Estimation d'une densité discrète sous contrainte de k-monotonie. Application à l'estimation du nombre d'espèces dans une population.

Jade Giguelay sous la direction de Christophe Giraud et Sylvie Huet

11 janvier 2016

Sommaire

- Densités $k-$monotones discrètes.
- Motivation : Estimation du nombre d'espèces dans une population.
- Estimation sous-contrainte de $k-$monotonie :
  - Propriétés.
  - Algorithme.
- Simulations :
  - Etude sur données simulées.
  - Application à la recherche du nombre d'espèces dans une population.
- Perspectives.

Notations

- $p = (p_0, p_1, ...)$ est une densité dans $\mathbb{N}$.
- Le maximum du support τ de p est défini par : $\tau = \min\{j \in \bar{\mathbb{N}}, \forall k > j, p_k = 0\}$.
- $X_1, .., X_n$ sont des variables aléatoires indépendantes de loi p de maximum de support τ (inconnu)
- $\tilde{p}_n$: $\tilde{p}_n(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i = j\}}$ est l'estimateur empirique.
- $k \geqslant 2$ est un entier fixé.

La convexité.

Définition

(Le Laplacien) $\Delta^2 p(i) = p_i - 2p_{i+1} + p_{i+2} = (p_{i+2} - p_{i+1}) - (p_{i+1} - p_i)$.

- La loi p est dite **convexe** si pour tout $i \in \mathbb{N}$, $\Delta^2 p(i) \geqslant 0$.
- Si $\Delta^2 p(i) > 0$ on dit que i est un **noeud** de p.

La convexité 2.

Une fonction convexe est mélange de fonctions triangulaires : si p est une fonction convexe discrète alors :

$$p(i) = \sum_{j=0}^{\infty} \frac{\Delta p(j+1)}{(j+1)(j+2)} T_j(i)$$

Où T_τ est la distribution triangulaire de support $\{0, .., \tau\}$:

$$T_\tau(i) = \begin{cases} \frac{2(\tau+1-i)}{(\tau+1)(\tau+2)} & \text{si j}\in \{0,..,\tau\} \\ 0 & \text{sinon,} \end{cases}$$

Fonction $k-$monotone

Une fonction p de $L^1(\mathbb{N})$ est :

- monotone ($1-$monotone) si pour tout i, $\Delta^1 p(i) := p_{i+1} - p_i \leqslant 0$
- convexe ($2-$monotone) si pour tout i,
  $\Delta^2 p(i) := p_i - 2p_{i+1} + p_{i+2} = \Delta^1 p_{i+1} - \Delta^1 p_i \geqslant 0.$

$\rightarrow$ On veut généraliser cette définition, on dira que p est :

- $3-$monotone si pour tout i, $\Delta^3 p(i) := \Delta^2 p_{i+1} - \Delta^2 p_i \leqslant 0$,

  .

  .
- $k-$monotone si pour tout i, $\Delta^k p(i) := \Delta^{k-1} p_{i+1} - \Delta^{k-1} p_i$ est du signe de $(-1)^k$.

Fonction $k-$monotone

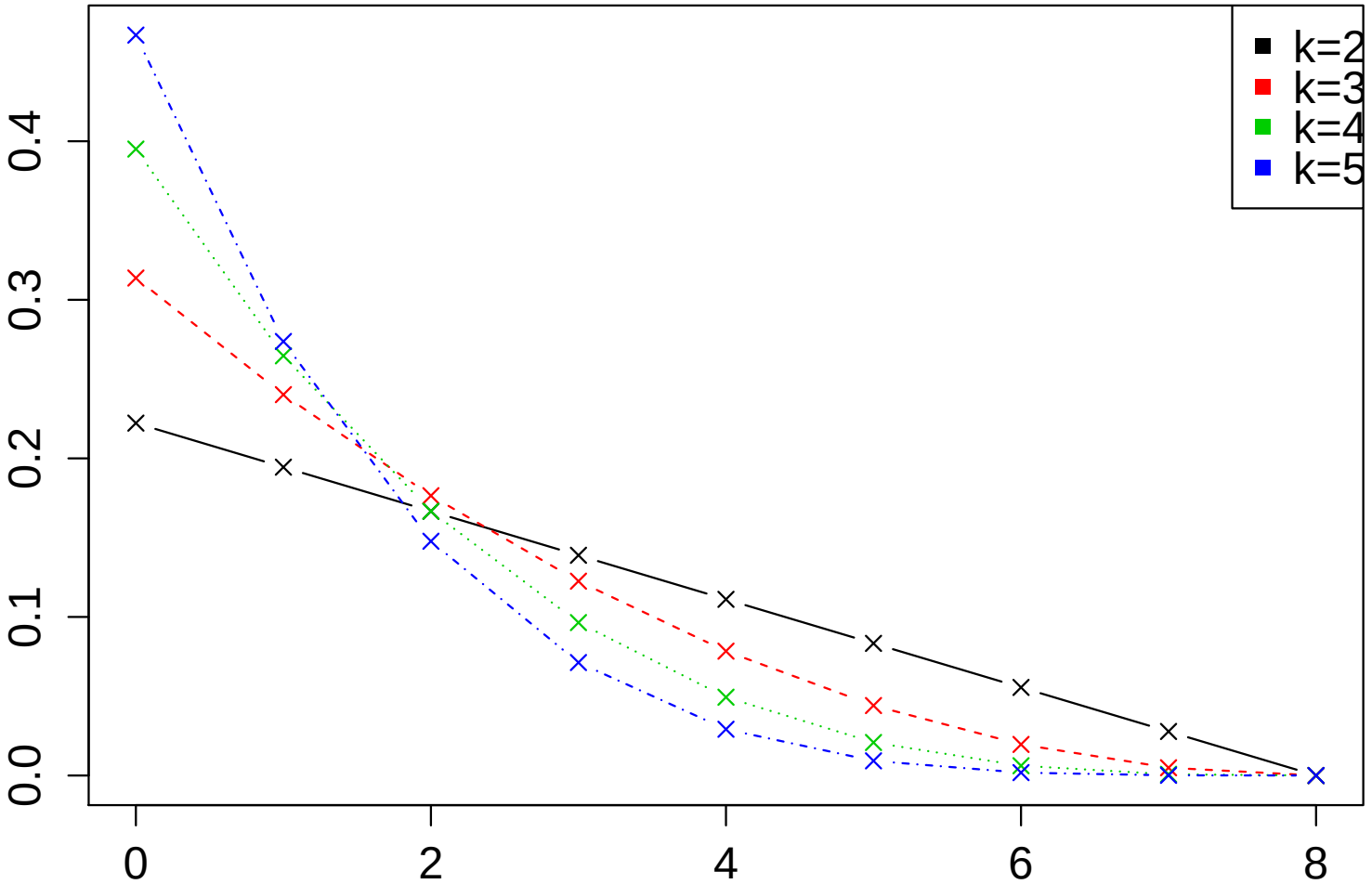
Définition

(Le j-ième Laplacien) $\Delta^j p(i) = \sum_{h=0}^{j} \binom{j}{h} (-1)^{j-h} p(h+i)$.

- La distribution p est $k-$monotone sur $\mathbb{N}$ si $(-1)^k \Delta^k p(i)$ est positive pour tout i.
- Si $(-1)^k \Delta^k p(i) > 0$ on dit que i est un $k-$**noeud** de p.

$\rightarrow$ $k-$monotone $\Rightarrow$ $j-$monotone strictement sur son support, $j = 1, .., k-1$

Exemple de fonctions $k-$monotones



Ecriture en base de spline

On définit une famille de splines $(Q_j^k)_{j\in\mathbb{N}^*}$ de la manière suivante :

$$Q_j^k(i) = C_{j-i+k-1}^{k-1} = \frac{(j-i+k-1)\dots(j-i+1)}{(k-1)!}\mathbb{I}_{j\geqslant i} \qquad (1)$$

(Lefèvre et Loisel (2012)).

Exemple :

- $Q_j^2(i) = (j+1-i)_+$ est la triangulaire T_j non-renormalisée.
- $Q_j^3(i) = \frac{1}{2}\left((j+1-i)_+^2 + (j+1-i)_+\right)$.
- $Q_j^4(i) = \frac{1}{6}\left((j+1-i)_+^3 - (j+1-i)_+\right)$.

Ecriture en base de splines

Une fonction $k-$monotone s'écrit comme un mélange de Q_j^k :

Theorem

Soit p une densité de masse finie sur $\mathbb{N}$. p est $k-$monotone si et seulement si p s'écrit :

$$p(i) = \sum_{j \geqslant i} (-1)^k \Delta^k p_j Q_j^k(i) = \sum_{j \geqslant i} \tilde{\pi}_j \frac{Q_j^k(i)}{m_j}, \tag{2}$$

où $m_j = \sum_{i=0}^{j} Q_j^k(i)$.
De plus, p est une probabilité $\Leftrightarrow$ $\tilde{\pi}$ est une probabilité.

La convexité 3.

Durot et al. (2013) : Estimation de p sous-contrainte de convexité par méthode des moindres carrés :

$$\hat{p}_n = argmin\{||f - \tilde{p}_n||_2, f \text{ fonction convexe}\}. \quad (3)$$

- $\hat{p}_n$ est de support fini.
- $\hat{p}_n$ est de masse 1.
- Algorithme exact via la Méthode de Réduction de Support (Groeneboom et al., (2008))

Estimateur sous contrainte de $k-$monotonie

L'estimateur $\hat{p}_{n,k}$ sous contrainte de $k-$monotonie est défini ainsi :

$$\hat{p}_{n,k} = argmin\{||f - \tilde{p}_n||_2, f \textbf{ probabilité } k-\text{monotone}\}. \quad (4)$$

$\rightarrow$ Il existe et il est unique.

Objectifs

- Etudier les propriétés théoriques de $\hat{p}_n$.
- Caractériser $\hat{p}_n$.
- Calculer $\hat{p}_n$ en pratique.
- Evaluer ses qualités d'estimation.

Propriétés de $\hat{p}_n$

Propriété

$\hat{p}_{n,k}$ est une densité à support fini.

Propriété

L'inégalité suivante est vérifiée pour toute probabilité f $k-$monotone :

$$||f - \hat{p}_{n,k}||_2 \leqslant ||f - \tilde{p}_n||_2$$

avec une inégalité stricte si $\tilde{p}_n$ n'est pas $k-$monotone.

Propriétés de $\hat{p}_n$

Propriété

Soit p une probabilité discrète quelconque.
Pour tout $r \in [2, +\infty]$ on a $\sqrt{n}||p_{S_k} - \hat{p}_{n,k}||_r = O_P(1)$
où $p_{S_k} = argmin\{||f - p||_2, f \in S_k\}$ est le projeté orthogonal de p sur l'ensemble des probabilités $k-$monotones.

Propriété

Soit p une densité $k-$monotone de support fini. Soit $r \in \mathbb{N}$ un $k-$noeud de p. Alors avec probabilité 1 il existe un rang n_0 tel que pour tout $n \geqslant n_0$, r est un $k-$noeud de $\hat{p}_n$.

Remarque sur la masse

On peut également définir comme estimateur :

$$\hat{p}^*_{n,k} = argmin\{||f - \tilde{p}_n||_2, f \textbf{ fonction } k-\text{monotone}\}. \quad (5)$$

Lemme

Soit δ_1 la masse de Dirac en 1. Le projeté de δ_1 sur les fonctions $3-$monotones est :

$$p = \frac{3}{238} Q_5^3 + \frac{1}{238} Q_6^3,$$

où (Q_j^3) désigne la base de spline usuelle :

$$Q_j^3(i) = \frac{1}{2}\left((j+1-i)_+^2 + (j+1-i)_+\right).$$

La masse de p est d'environ 1.14.

Résultat sur le support

Propriété

Si $k = 2$ alors on a l'ordre suivant sur les supports :

$$\tilde{s}_n \leqslant \hat{s}_{n,2}.$$

où $\tilde{s}_n$ est le support de l'estimateur empirique et $\hat{s}_{n,2}$ celui de $\hat{p}_{n,2}$.

Propriété

Soit p une densité $k-$monotone de support fini. On note s le maximum de son support et $\hat{s}_{n,k}$ celui de $\hat{p}_{n,k}$. Alors avec probabilité 1 :

- *Si k est impair : il existe un rang $n_0 \in \mathbb{N}$ tel que pour tout $n \geqslant n_0$ on ait $\hat{s}_{n,k} \leqslant s$.*
- *Si k est pair : il existe un rang $n_0 \in \mathbb{N}$ tel que pour tout $n \geqslant n_0$ on ait $\hat{s}_{n,k} \leqslant s + 1$.*

Heuristique de la preuve du support fini

Par l'absurde, Supposons que $\hat{p}_n$ est à support infini et construisons une autre densité $\bar{p}$ qui possède les propriétés suivantes :

i) $\bar{p} \leqslant \hat{p}_n$.

ii) $\forall i \leqslant \tilde{s}_n, \bar{p}(i) = \hat{p}_n(i)$.

iii) il existe i tel que $\bar{p}(i) < \hat{p}_n(i)$.

iv) $\bar{p}$ est k-monotone et positive,

Pour un tel $\bar{p}$ on aura $||\bar{p} - \tilde{p}_n||_2 < ||\hat{p}_n - \tilde{p}_n||_2$

Heuristique de la preuve du support fini 2

On définit pour $j \in \{1, .., k-1\}$:

$$\begin{aligned} q_1(i) &= -\hat{p}_n(i+1) + \hat{p}_n(i), \\ q_2(i) &= -q_1(i+1) + q_1(i), \\ &\vdots \\ q_{k-1}(i) &= -q_{k-2}(i+1) + q_{k-2}(i) \end{aligned}$$

pour tout $i \in \mathbb{N}$.
$\hat{p}_n$ est k-monotone $\Leftrightarrow$ q_{k-1} est décroissante.

Heuristique de la preuve du support fini 3

On va modifier q_{k-1} en $\bar{q}_{k-1}$ de sorte que si l'on pose de la même manière :

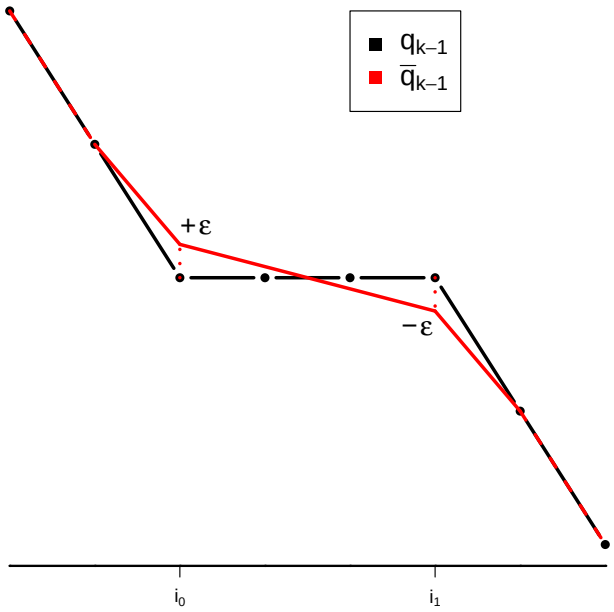
$$\forall j \in \{1, .., k-3\}, \bar{q}_j(i) = -\sum_{h=0}^{i-1} \bar{q}_{j+1}(h) + q_j(0),$$

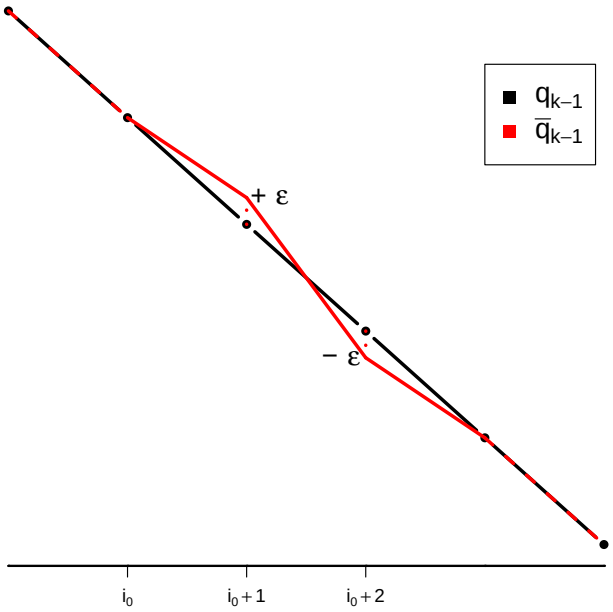
$$\bar{p}(i) = -\sum_{h=0}^{i-1} \bar{q}_1(h) + \hat{p}_n(0).$$

alors $\bar{p}$ vérifie les propriétés $i)ii)iii)iv)$.

Heuristique de la preuve 4

Cas où k pair : $\rightarrow$ Il existe une infinité de $1-$noeuds de q_{k-1}.





Un algorithme en deux temps

Step 1) $L \in \mathbb{N}^*$ est fixé,
on calcule l'estimateur $\hat{p}_n^L$:

$$\hat{p}_n^L = \text{argmin}\{||\tilde{p}_n - q||_2, q \in S_{k,L}\},$$

où $S_{k,L} = \{\text{probabilités } f \ k-\text{monotones, supp}(f) \subset \{0, .., L\}\}$.

Step 2) Trouver un critère d'arrêt : à partir de quel $L \in \mathbb{N}^*$, $\hat{p}_n = \hat{p}_n^L$?

Caractérisation de $\hat{p}_n$.

Notation : soit p une probabilité discrète.

$$\beta(p) = \sum_{i=0}^{\infty} p(i)\big(p(i) - \tilde{p}_n(i)\big).$$

Propriété

$$\beta(\hat{p}_n) \leqslant 0.$$

De plus $\beta(\hat{p}_n) = 0$ *si et seulement si :*

$$\hat{p}_{n,k} = \hat{p}^*_{n,k} = \{||f - \tilde{p}_n||_2, f \textbf{ fonction } k-monotone\}.$$

Caractérisation de $\hat{p}_n$.

$$\text{Notation (primitives)} : \forall l \in \mathbb{N}, \begin{cases} F_f^1(l) = F_f(l) = \sum_{i=0}^{l} f(i), \\ \forall j \geqslant 2, F_f^j(l) = \sum_{i=0}^{l} F_f^{j-1}(l). \end{cases}$$

Theorem

Soit p une probabilité. Soit m_l^k la masse de Q_l^k. Il y a équivalence entre :

1. $p = \hat{p}_n$.
2. - *Pour tout $l \in \mathbb{N}$ on a l'égalité :*

$$F_{\hat{p}_n}^k(l) - F_{\tilde{p}_n}^k(l) \geqslant m_l^k \beta(\hat{p}_n) \tag{6}$$

   - *il y a égalité si l est un noeud de $\hat{p}_n$.*

$\rightarrow$ Il faut trouver un entier L tel que pour tout $l \geqslant L$,

$$F_{\hat{p}_n}^k(l) - F_{\tilde{p}_n}^k(l) - m_l^k \beta(\hat{p}_n) \geqslant 0.$$

Critère d'arrêt

Résultats :

- On sait calculer un tel L en pratique pour tout $k \geqslant 3$.
- Pour les cas $k = 3$ et $k = 4$ on a même un résultat plus précis :

Propriété

Soit s le maximum entre le support de p et le support de $\tilde{p}_n$. On a équivalence entre :

1. $p = \hat{p}_n$.
2.
   1. $\forall l \leqslant s+1, F^k_{\hat{p}_n}(l) - F^k_{\tilde{p}_n}(l) \geqslant m^k_l \beta(\hat{p}_n)$.
   2. *Si l est un $k-$noeud de p il y a égalité dans l'inégalité précédente.*
   3. $\forall j \in \{1, .., k-1\}, F^j_p(s+1) \geqslant F^j_{\tilde{p}_n}(s+1)$

Distributions d'abondance

Une population comprend N espèces. A_l est le nombre d'individus de l'espèce i observés.

Les A_l sont i.i.d de loi p dite **distribution d'abondance** :

$$p_i = \mathbb{P}(A_l = i).$$

Observation : On n'observe pas les A_i mais $X_1, .., X_D \sim p^+$ la distribution d'abondance tronquée en 0 :

$$p_j^+ = \mathbb{P}(A_i = j / A_i > 0) = \frac{p_j}{1 - p_0}.$$

D est la VA du nombre d'espèces observées. $D \sim \mathcal{B}(N, 1 - p_0)$.

But : Estimation de N.

Distributions d'abondance

$D \sim \mathcal{B}(N, 1-p_0)$ et $\mathbb{E}[D] = N(1-p_0)$.

Un estimateur de N est $\hat{N} = \frac{D}{1-\hat{p}_0}$

$\rightarrow$ On se ramène à l'estimation de p_0. (Problème **non-identifiable**)

Définition

On dit qu'une probabilité $k-$monotone est $k-$monotone d'abondance si elle s'écrit ainsi :

$$p_j = \sum_{i=1}^{\infty} \pi_i T_i(j)$$

En d'autre termes, si dans son unique décomposition dans la base de spline (Q_j^k) on a $\pi_0 = 0$.

Estimation d'une distribution d'abondance

Idée : On veut estimer p_0 en supposant que p est $k-$monotone d'abondance. Alors p^+ est une distribution $k-$monotone.

$\rightarrow$ p $k-$monotone $\Rightarrow$ $p = \sum_{j=0}^{\infty} \pi_j Q_j^k$.

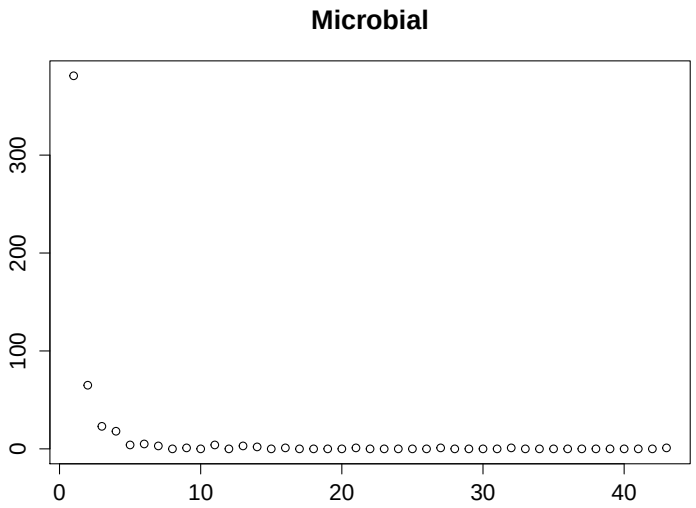
$Q_0^k = \delta_0$: les espèces qu'on observe 0 fois.

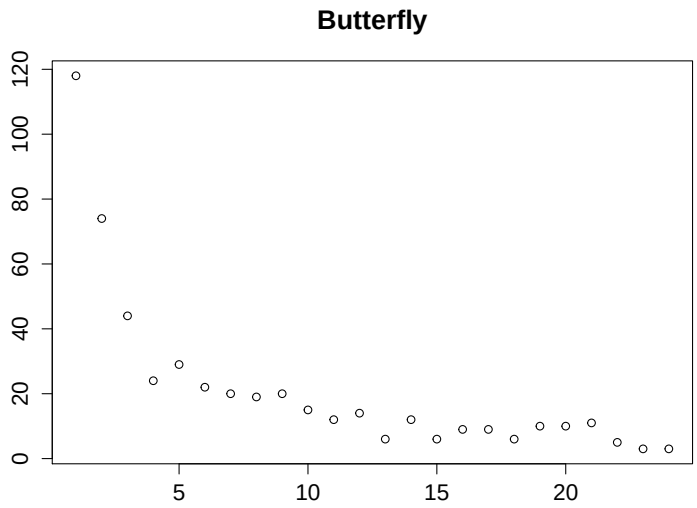
$\rightarrow$ On suppose $\pi_0 = 0$. On a alors :

$$p_0 = \sum_{h=1}^{k} \binom{k}{h} (-1)^{k-h} p_h = \sum_{h=1}^{k} \binom{k}{h} (-1)^{k-h} p_h^+.$$

$\rightarrow$ On se ramène à estimer p^+ sous-contrainte de $k-$monotonie.

Exemple de données (Wang and Lindsay (2005))





Exemple sur jeu de données réel

Butterfly :

k	$k = 2$	$k = 3$	$k = 4$
$\hat{p}_0$	0.244	0.26	0.277
$\hat{N}$	663	677	692

Microbial :

k	$k = 2$	$k = 3$	$k = 4$
$\hat{p}_0$	0.576	0.654	0.703
$\hat{N}$	1212	1485	1730

Bilan

- On a proposé un estimateur de p sous contrainte de $k-$monotonie.
- Il est de support fini, de masse 1.
- On sait le caractériser et le calculer en pratique.

Perspectives

- Inférence sur k.
  - Par sélection de modèle.
  - Par procédure de tests emboités.
- Application sur problématique concrète.

Bibliographie

Merci !

- Durot, Huet, Koladjo, and Robin, Computational Statistics and Data Analysis, **67**, (2013)
- Groeneboom et al., Scandinavian Journal of Statistics, **35**, (2008)
- Balabdaoui and Wellner, The Annals of Statistics, (2007)
- Lefevre and Loisel, Journal of Applied Probability, **50**, (2013)
- Wang and Lindsay, Journal of the American Statistical Association, **100**, (2005).