Covariance matrix estimation and error model in D&A

Aurélien Ribes

CNRM - GAME, Météo France - CNRS

Buenos Aires, 16 October 2012

METEO FRANCE
Toujours un temps d'avance

cnrs

INSTITUT de MATHEMATIQUES de TOULOUSE

Statistical model

Statistical model

$$Y_\ell = \sum_{i=1}^{I} \beta_i X_\ell^{(i)} + \varepsilon_\ell \tag{1}$$

- ℓ : the location (spatio-temporal),
- Y : observations (spatio-temporal vector),
- β_i : scaling factors (real number), unknown,
- $X^{(i)}$: expected response to the i$^{\text{th}}$ forcing (spatio-temporal vector), known,
- ε : internal variability (spatio-temporal vector).

Statistical model

$$Y_\ell = \sum_{i=1}^{I} \beta_i X_\ell^{(i)} + \varepsilon_\ell \quad (1)$$

- ℓ : the location (spatio-temporal),
- Y : observations (spatio-temporal vector),
- β_i : scaling factors (real number), unknown,
- $X^{(i)}$: expected response to the ith forcing (spatio-temporal vector), known,
- ε : internal variability (spatio-temporal vector).

2 key assumptions

- (H1) : The distribution of ε is known (from climate models) [D&A],
- (H2) : $X^{(i)}$ are known (from climate models) [A].

Optimal fingerprint : estimation

2 key assumptions

- (H1) : The distribution of ε is known (from climate models) [D&A],
- (H2) : $X^{(i)}$ are known (from climate models) [A].

Let $C = \mathrm{Cov}(\varepsilon)$; C is known according to (H2).

$$\widehat{\beta} = (X' C^{-1} X)^{-1} X' C^{-1} Y, \tag{2}$$

Discussed issues

- How can we estimate C to approximate $\widehat{\beta}$?
- Can we deal with uncertainties in X ?

High dimension in climate datasets

Let us assume that climat model are able to perfectly simulate the climate system dynamics.

Typical climate dataset (e.g. near-surface temperature)

- Spatial dimension : $5^o \times 5^o \sim 2600$ grid-points,
- Temporal dimension : 50 - 100 ans (instrumental period),
- Dimension of $Y \sim 10^5$.
- Internal variability is described by $C \sim 10^5 \times 10^5$.
- The estimation of C requires *at least* 10^5 realisations of ε, i.e. 10^7 yrs of control simulations (vs about $\sim 10^4$ yrs available).

Some options :

- Decrease the dimension of Y,
- Look for an estimator of C *accurate* in large dimension.

Decreasing the dimension

Statistical investigation of climate at the global scale requires to reduce the spatio-temporal dimension of datasets.

- Decadal means,
- Projection on principal components,
- Projection on spherical harmonics (e.g. truncation T4, $\sim$ spatial scales $>$ 5000 kms),
- Use of simple climate indices (globale mean, land-sea contrast, inter-hemispheric contrast, annual cycle, etc).

Estimation of C in large dimension

Let us assume that $x_1, \ldots, x_n \sim N(0, C)$ are available for estimating C $(p \times p)$.

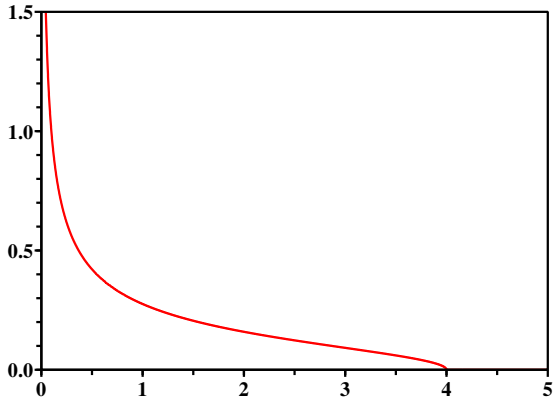
Estimation of C in large dimension

Let us assume that $x_1, \ldots, x_n \sim N(0, C)$ are available for estimating C $(p \times p)$.

What about $\widehat{C}$?

The sample estimate $\widehat{C}$ is a poor estimator of C in large dimension (n close to p).

Illustration : case $C = I$, distribution of the eigenvalues of $\widehat{C}$ when $n, p \to \infty$ (Marčenko-Pastur distribution).



Regularisation of C (1)
Principle

Principle

We use an estimator of C such as

$$\widetilde{C} = \gamma \widehat{C} + \rho I.$$

Regularisation of C (1)
Principle

Principle

We use an estimator of C such as

$$\widetilde{C} = \gamma \widehat{C} + \rho I.$$

LW estimate (Ledoit & Wolf, 2004)

- Introduction of estimators $\hat{\gamma}, \hat{\rho}$ of γ, ρ to minimise the mean square error

$$\mathrm{E}\left(\| \widetilde{C} - C \|_{\mathcal{M}}^2 \right).$$

- $$\widehat{C}_I = \hat{\gamma} \hat{C} + \hat{\rho} I.$$

Regularisation of C (1)
Principle

Principle

LW estimate (Ledoit & Wolf, 2004)

$$\widehat{C}_l = \hat{\gamma}\hat{C} + \hat{\rho}I.$$

New estimator

$$\widehat{\beta}_l = (X'\widehat{C}_l^{-1}X)^{-1}X'\widehat{C}_l^{-1}Y.$$

Regularisation of C (2)

Results : Mean square error

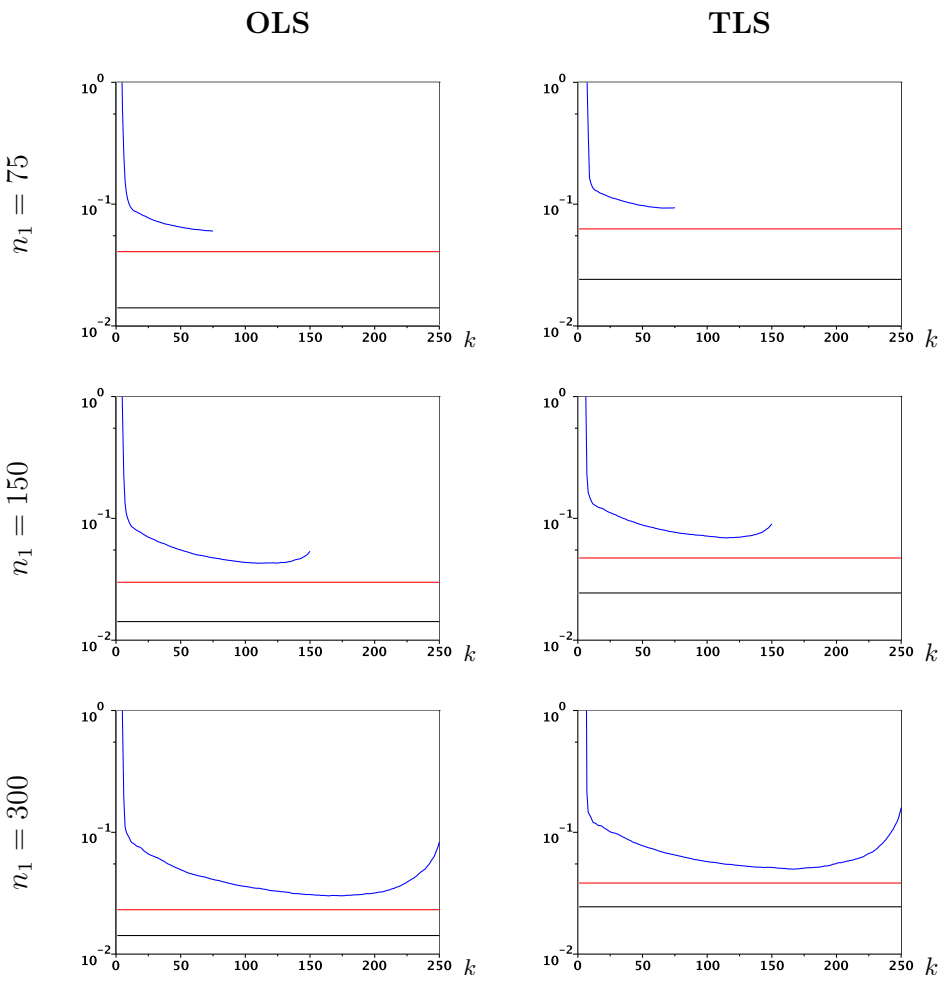


FIG. : Mean square error of β-estimates based on $\widehat{C}_l$ (red, regularisation), $\widehat{C}_q^+, q = 1 \dots p$ (blue, q-truncation) and C (black, perfect estimation). Estimation based on Monte-Carlo simulations, for three values of n ($p = 250$ here), and under OLS and TLS models.

Regularisation of C (3)

Resultats : Illustration

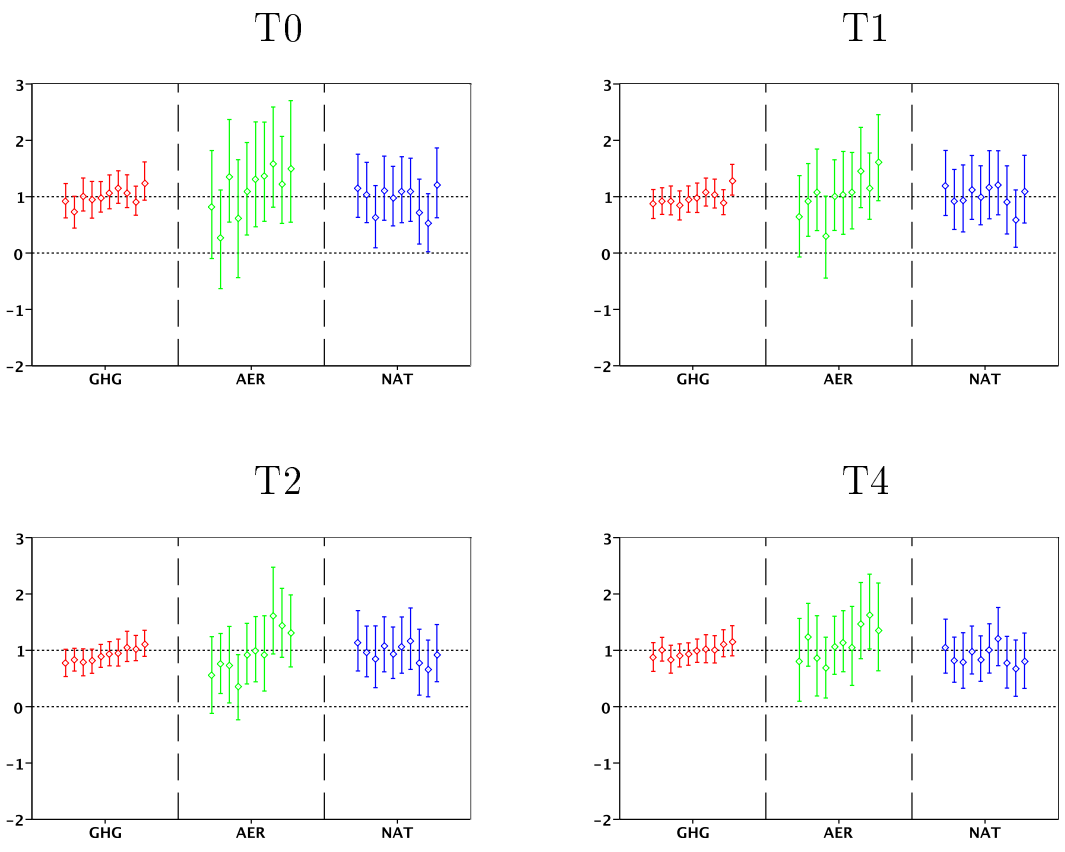


FIG. : Results obtained by regularising $\widehat{C}$ on pseudo-observations (historical simulations with the CNRM-CM5 climate model), based on different spatial resolutions in the pre-processing step (spherical harmonics T0, T1, T2 and T4).

Regularisation of C (3)

Resultats : Illustration

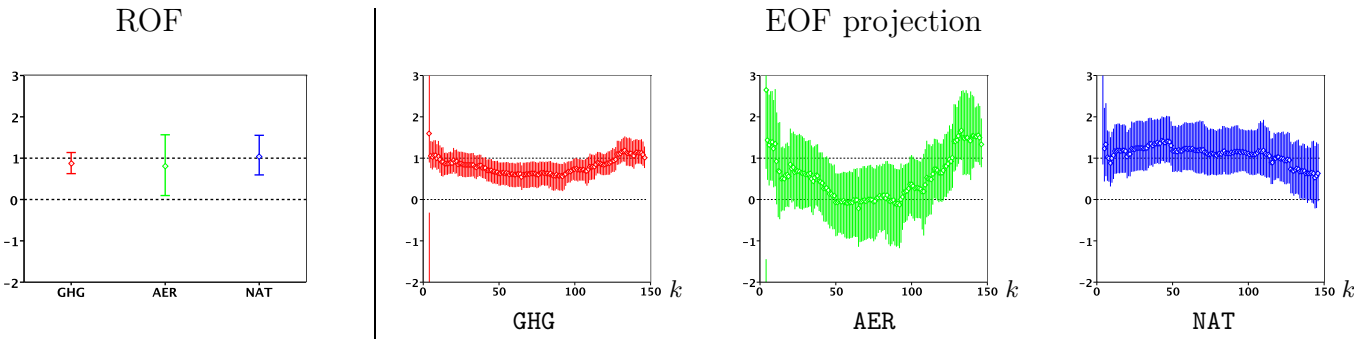


FIGURE 1: Sensitivity of the results to the choice k in the case of a projection on k principal components.

Questions

- How the LW estimate performs if $p \gg n$?
- Can we propose other way of regularising (eg with other priors than I) ?
- Can we use more recent results from random matrix theory ?

Key assumptions (recall)

2 key assumptions

- (H1) : The distribution of ε is known (from climate models) [D&A],
- (H2) : $X^{(i)}$ are known (from climate models) [A].

Model error in D&A primarily refers to the climate model error, i.e. error in X.

Internal varibility within the model

1. Problem :
   $X^{(i)}$ is estimated from climate model simulations, which do contain internal variability.
   What is actually observed is $X^{(i)} + \varepsilon_X^{(i)}$.

Internal varibility within the model

1. **Problem :**
   $X^{(i)}$ is estimated from climate model simulations, which do contain internal variability.
   What is actually observed is $X^{(i)} + \varepsilon_X^{(i)}$.

2. **Solution :**
   One may use a Total Least Squares algorithm instead of ordinary least squares (Allen & Stott, 2003).

TLS : Model, estimation

Revised statistical model (TLS)

Regression equation $Y_0 = X_0\beta$

One observes $\begin{cases} X = X_0 + \varepsilon_X \\ Y = Y_0 + \varepsilon_Y \end{cases}$

$$Cov(\varepsilon_X) = Cov(\varepsilon_Y) \quad (\text{or } Cov(\varepsilon_X) = Cov(\varepsilon_Y)/m)$$

TLS : Model, estimation

Revised statistical model (TLS)

Regression equation $Y_0 = X_0\beta$

One observes $\begin{cases} X = X_0 + \varepsilon_X \\ Y = Y_0 + \varepsilon_Y \end{cases}$

$$Cov(\varepsilon_X) = Cov(\varepsilon_Y) \quad (\text{or } Cov(\varepsilon_X) = Cov(\varepsilon_Y)/m)$$

- Alternative writing :

$$Y = (X - \varepsilon_X)\beta + \varepsilon_Y.$$

- Y and X play symetric roles,
- $\widehat{\beta}$, $\widehat{X_0}$ and $\widehat{Y_0}$ provided by the SVD of $[X, Y]$,
- Confidence intervals are not symetric in terms of β.

Modeling uncertainty (1)

1. **Problem :**
   Even if internal variability within the model is accounted for, $X^{(i)}$ is model-dependent (different models would provide different $X^{(i)}$).

2. **Solution :**
   Take into account the model discrepancies into the analysis : EIV methos (Errors In Variables, Huntigford & al., 2006).

Modeling uncertainty (1)

TLS model revised

Regression equation $$Y_0 = X_0 \beta$$

One observes $$\begin{cases} Y = Y_0 + \varepsilon_{Y,IV} \\ X = X_0 + \varepsilon_{X,IV} + \nu_{Mod} \end{cases}$$

With $$\begin{cases} \mathrm{Cov}(\varepsilon_{Y,IV}) = C, & \text{(I.V.)} \\ \mathrm{Cov}(\varepsilon_{X,IV} + \nu_{Mod}) = C + \Sigma_{Mod}, & \text{(I.V. + Mod. Uncert.)} \end{cases}$$

Modeling uncertainty (1)

TLS model revised

Regression equation $$Y_0 = X_0 \beta$$

One observes $$\begin{cases} Y = Y_0 + \varepsilon_{Y,IV} + \epsilon_{Obs} \\ X = X_0 + \varepsilon_{X,IV} + \nu_{Mod} \end{cases}$$

With $$\begin{cases} \mathrm{Cov}(\varepsilon_{Y,IV} + \epsilon_{Obs}) = C + \Sigma_{Obs}, & \text{(I.V. + Obs. Uncert.)} \\ \mathrm{Cov}(\varepsilon_{X,IV} + \nu_{Mod}) = C + \Sigma_{Mod}, & \text{(I.V. + Mod. Uncert.)} \end{cases}$$

Modeling uncertainty (1)

TLS model revised

Regression equation $$Y_0 = X_0\beta$$

One observes $$\begin{cases} Y = Y_0 + \varepsilon_{Y,IV} + \epsilon_{Obs} \\ X = X_0 + \varepsilon_{X,IV} + \nu_{Mod} \end{cases}$$

With $$\begin{cases} \text{Cov}(\varepsilon_{Y,IV} + \epsilon_{Obs}) = C + \Sigma_{Obs}, & \text{(I.V. + Obs. Uncert.)} \\ \text{Cov}(\varepsilon_{X,IV} + \nu_{Mod}) = C + \Sigma_{Mod}, & \text{(I.V. + Mod. Uncert.)} \end{cases}$$

- Estimation and hypothesis testing are more complicated than in TLS.
- Only a few studies have used this approach until now.

Modeling uncertainties (2)
Illustration

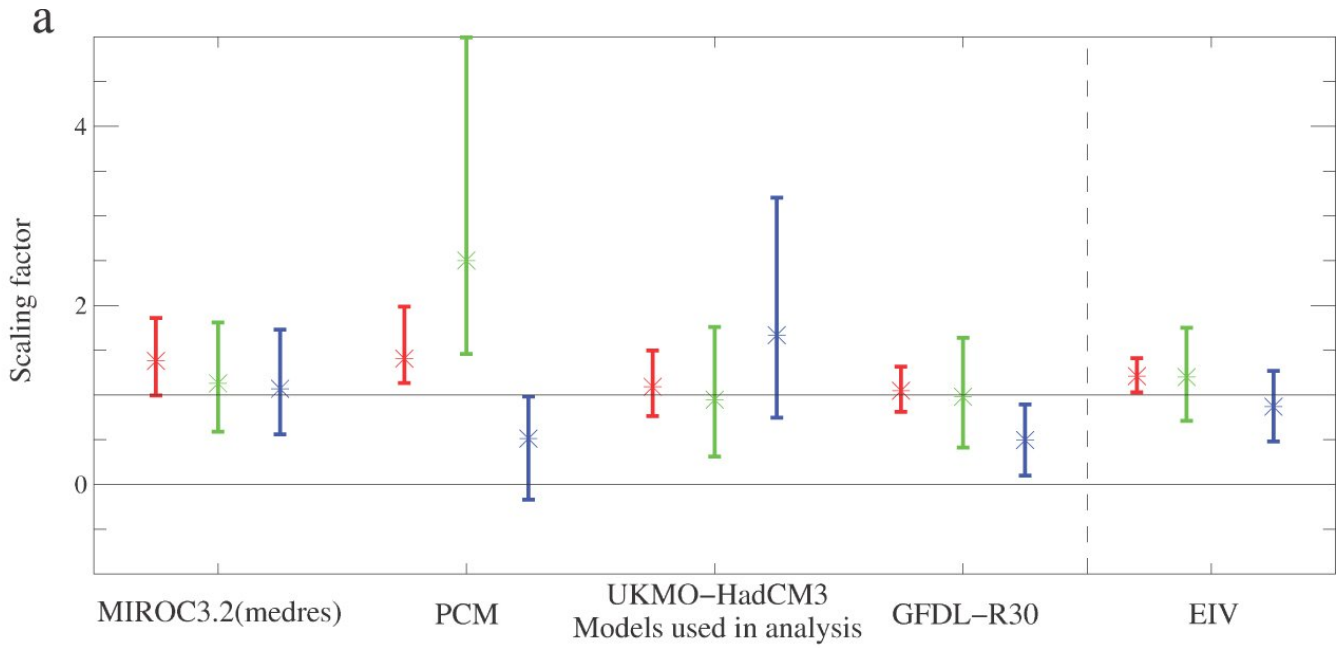


FIGURE 2: Illustration of the “EIV” method (IPCC, AR4, 2007)

Questions

- Can we improve the inference in EIV ?
- How can we estimate Σ_{Obs} and Σ_{Mod} ?
- ...

Conclusions

Covariance matrix estimation

- Estimation of large covariance matrix is required in D&A,
- The use of regularised estimates may help,
- Are such estimates potentially useful in DA ?

Model error

- Model error in D&A primarily refers to uncertainty on the climate response simulated by climate models,
- This error is typically taken into account in Errors In Variable regression models.