# Extraction of food consumption systems by non-negative matrix factorization (NMF) for the assessment of food choices.

Mélanie Zetlaoui, Max Feinberg, Philippe Verger, Stéphan Clémençon

# Extraction of food consumption systems by non-negative matrix factorization (NMF) for the assessment of food choices.

Mélanie Zetlaoui[1], Max Feinberg[1], Philippe Verger[1]
Stephan Clémençon[2]

**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 paper to examine how recent *nonnegative matrix factorization* (NMF) techniques can be applied to food consumption data in order to understand these combinations. Such data are nonnegative by nature and of high dimension. The NMF model provides a representation of consumption data through latent vectors with nonnegative coefficients, we call consumption systems, in a small number. As the NMF approach may encourage sparsity of the data representation produced, the resulting consumption systems are easily interpretable. Beyond the illustration of its properties we provide through a simple simulation result, the NMF method is applied to data issued from a french consumption survey. The numerical results thus obtained are displayed and thoroughly discussed. A clustering based on the $k$-means method is also achieved in the resulting latent consumption space, in order to recover food consumption patterns easily usable for nutritionists.

**Keywords**. Dimensionality reduction, food consumption patterns, NMF contribution clustering, Non-negative Matrix Factorization (NMF), sparse data.

## 1 Introduction

Food risk assessment has become an important issue for many national and international bodies in charge of public health. It requires combining several disciplines, such as epidemiology, nutrition, toxicology, and of course applied mathematics in order to develop rigorous methods for quantitative risk assessment. Hence, a growing scientific literature devoted to probabilistic and

[1]INRA-Métarisk
[2]Télécom-ParisTech

statistical methods applied to food risk assessment is now available in applied mathematics journals, see Bertail and Tressou (2006); Tressou (2006); Bertail et al. (2008) and the references therein for instance. Among topics that may be tackled through mathematical modeling, an important issue lies in understanding food choices and consumption behaviors. Some aspects of this question can be assessed by the means of food consumption surveys that are collected among a given population according to different data recording methods. For instance, a large food consumption survey, called INCA (Volatier, 2000), was conducted in 1999 by the Agence Française de Sécurité Sanitaire des Aliments (AFSSA) on a French population sample. It is an individual survey based on classical 7-day dietary records where respondents are required to report all their individual consumptions during one single week.

A classical way to understand consumption behavior by the means of individual food surveys consists in estimating the individual nutrient intakes, estimates being computed by combining consumption data with a food nutrient composition database. Hence intakes are calculated for several nutrients, this is a multivariate approach. Thereafter, homogeneous subgroups of consumers having comparable nutrient intakes are identified by using classical clustering statistical techniques. This operation is usually known as *dietary pattern clustering*. Applications of exploratory multidimensional techniques, in the purpose of deriving dietary pattern clusters, have recently been the subject of a good deal of attention. Among these, *cluster analysis* tries to divide the subjects into homogeneous non-overlapping subgroups with a similar pattern of mean food intake. Until now, iterative partitioning methods such as *k-means clustering* and hierarchical classification techniques are among the most widely used approaches in this context (James, 2009; Samieri et al., 2008). Though it is of practical simplicity, this approach presents several severe drawbacks:

- Whereas clustering is based on nutrient intakes, it is very difficult to a posteriori identify foods that contribute by a majority to a given pattern. Because consumers do not buy nutrients but foods, it is uneasy for the organizations in charge of nutrition policies to clearly establish recommendations that can be easily understood by the consumers.

- Classical clustering methods implicitly assume that average dietary patterns do exist and can be considered as quite representative for a given subgroup. Whether common dietary patterns may exist, individual behavior may also represent an important part in food choice.

Therefore, this paper proposes a different modeling of population consumption that can be directly applied to consumed food quantities without converting into nutrient intake estimates. Even though a very large number of different foods are involved in individual consumption patterns,

all possible food combinations are not observed in practice. Certain foods are preferentially combined or substituted as a function of hedonic choices and/or socio-cultural habits. One may then realistically expect that the vast majority of consumption data can be described by a few patterns, that are linear combinations of consumption vectors of specific foods. These underlying factors can be interpreted as latent variables that we call consumption systems (CS) in the specific context of this study. Therefore, according to this modeling, an individual diet must be seen as a linear superposition of several consumption systems.

Principal Component Analysis and Factor Analysis form part of a collection of statistical methods that aim at recovering such latent variables. However, they have been designed for dealing with Gaussian data. Latent variables and noise are indeed modeled as Gaussian random vectors, which considerably restricts the range of applications for these methods. Considering that observed episodically consumed foods data can be defined as *nonnegative data that have excess zeros and measurement error* (Kipnis et al., 2009), there is considerable empirical evidence for assuming non Normality of the data. Recently, a new latent variable based method has been proposed, called *Non-Negative Matrix Factorization* (NMF in abbreviated form) (Lee and Seung, 1999, 2001) in situations where measurements are nonnegative by nature, such as ours. Originally, this method was developed for image analysis by assuming that brain perception uses parts-based representations. This algorithm for non-negative matrix factorization is able to learn parts of data structures. This is in contrast to other methods, such as principal components analysis and vector quantization, that learn holistic, not parts-based, representations. NMF is distinguished from the other methods by its use of non-negativity constraints. This novel approach has recently been applied to a variety of applications in different fields, ranging from audio source separation (Ozerov and Févotte, 2010) in signal processing, to portfolio diversification (Drakakis et al., 2007) in mathematical finance, through clustering of scotch whiskies (Young et al., 2006).

Thus, it is expected that, here, NMF will permit to extract parts-based substructure, i.e. consumption systems, that are responsible for the whole structure observed in food survey datasets. In complement, each individual consumption pattern is a linear combination of several non-negative consumption systems, and these combinations may be sparse, which would facilitate their interpretation. In order to identify target groups or consumers having similar behavior, NMF analysis must be completed by a clustering of individuals in the consumption system space.

This paper is organized as follows. Section 2 describes consumption survey data. Section 3 presents the proposed NMF modeling of consumption data as combinations of latent consumption systems. A description of the constitutive components of the model is given, as well as its characteristics. The implementation method is developed and the theoretical properties are

also discussed in this section, through a toy simulation example in particular. These theoretical concepts are applied in section 4 to the INCA database. A statistical analysis is carried out in order to interpret the numerical results produced by the NMF procedure. A $k$-means clustering method is next applied in the NMF-derived latent space. Technical details are put together in the Appendix.

## 2 The INCA database

The "Individuelle et Nationale sur les Consommations Alimentaires" (INCA) survey was conducted in 1999 by AFSSA (Volatier, 2000). It was initially designed to assess a global description of French consumer behaviors and estimate nutrient intakes at national level. It consists of individual 7-day food records collected from 3003 French consumers over 11 months in order to take possible seasonality into account. It is a transversal survey performed on two independent samples: one sample of 1,985 French adults over 15 years, and the other of 1,018 children ranging 3-14 years. National representativeness was achieved by combining proportional stratified sampling (geographical zone, town size) and quota sampling (age, gender, social status, and household size). A group of 511 underreporting adults was excluded from the initial sample in order to avoid an under-estimation of food intakes. They were detected by using the method proposed by Schofield based on the ratio between the energy intake and the basal metabolism. They represent about 25 % of the adult sample; this is a classical proportion.

Each respondent was requested to self-report on a formatted record booklet, the quantities and quality of all foods and beverages that have been consumed during one week, including meals taken outside home. Consumed quantities or portion sizes were estimated by the means of a calibrated picture book. For children, record books were filled by the parents. Each type of meal (breakfast, lunch, diner, etc.) was recorded separately. When the booklet was completed, dietary records were validated and corrected if necessary by specialized surveyors who visited all participants.

A closed-ended list of 880 food names was used to coding dietary records. These foods were organized in 44 food groups displayed in Table 1, under the names that are used in the subsequent figures.

There are 37 groups for solid foods and 5 beverages groups (with or without alcohol). Consumed quantity for all foods were converted into grams and expressed as grams/kg of body weight/week, for they were divided by the body weight of the respondent. Some extra information was also reported, such as several socio-demographic variables, including gender and age, in order to make a possible classification of respondents.

For this study it was decided, for each respondent, to sum up all consumed quantities of each 44 groups over the 7 days of the survey. The

Table 1: Naming of the 44 food groups.

| Naming | Naming | Naming |
|---|---|---|
| 1 Breads | 16 Others fats | 31 Water |
| 2 Breakfast cereals | 17 Meats | 32 Non alcoholic drinks |
| 3 Pasta | 18 Fowls | 33 Alcoholic drinks |
| 4 Rice, semolina | 19 Offals | 34 Coffee |
| 5 Others cereals | 20 Cooked pork meats | 35 Hot drinks |
| 6 Vienna pastry | 21 Fishes | 36 Pizzas, quiches, pastries |
| 7 Biscuits | 22 Shellfishes | 37 Sandwiches |
| 8 Cakes | 23 Vegetables | 38 Soups |
| 9 Milk | 24 Potatoes | 39 Cooked dishes |
| 10 Ultra fresh dairy products | 25 Pulses | 40 Starters |
| 11 Cheeses | 26 Fruits | 41 Dessert |
| 12 Eggs | 27 Dried fruits, oilseeds | 42 Compots |
| 13 Butter | 28 Ice creams | 43 Condiments and sauces |
| 14 Oils | 29 Chocolate | 44 Meal substitutes |
| 15 Margarines | 30 Sugars | |

final working data is a 3003×44 data matrix with about 39% of zero values corresponding to a given food group that was never consumed during the 7-day survey. A very important variation range can be observed as well between food groups depending on usual portion sizes as among respondents. In order to minimize the influence of scaling effect due to portion size, data have been standardized. Each summed quantity was divided by its marginal standard deviation allowing then to reduce the influence of portion sizes.

# 3 NMF statistical model

It is the purpose of this section to set out the notations and list the model assumptions that shall be needed in the subsequent statistical analysis. A detailed description of the numerical procedure is also provided, together with a simulation experiment illustrating the methodology under study.

## 3.1 Assumptions and notations

Here and throughout, we denote by $\mathcal{M}_{m,q}(\mathbb{R})$ (by $\mathcal{M}_{m,q}(\mathbb{R}_+)$, respectively) the space of $m \times q$- dimensional matrices with real entries (with nonnegative entries, respectively), and by $||.||$ the Hilbert-Schmidt norm on this space (*i.e.* $||M||^2 = \sum_{i=1}^{m} \sum_{j=1}^{q} m_{ij}^2$ for $M = (m_{ij}) \in \mathcal{M}_{m,q}(\mathbb{R})$).

Classical latent variable analysis methods, such as principal components analysis and vector quantization, learn holistic, not parts-based, representations. NMF is distinguished from these methods by its use of non-negativity constraints that lead to a parts-based representation because they allow only additive, not subtractive, combinations.

Here, we assume that the food choices of an individual are described by a collection of $F$ foods indexed by $f = 1, \ldots, F$. More precisely, the whole diet of an individual is modeled by a vector of length $F$, $Q = (Q^{(1)}, \ldots, Q^{(F)})$ which takes its values in $\mathbb{R}_+^F$, the $f^{th}$ element $Q^{(f)}$ of $Q$ indicating the quantity of food $f$ consumed. Because food consumption widely depends on the nature of the food and its moisture contents, scaling effect can be observed between these $F$ variables. To avoid such effects, each $Q^{(f)}$ is normalized by its marginal standard deviation. We set $v^{(f)} = Q^{(f)}/\sigma^{(f)}$ where $(\sigma^{(f)})^2 = \mathbb{E}[(Q^{(f)} - \mathbb{E}[Q^{(f)}])^2]$, for $f = 1, \ldots, F$.

We assume that the vector $v = (v^{(1)}, \ldots v^{(F)})$ of renormalized food consumptions is drawn as

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

where $W \in \mathcal{M}_{F,K}(\mathbb{R}_+)$, $h = (h^1, \ldots, h^K)$ is a continuous random vector of length $K$ lying in $\mathbb{R}_+^K$ and $\epsilon = (\epsilon^{(1)}, \ldots, \epsilon^{(F)})$ a Gaussian random vector with mean 0 and covariance $\Gamma$, independent from $h$. The number $K$ of latent components is usually chosen much smaller than $F$, hence reducing the data dimension.

In addition, we assume that the matrix of consumption systems $W$ is full of rank $K$ and satisfies the normalization condition

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

This constraint allows for interpreting the columns of $W$. Consumption systems will be thought as *deterministic* combinations of foods in the social-cultural context of the consumption survey and a basis used by the consumer to organize his own diet. $W_{f\,k}$ being then referred to as the loading/weight of food $f$ within consumption system $k$. In order to guarantee a minimum amount of sparsity of the NMF representation and identifiability of the model as well (see subsection 3.2 below), we also assume that, for all pair $(k, l)$ such that $k \neq l$ in $\{1, \ldots, K\}^2$, there exists $f \in \{1, \ldots, F\}$ so that $W_{f,k} = 0$ whereas $W_{f,l} > 0$.

Hence, the model above stipulates that the variability of the (renormalized) consumption vector $v$ arises from two distinct sources, the one of the (non Gaussian) random weight vector $h$, reflecting the part of each consumption system in the diet, and the one of the (Gaussian) noise term $\epsilon$.

In the following, assume that we observe $N$ independent quantity vectors

$$Q_n = (\sigma^{(f)} v_n^{(f)})_{1 \leq f \leq F},$$

with $v_n = Wh_n + \epsilon_n$, for $n = 1, \ldots, N$, where $\epsilon_n$'s are i.i.d random vectors with Gaussian distribution $\mathcal{N}(0, \Gamma)$. The entire set of normalized food consumptions may be then represented by using a matrix notation:

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

where $V$ and $E$ are $F \times N$ matrices and $H$ a $K \times N$ matrix, the columns of which are $(v_1, \ldots, v_N)$, $(\epsilon_1, \ldots, \epsilon_N)$ and $(h_1, \ldots, h_N)$ respectively. Notice finally that the data $V$ are not observed, insofar as the normalization factors $\sigma^{(f)}$ are unknown. However, they can be replaced by their (consistent) empirical counterparts

$$\widehat{\sigma}_N^{(f)} = \left( \frac{2}{N(N-1)} \sum_{1 \leq m < n \leq N}^{N} \{Q_n^{(f)} - Q_m^{(f)}\}^2 \right)^{1/2},$$

as for PCA on pre-normalized data. Hence, the numerical procedure we next describe will be actually applied to the matrix $\widehat{V} = \left(\widehat{v}_n^{(f)}\right)$, where $\widehat{v}_n^{(f)} = Q_n^{(f)} / \widehat{\sigma}_N^{(f)}$ for $f = 1, \ldots, F$.

**Remark** (A MULTIPLICATIVE NMF MODEL) The original NMF method has been introduced by (Lee and Seung, 1999) in a deterministic setting. In this respect, we point out that other ways of "randomizing" the factorization $V = WH$ than the additive fashion could be considered. The noise term could be multiplicative and supposedly positive (drawn from a lognormal distribution for instance), leading to a statistical model of the form $v = \epsilon \times Wh$. In this setup, a statistical fit could be obtained by using the Kullback-Leibler divergence as criterion. Investigation of such a statistical model is beyond the scope of the present paper but will be tackled elsewhere.

### 3.2 The NMF procedure

Given the additive structure of the model, the principle of the NMF algorithm we consider here consists in minimizing the residual sum of squares

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

over the set of pairs $(W, H)$ in $\mathcal{M}_{F,K}(\mathbb{R}_+) \times \mathcal{M}_{K,N}(\mathbb{R}_+)$ subject to the constraints $\sum_{f=1}^{F} W_{fk} = 1$ for all $k = 1, \ldots, K$.

Based on this cost function, Lee and Seung (1999) proposed the following multiplicative algorithm:

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

In order to respect the constraint given by equation (2), matrices $W$ and $H$

are normalized at each step

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

The multiplicative algorithm is based on the gradient descent approach and guarantees the positivity of each component of $W$ and $H$ at each step. From a practical perspective, one starts with an initial choice for $(W, H)$ and the algorithm is iterated until the change in value of $D_K(V, (W, H))$ is negligible. Technical details of the achievement rules are developed in the Web Appendix A.

The convergence properties of the algorithm have been studied in Lee and Seung (2001). It has been proved that the criterion monitoring the convergence of the optimization algorithm decreases towards one solution. Nevertheless there is no evidence in absence of additional constraints on the underlying statistical model that the recovered solution is unique and asymptotically correct, a factorization in a product of nonnegative matrices being not unique in general. The topic of uniqueness/identifiability was tackled by Donoho and Stodden (2004) and Laurberg et al. (2008) by imposing conditions on the column vectors of $W$ or on the distribution of the random vector $h$; they concluded that uniqueness is achievable under specific conditions, see also (Cichocki et al., 2006). In particular, under special configuration when any two distinct column vectors $W_{.k}$ and $W_{.l}$ belong to different facets of the positive orthant $\mathbb{R}_+^F$, the uniqueness of the solution is proved (*cf* Theorem 8 in Laurberg et al. (2008) for instance): from a geometric angle, this clearly guarantees that the simplicial cone generated by the $W_{.k}$'s is the largest one which contains the support of the r.v. $Wh$ and is included in the positive orthant both at the same time.

### 3.3 NMF properties and simulations

The NMF model provides a new representation of consumption data as a basis of food consumption systems (CS) with non-negative entries. According to this model, up to an additive noise term, each individual consumption is represented by a nonnegative linear combination of the CS's:

$$\Pi_{v_n} = \sum_{k=1}^{K} h_{kn} W_{.k}.$$

Vector $\Pi_{v_n}$ belongs to the latent consumption subspace $\Gamma_K$ of $\mathbb{R}^F$, defined by the set of the non-negative linear combinations of CS's:

$$\Gamma_K = \left\{ \sum_{k=1}^{K} \lambda_k W_{.k} : \quad \lambda_k \geq 0 \text{ for } k = 1, \ \ldots, \ K \right\}.$$

It can also be viewed as the projection of the individual consumption $v_n$ into the latent space.

To illustrate the application of NMF statistical model for consumption data and its properties we used simulations. Here we considered the case $F = 3$ and $K = 2$, since data and results can easily be graphically represented in this situation. Given the $3\times2$ $W$ matrix, simulated data were located within the subspace $\Gamma_2$ defined by $W_{.1}$ and $W_{.2}$, which were assumed as belonging to two different facets of the positive orthant $\mathbb{R}^3_+$, that is $W_{.1} = (2/3, 1/3, 0)$ and $W_{.2} = (0, 2/3, 1/3)$. Concretely, 1000 points were generated, $\Pi_{v_n}$ for $n \in \{1, \ldots, 1000\}$, in the set $\Gamma_2$, given by:

$$\Pi_{v_n} = h_{1n}W_{.1} + h_{2n}W_{.2}.$$

In this toy simulation example, it was assumed that $h_{1n}$ and $h_{2n}$ are drawn independently from a lognormal distribution. The related parameters were chosen in such a way that the simulated data distribution be comparable to that of the observations in the INCA database, up to a point. The values for the mean and variance have been taken equal to 2 and 1 respectively, for both $h_{1n}$ and $h_{2n}$ and a white Gaussian noise $\epsilon$ with the identity as covariance matrix has been added.

Results of the simulation are presented in Figure 1.

Figure 1 depicts the simulated data in the original space of dimension $F = 3$. The estimates of the consumption systems produced by the algorithm described in subsection 3.2 are very close to the theoretical values $W_{.1}$ and $W_{.2}$. Indeed, after 1000 iterations, we obtained: $\widehat{W_{.1}} = (0.67, 0.33, 0.00)$ and $\widehat{W_{.2}} = (0.01, 0.66, 0.33)$. The normalized vectors $\widehat{W_{.1}}$ and $\widehat{W_{.2}}$ are graphically identified as the straight lines they generate. Guidelines for the interpretation of the graphic displayed in Fig. 1 can be the following:

1. a point close to the origin correspond to a weak or quasi-null consumption;

2. a point close to one of the CS straight line, $W_{.1}$ say, represents individual consumptions with a weak contribution on $W_{.2}$, and the further the point from the origin, the highest the contribution of the related CS to the global diet.

The parts of the data that are described by the CS's, $\Pi_v$, are represented in the plane containing $W_{.1}$ and $W_{.2}$ (see Web Figure 1) and lie in the interior of the cone generated by theses vectors.

The relevance of the model (1) in regards to prediction/interpretation can be measured by the fraction of the whole variability arising from the component $Wh$ with a parts-based representation, comparing the quantities

$$R^2_f = W_{f.}\Gamma_h\ {}^tW_{f.},$$

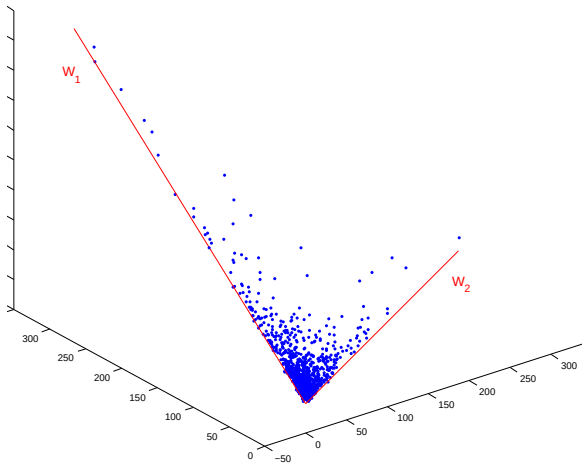


Figure 1: Representation of simulated data. Points are the data and the straight line are the estimated consumption systems.

$f \in \{1, \ldots, F\}$, to $var(v^{(f)})$, where ${}^tM$ denotes the transpose of any matrix $M$ and $\Gamma_h$ the covariance matrix of the r.v. $h$. From a statistical perspective, it can be naturally estimated by its empirical counterpart:

$$\widehat{R}^2_f = \widehat{W}_{f.} \widehat{\Gamma}_h \, {}^t\widehat{W}_{f.},$$

where $\widehat{\Gamma}_h = N^{-1} \sum_{n \leq N} (H_{.n} - \bar{H}_N) \, {}^t(H_{.n} - \bar{H}_N)$ and $\bar{H}_N = N^{-1} \sum_{n \leq N} H_{.n}$. For the simulation example above, $R^2 = (0.999, 0.999, 0.997)$.

# 4 Statistical results and discussion

In this section, the NMF procedure is implemented on the INCA database. Results are presented and discussed from the perspective of nutrition policy and consumer behavior assessment.

### 4.1 Model complexity

In the first place, it should be noticed that the NMF procedure requires to pick the number $K$ of underlying consumption systems, which can be viewed as a complexity parameter and is typically chosen in ranges such that $K \ll N$. As illustrated in Web Figure 2, the determination of the optimal number of consumption systems can be achieved by plotting the residual sum of squares as a function of $K$, typically ranging from 1 to $F$, and looking for a kink in this curve. Unfortunately, in spite of the slight diminution of the decreasing rate one observes between $K = 5$ and $K = 10$ with this data set, there is no clear indication for how to determine the optimal value of $K$. This regular decrease of the sum of squares when the number of CS's increases is confirmed by the regular growth of the percentage of prediction. Table 3 gives the percentage of prediction for $K$ equal to 5, 10 and 20 and for each food group. In average, the percentage of prediction increases by 54% for $K = 10$, compared to $K = 5$. The increasing of prediction percentage from $K = 10$ to $K = 20$ becomes equal to 90%, almost the double of the previous increasing. This fact perfectly illustrates the need for a trade-off between statistical fit and interpretability.

In absence of any theoretical results for grounding automatic selection procedures based on complexity penalization in the NMF setup (see the discussion in section 5), a strategy may consist in selecting $K$ in order to achieve a satisfactory sparsity rate for $W$ and $H$ in such a way that the resulting model can easily be interpretable for an end-user. Heuristically, if $K$ is too small, the matrix $W$ tends to be less sparse (while $H$ tends to be sparser in contrast) and vice versa when $K$ is too large. For the study, we are more interesting on achieving a high degree of sparsity for $W$ that is the most expected output of NMF modeling i.e. understanding the structure of consumption behavior by the means of the CS's. Thus, it was decided to compare the results derived from NMF procedure for different empirical values of $K$, namely 5, 10 and 20. Web Figures 3 and 4 and Figure 2 depict the order of magnitude of the $W$'s entries for the different values of $K$. A grayscale mapping table was used for this illustration. Consumption systems appear in columns and food groups in rows. The larger the loadings of the food group in a given CS, the darker is the cell of the table; for null loading, white is used.

As expected, the larger the number of CS's, the sparser the vectors representing the CS's. For $K$ sufficiently large, the CS's can be characterized by a small number of food, consumed in a same meal or corresponding to a same consumption behavior. For $K$ equal to 5, the CS's are not sufficiently sparse, while, in contrast, for $K$ equal to 20, the interpretation of the CS's obtained is poor in terms of food association. A number $K$ of CS's equal to 10 thus permits to achieve a good trade-off between sparsity and interpretability of the CS's.

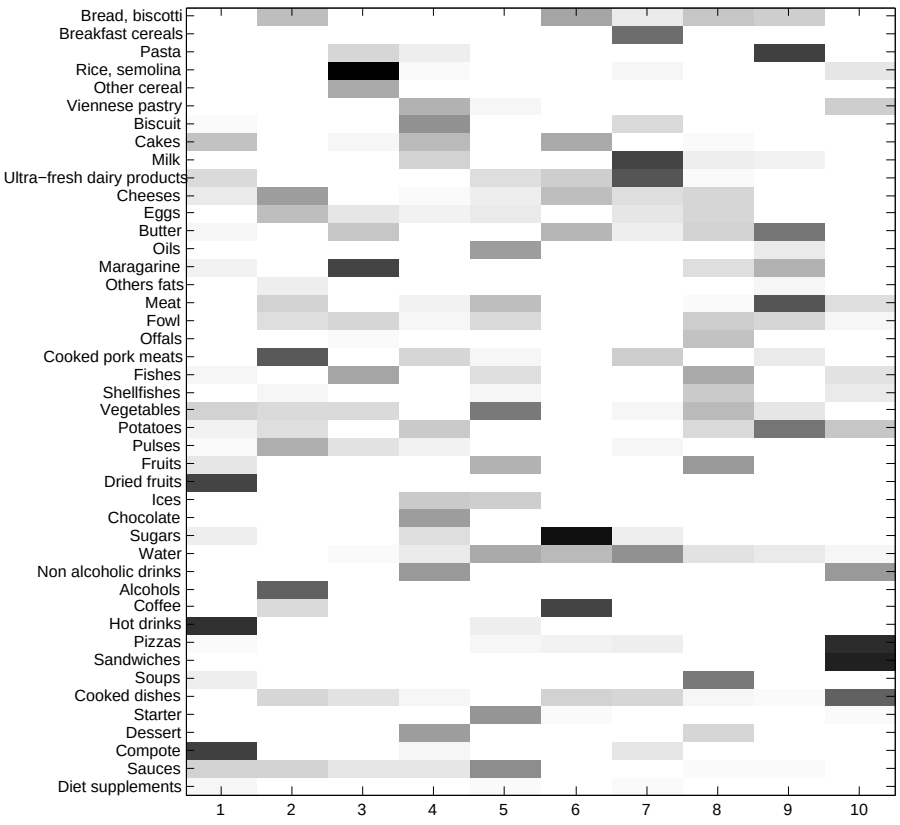


Figure 2: Graphical representation of the $W$ matrix which describes the loadings of each of the 44 food groups to each of the K consumption systems for $K = 10$.

### 4.2 Implementation of the NMF procedure on the INCA database

In this section, the NMF procedure is implemented on the INCA database with a number $K$ of CS's equal to 10. Interpretation of the matrix $W$ can be based upon Figure 2. For each consumption system, the larger the proportion of a food group, the more it contributes to the CS. For instance, CS $W_{\cdot 10}$ is mainly represented by pizzas, sandwiches and cooked dishes and non alcoholic beverages: this structure can be regarded as the part of the diet based on "fast food-like consumption". It can also be observed that certain food groups may appear in several consumption systems (e.g. meat) while some other groups may be involved in a few CS's and in a weak proportion (e.g. offals and diet supplements). This gives indication on food groups which are traditionally consumed and how they are associated according to the French consumer habits. Most consumption systems are characterized by a few strongly contributing products. This sparsity structure of $W$ was an objective when choosing the value of $K$, as it gives a synthetic view of the different consumption behaviors. We shall say that food groups are "associated", when they are consumed during a same meal and/or they describe similar food behaviors. For example, meat is often consumed with potatoes in $W_{\cdot 9}$, or eating pizzas also implies eating sandwiches in $W_{\cdot 10}$.

We also underline that the weights $h_k$ of the CS's in the global diet are not independent, as we can see from their correlation matrix in Table 2.

Table 2: Correlation matrix of the CS weights $h_1, \ldots, h_K$

| | $h_1$ | $h_2$ | $h_3$ | $h_4$ | $h_5$ | $h_6$ | $h_7$ | $h_8$ | $h_9$ | $h_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $h_1$ | 1.000 | 0.003 | 0.018 | -0.044 | 0.040 | 0.022 | -0.055 | 0.102 | -0.077 | -0.060 |
| $h_2$ | | 1.000 | -0.036 | -0.237 | 0.035 | 0.170 | -0.280 | 0.124 | 0.051 | -0.008 |
| $h_3$ | | | 1.000 | 0.014 | 0.040 | 0.048 | -0.016 | -0.081 | -0.016 | -0.069 |
| $h_4$ | | | | 1.000 | -0.139 | -0.177 | 0.090 | -0.209 | -0.036 | 0.018 |
| $h_5$ | | | | | 1.000 | 0.159 | -0.137 | -0.110 | -0.071 | -0.081 |
| $h_6$ | | | | | | 1.000 | -0.245 | 0.094 | 0.037 | -0.042 |
| $h_7$ | | | | | | | 1.000 | -0.124 | -0.050 | -0.118 |
| $h_8$ | | | | | | | | 1.000 | -0.070 | -0.185 |
| $h_9$ | | | | | | | | | 1.000 | 0.022 |
| $h_{10}$ | | | | | | | | | | 1.000 |

The strongest correlation (between $h_6$ and $h_7$) is equal to $-0.245$ and half of the absolute values are below 0.1. One may interpret the sign of the correlation between two $h_k$'s in terms of "opposition" or "complementarity" of the consumption behaviors related to the corresponding CS's. For in-

stance, $h_6$, corresponding to CS $W_{.6}$, where coffee, sugar, bread biscotti are predominant food groups, is negatively correlated to $h_7$ which corresponds to CS $W_{.7}$, mainly represented by breakfast cereals, milk, ultra-fresh dairy products: this can be interpreted as opposed consumption behaviors.

In geometrical terms, consumption systems form a new basis for consumption data. Data can be projected on any map generated by two consumption systems; whatever the selected couple, the graphical appearance of data is very similar to Web Figure 1.

Beyond global interpretability aspects, another attractive property of NMF modelling lies in the fact that it can be used, as a byproduct, for "prediction": the CS-based quantity $\hat{v}_n = \hat{W}\hat{h}_n$ can be viewed as a proxy for the individual consumption $v_n$. In order to assess the quality of the approximation, Table 3 provides elementary statistics for data and predictions. More precisely, it gives for each food group the mean, the standard deviation, the minimum and the maximum of the observed and the predicted quantities.

First of all, the average consumed quantity for each food group is recovered by the model. The mean of prediction error is approximately equal to zero as stipulated by the assumption of the model. Given a food group, the standard deviation of the predictions is, in average, about 50% of the standard deviation of data. The maximum value of the standard deviation, reached for "rice and semolina", is equal to 0.83. For about 70% food groups, the standard deviation is larger than 0.5. The minimum is reached for 'diet supplements' and is equal to 0.05, which food group is actually involved in the consumption of very few individuals. In contrast to the largest values (underestimated by the model) the minimum values (close to zero for each food group) are well captured by the CS-based component.

In order to explain how to interpret more precisely these results, we now focus on meat consumption. The fraction of variability explained by the CS-based component is equal to 70%. For this food group, we simultaneously plotted the distribution of the related consumption data and that of the predictions in Fig. 3. The range of the 3003 consumption values is split into 40 bins of length equal to 0.3. Notice that the overall shape of the distribution is well captured by the model. The maximum value for both distributions is equal to about 18% reached in the third class $(0.6 - 0.9)$ for the data and in the forth class $(0.9 - 1.2)$ for predictions. The greatest values are not well estimated by the CS-based part of the model but concern very few individuals. To some extent, this is also the case for the lowest values: the frequency of $v_n$'s in $(0 - 0.3)$ is equal to 2%, although the (normalized) meat consumption lies in this interval for 12% of the observed population, actually, 80% of them have a null meat consumption. This cannot be well captured by the CS-based component insofar as, for this part of the model, null meat consumption implies also zero consumption of products to which it is associated (for instance potatoes when considering CS No. 9, vegetables for CS No. 5 ...). Null or quasi-null consumptions thus produce large

Table 3: Elementary Statistics on the 44 food groups.

| Group | Observed | | | | Predicted for $K = 10$ | | | | Percentage | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Sdt Dev | Min | Max | Mean | Sdt Dev | Min | Max | $K = 5$ | $K = 10$ | $K = 20$ |
| 1 | 1.04 | 1.00 | 0.00 | 10.90 | 1.039 | 0.58 | 0.05 | 4.36 | 0.372 | 0.335 | 0.783 |
| 2 | 0.35 | 1.00 | 0.00 | 17.80 | 0.35 | 0.51 | 0.00 | 4.78 | 0.155 | 0.263 | 0.467 |
| 3 | 0.86 | 1.00 | 0.00 | 12.83 | 0.86 | 0.68 | 0.00 | 6.07 | 0.324 | 0.458 | 0.460 |
| 4 | 0.74 | 1.00 | 0.00 | 11.39 | 0.74 | 0.83 | 0.00 | 14.56 | 0.356 | 0.683 | 0.502 |
| 5 | 0.15 | 1.00 | 0.00 | 29.28 | 0.15 | 0.28 | 0.00 | 4.92 | 0.025 | 0.079 | 0.445 |
| 6 | 0.51 | 1.00 | 0.00 | 10.35 | 0.51 | 0.42 | 0.00 | 2.82 | 0.172 | 0.177 | 0.530 |
| 7 | 0.56 | 1.00 | 0.00 | 8.98 | 0.56 | 0.57 | 0.00 | 3.31 | 0.218 | 0.327 | 0.423 |
| 8 | 0.68 | 1.00 | 0.00 | 11.79 | 0.68 | 0.46 | 0.01 | 3.39 | 0.062 | 0.210 | 0.342 |
| 9 | 0.72 | 1.00 | 0.00 | 11.35 | 0.72 | 0.70 | 0.00 | 6.11 | 0.529 | 0.486 | 0.553 |
| 10 | 0.79 | 1.00 | 0.00 | 12.39 | 0.79 | 0.60 | 0.00 | 5.60 | 0.246 | 0.357 | 0.641 |
| 11 | 0.98 | 1.00 | 0.00 | 10.85 | 0.98 | 0.56 | 0.01 | 4.17 | 0.322 | 0.315 | 0.291 |
| 12 | 0.78 | 1.00 | 0.00 | 9.55 | 0.78 | 0.35 | 0.08 | 2.83 | 0.099 | 0.126 | 0.936 |
| 13 | 1.08 | 1.00 | 0.00 | 17.43 | 1.08 | 0.61 | 0.00 | 5.42 | 0.344 | 0.374 | 0.577 |
| 14 | 0.59 | 1.00 | 0.00 | 16.31 | 0.59 | 0.46 | 0.00 | 4.83 | 0.107 | 0.214 | 0.723 |
| 15 | 1.05 | 1.00 | 0.00 | 11.91 | 1.048 | 0.66 | 0.00 | 10.52 | 0.465 | 0.430 | 0.673 |
| 16 | 0.09 | 1.00 | 0.00 | 24.80 | 0.093 | 0.09 | 0.00 | 0.65 | 0.004 | 0.009 | 0.987 |
| 17 | 1.21 | 1.00 | 0.00 | 11.79 | 1.21 | 0.70 | 0.01 | 5.67 | 0.259 | 0.484 | 0.506 |
| 18 | 0.86 | 1.00 | 0.00 | 16.16 | 0.86 | 0.36 | 0.07 | 2.70 | 0.123 | 0.127 | 0.244 |
| 19 | 0.33 | 1.00 | 0.00 | 11.15 | 0.33 | 0.31 | 0.00 | 2.53 | 0.028 | 0.094 | 0.942 |
| 20 | 1.02 | 1.00 | 0.00 | 10.92 | 1.02 | 0.69 | 0.03 | 5.67 | 0.246 | 0.471 | 0.390 |
| 21 | 0.85 | 1.00 | 0.00 | 12.04 | 0.85 | 0.50 | 0.01 | 5.19 | 0.190 | 0.245 | 0.290 |
| 22 | 0.38 | 1.00 | 0.00 | 16.61 | 0.38 | 0.27 | 0.00 | 2.19 | 0.046 | 0.072 | 0.953 |
| 23 | 1.36 | 1.00 | 0.00 | 6.78 | 1.36 | 0.74 | 0.10 | 6.96 | 0.580 | 0.553 | 0.743 |
| 24 | 1.08 | 1.00 | 0.00 | 9.07 | 1.08 | 0.58 | 0.03 | 5.03 | 0.265 | 0.338 | 0.492 |
| 25 | 0.44 | 1.00 | 0.00 | 14.16 | 0.44 | 0.35 | 0.00 | 2.73 | 0.040 | 0.120 | 0.883 |
| 26 | 0.84 | 1.00 | 0.00 | 12.66 | 0.84 | 0.60 | 0.00 | 4.20 | 0.312 | 0.365 | 0.300 |
| 27 | 0.20 | 1.00 | 0.00 | 30.40 | 0.20 | 0.58 | 0.00 | 10.44 | 0.024 | 0.342 | 0.987 |
| 28 | 0.39 | 1.00 | 0.00 | 13.20 | 0.39 | 0.33 | 0.00 | 2.86 | 0.075 | 0.111 | 0.548 |
| 29 | 0.42 | 1.00 | 0.00 | 11.01 | 0.42 | 0.50 | 0.00 | 2.98 | 0.092 | 0.246 | 0.485 |
| 30 | 0.93 | 1.00 | 0.00 | 9.96 | 0.93 | 0.82 | 0.00 | 6.26 | 0.266 | 0.676 | 0.357 |
| 31 | 1.23 | 1.00 | 0.00 | 7.71 | 1.23 | 0.56 | 0.18 | 5.19 | 0.234 | 0.314 | 0.585 |
| 32 | 0.63 | 1.00 | 0.00 | 14.51 | 0.63 | 0.60 | 0.00 | 3.94 | 0.315 | 0.365 | 0.469 |
| 33 | 0.50 | 1.00 | 0.00 | 15.02 | 0.50 | 0.69 | 0.00 | 5.06 | 0.273 | 0.471 | 0.450 |
| 34 | 0.66 | 1.00 | 0.00 | 15.11 | 0.66 | 0.71 | 0.00 | 5.16 | 0.402 | 0.507 | 0.716 |
| 35 | 0.45 | 1.00 | 0.00 | 16.12 | 0.45 | 0.65 | 0.00 | 11.57 | 0.123 | 0.423 | 0.916 |
| 36 | 0.72 | 1.00 | 0.00 | 9.58 | 0.72 | 0.61 | 0.03 | 4.75 | 0.203 | 0.373 | 0.442 |
| 37 | 0.50 | 1.00 | 0.00 | 9.71 | 0.50 | 0.65 | 0.00 | 4.87 | 0.190 | 0.419 | 0.592 |
| 38 | 0.65 | 1.00 | 0.00 | 11.01 | 0.65 | 0.66 | 0.00 | 5.31 | 0.124 | 0.432 | 0.811 |
| 39 | 0.96 | 1.00 | 0.00 | 7.93 | 0.96 | 0.50 | 0.10 | 3.80 | 0.141 | 0.249 | 0.438 |
| 40 | 0.52 | 1.00 | 0.00 | 12.75 | 0.52 | 0.50 | 0.00 | 5.24 | 0.159 | 0.255 | 0.281 |
| 41 | 0.54 | 1.00 | 0.00 | 10.59 | 0.54 | 0.50 | 0.00 | 3.49 | 0.121 | 0.249 | 0.447 |
| 42 | 0.42 | 1.00 | 0.00 | 10.25 | 0.42 | 0.60 | 0.00 | 10.65 | 0.066 | 0.359 | 0.957 |
| 43 | 0.92 | 1.00 | 0.00 | 15.03 | 0.92 | 0.61 | 0.06 | 6.33 | 0.305 | 0.373 | 0.698 |
| 44 | 0.05 | 1.00 | 0.00 | 37.41 | 0.05 | 0.05 | 0.00 | 0.57 | 0.001 | 0.002 | 0.991 |

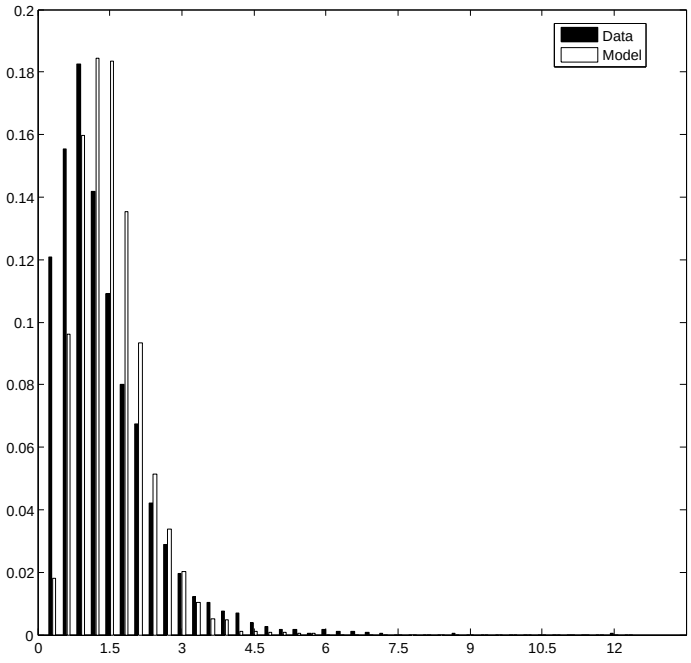


Figure 3: Distribution of data and predictions for the food group "meats".

residuals in general.

Investigating the statistical fit of the model food group by food group is informative, but must be however completed by analyzing the correlations accross the various food groups. Table 4 gives an extract from the correlation matrix of the empirical errors/residuals $\epsilon^{(f)}$ in model (1), displaying the ten highest correlations in absolute value.

It is noteworthy that certain empirical correlation coefficients are significant: the strongest correlation coefficient, between $\epsilon^{(13)}$ and $\epsilon^{(23)}$, is equal to 0.55 and the lowest one, between $\epsilon^{(10)}$ and $\epsilon^{(43)}$, is equal to 0.31. Notice also that all correlations are positive: a large residual for the consumption of one food group leads to a large residual for the consumption of another food

Table 4: An extract from the correlation matrix of the empirical errors.

| | $\epsilon^{(10)}$ | $\epsilon^{(13)}$ | $\epsilon^{(15)}$ | $\epsilon^{(17)}$ | $\epsilon^{(20)}$ | $\epsilon^{(23)}$ | $\epsilon^{(24)}$ | $\epsilon^{(30)}$ | $\epsilon^{(31)}$ | $\epsilon^{(43)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon^{(10)}$ | 1.00 | 0.38 | 0.35 | 0.37 | 0.32 | 0.40 | 0.36 | 0.44 | 0.35 | 0.31 |
| $\epsilon^{(13)}$ | | 1.00 | 0.32 | 0.35 | 0.43 | 0.55 | 0.35 | 0.44 | 0.41 | 0.40 |
| $\epsilon^{(15)}$ | | | 1.00 | 0.43 | 0.41 | 0.53 | 0.35 | 0.39 | 0.39 | 0.23 |
| $\epsilon^{(17)}$ | | | | 1.00 | 0.38 | 0.48 | 0.41 | 0.44 | 0.45 | 0.28 |
| $\epsilon^{(20)}$ | | | | | 1.00 | 0.44 | 0.42 | 0.41 | 0.48 | 0.32 |
| $\epsilon^{(23)}$ | | | | | | 1.00 | 0.44 | 0.49 | 0.43 | 0.53 |
| $\epsilon^{(24)}$ | | | | | | | 1.00 | 0.43 | 0.45 | 0.38 |
| $\epsilon^{(30)}$ | | | | | | | | 1.00 | 0.43 | 0.41 |
| $\epsilon^{(31)}$ | | | | | | | | | 1.00 | 0.37 |
| $\epsilon^{(43)}$ | | | | | | | | | | 1.00 |

group. In addition, observe that the pairs of food groups corresponding to strong correlation coefficients are not necessarily predominant in the same CS. Butter, food group No. 13, is predominant in CS's $W_3$, $W_6$ and $W_9$, while vegetables, food group No. 23, is predominant in CS's $W_5$ and $W_8$, and sauces, food group No. 43, is predominant in CS $W_5$. This can be interpreted as follows: large residuals are simultaneously observed for two food groups, when the additive superposition of CS's do not permit to describe the consumption habits of certain individuals, these individual choices concerning the combination of these food groups overtaking the average trends to some extent.

### 4.3 Consumer clustering based on consumption systems

NMF model actually provides a new representation of consumption data by latent CS's, characterizing most of the population consumption behavior. In the latent consumption space, each individual is represented by an additive combination of these consumption systems. Next, the aim of this section is to identify eventual food consumption patterns among individuals by using latent consumption space derived from NMF. It is expected that clusters based on NMF representation can then be easily interpretable in terms of consumer groups, such as "group at risk".

Recent clustering techniques using NMF (Xu et al., 2003; Ding et al., 2008) consist in considering a partition of the data straightforwardly connected to the CS's . More precisely, individuals are divided into, at most, $K$ clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$, where each individual, $v_n$, is assigned to cluster $\mathcal{C}_k$ with $k \in \{1, \ldots, K\}$, when:

$$\mathcal{C}_k = \text{argmax}\{h_{ln}; l = 1, \ldots, K\}. \tag{4}$$

According to this strategy, the CS's and the clusters are in one-to-one correspondence. An individual is assigned to the cluster for which she/he presents the highest contribution in her/his pattern $h_n$.

This clustering method has been applied to the contribution matrix $H$ derived from the INCA database by the NMF procedure. In order to illustrate graphically the results, we considered two specific CS's, those which explain the major part of the total variability of the data (once the additive noise has been removed) for instance, namely $W_{(1)}$ and $W_{(2)}$, where the indexes (1) and (2) are defined by:

$$(1) \quad = \quad \text{argmax}_{k\in\{1,\,...,\,K\}}\; W_{.k}\widehat{\Gamma}_h\;{}^tW_{.k}, \tag{5}$$

$$(2) \quad = \quad \text{argmax}_{k\in\{1,\,...,\,K\}\setminus\{(1)\}}\; W_{.k}\widehat{\Gamma}_h\;{}^tW_{.k}, \tag{6}$$

where $\widehat{\Gamma}_h$ denotes the empirical variance covariance matrix of the $h_{.n}$'s. In this sense, the two most discriminative CS's actually correspond to the fourth and seventh CS's respectively. Data assigned to cluster $\mathcal{C}_{(1)}$ or cluster $\mathcal{C}_{(2)}$ are projected on the map generated by $W_{(1)}$ and $W_{(2)}$. This is illustrated by Figure 4, where the points of cluster $\mathcal{C}_{(1)}$ are represented by full circles and the points of cluster $\mathcal{C}_{(2)}$ by crosses. In this case only a subsample of the initial population is represented: those having the highest contributions for both selected CS's.

Observe that most of the points lying in cluster $\mathcal{C}_{(1)}$ are closer to $W_{(1)}$ than to $W_{(2)}$, and vice-versa for $\mathcal{C}_{(2)}$. However, this is not always true. Indeed, it should be noticed that the contributions $h_{kn}$'s do not exactly reflect euclidian distances to the axis $W_{.k}$ insofar as the latent vectors extracted by NMF are neither orthogonal nor of unit euclidian norm. In addition, this clustering scheme does not take into account the fact that different CS's may strongly contribute to the diet behavior of many individuals, while the latter are forced to be assigned to only one cluster. Because it is easy to understand that a given consumer may combine several CS's in order to build her/his own diet, the strategy developed by Xu et al. (2003) and Ding et al. (2008), originally tailored for document clustering or audio source separation, seems to be unadapted to consumption data.

Consequently, it was decided to apply classical $k$-means clustering method using the contributions $H$, so as to calculate the distance between individuals. To identify the optimal number of cluster, $k$-means clustering algorithm was applied on the matrix of the individual contributions $H$ for a number of clusters $l$ running from 1 to 30. For each value $l$ the total within-sum of squares was computed. By the mean of his criterion it is possible to identify the approximate optimal number of clusters between 5 and 10: it was decided to select $l = 6$, leading to the six clusters, shown in Figure 5.

The six clusters were drawn in the subspace generated by $W_2$, $W_4$ and $W_5$. More precisely, for each cluster, each individual $n$ is represented by the vector of coordinates $(h_{2n}, h_{4n}, h_{5n})$. The choice of the three CS is guided

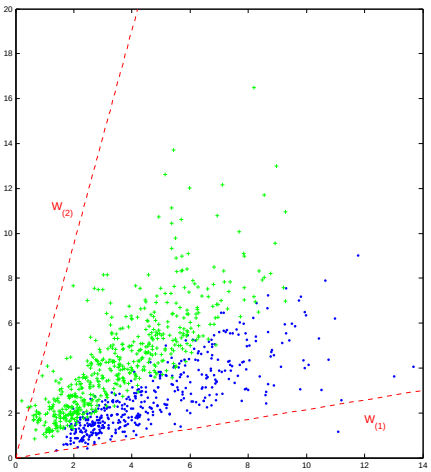


Figure 4: Data clustering representation in the map generated by $W_{(1)}$ and $W_{(2)}$. Full circles are the projected data assigned to cluster $\mathcal{C}_{(1)}$ and crosses to $\mathcal{C}_{(2)}$.

by Web Figure 6 given the coordinates of the centroid of each cluster : the CS $W_2$, $W_4$ and $W_5$ are predominant of cluster 3, 4 and 1 respectively. We remark that cluster 2 and cluster 3 are well separated. They group together individuals having a consumption behavior given predominantly by the CS $W_2$ and $W_5$ respectively. The structure of both clusters shows that both CS represent opposed consumption behavior in the population. Other clusters group together individuals having a lower or quasi-null contribution on both CS's.

To compare with a clustering method directly applied on data, clusters obtained by k-means through NMF are represented on the consumption base. Let us focus now on the three food groups that are predominant for CS $W_2$, $W_4$ and $W_5$, named cooked pork meats, biscuit and vegetables respectively. For each cluster, individuals are represented by their consumption of the three food groups in Web Figure 5. In the consumption space, clusters obtained by the NMF representation are not separated. The representation of the individual consumption in the space of food group does not allow to recover a consumer grouping similar to that provided by the NMF method.

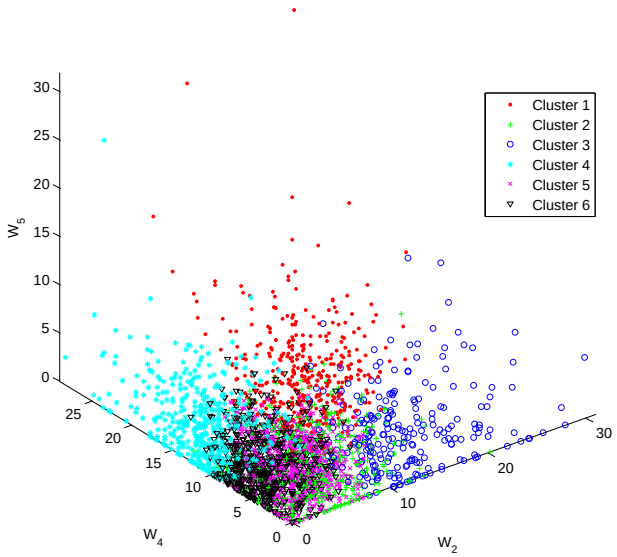


Figure 5: Representation of the six clusters obtained by $k$-means clustering in the base generated by CS $W_2$, $W_4$ and $W_5$.

# 5 Conclusion

The characterization of consumer's eating behaviour is a crucial issue to understand the trend in food consumption and its impact on agricultural production and public health. The clustering techniques are widely applied to reach this goal and to identify the similarities and differences in dietary patterns between countries and regions. As an example the World Health Organization (WHO) developed this approach to describe the various diets around the world and resulted in thirteen so-termed "cluster diets" (Wirfält et al., 2001). In this case the clustering was based on the economical data

collected by the Food and Agricultural Organization and known as the FAO Food Balance Sheets. The food balance sheet shows for each food item i.e. each of about 100 primary commodities available for human consumption which corresponds to the sources of supply and its utilization. The total quantity of foodstuffs produced in a country added to the total quantity imported and adjusted to any change in stocks gives the supply available during that period. On the utilization side a distinction is made between the quantities exported, fed to livestock and used for seed, losses during storage and transportation, and food supplies available for human consumption. The per capita supply of each such food item available for human consumption is then obtained by dividing the respective quantity by the related data on the population actually partaking in it.

In practice the clustering applies on average values for per capita consumption on a period of at least one year. Under these conditions the main problem observed in individual food consumption surveys which is the proportion of null values does not occur. On the other hand this averaged picture does not reflect correctly the complexity of the consumption structure within a considered country or region. The cluster diets are extremely useful to compare the overarching structure of the diets around the world and in particular for developing countries where processed food does not represent the main component of the dietary pattern. In developed countries where several thousands of foods are currently available on the market, estimating the risks and the benefits of a particular dietary pattern are unlikely to be observable from the consumption of agricultural commodities. This paper shows that new statistical techniques can allow extracting homogeneous consumption systems behind the cluster diets. For the future, these techniques should help in quantifying the health impact of food on particular consumer groups.

The NMF methodology appear as a feasible way to tackle this issue. The underlying model can be interpreted at two levels: it permits to extract consumption systems, accounting of the socio-cultural habits present at the population level, while describing the individual preferences of a given consumer through the individual weights assigned to the CS's. The part of the individual diet that cannot be expressed in terms of consumption systems forms the residual noise. When the number of consumption systems introduced in the model is $K = 10$, the percentage of prediction is in average about $R^2 = 32\%$. This value is small when compared to the results of the simulation study but seems rather pertinent when considering the importance of individual choice in the organization of an individual diet. according to this explanation, residuals can be explained as the mixture of individual preferences and measurement uncertainty. Measurement uncertainty also contains all sources of errors, such as aggregation of different foods, portion size estimation and consumer lack of memory.

Compared with other multivariate techniques, such as factorial analy-

sis, NMF presents very specific and interesting properties : factors are not orthogonal and the related weights are not necessarily independent. An individual can be then equally combine independent or not CS. That means that she/he can organize her/his own diet with opposed consumption behaviors or, inversely, not combined closed consumption behaviors. The benefit lies on the fact that individual consumption description is more realistic because it takes into account the diversity of consumption behaviors within a population.

Due to the diversity of food supply, the period of time when the same food is consumed twice can be very long, For consumption surveys which are based on dietary records, even when records are summed over 7 days, it seems obvious there is no frequent replication of the same meal for the same individual. Therefore the observed combination of CS for one individual cannot readily be interpreted as a complete diet. The clusters are more representative of some average diet components for a group of consumers. The results we observed here are also dependent on the INCA database and the application of NMF to other surveys collected in other countries where consumption behavior may differ would exhibit other consumption systems. But the feasibility of this approach is demonstrated by the consistency of the obtained results, in regard of nutritional knowledge.

Though very promising, the application of NMF techniques bring several open problems, listed below. In the future, the latter should be tackled and solved hopefully, for a better understanding of the application of NMF to consumption data in particular.

- Until now, no theoretical grounds have been set for the statistical consistency of NMF procedures. Investigating the asymptotic properties of such $M$-estimation techniques (provided that the underlying NMF representation is unique) is a challenging theoretical problem and will be tackled in a forthcoming article.

- One of the limitations of the additive NMF model considered in this paper lies in the fact that it stipulates that nonnegative data can be observed (the noise being Gaussian) with positive probability, while null values occur with probability zero. Such a modeling is naturally arguable in the dietary context. Building a more relevant NMF model for consumption data, where noise is incorporated in a multiplicative way for instance (see Remark 3.1), defines an ambitious direction for further investigation.

- From a practical perspective, the major question is the determination of the optimal number of latent vectors. The sparsity of factors and the value of errors crucially depend on this number. Nevertheless, the consequences of under- or over-estimating $K$ are unknown. This is

a typical model selection problem, not yet tackled in the literature (Donoho and Stodden, 2004; Laurberg et al., 2008).

- The study was restricted to 44 food groups. An application to a larger number of foods must be conducted in order to ascertain that NMF method is adapted to larger dimension consumption data and if results give a more precise description of the consumption behaviors.

- For dietitians, nutritionists and nutrition policy-makers the added-value of NMF can only be appreciated if it can be combined with a clustering technique that gives clues on the grouping of consumers as potential groups at risk that may similarly combine the same consumptions systems. The proposed solution based on k-means clustering is still insufficient. Better characterization of these groups must be achieved and this will be done in further developments.

## Acknowledgements

This research was supported by the project TAMIS from Agence Nationale de la Recherche (ANR) (`http://www.agence-nationale-recherche.fr`).

## Supplementary Materials

Web Appendix A referenced in Section 3.2 and Web Figures are available under the pdf-file untitled "Supplementary Materials".

## References

Bertail, P., Clémençon, S., Tressou, J., 2008. A storage model for modelling exposure to food contaminants. Mathematical Biosciences and Engineering 5 (1), 35–60.

Bertail, P., Tressou, J., 2006. Incomplete generalized U-Statistics for food risk assessment. Biometrics 62 (1), 66–74.

Cichocki, A., Zdunek, R., Amari, S., Mar. 2006. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In: 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06). Charleston SC, USA, pp. 32–39.

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.

Donoho, D., Stodden, V., 2004. When does non-negative matrix factorization give a correct decomposition into parts? In: Thrun, S., Saul, L., Schölkopf, B. (Eds.), Advances in Neural Information Processing Systems 16. MIT Press. Cambridge, MA.

Drakakis, K., Rickard, S., de Frein, R., Cichocki, A., 2007. Analysis of financial data using non-negative matrix factorization. International Journal of Mathematical Sciences 6 (2).

James, D. C. S., 2009. Cluster analysis defines distinct dietary patterns for african-american men and women. Journal of American Diet Association 109, 255–262.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J., Freedman, L. S., 2009. Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. BiometricsDoi: 10.1111/j.1541-0420.2009.01223.x.

Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K., Jensen, S. H., 2008. Theorems on positive data: On the uniqueness of NMF. Computational Intelligence and Neuroscience 2008, 9 pages.

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

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.

Ozerov, A., Févotte, C., 2010. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Transactions on Audio, Speech and Language Processing 18 (3), 550–563.

Samieri, C., Jutand, M.-A., Féart, C., Capuron, L., Letenneur, L., Barberger-Gateau, P., 2008. Dietary patterns derived by hybrid clustering method in older people: Association with cognition, mood, and self-rated health. Journal of American Diet Association 108, 1461–1471.

Tressou, J., 2006. Non parametric modelling of the left censorship of analytical data in food risk exposure assessment. J. Amer. Stat. Assoc. 101 (476), 1377–1386.

Volatier, J.-L., 2000. Enquête INCA (individuelle et nationale sur les consommations alimentaires), TEC&DOC Edition. Lavoisier, Paris.

Wirfält, E., Hedblad, B., Gullberg, B., Mattisson, I., Andrèn, C., Rosander, U., Janzon, L., Berglund, G., 2001. Food patterns and components of the

metabolic syndrome in men and women: A cross-sectional study within the Malmö diet and cancer cohort. American Journal of Epidemiology 154 (12), 1150–1159.

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.

Young, S. S., Fogel, P., Hawkins, D., 2006. Clustering scotch whiskies using non-negative matrix factorization. Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association 14 (1), 11–13.