Data Assimilation and Climate

AIMS Rwanda, March, 2020

Lecture Notes

Discrete time stochastic process

Climate models

We have considered data assimilation methods for the energy balance model

$$C_p \frac{\partial T}{\partial t} = S(1-\alpha) - 4\epsilon\sigma T^4.$$

Alternatively, General Circulation Models (GCM) typically involve solving a conservation of momentum equation (F=ma) such as

$$\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right)\mathbf{v} = \frac{1}{\rho}\nabla p - g - F_{fric} + F_{coriolis}.$$

The unknowns include velocity in three directions, and pressure evaluated on a three dimensional grid. These unknowns form the state vector

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

where the value of n is the product of the number of variables and the number of grid points.

Discrete time stochastic process

Define the process variables as

$$\mathbf{X}(k) \in \mathbb{R}^n \quad k = 0, 1, \ldots$$

and data as

$$\mathbf{Y}(k) \in \mathbb{R}^m \quad k = 0, 1, \ldots$$

where $\mathbf{X}(0)$ is a background state for which there are no observations available.

Let $\mathbf{X}(0{:}i)$ to represent the sequence $\{\mathbf{X}(k), k = 1, \ldots, i\}$, similarly for $\mathbf{Y}(0{:}i)$. With this notation we define

$$f_{X(0:i),Y(1:i)}(x(0{:}i), y(1{:}i))$$

as the joint density function of the process variables $\{X(0), ..., X(i)\}$ and observations $\{Y(1), ..., Y(i)\}$, while

$$f_{X(2:i)|y(1:i-1)}(x(2{:}i))$$

is the conditional pdf for $\{X(2), ..., X(i)\}$ given observations $\{y(1), ..., y(i-1)\}$.

Since observations arrive sequentially, we can find forecast, filter and reanalysis estimates sequentially.

Lecture Activities 1.

Notation for Reanalysis, Filtering, and Forecasting Posterior Distributions

- Filtering: we seek distributions $f_{\mathbf{X}(i)|y(1:i)}(\mathbf{x}(i))$, $i \in \{1, \ldots, N\}$.
  (i.e. stimate the current state based on current and past observations)

- Forecasting: we seek distributions $f_{\mathbf{X}(i)|y(1:i-1)}(\mathbf{x}(i))$, $i \in \{2, \ldots, N+1\}$.

- Re-analysis: we seek distributions $f_{\mathbf{X}(i)|y(1:N)}(\mathbf{x}(i))$, $i = N, N-1, ..., 0$.
  (i.e. estimate a past state from all available data)

Non-stationary stochastic process and data models

$$
\begin{aligned}
\mathbf{X}(i) &= \mathbf{M}_i \mathbf{X}(i-1) + \boldsymbol{\delta}_i \quad i = 1, \ldots, N \\
\mathbf{Y}(i) &= \mathbf{H}_i \mathbf{X}(i) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N
\end{aligned}
$$

with $\mathbf{X}(0) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$, $\boldsymbol{\delta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i)$.

In our previous process models the mean and covariance were constant in time. A Data Assimilation scheme could be used at different times based on these invariant statistics, and this does happen in practice for climate models. In particular, errors considered stationary over a one-month time scale is reasonable. However, for general environmental applications, the governing equations vary with time and we must take into account non-stationary processes with model means and covariances that are not constant in time.

Activities 2(a)(b)

Notation for Process Variables' Means and Covariances

$$
\begin{aligned}
\text{Forecasting}: &\quad \boldsymbol{\mu}_{i|i-1} = E[\mathbf{X}(i)|\mathbf{y}(1:i-1)], \quad \boldsymbol{\Sigma}_{i|i-1} = \mathrm{Var}[\mathbf{X}(i)|\mathbf{y}(1:i-1)] \\
\text{Filtering}: &\quad \boldsymbol{\mu}_{i|i} = E[\mathbf{X}(i)|\mathbf{y}(1:i)], \quad \boldsymbol{\Sigma}_{i|i} = \mathrm{Var}[\mathbf{X}(i)|\mathbf{y}(1:i)] \\
\text{Reanalysis}: &\quad \boldsymbol{\mu}_{i} = E[\mathbf{X}(i)|\mathbf{y}(1:N)], \quad \boldsymbol{\Sigma}_{i|i} = \mathrm{Var}[\mathbf{X}(i)|\mathbf{y}(1:N)],
\end{aligned}
$$

Consider

$$\mathbf{X}(i) \;=\; \mathbf{M}_i\mathbf{X}(i-1) + \boldsymbol{\delta}_i \quad i = 1, \dots, N$$

with $\mathbf{X}(0) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\delta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i)$. Then

$$
\begin{aligned}
\boldsymbol{\mu}_{0|0} = &\quad E[\mathbf{X}(0)] = \boldsymbol{\mu}, \quad \boldsymbol{\Sigma}_{0|0} = \mathrm{Var}[\mathbf{X}(0)] = \boldsymbol{\Sigma} \\
\boldsymbol{\mu}_{1|0} = &\quad E[\mathbf{X}(1)] = \mathbf{M}_1\boldsymbol{\mu}, \quad \boldsymbol{\Sigma}_{1|0} = \mathrm{Var}[\mathbf{X}(1)] = \mathbf{M}_1\boldsymbol{\Sigma}\mathbf{M}_1^T + \mathbf{Q}_1
\end{aligned}
$$

and

$$\boldsymbol{\mu}_{i|i-1} = \mathbf{M}_i\boldsymbol{\mu}_{i-1|i-1}, \quad \boldsymbol{\Sigma}_{i|i-1} = \mathbf{Q}_i + \mathbf{M}_i\boldsymbol{\Sigma}_{i-1|i-1}\mathbf{M}_i^T, \quad \imath = 1, \dots, N$$

Activities 2(c)