Statistics & Data Assimilation

An introduction

Andreas S. Stordal & Patrick N. Raanes

Outline

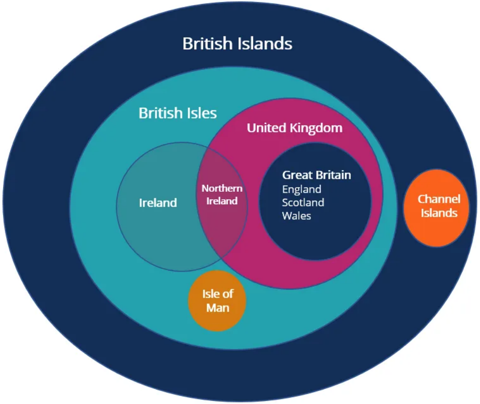
- Probability (crash course)
- Estimation (brief overview)
- State space models
- Monte Carlo methods
- Ensemble Kalman filter

Table of Contents

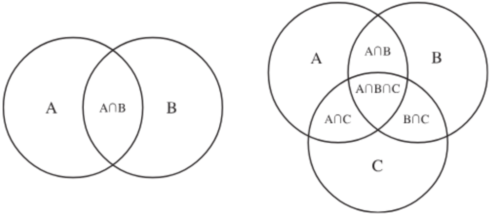
Probability of events

- A ***sample space***, S, is a (finite or countable) set of ***outcomes***, $s \in S$.
- Subsets $A, B, \cdots \subset S$ are called ***events***.
- The ***probability*** of some A is the number of outcomes in A relative to the total: $\mathbb{P}(A) = \frac{\#A}{\#S}$. More generally, $\mathbb{P}$ is defined by
  - $0 \leq \mathbb{P}(A) \leq 1$
  - $\mathbb{P}(S) = 1$
  - For any two *disjoint* events A, B, the probability of *either* one occurring, i.e. $\mathbb{P}(A \cup B)$, equals the sum $\mathbb{P}(A) + \mathbb{P}(B)$.
- The ***joint*** probability is that of both A *and* B occurring, i.e. $\mathbb{P}(A \cap B)$.
- We say that A and B are ***independent*** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ is the ***conditional*** probability of A given B
  - $\mathbb{P}(A) = \sum_{i=1}^{N} \mathbb{P}(A|B_i)\,\mathbb{P}(B_i)$ if $B_1, \ldots B_N$ is a *partition* of S.

Venn diagram examples



Venn diagrams exercise



Exercise:

- Express $\mathbb{P}(A \cup B)$ in terms of the labeled quantities.
- Then do the same for $\mathbb{P}(A \cup B \cup C)$ of the second panel.

Discrete random variables

Instead of asking '*Did* event X_n occur?' (for a family of X_n),
random variables enables the more convenient '*What* was the value of X?'

- Implies that the events (lowercase!) $x_1, \ldots, x_N$ *partition* the sample space.
- $\implies X$ is actually a function mapping any $s \in S$ to some $x_n \in \mathbb{R}$.
- Can have other random variables, e.g. Y, on the *same* prob. space.
- Tend to forget about underlying prob. space.

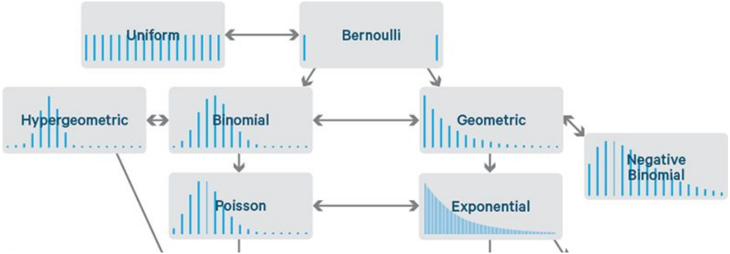
The probability *mass* function (***pmf***) of X is defined as

$$p(x) = \begin{cases} \mathbb{P}(X = x_n) & \text{if } x = x_n\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Clearly, $0 \leq p(x) \leq 1$ and $\sum_n p(x_n) = 1$.

Its ***cumulative distribution function*** (CDF) is: $F(x) = \mathbb{P}(X \leq x) = \sum_{x' \leq x} p(x')$.

Examples



- *Exercise:* What is $F(x)$ for the uniform (constant) dist., i.e. $p(x) = \frac{1}{N}$?

Joint pmf

The *joint* pmf of X and Y is defined as $p(x, y) = \mathbb{P}(X = x \cap Y = y)$

Example:

		Y			
		1	3	9	P(x)
X	2	0.02	0.19	0.08	0.29
	4	0.07	0.14	0.05	0.26
	6	0.05	0.21	0.19	0.45
	P(y)	0.14	0.54	0.32	1

- *Exercise:* What is $p(x|y = 1)$?
- Probabilities $\frac{0.02}{0.14}, \frac{0.07}{0.14}, \frac{0.05}{14}$ for each of $X = 2, 4, 6$.

Continuous random variables

NORCE

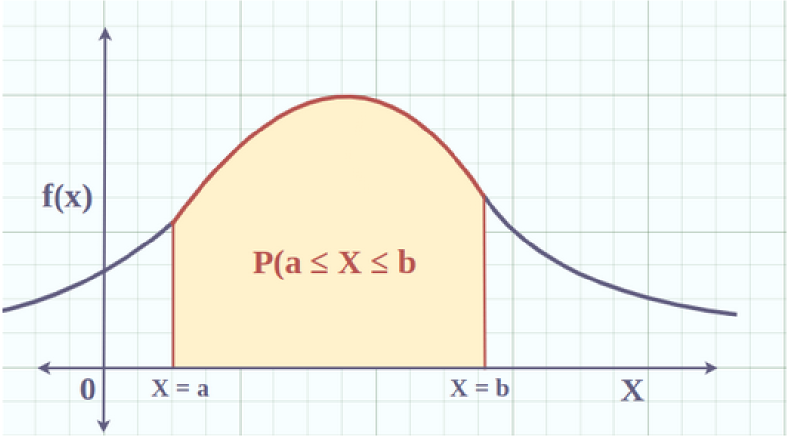
- A *continuous* random variable, X, taking values in $\mathbb{R}$ or some subset thereof, has a probability *density* function (***pdf***) $p(x) \geq 0$ such that

$$\mathbb{P}(X \in A) = \int_A p(x)\, dx.$$

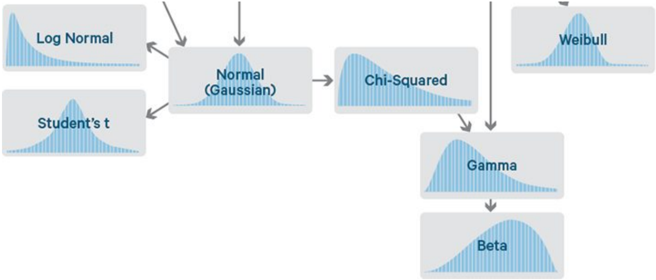
  - Can be derived from pmf by dividing by Δx and letting this $\to 0$.
  - Clearly, $\int p(x)\, dx = 1$.
- Its CDF, $F(x)$, is given by

$$F(x) = \int_{-\infty}^{x} f(z)\, dz$$

Example - probability density function



Example pdfs



- *Exercise:* What is $F(x)$ for the uniform dist. $U[0, a]$, i.e. $p(x) = \frac{1}{a}$ for $x \in [0, a]$.

Independence and conditional densities

- The *conditional* density of X given $Y = y$ is defined as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

- Furthermore, if *independent*,

$$p_{X,Y}(x,y) = p_X(x)\, p_Y(y),$$

Expectation

The ***expected value (first moment)*** of a of a random variable is defined by

$$\mathbb{E}[X] = \int x\, p(x)\, dx$$

[In the discrete case use $\sum \Delta x$]

The expectation is 'essentially/just' the *average/mean* of infinite draws of X:

$$\overline{X}_N := \frac{1}{N}(X_1 + \cdots + X_N) \xrightarrow[N\to\infty]{} \mathbb{E}[X] \,.$$

[law of large numbers (LLN)]

Transformations

- Let $Z = f(X)$ where f is a monotone function with inverse $x = f^{-1}(z)$, then

$$p_Z(z) = p_X\left(f^{-1}(z)\right)\ |\frac{d}{dz}f^{-1}(z)|$$

- *Exercise:* Prove this
- Note that

$$\mathbb{E}[Z] = \int z\, p_Z(z)\, dz = \int f(x)\, p_X(x)\, dx = \mathbb{E}[f(X)]$$

Moments

Similarity, the ***k***-th moment and ***central*** moment are defined by

$$\mathbb{E}[X^k] = \int x^k \, p(x) \, dx$$

$$\mathbb{E}[(X - \mathbb{E}[X])^k] = \int (x - \mathbb{E}[X])^k \, p(x) \, dx$$

- The first moment is simply the *expected value* $\mu_x = \mathbb{E}[X]$.
- The second central moment is the ***variance*** $\sigma_x^2 = \mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.
- The third central moment is the ***skewness*** $\mathbb{E}[(X - \mathbb{E}[X])^3]$.
- Note that the skewness is *zero* for symmetric distributions.
- The fourth central moment is the ***kurtosis*** $\mathbb{E}[(X - \mathbb{E}[X])^4]$.
- The kurtosis says something about how *heavy* the tails are.

Moment generating functions

- For a random variable X, the *moment generating function* (MGF) is define by
$$M_x(t) = \mathbb{E}[e^{tX}], \quad \text{must be finite for } t \in (-\epsilon, \epsilon)$$

- The k-th derivative at zero
$$M_x^{(k)}(0) = [X^k]$$

- MGF is unique
- MGF of a sum is the product of their MGF-s:
$M_{x+y+z}(t) = M_x(t)M_y(t)M_z(t)$
$\implies$ Facilitates finding the distributions of sums of random variables

Expectation properties

In general,

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

Exercise: If independent, then

- $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$,
- $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

Covariance

- Let X and Y be two random variables with
  - Expectations $\mathbb{E}[X] = \mu_x$ and $\mathbb{E}[Y] = \mu_x$
  - Variances $\mathbb{V}[X] = \sigma_x^2$ and $\mathbb{V}[Y] = \sigma_y^2$.
- We define the ***covariance*** between X and Y as

$$\begin{aligned}\mathbb{C}[X,Y] &= \mathbb{E}[(X-\mu_x)(Y-\mu_y)] \\ &= \mathbb{E}[XY] - \mu_x\mu_y \\ &= \mathbb{C}[Y,X]\end{aligned}$$

- *Example:* If $Y = HX$ for some number H, then $\mathbb{C}[Y,X] = H\sigma_x^2$ regardless of the distribution of X and Y.
- *Exercise*: Show that *if* X and Y are *independent*, then $\mathbb{C}[X,Y] = 0$
- Blackboard *exercise:* What is $\mathbb{V}[X+Y]$
  (X and Y not necessarily independent) ?

Correlation

Define the (unitless) ***correlation*** between X and Y as

$$\rho[X, Y] = \frac{\mathbb{C}[X, Y]}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

- Can show by Cauchy-Swartz that $-1 \leq \rho \leq 1$.
- *If* X and Y are *independent*, then $\rho[X, Y] = 0$
- ρ quantifies (defines) the ***linear dependence*** between X and Y.
- *Example:* for $Y = HX$ (as above), $\rho = \pm 1$.
- Blackboard *exercise:* Let X be a symmetric, zero mean random variable with variance one and let $Y = X^2$. What is $\rho[X, Y]$?

Multivariate (vector) case

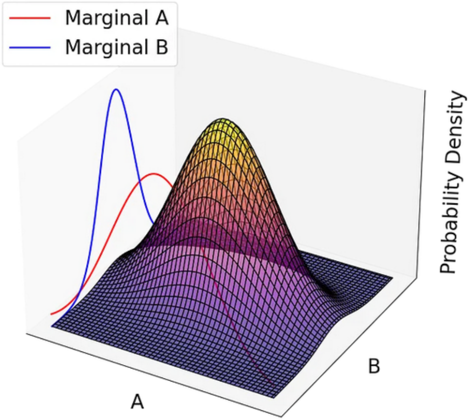
- A multivariate, continuous random variable, $X = (X_1, X_2, \dots, X_d)$, taking values in $\mathbb{R}^d$ or some subset, has a probability density function (pdf) $p(x) = p(x_1, x_2 \dots, x_d) \geq 0$ such that

$$\mathbb{P}(X \in A) = \int_A p(x_1, x_2 \dots, x_d)\, dx_1 dx_2 \dots dx_d,$$

  - A joint is a multivariate distribution.
- Its cumulative distribution function, $F(x)$ is given by

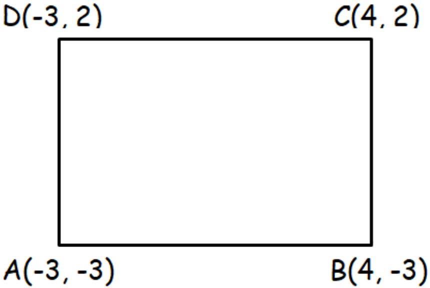
$$\begin{aligned} F(x) &= \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots X_d \leq x_d) \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_d} p(z_1, z_2, \dots, z_d)\, dz_1 dz_2 \dots dz_d \end{aligned}$$

Example - joint density function



Exercise: Multivariate integration (CDF)

Express the probability that a random variable X with CDF F lies within the rectangle



Marginal distributions

For multivariate continuous random variable, $X = (X_1, X_2, \ldots, X_d)$, the marginal distribution for any subset is given by (example:)

$$p(x_1, \ldots x_{k-1}, x_{k+1} \ldots x_d) = \int_{-\infty}^{\infty} p(x_1, x_2, \ldots x_k, \ldots, x_d)\, dx_k$$

$$p(x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots x_k, \ldots, x_d)\, dx_1 dx_2 \ldots dx_{k-1} dx_{k+1} \ldots dx_d$$

Factorization

Any joint density, $p(x_1, x_2, \ldots, x_d)$, can be factorized as

$$p(x_1, x_2, \ldots, x_d) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_2, x_1) \ldots p(x_d|x_1, x_2, \ldots x_d)$$

The ordering can be arbitrary and allows us to work only with marginal distributions

Covariance matrix

- Let X be a random vector.
  Its ***covariance matrix*** is defined by

$$\begin{aligned}\mathbf{\Sigma}_x &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X]^\top)] \\ &= \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top\end{aligned}$$

- Other frequently used notations: $\mathbf{C}_{xx}$ and $\mathbf{P}$,
  and we'll also encounter $\mathbf{R}$ and $\mathbf{Q}$!!!
- Σ_x is *symmetric* and *positive-definite*
- The *diagonal* elements are $[\mathbf{\Sigma}_x]_{ii} = \mathbb{V}[X_i]$
- The *off-diagonal* elements are $[\mathbf{\Sigma}_x]_{ij} = \mathbb{C}[X_i, X_j]$
- Σ_x is diagonal if all components of X are *independent*

More covariance

- If X has covariance matrix $\mathbf{\Sigma}_x$ and $Y = a + \mathbf{A}X$, then

$$\mathbf{\Sigma}_y = \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^\top$$

- The ***cross covariance*** matrix between two random vectors X and Y is:

$$\begin{aligned}\mathbf{\Sigma}_{xy} &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]^\top)] \\ &= \mathbb{E}[XY^\top] - \mathbb{E}[X]\mathbb{E}[Y]^\top\end{aligned}$$

- If $Z = [X,\ Y]$ then

$$\mathbf{\Sigma}_z = \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_y \end{bmatrix}$$

Table of Contents

Likelihood function

- For a given observation, y, we often refer to the *likelihood*
- What is a likelihood?
- For a statistical model with state X and/or parameter θ , we describe the model in terms of probability density functions, $p(y|x)$ or $p(y|\theta)$
- Yet we often refer to $p(y|x)$ or $p(y|\theta)$ as the *likelihood*
- The function $p(y|\theta)$ is a *density* w.r.t. y, and thus integrates to 1 for any fixed value of θ.
- However, for fixed y, we can define $\ell(\theta) = p(y|\theta)$ as a function of θ, a *likelihood* function. It does not integrate to 1
- For a given observation y and a given value θ, the value of the likelihood function tells us how 'likely' it is that the observation originates from a model with the given value for θ.

Likelihood in DA

- How do we go from observation to likelihood in Data Assimilation?
- We have observed $Y = y$
- We have assumed $y = \mathcal{H}(x) + \epsilon$
- It is the distribution of ϵ that defines the likelihood of the observation y, evaluated at the model output $\mathcal{H}(x)$

Likelihood in DA

- $Y = \mathcal{H}(x) + \epsilon$, hence $Y - \mathcal{H}(x) = \epsilon$
- As soon as we specify a probability density for ϵ, we have a likelihood
- $p(y|x) = p_\epsilon(y - \mathcal{H}(x))$
- We claim: an observation without uncertainty is infinitely less valuable than one with uncertainty specified.
- Tell your engineer!

Exercise

- Let Y be the time it takes for a patient to recover (in days) after surgery. Assume that Y is exponentially distributed with parameter θ. We start observations 1 weeks after surgery and observe the patients for 2 weeks.
- What is the likelihood function for θ?

Point Estimation

- An estimator of an unknown quantity, θ, is any function of the data, $\hat{\theta} = f(y_{1:n})$
- An estimator, $\hat{\theta}$, is unbiased if $\mathbb{E}_\theta(\hat{\theta}) = \theta$
- Most classical methods are the method of moments and maximum likelihood
- Bayesian point estimators are derived from the posterior, often the mean or mode depending on loss function used

Evaluating estimators

- For an estimator $\hat{\theta}$ we may evaluate the 'quality' by asking:
- Is the estimator precise?

$$\mathbb{B}[\hat{\theta}] = \mathbb{E}[\hat{\theta} - \theta], \quad \text{This is the bias}$$

- Is the estimator reliable?

$$\mathbb{V}[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2], \quad \text{This is the variance}$$

- The *mean squared error* defines the quality of the estimator

$$MSE[\hat{\theta}] = \mathbb{E}[(\theta - \hat{\theta})^2] = \mathbb{V}[\hat{\theta}] + \text{Bias}[\hat{\theta}]^2$$

Method of moments

- Assume we have N independent observations, $y_{1:n} = (y_1, y_2, \ldots, y_N)$ from a distribution/model with p unknown parameters $\theta = (\theta_1, \ldots, \theta_p)$
- Match p empirical and theoretical moments
  to estimate θ from $y_{1:n}$

$$\mathbb{E}_\theta(Y^k) = N^{-1} \sum_{i=1}^{n} y_i^k, \quad k = 1, \ldots p$$

- p equations, p unknowns
- Consistent due to S-LLN

Maximum likelihood

NORCE

- Given data $y = y_{1:n}$ from a likelihood model $p(y|\theta)$
- $\hat{\theta} =$ arg max$_\theta$ $\quad p(y|\theta)$
- If true likelihood is $\tilde{p}(y)$ then $p(y|\theta)$ asymptotically minimize

$$KL(\tilde{p}||p_\theta) = \int \log \frac{\tilde{p}(y)}{p(y|\theta)} \tilde{p}(y)\, dy$$

- Not always unbiased (restricted ML often alternative)
- IF a uniformly minimum variance unbiased estimator (UMVUE) exists, then it is a ML estimator
- ML is transformation invariant, $\widehat{g(\theta)} = g(\hat{\theta})$, where g is any function

Exercise

- Let $y_1, y_2, \ldots, y_N$ be an i.i.d. sample from a uniform density on $(0, \theta)$
- Find (1) the moment estimator and (2) maximum likelihood estimator for θ

Bayesian inference

- A model typically consists of unknown parameters,θ, and observations y from the likelihood $p(y|\theta)$
- Classical statistics treats θ as a fixed number that should be estimated from observations
- Bayesian statistics treats θ as a random variable whose density quantifies belief and is updated using observations

Prior and Posterior

- Bayesian statistics is conceptually simple
- For an unknown parameter θ, incorporate prior believes into a prior pdf $p(\theta)$
- Given data y from a likelihood model $p(y|\theta)$
- Update from prior to posterior using Bayes' rule

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{\int p(y|\theta)\, p(\theta)\, d\theta}$$

- The integral is the hard part in general

Predictions in Bayesian statistics

- Prior predictive distribution

$$p(y) = \int p(y|\theta)\, p(\theta)\, d\theta$$

- Posterior predictive

$$p(y'|y) = \int p(y'|\theta)\, p(\theta|y)\, d\theta$$

Monte Carlo versions are often used without referencing these equations

Hierarchical Bayes

- Often we have latent variables or hyperparameters in models

$$\text{Likelihood } p(y|x,\theta), \text{ prior } p(x|\theta), \text{ hyper prior } p(\theta)$$

$$\text{posterior } p(\theta, x|y) = \frac{p(y|x,\theta)\, p(x|\theta)\, p(\theta)}{p(y)}$$

- $p(y)$ is known as the model evidence, given two models: m_1 and m_2 we can compute the Bayes ratio

$$\frac{p(y|m_1)}{p(y|m_2)} = \frac{\int p(y|x,\theta_1)\, p(x|\theta_1)\, p_1(\theta_1)\, d\theta_1}{\int p(y|x,\theta_2)\, p(x|\theta_2)\, p_2(\theta_2)\, d\theta_2}$$

Table of Contents

State space models

Hidden Markov models

- Initial condition $X_0 \sim p(x_0)$. We will abuse the p notation
- Markov transitions: $X_t \sim p(x_t|x_{t-1})$, e.g. $X_t = \mathcal{M}(X_{t-1}, \eta_t)$
- Discrete time measurements Y_t, $t = 1, 2, \dots, T$
- Measurement operator $Y_t = \mathcal{H}(X_t) + \epsilon_t \to p(y_t|x_t)$

Our objective is either

- filtering $p(x_t|y_{1:t})$
- smoothing $p(x_t|y_{1:t})$
- forecasting $p(x_{t+1}|y_{1:t})$

Bayes' rule with several events

Typical formulation

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\,\mathbb{P}(B)}{\mathbb{P}(A)}$$

Often we condition on several events

$$\mathbb{P}(B|A,C) = \frac{\mathbb{P}(A|B,C)\,\mathbb{P}(B|C)}{\mathbb{P}(A|C)}$$

Frequently used in filtering and smoothing

Prediction step

Recall: $p(a,b) = \int_B p(a,b)\,db$. Similarly

$$\begin{aligned} p(x_k|y_{1:t-1}) &= \int p(x_t, x_{t-1}|y_{1:t-1})\,dx_{t-1} \\ &= \int p(x_t|x_{t-1}, y_{1:t-1})\,p(x_{t-1}|y_{1:t-1})\,dx_{t-1} \\ &= \int p(x_t|x_{t-1})\,p(x_{t-1}|y_{1:t-1})\,dx_{t-1} \end{aligned}$$

Chapman-Kolmogorov forward equation
Yesterdays forecast

Filter step

Using Bayes' rule:

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)\,p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})}$$

$$p(y_t|y_{1:t-1}) = \int p(y_t|x_t)\,p(x_t|y_{1:t-1})\,dx_t$$

Smoothing step

Hindcast step

$$p(x_t|y_{1:T}) = \int p(x_t, x_{t+1}|y_{1:T}) d_{x_{t+1}}$$
$$= \int p(x_t|x_{t+1}, y_{1:T})\, p(x_{t+1}|y_{1:T}) d_{x_{t+1}}$$

Exercise: Show that

$$p(x_t|x_{t+1}, y_{1:T}) = p(x_t|x_{t+1}, y_{1:t})$$

Joint and conditional Gaussian random variables

- Let Z be a Gaussian random vector
- Then all combinations of sub-vectors are also Gaussian random vector
- Moreover, all conditional distributions are also Gaussian
- If $Z = [X,\ Y]$ we have $\mu_z = [\mu_x, \mu_y]$ and

$$\boldsymbol{\Sigma}_z = \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y \end{bmatrix}$$

- What is the distribution if X given Y?

Joint and conditional Gaussian random variables

- $Z = [X, Y]$
- $p(x|y)$ is Gaussian with mean and covariance given by

$$\mu_{x|y} = \mu_x + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}(y - \mu_y),$$
$$\mathbf{\Sigma}_{x|y} = \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}$$

- Note that $\Sigma_{x|y}$ is independent of the actual value of Y
- These are the building blocks of the Kalman filter (and ensemble versions)

Kalman Filter

- Analytical solution to filter problem in linear/Gaussian state space models
- System of the form

$$
\begin{aligned}
&X_0 \sim \mathcal{N}(\mu_0, \mathbf{P}_0) \\
&X_t = \mathbf{M} X_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathbf{Q}) \\
&Y_t = \mathbf{H} X_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{R})
\end{aligned}
$$

- At each time step, we can use properties of Gaussian random vectors to derive the filtering solution (assuming independence between all combinations of η_t and ϵ_k)

Kalman Filter

- X_0 is Gaussian with mean μ_0 and $\mathbf{P}_0$
- Using affine properties of Gaussian random vectors, $[X_1, Y_1]$ is Gaussian with mean and covariance

$$
\begin{aligned}
\mu_1^f &= \mathbf{M}\mu_0, \\
\mu_{y_1} &= \mathbf{H}\mu_1^f, \\
\mathbf{P}_1^f &= \mathbf{M}\mathbf{P}_0\mathbf{M}^\top + \mathbf{Q}, \\
\mathbf{P}_{y_1} &= \mathbf{H}\mathbf{P}_1^f\mathbf{H}^\top + \mathbf{R}, \\
\mathbf{P}_{x_1,y_1} &= \mathbf{P}_1^f\mathbf{H}^\top
\end{aligned}
$$

Kalman Filter

- Using properties of conditional Gaussian random vectors, X_1 given $Y_1 = y_1$ is Gaussian with mean and covariance

$$\mu_1^a = \mu_1^f + \mathbf{P}_1^f \mathbf{H}^\top (\mathbf{H}\mathbf{P}_1^f \mathbf{H}^\top + \mathbf{R})^{-1} (y_1 - \mathbf{H}\mu_1),$$
$$\mathbf{P}_1^a = \mathbf{P}_1^f - \mathbf{P}_1^f \mathbf{H}^\top (\mathbf{H}\mathbf{P}_1^f \mathbf{H}^\top + \mathbf{R})^{-1} \mathbf{H}\mathbf{P}_1^f$$

- This is valid for all t by replacing 1 with t and 0 with $t-1$, by induction
- Defining $\mathbf{K}_t = \mathbf{P}_t^f \mathbf{H}^\top (\mathbf{H}\mathbf{P}_t^f \mathbf{H}^\top + \mathbf{R})^{-1}$ we have

$$\mu_t^a = \mu_t^f + \mathbf{K}_t (y_t - \mathbf{H}\mu_t),$$
$$\mathbf{P}_t^a = (\mathbf{I} - \mathbf{K}_t \mathbf{H}) \mathbf{P}_t^f$$

Table of Contents

Monte Carlo Sampling

- Let X be a random variable with probability density $p(x)$
- For any function f define the expectation $\mathbb{E}_p[f(X)] = \int f(x)\, p(x)\, dx$
- Assume $\{X^i\}_{i=1}^N$ is an i.i.d. sample from $p(x)$
- Then $N^{-1} \sum_{i=1}^n f(X^i)$ converges to $\mathbb{E}_p[f(X)]$, if the variance is finite
- Note: $\mathbb{E}_p[N^{-1} \sum_{i=1}^n f(X^i)] = \mathbb{E}_p[f(X)]$ (unbiased)

Ensemble representation in Data Assimilation

- In data assimilation we often work with Gaussian assumptions, i.e. first and second order moments
- Parameters and states are represented by an *initial* ensemble $\{X^i\}_{i=1}^N$, i.e. a *Monte Carlo sample* representing the distribution at the initial time and/or the prior distribution if parameters.
- $\mathbb{E}[X] \approx N^{-1} \sum_{i=1}^{n} X^i$
- $\mathbb{C}_x \approx (N-1)^{-1} \sum_{i=1}^{N} (X^i - \overline{X})(X_i - \overline{X})^\top = \mathbf{A}\mathbf{A}^\top$
- $\mathbf{A} = (N-1)^{-1/2}[X_1 - \overline{X}, X_2 - \overline{X} \ldots, X_N - \overline{X}]$ is often called the *ensemble anomaly* matrix

Predictions

- Given an initial ensemble $\{X^i\}_{i=1}^N$, we can compute first and second order moments of the *forecast ensemble* by 'applying' our model of interest, $\mathcal{M}$ to each ensemble member
- $\mathbb{E}[\mathcal{M}(X)] \approx N^{-1} \sum_{i=1}^N \mathcal{M}(X^i) = \overline{\mathcal{M}}$
- $\mathbb{C}_{\mathcal{M}} \approx (N-1)^{-1} \sum_{i=1}^N (\mathcal{M}(X^i) - \overline{\mathcal{M}})(\mathcal{M}(X_i) - \overline{\mathcal{M}})^\top$
- $\mathbb{C}_{\mathcal{M},x} \approx (N-1)^{-1} \sum_{i=1}^N (\mathcal{M}(X^i) - \overline{\mathcal{M}})(X_i - \overline{X})^\top$

More Monte Carlo

NORCE

- Let X be a random variable with probability density $p(x)$ and cumulative density function $F(x) = \int_{-\infty}^{x} p(u)\, du$
- Let U be a uniform random variable on $[0\,1]$
- Then $X = F^{-1}(U)$ has density $p(x)$ (*exercise:* prove this)
- U can easily be generated (pseudo) randomly on a computer
- F^{-1} is only known for some (simple) distributions

Importance Sampling

- What if I cannot sample from $p(x)$, but $q(x)$? (another density with at least same support)
- Since, for an arbitrary function f

$$\mathbb{E}_p[f(X)] = \int f(x)\,p(x)\,dx = \int f(x)\frac{p(x)}{q(x)}q(x)\,dx = \mathbb{E}_q\left[f(X)\frac{p(X)}{q(X)}\right]$$

- Sample $\{X^i\}_{i=1}^N$ from q and then $\mathbb{E}_q\left[N^{-1}\sum_{i=1}^N f(X^i)\frac{p(X^i)}{q(X^i)}\right] = \mathbb{E}_p[f(X)]$ (unbiased)
- $w(x) = \dfrac{p(x)}{q(x)}$ is the weight function

Proportionality

NORCE

- What if we can only evaluate p up to a constant, i.e. $p(x) = c^{-1}\tilde{p}(x)$ where the constant c is unknown and $\tilde{p}$ is known?
- Note that $\mathbb{E}_q\left[N^{-1}\sum_{i=1}^N f(X^i)\frac{\tilde{p}(X^i)}{q(X^i)}\right] = c\mathbb{E}_p[f(X)]$ multiplicative bias)
- However $\mathbb{E}_q\left[N^{-1}\sum_{i=1}^N \frac{\tilde{p}(X^i)}{q(X^i)}\right] = c$
- We can study the ratio
- Define the weight function $w(x) = \dfrac{\tilde{p}(x)}{q(x)}$

Importance sampling

- Sample $X_i, \ldots X_N$ from q
- Compute

$$\tilde{w}_i = \frac{\tilde{p}(X_i)}{q(X_i)}$$

$$w_i = \frac{\tilde{w}_i}{\sum_j \tilde{w}_j}$$

- Then $\sum_i f(X_i) w_i \to E_p[f(X)]$, but it is biased for finite N

Table of Contents

- Probability (crash course)
- Estimation (brief overview)
- State space models
- Monte Carlo methods
- Ensemble Kalman filter

Ensemble Kalman Filter

- Monte Carlo version of Kalman filter for nonlinear systems

$$
\begin{aligned}
X_0 &\sim \mathcal{N}(\mu_0, \mathbf{P}_0) \\
X_t &= \mathcal{M}(X_{t-1}) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathbf{Q}) \\
Y_t &= \mathcal{H}X_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{R})
\end{aligned}
$$

- How do we 'Kalman filter' this?
- Alternativ 1: Linearize model (and measurement operator)

$$
\begin{aligned}
\mu_t^f &= \mathcal{M}(\mu_{t-1}^a) \\
\mathbf{P}_t^f &= \mathbf{M}\mathbf{P}_{t-1}^a\mathbf{M}^\top + \mathbf{Q}
\end{aligned}
$$

- $\mathbf{M}$ is the *Jacobian* of the model evaluated at μ_{t-1}^a

Nonlinear Kalman Filter

- If the measurement operator is also nonlinear: $Y_t = \mathcal{H}(X_t) + \epsilon_t$ we get the update equation

$$
\begin{aligned}
\mu_t^a &= \mu_t^f + \mathbf{K}_t(y_t - \mathcal{H}(\mu_t^f)), \\
\mathbf{K}_t &= \mathbf{P}_t^f \mathbf{H}^\top (\mathbf{H}\mathbf{P}_t^f \mathbf{H}^\top + \mathbf{R})^{-1} \\
\mathbf{P}_t^a &= (\mathbf{I} - \mathbf{K}_t \mathbf{H}) \mathbf{P}_t^f
\end{aligned}
$$

- Where $\mathbf{H}$ is the Jacobian of the measurement operator evaluated at μ_t^f
- This is the classical *Extended* Kalman Filter
- Jacobians are often not available for complex models, and it might lead to unstable updates

Monte Carlo version

- In the Kalman Filter, how can we
- Sample a random variable from the forecast distribution at time t using a random variable from the analysis distribution at time $t-1$?
- How can we sample a random variable form the analysis distribution at time t using a random variable from the forecast distribution at time t?

Ensemble Kalman Filter

- Given a (assume independent) sample $\{X_{t-1}^{a,i}\}_{i=1}^{N}$ from the analysis distribution at $t-1$:
- Sample the forecast distribution:

$$X_t^{f,i} = \mathcal{M}(X_{t-1}^{a,i}) + \eta_t^i, \quad i = 1, \dots, N$$
$$Y_t^{f,i} = \mathcal{H}(X_t^{f,i}) + \epsilon_t^i$$

- Compute sample covariances $\mathbf{P}_x$,$\mathbf{P}_{xy}$ and $\mathbf{P}_y$
- *Update* the sample (ensemble)

$$X_t^{a,i} = X_t^{f,i} + \mathbf{P}_{xy}\mathbf{P}_y^{-1}(y_t^{\text{obs}} - Y_t^{f,i})$$

- This is *one version* of the Ensemble Kalman filter

Ensemble Kalman Filter v2

- Given a (assume independent) sample $\{X_{t-1}^{a,i}\}_{i=1}^{N}$ from the analysis distribution at $t-1$:
- Sample the forecast distribution:

$$
\begin{aligned}
X_t^{f,i} &= \mathcal{M}(X_{t-1}^{a,i}) + \eta_t^i, \quad i = 1, \ldots, N \\
Y_t^{f,i} &= \mathcal{H}(X_t^{f,i})
\end{aligned}
$$

- Compute sample covariances $\mathbf{P}_x$,$\mathbf{P}_{xy}$ and $\mathbf{P}_y$
- *Update* the sample (ensemble)

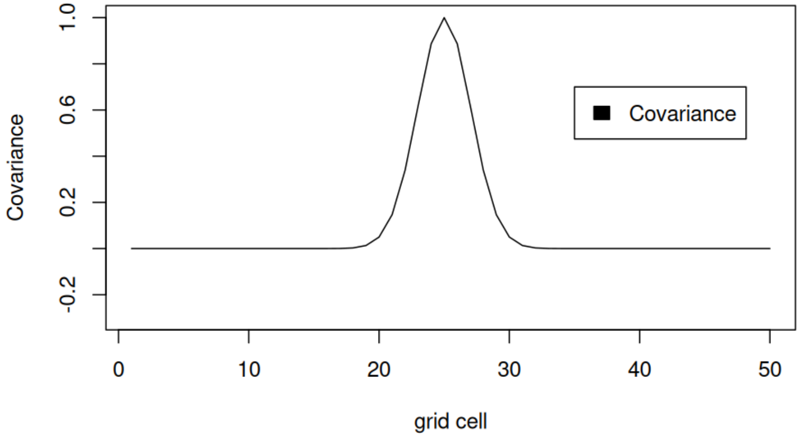
$$X_t^{a,i} = X_t^{f,i} + \mathbf{P}_{xy}(\mathbf{P}_y + \mathbf{R})^{-1}(y_t^{\text{obs}} - Y_t^{f,i} + \epsilon_t^i)$$

- This is *another version* of the Ensemble Kalman filter

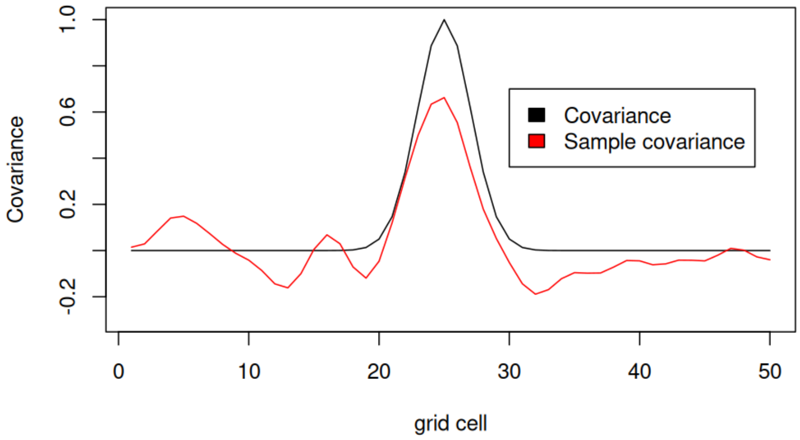
Localization

- One major challenge with EnKF is the poor estimation of high dimensional covariance matrices using a small sample size
- In addition to Monte Carlo errors, the fact that each ensemble member is updated using the sample covariances results in a positive correlation between ensemble members and hence a *under estimation* of the uncertainty
- To classical ways to deal with this is *localization* and *inflation*
- Localization is typically done either by multiplying the Kalman gain or covariance matrices with a *tapering* function, typically based on distances, or by doing *local updates*. Both works best if there is a physical distance between states and observations. For global parameters or non-local observations, covariance thresholding can be used.

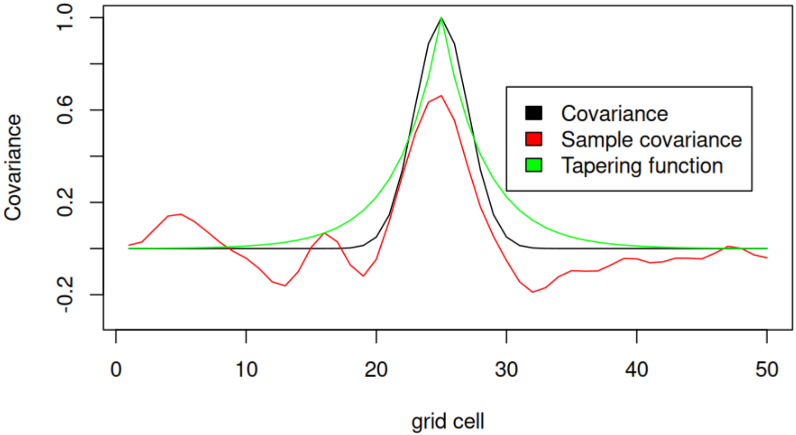
Covariance function at grid cell 25



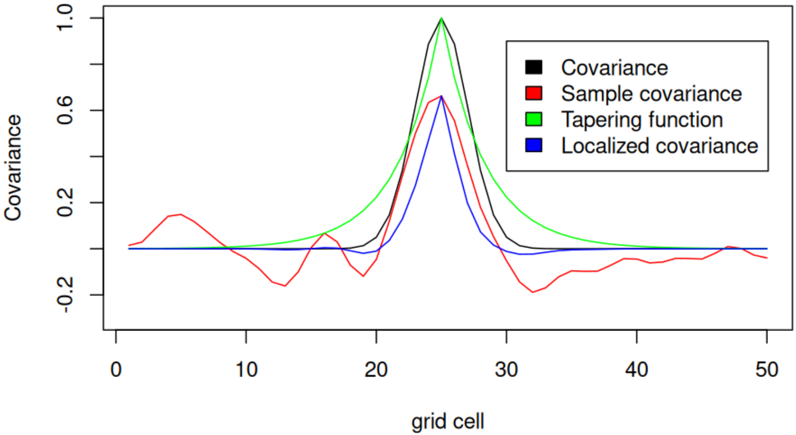
Sample covariance



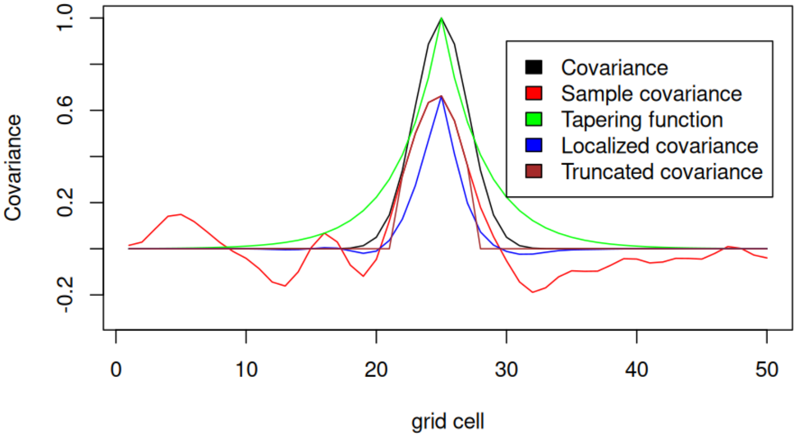
Tapering function



Localized covariance



Truncated covariance



Inflation

- Inflation is applied to either the forecast- or analysis anomalies (or anomaly matrix)
- $X = \overline{X} + \alpha(X - \overline{X})$
- $\alpha > 1$ is the *inflation factor* that 'compensates' for low rank/underestimated variance in the ensemble
- Benchmark studies on the Lorenz models shows that the optimal EnKF requires both inflation and localization, but they are both hard to tune

Square root ensemble methods

NORCE

- In stochastic EnKF, the simulated measurements are 'perturbed' with a random variable following the observation error distribution

$$X_t^{a,i} = X_t^{f,i} + \mathbf{K}(y_t - \mathbf{H}(X_t^{f,i}) + \epsilon_t^i), \qquad \epsilon_t^i \sim \mathcal{N}(0, \mathbf{R})$$

- This ensures that $\mathbf{P}_t^a \approx (\mathbf{I} - \mathbf{KH})\mathbf{P}_t^f$ (with equality as $N \to \infty$)
- Square root filter(s) forces $\mathbf{P}_t^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}_t^f$ for the ensemble
- $\mathbf{X}_t^f = \bar{X}_t{}^f + (N-1)^{1/2}\mathbf{A}_t^f$, where $\mathbf{A}_t^f$ is a matrix with column i equal to $(N-1)^{-1/2}(X_t^{f,i} - \bar{X}_t{}^f)$

Square root update

- Update $\bar{X}_t{}^a = \bar{X}_t{}^f + \mathbf{K}(y_t - H\bar{X}_t{}^f)$
- The updated anomaly matrix is given by

$$
\begin{aligned}
\mathbf{P}_t^a = \mathbf{A}_t^a(\mathbf{A}_t^a)^\top &= [\mathbf{I} - \mathbf{P}_t^f\mathbf{H}^\top(\mathbf{H}\mathbf{P}_t^f\mathbf{H}^\top + \mathbf{R})^{-1}\mathbf{H}]_t^f \\
&= \mathbf{A}_t^f[\mathbf{I} - (\mathbf{A}_t^f)^\top\mathbf{H}(\mathbf{H}\mathbf{A}_t^f(\mathbf{A}_t^f)^\top\mathbf{H}^\top + \mathbf{R})^{-1} + \mathbf{H}\mathbf{A}_t^f](\mathbf{A}_t^f)^\top \\
&= \mathbf{A}_t^f[\mathbf{I} - \mathbf{V}_t\mathbf{D}_t^{-1}\mathbf{V}_t^\top](\mathbf{A}_t^f)^\top,\ \mathbf{V}_t = (\mathbf{H}\mathbf{A}_t^f)^\top \text{ and } \mathbf{D}_t = \mathbf{V}_t^\top\mathbf{V}_t
\end{aligned}
$$

$$\mathbf{A}_t^a = \mathbf{A}_t^f[\mathbf{I} - \mathbf{V}_t\mathbf{D}_t^{-1}\mathbf{V}_t^\top]^{1/2}\mathbf{U}_t,\ \mathbf{U}_t \text{ is a random orthogonal matrix}$$

$$\mathbf{A}_t^a = \mathbf{A}_t^f\mathbf{B}\mathbf{\Gamma}^{1/2}\mathbf{B}\top\mathbf{U}_t,\ \text{ where } [\mathbf{I} - \mathbf{V}_t\mathbf{D}_t^{-1}\mathbf{V}_t^\top] = \mathbf{B}\mathbf{\Gamma}\mathbf{B}^\top$$

- This is known as the symmetric solution, others exist

Subspace/Transform variants

- Since the ensemble size is typically much smaller then the dimension of the state space, and since the update of EnKF is a linear combination of the ensemble, we never are actually working in an $N-1$ dimensional subspace/flat
- By re-writing the ensemble matrix, $\mathbf{E} = [X^1, X^2, \ldots, X^N]$ as

$$\mathbf{E} = \overline{X} + \mathbf{A}\mathbf{W}$$

- $\mathbf{W}$ is an $N \times N$ matrix, initially equal to the identity matrix, and is the one being updated in subspace/transform methods.

Ensemble smoother and iterative methods

- Used to update parameters/initial conditions or states in a given time window without stopping and starting simulations
- More data, longer time evolution of model between updates. More nonlinear/non Gaussian. Poor results
- Introduce iterations. Either by an annealing process or by recasting the problem as an optimization problem
- Methods to study: Iterative ensemble smoother (IES), Randomized maximum likelihood(RML/EnRML), multippel data assimilation (ESMDA)
- Alternative methods such as particle flow, Gaussian mixtures and optimal transport are also available in the DA literature