Le Filtrage Collaboratif vu comme un problème de Consensus d'Ordonnancements

Romaric Gaudel and **Stéphan Clémençon**

LTCI, UMR 5141
Télécom-ParisTech / CNRS

CAp, mai 2011

Filtrage Collaboratif pour le commerce en ligne

amazon.com ebay millemercis.com

- Site de commerce en ligne
  - $m = 10^6$ utilisateurs
  - $n = 10^6$ produits
  - Seule information : liste des achats

8h30	Alice	prod_01
8h31	Bob	prod_02
8h47	Bob	prod_03
8h50	Bob	prod_04
11h45	Charlie	prod_01
12h02	Alice	prod_05
12h34	Dave	prod_06

- Objectif
  - Prédire les achats futurs

Alice	prod_02
Alice	prod_07
Alice	prod_03
Bob	prod_08
Bob	prod_06
Bob	prod_07

Filtrage Collaboratif par complétion de matrice

- Données
  - Matrice utilisateur $\times$ produit : $M = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \end{pmatrix}$
- Approximation de rang k
  - $\hat{M} = \underset{A \cdot B^T : A \in \mathcal{M}_{m,k}, B \in \mathcal{M}_{n,k}}{\text{argmin}} \|M - A \cdot B^T\|^2$

$$\Longrightarrow \hat{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0.5 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.2 & 1 & 1 & 0.9 & 0.4 & 0.92 & 0.9 & 0.95 \end{pmatrix}$$

$$= \begin{pmatrix} 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.2 & 1 & 1 & 0.9 & 0.4 & 0.92 & 0.9 & 0.95 \\ 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.5 & 0.95 & 0.6 & 0.1 & 0.7 & 0.96 & 0.8 & 0.5 \end{pmatrix}$$

- Prédiction
  - Produits (non-achetés) de plus grand score

Filtrage Collaboratif par complétion de matrice

- Données
  - Matrice utilisateur $\times$ produit : $M = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \end{pmatrix}$
- Approximation de rang k
  - $\hat{M} = \underset{A \cdot B^T : A \in \mathcal{M}_{m,k}, B \in \mathcal{M}_{n,k}}{\operatorname{argmin}} \|M - A \cdot B^T\|^2$

$$\Longrightarrow \hat{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0.5 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.2 & 1 & 1 & 0.9 & 0.4 & 0.92 & 0.9 & 0.95 \end{pmatrix}$$

$$= \begin{pmatrix} 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.2 & 1 & 1 & 0.9 & 0.4 & 0.92 & 0.9 & 0.95 \\ 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.5 & 0.95 & 0.6 & 0.1 & 0.7 & 0.96 & 0.8 & 0.5 \end{pmatrix}$$

- Prédiction
  - Produits (non-achetés) de plus grand score

Filtrage Collaboratif par complétion de matrice

- Données
  - Matrice utilisateur $\times$ produit : $M = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \end{pmatrix}$
- Approximation de rang k
  - $\hat{M} = \underset{A \cdot B^T : A \in \mathcal{M}_{m,k}, B \in \mathcal{M}_{n,k}}{\operatorname{argmin}} \|M - A \cdot B^T\|^2$

$$\Longrightarrow \hat{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0.5 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.2 & 1 & 1 & 0.9 & 0.4 & 0.92 & 0.9 & 0.95 \end{pmatrix}$$

$$= \begin{pmatrix} 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.2 & 1 & 1 & 0.9 & 0.4 & 0.92 & 0.9 & 0.95 \\ 0.8 & 0.9 & 0.2 & 0.1 & 1 & 0.1 & 0.7 & 0.05 \\ 0.5 & 0.95 & 0.6 & 0.1 & 0.7 & 0.96 & 0.8 & 0.5 \end{pmatrix}$$

- Prédiction
  - Produits (non-achetés) de plus grand score

Ne tient pas compte de l'ordre des achats

Filtrage Collaboratif par consensus d'ordonnancements

- Objectif
  - Prendre en considération l'ordre des achats
- Un utilisateur = un ordonnancement / une permutation des produits

```
8h31   Bob   prod_02
8h47   Bob   prod_03
8h50   Bob   prod_04
```

$$\Longrightarrow \quad \pi_{\text{Bob}} = [2,\ 3,\ 4,\ \ldots]$$

- Des utilisateurs

[1, 5, . . .]	[2, 3, 4, . . .]	[1, 7, 9, . . .]
[5, 1, . . .]	[2, 3, . . .]	[11, 13, . . .]
[1, 4, 5, 3, 11, . . .]	[2, 4, 3, 8, 9, . . .]	[12, 4, 13, 8, 19, . . .]

- Objectifs intermédiaires
  - Regrouper les utilisateurs au comportement similaire
    $\Longrightarrow$ Partitionnement d'exemples (clustering)
  - Prédire le comportement futur d'un utilisateur à partir de son groupe
    $\Longrightarrow$ Consensus d'ordonnancements

Filtrage Collaboratif par consensus d'ordonnancements

- Objectif
  - Prendre en considération l'ordre des achats
- Un utilisateur = un ordonnancement / une permutation des produits

```
8h31   Bob   prod_02
8h47   Bob   prod_03
8h50   Bob   prod_04
```

$$\Longrightarrow \quad \pi_{\text{Bob}} = [2,\ 3,\ 4,\ \ldots]$$

- Des utilisateurs

[1, 5, . . .]	[2, 3, 4, . . .]	[1, 7, 9, . . .]
[5, 1, . . .]	[2, 3, . . .]	[11, 13, . . .]
[1, 4, 5, 3, 11, . . .]	[2, 4, 3, 8, 9, . . .]	[12, 4, 13, 8, 19, . . .]

- Objectifs intermédiaires
  - Regrouper les utilisateurs au comportement similaire
    $\Longrightarrow$ Partitionnement d'exemples (clustering)
  - Prédire le comportement futur d'un utilisateur à partir de son groupe
    $\Longrightarrow$ Consensus d'ordonnancements

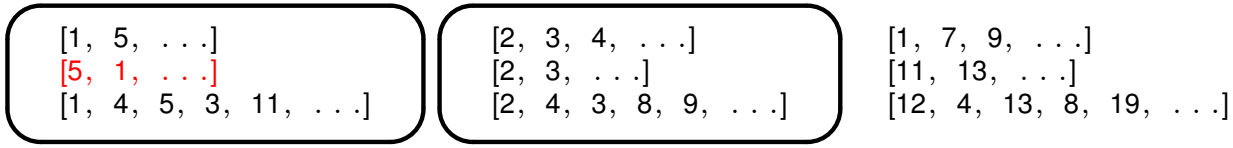
Filtrage Collaboratif par consensus d'ordonnancements

- Objectif
  - Prendre en considération l'ordre des achats
- Un utilisateur = un ordonnancement / une permutation des produits

```
8h31   Bob   prod_02
8h47   Bob   prod_03
8h50   Bob   prod_04
```

$\Longrightarrow \quad \pi_{\text{Bob}} = [2,\ 3,\ 4,\ \ldots]$

- Des utilisateurs



- Objectifs intermédiaires
  - Regrouper les utilisateurs au comportement similaire
    $\Longrightarrow$ Partitionnement d'exemples (clustering)
  - Prédire le comportement futur d'un utilisateur à partir de son groupe
    $\Longrightarrow$ Consensus d'ordonnancements

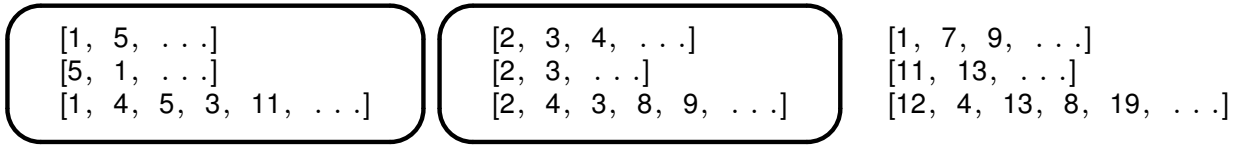
Filtrage Collaboratif par consensus d'ordonnancements

- Objectif
  - Prendre en considération l'ordre des achats
- Un utilisateur = un ordonnancement / une permutation des produits

```
8h31   Bob   prod_02
8h47   Bob   prod_03
8h50   Bob   prod_04
```

$\Longrightarrow \quad \pi_{\text{Bob}} = [2,\ 3,\ 4,\ \ldots]$

- Des utilisateurs



- Objectifs intermédiaires
  - Regrouper les utilisateurs au comportement similaire
    $\Longrightarrow$ Partitionnement d'exemples (clustering)
  - Prédire le comportement futur d'un utilisateur à partir de son groupe
    $\Longrightarrow$ Consensus d'ordonnancements

Solution proposée : un modèle statistique

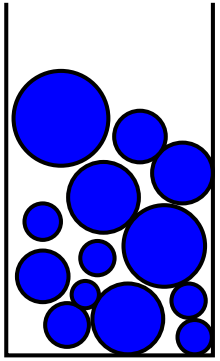
(utilisable avec un très grand nombre de produits et d'utilisateurs)

Plan

1. Modèle statistique
2. Inférence
3. Validation expérimentale

Modèle élémentaire

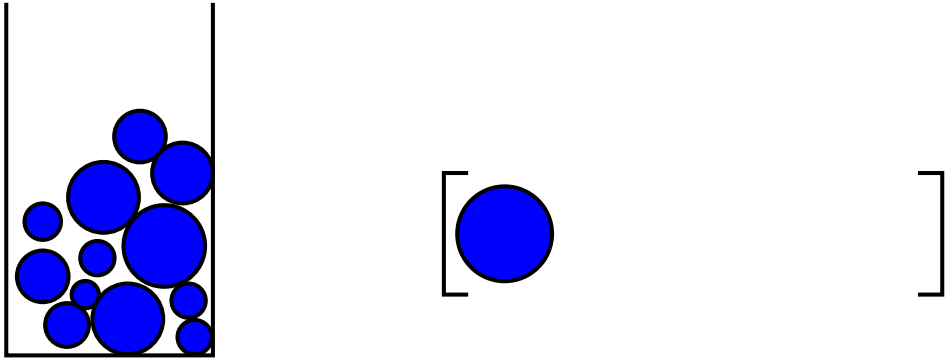
- Modèle de Plackett-Luce [Luce, 1959 ; Plackett, 1975]
  - Distribution élémentaire sur $\mathfrak{S}_n$ (ou top-k)
  - Paramètres
    - ⋆ Un poids q_i pour chaque produit i
  - Loi de probabilité
    - ⋆ $P_{\mathrm{PL}}(\mathfrak{S}_{n-k} \cdot \pi) = \prod_{\ell=1}^{k} \frac{q_{\pi^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q_{\pi^{-1}(\ell')}}$
- Interprétation
  - Tirages sans remise des produits
  - Probabilité de tirer un produit i proportionnelle à q_i



[ ]

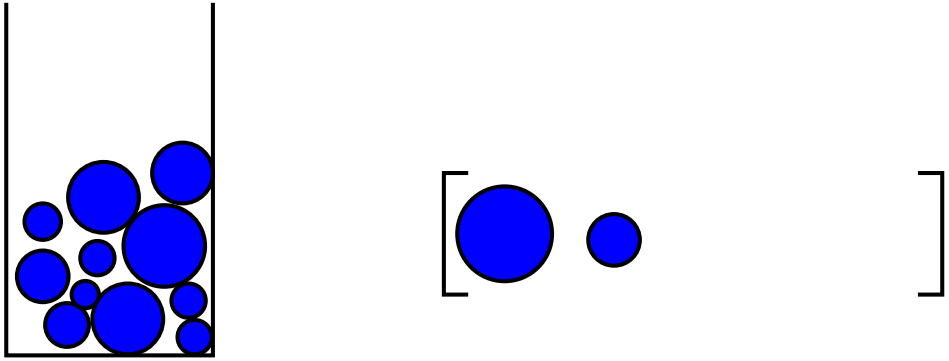
Modèle élémentaire

- Modèle de Plackett-Luce [Luce, 1959 ; Plackett, 1975]
  - Distribution élémentaire sur $\mathfrak{S}_n$ (ou top-k)
  - Paramètres
    - ⋆ Un poids q_i pour chaque produit i
  - Loi de probabilité
    - ⋆ $P_{\mathrm{PL}}(\mathfrak{S}_{n-k} \cdot \pi) = \prod_{\ell=1}^{k} \frac{q_{\pi^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q_{\pi^{-1}(\ell')}}$
- Interprétation
  - Tirages sans remise des produits
  - Probabilité de tirer un produit i proportionnelle à q_i



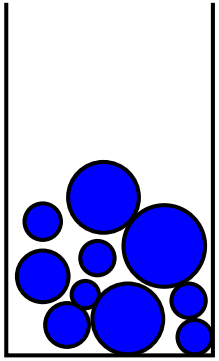
Modèle élémentaire

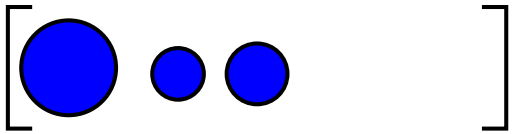
- Modèle de Plackett-Luce [Luce, 1959 ; Plackett, 1975]
  - Distribution élémentaire sur $\mathfrak{S}_n$ (ou top-k)
  - Paramètres
    - ⋆ Un poids q_i pour chaque produit i
  - Loi de probabilité
    - ⋆ $P_{\mathrm{PL}}(\mathfrak{S}_{n-k} \cdot \pi) = \prod_{\ell=1}^{k} \frac{q_{\pi^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q_{\pi^{-1}(\ell')}}$
- Interprétation
  - Tirages sans remise des produits
  - Probabilité de tirer un produit i proportionnelle à q_i



Modèle élémentaire

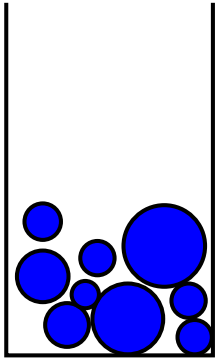
- Modèle de Plackett-Luce [Luce, 1959 ; Plackett, 1975]
  - Distribution élémentaire sur $\mathfrak{S}_n$ (ou top-k)
  - Paramètres
    - ⋆ Un poids q_i pour chaque produit i
  - Loi de probabilité
    - ⋆ $P_{\mathrm{PL}}(\mathfrak{S}_{n-k} \cdot \pi) = \prod_{\ell=1}^{k} \frac{q_{\pi^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q_{\pi^{-1}(\ell')}}$
- Interprétation
  - Tirages sans remise des produits
  - Probabilité de tirer un produit i proportionnelle à q_i

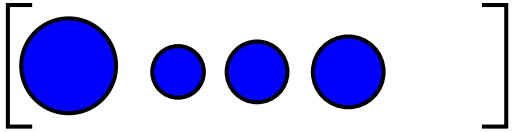




Modèle élémentaire

- Modèle de Plackett-Luce [Luce, 1959 ; Plackett, 1975]
  - Distribution élémentaire sur $\mathfrak{S}_n$ (ou top-k)
  - Paramètres
    - ★ Un poids q_i pour chaque produit i
  - Loi de probabilité
    - ★ $P_{\mathrm{PL}}(\mathfrak{S}_{n-k} \cdot \pi) = \prod_{\ell=1}^{k} \frac{q_{\pi^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q_{\pi^{-1}(\ell')}}$
- Interprétation
  - Tirages sans remise des produits
  - Probabilité de tirer un produit i proportionnelle à q_i

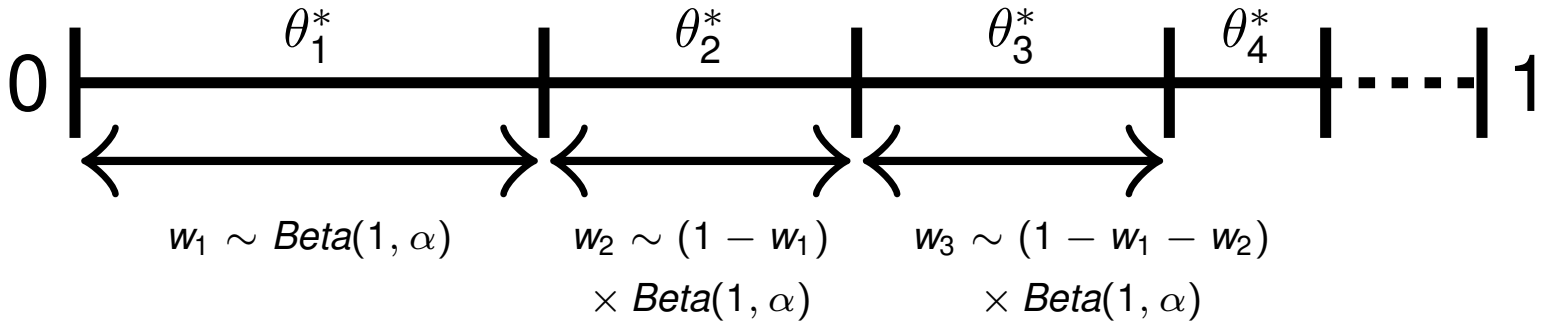




Mixture de modèles

- Processus de Dirichlet [Antoniak, 1974]
  - Distribution sur les lois de distribution sur un ensemble Θ
  - Paramètres
    - ⋆ $\alpha > 0$: paramètre de concentration
    - ⋆ H : loi de distribution moyenne sur Θ
  - Définition (propriété élémentaire)
    - ⋆ Pour toute partition finie mesurable $A_1, \ldots, A_r$ de Θ, $G \sim DP(\alpha, H)$ ssi $(G(A_1), \ldots, G(A_r)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_r))$
- Comportement
  - G est discrète
  - m tirages iid selon G donnent $\mathcal{O}(\alpha \log m)$ valeurs différentes de θ
  - stick-breaking process

$$\begin{array}{l} G \sim DP(\alpha, H) \\ \theta_i \sim G \end{array} \quad \equiv \quad \begin{array}{ll} \theta^*_\ell \sim H & \\ \beta_\ell \sim Beta(1, \alpha) & w_\ell = \beta_\ell \prod_{\ell'=1}^{\ell-1}(1-\beta_{\ell'}) \\ z_i \sim Mult(\boldsymbol{w}) & \theta_i = \theta^*_{z_i} \end{array}$$



Mixture de modèles

- Processus de Dirichlet [Antoniak, 1974]
  - Distribution sur les lois de distribution sur un ensemble Θ
  - Paramètres
    - $\alpha > 0$: paramètre de concentration
    - H : loi de distribution moyenne sur Θ
  - Définition (propriété élémentaire)
    - Pour toute partition finie mesurable $A_1, \ldots, A_r$ de Θ, $G \sim DP(\alpha, H)$ ssi $(G(A_1), \ldots, G(A_r)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_r))$
- Comportement
  - G est discrète
  - m tirages iid selon G donnent $\mathcal{O}(\alpha \log m)$ valeurs différentes de θ
  - stick-breaking process

$$\begin{array}{l} G \sim DP(\alpha, H) \\ \theta_i \sim G \end{array} \quad \equiv \quad \begin{array}{ll} \theta^*_\ell \sim H & \\ \beta_\ell \sim Beta(1, \alpha) & w_\ell = \beta_\ell \prod_{\ell'=1}^{\ell-1}(1-\beta_{\ell'}) \\ z_i \sim Mult(\boldsymbol{w}) & \theta_i = \theta^*_{z_i} \end{array}$$

Application

Partitionnement d'exemples

Modèle global pour le Filtrage Collaboratif

- Approche

$$G \sim DP(\alpha, Dir(\boldsymbol{\beta}))$$
$$\boldsymbol{q}_i \sim G$$
$$\mathfrak{S}_{n-k} \cdot \pi_i \sim PL(\boldsymbol{q}_i)$$

  - Processus de Dirichlet
    - Partitionne les exemples
    - Contrôle les paramètres de chaque partition
  - Modèle de Plackett-Luce
    - Génère les exemples dans un groupe donné

- Paramètres
  - $\alpha > 0$
    - Contrôle la taille de la partition
  - $\boldsymbol{\beta} = \beta \cdot (1, \ldots, 1), \beta > 0$
    - Contrôle la parcimonie des poids sur les produits
    - Tous les produits sont a priori équivalents

Plan

1 Modèle statistique

2 Inférence

3 Validation expérimentale

Partitionner les exemples

- Échantillonneur de Gibbs [Neal, 2000]

for $t = 1$ **to** T_{Gibbs} **do**
  /* Rééchantillonner l'affectation à un groupe */
  for $i = 1$ **to** m **do**
    Tirer z_i selon $\mathbb{P}(z_i = z \mid \pi_i, \boldsymbol{m}_{-i}, \{\boldsymbol{q}^*\})$

$$\propto \begin{cases} \frac{m_{-i,z}}{n-1+\alpha} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) & \text{si } m_{-i,z} \neq 0 \\ \frac{\alpha}{n-1+\alpha} \int_{\boldsymbol{q}} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}) \mathbb{P}_{\text{Dir}}(\boldsymbol{q} \mid \beta) d\boldsymbol{q} & \text{sinon} \end{cases}$$

    if $m_z = 1$ **then**
      Tirer $\boldsymbol{q}_i = \boldsymbol{q}_{z_i}^*$ selon $\mathbb{P}(\boldsymbol{q}_i \mid \pi_i, \beta) \propto \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_i) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_i \mid \beta)$
    end if
  end for
  /* Rééchantillonner les paramètres des groupes */
  for $z = 1$ **to** C **do**
    if $m_z > 1$ **then**
      Tirer $\boldsymbol{q}_z^*$ selon $\mathbb{P}(\boldsymbol{q}_z^* \mid \{\pi_i : z_i = z\}, \beta) \propto \prod_{i:z_i=z} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_z^* \mid \beta)$
    end if
  end for
end for

Partitionner les exemples

- Échantillonneur de Gibbs [Neal, 2000]

for $t = 1$ **to** T_{Gibbs} **do**
  /* Rééchantillonner l'affectation à un groupe */
  for $i = 1$ **to** m **do**
    Tirer z_i selon $\mathbb{P}(z_i = z \mid \pi_i, \boldsymbol{m}_{-i}, \{\boldsymbol{q}^*\})$

$$\propto \begin{cases} \frac{m_{-i,z}}{n-1+\alpha} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) & \text{si } m_{-i,z} \neq 0 \\ \frac{\alpha}{n-1+\alpha} \int_{\boldsymbol{q}} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}) \mathbb{P}_{\text{Dir}}(\boldsymbol{q} \mid \beta) d\boldsymbol{q} & \text{sinon} \end{cases}$$

    if $m_z = 1$ **then**
      Tirer $\boldsymbol{q}_i = \boldsymbol{q}_{z_i}^*$ selon $\mathbb{P}(\boldsymbol{q}_i \mid \pi_i, \beta) \propto \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_i) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_i \mid \beta)$
    end if
  end for
  /* Rééchantillonner les paramètres des groupes */
  for $z = 1$ **to** C **do**
    if $m_z > 1$ **then**
      Tirer $\boldsymbol{q}_z^*$ selon $\mathbb{P}(\boldsymbol{q}_z^* \mid \{\pi_i : z_i = z\}, \beta) \propto \prod_{i:z_i=z} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_z^* \mid \beta)$
    end if
  end for
end for

Partitionner les exemples

- Échantillonneur de Gibbs [Neal, 2000]

for $t = 1$ **to** T_{Gibbs} **do**
  /* Rééchantillonner l'affectation à un groupe */
  for $i = 1$ **to** m **do**
    Tirer z_i selon $\mathbb{P}(z_i = z \mid \pi_i, \boldsymbol{m}_{-i}, \{\boldsymbol{q}^*\})$

$$\propto \begin{cases} \frac{m_{-i,z}}{n-1+\alpha} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) & \text{si } m_{-i,z} \neq 0 \\ \frac{\alpha}{n-1+\alpha} \int_{\boldsymbol{q}} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}) \mathbb{P}_{\text{Dir}}(\boldsymbol{q} \mid \beta) d\boldsymbol{q} & \text{sinon} \end{cases}$$

    if $m_z = 1$ **then**
      Tirer $\boldsymbol{q}_i = \boldsymbol{q}_{z_i}^*$ selon $\mathbb{P}(\boldsymbol{q}_i \mid \pi_i, \beta) \propto \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_i) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_i \mid \beta)$
    end if
  end for
  /* Rééchantillonner les paramètres des groupes */
  for $z = 1$ **to** C **do**
    if $m_z > 1$ **then**
      Tirer $\boldsymbol{q}_z^*$ selon $\mathbb{P}(\boldsymbol{q}_z^* \mid \{\pi_i : z_i = z\}, \beta) \propto \prod_{i:z_i=z} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_z^* \mid \beta)$
    end if
  end for
end for

Partitionner les exemples

- Échantillonneur de Gibbs [Neal, 2000]

$$\frac{(n-k)!}{n!}$$

for $t = 1$ **to** T_{Gibbs} **do**
  /* Rééchantillonner l'affectation à un groupe */
  for $i = 1$ **to** m **do**
    Tirer z_i selon $\mathbb{P}(z_i = z \mid \pi_i, \boldsymbol{m}_{-i}, \{\boldsymbol{q}^*\})$

$$\propto \begin{cases} \frac{m_{-i,z}}{n-1+\alpha} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) & \text{si } m_{-i,z} \neq 0 \\ \frac{\alpha}{n-1+\alpha} \int_{\boldsymbol{q}} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}) \mathbb{P}_{\text{Dir}}(\boldsymbol{q} \mid \beta) d\boldsymbol{q} & \text{sinon} \end{cases}$$

    if $m_z = 1$ **then**
      Tirer $\boldsymbol{q}_i = \boldsymbol{q}_{z_i}^*$ selon $\mathbb{P}(\boldsymbol{q}_i \mid \pi_i, \beta) \propto \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_i) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_i \mid \beta)$
    end if
  end for
  /* Rééchantillonner les paramètres des groupes */
  for $z = 1$ **to** C **do**
    if $m_z > 1$ **then**
      Tirer $\boldsymbol{q}_z^*$ selon $\mathbb{P}(\boldsymbol{q}_z^* \mid \{\pi_i : z_i = z\}, \beta) \propto \prod_{i:z_i=z} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_z^* \mid \beta)$
    end if
  end for
end for

Partitionner les exemples

- Échantillonneur de Gibbs [Neal, 2000]

$$\frac{(n-k)!}{n!}$$

for $t = 1$ **to** T_{Gibbs} **do**
  /* Rééchantillonner l'affectation à un groupe */
  for $i = 1$ **to** m **do**
    Tirer z_i selon $\mathbb{P}(z_i = z \mid \pi_i, \boldsymbol{m}_{-i}, \{\boldsymbol{q}^*\})$

$$\propto \begin{cases} \frac{m_{-i,z}}{n-1+\alpha} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) & \text{si } m_{-i,z} \neq 0 \\ \frac{\alpha}{n-1+\alpha} \int_{\boldsymbol{q}} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}) \mathbb{P}_{\text{Dir}}(\boldsymbol{q} \mid \beta) d\boldsymbol{q} & \text{sinon} \end{cases}$$

    if $m_z = 1$ **then**
      Tirer $\boldsymbol{q}_i = \boldsymbol{q}_{z_i}^*$ selon $\mathbb{P}(\boldsymbol{q}_i \mid \pi_i, \beta) \propto \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_i) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_i \mid \beta)$
    end if
  end for
  /* Rééchantillonner les paramètres des groupes */
  for $z = 1$ **to** C **do**
    if $m_z > 1$ **then**
      Tirer $\boldsymbol{q}_z^*$ selon $\mathbb{P}(\boldsymbol{q}_z^* \mid \{\pi_i : z_i = z\}, \beta) \propto \prod_{i:z_i=z} \mathbb{P}_{\text{PL}}(\pi_i \mid \boldsymbol{q}_z^*) \mathbb{P}_{\text{Dir}}(\boldsymbol{q}_z^* \mid \beta)$
    end if
  end for
end for

Paramètres d'un groupe

Groupe contenant un utilisateur

- Loi a posteriori des paramètres
  - $\mathbb{P}(\boldsymbol{q}_i \mid \pi_i,\ \beta) \propto \prod_{\ell=1}^{k} \frac{q_{i,\pi_i^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q_{i,\pi_i^{-1}(\ell')}} \prod_{j=1}^{n} q_{i,j}^{\beta-1}$
- Marginales par rapport aux produits appartenant au top
  - Pour tout $\ell < k$, $q_{i,\pi_i^{-1}(\ell)} \mid \pi_i,\ q_{i,\pi_i^{-1}(1)},\ \ldots,\ q_{i,\pi_i^{-1}(\ell-1)},\ \beta$
    $\sim (1 - \sum_{\ell'=1}^{\ell-1} q_{i,\pi_i^{-1}(\ell')}) \times Beta(\beta+1, (M-\ell)\cdot\beta)$
- Loi conditionnelle des produits restants
  - $q_{i,\pi_i^{-1}(k+1)},\ \ldots,\ q_{i,\pi_i^{-1}(n)} \mid \pi_i,\ q_{i,\pi_i^{-1}(1)},\ \ldots,\ q_{i,\pi_i^{-1}(k)},\ \beta$
    $\sim (1 - \sum_{\ell=1}^{k} q_{i,\pi_i^{-1}(\ell)}) \times Dir(\beta,\ \ldots,\ \beta)$

Paramètres d'un groupe

Groupe contenant plusieurs utilisateurs

- Loi a posteriori des paramètres
  - $\nu_j = \#\{i : z_i = z,\ \pi_i(j) \leqslant k\}$
  - $J = \{j : \nu_j > 0\}$: produits présents dans au moins un top
  - $\mathbb{P}(\boldsymbol{q}_z^* \mid \{\pi_i : z_i = z\},\ \beta) \propto \prod_{i:z_i=z} \prod_{\ell=1}^{k} \frac{q^*_{z,\pi_i^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q^*_{z,\pi_i^{-1}(\ell')}} \prod_{j=1}^{n} {q^*_{z,j}}^{\beta-1}$
- Marginale des produits appartenant à J
  - $\{q^*_{z,j} : j \in J\} \mid \{\pi_i : z_i = z\},\ \beta \sim$
    $$\frac{\prod_{j\in J} {q^*_{z,j}}^{\nu_j+\beta-1}}{\prod_{i:z_i=z}\prod_{\ell=1}^{k}\left(1-\sum_{\ell'=1}^{\ell-1} q^*_{z,\pi_i^{-1}(\ell')}\right)} \left(1-\sum_{j\in J} q^*_{z,j}\right)^{(n-\#J)\beta-1} \mathbb{1}_{\left\{0<\sum_{j\in J} q^*_{z,j}<1\right\}}$$
- Loi conditionnelle des produits n'appartenant pas à J
  - $\{q^*_{z,j} : j \notin J\} \mid \{q^*_{z,j} : j \in J\},\ \beta \sim \left(1-\sum_{j\in J} q^*_{z,j}\right) \times Dir(\beta,\ \ldots,\ \beta)$

Paramètres d'un groupe

Groupe contenant plusieurs utilisateurs

- Loi a posteriori des paramètres
  - $\nu_j = \#\{i : z_i = z,\ \pi_i(j) \leqslant k\}$
  - $J = \{j : \nu_j > 0\}$: produits présents dans au moins un top
  - $\mathbb{P}(\boldsymbol{q}_z^* \mid \{\pi_i : z_i = z\},\ \beta) \propto \prod_{i:z_i=z} \prod_{\ell=1}^{k} \frac{q^*_{z,\pi_i^{-1}(\ell)}}{\sum_{\ell'=\ell}^{n} q^*_{z,\pi_i^{-1}(\ell')}} \prod_{j=1}^{n} {q^*_{z,j}}^{\beta-1}$
- Marginale des produits appartenant à J
  - $\{q^*_{z,j} : j \in J\} \mid \{\pi_i : z_i = z\},\ \beta \sim$

$$\frac{\prod_{j\in J} {q^*_{z,j}}^{\nu_j+\beta-1}}{\prod_{i:z_i=z} \prod_{\ell=1}^{k} \left(1 - \sum_{\ell'=1}^{\ell-1} q^*_{z,\pi_i^{-1}(\ell')}\right)} \left(1 - \sum_{j\in J} q^*_{z,j}\right)^{(n-\#J)\beta-1} \mathbb{1}_{\left\{0<\sum_{j\in J} q^*_{z,j}<1\right\}}$$

$\Longrightarrow$ Métropolis Hasting

- Loi conditionnelle des produits n'appartenant pas à J
  - $\{q^*_{z,j} : j \notin J\} \mid \{q^*_{z,j} : j \in J\},\ \beta \sim \left(1 - \sum_{j\in J} q^*_{z,j}\right) \times Dir(\beta,\ \ldots,\ \beta)$

Plan

1. Modèle statistique
2. Inférence
3. Validation expérimentale

Protocole expérimental

- Question
  - L'inférence retrouve-t-elle le vrai partitionnement ?
    - En présence d'un grand nombre d'utilisateurs ($m = 10^4$)
    - En présence d'un grand nombre de produits ($n = 10^6$)
- Données synthétiques
  - $C = 5$ groupes d'utilisateurs
  - Un groupe
    - Modèle de Plackett-Luce
    - Poids : $sort(\boldsymbol{w}) = \frac{\mu}{\mu Z}, \ldots, \frac{1}{\mu Z}, \frac{\varepsilon}{Z}, \ldots, \frac{\varepsilon}{Z}$
  - Intersection entre les sous-ensembles de produits intéressants
    - `disjoint` : intersection vide
    - `meme_debut` : même $\frac{\mu}{2}$ premiers produits intéressants pour les groupe 1 et 2 et les groupes 3 et 4
    - `meme_fin` : même $\frac{\mu}{2}$ derniers produits intéressants pour les groupe 1 et 2 et les groupes 3 et 4

Critère de réussite

- Variation d'importance (VI) [Meilă, 2007]

$$\mathrm{VI}(\mathcal{C}, \hat{\mathcal{C}}) \stackrel{\text{def}}{=} \mathrm{H}(\mathcal{C} \mid \hat{\mathcal{C}}) + \mathrm{H}(\hat{\mathcal{C}} \mid \mathcal{C})$$

  - Distance entre le partitionnement cible $\mathcal{C}$ et le partitionnement appris $\hat{\mathcal{C}}$
  - Fondée sur l'entropie conditionnelle H
  - À minimiser
  - Liée à l'information mutuelle
    - $2 \cdot \mathrm{IM}(\mathcal{C}, \hat{\mathcal{C}}) + \mathrm{VI}(\mathcal{C}, \hat{\mathcal{C}}) = \mathrm{H}(\mathcal{C}) + \mathrm{H}(\hat{\mathcal{C}})$

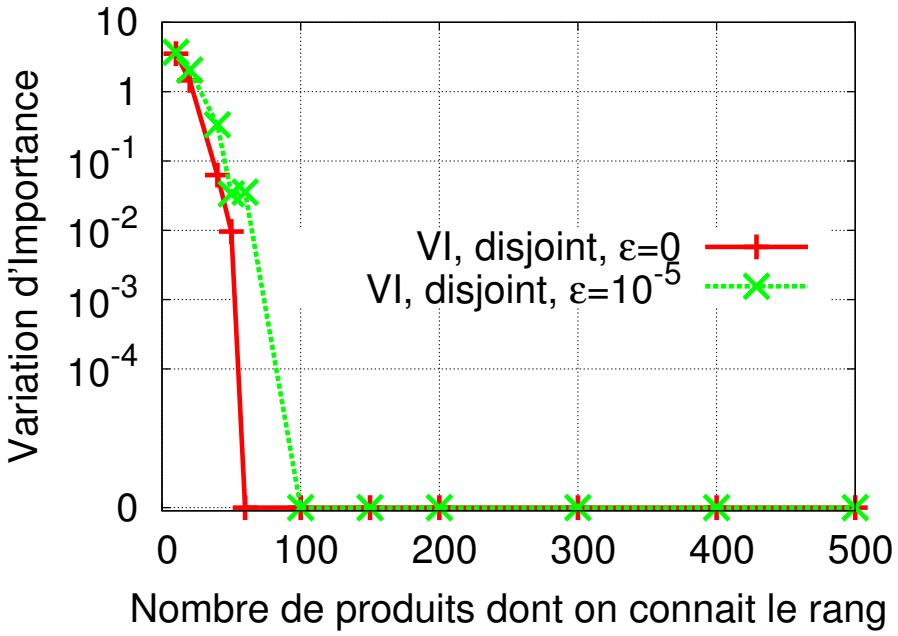
- Intersection non-vide entre les sous-ensembles de produits intéressants

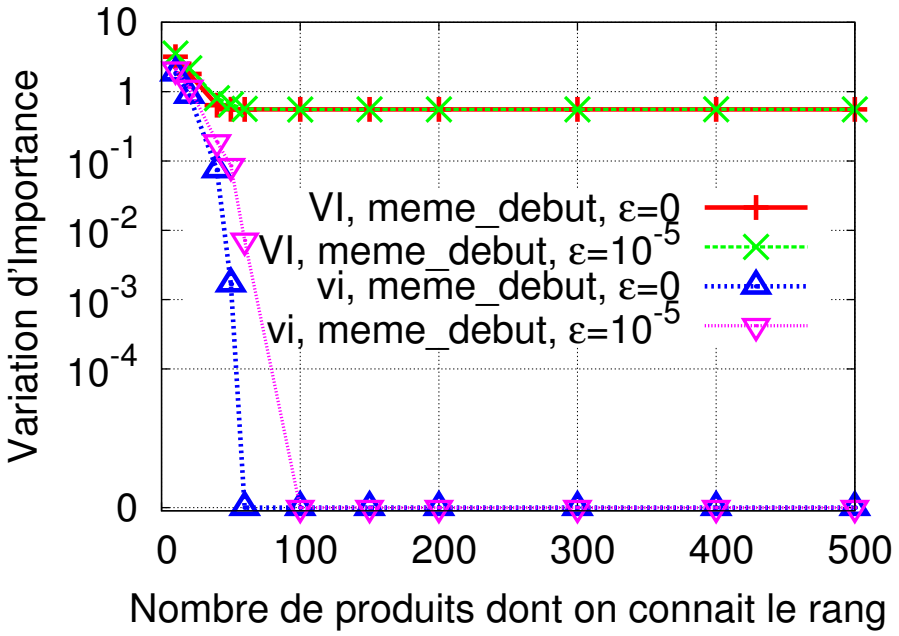
$$\mathrm{vi}(\mathcal{C}, \hat{\mathcal{C}}) \stackrel{\text{def}}{=} \mathrm{VI}(\mathcal{C}', \hat{\mathcal{C}})$$

  - $\mathcal{C}'$: partitionnement en trois groupes ($\{1, 2\}$, $\{3, 4\}$, et $\{5\}$)

Variation d'importance en fonction de la taille de top

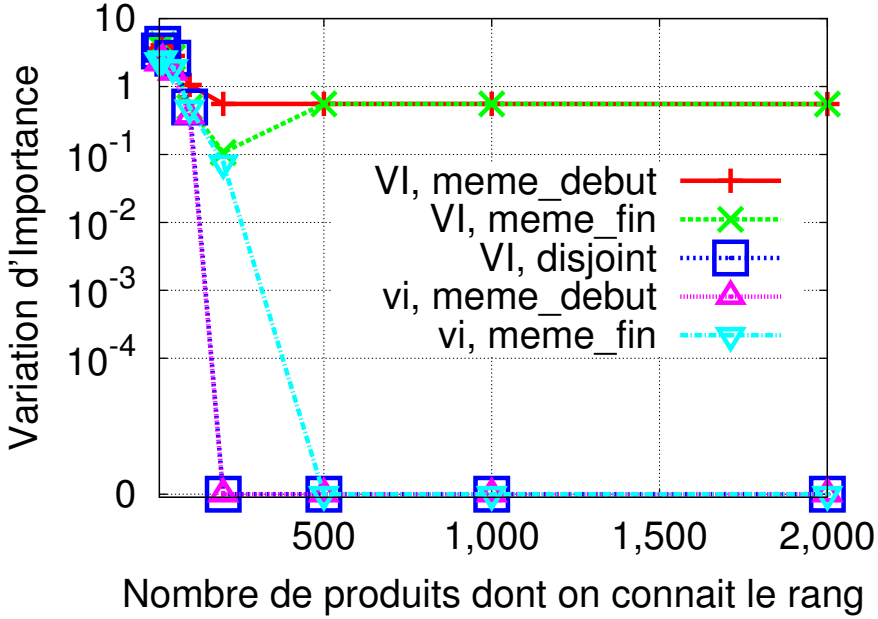
100 produits intéressants par groupe





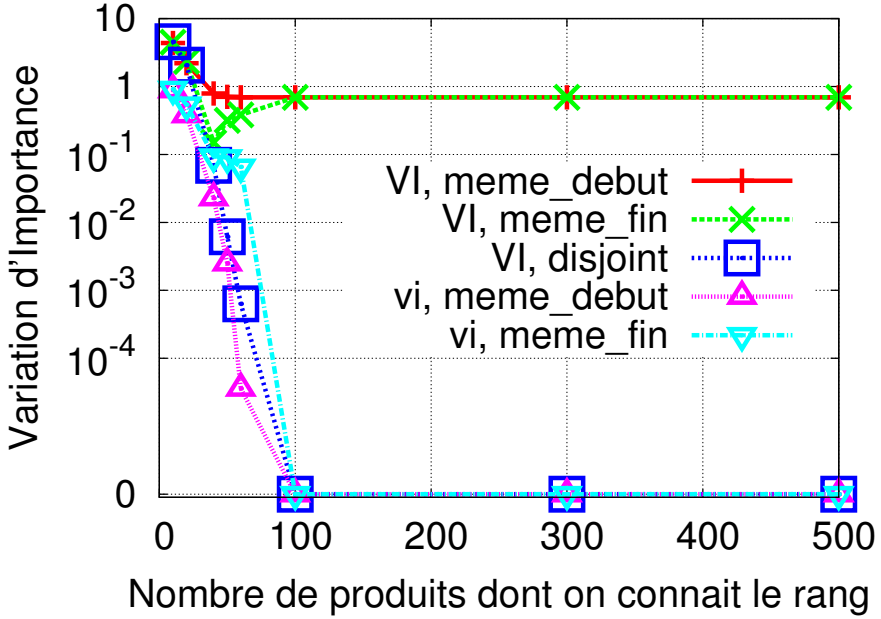
- Remarques
  - Retrouve le partitionnement initial dans le cas `disjoint`
  - Retrouve le partitionnement en trois groupes dans le cas `meme_debut`

Variation d'importance en fonction de la taille de top



1000 produits intéressants par groupe

$\varepsilon = 0$



$C = 10$ groupes

$\varepsilon = 0$

- Remarques
  - Passe à l'échelle en terme de nombre de produits intéressants
  - Retrouve le partitionnement initial avec $C = 10$ groupes

Conclusion et Perspectives

Contributions

- Formalisation du Filtrage Collaboratif comme un problème de consensus d'ordonnancements
  - Prend en compte l'ordre des achats
- Un modèle statistique de consensus d'ordonnancements
  - Fondé sur un processus de Dirichlet et le modèle de Plackett-Luce
  - Inféré par échantillonneur de Gibbs et Métropolis Hasting
  - Utilisable avec 10^4 utilisateurs / 10^6 produits
  - Faible coût de calcul (moins de 2 min. en conditions normales)

- Perspectives
  - Utilisation pour la prédiction de futurs achats
  - Couplage avec algorithmes de complétion de matrices

[Abernethy et al., 2009]

Bibliographie

J. Abernethy, F. Bach, T. Evgeniou, J.-P. Vert *A New Approach to Collaborative Filtering : Operator Estimation with Spectral Regularization*. JMLR'09
C. E. Antoniak *Mixtures of dirichlet processes with applications to bayesian nonparametric problems*. Annals of Statistics, 1974
R. D. Luce *Individual choice behavior : A theoretical analysis.*. Wiley, 1959
M. Meilă *Comparing clusterings – an information based distance*. J. of Multivariate Analysis, 2007
R. L. Plackett *The analysis of permutations*. J. of the Royal Stat. Soc., 1975

Fin du post-doc : 31 août 2011
$\Longrightarrow$ à la recherche d'un post-doc (hors région parisienne) pour après