Statistical methods in the D&A of long-term changes.

Aurélien Ribes, CNRM, Météo France - CNRS

Banff, 14th June 2016

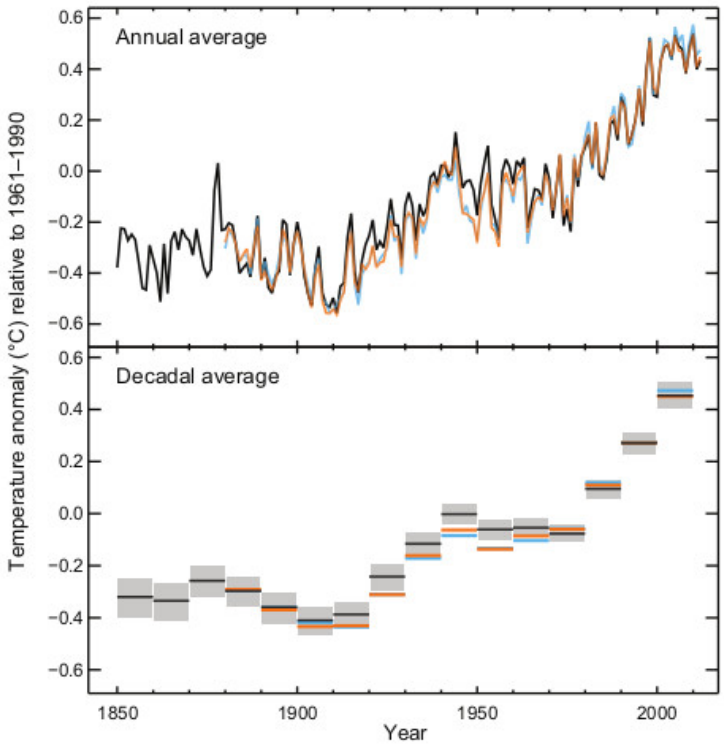
1. Introduction - Definitions

2. Statistical models and inference
   - OLS
   - TLS
   - EIV
   - No more regression

3. Common issues and challenges
   - Dimensionality
   - Estimating large covariance matrices
   - Estimating climate modeling uncertainty

4. Conclusion

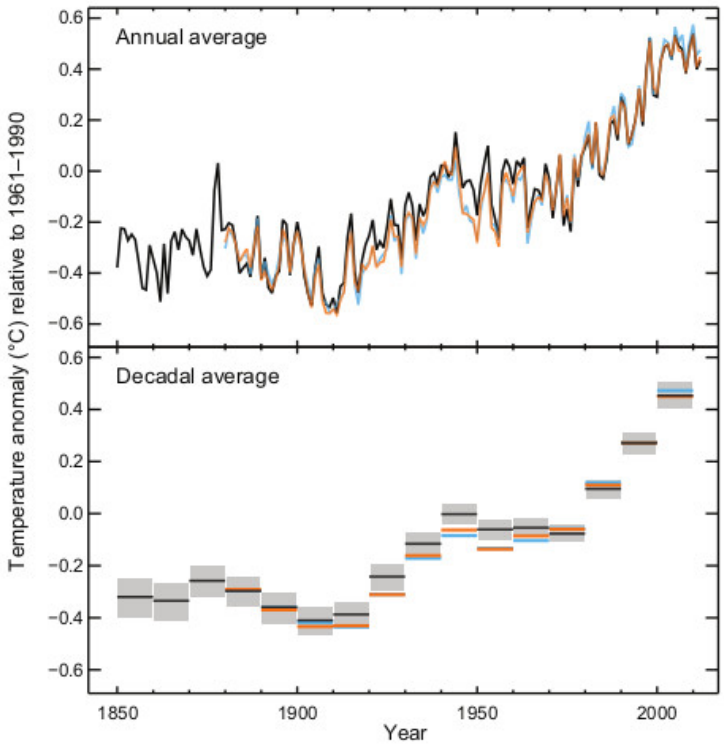
Introduction



Is there a change?

What are the causes?

Introduction



Is there a change?

Natural forcings

- Solar activity
- Volcanic eruptions
- Orbital variations (“astronomic”)
- ...

Anthropogenic forcings

- Greenhouse gases
- Aerosols
- Land use
- ...

Climate system

Atmosphere, Ocean, Cryosphere
Hydrosphere, Biosphere
& Interactions

+ Internal Variability

What are the causes?

Detection

D&A : assessment of the respective contributions from internal and forced variability.

Detection (of a change X) (personal rewording)

Demonstrating that (*the change*) X is not consistent with the internal variability only.

Detection requires to have some quantitative indication on the magnitude of internal variability (IV).

Attribution

Attribution (IPCC AR5)

"Evaluating the relative contributions of multiple causal factors to a change or event with an assignment of statistical confidence"

Attribution (of a change X to the cause Y) (IPCC AR4)

Demonstrating that (*the change*) X is :

- detectable,
- consistent with the expected response to the cause Y,
- not consistent with alternative, physically plausible explanations.

Requires some knowledge of the *expected responses* of the system to various forcings.

Key assumptions

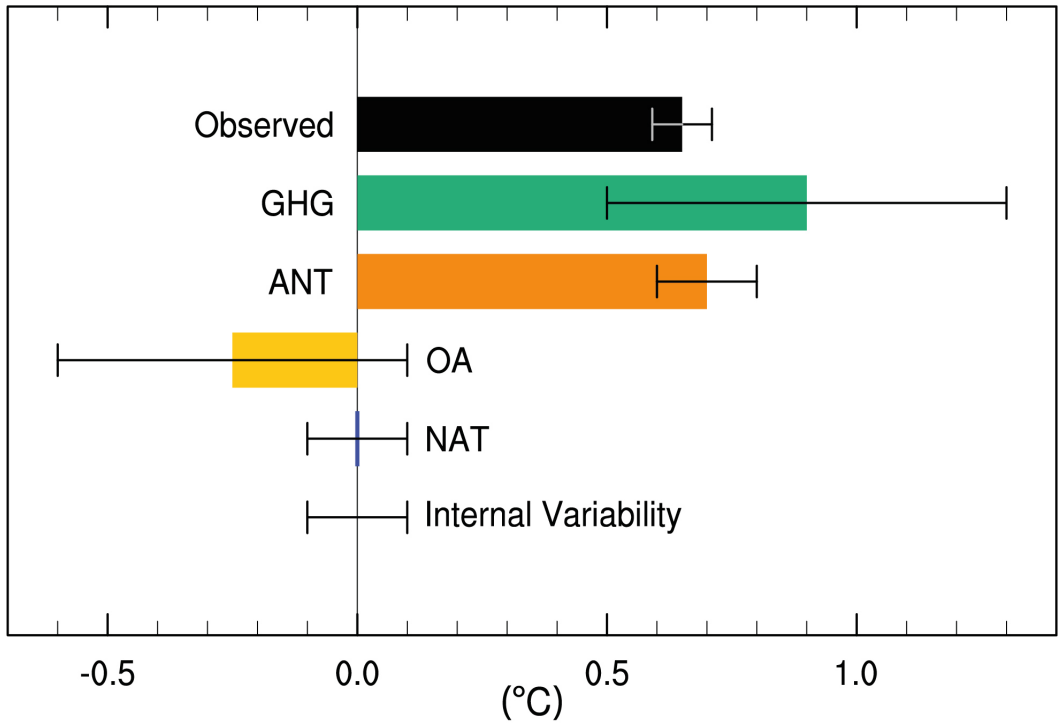
A couple of key assumptions in D&A

- (A1): The statistical distribution of the internal variability is known [D+A],
- (A2): The *expected responses* to each external forcing are known [A].

D&A involves

- Ideally: controled experiments on the climate system,
- In practice: a careful comparison of models and observations, in order to assess their consistency.

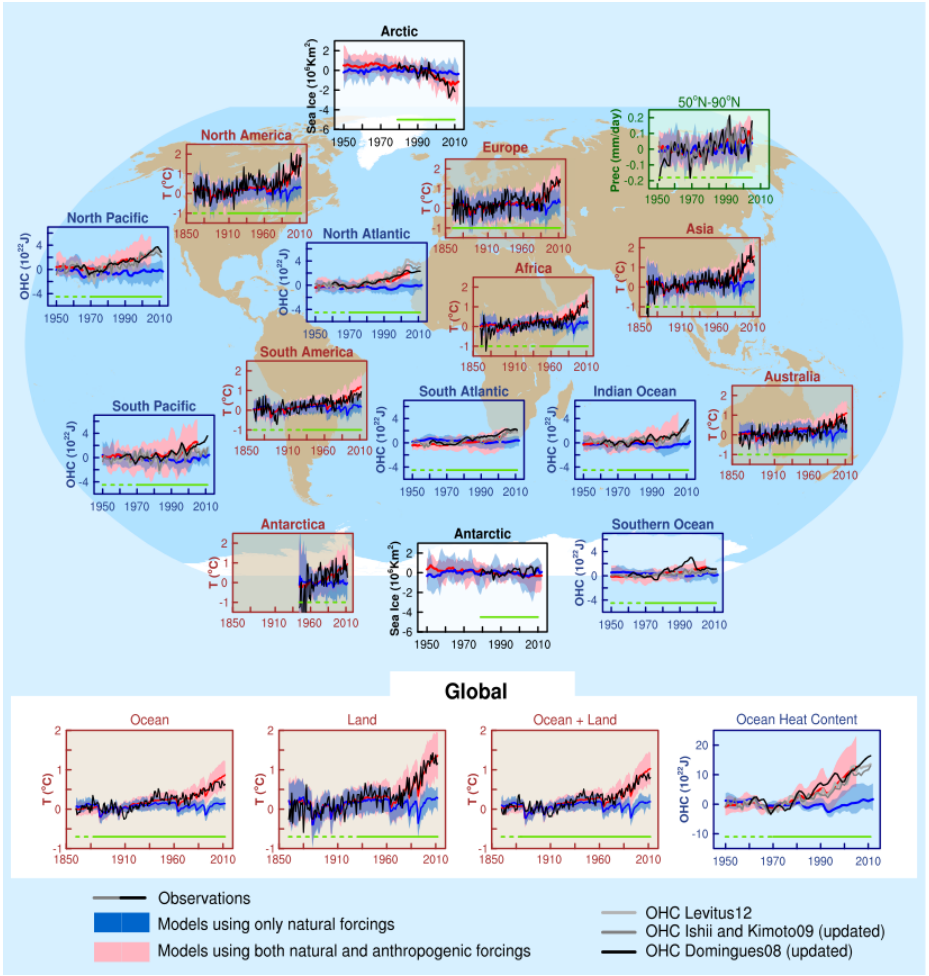
Broad Applications - Advertisement



GMT trend 1951-2010

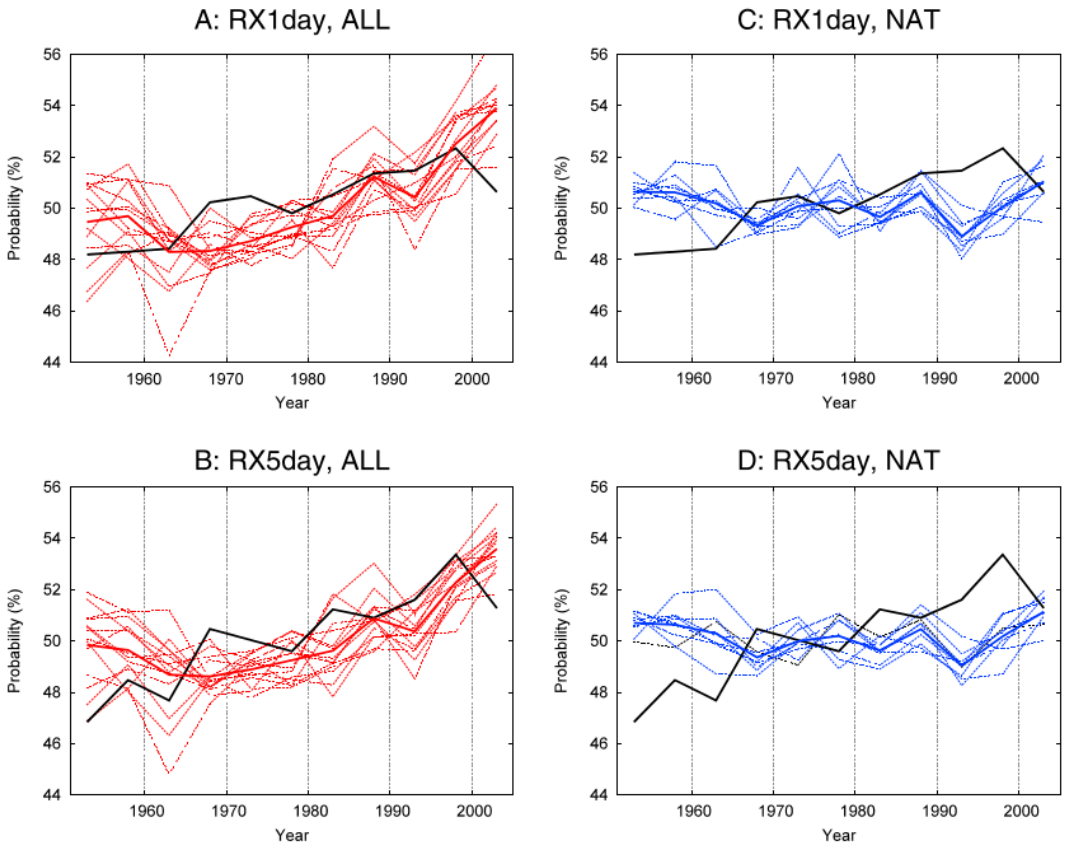
Fig 10.5, IPCC AR5, 2013

Broad Applications - Advertisement



Widespread human influence (IPCC AR5).

Broad Applications - Advertisement



Changes in extreme precipitation,
Source: Zhang et al. 2013; see also Min et al., 2011

1. Introduction - Definitions
2. Statistical models and inference
   - OLS
   - TLS
   - EIV
   - No more regression
3. Common issues and challenges
   - Dimensionality
   - Estimating large covariance matrices
   - Estimating climate modeling uncertainty
4. Conclusion

Regression based models

Philosophy: the response pattern is known, the magnitude is not.

$$Y_\ell = \sum_{i=1}^{k} \beta_i X_\ell^{(i)} + \varepsilon_\ell, \qquad Cov(\varepsilon) = \Sigma,$$

$$\underset{n\times 1}{Y} = \underset{n\times k,\ k\times 1}{X\ \beta} + \underset{n\times 1}{\varepsilon}, \qquad \underset{n\times n}{Cov(\varepsilon) = \Sigma},$$

- ℓ: location (space-time),
- Y: observations (space-time vector),
- β_i: scaling factor (scalar), unknown,
- $X^{(i)}$: expected response to forcing i (space-time vector), known,
- ε: internal variability (space-time vector),
- Σ: IV covariance matrix (matrix).

Attribution and hypothesis testing

Each step of the attribution process is related to some hypotheses testing (illustration here in a 2-forcing world).

Detection	"H_0": $(\beta_1, \beta_2) = (0, 0)$,
Forc. 1 only	"H_0": $\beta_2 = 0$,
Consistency	"H_0": $(\beta_1, \beta_2) = (1, 1)$,
	+ overall goodness-of-fit.

D&A requires

- Estimating β,
- Uncertainty analysis on β (i.e. confidence intervals),
- Testing goodness of fit.

OLS - Comparison to usual linear regression

Philosophy: The response patterns X are perfectly known.

$$\text{OLS}: \qquad Y = X\beta + \varepsilon, \qquad Cov(\varepsilon) = \Sigma,$$

σ^2 is (assumed to be) known

The residual consistency check is obtained by comparing (what would be) $\widehat{\sigma}^2$ to σ^2.

$\Sigma \neq I$

- If Σ known, then multiply by $\Sigma^{-1/2}$!
- $\widehat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$,
- The issue of how estimating Σ to efficiently approximate $\widehat{\beta}$ is uncommon.

OLS - Comparison to usual linear regression

Philosophy: The response patterns X are perfectly known.

$$\text{OLS :} \quad Y = X\beta + \varepsilon, \qquad Cov(\varepsilon) = \Sigma,$$
$$\text{Usually :} \quad Y = X\beta + \varepsilon, \qquad Cov(\varepsilon) = \sigma^2 I,$$

σ^2 is (assumed to be) known

The residual consistency check is obtained by comparing (what would be) $\widehat{\sigma}^2$ to σ^2.

$\Sigma \neq I$

- If Σ known, then multiply by $\Sigma^{-1/2}$!
- $\widehat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$,
- The issue of how estimating Σ to efficiently approximate $\widehat{\beta}$ is uncommon.

OLS model (Allen & Tett, 1999, and previous): Inference 1

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \Sigma), \qquad (\Sigma \text{ known})$$

Likelihood (-2 log-): $\ell_{\text{OLS}}(\beta) = (Y - X\beta)'\Sigma^{-1}(Y - X\beta).$

Estimation (optimal)	$\widehat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y,$
D&A Tests / CI	$\widehat{\beta} \sim N(\beta, (X'\Sigma^{-1}X)^{-1}),$
Goodness of fit	$\widehat{\varepsilon}'\Sigma^{-1}\widehat{\varepsilon} \sim_{H_0} \chi^2(n-k).$

OLS model (Allen & Tett, 1999, and previous): Inference 2

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \Sigma),$$

$\varepsilon_1, \ldots, \varepsilon_p$ iid $N(0, \Sigma)$ available to estimate Σ.

Derive 2 indep estimates $\widehat{\Sigma}_1$, $\widehat{\Sigma}_2$ ($\widehat{\Sigma}_1$ not necessarily sample estimate).

Estimation (optimal)	$\widehat{\beta} = (X'\widehat{\Sigma}_1^{-1}X)^{-1}X'\widehat{\Sigma}_1^{-1}Y,$
D&A Tests / CI	$\widehat{\beta} \sim N\left(\beta, (X'\widehat{\Sigma}_1^{-1}X)^{-1}X'\Sigma^{-1}X(X'\widehat{\Sigma}_1^{-1}X)^{-1}\right),$ $(\widehat{\beta}-\beta)'\left[(X'\widehat{\Sigma}^{-1}X)^{-1}X'\widehat{\Sigma}_2^{-1}X(X'\widehat{\Sigma}_1^{-1}X)^{-1}\right]^{-1}(\widehat{\beta}-\beta)$ $\sim kF(k, p_2),$
Goodness of fit	$\widehat{\varepsilon}'\widehat{\Sigma}_2^{-1}\widehat{\varepsilon} \sim_{H_0} \frac{p_2(n-k)}{p_2-n+1} F(n-k, p_2-n+1).$ (approximation).

1. Introduction - Definitions
2. Statistical models and inference
   - OLS
   - TLS
   - EIV
   - No more regression
3. Common issues and challenges
   - Dimensionality
   - Estimating large covariance matrices
   - Estimating climate modeling uncertainty
4. Conclusion

TLS model (Allen & Stott, 2003)

Philosophy: Taking into account the internal variability within the climate simulations leading to X.

$$Y = X^* \beta + \varepsilon, \quad \varepsilon \sim N(0, \Sigma) \tag{1}$$

$$X = X^* + \varepsilon_X, \quad \varepsilon_X \sim N(0, \Sigma / n_X) \tag{2}$$

TLS: model and likelihood

TLS model

Regression equation

$$Y^* = X^* \beta$$

One observes

$$\begin{cases} X = X^* + \varepsilon_X, & \mathrm{Cov}(\varepsilon_X) = \Sigma_X \\ Y = Y^* + \varepsilon_Y, & \mathrm{Cov}(\varepsilon_Y) = \Sigma_Y. \end{cases}$$

$$\Sigma_X = \lambda \Sigma_Y.$$

TLS: model and likelihood

TLS model

Regression equation $$Y^* = X^* \beta$$

One observes $$\begin{cases} X = X^* + \varepsilon_X, & \text{Cov}(\varepsilon_X) = \Sigma_X \\ Y = Y^* + \varepsilon_Y, & \text{Cov}(\varepsilon_Y) = \Sigma_Y. \end{cases}$$

$$\Sigma_X = \lambda \Sigma_Y.$$

Writing used by Allen & Stott (2003):

$$Y = (X - \varepsilon_X)\beta + \varepsilon_Y, \qquad \text{Cov}(\varepsilon_X) = \Sigma_X, \text{Cov}(\varepsilon_Y) = \Sigma_Y.$$

Misleading because:

- Suggests $X - \varepsilon_X \sim N(X, \Sigma_X)$, while $X - \varepsilon_X = X^*$,
- At least, must add $Cov(X, \varepsilon_X) = \Sigma_X$! (or better $\varepsilon_X | X = X^* - X$).
- X^* is also of interest !

TLS: model and likelihood

TLS model

Regression equation

$$Y^* = X^* \beta$$

One observes

$$\begin{cases} X = X^* + \varepsilon_X, & \mathrm{Cov}(\varepsilon_X) = \Sigma_X \\ Y = Y^* + \varepsilon_Y, & \mathrm{Cov}(\varepsilon_Y) = \Sigma_Y. \end{cases}$$

$$\Sigma_X = \lambda \Sigma_Y.$$

TLS: model and likelihood

TLS model

Regression equation $$Y^* = X^*\beta$$

One observes $$\begin{cases} X = X^* + \varepsilon_X, & \mathrm{Cov}(\varepsilon_X) = \Sigma_X \\ Y = Y^* + \varepsilon_Y, & \mathrm{Cov}(\varepsilon_Y) = \Sigma_Y. \end{cases}$$

$$\Sigma_X = \lambda \Sigma_Y.$$

$$\ell_{\mathrm{TLS}}(\beta, X^*) = (Y - X^*\beta)'\Sigma_Y^{-1}(Y - X^*\beta) + (X - X^*)'\Sigma_X^{-1}(X - X^*).$$

-> Geometry

TLS: model and likelihood

TLS model

Regression equation $$Y^* = X^*\beta$$

One observes $$\begin{cases} X = X^* + \varepsilon_X, & \text{Cov}(\varepsilon_X) = \Sigma_X \\ Y = Y^* + \varepsilon_Y, & \text{Cov}(\varepsilon_Y) = \Sigma_Y. \end{cases}$$

$$\Sigma_X = \lambda \Sigma_Y.$$

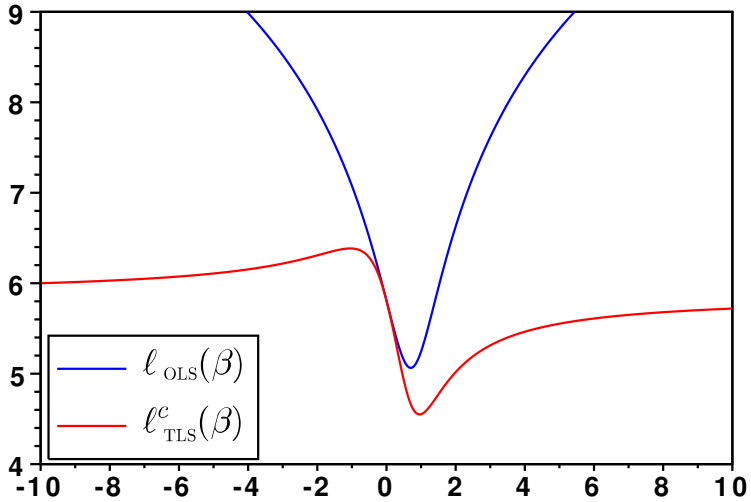
$$\ell_{\text{TLS}}(\beta, X^*) = (Y - X^*\beta)'\Sigma_Y^{-1}(Y - X^*\beta) + (X - X^*)'\Sigma_X^{-1}(X - X^*).$$

-> Geometry

$$\ell^c_{\text{TLS}}(\beta) = \frac{(Y - X\beta)'\Sigma^{-1}(Y - X\beta)}{1 + \beta^2} \qquad \text{(assuming } \lambda = 1\text{)}.$$

TLS: Likelihood

$$\ell^c_{\text{TLS}}(\beta) = \frac{(Y - X\beta)'\Sigma^{-1}(Y - X\beta)}{1 + \beta^2}. \tag{3}$$



If you assume the wrong model

- OLS is "the truth": $\widehat{\beta}_{TLS}$ is not optimal (too much variance),
- TLS is "the truth": $\widehat{\beta}_{OLS}$ is biased toward 0.

TLS model: Inference

$$[X, Y] = [X^*, X^*\beta] + [\varepsilon_X, \varepsilon_Y],$$

Estimation (optimal)

Explicit for $\widehat{\beta}$, $\widehat{X}^*$ and $\widehat{Y}^*$ from,

$$\text{SVD of } [X, Y] = U\Lambda V'.$$

D&A Tests and Goodness of fit: asymptotic results

- Approximated assuming $\lambda_k \gg \lambda_{k+1}$ (ie high s/n),
- e.g. $\widehat{v}' V(\Lambda^2 - \lambda_{k+1}^2 I) V' \widehat{v} \sim \chi_k^2$ (AS03, no proof),
- Mainly too small CI, too permissive RCT (too often accepted).

TLS: attributable trend

TLS model

Regression equation $$Y^* = X^* \beta$$

One observes $$\left\{ \begin{array}{ll} X = X^* + \varepsilon_X, & \mathrm{Cov}(\varepsilon_X) = \Sigma_X \\ Y = Y^* + \varepsilon_Y, & \mathrm{Cov}(\varepsilon_Y) = \Sigma_Y. \end{array} \right.$$

To assess the contribution of the forcing to a change:

- We usually consider: $[X\widehat{\beta}_{inf}, X\widehat{\beta}_{sup}]$,
- We should consider: $[(\widehat{X}^*\widehat{\beta})_{inf}, (\widehat{X}^*\widehat{\beta})_{sup}]$, (but difficult to estimate),

1. Introduction - Definitions
2. Statistical models and inference
   - OLS
   - TLS
   - EIV
   - No more regression
3. Common issues and challenges
   - Dimensionality
   - Estimating large covariance matrices
   - Estimating climate modeling uncertainty
4. Conclusion

EIV model (Huntingford et al., 2006; Hannart et al., 2014)

Philosophy: Taking into account the modeling uncertainty (ie different climate models provide different patterns X or X^*).

Regression equation $$Y^* = X^* \beta$$

One observes $$\begin{cases} Y = Y^* + \underbrace{\varepsilon_{Y,IV}}_{\varepsilon_Y} \\ X = X^* + \underbrace{\varepsilon_{X,IV}}_{\varepsilon_X} \end{cases}$$

With $$\begin{cases} \mathrm{Cov}(\varepsilon_Y) = \Sigma_Y = \Sigma_{IV}, & \text{(I.V.)} \\ \mathrm{Cov}(\varepsilon_X) = \Sigma_X = \Sigma_{IV}, & \text{(I.V.)} \end{cases}$$

EIV model (Huntingford et al., 2006; Hannart et al., 2014)

Philosophy: Taking into account the modeling uncertainty (ie different climate models provide different patterns X or X^*).

Regression equation $$Y^* = X^* \beta$$

One observes $$\begin{cases} Y = Y^* + \underbrace{\varepsilon_{Y,IV}}_{\varepsilon_Y} \\ X = X^* + \underbrace{\varepsilon_{X,IV} + \varepsilon_{Mod}}_{\varepsilon_X} \end{cases}$$

With $$\begin{cases} \mathrm{Cov}(\varepsilon_Y) = \Sigma_Y = \Sigma_{IV}, & \text{(I.V.)} \\ \mathrm{Cov}(\varepsilon_X) = \Sigma_X = \Sigma_{IV} + \Sigma_{Mod}, & \text{(I.V. + Mod. Uncert.)} \end{cases}$$

EIV model (Huntingford et al., 2006; Hannart et al., 2014)

Philosophy: Taking into account the modeling uncertainty (ie different climate models provide different patterns X or X^*).

Regression equation $$Y^* = X^* \beta$$

One observes $$\begin{cases} Y = Y^* + \underbrace{\varepsilon_{Y,IV} + \varepsilon_{Obs}}_{\varepsilon_Y} \\ X = X^* + \underbrace{\varepsilon_{X,IV} + \varepsilon_{Mod}}_{\varepsilon_X} \end{cases}$$

With $$\begin{cases} \mathrm{Cov}(\varepsilon_Y) = \Sigma_Y = \Sigma_{IV} + \Sigma_{Obs}, & \text{(I.V. + Obs. Uncert.)} \\ \mathrm{Cov}(\varepsilon_X) = \Sigma_X = \Sigma_{IV} + \Sigma_{Mod}, & \text{(I.V. + Mod. Uncert.)} \end{cases}$$

EIV model

Regression equation $$Y^* = X^* \beta$$

Observations $$\begin{cases} Y = Y^* + \varepsilon_Y, & \text{Cov}(\varepsilon_Y) = \Sigma_Y, \\ X = X^* + \varepsilon_X, & \text{Cov}(\varepsilon_X) = \Sigma_X. \end{cases}$$

Σ_Y and Σ_X have no relationship.

EIV: Likelihood

Regression equation $$Y^* = X^* \beta$$

Observations $$\begin{cases} Y = Y^* + \varepsilon_Y, & \text{Cov}(\varepsilon_Y) = \Sigma_Y, \\ X_i = X_i^* + \varepsilon_{X,i}, & \text{Cov}(\varepsilon_{X,i}) = \Sigma_{X,i}. \end{cases}$$

Assuming $k = 1$,

$$\begin{aligned} \ell(\beta, X^*) &= cte + (Y - X^*\beta)' \Sigma_Y^{-1} (Y - X^*\beta) + (X - X^*)' \Sigma_X^{-1} (X - X^*). \\ \ell_c(\beta) &= cte + (Y - X\beta)' (\beta^2 \Sigma_X + \Sigma_Y)^{-1} (Y - X\beta). \end{aligned}$$

No explicit maximum (i.e. MLE) !

Estimation (Hannart et al., 2014)

Philosophy: maximize $\ell(\beta, X^*)$ with a numerical algorithm (MLE).

Estimation: algorithm

- 1. Fix $\widehat{\beta}^0$,
- 2i. Compute $\widehat{X}_i^* = \text{Argmax}_{X^*}\, \ell(\widehat{\beta}^{i-1}, X^*)$,
- 3i. Compute $\widehat{\beta}_i = \text{Argmax}_{\beta}\, \ell(\beta, \widehat{X}_i^*)$,
- 4. When convergence occurs, you have $(\widehat{\beta}, \widehat{X^*})$.

Issue(s): may converge to a critical point or local maximum (not necessarily the global maximum, i.e. MLE).

Confidence intervals

Use asymptotic property of MLE.

Issue(s): too low coverage probability (i.e.: too small CI).

EIV model: Inference

Assuming Σ_X, Σ_Y known

Estimation (optimal)	Non-explicit (MLE algo, Hannart et al., 2014),
D&A Tests / CI	Approximated (too small CI),
Goodness of fit	??.

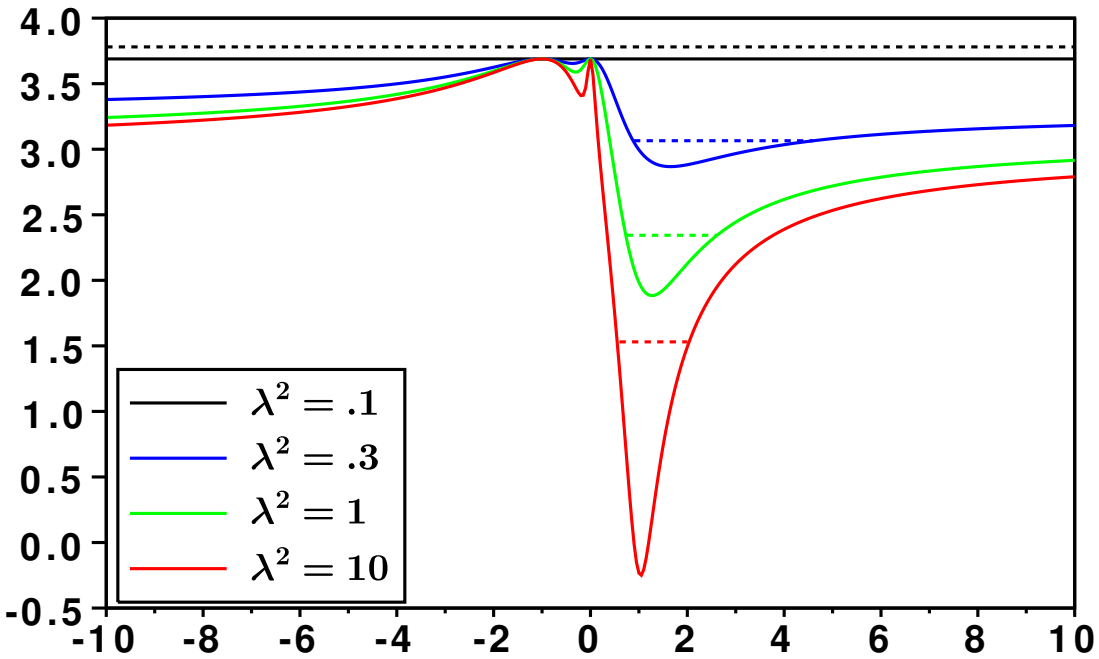
If Σ_X, Σ_Y are estimated

? ? ? ?

EIV: funny property

Using EIV instead of TLS may reduce uncertainty !

-> Toy exemple



1. Introduction - Definitions

2. Statistical models and inference
   - OLS
   - TLS
   - EIV
   - No more regression

3. Common issues and challenges
   - Dimensionality
   - Estimating large covariance matrices
   - Estimating climate modeling uncertainty

4. Conclusion

Motivation: Is linear regression suitable?

$$Y = \sum_{i=1}^{k} \beta_i X_i + \varepsilon$$

- Predominantly used for about 2 decades
- Assumes that
  - models are able to simulate response patterns,
  - response magnitudes are unknown.

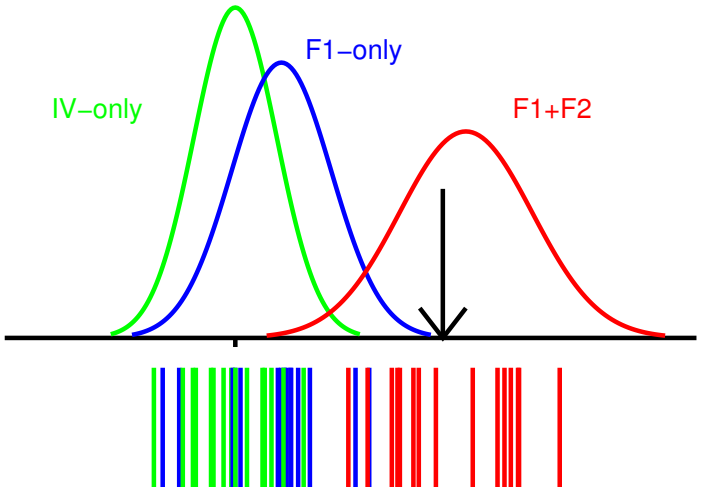
Motivation: Is linear regression suitable?

$$Y = \sum_{i=1}^{k} \beta_i X_i + \varepsilon$$

- Predominantly used for about 2 decades
- Assumes that
  - models are able to simulate response patterns,
  - response magnitudes are unknown.
- The reality is probably more balanced
  - Large uncertainty in the response magnitude (e.g. sensitivity), but also in the spatial response pattern (e.g. land sea warming ratio, amplitude of the Arctic amplification),
  - Unknown / Uncertain feedbacks are likely to modify spatial response pattern (e.g. the cloud feedback).

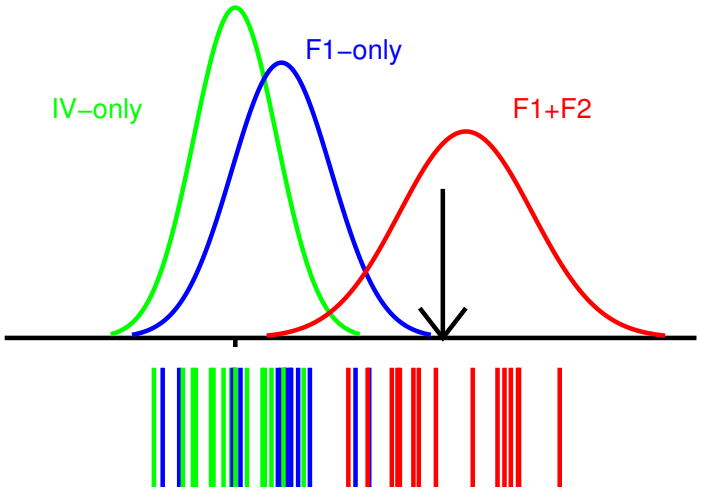
Wish to not discard the available physical knowledge on the response magnitudes.

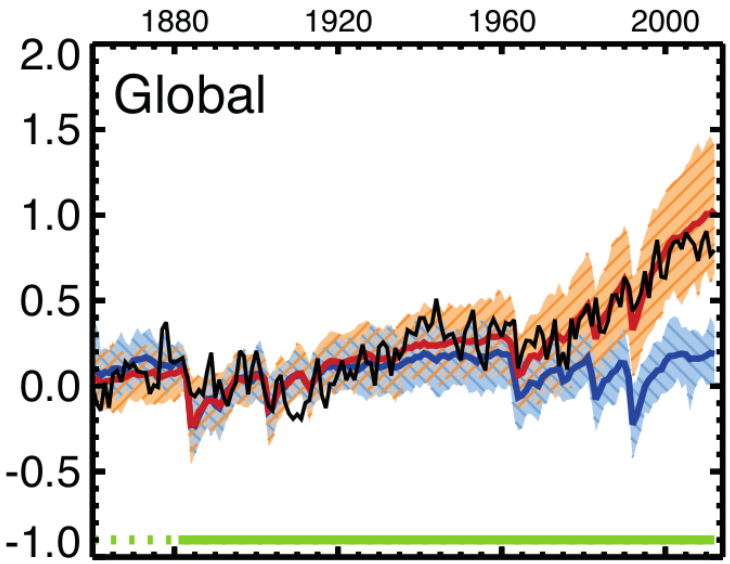
Possibility to apply D&A to single scalar variables



- Detection: inconsistency with IV-only,
- Attribution (1): consistency with F1+F2,
- Attribution (2): inconsistency with F1-only.

Possibility to apply D&A to single scalar variables





- Detection: inconsistency with IV-only,
- Attribution (1): consistency with F1+F2,
- Attribution (2): inconsistency with F1-only.

The new approach

$$\begin{cases} Y = Y^* + \varepsilon_Y, & \varepsilon_Y \sim N(0, \Sigma_Y), \\ X_i = X_i^* + \varepsilon_{X_i}, & \varepsilon_{X_i} \sim N(0, \Sigma_{X_i}), \quad i = 1, \ldots, k, \end{cases}$$

The new approach

$$Y^* = \sum_{i=1}^{k} X_i^*,$$

$$\begin{cases} Y = Y^* + \varepsilon_Y, & \varepsilon_Y \sim N(0, \Sigma_Y), \\ X_i = X_i^* + \varepsilon_{X_i}, & \varepsilon_{X_i} \sim N(0, \Sigma_{X_i}), \quad i = 1, \ldots, k, \end{cases}$$

The new approach

$$Y^* = \sum_{i=1}^{k} X_i^*,$$

$$\begin{cases} Y = Y^* + \varepsilon_Y, & \varepsilon_Y \sim N(0, \Sigma_Y), \\ X_i = X_i^* + \varepsilon_{X_i}, & \varepsilon_{X_i} \sim N(0, \Sigma_{X_i}), \quad i = 1, \ldots, k, \end{cases}$$

- Use identical assumptions, but remove the βs,Â
  response's magnitude and pattern are treated consistently
- Inference focuses on X_i^* (instead of β_i),
- Main assumption: additivity,
- Interpretation: models give information on each term X_i^*, then an additional constraint on their sum comes from observations.
- All inference can be made with maximum likelihood

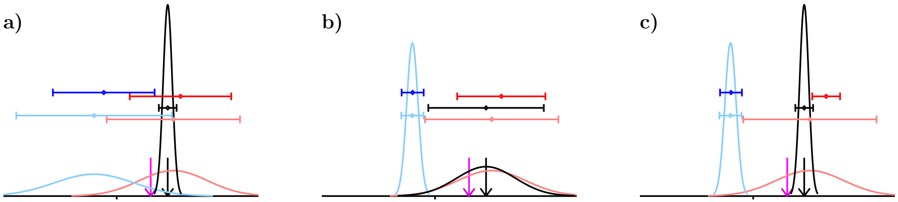
$$\widehat{X}_i^* = X_i + \Sigma_{X_i}(\Sigma_Y + \Sigma_X)^{-1}(Y - X) \sim N(X_i, \Sigma_{\widehat{X}_i^*}).$$

Comparing linear regression with this method

Linear Regression (EIV)	This method
• knowledge on magnitude ignored	• magnitude and pattern treated consistently
• estimators are non explicit and difficult to compute	• explicit estimators
• approximated CI on β, no CI on βX^* (attrib. trend), $([\widehat{\beta}_{inf} X, \widehat{\beta}_{sup} X] \neq [(\widehat{\beta X^*})_{inf}, (\widehat{\beta X^*})_{sup}])$	• exact CI.

How does this work (for scalars)?

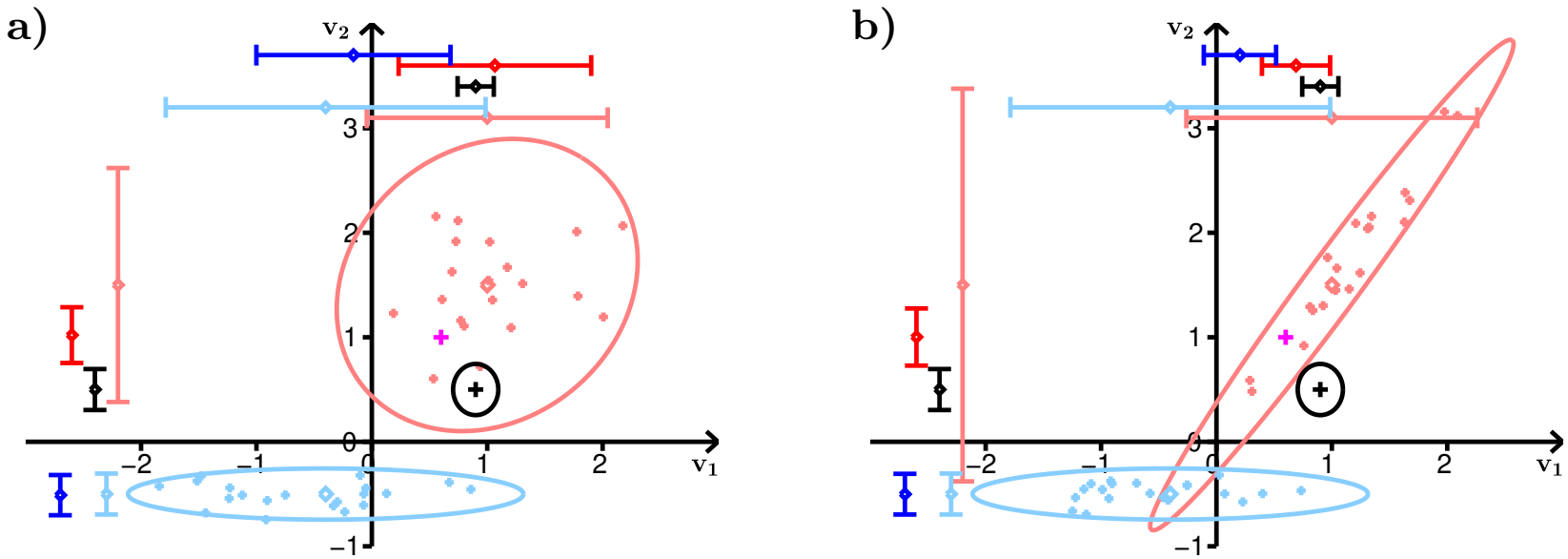
The method is efficient if all terms but one are well constrained



- a) large uncertainty in both F1 and F2: little gain.
- b) large uncertainty in both F1 and obs: little gain.
- c) limited uncertainty in both obs and F2: substantial gain on F1.

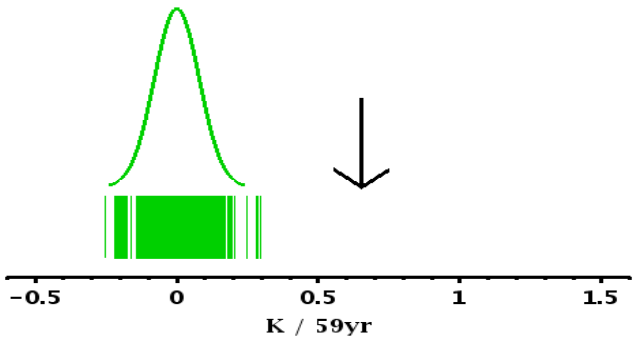
How does this work (2-D)?

The dependence structures $\Sigma_{X_1}, \Sigma_{X_2}$ influence the results

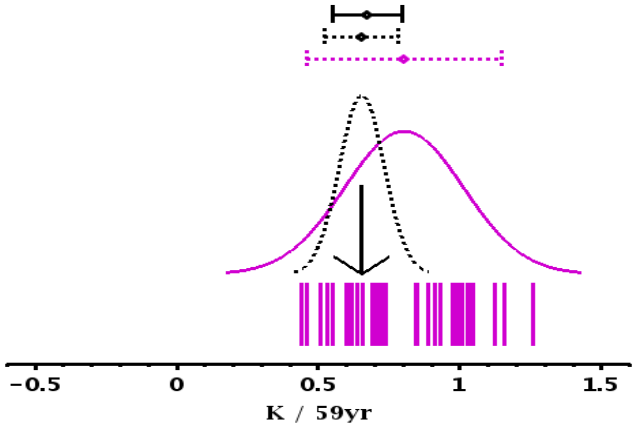


- a) Variables v_1 and v_2 are independent: weak obs. constrain,
- b) Strong dependence between v_1 and v_2: strong obs. constrain.

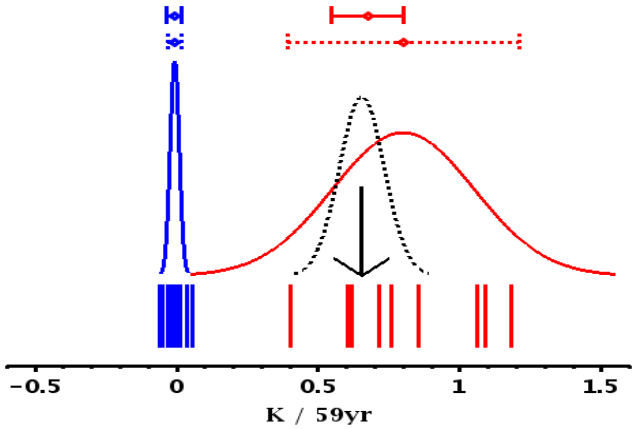
Analysis of the observed 1951-2010 GMT linear trend



Detection step



Consistency with all forcings



Attribution to ANT / NAT

Obs warming: +.65K,

ALL-induced: +.67K [+.55K,+.79K],

NAT-induced: -.01K [-.03K,+.02K],

ANT-induced: +.67K [+.55K,+.80K],

(consistent with Fig 10.5)

1. Introduction - Definitions

2. Statistical models and inference
   - OLS
   - TLS
   - EIV
   - No more regression

3. Common issues and challenges
   - Dimensionality
   - Estimating large covariance matrices
   - Estimating climate modeling uncertainty

4. Conclusion

High dimension in climate datasets

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 $\Sigma \sim 10^5 \times 10^5$.
- The estimation of Σ requires *at least* 10^5 realisations of ε, i.e. 10^7 yrs of control simulations (vs about $\sim 10^4$ yrs available).

High dimension in climate datasets

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 $\Sigma \sim 10^5 \times 10^5$.
- The estimation of Σ 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 Σ *accurate* in large dimension.

Decreasing the dimension (or pre-processing)

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

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

- Projection on EOFs,

This treatment is quite arbitrary and non optimal.

Problem statement

- Most inference methods assume that Σ is known.
  (and the full distribution of the internal variability ε).
- Usually, climate models are used to derive a few realisations of ε, say $\varepsilon_1, \ldots, \varepsilon_p$.
- The distribution, or at least, $\Sigma = \mathrm{Cov}(\varepsilon)$ is estimated from these.
- Optimal statistics requires to estimate Σ^{-1}
  (eg $\widehat{\beta}_{OLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$).

This is very challenging in high dimension.

Estimation of Σ in large dimension

Let us assume that $\varepsilon_1, \ldots, \varepsilon_n \sim N(0, \Sigma)$ are available for estimating Σ ($p \times p$).

Estimation of Σ in large dimension

Let us assume that $\varepsilon_1, \ldots, \varepsilon_n \sim N(0, \Sigma)$ are available for estimating Σ ($p \times p$).

What about $\widehat{\Sigma}$?

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

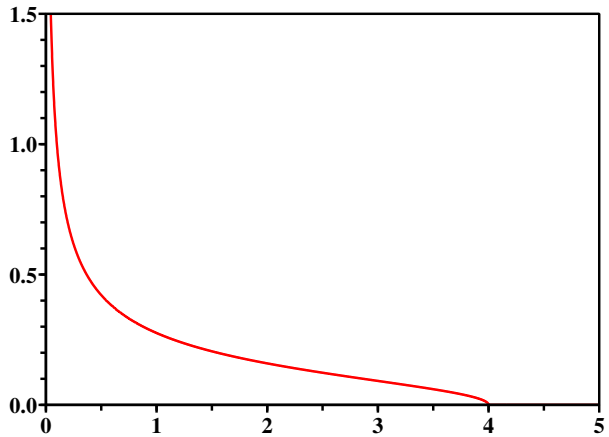
Estimation of Σ in large dimension

Let us assume that $\varepsilon_1, \ldots, \varepsilon_n \sim N(0, \Sigma)$ are available for estimating Σ ($p \times p$).

What about $\widehat{\Sigma}$?

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

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



EOF projections

EOF proj estimate $\widehat{\beta}_q^+$

$$\widehat{\beta}_q = (X'\widehat{\Sigma}_q^+ X)^{-1} X'\widehat{\Sigma}_q^+ Y.$$

$\widehat{\Sigma}_q^+ = P\,\mathrm{diag}(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_q}, 0, \ldots, 0)P'$ with $\widehat{\Sigma} = P\,\mathrm{diag}(\lambda_1, \ldots, \lambda_p)P'$, and $q < p$.

- There's no optimality result regarding the choice of k, and it may impact the results.

Regularising Σ (1)

Principle

Principle

We use an estimator of Σ such as

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

Regularising Σ (1)

Principle

Principle

We use an estimator of Σ such as

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

LW estimate (Ledoit & Wolf, 2004)

- Introduction of estimators $\hat{\gamma}, \hat{\rho}$ of γ, ρ to minimise the mean square error
$$\mathrm{E}\left(\|\widetilde{\Sigma} - \Sigma\|^2_{\mathcal{M}}\right).$$
- $$\widehat{\Sigma}_I = \hat{\gamma}\hat{\Sigma} + \hat{\rho} I.$$

Regularising Σ (1)

Principle

Principle

We use an estimator of Σ such as

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

LW estimate (Ledoit & Wolf, 2004)

- Introduction of estimators $\hat{\gamma}, \hat{\rho}$ of γ, ρ to minimise the mean square error
$$\mathrm{E}\left(\|\widetilde{\Sigma} - \Sigma\|_{\mathcal{M}}^2\right).$$
- $$\widehat{\Sigma}_I = \hat{\gamma}\hat{\Sigma} + \hat{\rho} I.$$

New estimator (Ribes et al., 2009)

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

Regularising Σ (2)

Results: Mean square error

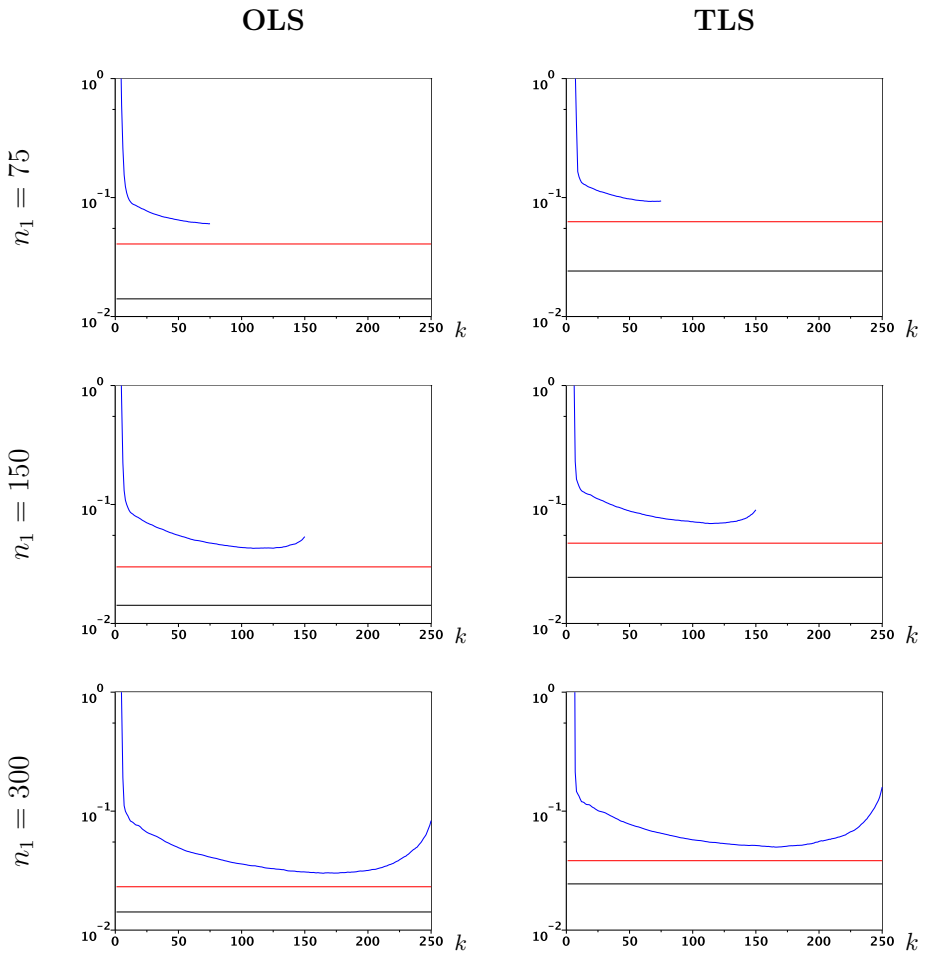


FIG.: Mean square error of β-estimates based on $\widehat{\Sigma}_l$ (red, regularisation), $\widehat{\Sigma}_q^+, q = 1 \dots p$ (blue, q-truncation) and Σ (black, perfect estimation). Estimation based on Monte-Carlo simulations, for three values of n ($p = 250$ here), and under OLS and TLS models.

By the way:
EOF projection is not optimal!

(Ribes et al., 2013)

Integrated Optimal Fingerprinting approach

- Regularization with a target $\Delta \neq I$ (Hannart et Naveau, 2014, JMVA).

  Use of a Bayesian prior: $\Sigma \sim \mathcal{W}^{-1}(\Delta, \alpha)$ (centered on Δ),

  Derive estimators $\widehat{\rho}_1, \widehat{\rho}_2$ leading to $\widetilde{\Sigma}_\Delta = \widehat{\rho}_1 \widehat{\Sigma} + \widehat{\rho}_2 \Delta$.

- Estimation of Σ (and therefore Σ^{-1}) and β in a joint statistical framework (Hannart, 2016, JClim).

  Uncertainty on Σ is partly taken into account in the estimation and CI on β.

$$\widehat{\beta}_\Delta = (X' \widehat{\Sigma}_\Delta^{-1} X)^{-1} X' \widehat{\Sigma}_\Delta^{-1} Y.$$

- The dimension reduction is no longer required - an appropriate prior has to be used.

How to estimate modeling uncertainty for D&A?

- Need to set a paradigm: how far are the models from the truth?
- We assume "models (m_i) are stat. indistinguishable from the truth (m^*)"

$$(m_i - m_j) \sim N(0, 2\Sigma_m), \quad (m_i - m^*) \sim N(0, 2\Sigma_m).$$

How to estimate modeling uncertainty for D&A?

- Need to set a paradigm: how far are the models from the truth?
- We assume "models (m_i) are stat. indistinguishable from the truth (m^*)"

$$(m_i - m_j) \sim N(0, 2\Sigma_m), \quad (m_i - m^*) \sim N(0, 2\Sigma_m).$$

How to estimate modeling uncertainty for D&A?

- Need to set a paradigm: how far are the models from the truth?
- We assume "models (m_i) are stat. indistinguishable from the truth (m^*)"

$$(m_i - m_j) \sim N(0, 2\Sigma_m), \quad (m_i - m^*) \sim N(0, 2\Sigma_m).$$

Or using a different point of view (μ: mean of the model population)

$$(m_i - \mu) \sim N(0, \Sigma_m), \quad (\mu - m^*) \sim N(0, \Sigma_m)$$

How to estimate modeling uncertainty for D&A?

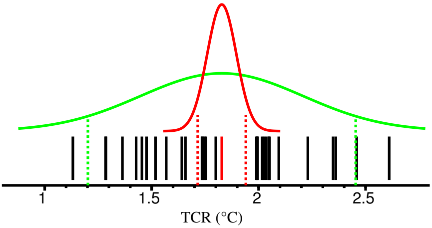
- Need to set a paradigm: how far are the models from the truth?
- We assume "models (m_i) are stat. indistinguishable from the truth (m^*)"

$$(m_i - m_j) \sim N(0, 2\Sigma_m), \quad (m_i - m^*) \sim N(0, 2\Sigma_m).$$

Or using a different point of view (μ: mean of the model population)

$$(m_i - \mu) \sim N(0, \Sigma_m), \quad (\mu - m^*) \sim N(0, \Sigma_m)$$

Illustration:



How to estimate modeling uncertainty for D&A?

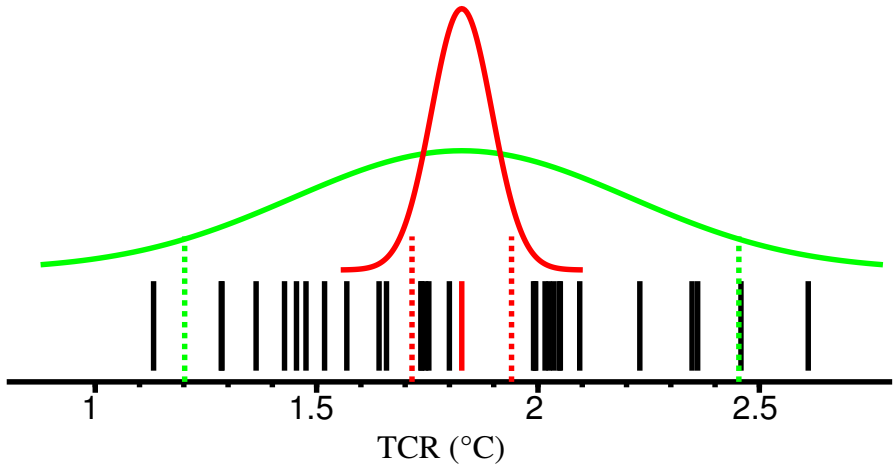
- Need to set a paradigm: how far are the models from the truth?
- We assume "models (m_i) are stat. indistinguishable from the truth (m^*)"

$$(m_i - m_j) \sim N(0, 2\Sigma_m), \quad (m_i - m^*) \sim N(0, 2\Sigma_m).$$

Or using a different point of view (μ: mean of the model population)

$$(m_i - \mu) \sim N(0, \Sigma_m), \quad (\mu - m^*) \sim N(0, \Sigma_m)$$

Illustration:



- Magnitude and pattern uncertainty are estimated consistently.
- Should we assume a larger distribution?

Modeling uncertainty vs internal variability in MME

- Simulated responses are affected by both model's error and internal variability,

Use of linear mixed models (model j, run k):

$$w_{jk} = \mu + m_j + \epsilon_{jk}, \qquad j = 1, \ldots, n_m, \; k = 1, \ldots, n_j,$$

Modeling uncertainty vs internal variability in MME

- Simulated responses are affected by both model's error and internal variability,

Use of linear mixed models (model j, run k):

$$\underset{\sim N(\mu,\, \Sigma_m + \Sigma_v)}{w_{jk}} = \mu + \underset{\sim N(0,\, \Sigma_m)}{m_j} + \underset{\sim N(0,\, \Sigma_v)}{\epsilon_{jk}}, \qquad j = 1, \ldots, n_m,\ k = 1, \ldots, n_j,$$

Modeling uncertainty vs internal variability in MME

- Simulated responses are affected by both model's error and internal variability,

Use of linear mixed models (model j, run k):

$$\underset{\sim N(\mu,\,\Sigma_m+\Sigma_v)}{w_{jk}} = \mu + \underset{\sim N(0,\,\Sigma_m)}{m_j} + \underset{\sim N(0,\,\Sigma_v)}{\epsilon_{jk}}, \qquad j=1,\ldots,n_m,\ k=1,\ldots,n_j,$$

Estimation of Σ_m

$$w_{j.} = 1/n_r \sum_{k=1}^{n_r} w_{jk}, \quad SSM = \sum_{j=1}^{n_m} (w_{j.} - \overline{w})^2,$$

$$\widehat{\Sigma}_m = \frac{1}{n_m - 1}\left(SSM - \frac{n_m - 1}{n_m} \sum_{j=1}^{n_m} \frac{1}{n_j}\Sigma_v \right)_+ .$$

Estimating modeling uncertainty : open issues

- Dimension:
  - about 40 models in CMIP5,
  - about 10 participating to DAMIP,
  - typical dimension of Y is > 30 (sometimes much larger)...
- Models are not independent,
- Ensemble design: CMIP not designed to sample uncertainty (e.g. physical parameters, forcing uncertainty).

Conclusions

- A wide range of statistical models and methods are used in D&A, with different levels of complexity.
  - Mainly regression based models (so far),
  - Climate modeling uncertainty is often not considered.
- Many statistical issues of interest in this area.
  - EIV models,
  - Estimation of large covariance matrices,
  - Estimation of climate modeling uncertainty,
- Hopefully, improving the methods could lead to improved observational constrain on future changes (e.g. climate sensitivity, changes in extreme events).