

Covariance localization in the approximated Karhunen-Loève basis

Le Duc (JAMSTEC), and Kazuo Saito (MRI/JMA)

Introduction

The forecast covariance plays an important role in data assimilation. The ensemble Kalman filter estimates this covariance matrix from N realizations of the atmospheric state

$$\widehat{cov}_{ij} = \frac{1}{N-1} \sum_k (x^k, e_i) (x^k, e_j) \quad (1)$$

where x^k is the k -th perturbation, e_i, e_j the canonical basis in which we usually estimate covariance ($e_i = (\delta_{ik}, k=1, n)$), and the symbol $\langle \cdot \rangle$ denotes the inner product. Under the assumption of the Gaussian distribution it can be proven that

$$E[(\widehat{cov}_{ij} - cov_{ij})^2] = \frac{1}{N} (cov_{ij}^2 + cov_{ii} cov_{jj}) \quad (2)$$

Clearly, the estimated errors of off-diagonal terms of a covariance matrix depend on the diagonal terms (variances). That means even if no correlations exist at distant points, the lower bounds of estimated correlations are constrained by variances.

The dependence of estimated errors of off-diagonal terms on variances is known as the sampling errors and localization is used to remove such spurious correlations. The common localization method is to use taper functions like Gaspari-Cohn functions. In this paper, we show that if localization is done in the basis that diagonalizes the covariance matrix (Karhunen-Loève basis) instead of the canonical basis, the estimated errors will reduce.

Localization in Karhunen-Loève basis

In fact covariance estimation is the estimation of a linear operator and we need a norm to evaluate the estimated error. Here we use the Frobenius norm for a linear operator T which assumes the matrix $\{t_{ij}\}$ in a basis.

$$\|T\|_F^2 = \sum_{ij} t_{ij}^2 \quad (3)$$

The estimated error now takes this form

$$E\|\widehat{cov} - cov\|_F^2 = \|E(\widehat{cov}) - cov\|_F^2 + E\|\widehat{cov} - E(\widehat{cov})\|_F^2$$

The terms in the right-hand side of Eq 4 are the bias and variance respectively. By using the estimation in Eq 1 the bias is always zero, however the sampling errors will occur in the variance term in Eq 4 as demonstrated in Eq 2.

To overcome the lower bounds of estimated errors for diagonal terms that are near-zero, we introduce the new estimation for such terms, which is equivalent to localization

$$\widehat{cov}_{ij}^e = \begin{cases} 0, & \text{if } cov_{ij} \approx 0 \ (ij \in K^c) \\ \widehat{cov}_{ij}, & \text{otherwise } (ij \in K) \end{cases} \quad (5)$$

with K^c is the set of index ij such that cov_{ij} is negligible. Apply Eq 3

for the new estimation, we have

$$E\|\widehat{cov}^e - cov\|_F^2 = \sum_{ij \in K^c} cov_{ij}^2 + \frac{1}{N} \sum_{ij \in K} (cov_{ij}^2 + cov_{ii} cov_{jj}) \geq \sum_{ij \in K^c} cov_{ij}^2 + \frac{2}{N} \sum_{ij \in K} cov_{ij}^2 = \frac{2}{N} \|cov\|_F^2 + \frac{N-2}{N} \sum_{ij \in K^c} cov_{ij}^2 \geq \frac{2}{N} \|cov\|_F^2 \quad (6)$$

Eq 6 proves that the estimated error attains its minimum value when the covariance operator has a diagonal form. In other words, the best estimation is obtained if we do localization in the Karhunen-Loève (KL) basis of the covariance operator. However the covariance is unknown and so its KL basis is

Approximated Karhunen-Loève basis

Apply the estimation in Eq 5 with the assumption that K is the diagonal of the covariance in any basis, Eq 6 becomes

$$E\|\widehat{cov}^e - cov\|_F^2 = \sum_{ij \in D^c} cov_{ij}^2 + \frac{1}{N} \sum_{ij \in D} (cov_{ij}^2 + cov_{ii} cov_{jj}) = \sum_{ij \in D^c} cov_{ij}^2 + \frac{2}{N} \sum cov_{ii}^2 = \|cov\|_F^2 - \frac{N-2}{N} \sum cov_{ii}^2 \quad (7)$$

That means we can approximate the KL basis with the basis that maximizes the sum of squared variances. We still do not know the true variances cov_{ii} . However we can estimate it through the sampling variances.

$$E[\sum cov_{ii}^2] = \frac{N+2}{N} \sum cov_{ii}^2 \quad (8)$$

Using a dictionary of bases like wavelets, we can select the best approximated KL basis. The test in 1-dimensional case (256 grid points, 50 ensemble members) with prescribed variances and correlations is depicted in Fig 1.

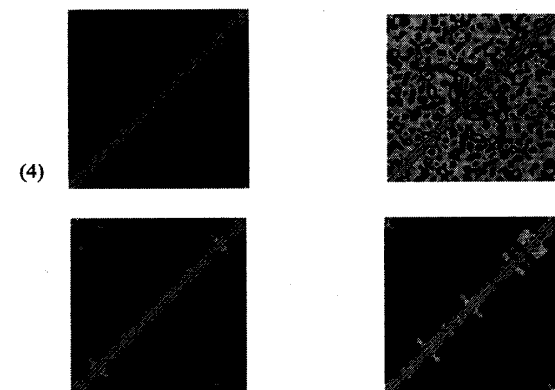


Fig. 1: True correlation (top left), and correlations estimated in the canonical basis (top right), in the KL basis (bottom left), in the approximated KL basis (bottom right) respectively.