Quality of the Analysis Step in EnKF

Master of Science Thesis in Reservoir Mechanics

Brede Rem Bergo

Center for Integrated Petroleum Research

Department of Mathematics
University of Bergen

June 2011

Acknowledgments

I would like to thank my supervisor Trond Mannseth for your guidance and help, and also for motivating me during the run. I would also like to thank Andrey Kovalenko for sharing his knowledge with me.

Thanks to my fellow students for making the years I have spent here absolutely worthwhile. Svenn and Kristian at room B, thanks for your efforts in keeping me on track of my studies, but most of all, thank you for all the hilarious moments.

Finally I want to express my gratitude to my closest family and friends. Especially to my parents for all your support.

Brede,
June 2011

Contents

Chapter 1

Introduction

In many physical applications we want to characterize the parameters of a system based on indirect observations or measurements.

In a reservoir simulator setting, the goal is to simulate the production of hydrocarbons from the reservoir. This way we can try out different production strategies and optimize the production plan before the reservoir is put on production. These decisions depend on good simulations of the flow of oil, gas and water in the porous rocks.

To achieve appropriate flow calculations, a good estimate of the flow properties of the rock is needed. The process of building an approximation to the reservoir itself and its properties is called reservoir modeling or reservoir characterization. For this, prior information is used, like well logs, analyzed core plugs from the appraisal wells and seismic data. This information gives us some estimate of our poorly known reservoir parameters, like the porosity and permeability fields.
The performance of the reservoir, given a recovery strategy, can be predicted by a reservoir simulator. After the field is put on production one may use the production data to improve the reservoir model. The basic idea is that predicted performance should match the observed performance. By tuning the parameters in the model, one tries to fit the output of the simulator to the production history. This is referred to as history matching, which is a nonlinear inverse problem.

A promising method to automatically perform the history matching is the Ensemble Kalman Filter. EnKF is a sequential data assimilation algorithm using Monte Carlo techniques where measurements and prior information about the system is combined to make the best weighted estimate based on their uncertainties. After the assimilation, the model is run forward in time using the reservoir simulator. When new observations or data are available, the next analysis step will incorporate the new observations to produce a new analyzed estimate.

A large number of data assimilated at the same time has proved to be a difficult challenge for EnKF. This could correspond to the use of e.g. 4D seismic data.

One computational advantage is that the covariance matrix of the system is never explicitly calculated, but rather approximated from the ensemble itself. However, spurious correlations in the ensemble sample covariance matrix is one problem to be addressed. In particular, properties in cells far away from the location of measurements are affected in too great scale.
EnKF is based on the Kalman Filter, which is a recursive filter for linear problems.

In this master thesis we consider the quality of the analysis step of the EnKF. Our main focus is the sampling errors caused by the approximated sample covariance matrix when a increasing number of measurements are assimilated.
The work here is inspired by [15, 14], where a probabilistic measure for the sampling error is derived under the assumptions of a normally distributed prior and negligible measurement errors.
Here we try a somewhat different approach using approximate calculations and Neumann series to asses the sampling error. We consider measurement errors of varying size.

Outline

Chapter 2 We introduce inverse problems, both linear and nonlinear. We start off defining the inverse problem, followed by classical theory for solving them. The Bayesian approach is introduced as an alternative framework. Also, the the history matching problem in reservoir engineering is mentioned.

Chapter 3 We describe sequential assimilation techniques, and introduce the Kalman Filter and the Extended Kalman Filter.We then concentrate on the formulation of the Ensemble Kalman Filter, practical implementations and some important challenges.

Chapter 4 This chapter provides the idea behind our approach and we define Neumann series which is utilized in the analytical calculations.

Chapter 5 Here we define our assumptions and approximations regarding the covariance matrix and the analysis step in the Ensemble Kalman Filter.

Chapter 6 We derive approximative analytical expressions for the norm of the sampling error, as well as some numerical results based on these.

Chapter 7 We summarize the work done in this thesis and make some remarks on the results from chapter 6

The Appendix contains some useful definitions and derivations to supplement the text. It is referred to the Appendix whenever appropriate with (A.#).

Theoretical Background

Chapter 2

Parameter estimation and inverse problems

Parameter estimation and inverse problems are introduced followed by the Bayesian approach. We use the notation $\mathbf{m}$ for the set of parameters in our model, $\mathbf{F}(.)$ is the forward model and $\mathbf{d}$ will be the output from the forward model.

In the forward problem the goal is to find $\mathbf{d}$ given $\mathbf{m}$

$$\mathbf{d} = \mathbf{F}(\mathbf{m}) \tag{2.0.1}$$

where $\mathbf{F}$ is the equations that govern our dynamical system and $\mathbf{m}$ is the set of parameters that characterize the system. In a reservoir case, $\mathbf{F}$ will be the reservoir simulator or forward model, $\mathbf{m}$ will be the properties that characterize our reservoir, like porosity and permeability, and $\mathbf{d}$ will be the predicted performance of the reservoir. In history matching, which is the inverse problem, one tries to estimate the parameters $\mathbf{m}$ based on the observed data $\mathbf{d}_{obs}$, that is

$$\mathbf{F}(\mathbf{m}) = \mathbf{d}_{obs} \tag{2.0.2}$$

Parameter estimation problems, or inverse problems, can be linear or nonlinear depending upon the forward operator $\mathbf{F}$. The fact that all observations or measurements are associated with uncertainty is one reason that makes inverse problems hard to solve. Because we know that noise is present in the data we should not try to fit the data perfectly because we may let the noise affect features in the model. We can rewrite (2.0.2) as

$$\mathbf{d}_{obs} = \mathbf{d}_{true} + \eta = \mathbf{F}(\mathbf{m}_{true}) + \eta \tag{2.0.3}$$

where we think of the observations as perfectly measured data with added noise. The data $\mathbf{d}_{true}$ would fit the true model $\mathbf{m}_{true}$ perfectly if we assume no modeling errors trough $\mathbf{F}$. Of course, the forward model $\mathbf{F}$ may not represent the physical system exactly. Thus, there may also be model errors present. The goal of the inverse problem is to recover the true model $\mathbf{m}$ given the noisy data $\mathbf{d}_{obs}$. Inverse problems are often ill-posed, that is they do not fulfill one or more of the following criteria:

- Existence: There may be no solution $\mathbf{m}$ that fits the data $\mathbf{d}$ perfectly.
- Uniqueness: There may be an infinite number of solutions that fit the data equally well.
- Stability: Small changes in the data may cause enormous changes in the estimated model, making the solution procedure unstable.

A problem can be well-posed, but still be ill-conditioned if it fails to honor the last criteria.

We usually distinguish between linear and nonlinear problems. We first look at linear inverse problems

2.1 Linear inverse problems

2.1.1 Least square problems

Considering a discrete linear inverse problem, we can write the problem as a linear system of equations

$$\mathbf{Fm} = \mathbf{d} \tag{2.1.1}$$

where we have a data vector $\mathbf{d}$ with N_d observations and a vector $\mathbf{m}$ of N model parameters. We assume that the linear operator $\mathbf{F}$ has full column rank, $rank(\mathbf{F}) = N$. Due to noise in the data, $\mathbf{d}$ frequently lies outside the range of $\mathbf{F}$, thus there is no solution $\mathbf{m}$ that satisfies 2.1.1 exactly. To find an approximate solution we can minimize the residual

$$\mathbf{r} = \mathbf{d} - \mathbf{Fm}$$

A model that minimizes the L_2-norm of the residual

$$min\|\mathbf{d} - \mathbf{Fm}\|_2 \tag{2.1.2}$$

is called a least square solution. This can be obtained by projecting $\mathbf{d}$ onto the range of $\mathbf{F}$. Let

$$\begin{aligned} \mathbf{Fm} &= \mathbf{p} \\ &= proj_{R(\mathbf{F})}\mathbf{d} \end{aligned} \tag{2.1.3}$$

Then the residual is perpendicular to the range of $\mathbf{F}$, so that

$$\mathbf{F}^T(\mathbf{Fm} - \mathbf{d}) = \mathbf{0} \tag{2.1.4}$$

and

$$\mathbf{F}^T\mathbf{Fm} = \mathbf{F}^T\mathbf{d} \tag{2.1.5}$$

This is called the normal equations. The least squares solution is given by

$$\mathbf{m}_{L_2} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{d} \tag{2.1.6}$$

It can be shown that if $\mathbf{F}$ has full column rank then $(\mathbf{F}^T\mathbf{F})^{-1}$ exists. Interestingly, if the data errors are normally distributed then the least squares solution turns out to be statistically the most likely solution.

2.1.2 Maximum Likelihood Estimation

Maximum likelihood estimation can be applied to problems where probability density functions can be associated with the data. Given a model $\mathbf{m}$, we have a probability density function $f_i(d_i \mid \mathbf{m})$ for every observation d_i. For independent data $\mathbf{d}$, the joint probability density is

$$f(\mathbf{d} \mid \mathbf{m}) = f_1(d_1 \mid \mathbf{m}) \cdot f_2(d_2 \mid \mathbf{m}) \cdot \cdots \cdot f_{N_d}(d_{N_d} \mid \mathbf{m}) \tag{2.1.7}$$

Given a set observed data points, what is the model $\mathbf{m}$ that most likely correspond to these data? The likelihood function is given as

$$L(\mathbf{m} \mid \mathbf{d}) = f(\mathbf{d} \mid \mathbf{m}) \tag{2.1.8}$$

where $\mathbf{d}$ is a fixed set of observations and $\mathbf{m}$ is to be estimated. The maximum likelihood principle tells us to choose the model that maximizes the value of the likelihood function. As mentioned before, when we have a discrete linear inverse problem with independent and normally distributed data errors, then the maximum likelihood principle solution is the least squares solution. The data are associated with given standard deviations σ_i. We can write the probability density for d_i as

$$f_i(d_i \mid \mathbf{m}) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(d_i - (\mathbf{Fm})_i)^2}{2\sigma_i^2}\right) \tag{2.1.9}$$

and the likelihood function for all the data is the product of the separate likelihoods

$$L(\mathbf{m} \mid \mathbf{d}) = \frac{1}{(2\pi)^{N_d/2}\prod_{i=1}^{N_d}\sigma_i} \prod_{i=1}^{N_d} \exp\left(-\frac{(d_i - (\mathbf{Fm})_i)^2}{2\sigma_i^2}\right) \tag{2.1.10}$$

Maximizing this expression is the same as maximizing the exponent or minimizing the exponent with opposite sign. Because the constants does not affect the maximization of L we end up with the minimization problem

$$min \sum_{i=1}^{N_d} \frac{(d_i - (\mathbf{Fm})_i)^2}{\sigma_i^2} \tag{2.1.11}$$

This is exactly the least squares solution in 2.1.2, but scaled with the individual standard deviations σ_i^2.

2.1.3 Rank deficiency and the Singular Value Decomposition

Until now we have assumed that the matrix $\mathbf{F} \in \mathbb{R}^{N_d \times N}$ has full column rank, thus the N columns of $\mathbf{F}$ are all linearly independent. If $rank(\mathbf{F}) < min(N_d, N)$ we say that $\mathbf{F}$ is rank deficient. One way to solve least squares problems in ill posed or rank deficient systems is by using the singular value decomposition (SVD). All matrices can be decomposed[8] into the following form

$$\mathbf{F} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{2.1.12}$$

where

- $\mathbf{U}$ is a $N_d \times N_d$ orthogonal matrix with columns that are unit basis vectors spanning the data space, $\mathbb{R}^{N_d}$.
- $\mathbf{V}$ is a $N \times N$ orthogonal matrix with columns that are unit basis vectors spanning the model space, $\mathbb{R}^N$.
- $\mathbf{S}$ is a $N_d \times N$ diagonal matrix with non negative diagonal elements called singular values.

The singular values in $\mathbf{S}$ are arranged in decreasing size, $s_1 \geq s_2 \geq \cdots s_{min(N_d,N)} \geq 0$. Some of the singular values may be zero. If that is the case, $\mathbf{S}$ can be written as

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{2.1.13}$$

with $\mathbf{S}_p$ containing the p nonzero singular values. Then the SVD of $\mathbf{F}$ can be simplified to its compact form

$$\mathbf{F} = \mathbf{U}_p\mathbf{S}_p\mathbf{V}_p^T \tag{2.1.14}$$

The SVD can be used to compute a generalized inverse of $\mathbf{F}$, the Moore-Penrose pseudo inverse[21, 23] which is given by

$$\mathbf{F}^\dagger = \mathbf{V}_p\mathbf{S}_p^{-1}\mathbf{U}_p^T \tag{2.1.15}$$

By use of 2.1.15, the pseudo inverse solution is defined as

$$\begin{aligned} \mathbf{m}^\dagger &= \mathbf{F}^\dagger\mathbf{d} \\ &= \mathbf{V}_p\mathbf{S}_p^{-1}\mathbf{U}_p^T\mathbf{d} \end{aligned} \tag{2.1.16}$$

This solution always exists, even if $\mathbf{F}$ is not of full column rank. In [2] it is shown that $\mathbf{m}^\dagger$ is a least squares solution. It is also shown that the solution given by the normal equations in 2.1.6 is a special case of $\mathbf{m}^\dagger$ when

- $\mathbf{F}$ has full column rank N
- but the range of $\mathbf{F}$ does not span the entire data space $\mathbb{R}^{N_d}$

This gave us a unique, but approximate solution in a least squares sense.

When $\mathbf{F}$ is not of full column rank, we get a solution that is not unique. This happens because $\mathbf{F}$ then has a nontrivial null space, $N(\mathbf{F})$[2]. Now the equation $\mathbf{Fm} = \mathbf{0}$ has more than one solution. In fact, there are infinitely many models, or solutions, $\mathbf{m}_0$ that satisfies $\mathbf{Fm}_0 = \mathbf{0}$. Adding such a solution to our solution would still satisfy $\mathbf{Fm} = \mathbf{d}$ because

$$\begin{aligned} \mathbf{F}(\mathbf{m} + \mathbf{m}_0) &= \mathbf{Fm} + \mathbf{Fm}_0 \\ &= \mathbf{d} + \mathbf{0} \\ &= \mathbf{d} \end{aligned} \tag{2.1.17}$$

The existence of a nontrivial null space leads to non uniqueness in the solution to a linear system of equations.

To see how very small singular values s_i can make the pseudo inverse solution unstable, they[2] also showed that

$$\begin{aligned} \mathbf{m}^\dagger &= \mathbf{V}_p \mathbf{S}_p^{-1} \mathbf{U}_p^T \mathbf{d} \\ &= \sum_{i=1}^{p} \frac{\mathbf{U}_{\cdot,i}^T \mathbf{d}}{s_i} \mathbf{V}_{\cdot,i} \end{aligned} \tag{2.1.18}$$

If the data vector contains random noise, small singular values in the denominator can blow up the corresponding coefficients, and thus let the noise dominate the solution. This gives us an unstable solution in the presence of noise in the data. As we know, all data or measurements comes with some amount of measurement errors, or noise. We need to stabilize the solution by regularization.

2.2 Regularization

We have seen that inverse problems often are hard to solve, and that the solution(s) may not reflect the nature of the true model. Different techniques called regularization are used in an attempt to obtain a meaningful solution to the inverse problem. We briefly look at the methods Truncated Singular Value Decomposition (TSVD) and the widely used Tikhonov regularization.

2.2.1 Truncated Singular Value Decomposition

As seen in (2.1.18), small singular values can create problems. An easy and straightforward way to make the solution more stable is to truncate the expression at some $p' < p$. Since the singular values are arranged in decreasing order this means that we discard the smallest values. The solution is then constructed by only the first p' singular values and corresponding model space basis vector. This stabilizes the solution, but this TSVD solution will not fit the data as well as solutions built from all the model space basis vectors. Stability is obtained at the expense of resolution. Understanding the sacrifice of resolution for stability is of fundamental importance when regularizing inverse problems.

2.2.2 Tikhonov Regularization

Here we present the Tikhonov regularization. We will look at the zeroth order Tikhonov regularization in particular, and show the idea for higher order Tikhonov regularization. The Tikhonov solution can be expressed in the terms of the singular value decomposition of $\mathbf{F}$. As we will see, this method does not discard any singular values. Instead, it gives greater weight to larger singular values and less weight to small singular values in the SVD.

We have considered the linear inverse problem $\mathbf{Fm} = \mathbf{d}$, and tried to find a solution that minimizes the misfit $\|\mathbf{Fm} - \mathbf{d}\|$. In zeroth order Tikhonov regularization, we add a regularization term and minimize the following expression

$$min \, \|\mathbf{Fm} - \mathbf{d}\|_2^2 + \alpha^2 \|\mathbf{m}\|_2^2 \tag{2.2.1}$$

where α is a regularization parameter. This is also called the damped least squares problem. To recognize the form of an ordinary least squares problem, we augment this and write

$$min \left\| \begin{bmatrix} \mathbf{F} \\ \alpha\mathbf{I} \end{bmatrix} \mathbf{m} - \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \right\|_2^2 \tag{2.2.2}$$

For $\alpha \neq 0$, equation 2.2.2 is a full rank least squares problem[2] that can be solved by the normal equations from 2.1.5. This gives us

$$\begin{bmatrix} \mathbf{F}^T \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{F} \\ \alpha\mathbf{I} \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{F}^T \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \tag{2.2.3}$$

which can be written as

$$(\mathbf{F}^T\mathbf{F} + \alpha^2\mathbf{I})\mathbf{m} = \mathbf{F}^T\mathbf{d} \tag{2.2.4}$$

This can be solved for different values of α. There are different ways of choosing the optimal α. We do not go into this here.

Using the SVD formulation, [2] shows that the solution $\mathbf{m}_\alpha$ is unique and can be expressed as

$$\mathbf{m}_\alpha = \sum_{i=1}^{min(N_d,N)} \frac{s_i^2}{s_i^2 + \alpha^2} \frac{\mathbf{U}_{.,i}^T\mathbf{d}}{s_i} \mathbf{V}_{.,i} \tag{2.2.5}$$

Here, all the singular values are included. The factors $f_i = \frac{s_i^2}{s_i^2+\alpha^2}$ are called filter factors, and determine the weighting of the singular values in the solution. We see that bigger values , $s_i \gg \alpha$ are weighted with $f_i \approx 1$, while smaller values $s_i \ll \alpha$ get $f_i \approx 0$. The filter factors decrease monotonically for decreasing singular values s_i.

Higher order Tikhonov regularization In the zeroth order Tikhonov regularization we minimize a function containing the term $\|\mathbf{m}\|_2^2$. This favors solutions where both the data misfit and the norm of the model $\mathbf{m}$ is minimized. Sometimes we may want to obtain solutions that minimize some other measure of $\mathbf{m}$, such as the norm of the first or second order derivative. In general, a matrix $\mathbf{L}$ is introduced in the damped least squares problem.

$$min\, \|\mathbf{Fm} - \mathbf{d}\|_2^2 + \alpha^2\|\mathbf{Lm}\|_2^2 \tag{2.2.6}$$

For first order Tikhonov regularization $\mathbf{L}$ can be a finite difference approximation to the first derivative. This would penalize solutions with high first derivatives, thus picking solutions that are relatively flat. Using second order Tikhonov regularization one may use a matrix $\mathbf{L}$ approximating the second derivative, thus favoring smooth solutions.

2.3 Nonlinear inverse problems

Until now we have only considered linear inverse problems. When solving nonlinear problems, we need other methods because the data are now related to the model parameters through a nonlinear system of equations. We will start off by introducing Newtons method, which provides the foundation for solving nonlinear problems. Then we will adapt the method so that we can solve nonlinear least squares problems where we try to minimize the data mismatch.

2.3.1 Newtons method

In Newtons method we look at a nonlinear system of n equations in n unknowns

$$\mathcal{F}(\mathbf{x}) = \mathbf{0} \tag{2.3.1}$$

The idea is to start with a initial guess $\mathbf{x}^0$ and compute a sequence of vectors, $\mathbf{x}^1$, $\mathbf{x}^2, \ldots$ that will converge to a solution $\mathbf{x}^\star$ of (2.3.1). If $\mathcal{F}$ is continuously differentiable, we can make a Taylor expansion of $\mathcal{F}$ about $\mathbf{x}^0$

$$\mathcal{F}(\mathbf{x}^0 + \Delta\mathbf{x}) \approx \mathcal{F}(\mathbf{x}^0) + \nabla\mathcal{F}(\mathbf{x}^0)\Delta\mathbf{x} \tag{2.3.2}$$

where $\nabla\mathcal{F}(\mathbf{x}^0)$ is the Jacobian of $\mathcal{F}$ evaluated at $\mathbf{x}^0$ with matrix entries

$$\big(\nabla\mathcal{F}(\mathbf{x}^0)\big)_{ij} = \frac{\partial\mathcal{F}_i(\mathbf{x}^0)}{\partial x_j} \tag{2.3.3}$$

From (2.3.2) we can compute a new approximate solution by

$$\mathcal{F}(\mathbf{x}^\star) = \mathbf{0} \approx \mathcal{F}(\mathbf{x}^0) + \nabla\mathcal{F}(\mathbf{x}^0)\Delta\mathbf{x} \tag{2.3.4}$$

with $\Delta\mathbf{x} = \mathbf{x}^\star - \mathbf{x}^0$. Solving this for the difference $\Delta\mathbf{x}$ gives

$$\nabla\mathcal{F}(\mathbf{x}^0)\Delta\mathbf{x} \approx -\mathcal{F}(\mathbf{x}^0) \tag{2.3.5}$$

Now the nonlinear system of equations is approximated with a linear system of equations. This can be solved using Gaussian elimination to produce the new vector $\mathbf{x}^1 = \mathbf{x}^0 + \Delta\mathbf{x}$. By performing this iteratively with the last solution as initial guess, one computes a sequence of vectors until it converges to a solution with $\mathcal{F}(\mathbf{x}) = \mathbf{0}$.

Newtons method often works very well. The assumptions is that $\mathcal{F}(\mathbf{x})$ is continuously differentiable about $\mathbf{x}^\star$ and that the matrix $\nabla\mathcal{F}(\mathbf{x})$ is non singular. But if these assumptions are not satisfied, or the initial guess $\mathbf{x}^0$ is not sufficiently close to $\mathbf{x}^\star$, the method may converge very slowly or even fail.

We will now look at Newtons method for minimizing. We consider a scalar valued function $f(\mathbf{x})$ that we want to minimize. If $f(\mathbf{x})$ is twice continuously differentiable, a Taylor series expansion is as follows

$$f(\mathbf{x}^0 + \Delta\mathbf{x}) \approx f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T\nabla^2 f(\mathbf{x}^0)\Delta\mathbf{x} \tag{2.3.6}$$

where $\nabla f(\mathbf{x}^0)$ is the gradient of f at $\mathbf{x}^0$ with vector entries

$$\big(\nabla f(\mathbf{x}^0)\big)_i = \frac{\partial f(\mathbf{x}^0)}{\partial x_i} \tag{2.3.7}$$

and $\nabla^2 f(\mathbf{x}^0)$ is the Hessian of f at $\mathbf{x}^0$ with matrix entries

$$\big(\nabla^2 f(\mathbf{x}^0)\big)_{ij} = \frac{\partial f(\mathbf{x}^0)}{\partial x_i \partial x_j} \tag{2.3.8}$$

To find a solution $\mathbf{x}^\star$ that minimizes $f(\mathbf{x})$ we demand that $\nabla f(\mathbf{x}^\star) = \mathbf{0}$. An approximation of the gradient ∇f nearby $\mathbf{x}^0$ is

$$\nabla f(\mathbf{x}^0 + \Delta\mathbf{x}) \approx \nabla f(\mathbf{x}^0) + \nabla^2 f(\mathbf{x}^0)\Delta\mathbf{x} \tag{2.3.9}$$

An approximate solution to $\nabla f(\mathbf{x}^0 + \Delta\mathbf{x}) = \mathbf{0}$ is then

$$\nabla^2 f(\mathbf{x}^0)\Delta\mathbf{x} = -\nabla f(\mathbf{x}^0) \tag{2.3.10}$$

We see that Newtons method for minimizing $f(\mathbf{x})$ is the same as equation (2.3.5) applied to $\nabla f(\mathbf{x}) = \mathbf{0}$. We will now see how nonlinear least squares problems can be solved using modified versions of Newtons minimization method.

2.3.2 Gauss-Newton and Levenberg-Marquardt methods

Consider now a general nonlinear inverse problem

$$\mathbf{F}(\mathbf{m}) = \mathbf{d} \tag{2.3.11}$$

In most problems there are not a equal number of data and parameters, and there may not be an exact solution to $\mathbf{F}(\mathbf{m}) = \mathbf{d}$. Next we use Newtons method to minimize a nonlinear least squares problem. A vector $\mathbf{d}$ of N_d data is given, along with a vector of the standard deviations $\boldsymbol{\sigma}$. The task is to find a solution $\mathbf{m}$ that minimizes the residuals $f_i(\mathbf{m}) = \frac{F(\mathbf{m})_i - d_i}{\sigma_i}$ in a 2-norm sense. Assume that the measurement errors are normally distributed. As in equation (2.1.11), the maximum likelihood principle tells us to minimize the sum of the squared errors divided by their standard deviations. We try to minimize what we in nonlinear problems call the objective function

$$f(\mathbf{m}) = \sum_{i=1}^{N_d} \left(\frac{F(\mathbf{m})_i - d_i}{\sigma_i} \right)^2 \tag{2.3.12}$$

where we let

$$f_i(\mathbf{m}) = \frac{F(\mathbf{m})_i - d_i}{\sigma_i} \tag{2.3.13}$$

so that

$$f(\mathbf{m}) = \sum_{i=1}^{N_d} f_i(\mathbf{m})^2 \tag{2.3.14}$$

To find the gradient, we write it as the sum of the gradients of the individual terms

$$\nabla f(\mathbf{m}) = \nabla \left(f_1(\mathbf{m})^2 \right) + \cdots + \nabla \left(f_{N_d}(\mathbf{m})^2 \right) \tag{2.3.15}$$

The $j'th$ entry of the gradient contains all the first derivatives of f with respect to the $j'th$ parameter m_j. Using the chain rule, we get

$$\nabla f(\mathbf{m})_j = \sum_{i=1}^{N_d} 2 \nabla f_i(\mathbf{m})_j \mathcal{F}(\mathbf{m})_j \tag{2.3.16}$$

where

$$\mathcal{F}(\mathbf{m}) = \begin{bmatrix} f_1(\mathbf{m}) \\ \vdots \\ f_{N_d}(\mathbf{m}) \end{bmatrix} \tag{2.3.17}$$

In matrix notation this becomes

$$\nabla f(\mathbf{m}) = 2\mathbf{J}(\mathbf{m})^T \mathcal{F}(\mathbf{m}) \tag{2.3.18}$$

where $\mathbf{J}(\mathbf{m})$ is the Jacobian

$$\mathbf{J}(\mathbf{m}) = \begin{bmatrix} \frac{\partial f_1(\mathbf{m})}{\partial m_1} & \cdots & \frac{\partial f_1(\mathbf{m})}{\partial m_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{N_d}(\mathbf{m})}{\partial m_1} & \cdots & \frac{\partial f_{N_d}(\mathbf{m})}{\partial m_N} \end{bmatrix} \tag{2.3.19}$$

We can also express the Hessian of $f(\mathbf{m})$ in a similar way.

$$\begin{aligned} \nabla^2 f(\mathbf{m}) &= \sum_{i=1}^{N_d} \nabla^2 \left(f_i(\mathbf{m})^2 \right) \\ &= \sum_{i=1}^{N_d} \mathbf{H}^i(\mathbf{m}) \end{aligned} \tag{2.3.20}$$

where $\mathbf{H}^i(\mathbf{m})$ is the Hessian of $f_i(\mathbf{m})^2$. Writing out the j , k element of $\mathbf{H}^i(\mathbf{m})$ would give us

$$\begin{aligned} H^i_{j,k} &= \frac{\partial^2 (f_i(\mathbf{m})^2)}{\partial m_j \partial m_k} \\ &= \frac{\partial}{\partial m_j}\left(2 f_i(\mathbf{m}) \frac{\partial f_i(\mathbf{m})}{\partial m_k}\right) \\ &= 2\left(\frac{\partial f_i(\mathbf{m})}{\partial m_j}\frac{\partial f_i(\mathbf{m})}{\partial m_k} + f_i(\mathbf{m})\frac{\partial^2 f_i(\mathbf{m})}{\partial m_j \partial m_k}\right) \end{aligned} \tag{2.3.21}$$

Note here that the second term of the Hessian involves $f_i(\mathbf{m})$.

Using the Jacobian from above, $\nabla^2 f(\mathbf{m})$ is

$$\nabla^2 f(\mathbf{m}) = 2\mathbf{J}(\mathbf{m})^T \mathbf{J}(\mathbf{m}) + \mathbf{Q}(\mathbf{m}) \tag{2.3.22}$$

with

$$\mathbf{Q}(\mathbf{m}) = 2\sum_{i=1}^{N_d} f_i(\mathbf{m}) \nabla^2 f_i(\mathbf{m}) \tag{2.3.23}$$

When minimizing $f(\mathbf{m})$ it seems reasonable to expect that the terms $f_i(\mathbf{m})$ are small as we get closer to the optimal solution $\mathbf{m}^\star$. In the Gauss-Newton method (GN), the last term of the Hessian is therefore ignored, and the Hessian is approximated by

$$\nabla^2 f(\mathbf{m}) \approx 2\mathbf{J}(\mathbf{m})^T \mathbf{J}(\mathbf{m}) \tag{2.3.24}$$

The equations to be solved iteratively in the GN method is then given as

$$\mathbf{J}(\mathbf{m}^k)^T \mathbf{J}(\mathbf{m}^k) \Delta \mathbf{m} = -\mathbf{J}(\mathbf{m}^k)^T \mathcal{F}(\mathbf{m}^k) \tag{2.3.25}$$

with $\Delta \mathbf{m} = \mathbf{m}^{k+1} - \mathbf{m}^k$. This method often works well. However, it is based on Newtons method, and therefore relies on similar assumptions. If the matrix $\mathbf{J}(\mathbf{m}^k)^T \mathbf{J}(\mathbf{m}^k)$ is singular, the method may fail. Also, the approximation of the Hessian in GN will not be valid if the terms $f_i(\mathbf{m})$ are large.

Another modification is introduced in the Levenberg-Marquardt (LM) method. Here, a positive parameter λ is adjusted during the iterations to ensure convergence when solving the linear systems. The equations in LM are given by

$$(\mathbf{J}(\mathbf{m}^k)^T \mathbf{J}(\mathbf{m}^k) + \lambda \mathbf{I}) \Delta \mathbf{m} = -\mathbf{J}(\mathbf{m}^k)^T \mathcal{F}(\mathbf{m}^k) \tag{2.3.26}$$

The $\lambda\mathbf{I}$ term makes the matrix non singular and stabilizes the the solution of the linear system in each iteration. It is not a form of regularization, since it only improves the convergence of the algorithm solving the linear systems. One challenge is to determine optimal values of λ. For large values of λ

$$\mathbf{J}(\mathbf{m}^k)^T\mathbf{J}(\mathbf{m}^k) + \lambda\mathbf{I} \approx \lambda\mathbf{I} \tag{2.3.27}$$

and

$$\Delta\mathbf{m} \approx -\frac{1}{\lambda}\nabla f(\mathbf{m}) \tag{2.3.28}$$

The gradient points in the direction of largest growth, so this is a steepest-descent step. The algorithm then moves a small step down-gradient to reduce $f(\mathbf{m})$. This gives slow but certain convergence.

For very small values of λ, the LM method approaches the GN method. This can provide potentially fast but uncertain convergence. A general strategy is to use small λ-values when the GN method works well, and change to larger values when the convergence properties of the LM method is required.

Both the GN- and the LM-method are designed to find a minimum of the objective function. When dealing with nonlinear inverse problems, there might be several local or global minimum of the objective function, and we can not be certain that the method converges to a global minimum. Different methods are developed to deal with this issue. One approach is to use a multistart method[2]. Several initial guesses are generated randomly, and the LM method is performed on each of these. The resulting local minimum solutions are then compared, and the one with the smallest value of $f(\mathbf{m})$ is selected.

2.3.3 Summary Classical Approach

Before we go on with the Bayesian approach we summarize some of the theory of the traditional approach to inverse problems.

For well-conditioned linear problems, with assumed independent and normally distributed data errors, the theory is well developed. Here the solution $\mathbf{m}_{L_2}$ is given by solving a least squares problem, minimizing the L_2-norm of the misfit, $\|\mathbf{Fm} - \mathbf{d}\|_2$.

When the linear problem is ill-conditioned the set of solutions may become large and diverse, and many of the models can be physically unreasonable. It is important to understand that the ill-posedness is the nature of the problem itself and not the solution procedure.

Producing a usable solution is possible by imposing additional constraints through regularization. These techniques penalizes certain properties of the solutions, giving a best pick from the solution set that fit the data sufficiently well. Regularization of

a problem stabilizes the solution, but at the cost of model resolution and introducing bias. We have to choose between a stable solution and fitting the data.

Both linear and nonlinear problems can be regularized. For solving linear problems one might use the SVD (singular value decomposition) or the CGLS[2] (conjugate gradient least squares) method. Nonlinear problems are more complicated to solve. Methods like GN (Gauss-Newton) or LM (Levenberg-Marquardt) can be used to find a minimum of the resulting nonlinear least square problem. However, there may be several local minimum solutions for a nonlinear problem, and finding the global minimum can be very hard. For more on inverse problems and solution methods, see [25, 2].

2.4 Bayesian approach

In the classical approach we assumed that there is one true model that we want to find. Bayesian techniques use a totally different view based on probability theory. We here consider a general inverse problem

$$\mathbf{F}(\mathbf{m}) = \mathbf{d} \tag{2.4.1}$$

We are still aware of that there are errors present, $\boldsymbol{\eta}$. We therefore write

$$\mathbf{d} = \mathbf{F}(\mathbf{m}) + \boldsymbol{\eta} \tag{2.4.2}$$

The error may originate from modeling through $\mathbf{F}$, from measurement errors in the data or from both.

It is called Bayesian because the solution approach is based on Bayes Theorem, see the Appendix (A.7). In the Bayesian approach the model is treated as a random variable, and the solution itself is a probability distribution for the model parameters. Once we have this probability distribution, we can use it to answer probabilistic questions about the model, such as, "What is the probability that m_5 is less than 1?". With the classical approach such questions do not make sense, since the model we try to find is not assumed to be a random variable.

Another important difference from the classical approach is that in Bayesian theory one can incorporate prior information or knowledge about the model using a prior distribution. To make it easier to use in computations it is often assumed that the distributions involved are Gaussian. Assuming that the prior is Gaussian distributed, and therefore completely described by its mean $\mathbf{m}_{pr}$ and covariance $\mathbf{C}_m$, we can write

$$p(\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{m}-\mathbf{m}_{pr})^T\mathbf{C}_m^{-1}(\mathbf{m}-\mathbf{m}_{pr})\right) \tag{2.4.3}$$

Given a model $\mathbf{m}$, a likelihood function $f(\mathbf{d} \mid \mathbf{m})$ for the observed data can be defined. If the modeling and measurement errors are assumed Gaussian and independent of each other it can be shown[26] that the likelihood function can be written as

$$f(\mathbf{d} \mid \mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{F}(\mathbf{m})-\mathbf{d})^T\mathbf{\Sigma}^{-1}(\mathbf{F}(\mathbf{m})-\mathbf{d})\right) \tag{2.4.4}$$

where the covariance matrix $\mathbf{\Sigma}$ combines the modeling and the measurement errors. We used a similar formulation in the maximum likelihood estimation section 2.1.2. But then prior information about $\mathbf{m}$ was ignored. If such a prior is available, Bayesian theory can be used to incorporate the information into the solution.

In a Bayesian framework the solution to the inverse problem is given by the posterior distribution, which takes into account both the likelihood function for the data and the prior. The collected data is combined with the prior trough Bayes Theorem to produce the posterior distribution for the model parameters. The prior distribution is here denoted by $p(\mathbf{m})$ and the likelihood by $f(\mathbf{d} \mid \mathbf{m})$. The latter is the probability that, given a model $\mathbf{m}$, the data $\mathbf{d}$ will be observed. It is assumed that this conditional distribution can be computed. The posterior distribution for the model, given the data is

$$q(\mathbf{m} \mid \mathbf{d}) = \frac{f(\mathbf{d} \mid \mathbf{m})p(\mathbf{m})}{\int_{all\,models} f(\mathbf{d} \mid \mathbf{m})p(\mathbf{m})d\mathbf{m}}. \tag{2.4.5}$$

The denominator here scales the posterior distribution $q(\mathbf{m} \mid \mathbf{d})$ so that its integral equals 1, integrating over all the data

$$\int f(\mathbf{d})\, d\mathbf{d} = \int_{all\,models} f(\mathbf{d} \mid \mathbf{m})p(\mathbf{m})d\mathbf{m} \tag{2.4.6}$$

with $p(\mathbf{m})$ being the prior. The integrals in the denominator may be very hard to compute, but it is not always necessary to do. Often we just write

$$q(\mathbf{m} \mid \mathbf{d}) \propto f(\mathbf{d} \mid \mathbf{m})p(\mathbf{m}) \tag{2.4.7}$$

If the distributions are Gaussian, and the errors are independent and normally distributed, then the posterior is Gaussian as well and can be written as

$$\begin{aligned} q(\mathbf{m} \mid \mathbf{d}) &\propto exp\left(-\frac{1}{2}(\mathbf{F}(\mathbf{m})-\mathbf{d})^T\mathbf{\Sigma}^{-1}(\mathbf{F}(\mathbf{m})-\mathbf{d}) - \frac{1}{2}(\mathbf{m}-\mathbf{m}_{pr})^T\mathbf{C}_m^{-1}(\mathbf{m}-\mathbf{m}_{pr})\right) \\ &= exp\left(-O(\mathbf{m})\right) \end{aligned} \tag{2.4.8}$$

where $O(\mathbf{m})$ is the objective function

$$O(\mathbf{m}) = \frac{1}{2}(\mathbf{F}(\mathbf{m})-\mathbf{d})^T\mathbf{\Sigma}^{-1}(\mathbf{F}(\mathbf{m})-\mathbf{d}) + \frac{1}{2}(\mathbf{m}-\mathbf{m}_{pr})^T\mathbf{C}_m^{-1}(\mathbf{m}-\mathbf{m}_{pr}) \tag{2.4.9}$$

A highly likely model $\mathbf{m}$ will give high values for the posterior distribution. In the same way it will give low values in the objective function. The model with the largest value of $q(\mathbf{m} \mid \mathbf{d})$ is referred to as the maximum a posteriori (MAP) model. This MAP model then also minimizes the objective function given in (2.4.9). In cases where we want to single out one model to be the answer we may use the MAP model. An alternative would be to use the mean of the posterior distribution. The MAP model and the posterior mean model are identical when the posterior distribution is Gaussian.

For a linear inverse problem, $\mathbf{Fm} = \mathbf{d}$, if both the prior and likelihood are Gaussian, it can be shown[12] that the posterior is Gaussian with mean value

$$\mathbf{m}_{MAP} = \mathbf{m}_{pr} + \mathbf{C}_m\mathbf{F}^T(\mathbf{F}\mathbf{C}_m\mathbf{F}^T + \mathbf{\Sigma})^{-1}(\mathbf{d} - \mathbf{F}\mathbf{m}_{pr}) \tag{2.4.10}$$

and covariance

$$\mathbf{C}_{MAP} = \mathbf{C}_m - \mathbf{C}_m\mathbf{F}^T(\mathbf{F}\mathbf{C}_m\mathbf{F}^T + \mathbf{\Sigma})^{-1}\mathbf{F}\mathbf{C}_m \tag{2.4.11}$$

where $\mathbf{F}$ is the linear operator.

When the forward model is nonlinear, it cannot longer be assumed that the posterior distribution is Gaussian. It may be the case that the posterior distribution has multiple modes, leading to multiple models $\mathbf{m}$ with high probabilities. This means that the corresponding objective function may have multiple local or global minima.

One way to explore the posterior distribution is by sampling. Sampling is done by randomly drawing(A.6) a large number of realizations from the appropriate distribution to form a suite or ensemble of realizations, or samples . These realizations are then used to represent an approximation to the posterior distribution. There are several different sampling techniques, but we do only mention one of them here, the Randomized Maximum Likelihood, RML. The reader may refer to [24, 22] for more on sampling methods.

2.4.1 Randomized Maximum Likelihood

If the prior covariance of the model parameters, $\mathbf{C}_m$, and the data error covariance $\mathbf{\Sigma}$ are known, then samples from the posterior distribution can be generated by the RML. To sample the posterior distribution $q(\mathbf{m} \mid \mathbf{d})$ we want to find a model $\mathbf{m}_c$ that minimizes the objective function (2.4.9) for sampled values from the prior distribution $p(\mathbf{m})$ and the data likelihood distribution $f(\mathbf{d} \mid \mathbf{m})$. First, instead of using $\mathbf{m}_{pr}$ in (2.4.9), we generate a sample $\mathbf{m}_{uc}$ from $\mathbf{p}(\mathbf{m})$ that is not conditioned to the data. Then a sample $\mathbf{d}_{uc}$ from $f(\mathbf{d} \mid \mathbf{m})$ with added measurement errors is generated. Do as follows

1. Generate an unconditional sample $\mathbf{m}_{uc} \sim N(\mathbf{m}_{pr}, \mathbf{C}_m)$ from the prior distribution
2. Generate an unconditional sample $\mathbf{d}_{uc} \sim N(\mathbf{d}, \mathbf{\Sigma})$ from the data likelihood distribution
3. Find the conditional sample model $\mathbf{m}_c$ that minimizes

$$O(\mathbf{m}) = \frac{1}{2}(\mathbf{F}(\mathbf{m}) - \mathbf{d}_{uc})^T \mathbf{\Sigma}^{-1} (\mathbf{F}(\mathbf{m}) - \mathbf{d}_{uc}) + \frac{1}{2}(\mathbf{m} - \mathbf{m}_{uc})^T \mathbf{C}_m^{-1} (\mathbf{m} - \mathbf{m}_{uc}) \tag{2.4.12}$$

This produces a single sample model from the posterior distribution. To generate additional samples, repeat the process with different sets of ($\mathbf{m}_{uc}$, $\mathbf{d}_{uc}$).

For linear problems, RML samples correctly when errors are added to the data as in step 2[22].

If the problem is nonlinear, then RML is an approximate sampling method. In the nonlinear case, the minimization process in step 3 has to be performed by an iterative minimization method.

2.5 History Matching; A Nonlinear Inverse Problem in Reservoir Engineering

We have now introduced inverse problems and seen that they can be solved by a classical approach, or within a Bayesian framework. From classical theory we arrive at an objective function with no prior included. The Bayesian formulation of the inverse problem however, includes the prior distribution for the parameters.

We now look briefly at a well known inverse problem in reservoir engineering, namely history matching. The task is to update the parameters and state variables of the reservoir model, based on data that are available. For illustrative purposes we shortly mention the reservoir model and write up a general version of the governing flow equations. For more on the subject we refer to [3, 16, 22].

2.5.1 The reservoir model and flow equations

The reservoir model is an approximation of the reservoir itself. It is discretized in space and may consist of several hundred thousand cells or grid blocks. Each cell is assigned with values for the different static properties, or parameters. Two of the most important parameters are permeability and porosity. Permeability is the inverse of flow resistance, and describes how easily fluids can flow through the porous medium. It is defined through Darcy's law, see (2.5.2), and must be determined experimentally. Porosity is the percentage of the rock volume that can be filled with

fluids, and is therefore the storage capacity of the porous medium. It is defined as the volume of the connected pore space V_p divided by the total volume V of the porous medium

$$\phi = \frac{V_p}{V}. \tag{2.5.1}$$

Most petroleum reservoirs are buried deep underground, making direct assessment of the reservoir properties difficult. Before production starts, we may have sparsely distributed local point measurements, such as core plugs and well log data from the wells. Additionally, we may have spatially distributed, but rather imprecise data from seismic surveys. Thus, the properties of the reservoir are associated with high uncertainties.

To compute the state variables; fluid pressures and saturations, fundamental flow equations are applied to each of the cells in the reservoir model.

The fluid flow in the reservoir can be described by Darcy's law and the principle of mass conservation. Darcy's law is an empirical equation, relating the filtration velocity, or Darcy velocity $\mathbf{u}$ to the pressure gradient ∇p, and can be written for one-phase flow as

$$\mathbf{u} = -\frac{1}{\mu}\mathbf{K}\left(\nabla p - \rho \mathbf{g}\right). \tag{2.5.2}$$

In (2.5.2), μ and ρ is the viscosity and the density of the fluid respectively, and $\mathbf{g}$ is the gravitational acceleration. $\mathbf{K}$ is the absolute permeability of the porous medium. In general, permeability may vary in direction and in space for a anisotropic and heterogeneous medium. Permeability is therefore represented by a second order tensor. In most practical applications, $\mathbf{K}$ is assumed to be diagonal [3]

$$\mathbf{K} = \begin{pmatrix} K_x(\mathbf{x}) & & \\ & K_y(\mathbf{x}) & \\ & & K_z(\mathbf{x}) \end{pmatrix}. \tag{2.5.3}$$

Darcy's law can be extended for multiphase flow. With several phases present, we define the saturation S_i and effective permeability $\mathbf{K}_i$ for each phase nr. i

$$S_i = \frac{V_i}{V_p}\,,\ \mathbf{K}_i = k_{r,i}\mathbf{K} \tag{2.5.4}$$

for $i = 1, \ldots n$.

Here V_p is the pore volume, V_i is the volume occupied by phase nr. i, $\mathbf{K}$ is the absolute permeability of the porous medium and $k_{r,i}$ is the relative permeability of phase nr. i. The relative permeability $k_{r,i}$ depends non linearly on the saturation S_i, and is therefore usually denoted as $k_{r,i}(S_i)$.

It is assumed that the pores of the medium are completely filled with fluids, i.e.

$$\sum_{i=1}^{n} S_i = 1. \tag{2.5.5}$$

The Darcy velocities for each phase is then

$$\begin{aligned}\mathbf{u}_i &= -\frac{k_{r,i}(S_i)}{\mu_i}\mathbf{K}\left(\nabla p_i - \rho_i \mathbf{g}\right) \\ &= -\lambda_i \mathbf{K}\left(\nabla p_i - \rho_i \mathbf{g}\right)\end{aligned} \tag{2.5.6}$$

with the mobility for each phase defined as $\lambda_i = \frac{k_{r,i}(S_i)}{\mu_i}$.

Using conservation of mass on each of the phases we can write[16]

$$\frac{\partial\left(\phi\rho_i S_i\right)}{\partial t} + \nabla\cdot\left(\rho_i \mathbf{u}_i\right) = G_i \tag{2.5.7}$$

where ϕ is the porosity and $\phi\rho_i S_i$ is the mass of phase nr. i relative to the volume of the cell. G_i is the source term for phase i. G is positive in the case of an injection well and negative in case of a production well. Without any sources or sinks, then $G = 0$.

Combining (2.5.6) and (2.5.7) produces a system of second order partial differential equations

$$\frac{\partial\left(\phi\rho_i S_i\right)}{\partial t} - \nabla\cdot\left(\rho_i\left(\lambda_i \mathbf{K}\left(\nabla p_i - \rho_i \mathbf{g}\right)\right)\right) = G_i. \tag{2.5.8}$$

Together with (2.5.5), state equations $\rho_i(p_i)$ and $\mu_i(p_i)$, relative permeability curves $k_{r,i}(S_i)$ and capillary pressure relations between the phase pressures p_i, this makes a complete system of equations. The relations for capillary pressure P_c are usually empirical.

Given specified initial and boundary conditions this can in theory be solved for the dynamic variables p_i and S_i. However, the equations constituting the mathematical model of the reservoir are almost always too complex to be solved analytically. Instead, they must be approximated by e.g. a finite difference formulation[3] to form a numerical model.

In reservoir simulation, the flow equations (here illustrated by (2.5.8)) are solved numerically by a reservoir flow simulator. It is assumed that the initial reservoir conditions are known. This is defined by the dynamic state variables, pressure and saturation of the different fluid phases, as well as the initial fluid contacts between water and oil (WOC) and between gas and oil (GOC).

The simulator takes the current reservoir state and the recovery strategy for the wells as input, and provides us with the computed state variables for all cells, as well as simulated data. The simulated data is typically bottom hole pressure in the wells, water cut, gas-oil ratio and total oil production over time.

2.5.2 History matching

To make good future predictions of the reservoir performance, it is important to quantify the reservoir properties. If the properties where known exactly we would expect that the observed data could be reproduced by running the simulator. Therefore it seems reasonable to condition the reservoir model to the observed data.

Originally, the history matching was done manually, by adjusting the parameters of the model and rerun the simulator from start to check if the history match improved. This work relied mainly on the experience of the reservoir engineer and only one or two parameters were changed at the same time. Normally one waited several years after production start, and performed the history matching on a campaign basis.

Automatic history matching has been subject to extensive research. Typically, one attempts to minimize the square of the mismatch between all computed data and observed data, and/or the square of the mismatch between the current model parameters and the prior model parameters, see (2.4.9). There are developed many minimization algorithms to optimize the objective function. Traditional methods are gradient-based methods where the gradient of the objective function is calculated, as well as the optimal length of the search step. Such methods include the Gauss-Newton and Levenberg-Marquardt methods, along with more sophisticated ones.

Automatic history matching is performed in a loop.

1. simulate the entire history matching period
2. minimize the objective function and evaluate it
   (a) if not satisfactory match is reached, return to 1.
   (b) if satisfactory match is reached, end loop

The normal procedure is to incorporate or assimilate the data simultaneously, instead of sequentially. This means that data are assimilated all at once, with regular intervals, rather than only assimilating the newest data as soon as they arrive. It also means that the simulator is run for the entire production history every time. Waiting a long time between each history matching, implies that the model is not consistent with the newest data. This may affect the quality of the predictions from the model.

Reasons for not assimilating data sequentially as they arrive is typically the large computationally effort involved when very large matrices has to be updated every time new data are incorporated. Performing the data assimilation sequentially would then require a considerable amount of work every time new data arrived, instead of history matching the data in batch.

EnKF is a promising method for performing the history matching sequentially. The fact that the covariance matrix never has to be computed explicitly makes the method feasible for sequential assimilation.

In the next chapter, we revisit the Bayesian formulation of the inverse problem, and see how the solution can be computed sequentially, which is done in the EnKF.

Chapter 3

Sequential data assimilation

EnKF is a sequential data assimilation method that can be used to update both the static parameters and the dynamic state variables of a system. It was originally used for updating only the state variables, but is now also used to update the combined problem. We present the EnKF at the end of this chapter. First we take a look at the combined parameter and state estimation problem, and how the Bayes theorem formulation of the inverse problem can be reformulated into a sequential form. Then we present the Kalman Filter for linear systems, which lays the foundation for EnKF. We also mention the Extended Kalman Filter.

3.1 Combined parameter and state estimation problem

The inverse problem can consist of estimating either the dynamical state variables or the static parameters of the system, or both. In some applications, only the dynamic variables are estimated, such as in weather forecasting. In other cases, the static parameters are estimated, and then used to compute the dynamic variables.

Following [7, 24] we now redefine the Bayesian formulation of the inverse problem to also involve the combined parameter and state estimation problem.

Let the augmented state vector $\mathbf{y}(\mathbf{x},t) = [\mathbf{u}(\mathbf{x},t)\,,\,\mathbf{m}(\mathbf{x})]^T$ contain the dynamical state variables $\mathbf{u}(\mathbf{x},t)$ and the static parameters $\mathbf{m}(\mathbf{x})$. The static parameters are assumed to be constant in time. A joint prior probability distribution $f(\mathbf{y})$ can be defined for the unknown parameters and state variables where

$$f(\mathbf{y}) = f(\mathbf{u}\,|\mathbf{m})f(\mathbf{m}) \tag{3.1.1}$$

Here $f(\mathbf{m})$ is the prior for the parameters and $f(\mathbf{u} \mid \mathbf{m})$ symbolizes that the dynamic variables are calculated when running the forward problem, given the parameters $\mathbf{m}$.

Going back to Bayes theorem in (2.4.7) we write

$$q(\mathbf{y}\,|\mathbf{d}) \propto f(\mathbf{d} \mid \mathbf{y})f(\mathbf{y}) \tag{3.1.2}$$

with $f(\mathbf{d} \mid \mathbf{y})$ being the likelihood distribution for the data.

3.1.1 Sequential formulation

We let the model state be discretized in time, $t = 0, 1, \ldots k$ so that $\mathbf{y}_i(\mathbf{x}) = \mathbf{y}(\mathbf{x}, t_i)$. $\mathbf{y}_0$ represents the specified initial conditions of the system. Instead of $f(\mathbf{y})$ we write $f(\mathbf{y}_k, \ldots \mathbf{y}_1, \mathbf{y}_0)$ and use the notation $f(\mathbf{y}_{k:0})$ for this. Written out as

$$\begin{aligned} f(\mathbf{y}_{k:0}) &= f(\mathbf{u}_k, \ldots \mathbf{u}_1, \mathbf{u}_0, \mathbf{m}) \\ &= f(\mathbf{u}_{k:1} \mid \mathbf{u}_0, \mathbf{m}) f(\mathbf{u}_0) f(\mathbf{m}) \end{aligned} \tag{3.1.3}$$

we see that this means the distribution for the state variables at all the discrete set of times given the static parameters and the initial conditions. From now on we only use the notation $\mathbf{y}$ and refer to this as the model state for the rest of this subsection.

Now assume that we have available data $\mathbf{d}_i$ at the same discrete set of times as the model state. Evensen showed in [5] that the general expression in (3.1.2) can be formulated in a sequential form

$$q(\mathbf{y}_{k:0} \mid \mathbf{d}_{k:1}) \propto f(\mathbf{d}_k \mid \mathbf{y}_{k:0}) f(\mathbf{y}_k \mid \mathbf{y}_{k-1:0}) f(\mathbf{y}_{k-1:0} \mid \mathbf{d}_{k-1:1}) \tag{3.1.4}$$

If the model state at time k only depends on the model state at the previous time $k-1$ it is called a first order Markov process. Propagating the model state forward in time from $k-1$ to k is then written as $f(\mathbf{y}_k \mid \mathbf{y}_{k-1})$. Assuming this, and also that the measurement errors are uncorrelated and that the data set $\mathbf{d}_i$ only depends on the corresponding model state $\mathbf{y}_i$, we can write this as

$$q(\mathbf{y}_{k:0} \mid \mathbf{d}_{k:1}) \propto f(\mathbf{d}_k \mid \mathbf{y}_k) f(\mathbf{y}_k \mid \mathbf{y}_{k-1}) f(\mathbf{y}_{k-1:0} \mid \mathbf{d}_{k-1:1}) \tag{3.1.5}$$

This is the solution for all the model states $\mathbf{y}_{k:0}$, updated with the data, for the time interval $t \in [0, k]$. It is called a smoother solution. It involves the state models defined from the initial time to the current update time. This way the variables are updated backwards in time. The first factor on the right side is the likelihood function. As mentioned, the second factor corresponds to propagating the solution forward in time, and the last factor is the posterior distribution from time $k-1$. Data can then be assimilated sequentially when they are available.

The filter solution arises when we want to evaluate only the current state at time k of the system, and is given by

$$q(\mathbf{y}_k \mid \mathbf{d}_{k:1}) \propto f(\mathbf{d}_k \mid \mathbf{y}_k) f(\mathbf{y}_k \mid \mathbf{d}_{k-1:1}) \tag{3.1.6}$$

This is a special case of the smoother solution, where the updates at previous times are left out. The filter equation can be derived by integrating over the solutions $\mathbf{y}_{k-1:0}$[24].

Given the data $\mathbf{d}_{k:1}$, we are often interested in estimating the current model state $\mathbf{y}_k$. Smoothing and prediction of $\mathbf{y}_k$ may also be done. We mention

- Filter solution; $q(\mathbf{y}_k \mid \mathbf{d}_{k:1})$
- Smoother solution; $q(\mathbf{y}_k \mid \mathbf{d}_{l:1})$ with $l > k$
- Predicter solution; $q(\mathbf{y}_k \mid \mathbf{d}_{j:1})$ with $j < k$

Thus, smoothing is estimating past states given past and present data, while prediction is the estimation of future states given past and present data.

From now on we focus on the filter solutions. We introduce the different Kalman filters. In the following we refer to the prior distribution as the forecasted estimate $\mathbf{y}_k^f$, and to the posterior distribution as the analyzed estimate $\mathbf{y}_k^a$.

3.2 Linear models - Kalman Filter

The Kalman Filter (KF)[13] is a data assimilating algorithm solving the Bayesian estimation update problem for systems with linear dynamics and assumed Gaussian distributions. It uses prior information about a linear system combined with measurements containing noise to produce the best estimate of the state of the system.

The method consists of two steps. The first is the forecast step where both the model and the model covariance is computed for the next time step k, given the analyzed estimate at the previous time step $k-1$. Also, the superscripts f and a denotes "forecasted" and "analyzed" respectively.

Forecast step

$$\mathbf{y}_k^f = \mathbf{A}\mathbf{y}_{k-1}^a + \mathbf{w}_{k-1} \tag{3.2.1}$$

This is the forward evolution in time of the state variables $\mathbf{y}$ by the linear relationship given by $\mathbf{A}$. The modeling error $\mathbf{w}_{k-1}$ is assumed to be independent in time and Gaussian distributed with zero mean and covariance $\mathbf{\Omega}_{k-1}$, thus we write $\mathbf{w}_{k-1} \sim N(0, \mathbf{\Omega}_{k-1})$. Since the two states are linear related we express the covariance of $\mathbf{y}_k^f$ as

$$\mathbf{C}_k^f = \mathbf{A}\mathbf{C}_{k-1}^a\mathbf{A}^T + \mathbf{\Omega}_{k-1}. \tag{3.2.2}$$

The result in (3.2.2) is shown in (A.4).

The measurements at time k are linearly related to the state vector through

$$\mathbf{d}_k = \mathbf{H}_k\mathbf{y}_k^f + \mathbf{v}_k \tag{3.2.3}$$

where $\mathbf{H}$ is the observation matrix and $\mathbf{v}_k$is the measurement error which is assumed to be independent in time and Gaussian distributed with zero mean and covariance $\mathbf{\Sigma}_k$. $\mathbf{v}_k \sim N(0, \mathbf{\Sigma}_k)$. Together with the Gaussian prior and assuming that $\mathbf{y}_k$ only depends on $\mathbf{y}_{k-1}$, this is enough to make the posterior distribution Gaussian as well. The model is advanced forward in time to the next step where we have available measurements. In the analysis step the observed data are assimilated and a new weighted estimate based on the uncertainties is made. The new estimate is made through the Kalman Gain matrix $\mathbf{K}_k$ such that the posterior error covariance is minimized[19].

Analysis step

$$\mathbf{y}_k^a = \mathbf{y}_k^f + \mathbf{K}_k(\mathbf{d}_k - \mathbf{H}_k\mathbf{y}_k^f) \tag{3.2.4}$$

where the Kalman Gain is given by

$$\mathbf{K}_k = \mathbf{C}_k^f\mathbf{H}_k^T(\mathbf{H}_k\mathbf{C}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k)^{-1} \tag{3.2.5}$$

We see that the filter expresses the posterior estimate as the the prior estimate adjusted by the $\mathbf{K}$-weighted deviation of the measured data from the predicted data.

To find the posterior error covariance $\mathbf{C}_k^a$ we want to see how $\mathbf{y}_k^a$ is distributed. Writing equation (3.2.4) as

$$\mathbf{y}_k^a = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{y}_k^f + \mathbf{K}_k\mathbf{d}_k \tag{3.2.6}$$

we see that $\mathbf{y}_k^a$ is distributed as

$$\mathbf{y}_k^a \sim N\left((\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{y}_k^f + \mathbf{K}_k\mathbf{d}_k\,,\,(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{C}_k^f(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{\Sigma}_k\mathbf{K}_k^T\right) \tag{3.2.7}$$

The covariance here can be simplified to

$$\begin{aligned}\mathbf{C}_k^a &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{C}_k^f(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{\Sigma}_k\mathbf{K}_k^T \\ &= \mathbf{C}_k^f - \mathbf{K}_k\mathbf{H}_k\mathbf{C}_k^f - \mathbf{C}_k^f\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{K}_k\mathbf{H}_k\mathbf{C}_k^f\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{K}_k\mathbf{\Sigma}_k\mathbf{K}_k^T \\ &= \mathbf{C}_k^f - \mathbf{K}_k\mathbf{H}_k\mathbf{C}_k^f - \mathbf{C}_k^f\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{K}_k(\mathbf{H}_k\mathbf{C}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k)\mathbf{K}_k^T\end{aligned} \tag{3.2.8}$$

where we insert the definition of the Kalman Gain in the fourth term of the last line to get

$$
\begin{aligned}
&= \mathbf{C}_k^f - \mathbf{K}_k\mathbf{H}_k\mathbf{C}_k^f - \mathbf{C}_k^f\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{C}_k^f\mathbf{H}_k^T(\mathbf{H}_k\mathbf{C}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k)^{-1}(\mathbf{H}_k\mathbf{C}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k)\mathbf{K}_k^T \\
&= \mathbf{C}_k^f - \mathbf{K}_k\mathbf{H}_k\mathbf{C}_k^f - \mathbf{C}_k^f\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{C}_k^f\mathbf{H}_k^T\mathbf{K}_k^T \\
\mathbf{C}_k^a &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{C}_k^f
\end{aligned}
\tag{3.2.9}
$$

See equation (2.4.11) together with (3.2.5) for comparison.

Equation(3.2.6) shows that the analysis step can be interpreted as an interpolation between the data and the forecasted state vector with the weights effected trough the respective uncertainties in the Kalman Gain matrix, $\mathbf{K}$.

We summarize the Kalman Filter forecast and analysis equations

$$
\begin{aligned}
\mathbf{y}_k^f &= \mathbf{A}\mathbf{y}_{k-1}^a + \mathbf{w}_{k-1} \\
\mathbf{C}_k^f &= \mathbf{A}\mathbf{C}_{k-1}^a\mathbf{A}^T + \mathbf{\Omega}_{k-1} \\
\mathbf{K}_k &= \mathbf{C}_k^f\mathbf{H}_k^T(\mathbf{H}_k\mathbf{C}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k)^{-1} \\
\mathbf{y}_k^a &= \mathbf{y}_k^f + \mathbf{K}_k(\mathbf{d}_k - \mathbf{H}_k\mathbf{y}_k^f) \\
\mathbf{C}_k^a &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{C}_k^f
\end{aligned}
\tag{3.2.10}
$$

3.2.1 Nonlinear models - Extended Kalman Filter

The Kalman Filter only applies to linear models. If the problem is nonlinear, then the state variables $\mathbf{y}_k$ and measurements $\mathbf{d}_k$ at time k must be written as

$$
\begin{aligned}
\mathbf{y}_k &= \mathbf{F}(\mathbf{y}_{k-1}, \mathbf{w}_{k-1}) \\
\mathbf{d}_k &= \mathbf{h}(\mathbf{y}_k, \mathbf{v}_k)
\end{aligned}
\tag{3.2.11}
$$

where $\mathbf{F}$ and $\mathbf{h}$ now are nonlinear functions with $\mathbf{w}_{k-1}$ as the modeling error and $\mathbf{v}_k$ as the measurement error.

One method which is developed to solve nonlinear updating problems is the Extended Kalman Filter (EKF). It linearizes about the current mean and covariance using partial derivatives of the process and the measurement function. The function $\mathbf{F}$ computes the new forecasted state estimate from the previous analyzed estimate, and the measurement function $\mathbf{h}$ computes the predicted measurements from the forecast. Since the true noise is not known, the forecast for the state vector and measurements are approximated by using zero noise

$$
\begin{aligned}
\mathbf{y}_k &\approx \mathbf{F}(\mathbf{y}_{k-1}, \mathbf{0}) \\
\mathbf{d}_k &\approx \mathbf{h}(\mathbf{y}_k, \mathbf{0})
\end{aligned}
\tag{3.2.12}
$$

F and **h** are not applied to the covariance directly. Instead, a matrix of partial derivatives; the Jacobian, is computed.

Here **A** and **W** are the Jacobi matrices containing the partial derivatives of **F** with respect to **y** and **w** respectively evaluated at the previous analyzed state and approximated zero noise,

$$\begin{aligned}\mathbf{A}(i,j) &= \frac{\partial \mathbf{F}(i)}{\partial \mathbf{y}(j)}(\mathbf{y}^a_{k-1}, \mathbf{0}) \\ \mathbf{W}(i,j) &= \frac{\partial \mathbf{F}(i)}{\partial \mathbf{w}(j)}(\mathbf{y}^a_{k-1}, \mathbf{0})\end{aligned} \tag{3.2.13}$$

and **H** and **V** are the Jacobi matrices with the partial derivatives of **h** with respect to **y** and **v** respectively evaluated at the current forecasted state and approximated zero noise.

$$\begin{aligned}\mathbf{H}(i,j) &= \frac{\partial \mathbf{h}(i)}{\partial \mathbf{y}(j)}(\mathbf{y}^f_{k-1}, \mathbf{0}) \\ \mathbf{V}(i,j) &= \frac{\partial \mathbf{h}(i)}{\partial \mathbf{v}(j)}(\mathbf{y}^f_{k-1}, \mathbf{0})\end{aligned} \tag{3.2.14}$$

The forecast and analysis equations for the EKF is as follows

Forecast equations

$$\begin{aligned}\mathbf{y}^f_k &= \mathbf{F}(\mathbf{y}^a_{k-1}, \mathbf{0}) \\ \mathbf{C}^f_k &= \mathbf{A}_k \mathbf{C}^a_{k-1} \mathbf{A}^T_k + \mathbf{W}_k \mathbf{\Omega}_{k-1} \mathbf{W}^T_k\end{aligned} \tag{3.2.15}$$

Analysis equations

$$\begin{aligned}\mathbf{K}_k &= \mathbf{C}^f_k \mathbf{H}^T_k (\mathbf{H}_k \mathbf{C}^f_k \mathbf{H}^T_k + \mathbf{V}_k \mathbf{\Sigma}_k \mathbf{V}^T_k)^{-1} \\ \mathbf{y}^a_k &= \mathbf{y}^f_k + \mathbf{K}_k(\mathbf{d}_k - \mathbf{h}(\mathbf{y}^f_k, \mathbf{0})) \\ \mathbf{C}^a_k &= (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{C}^f_k\end{aligned} \tag{3.2.16}$$

In many cases the EKF can give reasonable results as long as the nonentities of the dynamical system not are to severe. Also the EKF has problems handling systems with too many variables due to the computation and storage of very big matrices.

3.3 Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) was developed to address high dimensional problems with nonlinear models[5]. It was originally used to update the dynamical variables of an oceanic weather system. The first appearance of EnKF in the petroleum industry was presented in[17]. A broad overview of the EnKF in petroleum engineering can be found in[1].

3.3.1 Formulation

In the EnKF, one generates an initial ensemble of state vectors and runs each of the members forward in time. It is assumed that the mean and covariance are sufficient to describe the involved distributions, i.e that they are Gaussian. The statistics of the system, i.e the mean and the covariance, are approximated from the ensemble itself and are used in the following assimilation. The use of an ensemble avoids the computation of the real covariance matrix.

The method was originally designed to update only the dynamical variables of a system. In reservoir engineering, it has been used to update both static and dynamic parameters of the system. Typically, the state vector $\mathbf{y}_k$ contains both the static parameters $\mathbf{m} \in \mathbb{R}^{N_m}$and the dynamic variables $\mathbf{u}_k \in \mathbb{R}^{N_u}$, as well as the simulated data $\mathbf{h}_k \in \mathbb{R}^{N_d}$

$$\mathbf{y}_k = \begin{pmatrix} \mathbf{m} \\ \mathbf{u}_k \\ \mathbf{h}_k \end{pmatrix} \in \mathbb{R}^{N_m+N_u+N_d} \tag{3.3.1}$$

where $\mathbf{h}_k$ gives the generally nonlinear relationship between the variables and the predicted data at time k

$$\mathbf{h}_k = \mathbf{h}_k(\mathbf{u}_k, \mathbf{m}) \tag{3.3.2}$$

The observed data $\mathbf{d}_k$ is interpreted as the predicted data with added measurement errors $\mathbf{v}_k \sim N(\mathbf{0}, \mathbf{\Sigma}_k)$

$$\mathbf{d}_k = \mathbf{h}_k + \mathbf{v}_k \tag{3.3.3}$$

Because the augmented state vector contains the predicted data, we have the linear relationship

$$\mathbf{d}_k = \mathbf{H}_k\mathbf{y}_k + \mathbf{v}_k \tag{3.3.4}$$

where $\mathbf{H}_k$ is a matrix of zeros and ones. The linear relationship between the state vector and the data is achieved by adding the predicted data to form an augmented state vector containing the parameters, state variables and predicted data.

For the state vector we use the index i for the different ensemble members, and the total number of ensemble members is called N_e.

The initial ensemble is generated by specifying an initial ensemble mean, $\bar{\mathbf{y}}_0^a$, and an initial covariance matrix $\mathbf{C}_0$. The members are drawn randomly assuming a Gaussian distribution, (see (A.6))

$$\mathbf{y}_{0,i}^a = \bar{\mathbf{y}}_0^a + \mathbf{w}_{0,i} \tag{3.3.5}$$

where $\mathbf{w}_{0,i} \sim N(0, \mathbf{C}_0)$.

When the state vectors are propagated forward in time, the static parameters are kept constant, while the simulated data are computed using the latest state variables $\mathbf{u}_k^f$.

This gives the forecasted state vector at time k

$$\mathbf{y}_{k,i}^f = \mathbf{F}(\mathbf{y}_{k-1,i}^a) = \begin{pmatrix} \mathbf{m}_{k,i}^f \\ \mathbf{u}_{k,i}^f \\ \mathbf{h}_{k,i}^f \end{pmatrix} = \begin{pmatrix} \mathbf{m}_{k-1,i}^a \\ \mathbf{F}_k(\mathbf{u}_{k-1,i}^a\,,\, \mathbf{m}_{k-1,i}^a) \\ \mathbf{h}_k(\mathbf{u}_{k,i}^f\,,\, \mathbf{m}_{k-1,i}^a) \end{pmatrix} \tag{3.3.6}$$

where $i = 1...N_e$.

To compute the statistics from the ensemble we approximate the true state by the mean of the forecasted ensemble members

$$\mathbf{y}_k^{true} \approx \bar{\mathbf{y}}_k^f = \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{y}_{k,i}^f \tag{3.3.7}$$

The sample covariance matrix $\tilde{\mathbf{C}}_k^f$ is also computed from the ensemble. The sample covariance matrix is defined as

$$\tilde{\mathbf{C}}_k^f = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{y}_{k,i}^f - \bar{\mathbf{y}}_k^f)(\mathbf{y}_{k,i}^f - \bar{\mathbf{y}}_k^f)^T \tag{3.3.8}$$

The left factor of $\tilde{\mathbf{C}}_k^f$ can be written as

$$\mathbf{L}_k^f = \frac{1}{\sqrt{N_e - 1}} [(\mathbf{y}_{k,1}^f - \bar{\mathbf{y}}_k^f) \cdots (\mathbf{y}_{k,N_e}^f - \bar{\mathbf{y}}_k^f)] \tag{3.3.9}$$

and the sample covariance matrix may be written as

$$\tilde{\mathbf{C}}_k^f = \mathbf{L}_k(\mathbf{L}_k)^T \tag{3.3.10}$$

The full $\tilde{\mathbf{C}}$-matrix as defined in (3.3.10) never has to be computed explicitly during the analysis. Note that the rank of $\tilde{\mathbf{C}}$, i.e the minimum number of independent rows or columns, cannot be greater than $N_e - 1$ and is therefore limited by the size of the ensemble.

When new observations $\mathbf{d}_k \in \mathbb{R}^{N_d \times 1}$ are available at time step k, we have to treat them as random variables as well. If not, it has been shown that we get too low variance in our estimate[4].

Perturbed observations are generated using the actual measurements as reference measurements and adding errors from the same distribution as the measurement error.

$$\mathbf{d}_{k,i} = \mathbf{d}_k + \mathbf{v}_{k,i} \tag{3.3.11}$$

where $\mathbf{v}_{k,i} \sim N(0, \mathbf{\Sigma}_k)$

The analysis step is then carried out using the Kalman Filter equations, but with one update for each of the N_e ensemble members and the real covariance matrix $\mathbf{C}_k^f$ replaced by the sample covariance matrix $\tilde{\mathbf{C}}_k^f$.

$$\mathbf{y}_{k,i}^a = \mathbf{y}_{k,i}^f + \tilde{\mathbf{K}}_k(\mathbf{d}_{k,i} - \mathbf{H}_k\mathbf{y}_{k,i}^f) \tag{3.3.12}$$

where

$$\tilde{\mathbf{K}}_k = \tilde{\mathbf{C}}_k^f\mathbf{H}_k^T(\mathbf{H}_k\tilde{\mathbf{C}}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k)^{-1} \tag{3.3.13}$$

The analyzed estimate is the mean of the analyzed ensemble members, given as

$$\bar{\mathbf{y}}_k^a = \bar{\mathbf{y}}_k^f + \tilde{\mathbf{K}}_k(\bar{\mathbf{d}}_k - \mathbf{H}_k\bar{\mathbf{y}}_k^f) \tag{3.3.14}$$

In the analysis step, both the static and dynamic variables are updated.

The analyzed covariance matrix is given along the same lines as (3.3.8). If $N_e \to \infty$ and the distributions of the forecast and observation ensembles are independent, then the expression for $\tilde{\mathbf{C}}_k^a$ reduces to (3.2.8) equal to the Kalman Filter[1]. Since the number of ensemble members N_e is restricted to the order of tens or hundreds, this results in sampling errors making $\tilde{\mathbf{C}}_k^a \neq \mathbf{C}_k^a$.

We summarize the EnKF equations

$$\begin{aligned}
\bar{\mathbf{y}}_k^f &= \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{y}_{k,i}^f = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{F}(\mathbf{y}_{k-1,i}^a) \\
\tilde{\mathbf{C}}_k^f &= \frac{1}{N_e-1}\sum_{i=1}^{N_e}(\mathbf{y}_{k,i}^f - \bar{\mathbf{y}}_k^f)(\mathbf{y}_{k,i}^f - \bar{\mathbf{y}}_k^f)^T \\
\tilde{\mathbf{K}}_k &= \tilde{\mathbf{C}}_k^f\mathbf{H}_k^T(\mathbf{H}_k\tilde{\mathbf{C}}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k)^{-1} \\
\bar{\mathbf{y}}_k^a &= \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{y}_{k,i}^a = \bar{\mathbf{y}}_k^f + \tilde{\mathbf{K}}_k(\bar{\mathbf{d}}_k - \mathbf{H}_k\bar{\mathbf{y}}_k^f) \\
\tilde{\mathbf{C}}_k^a &= \frac{1}{N_e-1}\sum_{i=1}^{N_e}(\mathbf{y}_{k,i}^a - \bar{\mathbf{y}}_k^a)(\mathbf{y}_{k,i}^a - \bar{\mathbf{y}}_k^a)^T
\end{aligned} \tag{3.3.15}$$

3.3.2 Practical Implementation

We mentioned that the full covariance matrix never has to be formed explicitly when performing the update in the analysis step. This allows for efficient numerical implementation of the method. We also point out how the EnKF analyzed estimate can be interpreted as a linear combination of the forecasted ensemble.

Here we skip the time index k, since all variables refer to the same update time. Following [7] we start with defining the matrix $\mathbf{Y}^f$ which contains the ensemble of the forecasted state vectors $\mathbf{y}_i^f \in \mathbb{R}^N$ as its columns

$$\mathbf{Y}^f = \left(\mathbf{y}_1^f, \mathbf{y}_2^f, \ldots, \mathbf{y}_{N_e}^f\right) \in \mathbb{R}^{N\times N_e} \tag{3.3.16}$$

The mean of the forecasted ensemble is stored in all columns of the matrix

$$\bar{\mathbf{Y}}^f = \mathbf{Y}^f\mathbf{1}_{N_e} \tag{3.3.17}$$

where $\mathbf{1}_{N_e} \in \mathbb{R}^{N_e\times N_e}$ is a matrix with all entries equal to $1/N_e$. We then define $\Delta\mathbf{Y}$ as the ensemble perturbation matrix

$$\begin{aligned}
\Delta\mathbf{Y}^f &= \mathbf{Y}^f - \bar{\mathbf{Y}}^f \\
&= \mathbf{Y}^f(\mathbf{I} - \mathbf{1}_{N_e})
\end{aligned} \tag{3.3.18}$$

The ensemble approximation to the covariance matrix is then written as

$$\tilde{\mathbf{C}}^f = \frac{\Delta\mathbf{Y}^f(\Delta\mathbf{Y}^f)^T}{N_e - 1} \tag{3.3.19}$$

This is consistent with the equations (3.3.8)-(3.3.10).

We also define the matrix $\mathbf{D}$ containing the perturbed data vectors $\mathbf{d}_i \in \mathbb{R}^{N_d \times 1}$ from (3.3.11) as

$$\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{N_e}) \in \mathbb{R}^{N_d \times N_e} \tag{3.3.20}$$

An ensemble representation of the measurement error covariance matrix is defined from the perturbations $\mathbf{v}_i$ as

$$\tilde{\mathbf{\Sigma}} = \frac{\mathbf{E}\mathbf{E}^T}{N_e - 1} \tag{3.3.21}$$

where $\mathbf{E}$ is the ensemble of added data perturbations

$$\mathbf{E} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{N_e}) \in \mathbb{R}^{N_d \times N_e} \tag{3.3.22}$$

With these ensemble matrices we write the analysis step (3.3.12) as

$$\mathbf{Y}^a = \mathbf{Y}^f + \tilde{\mathbf{C}}\mathbf{H}^T(\mathbf{H}\tilde{\mathbf{C}}\mathbf{H}^T + \tilde{\mathbf{\Sigma}})^{-1}(\mathbf{D} - \mathbf{H}\mathbf{Y}^f) \tag{3.3.23}$$

Here we can use the exact measurement error covariance matrix $\mathbf{\Sigma}$, as well as the ensemble approximation $\tilde{\mathbf{\Sigma}}$ defined above.

Using the definitions of the ensemble error covariance matrices (3.3.19) and (3.3.21), we express the analysis step as

$$\mathbf{Y}^a = \mathbf{Y}^f + \Delta\mathbf{Y}^f(\Delta\mathbf{Y}^f)^T\mathbf{H}^T(\mathbf{H}\Delta\mathbf{Y}^f(\Delta\mathbf{Y}^f)^T\mathbf{H}^T + \mathbf{E}\mathbf{E}^T)^{-1}(\mathbf{D} - \mathbf{H}\mathbf{Y}^f) \tag{3.3.24}$$

In this formulation, the covariance matrices are not computed explicitly.

We now introduce some matrices to write the analysis in a simpler form. Let $\mathbf{S} \in \mathbb{R}^{N_d \times N_e}$ be

$$\mathbf{S} = \mathbf{H}\Delta\mathbf{Y}^f, \tag{3.3.25}$$

the matrix $\mathbf{C} \in \mathbb{R}^{N_d \times N_d}$

$$\mathbf{C} = \mathbf{S}\mathbf{S}^T + (N_e - 1)\mathbf{\Sigma}, \tag{3.3.26}$$

and the matrix $\Delta\mathbf{D} \in \mathbb{R}^{N_d \times N_e}$

$$\Delta\mathbf{D} = \mathbf{D} - \mathbf{H}\mathbf{Y}^f \tag{3.3.27}$$

Then (3.3.24) is given as

$$\begin{aligned}
\mathbf{Y}^a &= \mathbf{Y}^f + \Delta\mathbf{Y}^f\mathbf{S}^T\mathbf{C}^{-1}\Delta\mathbf{D} \\
&= \mathbf{Y}^f + \mathbf{Y}^f(\mathbf{I} - \mathbf{1}_{N_e})\mathbf{S}^T\mathbf{C}^{-1}\Delta\mathbf{D} \\
&= \mathbf{Y}^f(\mathbf{I} + (\mathbf{I} - \mathbf{1}_{N_e})\mathbf{S}^T\mathbf{C}^{-1}\Delta\mathbf{D}) \\
&= \mathbf{Y}^f(\mathbf{I} + \mathbf{S}^T\mathbf{C}^{-1}\Delta\mathbf{D})
\end{aligned} \tag{3.3.28}$$

where $\mathbf{1}_N\mathbf{S}^T \equiv \mathbf{0}$. The matrix $\mathbf{X} \in \mathbb{R}^{N_e \times N_e}$ is defined as

$$\mathbf{X} = \mathbf{I} + \mathbf{S}^T\mathbf{C}^{-1}\Delta\mathbf{D} \tag{3.3.29}$$

When the EnKF analysis is formulated as

$$\mathbf{Y}^a = \mathbf{Y}^f\mathbf{X} \tag{3.3.30}$$

we see that the analyzed ensemble $\mathbf{Y}^a$ must be a linear combination of the forecasted ensemble. Thus, the solution is searched for in the space spanned by the ensemble members. The fact that N_e typically is much smaller than the number of variables N in the state vector makes the generation of the initial ensemble an important issue. We want the distribution of the initial ensemble members to properly describe the uncertainty in the initial state, so that they span the space where the solution is.

3.3.3 Challenges with EnKF

Main advantages with the EnKF is that it is computationally efficient, fairly easy to implement, and the ability to approximate the covariance matrix without having to evolve it in time and store it. This makes it possible to efficiently update a large number of variables for nonlinear systems.

However, the fact that the covariance is approximated with $\tilde{\mathbf{C}}$ from a relatively small sized ensemble (N_e), leads to several issues. This could of course be solved by using a larger ensemble, but efficiency demands the ensemble size to be reduced as much as possible. Also, assimilating a large number of measurements (N_d), at the same time is problematic in the basic EnKF algorithm. This is closely related to the size of the ensemble N_e.

Other main challenges are problems with non-Gaussian distributions, strong non-linearities in the forward model and the application to large-scale field models, but we do not focus on this here.

Limited ensemble size As mentioned, the solution of the analysis step is confined to a smaller space spanned by the ensemble members, rather than the much bigger state space.

We also mentioned that a limited N_e results in sampling errors for $\tilde{\mathbf{C}}$. These spurious correlations may cause updating of variables in regions of no real influence. In particular, it is observed that observations affect variables far away from the measurement location in a too large degree. A spurious correlation between a predicted measurement and a variable leads to an artificial update of this variable for each of the ensemble members. This reduces the variance, leading to an updated ensemble variance which is underestimated. If the ensemble variance of the variables is underestimated, the filter "believes" that it performs better that it actually does, and the filter may eventually diverge.

Large number of data When assimilating a large number of data, N_d, at the same time, as for seismic data or 4D seismic data, two major problems may occur. The inversion of the $N_d \times N_d$-matrix $\mathbf{H}_k\mathbf{C}_k^f\mathbf{H}_k^T + \mathbf{\Sigma}_k$ from the Kalman Gain may be computationally much more demanding, requiring an efficient analysis scheme. In addition, since each analyzed ensemble member have to be a linear combination of the initial ensemble, it may be impossible to match large sets of data where $N_d > N_e$. Such a scheme to handle this is proposed in [6, 24].

The most common techniques for dealing with small ensembles and large amount of data are localization methods; covariance localization and local analysis.

Covariance localization To reduce the effects of spurious correlations, $\tilde{\mathbf{C}}$ can be multiplied element-wise by a compactly supported positive definite matrix to produce a localized covariance. The properties of this multiplication ensures that the covariance matrix achieves full rank[11, 9]. The most basic form of covariance localization simply uses a cut-off radius. This way only model parameters within a given distance of the observation will be updated, thus removing the long range correlations. Other approaches uses different correlation functions to compute the localized covariance.

Covariance localization is used in [11, 9, 20]. In [9], it is also used a technique called variance inflation to deal with the potentially underestimated variance.

Local analysis To avoid the problems associated with large data sets, one may assume that only measurements close to a grid cell, or grid point, should impact the analysis in that point. Then the analysis can be computed locally, grid point by grid point. In each local analysis, only measurements which are within some specified neighborhood are used in the update. The neighborhood should be defined big enough to include all relevant measurements, but small enough to keep the number of local measurements low, and also eliminate spurious correlations.

Performing the analysis locally implies that a small model state is now solved in a relatively large ensemble space. Also, the variables are now allowed to be updated by different linear combinations of the fore casted ensemble members, thus making it easier to obtain solutions that match a large data set.

It must be mentioned that the methods used for dealing with these issues creates an additional amount of work, and that it may not be straightforward to decide how to define "closeness" through local neighborhoods and different correlation functions.

3.3.3.1 Focus in this thesis

In this thesis, the focus lies on the errors made by approximating the covariance with the sample covariance, and how the error depends on increasing the number of measurements.

It is motivated by the recent work of Kovalenko et. al [14, 15]. Here, the distribution of the norm of the sampling error at one single analysis step is derived. Assumptions of a Gaussian distributed forecast ensemble together with zero measurement errors, $\mathbf{\Sigma} = \mathbf{0}$, are made. The analytical distribution and the parameters affecting it were studied through numerical experiments. The authors found that increasing the number of measurements led to an increase in the sampling error norm, even when the (positive) difference $N_e - N_d$ was held constant. They also found that spread measurement layouts gave a smaller sampling error norm than dense measurement layouts: As expected, increasing the ensemble size led to a smaller norm of the sampling error.

It could be interesting to look at the norm of the sampling error in the analysis step, with the assumption of zero measurement errors relaxed. To do this we try a different approach by using approximate calculations.

In the next chapter we introduce the ideas behind approximate calculations in general before we proceed to the analysis step in the EnKF.

Approximate calculations

Chapter 4

Approximation Theory

In this chapter we define some notation and useful tools like Neumann Series to perform order of magnitude calculations. In the next chapter we use approximate calculations to look at the analysis step in the EnKF.

4.1 Perturbation Theory

Very often, a mathematical problem cannot be solved exactly or, if the exact solution is available, it exhibits such an intricate dependency in the parameters that it is hard to use as such. It may be the case, however, that a parameter can be identified, say ϵ, such that the solution is available and reasonably simple for $\epsilon = 0$. Then, one may wonder how this solution is altered for non-zero, but small ϵ. Perturbation theory gives a systematic answer to this question. First, we define some notation to keep track of orders of magnitude.

Big-O Notation We write $f(\epsilon) = \mathcal{O}(u(\epsilon))$ as $\epsilon \to 0$ if there exists a positive constant K such that

$$| f(\epsilon) | \leq K | u(\epsilon) | \tag{4.1.1}$$

whenever ϵ is sufficiently close to 0.

For example, $sin(\epsilon) = \mathcal{O}(\epsilon)$ as $\epsilon \to 0$ because $| sin(\epsilon) | \leq | \epsilon |$ when ϵ approaches zero.

Example We now look at a simple example to show the basic idea in perturbation theory. Consider the quadratic equation

$$x^2 - 1 = \epsilon x \tag{4.1.2}$$

which we easily can confirm have the two roots

$$\begin{aligned} x_1 &= \sqrt{1 + \epsilon^2/4} + \epsilon/2 \\ x_2 &= -\sqrt{1 + \epsilon^2/4} + \epsilon/2 \end{aligned} \tag{4.1.3}$$

Imagine now that we did not know how to solve this equation analytically, but we recognized that the simpler form $x^2 - 1 = 0$, with $\epsilon = 0$, would give us the roots $x_{1,2} = \pm 1$. If ϵ is small, the real solution may not deviate far from this simpler one. Assuming that we can write the solution(s) x as

$$x = X_0 + \epsilon X_1 + \epsilon^2 X_2 + \mathcal{O}(\epsilon^3) \tag{4.1.4}$$

for X_0, X_1, X_2 to be determined. This can be done by substituting for x into the original equation written as $x^2 - 1 - \epsilon x = 0$, expanding the left hand side and collecting the terms in orders of ϵ, giving

$$X_0^2 - 1 + \epsilon(2X_0X_1 - X_0) + \epsilon^2(X_1^2 + 2X_0X_2 - X_1) + \mathcal{O}(\epsilon^3) = 0 \tag{4.1.5}$$

Equating the successive terms of this series to zero

$$\mathcal{O}(\epsilon^0) : X_0^2 - 1 = 0 \tag{4.1.6}$$
$$\mathcal{O}(\epsilon^1) : 2X_0X_1 - X_0 = 0 \tag{4.1.7}$$
$$\mathcal{O}(\epsilon^2) : X_1^2 + 2X_0X_2 - X_1 = 0 \tag{4.1.8}$$
$$\mathcal{O}(\epsilon^3) : \cdots \tag{4.1.9}$$

and solving for X_0, X_1, X_2 gives

$$\begin{aligned} x_1 &= 1 + \epsilon/2 + \epsilon^2/8 + \mathcal{O}(\epsilon^3) \\ x_2 &= -1 + \epsilon/2 - \epsilon^2/8 + \mathcal{O}(\epsilon^3) \end{aligned} \tag{4.1.10}$$

Since $X_0^2 - 1$ has two roots we get to different solutions. This is actually the Taylor series expansion of x in terms of ϵ. This one consists of the simple solution of $x^2 - 1 = 0$ and two correction terms for the small perturbation. For small ϵ, these roots are well approximated by the first few terms of their Taylor series expansion. The truncated Taylor series here is called a second order correction to the solution, and is valid only for small values of ϵ.

We use this idea to approximate the Kalman Gain matrix by expressing it as a truncated series in some small parameter ϵ.

4.2 Neumann Series

The inverse of the expression $(\mathbf{I}-\mathbf{A})^{-1}$ can be written as an infinite series,

$$(\mathbf{I}-\mathbf{A})^{-1}=\sum_{n=0}^{\infty}\mathbf{A}^n \tag{4.2.1}$$

provided that $\|\mathbf{A}\|<1$. This is the generalization of the well known scalar geometric series where $1/1-x=\sum_{n=0}^{\infty}x^n$ if $x<1$. A proof of this result can be found in the Appendix, (A.8).

This way, the inverse of a sum of two matrices can be expanded into a Neumann Series. To invert the sum $(\mathbf{A}+\mathbf{B})^{-1}$ we can write

$$\begin{aligned}(\mathbf{A}+\mathbf{B})^{-1}&=((\mathbf{I}-(-\mathbf{B}\mathbf{A}^{-1}))\mathbf{A})^{-1}\\&=\mathbf{A}^{-1}(\mathbf{I}-(-\mathbf{B}\mathbf{A}^{-1}))^{-1}\\&=\mathbf{A}^{-1}\left[\sum_{n=0}^{\infty}\left(-\mathbf{B}\mathbf{A}^{-1}\right)\right]\\&=\mathbf{A}^{-1}\left[\mathbf{I}-\mathbf{B}\mathbf{A}^{-1}+(\mathbf{B}\mathbf{A}^{-1})^2-\ldots\right]\end{aligned} \tag{4.2.2}$$

as long as $\|\mathbf{B}\mathbf{A}^{-1}\|<1$. This holds given that $\|\mathbf{A}\|>\|\mathbf{B}\|$. If we have the opposite situation, i.e $\|\mathbf{A}\|<\|\mathbf{B}\|$, we just factor out $\mathbf{B}$ instead giving

$$\begin{aligned}(\mathbf{A}+\mathbf{B})^{-1}&=((\mathbf{I}-(-\mathbf{A}\mathbf{B}^{-1}))\mathbf{B})^{-1}\\&=\mathbf{B}^{-1}(\mathbf{I}-(-\mathbf{A}\mathbf{B}^{-1}))^{-1}\\&=\mathbf{B}^{-1}\left[\sum_{n=0}^{\infty}\left(-\mathbf{A}\mathbf{B}^{-1}\right)\right]\\&=\mathbf{B}^{-1}\left[\mathbf{I}-\mathbf{A}\mathbf{B}^{-1}+(\mathbf{A}\mathbf{B}^{-1})^2-\ldots\right]\end{aligned} \tag{4.2.3}$$

where now $\|\mathbf{A}\mathbf{B}^{-1}\|<1$.

The sum we want to invert in our case is the one appearing in the Kalman Gain matrix from the analysis step in the Kalman Filter, namely $(\mathbf{H}\mathbf{C}^f\mathbf{H}^T+\mathbf{\Sigma})^{-1}$. Depending upon whether we have dominating measurement errors or model errors, we will express the inverse of the sum as a truncated Neumann Series so that $\|\mathbf{B}\mathbf{A}^{-1}\|<1$ or $\|\mathbf{A}\mathbf{B}^{-1}\|<1$ is ensured.

The inverted matrix can be expressed by power series in a small parameter ϵ, such that it may be written as

$$\mathbf{\Gamma}=\sum_{n=0}^{\infty}\epsilon^n\mathbf{\Gamma}_n \tag{4.2.4}$$

where each matrix $\mathbf{\Gamma}_n$ has elements defined by the coefficients of ϵ^k in the power series representing the corresponding elements in $\mathbf{\Gamma}$. Alternatively, it could have been expressed as

$$\mathbf{\Gamma} = \sum_{n=0}^{\infty} \widehat{\mathbf{\Gamma}}_n \tag{4.2.5}$$

such that each element in $\widehat{\mathbf{\Gamma}}_0$ equals the lowest order term occurring in the series representing the corresponding element in $\mathbf{\Gamma}$, each element in $\widehat{\mathbf{\Gamma}}_1$ equals the second lowest order term and so on[18]. Here, we want to keep terms of similar orders of magnitude in our calculations, so we use the first series representation. We approximate the inverse matrix with

$$\mathbf{\Gamma} \approx \sum_{n=0}^{N} \epsilon^n \mathbf{\Gamma}_n \tag{4.2.6}$$

where we truncate the series at some $N \geq 1$. Truncating for $n = 1$ gives a first order correction. It is normal to use either a first or second order correction. For very small parameters ϵ, a first order correction can be good enough.

Chapter 5

The EnKF analysis step

In this chapter we focus on the analysis step in EnKF. We are interested in the errors that arises because EnKF approximates the covariance with the sample covariance matrix $\tilde{\mathbf{C}}$. It is natural to compare the update in EnKF with the update in the Kalman Filter which uses the real covariance matrix. Even though our expressions will be simplified, we hope to gain some insight into the problem.

5.1 Approximations and assumptions

In our case we will look at one time step only, thus ignoring the k index from now on. This is done to isolate the effect of the sample covariance matrix in the analysis step, from other difficulties that the EnKF suffer from.

Thus, we want to eliminate all other sources of error than $\tilde{\mathbf{C}} \neq \mathbf{C}$. In our simplified calculations, the Kalman Filter would provide the correct answer. The analysis step in EnKF will be compared with the KF analysis step by subtracting the expressions.

5.1.1 Analysis difference

In the basic EnKF every ensemble member is updated linearly with the same Kalman gain matrix, but with different perturbed measurements $\mathbf{d}_i$ see (3.3.11). Because the added measurement perturbations $\mathbf{v}_i$, are assumed Gaussian distributed around zero, we approximate the ensemble mean of all the $\mathbf{d}_i$, equal to the actual measurements $\mathbf{d}$ (see (3.3.11)).

The analysis step with the Kalman Gain written out is given as

$$\mathbf{y}^a_{KF} = \mathbf{y}^f + \mathbf{C}^f\mathbf{H}^T(\mathbf{H}\mathbf{C}^f\mathbf{H}^T + \mathbf{\Sigma})^{-1}(\mathbf{d} - \mathbf{H}\mathbf{y}^f) \tag{5.1.1}$$

$$\mathbf{y}^a_{EnKF,i} = \mathbf{y}^f_i + \tilde{\mathbf{C}}^f\mathbf{H}^T(\mathbf{H}\tilde{\mathbf{C}}^f\mathbf{H}^T + \mathbf{\Sigma})^{-1}(\mathbf{d}_i - \mathbf{H}\mathbf{y}^f_i) \tag{5.1.2}$$
$for\, i = 1, \cdots N_e$

in KF and EnKF respectively. We assume that at our time step, we start out with the same forecasts in both the EnKF and the KF before the update, i.e.

$$\mathbf{y}^f = \mathbf{y}^f_{KF} = \bar{\mathbf{y}}^f_{EnKF} = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{y}^f_i \tag{5.1.3}$$

Then, from these equations we define the error difference

$$\begin{aligned}\mathbf{y}^a_{KF} - \bar{\mathbf{y}}^a_{EnKF} =& \mathbf{C}^f\mathbf{H}^T(\mathbf{H}\mathbf{C}^f\mathbf{H}^T + \mathbf{\Sigma})^{-1}(\mathbf{d} - \mathbf{H}\mathbf{y}^f) \\ &- \tilde{\mathbf{C}}^f\mathbf{H}^T(\mathbf{H}\tilde{\mathbf{C}}^f\mathbf{H}^T + \mathbf{\Sigma})^{-1}(\mathbf{d} - \mathbf{H}\mathbf{y}^f) \\ =& \mathbf{C}\mathbf{U} - \tilde{\mathbf{C}}\tilde{\mathbf{U}} \\ =& (\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}} + \mathbf{C}(\mathbf{U} - \tilde{\mathbf{U}})\end{aligned} \tag{5.1.4}$$

with $\mathbf{y}^f$ canceled out, and $\mathbf{U} \equiv \mathbf{H}^T(\mathbf{H}\mathbf{C}^f\mathbf{H}^T + \mathbf{\Sigma})^{-1}(\mathbf{d} - \mathbf{H}\mathbf{y}^f)$.

This way we can split the effect in two parts; the difference between the covariance matrices and the difference between the update vectors $\mathbf{U}$.
From now on we refer to the difference $(\mathbf{d} - \mathbf{H}\mathbf{y}^f)$ as $\mathbf{\Delta}$. To make it easy to work with, we let all the entries Δ_i in the vector $\mathbf{\Delta} \in \mathbb{R}^{N_d \times 1}$ be equal, that is $\Delta_i = \Delta$ for all $i = 1, 2, \ldots N_d$.

We are particularly interested in how the norm of the error depends on increasing the number of assimilated data, N_d. That is, we want to look at the L_2-norm of the analysis difference

$$\|\mathbf{y}^a_{KF} - \bar{\mathbf{y}}^a_{EnKF}\|_2 = \|(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}} + \mathbf{C}(\mathbf{U} - \tilde{\mathbf{U}})\|_2 \tag{5.1.5}$$

for increasing N_d.

5.1.2 Structure of the covariance matrices

We make some assumptions on the structures of the covariance matrices $\mathbf{C}$ and $\tilde{\mathbf{C}}$, and define some useful parameters. To obtain a simple structure for $\mathbf{C}$, we simplify the model.

In reality, the state vector contains parameters $\mathbf{m}$, state variables $\mathbf{u}(\mathbf{m})$ and simulated data $\mathbf{h}(\mathbf{u}, \mathbf{m})$, making the covariance matrix contain the covariances between all the components of the state vector

$$\begin{pmatrix} Cov(\mathbf{m}, \mathbf{m}) & Cov(\mathbf{m}, \mathbf{u}) & Cov(\mathbf{m}, \mathbf{h}) \\ Cov(\mathbf{u}, \mathbf{m}) & Cov(\mathbf{u}, \mathbf{u}) & Cov(\mathbf{u}, \mathbf{m}) \\ Cov(\mathbf{h}, \mathbf{m}) & Cov(\mathbf{h}, \mathbf{u}) & Cov(\mathbf{h}, \mathbf{h}) \end{pmatrix} \tag{5.1.6}$$

We consider a simplified version of the covariance matrix, namely just the parameter-parameter variance $\mathbf{C} = Cov(\mathbf{m}, \mathbf{m})$. This is similar to assumptions made in [14, 15].The reason for this simplification in our case will soon be apparent.

For a one dimensional grid, we assume that we have an exponential, distance-dependent covariance model (see (A.5)). The entries of $\mathbf{C}$, i.e the covariance between cells i_1 and i_2 will be defined by

$$C(i_1, i_2) = \xi^2 \exp\left(-\frac{\mid i_1 - i_2 \mid}{l}\right) \tag{5.1.7}$$

where ξ^2 is the variance in the model variables and l is the correlation length given with respect to number of grid cells. Using this exponential correlation model we define the parameters

$$r \equiv \frac{C(i_1, i_2+1)}{C(i_1, i_2)} = \frac{\xi^2 \exp\left(-\frac{|i_1 - i_{2+1}|}{l}\right)}{\xi^2 \exp\left(-\frac{|i_1 - i_2|}{l}\right)} = \exp\left(-\frac{1}{l}\right) \text{ ,} 0 < r < 1 \tag{5.1.8}$$

$$\begin{aligned} \varepsilon &\equiv \left(\frac{\xi}{\sigma}\right)^2 \\ \delta &\equiv \left(\frac{\sigma}{\xi}\right)^2 \end{aligned} \tag{5.1.9}$$

where σ^2 is the measurement variance and ξ^2 is the model variance. We assume that the errors in the measurements are independent of each other, so that $\mathbf{\Sigma} = \sigma^2\mathbf{I}$.

$$\mathbf{C} = \xi^2\mathbf{Q} = \xi^2 \begin{pmatrix} 1 & r & r^2 & \cdots & r^{N-1} \\ r & 1 & r & \ddots & \\ r^2 & r & 1 & r & r^2 \\ & \ddots & r & 1 & r \\ r^{N-1} & & r^2 & r & 1 \end{pmatrix} \tag{5.1.10}$$

where $\mathbf{Q}$ is the scaled version.

The main reason for picking this $\mathbf{C}$-matrix is that we can find the inverse of $\mathbf{Q}$ quite easily. In [10] it is shown that the matrix $\mathbf{Q}$ has an inverse $\mathbf{Q}^{-1}$, and that the matrix $(1-r^2)\mathbf{Q}^{-1}$ has the entries $-r$ in every position of the sub- and super diagonal, and has main diagonal entries $1,\ 1+r^2, \ldots, 1+r^2,\ 1$. To see this we can check that

$$[(1-r^2)\mathbf{Q}^{-1}]\mathbf{Q} = (1-r^2)\mathbf{I} \tag{5.1.11}$$

Thus the inverse of $\mathbf{Q}$ is given as

$$\mathbf{Q}^{-1} = \frac{1}{1-r^2}\begin{pmatrix} 1 & -r & 0 & \cdots & & 0 \\ -r & 1+r^2 & -r & \ddots & & \\ 0 & -r & 1+r^2 & -r & & \\ \vdots & \ddots & & \ddots & & 0 \\ & & & -r & 1+r^2 & -r \\ 0 & & & 0 & -r & 1 \end{pmatrix} \tag{5.1.12}$$

To approximate the sample covariance matrix $\tilde{\mathbf{C}}$, we interpret it as the true $\mathbf{C}$ with a perturbed part added to it.

The perturbations in the elements of $\tilde{\mathbf{C}}$ are caused by random perturbations from the ensemble members. Therefore, the perturbations in $\tilde{\mathbf{C}}$ are random as well.

If N_e is sufficiently large, then $\tilde{\mathbf{C}}$ will be close to $\mathbf{C}$. A too small N_e, leads to larger deviations from $\mathbf{C}$. We write $\tilde{\mathbf{C}}$ as

$$\begin{aligned} \tilde{\mathbf{C}} &= \xi^2 \tilde{\mathbf{Q}} \\ &= \xi^2(\mathbf{Q} + \gamma\mathbf{R}) \\ &= \mathbf{C} + \gamma\xi^2\mathbf{R} \end{aligned} \tag{5.1.13}$$

and

$$\tilde{\mathbf{Q}} = \mathbf{Q} + \gamma\mathbf{R} \tag{5.1.14}$$

with

$$\mathbf{R} \equiv \begin{pmatrix} \mu_{1,1} & \mu_{1,2}r & & \cdots & \mu_{1,n}r^{N-1} \\ \mu_{2,1}r & \ddots & \ddots & & \\ \vdots & \ddots & & & \\ & & & \ddots & \\ \mu_{N,1}r^{N-1} & & & & \mu_{N,N} \end{pmatrix} \tag{5.1.15}$$

where $\mathbf{R} \in \mathbb{R}^{N\times N}$ is a matrix with random elements. The γ allows for varying the deviation from $\mathbf{C}$ with the ensemble size. A very small $\gamma \ll 1$ symbolizes that N_e is big enough to make $\tilde{\mathbf{C}}$ close to $\mathbf{C}$, while a bigger value of γ represents a smaller ensemble, leading to larger deviations in the sample covariance matrix.

The diagonal entries in $\mathbf{R}$ are assumed to be distributed as $\mu_{ii} \sim (0, \zeta)$ with variance ζ small. Then, from (5.1.14) we have that the diagonal entries for $\tilde{\mathbf{Q}}$ are assumed to be distributed as $\tilde{q}_{ii} = q_{ii} + \gamma\mu_{ii} \sim (1, \gamma^2\zeta)$.

5.1.3 Measurement patterns

For convenience, we work with measurements or data that are distributed along our one dimensional "reservoir" in simple patterns, or layouts. The measurements are collected in equidistant locations which are spaced h cells apart from each other. Then we can vary the distance h to get either a dense distribution of measurements or a sparse distribution of measurements.

For example, for a dense distribution with data from all the N_d first cells we use a linear observation matrix $\mathbf{H}_{dense} \in \mathbb{R}^{N_d \times N}$ defined with $h_{dense} = 1$, making the first N_d columns of $\mathbf{H}_{dense}$ equal to the $N_d \times N_d$ identity matrix

$$\mathbf{H}_{dense} = [\mathbf{I}_{N_d} \mid \mathbf{0}] \in \mathbb{R}^{N_d \times N} \Longrightarrow (\mathbf{H}^T\mathbf{H})_{dense} = \begin{bmatrix} \mathbf{I}_{N_d} & | & \mathbf{0} \\ - & - & - \\ \mathbf{0} & | & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{N \times N} \quad (5.1.16)$$

For a sparse distribution with $h_{sparse} > 1$, the ones will appear in every $h'_{sparse}th$ position along the upper left part of the diagonal of $(\mathbf{H}^T\mathbf{H})_{sparse}$.

We introduce some simplified notation for dense and sparse measurements; If we have observations in every $h'th$ cell, we write the subscript $\mathbf{X}_h$ to denote this. In particular, we use the notation $\mathbf{Q}_h$ instead of $\mathbf{H}_h\mathbf{Q}\mathbf{H}_h^T$. This way $\mathbf{Q}_h$ is given as

$$\begin{bmatrix} 1 & r^h & r^{2h} & \cdots & r^{(N_d-1)h} \\ r^h & \ddots & \ddots & & \vdots \\ r^{2h} & \ddots & & & \\ \vdots & & & \ddots & r^h \\ r^{(N_d-1)h} & & \cdots & r^h & 1 \end{bmatrix} \quad (5.1.17)$$

and $\mathbf{Q}_h^{-1}$ is

$$\frac{1}{1-r^{2h}} \begin{bmatrix} 1 & -r^h & 0 & \cdots & 0 \\ -r^h & 1+r^{2h} & \ddots & & \vdots \\ 0 & \ddots & \ddots & & 0 \\ \vdots & & & 1+r^{2h} & -r^h \\ 0 & \cdots & 0 & -r^h & 1 \end{bmatrix}$$

and so on.

Chapter 6

Analysis and results

In the following sections we use the approximations and assumptions made earlier to produce analytical, approximate expressions for the norm of the analysis difference. Then we test these expressions through numerical experiments. Even though we use a very simplistic setting for our calculations, we hope that we can capture some of the dominating trends or effects.

Factors that are considered are

- the relationship between the errors of the data and the predicted data through the parameters ε and δ from (5.1.9),
- the number of assimilated data N_d,
- the distribution of the data through the spacing distances h_{sparse} and h_{dense}
- the correlation length l.
- the number of ensemble members N_e through the parameter γ,

We now consider two main cases; when we have dominating measurement errors, and when we have dominating model errors [in the predicted measurements]. For each of these cases we look into different situations.
We repeat that we consider the L_2-norm. From now on, $\|\cdot\|$ should be understood as $\|\cdot\|_2$.

6.1 Dominating measurement errors

In this case we have that $\sigma^2 \gg \xi^2$, making the parameter $\varepsilon \ll 1$. This means that the uncertainty in the model is much smaller than the uncertainty in the measurements.

In the extreme case with $\xi \to 0$, the analysis step will ignore the contribution from the data.

We now use the Neumann series from (4.2) to approximate the inverse matrix in the Kalman Gain

$$(\mathbf{C}_h + \mathbf{\Sigma})^{-1} = \mathbf{\Sigma}^{-1}[\mathbf{I} - \mathbf{\Sigma}^{-1}(\mathbf{C}_h) + \mathbf{\Sigma}^{-2}(\mathbf{C}_h)^2 - \cdots] \tag{6.1.1}$$

The matrix $\mathbf{\Sigma}$ is easily inverted because it is a diagonal matrix with our assumption of independent measurements. The entries in $\mathbf{\Sigma}^{-1}$ are $1/\sigma^2$. By using the scaled version of the covariance, factoring out ξ^2, we get the series in the parameter ε as defined in (5.1.9)

$$\begin{aligned}(\mathbf{C}_h + \mathbf{\Sigma})^{-1} &= \frac{1}{\sigma^2}[\mathbf{I} - \frac{\xi^2}{\sigma^2}\mathbf{Q}_h + \frac{\xi^4}{\sigma^4}\mathbf{Q}_h^2 - \cdots] \\ &= \frac{1}{\sigma^2}[\mathbf{I} - \varepsilon\mathbf{Q}_h + \varepsilon^2\mathbf{Q}_h^2 - \cdots]\end{aligned} \tag{6.1.2}$$

for $\varepsilon \ll 1$.

Here we want to approximate the inverse in (6.1.2) with the first two terms to get a first order correction. This is under the assumption that elements in the matrices $\mathbf{Q}_h^n$, associated with ε^n, do not grow too large. Thus, we truncate the series at $n = 1$.

$$(\mathbf{C}_h + \mathbf{\Sigma})^{-1} \approx \frac{1}{\sigma^2}[\mathbf{I} - \varepsilon\mathbf{Q}_h] \tag{6.1.3}$$

Next we compute the two terms in the analysis difference $(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}} + \mathbf{C}(\mathbf{U} - \tilde{\mathbf{U}})$

$$\begin{aligned}(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}} &\approx -\gamma\xi^2\mathbf{R}\mathbf{H}_h^T\frac{1}{\sigma^2}[\mathbf{I} - \varepsilon\tilde{\mathbf{Q}}_h]\mathbf{\Delta} \\ &= -\gamma\varepsilon\mathbf{R}\mathbf{H}_h^T[\mathbf{I} - \varepsilon(\mathbf{Q}_h + \gamma\mathbf{R}_h)]\mathbf{\Delta} \\ &= \varepsilon\gamma\mathbf{R}\mathbf{H}_h^T\left\{-\mathbf{I} + \varepsilon\left[\mathbf{Q}_h + \gamma\mathbf{R}_h\right]\right\}\mathbf{\Delta}\end{aligned} \tag{6.1.4}$$

and

$$\begin{aligned}\mathbf{C}(\mathbf{U} - \tilde{\mathbf{U}}) &= \xi^2\mathbf{Q}\mathbf{H}_h^T\left[(\mathbf{C}_h + \mathbf{\Sigma})^{-1} - (\tilde{\mathbf{C}}_h + \mathbf{\Sigma})^{-1}\right]\mathbf{\Delta} \\ &\approx \xi^2\mathbf{Q}\mathbf{H}_h^T\frac{1}{\sigma^2}\left[[\mathbf{I} - \varepsilon\mathbf{Q}_h] - [\mathbf{I} - \varepsilon\tilde{\mathbf{Q}}_h]\right]\mathbf{\Delta} \\ &= \varepsilon\mathbf{Q}\mathbf{H}_h^T\left[[\mathbf{I} - \varepsilon\mathbf{Q}_h] - [\mathbf{I} - \varepsilon(\mathbf{Q}_h + \gamma\mathbf{R}_h)]\right]\mathbf{\Delta} \\ &= \varepsilon^2\gamma\mathbf{Q}\mathbf{H}_h^T\mathbf{R}_h\mathbf{\Delta}\end{aligned} \tag{6.1.5}$$

The norm of the two lowest order terms from the analysis difference $(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}} + \mathbf{C}(\mathbf{U} - \tilde{\mathbf{U}})$ is then produced by adding (6.1.4) and (6.1.5)

$$\|\varepsilon\gamma\left\{-\mathbf{R}\mathbf{H}_h^T + \varepsilon[\mathbf{R}\mathbf{H}_h^T(\mathbf{Q}_h + \gamma\mathbf{R}_h) + \mathbf{Q}\mathbf{H}_h^T\mathbf{R}_h]\right\}\boldsymbol{\Delta}\| \tag{6.1.6}$$

with $\varepsilon \ll 1$ for dominating measurement errors.

This expression is dominated by the lowest order term $-\varepsilon\gamma\mathbf{R}\mathbf{H}_h^T\boldsymbol{\Delta}$. This term originates from the part $(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}}$. Note that there are two small parameters ε and γ outside the whole expression. This suggests that dominating measurement errors, $\varepsilon \ll 1$, may prevent a large error growth in $\mathbf{y}_{KF}^a - \bar{\mathbf{y}}_{EnKF}^a$ under our assumptions.

6.1.1 Analytic error growth

To get an idea of how the analysis difference depends on the number of measurements, we consider the dominating term $\varepsilon\gamma\mathbf{R}\mathbf{H}_h^T\boldsymbol{\Delta} \in \mathbb{R}^{N\times 1}$. The goal is to obtain simple, analytic expressions for the 2-norm of the error as a function of N_d.

To do this, we assume that the correlation length is short. This makes the factor r in $\mathbf{R}$ small.

The random correlation matrix $\mathbf{R}$ can be written as a series

$$\mathbf{R} = \sum_{i=0}^{N-1} \mathbf{R}_i r^i \tag{6.1.7}$$

where $\mathbf{R}_0$ is diagonal with elements μ_{ii}, $\mathbf{R}_1$ contains the first super- and sub diagonals of $\mathbf{R}$ with corresponding elements μ_{ij}, and so on.

When l is small, we choose to approximate $\mathbf{R}$ with $\mathbf{R}^\star$, which consists of the first two terms in the series

$$\begin{aligned}\mathbf{R}^\star &\approx \mathbf{R}_0 + \mathbf{R}_1 r \\ &= \begin{pmatrix} \mu_{1,1} & \mu_{1,2}r & 0 & \cdots & 0 \\ \mu_{2,1}r & \ddots & \ddots & & \vdots \\ 0 & \ddots & & & 0 \\ \vdots & & & \ddots & \mu_{N-1,N}r \\ 0 & \cdots & 0 & \mu_{N,N-1}r & \mu_{N,N} \end{pmatrix}\end{aligned} \tag{6.1.8}$$

The index-notation on μ_{ij} will be relaxed to simply μ, still keeping in mind that they are random and indeed different.

Let

$$\|\mathbf{S}\| = \|\mathbf{R}\mathbf{H}_h^T\boldsymbol{\Delta}\| = \sqrt{\sum_{i=1}^{N} S_i^2} \tag{6.1.9}$$

We need to find the sum of all the squared elements S_i^2. Note that the parameters ε and γ are left out here.

We consider two very simplified cases; a dense and a sparse measurement layout.

Dense measurements In this case we consider a dense layout where we have measurements in cells next to each other, so that $h_{dense} = 1$. Furthermore, we let the first measurement be in cell number 2 so that the matrix $\mathbf{R}^{\star}\mathbf{H}_{dense}^{T} \in \mathbb{R}^{N \times N_d}$ takes the form

$$\begin{pmatrix} \mu r & 0 & & & 0 \\ \mu & \mu r & & & \\ \mu r & \mu & \mu r & & \\ 0 & \mu r & \ddots & \ddots & \\ & & \ddots & & \mu r \\ & & & & \mu \\ 0 & & & 0 & \mu r \\ - & & - & & - \\ & & \mathbf{0} & & \end{pmatrix} \tag{6.1.10}$$

The elements of $\mathbf{S}$ can be summarized as follows; we have $N_d - 2$ elements equal to [1]$(\mu + 2\mu r)\Delta$, 2 elements equal to $(\mu + \mu r)\Delta$ and 2 elements equal to $(\mu r)\Delta$. The rest of the S_i-elements are zero.

The sum of the squares can be expressed as

$$\sum_{i=1}^{N} S_i^2 = N_d \left((\mu + 2\mu r)^2 \Delta^2 \right) \tag{6.1.11}$$

We have

$$\begin{aligned} \|\mathbf{S}_{dense}\| &= \sqrt{N_d \left((\mu + 2\mu r)^2 \Delta^2 \right)} \\ &= \Delta \sqrt{N_d \left(\mu^2 + 4\mu^2 r + 4\mu^2 r^2 \right)} \end{aligned} \tag{6.1.12}$$

After summing up the individual μ^2 values for a large N_d, we represent the μ values by some value μ_{mean} outside the square root

$$\|\mathbf{S}_{dense}\| = \mu_{mean} \Delta \sqrt{N_d \left(1 + 4r + 4r^2 \right)} \tag{6.1.13}$$

In this simplified setting, we see that the norm of the error increases slightly more than the square root of N_d.

Sparse measurements In this case we consider a sparse layout where we have measurements in every h_{sparse} cell, so that $h_{sparse} > 2$. Again, we let the measure-

[1] We defined all the entries in the vector $\mathbf{\Delta} \in \mathbb{R}^{N_d \times 1}$ equal; $\Delta_i = \Delta$ for all $i = 1, 2, \ldots N_d$ (see end of section 5.1.1)

ments start in cell 2, and obtain the matrix $\mathbf{R}^{\star}\mathbf{H}_{sparse}^{T} \in \mathbb{R}^{N\times N_d}$

$$\begin{pmatrix} \mu r & 0 & & & 0 \\ \mu & & & & \\ \mu r & \vdots & & & \\ 0 & \mu r & & & \\ & \mu & \ddots & & \\ & \mu r & & & \vdots \\ & 0 & \ddots & & \mu r \\ & & & & \mu \\ 0 & & & 0 & \mu r \\ - & & - & & - \\ & & \mathbf{0} & & \end{pmatrix} \tag{6.1.14}$$

Here, the nonzero elements of $\mathbf{S}$ will be either $\mu\Delta$ or $\mu r\Delta$. In particular, we have N_d of the $\mu\Delta$-elements and $2N_d$ of the $\mu r\Delta$-elements. The sum is therefore

$$\sum_{i=1}^{N} S_i^2 = N_d(\mu\Delta)^2 + 2N_d(\mu r\Delta)^2 \tag{6.1.15}$$

and the norm is then

$$\begin{aligned} \|\mathbf{S}_{sparse}\| &= \sqrt{N_d\Delta^2\left((\mu^2 + 2(\mu r)^2\right)} \\ &= \mu_{mean}\Delta\sqrt{N_d\left(1+2r^2\right)} \end{aligned} \tag{6.1.16}$$

Again, the norm of the error increases approximately with the square root of N_d. From the expressions (6.1.13) and (6.1.16) we expect that a sparse measurement layouts will result in a slightly smaller error norm than a dense layout when r is small Our illustrative case here is restricted to correlation lengths that are shorter than the spacing h_{sparse}. In figure 6.1 we compare the two expressions from (6.1.13) and (6.1.16).

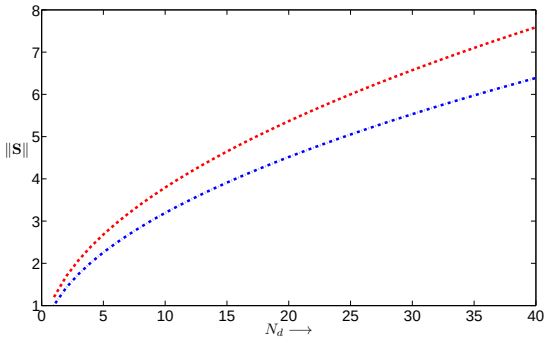


Figure 6.1: The expressions for dense(red) and sparse(blue) data layouts and number of measurements $N_d = [1, 40]$. It is assumed that we have short correlation.

6.1.2 Numerical experiments

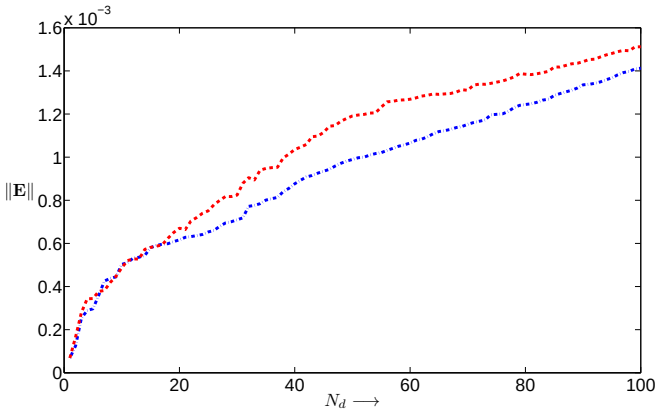
We present some numerical results in between the different cases we consider. In figures, we always use the notation $\|\mathbf{E}\|$ for the appropriate norm-expression that is considered.

In all the numerical experiments we use two different measurement layouts; a dense layout and a sparse layout. The dense layout has measurements in every second cell within the area, or line in this one dimensional, covered by the measurement layout. The sparse layout has measurements in every tenth cell within its "area", or line.

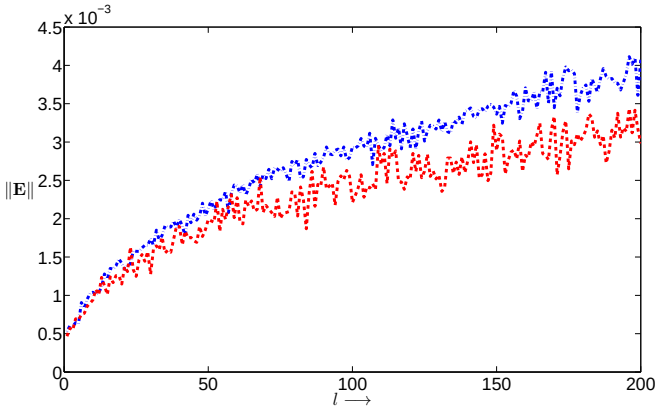
In this section we calculate the norm of the error from (6.1.6) for both dense and sparse measurement layouts. The expression is valid only when $\varepsilon \ll 1$.

Case 1: Increasing number of measurements, $\varepsilon \ll 1$ In this example we use a one dimensional field with $N = 1100$ cells. The correlation length l is given in number of cells, and is fixed at $l = 6$ in this case. We let the number of measurements increase from 1 measurement to 100 measurements, and calculated the norm-expression in (6.1.6). We included the terms of $\mathcal{O}(\varepsilon^2, \varepsilon\gamma)$ in the calculations, but they were too small to affect the result. A plot of this is shown in 6.2a.

Case 2: Increasing correlation length, $\varepsilon \ll 1$ Here, $N = 400$ and N_d was set to $N_d = 30$. The correlation length was increased from $l = 1$ up to $l = 200$ cells. The norm in (6.1.6) was calculated at each step and plotted, similar as in case 1.



(a) Plotted against increasing number of measurements.



(b) Plotted against increasing correlation length.

Figure 6.2: The norm-expression in (6.1.6) plotted for the dense(red) and the sparse(blue) measurement distribution.

Both plots suggests that the norm of the error is very small in the situation with dominating measurement errors.

6.2 Small measurement errors

In this section we look at the case where we have very accurate data or measurements, such that the model error is the dominating one. Thus, $\xi \gg \sigma$. The way the filter handles the uncertainties, means that the observed data will be weighted more than the predicted data. This situation is expected to be more challenging for the analysis step in EnKF.

We start off with the approximation of the inverse from

$$\mathbf{U} = \mathbf{H}_h^T(\mathbf{C}_h + \mathbf{\Sigma})^{-1}\mathbf{\Delta} \tag{6.2.1}$$

Now, from (4.2), the inverse can be developed into the Neumann series

$$(\mathbf{C}_h + \mathbf{\Sigma})^{-1} \approx (\mathbf{C}_h)^{-1}\left[\mathbf{I} - {\mathbf{C}_h}^{-1}\mathbf{\Sigma} + ({\mathbf{C}_h}^{-1})^2\mathbf{\Sigma}^2 - \cdots\right] \tag{6.2.2}$$

In (6.2.2) we have to invert the full matrix $\mathbf{C}_h$. Recall the simplification of the covariance structure we made in section 5.1.2. This gave us a matrix which we know the inverse of, namely $\mathbf{Q}^{-1}$ given in (5.1.12).

We use the parameter δ defined as

$$\delta \equiv \left(\frac{\sigma}{\xi}\right)^2 \ll 1$$

for small measurement variance σ^2. The Neumann series becomes

$$\begin{aligned}
(\mathbf{C}_h + \mathbf{\Sigma})^{-1} &= {\mathbf{C}_h}^{-1}\left[\mathbf{I} - {\mathbf{C}_h}^{-1}\mathbf{\Sigma} + ({\mathbf{C}_h}^{-1})^2\mathbf{\Sigma}^2 - \cdots\right] \\
&= \frac{1}{\xi^2}\mathbf{Q}_h^{-1}\left[\mathbf{I} - \frac{1}{\xi^2}\mathbf{Q}_h^{-1}\sigma^2\mathbf{I} + (\frac{1}{\xi^2}\mathbf{Q}_h^{-1})^2(\sigma^2\mathbf{I})^2 - \cdots\right] \\
&= \frac{1}{\xi^2}\mathbf{Q}_h^{-1}\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1} + \delta^2(\mathbf{Q}_h^{-1})^2 - \cdots\right] \\
&\approx \frac{1}{\xi^2}\mathbf{Q}_h^{-1}\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1}\right]
\end{aligned} \tag{6.2.3}$$

We approximate the inverse by truncating the series at $n = 1$ under similar assumptions as in (6.1.2). This now implies that elements in the matrix $\left(\mathbf{Q}_h^{-1}\right)^n$, associated with δ^n, are assumed not too large. Inserting the last line of (6.2.3) into $(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}}$ and $\mathbf{C}(\mathbf{U} - \tilde{\mathbf{U}})$ we get

$$\begin{aligned}
(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}} &= -\xi^2\gamma\mathbf{R}\mathbf{H}_h^T\frac{1}{\xi^2}\tilde{\mathbf{Q}}_h^{-1}\left[\mathbf{I} - \frac{1}{\xi^2}\tilde{\mathbf{Q}}_h^{-1}\sigma^2\mathbf{I} + \cdots\right]\mathbf{\Delta} \\
&\approx -\gamma\mathbf{R}\mathbf{H}_h^T\tilde{\mathbf{Q}}_h^{-1}\left[\mathbf{I} - \delta\tilde{\mathbf{Q}}_h^{-1}\right]\mathbf{\Delta} \\
&= -\gamma\mathbf{R}\mathbf{H}_h^T\left(\mathbf{Q}_h + \gamma\mathbf{R}_h\right)^{-1}\left[\mathbf{I} - \delta\left(\mathbf{Q}_h + \gamma\mathbf{R}_h\right)^{-1}\right]\mathbf{\Delta}
\end{aligned} \tag{6.2.4}$$

and

$$\begin{aligned}
\mathbf{C}(\mathbf{U} - \tilde{\mathbf{U}}) &\approx \xi^2\mathbf{Q}\mathbf{H}_h^T\left(\frac{1}{\xi^2}\mathbf{Q}_h^{-1}\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1}\right] - \frac{1}{\xi^2}\tilde{\mathbf{Q}}_h^{-1}\left[\mathbf{I} - \delta\tilde{\mathbf{Q}}_h^{-1}\right]\right)\mathbf{\Delta} \\
&= \xi^2\mathbf{Q}\mathbf{H}_h^T\frac{1}{\xi^2}\left(\mathbf{Q}_h^{-1}\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1}\right] - \left(\mathbf{Q}_h + \gamma\mathbf{R}_h\right)^{-1}\left[\mathbf{I} - \delta\left(\mathbf{Q}_h + \gamma\mathbf{R}_h\right)^{-1}\right]\right)\mathbf{\Delta} \\
&= \mathbf{Q}\mathbf{H}_h^T\left(\mathbf{Q}_h^{-1}\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1}\right] - \left(\mathbf{Q}_h + \gamma\mathbf{R}_h\right)^{-1}\left[\mathbf{I} - \delta\left(\mathbf{Q}_h + \gamma\mathbf{R}_h\right)^{-1}\right]\right)\mathbf{\Delta}
\end{aligned} \tag{6.2.5}$$

Notice that we have do not have any small parameter δ outside these expressions, compared to (6.1.6). We also notice that $\tilde{\mathbf{Q}}_h^{-1} = (\mathbf{Q}_h + \gamma\mathbf{R}_h)^{-1}$, appears twice in both expressions, associated with the term $\tilde{\mathbf{U}}$. The way we have defined this, we must consider the inverse of

$$(\mathbf{Q}_h + \gamma\mathbf{R}_h)^{-1} \equiv (\mathbf{HQH}^T + \gamma\mathbf{HRH}^T)^{-1} \tag{6.2.6}$$

The form of (6.2.6) is of the same form as (6.1.2) and (6.2.2). We develop this into a new Neumann series within the first one.

To obtain convergence of the Neumann series, we need to consider two separate situations, depending upon which of $\mathbf{Q}_h$ and $\gamma\mathbf{R}_h$ having the larger norm, see (4.2).

The matrices $\mathbf{Q}_h$ and $\mathbf{R}_h$ are both scaled covariance matrices with main diagonal entries equal to unity, but in $\mathbf{R}_h$, all the elements are multiplied by random numbers $\mu_{ii} \sim (0, \zeta)$ with variance ζ small. This "shifts" the elements of $\mathbf{R}_h$ to be distributed around zero, while the elements of $\mathbf{Q}_h$ ranges between 0 and 1. With this in mind, we may say that $\|\mathbf{Q}_h\| = \mathcal{O}(\|\mathbf{R}_h\|)$ or slightly bigger.
If, in addition, $\gamma \ll 1$ then we are confident that the norm of $\mathbf{Q}_h$ is significantly bigger than the the norm of $\gamma\mathbf{R}_h$. This case is now considered in (6.2.1).

On the other hand, if $\gamma \gg 1$ then we might say that the norm of $\gamma\mathbf{R}_h$ is the largest. This is covered later in (6.2.2).

6.2.1 Large ensemble size

Here we assume that we have very accurate data, $\delta \ll 1$, and that we have a sufficiently large ensemble in the EnKF. In this case γ is very small, $\gamma \ll 1$, because we have a sufficiently number of ensemble members, so that the sample covariance matrix $\tilde{\mathbf{C}}$ is not too different from the real covariance matrix $\mathbf{C}$, see (5.1.13).

Using Neumann Series to insert for $(\mathbf{Q}_h + \gamma\mathbf{R}_h)^{-1}$ when $\gamma \ll 1$ we get

$$\begin{aligned}
(\mathbf{Q}_h + \gamma\mathbf{R}_h)^{-1} &= \left(\left[\mathbf{I}+\gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\mathbf{Q}_h\right)^{-1} \\
&= \mathbf{Q}_h^{-1}\sum_{i=0}^{\infty}\left(-\gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right)^i \\
&\approx \mathbf{Q}_h^{-1}\left[\mathbf{I} - \gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]
\end{aligned} \tag{6.2.7}$$

Inserting this into (6.2.4) and (6.2.5) gives

$$\begin{aligned}
(\mathbf{C} - \tilde{\mathbf{C}})\tilde{\mathbf{U}} &\approx (-\xi^2\gamma\mathbf{R})\mathbf{H}_h^T\frac{1}{\xi^2}\mathbf{Q}_h^{-1}\left[\mathbf{I} - \gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1}\left[\mathbf{I} - \gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\right]\boldsymbol{\Delta} \\
&= -\gamma\mathbf{R}\mathbf{H}_h^T\mathbf{Q}_h^{-1}\left[\mathbf{I} - \gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1} + \delta\gamma\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\boldsymbol{\Delta} \\
&= -\gamma\mathbf{R}\mathbf{H}_h^T\mathbf{Q}_h^{-1}\left[\mathbf{I} - \delta\mathbf{Q}_h^{-1} - \gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\Delta + \mathcal{O}(\delta\gamma^2) + \mathcal{O}(\delta\gamma^3)
\end{aligned} \tag{6.2.8}$$

and

$$\begin{aligned}\mathbf{C}(\mathbf{U}-\tilde{\mathbf{U}}) \approx & (\xi^2\mathbf{Q})\mathbf{H}_h^T\frac{1}{\xi^2}\left(\mathbf{Q}_h^{-1}\left[\mathbf{I}-\delta\mathbf{Q}_h^{-1}\right]\right. \\ & \left.-\mathbf{Q}_h^{-1}\left[\mathbf{I}-\gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\left[\mathbf{I}-\delta\mathbf{Q}_h^{-1}\left[\mathbf{I}-\gamma\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\right]\right)\boldsymbol{\Delta} \\ = & \mathbf{Q}\mathbf{H}_h^T\left(\left[\mathbf{Q}_h^{-1}-\delta\mathbf{Q}_h^{-1}\mathbf{Q}_h^{-1}\right]\right. \\ & \left.-\left[\mathbf{Q}_h^{-1}-\gamma\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\left[\mathbf{I}-\delta\mathbf{Q}_h^{-1}+\delta\gamma\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}\right]\right)\boldsymbol{\Delta} \\ = & \gamma\mathbf{Q}\mathbf{H}_h^T\left(\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}\right. \\ & \left.-\delta\left(\mathbf{Q}_h^{-1}\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}+\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}\mathbf{Q}_h^{-1}\right)\right)\Delta+\mathcal{O}(\delta\gamma^2)\end{aligned} \tag{6.2.9}$$

where two terms canceled out from the $(\mathbf{U}-\tilde{\mathbf{U}})$-difference, namely $\mathbf{Q}_h^{-1}$ and $\delta(\mathbf{Q}_h^{-1})^2$.

Adding the two last equations produces the expression for the analysis difference $\|(\mathbf{C}-\tilde{\mathbf{C}})\tilde{\mathbf{U}}+\mathbf{C}(\mathbf{U}-\tilde{\mathbf{U}})\|$ in the case of small measurement errors and a sufficiently large number of ensemble members.

$$\begin{aligned}= & \left\|\gamma\left\{-\mathbf{R}\mathbf{H}^T+\mathbf{Q}\mathbf{H}^T\mathbf{Q}_h^{-1}\mathbf{R}_h+\gamma\left[\mathbf{R}\mathbf{H}^T\mathbf{Q}_h^{-1}\mathbf{R}_h\right]\right.\right. \\ & \left.\left.+\delta\left[\mathbf{R}\mathbf{H}^T\mathbf{Q}_h^{-1}-\mathbf{Q}\mathbf{H}^T\left(\mathbf{Q}_h^{-1}\mathbf{Q}_h^{-1}\mathbf{R}_h+\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}\right)\right]\right\}\mathbf{Q}_h^{-1}\boldsymbol{\Delta}\right\| \\ & +\mathcal{O}(\delta\gamma^2,\delta\gamma^3)\end{aligned} \tag{6.2.10}$$

We note that γ is the only parameter which can make this small. The expression (6.2.10) is valid only for $\delta,\gamma\ll 1$. We see that it is dominated by the two terms

$$-\gamma\mathbf{R}\mathbf{H}^T\mathbf{Q}_h^{-1}\boldsymbol{\Delta}+\gamma\mathbf{Q}\mathbf{H}^T\mathbf{Q}_h^{-1}\mathbf{R}_h\mathbf{Q}_h^{-1}\boldsymbol{\Delta} \tag{6.2.11}$$

and that the entire expression is multiplied with $\mathbf{Q}_h^{-1}$.

6.2.1.1 Analytic error growth

Before we perform numerical experiments with the norm of the error in (6.2.10), we have an analytical view of the term $\mathbf{R}\mathbf{H}^T\mathbf{Q}_h^{-1}\boldsymbol{\Delta}$ for illustrative purposes.

As in section 6.1.1, this is done in hope of getting an analytical insight into the error growth.

We now consider only the case with dense measurements, using an observation matrix $\mathbf{H}_{dense}\in\mathbb{R}^{N_d\times N}$ together with the approximation of $\mathbf{R}$ as done in (6.1.1). Then we again assume a small correlation length.

This results in the same matrix $\mathbf{R}^{\star}\mathbf{H}_{dense}^{T} \in \mathbb{R}^{N \times N_d}$ given in (6.1.10). We write $\mathbf{R}\mathbf{H}^{T}\mathbf{Q}_{h}^{-1}$ as

$$\begin{pmatrix} \mu r & 0 & & & 0 \\ \mu & \mu r & & & \\ \mu r & \mu & \mu r & & \\ 0 & \mu r & \ddots & \ddots & \\ & & \ddots & & \mu r \\ & & & & \mu \\ 0 & & & 0 & \mu r \\ - & & - & & - \\ & & \mathbf{0} & & \end{pmatrix} \frac{1}{1-r^2} \begin{pmatrix} 1 & -r & 0 & & 0 \\ -r & 1+r^2 & \ddots & & \\ 0 & -r & \ddots & -r & 0 \\ & & \ddots & 1+r^2 & -r \\ 0 & & 0 & -r & 1 \end{pmatrix} \tag{6.2.12}$$

Again, we have relaxed the index-notation on the random values for μ_{ij}. We get a matrix of the form

$$\frac{1}{1-r^2} \begin{pmatrix} \mu r & -\mu r^2 & 0 & & 0 \\ x & \beta & \ddots & & \\ \alpha & x & \ddots & -\mu r^2 & 0 \\ -\mu r^2 & \beta & \ddots & \beta & -\mu r^2 \\ 0 & -\mu r^2 & \ddots & x & \alpha \\ & & \ddots & \beta & x \\ 0 & & 0 & -\mu r^2 & \mu r \\ - & & - & & - \\ & & \mathbf{0} & & \end{pmatrix} \tag{6.2.13}$$

with elements x, α and β given as

$$\begin{aligned} x &= \mu - \mu r^2 \\ \alpha &= \mu r - \mu r \\ \beta &= \mu r - \mu r + \mu r^3 \\ &= \alpha + \mu r^3 \end{aligned} \tag{6.2.14}$$

These expressions contain values of μ that are random perturbations distributed around zero. This gives a rather unpredictable expression to handle. We expect that the values μ alternate the elements. They may cancel each other out, or they may double up when they are of same order of magnitude.

To get any further we need to ease the handling of μ.
If we allow the assumption that we can factor out the values μ in (6.2.14) then

$$\begin{aligned} x &= \mu(1 - r^2) \\ \alpha &= (\mu - \mu)r \\ \beta &= (\mu - \mu)r + \mu r^3 \end{aligned} \tag{6.2.15}$$

For values $\mu$of same order of magnitude, we may have that α, β are small enough to be neglected. This would be a best case scenario with errors canceling each other out. Furthermore,

$$\begin{aligned} \frac{x}{1-r^2} &= \mu\frac{1-r^2}{1-r^2} = \mu \\ \frac{\mu r}{1-r^2} &\to \mu r \\ -\frac{\mu r^2}{1-r^2} &\to -\mu r^2 \end{aligned} \tag{6.2.16}$$

for small values of r.

In that case, the elements of (6.2.13), together with the factor $\frac{1}{1-r^2}$, becomes

$$\begin{pmatrix} \mu r & -\mu r^2 & 0 & & 0 \\ \mu & 0 & \ddots & & \\ 0 & \mu & \ddots & -\mu r^2 & 0 \\ -\mu r^2 & 0 & \ddots & 0 & -\mu r^2 \\ 0 & -\mu r^2 & \ddots & \mu & 0 \\ & & \ddots & 0 & \mu \\ 0 & & 0 & -\mu r^2 & \mu r \\ - & & - & & - \\ & & \mathbf{0} & & \end{pmatrix} = \mu \begin{pmatrix} r & -r^2 & 0 & & 0 \\ 1 & 0 & \ddots & & \\ 0 & 1 & \ddots & -r^2 & 0 \\ -r^2 & 0 & \ddots & 0 & -r^2 \\ 0 & -r^2 & \ddots & 1 & 0 \\ & & \ddots & 0 & 1 \\ 0 & & 0 & -r^2 & r \\ - & & - & & - \\ & & \mathbf{0} & & \end{pmatrix} \tag{6.2.17}$$

With the norm

$$\|\mathbf{S}\| = \|\mathbf{R}\mathbf{H}_h^T\mathbf{Q}_h^{-1}\mathbf{\Delta}\| = \sqrt{\sum_{i=1}^{N} S_i^2} \tag{6.2.18}$$

we summarize the nonzero S_i-elements as follows: We have N_d elements equal to $\mu(1-r^2)\Delta$ and 2 elements equal to $\mu(r-r^2)\Delta$. The norm is then given as

$$\begin{aligned} \|\mathbf{S}\| &= \sqrt{\sum_{i=1}^{N} S_i^2} \\ &= \sqrt{N_d\left(\mu(1-r^2)\Delta\right)^2 + 2\left(\mu(r-r^2)\Delta\right)^2} \\ &\approx \sqrt{(\mu\Delta)^2\left(N_d(1-2r^2)+2r^2\right)} \\ &= \mu_{mean}\Delta\sqrt{N_d(1-2r^2)+2r^2} \end{aligned} \tag{6.2.19}$$

where we discarded terms of $\mathcal{O}(r^3)$ and smaller.

This expression of the norm is yet again grossly simplified, and the values of μ are not handled correctly. We may interpret this error growth as part of the dominating

trend, keeping in mind that the random perturbations from μ will alternate the errors. It also represents a best case scenario, since we let values of μ cancel each other out. This best case scenario for the term $\mathbf{R}\mathbf{H}_h^T\mathbf{Q}_h^{-1}\mathbf{\Delta}$ with $\xi \gg \sigma$ is shown in figure 6.3. We recognize the trend from figure 6.1, with an error growth similar to the square root of N_d.

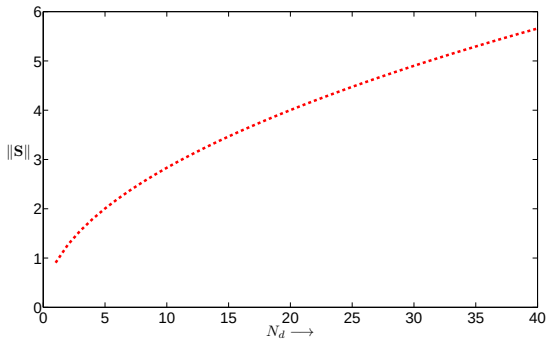
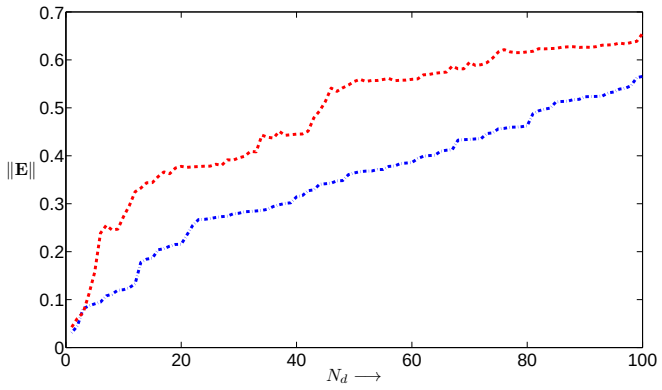


Figure 6.3: Best case scenario error growth from (6.2.19) of the term $\mathbf{R}\mathbf{H}_h^T\mathbf{Q}_h^{-1}\mathbf{\Delta}$.

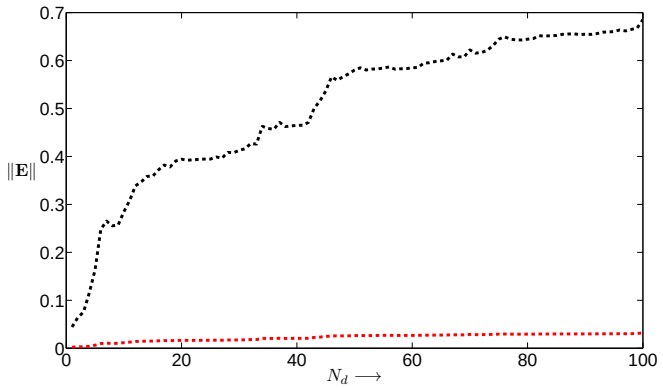
6.2.1.2 Numerical experiments

In this section we test out the norm-expression from (6.2.10) for both dense and sparse measurement layouts. The expression is valid only if $\delta, \gamma \ll 1$. This means that we have small measurement errors compared to the model errors, and that we have a relatively large ensemble.

Case 3: Increasing number of measurements, $\delta, \gamma \ll 1$ We used the same set up for this case as in case 1, namely $N = 1100$ cells, correlation length set as $l = 6$ and number of assimilated data N_d increasing from 1 to 100.



(a) The dense(red) and sparse(blue) measurement layouts against increasing number of measurements.

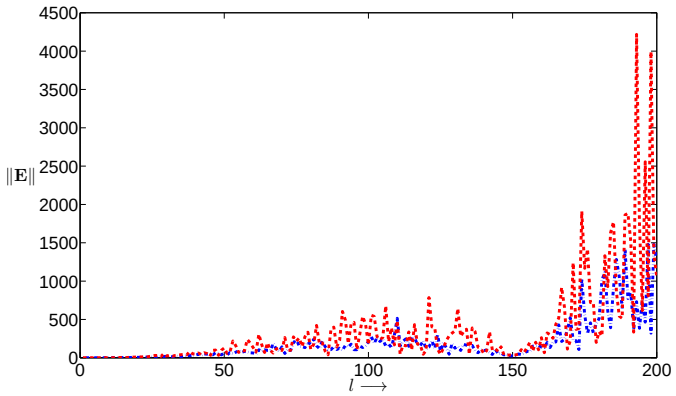


(b) The dense measurement layout from a). Black is the terms of $\mathcal{O}(\gamma)$, red is $\mathcal{O}(\gamma^2, \delta\gamma)$.

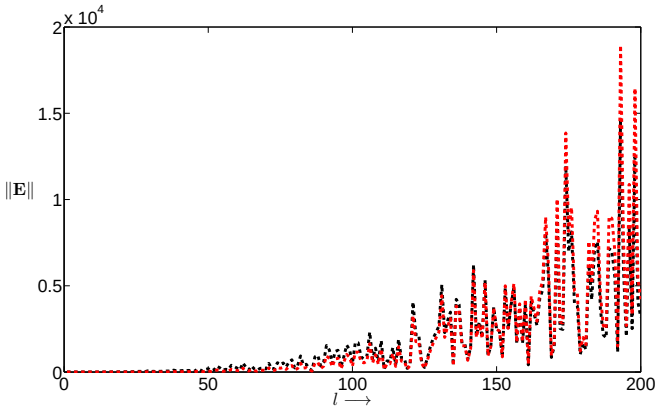
Figure 6.4: The norm of (6.2.10) plotted for increasing N_d.

Figure 6.4 suggests that a dense measurement layout results in a larger norm of the error. We also see that the norm of the lowest order term, $\mathcal{O}(\gamma)$ from (6.2.10) is dominating the norm of the higher order terms $\mathcal{O}(\gamma^2, \delta\gamma)$ as expected.

Case 4: Increasing correlation length, $\delta, \gamma \ll 1$



(a) Dense(red) and sparse(blue) layout from (6.2.10)



(b) The two lowest order terms from a). Black is the terms of $\mathcal{O}(\gamma)$, red is $\mathcal{O}(\gamma^2, \delta\gamma)$.

Figure 6.5: The norm of (6.2.10) plotted for increasing correlation length.

In figure 6.5(a), the norm grows rapidly as the correlation length gets beyond a certain value. In 6.5(b) we observe that the norm of the $\mathcal{O}(\gamma^2, \delta\gamma)$-terms is approximately as big as the $\mathcal{O}(\gamma)$-term, at least for long correlation lengths. This may suggest that such long correlation lengths violates our assumption in (6.2.3).

6.2.2 Small ensemble size

Here we still assume that we have very accurate data, $\delta \ll 1$, but that we approximate the covariance matrix with a relatively small ensemble represented by $\gamma \gg 1$.
We will put $\gamma \mathbf{R}_h$ outside the brackets to give an approximation. We do not know

how $\mathbf{R}^{-1}$ will behave.

$$\begin{aligned}(\mathbf{Q}_h + \gamma\mathbf{R}_h)^{-1} &= \left(\left[\mathbf{I}+\mathbf{Q}_h\left(\gamma\mathbf{R}_h\right)^{-1}\right]\gamma\mathbf{R}_h\right)^{-1} \\ &= \left(\gamma\mathbf{R}_h\right)^{-1}\sum_{i=0}^{\infty}\left(-\mathbf{Q}_h\left(\gamma\mathbf{R}_h\right)^{-1}\right)^i \\ &\approx \frac{1}{\gamma}\mathbf{R}_h^{-1}\left[\mathbf{I}-\frac{1}{\gamma}\mathbf{Q}_h\mathbf{R}_h^{-1}\right]\end{aligned} \tag{6.2.20}$$

Inserting this into the full expression for the analysis error leads to

$$\begin{aligned}(\mathbf{C}-\tilde{\mathbf{C}})\tilde{\mathbf{U}} &= (-\xi^2\gamma\mathbf{R})\mathbf{H}^T\frac{1}{\xi^2}\frac{1}{\gamma}\mathbf{R}_h^{-1}\left[\mathbf{I}-\frac{1}{\gamma}\mathbf{Q}_h\mathbf{R}_h^{-1}\right]\left[\mathbf{I}-\delta\frac{1}{\gamma}\mathbf{R}_h^{-1}\left[\mathbf{I}-\frac{1}{\gamma}\mathbf{Q}_h\mathbf{R}_h^{-1}\right]\right]\boldsymbol{\Delta} \\ &= -\mathbf{R}\mathbf{H}^T\mathbf{R}_h^{-1}\left(\mathbf{I}-\frac{1}{\gamma}\mathbf{Q}_h\mathbf{R}_h^{-1}-\frac{\delta}{\gamma}\mathbf{R}_h^{-1}\right. \\ &\left.+\frac{\delta}{\gamma^2}(\mathbf{R}_h^{-1}\mathbf{Q}_h\mathbf{R}_h^{-1}+\mathbf{Q}_h\mathbf{R}_h^{-1}\mathbf{R}_h^{-1})-\frac{\delta}{\gamma^3}\mathbf{Q}_h\mathbf{R}_h^{-1}\mathbf{R}_h^{-1}\mathbf{Q}_h\mathbf{R}_h^{-1}\right)\boldsymbol{\Delta}\end{aligned} \tag{6.2.21}$$

and

$$\begin{aligned}\mathbf{C}(\mathbf{U}-\tilde{\mathbf{U}}) &= (\xi^2\mathbf{Q})\mathbf{H}^T\frac{1}{\xi^2}\left(\mathbf{Q}_h^{-1}\left[\mathbf{I}-\delta\mathbf{Q}_h^{-1}\right]\right. \\ &\left.-\frac{1}{\gamma}\mathbf{R}_h^{-1}\left[\mathbf{I}-\frac{1}{\gamma}\mathbf{Q}_h\mathbf{R}_h^{-1}\right]\left[\mathbf{I}-\delta\frac{1}{\gamma}\mathbf{R}_h^{-1}\left[\mathbf{I}-\frac{1}{\gamma}\mathbf{Q}_h\mathbf{R}_h^{-1}\right]\right]\right)\boldsymbol{\Delta} \\ &= \mathbf{Q}\mathbf{H}^T\left(\mathbf{Q}_h^{-1}-\delta\mathbf{Q}_h^{-1}\mathbf{Q}_h^{-1}-\frac{1}{\gamma}\mathbf{R}_h^{-1}+\frac{1}{\gamma^2}\mathbf{R}_h^{-1}\mathbf{Q}_h\mathbf{R}_h^{-1}\right. \\ &\left.+\frac{\delta}{\gamma^2}\mathbf{R}_h^{-1}\mathbf{R}_h^{-1}-\frac{\delta}{\gamma^3}(\mathbf{R}_h^{-1}\mathbf{R}_h^{-1}\mathbf{Q}_h\mathbf{R}_h^{-1}+\mathbf{R}_h^{-1}\mathbf{Q}_h\mathbf{R}_h^{-1}\mathbf{R}_h^{-1})+\frac{\delta}{\gamma^4}(\mathbf{R}_h^{-1}\mathbf{Q}_h\mathbf{R}_h^{-1})^2\right)\boldsymbol{\Delta}\end{aligned} \tag{6.2.22}$$

Summing up the terms results in the analysis difference $\|(\mathbf{C}-\tilde{\mathbf{C}})\tilde{\mathbf{U}}+\mathbf{C}(\mathbf{U}-\tilde{\mathbf{U}})\|$

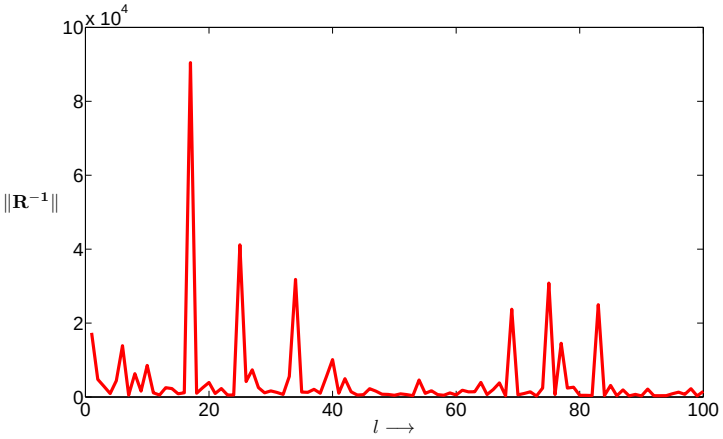
$$\begin{aligned}&= \left\|\left\{\mathbf{Q}\mathbf{H}^T\left(\mathbf{I}-\delta\mathbf{Q}_h^{-1}\right)\right\}\mathbf{Q}_h^{-1}\boldsymbol{\Delta}\right. \\ &\quad+\left\{-\mathbf{R}\mathbf{H}^T+\frac{1}{\gamma}\left[\mathbf{R}\mathbf{H}^T\mathbf{R}_h^{-1}\mathbf{Q}_h-\mathbf{Q}\mathbf{H}^T\right]\right. \\ &\quad\left.\left.+\frac{1}{\gamma^2}\left[\mathbf{Q}\mathbf{H}^T\mathbf{R}_h^{-1}\mathbf{Q}_h\right]+\frac{\delta}{\gamma}\left[\mathbf{R}\mathbf{H}^T\mathbf{R}_h^{-1}\right]\right\}\mathbf{R}_h^{-1}\boldsymbol{\Delta}\right\| \\ &\quad+\mathcal{O}(\delta/\gamma^2,\delta/\gamma^3,\delta/\gamma^4)\end{aligned} \tag{6.2.23}$$

In this situation there is no small parameter to limit the error. Both $\mathbf{Q}_h^{-1}$ and $\mathbf{R}_h^{-1}$ appear in this expression, as opposed to (6.2.10) where we only had $\mathbf{Q}_h^{-1}$-terms.

The difference between (6.2.10) and (6.2.23) is the dependence on $\mathbf{Q}_h^{-1}$ and $\mathbf{R}_h^{-1}$ respectively, and the presence of the parameter γ.
The behavior of these two inverse matrices $\mathbf{Q}_h^{-1}$ and $\mathbf{R}_h^{-1}$ are important. Clearly, we expect the latter to be the worst case because this represents a situation with relatively few ensemble members.

6.2.2.1 Numerical experiments

First we look at the size of the norm of $\mathbf{R}_h^{-1}$. In figure 6.6 we observe that the norm of $\mathbf{R}^{-1}$ varies in an extremely and irregularly manner. The behavior of $\mathbf{R}^{-1}$ seems to cause large errors in the norm



(a)

Figure 6.6: Plot of $\|\mathbf{R}^{-1}\|$ for increasing correlation length l.

Case 5: Increasing number of measurements, $\delta \ll 1$, $\gamma \gg 1$ In this experiment we used $N = 550$, $l = 6$ and N_d ranged from 1 to 50.

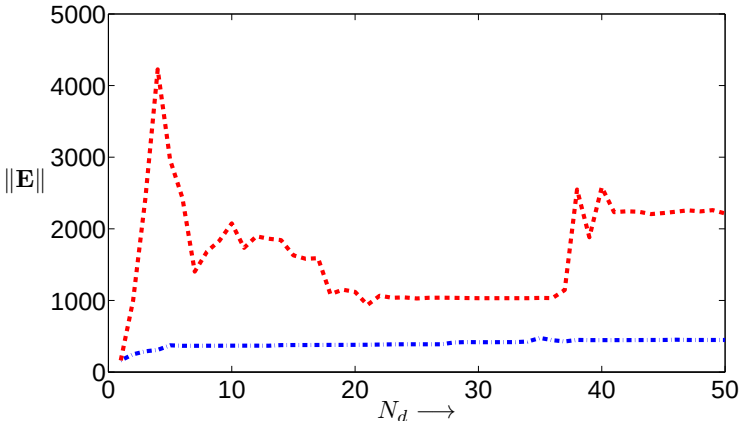


Figure 6.7: Expression (6.2.2) computed with dense measurements(red) and sparse measurements(blue) for increasing N_d.

In this case we see that the error grows large very fast, which is what to be expected. This was the result in all plots of (6.2.23). The case with increasing correlation length gave similar results. The plots from this case The norm of (6.2.23) then resembled the behavior of figure 6.6 of $\|\mathbf{R}^{-1}\|$. In addition, the terms associated with $1/\gamma$, $1/\gamma^2$ and δ/γ from (6.2.23) dominated the expression. This may suggest that our assumption on orders of magnitude is violated her.

This scenario is clearly the worst, as we expect with few ensemble members.

Chapter 7

Summary and Conclusions

In chapter 2 we presented both linear and non linear inverse problems together with regularization techniques and non linear solution methods like Gauss Newton and Levenberg Marquardt. Then Bayesian inference were introduced as well as the Kalman Filter and Extended Kalman Filter and the Ensemble Kalman Filter.

The focus in this thesis has been on the Ensemble Kalman Filter analysis step, and on the norm of the sampling error. We have used truncated Neumann series to approximate the inverse in the Kalman Gain matrix. Under assumptions on the structures of the covariance matrix and the sample covariance matrix we found approximate expressions of the sampling error. We have considered varying measurement errors.

Dominating measurement errors relative to the model error seems to be handled well by the analysis step in EnKF. In section 6.1, the small parameter ε seemed to balance even a poorly estimated covariance matrix.
At the end of the Kalman Filter section 3.2, we mentioned that when the uncertainty in the assimilated data is large compared to the uncertainty in the model errors, the analysis step gives less weight to the data. Thus, the updating of the variables is limited in this case, which makes the difference between the Kalman update and the Ensemble Kalman update relatively small.
Here we also arrived at simple analytic expressions where the norm increases as the square root of number of measurements.

When we considered small measurement errors we looked at two cases.
This is the most interesting results since the situation of small measurement errors may be close to the assumption of negligible errors in [14, 15] .
In the first we assumed that we used a large ensemble. We observed that the norm of the analysis difference became larger than with the case $\varepsilon \ll 1$. The expression we derived suggests that we are more dependent on a large ensemble in the case of small measurement errors. Accurate measurements are weighted more in the Kalman Gain, leading to stronger updating of correlated variables near the measurement location. The sample covariance matrix may then update physically uncorrelated variables due to spurious correlations.

Increasing the number of measurements made the norm of the sampling error larger. We also experienced that a dense distribution of the measurements resulted in a larger norm of the sampling error, at least for relative short correlation lengths.

The situation considered where we assumed few ensemble members gave very large norm of the sampling errors.

Future work

Our approach is limited to big contrasts between the measurement error and the model error. And in the case of small measurement errors, our expressions are restricted to either very big or very small numbers of ensemble members.

Also, these calculations was only done in one dimension. It could be interesting to expand to two dimensions and compare with [14, 15]

Appendix A

A.1 Random vector

A random vector $\mathbf{X} \in \mathbb{R}^N$ consists of the random variables X_i for $i = 1, 2, \ldots N$.

A random variable X is a function $X(s)$ that assigns a value to each outcome s in the sample space S. A realization of X is a particular value, obtained by evaluating the random variable.

A.2 Covariance

The covariance of the elements of a random vector $\mathbf{X}$ with expected value $\boldsymbol{\mu}_X$ is defined as

$$\begin{aligned}\mathbf{C}_x &= E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T] \\ &= E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T\end{aligned}$$

This forms a matrix with the elements $E[X_iX_j] - \mu_i\mu_j$. If the components of $\mathbf{X}$ are uncorrelated, the covariance matrix is a diagonal matrix.

A.3 Multivariate Gaussian probability density function

A random N-dimensional vector $\mathbf{X}$ is said to be multivariate Gaussian with with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$ if the probability density for $\mathbf{X}$ is

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{det(\mathbf{C})}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu}$ is a vector containing the expected values of $X_1, \cdots, X_n$ and the matrix $\mathbf{C}$ contains the covariances between the random variables

$$\begin{aligned}\boldsymbol{\mu} &= E[\mathbf{X}] \\ C_{i,j} &= Cov(X_i, X_j)\end{aligned}$$

A.4 Covariance of a linear transformation

Let $\mathbf{Y} = \mathbf{AX}$.

Then the expectation of $\mathbf{Y}$ is

$$\boldsymbol{\mu}_\mathbf{Y} = E[\mathbf{Y}] = E[\mathbf{AX}] = \mathbf{A}E[\mathbf{X}] = \mathbf{A}\boldsymbol{\mu}_\mathbf{X}$$

Then the covariance of $\mathbf{Y}$ is found by insertion

$$\begin{aligned}
\mathbf{C}_\mathbf{Y} &= E[(\mathbf{Y} - E[\mathbf{Y}])(\mathbf{Y} - E[\mathbf{Y}])^T] \\
&= E[(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu}_\mathbf{X})(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu}_\mathbf{X})^T] \\
&= E[\mathbf{AX}\mathbf{X}^T\mathbf{A}^T - \mathbf{AX}\boldsymbol{\mu}_\mathbf{X}^T\mathbf{A}^T - \mathbf{A}\boldsymbol{\mu}_\mathbf{X}\mathbf{X}^T\mathbf{A}^T + \mathbf{A}\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T\mathbf{A}^T] \\
&= \mathbf{A}(E[\mathbf{X}\mathbf{X}^T - \mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T - \boldsymbol{\mu}_\mathbf{X}\mathbf{X}^T + \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T])\mathbf{A}^T \\
&= \mathbf{A}(E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T + \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T)\mathbf{A}^T \\
&= \mathbf{A}(E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^T)\mathbf{A}^T \\
&= \mathbf{A}\mathbf{C}_\mathbf{X}\mathbf{A}^T
\end{aligned}$$

A.5 Covariance models

The covariance between X_i and X_j may be assumed to be a function of the distance h only. Examples of such models are the spherical model and the exponential family of covariance functions respectively:

$$C(h) = \xi^2 \begin{cases} 1 - \frac{3h}{2l} + \frac{h^3}{2l^3} & for\, 0 \leq h \leq l \\ 0 & for\, h > l \end{cases}$$

$$C(h) = \xi^2 \exp(-3(h/l)^v)$$

The last one is called the exponential covariance function when $v = 1$ and the Gaussian covariance when $v = 2$. l is the correlation range. All covariance matrices are semi positive definite and symmetric, thus

$$\begin{aligned}
C_{i,j} &\geq 0 \\
\mathbf{C} &= \mathbf{C}^T
\end{aligned}$$

for all i, j.

A.6 Random realizations

A random realization $\mathbf{X}$ from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$ can be achieved by adding a term $\mathbf{LZ}$ to the mean value $\boldsymbol{\mu}$

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{LZ}$$

where $\mathbf{Z}$ is a vector of independent identically distributed random variables with zero mean and variance one, and $\mathbf{L}$ is the Cholesky composition of $\mathbf{C}$ such that $\mathbf{C} = \mathbf{L}\mathbf{L}^T$.

This can be shown in the following way:

The expected value of $\mathbf{X}$ is

$$E[\mathbf{X}] = E[\boldsymbol{\mu} + \mathbf{LZ}] = \boldsymbol{\mu} + \mathbf{L}E[\mathbf{Z}] = \boldsymbol{\mu} + \mathbf{0} = \boldsymbol{\mu}$$

and the covariance of $\mathbf{X}$ is

$$E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = E[\mathbf{LZ}(\mathbf{LZ})^T] = \mathbf{L}E[\mathbf{Z}\mathbf{Z}^T]\mathbf{L}^T = \mathbf{L}\mathbf{I}\mathbf{L}^T = \mathbf{L}\mathbf{L}^T$$

Thus, the vector $\mathbf{X}$ is a random realization from the appropriate distribution.

A.7 Bayes Theorem

The discrete case of Bayes Theorem states that the conditional probability of event A, given event B, is given as

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

The factors on the right hand side are often called

- $P(A)$ is the prior probability, as it does not take into account any information about B.
- $P(B \mid A)$ is the conditional probability of B given A, also called the likelihood.
- $P(B)$ is the marginal probability of B, and acts as a normalizing constant.

We call the left hand side the posterior probability. This is the conditional probability of A given B.

A.8 Proof of convergence on Neumann Series

We must show that the series $\sum_{i=0}^{\infty} \mathbf{A}^i = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \ldots + \mathbf{A}^k + \ldots$ converges to the inverse of $(\mathbf{I} - \mathbf{A})$ when $\|\mathbf{A}\| < 1$.

By the associative and distributive properties of matrices we have

$$\begin{aligned}
&(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A} + \ldots \mathbf{A}^k) \\
&= \mathbf{I}(\mathbf{I} + \mathbf{A} + \ldots \mathbf{A}^k) - \mathbf{A}(\mathbf{I} + \mathbf{A} + \ldots \mathbf{A}^k) \\
&= \mathbf{I} - \mathbf{A}^{k+1}
\end{aligned}$$

Multiply both sides by the inverse of $(\mathbf{I} - \mathbf{A})$ to get

$$\begin{aligned}
(\mathbf{I} + \mathbf{A} + \ldots \mathbf{A}^k) &= (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^{k+1}) \\
&= (\mathbf{I} - \mathbf{A})^{-1}\mathbf{I} - (\mathbf{I} - \mathbf{A})^{-1}\mathbf{A}^{k+1}
\end{aligned}$$

which gives

$$(\mathbf{I} + \mathbf{A} + \ldots \mathbf{A}^k) - (\mathbf{I} - \mathbf{A})^{-1} = -(\mathbf{I} - \mathbf{A})^{-1}\mathbf{A}^{k+1}$$

Now we use the fact that $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$

$$\begin{aligned}
\|(\mathbf{I} + \mathbf{A} + \ldots \mathbf{A}^k) - (\mathbf{I} - \mathbf{A})^{-1}\| &= \| - (\mathbf{I} - \mathbf{A})^{-1}\mathbf{A}^{k+1}\| \\
&\leq \| - (\mathbf{I} - \mathbf{A})^{-1}\|\|\mathbf{A}^{k+1}\| \\
&\leq \| - (\mathbf{I} - \mathbf{A})^{-1}\|\|\mathbf{A}\|^{k+1}
\end{aligned}$$

Since the norm of $\mathbf{A}$ is less than one, the right side must go to zero. Thus, the series must converge to the inverse of $(\mathbf{I} - \mathbf{A})$

Bibliography

[1] S. Aanonsen, G. Nævdal, D. Oliver, A. Reynolds, and B. Vallès, *The Ensemble Kalman Filter in Reservoir Engineering–a Review*, SPE Journal, 14 (2009), pp. 393–412.

[2] R. Aster, B. Borchers, C. Thurber, and E. Corporation, *Parameter estimation and inverse problems*, Elsevier Academic Press Amsterdam, The Netherlands, 2005.

[3] K. Aziz and A. Settari, *Petroleum reservoir simulation*, Chapman & Hall, 1979.

[4] G. Burgers, P. van Leeuwen, G. Evensen, and K. N. M. Instituut, *Analysis scheme in the ensemble Kalman filter*, Monthly Weather Review, 126 (1998), pp. 1719–1724.

[5] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, Journal of geophysical research, 99 (1994), p. 10143.

[6] ——, *Sampling strategies and square root analysis schemes for the EnKF*, Ocean Dynamics, 54 (2004), pp. 539–560.

[7] ——, *The ensemble kalman filter for combined state and parameter estimation*, Control Systems Magazine, IEEE, 29 (2009), pp. 83–104.

[8] G. Golub and C. Van Loan, *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)*, (1996).

[9] T. Hamill, J. Whitaker, and C. Snyder, *Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter*, Monthly Weather Review, 129 (2001), pp. 2776–2790.

[10] R. Horn and C. Johnson, *Matrix analysis*, Cambridge university press, 2005.

[11] P. Houtekamer and H. Mitchell, *Data assimilation using an ensemble Kalman filter technique*, Monthly Weather Review, 126 (1998), pp. 796–811.

[12] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, vol. 5, Prentice Hall Upper Saddle River, NJ, 2002.

[13] R. Kalman et al., *A new approach to linear filtering and prediction problems*, Journal of basic Engineering, 82 (1960), pp. 35–45.

[14] A. Kovalenko, T. Mannseth, and G. Nævdal, *Error estimate for the ensemble kalman filter update step*, In:Proceedings of the 12th European Conference on the Mathematics of Oil Recovery (ECMOR XII)., (2010).

[15] ——, *Sampling error distribution for the ensemble kalman filter update step*, submitted to SIAM Journal on Matrix Analysis and Applications, (2011).

[16] J. R. Lien, *Reservoarteknikk, PTEK212*, Institutt fir fysikk og teknologi, Universitetet i Bergen, 2009.

[17] R. Lorentzen, K. Fjelde, J. Frøyen, A. Lage, G. Nævdal, and E. Vefring, *Underbalanced and low-head drilling operations: Real time interpretation of measured data and operational support*, in SPE Annual Technical Conference and Exhibition, 2001.

[18] T. Mannseth, *An analysis of the robustness of some incomplete factorizations*, SIAM Journal on Scientific Computing, 16 (1995), p. 1428.

[19] P. Maybeck, *Stochastic models, estimation, and control*, Academic press, 1979.

[20] H. Mitchell, P. Houtekamer, and G. Pellerin, *Ensemble size, balance, and model-error representation in an ensemble kalman filter*, Monthly weather review, 130 (2002), pp. 2791–2811.

[21] E. Moore, *On the reciprocal of the general algebraic matrix*, Bull. Amer. Math. Soc, 26 (1920), pp. 394–395.

[22] D. Oliver, A. Reynolds, and N. Liu, *Inverse theory for petroleum reservoir characterization and history matching*, Cambridge Univ Pr, 2008.

[23] R. Penrose, *A generalized inverse for matrices*, in Mathematical proceedings of the Cambridge philosophical society, vol. 51, Cambridge Univ Press, 1955, pp. 406–413.

[24] J. Skjervheim, *Continuous updating of a coupled reservoir-seismic model using an ensemble kalman filter technique*, (2007).

[25] A. Tarantola, *Inverse problem theory*, vol. 130, Elsevier Amsterdam etc., 1987.

[26] ——, *Inverse problem theory and methods for model parameter estimation*, Society for Industrial Mathematics, 2005.