Data Assimilation

Andrew M. Stuart
Mathematics Institute
Warwick University
CV4 7AL, UK
a.m.stuart@warwick.ac.uk

Abstract: These lecture notes provide an introduction to the subject of data assimilation, based on an underlying formulation as a Bayesian inverse problem. Various standard methods are then derived and discussed from this standpoint.

Contents

1. Discrete Time: Formulation

Throughout $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^{\grave{}}$, and, for any positive-definite $A \in \mathbb{R}^{\grave{}\times\grave{}}$, $|\cdot|_A = |A^{-\frac{1}{2}}\cdot|$: The symbol $\mathbb{N} = \{0; 1; 2; :::\}$ denotes the natural numbers and $\mathbb{Z}^+ = \{1; 2; :::\}$ the positive integers.

1.1. Set-up

Let $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ and consider the Markov chain $v = \{v_j\}_{j\in\mathbb{N}}$ defined by the random map

$$v_{j+1} = \Psi(v_j) + \ _j; \ j \in \mathbb{N}; \tag{1.1a}$$
$$v_0 = \ u \sim N(m_0; C_0); \tag{1.1b}$$

where $= \{\ _j\}_{j\in\mathbb{N}}$ is an i.i.d. sequence, independent of u, with $_0 \sim N(0; \Sigma)$. Because $(u; \ )$ is a random variable, so too is the solution sequence $\{v_j\}_{j\in\mathbb{N}}$.

In many applications, models such as (1.1a) are supplemented by observations of the system as it evolves. We assume that we are given data $y = \{y_j\}_{j\in\mathbb{Z}^+}$ defined as

$$y_{j+1} = h(v_{j+1}) + \ _{j+1}; \ j \in \mathbb{N}; \tag{1.2}$$

where $h : \mathbb{R}^n \to \mathbb{R}^m$ and $= \{\ _j\}_{j\in\mathbb{Z}^+}$ is an i.i.d. sequence, independent of $(u; \ )$, with $_1 \sim N(0; \Gamma)$. The function h is known as the *observation operator*. The objective of data assimilation is to determine information about the state v of the system, given data y. Mathematically we wish to solve the problem of conditioning the random variable v on the observed data y, or problems closely related to this. The next section describes two key problems of this type: smoothing and Itering.

On occasion we will be interested in the case where the dynamics is deterministic and (1.1a) becomes

$$v_{j+1} = \ \Psi(v_j); \ j \in \mathbb{N}; \tag{1.3a}$$
$$v_0 = \ u \sim N(m_0; C_0): \tag{1.3b}$$

In this case we are interested in the random variable v_0, given the observed data y, since this determines all subsequent values of the state v.

1.2. Smoothing Problem

Here we consdider conditioning the state of the model on a discrete time interval $\mathbb{J}_0 = \{0; :::; J\}$, given data on the discrete time interval $\mathbb{J} = \{1; :::; J\}$. We define $v = \{v_j\}_{j\in\mathbb{J}_0}; y = \{y_j\}_{j\in\mathbb{J}}; \ = \{\ _j\}_{j\in\mathbb{J}_0}$ and $= \{\ _j\}_{j\in\mathbb{J}_0}$: The smoothing problem is to find v from y. We provide a Bayesian formulation of this problem. Recall that we have assumed that $u; \ $ and $\ $ are mutually independent random variables.

Prior The prior on v is specified by (1.1a), together with the independence of u and $\ $. By conditional independence we have

$$\begin{aligned}
\mathbb{P}(v) &= \mathbb{P}(v_J|v_{J-1})\mathbb{P}(v_{J-1}|v_{J-2})\cdots\mathbb{P}(v_1|v_0)\mathbb{P}(v_0) \\
&= \Big(\Pi_{j=0}^{J-1}\mathbb{P}(v_{j+1}|v_j)\Big)\mathbb{P}(v_0) \\
&\propto (\Pi_{j=0}^{J-1}\exp\Big(-\frac{1}{2}\Big|\Sigma^{-\frac{1}{2}}\big(v_j - \Psi(v_{j-1})\big)\Big|^2\Big) \times \exp\Big(-\frac{1}{2}\Big|C_0^{-\frac{1}{2}}(v_0 - m_0)\Big|^2\Big) \\
&= \exp(-I_0(v))
\end{aligned}$$

where

$$I_0(v) := \frac{1}{2}\Big|C_0^{-\frac{1}{2}}(v_0 - m_0)\Big|^2 + \sum_{j=0}^{J-1}\frac{1}{2}\Big|\Sigma^{-\frac{1}{2}}\big(v_{j+1} - \Psi(v_j)\big)\Big|^2:$$

The probability density function (pdf) $\mathbb{P}(v) = \ _0(v)$ proportional to $\exp(-I_0(v))$ determines a prior measure $_0$ on $\mathbb{R}^N$; $N = |\mathbb{J}_0| \times n$.

Likelihood The likelihood of the data $y|v$, is a (Gaussian) probability measure on $\mathbb{R}^M$; $M = |\mathbb{J}| \times m$, with density $\mathbb{P}(y|v)$ proportional to $\exp(-\Phi(v;y))$, where

$$\Phi(v;y) = \sum_{j=0}^{J-1} \frac{1}{2} \left|\Gamma^{-\frac{1}{2}} \left(y_{j+1} - h(v_{j+1})\right)\right|^2. \tag{1.4}$$

To see this note that

$$\begin{aligned}
\mathbb{P}(y|v) &= \Pi_{j=1}^{J} \mathbb{P}(y_j|v_j) \\
&\propto \Pi_{j=1}^{J} \exp\left(-\frac{1}{2}\left|\Gamma^{-\frac{1}{2}} \left(y_{j+1} - h(v_{j+1})\right)\right|^2\right) \\
&= \exp(-\Phi(v;y)).
\end{aligned}$$

Theorem 1.1. *The posterior smoothing distribution on $v|y$ is a probability measure on $\mathbb{R}^N$ with density $\mathbb{P}(v|y) = \ (v)$ proportional to $\exp(-I(v;y))$ where*

$$I(v;y) = I_0(v) + \Phi(v;y) \tag{1.5}$$

Proof. Bayes' Theorem states that

$$\mathbb{P}(v|y) = \frac{\mathbb{P}(y|v)\mathbb{P}(v)}{\mathbb{P}(y)}.$$

Thus, ignoring constants of proportionality which depend only on y,

$$\begin{aligned}
\mathbb{P}(v|y) &\propto \mathbb{P}(y|u)\mathbb{P}(v) \\
&\propto \exp(-\Phi(v;y)\exp(-I_0(v)) \\
&= \exp(-I(v;y)).
\end{aligned}$$

□

In the case where the model dynamics contains no noise, and is given by (2.9a), the posterior distribution for $v_0 \in \mathbb{R}^n$, given y, is measure $'$ with density $'$ proportional to $I'(v;y)$ where

$$I'(v_0;y) := \frac{1}{2}\left|C_0^{-\frac{1}{2}}(v_0 - m_0)\right|^2 + \sum_{j=0}^{J-1} \frac{1}{2}\left|\Gamma^{-\frac{1}{2}} \left(y_{j+1} - h(v_{j+1})\right)\right|^2. \tag{1.6}$$

The posterior distribution on v_j is simply found as the push forward of this measure under j applications of Ψ.

1.3. Filtering Problem

Let $Y_j = \{y_l\}_{l=1}^{j}$ denote the accumulated data up to time j. Filtering is concerned with the sequential updating of $\mathbb{P}(v_j|Y_j)$, the probability density of $v_j|Y_j$. This update is defined by the following two-step procedure which provides a prescription for computing $\mathbb{P}(v_{j+1}|Y_{j+1})$ from $\mathbb{P}(v_j|Y_j)$ via a two step-procedure: **prediction** which computes the mapping $\mathbb{P}(v_j|Y_j) \mapsto \mathbb{P}(v_{j+1}|Y_j)$ and **analysis** which computes $\mathbb{P}(v_{j+1}|Y_j) \mapsto \mathbb{P}(v_{j+1}|Y_j$:

Prediction This is governed by (1.1a). Here we note that

$$\mathbb{P}(v_{j+1}|Y_j) = \int_{\mathbb{R}^n} \mathbb{P}(v_{j+1}|Y_j; v_j)\mathbb{P}(v_j|Y_j)dv_j \tag{1.7a}$$

$$= \int_{\mathbb{R}^n} \mathbb{P}(v_{j+1}|v_j)\mathbb{P}(v_j|Y_j)dv_j \tag{1.7b}$$

Note that, since the forward model equation (1.1a) determines $\mathbb{P}(v_{j+1}|v_j)$, this prediction step provides the map from $\mathbb{P}(v_j|Y_j)$ to $\mathbb{P}(v_{j+1}|Y_j)$:

Analysis This is governed by (1.2). We apply Bayes' Theorem and deduce that

$$\begin{aligned}\mathbb{P}(v_{j+1}|Y_{j+1}) &= \mathbb{P}(v_{j+1}|Y_j; y_{j+1}) \\ &= \frac{\mathbb{P}(y_{j+1}|v_{j+1}; Y_j)\mathbb{P}(v_{j+1}|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)} \\ &= \frac{\mathbb{P}(y_{j+1}|v_{j+1})\mathbb{P}(v_{j+1}|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)}: \end{aligned} \tag{1.8}$$

Since the observation equation (1.2) determines $\mathbb{P}(y_{j+1}|v_{j+1})$, this analysis step provides a map from $\mathbb{P}(v_{j+1}|Y_j)$ to $\mathbb{P}(v_{j+1}|Y_{j+1})$:

Filtering Update Together, then, the prediction and analysis step provide a mapping from $\mathbb{P}(v_j|Y_j)$ to $\mathbb{P}(v_{j+1}|Y_j)$: However, there is, in general, no easily usable closed form expression for $\mathbb{P}(v_j|Y_j)$. However, formulae (1.7), (1.8) form the starting point for numerous algorithms to approximate it.

1.4. Filtering and Smoothing are Related

Theorem 1.2. *Let $\mathbb{P}(v|y)$ denote the smoothing distribution on the discrete time interval $j \in \mathbb{J}_0$, and $\mathbb{P}(v_J|Y_J)$ the Itering distribution at time $j = J$. Then the marginal of the smoothing distribution on v_J is the same as the Itering distribution:*

$$\int \mathbb{P}(v|y)dv_0dv_1:::dv_{J-1} = \mathbb{P}(v_J|Y_J):$$

Proof. Note that $y = Y_J$. Since $v = (v_0; :::; v_{J-1}; v_J)$ the result follows trivially. □

Remark 1.3. *Note that the marginal of the smoothing distribution on say v_j; $j < J$ is* not *equal to the Iter $\mathbb{P}(v_j|Y_j)$. This is because the smoother induces a distribution on v_j which is in uenced by the entire data stream $Y_J = y = \{y_l\}_{l\in\mathbb{J}}$; in contrast the Iter at j involves only the data $Y_j = \{y_l\}_{l\in\{1;:::;j\}}$.*

1.5. Well-Posedness

Let and $'$ denote two probability measures which have strictly positive Lebesgue densities and $'$ on $\mathbb{R}^`$. Throughout this subsection, all integrals are over $\mathbb{R}^`$: Define the Hellinger distance between and $'$ as

$$\begin{aligned} d_{\text{Hell}}(\ ; \ ') &= \sqrt{\frac{1}{2}\int \left(1 - \sqrt{\frac{'(x)}{(x)}}\right)^2 (x)dx} \\ &= \left(\frac{1}{2}\mathbb{E}\ \left(1 - \sqrt{\frac{'(x)}{(x)}}\right)^2\right)^{\frac{1}{2}}: \end{aligned} \tag{1.9}$$

We also define the TV distance

$$\begin{aligned} d_{\text{TV}}(\ ; \ ') &= \frac{1}{2}\int \left|1 - \frac{'(x)}{(x)}\right| \ (x)dx \\ &= \frac{1}{2}\mathbb{E}\ \left|1 - \frac{'(x)}{(x)}\right|: \end{aligned} \tag{1.10}$$

Notice that, for $f : \mathbb{R}^{`} \to \mathbb{R}^{p}$,

$$\begin{aligned}
|\mathbb{E}\ f(x) - \mathbb{E}\ ' f(x)| &\le \int |f(x)||\ (x) -\ '(x)|dx \\
&= \int \sqrt{2}|f(x)||\sqrt{\ (x)} + \sqrt{\ '(x)}| \cdot \frac{1}{\sqrt{2}}|\sqrt{\ (x)} - \sqrt{\ '(x)}|dx \\
&\le \left(\int 2|f(x)|^2|\sqrt{\ (x)} + \sqrt{\ '(x)}|^2 dx\right)^{\frac{1}{2}} \left(\frac{1}{2}\int |\sqrt{\ (x)} - \sqrt{\ '(x)}|^2 dx\right)^{\frac{1}{2}} \\
&\le \left(\int 4|f(x)|^2(\ (x) +\ '(x))dx\right)^{\frac{1}{2}} \left(\frac{1}{2}\int \left(1 - \frac{\sqrt{\ '(x)}}{\sqrt{\ (x)}}\right)^2\ (x)dx\right)^{\frac{1}{2}} \\
&= 2(\mathbb{E}\ |f(x)|^2 + \mathbb{E}\ '|f(x)|^2)^{\frac{1}{2}} d_{\mathrm{Hell}}(\ ;\ ');
\end{aligned}$$

Thus the Hellinger metric provides a direct way of estimating changes in expectation of square integrable functions:

$$|\mathbb{E}\ f(x) - \mathbb{E}\ ' f(x)| \le 2(\mathbb{E}\ |f(x)|^2 + \mathbb{E}\ '|f(x)|^2)^{\frac{1}{2}} d_{\mathrm{Hell}}(\ ;\ '); \tag{1.11}$$

We let $_0$ denote the prior measure on v for the smoothing problem, and and $'$ the posterior measures resulting from two different instances of the data, y and y' respectively.

Theorem 1.4. *Assume that* $\mathbb{E}\ ^0(\sum_{j=0}^{J-1} 1 + |h(v_{j+1})|^2)^{\frac{1}{2}} < \infty$. *Then, for* $|y|; |y'| \le r$ *there exists* $c = c(r)$ *such that*

$$d_{\mathrm{Hell}}(\ ;\ ') \le c|y - y'|;$$

Proof. Let $_0$; and $'$ denote the Lebesgue densities on $_0$; and $'$ respectively. Then

$$\begin{aligned}
_0(v) &= \frac{1}{Z_0}\exp(-I_0(v)); \\
(v) &= \frac{1}{Z}\exp(-I_0(v) - \Phi(v;y)); \\
'(v) &= \frac{1}{Z'}\exp(-I_0(v) - \Phi(v;y'));
\end{aligned}$$

where

$$\begin{aligned}
Z_0 &= \int \exp(-I_0(v))dv; \\
Z &= \int \exp(-I_0(v) - \Phi(v;y))dv; \\
Z' &= \int \exp(-I_0(v) - \Phi(v;y'))dv;
\end{aligned}$$

Thus we have

$$\begin{aligned}
d_{\mathrm{Hell}}(\ ;\ ')^2 &= \frac{1}{2}\int |\sqrt{\ (x)} - \sqrt{\ '(x)}|^2 dx \\
&= \frac{1}{2}\int Z_0 \left|\frac{1}{\sqrt{Z}}e^{-\frac{1}{2}\Phi(v;y)} - \frac{1}{\sqrt{Z'}}e^{-\frac{1}{2}\Phi(v;y')}\right|^2\ _0(dv) \\
&\le I_1 + I_2;
\end{aligned}$$

where

$$I_1 = Z_0 \int \frac{1}{Z}|e^{-\frac{1}{2}\Phi(v;y)} - e^{-\frac{1}{2}\Phi(v;y')}|^2\ _0(dv)$$

and

$$I_2 = Z_0\left|\frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z'}}\right|^2 \int e^{-\Phi(v;y')}\ {}_0(dv)$$
$$= Z'\left|\frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z'}}\right|^2:$$

Since $\Phi(v;y) \geq 0$ and $\Phi(v;y') \geq 0$ we have

$$|Z - Z'| \leq Z_0 \int |e^{-\Phi(v;y)} - e^{-\Phi(v;y')}|\ {}_0(dv)$$
$$\leq Z_0 \int |\Phi(v;y) - \Phi(v;y')|\ {}_0(dv):$$

By definition

$$|\Phi(v;y) - \Phi(v;y')| \leq \frac{1}{2}\sum_{j=0}^{J-1} |y_{j+1} - y'_{j+1}|_\Gamma |y_{j+1} + y'_{j+1} - 2h(v_{j+1})|_\Gamma$$
$$\leq \frac{1}{2}\left(\sum_{j=0}^{J-1} |y_{j+1} - y'_{j+1}|_\Gamma^2\right)^{\frac{1}{2}} \left(\sum_{j=0}^{J-1} |y_{j+1} + y'_{j+1} - 2h(v_{j+1})|_\Gamma^2\right)^{\frac{1}{2}}$$
$$\leq c(r)|y - y'| \left(\sum_{j=0}^{J-1} 1 + |h(v_{j+1})|^2\right)^{\frac{1}{2}}:$$

Thus

$$|Z - Z'| \leq c(r)|y - y'|:$$

Hence, since $Z; Z' > 0$, $I_2 \leq c(r)|y - y'|$. A similar argument shows that $I_1 \leq c(r)|y - y'|$ and the proof is complete. □

Corollary 1.5. *Let $f : \mathbb{R}^N \to \mathbb{R}^p$ be such that $\mathbb{E}\ {}_0|f(v)|^2 < \infty$ and assume further that $\mathbb{E}\ {}_0\left(\sum_{j=0}^{J-1} 1 + |h(v_{j+1})|^2\right)^{\frac{1}{2}} < \infty$. Then*

$$|\mathbb{E}\ \ f(x) - \mathbb{E}\ {}'f(x)| \leq c|y - y'|:$$

Proof. First note that, since $\Phi(v;y) \geq 0$, $\mathbb{E}\ \ |f(x)|^2 \leq c\mathbb{E}\ {}_0|f(x)|^2$, and similarly for ${}'$. The result follows from (1.11) and Theorem 1.4. □

Corollary 1.6. *Let $g : \mathbb{R}^n \to \mathbb{R}^p$ be such that $\mathbb{E}\ {}_0|g(v_J)|^2 < \infty$ and assume further that $\mathbb{E}\ {}_0\left(\sum_{j=0}^{J-1} 1 + |h(v_{j+1})|^2\right)^{\frac{1}{2}} < \infty$: If ${}_J$ and ${}'_J$ denote the Itering distributions at time J corresponding to data $Y_J; Y'_J$ respectively, then*

$$|\mathbb{E}\ {}_J g(x) - \mathbb{E}\ {}'_J g(x)| \leq c|Y_J - Y'_J|:$$

Proof. Since, by Theorem 1.2, ${}_J$ is the marginal of the smoother on the v_J coordinate, the result follows from Corollary 1.5 by choosing $f(v) = g(v_J)$. □

2. Discrete Time: Smoothing Algorithms

The formulation of the data assimilation problem described in the previous chapter is probabilistic, and its computational resolution requires the probing of a probability distribution. This can be computationally infeasible for very large problems but, where possible, provides an important benchmark solution. In section 2.1 we provide some background concerning Monte Carlo Markov Chain (MCMC) methods for this problem. Then, in section 2.2, we describe some optimization algorithms which relate to maximising the posterior probability.

2.1. MCMC Methods

The posterior distribution of interest is the measure with density given in the proof of Theorem 1.4. We will describe the Metropolis-Hastings methodology for creating a Markov chain which is invariant for a general meaasure on $\mathbb{R}^{`}$ with pdf . We then describe two specific instances of this method.

We are given a probability density function $: \mathbb{R}^{`} \to \mathbb{R}^{+}$, with $\int (x)dx = 1$. Now consider a Markov transition kernel $q : \mathbb{R}^{`} \times \mathbb{R}^{`} \to \mathbb{R}^{+}$ with the property that $\int q(x;y)dy = 1$ for every $x \in \mathbb{R}^{`}$. We create a Markov chain $\{x^{(k)}\}_{k\in\mathbb{N}}$ which is invariant for as follows. To this end we define

$$a(x;y) = 1 \wedge \frac{(y)q(y;x)}{(x)q(x;y)} \tag{2.1}$$

The algorithm is:

1. Set $k = 0$ and choose $x^{(0)} \in \mathbb{R}^{`}$.
2. $k \to k + 1$.
3. Draw $y^{(k)} \sim q(x^{(k-1)};dy)$.
4. Set $x^{(k)} = y^{(k)}$ with probability $a(x^{(k-1)};y^{(k)})$, $x^{(k)} = x^{(k-1)}$ otherwise.
5. Go to step 2.

The draw of $y^{(k)}$ in step 3. is, given $x^{(k-1)}$, independent of all previous randomness. The probability of accepting $x^{(k)} = y^{(k)}$ in step 4 is independent of the randomness in step 4, and of all preceding randomness. Thus the whole procedure gives a Markov chain. If $\mathsf{u} = \{\mathsf{u}^{(j)}\}_{j\in\mathbb{N}}$ is an i.i.d. sequence of $U[0;1]$ random variables then we may write the algorithm as follows:

$$\begin{aligned} y^{(k)} &\sim q(x^{(k-1)};dy) \\ x^{(k)} &= y^{(k)}\mathbb{I}\big(\mathsf{u}^{(j)} \le a(x^{(k-1)};y^{(k)})\big) + x^{(k-1)}\mathbb{I}\big(\mathsf{u}^{(j)} > a(x^{(k-1)};y^{(k)})\big): \end{aligned}$$

We let $p : \mathbb{R}^{`} \times \mathbb{R}^{`} \to \mathbb{R}^{+}$ denote the transition kernel of the resulting Markov chain, and we let p^k denote the transition kernel over k steps. Thus $p^k(x;A) = \mathbb{P}(x^{(k)} \in A | x^{(0)} = x)$ and $p^k(x;\cdot)$ is a probability measure on $\mathbb{R}^{`}$: The resulting algorith is known as a **Metropolis-Hastings** MCMC algorithm.

Remark 2.1. *The following two observations are central to Metropolis-Hastings MCMC methods.*

- *The construction of Metropolis-Hastings MCMC methods is designed to ensure* detailed balance*:*

$$(x)p(x;y) = (y)p(y;x): \tag{2.2}$$

  Once this condition is obtained it follows trivially that measure with density is invariant since, integrating over x, we obtain

$$\begin{aligned} \int (x)p(x;y)dy &= \int (y)p(y;x)dx \\ &= (y)\int p(y;x)dx \\ &= (y): \end{aligned}$$

- *In order to implement Metropolis-Hastings MCMC methods it is not necessary to know the normalisation constants for $(\cdot)$ and $q(x;\cdot)$ since only their ratios appear in a.*

The **Metropolis-Hastings** algorithm defined above satisfies the following:

Theorem 2.2. *If $x^{(0)} \sim$ with Lebesque density , then $x^{(k)} \sim$ for all $k \in \mathbb{Z}^{+}$. Thus, if the Markov chain is ergodic, then for any bounded continuous $' : \mathbb{R}^{`} \to \mathbb{R}$,*

$$\frac{1}{K}\sum_{k=1}^{K} '(x^{(k)}) \xrightarrow{a.s.} \mathbb{E}\ '(x)$$

for a.e. initial condition $x^{(0)}$. In particular, if there is probability measure on $\mathbb{R}^{\grave{}}$ and > 0 such that, for all $x \in \mathbb{R}^{\grave{}}$, $p(x;A) \geq (A)$ then, for all $x \in \mathbb{R}^{\grave{}}$,

$$d_{\mathrm{TV}}\big(p^k(x;\cdot);\ \big) \leq 2(1- \ )^n: \tag{2.3}$$

Furthermore, there is $C > 0$ such that

$$\frac{1}{K}\sum_{k=1}^{K} '(x^{(k)}) = \mathbb{E}\ '(x) + C\ _K K^{-\frac{1}{2}} \tag{2.4}$$

where $_K$ converges weakly to $N(0;1)$ as $K \to \infty$:

We now apply the Metropolis-Hastings methodology to the data assimilation smoothing problem. In addition to the measures $_0$ and , with densities $_0$ and , it is helpful in what follows to introduce the measure $_0$ with density $_0$ found from $_0$ and $_0$ in the case where $\Psi \equiv 0$. Thus

$$_0(v) \propto \exp\Big(-\frac{1}{2}\big|C_0^{-\frac{1}{2}}(v_0 - m_0)\big|^2 - \sum_{j=0}^{J-1}\frac{1}{2}|\Sigma^{-\frac{1}{2}}v_{j+1}|^2\Big) \tag{2.5}$$

and is a Gaussian measure, independent in each component v_j for $j = 0;\cdots;J$: We denote the mean by m and covariance by C. Thus $_0 = N(m;C)$:

It is also useful to rewrite $_0$ as follows:

$$_0(v) \propto \exp(-I_0(v) + G(v));$$

where

$$G(v) = \sum_{j=0}^{J-1}\Big(\frac{1}{2}|\Sigma^{-\frac{1}{2}}\Psi(v_j)|^2 - \langle\Sigma^{-\frac{1}{2}}v_{j+1};\Sigma^{-\frac{1}{2}}\Psi(v_j)\rangle\Big):$$

Important in what follows are the observations that

$$\frac{(v)}{_0(v)} \propto \exp(-\Phi(v;y)); \tag{2.6}$$

$$\frac{(v)}{_0(v)} \propto \exp(-\Phi(v;y) - G(v)): \tag{2.7}$$

We now construct two Markov chains $\{v^{(k)}\}_{k\in\mathbb{N}}$ which are invariant with respect to . These will both be Metropolis-Hastings methods and hence we need only specify the transition kernel $q(v;w)$, and identify the resulting acceptance probability $a(v;w)$. The sequence $\{w^{(k)}\}_{k\in\mathbb{Z}^+}$ will denote the proposals.

Independence Sampler Here we choose the proposal $w^{(k)}$, independently of the current state $v^{(k-1)}$, from the prior $_0$. Thus $q(v;w) \propto\ _0(w)$ and

$$\begin{aligned} a(v;w) &= 1 \wedge \frac{(w)q(w;v)}{(v)q(v;w)} \\ &= 1 \wedge \frac{(w)=\ _0(w)}{(v)=\ _0(v)} \\ &= 1 \wedge \exp(\Phi(v;y) - \Phi(w;y)): \end{aligned}$$

In particular, the resulting MCMC method always accepts moves which decrease the model-data misfit functional $\Phi(:;y)$ given in (1.4). For simplicity assume that the observation operator h is bounded so that,

for all $u \in \mathbb{R}^m$, $|h(u)| \le h_{\max}$: Then

$$\begin{aligned}\Phi(u;y) &\le \sum_{j=0}^{J-1}\left(|\Gamma^{-\frac{1}{2}}y_{j+1}|^2 + |\Gamma^{-\frac{1}{2}}h(v_{j+1})|2\right)\\ &\le |\Gamma^{-\frac{1}{2}}|^2\Big(\sum_{j=0}^{J-1}|y_{j+1}|^2 + Jh_{\max}^2\Big)\\ &\le |\Gamma^{-\frac{1}{2}}|^2\Big(|Y_J|^2 + Jh_{\max}^2\Big)\\ &:= \Phi_{\max}:\end{aligned}$$

Since $\Phi \ge 0$ this shows that every proposed step is accepted with probability exceeding $e^{-\Phi_{\max}}$ and hence that

$$p(x;A) \ge e^{-\Phi_{\max}} \ {}_0(A):$$

Thus Theorem 2.2 applies and, in particular, eqrefeq:tv and (2.4), with $= e^{-\Phi_{\max}}$: The independence sampler relies on draws from the prior matching the data well. Where the data set is large ($J \gg 1$) or the noise covariance small ($|\Gamma| \ll 1$) this will happen infrequently and the MCMC method will reject frequently and be inefficient. This can be see in Theorem 2.2 in the case where $= e^{-\Phi_{\max}}$ and $\Phi_{\max} \gg 1$. To overcome such problems local proposals, which do not move far from the current state, are useful. These are exemplified in the following.

The pCN Method Recall the Gaussian measure ${}_0 = N(m;C)$ defined via its pdf in (2.5). The pcN method is a variant of random walk type methods, based on the following proposal

$$w^{(k)} = m + (1 - {}^2)^{\frac{1}{2}}\left(v^{(k-1)} - m\right) + {}^{(k-1)};$$

$${}^{(k-1)} \sim N(0;C):$$

Here ${}^{(k-1)}$ is assumed to be independent of $v^{(k-1)}$. This proposal preserves ${}_0$ and would be accepted all the time in the absence of data, and if $\Psi \equiv 0$. To see this preservation of ${}_0$ notice that if $v^{(k-1)} \sim {}_0$ then $\mathbb{E}w^{(k)} = m$ and

$$\begin{aligned}\mathbb{E}\left(w^{(k)} - m\right) \otimes \left(w^{(k)} - m\right) &= (1 - {}^2)\mathbb{E}\left(v^{(k-1)} -\right) \otimes \left(v^{(k-1)} - m\right) + {}^2\mathbb{E}\ {}^{(k-1)} \otimes {}^{(k-1)}\\ &= (1 - {}^2)C + {}^2C\\ &= C;\end{aligned}$$

where C is the covariance under ${}_0$. This shows that the proposal preserves ${}_0$ and, in fact that

$$\frac{{}_0(w)q(w;v)}{{}_0(v)q(v;w)} = 1; \tag{2.8}$$

reflecting the fact that the Markov chain

$$v^{(k)} = m + (1 - {}^2)^{\frac{1}{2}}\left(v^{(k-1)} - m\right) + {}^{(k-1)}$$

is in fact reversible with respect to ${}_0$.
By use of (2.8) and (2.7) we deduce that the acceptance probability for this method is

$$\begin{aligned}a(v;w) &= 1 \wedge \frac{(w)q(w;v)}{(v)q(v;w)}\\ &= 1 \wedge \frac{(w)= {}_0(w)}{(v)= {}_0(v)}\\ &= 1 \wedge \exp(\Phi(v;y) - \Phi(w;y) + G(v) - G(w)):\end{aligned}$$

By choosing small, so that $w^{(k)}$ is close to $v^{(k-1)}$, we can make $a(v^{(k-1)};w^{(k)})$ reasonably large and obtain a useable algorithm.

2.2. Variational Methods

Sampling the posterior using MCMC methods can be prohibitively expensive. Furthermore, if the probability is peaked at one, or a small number of places, then simply locating these peaks may be sufficient in an applied context. This is the basis for variational methods which seek to maximize the posterior probability, thereby locating such peaks.

The methods lead to problems in the calculus of variations, and are hence termed **variational methods**. In the atmospheric sciences they are called **4DVAR** since they incorporate data over three spatial dimensions and one temporal dimension, in order to estimate the state. In Bayesian statistics the methods are called **MAP estimators**: maximum *a posteriori* estimators.

Theorem 2.3. *The density associated with the posterior probability , is maximized where $I(v;y)$ given in (1.5) is minimized. Furthermore, if $B(z)$ denotes a ball in $\mathbb{R}^N$ of radius , centred at z, then if $\Psi; h$ are continuous,*

$$\lim_{\to 0} \frac{\mathbb{P}(B(z_1))}{\mathbb{P}(B(z_2))} = \exp(I(z_2;y) - I(z_1;y)):$$

Proof. Since

$$\begin{aligned}(dv) &= \frac{1}{Z}\exp(-I(v;y))dv \\ &= (v)dv\end{aligned}$$

the first result is clear. For the second note that

$$\begin{aligned}\mathbb{P}(B(z)) &= \frac{1}{Z}\int_{|v-z|<} \exp\big(-I(v;y)\big)dv \\ &= \frac{1}{Z}\int_{|v-z|<} \Big(\exp(-I(z;y)) + e(;v)\Big)dv\end{aligned}$$

where $e(;v) \to 0$ as $\to 0$, uniformly for $y \in B(z)$. This is because $I(\cdot;y)$ inherits continuity from $\Psi(\cdot)$ and $h(\cdot)$. The result follows. □

Remark 2.4. *The second statement in Theorem 2.3 may appear a little abstract. However, unlike the rst statement, it can be generalised to in nite dimensions, as is required in continuous time. We state the result as in the second statement for precisely this reason.*

In applications to metereology the variational method just described is known as **weak constraint 4DVAR**. This generalizes the standard **4DVAR** method which may be derived in the limit $\Sigma \to 0$ so that the prior on the model dynamics (1.1a) is deterministic, but with a random initial condition, as in (2.9a). In this case the appropriate minimization if of $I'(v_0;y)$ given by (1.6). This has the advantage of being a lower dimensional minimization problem than weak constraint 4DVAR; however it is often a harder minimization problem, especially when the dynamics is chaotic.

$$v_{j+1} = \Psi(v_j);\ j \in \mathbb{N}; \tag{2.9a}$$

$$v_0 = u \sim N(m_0; C_0): \tag{2.9b}$$

The resulting optimization problem involves optimizing only over v_0.

3. Discrete Time: Filtering Algorithms

3.1. The Kalman Filter

This algorithm provides a sequential method for updating the filtering distribution $\mathbb{P}(v_j|Y_j)$ from time j to time $j+1$, when Ψ and h are linear maps. In this case the filtering distribution is Gaussian and it can be characterized entirely through the mean and covariance. We let

$$\Psi(v) = Mv;\ h(v) = Hv \tag{3.1}$$

for matrices $M \in \mathbb{R}^{n\times n}; H \in \mathbb{R}^{m\times n}$. We assume that $m \leq n$ and $\mathrm{Rank}(H) = m$. We let $(m_j; C_j)$ denote the mean and covariance of $\mathbb{P}(v_j|Y_j)$, noting that this random variable is Gaussian for each j since all maps are linear and all noise is Gaussian additive. We let $(\widehat{m}_{j+1}; \widehat{C}_{j+1})$ denote the mean and covariance of $\mathbb{P}(v_{j+1}|Y_j)$, noting that this too is a Gaussian random variable. We now derive the map $(m_j; C_j) \mapsto (m_{j+1}; C_{j+1})$, using the intermediate variables $(\widehat{m}_{j+1}; \widehat{C}_{j+1})$.

Theorem 3.1. *Assume that $C_0; \Sigma > 0$. Then $C_j > 0$ for all $j \in \mathbb{Z}^+$ and*

$$C_{j+1}^{-1} = (MC_jM^T + \Sigma)^{-1} + H^T\Gamma^{-1}H$$
$$C_{j+1}^{-1}m_{j+1} = (MC_jM^T + \Sigma)^{-1}Mm_j + H^T\Gamma^{-1}y_{j+1}:$$

Proof. The prediction step is determined by (1.1a) in the case $\Psi(\cdot) = M\cdot$:

$$v_{j+1} = Mv_j + \ _j; \quad \ _j \sim N(0; \Sigma):$$

From this it is clear that

$$\mathbb{E}(v_{j+1}|Y_j) = \mathbb{E}(Mv_j|Y_j) + \mathbb{E}(\ _j|Y_j):$$

Since $\ _j$ is independent of Y_j we have

$$\widehat{m}_{j+1} = Mm_j: \tag{3.2}$$

Similarly

$$\begin{aligned}\mathbb{E}\big((v_{j+1} - \widehat{m}_{j+1}) \otimes (v_{j+1} - \widehat{m}_{j+1})|Y_j\big) &= \mathbb{E}\big(M(v_j - m_j) \otimes M(v_j - m_j)|Y_j\big) + \mathbb{E}\big(\ _j \otimes \ _j|Y_j\big) \\ &+ \mathbb{E}\big(M(v_j - m_j) \otimes \ _j|Y_j\big) + \mathbb{E}\big(\ _j \otimes M(v_j - m_j)|Y_j\big):\end{aligned}$$

Again, since $\ _j$ is independent of Y_j and of v_j, we have

$$\begin{aligned}\widehat{C}_{j+1} &= M\mathbb{E}(v_j \otimes v_j|Y_j)M^T + \Sigma \\ &= MC_jM^T + \Sigma:\end{aligned} \tag{3.3}$$

Now we consider the analysis step. By (1.8), which is just Bayes' rule, and using Gaussianity, we have

$$\exp\Big(-\frac{1}{2}\big|C_{j+1}^{-\frac{1}{2}}(v - m_{j+1})\big|^2\Big) \propto \exp\Big(-\frac{1}{2}\big|\Gamma^{-\frac{1}{2}}(y_{j+1} - Hv)\big|^2 - \frac{1}{2}\big|\widehat{C}_{j+1}^{-\frac{1}{2}}(v - \widehat{m}_{j+1})\big|^2\Big):$$

Equating quadratic terms in v, gives

$$C_{j+1}^{-1} = \widehat{C}_{j+1}^{-1} + H^T\Gamma^{-1}H \tag{3.4}$$

and equating linear terms in v gives

$$C_{j+1}^{-1}m_{j+1} = \widehat{C}_{j+1}^{-1}\widehat{m}_{j+1} + H^T\Gamma^{-1}y_{j+1} \tag{3.5}$$

Substituting the expressions (3.2) and (3.3) for $(\widehat{m}_{j+1}; \widehat{C}_{j+1})$ gives the desired result. It remains to verify that $C_j > 0$ for all $j \in \mathbb{Z}^+$ so that the formal calculations above make sense. To this end we notice that, since Σ and $C_j > 0$ (inductive hypothesis, true for $j = 0$), we have $MC_jM^T + \Sigma > 0$ and hence $(MC_jM^T + \Sigma)^{-1} > 0$. Thus $C_{j+1}^{-1} > 0$ and hence $C_{j+1} > 0$. □

Corollary 3.2. *The formulae for the Kalman Iter from Theorem 3.1 may be rewritten as follows:*

$$\begin{aligned}d_{j+1} &= y_{j+1} - H\widehat{m}_{j+1} \\ S_{j+1} &= H\widehat{C}_{j+1}H^T + \Gamma \\ K_{j+1} &= \widehat{C}_{j+1}H^TS_{j+1}^{-1} \\ m_{j+1} &= \widehat{m}_{j+1} + K_{j+1}d_{j+1} \\ C_{j+1} &= (I - K_{j+1}H)\widehat{C}_{j+1};\end{aligned}$$

with $(\widehat{m}_{j+1}; \widehat{C}_{j+1})$ given in (3.2), (3.3).

Proof. By (3.4) we have

$$C_{j+1}^{-1} = \widehat{C}_{j+1}^{-1} + H^T \Gamma^{-1} H$$

and application of Lemma 3.4 below gives

$$\begin{aligned} C_{j+1} &= \widehat{C}_{j+1} - \widehat{C}_{j+1} H^T (\Gamma + H \widehat{C}_{j+1} H^T)^{-1} H \widehat{C}_{j+1} \\ &= \left(I - \widehat{C}_{j+1} H^T (\Gamma + H \widehat{C}_{j+1} H^T)^{-1} H \right) \widehat{C}_{j+1} \\ &= (I - \widehat{C}_{j+1} H^T S_{j+1}^{-1} H) \widehat{C}_{j+1} \\ &= (I - K_{j+1} H) \widehat{C}_{j+1} \end{aligned}$$

as required. Then the identity (3.5) gives

$$\begin{aligned} m_{j+1} &= C_{j+1} \widehat{C}_{j+1}^{-1} \widehat{m}_{j+1} + C_{j+1} H^T \Gamma^{-1} y_{j+1} \\ &= (I - K_{j+1} H) \widehat{m}_{j+1} + C_{j+1} H^T \Gamma^{-1} y_{j+1} \end{aligned} \tag{3.6}$$

Now note that, again by (3.4),

$$C_{j+1} (\widehat{C}_{j+1}^{-1} + H^T \Gamma^{-1} H) = I$$

so that

$$\begin{aligned} C_{j+1} H^T \Gamma^{-1} H &= I - C_{j+1} \widehat{C}_{j+1}^{-1} \\ &= I - (I - K_{j+1} H) \\ &= K_{j+1} H. \end{aligned}$$

Since H has rank m we deduce that

$$C_{j+1} H^T \Gamma^{-1} = K_{j+1}.$$

Hence (3.6) gives

$$m_{j+1} = (I - K_{j+1} H) \widehat{m}_{j+1} + K_{j+1} y_{j+1} = \widehat{m}_{j+1} + K_{j+1} d_{j+1}$$

as required. □

Remark 3.3. *The key difference between the Kalman update formulae in Theorem 3.1 and in Corollary 3.2 is that, in the former matrix inversion takes place in the state space, with dimension n, whilst in the latter matrix inversion takes place in the data space, with dimension m. In many applications $m \ll n$, as the observed subspace dimension is much less than the state space dimension, and thus the formulation in Corollary 3.2 is frequently employed in practice.*

Lemma 3.4. Woodbury Matrix Identity *Let $A \in \mathbb{R}^{p \times p}; U \in \mathbb{R}^{p \times q}; C \in \mathbb{R}^{q \times q}$ and $V \in \mathbb{R}^{q \times p}$. If A and C are invertible then $A + UCV$ is invertible and*

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U \left(C^{-1} + V A^{-1} U \right)^{-1} V A^{-1}.$$

3.2. Non-Gaussian Filters

The update equation for the Kalman filter mean, (3.5), can be derived by minimizing the following model/data compromise functional, derived from (3.3),

$$J(v) := \frac{1}{2} |\Gamma^{-\frac{1}{2}} (y_{j+1} - Hv)|^2 + \frac{1}{2} |\widehat{C}_{j+1}^{-\frac{1}{2}} (v - \widehat{m}_{j+1})|^2. \tag{3.7}$$

Whilst the Kalman filter itself is restricted to linear, Gaussian problems, the formulation via minimization generalizes to nonlinear problems. Noting that $\widehat{m}_{j+1} = M m_j$, and that $\Psi(\cdot) = M \cdot$, $h(\cdot) = H \cdot$, we see that a natural generalization of (3.7) to the nonlinear case is to minimize

$$J(v) := \frac{1}{2} |\Gamma^{-\frac{1}{2}} (y_{j+1} - h(v))|^2 + \frac{1}{2} |\widehat{C}_{j+1}^{-\frac{1}{2}} (v - \Psi(m_j))|^2$$

and then to set

$$m_{j+1} = \arg\min_v J(v).$$

For simplicity we consider the case where observations are linear and $h(v) = Hv$ leading to the update algorithm $\widehat{m}_j \mapsto \widehat{m}_{j+1}$ defined by

$$\begin{aligned} J(v) &= \frac{1}{2}|\Gamma^{-\frac{1}{2}}(y_{j+1} - Hv)|^2 + \frac{1}{2}|\widehat{C}_{j+1}^{-\frac{1}{2}}(v - \Psi(\widehat{m}_j))|^2 \\ m_{j+1} &= \arg\min_v J(v). \end{aligned} \tag{3.8}$$

By the arguments used in Corollary 3.2 we deduce the following update formulae:

$$m_{j+1} = (I - K_{j+1}H)\Psi(m_j) + K_{j+1}y_{j+1} \tag{3.9a}$$

$$K_{j+1} = \widehat{C}_{j+1}H^T S_{j+1}^{-1} \tag{3.9b}$$

$$S_{j+1} = H\widehat{C}_{j+1}H^T + \Gamma \tag{3.9c}$$

The next three sections each correspond to algorithms derived in this way, namely by minimizing $J(v)$, but corresponding to different choices of the model covariance $\widehat{C}_{j+1}$.

3.3. 3DVAR

This algorithm is derived from (3.9) by simply fixing the model covariance $\widehat{C}_{j+1} \equiv \widehat{C}$ for all j. Thus we obtain

$$m_{j+1} = (I - KH)\Psi(m_j) + Ky_{j+1} \tag{3.10a}$$

$$K = \widehat{C}H^T S^{-1}; \quad S = H\widehat{C}H^T + \Gamma \tag{3.10b}$$

It is natural to ask when this filter will recover the true signal. To this end we assume that

$$y_{j+1} = Hv_{j+1}^\dagger + \ _j \tag{3.11}$$

where the true signal $\{v_j^\dagger\}_{j\in\mathbb{N}}$ satisfies

$$v_{j+1}^\dagger = \Psi(v_j^\dagger); \ j \in \mathbb{N} \tag{3.12a}$$

$$v_0^\dagger = u \tag{3.12b}$$

and, for simplicity, we assume that

$$\sup_{j\in\mathbb{N}} |\ _j| = \ :$$

We have the following result.

Theorem 3.5. *Assume that $(I - KH)\Psi : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz with constant $a < 1$ in some norm $\|\cdot\|$, then*

$$\limsup_{j\to\infty} \|m_j - v_j^\dagger\| \le c \ :$$

Proof. We may write (3.10), (3.12), using (3.11), as

$$\begin{aligned} m_{j+1} &= (I - KH)\Psi(m_j) + KH\Psi(v_j^\dagger) + K\ _j \\ v_{j+1}^\dagger &= (I - KH)\Psi(v_j^\dagger) + KH\Psi(v_j^\dagger). \end{aligned}$$

Subtracting, and letting $e_j = \widehat{m}_j - v_j^\dagger$, gives

$$\begin{aligned} \|e_{j+1}\| &\le \|(I - KH)\Psi(m_j) + (I - KH)\Psi(v_j^\dagger)\| + \|K\ _j\| \\ &\le a\|e_j\| + c \end{aligned}$$

Applying Gronwall gives the desired result. □

Example. *Assume that $H = I$, so that the whole system is observed, that $\Gamma = \ ^2 I$ and $\widehat{C} = \ ^2 I$. Then, for $^2 = \frac{^2}{^2}$*

$$S = (\ ^2 + \ ^2) I; \quad K = \frac{^2}{(\ ^2 + \ ^2)} I$$

and

$$(I - KH) = \frac{^2}{(\ ^2 + \ ^2)} I = \frac{^2}{(1 + \ ^2)} I:$$

Thus, if $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz with constant $\ > 0$ in the Euclidean norm, $|\cdot|$, then $(I - KH)\Psi$ is globally Lipschitz with constant $a < 1$, if $\ $ is chosen so that $\frac{^2}{1+\ ^2} < 1$. Thus, by choosing $\ $ su ciently small the $\ $lter can be made to contract. This corresponds to trusting the data su ciently in comparison to the model.

3.4. Extended Kalman Filter

The idea of the extended Kalman filter (ExKF) is to propagate covariances according to the linearization of (1.1a), and propagate the mean, using (1.1a). Thus we obtain, from modification of Corollary 3.2 and (3.2), (3.3)

$$\text{Prediction} \begin{cases} \widehat{m}_{j+1} &= \Psi(m_j) \\ \widehat{C}_{j+1} &= D\Psi(m_j) C_j D\Psi(m_j)^T + \Sigma \end{cases}$$

$$\text{Analysis} \begin{cases} S_{j+1} &= H\widehat{C}_{j+1} H^T + \Gamma \\ K_{j+1} &= \widehat{C}_{j+1} H^T S_{j+1}^{-1} \\ m_{j+1} &= (I - K_{j+1} H)\widehat{m}_{j+1} + K_{j+1} y_{j+1} \\ C_{j+1} &= (I - K_{j+1} H)\widehat{C}_{j+1} \end{cases}$$

3.5. Ensemble Kalman Filter

The idea of the ensemble Kalman filter (EnKF) is to propagate covariances and mean by maintaining an ensemble of particles, and using this ensemble to estimate the covariance and mean. Then EnKF is executed in a variety of ways and we describe one of these, the perturbed observation EnKF.

$$\text{Prediction} \begin{cases} \widehat{v}_{j+1}^{(k)} &= \Psi(v_j^{(k)});\ k = 1;:::;K \\ \widehat{m}_{j+1} &= \frac{1}{K}\sum_{k=1}^{K} \widehat{v}_{j+1}^{(k)}; \\ \widehat{C}_{j+1} &= \frac{1}{K}\sum_{k=1}^{K} (\widehat{v}_{j+1}^{(k)} - \widehat{m}_{j+1})(\widehat{v}_{j+1}^{(k)} - \widehat{m}_{j+1}) \end{cases}$$

$$\text{Analysis} \begin{cases} S_{j+1} &= H\widehat{C}_{j+1} H^T + \Gamma \\ K_{j+1} &= \widehat{C}_{j+1} H^T S_{j+1}^{-1} \\ v_{j+1}^{(k)} &= (I - K_{j+1} H)\widehat{v}_{j+1}^{(k)} + K_{j+1} y_{j+1}^{(k)} \\ y_{j+1}^{(k)} &= y_{j+1} + \ _{j+1}^{(k)} \end{cases}$$

Here $\ _j^{(k)}$ are i.i.d. draws from $N(0;\Gamma)$ and perturbed observation refers to the fact that each particle sees an observation perturbed by an independent draw from $N(0;\Gamma)$.

4. Bibliography

- Subsection 1.1 Data Assimilation is a subject which has its roots in the geophysical sciences, and is driven by the desire to improve inaccurate models of complex dynamically evolving phenomena by means of incorporation of data. The book [Kal03] describes data assimilation from the viewpoint of atmospheric weather prediction, whilst the book [Ben02] describes the subject from the viewpoint of oceanography. These two subjects were the initial drivers for evolution of the field. However, other applications are increasingly using the methodology of data assimilation, and the oil industry in particular is heavily involved in the use, and devlopment, of algorithms in this area [ORL08]. The article [ICGL97] is a useful one to read because it establishes a notation which is now widely used in the applied communities and the articles [Nic03, AJSV08] provide simple introductions to various aspects of the subject from a mathematical perspective. The special edition of the journal PhysicaD, devoted to Data Assimilation, [IJ07], provides an overview of the state of the art around a decade ago.
- Subsection 1.2 contains the formulation of Data Assimilation as a fully nonlinear and non-Gaussian problem in Bayesian statistics. This formulation is not yet the basis of practical algorithms in the geophysical systems such as weather forecasting. This is because global weather forecast models involve $n = \mathcal{O}(10^9)$ unknowns, and incorporate $m = \mathcal{O}(10^6)$ data points daily; sampling the posterior on $\mathbb{R}^n$ given data in $\mathbb{R}^m$ in an online fashion, useable for forecasting, is beyond current algorithmic and computational capability. However the fully Bayesian perspective provides a fundamental mathematical underpinning of the subject, from which other more tractable approaches can be systematically derived. See [Stu10] for discussion of the Bayesian approach to inverse problems. Historically, data assimilation has not evolved from this Bayesian perspective, but has rather evolved out of the control theory perspective. This perspective is summarized well in the book [Jaz70]. However, the importance of the Bayesian perspective is increasingly being recognized in the applied communitiess.
- Subsection 1.3 describes the filtering, or sequential, approach to data assimilation, within the fully Bayesian framework. For low dimensional systems the use of particle filters, which may be shown to rigorously approximate the required filtering distribution as it evolves in discrete time, has been enormously succesful; see [DG01] for an overview. Unfortunately, these filters can behave poorly in high dimensions [BLB08, BBL08, SBBA08]. Whilst there is ongoing work to overcome these problems with high-dimensional particle filtering, see [BCJ11, vL10] for example, this work has yet to impact practical data assimilation in, for example, operational weather forecasting. For this reason the *ad hoc* filters, such as 3DVAR, Extended Kalman Filter and Ensemble Kalman Filter, described in section 3, are of great practical importance. And their analysis is an important challenge for applied mathematicians.
- Subsection 2.1 Monte Carlo Markov Chain methods have a long history, initiated through the 1953 paper [MRTT53] and then generalized to an abstract formulation in the 1970 papr [Has70]. The subject is overviewed from an algorithmic point of view in [Liu01]. Theorem 2.2 is contained in [MT93], and that reference also contains many other converegence theorems for Markov chains. The specific forms of the MCMC methods which we introduce here have been chosen to be particularly effective in high dimensions, especially the pCN method; see [CRSW12].
- Subsection 2.2 Variational Methods, known as 4DVAR in the metereology community, have the distinction, when compared with the *ad hoc* non-Gaussian filters described in later sections, of being well-founded statistically: they correspond to the maximum *a posteriori* estimator for the fully Bayesian posterior distribution on model state given data. See [Zup97] and the references therein for a discussion of the applied context. See [CDRS09] for a more theoretical presentation. Currently the European Centre for Medium-Range Weather Forecasts (ECMWF) weather prediction code, which is based on 4DVAR, is the best weather predictor, worldwide. Theorem 3.5 is prototypical of a more sophisticated result, concerning infinite dimensional filtering of dissipative PDEs such as the Navier-Stokes equation, which appears in [BLL+12]. See also [PMLvL12] and [MLPvL12].
- Subsection 3.1 The Kalman Filter has found wide-ranging application to low dimensional engineering applications where the linear Gaussian model is appropriate, especially in econometric time-series analysis, and signal processing [Har91]. It is also important because it plays a key role in the development of the *ad hoc* non-Gaussian filters which are the subject of the next section.
- Subsection 3.2 All the non-Gaussian Filters we discuss are based on modifying the Kalman filter so

that it may be applied to non-linear problems. The development of new filters is a very active area of research and the reader is directed to [MH12, MHG10] and [VL09].

- Subsection 3.3 The 3DVAR algorithm was proposed at the UK Met Office in 1986 [Lor00, LBB+00], and was subsequently developed by the US National Oceanic and Atmospheric Administration [PD92] and by the European Centre for Medium-Range Weather Forecasts (ECMWF) in [CAH+98]. The 3DVAR algorithm is prototypical of the many more sophisticated filters which are now widely used in practice and it is thus natural to study it.
- Subsection 3.4 The extended Kalman filter was developed in the control theory community and is discussed at length in [Jaz70]. It is not practical to implement in high dimensions, and low-rank extended Kalman filters are then used instead; see [LS11] for a recent discussion.
- Subsection 3.5 The ensemble Kalman filter uses a set of particles to estimate covariance information, and may be viewed as an approximation of the extended Kalman filter, designed to be suitable in high dimensions. See [Eve06] for an overview of the methodology, written by one of its originators, and [VLE96] for an early example of the power of the method.

Acknowledgements AMS is grateful to Sergios Agapiou and to Yuan-Xiang Zhang for help in the preparation of these lecture notes. He is also grateful to EPSRC, ERC and ONR for financial support.

References

[AJSV08] A. Apte, C.K.R.T Jones, A.M. Stuart, and J. Voss. Data assimilation: mathematical and statistical perspectives. *Int. J. Num. Meth. Fluids*, 56:1033–1046, 2008.

[BBL08] T. Bengtsson, P. Bickel, and B. Li. Curse of dimensionality revisited: the collapse of importance sampling in very large scale systems. *IMS Collections: Probability and Statistics: Essays in Honor of David Freedman*, 2:316–334, 2008.

[BCJ11] A. Beskos, D. Crisan, and A. Jasra. On the stability of sequential monte carlo methods in high dimensions. *Arxiv preprint arXiv:1103.3965*, 2011.

[Ben02] A. Bennett. *Inverse Modeling of the ocean and Atmosphere*. Cambridge, 2002.

[BLB08] P. Bickel, B. Li, and T. Bengtsson. Sharp failure rates for the bootstrap particle filter in high dimensions. *IMS Collections: Pushing the Limits of Contemporary Statistics*, 3:318–329, 2008.

[BLL+12] CEA Brett, KF Lam, KJH Law, DS McCormick, MR Scott, and AM Stuart. Accuracy and stability of filters for dissipative pdes. *Arxiv preprint arXiv:1203.5845*, 2012.

[CAH+98] P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3d-Var). I: Formulation. *Quart. J. R. Met. Soc.*, 124(550):1783–1807, 1998.

[CDRS09] S.L. Cotter, M. Dashti, J.C. Robinson, and A.M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25:doi:10.1088/0266–5611/25/11/115008, 2009.

[CRSW12] S. Cotter, G. Roberts, A. Stuart, and D. White. Mcmc methods for functions: modifying old algorithms to make them faster. *Arxiv preprint arXiv:1202.0709*, 2012.

[DG01] N. Doucet, A. de Frietas and N. Gordon. *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.

[Eve06] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer, 2006.

[Har91] A.C. Harvey. *Forecasting, structural time series models and the Kalman Iter*. Cambridge Univ Pr, 1991.

[Has70] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[ICGL97] K. Ide, M. Courier, M. Ghil, and A. Lorenc. Unified notation for assimilation: Operational, sequential and variational. *J. Met. Soc. Japan*, 75:181–189, 1997.

[IJ07] K. Ide and C.K.R.T. Jones. Special issue on the mathematics of data assimilation. *PhysicaD*, 230:vii–viii, 2007.

[Jaz70] A.H. Jazwinski. *Stochastic processes and Itering theory*, volume 63. Academic Pr, 1970.

[Kal03] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge, 2003.

[LBB+00] A. C. Lorenc, S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne, and F. W. Saunders. The Met. Office global three-dimensional variational data assimilation scheme. *Quart. J. R. Met. Soc.*, 126(570):2991–3012, 2000.

[Liu01] Jun S. Liu. *Monte Carlo strategies in scienti c computing*. Springer Series in Statistics. Springer, 2001.

[Lor00] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quart. J. R. Met. Soc.*, 112(474):1177–1194, 2000.

[LS11] K.J.H. Law and A.M. Stuart. Evaluating data assimilation algorithms. *Arxiv preprint arXiv:1107.4118*, 2011.

[MH12] A.J. Majda and J. Harlim. Filtering complex turbulent systems. *Recherche*, 67:02, 2012.

[MHG10] A.J. Majda, J. Harlim, and B. Gershgorin. Mathematical strategies for filtering turbulent dynamical systems. *Disc. Cont. Dyn. Sys.*, 2010.

[MLPvL12] A.J.F. Moodey, A.S. Lawless, R.W.E. Potthast, and P.J. van Leeuwen. Nonlinear error dynamics for cycled data assimilation methods. 2012.

[MRTT53] N. Metropolis, R.W. Rosenbluth, M.N. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[MT93] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[Nic03] N.K. Nichols. Data assimilation: aims and basic concepts. In *Data Assimilation for the Earth System, Editors R. Swinbank, V.Shutyaev, W.A.Lahoz*, pages 9–20. Kluwer, 2003.

[ORL08] D.S. Oliver, A.C. Reynolds, and N. Liu. *Inverse theory for petroleum reservoir characterization and history matching*. Cambridge Univ Pr, 2008.

[PD92] D. F. Parrish and J. C. Derber. The national meteorological centers spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120(8):1747–1763, 1992.

[PMLvL12] R.W.E. Potthast, A.J.F. Moodey, A.S. Lawless, and P.J. van Leeuwen. On error dynamics and instability in data assimilation. *Preprint*, 2012.

[SBBA08] T. Snyder, T. Bengtsson, P. Bickel, and J. Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review.*, 136:4629–4640, 2008.

[Stu10] A.M. Stuart. Inverse problems: a Bayesian approach. *Acta Numerica*, 19, 2010.

[VL09] P.J. Van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137:4089–4114, 2009.

[vL10] P.J. van Leeuwen. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999, 2010.

[VLE96] P.J. Van Leeuwen and G. Evensen. Data assimilation and inverse methods in terms of a probabilistic formulation. *Monthly Weather Review*, 124:2892–2913, 1996.

[Zup97] D. Zupanski. A general weak constraint applicable to operational 4dvar data assimilation systems. *Monthly Weather Review*, 125:2274–2292, 1997.