

Data Assimilation: Estimation and Prediction (group meeting)

Sujit Ghosh and Mariana Pensky

Time and place: February 23, 2005, 3:30-5:00, Rm.# 203 SAMSI

Summary report:

In our meeting we mainly discussed the first topic on our list of questions:

What distribution should we choose to use for D_1 and D_2 .

E. Kalnay (2003) suggests to use normal distributions for both D_1 and D_2 . This, however, does not solve the problem completely, since the covariance structure of vector x_b is unknown and is impossible to estimate from observations. The reason for this is that x_b is $\sim 10^7$ -dimensional vector, so that B is $10^7 \times 10^7$ dimensional matrix, hence thousands of years of observations are necessary in order the sample covariance matrix is well-posed.

One of the ways to deal with this problem is to use *simultaneous autoregressive models (SAR)* or *conditionally autoregressive models (CAR)*.

To describe SAR models consider a set of random variables z_i in d -dimensional space, i.e. $i = (i_1, \dots, i_d)$. Define a set of simultaneous equations

$$z_i = \sum_{i \neq j} g_{i,j} z_j + \epsilon_i, \quad i = 1, \dots, N, \quad (1)$$

where $g_{i,j}$ are deterministic weights and ϵ_i are zero-mean random errors. Introducing vectors $z = (z_1, \dots, z_N)^T$ and $\epsilon = (\epsilon_1, \dots, \epsilon_N)^T$ and a matrix G with elements $g_{i,j}$, $i, j = 1, \dots, N$, we rewrite (1) as

$$(I - G)z = \epsilon. \quad (2)$$

If $\epsilon \sim N(0, \Sigma)$, then

$$z \sim N(0, (I - G)^{-1} \Sigma (I - G)^{-T}), \quad (3)$$

i.e. the covariance structure of z is completely determined by the weights G and Σ . Covariance matrix Σ is usually chosen to be diagonal. Note that if vector ϵ does not have a normal distribution, then Z is not normally distributed either.

The CAR models are very similar to SAR models described by (1): it is assumed that conditionally on $z_j, j \neq i$, random variables z_i are normally distributed

$$z_i | (z_j, j \neq i) \sim N\left(\sum_{i \neq j} g_{i,j} z_j, \sigma_i^2\right). \quad (4)$$

Under assumption (4) vector z is normally distributed with the mean zero and covariance matrix $(I - G)^{-1}\Sigma$ where $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_N)$. In CAR models, normality assumption is unavoidable and the weight matrix G is usually chosen to be symmetric.

We thought that SAR models provide more flexibility and may be a more appropriate way to define covariance structure of background vector x^b . In reality, vector x^b represents array of variables on a 3-dimensional grid. Define neighbors $N_{i,j,k}$ of $x_{i,j,k}$ as

$$N_{i,j,k} = \{(i', j', k') : |i - i'| \leq M_1, |j - j'| \leq M_2, |k - k'| \leq M_3\}.$$

Then SAR model is the set of simultaneous equations

$$x_{i,j,k}^b = \varrho \sum_{(i',j',k') \in N(i,j,k)} w_{i,j,k,i',j',k'} x_{i',j',k'}^b + \epsilon_{i,j,k} \quad (5)$$

where ϱ is the global weight (usually treated as unknown) and $w_{i,j,k,i',j',k'}$ are local weights which are supposed to be completely or partially known (may depend on a small number of unknown parameters). One of the possible choices is

$$w_{i,j,k,i',j',k'} = w(|i - i'|, |j - j'|, |k - k'|).$$

The errors $\epsilon_{i,j,k}$ in (5) are assumed to be independent for different indices and we assume that they are normally distributed with the variance inversely proportional to the number $n_{i,j,k}$ of neighbors of $x_{i,j,k}^b$:

$$\epsilon_{i,j,k} \sim N(0, \sigma^2/n_{i,j,k}). \quad (6)$$

Then arranging weights in the matrix form as W and using formula (3), we can derive that the covariance matrix B of x^b as

$$B = \sigma^2[(I - \varrho W)\Upsilon(I - \varrho W)]^{-1}.$$

Unknown parameter ϱ and various unknown parameters of matrix W can be then estimated from the data.

Another question which was discussed is the choice and properties of observation operator H . We are planning to study this question as well as SAR and CAR models more extensively.

Participants: Mariana Pensky, Sujit Ghosh and Dave Holland.