

Data Assimilation: Estimation and Prediction (group meeting)

Sujit Ghosh and Mariana Pensky

Time and place: February 16, 2005, 3:30-5:00, Rm.# 203 SAMSI

Summary report:

The first meeting held on Feb 11th, 2005 at SAMSI generated several issues related to weather system *estimation and prediction* problems. One of the general problems that this group decided to consider is the following:

“Given that there are several computer model outputs generated by different weather forecasting agencies around the world, how do we efficiently combine these models forecasts and the observation collected from different sites to come up with better initial conditions for computer forecasts to be used by the computers in near future?”

Lenny and Shree discussed several general issues that need to be resolved statistically and also by using climatology. Others (mainly statisticians) tried to learn the the problem and offered to develop a statistical perspective of the general problem. However a general consensus was not reached!

During our second meeting we decided to explore issues that are formulated more specifically toward solving only a few aspects of the general problem of estimation and prediction. Even this turned out to be a fairly complicated and high-dimensional problems and there are several existing methods in the literature (Kalney, 2003, Chapters 1 and 5). So we started with a very simple framework and decided to generalize the setup as needed later.

A simplified setup:

In most practical settings, the computer forecasts are available on a regular grid (of approximately 10^6 to 10^7 grid points) and are denoted by the x . The observations obtained at various sites (significantly less than the number of grid points) within the grid are denoted by y^o

Let x^b denote the *background or prior forecasts* available at the current stage (which is usually based on climatology and previous observations). Suppose x^a denote the *analysis forecasts* that we'd like to determine based on the current observation y^o and background information x^b .

Research question: *How do we obtain the “best value” of x^a based on the current values of y^o and x^b ?*

A Bayesian framework:

$$\begin{aligned} X^a | X^b = x^b &\sim D_1(x^b, B) \\ Y^o | X^a = x^a &\sim D_2(H(x^a), R) \\ \text{which implies} &\quad \text{using Bayes rule} \\ X^a | Y^o = y^o, X^b = x^b &\sim D_3(?, ?) \end{aligned}$$

where D_1, D_2 are some location-scale family and D_3 denote some distribution that is determined based on D_1, D_2 . Based on some loss (cost) functions, we can determine the “best” value of the analysis, x^a , from the posterior distribution D_3 , which will be used as prior for next analysis.

Several questions now arise as a part of estimation and prediction:

1. What distributions should we choose to use for D_1 and D_2 ?
2. What *observation operator*, $H(\cdot)$ should we use to map model variables x 's to the observations y 's?
3. Should we estimate the scale matrices B and R ? If yes, how? If no, what are “good” values of B and R ?
4. Once we obtain the posterior distribution, D_3 , how do we come up with a summary value to be used as the “best” value of x^a ? In other word, what loss function should we choose to use for this purpose? This seems to be one of the most crucial step for the updating scheme that we develop here.

We briefly discussed about several options to address the questions 1 and 4 above. To address 1, it was proposed if we can use models that uses local neighborhoods in order to reduce the dimensionality of the problem, e.g., can we use Conditionally AutoRegressive (CAR) model for D_1 ? How about using wavelets to model the covariance structure? One possibility to address 4 is to use intrinsic loss functions, such Kullback-Liebler Information (KLI), so that the best predictions are equivariant under monotone transformations. In our future meetings we plan to address most of these issues.

Participants: Mariana Pensky, Sujit Ghosh, Minjung Kyung and Dave Holland.