HAL open science

Extraction de systèmes de consommation alimentaire utilisant la Nonnegative Matrix Factorization (NMF) pour l'évaluation des choix alimentaires.

Mélanie Zetlaoui, Stéphan Clémençon, Max Feinberg, Philippe Verger

▶ **To cite this version:**

Mélanie Zetlaoui, Stéphan Clémençon, Max Feinberg, Philippe Verger. Extraction de systèmes de consommation alimentaire utilisant la Nonnegative Matrix Factorization (NMF) pour l'évaluation des choix alimentaires.. 42èmes Journées de Statistique, 2010, Marseille, France, France. inria-00494768

HAL Id: inria-00494768

https://inria.hal.science/inria-00494768v1

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Extraction de systèmes de consommation alimentaire utilisant la *Nonnegative Matrix Factorization* (NMF) pour l'évaluation des choix alimentaires.

Mélanie Zetlaoui[1] & Stéphan Clémençon[2] & Max Feinberg[1] & Philippe Verger[1]

[1] *Institut de la Recherche Agronomique (INRA), Métarisk, 16 rue Claude Bernard, 75231 Paris cedex 05 (France)*

[2] *Telecom-ParisTech, 46 rue Barrault, 75634 Paris cedex 13 (France)*

Résumé

Dans les pays occidentaux où l'approvisionnement alimentaire est satisfaisant, les consommateurs agencent leur régime au moyen d'un grand nombre d'aliments. L'objectif de ce travail est d'étudier comment une technique récente en analyse de variables latentes, la *Nonnegative Matrix Factorization* (NMF), peut être appliquée aux données de consommation pour comprendre cet agencement. De telles données sont positives par nature et de grande dimension. Le modèle statistique NMF ici construit fournit alors une représentation des données par des variables latentes positives, appelées systèmes de consommation, qui sont en nombre très petit. L'approche NMF favorisant la sparsité, les systèmes de consommations obtenus sont de plus facilement interprétables. En application, des résultats numériques à partir d'une enquête française de consommation, sont donnés. Une méthode de clustering, basée sur la méthode des k-means dans le sous-espace des systèmes de consommation, permet de construire des groupes de consommateurs facilement interprétables par les nutritionnistes.

Abstract

In Western countries where food supply is satisfactory, consumers organize their diets around a large combination of foods. It is the purpose of this work to examine how recent *nonnegative matrix factorization* (NMF) techniques can be applied to food consumption data in order to understand this combination. Such data are nonnegative by nature and of high dimension. The NMF model provides a representation of data by nonnegative latent variables, called consumption systems, in a small number. As the NMF approach may encourage the sparsity, the resulting consumption systems are easily interpretable. As application, numerical results based on a french survey, are displayed. A clustering based on k-means method is also achieved in the obtained latent consumption space, in order to recover food consumption patterns easily usable for nutritionists.

Mots-clés : Non-negative Matrix Factorization (NMF), modèle à variables latentes, données de consommation, réduction de la dimension, données sparses, algorithme de descente de gradient, clustering.

1 Contexte

L'évaluation du risque alimentaire est un problème important dans le cadre de la santé publique. Ce problème nécessite en particulier de comprendre les choix et les comportements alimentaires des consommateurs. Certains aspects de cette question peuvent être mesurés par le biais d'enquêtes de consommation menées auprès d'un échantillon d'une population. Notre travail se base sur une enquête de consommation s'appuyant sur de nombreux aliments (880 aliments regroupés en 44 groupes), appelée INCA (Volatier, 2000). Elle a été conduite en 1999 par l'Agence Fraçaise de Sécurité Sanitaire des Aliments (AFSSA) auprès d'un échantillon représentatif de 3003 individus de la population française. C'est une enquête individuelle puisque les participants reportent leur propre consommation pendant une semaine, chaque type de repas étant reporté séparément.

Même si un grand nombre d'aliments peuvent potentiellement intervenir dans le régime d'un individu, tous les comportements ne sont pas observés en pratique. Certains aliments sont combinés ou bien alternés suivant les goûts et/ou les habitudes socio-culturelles. Il est donc réaliste de supposer que le régime d'un individu est une combinaison linéaire d'aliments spécifiques. Ces facteurs sous-jacents peuvent être interprétés comme des variables latentes, qu'on appelle ici des systèmes de consommations (SC).

L'*analyse en composantes principales* (ACP) ou l'*analyse factorielle* font partie des méthodes qui permettent d'obtenir de telles variables latentes. Elles sont conçues pour traiter des données supposées Gaussiennes (les variables latentes et le bruit sont supposés gaussiens) ce qui réduit considérablement le champ d'application de ces méthodes. Les données de consommation, étant positives et contenant beaucoup de zéros, ne sont de tout évidence pas gaussiennes, cette hypothèse sur les facteurs latents compromettrait de plus leur interprétabilité.

Pour ces raisons, une méthode, appelée *Nonnegative Matrix Factorization* (NMF), a récemment émergée (Lee et Seung, 1999) comme une alternative aux techniques traditionnelles en analyse de variables latentes. Ici, cette méthode est développée pour permettre d'extraire une base de consommation sous-jacente, i.e les systèmes de consommations, en nombre très petit, donnant ainsi une nouvelle représentation des consommations individuelles. De plus, la procédure NMF favorise la sparsité, ce qui rend les variables latentes facilement interprétables au moyen d'un petit nombre d'aliments. Afin d'identifier des groupes de consommateurs ayant des comportements similaires, la méthode NMF sur les données de consommation doit être complétée par un clustering sur les individus dans l'espace latent. Dans ce papier, nous présentons dans la section 2 le modèle NMF et la procédure permettant d'obtenir les facteurs latents. Nous implémentons ensuite, dans la section 3, la méthode sur la base de donnée INCA, permettant d'illustrer la méthode et ses propriétés.

2 Le modèle NMF

2.1 Hypothèses et notations

Nous supposons ici que les choix alimentaires d'un individu sont décrits par une collection de F aliments indexés par $f = 1, \ldots, F$. Plus précisément, le régime d'un individu peut être modélisé par un vecteur de longueur F, $Q = (Q^{(1)}, \ldots, Q^{(F)})$, dont les valeurs appartiennent à $\mathbb{R}_+^F$, le $f^{\text{ième}}$ élement $Q^{(f)}$ de Q indiquant la quantité de l'aliment f consommée. La consommation des aliments dépendant largement de leur nature et de leur consistance (consommer un litre d'eau n'est pas équivalent à un kilogramme de viande), un effet d'échelle peut être observé entre les F variables. Pour éviter cet effet d'échelle, chaque élément $Q^{(f)}$ est normalisé par sa variance. Pour chaque $f = 1, \ldots, F$, on pose $v^{(f)} = \frac{Q^{(f)}}{\sigma^{(f)}}$ où $\sigma^{(f)\,2} = E[(Q^{(f)} - E[Q^{(f)}])^2]$.

Supposons que :

$$v = Wh + \epsilon \tag{1}$$

où $v = (v^{(1)}, \ldots v^{(F)})$, West une matrice de taille $F \times K$ dont les éléments sont positifs, $h = (h^1, \ldots, h^K)$ un vecteur de longueur K dans $\mathbb{R}_+^K$ et $\epsilon = (\epsilon^{(1)}, \ldots, \epsilon^{(F)})$ un vecteur aléatoire Gaussien de moyenne 0 et de covariance Γ. Le nombre K de composantes latentes est habituellement choisi plus petit que F, réduisant ainsi la dimension des données.

De plus, nous supposons que la matrice des systèmes de consommation W est de rang plein K et satisfait la contrainte de normalité

$$\sum_{f=1}^{F} W_{f\,k} = 1 \tag{2}$$

Cette contrainte permet d'interpréter les colonnes de W. Les systèmes de consommation peuvent être vus comme des combinaisons deterministes d'aliments dans un contexte socio-économique d'une enquête de consommation et d'une base utilisée par les consommateurs pour agencer leur propre régime. $W_{f\,k}$ peut ainsi être renvoyé à la proportion de l'aliment f au sein du système de consommation k.

Dans la suite, nous supposons qu'on observe N vecteurs indépendant des quantités normalisées, $v_n = Wh_n + \epsilon_n$, for $n = 1, \ldots, N$, où les ϵ_n sont i.i.d et suivent une loi normale $\mathcal{N}(0, \Gamma)$. L'ensemble de la consommation alimentaire des N individus peut être alors représenté en untilisant la notation matricielle :

$$V = WH + E \tag{3}$$

où V et E sont des matrices de taille $F \times N$ et H une matrice $K \times N$, dont les colonnes sont $(v_1, \ldots, v_N)$, $(\epsilon_1, \ldots, \epsilon_N)$ et $(h_1, \ldots, h_N)$ respectivement.

2.2 La procédure NMF

Étant donnée la structure additive du modèle, le principe de l'algorithme NMF consiste à minimiser la somme des carrés résiduelle

$$\begin{aligned} D_K(V,(W,H)) &= \|V - WH\|^2 \\ &= \sum_{n=1}^{N}\sum_{f=1}^{F}\left(v_{fn} - \sum_{k=1}^{K} W_{fk}H_{kn}\right)^2 \end{aligned}$$

sur l'ensemble des couples (W,H) dans l'ensemble des matrices $\mathcal{M}_{F,K}(\mathbb{R}_+) \times \mathcal{M}_{K,N}(\mathbb{R}_+)$ sous la contrainte $\sum_{f=1}^{F} W_{fk} = 1$pour tout $k = 1,\ldots,K$.

Se basant sur cette fonction de coût, Lee et Seung (2001) proposent l'algorithme multiplicatif suivant :

$$W_{fk} \leftarrow W_{fk}\frac{[VH^t]_{fk}}{[WHH^t]_{fk}}, \qquad H_{kn} \leftarrow H_{kn}\frac{[W^tV]_{kn}}{[W^tWH]_{kn}}$$

Pour vérifier la contrainte donnée par l'équation (2), les matrices W et H sont renormalisées à chaque étape de la façon suivante :

$$\begin{aligned} H_{kn} &\leftarrow \frac{H_{kn}}{\sum_{f'=1}^{F} W_{f'k}} \\ W_{fk} &\leftarrow \frac{W_{fk}}{\sum_{f'=1}^{F} W_{f'k}} \end{aligned}$$

L'algorithme multiplicatif est une méthode de descente de gradient assurant la positivité des composantes de W et H à chaque étape.

3 Résultats

3.1 Implémentation de la procédure NMF sur la base INCA

Dans cette partie, la méthode NMF est appliquée à la base INCA. Cette méthode nécessite de choisir au préalable le nombre K de systèmes de consommation, qui est typiquement très petit par rapport au nombre de produits F. Pour déterminer K, un premier travail a consisté à tracer la courbe des sommes des carrés résiduelles en fonction de K. Celle-ci ne donnant pas d'indication claire (la courbe décroissant de façon continue vers zéro), il a été décidé de choisir K en fonction du taux de sparsité des systèmes de consommation obtenus, et donc de leur interprétabilité. En comparant les résultats pour plusieurs valeurs de K, on a choisi K égal à 10.

La figure 1 représente les éléments de la matrice W pour K égale à 10. Les systèmes de consommation apparaissent en colonne et les groupes d'aliments en ligne. Plus la

proportion du groupe d'aliment dans un SC est grande plus la cellule correspondante dans la table est foncée. Les proportions nulles sont représentées par du blanc. Les SC obtenus sont déterminés par quelques produits, donnant une vue synthétique des différents comportements alimentaires présents dans la population française. Les produits intervenant au sein de chaque SC, sont ainsi associés dans le régime d'un individu. Ces produits sont soit consommés en même temps ou bien alternés tout en faisant d'un même comportement alimentaire.

3.2 Clustering

Cette nouvelle représentation des consommations individuelles par des SC permet de fournir des clusters facilement interprétables en termes de groupes de consommateurs et donc de groupes à risque. Des techniques de clustering utilisant la représentation NMF (Xu et Al. 2003; Ding et Al. 2008) consistent à regrouper les individus en K clusters, $\mathcal{C}_1, \ldots, \mathcal{C}_K$, où chaque individu, v_n, est affecté au cluster $\mathcal{C}_k$ si :

$$k = \text{argmax}\{h_{k'n}; k' = 1, \ldots, K\} \tag{4}$$

Cette méthode fait correspondre à chaque individu le SC qui a le plus de poids dans son régime. Or, en représentant ces groupes en projetant das un plan contenant deux SC choisis, on constate que le régime d'un individu n'est pas, pour beaucoup d'entre eux, régi par un seul régime.

C'est pourquoi on opte plutôt pour la méthode des k-means dans la base latente des SC, qui regroupe les individus les plus proches en terme de distance euclidienne. Les résultats montrent que pourchaque cluster, un seul SC est prédominant mais certains d'entre eux sont représentés par deux SC.

Bibliographie

[1] Ding, C., Li, T., Luo, D., Peng, W. (2008). Posterior probabilistic clustering using nmf. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, NY, USA, pp. 831–832.

[2] Lee, D. D., Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. Nature 401, 788–791.

[3] Lee, D. D., Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In: Advances in Neural and Information Processing Systems 13. pp. 556–562.

[4] Volatier, J.-L. (2000). Enquête INCA (Individuelle et Nationale sur les Consommations Alimentaires), TEC&DOC Edition. Lavoisier, Paris.

[5] Xu, W., Liu, X., Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, New York, NY, USA, pp. 267–273.

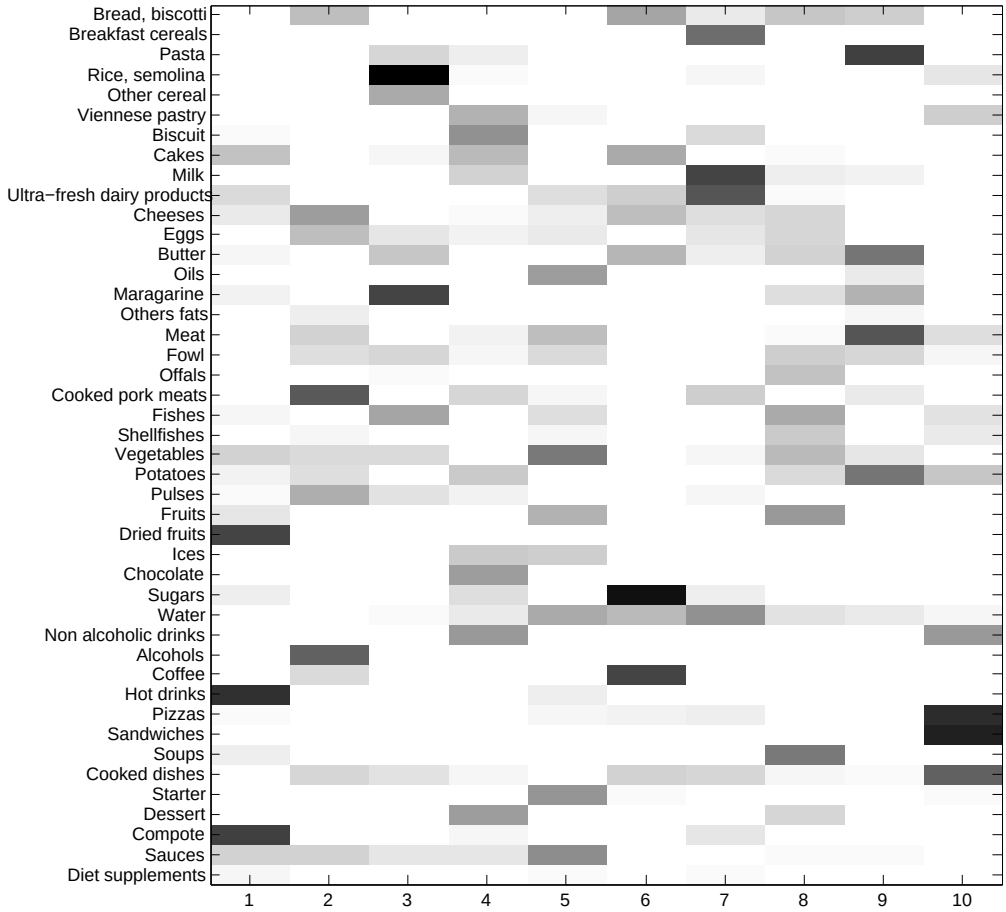

Figure 1: Représentation graphique de la matrice W pour $K = 10$ décrivant les proportions de chaque groupe d'aliment dans chaque système de consommation.