal{H} \times \mathcal{H} \to \mathrm{E}(\langle u, X - \mathrm{E}X\rangle \langle v, Y - \mathrm{E}Y\rangle) \in \mathbb{R}.$$

This bilinear form is continuous because

$$\begin{aligned}|\mathrm{B}(u, v)| \leq& \mathrm{E}|\langle u, X - \mathrm{E}X\rangle \langle v, X - \mathrm{E}X\rangle| \\ \leq& |u||v| \mathrm{E}(|X - \mathrm{E}X||Y - \mathrm{E}Y|)\end{aligned}$$

for any $u, v \in \mathcal{H}$ using Jensen's inequality and the Cauchy-Schwarz inequality. Again, by the Riesz's representation theorem, notably, by Corollary 1, there exists a unique bounded linear operator $\mathrm{C} \in [\mathcal{H}]$ such that

$$\langle u, \mathrm{C}v\rangle = \mathrm{E}(\langle u, X - \mathrm{E}X\rangle \langle v, Y - \mathrm{E}Y\rangle) \tag{3.5}$$

for all $u, v \in \mathcal{H}$, and we call C the covariance between X and Y. We denote the covariance between X and Y $\mathrm{cov}(X, Y)$, and define $\mathrm{cov}(X) = \mathrm{cov}(X, X)$.

We show that the usual formulas for covariances hold, and also some of its basic properties.

Theorem 12. *If $X, Y \in \mathcal{L}^2(\mathcal{H})$, then*

$$\mathrm{cov}(X, Y) = \mathrm{E}((X - \mathrm{E}X) \otimes (Y - \mathrm{E}Y)) \tag{3.6}$$

$$= \mathrm{E}(X \otimes Y) - (\mathrm{E}X) \otimes (\mathrm{E}Y) \tag{3.7}$$

Proof. Using (3.5) and the definition of a tensor product,

$$\begin{aligned}\langle u, \mathrm{cov}(X, Y) v\rangle =& \mathrm{E}(\langle u, X - \mathrm{E}X\rangle \langle v, Y - \mathrm{E}Y\rangle) \\ =& \mathrm{E}\langle u, (X - \mathrm{E}X)\langle v, Y - \mathrm{E}Y\rangle\rangle \\ =& \mathrm{E}\langle u, ((X - \mathrm{E}X) \otimes (Y - \mathrm{E}Y)) v\rangle \\ =& \langle u, \mathrm{E}((X - \mathrm{E}X) \otimes (Y - \mathrm{E}Y)) v\rangle\end{aligned}$$

for any $u, v \in \mathcal{H}$, which proves (3.6). Using (3.3), (3.4), linearity of a mean value operator and the fact that a tensor product is bilinear,

$$\begin{aligned}\mathrm{E}((X - \mathrm{E}X) \otimes (Y - \mathrm{E}Y)) =& \mathrm{E}(X \otimes Y) - \mathrm{E}(\mathrm{E}X \otimes Y) \\ & - \mathrm{E}(X \otimes \mathrm{E}Y) + \mathrm{E}X \otimes \mathrm{E}Y \\ =& \mathrm{E}(X \otimes Y) - \mathrm{E}X \otimes \mathrm{E}Y,\end{aligned}$$

and the last identity proves (3.7). □

Theorem 13. *If $X \in \mathcal{L}^2(\mathcal{H})$, then the covariance operator* $\operatorname{cov}(X)$ *is selfadjoint, positive semidefinite and trace class.*

Proof. A symmetry of $\operatorname{cov}(X)$ follows immediately from (3.5).

Because for any $u \in \mathcal{H}$

$$\langle u, \operatorname{cov}(X)\, u\rangle = \mathrm{E}\,|\langle u, X - \mathrm{E}X\rangle|^2 \geq 0,$$

it follows that the covariance is a positive semidefinite operator.

Denote $\{e_i\}_{i=1}^{\infty}$ a total orthonormal set of $\mathcal{H}$. Using the monotone convergence theorem [Aliprantis and Border, 1999, Theorem 11.18] and the Parseval's identity [Bogachev, 2007, Theorem 4.3.6] we obtain

$$\operatorname{Tr}(\operatorname{cov}(X)) = \sum_{i=1}^{\infty} \mathrm{E}\,|\langle e_i, X - \mathrm{E}X\rangle|^2 = \mathrm{E}\,|X - \mathrm{E}X|^2 < \infty,$$

where the last inequality holds since $X \in \mathcal{L}^2(\mathcal{H})$. □

If $X \in \mathcal{L}^2(\mathcal{H})$ and a kernel of $\operatorname{cov}(X)$,

$$\ker(\operatorname{cov}(X)) = \{u \in \mathcal{H} : \langle u, \operatorname{cov}(X)\, u\rangle = 0\},$$

contains only zero element, then we say that the random variable X is non-degenerate. If $\ker(\operatorname{cov}(X))$ contains also non zero elements, then we say that X is a degenerate random variable.

The kernel of a linear operator is always a subspace. Therefore, if X is degenerate, we can work with its projection onto the orthogonal complement of the kernel of $\operatorname{cov}(X)$, which is obviously a non-degenerate random variable. Hence, unless explicitly noted, we always assume that a random variable is non-degenerate.

3.1.3 Characteristic function

For $X \in \mathcal{L}(\mathcal{H})$

$$\psi_X :\ h \in \mathcal{H} \to \mathrm{E}\left(e^{i\langle h, X\rangle}\right) = \int_{\mathcal{H}} e^{i\langle h, x\rangle} d\mu_X(x) \in \mathbb{C}$$

is the characteristic function of X, and the function ψ_X is also called the Fourier transform of X or the Fourier transform of μ_X.

Theorem 14. *Assume that $X, Y \in \mathcal{L}(\mathcal{H})$ have characteristic functions ψ_X and ψ_Y respectively. If $\psi_X = \psi_Y$, then $X = Y$ almost surely.*

3.1.4 Sample statistics

For a finite collection of random variables $X_1, \ldots, X_N \in \mathcal{L}(\mathcal{H})$ we define a sample mean

$$\overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i, \tag{3.8}$$

a sample covariance

$$\widehat{\mathrm{C}}_N = \frac{1}{N-1}\sum_{i=1}^{N}\left(\left(X_i - \overline{X}_N\right)\otimes\left(X_i - \overline{X}_N\right)\right), \tag{3.9}$$

and for any $p \geq 1$ p^{th} empirical moment, or empirical moment of order p,

$$\widehat{X}_{N,p} = \left(\frac{1}{N}\sum_{i=1}^{N}|X_i|^p\right)^{1/p}.$$

By putting (3.8) into (3.9) we immediately get identity

$$\widehat{\mathrm{C}}_N = \frac{1}{N}\sum_{i=1}^{N}\left(X_i\otimes X_i\right) - \left(\overline{X}_N\otimes\overline{X}_N\right), \tag{3.10}$$

and by the Hölder inequality for any $p < q$ we obtain

$$\widehat{X}_{N,p} = \left(\frac{1}{N}\sum_{i=1}^{N}\left||X_i|^p\,1\right|\right)^{1/p} \leq \widehat{X}_{N,q}. \tag{3.11}$$

Lemma 15. *Assume that $X_1,\ldots,X_N$ and $Y_1,\ldots,Y_N$ are two collections of random variables on $\mathcal{H}$. Then,*

$$\left|\overline{X}_N\right| \leq \widehat{X}_{N,1}, \quad \left|\overline{Y}_N\right| \leq \widehat{Y}_{N,1},$$

and

$$\left|\frac{1}{N}\sum_{i=1}^{N}\left(X_i\otimes Y_i\right) - \left(\overline{X}_N\otimes\overline{Y}_N\right)\right| \leq 2\widehat{X}_{N,2}\widehat{Y}_{N,2}.$$

Proof. Using the triangle inequality,

$$\left|\overline{X}_N\right| \leq \frac{1}{N}\sum_{i=1}^{N}|X_i| = \widehat{X}_{N,1},$$

and the first and second inequality in the lemma is proved.

Similarly,

$$\left|\frac{1}{N}\sum_{i=1}^{N}\left(X_i\otimes Y_i\right) - \left(\overline{X}_N\otimes\overline{Y}_N\right)\right| \leq \frac{1}{N}\sum_{i=1}^{N}|X_i\otimes Y_i| - \left|\overline{X}_N\otimes\overline{Y}_N\right|, \tag{3.12}$$

and using (2.7),

$$\left|\overline{X}_N\otimes\overline{Y}_N\right| = \left|\overline{X}_N\right|\left|\overline{Y}_N\right| \leq \widehat{X}_{N,1}\widehat{Y}_{N,1}. \tag{3.13}$$

Using the same approach together with the Cauchy-Schwarz inequality gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}|X_i\otimes Y_i| &= \frac{1}{N}\sum_{i=1}^{N}\left(|X_i|\,|Y_i|\right)\\ &\leq \left(\frac{1}{N}\sum_{i=1}^{N}|X_i|^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}|Y_i|^2\right)^{1/2} = \widehat{X}_{N,2}\widehat{Y}_{N,2}.\end{aligned} \tag{3.14}$$

Finally, using (3.12) together with (3.13) and (3.14) yields

$$\left|\frac{1}{N}\sum_{i=1}^{N}(X_i\otimes Y_i)-\left(\overline{X}_N\otimes\overline{Y}_N\right)\right|\leq\widehat{X}_{N,1}\widehat{Y}_{N,1}+\widehat{X}_{N,2}\widehat{Y}_{N,2}\leq 2\widehat{X}_{N,2}\widehat{Y}_{N,2}.$$

□

Lemma 16. *Assume that $X_1,\dots,X_N$ and $Y_1,\dots,Y_N$ are two sets of samples from distributions μ_X and μ_Y respectively. Denote $\widehat{\mathrm{C}}_N$ a sample covariance of the first set, $\widehat{\mathrm{Q}}_N$ a sample covariance of the the second set and*

$$Z_i=X_i-Y_i$$

for $i=1,\dots,N$. Then

$$\left|\widehat{\mathrm{C}}_N-\widehat{\mathrm{Q}}_N\right|\leq 2\left(\widehat{Z}_{N,2}\right)^2+4\widehat{Z}_{N,2}\widehat{Y}_{N,2}.$$

Proof. From (3.10) and the bilinearity of the tensor product follows

$$\begin{aligned}
\widehat{\mathrm{C}}_N-\widehat{\mathrm{Q}}_N=&\frac{1}{N}\sum_{i=1}^{N}(X_i\otimes X_i)-\left(\overline{X}_N\otimes\overline{X}_N\right)-\frac{1}{N}\sum_{i=1}^{N}(Y_i\otimes Y_i)+\left(\overline{Y}_N\otimes\overline{Y}_N\right)\\
=&\frac{1}{N}\sum_{i=1}^{N}\left((X_i\otimes X_i)-(X_i\otimes Y_i)\right)-\left(\overline{X}_N\otimes\overline{X}_N\right)+\left(\overline{X}_N\otimes\overline{Y}_N\right)\\
&+\frac{1}{N}\sum_{i=1}^{N}\left((X_i\otimes Y_i)-(Y_i\otimes Y_i)\right)-\left(\overline{X}_N\otimes\overline{Y}_N\right)+\left(\overline{Y}_N\otimes\overline{Y}_N\right)\\
=&\frac{1}{N}\sum_{i=1}^{N}(X_i\otimes Z_i)-\left(\overline{X}_N\otimes\overline{Z}_N\right)\\
&+\frac{1}{N}\sum_{i=1}^{N}(Z_i\otimes Y_i)-\left(\overline{Z}_N\otimes\overline{Y}_N\right),
\end{aligned}\tag{3.15}$$

and using the previous Lemma 15 gives

$$\left|\frac{1}{N}\sum_{i=1}^{N}(X_i\otimes Z_i)-\left(\overline{X}_N\otimes\overline{Z}_N\right)\right|\leq 2\widehat{X}_{N,2}\widehat{Z}_{N,2},\tag{3.16}$$

$$\left|\frac{1}{N}\sum_{i=1}^{N}(Z_i\otimes Y_i)-\left(\overline{Z}_N\otimes\overline{Y}_N\right)\right|\leq 2\widehat{Z}_{N,2}\widehat{Y}_{N,2}.\tag{3.17}$$

By the triangle inequality

$$\widehat{X}_{N,2}=\left(\frac{1}{N}\sum_{i=1}^{N}|Z_i+Y_i|^2\right)^{1/2}\leq\widehat{Z}_{N,2}+\widehat{Y}_{N,2},\tag{3.18}$$

and using (3.15), (3.16), (3.17), (3.18), and the triangle inequality again concludes the proof since

$$\begin{aligned}
\left|\widehat{\mathrm{C}}_N-\widehat{\mathrm{Q}}_N\right|\leq&2\left(\widehat{Z}_{N,2}+\widehat{Y}_{N,2}\right)\widehat{Z}_{N,2}+2\widehat{Z}_{N,2}\widehat{Y}_{N,2}\\
\leq&2\left(\widehat{Z}_{N,2}\right)^2+4\widehat{Z}_{N,2}\widehat{Y}_{N,2}.
\end{aligned}$$

□

Lemma 17. *Let $X_1, \ldots, X_N \sim \mu_X \in \mathcal{L}^{\max\{p,q\}}(\mathcal{H})$ with $p, q \geq 1$, then*

$$\left\| \left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right)^{1/p} \right\|_q \leq \|X_1\|_{\max\{p,q\}},$$

and the inequality change to equality if $p = q$.

Proof. If $p = q$, then

$$\mathrm{E} \left| \left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right)^{1/p} \right|^p = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E} |X_i|^p = \|X_1\|_p^p$$

because $X_1, \ldots, X_N$ are identically distributed.

If $p > q$, then the function

$$\varphi : \ x \in \mathbb{R} \to x^{q/p}$$

is concave, so the function $-\varphi$ is convex, and

$$\mathrm{E} \left(\left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right)^{q/p} \right) \leq \left(\mathrm{E} \left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right) \right)^{q/p}$$

using Jensen's inequality for real valued random variables. Therefore,

$$\left\| \left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right)^{1/p} \right\|_q = \left(\mathrm{E} \left(\left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right)^{q/p} \right) \right)^{1/q} \leq \left(\mathrm{E} \left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right) \right)^{q/pq}$$

and using the linearity of the mean value operator,

$$\left(\mathrm{E} \left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right) \right)^{1/p} = \|X_1\|_p .$$

If $p < q$, then the function

$$\varphi : \ x \in \mathbb{R} \to x^{q/p}$$

is convex, so using the classical form of Jensen's inequality,

$$\left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right)^{q/p} \leq \frac{1}{N} \sum_{i=1}^{N} \left(|X_i|^p \right)^{q/p} = \frac{1}{N} \sum_{i=1}^{N} |X_i|^q .$$

To conclude, we use the linearity of the mean value operator,

$$\left\| \left(\frac{1}{N} \sum_{i=1}^{N} |X_i|^p \right)^{1/p} \right\|_q^q \leq \mathrm{E} \left(\frac{1}{N} \sum_{i=1}^{N} \left(|X_i|^p \right)^{q/p} \right) = \mathrm{E} |X_1|^q .$$

□

3.2 Weak random variables

Recall that we have defined cylindrical sets in Section 2.3, and shown that these sets establish an algebra on $\mathcal{H}$.

Definition 2. *Let Ω be a nonempty abstract space, $\widehat{\mathcal{A}} \subset 2^{\Omega}$ be an algebra containing all cylindrical sets on Ω, and $\widehat{P}$ be a cylindrical measure on Ω such that $\widehat{P}(\Omega) = 1$. Additionally, let $\mathcal{H}$ be a separable Hilbert space, and $\mathcal{T} \subset 2^{\mathcal{H}}$ be an algebra which contains all cylindrical sets on $\mathcal{H}$. A mapping*

$$W: \left(\Omega, \widehat{\mathcal{A}}, \widehat{P}\right) \to (\mathcal{H}, \mathcal{T})$$

is a weak random variable if satisfies two conditions:

1. *for any $S \in \mathcal{T}$ is $W^{-1}(S) \in \widehat{\mathcal{A}}$, and*
2. *for any $n \in \mathbb{N}$ and any $u_1, \ldots, u_n \in \mathcal{H}$*

$$V: \omega \in \Omega \to \left(\langle u_1, W(\omega)\rangle, \ldots, \langle u_n, W(\omega)\rangle\right)^* \in \mathbb{R}^n$$

is n dimensional random vector.

We denote $\mathcal{L}_w(\mathcal{H})$ the space of all weak random variables on $\mathcal{H}$.

Example 9. Denote μ the cylindrical Gaussian measure introduced in Example 8, and denote $\mathcal{T}$ the smallest algebra containing all cylindrical sets in $\mathcal{H}$. An identity mapping

$$W: u \in \mathcal{H} \to u \in \mathcal{H}$$

is clearly a weak random variable since

$$\forall S \in \mathcal{T}: W^{-1}(S) = S \in \mathcal{T},$$

and for any $u_1, \ldots u_n$ is

$$\left(\langle u_1, W(u)\rangle, \ldots, \langle u_n, W(u)\rangle\right)^*$$

a Gaussian random vector by the construction of the cylindrical measure μ.

The previous example shows the simplest way to create a weak random variable. However, in our further applications we assume that

$$\left(\Omega, \widehat{\mathcal{A}}, \widehat{P}\right) = (\Omega, \mathcal{A}, P),$$

and we are mainly interested in conditions when the cylindrical measure induced by a weak random variable can be extended to be σ-additive on $\mathcal{B}(\mathcal{H})$.

If W is a weak random variable, then $\langle u, W\rangle$ is a real-valued random variable for any $u \in \mathcal{H}$, and we denote $\mu_{W,u}$ the distribution on $\mathbb{R}$ of the random variable $\langle u, W\rangle$. Hence, similarly to a measurable random variable, we can define a characteristic function of a weak random variable W by

$$\psi_W: u \in \mathcal{H} \to \mathrm{E}\left(e^{i\langle u, W\rangle}\right) = \int_{\mathbb{R}} e^{ix} d\mu_{W,u}(x) \in \mathbb{C}.$$

The next lemma shows that weak random variables are interesting only on infinite dimensional spaces.

Lemma 18. *If $X \in \mathcal{L}(\mathcal{H})$, then $X \in \mathcal{L}_w(\mathcal{H})$. Conversely, if $X \in \mathcal{L}_w(\mathcal{H})$ and $\dim(\mathcal{H}) < \infty$, then $X \in \mathcal{L}(\mathcal{H})$.*

Proof. If $X \in \mathcal{L}(\mathcal{H})$, it is obvious that X satisfies Definition 2 because all cylindrical sets are contained in $\mathcal{B}(\mathcal{H})$.

On the other hand, if $\dim(\mathcal{H}) < \infty$, there is a total orthonormal set

$$\{e_1, \ldots, e_n\} \subset \mathcal{H}.$$

Using the definition, $(\langle e_1, W\rangle, \ldots, \langle e_n, W\rangle)^*$ is a random vector, so

$$W = \sum_{i=1}^{n} e_i \langle e_i, W\rangle$$

is measurable as well. □

Additionally, it can be shown [Vakhania et al., 1987, Theorem 2.1] that

$$\mathcal{L}_w(\mathcal{H}) = \mathcal{L}(\mathcal{H}) \Leftrightarrow \dim(\mathcal{H}) < \infty.$$

We use term random elements to denote both measurable random variables and weak random variables.

3.2.1 Weak stochastic norm

If W is a weak random variable on $\mathcal{H}$, then for any $p \geq 1$,

$$\|W\|_{p,w} = \sup_{u\in\mathcal{H},|u|\leq 1} \left(\mathrm{E}\,|\langle u, W\rangle|^p\right)^{1/p} = \sup_{u\in\mathcal{H},|u|\leq 1} \|\langle u, W\rangle\|_p \tag{3.19}$$

defines a norm, and

$$\mathcal{L}_w^p(\mathcal{H}) = \left\{W \in \mathcal{L}_w(\mathcal{H}) :\ \|W\|_{p,w} < \infty\right\}$$

is a linear space. The norm (3.19) is called the weak stochastic norm, e.g., Vakhania et al. [1987].

It it straightforward to check that Equation (3.19) defines a norm:

Firstly, if $\|W\|_{p,w} = 0$, then $\langle u, W\rangle = 0$ almost surely for all $u \in \mathcal{H}$, so $W = 0$ almost surely.

Secondly,

$$\|\alpha W\|_{p,w} = \sup_{u\in\mathcal{H},|u|\leq 1} \|\langle u, \alpha W\rangle\|_p = |\alpha| \sup_{u\in\mathcal{H},|u|\leq 1} \|\langle u, W\rangle\|_p = |\alpha|\, \|\langle u, W\rangle\|_{p,w}$$

for all $\alpha \in \mathbb{R}$.

Finally,

$$\|\langle u, W+V\rangle\|_p = \|\langle u, W\rangle + \langle u, V\rangle\|_p \leq \|\langle u, W\rangle\|_p + \|\langle u, V\rangle\|_p, \tag{3.20}$$

and the triangle inequality for weak stochastic norm yields from taking supremum of each side of (3.20).

We already know, from Lemma 18, that $\mathcal{L}(\mathcal{H}) \subset \mathcal{L}_w(\mathcal{H})$, and the next lemma shows that the weak norm of a random variable is always bounded by its stochastic norm.

Lemma 19. *If $X \in \mathcal{L}^p(\mathcal{H})$, then*

$$\|X\|_{p,w}^p \leq \|X\|_p^p.$$

Proof. The lemma can be proved simply by using the definition of the weak $\mathcal{L}^p$ norm together with with Cauchy-Schwarz inequality:

$$\begin{aligned}
\|X\|_{p,w}^p &= \sup_{u\in\mathcal{H},|u|\leq 1} \left\{ \int_{\mathcal{H}} |\langle u, x\rangle|^p \, d\mu_X(x) \right\} \\
&\leq \int_{\mathcal{H}} \sup_{u\in\mathcal{H},|u|\leq 1} \left\{ |\langle u, x\rangle|^p \right\} d\mu_X(x) \\
&\leq \int_{\mathcal{H}} \sup_{u\in\mathcal{H},|u|\leq 1} \left\{ \left| \langle u, u\rangle^{1/2} \langle x, x\rangle^{1/2} \right|^p \right\} d\mu_X(x) \\
&\leq \int_{\mathcal{H}} |x|^p \, d\mu_X(x) = \|X\|_p^p.
\end{aligned}$$

□

A very important property is that an inner product of a measurable random variable with a finite first moment and weak random variable is measurable.

Lemma 20. *Suppose that $X \in \mathcal{L}^1(\mathcal{H})$ and $W \in \mathcal{L}_w(\mathcal{H})$. Then, $\langle X, W\rangle$ is a real-valued random variable.*

Proof. Since $X \in \mathcal{L}^1(\mathcal{H})$, there exists a sequence of step functions

$$X_n : (\Omega, \mathcal{A}, P) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$$

such that $X_n \to X$ in $\mathcal{L}^1(\mathcal{H})$, e.g., [Lang, 1993, page 211], and, therefore, $X_n \to X$ in $\mathcal{H}$ a.s. For each n, the step function X_n is of the form

$$X_n : \omega \mapsto \sum_{j=1}^{J_n} x_{j,n} \mathbf{1}_{A_{j,n}}(\omega),$$

where $x_{j,n} \in \mathcal{H}$,

$$\mathbf{1}_{A_{j,n}}(\omega) = \begin{cases} 1 & \text{if } \omega \in A_{j,n}, \\ 0 & \text{otherwise,} \end{cases}$$

the sets $A_{j,n}$ are measurable, pairwise disjoint, and

$$\bigcup_{j=1}^{J_n} A_{j,n} = \Omega.$$

Consequently, the function

$$\omega \to \langle X_n(\omega), W(\omega)\rangle$$

is measurable, since

$$\langle X_n, W\rangle : \omega \to \sum_{j=1}^{J_n} \left(\langle x_{j,n}, W(\omega)\rangle \mathbf{1}_{A_{j,n}}(\omega) \right)$$

where $\omega \to \langle x_{j,n}, W(\omega)\rangle$ is measurable by the definition of a weak random variable, the indicator function $\mathbf{1}_{A_{j,n}}(\omega)$ of a measurable set is measurable, and the product and sum of measurable functions is measurable as well .

Now fix $\omega \in \Omega$ such that $X_n(\omega) \to X(\omega)$ in $\mathcal{H}$. Then,

$$\langle X_n(\omega), W(\omega)\rangle \to \langle X(\omega), W(\omega)\rangle$$

in $\mathbb{R}$. Consequently, $\langle X_n, W\rangle \to \langle X, W\rangle$ almost surely, and since a.s. limit of measurable functions is measurable, $\langle X, W\rangle$ is measurable. □

We say that two weak random variables $W_1, W_2 \in \mathcal{L}_w(\mathcal{H})$ are independent if for any $n \in \mathbb{N}$ and any projection operator

$$\Pi : \mathcal{H} \to \mathbb{R}^n,$$

the random vectors ΠW_1 and ΠW_2 are independent.

The following lemma gives us an estimate of the stochastic norm of an inner product of a measurable and a week random variable.

Lemma 21. *Assume that $X \in \mathcal{L}^p(\mathcal{H})$, $W \in \mathcal{L}^p_w(\mathcal{H})$, $p \geq 1$, and X and W are independent. Then $\langle X, W\rangle \in \mathcal{L}^p(\mathbb{R})$ and*

$$\|\langle X, W\rangle\|_p \leq \|X\|_p \|W\|_{p,w}.$$

Proof. Using Lemma 20, the function

$$\langle X, W\rangle : \ (\omega_1, \omega_2) \in \Omega \times \Omega \to \langle X(\omega_1), W(\omega_2)\rangle$$

is measurable. Using the independence of X and W,

$$\begin{aligned}\|\langle X, W\rangle\|_p^p &= \int_{\mathcal{H}}\int_{\mathcal{H}} |\langle X(\omega_1), W(\omega_2)\rangle|^p \, dP(\omega_2)\, dP(\omega_1) \\ &= \int_{\mathcal{H}}\int_{\mathcal{H}} \left|\left\langle \frac{X(\omega_1)}{|X(\omega_1)|}, W(\omega_2)\right\rangle\right|^p |X(\omega_1)|^p \, dP(\omega_2)\, dP(\omega_1),\end{aligned}$$

where we take $\frac{X(\omega_1)}{|X(\omega_1)|}$ a fixed vector of unit length if $X(\omega_1) = 0$, and the use of Fubini's theorem gives

$$\|\langle X, W\rangle\|_p^p = \int_{\mathcal{H}}\int_{\mathcal{H}} \left|\left\langle \frac{X(\omega_1)}{|X(\omega_1)|}, W(\omega_2)\right\rangle\right|^p dP(\omega_2)\, |X(\omega_1)|^p \, dP(\omega_1).$$

Note that

$$\left|\frac{X(\omega_1)}{|X(\omega_1)|}\right| \leq 1,$$

so

$$\begin{aligned}\|\langle X, W\rangle\|_p^p &\leq \int_{\mathcal{H}} \sup_{h\in\mathcal{H}, |h|\leq 1} \int_{\mathcal{H}} |\langle h, W(\omega_2)\rangle|^p \, dP(\omega_2)\, |X(\omega_1)|^p \, dP(\omega_1) \\ &= \int_{\mathcal{H}} \|W\|_{p,w}^p |X(\omega_1)|^p \, dP(\omega_1) \\ &= \|W_i\|_{p,w}^p \int_{\mathcal{H}} |X(\omega_1)|^p \, dP(\omega_1) = \|W\|_{p,w}^p \|X\|_p^p.\end{aligned}$$

□

3.3 Gaussian distributions

Definition 3 (Gaussian random vector)**.** *Assume that $m \in \mathbb{R}^n$, and $\mathrm{C} \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix. Then a random variable $X \in \mathcal{L}(\mathbb{R}^n)$ has Gaussian distribution with the mean m and the covariance C, $X \sim \mathcal{N}(m, \mathrm{C})$, if its characteristic function is*

$$\psi_X(t) = \exp\left(im^*t - \frac{1}{2}t^*\mathrm{C}t\right),\ t \in \mathbb{R}.$$

Additionally, if the covariance matrix C is positive definite, then X has a density

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\left(\det(\mathrm{C})\right)^{1/2}} \exp\left(-\frac{1}{2}|x-m|_{\mathrm{C}^{-1}}\right)$$

where

$$|x|_{\mathrm{C}^{-1}} = x\mathrm{C}^{-1}x^* = \left\langle \mathrm{C}^{-1/2}x, \mathrm{C}^{-1/2}x\right\rangle.$$

Definition 4 (Gaussian random variable)**.** *Assume that $m \in \mathcal{H}$, and an operator $\mathrm{C} \in [\mathcal{H}]$ is symmetric, positive semidefinite and trace class, then a random variable $X \in \mathcal{L}(\mathcal{H})$ has Gaussian distribution with the mean m and the covariance C if for all $n \in \mathbb{N}$ and all $\mathrm{T} \in [\mathcal{H}, \mathbb{R}^n]$,*

$$\mathrm{T}X \sim \mathcal{N}(\mathrm{T}m, \mathrm{TCT}^*).$$

Equivalently, $X \sim \mathcal{N}(m, \mathrm{C})$ if its characteristic function is

$$\psi_X(u) = \exp\left(i\langle m, u\rangle + \langle u, \mathrm{C}u\rangle\right),\ u \in \mathcal{H}.$$

Obviously, if $X \sim \mathcal{N}(m, \mathrm{C})$, then $\mathrm{E}X = m$ and $\mathrm{cov}(X) = \mathrm{C}$.

Definition 5 (Gaussian weak random variable)**.** *Assume that $m \in \mathcal{H}$, and an operator $\mathrm{C} \in [\mathcal{H}]$ is symmetric and positive semidefinite, then a weak random variable $X \in \mathcal{L}_w(\mathcal{H})$ has weak Gaussian distribution $\mathcal{N}_w(m, \mathrm{C})$, if for all $n \in \mathbb{N}$ and all $\mathrm{T} \in [\mathcal{H}, \mathbb{R}^n]$*

$$\mathrm{T}X \sim \mathcal{N}(\mathrm{T}m, \mathrm{TCT}^*),$$

or, equivalently, $X \sim \mathcal{N}_w(m, \mathrm{C})$ if

$$\psi_X(u) = \exp\left(i\langle m, u\rangle + \langle u, \mathrm{C}u\rangle\right),\ u \in \mathcal{H},$$

is the characteristic function of the weak random variable X.

The only difference between the definition of a Gaussian random variable and a weak Gaussian random variable is that the first one assumes that C is trace class. It can be shown that, this is a necessary and sufficient condition for a Gaussian weak random variable to be measurable.

Theorem 22 ([Balakrishnan, 1976, Theorem 6.2.2])**.** *Assume that $m \in \mathcal{H}$, and $\mathrm{C} \in [\mathcal{H}]$ is a symmetric positive semidefinite operator. A cylindrical measure induced by a weak random variable $X \sim \mathcal{N}_w(m, \mathrm{C})$ can be extended to be σ-additive on $\mathcal{H}$, i.e., $X \sim \mathcal{N}(m, \mathrm{C})$, if and only if the operator C is trace class.*

The next example shows an example of a Gaussian weak random variable that is not measurable.

Example 10. Assume that $X \sim \mathcal{N}_w(0, \mathrm{I})$. The weak random variable induces a cylindrical measure μ_X on $\mathcal{H}$, and we have already shown in Example 8 that this measure cannot be extended on all Borel sets in $\mathcal{H}$.

Through the thesis, we strictly use the convention that by writing

$$X \sim \mathcal{N}(m, \mathrm{C}) \quad \text{or} \quad X \sim \mathcal{N}_w(m, \mathrm{C})$$

we silently assume that the operator C has the appropriate properties given by Definition 4 or Definition 5 respectively. Also, unless explicitly noted, we assume that the random elements are non-degenerate, i.e., that 0 is the only element of the kernel of the operator C.

3.3.1 Basic properties

Theorem 23. *If* $X \sim \mathcal{N}_w(m, \mathrm{C})$ *and* $\mathrm{K} \in [\mathcal{H}]$, *then random element* $\mathrm{K}X \sim \mathcal{N}_w(\mathrm{K}m, \mathrm{KCK}^*)$.

Proof. The statement immediately follows from the properties of characteristic functions. □

Theorem 24. *Let* $X \sim \mathcal{N}_w(m, \mathrm{C})$. *If* $\mathrm{K} \in [\mathcal{H}]$ *is a Hilbert-Schmidt operator, then* $\mathrm{K}W$ *is a Gaussian measurable random variable, and* $\mathrm{K}W \sim \mathcal{N}(\mathrm{K}m, \mathrm{KCK}^*)$.

Proof. Using Theorem VI.19 from Reed and Simon [1980], the operator KCK^* is trace class, and the statement now immediately yields from Theorem 22 and Theorem 23. □

If $X \sim \mathcal{N}(m, \mathrm{C})$, we define function $M_X : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$,

$$M_X(\varepsilon) = \int_{\mathcal{H}} \exp\left(\frac{\varepsilon}{2}|x|^2\right) d\mu_X(x), \tag{3.21}$$

and it can be shown [Da Prato, 2006, Proposition 1.13] that

$$M_X(\varepsilon) = \begin{cases} \left(\prod_{i=1}^{\infty}(1-\varepsilon c_i)\right)^{-1/2} \exp\left(-\frac{\varepsilon}{2}\left\langle(\mathrm{I}-\varepsilon\mathrm{C})^{-1}m, m\right\rangle\right) & \text{for } \varepsilon < \frac{1}{c_{\max}} \\ \infty & \text{otherwise,} \end{cases} \tag{3.22}$$

where $c_i, i \in \mathbb{N}$, are the eigenvalues of the operator C, and

$$c_{\max} = \max_{i \in \mathbb{N}} c_i.$$

The function M_X is useful for a computation of even moments of the random variable X because using the identities

$$M_X'(\varepsilon) = \frac{1}{2}\int_{\mathcal{H}} |x|^2 \exp\left(\frac{\varepsilon}{2}|x|^2\right) d\mu_X(x),$$

and

$$M_X^{(p)}(\varepsilon) = \frac{1}{2^p}\int_{\mathcal{H}} |x|^{2p} \exp\left(\frac{\varepsilon}{2}|x|^2\right) d\mu_X(x)$$

for any $p \in \mathbb{N}$ yields

$$M_X^{(p)}(0) = \frac{1}{2^p} \int_{\mathcal{H}} |x|^{2p} \, d\mu_X(x) = \frac{1}{2^p} \|X\|_{2p}^{2p}. \tag{3.23}$$

Obviously,

$$M_X(0) = 1, \tag{3.24}$$

and Equation (3.23) allows us to compute moments of a centered Gaussian random variable easily if we are able to evaluate the derivatives of M_X. The following lemma becomes very useful for this purpose.

Lemma 25. *Let $\{c_i\}_{i=1}^{\infty}$ be a sequence of positive numbers such that $\sum_{i=1}^{\infty} c_i < \infty$, and define function*

$$M : \varepsilon \in (-1/c, 1/c) \to \left(\prod_{i=1}^{\infty} (1 - \varepsilon c_i) \right)^{-1/2},$$

where $c = \max_{i \in \mathbb{N}} c_i$. Then for any $p \in \mathbb{N}$

$$M^{(p)}(\varepsilon) = \frac{1}{2} \sum_{j=0}^{p-1} \frac{(p-1)!}{j!} M^{(j)}(\varepsilon) \, S_{p-j}(\varepsilon) \tag{3.25}$$

where

$$S_k(\varepsilon) = \sum_{i=1}^{\infty} \left(\frac{c_i}{1 - \varepsilon c_i} \right)^k. \tag{3.26}$$

Proof. Firstly,

$$\log \left(\prod_{i=1}^{\infty} (1 - \varepsilon c_i) \right) = \sum_{i=1}^{\infty} \log(1 - \varepsilon c_i),$$

and, using the limit comparison criteria,

$$0 < \prod_{i=1}^{\infty} (1 - \varepsilon c_i) \;\Leftrightarrow\; \sum_{i=1}^{\infty} c_i < \infty,$$

so the function M is well defined.

Secondly, for any $j \in \mathbb{N}$

$$\begin{aligned} S_j'(\varepsilon) = &\left(\sum_{i=1}^{\infty} \frac{c_i^j}{(1 - \varepsilon c_i)^j} \right)' = \sum_{i=1}^{\infty} \left(\frac{c_i^j}{(1 - \varepsilon c_i)^j} \right)' \\ = &j \sum_{i=1}^{\infty} \left(\frac{c_i}{1 - \varepsilon c_i} \right)^{j+1} = j S_{j+1}(\varepsilon). \end{aligned}$$

Now, we use an induction to prove (3.25). If $p = 1$,

$$\begin{aligned} (M(\varepsilon))' = &\frac{1}{2} \left(\prod_{i=1}^{\infty} (1 - \varepsilon c_i) \right)^{-3/2} \sum_{i=1}^{\infty} \left(\frac{c_i}{1 - \varepsilon c_i} \prod_{j=1}^{\infty} (1 - \varepsilon c_j) \right) \\ = &\frac{1}{2} \left(\prod_{i=1}^{\infty} (1 - \varepsilon c_i) \right)^{-1/2} \sum_{i=1}^{\infty} \frac{c_i}{1 - \varepsilon c_i} \\ = &\frac{1}{2} M(\varepsilon) \, S_1(\varepsilon). \end{aligned}$$

When (3.25) holds for some $p \in \mathbb{N}$, then

$$\begin{aligned} M^{(p+1)} =& \frac{1}{2} \sum_{j=0}^{p-1} \left(\frac{(p-1)!}{j!} M^{(j)} S_{p-j} \right)' \\ =& \frac{1}{2} \sum_{j=0}^{p-1} \left(\frac{(p-1)!}{j!} M^{(j+1)} S_{p-j} + \frac{(p-1)!}{j!} (p-j) M^{(j)} S_{p-j+1} \right), \end{aligned}$$

and we conclude the proof by rearranging terms in the last sum,

$$\begin{aligned} M^{(p+1)} =& \frac{1}{2} \left(M^{(p)} S_1 + p! M^{(0)} S_{p+1} \right) \\ & + \frac{1}{2} \sum_{j=1}^{p-1} \left(\frac{(p-1)!}{(j-1)!} M^{(j)} S_{p-j+1} + \frac{(p-1)!}{j!} (p-j) M^{(j)} S_{p-j+1} \right) \\ =& \frac{1}{2} \left(M^{(p)} S_1 + p! M^{(0)} S_{p+1} + \sum_{j=1}^{p-1} \frac{p!}{j!} M^{(j)} S_{p+1-j} \right) \\ =& \frac{1}{2} \sum_{j=0}^{p} \frac{p!}{j!} M^{(j)} S_{p+1-j}. \end{aligned}$$

□

The functions $S_p(\varepsilon)$ defined by (3.26) have one very important property: for a given $X \sim \mathcal{N}(0, \mathrm{C})$ and any $p \in \mathbb{N}$ is

$$S_p(0) = \sum_{i=1}^{\infty} (c_i)^p = |\mathrm{C}|_p^p \tag{3.27}$$

where $|\mathrm{C}|_p$ stands for the Schatten norm of C, Equation (2.13). This identity allows us to prove the following, very important, estimate.

Lemma 26. *Let X be $\mathcal{N}(0, \mathrm{C})$ random variable. Then there are positive constants k_p, $p \in \mathbb{N}$, such that*

$$\|X\|_{2p} \leq k_p \left(\mathrm{Trace}\,(\mathrm{C}) \right)^{1/2} = k_p |\mathrm{C}|_1^{1/2},$$

and these constants only depend on p.

Proof. We use the same notation as in Lemma 25, and, similarly as in the already mentioned lemma, we use induction to show the existence of constants $\widetilde{k}_p$ such that

$$M_X^{(p)}(0) \leq \widetilde{k}_p |\mathrm{C}|_1^p \tag{3.28}$$

for any $p \in \mathbb{N}$.

When $p = 1$,

$$M_X^{(1)}(0) = \frac{1}{2} M_X(0) S_1(0) = \frac{1}{2} |\mathrm{C}|_1$$

using equations (3.24), (3.23), (3.27), and Lemma 25, so $\widetilde{k}_1 = 1/2$.

If the inequality (3.28) holds for all $j \in \{1, \ldots, p\}$ with constants $\widetilde{k}_1, \ldots, \widetilde{k}_p$, then

$$\begin{aligned} M_X^{(p+1)}(0) =& \frac{1}{2} \sum_{j=0}^{p} \frac{p!}{j!} M^{(j)} S_{p+1-j} \\ \leq& \frac{1}{2} \sum_{j=0}^{p} \frac{p!}{j!} \widetilde{k}_j \left|\mathrm{C}\right|_1^j \left|\mathrm{C}\right|_{p+1-j}^{p+1-j} \end{aligned}$$

where we define $\widetilde{k}_0 = 1$. Using (2.14) and the properties of Schatten norm,

$$M_X^{(p+1)}(0) \leq \frac{1}{2} \sum_{j=0}^{p} \frac{p!}{j!} \widetilde{k}_j \left|\mathrm{C}\right|_1^j \left|\mathrm{C}\right|_1^{p+1-j} = \frac{1}{2} \sum_{j=0}^{p} \frac{p!}{j!} \widetilde{k}_j \left|\mathrm{C}\right|_1^{p+1},$$

and inequality (3.28) is proved by defining

$$\widetilde{k}_{p+1} = \frac{1}{2} \sum_{j=0}^{p} \frac{p!}{j!} \widetilde{k}_j.$$

Using Equation (3.23) yields

$$\left\|X\right\|_{2p}^{2p} = 2^p M_X^{(p)}(0) \leq 2^p \widetilde{k}_{p+1} \left|\mathrm{C}\right|_1^p,$$

so defining $k_p = \sqrt{2} \left(\widetilde{k}_p\right)^{\frac{1}{2p}}$, $p \in \mathbb{N}$, proves the statement. □

3.3.2 Cameron-Martin space

A Cameron-Martin space of random variable $X \sim \mathcal{N}(0, \mathrm{C})$ is

$$\mathrm{C}^{1/2}(\mathcal{H}) = \left\{u \in \mathcal{H} : \ u = \mathrm{C}^{1/2} v, \ v \in \mathcal{H}\right\}.$$

The Cameron-Martin space of a Gaussian random variable X can be equivalently defined as an intersection of all linear subspaces $\mathcal{G} \subset \mathcal{H}$ such that

$$\mu_X(\mathcal{G}) = 1,$$

e.g., [Bogachev, 1998, Theorem 2.4.7]. Hence, it may be a surprise that a measure of a Cameron-Martin space is positive only when $\mathcal{H}$ is finite dimensional.

Theorem 27. *Assume that $X \sim \mathcal{N}(0, \mathrm{C})$ on a separable Hilbert space $\mathcal{H}$.*
If $\dim(\mathcal{H}) < \infty$, then $\mu_X\left(\mathrm{C}^{1/2}(\mathcal{H})\right) = 1$.
If $\dim(\mathcal{H}) = \infty$, then $\mu_X\left(\mathrm{C}^{1/2}(\mathcal{H})\right) = 0$.

Proof. See Da Prato [2006], Proposition 1.27. □

Theorem 28 (Camerom-Martin theorem)**.** *Assume that $X \sim \mathcal{N}(m_X, \mathrm{C})$ and $Y \sim \mathcal{N}(m_Y, \mathrm{C})$, then measures μ_X and μ_Y are equivalent if and only if*

$$m_X - m_Y \in \mathrm{C}^{1/2}(\mathcal{H}).$$

If $(m_X - m_Y) \notin \mathrm{C}^{1/2}(\mathcal{H})$, then measures μ_X and μ_Y are singular.

Proof. See Da Prato and Zabczyk [2002], Theorem 1.3.6. □

The following system example illustrates the use of the previous theorem when $\mathcal{H} = \mathbb{R}^2$.

Example 11. Define matrix

$$\mathrm{C} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

and assume that $X \sim \mathcal{N}(0, \mathrm{C})$ and $Y \sim \mathcal{N}((1,1)^*, \mathrm{C})$. The Cameron-Martin space of X is

$$\mathrm{C}^{1/2}\left(\mathbb{R}^2\right) = \operatorname{span}\{(1,0)^*\},$$

so the measures μ_X and μ_Y are singular by the Cameron-Martin theorem.

We can check that μ_X and μ_Y are singular using the definition. Obviously,

$$\mu_X(\operatorname{span}\{(1,0)^*\}) = 1$$

and

$$\mu_Y(\operatorname{span}\{(1,1)^*\} \setminus \{(0,0)^*\}) = 1,$$

so μ_X and μ_Y are singular because

$$\operatorname{span}\{(1,0)^*\} \cap (\operatorname{span}\{(1,1)^*\} \setminus \{(0,0)^*\}) = \emptyset.$$

3.3.3 Feldman-Hájek theorem

The Feldman-Hájek theorem reveals an interesting fact that any two centered Gaussian measures are either equivalent or singular.

Theorem 29 (Feldman-Hájek theorem)**.** *Assume that*

$$\mu_{\mathrm{C}} \sim \mathcal{N}(0, \mathrm{C}) \text{ and } \mu_{\mathrm{R}} \sim \mathcal{N}(0, \mathrm{R}).$$

If the measures μ_{C} and μ_{R} are not singular, there exists a selfadjoint Hilbert-Schmidt operator $\mathrm{S} \in [\mathcal{H}]$ such that

$$\mathrm{C} = \mathrm{R}^{1/2}(\mathrm{I} - \mathrm{S})\,\mathrm{R}^{1/2}.$$

If such operator S exists, the measures μ_{C} and μ_{R} are equivalent.

Proof. The first implication is Theorem 1.3.9 in Da Prato and Zabczyk [2002], and the second one is Theorem 1.3.10 in the same book. □

The last theorem may be easily illustrated when the measures are defined on a finite dimensional space.

Example 12. Assume that matrices P, Q and R belongs to $\mathbb{R}^{n\times n}$. If

$$\operatorname{rank}(\mathrm{P}) = \operatorname{rank}(\mathrm{Q}) = n,$$

then measures $\mu_{\mathrm{P}} \sim \mathcal{N}(0, \mathrm{P})$ and $\mu_{\mathrm{Q}} \sim \mathcal{N}(0, \mathrm{Q})$ are equivalent. If

$$\operatorname{rank}(\mathrm{R}) = m < n,$$

then the dimension of the Cameron-Martin space $\mathrm{R}^{1/2}(\mathbb{R}^n)$ is m, and we already know that

$$\mu_{\mathrm{R}}\left(\mathrm{R}^{1/2}(\mathbb{R}^n)\right) = 1,$$

where $\mu_{\mathrm{R}} \sim \mathcal{N}(0, \mathrm{R})$. On the other side,

$$\mu_{\mathrm{P}}\left(\mathbb{R}^n \setminus \mathrm{R}^{1/2}(\mathbb{R}^n)\right) = 1,$$

so μ_{P} and μ_{R} are singular.

Additionally, it can be shown [Da Prato, 2006, Theorem 2.9] that if C and R commute, then the Feldman-Hájek theorem implies that C and R are equivalent if and only if

$$\sum_{i=1}^{\infty} \frac{(c_i - r_i)^2}{(c_i + r_i)^2} < \infty \tag{3.29}$$

where $\{c_i\}$ and $\{r_i\}$ are eigenvalues of C and R respectively. This consequence of the the Feldman-Hájek theorem has a very interesting corollary.

Corollary 3. Assume that $\alpha \in \mathbb{R}$. The measures $\mathcal{N}(0, \mathrm{C})$ and $\mathcal{N}(0, \alpha\mathrm{C})$ defined on an infinite dimensional Hilbert space are equivalent if and only if $\alpha = 1$. If $\alpha \neq 1$, the measures are singular.

Proof. Operators C and αC commute, so the measures commute if and only if

$$\sum_{i=1}^{\infty} \frac{(\alpha c_i - c_i)^2}{(\alpha c_i + c_i)^2} < \infty$$

where c_i, $i \in \mathbb{N}$, are eigenvalues of C. Obviously,

$$\sum_{i=1}^{\infty} \frac{(\alpha c_i - c_i)^2}{(\alpha c_i + c_i)^2} = \sum_{i=1}^{\infty} \frac{(\alpha - 1)^2}{(\alpha + 1)^2} < \infty \quad \Leftrightarrow \quad \alpha = 1,$$

and the corollary follows immediately from the Feldman-Hájek theorem.

Obviously, Camerom-Martin theorem and Feldman-Hájek theorem give important corollary that two Gaussian measures defined on a Hilbert space are either equivalent or singular. □

3.4 Markinciewicz-Zygmund inequality

There are four types of convergence that are usually of primary interest in the area of probability theory: the convergence almost surely, the convergence in probability, the convergence in distribution and the convergence in $\mathcal{L}^p$. All of them are well defined on an infinite dimensional space. Additionally, the infinite dimension brings many more interesting types of convergence. However, we are interested only in the convergence in $\mathcal{L}^p$. Random variables $X_n \in \mathcal{L}(\mathcal{H})$, $n \in \mathbb{N}$, converge to $X \in \mathcal{L}^p(\mathcal{H})$ in $\mathcal{L}^p$ if

$$\lim_{n\to\infty} \|X_n - X\|_p = 0.$$

The next theory is very important, because it allows us to prove the weak law of large numbers for infinite dimensional random variables. This formulation can be found as Lemma 5.1 in Kwiatkowski and Mandel [2015], and its proof may be found in Chow and Teicher [1997] and Woyczyński [1980].

Theorem 30 (Marcinkiewicz-Zygmund inequality)**.** *Let $\mu_X \in \mathcal{L}^p(\mathcal{H})$ where $\mathcal{H}$ is a separable Hilbert space and $p \geq 1$, and let $X_1, \ldots, X_N \sim \mu_X$ be independent, then*

$$\mathrm{E}\left|\sum_{i=1}^{N} X_i\right|^p \leq b_p \mathrm{E}\left(\sum_{i=1}^{N} |X_i|^2\right)^{p/2}$$

where b_p is positive real constant which depends on p only.

The previous theorem is only a special case of the Markinciewicz-Zygmund inequality, which holds on each separable Banach space $\mathcal{B}$ such that for any finite set $\{y_1, \ldots, y_N\} \subset \mathcal{B}$

$$\mathrm{E}\left|\sum_{i=1}^{N} U_i y_i\right| \leq b\left(\sum_{i=1}^{N} |y_i|^2\right)$$

where $U_1, \ldots, U_N$ are i.i.d. random variables,

$$P\left(U_1 = \frac{1}{2}\right) = P\left(U_1 = -\frac{1}{2}\right) = \frac{1}{2},$$

and constant $b \in \mathbb{R}$ depends neither on $\{y_1, \ldots, y_N\}$ nor on N. Such Banach space is called Rademacher type 2.

3.4.1 Law of Large numbers

Theorem 31. *Assume that $\mu_X \in \mathcal{L}^p(\mathcal{H})$, $p \geq 2$, and $X_1, \ldots, X_N \sim \mu_X$ are i.i.d. random variables. Then*

$$\left\|\overline{X}_N - \mathrm{E}X_1\right\|_p \leq \frac{c_p}{\sqrt{N}} \|X_1\|_p,$$

where real constant c_p only depends on p.

Proof. First, assume that $\mathrm{E}X_1 = 0$. If $p = 2$, then from the Cauchy-Schwarz inequality

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^{N} X_i\right\|_2 &= \left(\mathrm{E}\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right|^2\right)^{1/2} = \left(\frac{1}{N^2}\sum_{i,j=1}^{N} \mathrm{E}\langle X_i, X_j\rangle\right)^{1/2} \\ &= \left(\frac{1}{N^2}\sum_{i=1}^{N} \mathrm{E}|X_i|^2\right)^{1/2} = \frac{1}{\sqrt{N}}\|X_1\|_2. \end{aligned} \tag{3.30}$$

If $p > 2$, then, using the Hölder inequality,

$$\begin{aligned}\sum_{i=1}^{N} |1|\,|X_i|^2 \leq &\left(\sum_{i=1}^{N} |1|^{p/(p-2)}\right)^{(p-2)/p} \left(\sum_{i=1}^{N} \left(|X_i|^2\right)^{p/2}\right)^{2/p} \\ = &N^{(p-2)/p}\left(\sum_{i=1}^{N} |X_i|^p\right)^{2/p},\end{aligned}$$

and, using the Markinciewicz-Zygmund inequality,

$$\begin{aligned}\left\|\sum_{i=1}^{N} X_i\right\|_p^p \leq & b_p \mathrm{E}\left(\sum_{i=1}^{N} |X_i|^2\right)^{p/2} \\ \leq & b_p \mathrm{E}\left(N^{(p-2)/p}\left(\sum_{i=1}^{N} |X_i|^p\right)^{2/p}\right)^{p/2} \\ = & b_p N^{p/2-1} \sum_{i=1}^{N} \mathrm{E}\,|X_i|^p = b_p N^{p/2} \|X_1\|_p^p .\end{aligned}$$

Therefore,

$$\left\|\frac{1}{N}\sum_{i=1}^{N} X_i\right\|_p \leq \frac{b_p}{\sqrt{N}} \|X_1\|_p . \tag{3.31}$$

Now, if $\mathrm{E}X_1 \neq 0$, then $X_i - \mathrm{E}X_1$, $i = 1, \ldots, N$, are i.i.d. centered random variables, and we can apply (3.30) amd (3.31) to obtain

$$\begin{aligned}\left\|\overline{X}_N - \mathrm{E}X_1\right\|_p = & \left\|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathrm{E}X_1)\right\|_p \leq \frac{b_p}{\sqrt{N}} \|X_1 - \mathrm{E}X_1\|_p \\ \leq & \frac{b_p}{\sqrt{N}}\left(\|X_1\|_p + \|X_1\|_1\right) \leq \frac{2b_p}{\sqrt{N}} \|X_1\|_p ,\end{aligned}$$

where we used the identity

$$\|\mathrm{E}X_1\|_p = \mathrm{E}X_1 = \|X_1\|_1 .$$

□

Theorem 32. *Assume that $\mu_X \in \mathcal{L}^{2p}(\mathcal{H})$, $p \geq 2$, and $X_1, \ldots, X_N \sim \mu_X$ are i.i.d. random variables. If $\widehat{\mathrm{C}}_N$ denotes the sample covariance, i.e.,*

$$\widehat{\mathrm{C}}_N = \frac{1}{N-1}\sum_{i=1}^{N}\left(\left(X_i - \overline{X}_N\right) \otimes \left(X_i - \overline{X}_N\right)\right),$$

Then

$$\left\|\widehat{\mathrm{C}}_N - \mathrm{cov}\,(X_1)\right\|_p \leq \left(\frac{c_p}{\sqrt{N}} + \frac{c_{2p}^2}{N}\right) \|X_1\|_{2p}^2 ,$$

where real constants c_p and c_{2p} depend on p only.

Proof. By Theorem 13 a covariance of a random variable is a trace class operator, so it is also a Hilber-Schmidt operator. A sample covariance is a finite sum of tensor products, which are trace class operators, so the sample covariance is also trace class. Therefore, to prove the theorem we use the fact that

$$\{\mathrm{T} \in [\mathcal{H}] : \ |\mathrm{T}|_{\mathrm{HS}} < \infty\}$$

is a Hilbert space, and we can use the Markinciewicz-Zygmund inequality.

Covariance operators of random variables X_i and $X_i - \mathrm{E}X_1$ are identical, so, without loss of generality, we can assume that $\mathrm{E}X_1 = 0$. By the triangle inequality

$$\left\|\widehat{\mathrm{C}}_N - \mathrm{cov}\,(X_1)\right\|_p \leq \left\|\frac{1}{N}\sum_{i=1}^{N}(X_i \otimes X_i) - \mathrm{cov}\,(X_1)\right\|_p + \left\|\overline{X}_N \otimes \overline{X}_N\right\|_p . \quad (3.32)$$

Random variables $X_i \otimes X_i$, $i = 1, \ldots, N$, are i.i.d., and we can use Theorem 31 to obtain

$$\left(\mathrm{E}\left|\frac{1}{N}\sum_{i=1}^{N}(X_i \otimes X_i) - \mathrm{E}\,(X_1 \otimes X_1)\right|_{\mathrm{HS}}^{p}\right)^{1/p} \leq \frac{c_p}{\sqrt{N}}\left(\mathrm{E}\,|X_1 \otimes X_1|_{\mathrm{HS}}^{p}\right)^{1/p} \quad (3.33)$$

for some $c_p > 0$. Using Lemma 5,

$$\left(\mathrm{E}\,|X_1 \otimes X_1|_{\mathrm{HS}}^{p}\right)^{1/p} = \left(\mathrm{E}\,|X_1|^{2p}\right)^{1/p} = \|X_1\|_{2p}^{2}, \quad (3.34)$$

and using (2.14), (3.33) and (3.34) gives

$$\left\|\frac{1}{N}\sum_{i=1}^{N}(X_i \otimes X_i) - \mathrm{cov}\,(X_1)\right\|_p \leq \frac{c_p}{\sqrt{N}}\|X_1\|_{2p}^{2}. \quad (3.35)$$

For any $x \in \mathcal{H}$

$$|x \otimes x| = |x|\,|x|,$$

so the second term on the right side of (3.32) is bounded,

$$\begin{aligned}\left\|\overline{X}_N \otimes \overline{X}_N\right\|_p &= \left(\mathrm{E}\left|\overline{X}_N\right|^p \left|\overline{X}_N\right|^p\right)^{1/p} \\ &\leq \left\|\overline{X}_N\right\|_{2p}\left\|\overline{X}_N\right\|_{2p}, \end{aligned} \quad (3.36)$$

and again Theorem 31 gives an existence of a constant $c_{2p} > 0$ such that

$$\left\|\overline{X}_N\right\|_{2p} \leq \frac{c_{2p}}{\sqrt{N}} \quad (3.37)$$

since $\mathrm{E}X_1 = 0$. To finish the proof, we just put together (3.32), (3.35), (3.36) and (3.37), so

$$\left\|\widehat{\mathrm{C}}_N - \mathrm{cov}\,(X_1)\right\|_p \leq \left(\frac{c_p}{\sqrt{N}} + \frac{c_{2p}^2}{N}\right)\|X_1\|_{2p}^{2}.$$

□

Obviously, if a, b are two positive numbers, then

$$\frac{a}{\sqrt{N}} + \frac{b}{N} \leq \frac{2\max\{a,b\}}{\sqrt{N}}$$

for any $N \in \mathbb{N}$, and this simple observation gives us an important corollary of the previous theorem.

Corollary 4. Using the assumptions and notation of Theorem 32, there is a real constants c_p such that

$$\left\|\widehat{\mathrm{C}}_N - \mathrm{cov}\,(X_1)\right\|_p \leq \frac{c_p}{\sqrt{N}}\|X_1\|_{2p}^{2}.$$

3.5 Bayes theorem

Assume that $X \in \mathcal{L}(\mathbb{R}^n)$ and $Y \in \mathcal{L}(\mathbb{R}^m)$ are random vectors with densities f_X and f_Y respectively. If

$$f_{Y|x} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}_+ \tag{3.38}$$

is a conditional density of Y given a condition $X = x$, then the Bayes theorem states that

$$f_{X|y}(x|y) = \frac{1}{c(y)} f_{Y|x}(y|x) f_X(x) \quad \text{if} \quad c(y) > 0, \tag{3.39}$$

where

$$c(y) = \int_{\mathbb{R}^n} f_{Y|x}(y|x) f_X(x) \, d\lambda^n(x),$$

is a conditional density of X given $Y = y$. The distribution of X is called a prior distribution, and the conditional distribution of X given $Y = y$ is called a posterior distribution. Note that the posterior distribution is absolutely continuous with respect to the prior distribution, and the function (3.38) is Radon-Nikodym derivative, i.e.,

$$\frac{d\mu_{X|y}}{d\mu_X} = f_{Y|x}.$$

Usually, the derivative is called a data likelihood.

When $X, Y \in \mathcal{L}(\mathcal{H})$ and $\dim(\mathcal{H}) = \infty$, the formula (3.39) is inapplicable because the Lebesgue measure on $\mathcal{H}$ does not exist, Theorem 9. However, Bayes theorem still holds, and we can define the posterior measure

$$\mu_{X|y}(A) = P(X \in A|Y = y) = \frac{1}{c(y)} \int_A d(y|x) \, d\mu_X(x) \quad \forall A \in \mathcal{B}(\mathcal{H})$$

if

$$c(y) = \int_{\mathcal{H}} d(y|x) \, d\mu_X(x) > 0.$$

The function

$$d : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+, \quad d = \frac{d\mu_{X|y}}{d\mu_X},$$

is also called data likelihood.

3.6 Additional notes and references

The presentation in this chapter is partially based on Mandel [2016]. The majority of definitions and statements introduced in this section also holds when $\mathcal{H}$ is a separable Banach space, but we do not need such a general approach, which brings additional problems with measurability. The probability on Banach spaces is excellently covered by, for example, Ledoux and Talagrand [1991]. Bogachev [1998] and Vakhania et al. [1987] cover the theory of weak random variables, and the latter discusses weak stochastic norms in much greater detail.

There are many great books that deal with Gaussian measure on Banach or Hilbert spaces such as Bogachev [1998] or Da Prato [2006]. This topic is closely related to stochastic differential equations, so many useful details may be found

in Da Prato and Zabczyk [1992]. A more statistical approach to an infinite dimensional random variable is presented by Ramsay and Silverman [2002, 2005].

The existence of the Cameron-Martin space was identified in R. H. Cameron [1944], and the Feldman-Hájek theorem was firstly proved in Feldman [1958]. More recently, both these theorems are covered by Bogachev [1998], Da Prato and Zabczyk [1992] and Da Prato [2006].

The Rademacher spaces, mentioned at the end of Section 3.4, are studied in Hoffmann-Jørgensen [1974] and Ledoux and Talagrand [1991]. More details about the Markinciewicz-Zygmund inequality can be found in Chow and Teicher [1997] or Woyczyński [1980].

Finally, the Bayes theorem may be found in nearly any statistical or probability textbook. The definition of the Bayes theorem on infinite dimensional spaces may be found in Stuart [2010], and will be later discussed in Section 6.3.

4. State space model and data assimilation

This chapter introduces a state space model and a data assimilation, and contains multiple examples of state space models. Readers familiar with these topics may skip this entire section.

The chapter is organized as follows. Section 4.1 defines a state space model, and Definition 7 introduces the notation used in all subsequent chapters of the thesis. Section 4.2 defines a data assimilation and its relations to filtering, and Section 4.3 contains important references.

4.1 State space model

A state space model usually represents an evolution of a physical phenomena in nature. It consist of two parts: a dynamical system and noisy observations. While the dynamical system fully describes the modeled phenomena, the observations represent the part of the system that can be detected by available measuring devices. A typical example is the atmosphere of the Earth, where the underlying dynamical system represent the evolution of the atmosphere over the the whole globe, and the observations may represent satellite images of clouds, radiosonde measurements at different locations, etc.

4.1.1 Dynamical system

In general, a dynamical system is a random process defined on a Banach space whose evolution is governed by a known mapping, and this mapping is usually a solution to a differential equation. Such general definition may bring problems with a measurability of individual states of the system. To avoid these problems, we limit ourselves to systems defined on a separable Hilbert space with a Gaussian initial condition and Gaussian errors.

Definition 6. *Assume that the following premises hold.*

1. $\mathcal{H}$ *is a separable Hilbert space.*
2. *A mapping* $\Psi : \mathcal{H} \to \mathcal{H}$ *is measurable.*
3. *A random variable* $X^{(0)} \sim \mathcal{N}\left(m^{(0)}, \mathrm{P}^{(0)}\right)$, *and* $\mathrm{P}^{(0)}$ *is positive semidefinite.*
4. *Random variables* $V^{(t)}$, $t \in \mathbb{N}$, *defined on* $\mathcal{H}$ *are mutually independent random variables on* $\mathcal{H}$, *and each* $V^{(t)} \sim \mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$ *with* $\mathrm{Q}^{(t)}$ *being positive semidefinite.*

Then, a random process $\left\{X^{(t)}\right\}_{t\in\mathbb{N}}$ *defined by*

$$X^{(t)} = \Psi\left(X^{(t-1)}\right) + V^{(t)} \quad \forall t \in \mathbb{N}$$

is called a discrete time dynamical system.

The mapping Ψ is called an iterated map; $X^{(t)}$ are called states of the dynamical system; $V^{(t)}$ are called model errors; $X^{(0)}$ is an initial condition of the dynamical system, and the space $\mathcal{H}$ is called a state space.

The definition admits that all $\mathrm{Q}^{(t)}$ are null operators, i.e.,

$$V^{(t)} = 0 \quad \text{a.s.} \quad \forall t \in \mathbb{N}. \tag{4.1}$$

Hence, we can the divide dynamical systems into two categories. If $\left\{X^{\{t\}}\right\}$ is a dynamical system such that condition (4.1) is satisfied, then we say that the system is governed by deterministic dynamics. Conversely, if $\mathrm{Q}^{(t)}$ are nonzero operators, then we say that the system is governed by stochastic dynamics.

To summarize, a dynamical system $\left\{X^{\{t\}}\right\}$ is completely determined by the distribution of its initial condition, a set of covariance operators $\left\{\mathrm{Q}^{(t)}\right\}$, and the iterated map Ψ. Before showing multiple examples of different dynamical systems we formulate an obvious lemma that shows an important property of any dynamical system.

Lemma 33. *Let $\left\{X^{(t)}\right\}$ be the dynamical system from Definition 6, then the random process $\left\{X^{(t)}\right\}$ is Markov, i.e.,*

$$P\left(X^{(t)} \in B \middle| X^{(t-1)}, \ldots, X^{(0)}\right) = P\left(X^{(t)} \in B \middle| X^{(t-1)}\right)$$

for any $B \in \mathcal{B}(\mathcal{H})$ and all $t \in \mathbb{N}$.

Proof. The statement follows directly from the definition because

$$P\left(X^{(t)} \in B \middle| X^{(t-1)}, \ldots, X^{(0)}\right) = P\left(V^{(t)} \in A \middle| X^{(t-1)}\right)$$

where

$$A = \left\{x \in \mathcal{H} : \ \left(\Psi\left(X^{(t-1)}\right) + x\right) \in B\right\}.$$

□

Example 13. Let Ψ be a real function,

$$\Psi : \ x \in \mathbb{R} \to \lambda x \in \mathbb{R}$$

for a given $\lambda \in \mathbb{R}$, and define

$$X^{(0)} = k \quad \text{a.s.}$$

for some $k \in \mathbb{R}$. Obviously, all states of dynamical system $\left\{X^{(t)}\right\}$,

$$X^{(t)} = \Psi\left(X^{(t-1)}\right),$$

are deterministic,

$$X^{(t)} = \lambda^t k \quad \text{a.s.},$$

and $\left\{X^{(t)}\right\}$ is just a sequence of real numbers.

Example 14. Using the function Ψ from the previous example with initial condition

$$X^{(0)} \sim \mathcal{N}\left(m, \sigma^2\right),\ m \in \mathbb{R}, \sigma \geq 0,$$

we get another dynamical system $\left\{X^{(t)}\right\}$. In this case

$$X^{(t)} \sim \mathcal{N}\left(\lambda^t m, \lambda^{2t}\sigma^2\right)$$

for each $t \in \mathbb{N}$, and, clearly, this system coincides with the system in Example 13 if $\sigma = 0$ and $m = k$.

In the previous examples it was straightforward to determine the distribution of the state at any given time using the linearity of the mapping Ψ. However, that may not be the case in many useful applications as shown in the next example.

Example 15. Given a continuous functions

$$f:\ \mathbb{R} \to \mathbb{R}$$

and a first order differential equation

$$\frac{\partial v}{\partial \tau} = f\left(v\right) \tag{4.2}$$

with an assumption that

$$v : [0,1] \to \mathbb{R}$$

is continuous we define the function

$$\Psi:\ \mathbb{R} \to \mathbb{R}$$

by

$$\Psi\left(x\right) = \widehat{v}_x\left(1\right)$$

where $\widehat{v}_x$ is a solution of Equation (4.2) given an initial condition $v\left(0\right) = x$. The distribution of the states of the dynamical system

$$\begin{aligned} X^{(0)} \sim& \mathcal{N}\left(0, \sigma^2\right),\ \sigma \geq 0, \\ X^{(t)} =& \Psi\left(X^{(t)}\right),\ t \in \mathbb{N}, \end{aligned}$$

is generally unknown unless $\sigma = 0$, when the system is fully degenerate.

All three examples above contain systems with deterministic dynamics, so all states of these systems are completely determined by its initial condition. Examples of systems with stochastic dynamics follow.

Example 16. Similarly as in Example 14, let Ψ be a real function,

$$\Psi:\ x \in \mathbb{R} \to \lambda x \in \mathbb{R}$$

for a given $\lambda \in \mathbb{R}$, and $V^{(t)} \sim \mathcal{N}\left(0,1\right)$, $t \in \mathbb{N}$, be i.i.d. random variables. Then, equations

$$\begin{aligned} X^{(0)} \sim& \mathcal{N}\left(m, \sigma^2\right),\ m \in \mathbb{R}, \sigma > 0, \\ X^{(t)} =& \lambda X^{(t-1)} + V^{(t)},\ t \in \mathbb{N}, \end{aligned}$$

define a system with a stochastic dynamics. Clearly, using the properties of Gaussian distribution,

$$X^{(1)} \sim \mathcal{N}\left(\lambda m, \lambda^2\sigma^2 + 1\right), \quad X^{(2)} \sim \mathcal{N}\left(\lambda^2 m, \lambda^4\sigma^2 + \lambda^2 + 1\right),$$

and, using induction,

$$X^{(t)} \sim \mathcal{N}\left(\lambda^t m, \lambda^{2t}\sigma^2 + \sum_{i=0}^{t-1}\lambda^{2i}\right)$$

for each $t \in \mathbb{N}$.

Of course, in real world application the dimension of the state space is much higher. The next example shows a dynamical system defined on an infinite dimensional Hilbert space.

Example 17. The heat equation

$$\frac{\partial v}{\partial \tau} = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2},$$

where v is a four dimensional real function, describe the distribution of heat in a three dimensional area. Let

$$u_f : \ [0,\infty) \times [0,1]^3 \to \mathbb{R}$$

be a solution of the heat equation satisfying an initial condition

$$u_f(0,x,y,z) = f(x,y,z)$$

for a continuous function $f : \ [0,1]^3 \to \mathbb{R}$. For a given $\Delta > 0$ define mapping

$$\Psi : \ \mathcal{L}^2\left([0,1]^3, \mathbb{R}\right) \to \mathcal{L}^2\left([0,1]^3, \mathbb{R}\right) \tag{4.3}$$

by

$$\Psi(f)(x,y,z) = u_f(\Delta,x,y,z) \tag{4.4}$$

for all $x, y, z \in [0,1]$. Now, we may define a dynamical by

$$\begin{aligned} X^{(0)} \sim & \mathcal{N}\left(m^{(0)}, \mathrm{P}^{(0)}\right), \\ X^{(t)} = & \Psi\left(X^{(t-1)}\right) + V^{(t)}, \ t \in \mathbb{N} \end{aligned}$$

with $\mathrm{P}^{(0)}$ being a non degenerate covariance operator on

$$\mathcal{H} = L^2\left([0,1]^3, \mathbb{R}\right),$$

$m^{(0)} \in \mathcal{H}$, and $V^{(t)} \sim \mathcal{N}(0, \mathrm{Q})$, $t \in \mathbb{N}$, being i.i.d. random variables.

Clearly, it is not clear whether the solution (4.4) exists for any continuous f. However, since the discussion about existence of the solution exceeds the scope of this thesis, we silently assume that the solution exists at least in some weak sense, and that the function Ψ is measurable.

The resulting dynamical system $\left\{X^{(t)}\right\}$ represents the distribution of heat in a three dimensional unit cube in discrete time steps Δ. The random variables $V^{(t)}$ may represent, for example, an unknown heat flow between the cube and a surrounding area.

A discretization of the dynamical system from the previous example shows us another potential source of the model error.

Example 18. Let $\{x_1, \ldots, x_n\}$ be a discretization of a three dimensional unit cube, and define a linear operator Π by

$$\Pi : \ f \in L^2\left([0,1]^3, \mathbb{R}\right) \to \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \in \mathbb{R}^n.$$

Using the system $\left\{X^{(t)}\right\}$ introduced in Example 17 we can define the new dynamical system $\left\{\widetilde{X}^{(t)}\right\}$,

$$\widetilde{X}^{(t)} = \Pi X^{(t)}, \ t = 0, 1, 2, \ldots$$

Obviously, the initial condition of the new system is

$$\widetilde{X}^{(0)} \sim \mathcal{N}\left(\Pi m^{(0)}, \Pi \mathrm{P}^{(0)} \Pi^*\right),$$

and the evolution of the system may be described by

$$\widetilde{X}^{(t)} = \widetilde{\Psi}\left(\widetilde{X}^{(t-1)}\right) + \widetilde{V}^{\{t\}}$$

with $\widetilde{\Psi}$ being an approximation of Ψ and

$$\widetilde{V}^{(t)} = \Pi V^{(t)} + \widehat{V}^{(t)}$$

where $\widehat{V}^{(t)}$ represents the error of approximation Ψ using $\widetilde{\Psi}$. When the mesh $\{x_1, \ldots, x_n\}$ is well designed, we can assume that the error $\widehat{V}^{(t)}$ is centered, and it has a Gaussian distribution with a known covariance $\widehat{\mathrm{Q}}^{(t)}$. Hence, the random variable $\widetilde{V}^{(t)}$ has the centered Gaussian distribution with covariance operator

$$\widetilde{\mathrm{Q}}^{(t)} = \Psi \mathrm{Q}^{(t)} \Psi^* + \widehat{\mathrm{Q}}^{(t)}.$$

Examples 17 and 18 show two usual reasons for stochastic dynamics of a dynamical system. The first one is an imperfect model, which, for example, do not capture all interactions of the system with a surrounding environment. The second reason is an error of a necessary approximation of an iterated map, e.g., when the iterated map is a solution of a differential equation, this equation may be solvable only using numerical methods.

4.1.2 Observations

Assume that $\left\{X^{(t)}\right\}$ is a dynamical system defined on a separable Hilbert space $\mathcal{H}$, and $\mathcal{G}$ is another separable Hilbert space. If

$$h : \ \mathcal{H} \to \mathcal{G}$$

is a measurable mapping, and $W^{(t)}$, $t \in \mathbb{N}$, are pairwise independent weak random variables, each $W^{(t)} \sim \mathcal{N}_w\left(0, \mathrm{R}^{(t)}\right)$, then weak random variables

$$Y^{(t)} = h\left(X^{(t)}\right) + W^{(t)},$$

are called observations of the dynamical system. The mapping h is called an observation operator, and the space $\mathcal{G}$ is called an observation space. If the dimension of the observational space is finite, $Y^{(t)}$ is also called an observational vector.

The next example shows that a set of all measurable mappings from the state space to the observation space is an unnecessarily big class of operators, and we can limit ourselves to only linear observation operators.

Example 19. Assume that $\left\{X^{(t)}\right\}$ is the dynamical system from Definition 6, and

$$h:\ \mathcal{H}\to\mathcal{G}$$

is an observational operator with $\mathcal{G}$ being a separable Hilbert space. Let

$$Y^{(t)}=h\left(X^{(t)}\right)+W^{(t)}$$

be observations with $W^{(t)}\sim\mathcal{N}_w\left(0,\mathrm{R}^{(t)}\right),\ t\in\mathbb{N}$, being pairwise independent. The space $\widetilde{\mathcal{H}}=\mathcal{H}\oplus\mathcal{G}$ is also Hilbert, and a function

$$\widetilde{\Psi}:\ \widetilde{\mathcal{H}}\to\widetilde{\mathcal{H}}$$

defined for all

$$x=(x_{\mathcal{H}},x_{\mathcal{G}})\in\widetilde{\mathcal{H}},\ x_{\mathcal{H}}\in\mathcal{H},\ x_{\mathcal{G}}\in\mathcal{G},$$

by

$$\widetilde{\Psi}\left(x\right)=\left(\Psi\left(x_{\mathcal{H}}\right),h\left(\Psi\left(x_{\mathcal{H}}\right)\right)\right)$$

is measurable. Hence, we can define the new dynamical system $\left\{\widetilde{X}^{(t)}\right\}$ with the initial condition

$$\widetilde{X}^{(0)}=\left(X^{(0)},0\right)$$

and the iterated map $\widetilde{\Psi}$. If we define

$$\mathrm{H}:\ (x_{\mathcal{H}},x_{\mathcal{G}})\in\widetilde{\mathcal{H}}\to x_{\mathcal{G}}\in\mathcal{G},$$

then for each $t\in\mathbb{N}$

$$Y^{(t)}=\mathrm{H}\widetilde{X}^{(t)}+W^{(t)}.$$

Now, instead of working with the original dynamical system $\left\{X^{(t)}\right\}$ with the measurable observation operator h, we can work with the augmented dynamical system $\left\{\widetilde{X}^{(t)}\right\}$ with the linear observation operator H.

The previous example allow us, without loss of generality, to assume that an observation operator is always linear.

We show a typical example of observations using the dynamical system from Example 17.

Example 20. Recall that the dynamical system $\left\{X^{(t)}\right\}$ introduced in Example 17 describes the evolution of the temperature of a three dimensional homogeneous unit cube. If $B\in\mathcal{B}\left([0,1]^3\right)$, then mapping

$$\mathrm{H}:\ f\in L^2\left([0,1]^3,\mathbb{R}\right)\to\frac{\int_B f\left(x\right)d\lambda^3\left(x\right)}{\lambda^3\left(B\right)}$$

is a linear observational operator from the state space to the observational space $\mathbb{R}$, and observations

$$Y^{(t)} = \mathrm{H}X^{(t)} + W^{(t)},\ W^{(t)} \sim \mathcal{N}\left(0, \sigma^2\right),\ \sigma > 0,$$

represent the mean temperature of the area described by B with $W^{(t)}$ standing for a measurement device error.

4.1.3 Summary

To summarize, a state space model is a couple consisting of a dynamical system and a set of noisy observations. Since the definition of a state space model is a crucial part of the thesis, we summarize it in the next definition. The notation introduced in this definition is used in all subsequent chapters.

To avoid unnecessary obstacles with a kernel of a random variable we assume, without loss of generality, that all random variables in the next definition are non-degenerate.

Definition 7 (State space model)**.** *Assume that $\mathcal{H}$ and $\mathcal{G}$ are separable Hilbert spaces, and the following statements hold.*

1. *A random variable $X^{(0)} \sim \mathcal{N}\left(m^{(0)}, \mathrm{P}^{(0)}\right)$ with $m^{(0)} \in \mathcal{H}$, and the operator $\mathrm{P}^{(0)} \in \mathcal{H}$ is positive definite, symmetric and trace class.*

2. *A dynamical system $\left\{X^{(t)}\right\}_{t\in\mathbb{N}_0}$ defined on $\mathcal{H}$ is governed by an iterated map*

$$\Psi :\ \mathcal{H} \to \mathcal{H}$$

with independent model errors $V^{(t)} \sim \mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$, $\mathrm{Q}^{(t)}$ positive definite, $t \in \mathbb{N}$, i.e.,

$$X^{(t)} = \Psi\left(X^{(t-1)}\right) + V^{(t)},\ t \in \mathbb{N}.$$

3. *An observation operator $\mathrm{H} \in [\mathcal{H}, \mathcal{G}]$.*

4. *For each $t \in \mathbb{N}$*

$$Y^{(t)} = \mathrm{H}X^{(t)} + W^{(t)},$$

where $W^{(t)} \sim \mathcal{N}_w\left(0, \mathrm{R}^{(t)}\right)$, $t \in \mathbb{N}$, are mutually independent week Gaussian random variables on $\mathcal{G}$.

We say that the dynamical system $\left\{X^{(t)}\right\}$ and the observations $\left\{Y^{(t)}\right\}$ establish the state space model.

We always write that there is a state space model with an underlying dynamical system $\left\{X^{(t)}\right\}$ and a set of noisy observations $\left\{Y^{(t)}\right\}$, and by saying this we assume that everything that determines the model, i.e., the initial condition, the iterated map, the observation operator and the distribution of the model and observational error, is known. Unless explicitly noted, we always use the same notation as in the last definition.

By saying that a state space model is infinite dimensional we always mean that $\dim(\mathcal{H}) = \infty$. Additionally, if $\dim(\mathcal{G}) = \infty$ as well, then we usually assume that $\mathcal{H} = \mathcal{G}$, since any two separable Hilbert spaces are isometric.

4.2 Data assimilation

Given the state space model from Definition 7, a data assimilation is a sequential process of estimating the state $X^{(t)}$ of the underlying dynamical system using the observations

$$Y^{(t)}, Y^{(t-1)}, \ldots, Y^{(1)}$$

and a prior estimate of the state $X^{(t)}$, which is based on the observations up to time $t-1$. Hence, the data assimilation is a filtering problem, and its goal is to produce the best estimate of the true states of the system as observations become available. The step by step definition of data assimilation procedure follows.

Definition 8. *Using the notation from Definition 7, a data assimilation algorithm consists of the following steps.*

1. *An initialization, when the initial estimate is generated from the true distribution of $X^{(0)}$. We denote this estimate $X^{(0),a}$, and call it a first guess.*

2. *Recursively repeated data assimilation cycles. One data assimilation cycle for a given t consists of two steps:*

   (a) *forecast step,*

   $$X^{(t),f} = \Psi\left(X^{(t-1),a}\right) + V^{(t),f} \tag{4.5}$$

   with $V^{(t),f}$ being sampled independently from $\mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$ distribution, and

   (b) *analysis step, when the prior estimate $X^{(t),f}$ is combined with the observation $Y^{(t)}$ to produce new estimate $X^{(t),a}$.*

The prior estimate $X^{(t),f}$ is called a forecast, and the estimate $X^{(t),a}$ is called an analysis.

Obviously, the distribution of the forecast $X^{(t),f}$, for a given $t \in \mathbb{N}$, depends on the particular value of the observations up to time $t-1$, so using the previous definition

$$P\left(X^{(t),f} \in B\right) = P\left(X^{(t)} \in B \middle| Y^{(t-1)} = y^{(t-1)}, \ldots, Y^{(1)} = y^{(1)}\right),$$

for any $B \in \mathcal{B}(\mathcal{H})$, where $y^{(t-1)}, \ldots, y^{(1)}$ are the already assimilated observations. Similarly,

$$P\left(X^{(t),a} \in B\right) = P\left(X^{(t)} \in B \middle| Y^{(t)} = y^{(t)}, \ldots, Y^{(1)} = y^{(1)}\right)$$

for any $B \in \mathcal{B}(\mathcal{H})$.

It is important to emphasize that both forecast and analysis distributions are conditional, and they are conditioned on the values of the observations. Hence, when we say that $X^{(t),f}$ and another random variable Z are independent, it always means that they are conditionally independent given observations up to time $t-1$. The same applies on the analysis $X^{(t),a}$.

We conclude the section with an obvious, yet very important, lemma.

Lemma 34. *For a given t, the forecast state $X^{(t),f}$ and observations $Y^{(t)}$ are conditionally independent given $X^{(t-1)}$.*

Proof. Let $B_1, B_2 \in \mathcal{B}(\mathcal{H})$ and $D \in \mathcal{B}(\mathcal{G})$. Using the chain rule for conditional probabilities,

$$\begin{aligned} P\left(X^{(t),f} \in B_1, Y^{(t)} \in D \middle| X^{(t-1)} \in B_2\right) =& P\left(X^{(t),f} \in B_1 \middle| Y^{(t)} \in D, X^{(t-1)} \in B_2\right) \\ & \cdot P\left(Y^{(t)} \in D \middle| X^{(t-1)} \in B_2\right), \end{aligned}$$

and the statement of the lemma follows because

$$P\left(X^{(t),f} \in B_1 \middle| Y^{(t)} \in D, X^{(t-1)} \in B_2\right) = P\left(X^{(t),f} \in B_1 \middle| X^{(t-1)} \in B_2\right).$$

□

4.3 Additional notes and references

Clearly, assumptions that both types of errors in a state space model are Gaussian may be omitted, and one can work with more general state space models. Such models may be found in Cressie [1993], Durbin and Koopman [2012], etc. However, we limit our scope to the model introduced in Definition 7 because this is the standard model used in atmospheric sciences. A physical motivation for this type of model and multiple examples of dynamical systems in the atmosphere may be found in Jacobson [2005], Kalnay [2003], etc. Examples of dynamical processes in oceans may be found in Bennett [1992, 2002].

One data assimilation cycle may also be understand as an inverse problem, and multiple books studying this approach have been published recently, e.g., Nakamura and Potthast [2015], and van Leeuwen et al. [2015]. Extensive mathematical backgrounds of the data assimilation are discussed in, for example, Banks et al. [2014] or Law et al. [2015], and both books also contain an extension of a state space model to a case when the time is a continuous variable. As already mentioned, the data assimilation is a filtering problem, and extensive study of filtering techniques is provided by Anderson and Moore [1979], Jazwinski [1970], etc.

Obviously, one may think of assimilating multiple observations at once, i.e., the data assimilation procedure may be adjusted, so that the assimilation is provided only every L time steps. In this situation, the assimilation step consists of combining forecasts $X^{(t),f}, \ldots, X^{(t+L),f}$, where now

$$\begin{aligned} X^{(t),f} &= X^{(t)} \middle| Y^{(t-1)} = y^{(t-1)}, \ldots, Y^{(1)} = y^{(1)}, \\ X^{(t+L),f} &= X^{(t+L)} \middle| Y^{(t-1)} = y^{(t-1)}, \ldots, Y^{(1)} = y^{(1)}, \end{aligned}$$

and observations

$$Y^{(t)} = y^{(t)}, \ldots, Y^{(t+L)} = y^{(t+L)}$$

to produce analysis $X^{(t),a}, \ldots, X^{(t+L),a}$, where

$$\begin{aligned} X^{(t),a} &= X^{(t)} \middle| Y^{(t+L)} = y^{(t+L)}, \ldots, Y^{(1)} = y^{(1)}, \\ X^{(t+L),a} &= X^{(t+L)} \middle| Y^{(t+L)} = y^{(t+L)}, \ldots, Y^{(1)} = y^{(1)}. \end{aligned}$$

Hence, the analysis state $X^{(t),a}$ is conditioned by the values of future observations $y^{(t+1)}, \ldots, y^{(t+L)}$ unless $L = 1$. The number L is called a length of a assimilation window, and assimilation method with $L > 1$ are called smoothing methods. Conversely, methods with length of assimilation windows equal to one are called filtering methods.

5. Data assimilation in finite dimension

This section introduces four well known data assimilation methods available when a state space model is finite dimensional. The methods are:

- the 3DVAR in Section 5.1,
- the Kalman filter (KF) in Section 5.2,
- the ensemble Kalman filter (EnKF) in Section 5.3 and
- the Bayesian filtering (BF) in Section 5.4.

Through the whole section we use the state space model from Definition 7 with two additional assumptions:

1. we denote n the dimension of the state space $\mathcal{H}$, so, without loss of generality, $\mathcal{H} = \mathbb{R}^n$, and
2. we denote m the dimension of the observation space $\mathcal{G}$, so $\mathcal{G} = \mathbb{R}^m$.

5.1 3DVAR

The 3DVAR is a variational method, and it produces point estimates of the states of the underlying dynamical system. In general, variational methods are based on a minimization of a defined cost function.

For a given $t \in \mathbb{N}$, forecast $X^{(t),f} = x^{(t),f}$, and observation vector $Y^{(t)} = y^{(t)}$ the 3DVAR cost function is

$$J^{\text{3DVAR}}(x) = \left|x - x^{(t),f}\right|^2_{\mathrm{B}^{-1}} + \left|y^{(t)} - \mathrm{H}x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}, \tag{5.1}$$

where $\mathrm{B} \in \mathrm{R}^{n\times n}$ is a prescribed covariance matrix, called the background covariance, and for any $x \in \mathbb{R}^n$

$$|x|^2_{\mathrm{B}^{-1}} = \left\langle x, \mathrm{B}^{-1}x\right\rangle = \left\langle \mathrm{B}^{-1/2}x, \mathrm{B}^{-1/2}x\right\rangle = x^*\mathrm{B}^{-1}x.$$

Similarly, for any $x \in \mathbb{R}^m$

$$|x|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}} = \left\langle x, \left(\mathrm{R}^{(t)}\right)^{-1} x\right\rangle = \left\langle \left(\mathrm{R}^{(t)}\right)^{-1/2} x, \left(\mathrm{R}^{(t)}\right)^{-1/2} x\right\rangle = x^* \left(\mathrm{R}^{(t)}\right)^{-1} x.$$

Using the fact that both matrices B and $\mathrm{R}^{(t)}$ are positive definite it is easy to check that both functionals

$$x \in \mathbb{R}^n \to |x|_{\mathrm{B}^{-1}}$$

and

$$x \in \mathbb{R}^m \to |x|_{\left(\mathrm{R}^{(t)}\right)^{-1}}$$

define norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. The analysis is obtained by minimizing the cost function over the whole state space. The whole algorithm is summarized in the following definition.

Definition 9 (3DVAR). *Using the state space model from Definition 7, the 3DVAR assimilation algorithm consists of the following steps.*

1. *Generate a first guess $X^{(0),a}$ from the distribution of the initial condition $X^{(0)}$, or obtain the first guess using expert knowledge.*

2. *For $t \in \mathbb{N}$ recursively repeat the following steps.*

   (a) *Advance the analysis from the previous cycle,*

   $$X^{(t),f} = \Psi\left(X^{(t-1),a}\right) + V^{(t)}$$

   using $V^{(t)}$ independently generated from $\mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$.

   (b) *Given the forecast $X^{(t),f} = x^{(t),f}$ and the observation vector $Y^{(t)} = y^{(t)}$ update the analysis by minimizing the 3DVAR cost function:*

$$X^{(t),a} = \underset{x \in \mathbb{R}^n}{\arg\min} J^{3DVAR}(x). \tag{5.2}$$

The background covariance matrix B in (5.1) does not change or evolve in time, so the choice of this matrix is crucial. We briefly discuss this topic in the last section of this chapter.

5.2 Kalman filter

The Kalman filter updates not only the forecast state of the system, but also its mean and covariance. Recall that mean and covariance form together a sufficient statistic of a Gaussian distributed random variable. We introduce the KF algorithm in the next definition.

Definition 10 (Kalman filter). *Using the state space model and the notation from Definition 7, assume that the iterated map Ψ is linear, so there exist a matrix $\mathrm{A} \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$ such that*

$$\Psi(x) = \mathrm{A}x + b. \tag{5.3}$$

Then, the Kalman filter assimilation algorithm consists of the following steps.

1. *Generate the first guess $X^{(0),a}$ from the distribution of the initial condition $X^{(0)}$, and define*

   $$\mathrm{P}^{(0),a} = \mathrm{cov}\left(X^{(0)}\right) = \mathrm{P}^{(0)}.$$

2. *For $t \in \mathbb{N}$ recursively repeat the following steps.*

   (a) *Advance the analysis from the previous cycle,*

   $$X^{(t),f} = \Psi\left(X^{(t-1),a}\right) + V^{(t)}$$

   using $V^{(t)}$ independently generated form $\mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$.

(b) Propagate the analysis covariance in time,

$$\mathrm{P}^{(t),f} = \mathrm{A}\mathrm{P}^{(t-1),a}\mathrm{A}^* + \mathrm{Q}^{(t)}. \tag{5.4}$$

(c) Evaluate the Kalman gain matrix,

$$\mathrm{K}^{(t)} = \mathrm{P}^{(t),f}\mathrm{H}^* \left(\mathrm{H}\mathrm{P}^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1}. \tag{5.5}$$

(d) Using the observation $Y^{(t)} = y^{(t)}$, update the forecast,

$$X^{(t),a} = X^{(t),f} + \mathrm{K}^{(t)}\left(y^{(t)} - \mathrm{H}X^{(t),f}\right).$$

(e) Update the forecast covariance,

$$\mathrm{P}^{(t),a} = \left(\mathrm{I} - \mathrm{K}^{(t)}\mathrm{H}\right)\mathrm{P}^{(t),f}. \tag{5.6}$$

Additionally, if we denote the forecast and analysis mean by

$$\begin{aligned} m^{(t),f} =& \mathrm{E}X^{(t),f}, \\ m^{(t),a} =& \mathrm{E}X^{(t),a}, \end{aligned}$$

then from the definition of the KF immediately follows the recursive relations

$$\begin{aligned} m^{(t),f} =& \mathrm{A}m^{(t-1),a} + b, \\ m^{(t),a} =& m^{(t),f} + \mathrm{K}^{(t)}\left(y^{(t)} - \mathrm{H}m^{(t),f}\right) \end{aligned}$$

for all $t \in \mathbb{N}$ with $\mathrm{K}^{(t)}$ defined in (5.5). Similarly, one can directly show that

$$\mathrm{cov}\left(X^{(t),f}\right) = \mathrm{P}^{(t),f}$$

and also

$$\mathrm{cov}\left(X^{(t),a}\right) = \mathrm{P}^{(t),a}$$

for all $t \in \mathbb{N}$.

For a given $t \in \mathbb{N}$ the KF may be derived as the best linear unbiased estimate of the state $X^{(t)}$ given the observations $Y^{(t)} = y^{(t)}$ [Durbin and Koopman, 2012, Chapter 4]. When the distribution of $X^{(t),f}$ is Gaussian, the analysis obtained by the KF is also the estimate with the minimal mean square error and maximal likelihood estimate of $X^{(t)}$.

There are two big obstacles with application of the KF, especially in the area of atmospheric physics.

1. The KF equations assume that the dynamics of the underlying system is linear. When the iterated map Ψ is not linear, one needs to replace matrix A in Equation (5.4) with some linearization of Ψ. This may be done, for example, using adjoint and tangent operators of Ψ, which leads to the extended Kalman filter. However, the computation of these operators is usually very difficult.

2. The dimension of a modeled system may be huge, even a few billion, so working with the state covariance may easily become impractical, as manipulation with the matrix of size $10^9 \times 10^9$ is impossible even when one can use a supercomputer.

5.3 Ensemble Kalman filter

The ensemble Kalman filter resolves both obstacles of the KF mentioned at the end of the previous section. The basic idea is to represent the distribution of the underlying system using multiple samples, which may be interpreted as possible scenarios of an evolution of a modeled system, and use a sample covariance in place of the forecast covariance in the KF update equations.

The method was first published in Evensen [1994], and later improved in Burgers et al. [1998]. The next definition presents the version from the second cited paper.

Definition 11 (Ensemble Kalman filter)**.** *The ensemble Kalman filter consists of the following steps.*

1. *For a given $N \in \mathbb{N}$ generate i.i.d. random variables*

$$X_1^{(0),a}, \dots, X_N^{(0),a}$$

from the distribution of $X^{(0)}$.

2. *For $t \in \mathbb{N}$ recursively repeat the following steps.*

   (a) *Advance each ensemble member in time,*

$$X_i^{(t),f} = \Psi\left(X_i^{(t-1),a}\right) + V_i^{(t)},\ i = 1, \dots, N,$$

   using independently generated random variables

$$V_1^{(t)}, \dots, V_N^{(t)} \sim \mathcal{N}\left(0, \mathrm{Q}^{(t)}\right).$$

   (b) *Compute the forecast sample mean*

$$\overline{X}_N^{(t),f} = \frac{1}{N}\sum_{i=1}^{N} X_i^{(t),f},$$

   and the forecast sample covariance

$$\begin{aligned}\widehat{\mathrm{P}}_N^{(t),f} =& \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i^{(t),f} - \overline{X}_N^{(t),f}\right)\otimes\left(X_i^{(t),f} - \overline{X}_N^{(t),f}\right)\\ =& \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i^{(t),f} - \overline{X}_N^{(t),f}\right)\left(X_i^{(t),f} - \overline{X}_N^{(t),f}\right)^*.\end{aligned}$$

   (c) *Compute the sample Kalman gain*

$$\widehat{\mathrm{K}}_N^{(t)} = \widehat{\mathrm{P}}_N^{(t),f}\mathrm{H}^*\left(\mathrm{H}\widehat{\mathrm{P}}_N^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1}. \tag{5.7}$$

   (d) *Add additional perturbation to the observational vector $Y^{(t)} = y^{(t)}$,*

$$Y_i^{(t)} = y^{(t)} + W_i^{(t)},\ i = 1, \dots N, \tag{5.8}$$

   using independently generated random variables

$$W_1^{(t)}, \dots, W_N^{(t)} \sim \mathcal{N}\left(0, \mathrm{R}^{(t)}\right).$$

(e) Update each forecast ensemble member

$$X_i^{(t),a} = X_i^{(t),f} + \widehat{\mathrm{K}}_N^{(t)} \left(Y_i^{(t)} - \mathrm{H} X_i^{(t),f} \right), \ i = 1, \ldots, N. \tag{5.9}$$

We call the collection of random variables

$$X_1^{(t),f}, \ldots, X_N^{(t),f} \tag{5.10}$$

the forecast ensemble, and, similarly,

$$X_1^{(t),a}, \ldots, X_N^{(t),a} \tag{5.11}$$

the analysis ensemble. We avoid saying that random variables (5.11) are samples because it may evoke that they are independent, and that is not true as shown in the next example.

Example 21. For a given t define

$$Z_j = X_j^{(t),f} - \overline{X}_N^{(t),f}, \ j = 1, \ldots, N,$$

and for each combination of $i, j \in \{1, \ldots, N\}$ define real coefficients

$$w_{j,i} = \frac{1}{N-1} Z_j^* \mathrm{H}^* \left(\mathrm{H} \widehat{\mathrm{P}}_N^{(t),f} \mathrm{H}^* + \mathrm{R}^{(t)} \right)^{-1} \left(Y_i^{(t)} - \mathrm{H} X_i^{(t),f} \right).$$

Using this notation, the analysis update for the i^{th} ensemble member, Equation (5.9), may be written in the form

$$\begin{aligned} X_i^{(t),a} =& X_i^{(t),f} + \sum_{j=1}^{N} \left(X_j^{(t),f} - \overline{X}_N^{(t),f} \right) w_{j,i} \\ =& \overline{X}_N^{(t),f} + \sum_{j=1}^{N} \left(\left(X_j^{(t),f} - \overline{X}_N^{(t),f} \right) (\delta_{ij} + w_{j,i}) \right), \end{aligned}$$

where δ_{ij} is the Kronecker delta.

Hence, each analysis ensemble member can be written as a linear combination of all forecast ensemble members. Therefore, all analysis ensemble members are dependent. It also follows that

$$\operatorname{span}\left(\left\{ X_1^{(t),f}, \ldots, X_N^{(t),f} \right\} \right) = \operatorname{span}\left(\left\{ X_1^{(t),a}, \ldots, X_N^{(t),a} \right\} \right) \tag{5.12}$$

for all $t \in \mathbb{N}$.

Burgers et al. [1998] shows that without the data perturbation, Equation (5.8), the covariance of the ensemble would go to zero matrix as t goes to infinity. The data perturbation also guarantees that the relation between the forecast sample covariance

$$\widehat{\mathrm{P}}_N^{(t),f} = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i^{(t),f} - \overline{X}_N^{(t),f} \right) \left(X_i^{(t),f} - \overline{X}_N^{(t),f} \right)^*$$

and the analysis sample covariance

$$\widehat{\mathrm{P}}_N^{(t),a} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i^{(t),a} - \overline{X}_N^{(t),a}\right)\left(X_i^{(t),a} - \overline{X}_N^{(t),a}\right)^*$$

is analogous to the relation between the forecast and analysis covariances in the KF, Equation (5.6), i.e.,

$$\widehat{\mathrm{P}}_N^{(t),a} = \left(\mathrm{I} - \widehat{\mathrm{K}}_N^{(t)}\mathrm{H}\right)\widehat{\mathrm{P}}_N^{(t),f}. \tag{5.13}$$

One of the biggest advantages of the EnKF is that there is no need to store the full forecast sample covariance in the computer memory, when one wants to evaluate sample Kalman gain matrix $\widehat{\mathrm{K}}_N^{(t)}$, and different implementations may be found in Evensen [2009]. On the other hand, the method has two main obstacles.

1. In usual application the size of the ensemble N is much lower low than the dimension of the system. The true forecast covariance is regular, so its rank is n, but the rank of the sample covariance is $N-1$ at most. Hence, the EnKF uses a very low rank approximation of the forecast covariance, and this low rank approximation often leads to spurious correlations. Also, each analysis ensemble member lies in the subspace generated by the forecast ensemble, Equation (5.12), and the dimension of this subspace is usually incomparably smaller than the dimension of the original space.

2. The additional data perturbation, Equation (5.8), brings additional noise, and many authors consider this to be unwilling.

We discuss the first obstacle with a possible solution in Chapter 8, and the second obstacle in the last section of this chapter. The distribution and other statistical properties of the EnKF are discussed in Chapter 7.

5.4 Bayesian filtering

Undoubtedly, the forecast distribution of $X^{(t),f}$ may be understood as a prior distribution of the state $X^{(t)}$, and the analysis distribution corresponds to a posterior distribution of $X^{(t)}$. These observations immediately evoke the use of the Bayes theorem, and we discuss this approach in this section.

Similar to Section 5.2, we fully formulate the BF algorithm only in the case that the underlying dynamics is linear. The BF estimate the whole distribution of forecast and analysis states, and for each $t \in \mathbb{N}$ we denote $\phi^{(t),f}$ the density of $X^{(t),f}$ and $\phi^{(t),a}$ the density of $X^{(t),a}$.

Definition 12 (Bayesian filtering)**.** *Using the state space model and the notation from Definition 7, assume that the iterated map Ψ is linear, so there exists a matrix $\mathrm{A} \in \mathbb{R}^{n\times n}$ and a vector $b \in \mathbb{R}^n$ such that*

$$\Psi(x) = \mathrm{A}x + b.$$

The Bayesian filtering consists of the following steps.

1. *Define* $\phi^{(0),a}$ *to be a density of the* $X^{(0)}$*, i.e., a density of a* $\mathcal{N}\left(m^{(0)}, \mathrm{P}^{(0)}\right)$ *distributed random variable.*

2. *For* $t \in \mathbb{N}$ *recursively repeat the following steps.*

   (a) *Define* $\phi^{(t),f}$ *to be a density of a* $\mathcal{N}\left(m^{(t),f}, \mathrm{P}^{(t),t}\right)$ *distributed random variable with*

   $$\begin{aligned} m^{(t),f} =& \Psi\left(m^{(t-1),a}\right), \\ \mathrm{P}^{(t),f} =& \mathrm{A}\mathrm{P}^{(t-1),a}\mathrm{A}^* + \mathrm{Q}^{(t)}, \end{aligned}$$

   where $m^{(t-1),a}$ *and* $\mathrm{P}^{(t-1),a}$ *are mean and covariance of the analysis from the previous cycle.*

   (b) *Update the forecast density using the observation* $Y^{(t)} = y^{(t)}$*. For all* $x \in \mathbb{R}^n$ *define*

   $$\phi^{(t),a}(x) = \frac{1}{c\left(y^{(t)}\right)} d\left(y^{(t)} \middle| x\right) \phi^{(t),f}(x)$$

   with the data likelihood

   $$d\left(y^{(t)} \middle| x\right) \propto \exp\left(-\frac{1}{2}\left|y^{(t)} - \mathrm{H}x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}\right),$$

   and the normalization constant

   $$c\left(y^{(t)}\right) = \int_{\mathbb{R}^n} d\left(y^{(t)} \middle| x\right) \phi^{(t),f}(x)\, d\lambda^n(x).$$

Under the assumptions of the previous definition, the analysis distributions remain Gaussian and their means and covariances are

$$\begin{aligned} m^{(t),a} =& m^{(t),f} + \mathrm{P}^{(t),f}\mathrm{H}^*\left(\mathrm{H}\mathrm{P}^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1}\left(y^{(t)} - \mathrm{H}m^{(t),f}\right), \\ \mathrm{P}^{(t),a} =& \left(\mathrm{I} - \mathrm{P}^{(t),f}\mathrm{H}^*\left(\mathrm{H}\mathrm{P}^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1}\mathrm{H}\right)\mathrm{P}^{(t),f}. \end{aligned}$$

Yet, we see that when the iterated map is linear, the analysis distributions obtained by the KF and by the BF are identical.

When the iterated map Ψ is not linear, the forecast density has to be evaluated using the theorem about transformations of random variables, which requires the integration of the gradient of the iterated map Ψ, and this integration may easily become impossible. Additionally, when the iterated map is not linear, the forecast distribution is usually not Gaussian anymore, and one has to make sure that the normalization constant $c\left(y^{(t)}\right)$ is positive for all possible values of data. Otherwise, the Bayesian filtering algorithm would not be well defined.

5.5 Additional notes and references

Four assimilation methods presented in this chapter form only a small subset off all assimilation techniques. Kalnay [2003] contains an extensive list of known assimilation methods, and also describes the historical development in this area.

Lahoz et al. [2010] uses a more mathematical approach to describe data assimilation algorithms, and also discuss the relations between analysis obtained by different assimilation methods. The relations between all presented assimilation methods are also studied in Law et al. [2015], Nakamura and Potthast [2015], van Leeuwen et al. [2015], etc.

The 3DVAR, introduced in Section 5.1, was developed in the second half of the twentieth century (Courtier and Talagrand [1987], Talagrand and Courtier [1987]), and it is well known that when the prescribed background covariance corresponds to the forecast covariance and the forecast distribution is Gaussian, the analysis obtained by the 3DVAR is a maximal a posteriori estimate of the hidden state and also it is equal to the analysis obtained by the Kalman filter (Lorenc [1986], Law et al. [2015]).

As it was already mentioned, the choice of the background covariance B in the 3DVAR update equation is crucial. Since the usual dimension of a modeled state in the area of atmospheric physics is a few billion, and it is never possible to observe the whole state, well known statistical methods for an estimation of a covariance are commonly inapplicable. Hence, Parrish and Derber [1992] proposes an NMC method that estimates the background covariance from the differences of multiple forecasts for the same time using different initial conditions. The background covariance does not develop in time, i.e., it is stationary in time, and this is usually considered to be the biggest obstacle of the 3DVAR. Thus, Hamill and Snyder [2000] proposed to combined the background covariance B with an sample covariance of an ensemble, and this paper boosted a whole new group of hybrid methods which combine both variational and ensemble approaches, e.g., Desroziers et al. [2014], Liu et al. [2008], or Lorenc et al. [2014].

The Kalman filter was first described in Kalman [1960], and Kalman and Bucy [1961]. One of its first uses was a trajectory estimation for the Apollo space program. The Kalman filter is often used for a time series analysis, so its properties and derivation may also be found in books covering this topic such as Durbin and Koopman [2012]. The need for adjoint and tangent operators for covariance matrix propagations when an iterated map is not linear leads to the extended Kalman filter (Jazwinski [1970]) and unscented Kalman filter (Julier and Uhlmann [1997, 2004]).

Whitaker and Hamill [2002] show that under some conditions the perturbation of data can cause systematic errors in estimation of the analysis covariance. The square-root ensemble Kalman filter (SREnKF) (Tippett et al. [2003]) updates the ensemble mean and ensemble deviations from the mean separately without the additional data perturbation. The general idea of the SREnKF is as follows.

Using the notation from Definition 11, denote for a given t an ensemble of deviations

$$Z_i^f = X_i^{(t),f} - \overline{X}_N^{(t),f}, \quad Z_i^a = X_i^{(t),a} - \overline{X}_N^{(t),a}, \quad i = 1, \dots, N,$$

and denote Z^f and Z^a the matrices with the deviations in columns, i.e.,

$$\mathrm{Z}^f = \begin{pmatrix} Z_1^f & \cdots & Z_N^f \end{pmatrix}, \quad \mathrm{Z}^a = \begin{pmatrix} Z_1^a & \cdots & Z_N^a \end{pmatrix}.$$

Obviously,

$$\widehat{\mathrm{P}}_N^{(t),f} = \frac{1}{N-1}\mathrm{Z}^{(t),f}\left(\mathrm{Z}^{(t),f}\right)^*, \quad \widehat{\mathrm{P}}_N^{(t),a} = \frac{1}{N-1}\mathrm{Z}^{(t),a}\left(\mathrm{Z}^{(t),a}\right)^*, \tag{5.14}$$

and the EnKF covariance update, Equation (5.13), may be written in the form

$$\mathrm{P}^{(t),a} = \mathrm{P}^{(t),f} - \mathrm{P}^{(t),f}\mathrm{H}^* \left(\mathrm{H}\mathrm{P}^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1} \mathrm{H}\mathrm{P}^{(t),f}. \tag{5.15}$$

Putting (5.14) into (5.15) gives

$$\mathrm{Z}^a \left(\mathrm{Z}^a\right)^* = \mathrm{Z}^f \left(\mathrm{I} - \frac{1}{N-1} \left(\mathrm{Z}^f\right)^* \mathrm{H}^* \left(\mathrm{H}\mathrm{P}^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1} \mathrm{H}\mathrm{Z}^f\right) \left(\mathrm{Z}^f\right)^*,$$

and it immediately follows that

$$\mathrm{Z}^a \left(\mathrm{Z}^a\right)^* = \mathrm{Z}^f \mathrm{A}\mathrm{A}^* \left(\mathrm{Z}^f\right)^*$$

where A is $n \times N$ matrix such that

$$\mathrm{A}\mathrm{A}^* = \left(\mathrm{I} - \frac{1}{N-1} \left(\mathrm{Z}^f\right)^* \mathrm{H}^* \left(\mathrm{H}\mathrm{P}^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1} \mathrm{H}\mathrm{Z}^f\right). \tag{5.16}$$

The SREnKF update consists of two steps.

1. Using the unperturbed observations and the Kalman gain from the EnKF, Equation (5.7), update the ensemble mean,

$$\overline{X}_N^{(t),a} = \overline{X}_N^{(t),f} + \widehat{\mathrm{K}}_N^{(t)} \left(y^{(t)} - \mathrm{H}\overline{X}_N^{(t),f}\right).$$

2. Find matrix A as a solution of Equation (5.16), and update the differences

$$\mathrm{Z}^a = \mathrm{Z}^f \mathrm{A}.$$

The SREnKF does not, in general, produce an unbiased estimate of the true state, and necessary conditions when the SREnKF is unbiased are discussed in, for example, Livings et al. [2008], and Sakov and Oke [2008]. The convergence of the SREnKF to the Kalman filter when an iterated map of an underlying dynamical system is linear is proved in Kwiatkowski and Mandel [2015]. The class of SREnKF filters covers many filters used in real world applications such as the ensemble transform Kalman filter (Bishop et al. [2001]), the local ensemble transform Kalman filter (Hunt et al. [2007]), the ensemble adjusment Kalman filter (Anderson [2001]), and the filter proposed in Whitaker and Hamill [2002].

The particle filter (Doucet et al. [2001], Mandel and Beezley [2009]) use also an ensemble to develop the forecast distribution, but instead of updating the ensemble members directly, it updates its weight using the Bayes theorem. This filter precisely reconstructs the analysis distribution regardless of the forecast distribution, but its computation is usually unfeasible.

6. Data assimilation in infinite dimension

In the previous chapter we introduced four assimilation methods that can be used when a state space is finite dimensional, and there is a natural question whether these methods may be used when the state is infinite dimensional. In this section we look for sufficient conditions when these methods are well defined and well posed.

Hadamard [1902] stated that a mathematical problem based on a differential equation is well posed if three conditions are fulfilled:

1. it has a solution,
2. its solution is unique,
3. the solution changes smoothly when the initial condition changes.

In the chapter we are mainly interested whether assimilation methods satisfy the first two conditions.

Through the whole chapter we use the state space model from Definition 7 with assumption that both $\mathcal{H}$ and $\mathcal{G}$ are infinite dimensional separable Hilbert spaces. Hence, without loss of generality, we assume that

$$\mathcal{H} = \mathcal{G}.$$

6.1 3DVAR

The 3DVAR algorithm, introduced in Section 5.1, is based on a minimization of the cost function

$$J^{\mathrm{3DVAR}}(x) = \left|x - x^{(t),f}\right|^2_{\mathrm{B}^{-1}} + \left|y^{(t)} - \mathrm{H}x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}} \tag{6.1}$$

for a given $t \in \mathbb{N}$ with B being a known background covariance operator. When B is trace class, i.e., it is a covariance of a measurable random variable, the norm $|x|_{\mathrm{B}^{-1}}$ is defined only on a proper dense subset of the state space, so the definition of $|x|^2_{\mathrm{B}^{-1}}$ must be extended

$$|x|^2_{\mathrm{B}^{-1}} = \begin{cases} \left\langle \mathrm{B}^{-1/2}x, \mathrm{B}^{-1/2}x \right\rangle & \text{if } x \in \mathrm{B}^{1/2}(\mathcal{H}), \\ \infty & \text{if } x \notin \mathrm{B}^{1/2}(\mathcal{H}), \end{cases} \tag{6.2}$$

and similarly,

$$|x|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}} = \begin{cases} \left\langle \left(\mathrm{R}^{(t)}\right)^{-1/2} x, \left(\mathrm{R}^{(t)}\right)^{-1/2} x \right\rangle & \text{if } x \in \left(\mathrm{R}^{(t)}\right)^{1/2}(\mathcal{H}), \\ \infty & \text{if } x \notin \left(\mathrm{R}^{(t)}\right)^{1/2}(\mathcal{H}). \end{cases} \tag{6.3}$$

Hence, both functionals

$$x \in \mathcal{H} \to |x|_{\mathrm{B}^{-1}},$$
$$x \in \mathcal{H} \to |x|_{\left(\mathrm{R}^{(t)}\right)^{-1}}$$

are unbounded. Equation (6.2) and Equation (6.2) allow us to formulate the following obvious, yet very important, corollary.

Corollary 5. When a state space model is such that the state and data spaces are both infinite dimensional, then the 3DVAR cost function

$$J^{\mathrm{3DVAR}}(x): \ x \in \mathcal{H} \to [0, \infty]$$

is defined for all $x \in \mathcal{H}$ and any possible values of $X^{(t),f}$ and $Y^{(t)}$ from $\mathcal{H}$.

Even though the last corollary shows that the 3DVAR cost function is defined for all $x \in \mathcal{H}$, to state that its minimization is well posed, it is necessary that for all possible values of $X^{(t),f}$ and $Y^{(t)}$ there is at least one $x \in \mathcal{H}$ such that

$$J^{\mathrm{3DVAR}}(x) < \infty. \tag{6.4}$$

In the next example we show that, in general, the 3DVAR cost function does not satisfy this condition.

Example 22. Assume that the observation operator $\mathrm{H} = \mathrm{I}$ and that

$$\mathrm{R}^{(t)} = \mathrm{B}.$$

Recall that if B is trace class, then $\mathrm{B}^{1/2}(\mathcal{H})$ is the Cameron-Martin space of a $\mathcal{N}(0, \mathrm{B})$-distributed random variable, and this space is an intersection of all linear subspaces of the full measure (Section 3.3.2).

Suppose that $X^{(t),f} = x^{(t),f}$ and $Y^{(t)} = y^{(t)}$, i.e., $x^{(t),f}$ and $y^{(t)}$ are realizations of $X^{(t),f}$ and $Y^{(t)}$ respectively, and

$$x^{(t),f} \in \mathrm{B}^{1/2}(\mathcal{H}), \tag{6.5}$$

$$y^{(t)} \notin \mathrm{B}^{1/2}(\mathcal{H}). \tag{6.6}$$

Now, pick $x \in \mathcal{H}$ arbitrary. If $x \notin \mathrm{B}^{1/2}(\mathcal{H})$, then

$$\left(x - x^{(t),f}\right) \notin \mathrm{B}^{1/2}(\mathcal{H})$$

because $\mathrm{B}^{1/2}(\mathcal{H})$ is an linear subspace of $\mathcal{H}$ and (6.5), so

$$\left|x - x^{(t),f}\right|^2_{\mathrm{B}^{-1}} = \infty,$$

and $J^{\mathrm{3DVAR}}(x) = \infty$. Conversely, if $x \in \mathrm{B}^{1/2}(\mathcal{H})$, then

$$\left(y^{(t),f} - x\right) \notin \mathrm{B}^{1/2}(\mathcal{H})$$

using (6.6), so

$$\left|y^{(t)} - x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}} = \infty,$$

and, again, $J^{\mathrm{3DVAR}}(x) = \infty$. Since we picked x arbitrary,

$$J^{\mathrm{3DVAR}}(x) = \infty$$

for all $x \in \mathcal{H}$.

Theorem 35. *Suppose that a state space model is such that both state and observation spaces are infinite dimensional.*

1. *The minimization of the 3DVAR cost function may not have a unique solution for certain realizations of $X^{(t),f}$ and $Y^{(t)}$ if both operators B and $\mathrm{R}^{(t)}$ are trace class.*

2. *The minimization of the 3DVAR cost function has a unique solution for all possible realization of $X^{(t),f}$ and $Y^{(t)}$ if at least one operator is bounded from bellow.*

Proof. The proof of the first statement follows immediately from the previous example.

To prove the second part, assume, without loss of generality, that $\mathrm{R}^{(t)}$ is bounded from bellow. Therefore, for all $x \in \mathcal{H}$

$$\left\langle \left(\mathrm{R}^{(t)}\right)^{-1/2} x, \left(\mathrm{R}^{(t)}\right)^{-1/2} x \right\rangle < \infty,$$

so if the forecast $X^{(t),f} = x^{(t),f}$ is given, then

$$J^{\mathrm{3DVAR}}(x) < \infty$$

for all $x \in \mathcal{H}$ such that $\left(x - x^{(t),f}\right) \in \mathrm{B}^{1/2}(\mathcal{H})$. □

6.2 Ensemble Kalman filter

Recall the ensemble Kalman filter update equation

$$X_i^{(t),a} = X_i^{(t),f} + \widehat{\mathrm{K}}_N \left(Y_i^{(t)} - \mathrm{H} X_i^{(t),f}\right)$$

with

$$\begin{aligned}
\widehat{\mathrm{K}}_N =& \widehat{\mathrm{P}}_N^{(t),f} \mathrm{H}^* \left(\mathrm{H} \widehat{\mathrm{P}}_N^{(t),f} \mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1}, \\
\widehat{\mathrm{P}}_N^{(t),f} =& \frac{1}{N-1} \sum_{i=1}^{N} \left[\left(X_i^{(t),f} - \overline{X}_N^{(t),f}\right) \otimes \left(X_i^{(t),f} - \overline{X}_N^{(t),f}\right)\right], \\
\overline{X}_N^{(t),f} =& \frac{1}{N} \sum_{i=1}^{N} X_{N,i}^{(t),f}.
\end{aligned}$$

Obviously, if the dimension of the state and observations are infinite, then the ensemble Kalman filter equation make sense if and only if the Kalman gain operator $\widehat{\mathrm{K}}_N$,

$$\widehat{\mathrm{K}}_N : \mathcal{H} \to \mathcal{H},$$

is well defined for all $x \in \mathcal{H}$. We formulate a sufficient condition when $\widehat{\mathrm{K}}_N$ is defined on the whole space, and postpone further exploration of the operator to the next chapter.

Theorem 36. *Suppose that a state space model is such that both state and observation spaces are infinite dimensional.*

1. *If* $\mathrm{R}^{(t)}$ *is a trace class operator, i.e., it is a covariance of a measurable random variable, the ensemble Kalman filter equation is not defined for some realization of* $X^{(t),f}$ *and* $Y^{(t)}$.

2. *If* $\mathrm{R}^{(t)}$ *is bounded from below, the ensemble Kalman filter equation is defined for all possible realization of* $X^{(t),f}$ *and* $Y^{(t)}$.

Proof. When $\mathrm{R}^{(t)}$ is a covariance of a measurable random variable, it is trace class, Theorem 13. The sample covariance $\widehat{\mathrm{P}}_N^{(t),f}$ is a finite rank operator, and so it is $\mathrm{H}\widehat{\mathrm{P}}_N^{(t),f}\mathrm{H}^*$. Therefore, also

$$\left(\mathrm{H}\widehat{\mathrm{P}}_N^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)$$

is trace class, so its inverse is only densely defined. Hence the Kalman gain operator is also only densely defined.

If $\mathrm{R}^{(t)}$ is bounded from below, then also

$$\left(\mathrm{H}\widehat{\mathrm{P}}_N^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)$$

is bounded from bellow because $\widehat{\mathrm{P}}_N^{(t),f}$ is positive semidefinite. The second statement of the theorem now yields from the fact that inverse of an operator bounded from below is defined on the whole space. □

6.3 Bayesian filtering

We already know that the Lebesgue measure on an infinite dimensional Hilbert space $\mathcal{H}$ does not exist, so the algorithm from Definition 12 is inapplicable. However, the Bayes theorem holds even on $\mathcal{H}$, as we noticed in Section 3.5, so we may formulate an infinite dimensional version of the Bayesian filtering.

For a given $t \in \mathbb{N}$ we denote $\mu^{(t),f}$ the measure induced by the forecast $X^{(t),f}$, and $\mu^{(t),a}$ the measure induced by the analysis $X^{(t),a}$. From the definition of the state space model it follows that the data likelihood is

$$d\left(y\,|x\right) \propto \exp\left(-\frac{1}{2}\left|y-x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}\right) \tag{6.7}$$

(Stuart [2010]) with, similarly to Section 6.1, the norm defined by

$$|x|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}} = \begin{cases} \left\langle \left(\mathrm{R}^{(t)}\right)^{-1/2} x, \left(\mathrm{R}^{(t)}\right)^{-1/2} x \right\rangle & \text{if } x \in \left(\mathrm{R}^{(t)}\right)^{1/2}(\mathcal{H}), \\ \infty & \text{if } x \notin \left(\mathrm{R}^{(t)}\right)^{1/2}(\mathcal{H}), \end{cases}$$

and a natural convention that

$$\exp\left(-\infty\right) = 0.$$

Now, if the condition

$$c\left(y^{(t)}\right) = \int_{\mathcal{H}} d\left(y^{(t)}\,|x\right) d\mu^{(t),f}\left(x\right) > 0 \tag{6.8}$$

is fulfilled for a given $t \in \mathbb{N}$ and the observation $Y^{(t)} = y^{(t)}$, then from the Bayes theorem

$$\mu^{(t),a}(B) = \frac{1}{c\left(y^{(t)}\right)} \int_B d\left(y^{(t)} | x\right) d\mu^{(t),f}(x) \tag{6.9}$$

for every $B \in \mathcal{B}(\mathcal{H})$.

Remark 1. From a purely mathematical perspective, Equation 6.9 is not a proper definition, because it is not clear what the analysis distribution is if condition (6.8) is not satisfied. Stuart [2010] proposes using another data likelihood

$$\widetilde{d}\left(y^{(t)} | x\right) = \begin{cases} d\left(y^{(t)} | x\right) & \text{if } c\left(y^{(t)}\right) > 0, \\ 1 & \text{otherwise,} \end{cases}$$

which leads to the analysis distribution

$$\widetilde{\mu}^{(t),a}(B) = \begin{cases} \frac{1}{c\left(y^{(t)}\right)} \int_B d\left(y^{(t)} | x\right) d\mu^{(t),f}(x) & \text{if } c\left(y^{(t)}\right) > 0, \\ \mu^{(t),f}(B) & \text{if } c\left(y\left(t\right)\right) = 0 \end{cases} \tag{6.10}$$

for all $B \in \mathcal{B}(\mathcal{H})$. Although $\widetilde{\mu}^{(t),a}$ is mathematically well defined, it is not very useful because when condition (6.8) is not fulfilled, the filter just completely ignores the observed data $Y^{(t)} = y^{(t)}$.

The previous remark shows that, for a given $t \in \mathbb{N}$, the set

$$A^{(t)} = \left\{ y \in \mathcal{H} : \quad \int_{\mathcal{H}} d\left(y | x\right) d\mu^{(t),f}(x) = 0 \right\} \tag{6.11}$$

should be of our primary interest, and, ideally, we would like to show that this set is empty. Unfortunately, when $\mathrm{R}^{(t)}$ is trace class, i.e., the data noise $W^{(t)}$ is measurable, this set is not empty, as shown in the next example.

Example 23. Recall that we denote by $m^{(t),f}$ the mean and by $\mathrm{P}^{(t),f}$ the covariance of $X^{(t),f}$, and assume that $m^{(t),f}$ belongs to the Cameron-Martin space of $X^{(t),f}$. Additionally, assume that measures $\mu_{\mathrm{R}} \sim \mathcal{N}\left(0, \mathrm{R}^{(t)}\right)$ and $\mu^{(t),f}$ are equivalent.

The set $\mathrm{R}^{1/2}(\mathcal{H})$ is the Cameron-Martin space of μ_{R} and

$$\mu_{\mathrm{R}}\left(\mathrm{R}^{1/2}(\mathcal{H})\right) = 0,$$

so

$$\begin{aligned} \int_{\mathcal{H}} d\left(0 | x\right) d\mu_{\mathrm{R}}(x) = & \int_{\mathrm{R}^{1/2}(\mathcal{H})} \exp\left(-\frac{1}{2} |x|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}\right) d\mu_{\mathrm{R}}(x) \\ & + \int_{\mathcal{H} \setminus \mathrm{R}^{1/2}(\mathcal{H})} \exp\left(-\infty\right) d\mu_{\mathrm{R}}(x) = 0 \end{aligned}$$

Therefore, the data value $y = 0$ belongs to the set $A^{(t)}$.

The sufficient condition when the set $A^{(t)}$ is empty is similar to conditions when the previously mentioned assimilation techniques 3DVAR (Section 6.1) and EnKF (Section 6.2) are well defined.

Theorem 37. *When a state space model is such that both spaces state and observation are infinite dimensional, then the set $A^{(t)}$, defined by Equation (6.11), is empty if the operator $\mathrm{R}^{(t)}$ is bounded from below.*

Proof. The operator R is bounded from below, so the data likelihood function

$$d\left(y\,|x\right) \propto \exp\left(-\frac{1}{2}\left|y-x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}\right)$$

is positive for any $x, y \in \mathcal{H}$, and it follows that

$$\int_{\mathcal{H}} d\left(y\,|x\right) d\mu^{(t),f}\left(x\right) > 0$$

for all $y \in \mathcal{H}$. □

In the special case when both forecast and data covariances commute, we can show that this condition is also necessary. Recall that operators $\mathrm{P}^{(t),f}$ and $\mathrm{R}^{(t)}$ commute when

$$\mathrm{P}^{(t),f}\mathrm{R}^{(t)} - \mathrm{R}^{(t)}\mathrm{P}^{(t),f} = 0.$$

Lemma 38. *Assume that* $\mu^{(t),f} \sim \mathcal{N}\left(m^{(t),f}, \mathrm{P}^{(t),f}\right)$*, and operators* $\mathrm{P}^{(t),f}$ *and* $\mathrm{R}^{(t)}$ *commute. Then,*

$$\int_{\mathcal{H}} \exp\left(-\frac{1}{2}\left|y-x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}\right) d\mu^{(t),f}\left(x\right) > 0$$

for all $y \in \mathcal{H}$ *if and only if the operator* $\mathrm{R}^{(t)}$ *is bounded from below.*

Proof. Without loss of generality assume that $m^{(t),f} = 0$. The operators $\mathrm{P}^{(t),f}$ is compact, and commutes with the operator $\mathrm{R}^{(t)}$, so, using Lemma 8, they both have countable sets of eigenvalues $\left\{p_i^f\right\}$ and $\{r_i\}$ respectively, and there exist an orthonormal set $\{e_i\}_{i=1}^{\infty}$, $e_i \in \mathcal{H}$, such that

$$\mathrm{P}^{(t),f}e_i = p_i^f e_i \quad \text{and} \quad \mathrm{R}^{(t)}e_i = r_i e_i.$$

Recall that, through the whole thesis, we assume that the kernels of $\mathrm{P}^{(t),f}$ and $\mathrm{R}^{(t)}$ contain only the zero element, so all eigenvalues $\left\{p_i^f\right\}$ and $\{r_i\}$ are strictly positive.

For any $z \in \mathcal{H}$ we denote $\{z_i\}$ its coefficient with respect to the set $\{e_i\}$,

$$z_i = \left\langle z, e_i \right\rangle, \ i \in \mathbb{N}.$$

Using this notation,

$$\begin{aligned} d\left(y|x\right) &= \exp\left(-\frac{1}{2}\left|y-x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}\right) \\ &= \exp\left(-\sum_{i=1}^{\infty}\frac{\left(y_i-x_i\right)^2}{2r_i}\right) = \prod_{i=1}^{\infty}\exp\left(-\frac{\left(y_i-x_i\right)^2}{2r_i}\right), \end{aligned}$$

and

$$\int_{\mathcal{H}} d\left(y|x\right) d\mu^{(t),f}\left(x\right) = \int_{\mathcal{H}} \lim_{n\to\infty}\left(\prod_{i=1}^{n}\exp\left(-\frac{\left(y_i-x_i\right)^2}{2r_i}\right)\right) d\mu^{(t),f}\left(x\right). \tag{6.12}$$

The function

$$z \in \mathbb{R} \to \exp\left(-z^2\right)$$

is bounded since

$$0 \leq \exp\left(-z^2\right) \leq 1$$

for all $z \in \mathbb{R}$, so

$$f_n\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} \exp\left(-\frac{\left(y_i - x_i\right)^2}{2r_i}\right), \ n \in \mathbb{N},$$

is a monotone sequence of measurable functions, and we can use the monotone convergence theorem to swap the limit and the integral sign in Equation (6.12),

$$\int_{\mathcal{H}} d\left(y|x\right) d\mu^{(t),f}\left(x\right) = \lim_{n\to\infty} \left(\int_{\mathcal{H}} \prod_{i=1}^{n} \exp\left(-\frac{\left(y_i - x_i\right)^2}{2r_i}\right) d\mu^{(t),f}\left(x\right) \right). \tag{6.13}$$

Using the properties of the Gaussian distribution,

$$\left\langle X^f, e_i \right\rangle \sim \mathcal{N}\left(0, p_i^f\right)$$

for each $i \in \mathbb{N}$, and

$$\mathrm{E}\left(\left\langle X^f, e_i \right\rangle \left\langle X^f, e_j \right\rangle\right) = \delta_{ij}, \ i, j \in \mathbb{N},$$

i.e., random variables $\left\langle X^f, e_i \right\rangle$ and $\left\langle X^f, e_j \right\rangle$ are independent unless $i = j$. Therefore, if we denote by μ_i^f the measure induced by $\left\langle X^f, e_i \right\rangle$, then

$$\int_{\mathcal{H}} f_n\left(x_1, \ldots, x_n\right) d\mu^{(t),f}\left(x\right) = \int_{\mathcal{H}} f_n\left(x_1, \ldots, x_n\right) d\mu_1^f\left(x_1\right) \times \cdots \times d\mu_n^f\left(x_n\right)$$

for all $n \in \mathbb{N}$, and using Fubini's theorem we obtain

$$\int_{\mathcal{H}} f_n\left(x_1, \ldots, x_n\right) d\mu^{(t),f}\left(x\right) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_n\left(x_1, \ldots, x_n\right) d\mu_1^f\left(x_1\right) \cdots d\mu_n^f\left(x_n\right).$$

Putting the last equation into (6.13) yields

$$\int_{\mathcal{H}} d\left(y|x\right) d\mu^f\left(x\right) = \lim_{n\to\infty} \prod_{i=1}^{n} \int_{\mathbb{R}} \exp\left(-\frac{\left(y_i - x_i\right)^2}{2r_i}\right) d\mu_i^f\left(x_i\right),$$

and, using the fact that μ_i^f is absolutely continuous with respect to the Lebesgue measure,

$$\int_{\mathcal{H}} d\left(y|x\right) d\mu^f\left(x\right) = \prod_{i=1}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{\left(y_i - x_i\right)^2}{2r_i}\right) \psi\left(x_i\right) d\lambda^1\left(x_i\right) \tag{6.14}$$

where

$$\psi_i\left(x\right) = \frac{1}{\sqrt{2\pi p_i^f}} \exp\left(-\frac{x_i^2}{2p_i^f}\right),$$

i.e., ψ_i is the density of a $\mathcal{N}\left(0, p_i^f\right)$-distributed random variable.

Now, we use identities

$$\begin{aligned}\frac{(y_i - x_i)^2}{r_i} + \frac{x_i^2}{p_i^f} &= \left(\frac{1}{p_i^f} + \frac{1}{r_i}\right) x_i^2 - 2\frac{x_i y_i}{r_i} + \frac{y_i^2}{r_i} \\ &= \frac{(x_i - m_i^a)^2}{p_i^a} + \frac{y_i^2}{r_i + p_i^f},\end{aligned}$$

with

$$m_i^a = \frac{p_i^f}{r_i + p_i^f} y_i \quad \text{and} \quad p_i^a = \left(\frac{1}{p_i^f} + \frac{1}{r_i}\right)^{-1}$$

to write Equation (6.14) in the form

$$\int_{\mathcal{H}} d\left(y|x\right) d\mu^f\left(x\right) = \prod_{i=1}^{\infty} \frac{1}{\sqrt{2\pi p_i^f}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x_i - m_i^a)^2}{2p_i^a} - \frac{y_i^2}{2\left(r_i + p_i^f\right)}\right) d\lambda\left(x_i\right),$$

and because

$$\int_{-\infty}^{\infty} \exp\left(-\frac{(x_i - m_i^a)^2}{2p_i^a}\right) dx_i = \sqrt{2\pi p_i^a}$$

for each $i \in \mathbb{N}$, it follows that

$$\begin{aligned}\int_{\mathcal{H}} d\left(y|x\right) d\mu^f\left(x\right) &= \prod_{i=1}^{\infty} \left(\frac{p_i^a}{p_i^f}\right)^{1/2} \exp\left(-\frac{y_i^2}{2\left(r_i + p_i^f\right)}\right) \\ &= \prod_{i=1}^{\infty} \left(1 + \frac{p_i^f}{r_i}\right)^{-1/2} \exp\left(-\frac{y_i^2}{2\left(r_i + p_i^f\right)}\right), \qquad (6.15)\end{aligned}$$

where we used that

$$\frac{p_i^a}{p_i^f} = \frac{1}{p_i^f\left(\frac{1}{p_i^f} + \frac{1}{r_i}\right)} = \frac{1}{1 + \frac{p_i^f}{r_i}}$$

in the second step. The infinite product on the right side of (6.15) is nonzero if and only if the sum

$$\sum_{i=1}^{\infty} \log\left(\left(1 + \frac{p_i^f}{r_i}\right)^{-1/2} \exp\left(-\frac{y_i^2}{2\left(r_i + p_i^f\right)}\right)\right)$$

converges, and this sum can be written in the form

$$-\frac{1}{2}\left(\sum_{i=1}^{\infty} \log\left(1 + \frac{p_i^f}{r_i}\right)\right) - \left(\sum_{i=1}^{\infty} \frac{y_i^2}{r_i + p_i^f}\right). \qquad (6.16)$$

To finish the proof we only need to show that (6.16) is finite if and only if

$$r = \inf_{i \in \mathbb{N}} \{r_i\} > 0, \qquad (6.17)$$

and we examine the convergence of each summand in (6.16) individually.

First, the equivalence

$$\sum_{i=1}^{\infty} \ln\left(1+\frac{p_i^f}{r_i}\right) < \infty \quad \Leftrightarrow \quad \sum_{i=1}^{\infty} \frac{p_i^f}{r_i} < \infty \tag{6.18}$$

follows from the limit comparison test since

$$\lim_{i\to\infty} \frac{\ln\left(1+\frac{p_i^f}{r_i}\right)}{\frac{p_i^f}{r_i}} = 1$$

when

$$\lim_{i\to\infty} \frac{p_i^f}{r_i} = 0. \tag{6.19}$$

If condition (6.19) is not satisfied, than both sums in (6.18) obviously diverge. Conversely, if $r > 0$, then

$$\sum_{i=1}^{\infty} \frac{p_i^f}{r_i} \leq \sum_{i=1}^{\infty} \frac{p_i^f}{r} \leq r^{-1} \sum_{i=1}^{\infty} p_i^f < \infty$$

because $\mathrm{P}^{(t),f}$ is trace class.

Further, if $r > 0$, then

$$\sum_{i=1}^{\infty} \frac{y_i^2}{r_i + p_i^f} \leq \sum_{i=1}^{\infty} \frac{y_i^2}{r} \leq r^{-1} \sum_{i=1}^{\infty} y_i^2 = r^{-1} |y|^2 < \infty$$

since $\{y_i\}$ are Fourier coefficients of y. Conversely, if $r = 0$, we will construct $\widetilde{y} \in \mathcal{H}$ such that $|\widetilde{y}| \leq 1$ and

$$\sum_{i=1}^{\infty} \frac{\widetilde{y}_i^2}{r_i + p_i^f} = \infty.$$

Since $r = 0$, there exists a subsequence $\{r_{i_k}\}_{k=1}^{\infty}$ such that

$$r_{i_k} \leq \frac{1}{2^k}, \quad k \in \mathbb{N},$$

and we define

$$\widetilde{y} = \sum_{i=1}^{\infty} \widetilde{y}_i e_i$$

with

$$\widetilde{y}_i = \begin{cases} r_i^{1/2} & \text{if } i \in \{i_k\}_{k\in\mathbb{N}}\,, \\ 0 & \text{if } i \notin \{i_k\}_{k\in\mathbb{N}}. \end{cases}$$

The element $\widetilde{y}$ lies in the unit circle because

$$|\widetilde{y}|^2 = \sum_{i=1}^{\infty} \widetilde{y}_i^2 = \sum_{k=1}^{\infty} r_{i_k} \leq \sum_{k=1}^{\infty} \frac{1}{2^k} = 1,$$

while

$$\sum_{i=1}^{\infty} \frac{\widetilde{y}_i^2}{r_i + p_i^f} = \sum_{k=1}^{\infty} \frac{r_{i_k}}{r_{i_k} + p_{i_k}^f} = \sum_{k=1}^{\infty} \frac{1}{1 + \frac{p_{i_k}^f}{r_{i_k}}} = \infty$$

where the last equality follows immediately from condition (6.19).

Therefore, the sum (6.16) is finite for all $y \in \mathcal{H}$ if and only if $r > 0$. □

The construction of the element $\widetilde{y}$ at the end of the previous proof may be generalized, and it implies the following interesting result.

Lemma 39. *Under the same assumptions as in Lemma 38, if*

$$r = \inf_{i \in \mathbb{N}} \{r_i\} = 0,$$

then the set

$$A^{(t)} = \left\{ y \in \mathcal{H} : \int_{\mathcal{H}} \exp\left(-\frac{1}{2} \left| y - x \right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}} \right) d\mu^{(t),f}(x) = 0 \right\}$$

is dense in $\mathcal{H}$.

Proof. To show that A is dense it is sufficient to show that for each $z \in \mathcal{H}$ and any $\delta > 0$

$$A^{(t)} \cap \{u \in \mathcal{H} : |z - u| < \delta\} \neq \emptyset.$$

Let $z \in \mathcal{H}$ and $\delta > 0$. Similarly as in the previous proof, denote by $\{e_i\}$ the common eigenvector basis of operators $\mathrm{R}^{(t)}$ and $\mathrm{P}^{(t),f}$. Because $r = 0$, there exists a subsequence $\{r_{i_k}\}_{k=1}^{\infty}$ such that

$$r_{i_k} \leq \frac{\delta^2}{2^k}$$

for all $k \in \mathbb{N}$. Define $\widetilde{z} = \sum_{i=1}^{\infty} \widetilde{z}_i e_i$ where

$$\widetilde{z}_i = \begin{cases} \langle z, e_i \rangle + r_i^{1/2} & \text{if } i \in \{i_k\}_{k \in \mathbb{N}}, \\ \langle z, e_i \rangle & \text{if } i \notin \{i_k\}_{k \in \mathbb{N},} \end{cases}$$

and obviously

$$|z - \widetilde{z}| = \left(\sum_{i=1}^{\infty} |\langle z - \widetilde{z}, e_i \rangle|^2 \right)^{1/2} = \left(\sum_{k=1}^{\infty} r_{i_k} \right)^{1/2} \leq \delta.$$

From proof of the previous Lemma we know that

$$\int_{\mathcal{H}} \exp\left(-\frac{1}{2} \left| \widetilde{z} - x \right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}} \right) d\mu^f(x) > 0$$

if and only if

$$\left(\sum_{i=1}^{\infty} \ln\left(1 + \frac{p_i^f}{r_i} \right) \right) + \left(\sum_{i=1}^{\infty} \frac{\widetilde{z}_i^2}{r_i + p_i^f} \right) < \infty, \tag{6.20}$$

but

$$\sum_{i=1}^{\infty} \frac{\widetilde{z}_i^2}{r_i + p_i^f} \geq \sum_{k=1}^{\infty} \frac{\widetilde{z}_{i_k}^2}{r_{i_k} + p_{i_k}^f} \geq \sum_{k=1}^{\infty} \frac{r_{i_k}}{r_{i_k}\left(1 + \frac{p_{i_k}^f}{r_{i_k}}\right)} = \infty$$

when

$$\lim_{i \to \infty} \frac{p_i^f}{r_i} = 0,$$

and, again from the previous proof, if this condition is not satisfied, then

$$\sum_{i=1}^{\infty} \ln\left(1 + \frac{p_i^f}{r_i}\right) = \infty.$$

Therefore, the sum at the left side of Equation (6.20) diverges, and $\widetilde{z} \in A^{(t)}$. $\square$

The last two lemmas may be summarized in the following theorem.

Theorem 40. *Assume that* $X^{(t),f}$ *has the* $\mathcal{N}\left(m^{(t),f}, \mathrm{P}^{(t),f}\right)$ *distribution and operators* $\mathrm{HP}^{(t),f}\mathrm{H}^*$ *and* $\mathrm{R}^{(t)}$ *commute. Then, the set*

$$A^{(t)} = \left\{ y \in \mathcal{H} : \int_{\mathcal{H}} \exp\left(-\frac{1}{2}\left|y - \mathrm{H}x\right|^2_{\left(\mathrm{R}^{(t)}\right)^{-1}}\right) d\mu^f(x) = 0 \right\}$$

is empty if and only if $\mathrm{R}^{(t)}$ *is bounded away from zero.*

Additionally, if the eigenvalues of $\mathrm{R}^{(t)}$ *converge to zero, then* $\mathrm{A}^{(t)}$ *is dense in* $\mathcal{H}$.

6.4 Summary

We have reviewed the definitions of three assimilation methods in a situation when both state and observation space are infinite dimensional. All three methods are either not well posed or even not well defined if a data noise is a measurable random variable. On the other hand, all methods are well posed when the data noise covariance operator is bounded from below.

At first sight, these observations may look surprising, but, from another point of view, it is quite a natural consequence of a "size" of an infinite dimensional space. When one defines an assimilation method, a natural requirement should be that an analysis distribution is absolutely continuous with respect to a forecast distribution. However, Example 3 shows that even in the simplest case when X and Y are identically Gaussian distributed random variables on $\mathcal{H}$, the measures induced by X and by $X + Y$ are singular. Therefore any assimilation method that is based on a linear combination of forecast states and observations must be expected to fail when observations are measurable random variables.

A typical example of a weak random variable is the white noise, i.e., weak random variable with the $\mathcal{N}_w(0, \mathrm{I})$ distribution. Some people may argue that expecting data to contain continuous white noise is unrealistic. However, when one thinks about a high resolution picture with every pixel created by a separate sensor, then the noise contained in this picture is uncorrelated, and as the resolution of the picture increases, the noise converges to the continuous white noise. Hence, at least in some applications, the assumptions that data noise is only weakly measurable is reasonable.

6.5 Additional notes and references

The problem with the posedness of assimilation algorithms on an infinite dimensional space is not new, and many authors have recently studied the coincidental inverse problem on such spaces.

Lasanen [2007] observes that the Bayesian filtering is well defined only for some values of observations when the observation space is infinite dimensional, and formulates a necessary and sufficient condition on the observation so that the posterior distribution is well defined. He also observes that the Bayes solution minimize the 3DVAR cost function if the forecast distribution is Gaussian.

Cotter et al. [2009] describes mathematical framework for an inverse problem on functional spaces, and formulates the Bayes theorem on an infinite dimensional space. The paper also contains sufficient conditions for a data likelihood function so that the posterior distribution is well defined, and a proof of existence of an MAP estimate. Additionally, Dashti et al. [2013] shows consistency of this estimate, and Dashti et al. [2012] studies possibilities of using more general prior distribution.

Finally, Stuart [2010, 2013] provides an overview of the Bayesian approach to an inverse problem on a separable Hilbert space.

7. Convergence of ensemble Kalman filter in Hilbert space

In this chapter, we study statistical properties of the EnKF in the large ensemble limit. The chapter extends the work done in Le Gland et al. [2011] and Mandel et al. [2011], where the convergence of the EnKF is proved if the state space model is finite dimensional.

The chapter is organized as follows. Section 7.1 introduces notations and assumptions used through out the whole chapter. Section 7.2 shows the continuity of the Kalman gain operator. Section 7.3 shows some statistical properties on the ensemble. Section 7.4 contains auxiliary estimates and lemmas, and Section 7.5 contains the proof of convergence.

7.1 Assumptions and definitions

As usual, we assume that $\mathcal{H}$ is a separable Hilbert space, and $\left\{X^{(t)}\right\}$ is a system with a stochastic dynamics defined on $\mathcal{H}$, i.e.,

$$\begin{aligned} X^{(0)} \sim & \mathcal{N}\left(m^{(t)}, \mathrm{P}^{(0)}\right), \\ X^{(t)} = & \Psi\left(X^{(t-1)}\right) + V^{(t)}, \ t \in \mathbb{N}, \end{aligned}$$

with the measurable iterated map

$$\Psi : \mathcal{H} \to \mathcal{H}$$

and $V^{(t)} \sim \mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$.

Assumption 1. *Through the whole chapter, we assume that the following statements hold.*

1. *An observation operator* $\mathrm{H} \in [\mathcal{H}]$.
2. *The iterated map* $\Psi : \mathcal{H} \to \mathcal{H}$ *is locally Lipschitz continuous with at most polynomial growth in infinity, i.e., there exist positive constants* s, l *such that*
$$|\Psi(x) - \Psi(y)| \leq l\,|x-y|\left(1 + |x|^s + |y|^s\right)$$
for all $x, y \in \mathcal{H}$.
3. *Observations of the system* $y^{(t)}$, $t \in \mathbb{N}$, *are deterministic.*
4. *Random variables* $X_i^{(0),(a)} \sim \mathcal{N}\left(m^{(0)}, \mathrm{P}^{(0)}\right)$, $i \in \mathbb{N}$, *are i.i.d., i.e., they are independently generated from the distribution of the initial condition of the underlying dynamical system.*
5. *Random variables* $V_i^{(t)}$, $i \in \mathbb{N}$, $t \in \mathbb{N}$, *are i.i.d. samples from* $\mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$ *distribution.*
6. *Weak random variables* $W_i^{(t)} \sim \mathcal{N}_w\left(0, \mathrm{R}^{(t)}\right)$, $i \in \mathbb{N}$, $t \in \mathbb{N}$, *are independent, and all operators* $\mathrm{R}^{(t)}$, $t \in \mathbb{N}$, *are bounded from below.*

Additionally for each $t \in \mathbb{N}$ we define the Kalman gain operator

$$\mathcal{K}^{(t)} : \mathrm{Q} \to \mathrm{QH}^* \left(\mathrm{HQH}^* + \mathrm{R}^{(t)}\right)^{-1} \tag{7.1}$$

for any $\mathrm{Q} \in [\mathcal{H}]$.

Using the previous list of assumptions and the introduced notation, we can summarize the EnKF and mean field EnKF algorithms. The latter one evolves the ensemble in time according to the original Kalman filter equations, and is a candidate to represent the limit distribution of the EnKF.

Definition 13 (Ensemble Kalman filter)**.** *Given $N \in \mathbb{N}$, $N \geq 2$, the EnKF algorithm consists of these consecutive steps.*

1. *Initialize the first guess:*

$$\forall i = 1, \ldots, N : \quad X_{N,i}^{(0),a} = X_i^{(0),a}. \tag{7.2}$$

2. *For $t = 1, 2, \ldots$ repeat the following steps.*

   (a) *Advance the analysis ensemble from the previous cycle:*

$$\forall i = 1, \ldots, N : \quad X_{N,i}^{(t),f} = \Psi\left(X_{N,i}^{(t-1),a}\right) + V_i^{(t)}. \tag{7.3}$$

   (b) *Evaluate the sample mean and the sample covariance:*

$$\begin{aligned} \overline{X}_N^{(t),f} =& \frac{1}{N} \sum_{i=1}^{N} X_{N,i}^{(t),f}, \\ \widehat{\mathrm{P}}_N^{(t),f} =& \frac{1}{N-1} \sum_{i=1}^{N} \left[\left(X_{N,i}^{(t),f} - \overline{X}_N^{(t),f}\right) \otimes \left(X_{N,i}^{(t),f} - \overline{X}_N^{(t),f}\right)\right]. \end{aligned}$$

   (c) *Evaluate the sample Kalman gain operator:*

$$\widehat{\mathrm{K}}_N^{(t),X} = \mathcal{K}^{(t)}\left(\widehat{\mathrm{P}}_N^{(t),f}\right).$$

   (d) *Update the forecast ensemble:*

$$\forall i = 1, \ldots, N : \quad X_{N,i}^{(t),a} = X_{N,i}^{(t),f} + \widehat{\mathrm{K}}_N^{(t),X}\left(y^{(t)} - \mathrm{H}X_{N,i}^{(t),f} - W_i^{(t)}\right). \tag{7.4}$$

Definition 14 (Mean field EnKF)**.** *Given $N \in \mathbb{N}$, $N \geq 2$, the mean field EnKF algorithm consists of these consecutive steps.*

1. *Initialize the first guess:*

$$\forall i = 1, \ldots, N : \quad U_{N,i}^{(0),a} = X_i^{(0),a}. \tag{7.5}$$

2. *For $t = 1, 2, \ldots$ recursively repeat the following steps.*

(a) Advance the analysis ensemble from the previous cycle:

$$\forall i = 1, \ldots, N: \quad U_{N,i}^{(t),f} = \Psi^{(t)}\left(U_{N,i}^{(t-1),a}\right) + V_i^{(t)}. \tag{7.6}$$

(b) Evaluate the forecast covariance:

$$\mathrm{C}^{(t),f} = \mathrm{E}\left(\left(U_{N,1}^{(t),f} - \mathrm{E}U_{N,1}^{(t),f}\right) \otimes \left(U_{N,1}^{(t),f} - \mathrm{E}U_{N,1}^{(t),f}\right)\right).$$

(c) Evaluate the Kalman gain operator:

$$\mathrm{K}^{(t),U} = \mathcal{K}^{(t)}\left(\mathrm{C}^{(t),f}\right).$$

(d) Update the forecast ensemble:

$$\forall i = 1, \ldots, N: \quad U_{N,i}^{(t),a} = U_{N,i}^{(t),f} + \mathrm{K}^{(t),U}\left(y^{(t)} - \mathrm{H}U_{N,i}^{(t),f} - W_i^{(t)}\right). \tag{7.7}$$

For future reference we define the sample mean

$$\overline{U}_N^{(t),f} = \frac{1}{N}\sum_{i=1}^{N} U_i^{(t),f},$$

and the sample covariance

$$\widehat{\mathrm{C}}_N^{(t),f} = \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(U_i^{(t),f} - \overline{U}_N^{(t),f}\right) \otimes \left(U_i^{(t),f} - \overline{U}_N^{(t),f}\right)\right],$$

of the mean field ensemble, and, to simplify the proofs, we define forecast and analysis ensembles of differences

$$\begin{aligned} Z_{N,i}^{(t),f} &= X_{N,i}^{(t),f} - U_i^{(t),f}, \\ Z_{N,i}^{(t),a} &= X_{N,i}^{(t),a} - U_i^{(t),a} \end{aligned}$$

for any possible values of t, N and i. We also recall that we have defined the empirical moments in Section 3.1.4, and for all six ensembles and for any $p \geq 1$ we define the real valued random variables

$$\widehat{X}_{N,p}^{(t),f} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|X_{N,i}^{(t),f}\right|^p\right)^{1/p}, \qquad \widehat{X}_{N,p}^{(t),a} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|X_{N,i}^{(t),a}\right|^p\right)^{1/p},$$

$$\widehat{U}_{N,p}^{(t),f} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|U_{N,i}^{(t),f}\right|^p\right)^{1/p}, \qquad \widehat{U}_{N,p}^{(t),a} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|U_{N,i}^{(t),a}\right|^p\right)^{1/p},$$

$$\widehat{Z}_{N,p}^{(t),f} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|Z_{N,i}^{(t),f}\right|^p\right)^{1/p}, \qquad \widehat{Z}_{N,p}^{(t),a} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|Z_{N,i}^{(t),a}\right|^p\right)^{1/p}.$$

Finally, in many cases, the same statements hold for both forecast and analysis ensemble, and we use upper index $\bullet$ instead of f or a if we do not distinguish between forecast and analysis ensemble. Hence, by writing $X_{N,i}^{(t),\bullet}$ for any $i = 1, \ldots, N$ we mean both $X_{N,i}^{(t),f}$ and $X_{N,i}^{(t),a}$, and we use the same convention with $U_{N,i}^{(t),\bullet}$ and $Z_{N,i}^{(t),\bullet}$.

7.2 Continuity of Kalman gain operator

This section contains an important proof of the fact that the Kalman gain operator is continuous and even locally Lipschitz continuous, which was originally published as Lemma 4.1 in Kwiatkowski and Mandel [2015]. We include the proof from the cited article for completeness.

Theorem 41 ([Kwiatkowski and Mandel, 2015, Lemma 4.1])**.** *Assume that operators* $\mathrm{P}, \mathrm{Q} \in [\mathcal{H}]$ *are semidefinite and selfadjoint, and* $\mathcal{K}^{(t)}$ *is, for a given* $t \in \mathbb{N}$*, the Kalman gain operator defined by Equation* (7.1)*. Then*

$$\left|\mathcal{K}^{(t)}(\mathrm{P}) - \mathcal{K}^{(t)}(\mathrm{Q})\right| \leq c\,|\mathrm{P}-\mathrm{Q}|\,(1+\min\{|\mathrm{P}|,|\mathrm{Q}|\})$$

where real positive constant c *depends on operators* $\mathrm{R}^{(t)}$ *and* H *only.*

Proof. We denote $\mathrm{R} = \mathrm{R}^{(t)}$ and $\mathcal{K} = \mathcal{K}^{(t)}$, i.e., we drop the time index. Using Lemma 7

$$\begin{aligned}\left|(\mathrm{HPH}^* + \mathrm{R})^{-1} - (\mathrm{HQH}^* + \mathrm{R})^{-1}\right| &\leq |\mathrm{HQH}^* - \mathrm{HPH}^*|\left|\mathrm{R}^{-1}\right|^2 \\ &\leq \left|\mathrm{R}^{-1}\right|^2 |\mathrm{H}|^2 |\mathrm{Q}-\mathrm{P}|,\end{aligned}$$

and by the triangle inequality

$$\begin{aligned}|\mathcal{K}(\mathrm{P}) - \mathcal{K}(\mathrm{Q})| &= \left|\mathrm{PH}^*(\mathrm{HPH}^*+\mathrm{R})^{-1} - \mathrm{QH}^*(\mathrm{HPH}^*+\mathrm{R})^{-1}\right| \\ &\leq \left|\mathrm{QH}^*(\mathrm{HQH}^*+\mathrm{R})^{-1} - \mathrm{PH}^*(\mathrm{HQH}^*+\mathrm{R})^{-1}\right| \\ &\quad + \left|\mathrm{PH}^*(\mathrm{HQH}^*+\mathrm{R})^{-1} - \mathrm{PH}^*(\mathrm{HPH}^*+\mathrm{R})^{-1}\right| \\ &\leq |\mathrm{Q}-\mathrm{P}|\,|\mathrm{H}|\left|\mathrm{R}^{-1}\right| + |\mathrm{P}|\left|\mathrm{R}^{-1}\right|^2|\mathrm{H}|^3|\mathrm{Q}-\mathrm{P}|. \\ &\leq |\mathrm{Q}-\mathrm{P}|\left(|\mathrm{H}|\left|\mathrm{R}^{-1}\right| + |\mathrm{P}|\left|\mathrm{R}^{-1}\right|^2|\mathrm{H}|^3\right)\end{aligned}$$

Swapping the roles of P and Q yields

$$|\mathcal{K}(\mathrm{P}) - \mathcal{K}(\mathrm{Q})| \leq |\mathrm{Q}-\mathrm{P}|\left(|\mathrm{H}|\left|\mathrm{R}^{-1}\right| + |\mathrm{Q}|\left|\mathrm{R}^{-1}\right|^2|\mathrm{H}|^3\right), \tag{7.8}$$

and hence

$$|\mathcal{K}(\mathrm{P}) - \mathcal{K}(\mathrm{Q})| \leq c\,|\mathrm{P}-\mathrm{Q}|\,(1+\min\{|\mathrm{P}|,|\mathrm{Q}|\})$$

where $c = \max\left\{|\mathrm{H}|\,|\mathrm{R}^{-1}|, |\mathrm{H}|^3\,|\mathrm{R}^{-1}|^2\right\}$. □

The previous theorem has an important corollary.

Corollary 6. Under the same assumption as Theorem 41

$$\left|\mathcal{K}^{(t)}(\mathrm{P})\right| \leq |\mathrm{H}|\left|\left(\mathrm{R}^{(t)}\right)^{-1}\right||\mathrm{P}|. \tag{7.9}$$

Proof. Take $\mathrm{Q} = 0$, which is obviously positive semidefinite and selfadjoint. Using Equation (7.8) from the previous proof,

$$\left|\mathcal{K}^{(t)}(\mathrm{P})\right| \leq |\mathrm{P}|\left(|\mathrm{H}|\left|\left(\mathrm{R}^{(t)}\right)^{-1}\right| + |\mathrm{Q}|\left|\left(\mathrm{R}^{(t)}\right)^{-1}\right|^2|\mathrm{H}|^3\right),$$

and (7.9) yields from the fact that $|\mathrm{Q}| = 0$. □

7.3 Ensemble properties

We have already shown in Example 21 that ensemble members

$$X_{N,1}^{(t),\bullet}, \ldots, X_{N,N}^{(t),\bullet}$$

are not independent, but the following theorem shows that they are exchangeable, and also shows statistical properties of other ensembles used in this chapter.

Theorem 42. *Given positive integers t, N, M, with $M, N \geq 2$, the following statements hold.*

1. *Random variables $U_{N,1}^{(t),\bullet}, \ldots, U_{N,N}^{(t),\bullet}$ are all identically distributed, and are independent given the observations.*
2. *The distribution of $U_{N,1}^{(t),\bullet}$ does not depend on the size N of the ensemble, i.e.,*
$$U_{N,1}^{(t),\bullet} = U_{M,1}^{(t),\bullet}$$
for any combination of M and N.
3. *Random variables $X_{N,1}^{(t),\bullet}, \ldots, X_{N,N}^{(t),\bullet}$ are exchangeable.*
4. *The distribution of $X_{N,1}^{(t),\bullet}$ depends on N, i.e.,*
$$X_{N,1}^{(t),\bullet} \neq X_{M,1}^{(t),\bullet}$$
unless $N = M$.
5. *Random variables $Z_{N,1}^{(t),\bullet}, \ldots, Z_{N,N}^{(t),\bullet}$ are exchangeable.*
6. *The distribution of $Z_{N,1}^{(t),\bullet}$ depends on N.*

Proof. The first two statements are trivial, and follow immediately using Definition 14.

The third statement can be proved using induction. For any $N \geq 2$ the initial ensemble

$$X_{N,1}^{(0),a}, \ldots, X_{N,N}^{(0),a}$$

consists of independent random variables by assumption, and hence the members are exchangeable. Now, if random variables

$$X_{N,1}^{(t-1),a}, \ldots, X_{N,N}^{(t-1),a}$$

are exchangeable, then the members of the forecast ensemble

$$X_{N,1}^{(t),f}, \ldots, X_{N,N}^{(t),f}$$

are also exchangeable, since they are the sum of exchangeable random variables

$$\Psi\left(X_{N,1}^{(t-1),a}\right), \ldots, \Psi\left(X_{N,N}^{(t-1),a}\right)$$

and independent, and hence exchangeable, random variables $V_i^{(t)}$, $i = 1, \ldots, N$. Conversely, if ensemble

$$X_{N,1}^{(t),f}, \ldots, X_{N,N}^{(t),f}$$

consists of exchangeable random variables, then random variables

$$\widehat{\mathrm{K}}_N^{(t),X}\left(y^{(t)} - \mathrm{H}X_{N,i}^{(t),f} - W_i^{(t)}\right),\ i = 1, \ldots, N,$$

are exchangeable as well, and it follows that

$$X_{N,1}^{(t),a}, \ldots, X_{N,N}^{(t),a}$$

are exchangeable.

If N and M are not equal, then the distributions of $\widehat{\mathrm{P}}_N^{(t),f}$ and $\widehat{\mathrm{P}}_N^{(t),f}$ are different, which prove the fourth statement.

The fifth and sixth statements are immediate consequences of statements number three and four. □

The second statement of the last lemma shows that using a lower index N in $U_{N,1}^{(t),\bullet}$ is unnecessary, so we write $U_1^{(t),\bullet}$ instead of $U_{N,1}^{(t),\bullet}$ from now on.

Theorem 43. *For every $t \in \mathbb{N}$ and any $p \geq 1$ are $U_1^{(t),f}$ and $U_1^{(t),a}$ elements of $\mathcal{L}^p(\mathcal{H})$, i.e., they have finite moments of order p.*

Proof. We use induction to prove the statement of the theorem.

Firstly, random variable $U_1^{(0),a} \in \mathcal{L}^p(\mathcal{H})$ for all $p \geq 1$ by definition.

Secondly, assume that

$$U_1^{(t-1),a} \in \mathcal{L}^p(\mathcal{H}) \quad \forall p \geq 1. \tag{7.10}$$

Function Ψ is locally Lipschitz continuous with at most polynomial growth in infinity, so, using the triangle inequality and Lemma 3, there exist $s, l > 0$ such that

$$\begin{aligned}\left|U_1^{(t),f}\right| \leq& \left|\Psi\left(U_1^{(t-1),a}\right)\right| + \left|V_1^{(t)}\right| \\ \leq& l\left(1 + \left|U_1^{(t-1),a}\right|^{s+1}\right) + \left|V_1^{(t)}\right|.\end{aligned}$$

Because $V_1^{(t)}$ is a Gaussian random variable, it has finite moments of all orders, and it follows that

$$\left\|U_1^{(t),f}\right\|_p \leq l\left(1 + \left\|U_1^{(t-1),a}\right\|_{p(s+1)}^{s+1}\right) + \left\|V_1^{(t)}\right\|_p < \infty$$

where the last inequality follows from Lemma 11 and assumption (7.10).

Thirdly, assume that

$$U_1^{(t),f} \in \mathcal{L}^p(\mathcal{H}) \quad \forall p \geq 1.$$

Using the triangle inequality,

$$\left\|U_1^{(t),a}\right\|_p \leq \left\|U_1^{(t),f}\right\|_p + \left\|\mathrm{K}^{(t),U}y^{(t)}\right\|_p + \left\|\mathrm{K}^{(t),U}\mathrm{H}U_1^{(t),f}\right\|_p + \left\|\mathrm{K}^{(t),U}W_1^{(t)}\right\|_p,$$

and to finish the proof just recall that $\left\|U_1^{(t),a}\right\|_p$ is finite by assumption; term $\mathrm{K}^{(t),U}y^{(t)}$ is deterministic; $\left\|\mathrm{K}^{(t),U}W_1^{(t)}\right\|_p$ is finite because $\mathrm{K}^{(t),U}W_1^{(t)}$ is a Gaussian random variable by Lemma 24, and

$$\left\|\mathrm{K}^{(t),U}\mathrm{H}U_1^{(t),f}\right\|_p \leq |\mathrm{H}|\left|\left(\mathrm{R}^{(t)}\right)^{-1}\right|\left|\mathrm{C}^{(t),f}\right|\left\|U_1^{(t),f}\right\|_p < \infty$$

by Corollary 6. □

7.4 Auxiliary estimates

This section contains various estimates, which are necessary to prove main theorems of the chapter. We divide the estimates into three categories:

1. empirical moments estimates,
2. covariance distance estimates, and
3. data noise estimates.

Empirical moment estimates

Lemma 44. *There are positive constants l, s depending only on the iterated map Ψ such that*

$$\widehat{Z}_{N,p}^{(t),f} \leq l\left(\widehat{Z}_{N,p}^{(t-1),a} + \widehat{Z}_{N,2p}^{(t-1),a}\left(\widehat{U}_{N,2ps}^{(t-1),a}\right)^{s} + \left(\widehat{Z}_{N,p(s+1)}^{(t-1),a}\right)^{s+1}\right)$$

for every $t \in \mathbb{N}$ and any $p \geq 1$.

Proof. For given t, N and any $i = 1, \ldots, N$ we define

$$X_i^a = X_{N,i}^{(t-1),a},\ U_i^a = U_i^{(t-1),a} \text{ and } Z_i^a = Z_{N,i}^{(t-1),a}.$$

The iterated map Ψ is locally Lipschitz continuous with at most polynomial growth in infinity by Assumption 1. Therefore, using Lemma 3, there exist positive constants s and l such that

$$\begin{aligned}\left|\Psi\left(X_i^a\right) - \Psi\left(U_i^a\right)\right| \leq & l\left(\left|X_i^a - U_i^a\right| + \left|X_i^a - U_i^a\right|\left|U_i^a\right|^s + \left|X_i^a - U_i^a\right|^{s+1}\right) \\ = & l\left(\left|Z_i^a\right| + \left|Z_i^a\right|\left|U_i^a\right|^s + \left|Z_i^a\right|^{s+1}\right) \end{aligned} \tag{7.11}$$

for any $i = 1, \ldots, N$, and these constants depend only on Ψ. Using (7.11) and the triangle inequality,

$$\begin{aligned}\widehat{Z}_{N,p}^{(t),f} = & \left(\frac{1}{N}\sum_{i=1}^{N}\left|\Psi\left(X_i^a\right) - \Psi\left(U_i^a\right)\right|^p\right)^{1/p} \\ \leq & l\left(\frac{1}{N}\sum_{i=1}^{N}\left|Z_i^a\right|^p\right)^{1/p} + l\left(\frac{1}{N}\sum_{i=1}^{N}\left(\left|Z_i^a\right|\left|U_i^a\right|^s\right)^p\right)^{1/p} \\ & + l\left(\frac{1}{N}\sum_{i=1}^{N}\left|Z_i^a\right|^{p(s+1)}\right)^{1/p} \\ = & l\left(\widehat{Z}_{N,p}^{(t-1),a} + \left(\frac{1}{N}\sum_{i=1}^{N}\left(\left|Z_i^a\right|\left|U_i^a\right|^s\right)^p\right)^{1/p} + \left(\widehat{Z}_{N,p(s+1)}^{(t-1),a}\right)^{s+1}\right). \end{aligned}$$

To finish the proof, we just use Cauchy-Schwarz inequality to obtain

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(\left|Z_i^a\right|^p\left|U_i^a\right|^{ps}\right) \leq & \left(\frac{1}{N}\sum_{i=1}^{N}\left|Z_i^a\right|^{2p}\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\left|U_i^a\right|^{2ps}\right)^{1/2} \\ = & \left(\widehat{Z}_{N,2p}^{(t-1),a}\right)^p\left(\widehat{U}_{N,2ps}^{(t-1),a}\right)^{ps} \end{aligned}$$

and the inequality in the statement of the lemma follows. □

Lemma 45. *There exist real positive constants $c^{(t)}$, $t \in \mathbb{N}$, such that*

$$\begin{aligned}\widehat{Z}_{N,p}^{(t),a} \leq & c^{(t)} \left(\widehat{Z}_{N,p}^{(t),f} + \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \widehat{Z}_{N,p}^{(t),f} \right) \\ & + c^{(t)} \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \left(\frac{1}{N} \sum_{i=1}^{N} \left| y^{(t)} - \mathrm{H} U_i^{(t),f} \right|^p \right)^{1/p} \\ & + \left(\frac{1}{N} \sum_{i=1}^{N} \left| \widehat{\mathrm{K}}_N^{(t),X} W_i^{(t)} - \mathrm{K}^{(t),U} W_i^{(t)} \right|^p \right)^{1/p}\end{aligned}$$

any $p \geq 1$. These constants depend on t only.

Proof. Similar to the proof of the previous lemma we do note use unnecessary indexes, and for given t and N we define

$$\begin{aligned} X_i^\bullet = & X_{N,i}^{(t),\bullet}, & U_i^\bullet = & U_i^{(t),\bullet}, & Z_i^\bullet = & Z_{N,i}^{(t),\bullet}, & \widehat{\mathrm{K}}_N^X = & \widehat{\mathrm{K}}_N^{(t),X}, & \mathrm{K}^U = & \mathrm{K}^{(t),U}, \\ y = & y^{(t)} & W_i = & W_i^{(t)}, & \widehat{\mathrm{P}}_N^f = & \widehat{\mathrm{P}}_N^{(t),f}, & \mathrm{C}^f = & \mathrm{C}^{(t),f}, & \mathrm{R} = & \mathrm{R}^{(t)}, \end{aligned}$$

where $\bullet$ stands for either f or a.

For each $i = 1, \ldots, N$,

$$\begin{aligned} Z_i^a = & X_i^a - U_i^a \\ = & X_i^f + \widehat{\mathrm{K}}_N^X \left(y - \mathrm{H} X_i^f - W_i \right) - U_i^f - \mathrm{K}^U \left(y - \mathrm{H} U_i^f - W_i \right) \\ = & \left(\mathrm{I} - \mathrm{K}^U \mathrm{H} \right) Z_i^f - \left(\widehat{\mathrm{K}}_N^X - \mathrm{K}^U \right) \mathrm{H} Z_i^f + \left(\widehat{\mathrm{K}}_N^X - \mathrm{K}^U \right) \left(y - \mathrm{H} U_i^f \right) \\ & - \widehat{\mathrm{K}}_N^X W_i + \mathrm{K}^U W_i, \end{aligned} \tag{7.12}$$

and we need to estimate the norm of each term on the right side of the last equation. Using the triangle inequality and (7.9),

$$\begin{aligned} \left| \left(\mathrm{I} - \mathrm{K}^U \mathrm{H} \right) Z_i^f \right| \leq & \left| \mathrm{I} - \mathrm{K}^U \mathrm{H} \right| \left| Z_i^f \right| \\ \leq & \left(1 + \left| \mathrm{C}^f \right| |\mathrm{H}|^2 \left| \mathrm{R}^{-1} \right| \right) \left| Z_i^f \right|, \end{aligned} \tag{7.13}$$

and from Theorem 41 there is $k_1 > 0$ such that

$$\begin{aligned} \left| \left(\widehat{\mathrm{K}}_N^X - \mathrm{K}^U \right) \mathrm{H} Z_i^f \right| \leq & \left| \widehat{\mathrm{K}}_N^X - \mathrm{K}^U \right| |\mathrm{H}| \left| Z_i^f \right| \\ \leq & k_1 \left(1 + \left| \mathrm{C}^f \right| \right) \left| \widehat{\mathrm{P}}_N^f - \mathrm{C}^f \right| |\mathrm{H}| \left| Z_i^f \right|. \end{aligned} \tag{7.14}$$

Using the same theorem again gives

$$\begin{aligned} \left| \left(\widehat{\mathrm{K}}_N^X - \mathrm{K}^U \right) \left(y - \mathrm{H} U_i^f \right) \right| \leq & \left| \widehat{\mathrm{K}}_N^X - \mathrm{K}^U \right| \left| y - \mathrm{H} U_i^f \right| \\ \leq & k_1 \left(1 + \left| \mathrm{C}^f \right| \right) \left| \widehat{\mathrm{P}}_N^f - \mathrm{C}^f \right| \left| y - \mathrm{H} U_i^f \right|. \end{aligned} \tag{7.15}$$

Putting (7.13), (7.14) and (7.15) into (7.12) yields

$$\begin{aligned} |Z_i^a| \leq & c \left(\left| Z_i^f \right| + \left| \widehat{\mathrm{P}}_N^f - \mathrm{C}^f \right| \left| Z_i^f \right| + \left| \widehat{\mathrm{C}}_N^f - \mathrm{C}^f \right| \left| y - \mathrm{H} U_i^f \right| \right) \\ & + \left| \widehat{\mathrm{K}}_N^X W_i + \mathrm{K}^U W_i \right| \end{aligned} \tag{7.16}$$

where we define

$$c = \max\left\{k_1\left(|H|+1\right)\left(1+\left|\mathrm{C}^f\right|\right), 1+\left|\mathrm{C}^f\right|\left|\mathrm{H}\right|\left|\mathrm{R}^{-1}\right|\right\}.$$

We finish the proof by bounding each term of the sum

$$\left(\frac{1}{N}\sum_{i=1}^{N}|Z_i^a|^p\right)^{1/p}$$

by (7.16) and using the triangle inequality. □

Covariance distance estimates

Now, we derive a bound on the distance of the forecast sample covariance $\widehat{\mathrm{P}}_N^{(t),f}$ and the covariance $\mathrm{C}^{(t),f}$, where both operators are defined in Definition 13 and Definition 14 respectively.

Lemma 46. *Assume that for a fixed $t \in \mathbb{N}$ and every $p \geq 1$ there is a positive constants k_p such that*

$$\left\|Z_{N,1}^{(t),f}\right\|_p \leq \frac{k_p}{\sqrt{N}} \tag{7.17}$$

for all $N \in \mathbb{N}$. Then, for every $p \geq 1$ there is a positive constant c_p such that

$$\left\|\widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f}\right\|_p \leq \frac{c_p}{\sqrt{N}}.$$

Proof. First, notice that from assumption (7.17) follows that $Z_{N,1}^{(t),f} \in \mathcal{L}^p\left(\mathcal{H}\right)$ for all $p \geq 1$. Let $p \geq 2$. By the triangle inequality

$$\left\|\widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f}\right\|_p \leq \left\|\widehat{\mathrm{P}}_N^{(t),f} - \widehat{\mathrm{C}}_N^{(t),f}\right\|_p + \left\|\widehat{\mathrm{C}}_N^{(t),f} - \mathrm{C}^{(t),f}\right\|_p. \tag{7.18}$$

The first term on the right side of (7.18) can be bounded using Lemma 16 and Cauchy-Schwarz inequality,

$$\begin{aligned}\left\|\widehat{\mathrm{P}}_N^{(t),f} - \widehat{\mathrm{C}}_N^{(t),f}\right\|_p \leq& 2\left\|\left(\widehat{Z}_{N,2}^{(t),f}\right)^2\right\|_p + 4\left\|\widehat{Z}_{N,2}^{(t),f}\widehat{U}_{N,2}^{(t),f}\right\|_p \\ \leq& 2\left\|\widehat{Z}_{N,2}^{(t),f}\right\|_{2p}^2 + 4\left\|\widehat{Z}_{N,2}^{(t),f}\right\|_{2p}\left\|\widehat{U}_{N,2}^{(t),f}\right\|_{2p},\end{aligned}$$

and the second term on the right side of (7.18)

$$\left\|\widehat{\mathrm{C}}_N^{(t),f} - \mathrm{C}^{(t),f}\right\|_p \leq \frac{\widetilde{k}_p}{\sqrt{N}}\left\|U_1^{(t),f}\right\|_{2p}$$

where the existence of such $\widetilde{k}_p \in \mathbb{R}$ follows from Corollary 4. These bounds together with Lemma 17 give

$$\begin{aligned}\left\|\widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f}\right\|_p \leq& 2\left\|\widehat{Z}_{N,2}^{(t),f}\right\|_{2p}^2 + 4\left\|\widehat{Z}_{N,2}^{(t),f}\right\|_{2p}\left\|\widehat{U}_{N,2}^{(t),f}\right\|_{2p} + \frac{\widetilde{k}_p}{\sqrt{N}}\left\|U_1^{(t),f}\right\|_{2p} \\ \leq& 2\left\|Z_1^{(t),f}\right\|_{2p}^2 + 4\left\|Z_1^{(t),f}\right\|_{2p}\left\|U_1^{(t),f}\right\|_{2p} + \frac{\widetilde{k}_p}{\sqrt{N}}\left\|U_1^{(t),f}\right\|_{2p} \\ \leq& \frac{\max\left\{k_{2p}^2, \widetilde{k}_p\right\}}{\sqrt{N}}\left(2 + 4\left\|U_1^{(t),f}\right\|_{2p} + \left\|U_1^{(t),f}\right\|_{2p}\right),\end{aligned}$$

and to finish the proof we define

$$c_p = \begin{cases} 2\max\left\{k_p^2, \widetilde{k}_p\right\}\left(1 + 5\left\|U_1^{(t),f}\right\|\right) & \text{for } p \geq 2, \\ c_2 & \text{for } p < 2. \end{cases}$$

□

Data noise estimates

The last group of estimates bound the expression

$$\left\|\widehat{\mathrm{K}}_N^{(t),X} W_1^{(t)} - \mathrm{K}^{(t),U} W_1^{(t)}\right\|_p. \tag{7.19}$$

If the data noise $W_1^{(t)}$ is a finite dimensional random variable, then (7.19) can be estimated using the Hölder inequality,

$$\left\|\widehat{\mathrm{K}}_N^{(t),X} W_1^{(t)} - \mathrm{K}^{(t),U} W_1^{(t)}\right\|_p \leq \left\|\widehat{\mathrm{K}}_N^{(t),X} - \mathrm{K}^{(t),U}\right\|_{2p} \left\|W_1^{(t)}\right\|_{2p} \tag{7.20}$$

similarly as in Proposition 22.1 in Le Gland et al. [2011]. When $W_1^{(t)}$ is only a weak random variable, the norm

$$\left\|W_1^{(t)}\right\|_p$$

is not defined, so the right-hand side of (7.20) is undefined. On the other hand, we can use the fact that $W_1^{(t)}$ is a Gaussian weak random variable, and find a reasonable upper bound.

We divide the process of finding the upper bound to (7.19) in four lemmas, and we formulate these lemmas for a fixed $t \in \mathbb{N}$. Therefore, just for this subsection, we simplify the notation from Definition 13 and Definition 14, and we drop the time index, e.g.,

$$X_{N,i}^f = X_{N,i}^{(t),f},$$

Using this convention, we define additional operators

$$\begin{aligned} \widehat{\mathrm{A}}_N^X =& \mathrm{H}^* \left(\mathrm{H}\widehat{\mathrm{P}}_N^f \mathrm{H}^* + \mathrm{R}\right)^{-1}, \\ \mathrm{A}^U =& \mathrm{H}^* \left(\mathrm{H}\mathrm{C}^f \mathrm{H}^* + \mathrm{R}\right)^{-1}. \end{aligned}$$

Obviously,

$$\begin{aligned} \widehat{\mathrm{K}}_N^X =& \widehat{\mathrm{P}}_N^f \widehat{\mathrm{A}}_N^X, \\ \mathrm{K}^U =& \mathrm{C}^f \mathrm{A}^U, \end{aligned}$$

and the norms of both operators are bounded by the value $|\mathrm{H}|\,|\mathrm{R}^{-1}|$, i.e.,

$$\left|\widehat{\mathrm{A}}_N^X\right| \leq |\mathrm{H}|\left|\mathrm{R}^{-1}\right| \text{ and } \left|\mathrm{A}^U\right| \leq |\mathrm{H}|\left|\mathrm{R}^{-1}\right|. \tag{7.21}$$

The next four lemmas have the following assumptions in common.

Assumption 2. *Assume that for any $p \geq 1$, there is a positive constant k_p such that*

$$\left\| Z_{N,1}^f \right\|_p \leq \frac{k_p}{\sqrt{N}}$$

for all $N \in \mathbb{N}$.

Lemma 47. *If Assumption 2 holds, then for any $p \geq 1$, there is a positive constant c_p such that*

$$\left\| \widehat{\mathrm{P}}_N^f \widehat{\mathrm{A}}_N^X W_1 - \widehat{\mathrm{C}}_N^f \widehat{\mathrm{A}}_N^X W_1 \right\|_p \leq \frac{c_p}{\sqrt{N}}$$

for all $N \in \mathbb{N},\ N > 1$.

Proof. The identities

$$X_{N,i}^f \otimes X_{N,i}^f - U_i^f \otimes U_i^f = Z_{N,i}^f \otimes X_{N,i}^f - U_i^f \otimes Z_{N,i}^f,\ i = 1, \ldots, N,$$

and

$$\overline{X}_N^f \otimes \overline{X}_N^f - \overline{U}_N^f \otimes \overline{U}_N^f = \overline{Z}_N^f \otimes \overline{X}_N^f - \overline{U}_N^f \otimes \overline{Z}_N^f$$

together with the triangle inequality give

$$\begin{aligned} \left\| \widehat{\mathrm{P}}_N^f \widehat{\mathrm{A}}_N^X W_1 - \widehat{\mathrm{C}}_N^f \widehat{\mathrm{A}}_N^X W_1 \right\|_p \leq & \left\| \frac{1}{N} \sum_{i=1}^N \left(Z_{N,i}^f \otimes X_{N,i}^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p \\ & + \left\| \frac{1}{N} \sum_{i=1}^N \left(U_i^f \otimes Z_{N,i}^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p \\ & + \left\| \left(\overline{Z}_N^f \otimes \overline{X}_N^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p + \left\| \left(\overline{U}_N^f \otimes \overline{Z}_N^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p . \end{aligned} \tag{7.22}$$

The first term on the right side of (7.22) can be bounded again by the triangle inequality,

$$\begin{aligned} \left\| \frac{1}{N} \sum_{i=1}^N \left(Z_{N,i}^f \otimes X_{N,i}^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p \leq & \frac{1}{N} \sum_{i=1}^N \left\| \left(Z_{N,i}^f \otimes X_{N,i}^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p \\ = & \left\| Z_1^f \left\langle X_{N,1}^f, \widehat{\mathrm{A}}_N^X W_1 \right\rangle \right\|_p , \end{aligned}$$

and, using Cauchy-Schwarz inequality and Lemma 21,

$$\begin{aligned} \left\| Z_{N,1}^f \left\langle X_{N,1}^f, \widehat{\mathrm{A}}_N^X W_1 \right\rangle \right\|_p &= \left\| Z_{N,1}^f \left\langle \left(\widehat{\mathrm{A}}_N^X \right)^* X_{N,1}^f, W_1 \right\rangle \right\|_p \\ &\leq \left\| Z_{N,1}^f \right\|_{2p} \left\| \left\langle \left(\widehat{\mathrm{A}}_N^X \right)^* X_{N,1}^f, W_1 \right\rangle \right\|_{2p} \\ &\leq \left\| Z_{N,1}^f \right\|_{2p} \left\| \left(\widehat{\mathrm{A}}_N^X \right)^* X_{N,1}^f \right\|_{2p} \|W_1\|_{2p,w} \\ &\leq |\mathrm{R}|\,|\mathrm{H}| \left\| Z_{N,1}^f \right\|_{2p} \left\| X_{N,1}^f \right\|_{2p} \|W_1\|_{2p,w} , \end{aligned} \tag{7.23}$$

where the last inequality follows from (7.21). By the same approach it is possible to show that

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \left(U_i^f \otimes Z_{N,i}^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p \leq |\mathrm{R}| \, |\mathrm{H}| \left\| Z_{N,1}^f \right\|_{2p} \left\| U_1^f \right\|_{2p} \|W_1\|_{2p,w} . \tag{7.24}$$

Similarly,

$$\begin{aligned} \left\| \left(\overline{Z}_N^f \otimes \overline{X}_N^f \right) \widehat{\mathrm{A}}_N^X W_1 \right\|_p &= \left\| \overline{Z}_N^f \left\langle \left(\widehat{\mathrm{A}}_N^X \right)^* \overline{X}_N^f, W_1 \right\rangle \right\|_p \\ &\leq |\mathrm{R}| \, |\mathrm{H}| \left\| \overline{Z}_N^f \right\|_{2p} \left\| \overline{X}_N^f \right\|_{2p} \|W_1\|_{2p,w} \\ &\leq |\mathrm{R}| \, |\mathrm{H}| \left\| Z_{N,1}^f \right\|_{2p} \left\| X_{N,1}^f \right\|_{2p} \|W_1\|_{2p,w} , \end{aligned} \tag{7.25}$$

where the last inequality follows from the triangle inequality applied on terms $\left\| \overline{Z}_N^f \right\|_{2p}$ and $\left\| \overline{X}_N^f \right\|_{2p}$, and

$$\left\| \left(\overline{U}_N^f \otimes \overline{Z}_N^f \right) \hat{\mathrm{A}}_N^X W_1 \right\|_p \leq |\mathrm{R}| \, |\mathrm{H}| \left\| Z_{N,1}^f \right\|_{2p} \left\| U_1^f \right\|_{2p} \|W_1\|_{2p,w} . \tag{7.26}$$

Now, putting (7.23), (7.24), (7.25) and (7.26) into (7.22) gives

$$\begin{aligned} \left\| \widehat{\mathrm{P}}_N^f \widehat{\mathrm{A}}_N^X W_1 - \widehat{\mathrm{C}}_N^f \widehat{\mathrm{A}}_N^X W_1 \right\|_p \leq & 2 \, |\mathrm{R}| \, |\mathrm{H}| \left\| Z_{N,1}^f \right\|_{2p} \|W_1\|_{2p,w} \left(\left\| X_{N,1}^f \right\|_{2p} + \left\| U_{N,1}^f \right\|_{2p} \right) \\ \leq & 2 \, |\mathrm{R}| \, |\mathrm{H}| \left\| Z_{N,1}^f \right\|_{2p}^2 \|W_1\|_{2p,w} \left(1 + 2 \frac{\left\| U_{N,1}^f \right\|_{2p}}{\left\| Z_{N,1}^f \right\|_{2p}} \right) , \end{aligned}$$

where the last inequality is obtained by applying the triangle inequality, and we define

$$c_p = 2 k_{2p}^2 \, |\mathrm{R}| \, |\mathrm{H}| \, \|W_1\|_{2p,w} \left(1 + 2 \left\| U_{N,1}^f \right\|_{2p} k_{2p}^{-1} \right)$$

to conclude the proof. □

Lemma 48. *If Assumption 2 holds, then for any $p \geq 1$, there exists positive constant c_p such that*

$$\left\| \widehat{\mathrm{C}}_N^f \widehat{\mathrm{A}}_N^X W_1 - \widehat{\mathrm{C}}_N^f \mathrm{A}^U W_1 \right\|_p \leq \frac{c_p}{\sqrt{N}}$$

for all $N \in \mathbb{N}, \ N > 1$.

Proof. Pick $p \geq 1$ arbitrary. Define operator

$$\mathrm{G}_N = \widehat{\mathrm{A}}_N^X - \mathrm{A}^U ,$$

and, using (7.21), this operator is bounded.

Using Lemma 7,

$$\begin{aligned} \|\mathrm{G}_N\|_p &= \left\| \mathrm{H}^* \left(\left(\mathrm{H} \widehat{\mathrm{P}}_N^f \mathrm{H}^* + \mathrm{R} \right)^{-1} - \left(\mathrm{H} \mathrm{C}^f \mathrm{H}^* + \mathrm{R} \right)^{-1} \right) \right\|_p \\ &\leq \left\| |\mathrm{H}| \left| \mathrm{H} \widehat{\mathrm{P}}_N^f \mathrm{H}^* - \mathrm{H} \mathrm{C}^f \mathrm{H}^* \right| \left| \mathrm{R}^{-1} \right|^2 \right\|_p \\ &\leq \left| \mathrm{R}^{-1} \right|^2 |\mathrm{H}|^3 \left\| \widehat{\mathrm{P}}_N^f - \mathrm{C}^f \right\|_p , \end{aligned}$$

and, using Lemma 46,

$$\|\mathrm{G}_N\|_p \leq \frac{\widetilde{k}_p}{\sqrt{N}} \left|\mathrm{R}^{-1}\right|^2 |\mathrm{H}|^3 \tag{7.27}$$

for some positive constant $\widetilde{k}_p$.

Using the definition of a sample covariance and the triangle inequality,

$$\begin{aligned}\left\|\widehat{\mathrm{C}}_N^f \mathrm{G}_N W_1\right\|_p \leq & \frac{1}{N-1}\sum_{i=1}^{N}\left\|\left(U_i^f - \overline{U}_N^f \otimes U_i^f - \overline{U}_N^f\right)\mathrm{G}_N W_1\right\|_p \\ \leq & \frac{N}{N-1}\left\|\left(U_1^f - \overline{U}_N^f\right)\left\langle U_1^f - \overline{U}_N^f, \mathrm{G}_N W_1\right\rangle\right\|_p,\end{aligned}$$

and using the Cauchy-Schwarz inequality on the right side of the last inequality gives

$$\left\|\left(U_1^f - \overline{U}_N^f\right)\left\langle U_1^f - \overline{U}_N^f, \mathrm{G}_N W_1\right\rangle\right\|_p \leq 2\left\|U_1^f\right\|_{2p}\left\|\left\langle U_1^f - \overline{U}_N^f, \mathrm{G}_N W_1\right\rangle\right\|_{2p} \tag{7.28}$$

because

$$\left\|U_1^f - \overline{U}_N^f\right\|_{2p} \leq 2\left\|U_1^f\right\|_{2p}.$$

By Lemma 21

$$\left\|\left\langle \mathrm{G}_N^*\left(U_1^f - \overline{U}_N^f\right), W_1\right\rangle\right\|_{2p} \leq \left\|\mathrm{G}_N^*\left(U_1^f - \overline{U}_N^f\right)\right\|_{2p}\|W_1\|_{2p,w},$$

and again by the Cauchy-Schwarz inequality

$$\begin{aligned}\mathrm{E}\left|\mathrm{G}_N^*\left(U_1^f - \overline{U}_N^f\right)\right|^{2p} \leq & \mathrm{E}\left(|\mathrm{G}_N|^{2p}\left|\left(U_1^f - \overline{U}_N^f\right)\right|^{2p}\right) \\ & \leq \left(\mathrm{E}\,|\mathrm{G}_N|^{4p}\right)^{1/2}\left(\mathrm{E}\left|\left(U_1^f - \overline{U}_N^f\right)\right|^{4p}\right)^{1/2}.\end{aligned}$$

By putting all equations together we obtain

$$\left\|\widehat{\mathrm{C}}_N^f \mathrm{G}_N W_1\right\|_p \leq 4\frac{N}{N-1}\|\mathrm{G}_N\|_{4p}\left\|U_1^f\right\|_{2p}\left\|U_1^f\right\|_{4p}\|W_1\|_{2p,w}$$

and, using (7.27),

$$\left\|\widehat{\mathrm{C}}_N^f \mathrm{G}_N W_1\right\|_p \leq \frac{c_p}{\sqrt{N}}$$

where we define

$$c_p = 4\widetilde{k}_{4p}\left|\mathrm{R}^{-1}\right|^2|\mathrm{H}|^3\left\|U_1^f\right\|_{2p}\left\|U_1^f\right\|_{4p}\|W_1\|_{2p,w}.$$

□

Lemma 49. *For any $p \geq 1$, there is a positive constant c_p such that*

$$\left\|\widehat{\mathrm{C}}_N^f \mathrm{A}^U W_1 - \mathrm{C}^f \mathrm{A}^U W_1\right\|_p \leq \frac{c_p}{\sqrt{N}}$$

for all $N \in \mathbb{N},\ N > 1$.

Proof. Firstly, $\mathrm{A}^U W_1$ is a weak random variable, and

$$\mathrm{A}^U W_1 \sim \mathcal{N}_w\left(0, \mathrm{A}^U \mathrm{R}\left(\mathrm{A}^U\right)^*\right)$$

by Lemma 24.

Pick $p \in \mathbb{N}$ arbitrary. Then,

$$\left\|\left(\widehat{\mathrm{C}}_N^f - \mathrm{C}^f\right) \mathrm{A}^U W_1\right\|_{2p}^{2p} = \int_\Omega \int_\Omega \left|\left(\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right) \mathrm{A}^U W_1(\omega_2)\right|^{2p} dP(\omega_2)\, dP(\omega_1). \tag{7.29}$$

From Lemma 24 follows that for any fixed $\omega_1 \in \Omega$,,

$$S_{\omega_1} = \left(\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right) \mathrm{A}^U W_1$$

is a measurable random variable with $\mathcal{N}(0, \Sigma_N(\omega_1))$ distribution where

$$\Sigma_N(\omega_1) = \left(\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right) \mathrm{A}^U \mathrm{R}\left(\mathrm{A}^U\right)^* \left(\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right)^*,$$

and from Lemma 26 yields existence of positive constant $\widetilde{k}_p$ such that

$$\int_\Omega \left|\left(\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right) \mathrm{A} W_2(\omega_2)\right|^{2p} dP(\omega_2) \leq \widetilde{k}_p \left|\Sigma_N(\omega_1)\right|_{\mathrm{Tr}}^p. \tag{7.30}$$

The operator $\left(\widehat{\mathrm{C}}_N^f - \mathrm{C}^f\right)$ is Hilbert-Schmidt and the operator $\mathrm{A}^U \mathrm{R}\left(\mathrm{A}^U\right)^*$ is bounded, so the operator

$$\left(\widehat{\mathrm{C}}_N^f - \mathrm{C}^f\right) \mathrm{A}^U \mathrm{R}\left(\mathrm{A}^U\right)^*$$

is Hilbert-Schmidt as well, and, using Lemma 6,

$$\begin{aligned}\left|\Sigma_N(\omega_1)\right|_{\mathrm{Tr}} \leq & 2 \left|\left(\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right) \mathrm{A}^U \mathrm{R}\left(\mathrm{A}^U\right)^*\right|_{\mathrm{HS}} \left|\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right|_{\mathrm{HS}} \\ \leq & 2 \left|\mathrm{A}^U \mathrm{R}\left(\mathrm{A}^U\right)^*\right| \left|\widehat{\mathrm{C}}_N^f(\omega_1) - \mathrm{C}^f\right|_{\mathrm{HS}}^2. \end{aligned} \tag{7.31}$$

Putting (7.30) and (7.31) into (7.29) gives

$$\begin{aligned}\left\|\widehat{\mathrm{C}}_N^f \mathrm{A} W_1 - \mathrm{C}^f \mathrm{A} W_1\right\|_{2p}^{2p} \leq & \int_\Omega \widetilde{k}_p \left|\Sigma_N(\omega_1)\right|_{\mathrm{Tr}}^p dP(\omega_1) \\ \leq & \widetilde{k}_p 2 \left|\mathrm{A}^U \mathrm{R}\left(\mathrm{A}^U\right)^*\right|^{2p} \mathrm{E} \left|\widehat{\mathrm{C}}_N^f - \mathrm{C}^f\right|_{\mathrm{HS}}^{2p}, \end{aligned}$$

and, similarly as in the proof of Theorem 32, the Marcinkiewicz-Zygmund inequality, Theorem 30, gives existence of positive constant b_{2p} such that

$$\left(\mathrm{E} \left|\widehat{\mathrm{C}}_N^f - \mathrm{C}^f\right|_{\mathrm{HS}}^{2p}\right)^{1/(2p)} \leq \frac{b_{2p}}{\sqrt{N}}.$$

To conclude the proof we define

$$c_p = \begin{cases} \widetilde{k}_{p/2}^{1/(2p)} \left|\mathrm{A}^{\mathrm{U}} \mathrm{R}\left(\mathrm{A}^{\mathrm{U}}\right)^*\right| b_{2p} & \text{for } p = 2r,\ r \in \mathbb{N} \\ c_{2r} & \text{for } p \in (2r-2, 2r),\ r \in \mathbb{N}. \end{cases}$$

□

Lemma 50. *If Assumption 2 holds, then for any $p \geq 1$, there exists positive constant c_p such that*

$$\left\| \widehat{\mathrm{K}}_N^{(t),X} W_1^{(t)} - \mathrm{K}^{(t),U} W_1^{(t)} \right\|_p \leq \frac{c_p}{\sqrt{N}}$$

for all $N \in \mathbb{N},\ N > 1$.

Proof. Using the triangle inequality,

$$\begin{aligned}\left\| \widehat{\mathrm{K}}_N^X W_1 - \mathrm{K}^U W_1 \right\|_p \leq & \left\| \widehat{\mathrm{P}}_N^f \hat{\mathrm{A}}_N^X W_1 - \mathrm{C}^f \hat{\mathrm{A}}_N^X W_1 \right\|_p + \left\| \widehat{\mathrm{C}}_N^f \hat{\mathrm{A}}_N^X W_1 - \widehat{\mathrm{C}}_N^f \mathrm{A}^U W_1 \right\|_p \\ & + \left\| \widehat{\mathrm{C}}_N^f \mathrm{A}^U W_1 - \mathrm{C}^f \mathrm{A}^U W_1 \right\|_p \\ \leq & \frac{k_p^1 + k_p^2 + k_p^3}{\sqrt{N}}\end{aligned}$$

where the existence of positive constants k_p^1, k_p^2 and k_p^3 follows from Lemma 47, Lemma 48 and Lemma 49. □

7.5 Convergence of ensembles

The next three theorems form the main result of this chapter, they show that $X_{N,1}^{(t),f}$ converges to $U_1^{(t),f}$ in $\mathcal{L}^p$ as the size of the ensemble goes to infinity.

Recall that, in all statements, we are using the notation introduced in Section 7.1.

We estimate the difference of forecast members using the properties of the iterated map Ψ and a prior estimate of a difference of analysis members from the previous time step.

Lemma 51. *Assume that for a fixed $t \in \mathbb{N}$ and any $p \geq 1$, there is a positive constant $k_p^{(t-1),a}$ such that*

$$\left\| Z_{N,1}^{(t-1),a} \right\|_p \leq \frac{k_p^{(t-1),a}}{\sqrt{N}}.$$

Then, for any $p \geq 1$, there is a positive constant $k_p^{(t),f}$ such that

$$\left\| Z_{N,1}^{(t),f} \right\|_p \leq \frac{k_p^{(t),f}}{\sqrt{N}}.$$

Proof. First, using Theorem 43,

$$\left\| U_1^{(t-1),a} \right\|_p < \infty$$

for any $p > 1$.

Let $p \geq 1$. Using Lemma 17, Lemma 44, and the triangle inequality,

$$\begin{aligned}\left\| Z_{N,1}^{(t),f} \right\|_p = & \left\| \left(\frac{1}{N} \sum_{i=1}^{N} \left| Z_{N,i}^{(t),f} \right|^p \right)^{1/p} \right\|_p = \left\| \widehat{Z}_{N,p}^{(t),f} \right\|_p \\ & \leq l \left\| \widehat{Z}_{N,p}^{(t-1),a} \right\|_p + l \left\| \widehat{Z}_{N,2p}^{(t-1),a} \left(\widehat{U}_{N,2ps}^{(t-1),a} \right)^s \right\|_p + l \left\| \left(\widehat{Z}_{N,p(s+1)}^{(t-1),a} \right)^{s+1} \right\|_p, \quad (7.32)\end{aligned}$$

where s and l are some positive constants that depend on Ψ only, and using the Cauchy-Schwarz inequality,

$$\begin{aligned}\left\|\widehat{Z}_{N,2p}^{(t-1),a}\left(\widehat{U}_{N,2ps}^{(t-1),a}\right)^{s}\right\|_{p}^{p} &= \left\|\left(\widehat{Z}_{N,2p}^{(t-1),a}\right)^{p}\left(\widehat{U}_{N,2ps}^{(t-1),a}\right)^{ps}\right\|_{1} \\ &\leq \left\|\left(\widehat{Z}_{N,2p}^{(t-1),a}\right)^{p}\right\|_{2}\left\|\left(\widehat{U}_{N,2ps}^{(t-1),a}\right)^{ps}\right\|_{2} \\ &= \left\|\widehat{Z}_{N,2p}^{(t-1),a}\right\|_{2p}^{p}\left\|\widehat{U}_{N,2ps}^{(t-1),a}\right\|_{2ps}^{ps} = \left\|Z_{1}^{(t-1),a}\right\|_{2p}^{p}\left\|U_{1}^{(t-1),a}\right\|_{2ps}^{ps}\end{aligned} \tag{7.33}$$

where the last equality follows again from Lemma 17. This lemma also yields

$$\left\|\left(\widehat{Z}_{N,p(s+1)}^{(t-1),a}\right)^{s+1}\right\|_{p} = \left\|\widehat{Z}_{N,p(s+1)}^{(t-1),a}\right\|_{p(s+1)}^{s+1} = \left\|Z^{(t-1),a}\right\|_{p(s+1)}^{s+1}. \tag{7.34}$$

To conclude the proof just put (7.33) and (7.34) into (7.32) to obtain

$$\begin{aligned}\left\|Z_{N,1}^{(t),f}\right\|_{p} \leq & l\left(\frac{k_{p}^{(t-1),a}}{\sqrt{N}} + \frac{k_{2p}^{(t-1),a}}{\sqrt{N}}\left\|U_{1}^{(t-1),a}\right\|_{2ps}^{s} + \left(\frac{k_{p(s+1)}^{(t-1),a}}{\sqrt{N}}\right)^{s+1}\right) \\ \leq & \frac{k_{p}^{(t),f}}{\sqrt{N}},\end{aligned}$$

where

$$k_{p}^{(t),f} = l\left(k_{p}^{(t-1),a} + k_{2p}^{(t-1),a}\left\|U_{1}^{(t-1),a}\right\|_{2ps}^{s} + \left(k_{p(s+1)}^{(t-1),a}\right)^{s+1}\right).$$

□

Next, we bound the difference of forecast members.

Lemma 52. *Assume that for a fixed $t \in \mathbb{N}$ and any $p \geq 1$, there is a positive constant $k_{p}^{(t),f}$ such that*

$$\left\|Z_{N,1}^{(t),f}\right\|_{p} \leq \frac{k_{p}^{(t),f}}{\sqrt{N}}. \tag{7.35}$$

Then, for any $p \geq 1$, there is a positive constant $k_{p}^{(t),a}$ such that

$$\left\|Z_{N,1}^{(t),a}\right\|_{p} \leq \frac{k_{p}^{(t),a}}{\sqrt{N}}.$$

Proof. Again, using Theorem 43,

$$\left\|U_{1}^{(t),f}\right\|_{p} < \infty$$

for any $p > 1$.

Let $p \geq 1$. Lemma 17 gives the identity

$$\left\|Z_{N,1}^{(t),a}\right\|_{p} = \left\|\widehat{Z}_{N,p}^{(t),a}\right\|_{p},$$

and from this identity, Lemma 45 and the triangle inequality yield the existence of $k \in \mathbb{R}$ such that

$$\begin{aligned}\left\| Z_{N,1}^{(t),a} \right\|_p \leq & k \left\| \widehat{Z}_{N,p}^{(t),f} \right\|_p + k \left\| \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \widehat{Z}_{N,p}^{(t),f} \right\|_p \\ & + k \left\| \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \left(\frac{1}{N} \sum_{i=1}^{N} \left| y^{(t)} - \mathrm{H}^{(t)} U_i^{(t),f} \right|^p \right)^{1/p} \right\|_p \\ & + \left\| \widehat{\mathrm{K}}_N^{(t),X} W_1^{(t)} - \mathrm{K}^{(t),U} W_1^{(t)} \right\|_p . \end{aligned} \tag{7.36}$$

Using the Cauchy-Schwarz inequality, we have

$$\left\| \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \widehat{Z}_{N,p}^{(t),f} \right\|_p \leq \left\| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right\|_{2p} \left\| \widehat{Z}_{N,p}^{(t),f} \right\|_{2p},$$

and Lemma 46 gives existence of constant $c_1 \in \mathbb{R}$ such that

$$\left\| \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \widehat{Z}_{N,p}^{(t),f} \right\|_p \leq \frac{c_1}{\sqrt{N}} \left\| Z_{N,1}^{(t),f} \right\|_p \leq \frac{c_1}{\sqrt{N}} \frac{k_p^{(t),f}}{\sqrt{N}}. \tag{7.37}$$

If we denote

$$\widehat{D}_{N,p}^{(t)} = \left(\frac{1}{N} \sum_{i=1}^{N} \left| y^{(t)} - \mathrm{H}^{(t)} U_i^{(t),f} \right|^p \right)^{1/p},$$

then, using the Cauchy-Schwarz inequality,

$$\left\| \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \widehat{D}_{N,p}^{(t)} \right\|_p \leq \left\| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right\|_{2p} \left\| \widehat{D}_{N,p}^{(t)} \right\|_{2p},$$

and Lemma 46 together with Lemma 17 yield

$$\left\| \left| \widehat{\mathrm{P}}_N^{(t),f} - \mathrm{C}^{(t),f} \right| \widehat{D}_{N,p}^{(t)} \right\|_p \leq \frac{c_1}{\sqrt{N}} \left\| y^{(t)} - \mathrm{H}^{(t)} U_1^{(t),f} \right\|_{2p}. \tag{7.38}$$

Lemma 50 states that there exists $c_2 \in \mathbb{R}$ such that

$$\left\| \widehat{\mathrm{K}}_N^{(t),X} W_1^{(t)} - \mathrm{K}^{(t),U} W_1^{(t)} \right\|_p \leq \frac{c_2}{\sqrt{N}}, \tag{7.39}$$

and, to finalize the proof, put inequalities (7.35), (7.37), (7.38) and (7.39) into (7.36) to obtain

$$\left\| Z_{N,1}^{(t),a} \right\|_p \leq m \left(\frac{k_p^{(t),f}}{\sqrt{N}} + \frac{c_1 k_p^{(t),f}}{N} + \frac{c_1}{\sqrt{N}} \left\| y^{(t)} - \mathrm{H}^{(t)} U_i^{(t),f} \right\|_{2p} \right) + \frac{c_2}{\sqrt{N}} \leq \frac{k_p^{(t),a}}{\sqrt{N}}$$

where the choice of $k_p^{(t),a}$ is obvious,

$$k_p^{(t),a} = k \left(k_p^{(t),f} + c_1 k_p^{(t),f} + c_1 \left\| y^{(t)} - \mathrm{H}^{(t)} U_i^{(t),f} \right\|_{2p} \right) + c_2.$$

□

The last two lemmas give the main theorem of this chapter.

Theorem 53. *If Assumption 1 holds, then, using notation from Definition 13 and Definition 14, for each $t \in \mathbb{N}$,*

$$\left\| X_{N,1}^{(t),f} - U_1^{(t),f} \right\|_p \to 0,$$
$$\left\| X_{N,1}^{(t),a} - U_1^{(t),a} \right\|_p \to 0$$

for all $p \geq 1$ as N goes to infinity. Additionally,

$$\left\| X_{N,1}^{(t),f} - U_1^{(t),f} \right\|_p = \mathcal{O}\left(N^{-1/2}\right), \tag{7.40}$$
$$\left\| X_{N,1}^{(t),a} - U_1^{(t),a} \right\|_p = \mathcal{O}\left(N^{-1/2}\right). \tag{7.41}$$

Proof. We use an induction to prove the theorem.

Firstly, for $t = 0$, the ensemble members $X_{N,1}^{(0),a}$ and $U_{N,1}^{(0),a}$ identical, so

$$\left\| X_{N,1}^{(0),a} - U_1^{(0),a} \right\|_p = 0$$

for all $p \geq 1$.

Secondly, if equations (7.40) and (7.41) hold for a given $t \in \mathbb{N}$, Lemma 51 and Lemma 52 immediately prove that these equations hold also for $t + 1$. □

7.6 Additional notes and references

Although the ensemble Kalman filter was first published in the early nineties, there was a lack of rigorous probabilistic studies on the convergence of the EnKF until both Mandel et al. [2011] and Le Gland et al. [2011] independently showed that the EnKF converges to the solution obtained using the Kalman filter. Both cited papers show the convergence in $\mathcal{L}^p$ norm for the case when a state space model is finitely dimensional. Almost sure convergence is proved in Le Gland et al. [2011], but no convergence rate is available.

The same results are obtained for a dynamical system with a continuous time in Law et al. [2016]. The paper also provides multiple numerical experiments of different versions of mean field filters, but the paper still assumes that a state space is finitely dimensional.

The convergence of a square root filter, briefly mentioned in Section 5.5, may be found in Kwiatkowski and Mandel [2015], and this paper also provides a proof that the Kalman gain operator is continuous, which is also proved in Section 7.2 for completeness.

8. Spectral diagonal ensemble Kalman filter

As mentioned at the end of Section 5.3, one of the biggest disadvantages of the ensemble Kalman filter is a low rank approximation of a forecast covariance. This obstacle is often suppressed by a localization technique. In this section we presented a modified version of the EnKF that allows a natural localization.

This chapter contains work that has been published in Kasanický et al. [2015], but has more detailed proofs.

The chapter is organized as follows. Section 8.1 proposes another estimate of a covariance and shows that this estimate is better than the sample covariance when samples are taken from a Gaussian distribution. In Section 8.2 this estimate is used to define a new assimilation method called spectral diagonal ensemble Kalman filter (SDEnKF). Section 8.3 shows efficient implementation of the proposed method. Finally, Section 8.4 shows results from multiple experiments using different state space models.

8.1 Spectral diagonal sample covariance

When $X \sim \mathcal{N}(0, \mathrm{P})$, $\mathrm{P} \in \mathbb{C}^{n\times n}$, $n \in \mathbb{N}$, then the Karhunen-Loève expansion, Theorem 1.4.1 in Ash and Gardner [1975], guarantees that there are orthonormal vectors $u_1, \ldots, u_n \in \mathbb{C}^n$ and positive numbers $\lambda_1, \ldots, \lambda_n$ such that

$$X = \sum_{i=1}^{n} \lambda_i^{1/2} \theta_i u_i \tag{8.1}$$

where $\theta_1, \ldots, \theta_n \sim \mathcal{N}(0, 1)$ are i.i.d. random variables. If we denote

$$\mathrm{F}^* = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}, \tag{8.2}$$

i.e., matrix F contains vectors $u_1^*, \ldots, u_n^*$ as rows, then

$$U = \mathrm{F}X = \sum_{i=1}^{n} \lambda_i^{1/2} \theta_i \mathrm{F} u_i = \sum_{i=1}^{n} \lambda_i^{1/2} \theta_i e_i$$

where $e_1, \ldots, e_n$ are unit vectors such that $(e_i)_j = \delta_{i,j},\ i = 1, \ldots, n$. Using Theorem 23, the vector U has $\mathcal{N}(0, \mathrm{D})$ distribution with

$$\mathrm{D} = \mathrm{FPF}^* = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Since u_i, $i = 1, \ldots, n$, are orthonormal,

$$\mathrm{FF}^* = \mathrm{F}^*\mathrm{F} = \mathrm{I}.$$

Now, assume that $X_1, \ldots, X_N$ are i.i.d. samples from $\mathcal{N}(0, \mathrm{P})$ distribution. Each sample may be written in the form

$$X_j = \sum_{i=1}^{n} \lambda_i^{1/2} \theta_{i,j} u_i,$$

where $\theta_{i,j} \sim \mathcal{N}(0,1)$, $i = 1, \ldots, n$, $j = 1, \ldots, N$, are i.i.d. We denote by $U_1, \ldots, U_N$ samples such that

$$U_i = \mathrm{F} X_i, \quad i = 1, \ldots N,$$

where matrix F is defined by (8.2), and we denote $\widehat{\mathrm{P}}_N$ and $\widehat{\mathrm{D}}_N$ the sample covariances of both sets respectively,

$$\widehat{\mathrm{P}}_N = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \overline{X}\right) \left(X_i - \overline{X}\right)^*, \qquad \overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i, \tag{8.3}$$

$$\widehat{\mathrm{D}}_N = \frac{1}{N-1} \sum_{i=1}^{N} \left(U_i - \overline{U}\right) \left(U_i - \overline{U}\right)^*, \qquad \overline{U}_N = \frac{1}{N} \sum_{i=1}^{N} U_i. \tag{8.4}$$

It is obvious that

$$\overline{U}_N = \frac{1}{N} \sum_{i=1}^{N} U_i = \frac{1}{N} \sum_{i=1}^{N} \mathrm{F} X_i = \frac{1}{N} \mathrm{F} \sum_{i=1}^{N} X_i = \mathrm{F} \overline{X}_N$$

and

$$\widehat{\mathrm{D}}_N = \frac{1}{N-1} \sum_{i=1}^{N} \mathrm{F} \left(X_i - \overline{X}\right) \left(\mathrm{F} \left(X_i - \overline{X}\right)\right)^* = \mathrm{F} \widehat{\mathrm{P}}_N \mathrm{C}^*.$$

Using the knowledge of the Karhunen-Loève expansion of the random variable X, we can work with an estimate of the form

$$\widetilde{\mathrm{D}}_N = \left(\widehat{\mathrm{D}}_N \circ \mathrm{I}\right) = \begin{pmatrix} \widehat{\lambda}_1 & & 0 \\ & \ddots & \\ 0 & & \widehat{\lambda}_n \end{pmatrix}, \tag{8.5}$$

and we show that this intuitive estimate has a smaller expected error than the sample covariance. Additionally, we define estimate

$$\widetilde{\mathrm{P}}_N = \mathrm{F}^* \widetilde{\mathrm{D}}_N \mathrm{F}. \tag{8.6}$$

8.1.1 Variance of sample covariance

Using the properties of sample covariance and the Karhunen-Loève expansion, we can evaluate the element of $\widehat{\mathrm{D}}_N$ in the i^{th} row and j^{th} column,

$$\begin{aligned}
\left(\widehat{\mathrm{D}}_N\right)_{i,j} =& \frac{1}{N-1} \sum_{k=1}^{N} \left(U_k - \overline{U}_N\right)_i \left(U_k - \overline{U}_N\right)_j^* \\
=& \frac{1}{N-1} \sum_{k=1}^{N} \left(\lambda_i^{1/2} \theta_{i,k} - \frac{1}{N} \sum_{l=1}^{N} \lambda_i^{1/2} \theta_{i,l}\right) \left(\lambda_j^{1/2} \theta_{j,k} - \frac{1}{N} \sum_{l=1}^{N} \lambda_j^{1/2} \theta_{j,l}\right) \\
=& \frac{(\lambda_i \lambda_j)^{1/2}}{N-1} \left(\sum_{k=1}^{N} \theta_{i,k} \theta_{j,k} - \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \theta_{i,k} \theta_{j,l}\right),
\end{aligned} \tag{8.7}$$

and this expression allows us to formulate the following lemma.

Lemma 54. *The variance the element in the i^{th} row and the j^{th} column of the matrix $\widehat{\mathrm{D}}_N$, defined by* (8.4)*, is*

$$\operatorname{var}\left(\left(\widehat{\mathrm{D}}_N\right)_{i,j}\right) = \begin{cases} \frac{2\lambda_i}{N-1} & \text{if } i = j, \\ \frac{\lambda_i \lambda_j}{N-1} & \text{if } i \neq j. \end{cases}$$

Proof. The sample covariance is an unbiased estimate of a true covariance. Using (8.7), the variance of the diagonal terms of the sample covariance is

$$\begin{aligned} \operatorname{var}\left(\left(\widehat{\mathrm{D}}_N\right)_{i,i}\right) =& \mathrm{E}\left(\frac{\lambda_i}{N-1}\left(\sum_{k=1}^{N}\theta_{i,k}^2 - \frac{1}{N}\sum_{k,l=1}^{N}\theta_{i,k}\theta_{i,l}\right) - \lambda_i\right)^2 \\ =& \frac{\lambda_i^2}{(N-1)^2}\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}^2\right)^2 - \frac{2\lambda_i^2}{N(N-1)^2}\mathrm{E}\left(\sum_{k,l,m=1}^{N}\theta_{i,k}^2\theta_{i,l}\theta_{i,m}\right) \\ &+ \frac{\lambda_i^2}{N^2(N-1)^2}\mathrm{E}\left(\sum_{k,l=1}^{N}\theta_{i,k}\theta_{i,l}\right)^2 - \frac{2\lambda_i^2}{(N-1)}\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}^2\right) \\ &+ \frac{2\lambda_i^2}{N(N-1)}\mathrm{E}\left(\sum_{k,l=1}^{N}\theta_{i,k}\theta_{i,l}\right) + \lambda_i^2. \end{aligned} \tag{8.8}$$

Because $\theta_{k,l}$ are $\mathcal{N}(0,1)$ i.i.d. random variables

$$\mathrm{E}\left(\theta_{i,k}\theta_{i,l}\theta_{i,m}\theta_{i,n}\right) = \begin{cases} 3 & \text{if } k = l = m = n, \\ 1 & \text{if } k = l, m = n, k \neq m, \\ 1 & \text{if } k = m, l = n, k \neq l, \\ 1 & \text{if } k = n, l = m, k \neq l, \\ 0 & \text{otherwise.} \end{cases} \tag{8.9}$$

Hence, we can evaluate all terms on the right side of Equation (8.8):

$$\begin{aligned} \mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}^2\right)^2 =& \sum_{k=1}^{N}\mathrm{E}\theta_{i,k}^4 + \sum_{k,l=1,l\neq k}^{N}\mathrm{E}\left(\theta_{i,k}^2\theta_{i,l}^2\right) \\ =& 3N + N(N-1) = N(N+2), \\ \mathrm{E}\left(\sum_{k,l,m=1}^{N}\theta_{i,k}^2\theta_{i,l}\theta_{i,m}\right) =& \sum_{k=1}^{N}\mathrm{E}\theta_{i,k}^4 + \sum_{k,l=1,l\neq k}^{N}\mathrm{E}\left(\theta_{i,k}^2\theta_{i,l}^2\right) \\ =& N(N+2), \\ \mathrm{E}\left(\sum_{k,l=1}^{N}\theta_{i,k}\theta_{i,l}\right)^2 =& \sum_{k,l,m,n=1}^{N}\mathrm{E}\left(\theta_{i,k}\theta_{i,l}\theta_{i,m}\theta_{i,n}\right) \\ =& \sum_{k=1}^{N}\mathrm{E}\theta_{i,k}^4 + 3\sum_{k,l=1,l\neq k}^{N}\mathrm{E}\left(\theta_{i,k}^2\theta_{i,l}^2\right) = 3N^2, \\ \mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}^2\right) =& \sum_{k=1}^{N}\mathrm{E}\theta_{i,k}^2 = N, \end{aligned}$$

and

$$\mathrm{E}\left(\sum_{k,l=1}^{N}\theta_{i,k}\theta_{i,l}\right)=\sum_{k,l=1}^{N}\mathrm{E}\left(\theta_{i,k}\theta_{i,l}\right)=N.$$

Therefore, for any $i=1,\ldots,n$,

$$\begin{aligned}\mathrm{var}\left(\left(\widehat{\mathrm{D}}_N\right)_{i,i}\right)=&\frac{\lambda_i^2}{(N-1)}\left(\frac{N\left(N+2\right)}{(N-1)}-\frac{2\left(N+2\right)}{(N-1)}+3-2N+2+1\right)\\=&\frac{2\lambda_i^2}{(N-1)}.\end{aligned}$$

The variance of the off diagonal element is

$$\begin{aligned}\mathrm{var}\left(\left(\widehat{\mathrm{D}}_N\right)_{i.j}\right)=&\mathrm{E}\left(\frac{\left(\lambda_i\lambda_j\right)^{1/2}}{N-1}\left(\sum_{k=1}^{N}\theta_{i,k}\theta_{j,k}-\frac{1}{N}\sum_{k,l=1}^{N}\theta_{i,k}\theta_{j,l}\right)\right)^2\\=&\frac{\lambda_i\lambda_j}{\left(N-1\right)^2}\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}\theta_{j,k}\right)^2\\&-\frac{2\lambda_i\lambda_j}{N\left(N-1\right)^2}\mathrm{E}\left(\sum_{k,l,m=1}^{N}\theta_{i,k}\theta_{j,k}\theta_{i,l}\theta_{j,m}\right)\\&+\frac{\lambda_i\lambda_j}{N^2\left(N-1\right)^2}\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}\theta_{j,k}\right)^2,\end{aligned}$$

and, using (8.9), we can evaluate each summand on the right side of the last equation:

$$\begin{aligned}\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}\theta_{j,k}\right)^2=&\sum_{k,l=1}^{N}\mathrm{E}\left(\theta_{i,k}\theta_{j,k}\theta_{i,l}\theta_{j,l}\right)=\\=&\sum_{k,l=1}^{N}\mathrm{E}\left(\theta_{i,k}\theta_{i,l}\right)\mathrm{E}\left(\theta_{j,k}\theta_{j,l}\right)=N,\\\mathrm{E}\left(\sum_{k,l,m=1}^{N}\theta_{i,k}\theta_{j,k}\theta_{i,l}\theta_{j,m}\right)=&\sum_{k,l,m=1}^{N}\mathrm{E}\left(\theta_{i,k}\theta_{i,l}\right)\mathrm{E}\left(\theta_{j,k}\theta_{j,m}\right)\\=&\sum_{k=1}^{N}\mathrm{E}\left(\theta_{i,k}^2\right)\mathrm{E}\left(\theta_{i,k}^2\right)=N,\\\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}\theta_{j,k}\right)^2=&\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}\sum_{k=1}^{N}\theta_{j,k}\right)^2\\=&\mathrm{E}\left(\sum_{k=1}^{N}\theta_{i,k}\right)^2\mathrm{E}\left(\sum_{k=1}^{N}\theta_{j,k}\right)^2=N^2.\end{aligned}$$

Hence, for any $i,j=1,\ldots,n$ such that $i\neq j$,

$$\mathrm{var}\left(\left(\widehat{\mathrm{D}}_N\right)_{i,j}\right)=\frac{\lambda_i\lambda_j}{\left(N-1\right)^2}\left(N-1+1\right)=\frac{\lambda_i\lambda_j}{N-1}.$$

□

8.1.2 Error estimates

The error of a covariance estimate is measured using the Hilbert-Schmidt norm of its difference from the covariance. Recall that the Hilbert-Schmidt norm of an bounded linear operator from $\mathbb{C}^n$ to $\mathbb{C}^n$, i.e., a $n \times n$ complex matrix, is also called a Frobenius norm, and it can be shown that

$$|\mathrm{A}|_{\mathrm{HS}} = \sqrt{\sum_{i,j=1}^{n} \left|(\mathrm{A})_{i,j}\right|^2}. \tag{8.10}$$

The following lemma shows that this norm is invariant under unitary transformations.

Lemma 55. *Assume that* $\mathrm{A} \in \mathbb{C}^{n\times n}$, *and* $\mathrm{F} \in \mathbb{C}^{n\times n}$ *is unitary, i.e.,*

$$\mathrm{FF}^* = \mathrm{F}^*\mathrm{F} = \mathrm{I}.$$

Then,

$$|\mathrm{A}|_{\mathrm{HS}} = |\mathrm{FAF}^*|_{\mathrm{HS}}\,.$$

Proof. Denote a_j the the j^{th} column of the matrix A. Then,

$$|\mathrm{A}|_{\mathrm{HS}}^2 = \sum_{j=1}^{n} |a_j|^2\,,$$

where $|a_j|$ denotes the standard Euclidean norm of the vector a_j, and

$$|\mathrm{FA}|_{\mathrm{HS}}^2 = \sum_{j=1}^{n} |\mathrm{F}a_j|^2 = |\mathrm{FA}|_{\mathrm{HS}}^2\,.$$

Obviously, using (8.10),

$$|\mathrm{A}|_{\mathrm{HS}} = |\mathrm{A}^*|_{\mathrm{HS}}\,,$$

and this simple observation concludes the proof because

$$|\mathrm{A}|_{\mathrm{HS}} = |\mathrm{FA}|_{\mathrm{HS}} = |\mathrm{F}(\mathrm{FA})^*|_{\mathrm{HS}} = |\mathrm{FAF}^*|_{\mathrm{HS}}\,.$$

□

The previous lemma has an important corollary.

Corollary 7. The following identities hold:

$$\begin{aligned}
\mathrm{E}\left|\mathrm{P} - \widetilde{\mathrm{P}}_N\right|_{\mathrm{HS}}^2 &= \mathrm{E}\left|\mathrm{D} - \widetilde{\mathrm{D}}_N\right|_{\mathrm{HS}}^2, \\
\mathrm{E}\left|\mathrm{P} - \widehat{\mathrm{P}}_N\right|_{\mathrm{HS}}^2 &= \mathrm{E}\left|\mathrm{D} - \widehat{\mathrm{D}}_N\right|_{\mathrm{HS}}^2.
\end{aligned}$$

Proof. Both identities follows directly from Lemma 55 because

$$\mathrm{P} - \widetilde{\mathrm{P}}_N = \mathrm{F}^*\left(\mathrm{D} - \widetilde{\mathrm{D}}_N\right)\mathrm{F}$$

and

$$\mathrm{P} - \widehat{\mathrm{P}}_N = \mathrm{F}^*\left(\mathrm{D} - \widehat{\mathrm{D}}_N\right)\mathrm{F}.$$

□

Now, we may establish the main statistical results of the chapter.

Theorem 56. *Assume that $X_1, \dots, X_N$ are i.i.d. samples from distribution $\mathcal{N}(0, \mathrm{P})$, and matrices $\widehat{\mathrm{P}}_N$ and $\widetilde{\mathrm{P}}_N$ are defined by* (8.3) *and* (8.6) *respectively. Then,*

$$\mathrm{E}\left|\mathrm{P}-\widetilde{\mathrm{P}}_N\right|_{\mathrm{HS}}^2=\frac{2}{N-1}\left|\mathrm{P}\right|_{\mathrm{HS}}^2=\frac{2}{N-1}\sum_{i=1}^n\lambda_i^2$$

and

$$\mathrm{E}\left|\mathrm{P}-\widehat{\mathrm{P}}_N\right|_{\mathrm{HS}}^2=\frac{1}{N-1}\left(\sum_{i=1}^n\lambda_i^2+\sum_{i,j=1}^n\lambda_i\lambda_j\right)$$

where $\lambda_1, \dots, \lambda_n$ are eigenvalues of the matrix P.

Proof. Using Corollary 7, we immediately obtain identities

$$\mathrm{E}\left|\mathrm{P}-\widetilde{\mathrm{P}}_N\right|_{\mathrm{HS}}^2=\mathrm{E}\left|\mathrm{D}-\widetilde{\mathrm{D}}_N\right|_{\mathrm{HS}}^2 \tag{8.11}$$

and

$$\mathrm{E}\left|\mathrm{P}-\widehat{\mathrm{P}}_N\right|_{\mathrm{HS}}^2=\mathrm{E}\left|\mathrm{D}-\widehat{\mathrm{D}}_N\right|_{\mathrm{HS}}^2 \tag{8.12}$$

where matrices $\widehat{\mathrm{D}}_N$ and $\widetilde{\mathrm{D}}_N$ are defined by (8.4) and (8.5) respectively.

Both D and $\widetilde{\mathrm{D}}_N$ are diagonal matrices, so

$$\mathrm{E}\left|\mathrm{D}-\widetilde{\mathrm{D}}_N\right|_{\mathrm{HS}}^2=\mathrm{E}\sum_{i=1}^n\left|(\mathrm{D})_{i,i}-\left(\widetilde{\mathrm{D}}_N\right)_{i,i}\right|^2,$$

and because

$$\mathrm{E}\left(\widetilde{\mathrm{D}}_N\right)_{i,j}=(\mathrm{D})_{i,j}$$

for all $i, j = 1, \dots, n$,

$$\mathrm{E}\left|\mathrm{D}-\widetilde{\mathrm{D}}_N\right|_{\mathrm{HS}}^2=\sum_{i=1}^n\mathrm{var}\left(\left(\widetilde{\mathrm{D}}_N\right)_{i,i}\right).$$

Now, using Lemma 54 together with (8.11) finish the proof of the first statement in the theorem.

Because $\widehat{\mathrm{D}}_N$ is also an unbiased estimate of D we can similarly show that

$$\mathrm{E}\left|\mathrm{D}-\widehat{\mathrm{D}}_N\right|_{\mathrm{HS}}^2=\sum_{i,j=1}^n\mathrm{var}\left(\widehat{\mathrm{D}}_N\right)_{i,j},$$

and, using Lemma 54,

$$\mathrm{E}\left|\mathrm{D}-\widehat{\mathrm{D}}_N\right|_{\mathrm{HS}}^2=\frac{2}{N-1}\sum_{i=1}^n\lambda_i^2+\frac{1}{N-1}\sum_{i,j=1,i\neq j}^n\lambda_i\lambda_j.$$

The last identity together with (8.12) concludes the proof. □

Theorem 57. *Using the same assumptions and notation as in the previous theorem,*

$$\mathrm{E}\left|\mathrm{P}-\widetilde{\mathrm{P}}_N\right|_{\mathrm{HS}}^2<\mathrm{E}\left|\mathrm{P}-\widehat{\mathrm{P}}_N\right|_{\mathrm{HS}}^2.$$

Proof. Using the previous theorem and 8.10,

$$\begin{aligned}\left|\mathrm{P}-\widehat{\mathrm{P}}_N\right|_{\mathrm{HS}}^2=&\frac{2}{N-1}\sum_{i=1}^{n}\lambda_i^2+\frac{1}{N-1}\sum_{i,j=1,i\neq j}^{n}\lambda_i\lambda_j\\=&\frac{2}{N-1}\left|\mathrm{P}\right|_{\mathrm{HS}}^2+\frac{1}{N-1}\sum_{i,j=1,i\neq j}^{n}\lambda_i\lambda_j,\end{aligned}$$

and the second term is always positive because $\lambda_1,\ldots,\lambda_n$ are eigenvalues of the symmetric positive definite operator C, i.e., they are all larger than zero. □

We illustrate the previous theorem in the next example.

Example 24. Assume that $X\in\mathcal{L}^2\left(\mathbb{R}^{64}\right)$ is a random vector with covariance

$$\mathrm{P}=\mathrm{cov}\left(X\right)=\mathrm{F}^*\mathrm{DF}$$

where D is a diagonal matrix with elements $\lambda_1,\ldots,\lambda_{64}$ on the main diagonal, and F is a real matrix which correspond to the discrete sine transformation (DST). The DST is equivalent to the imaginary part of the discrete Fourier transformation (Martucci [1994]), and for $n=64$, the matrix F consists of elements

$$\left(\mathrm{F}\right)_{k,l}=\frac{2}{64+1}\sin\left(\pi\frac{kl}{64+1}\right),\quad k,l=1,\ldots,64.$$

We define

$$\lambda_k=\exp\left(-k\right),\quad k=1,\ldots,64$$

and the covariance P is illustrated in Figure 8.1. Now, if we generate four samples $X_1,\ldots,X_4$ from the distribution of the random vector X, then a sample covariance $\widehat{\mathrm{P}}_4$ should have a larger error than the spectral diagonal sample covariance

$$\widetilde{\mathrm{P}}_4=\mathrm{F}^*\left(\mathrm{F}\widehat{\mathrm{P}}_4\mathrm{F}^*\circ\mathrm{I}\right)\mathrm{F}.$$

Figure 8.2 shows that the sample covariance using only four samples is a very poor estimate of the true covariance P. In fact, the sample covariance does not even catch even the basic shape of P, and, additionally, the maximal values in the upper left corner are 1.5 higher than the true values. On the other hand, the matrix $\widetilde{\mathrm{P}}_4$ looks really similar to the true covariance P, and we see that even such small sample can capture the basic shape of covariance.

8.1.3 Spectral transformations

The main results in the previous subsection state that spectral diagonal sample covariance always has a smaller error than a sample covariance. However, when one wants to use this estimate, one needs to know how to pick the transformation matrix, and this choice is strongly related to the properties of the true covariance

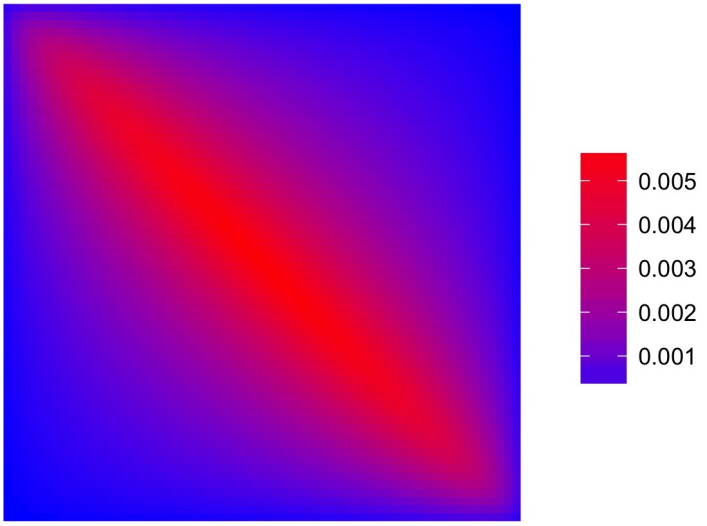

Figure 8.1: True covariance of the random vector X from Example 24; size of random vector is 64.

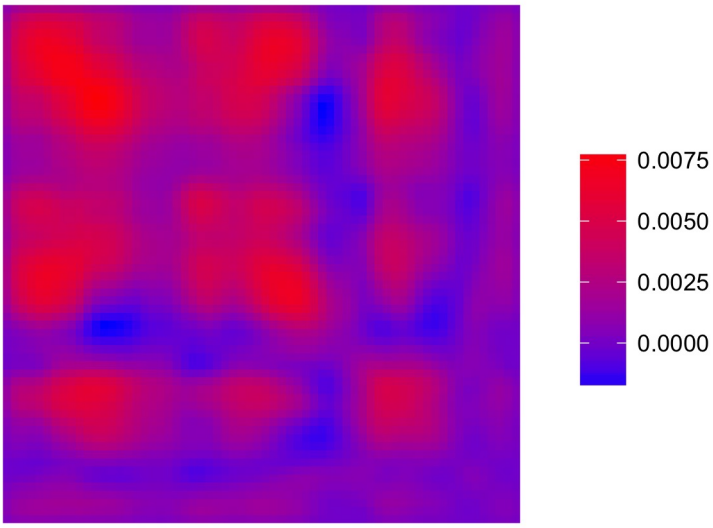

Figure 8.2: Sample covariance using four samples from Example 24; size of random vector is 64.

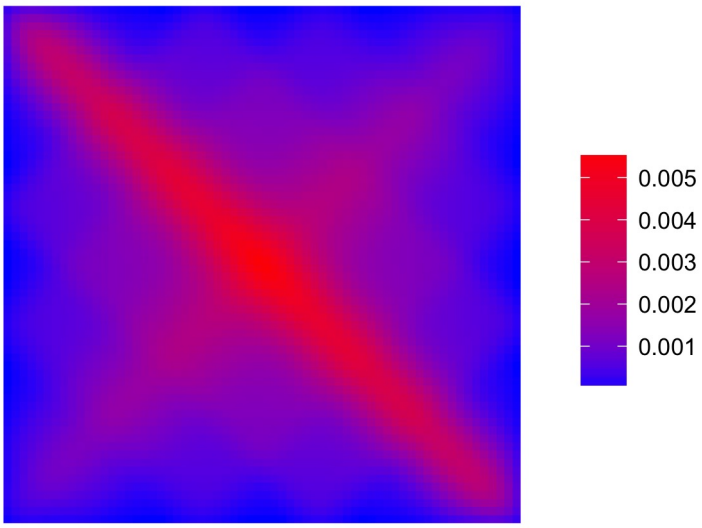

Figure 8.3: Spectral diagonal sample covariance using four samples and DST from Example 24; size of random vector is 64.

of X. In the real world applications, the random vector $X \in \mathcal{L}(\mathbb{C}^n)$, which represent one state of a used state space model, is usually a discretization of a continuous d dimensional random field defined on a discrete mesh with n grid points. In the next example we show how the Karhunen-Loève decomposition of X is related to the shape of the covariance of X. The following two examples are motivated by Pannekoucke et al. [2007], but they are presented with greater details.

Example 25. Assume that $X \in \mathcal{L}(\mathbb{C}^n)$ is a discretization of a real random function defined on a unite circle, and a mesh, on which the function is discretized, is equidistant. For each $k = 1, \ldots, n$, we define a local covariance function

$$\Gamma^k(l) = \operatorname{cov}\left((X)_k, (X)_{k+l}\right) = (\operatorname{cov}(X))_{k,k+l}, \quad l = -n, \ldots, n, \tag{8.13}$$

where we periodically extend the vector X, i.e., we set

$$\begin{aligned} (X)_{-k} &= (X)_{n-k+1}, & k &= 1, \ldots n, \\ (X)_{n+k} &= (X)_k, & k &= 1, \ldots n. \end{aligned}$$

We already know that, using the Karhunen-Loève decomposition, the vector X may be written in the form

$$X = \sum_{i=1}^{n} \lambda_i^{1/2} \theta_i u_i = \sum_{i=1}^{n} \langle X, u_i \rangle u_i.$$

Using this expansion,

$$\begin{aligned}\mathrm{P} =&\mathrm{cov}\,(X) = \mathrm{cov}\left(\sum_{i=1}^{n}\left(\langle X, u_i\rangle\, u_i\right)\right)\\ =&\mathrm{E}\left(\left(\sum_{i=1}^{n}\langle X, u_i\rangle\, u_i\right)\left(\sum_{i=1}^{n}\langle X, u_i\rangle\, u_i\right)^{*}\right)\\ =&\sum_{i,j=1}^{n}\mathrm{E}\left(\langle X, u_i\rangle\,\overline{\langle X, u_j\rangle}\right)u_i u_j^{*} = \sum_{i,j=1}^{n}(\mathrm{D})_{i,j}\,u_i u_j^{*}\end{aligned}$$

where D is the covariance of

$$U = \mathrm{F}X = \begin{pmatrix} u_1^{*}X \\ \vdots \\ u_n^{*}X \end{pmatrix} = \begin{pmatrix} \langle X, u_1\rangle \\ \vdots \\ \langle X, u_n\rangle \end{pmatrix},$$

and we know that the matrix D is diagonal,

$$\mathrm{D} = \mathrm{cov}\,(U) = \mathrm{E}\,(UU^{*}) = \mathrm{FPF}^{*}.$$

Therefore,

$$\mathrm{P} = \sum_{i=1}^{n}\left((\mathrm{D})_{i,i}\,u_i u_i^{*}\right), \tag{8.14}$$

and for its element in the k^{th} row and the l^{th} row follows

$$(\mathrm{P})_{k,l} = \sum_{i=1}^{n}\left(\mathrm{E}\,|\langle X, u_i\rangle|^2\,(u_i)_k\,\overline{(u_i)}_l\right). \tag{8.15}$$

Using the definition of the inner product in $\mathbb{C}^n$,

$$\begin{aligned}\mathrm{E}\,|\langle X, u_i\rangle|^2 =&\mathrm{E}\left(\langle X, u_i\rangle\,\overline{\langle X, u_i\rangle}\right)\\ =&\mathrm{E}\left(\left(\sum_{k=1}^{n}(X)_k\,\overline{(u_i)}_k\right)\overline{\left(\sum_{q=1}^{n}(X)_q\,\overline{(u_i)}_q\right)}\right)\\ =&\sum_{k=1}^{n}\sum_{q=1}^{n}\mathrm{E}\left((X)_k\,\overline{(X)}_q\right)\overline{(u_i)}_k\,(u_i)_q\\ =&\sum_{k=1}^{n}\sum_{l=-k+1}^{n-k}\mathrm{E}\left((X)_k\,\overline{(X)}_{k+l}\right)\overline{(u_i)}_k\,(u_i)_{k+l}.\end{aligned} \tag{8.16}$$

Applying (8.14), (8.15) and (8.16) into the definition of the local covariance we obtain

$$\begin{aligned}\Gamma^{k}(l) =&\sum_{i=1}^{n}\sum_{\widetilde{k}=1}^{n}\sum_{\widetilde{l}=-\widetilde{k}+1}^{n-\widetilde{k}}\mathrm{E}\left((X)_{\widetilde{k}}\,\overline{(X)}_{\widetilde{k}+\widetilde{l}}\right)\overline{(u_i)}_{\widetilde{k}}\,(u_i)_{\widetilde{k}+\widetilde{l}}(u_i)_k\,\overline{(u_i)}_{k+l}\\ =&\sum_{\widetilde{k}=1}^{n}\sum_{\widetilde{l}=-\widetilde{k}+1}^{n-\widetilde{k}}\Gamma^{\widetilde{k}}\left(\widetilde{l}\right)\Phi^{k,l}\left(\widetilde{k},\widetilde{l}\right)\end{aligned} \tag{8.17}$$

where we define

$$\Phi^{k,l}\left(\widetilde{k},\widetilde{l}\right)=\sum_{i=1}^{n}\overline{(u_i)_{\widetilde{k}}}\,(u_i)_{\widetilde{k}+\widetilde{l}}(u_i)_k\,\overline{(u_i)}_{k+l}.$$

The previous example shows that a covariance between any two elements of random vector X is a weighted sum of covariance between any two elements with the same relative positions. The next example uses Fourier vectors in place of $u_1,\ldots,u_n$.

Example 26. Using the notation from Example 25, assume that

$$u_k=\begin{pmatrix}\frac{1}{\sqrt{n}}\exp\left(\frac{-ik2\pi 0}{n}\right)\\ \vdots\\ \frac{1}{\sqrt{n}}\exp\left(\frac{-ik2\pi(n-1)}{n}\right)\end{pmatrix},\ k=1,\ldots,n,$$

i.e., $u_1,\ldots,u_n$ are discrete Fourier basis vectors. Using (8.17),

$$\Gamma^k(l)=\sum_{\widetilde{k}=1}^{n}\sum_{\widetilde{l}=-\widetilde{k}+1}^{n-\widetilde{k}}\Gamma^{\widetilde{k}}\left(\widetilde{l}\right)\Phi^{k,l}\left(\widetilde{k},\widetilde{l}\right)$$

with

$$\begin{aligned}\Phi^{k,l}\left(\widetilde{k},\widetilde{l}\right)=&\sum_{i=1}^{n}\overline{(u_i)_{\widetilde{k}}}\,(u_i)_{\widetilde{k}+\widetilde{l}}(u_i)_k\,\overline{(u_i)}_{k+l}.\\ =&\frac{1}{n^2}\sum_{k=1}^{n}\exp\left(ik2\pi\frac{\widetilde{k}-\left(\widetilde{k}+\widetilde{l}\right)-k+k+l}{n}\right).\end{aligned}$$

The function

$$y\in\mathbb{R}\to\exp(iy)$$

is periodic, so

$$\Phi^{k,l}\left(\widetilde{k},\widetilde{l}\right)=\frac{1}{n}\delta_{k,\widetilde{k}}.$$

Therefore,

$$\begin{aligned}(\mathrm{P})_{k,k+l}=&\Gamma^k(l)=\frac{1}{N}\sum_{\widetilde{k}=1}^{n}\sum_{\widetilde{l}=-\widetilde{k}+1}^{n-\widetilde{k}}\Gamma^{\widetilde{k}}\left(\widetilde{l}\right)\delta_{l,\widetilde{l}}\\ =&\frac{1}{N}\sum_{\widetilde{k}=1}^{n}\Gamma^{\widetilde{k}}(l).\end{aligned}$$

and we see that the covariance of X is a function of relative positions between the nodes of the mesh.

Last two examples may be extended to a case when X is a discretization of a multidimensional random field, and the similar results hold [Pannekoucke et al., 2007, Appendix A]. It may even be shown that the matrix F contains discrete Fourier basis vectors if and only if X is second order stationary random field, i.e., the covariance between values of the field in two different locations is a function of their relative positions.

8.2 Spectral diagonal EnKF

A natural idea is to use the estimate introduced in the previous section in place of a sample covariance in the EnKF update equation. This is the main idea of the spectral diagonal ensemble Kalman filter.

In this section we use the state space model and the notation from Definition 7 with additional assumptions that the spate space is $\mathbb{R}^n$, and the observation space is $\mathbb{R}^m$. Additionally, we assume that there is a matrix $\mathrm{F} \in \mathbb{R}^{n\times n}$ with orthonormal columns, i.e., $\mathrm{FF}^* = \mathrm{I}$ and $\mathrm{F}^*\mathrm{F} = \mathrm{I}$, such that

$$\operatorname{cov}\left(\mathrm{F}X^{(t)}\right) = \mathrm{D}^{(t)},\ t \in \mathbb{N}_0, \tag{8.18}$$

where $\mathrm{D}^{(t)}$, $t \in \mathbb{N}$, are diagonal matrices.

Definition 15 (Spectral diagonal EnKF)**.** *Given $N \geq 2$ the spectral diagonal EnKF consists of the following steps.*

1. *Initialization. Generate N independent samples from the distribution of the initial condition:*

$$X_1^{(0),a}, \ldots, X_N^{(0),a} \sim \mathcal{N}\left(m^{(0)}, \mathrm{P}^{(0)}\right).$$

2. *For $t = 1, 2, \ldots$, repeat the following steps.*

   (a) *Forecast step. Advance the analysis ensemble from the previous cycle:*

$$X_i^{(t),f} = \Psi\left(X_i^{(t-1),a}\right) + V_i^{(t)}, \quad i = 1, \ldots, N$$

   where

$$V_1^{(t)}, \ldots, V_N^{(t)} \sim \mathcal{N}\left(0, \mathrm{Q}^{(t)}\right)$$

   are independently generated random variables.

   (b) *Transform the ensemble into the spectral space:*

$$U_i^{(t),f} = \mathrm{F}X_i^{(t),f},\ i = 1, \ldots, N.$$

   (c) *Evaluate the sample statistics of the spectral ensemble:*

$$\begin{aligned}
\overline{U}_N^{(t),f} =& \frac{1}{N}\sum_{i=1}^{N} U_i^{(t),f}, \\
\widehat{\mathrm{D}}_N^{(t),f} =& \frac{1}{N-1}\sum_{i=1}^{N}\left(U_{N,i}^{(t),f} - \overline{U}_N^{(t),f}\right)\left(U_{N,i}^{(t),f} - \overline{U}_N^{(t),f}\right)^*.
\end{aligned}$$

   (d) *Delete all off diagonal elements of spectral the sample covariance:*

$$\widetilde{\mathrm{D}}_N^{(t),f} = \widehat{\mathrm{D}}_N^{(t),f} \circ \mathrm{I}. \tag{8.19}$$

   (e) *Transform this new estimate back to the original space:*

$$\widetilde{\mathrm{P}}_N^{(t),f} = \mathrm{F}^*\widetilde{\mathrm{D}}_N^{(t),f}\mathrm{F}. \tag{8.20}$$

(f) Use this estimate instead of the sample covariance in the update equation:

$$X_i^{(t),a} = X_i^{(t),f} + \widetilde{\mathrm{P}}_N^{(t),f}\mathrm{H}^*\left(\mathrm{H}\widetilde{\mathrm{P}}_N^{(t),f}\mathrm{H}^* + \mathrm{R}^{(t)}\right)^{-1}\left(Y_i^{(t)} - \mathrm{H}X_i^{(t),f}\right), \tag{8.21}$$

$i = 1, \ldots, N$ where

$$Y_i^{(t)} = Y^{(t)} + W_i^{(t)}$$

and

$$W_1^{(t)}, \ldots, W_N^{(t)} \sim \mathcal{N}\left(0, \mathrm{R}^{(t)}\right)$$

are independently generated random variables.

Using the results from the previous section, one should use the proposed method every time when the matrix F exists. However, even when the matrix is not known one may use some of the well known spectral transformations such as Fourier or wavelet transformation (Daubechies [1992], Strang and Nguyen [1996]).

8.3 Efficient implementation

One of the biggest advantages of the SDEnKF is that it can be implemented very efficiently in many cases even if a dynamical system has a huge dimension, and we show some implementations in this section. Through this section we use the same notation as in Definition 15, but since this section is devoted to efficient implementation of one update, we do not use the time index t.

8.3.1 One variable, completely observed

When the state of the system only consists of one variable that is completely observed, i.e., the observation operator is identity, and the observation error is

$$\mathrm{R} = c\mathrm{I},\ \mathrm{c} \in \mathbb{C}, \tag{8.22}$$

then the update equation, (8.21) simplifies to

$$X_i^a = X_i^{(t),f} + \widetilde{\mathrm{P}}_N^f\left(\widetilde{\mathrm{P}}_N^f + \mathrm{R}\right)^{-1}\left(Y_i - X_i^f\right),\ i = 1, \ldots N.$$

Recall that by the definition, Equation (8.20),

$$\widetilde{\mathrm{P}}_N^f = \mathrm{F}^*\widetilde{\mathrm{D}}_N^f\mathrm{F},$$

and, using (8.22),

$$\begin{aligned}\left(\widetilde{\mathrm{P}}_N^f + \mathrm{R}\right)^{-1} &= \left(\mathrm{F}^*\widetilde{\mathrm{D}}_N^f\mathrm{F} + c\mathrm{F}^*\mathrm{IF}\right)^{-1}\\ &= \mathrm{F}^*\left(\widetilde{\mathrm{D}}_N^f + c\mathrm{I}\right)^{-1}\mathrm{F}\\ &= \mathrm{F}^*\left(\widetilde{\mathrm{D}}_N^f + \mathrm{R}\right)^{-1}\mathrm{F}.\end{aligned}$$

Applying the last identity into the update equation gives for each $i = 1, \dots n$,

$$\begin{aligned} X_i^a =& X_i^f + \mathrm{F}^* \widetilde{\mathrm{D}}_N^f \mathrm{F} \mathrm{F}^* \left(\widetilde{\mathrm{D}}_N^f + \mathrm{R} \right)^{-1} \mathrm{F} \left(Y_i - X_i^f \right) \\ =& X_i^{(t),f} + \mathrm{F}^* \widetilde{\mathrm{D}}_N^{(t),f} \left(\widetilde{\mathrm{D}}_N^f + \mathrm{R} \right)^{-1} \mathrm{F} \left(Y_i - X_i^f \right). \end{aligned}$$

The matrices $\widetilde{\mathrm{D}}_N^{(t),f}$, $\mathrm{R}^{(t)}$ and $\left(\widetilde{\mathrm{D}}_N^{(t),f} + \mathrm{R}^{(t)} \right)$ are diagonal, so any operation, e.g., multiplication or inversion, is extremely fast. Additionally, multiplication by the transformation matrix F can be usually efficiently done using algorithms such as a fast Fourier transform or a discrete wavelet transform (Strang and Nguyen [1996]).

8.3.2 Multiple variables, one completely observed

Another example is when the state consists of m variables measured on the same mesh,

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^m \end{pmatrix}$$

where $X^k \in \mathcal{L}\left(\mathbb{C}^d\right)$, $k = 1, \dots, m$, $d \in \mathbb{N}$. In this case we apply the spectral transformation separately on each variable, so the transformation matrix has block structure,

$$\mathrm{F} = \left.\begin{pmatrix} \widetilde{\mathrm{F}} & & 0 \\ & \ddots & \\ 0 & & \widetilde{\mathrm{F}} \end{pmatrix}\right\} m \text{ times},$$

where $\widetilde{\mathrm{F}} \in \mathbb{C}^{d \times d}$, and we replace the sample covariance in the EnKF update equation by the block diagonal matrix

$$\widetilde{\mathrm{D}}_N^f = \begin{pmatrix} \widetilde{\mathrm{D}}_N^{1,1} & \cdots & \widetilde{\mathrm{D}}_N^{1,N} \\ \vdots & \ddots & \vdots \\ \widetilde{\mathrm{D}}_N^{N,1} & \cdots & \widetilde{\mathrm{D}}_N^{N,N} \end{pmatrix}$$

where for each $i, j \in \{1, \dots, m\}$,

$$\widetilde{\mathrm{D}}_N^{i,j} = \widehat{\mathrm{D}}_N^{i,j} \circ \mathrm{I},$$

and $\widehat{\mathrm{D}}_N^{i,j}$ is the sample covariance between the i^{th} and j^{th} variable in the spectral space. Now, if we assume that we observe the first variable completely, then the observation matrix $\mathrm{H} \in \mathbb{R}^{d \times n}$ has also special form

$$\mathrm{H} = \begin{pmatrix} \mathrm{I} & 0 & \cdots & 0 \end{pmatrix}$$

where I is $d \times d$ identity matrix. Obviously,

$$\mathrm{HF}^* = \begin{pmatrix} \widetilde{\mathrm{F}}^* & 0 & \cdots & 0 \end{pmatrix},$$

and thus

$$\mathrm{H}\mathrm{F}^*\widetilde{\mathrm{D}}_N^f\mathrm{F}\mathrm{H}^* = \widetilde{\mathrm{F}}^*\widetilde{\mathrm{D}}_N^{1,1}\widetilde{\mathrm{F}}.$$

Therefore , similar to the previous case, if

$$\mathrm{R} = c\mathrm{I},\ \mathrm{c} \in \mathbb{C},$$

then the update equation can be written in the form

$$\begin{aligned} X_i^a =& X_i^f + \mathrm{F}^*\widetilde{\mathrm{D}}_N^f\mathrm{F}\mathrm{H}^*\left(\widetilde{\mathrm{F}}^*\widetilde{\mathrm{D}}_N^{1,1}\widetilde{\mathrm{F}} + \mathrm{R}\right)^{-1}\left(Y_i - \mathrm{H}X_i^f\right) \\ =& X_i^f + \mathrm{F}^*\begin{pmatrix} \widetilde{\mathrm{D}}_N^{1,1}\widetilde{\mathrm{F}}\widetilde{\mathrm{F}}^* \\ \vdots \\ \widetilde{\mathrm{D}}_N^{N,1}\widetilde{\mathrm{F}}\widetilde{\mathrm{F}}^* \end{pmatrix}\left(\widetilde{\mathrm{D}}_N^{1,1} + \mathrm{R}\right)^{-1}\widetilde{\mathrm{F}}\left(Y_i - \mathrm{H}X_i^f\right), \end{aligned}$$

and hence

$$X_i^a = X_i^f + \mathrm{F}^*\begin{pmatrix} \widetilde{\mathrm{D}}_N^{1,1} \\ \vdots \\ \widetilde{\mathrm{D}}_N^{N,1} \end{pmatrix}\left(\widetilde{\mathrm{D}}_N^{1,1} + \mathrm{R}\right)^{-1}\widetilde{\mathrm{F}}\left(Y_i - \mathrm{H}X_i^f\right).$$

for $i = 1, \ldots, N$. Again, all matrices in the final update equation are diagonal, so any operations are very efficient.

Even if this case looks trivial, it is quite common in real applications. For example, when modeling the state of the atmosphere and satellite images of the whole domain are available.

8.3.3 Small size of observations

One often finds out that the size of the observation space, i.e. the number of observations assimilated in one cycle, is not so huge, and directly the SDEnKF may be used. The inversion needed to evaluate the Kalman gain may be computed using some software library for a matrix computation on parallel distributed memory machines such as ScaLAPACK (Blackford et al. [1997]).

8.4 Computational experiments

Theorem 8.1.3 states that a sample spectral diagonal covariance is a better estimate than a classical sample covariance. However, in real world applications, a better approximation of a forecast covariance does not always provide a better analysis. Therefore, we test the SDEnKF algorithm using multiple chaotic models.

We perform a usual twin model experiment with two toy models: Lorenz 96 and shallow water equation. We use one trajectory of the model as a truth $\left\{X^{(t)}\right\}$, and assimilate observations derived from this trajectory to an independently initialized ensemble. Hence, we have full control over the experiment, and can evaluate whether an assimilation method pushes the ensemble towards the truth states, i.e, where the method decrease the error of the forecast. We use a

root mean square error (RMSE) of an ensemble mean to measure the error of an ensemble. For a given $t \in \mathbb{N}$ and a given value of a state $X^{(t)} = x^{(t)} \in \mathbb{R}^n$,

$$\mathrm{RMSE} = \frac{1}{n} \sum_{i=1}^{n} \left| \left(x^{(t)}\right)_i - \left(\overline{x}_N^{(t),\bullet}\right)_i \right|^2,$$

where $\overline{X}_N^{(t),\bullet} = \overline{x}_N^{(t),\bullet}$ is the value of the ensemble mean

$$\overline{X}_N^{(t),\bullet} = \sum_{i=1}^{N} X_i^{(t),\bullet},$$

and $\bullet$ stands either for forecast or analysis.

We use 3 different spectral transformations:

1. the discrete sine transform (DST), which can be derived taking an imaginary part of the discrete Fourier transform,
2. the discrete cosine transform (DCT), which can be derived taking a real part of the discrete Fourier transform, and
3. the discrete wavelet transform (DWT) with Coiflet (2,4) wavelet basis functions (Daubechies [1992]).

We compare these methods with an analysis obtained by the standard EnKF, and with an error of an ensemble without no assimilation, marked as free run.

8.4.1 Lorenz 96

The Lorenz 96 model (Lorenz [2006]) is a very popular toy model in a data assimilation area, as the dynamical system produced by this model is very chaotic and still easy to compute. A state of the model evolves in time, and its evolution during 0.05 time units roughly corresponds to the evolution of a real climatological model during 6 hours.

The state of the model is the vector

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix} \in \mathbb{R}^K$$

and its evolution is governed by the differential equation

$$\frac{dx_j}{dt} = x_{j-1}x_{j+1} - x_{j-1}x_{j-2} - x_j + F, \quad j = 1, \ldots, K, \tag{8.23}$$

where by definition

$$x_{j-K} = x_{j+K} = x_j, \quad j = 1, \ldots, K,$$

and the parameter $F \in \mathbb{R}$ controls the chaotic behavior of the model. It can be shown that when $F = 8$, the model is strongly chaotic.

Our experiments' setup follows the one used by Lorenz and Emanuel [1998]. We set $F = 8$, and sample initial conditions from $\mathcal{N}\left(m^{(0)}, \mathrm{P}^{(0)}\right)$ distribution with

$$m^{(0)} = \begin{pmatrix} F/4 \\ \vdots \\ F/4 \end{pmatrix} \text{ and } \mathrm{P}^{(0)} = \begin{pmatrix} F^2/4 & & 0 \\ & \ddots & \\ 0 & & F^2/4 \end{pmatrix}.$$

After the initialization, an additional spinup for 18 time units, equivalent of 90 days, is performed. The goal of the spinup is to relax the state, and to develop the covariances between the elements of the stare vector. The initialization and the spinup are performed independently for a truth state $X^{(18)}$ and for a first guess $X^{(18),a}$, and an initial ensemble is created by adding $\mathcal{N}(0, 0.0001 \cdot \mathrm{I})$ distributed noise to the first guess. We assume that an observation operator is identity, $\mathrm{H} = \mathrm{I}$, and $\mathrm{R} = \mathrm{I}$. To add an additional source of model error to the experiment, we advance the ensemble in time using the value $0.95F$ in Equation (8.23).

Firstly, we test the dependence of the selected method on the ensemble size with the state dimension $K = 64$. For the given size of the ensemble, we initialize the experiment and perform the first assimilation for 10 times. The RMSE of the first analysis is shown on Figure 8.4. The figure shows that the proposed method decreases the RMSE even when the ensemble size $N = 2$ and the errors decrease as the size of the ensemble increases with a typical rate $N^{-1/2}$.

On the other hand, the error of the analysis obtained by the EnKF is comparable to the error of the free run unless the size of the ensemble is higher than K. When the size of the ensemble is larger than the dimension of the state space, the error of the EnKF decreases. However, even when $N = 130$, i.e., the size of the ensemble is two times larger than the size of the state space, the error of the EnKF is significantly higher than the error of any SDEnKF.

In the second experiment we perform 20 assimilation cycles using the Lorenz 96 model with the dimension of the state space $K = 64$ and the ensemble size $N = 4$. We assume that the whole state is observed, and we perform the assimilation every 0.05 time units with the first assimilation at 18.05. The results (Figure 8.5) shows that the proposed methods decrease the RMSE immediately after the first assimilation. Conversely, the EnKF even increase the error as multiple assimilations are performed. This observation is not completely surprising, because Kelly et al. [2014] shows that, for a class of dynamical systems, the EnKF remains within a bounded distance of truth unless sufficiently large covariance inflation is used.

8.4.2 Shallow water equations

The shallow water equations is another popular toy model, and it can serve as a simplified model of atmospheric flow. The state

$$x = \begin{pmatrix} h \\ u \\ v \end{pmatrix}$$

consist of three vectors: a water level height h, and velocities in x and y directions u and v respectively. The evolution of the state is governed by the differential

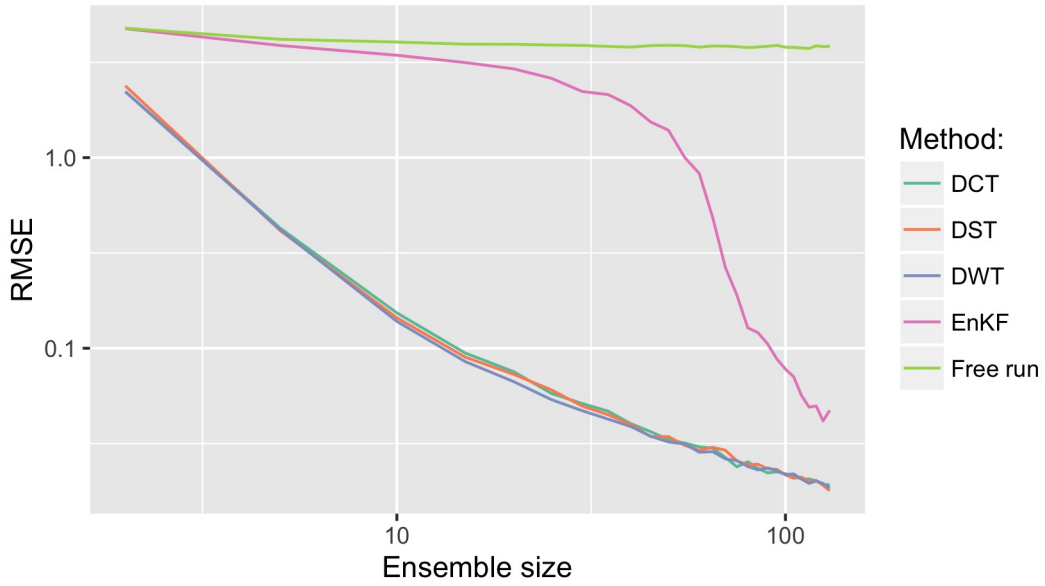

Figure 8.4: Mean RMSE from 10 realization of the Lorenz 96 problem with the whole state observed and the size of the state $K = 64$. The RMSE is measured after the first data assimilation cycle, and DCT, DST and DWT stand for different spectral transformations. Free run stands for the ensemble without any assimilation.

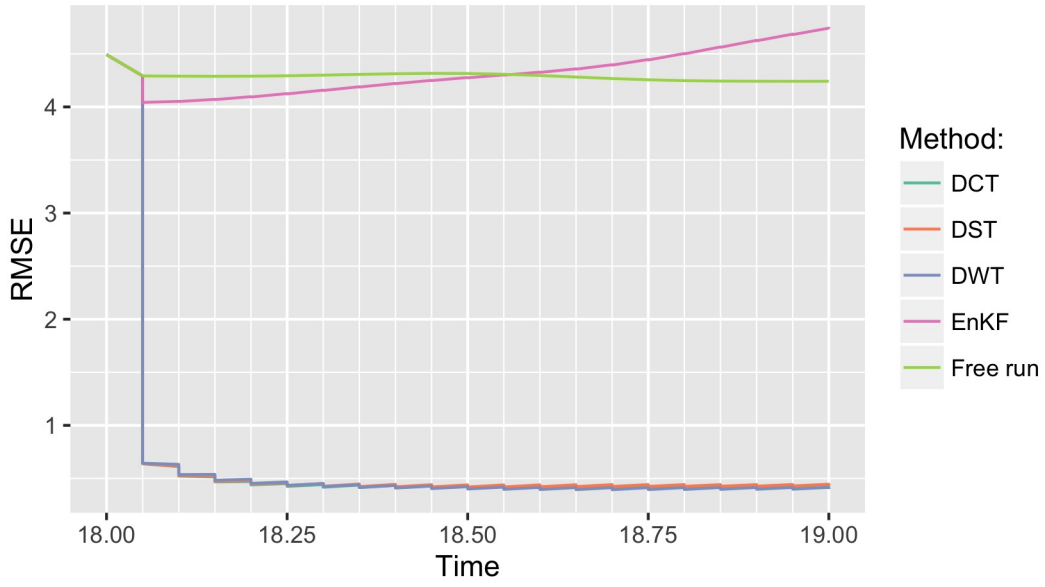

Figure 8.5: Mean RMSE form 10 realization of the Lorenz 96 problem with the whole state observed. The size of the state $K = 64$, and the size of the ensemble $N = 4$. The first assimilation is performed at 18.05, and then the assimilation is performed every 0.05 time units. DCT, DST and DWT stand for different spectral transformations, and Free run stands for the ensemble without any assimilation.

equations of conservation of mass and momentum,

$$\begin{aligned}
&\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} + \frac{\partial (vh)}{\partial y} = 0,\\
&\frac{\partial (hu)}{\partial t} + \frac{\partial}{\partial x}\left(hu^2 + \frac{1}{2}gh^2\right) + \frac{\partial (huv)}{\partial y} = 0,\\
&\frac{\partial (hv)}{\partial t} + \frac{\partial (huv)}{\partial x} + \frac{\partial}{\partial y}\left(hv^2 + \frac{1}{2}gh^2\right) = 0,
\end{aligned}$$

where g is gravity acceleration, with reflective boundary conditions, and without Coriolis force or viscosity. For computations, the equations are discretized on a rectangular grid size 64×64 with horizontal distance between grid points 150 km and advanced by the Lax–Wendroff method with the time step 1 s [Moler, 2011, Chapter 18]. Hence, the dimension of the state is

$$64 \times 64 \times 3 = 12\,288.$$

The initial values are set as follows. Water level $h = 10$ km with an additional Gaussian shaped water rise of a height of 1 km and a width of 32 nodes, located in the center of the domain, and $u = v = 0$. These initial values are moved for a 3 hour spinup. An ensemble is created by adding random noise with a prescribed background covariance, described in the next paragraph.

The background covariance

$$\mathrm{B} = \widehat{\mathrm{C}} \circ \mathrm{T},$$

where $\widehat{\mathrm{C}}$ is a sample covariance from samples taken every second from time $t_{\mathrm{start}} = 3$ h to time $t_{\mathrm{end}} = 6$ h, and T is a tapering matrix with a block structure

$$\mathrm{T} = \begin{pmatrix} \mathrm{A} & 0 & 0 \\ 0 & \mathrm{A} & 0 \\ 0 & 0 & \mathrm{A} \end{pmatrix} + 0.9 \begin{pmatrix} 0 & \mathrm{A} & \mathrm{A} \\ \mathrm{A} & 0 & \mathrm{A} \\ \mathrm{A} & \mathrm{A} & 0 \end{pmatrix}$$

with $\mathrm{A} \in \mathbb{R}^{64^2 \times 64^2}$ such that

$$(\mathrm{A})_{k,l} = \exp(-|x_k - x_l|) \exp(-|y_k - y_l|),$$

where (x_k, y_k) and (x_j, y_j) are coordinates of grid points corresponding the k^{th} row and the j^{th} column of the matrix A respectively. Equivalently, the matrix T can be written in the form

$$\mathrm{T} = \mathrm{K} \otimes \mathrm{M} \otimes \mathrm{M}$$

with

$$\mathrm{K} = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix}$$

and $\mathrm{M} \in \mathbb{R}^{64 \times 64}$ such that

$$(\mathrm{M})_{i,j} = \exp(-|i - j|), \quad i, j = 1, \ldots, 64.$$

The tapering is performed due to the fact that the number of samples is lower than the dimension of the state, and the sample covariance is therefore singular.

We use 2-D tensor product DST, DCT and DWT in place of a spectral transformation in the SDEnKF. The observation error is assumed to be zero mean with variance 1000 m in h and 1000 kg m s^{-1} in u and v.

After the ensemble perturbation, an additional 3 hour spinup is performed to relax the members. The spinup also guarantees that the first forecast is physically reasonable. We use 20 ensemble members in both experiments, and we assume that the first observation is available 6 h after the initial drop.

The first experiment assumes that the full state is observed, and observations are available every hour from 6 h till 10 h, i.e., there is a total number of 5 observations. The results (Figure 8.6) show that the spectral methods significantly decrease RMSE immediately after the first assimilation, but the RMSE of the analysis obtained by the EnKF is nearly indistinguishable from the RMSE of the free run.

The second experiment assumes that only the water level height is observed, so the implementation described in Section 8.3.2 is used. We perform three assimilation cycles with the first one at 6 h. The SDEnKF decreases not only the RMSE of the water level height, but also the RMSE of velocities in both directions, although they are not observed, Figure 8.7.

The previous experiment confirms that, at least for the shallow water equation model, the method proposed in Section 8.3.2 gives reasonable analysis, which update the forecast, i.e., decrease the error of the forecast, and

8.4.3 WRF model

The last experiment is the , so called, pseudo observation test (PSOT) using the Weather and Research and Forecast (WRF) model.

The PSOT test the response of a selected assimilation method to a single pseudo observation. Hence, outcomes of the test are pictures of an innovation, which is a difference between the forecast and the analysis. Therefore, the test results are interesting mainly for experts in a modeled area as they can evaluate whether the results are reasonable.

The WRF model is a numerical weather prediction system, which can be used for both atmospheric research and operating weather forecasting. It can be used in many meteorological applications across scales from hundreds of meters to hundreds of kilometers. The details about the model may be found in Skamarock et al. [2008]. The state of the model consists of seven variables: potential temperature, wind in x and y directions, perturbation geopotential, perturbation dry air mass in column and water vapor mixing ratio. These variables are discretized on a three dimensional mesh.

In our experiment we use a mesh with a 27 km horizontal resolution and 39 vertical levels covering Central Europe. Although this domain is, in comparison with real applications of the WRF model, very small, the size of the state is nearly 1.2 million.

We initial a forecast ensemble using the analysis provided by the GFS model for May 30$^{\text{th}}$, 2013 00:00 AM, and then perform 4 hour spinnup. The ensemble is created adding noise with a covariance obtained by the NCM method (Parrish and Derber [1992]). We assume that we observe a potential temperature in one point located in the middle of the domain in the third vertical level, and the

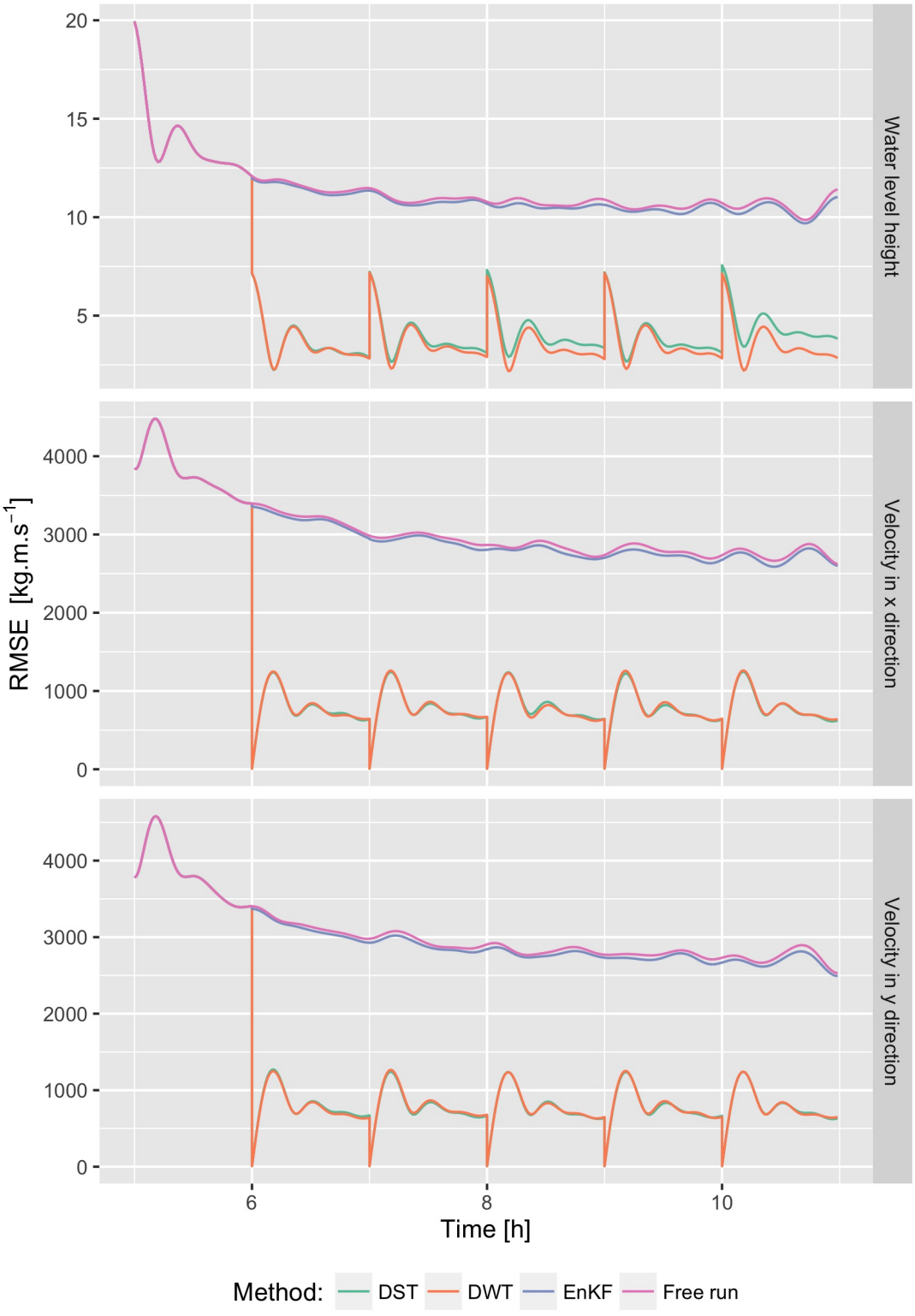

Figure 8.6: RMSE from one realization of five assimilation cycles using shallow water equations. The size of the ensemble is 20, and the observations are available every hour from 6 h until 10 h. The full state is observed.

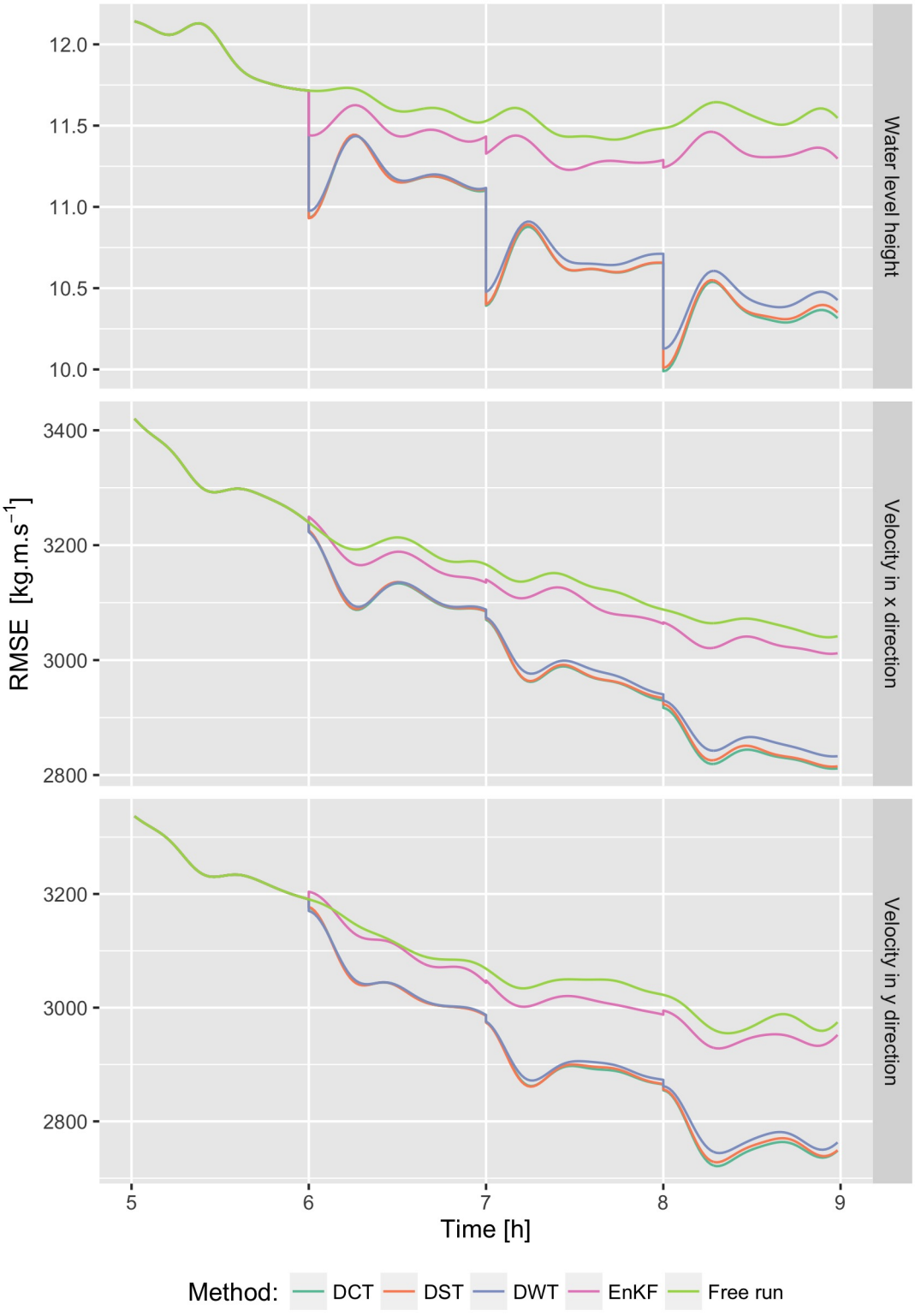

Figure 8.7: RMSE from one realization of three assimilation cycles using shallow water equations. The size of the ensemble is 20, and the observations are available every hour from 6 h until 8 h. Only the water level height is observed.

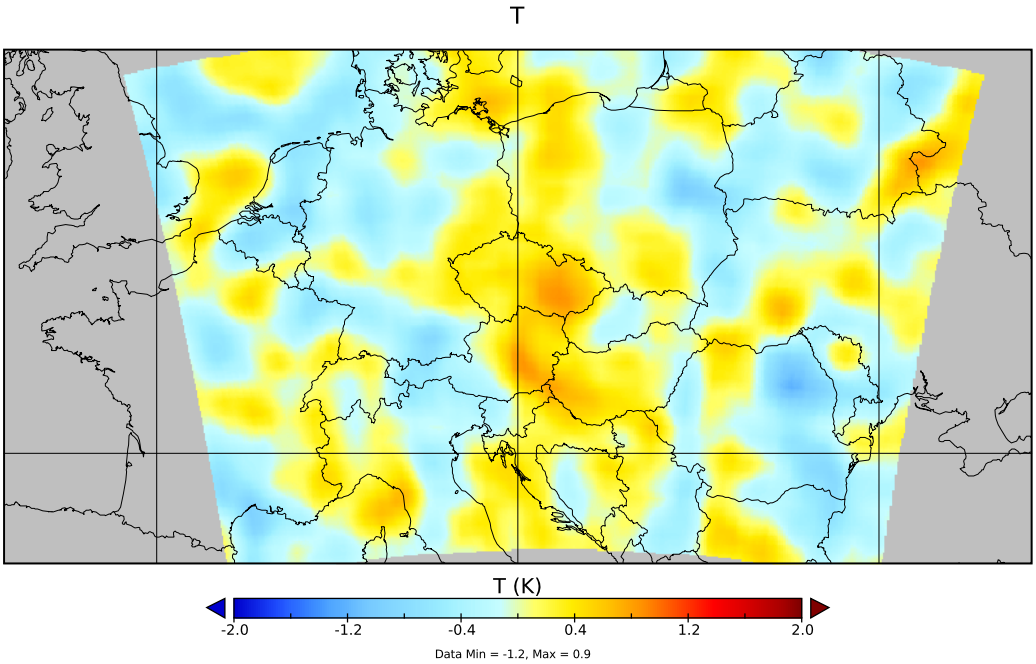

Figure 8.8: Difference of the potential temperature in the third vertical level between the analysis ensemble mean obtained using the EnKF and the forecast ensemble mean. The pseudo observation is located in the middle of the domain, and its value is 2 K higher than the forecast ensemble mean at the same point.

observed value is 2 degrees higher than the forecast ensemble mean. We assimilate this observation using the EnKF and the SDEnKF with discrete sine transform (DST). We use an ensemble of 20 members, and set the observation error variance to 0.5 K.

Innovation using the EnKF, Figure 8.8, shows spurious correlations caused by the small size of the ensemble. There are significant changes in grid points located far away from the location of the pseudo observation, which are clearly not realistic. On the other hand, the innovation using SDEnKF, Figure 8.9, is smooth, and centered around the location of the pseudo observation.

8.5 Summary

We have proposed an evolution of a standard ensemble Kalman filter, which uses a different estimate of a covariance in place of a forecast covariance the EnKF update equation. We have shown that, under reasonable conditions, the proposed estimate always has a smaller error than a classical sample covariance.

We have tested the proposed method with three different models, and we can summarize our observations in the following points.

1. Analysis states obtained using SDEnKF has always been better than analysis states obtained using the EnKF.

2. When an ensemble is really small, the analysis obtained using EnKF may be even worse than a forecast without assimilation. On the other hand, SDEnKF has decreased the error of the forecast even in this situation.

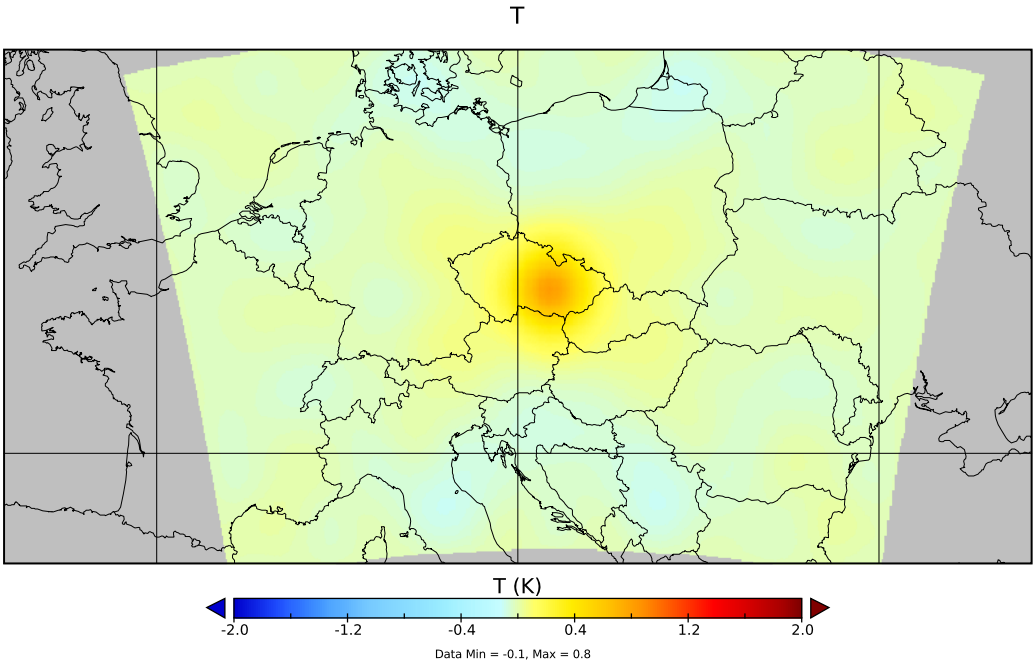

Figure 8.9: Difference of the potential temperature in the third vertical level between the analysis ensemble mean obtained using the SDEnKF with DST and the forecast ensemble mean. The pseudo observation is located in the middle of the domain, and its value is 2 K higher than the forecast ensemble mean at the same point.

3. The analysis obtained by SDEnKF has always been physically reasonable, i.e., the model has been able to be restarted from analysis states.

8.6 Additional notes and references

Then need of huge ensembles to avoid spurious correlations has been known for some time, and the extensive discussion of this topic may be found in Evensen [2009]. Therefore, many papers, e.g., Anderson [2001], Furrer and Bengtsson [2007], Hunt et al. [2007], Sakov and Bertino [2011], propose different types of covariance localizations, which should suppress these correlations.

Using FFT in the EnKF is proposed in Mandel et al. [2010a,b] as an alternative approach to a localization. This approach is motivated by the fact that Fourier basis vectors are eigenvectors of a covariance of a second order stationary random field (Pannekoucke et al. [2007]). On a sphere, an isotropic random field has diagonal covariance in the basis of spherical harmonics, as shown in Boer [1983], so similar algorithms can be developed there as well. On the other hand, the stationary assumption does not allow the covariance to vary spatially. Therefore, using wavelets instead of FFT is proposed in Beezley et al. [2011].

Sparse approximation of a covariance in data assimilation have been studied for some time. Parrish and Derber [1992] proposes to model the background covariance using diagonal approximation in spherical harmonics. The ECMWF 3DVAR system also uses diagonal covariance in spherical harmonics (Courtier et al. [1998]). Diagonal approximation in the Fourier space for homogeneous 2D

error fields is proposed in Berre [2000], Pannekoucke et al. [2007], and Buehner and Charron [2007] combines spatial and spectral localization. The error of a different covariance estimates used in data assimilation is studied also in Furrer and Bengtsson [2007].

9. Summary

We have obtained three main scientific results in this thesis.

Firstly, the assumption that the data noise is only weakly measurable with the covariance bounded from below is crucial, and when this assumption is fulfilled, then all three studied assimilation method, i.e., the 3DVAR, the ensemble Kalman filter and the Bayesian filtering, are well defined and well posed. In fact, we have shown that when the data covariance and a forecast covariance commute, the Bayes formula is well posed if and only if the data covariance is bounded from below. From these statements follows that if the data are contaminated by a white noise, then all three methods are well posed.

Secondly, the ensemble Kalman filter on an infinite dimensional space has the same properties as the EnKF on a finite dimensional space. We have shown that, regardless of the dimension, the ensemble converge in $\mathcal{L}^p$ to the mean field ensemble, whose evolution is governed by the original Kalman filter equations. We have also shown that this convergence has a usual rate $\mathcal{O}\left(N^{-1/2}\right)$. This result is an extension of the already known fact that EnKF converges to the solution of the Kalman filter in large ensemble limit when the state space is finitely dimensional.

Thirdly, the idea of the spectral diagonal ensemble Kalman filter is not completely new, but we have computed the expected error of both covariance estimates, and shown that the spectral diagonal sample covariance has a smaller expected error than the sample covariance. We have also tested this algorithm using multiple toy models. The results have shown that this algorithm produces a better analysis than the classical EnKF especially when only a small ensemble is available.

Bibliography

Ben Adcock and Anders C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. *Foundations of Computational Mathematics*, pages 1–61, 2015. ISSN 1615-3383. doi: 10.1007/s10208-015-9276-6. URL `http://dx.doi.org/10.1007/s10208-015-9276-6`.

Charalambos D. Aliprantis and Kim C. Border. *Infinite-dimensional analysis*. Springer-Verlag, Berlin, second edition, 1999. ISBN 3-540-65854-8. doi: 10.1007/978-3-662-03961-8. A hitchhiker's guide.

Brian D. O. Anderson and John B. Moore. *Optimal filtering*. Prentice-Hall, Englewood Cliffs, N.J., 1979.

Jeffrey L. Anderson. An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, 129:2884–2903, 2001. doi: 10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2.

Robert B. Ash and Melvin F. Gardner. *Topics in stochastic processes*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Probability and Mathematical Statistics, Vol. 27.

A. V. Balakrishnan. *Applied functional analysis*. Springer-Verlag, New York, 1976.

H. T. Banks, Shuhua Hu, and W. Clayton Thompson. *Modeling and inverse problems in the presence of uncertainty*. Monographs and Research Notes in Mathematics. CRC Press, Boca Raton, FL, 2014. ISBN 978-1-4822-0642-5.

Jonathan D. Beezley, Jan Mandel, and Loren Cobb. Wavelet ensemble Kalman filters. In *Proceedings of IEEE IDAACS'2011, Prague, September 2011*, volume 2, pages 514–518. IEEE, 2011. ISBN 978-1-4577-1423-8. doi: 10.1109/IDAACS.2011.6072819.

P. Bénard, J. Vivoda, J. Mašek, P. Smolíková, K. Yessad, Ch. Smith, R. Brožková, and J.-F. Geleyn. Dynamical kernel of the Aladin-NH spectral limited-area model: Revised formulation and sensitivity experiments. *Quarterly Journal of the Royal Meteorological Society*, 136(646):155–169, 2010. doi: 10.1002/qj.522.

Andrew F. Bennett. *Inverse Methods in Physical Oceanography*. Cambridge University Press, 1992.

Andrew F. Bennett. *Inverse modeling of the ocean and atmosphere*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-81373-5. doi: 10.1017/CBO9780511535895. URL `http://dx.doi.org/10.1017/CBO9780511535895`.

Loïk Berre. Estimation of synoptic and mesoscale forecast error covariances in a limited-area model. *Monthly Weather Review*, 128(3):644–667, 2000. doi: 10.1175/1520-0493(2000)128⟨0644:EOSAMF⟩2.0.CO;2.

Craig H. Bishop, Brian J. Etherton, and Sharanya J. Majumdar. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly Weather Review*, 129:420–436, 2001. doi: 10.1175/1520-0493(2001)129⟨0420:ASWTET⟩2.0.CO;2.

L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. *ScaLAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997. ISBN 0-89871-397-8 (paperback).

G. J. Boer. Homogeneous and isotropic turbulence on the sphere. *Journal of the Atmospheric Sciences*, 40(1):154–163, 1983. doi: 10.1175/1520-0469(1983)040⟨0154:HAITOT⟩2.0.CO;2.

V. I. Bogachev. *Measure theory. Vol. I, II*. Springer-Verlag, Berlin, 2007. ISBN 978-3-540-34513-8; 3-540-34513-2. doi: 10.1007/978-3-540-34514-5.

Vladimir I. Bogachev. *Gaussian measures*. Mathematical Surveys and Monographs, Vol. 62. American Mathematical Society, Providence, RI, 1998. ISBN 0-8218-1054-5.

Mark Buehner. Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. *Monthly Weather Review*, 140(2):617–636, 2011. doi: 10.1175/MWR-D-10-05052.1.

Mark Buehner and Martin Charron. Spectral and spatial localization of background-error correlations for data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 133(624):615–630, 2007. ISSN 1477-870X. doi: 10.1002/qj.50.

Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126:1719–1724, 1998.

Yuan Shih Chow and Henry Teicher. *Probability theory. Independence, interchangeability, martingales*. Springer-Verlag, New York, third edition, 1997. ISBN 0-387-98228-0.

Philippe G. Ciarlet. *Linear and nonlinear functional analysis with applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013. ISBN 978-1-611972-58-0.

S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25(11):115008, 43, 2009. doi: 10.1088/0266-5611/25/11/115008.

S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of Bayesian inverse problems for PDEs. *SIAM J. Numer. Anal.*, 48(1):322–345, 2010. doi: 10.1137/090770734.

P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998. ISSN 1477-870X. doi: 10.1002/qj.49712455002.

Philippe Courtier and Olivier Talagrand. Variational assimilation of meteorological observations with the adjoint vorticity equation. II: Numerical results. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1329–1347, 1987. ISSN 1477-870X. doi: 10.1002/qj.49711347813. URL `http://dx.doi.org/10.1002/qj.49711347813`.

Noel A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons Inc., New York, 1993. ISBN 0-471-00255-0.

Giuseppe Da Prato. *An introduction to infinite-dimensional analysis*. Springer-Verlag, Berlin, 2006. ISBN 978-3-540-29020-9; 3-540-29020-6. doi: 10.1007/3-540-29021-4.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992. ISBN 0-521-38529-6.

Giuseppe Da Prato and Jerzy Zabczyk. *Second order partial differential equations in Hilbert spaces*, volume 293 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-77729-1.

M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29 (9):095017, 27, 2013. ISSN 0266-5611. doi: 10.1088/0266-5611/29/9/095017.

Masoumeh Dashti, Stephen Harris, and Andrew Stuart. Besov priors for Bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200, 2012. ISSN 1930-8337. doi: 10.3934/ipi.2012.6.183.

Ingrid Daubechies. *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. ISBN 0-89871-274-2. doi: 10.1137/1.9781611970104.

G. Desroziers, L. Berre, B. Chapnik, and P. Poli. Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3385–3396, 2005. ISSN 1477-870X. doi: 10.1256/qj.05.108. URL `http://dx.doi.org/10.1256/qj.05.108`.

Gérald Desroziers, Jean-Thomas Camino, and Lo ik Berre. 4DEnVar: link with 4D state formulation of variational assimilation and different possible implementations. *Quarterly Journal of the Royal Meteorological Society*, 140(684):2097–2110, 2014. ISSN 1477-870X. doi: 10.1002/qj.2325.

Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo in Practice*. Springer, 2001. ISBN 0-387-95146-6.

J Durbin and Siem Jan Koopman. *Time series analysis by state space methods*. Oxford University Press, Oxford, 2nd ed. edition, 2012. ISBN 9780199641178.

Geir Evensen. Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99 (C5)(10):143–162, 1994.

Geir Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer, 2nd edition, 2009. ISBN 978-3-642-03710-8. doi: 10.1007/978-3-642-03711-5.

Jacob Feldman. Equivalence and perpendicularity of Gaussian processes. *Pacific J. Math.*, 8:699–708, 1958. ISSN 0030-8730.

Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.*, 98(2):227–255, 2007. ISSN 0047-259X. doi: 10.1016/j.jmva.2006.08.003.

Jacques Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

Paul R. Halmos. *Measure Theory*. Graduate Texts in Mathematics. D. Van Nostrand Company, Inc., New York, N. Y., 1950. ISBN 9780387900889. doi: 10.1007/978-1-4684-9440-2.

Thomas M. Hamill and Chris Snyder. A hybrid ensemble Kalman filter–3D variational analysis scheme. *Monthly Weather Review*, 128(8):2905–2919, 2000. doi: 10.1175/1520-0493(2000)128⟨2905:AHEKFV⟩2.0.CO;2.

Kenneth Hoffman and Ray Kunze. *Linear algebra*. Second edition. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1971.

Jørgen Hoffmann-Jørgensen. Sums of independent Banach space valued random variables. *Studia Math.*, 52:159–186, 1974. ISSN 0039-3223.

P.L. Houtekamer and Herschel L. Mitchell. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796–811, 1998.

B. R. Hunt, E. J. Kostelich, and I. Szunyogh. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230:112–126, 2007. doi: 10.1016/j.physd.2006.11.008.

Mark Z. Jacobson. *Fundamentals of atmospheric modeling*. Cambridge Univ Press, 2005. ISBN 052183970X.

Andrew H. Jazwinski. *Stochastic processes and filtering theory*. Academic Press, New York, 1970.

Simon .J. Julier and Jeffrey K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defense Sensing, Simulations and Controls*, 1997. doi: 10.1117/12.280797.

S.J. Julier and J. K. Uhlmann. Corrections to uscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(12):1958–1958, 2004.

R. E. Kalman and R. S. Bucy. New results in filtering and prediction theory. *Transactions of the ASME – Journal of Basic Engineering*, 83:95–108, 1961.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering, Series D*, 82: 35–45, 1960. doi: 10.1115/1.3662552.

Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003. ISBN 0-521-79629-6, 0-521-79179-0.

I. Kasanický, J. Mandel, and M. Vejmelka. Spectral diagonal ensemble Kalman filters. *Nonlinear Processes in Geophysics*, 22(4):485 – 497, 2015. doi: 10.5194/npg-22-485-2015.

D. T. B. Kelly, K. J. H. Law, and A. M. Stuart. Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, 27 (10):2579–2603, 2014. doi: 10.1088/0951-7715/27/10/2579.

Jeffrey D. Kepert. Covariance localisation and balance in an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 135(642):1157–1176, 2009. doi: 10.1002/qj.443.

Erwin Kreyszig. *Introductory functional analysis with applications*. Wiley Classics Library. John Wiley & Sons Inc., New York, 1989. ISBN 0-471-50459-9.

Evan Kwiatkowski and Jan Mandel. Convergence of the square root ensemble Kalman filter in the large ensemble limit. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1–17, 2015. doi: 10.1137/140965363.

William Lahoz, Boris Khattatov, and Richard Menard, editors. *Data Assimilation: Making Sense of Observations*. Springer, 2010. doi: 10.1007/978-3-540-74703-1.

Serge Lang. *Real and functional analysis*, volume 142 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 1993. ISBN 0-387-94001-4.

S. Lasanen. Measurements and infinite-dimensional statistical inverse theory. *PAMM*, 7(1):1080101–1080102, 2007. ISSN 1617-7061. doi: 10.1002/pamm.200700068.

Kody Law, Andrew Stuart, and Konstantinos Zygalakis. *Data assimilation*, volume 62 of *Texts in Applied Mathematics*. Springer, Cham, 2015. ISBN 978-3-319-20324-9; 978-3-319-20325-6. doi: 10.1007/978-3-319-20325-6. URL `http://dx.doi.org/10.1007/978-3-319-20325-6`. A mathematical introduction.

Kody J. H. Law, Hamidou Tembine, and Raul Tempone. Deterministic mean-field ensemble kalman filtering. *SIAM Journal on Scientific Computing*, 38(3), 2016. ISSN 1064-8275. doi: 10.1137/140984415. URL `http://hdl.handle.net/10754/608649`.

F. Le Gland, V. Monbet, and V.-D. Tran. Large sample asymptotics for the ensemble Kalman filter. In Dan Crisan and Boris Rozovskiĭ, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 598–631. Oxford University Press, 2011.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete (3), Vol. 23. Springer-Verlag, Berlin, 1991. ISBN 3-540-52013-9.

F. S. Levin. *An introduction to quantum theory*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-59161-9.

Chengsi Liu, Qingnong Xiao, and Bin Wang. An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Monthly Weather Review*, 136(9):3363–3373, 2008. doi: 10.1175/2008MWR2312.1.

David M. Livings, Sarah L. Dance, and Nancy K. Nichols. Unbiased ensemble square root filters. *Phys. D*, 237(8):1021–1028, 2008. doi: 10.1016/j.physd.2008.01.005.

A. C. Lorenc. Analysis methods for numerical weather prediction. *Qart. J. R. Met. Soc.*, 112:1177–1194, 1986.

Andrew C. Lorenc, Neill E. Bowler, Adam M. Clayton, Stephen R. Pring, and David Fairbairn. Comparison of Hybrid-4DEnVar and Hybrid-4DVar data assimilation methods for global NWP. *Monthly Weather Review*, 143(1):212–229, 2015/09/12 2014. doi: 10.1175/MWR-D-14-00195.1.

E. N. Lorenz. Predictability - a problem partly solved. In Tim Palmer and Renate Hagendorn, editors, *Predictability of Weather and Climate*, pages 40–58. Cambridge University Press, 2006.

Edward N. Lorenz and Kerry A. Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, 55(3):399–414, 1998. doi: 10.1175/1520-0469(1998)055⟨0399:OSFSWO⟩2.0.CO;2.

Jan Mandel. Introduction to infinite dimensional statistics and applications. Unpublished lecture notes, 2016.

Jan Mandel and Jonathan D. Beezley. An ensemble Kalman-particle predictor-corrector filter for non-gaussian data assimilation. In Gabrielle Allen, Jaroslaw Nabrzyski, Edward Seidel, Geert van Albada, Jack Dongarra, and Peter Sloot, editors, *Computational Science - ICCS 2009*, volume 5545 of *Lecture Notes in Computer Science*, pages 470–478. Springer Berlin / Heidelberg, 2009. doi: 10.1007/978-3-642-01973-9_53.

Jan Mandel, Jonathan D. Beezley, Kryštof Eben, Pavel Juruš, Volodymyr Y. Kondratenko, and Jaroslav Resler. Data assimilation by morphing fast Fourier transform ensemble Kalman filter for precipitation forecasts using radar images. CCM Report 289, University of Colorado Denver, April 2010a. `http://ccm.ucdenver.edu/reports/rep289.pdf`, retrieved December 2011.

Jan Mandel, Jonathan D. Beezley, and Volodymyr Y. Kondratenko. Fast Fourier transform ensemble Kalman filter with application to a coupled atmosphere-wildland fire model. In A. M. Gil-Lafuente and J. M. Merigo, editors, *Computational Intelligence in Business and Economics, Proceedings of MS'10*, pages 777–784. World Scientific, 2010b. doi: 10.1142/9789814324441_0089.

Jan Mandel, Loren Cobb, and Jonathan D. Beezley. On the convergence of the ensemble Kalman filter. *Applications of Mathematics*, 56:533–541, 2011. doi: 10.1007/s10492-011-0031-2.

Stephen A. Martucci. Symmetric convolution and the discrete sine and cosine transforms. *IEEE Transactions on Signal Processing*, 42(5):1038–1051, 1994. doi: 10.1109/78.295213.

Takemasa Miyoshi, Keiichi Kondo, and Toshiyuki Imamura. The 10,240-member ensemble kalman filtering with an intermediate agcm. *Geophysical Research Letters*, 41(14):5264–5271, 2014. ISSN 1944-8007. doi: 10.1002/2014GL060863. URL `http://dx.doi.org/10.1002/2014GL060863`. 2014GL060863.

Cleve Moler. Experiments with MATLAB. `http://www.mathworks.com/moler/exm`, 2011. Accessed December 2014.

Gen Nakamura and Roland Potthast. *Inverse Modeling*. 2053-2563. IOP Publishing, 2015. ISBN 978-0-7503-1218-9. doi: 10.1088/978-0-7503-1218-9. URL `http://dx.doi.org/10.1088/978-0-7503-1218-9`.

Lars Nerger, Tijana Janjić, Jens Schröter, and Wolfgang Hiller. A regulated localization scheme for ensemble-based kalman filters. *Quarterly Journal of the Royal Meteorological Society*, 138(664):802–812, 2012. ISSN 1477-870X. doi: 10.1002/qj.945. URL `http://dx.doi.org/10.1002/qj.945`.

Olivier Pannekoucke, Loïk Berre, and Gerald Desroziers. Filtering properties of wavelets for local background-error correlations. *Quarterly Journal of the Royal Meteorological Society*, 133(623, Part B):363–379, 2007. doi: 10.1002/qj.33.

David F. Parrish and John C. Derber. The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120(8):1747–1763, 1992. doi: 10.1175/1520-0493(1992)120⟨1747:TNMCSS⟩2.0.CO;2.

B. J. Pettis. On integration in vector spaces. *Trans. Amer. Math. Soc.*, 44(2):277–304, 1938. ISSN 0002-9947.

W. T. Martin R. H. Cameron. Transformations of weiner integrals under translations. *Annals of Mathematics*, 45(2):386–396, 1944. ISSN 0003486X. doi: 10.2307/1969276.

J. O. Ramsay and B. W. Silverman. *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. ISBN 0-387-95414-7. doi: 10.1007/b98886. URL `http://dx.doi.org/10.1007/b98886`. Methods and case studies.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005. ISBN 978-0387-40080-8; 0-387-40080-X. doi: 10.1007/b98888.

Michael Reed and Barry Simon. *Methods of modern mathematical physics. I.* Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, second edition, 1980. ISBN 0-12-585050-6. Functional analysis.

P. Sakov and L. Bertino. Relation between two common localisation methods for the EnKF. *Computational Geosciences*, 10:225–237, 2011. ISSN 1420-0597. doi: 10.1007/s10596-010-9202-6.

Pavel Sakov and Peter R. Oke. Implications of the form of the ensemble transformation in the ensemble square root filters. *Monthly Weather Review*, 136(3): 1042–1053, 2008. doi: 10.1175/2007MWR2021.1.

William C. Skamarock, Joseph B. Klemp, Jimy Dudhia, David O. Gill, Dale M. Barker, Michael G. Duda, Xiang-Yu Huang, Wei Wang, and Jordan G. Powers. A description of the Advanced Research WRF version 3. NCAR Technical Note 475, 2008. `http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf`, retrieved December 2011.

Gilbert Strang and Truong Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, Wellesley, MA, 1996. ISBN 0-9614088-7-1.

A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

Andrew M. Stuart. The Bayesian approach to inverse problems. arXiv:1302.6989, 2013.

Olivier Talagrand and Philippe Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328, 1987. ISSN 1477-870X. doi: 10.1002/qj.49711347812.

Michael K. Tippett, Jeffrey L. Anderson, Craig H. Bishop, Thomas M. Hamill, and Jeffery S. Whitaker. Ensemble square root filters. *Monthly Weather Review*, 131:1485–1490, 2003.

N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability distributions on Banach spaces*, volume 14 of *Mathematics and its Applications (Soviet Series)*. D. Reidel Publishing Co., Dordrecht, 1987. ISBN 90-277-2496-2. doi: 10.1007/978-94-009-3873-1. Translated from the Russian and with a preface by Wojbor A. Woyczynski.

Peter Jan van Leeuwen, Yuan Cheng, and Sebastian Reich. *Nonlinear data assimilation*, volume 2 of *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*. Springer, Cham, 2015. ISBN 978-3-319-18346-6; 978-3-319-18347-3. doi: 10.1007/978-3-319-18347-3. URL `http://dx.doi.org/10.1007/978-3-319-18347-3`.

Joachim Weidmann. *Linear operators in Hilbert spaces*, volume 68 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Berlin, 1980. ISBN 0-387-90427-1. doi: 10.1007/978-1-4612-6027-1. Translated from the German by Joseph Szücs.

J. S. Whitaker and T. M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130:1913–1924, 2002.

Wojbor A. Woyczyński. On Marcinkiewicz-Zygmund laws of large numbers in Banach spaces and related rates of convergence. *Probab. Math. Statist.*, 1(2): 117–131, 1980. ISSN 0208-4147.

List of Figures

List of Abbreviations

BF	Bayes' filter
DCT	discrete cosine transfom
DST	discrete sine transfom
DWT	discrete wavelet transfom
EnKF	ensemble Kalman filter
FFT	fast Fourier transform
KF	Kalman filter
SDEnKF	spectral diagonal ensemble Kalman filter