Méthodologie pour la détection statistique du changement climatique

Aurélien RIBES, Jean-Marc AZAÏS, Serge PLANTON

9 Octobre 2007

Réunion de rentrée de l'équipe MAFIA

INSTITUT de MATHEMATIQUES de TOULOUSE
UNIVERSITE PAUL SABATIER TOULOUSE III
METEO FRANCE Toujours un temps d'avance

Problématique

Hypothèses

- $(\psi_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^p$ va iid de loi $N(0, C)$
- $\psi_{n+1} \in \mathbb{R}^p$, independante des $(\psi_i)_{i\in[\![1,n]\!]}$, de loi $N(\mu g, C)$

Avec $\mu \in \mathbb{R}$, $g \in \mathbb{S}^{p-1} \subset \mathbb{R}^p$, et $C \in \mathcal{M}_p(\mathbb{R})$

Problématique

Hypothèses

- $(\psi_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^p$ va iid de loi $N(0, C)$
- $\psi_{n+1} \in \mathbb{R}^p$, independante des $(\psi_i)_{i\in[\![1,n]\!]}$, de loi $N(\mu g, C)$

Avec $\mu \in \mathbb{R}$, $g \in \mathbb{S}^{p-1} \subset \mathbb{R}^p$, et $C \in \mathcal{M}_p(\mathbb{R})$

Test

On souhaite tester :

$$H_0 : "\mu = 0" \text{ vs } H_1 : "\mu > 0"$$

dans un contexte de grande dimension, où $n \sim p$.

1 Introduction

2 Présentation du problème
- Cas C connu
- Cas C inconnu

3 Solution proposée
- Difficultés pratiques
- Régularisation
- Bootstrap

4 Exemples d'application
- Scenario climatique
- Observations

Famille de tests étudiée

On s'intéresse aux tests $(T_f)_{f \in \mathbb{R}^p}$ suivants :

Variable de test d_f

$$d_f = \langle \psi_{n+1}, f \rangle \quad \sim_{H_0} N(0, f'Cf)$$

Région de rejet

$$W_f = \left\{ \psi_{n+1} \in \mathbb{R}^p, d_f = \langle \psi_{n+1}, f \rangle \geq d_f^{(\alpha)} \right\}$$

avec

$$d_f^{(\alpha)} = \Phi^{-1}(1-\alpha)\sqrt{f'Cf}$$

T_f optimal

Question

Connaissant C et g, existe-t-il un T_f optimal ?

T_f optimal

Question

Connaissant C et g, existe-t-il un T_f optimal ?

Réponse : oui

$T_{C^{-1}g}$ est optimal, parmi les (T_f), dans les sens suivants

- $d_{C^{-1}g}$ maximise le rapport signal sur bruit
- $T_{C^{-1}g}$ est le test le plus puissant
- $T_{C^{-1}g}$ est le test du rapport de vraisemblance

Notations

Dans le cas "C inconnu", on cherche à approcher $T_{C^{-1}g}$.
On considère deux nouvelles familles de tests

Tests $\mathcal{T}_f$

- f est "estimé" : dépend des $(\psi_i)_{i\in[\![1,n]\!]}$, aléatoires, et de g
- Le niveau, conditionnellement aux $(\psi_i)_{i\in[\![1,n]\!]}$, est nominal ;
  C connu pour le calcul de $d_f^{(\alpha)} = \Phi^{-1}(1-\alpha)\sqrt{f'Cf}$

Notations

Dans le cas "C inconnu", on cherche à approcher $T_{C^{-1}g}$.
On considère deux nouvelles familles de tests

Tests $\mathcal{T}_f$

Tests $\mathbb{T}_f$

- f est "estimé"
- Le seuil $d_f^{(\alpha)}$ est estimé et dépend de g et des $(\psi_i)_{i\in[\![1,n]\!]}$; de même pour la p-value
- Le niveau total n'est pas nécessairement nominal

Notations

Dans le cas "C inconnu", on cherche à approcher $T_{C^{-1}g}$.
On considère deux nouvelles familles de tests

Tests $\mathcal{T}_f$

Tests $\mathbb{T}_f$

Test naïf : T_g

Notations

Dans le cas "C inconnu", on cherche à approcher $T_{C^{-1}g}$.
On considère deux nouvelles familles de tests

Tests $\mathcal{T}_f$

Tests $\mathbb{T}_f$

Test naïf : T_g

$T_g = \mathcal{T}_g$

Problématique dans le cas C inconnu

Objectifs

On souhaite construire un test $\mathbb{T}_f$ ayant de bonnes propriétés :

- Niveau nominal
- Puissance supérieure à celle du test naïf T_g

Problématique dans le cas C inconnu

Objectifs

On souhaite construire un test $\mathbb{T}_f$ ayant de bonnes propriétés :

- Niveau nominal
- Puissance supérieure à celle du test naïf T_g

Remarque

On peut commencer par étudier les $\mathcal{T}_f$, en sachant "$\mathcal{T}_f > \mathbb{T}_f$"

Utilisation de la covariance empirique

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} \psi_i \psi_i'$$

$\mathcal{T}_{\hat{C}^{-1}g}$ a-t-il de bonnes propriétés ?

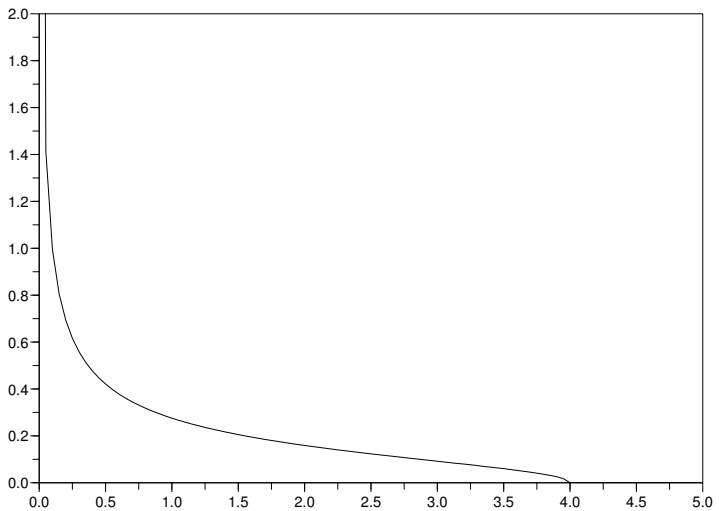

Fig.: Loi de Marčenko-Pastur : densité limite des valeurs propres de $\hat{C}$, lorsque $C = I_p$, $n = p$, et $n, p \to \infty$

Utilisation de la covariance empirique

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} \psi_i \psi_i'$$

$\mathcal{T}_{\hat{C}^{-1}g}$ a-t-il de bonnes propriétés ?

Non ! ($\hat{C}$ est quasi-singulière)
L'utilisation de la pseudo-inverse d'une troncation de $\hat{C}$ ne donne pas de résultats satisfaisants.

La régularisation dans notre cas

Principe

On utilise $\gamma\hat{C} + \rho I_p$ au lieu de $\hat{C}$.

Justification

- Interpolation avec le test naïf T_g
- Point de vue ridge-régression
- Puissance des tests $\mathcal{T}_{(\hat{C}+\rho I_p)^{-1}g}$

Estimation des paramètres (γ, ρ)

(O. Ledoit, *A Well-Conditioned Estimator for Large Dimensional Covariance Matrices*)

Cadre :

- On raisone sur $\mathcal{M}_p(\mathbb{R})$ munit de la norme

$$\|A\|^2_{\mathcal{M}_p} = \frac{\mathrm{Tr}(AA')}{p}$$

- On se place en "asymptotique générale" : n, p_n, $\frac{p_n}{n} \leq K$
- On cherche un estimateur de la forme $C^* = \gamma \hat{C} + \rho I_p$, qui va minimiser :

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

Estimation des paramètres (γ, ρ)

(O. Ledoit, *A Well-Conditioned Estimator for Large Dimensional Covariance Matrices*)

n fixé : γ_n^0 et ρ_n^0 optimaux, en fonction de C

$$\gamma_n^0 = \frac{\alpha^2}{\delta^2}, \qquad \text{and} \qquad \rho_n^0 = \frac{\beta^2 \nu}{\delta^2}$$

où

$$\nu = \langle C, I_p \rangle_{\mathcal{M}_p} = \frac{\mathrm{Tr}(C)}{p}, \qquad \alpha^2 = \| C - \nu I_p \|^2_{\mathcal{M}_p},$$

$$\beta^2 = \mathrm{E}\left(\| \hat{C} - C \|^2_{\mathcal{M}_p} \right), \qquad \delta^2 = \mathrm{E}\left(\| \hat{C} - \nu I_p \|^2_{\mathcal{M}_p} \right),$$

Estimation des paramètres (γ, ρ)

(O. Ledoit, *A Well-Conditioned Estimator for Large Dimensional Covariance Matrices*)

n fixé : γ_n^0 et ρ_n^0 optimaux, en fonction de C

Sous l'asymptotique générale : estimateurs convergents $\hat{\gamma}_n^0$ et $\hat{\rho}_n^0$ de γ_n^0 et ρ_n^0

$$\hat{\gamma} = \frac{\hat{\alpha}^2}{\hat{\delta}^2}, \qquad \text{and} \qquad \hat{\rho} = \frac{\hat{\beta}^2 \hat{\nu}}{\hat{\delta}^2}$$

Avec

$$\hat{\nu} = \langle \hat{C}, I_p \rangle_{\mathcal{M}_p} = \frac{\mathrm{Tr}(\hat{C})}{p}, \qquad \hat{\beta}^2 = \min\left(\hat{\delta}^2, \frac{1}{n^2} \sum_{i=1}^{n} \| \psi_i \psi_i' - \hat{C} \|_{\mathcal{M}_p}^2 \right),$$

$$\hat{\delta}^2 = \| \hat{C} - \hat{\nu} I_p \|_{\mathcal{M}_p}^2, \qquad \hat{\alpha}^2 = \hat{\delta}^2 - \hat{\beta}^2.$$

Estimation des paramètres (γ, ρ)

(O. Ledoit, *A Well-Conditioned Estimator for Large Dimensional Covariance Matrices*)

n fixé : γ_n^0 et ρ_n^0 optimaux, en fonction de C

Sous l'asymptotique générale : estimateurs convergents $\hat{\gamma}_n^0$ et $\hat{\rho}_n^0$ de γ_n^0 et ρ_n^0

Un nouvel estimateur de C peut être déduit :

$$\hat{C}_I = \hat{\gamma}_n^0 \hat{C} + \hat{\rho}_n^0 I_p$$

Problématique

On souhaite finir de construire un $\mathbb{T}_{\hat{C}_I^{-1} g}$.

Il reste à :

- estimer le seuil (ou plus généralement la p-value).
- vérifier que l'estimation est bonne, ie le test est de niveau nominal
- calculer la puissance

Normalisation

Estimation de la variance

En notant $\hat{f}_o = \hat{C}_I^{-1} g$,

$$d_{\hat{f}_o} \left| (\psi_i)_{1 \leq i \leq n} \right. \sim_{H_0} N(0, \hat{f}_o' \, C \, \hat{f}_o)$$

La variance $\hat{f}_o' \, C \, \hat{f}_o$ peut être estimée par $\hat{f}_o' \, \hat{C}_I \, \hat{f}_o$

Normalisation

Estimation de la variance

En notant $\hat{f}_o = \hat{C}_I^{-1} g$,

$$d_{\hat{f}_o} \big| (\psi_i)_{1 \leq i \leq n} \sim_{H_0} N(0, \hat{f}_o' \, C \, \hat{f}_o)$$

La variance $\hat{f}_o' \, C \, \hat{f}_o$ peut être estimée par $\hat{f}_o' \, \hat{C}_I \, \hat{f}_o$

Variable normalisée

Une variable normalisée est déduite : $\delta = \dfrac{d_{\hat{f}_o}}{\sqrt{\hat{f}_o' \, \hat{C}_I \, \hat{f}_o}}$

$$\delta \big| (\psi_i)_{1 \leq i \leq n} \sim_{H_0} N\left(0, \frac{\hat{f}_i^{*\prime} \, C \, \hat{f}_i^*}{\hat{f}_i^{*\prime} \, \hat{C}_I \, \hat{f}_i^*}\right)$$

Normalisation

Estimation de la variance

En notant $\hat{f}_o = \hat{C}_I^{-1} g$,

$$d_{\hat{f}_o} \big| (\psi_i)_{1 \leq i \leq n} \sim_{H_0} N(0, \hat{f}_o' \, C \, \hat{f}_o)$$

La variance $\hat{f}_o' \, C \, \hat{f}_o$ peut être estimée par $\hat{f}_o' \, \hat{C}_I \, \hat{f}_o$

Variable normalisée

Une variable normalisée est déduite : $\delta = \dfrac{d_{\hat{f}_o}}{\sqrt{\hat{f}_o' \, \hat{C}_I \, \hat{f}_o}}$

$$\delta \big| (\psi_i)_{1 \leq i \leq n} \sim_{H_0} N\left(0, \frac{\hat{f}_i^{*\prime} \, C \, \hat{f}_i^*}{\hat{f}_i^{*\prime} \, \hat{C}_I \, \hat{f}_i^*}\right) \qquad \delta \sim_{H_0} \mathcal{D}_{C,g}$$

Bootstrap

Step 1 : Estimating the threshold

- Simulation of $\mathcal{D}_{\hat{C}_I,g}$ (instead of $\mathcal{D}_{C,g}$) ;
- "Candidate" threshold $\mathcal{D}^{(\alpha)}_{\hat{C}_I,g}$.

Bootstrap

Step 1 : Estimating the threshold

- Simulation of $\mathcal{D}_{\hat{C}_I,g}$ (instead of $\mathcal{D}_{C,g}$) ;
- "Candidate" threshold $\mathcal{D}^{(\alpha)}_{\hat{C}_I,g}$.

Step 2 : Validating

- Validation of the use of $\mathcal{D}^{(\alpha)}_{\hat{C}_I,g}$ instead of $\mathcal{D}^{(\alpha)}_{C,g}$;
- Computation of the level and the power.

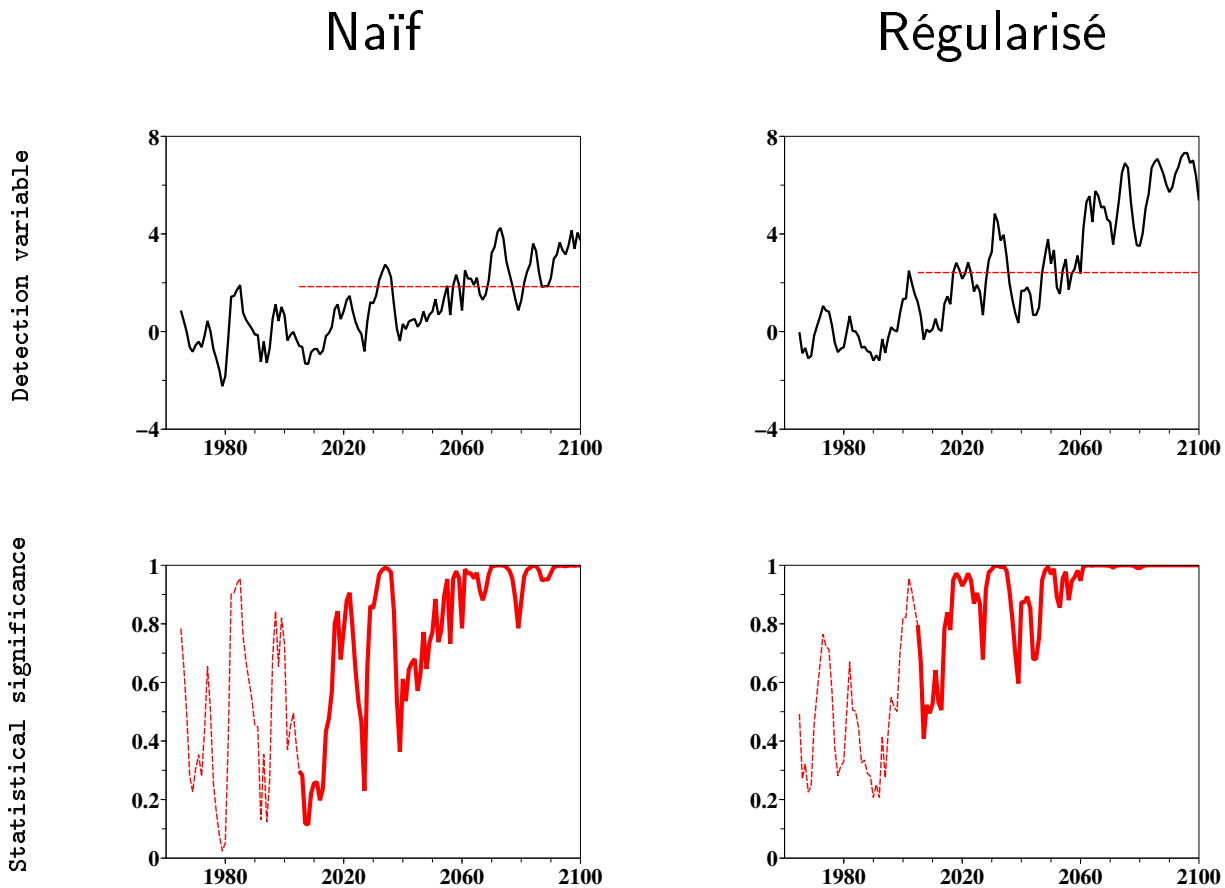

Fig.: Températures minimales d'été dans le modèle : Comparison bentre le test naïf T_g et le test régularisé $\mathbb{T}_{\hat{C}_I^{-1}g}$ pour des températures minimales estivales issues d'un scénario climatique.

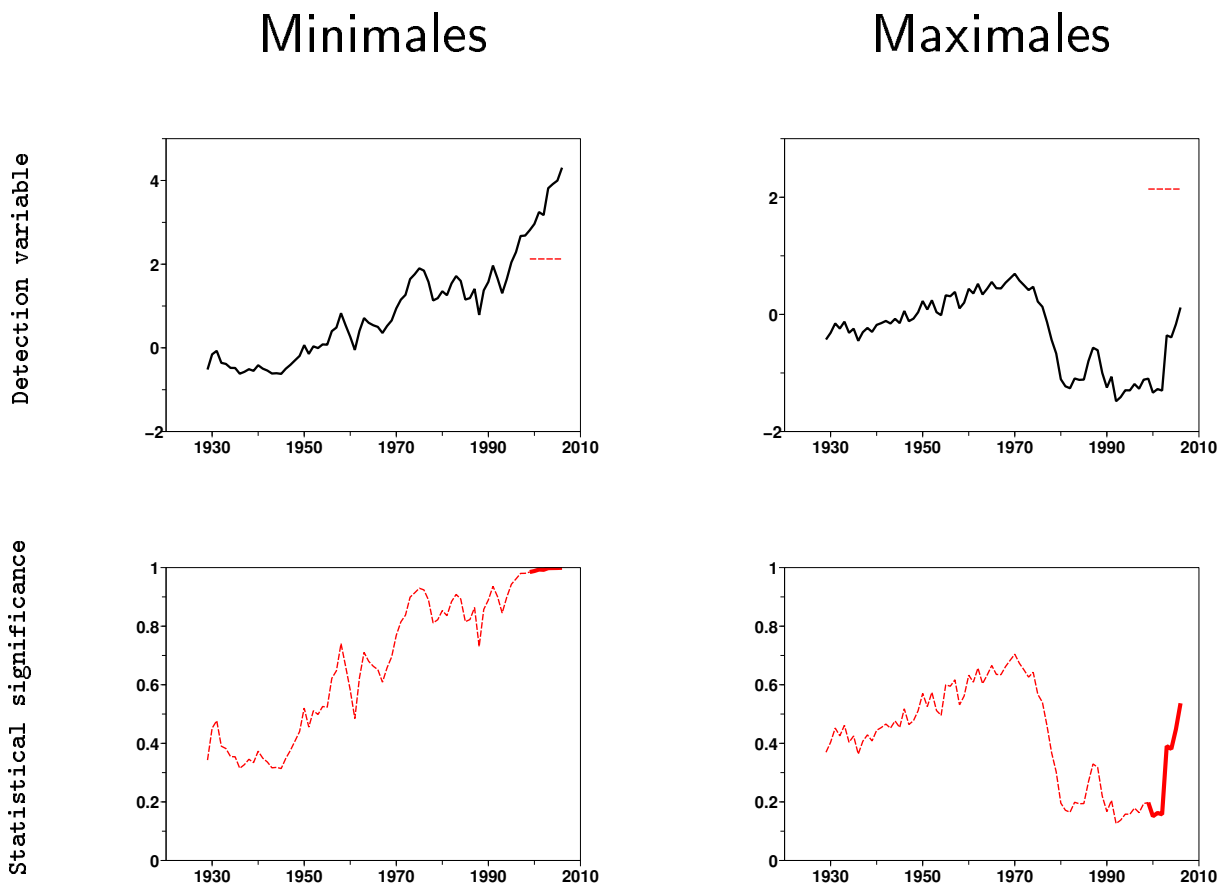

Fig.: Summer temperatures in observations : Results for minimum and maximum temperatures

Méthodologie pour la détection statistique du changement climatique

Aurélien RIBES, Jean-Marc AZAÏS, Serge PLANTON

9 Octobre 2007

Questions