The New England Journal of Statistics in Data Science Volume 2, 30–53 (2024)
DOI: https://doi.org/10.51387/23-NEJSDS35

Dietary Patterns and Cancer Risk: An Overview with Focus on Methods

Valeria EDEFONTI[1,2], Roberta DE VITO[3,4,5,*], Maria PARPINEL[6], and Monica FERRARONI[1,2]

to Adriano Decarli, Honorary Professor of Medical Statistics, Università degli Studi di Milano

Abstract

Traditionally, research in nutritional epidemiology has focused on specific foods/food groups or single nutrients in their relation with disease outcomes, including cancer. Dietary pattern analysis have been introduced to examine potential cumulative and interactive effects of individual dietary components of the overall diet, in which foods are consumed in combination. Dietary patterns can be identified by using evidence-based investigator-defined approaches or by using data-driven approaches, which rely on either response independent (also named "a posteriori" dietary patterns) or response dependent (also named "mixed-type" dietary patterns) multivariate statistical methods. Within the open methodological challenges related to study design, dietary assessment, identification of dietary patterns, confounding phenomena, and cancer risk assessment, the current paper provides an updated landscape review of novel methodological developments in the statistical analysis of a posteriori/mixed-type dietary patterns and cancer risk. The review starts from standard a posteriori dietary patterns from principal component, factor, and cluster analyses, including mixture models, and examines mixed-type dietary patterns from reduced rank regression, partial least squares, classification and regression tree analysis, and least absolute shrinkage and selection operator. Novel statistical approaches reviewed include Bayesian factor analysis with modeling of sparsity through shrinkage and sparse priors and frequentist focused principal component analysis. Most novelties relate to the reproducibility of dietary patterns across studies where potentialities of the Bayesian approach to factor and cluster analysis work at best.

keywords and phrases: Dietary patterns, Cluster analysis, Factor analysis, Reduced rank regression, Multi-study factor analysis, Robust profile clustering.

1. INTRODUCTION

There were an estimated 18.1 (9.3 in men and 8.8 in women) million new cancer cases and 9.9 (5.5 in men and 4.4 in women) million cancer deaths worldwide in 2020. Breast and lung cancers were the most common cancers worldwide (12.5% and 12.2% of the total number of new cases diagnosed in 2020, excluding non-melanoma skin cancer), followed by colorectal cancer (10.7% of new cases). With the burden growing in almost every country, preventing cancer is one of the most significant public health challenges of the 21st century. It has been estimated that 30–50% of cancer cases could be prevented by tackling risk factors relating to diet, nutrition, and physical activity [117]. According to the World Cancer Research Fund Prevention Recommendations, changing dietary patterns (i.e., eating whole grains, vegetables, fruit, and limiting consumption of red and processed meat, fast-food products, and sugary drinks), reducing alcohol consumption, increasing physical activity, and achieving and maintaining a healthy body weight can impact people's likelihood of developing cancer and other noncommunicable diseases over their lifetimes [19]. Tailored statistical methods are essential to support the collection of a sound evidence base for cancer prevention.

Traditionally, research in nutritional epidemiology has focused on specific foods/food groups or single nutrients. However, dietary determinants of non-communicable diseases differ from those of undernutrition and nutrient deficiencies, which result from insufficient intake or absorption of a particular nutrient. Multiple dietary determinants act interactively and cumulatively affect disease risk over decades. In addition, when one component of the diet changes, it is typically substituted by another [118]. Consequently, nutritional epidemiologic investigations of non-communicable

[*] Corresponding author.

[1] Branch of Medical Statistics, Biometry, and Epidemiology "G. A. Maccacaro", Department of Clinical Sciences and Community Health, Università degli Studi di Milano.

[2] Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico.

[3] Department of Biostatistics, Brown University.

[4] Data Science Initiative, Brown University.

[5] Center for Computational Molecular Biology, Brown University.

[6] Department of Medicine - DAME, Università degli Studi di Udine.

diseases have integrated the single-nutrient and the single-food approaches (named single-component approach) with the overall diet evaluation through dietary patterns. Dietary patterns can be broadly defined as the foods, food groups, or nutrients included; their combination and variety; and the frequency and quantity with which they are habitually consumed.

Dietary patterns can be identified by using evidence-based investigator-defined approaches (also named "a priori" dietary patterns) or by using data-driven approaches, which rely on either response independent (also named "a posteriori" dietary patterns) or response dependent (also named "mixed-type" dietary patterns) multivariate statistical methods. A priori dietary patterns express adherence to benchmark diets, including those suggested by dietary guidelines, using simple mathematical expressions, like sums or ratios. A posteriori dietary patterns are mostly derived with variants of principal component analysis (PCA), exploratory factor analysis (EFA), or cluster analysis (CA). The mixed-type dietary patterns – which combine elements of both previous approaches – are traditionally identified by using the reduced rank regression, which directly allows them to explain the most variability in (intermediate) response variables [67, 68, 89, 85, 33, 92, 70, 47, 135, 107].

In recent years, innovation has been observed in methods identifying a posteriori dietary patterns across different studies or known subgroups (e.g., by center or ethnicity) within the same study [24, 27]. These dietary patterns have been successfully related to cancer risk [24].

The current paper will provide an updated landscape review of novel methodological developments in the statistical analysis of a posteriori/mixed-type dietary patterns, within the open methodological challenges related to study design, dietary assessment, identification of dietary patterns, confounding phenomena, and cancer risk assessment. The strengths and limitations of standard and novel approaches will be carefully described and evaluated.

2. RELIABLY MEASURE DIETARY INTAKES

Measuring diet in free-living populations is challenging. Diet is a complex exposure, with numerous and (sometimes) poorly characterized components consumed in varying amounts and combinations by individuals. Diet is also a time-varying exposure, with dietary habits and food composition changing over time [104].

Several techniques (i.e., dietary assessment methods) have been developed to ascertain dietary intake in free-living populations. Although each assessment method has its own set of limitations and is prone to some form of (random or systematic) error, its unique strengths make it appropriate for use in specific applications [131].

2.1 Dietary Assessment Methods

Multiple-week diet records require participants to record everything they eat/drink over several weeks. They are considered the gold standard for collecting dietary information because, unlike other methods, they do not rely on memory. The high costs and participant burden have severely limited their use in large-scale epidemiological studies; however, their ability to accurately provide detailed dietary information makes them useful in validation studies of other dietary assessment tools and in monitoring compliance in trials. In addition, recording can worsen as recording days increase [131].

Multiple 24-hour recalls involve reporting all foods and beverages consumed in the previous 24 hours (or in a calendar day) to a trained interviewer in person or over the phone for more than one assessment. Although reliance on the participant's memory leaves room for measurement error, a skilled interviewer can produce highly detailed and useful nutritional data, (almost) comparable to a diet record [22]. This method has been widely used in clinical nutrition and dietary intervention trials. It is also employed in national surveys to monitor trends in nutritional intakes [131].

A food-frequency questionnaire (FFQ) consists of a structured food list and a frequency response section in which the participant indicates his/her usual frequency of intake of each food or beverage over a certain period in the past, usually one year. Portion sizes, with the indication of a standard portion size (in grams or natural units), are also generally queried [131]. Because of their relying on memory, FFQs are more prone to a biased and/or partial recording of dietary information, compared with 24-hour recalls and diet records administered for many days spread out over the entire reference period of an FFQ [104]. The FFQ is, however, the most common option for measuring intake in extensive observational studies. Indeed, it is easy to administer, has a low participant burden, and well captures usual long-term dietary intake. These features also allow for repeated assessments of dietary habits via FFQ over time. This is crucial for capturing longer-term diet variation [131].

Country-specific, complete, and up-to-date food composition databases are essential to convert food consumption (as collected with any of the previous dietary assessment tools) into macro-nutrients, micro-nutrients, bioactive substances (e.g., polyphenols), and non-nutrient (e.g., ethanol or contaminants) intakes. The difference between macro- and micro-nutrients is based on their different function: the former create energy and then promote the organism's growth, and the latter contribute to other (i.e., optimal cell) functions. The former are expressed in grams, the latter in milligrams or micrograms. Estimating nutrient composition from food intake data poses additional challenges. Among others, a food's nutrient content varies with production location, season, growing conditions, storage, processing, and cooking techniques. Some of these factors are unaccounted for in food composition databases, generating estimates of nutritional content that can be affected by errors. However, the degree to which this is problematic differs from nutrient to nutrient. Food composition databases represent the current composition of foods consumed by a given population as far as they are regularly updated. Continuous updates allow targeting and covering newly manufactured prepackaged products, whose formulation and nutrient composition is swiftly changing due to policies, customers' demand, and marketing strategies. The incomplete coverage of foods and the lack of information on some nutrients or bioactive substances of health interest in a food composition database might be potential sources of error. While these sources of error do not substantially compromise the ability to rank individuals based on nutrient intake and, therefore, to evaluate associations with health outcomes [131], estimating nutrient composition from food intake data requires continuous improvements in the accuracy of food composition databases, especially given the changing food landscape [128].

When participants provide biological specimens, researchers can also assay biomarkers to integrate information on foods/beverages from the dietary assessment tool or nutrients from the food composition databases. Examples of biomarkers include doubly labeled water (for total energy intake), urinary nitrogen (for protein intake), 24-hour urinary sodium and potassium, blood lipid profiles, and serum and plasma folate. In principle, biomarkers provide objective intake measurements without bias due to self-reporting. In practice, the use of biomarkers in investigating nutrient-disease relations has been limited to nested case-control studies, small trials, and validation studies of dietary assessment tools. This is due to their inherent limitations, including a lack of sensitive or specific biomarkers for many foods and nutrients, their expensive collection, and assessment errors from multiple sources. In addition, biomarkers may not be indicators of individual long-term intake [104].

2.2 Measurement Error

While the appropriate application of different diet assessment methods, alone or in combination, and of food composition databases allows for a reasonably comprehensive assessment of the diet in free-living populations, awareness of sources/types of measurement error in dietary intake data is crucial in the analysis of the association between diet and non-communicable diseases, including cancer [131, 52].

Usual intake (in the short or the long term) is estimated via the reported intake, as derived from dietary assessment methods. This estimation procedure poses critical challenges. In its simplest form, the following additive model expresses the relation between true, T_i, and reported, R_{it}, intake:

$$R_{it} = \beta_0 + \beta_1 T_i + u_i + \epsilon_{i,t},$$

where the reported intake R_{it} for the i-th individual on the t-th time (e.g., day) linearly depends on one source of random error, $\epsilon_{i,t}$, and three different sources of systematic error, β_0, β_1, and u_i. In detail, the random measurement error $\epsilon_{i,t}$ of the i-th individual on the t-th time models individual consumption variation over time and usually reflects day-to-day changes in consumption; β_0 represents systematic error occurring in the same way to all individuals, and therefore the effect is constant; β_1 depends on the true intake, T_i, and its magnitude is proportional to or multiplied by the true intake; u_i is the subject-specific bias, depending on individual characteristics (i.e., age, sex, or education) that lead to systematic under- or over-reporting consumption.

Validation studies allow for assessing the relationship between reported and true intake. In the absence of a "true" gold standard, an alloyed gold standard (i.e., a reference dietary assessment method with random error only) is used to estimate the true intake. This gold standard is generally a multiple-week diet record or a recovery biomarker, such as a specific biological product directly related to intake and not subject to homestasis or substantial inter-individual differences in metabolism. The gold standard is needed to distinguish the systematic error components and correct the intake estimates.

The 24-hour recalls have generally larger random within-person error than an FFQ, but smaller systematic error, when the two tools are compared with a reference recovery biomarker [105]. The random error in 24-hour recalls is mostly driven by day-to-day variation in intake and other random errors that affect reporting from day to day. It can be mitigated by averaging over many repeats. FFQs usually have lower random within-person variation than the other dietary assessment methods, because they are designed to assess the usual average intake over a longer time period. In FFQs, the error is driven by inaccuracies associated with recalling long-term intakes and features of the instrument, such as the finite food list and the (relative) lack of detail about foods consumed. These biases are systematic and are not mitigated by averaging across repeated measures. On average, at the population or group level, the Observing Protein and Energy Nutrition [105] and other validation studies suggest serious energy underreporting, of approximately 10% when using 24-hour recalls and about 30% when using FFQs. This underreporting arises from sources of group-level bias (u_i), constant additive bias β_0, and intake-related bias β_1.

Although FFQs may suffer from greater measurement errors, they have been shown to have acceptable validity when compared to reference measures [12, 116]: typical correlation coefficients for individual nutrients or foods range from 0.4 to 0.7 [131]. Along with repeated FFQs, adjustment for total energy intake in long-term prospective cohort studies further improves these validity coefficients. When biomarkers are available together with dietary records, triangulation methods can be used to obtain improved estimates of correlations of FFQ intake with true intake [91]. These validity coefficients can be used to correct for measurement error in epidemiologic analyses, and the application of these measurement error correction methods is increasingly being extended to more complex analyses [97, 102, 112]. These techniques have allowed for valid inferences to be drawn from large cohort studies with the use of FFQ data.

In conclusion, the considerable progress made – especially with the use of repeated measures of diet over time – has enabled nutritional epidemiologists to increase reliability in collecting and using dietary information at the individual and population levels. However, continued improvements in dietary assessment methodology and measurement error correction are needed to support the current understanding of the relationship between diet and non-communicable diseases, including cancer [104].

3. STUDY DESIGNS AND RELATED ISSUES IN NUTRITIONAL EPIDEMIOLOGY

The most suitable dietary assessment method for collecting dietary exposure depends heavily on the overall study design that is feasible to adopt for answering the research question. Similarly, additional biases and forms of measurement error not related to dietary exposure are relevant in assessing the relationship between dietary habits and disease, including, but not limited to, cancer.

3.1 Study Designs in Nutritional Epidemiology

One of the major remarks against nutritional epidemiology is that it mostly relies on observational data, which is deemed inferior to experimental data in determining causality. While randomized controlled trials (RCTs) with hard clinical endpoints occupy the highest position in this hierarchy, RCTs are usually not the most suitable/feasible study design to answer nutritional epidemiologic questions regarding long-term effects on health/disease (e.g., cancer) of specific foods or nutrients [104]. Unlike classic drug trials, in RCTs of dietary interventions typically [104, 108]:

- blinding is not feasible, leading to the possibility that the intervention effect is due to knowledge of treatment assignment instead of the dietary component of the intervention;
- higher dropout rates are more likely, especially if the intervention is very demanding, including if it lasts for long periods. When substantial, this high dropout will reduce power in the presence of random losses to follow-up. But, if the dropout is differential to treatment and outcome, it may also introduce systematic bias in the effect estimate, usually in unpredictable directions;
- insufficient adherence of participants to their assigned intervention (i.e., noncompliance) is a major issue, which may become severe in trials of longer duration;
- choosing the control group, when present, is more complicated. Indeed:
    1. when one control group is selected, this group is asked to follow its usual diet; when more than one control group is selected, different dietary regimens are compared, and each is designed to differ from the others in some respect; this makes results interpretation more difficult in dietary interventions trials;
    2. since decreasing the intake of one nutrient/food usually entails increasing the intake of another nutrient/food to compensate for the reduction in calories in isocaloric trials, the choice of the comparison group can influence the observed effect of dietary intervention, further complicating its interpretation.

In the absence of evidence from large RCTs on hard endpoints, nutritional epidemiologists typically rely on prospective cohort studies, the strongest available observational design, to infer causality. Being prospective, cohort studies are less affected by the typical biases (i.e., reverse causation, recall bias, and selection bias) of retrospective or cross-sectional study designs.

Reverse causation describes the situation in which the outcome affects the exposure rather than the other way around. This is a common concern with cross-sectional and retrospective case-control studies as they assess exposure and outcome simultaneously (although in case-control studies, exposure information concerns the past). Prospective cohort studies can minimize the possibility of reverse causation because participants are followed forward in time; these studies can also examine the extent of reverse causation from subclinical disease by lagged analyses [104].

Compared to retrospective case-control studies, prospective cohort studies begin with a disease-free population at baseline that is followed up to ascertain incident cases that develop over time and can minimize both selection bias (controls not being representative of the underlying population that gave rise to cases) and recall bias (knowledge of disease status affecting recall of diet) [104].

3.2 Confounding

A major challenge when working with observational data is confounding. A confounder is a variable associated with both the exposure (i.e., diet) and the outcome (i.e., cancer), and, when unaccounted for, introduces bias into the exposure-outcome relation. The main reason why randomized trials are considered superior in inferring causality is that, as far as the sample size is large enough, the random allocation of participants to treatment groups nullifies measured and unmeasured confounding.

To account for this bias in an observational study design, researchers must identify all relevant confounders based on existing evidence/theory on the association between dietary habits and the disease (e.g., cancer type) under consideration. Once data are collected, the investigator can statistically adjust for confounders in a multiple regression model, including the main exposure (i.e., diet) and/or restrict the analysis to a specific subgroup to minimize residual confounding. Sensitivity analyses further strengthen results by suggesting the magnitude of unmeasured confounding needed to neutralize an effect completely [104]. A prospective design additionally allows for up-to-date tracking of confounders and, in this way, limits the risk of residual confounding. While updated information may reduce measurement errors in the assessment of confounders, additional information can be collected when needed.

Although there are several ways to account for confounding in prospective cohort studies, the critical assumption of no unmeasured or residual confounding needed to infer causality cannot be empirically verified in observational epidemiology [53]. For this reason, prospective cohort studies are often seen as providing statistical associations but not causality. However, when satisfied, the Hill criteria [54] supports the possibility of inferring causality from observational data when randomized trials of hard endpoints are not feasible [104].

In addition, evidence from prospective cohort studies should be integrated with results from randomized trials of intermediate responses, in vitro and in vivo studies, to arrive at a consensus on diet and a health/disease (e.g., cancer). The inference of causality is strengthened when these different types of studies provide consistent evidence in the context of the larger evidence base.

In conclusion, when randomized trials of hard endpoints are unfeasible, well-conducted prospective cohort studies can be used to infer causality with a high degree of certainty. Sources of bias, including confounding, can be minimized by relying on high-quality study design, careful statistical analysis and interpretation, and replications of the findings across different populations. Corroborating data from multiple study types and populations can enhance the weight of evidence [104].

4. DIFFERENT TYPES OF EXPOSURES: FOOD ITEMS, FOOD GROUPS, NUTRIENTS, AND BIOMARKERS

Given the complex nature of the human diet, another way of inferring causality is to consider different types of exposure (i.e., food items, food groups, nutrients, and biomarkers, if any) simultaneously within the same study [62, 58]. This is one of the unique features of nutritional epidemiology, where dietary information is disentangled into different sets of potentially related variables.

While food items are available from the dietary assessment tools, and they are several (i.e., from 50 to 200, depending on the dietary assessment and/or variety of the collected diet), a smaller set of food groups – between 20 and 40 – may be created by the researchers to summarize key components of the overall diet by summing up food items based on similarities in nutrient content, consumption at meal, or culinary use. Examples of food groups include citrus and non-citrus fruit to summarize fruit consumption, raw and cooked vegetables to summarize vegetable consumption, or whole and refined grains to summarize grain consumption.

Measurements for food items and groups depend on how foods are recorded in the dietary assessment tool (directly as raw frequencies per day or week or in pre-specified consumption categories roughly converted into frequencies for the analysis). Generally, food items/groups are expressed as counts, including frequency fractions to account for data collection in consumption categories (e.g., 1–2 times/week converted into 1.5 times/week). Nutrients are continuous variables derived from country-specific food composition databases. As derived by laboratory processing of collected blood or urine samples, biomarkers are continuous variables too. Differently from the previous dietary components, biomarkers are not generally used alone to represent diet but are typically used jointly with other sources of dietary information to validate them.

In conclusion, analyses in nutritional epidemiology can rely on different sets of potentially correlated variables (i.e., food items, food groups, nutrients, biomarkers, if any) expressed in different scales to provide a comprehensive representation of dietary exposure's complexity.

5. SINGLE DIETARY COMPONENTS AND DIETARY PATTERNS

Traditional analyses in nutritional epidemiology examine diseases in relation to a single dietary component (i.e., nutrient, food item, or food group). Typically, a series of regression models are fitted to consider the same confounding factors and one nutrient or food item/group at a time. The single dietary components are categorized according to specific quantiles (e.g., tertiles, quartiles, or quintiles, depending on the sample size). Point estimates and corresponding confidence intervals for measures of disease (e.g., cancer) risk are related to the highest quantile-based categories of consumption (vs. the lowest one); an additional p-value provides an analysis of the trend in risk for each nutrient or food item/group. Quantiles are chosen in studies based on FFQs as FFQs are valuable tools to rank individual consumption in a population but do not necessarily provide accurate estimates of absolute intakes of food items.

Total energy intake represents a major confounding factor in assessing the association between dietary habits and disease risk. Differences in working and leisure time physical activity, body size, and metabolism may be roughly captured by total energy intake [98, 129], which is directly available from food composition databases. In addition, adjusting for energy intake diminishes extraneous sources of variation in dietary intake and, to some extent, also reduces systematic sources of under- and over-reporting [131, 130, 1].

Correction for total energy intake can be accomplished in different ways:

- total energy intake is entered together with other relevant confounding factors in the regression model including quantile-based categories of single nutrients or food items/groups with no previous preprocessing of the dietary component considered as the main exposure;

- single nutrients or food items/groups are preprocessed to account for total energy intake using the residual, the energy partition, or the density methods [69, 129, 120, 88].

In diseases where one nutrient/food is the predominant etiologically relevant dietary component (e.g., folate intake for the prevention of neural tube defects [131] or trans fatty acids from partially hydrogenated oils and heart disease risk [87]), the single-component approach has (likely) the greatest power to identify the effect of a dietary component. In addition, it allows for easier comparisons of results across populations and studies, especially when nutrients or food groups common to most diets are considered [34].

The single-component approach has also conceptual and methodological limitations [56]. First, free-living individuals do not eat isolated nutrients or foods. They eat meals consisting of several foods [60] with complex combinations of nutrients that are likely to be interactive or synergistic; the "single-nutrient" approach may not take into account complicated interactions among nutrients in studies of free-living individuals [61, 59]. Even if one wishes to consider such interactions, enormous sample sizes may be required to assess even a few.

Second, the high correlation among some nutrients (such as potassium and magnesium) makes it difficult to examine their separate effects: the degree of independent variation of each nutrient is markedly reduced when correlated nutrients are entered into a regression model simultaneously.

Third, focusing only on one nutrient often fails to consider substitution effects between nutrients and the potential additional role of the associated food sources. In weight-stable populations - in which changes in macro-nutrient composition occur in isocaloric conditions - when testing the effects of reducing a dietary macro-nutrient, one must consider the alternative macro-nutrient and its food sources. Hence, there is no effect of a macro-nutrient in an absolute sense, because this may change based on the replacement nutrient and the foods that deliver them [118].

Fourth, the effect of a single nutrient may be too small to detect, but the cumulative effects of multiple nutrients may be sufficiently large to be detectable [56]. In addition, by trying to parse the effect of dietary components, one might miss associations between diet and disease because the effects of the individual components are examined against the background of average risk associated with other nutrients or foods. Adjustment for the other nutrients would provide little help in this case.

Fifth, single-component analyses examining a large set of nutrients/food groups may produce statistically significant associations simply by chance, without a proper adjustment for multiple comparisons [56].

Sixth, because nutrient intakes are commonly associated with specific dietary patterns, the "single component" analysis may potentially be confounded by the effect of the overall dietary pattern. For example, diets high in fiber tend to be high in vitamin C, folate, various carotenoids, magnesium, and potassium. When one sees a protective significant effect of fiber on disease risk, can she/he be sure that the relationship is not a consequence of a higher intake of the other nutrients altogether? Even if we adjust for intakes of other nutrients or foods, our ability to accomplish the adjustment can be limited when these intakes are highly correlated. Even when the models fit well, adjustment for the single nutrients/food groups summarized in the dietary pattern may not remove all the confounding effects because these dietary components may interact with each other [56].

Since the mid-Nineties [122, 63, 79, 109], several authors have proposed to integrate the single-component analysis with the analysis of dietary patterns. Within this comprehensive approach to disease prevention or treatment, the collinearity of nutrients and foods is taken into account and exploited [67, 68, 89, 85, 33], in parallel with their separate effects.

6. AN INTEGRATED APPROACH TO THE ANALYSIS OF DIETARY DATA IN RELATION TO DISEASE RISK

From the public health perspective, examining dietary patterns would parallel the real world more closely. As is often stated by our colleagues in nutrition: "We don't eat nutrients, we eat foods." [60]. This statement can be amended to say that we eat foods and eat them in certain combinations or "patterns" [63]. This is why dietary patterns should come first, followed and integrated by evidence on single foods or food groups and then by nutrient-based research findings.

This approach has been recently proposed again as a systematic strategy for the review of evidence underpinning dietary guidelines [118, 132]: evidence supporting healthy dietary patterns provides the foundation for the development of dietary guidelines, whereas further reference to individual foods and nutrients follows from the foundation of healthy dietary patterns. To put this in context, the 2015 Dietary Guidelines Advisory Committee in support of the 2015 Dietary Guidelines for Americans focused its evidence review and recommendations on healthful dietary patterns instead of individual nutrients or foods [30]. Due to remarkable consistency in the findings over different disease outcomes and dietary pattern identification methods, the Committee review showed that a "healthy dietary pattern" is higher in vegetables, fruits, whole grains, low- or non-fat dairy, seafood, legumes, and nuts; moderate in alcohol (among adults), lower in red and processed meat, and low in sugar-sweetened foods/drinks, and refined grains. in addition, the core features of this healthy diet can be obtained through many different healthy dietary patterns, potentially accommodating varying individual needs and socio-cultural preferences [30].

As a final note, dietary recommendations based on foods and dietary patterns are likely to be more accessible to the general audience. It would be easier for people to understand and adopt recommendations regarding cohesive dietary patterns, as opposed to those regarding several different nutrients. Additionally, dietary patterns have been shown to be more stable than single food groups over time [37], and this should help to communicate public health messages that remain consistent as much as possible. However, it is still essential to integrate information from dietary patterns and single components because the former approach cannot be specific about the particular dietary components responsible for the observed differences in disease risk, and it may thus not be very informative about biological associations between these components and disease risk [56].

7. HOW TO DEFINE DIETARY PATTERNS: AVAILABLE APPROACHES

Dietary patterns are combinations of dietary components (food items, food groups, or nutrients) intended to summarize the total diet or key aspects of the diet in free-living individuals. The majority of published reviews organize the statistical methods for dietary pattern analysis into three categories that are described in the following [67, 68, 89, 85, 33, 92, 70, 47, 135, 107]:

- **a priori**, **investigator-driven**, **investigator-defined**, or **dietary indexes/scores**: patterns are specified by researchers a priori based upon scientific evidence or theory for specific diseases and, generally, include foods or nutrients supported by current nutrition guidelines, recommendations, and/or a specific dietary composition that is considered healthful;
- **a posteriori**, **exploratory**, **empirically-derived**, **data-driven**, or **"data-driven, response-independent"**: the patterns emerge a posteriori from an analysis of dietary data – generally based on multivariate statistics – (i.e., data-driven) and the patterns are derived independent of their potential relationship to a health outcome (i.e., response-independent);
- **mixed-type**, **hybrid**, or **"data-driven, response-dependent"**: the patterns emerge from an analysis of dietary data – generally based on multivariate statistics – (i.e., data-driven) expressly used to examine the relationship between dietary patterns and a health outcome (i.e., response-dependent).

Note that previous names are simply shorthand notations to refer to how the patterns are derived. Data-driven does not mean that a method is more evidence-based, and investigator-defined does not mean a method includes more subjectivity. Each method is built on evidence and includes some degree of subjectivity [70].

All approaches allow assessing and/or ranking and quantifying adherence of study participants to these patterns, which is needed to evaluate their association with disease (e.g., cancer) risk within a multiple regression model, including confounding variables.

8. A PRIORI OR INVESTIGATOR-DEFINED APPROACHES

A priori or "investigator-defined" methods compare subjects' diet against a pre-specified evidence-based benchmark diet and express how individuals adhere to the benchmark diet with a score [67, 85, 92]. Benchmark diets are built upon scientific evidence/theory for specific diseases or include foods or nutrients supported by current dietary guidance, recommendations, and/or a specific dietary composition (for instance, Mediterranean, vegetarian, vegan, or gluten-free diets) that is considered healthful. The subject's dietary intake (from food items/groups or nutrients) is scored on the basis of each component of the benchmark diet following the adopted scoring system; single scores are then combined into a total score using the proper mathematical expression (e.g., sum or ratio). Typical examples in the literature include the Diet Quality Index, Healthy Eating Index, Recommended Food Score, Dietary Approaches to Stop Hypertension Index, World Cancer Research Fund Index, Mediterranean Diet Score, and total Plant-based Diet Index [135]. Using these indices answers the question "How close is the population to meeting a certain benchmark diet, expressed as a dietary recommendation or a specific dietary composition?" For this reason, they are sometimes called measures of diet quality [70].

Among major advantages, "investigator-defined" dietary patterns generally characterize overall diet, they are intuitively appealing, analytically simple to compute (e.g., primarily sums), easily reproducible and comparable. In index-based summary analysis, the dimensions of the pattern and how those dimensions are scaled are specified (and thus standardized) by the researcher based on external evidence regarding what constitutes a healthy diet [124]. Results can be meaningful, interpretable and are generally well associated with health outcomes, including cancer [101, 38, 46].

The major limitation of "investigator-defined" dietary patterns is that scores are defined on, thus reduced to, current knowledge and understanding of diet-disease association. In addition, it is challenging to translate the inherently qualitative concept of diet quality and its variants into quantitative mathematical formulas. While dietary scores are multidimensional in design, the end product is generally one number – the summary score – that may provide little information about the contributing components. This is especially true for individuals with a middle-range score, who might have different dietary behaviors. Construct and content validity should be assessed for newly developed "investigator-defined" dietary patterns. Subjectivity is introduced in the interpretation of the guidelines (if any) and in the construction of the scores (which foods are selected for inclusion in each component). However, a big effort in standardizing index construction has been recently made within the Dietary Pattern Pooling Project sponsored by the US National Institutes of Health [74]. No research has established the preferable scoring system for specific situations. The summation of equally weighted dietary component scores implies that each component is equally important and additively related to health. This might not be nutritionally meaningful and may be different for different diseases. More than one "investigator-defined" dietary pattern is sometimes available to measure the same benchmark diet, with differences in dietary components included, structure (e.g., sum or ratio), processing of dietary variables, component weighting (i.e., equal weights or not), and cut-off points (i.e., population-specific or absolute). The different (recognized) variants of Mediterranean diet have prevented so far a meaningful recommendation in favor of Mediterranean diet to fight against cancer, although a strong evidence of an effect on cancer risk was recognized and the association was judged causal [19, 84].

9. A POSTERIORI OR "DATA-DRIVEN, RESPONSE-INDEPENDENT" APPROACHES

In nutritional epidemiological settings, data-driven methods estimate dietary patterns directly on dietary data [89, 47], and do not explicitly refer to a priori information in the identification of dietary patterns. To highlight this aspect, these methods are also indicated as exploratory approaches. Variability in dietary habits might be explored in the overall population, with no further reference to disease outcomes. In this case, we have "data-driven, response-independent" methods. Among these "data-driven, response-independent" methods, the most used in nutritional epidemiology are PCA, EFA, and CA [135].

In most applications in nutritional epidemiology, PCA, EFA, and CA have been applied directly to the food-group data matrix. While food groups provide the most immediate dietary pattern interpretation [55, 131], they are far from continuous variables, as PCA, EFA, and most standard clustering approaches would require. Fewer applications have identified nutrient-based dietary patterns by following standard input data requirements [32, 35, 23]. From the epidemiological point of view, nutrient-based dietary patterns have been suggested instead of the food-group-based ones when the study target is the comparison of dietary patterns across populations from different countries [86, 24]. A polychoric correlation matrix has sometimes been adopted as the input data matrix for standard PCA/EFA to account for binary (i.e., non-consumption versus consumption) and ordinal food groups. This opens the possibility of tailoring PCA, EFA, and CA variants to discrete metrical data instead of scale/metrical data.

9.1 Principal Component Analysis and Factor Analysis

The aim of PCA and EFA is to reduce the dimensionality of the data by transforming an original, more extensive set of correlated food groups/nutrients into a smaller and more easily interpretable set of uncorrelated variables, called principal components or factors. Both approaches answer the following question: "What are the major components/factors in a population under study, i.e., those contributing most to the variation of nutrient/food group intakes reported by study participants?" [70].

To answer this question, PCA uses the singular value decomposition and identifies principal components based on the covariance/correlation matrix of the input variables (nutrients or food groups). The resulting components are linear combinations of the original variables with suitable weights (loadings) that explain as much of the variation in the original variables as possible [47]. EFA starts from the same covariance/correlation matrix and shares the data reduction rationale of PCA, but it is based on a statistical model where the random vector of observations (i.e., individual's dietary data) is explained in terms of some latent common and specific factors. The definition of a statistical model allows rotation of the factor loading matrix, improving the interpretation of the identified factors. EFA may use different estimation methods for model parameter estimation, including PCA and maximum likelihood. Therefore, a principal component factor analysis is defined as an EFA where the PCA method is adopted for parameter estimation. Principal component factor analysis is the more common method used so far to derive a posteriori dietary patterns in nutritional epidemiology [33].

The most followed approaches to select the appropriate number of principal components/factors are eigenvalue greater than 1 criterion (when the correlation matrix is adopted), visual inspection of the scree-plot, and a sensible interpretation of the dietary patterns. A fixed threshold (generally 5%) can also be decided, and only the components/factors whose explained variance exceeds the chosen threshold are incorporated in the analysis [47].

Following both approaches, individuals are ranked based on the degree to which their diets conform to each of the identified factors; this is done by adopting continuously scaled scores either obtained by simple matrix algebra (PCA) or by different estimating procedures (EFA).

Scores are further entered into a regression model for disease (e.g., cancer) risk estimation. In this case, scores categorization into quantiles improves results interpretation because scores, although continuous, have a restricted scale and no measurement unit. Unlike most a priori indexes, any one principal component/factor does not represent the entire eating pattern for any individual or group because the principal components/factors are not mutually exclusive. However, each person's overall eating pattern can be inferred by assessing his/her multiple principal component/factor scores [70]. The simultaneous inclusion of all dietary patterns in the same regression model without additional multicollinearity issues is granted by scores being uncorrelated by design (PCA) or by additional orthogonal rotation (EFA).

This implies that the effect on disease (e.g., cancer) risk of each dietary pattern can be easily adjusted for the remaining dietary patterns.

In disentangling the overall diet into a few summarized profiles, PCA and EFA provide an immediate representation of how food groups/nutrients interact within each dietary pattern. This happens through the principal component or factor loadings. Indeed, the magnitude of each loading for a given component/factor measures the importance of the corresponding food group/nutrient to that component/factor [33, 47]. If a few loadings are high (in absolute value) on a pattern, this pattern will be mostly characterized by pairwise interactions of the corresponding food group/nutrients. Results are generally meaningful, interpretable, associated with health outcomes, including cancer [33, 6, 45], and show modest but raw reproducibility across populations [4, 77, 86].

PCA and EFA have also limitations and challenges. Subjectivity is introduced in constructing food groups or selecting nutrients, in preprocessing input variables, in choosing which data matrix to work on (e.g., covariance or correlation matrix, separate analyses for known subgroups in the data or not) or the number of factors to retain, in the opportunity for factor rotation and which rotation to choose (in EFA only), and in the identification of some criteria for labelling the factors. Unless data collection methods are comparable, input variable choice/preprocessing is standardized, and the dietary pattern identification method is the same, results are not comparable across studies. Even when the previous aspects are standardized, labelling of principal components/factors is still very subjective. In principle, variables corresponding to "large" loadings are interpreted as being important for describing the original data; variables corresponding to "small" loadings can be discarded. However, such interpretation is complicated by the fact that all component/factor loadings are nonzero. Various cut-off rules, rotation strategies, and other procedures have been developed to simplify interpretation, but these largely ad hoc procedures do not contribute to the transparency or objectivity of PCA/EFA. Alternatives to PCA/EFA that offer more interpretable components/factors by forcing loading patterns to include many loadings exactly equal to zero (i.e., by forcing the identification of "sparse" components) are interesting because they reflect current nutritional knowledge where each pattern is typically described by a small subset of food groups/nutrients [65]. In addition, since the label generally needs to be short, often they do not adequately convey to what the underlying principal component/factor is [108]. This further makes comparison of results across studies more challenging. Few rigorous statistical procedures have been implemented to examine internal consistency and validity of the identified solutions [85], although some efforts have been made to assess reproducibility and validity of PCA-based or EFA-based dietary patterns in more recent years [36].

Confirmatory factor analysis (CFA) [50] has not been so much used in nutritional epidemiology so far [126]. The main difference with EFA is that CFA involves specifying the number of factors, which variables will load on each factor (i.e., by putting some 0s in the factor-loading matrix structure instead of estimating each loading) and which relationship exists between each pair of latent factors [66]. In this way, CFA is able to perform hypothesis testing on the overall factor structure and on factor loadings of food groups/nutrients to estimate rigorously the number of factors and identify food groups/nutrients contributing significantly to those factors [85, 83]. Alternatively, CFA can be employed to test the goodness of fit and validity of the factor structure of dietary patterns in a second step, i.e., after PCA or EFA was performed [92, 103]. It is not scientifically confirmed if the results are more accurate than those obtained in a single-step process [90]. With this motivation, some studies performed CFA as a one-step approach instead of using previous results from PCA or EFA.

The advantage of CFA compared to the two methods previously presented is that a latent structure can be specified and tested, and if the researcher has some additional a priori knowledge, this can be incorporated into the model [5].

9.2 Cluster Analysis

While PCA and EFA work on data matrix columns (i.e., food groups/nutrients), CA [9] is traditionally used in nutritional epidemiology to explore data matrix rows (i.e., individuals), to identify groups of participants (clusters) with a specific dietary behavior, based on a pre-specified measure of similarity/difference in food group/nutrient intake among individuals [28]. So, CA replies to the following question: "Are there groups of individuals characterized by distinct dietary patterns?" [70].

On the opposite of PCA and EFA, where the subjects can belong to more than one principal component/factor, CA provides one group belonging indicator for each subject. The group belonging indicator is then entered into a multiple regression model with confounding factors to estimate adjusted disease (e.g., cancer) risk related to specific group belongings.

Hard clustering, where study participants are grouped into mutually exclusive clusters, and individuals only belong to one cluster, is by far the most followed approach in nutritional epidemiology [47]. Among available methods, K-means and Ward's minimum-variance method are widely used, with one paper only [75] comparing previous approaches with the flexible beta one.

The K-means algorithm partitions observations into K clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or centroids) by minimizing within-cluster variances (squared Euclidean distances). The advantages of K-means clustering include its simple interpretation, low computation complexity, fast calculation speed, and suitability for large samples.

Ward's minimum-variance method is an agglomerative hierarchical clustering algorithm, with the number of clusters changing at each step [49]. Indeed, at each step, one has to find the pair of clusters that leads to a minimum increase (i.e., a weighted squared distance between cluster centers) in total within-cluster variance after merging. Thus, the calculation is slow, and the Ward's approach is hard to apply in large samples [135].

Among the advantages of hard CA, distinct subgroups of individuals – where everyone belongs only to one specific dietary pattern group – are easy to interpret and relate to disease (e.g., cancer) risk. The dendrogram from Ward's method shows the clustering process and results visually [47].

Among limitations, uncertainty in individuals' classification is removed, and each individual is assigned to a cluster with a probability of 1 or 0 [49]. Second, subjective decisions are required at several steps and include selection of input variables (i.e., nutrients, food groups, or factor scores) and data preprocessing, similarity measure, clustering algorithm, initial values, and number of clusters [47]. While some objective methods for selecting appropriate clustering algorithms and the number of clusters exist, the reproducibility of results cannot ensure their ability in representing actual dietary behavior and its relation with cancer risk [75]. Third, sensitivity of the CA to outliers is also an issue. Fourth, formal comparison of results from different clustering algorithms is not as easy as with PCA/EFA, where congruence coefficients between factor loadings are recognized as the preferred method to compare different solutions [16]. Fifth, the CA output is just a group belonging indicator, and there is no PCA-based or FA-based loading matrix to provide an immediate representation of how food groups/nutrients interact within each dietary pattern. So, when the CA is completed, further analyses to compare dietary, socio-demographic, or socio-economic profiles across clusters must be used to interpret the identified patterns [56]. Sixth, a major drawback of both K-means and Ward's method is their tendency to create spherical clusters of equal volume [49], which leads to biased clustering solutions when this assumption is not met by the data. In addition, another limitation of Ward's method is its tendency to create clusters with an equal number of observations, which is an unrealistic scenario in nutritional epidemiology [49]. Seventh, the reference category to estimate the effect of each dietary pattern on disease (e.g., cancer) risk is not as naturally identified as in regression models based on PCA/EFA-based dietary patterns. As one group has to be chosen as the reference group to express disease risk related to the remaining groups, what are the main characteristics this reference group should show? Should it be the bigger one or the one with a more balanced diet? No consensus exists on this issue. Floating absolute risk method has been used once to analyze the association between CA-based patterns and breast and ovarian cancers, to overcome this issue [32].

The term finite mixture model refers to a convex combination of a finite number of probability distributions, each of these commonly designated as mixture component. Mixture models can therefore be viewed as model-based clustering, where clusters are groups of individuals in the data (with a similar dietary behavior) induced by mixture components. Classification uncertainty is taken into account by estimating individual probabilities of belonging to each identified cluster, based on available data. For this reason, this approach provides an example of soft clustering, where subjects are assigned to each class with a weight equal to posterior membership probability for that class. The unknown true parameter vector (including parameters of each mixture component and the mixture weights) as well as the unknown allocation variables are estimated by the maximum likelihood method using the Expectation-Maximization (EM) algorithm [94, 49] in the frequentist framework. As an alternative to the EM algorithm, the mixture model parameters can be deduced using posterior sampling as indicated by Bayes' theorem. This is still regarded as an incomplete data problem whereby membership of data points is the missing data. A two-step iterative procedure known as Gibbs sampling can be used. In both cases, a "hard" clustering solution can then be obtained by simply assigning each observation to the cluster to which it belongs with the highest probability, following Bayes' Theorem.

In finite mixture models, the number of components is still pre-specified to be equal to K. Covariates can be accommodated [41, 99] or not [49] within the model to describe dependence of main dietary variables on other lifestyle or anthropometric variables. In the former case we have the general multivariate mixture model, also denoted by regression mixture model. In the latter case, standard mixture models are assumed and fitted. Both approaches have been applied in nutritional epidemiology in the frequentist framework [96, 94, 111, 41, 99, 49, 20, 21].

The paper by Greve et al. [49] shows the higher flexibility of Gaussian mixture models as implemented by Fraley and Raftery [44] in comparison with K-means and Ward's hierarchical clustering on simulated and real-life data. Even on simulated data with spherical clusters of equal volume, the clustering solutions obtained from this Gaussian mixture model were more similar to the true cluster structure than those obtained from the K-means algorithm or Ward's method in more than 72% of all simulated data sets. For simulated data sets with clusters of variable volume, shape and orientation, the Gaussian mixture model achieved a higher agreement with the true cluster structure in more than 90% of data sets [49]. The decomposition of the variance-covariance matrix proposed in this approach enables the researchers to place constraints on the geometrical properties of the clusters and thus to specify a desired degree of flexibility in terms of cluster volume, shape, and orientation. The choice of the number of clusters and of the available models (i.e., different parameterizations of the variance-covariance matrix) is transformed into a model selection problem. The final model is then identified according to information criteria after the finite mixture model is fitted by setting different values for the number of clusters or imposing different restrictions on the variance-covariance matrix [49].

Because finite mixture models has many parameters, large samples are generally required, especially when the number of selected clusters is moderate-to-high. Thus, a restricted mixture model is proposed that reduces the number of parameters and is suitable for small-to moderately-sized samples [99]. This regression mixture model has been shown to generalize the one by Fahey et al., also applied in nutritional epidemiology [41], and includes in the same model the Gaussian and binomial distributions for modeling different forms of food group consumption within the exponential family.

The finite mixture model method can also be used to classify the population according to the factor scores derived from EFA-based CFA. This so called two-step classification approach combines the advantages of both finite mixture models on food items to identify mutually exclusive clusters and CFA to understand which foods are eaten in combination [111].

Among major advantages, finite mixture models are more oriented towards capturing real-life dietary behavior because they can account for the within-cluster correlation between dietary variables [49], allow the variances of dietary components to vary within and between clusters, and enable covariate adjustment (e.g., age, sex, non-alcohol or total energy intake) for food intake simultaneously with the fitting process [41, 99, 20, 111].

Among major issues, the observed data may violate the distributional hypotheses, which are *per se* difficult to adhere to in a multivariate setting. When there are many 0 values – indicating non consumption – the need to deal with them increases the models complexity, as does the high number of parameters to be estimated [41, 99]. In addition, sensitivity to the initial values, convergence to local extremum, and slow convergence speed have been reported for most approaches [135]. Finally, although finite mixture models improve on selection of the optimal number of clusters compared to traditional CA approaches, still this selection is not made simultaneously within an overall parameters' estimation process [94, 49, 20, 21].

9.3 Treelet Transform

A composite of hierarchical CA and traditional PCA, treelet transform [72] provides an improvement over traditional PCA and an important contribution to clustering methodology. For clustering methodology, it provides a framework which actively searches for the correct underlying correlation structure of the data. Its improvement over PCA happens especially when the correlation matrix is believed to be sparse, as in the analysis of dietary patterns, and is generally worth note [123].

Firstly introduced in nutritional epidemiology in 2011 [48], treelet transform is a dimensionality reduction technique aimed at converting a set of observations of possibly correlated variables into orthogonal components. Similarly to PCA/EFA, identifying the optimal number of retained components is based on scree-plot inspection (and related percentage of explained variance) and interpretability, and interpretation/labelling of components are based on loadings. Scores are determined for each component of the treelet transform and measure adherence to a given component. Unlike PCA scores which are always uncorrelated, treelet transform scores generally have a small degree of correlation.

Treelet transform combines the quantitative pattern extraction of PCA with the interpretational advantages of hierarchical clustering of variables. The two variables showing the highest covariance/correlation are identified in the treelet transform, and a PCA is performed on them. They are then replaced with the score of their first principal component, and a merge is indicated in the cluster tree. This operation is re-iterated until all variables have joined the cluster tree. In this way, the treelet transform produces a hierarchical grouping of variables that may reveal the data structure's intrinsic characteristics.

By combining PCA and hierarchical clustering, treelet transform introduces sparsity into principal components (i.e., making many loadings equal to 0), thus potentially simplifying the interpretation. While EFA achieves this structure in a post-hoc analysis with the use of factor rotation and loading truncations (in which the factor loadings with absolute values smaller than an arbitrary threshold are ignored) [65], treelet transform directly derives sparse components that, similarly to PCA components, account for a large part of the variation in the original data and can be used analogously; treelet transform leads to an associated cluster tree that provides a concise visual representation of loading sparsity patterns and the general dependency structure of the data [48].

Alongside the cluster tree, treelet transform yields a coordinate system for the data at each level of the cluster tree. Selecting a cluster tree level (cut-level) amounts to choosing the level of detail desired in the data dimensionality reduction [2]. As pointed out in one of the Discussions in the original paper, the use of treelet transform leads to a tradeoff between the amount of variability explained and sparsity. The objective is to "make the results as sparse as possible but not any sparser" [2]. Treelet transform has been used to derive dietary patterns using nutrients [2], food items [106], and food groups [48, 93].

Treelet transform works best for dimensionality reduction and/or feature selection when sample sizes are relatively small, and the data are sparse, with unknown groupings of correlated variables. Unlike PCA and its subjective choice of components to retain, treelet transform has a fast and efficient algorithm for determining the optimal number of dietary patterns to be retained and assessing dietary pattern internal reproducibility [107]. Each derived component can involve only a small number of input dietary data, so pattern structure is generally simpler than in PCA; the corresponding cluster tree for all variables also supports pattern interpretation. Not only treelet transform provides a concise visual representation of loading sparsity patterns, but it also shows the data's general dependency structure [48].

In line with other CA techniques, a degree of subjectivity exists in choosing the cut-level of the cluster tree before extracting components. When the cut-level is close to the root, most variables are included in the components, with potential interpretation issues; the information is comparable to PCA output when all variables contribute to treelet components. As the cut-level moves away from the root, the component loadings become sparse (as many are equal to 0), and the components possibly become more interpretable [107]; however, this may lead to components that do not capture dietary complexity and are therefore not informative [2]. Cross-validation can be used to identify an optimal cut-level. Once the cut-level is chosen, the loadings computed are invariant to the number of retained components; hence, the number of components is an a priori parameter to be specified in the cross-validation step. In addition, if the correlation of some nutrients/food groups is too strong, then the sparsity hypothesis may not hold [57]. It also remains debatable whether patterns derived by treelet transform are more effective than those from PCA/EFA or CA methods in exploring the relationship with health outcomes, including cancer [106].

10. MIXED-TYPE OR "DATA-DRIVEN, RESPONSE-DEPENDENT" APPROACHES

Although previous multivariate statistical methods are frequently used to identify a posteriori dietary patterns and study their relationship with health outcomes, neither is expressly designed to derive dietary patterns that are predictors of disease. Therefore, the resulting dietary patterns may be significantly associated with a disease (e.g., cancer), but their interpretation is not that such patterns are the best (or worst) possible for disease prediction.

Disease prediction may be achieved in two ways:

- the disease is directly modeled as the response variable;
- other intermediate variables related to the disease are chosen as response variables in the path between dietary exposure and disease.

Both cases are included under the "data-driven, response-dependent" methods developed and applied to nutritional epidemiology.

10.1 Reduced Rank Regression

Reduced rank regression was formally introduced in nutritional epidemiology in 2004 to combine the advantages of the a priori and a posteriori approaches [55]. A posteriori pattern method can be applied to investigate if major consumption patterns of a specific population have relevance for health outcomes; a priori patterns can help clarify if adherence to specific benchmark diets is related to reduced disease risk. Reduced rank regression fills a gap as it more directly relates the step of data-driven pattern identification to the health/disease outcome of interest. In detail, after about 20 years, researchers in nutritional epidemiology are still increasingly interested in identifying specific dietary patterns related to established and new pathways in the development of major chronic diseases, including cancer [127]. So, reduced rank regression replies to the following question: "What combinations of dietary components explain the most variation in a set of intermediate health markers? And then, in further analysis, does that pattern explain the disease outcome of interest?" [70].

The method has similarities with PCA, but it uses two different sets of variables: a set of independent variables or predictors, generally dietary variables, and a set of response variables, expected to be associated with the disease under examination based on a priori knowledge [47]. Food groups are generally used as predictors. Nutrient intakes, contaminants, and endogenous biomarkers or intermediate disease phenotypes are generally used as response variables according to the specific research question [135].

The identified dietary patterns are projections of the principal components obtained from the covariance/correlation matrix of the responses onto the space of predictors. They can be interpreted as those linear functions of the original dietary variables that maximally explain variation in the response variables. Reduced rank regression can therefore be interpreted as a PCA applied to responses and subsequent linear regression of principal components on predictors, although it is somewhat more efficient and sophisticated than this two-step procedure [127].

Reduced rank regression starts from a linear function of responses called response score that will then be projected onto the space of predictors to produce a factor score, that is, a linear function of predictors. Both scores form an inseparable pair reflecting the same latent variable in different sets of original variables. Because the first aim of this method is to explain a high proportion of response variation, the evaluation of factors extracted by reduced rank regression should be based on response scores rather than factor scores. However, similarly to PCA and EFA, factor scores represent the comprehensive variables used in subsequent statistical analysis, to assess the association between the identified dietary patterns and disease outcome [55].

Given that linear functions of predictors are defined as principal components of the responses, there will be as many dietary patterns as were selected responses. This is different from PCA, in which the analysis cab identify as many principal components as predictors. Still, similar to PCA, only a subset of patterns might be informative [47]. In reduced rank regression, only one or a few principal components of responses might account for most of the responses' variation, and the corresponding pattern would be the strongest candidate to be selected [127].

The exploratory nature of the approach hampers the generalization of the results from one population to others. However, compared to PCA and EFA, reduced rank regression allows the use of the same set of response variables in different study populations and thus improves results comparison across different populations [33]. In addition, the so-called simplified patterns (i.e., scores are calculated based on only those food items strongly contributing to the pattern, to which all contribute with equal weight) ease the step of external validation of pattern-disease associations. Such validation is highly recommended given that pattern identification (using informative responses) is usually carried out within the same study in which the pattern-disease association is evaluated [127].

Major challenges in performing reduced rank regression include the choice of predictors and response variables to work on, the number of response variables to consider and their relationship, the number of principal components to retain, and their labelling. Most studies try to cover the whole diet including all available food items or food groups and very few studies investigated the influences of the selection or building of food groups on the extraction of dietary patterns [127]. The value of the application of reduced rank regression is considerably dependent on a good selection of response variables. If responses with no or little intercorrelation are selected for analysis, they will unlikely be well reflected by a single response score. Instead, it is more likely that only a fraction of responses is accounted for – in the worst-case scenario, single response scores (and consequently single reduced rank regression patterns) reflect only a single response variable. In this case, more than a dietary pattern have to be retained and these dietary patterns might reflect different pathways relevant to disease risk [127]. Confounding for responses is a second crucial issue. The derived pattern may considerably differ depending on an adequate consideration of confounders [127]. So far, medication and anthropometric characteristics have been considered in some studies as confounding factors, but there could be scenarios where the role of confounding factors is less obvious.

10.2 Partial Least Squares

A method similar to reduced rank regression is partial least squares, a regression model of multiple predictor variables on multiple response variables, sometimes used in nutritional epidemiology in comparison with PCA and reduced rank regression (e.g., [55, 29, 81]).

The partial least squares method is a compromise between PCA and reduced rank regression [55, 29, 81]. Indeed, while PCA selects factors that explain as much predictor variation as possible and reduced rank regression extracts factors that explain as much response variation as possible, partial least squares balances the two goals of explaining predictor variation and explaining response variation [92]. So, partial least squares replies to the following question: "What combinations of dietary components explain the most variation in dietary components and in a set of intermediate health markers? And then, in further analysis, does that pattern explain the disease outcome of interest?" [70].

The three methods are similar in terms of their mathematical foundation and their technique of deriving factors. For each method, the coefficient vectors of the extracted linear functions are eigenvectors of a covariance matrix. PCA uses the covariance matrix of predictors, whereas reduced rank regression starts from the covariance matrix of responses. Partial least squares uses the matrix of covariances between predictors and responses. However, as the number of factors cannot be greater than the rank of the corresponding covariance matrix, partial least squares can extract as many dietary patterns as were the selected predictor variables, like in PCA, but differently from reduced rank regression. The eigenvalue belonging to an eigenvector quantifies the percentage of variation explained by the corresponding linear function of the original variables (i.e., predictors or responses depedenign on the method). The factors obtained by PCA, partial least squares, and reduced rank regression usually are sorted by decreasing eigenvalues. While the first factor of PCA is the linear function of predictors that maximizes the explained variation in predictors, it is, in general, not optimal in terms of response variation. In contrast, the first factor of reduced rank regression explains more variation in response than any other linear function of predictors, but possibly explains only a moderate fraction of predictor variation. The first factor from partial least squares maximizes the covariance between linear combinations of predictors and responses [55].

Due to the orthogonality of eigenvectors, successive extracted factors from all three methods are uncorrelated. Therefore, the variation in the original variables (i.e., predictors or responses, depending on the method), can be decomposed into percentages of variation explained by the obtained factors. These uncorrelated factors can simultaneously be chosen as independent variables in a regression model for predicting disease (e.g., cancer) risk without confounding each other [55].

Similarly to reduced rank regression, major challenges in performing partial least squares include the choice of both predictors and response variables to work on, the number of response variables and food groups to consider and their relationships, as well as the number of factors to retain, and their labelling.

10.3 Classification and Regression Tree

CA has a parallel methodology that defines distinct subgroups in a population while making full use of information on a response variable. This is "classification and regression tree" analysis [11]. Classification and regression tree is a non-parametric decision tree procedure (i.e., selection of nodes with successive splitting to produce different subgroups) that identifies mutually exclusive and exhaustive subgroups of subjects sharing common characteristics that are associated with the response variable of interest [73].

Compared to reduced rank regression, decision tree analysis uses one response variable only. This may be a disease risk factor, including also overall measures of diet quality [51], or the disease outcome [13]. Therefore classification and regression tree could be used to answer the question: “What combinations of dietary components explain the most variation in one response variable, like the disease outcome or a selected risk factor for the disease?” [70]. The response variable can be either categorical (in the so called classification tree analysis) or continuous (in the so called regression tree analysis), whereas independent variables can be any combination of categorical and continuous variables. No data assumptions are required [73, 92].

Decision tree analysis ends up with a graphical output that is a multi-level structure that resembles branches of a tree, with the root node and several leaf nodes. A classification rule is a path from the root node to a leaf node associated with the response variable. The results can thus be interpreted as “hierarchical” dietary patterns. For example, one might find that in predicting a health outcome the most important variable is the amount of added sugars consumed. If intake of added sugars is high, the next most important factor in predicting the health outcome might be the amount of, say, solid fats, but, if added sugars are low, the next most important factor might be fruit intake. The terminal nodes show the specific pattern features of the sub-populations in percentage, including the number of participants and the probability or mean values of the response variable in the terminal node [119].

Until now, decision tree analysis was seldom applied to derive dietary patterns [51] or risk-related patterns, including dietary and other risk factors [13].

Among major advantages, classification and regression tree can be used to reveal heterogeneity in the dietary behavior of a population, when present, and to develop preventive measures tailored to specific sub-populations. In principle, the output is very intuitive and, due to its transparent nature, users can trace back through the generated model. When selected rules involve a few variables, this approach may allow to demonstrate the effect (on disease risk) associated to modifications of single dietary habits, prompting to an individualized approach to public health messages [71]. Decision tree analysis is also able to generate new aetiological hypotheses, without prior assumptions on potential risk factors. It might be particularly suitable in identifying disease risk based on a combination of food groups and other non-dietary risk factors. In this sense, it is crucial that this approach allows for independent variables of any kind (i.e., categorical or continuous variables).

Major disadvantages have likely avoided a wider diffusion of classification and regression tree in nutritional epidemiology. In particular, one key variable can dominate the model [135], misclassification can be rather large [70], and overfitting may be a serious issue [51]. If many classification rules are generated, the selection of meaningful rules will require considerable professional knowledge. Rules containing many variables can be long and/or complex even if they are meaningful, making it difficult to translate them into simple health recommendations [135].

Other data mining techniques, such as random forest, artificial neural networks, and Naïve Bayes Classifiers, have also been used to analyze the relationship between dietary patterns and diseases in a few applications [51, 31, 8].

10.4 Least Absolute Shrinkage and Selection Operator

In nutritional epidemiology, multicollinearity may be strong between nutrients, as well as between food groups. In observational studies, several confounding factors potentially measure slightly different variants of the same lifestyle characteristics (e.g., physical activity or alcohol drinking consumption). Interactions may also exist between dietary exposures and confounding factors (e.g., between dietary patterns and alcohol consumption or education). This makes fitting a standard regression model for assessing disease risk challenging.

Least absolute shrinkage and selection operator (LASSO) is a regression-based methodology that allows a large number of covariates to be included in the model and penalizes the absolute value of the regression coefficients, thus, regulating the impact a coefficient may have on the overall regression. The greater the penalization, the greater the shrinkage of coefficients (some reaching 0), thus automatically removing unnecessary/uninfluential covariates. Thus the LASSO minimizes regression coefficients in order to reduce the likelihood of overfitting. The algorithm shrinks the sum of the absolute value of regression coefficients, producing coefficients that are exactly 0, and thus selecting the nonzero variables to remain in the model. The shrinkage amount is controlled by the shrinkage parameter λ.

A critical choice in the LASSO method is selecting the appropriate amount of shrinkage, as controlled by λ.

We identified two papers [134, 80] applying LASSO in nutritional epidemiology. The former paper [80] adopts the logistic LASSO in the National Health and Nutrition Examination Survey (NHANES) 1999-2010 to estimate the association between a composite risk pattern, including diet and other risk factors, and self-reported breast cancer (binary variable, no/yes), in females ≥ 50 years. Based on 29 variables, including 21 macro- (density method) and micronutrients, alcohol, and coffee, the following variables: age, parity, vitamin B12, caffeine, and alcohol remain in the model, as far as the penalty parameter λ increases in the

LASSO. This paper adopts the cross-validation technique to choose the optimal λ value.

In the latter paper, LASSO is not directly used to identify a disease-related risk profile, but as an alternative approach for assessing disease risk. Based on FFQ data from the NHANES 2005-2006, including healthy US adults ($n = 2609$), ten PCA-based dietary patterns (65% of the total variation) are associated with cardiovascular disease risk factors via the LASSO method. It is shown that the LASSO method outperforms the traditional linear regression model, by better predicting the levels of triglycerides, LDL cholesterol, HDL cholesterol, and total cholesterol (LASSO adjusted $R^2 = 0.861$, versus traditional linear model adjusted $R^2 = 0.163$). The study adjusts for confounding factors such as age and body mass index. The authors also test the prediction accuracy of the model performance using an independent test set.

11. NOVEL STATISTICAL APPROACHES IN THE ANALYSIS OF A POSTERIORI DIETARY PATTERNS

Maximum likelihood may be used in EFA, and is generally used in finite mixture models in nutritional epidemiology. However, in some applications and especially for the EFA, in order to regularize the factor loadings, priors or penalties are used to induce sparsity [76, 15, 40]. This might be particularly interesting in the identification of a posteriori dietary patterns. Sparse factor loadings can be used to identify specific subsets of nutrients and interpret them as interacting subsets, helping elucidate the name chosen for a dietary pattern. In contrast to the frequentist approach, Bayesian methods model sparsity through the introduction of shrinkage and sparse priors on EFA. Sparse latent factor models therefore exploit sparsity-inducing priors as an integral part of the identification of dietary patterns. In the Bayesian approach to sparsity, two main priors have been widely used and developed.

The first one is the Bayesian LASSO prior, introduced by Park and Casella (2008) [95], and developed in different settings [15, 100]. Based on the LASSO penalty of Tibshirani (1996) [121], the Bayesian LASSO prior is a conditional Laplace prior for the loadings λ_{pj} with $p = 1, \ldots, P$ variables and $j = 1, \ldots, J$ factors

$$\lambda_{pj}|\psi_p \sim \frac{\tau}{2\sqrt{\psi_p}} e^{-\tau|\lambda_{pj}|/\sqrt{\psi_p}},$$

where ψ_p is the diagonal element of the covariance error matrix and $\tau > 0$ is the scale hyper parameter. In this modeling setting, the posterior mode of λ_{pj} is the LASSO estimate with the penalty equal to $2\tau\psi_p$, which regulates the amount of shrinkage. Posterior inference is developed via Gibbs sampling. This approach's major limitation lies in the lack of unimodality for the posterior distribution of λ_{pj}. Indeed, the posterior distribution of the factor loadings could present a bimodality, and this problem leads to point estimates less meaningful [95]. This problem can also occur considering the prior error variance ψ_p as proper.

The second approach is focused on a mixture prior [15] defined by the random variable δ_{pj} assigned to each element of the loadings $\lambda_{pj}, p = 1, \ldots, P$, $k = 1, \ldots, K$, of the factor loadings matrix $\mathbf{\Lambda}$:

$$\lambda_{pj}|\delta_{pj} \sim (1 - \delta_{pj})N(0, \zeta_{pj}^2) + \delta_{pj}N(0, c_{\lambda_{pj}}^2 \zeta_{pj}^2),$$

and

$$P(\delta_{pj} = 1) = 1 - P(\delta_{pj} = 0) = p_{pj}.$$

The priors for the factor loadings belong to the class of absolutely continuous spike and slab priors where ζ_{pj}^2 is a small constant, thus representing the spike of the factors, respectively, and so the distributions are concentrated on zero. Instead, $c_{\lambda_{pj}}^2$ are large constants ($\gg 1$), thus representing the slab part of the mixture of the factor loadings. This prior was used on identifying sparse latent factor models in dietary pattern analysis on 102 food items in young American adults [65]. This paper illustrates the potential of using EFA in a Bayesian perspective, by shrinking some loadings and improving dietary pattern interpretation, while potentially allowing for the incorporation of covariates that may provide important information when exploring dietary patterns or measurement error. Sparse latent factor models exploit sparsity-inducing priors as an integral part of the identification of dietary patterns. Indeed, prior distributions over individual probabilities are chosen to have substantial probability mass at zero to induce shrinkage of negligible loadings to zero. On the other hand, they ensure that the probability mass is spread over a wide range of plausible values so that important loadings escape shrinkage and take nonzero values [65]. In addition, the proposed sparse latent factor analysis robustly derives dietary patterns while simultaneously controlling for potential interaction with other variables, including total energy intake. Covariates are directly included as additional regressors in the model, instead of proposing preprocessing of input data to account for them (e.g., residual method for energy intake) or separate EFAs by relevant covariates to be ad-hoc combined. While controlling for influence of covariates, their information is also jointly used to derive dietary patterns [65]. Other possible extensions rely on the spike and slab with a LASSO prior [100, 3], but to our knowledge they have not been applied in nutritional epidemiology so far.

In many circumstances, researchers aim to provide insight into dietary patterns that emerge based on a given characteristic of the sample, for example a socio-demographic characteristic (e.g., age, education, or income) and a method called Focused Principal Component Analysis (FPCA) [14] is available. This method derives principal components, so it is a data-driven approach, but it is more similar to a "data-driven, response-dependent" method, where, however, the response is generally a confounding factor.

Unlike in PCA, dietary patterns focusing on a particular variable of interest (i.e., a population characteristic) are formed, and are presented exclusively in graphical format. Applying FPCA to dietary data makes it possible to view the correlation between each dietary variable and a given variable of interest, at the same time as enabling detection of correlations between the different dietary variables themselves. So, FPCA replies to the following research question: "How to represent the relation between food/nutrient consumption and one selected population characteristic, without loosing the relationship that the different food groups/nutrients have with each other?".

The FPCA method considers a set of P variables measured on n subjects. The variables are n-dimensional column vectors $\mathbf{x}_p$. The correlation matrix of each column vector $\mathbf{x}_p$ can be geometrically represented thus by the P points $\mathbf{p}_p$ on the unit hypersphere of an n-dimensional Euclidean space. The correlation of $\mathbf{x}_1$ and $\mathbf{x}_2$ is close to 0 if and only if $\mathbf{p}_1$ and $\mathbf{p}_2$ lie on perpendicular radii; on the opposite the correlation of $\mathbf{x}_1$ and $\mathbf{x}_2$ is close to 1 if $\mathbf{p}_1$ and $\mathbf{p}_2$ are neighbours; finally the correlation of $\mathbf{x}_1$ and $\mathbf{x}_2$ is close to -1 if $\mathbf{p}_1$ and $\mathbf{p}_2$ are diametrically opposed [43]. The smaller the radius, the stronger the correlation.

In order to do that, FPCA projects all the variables vectors onto the hyperplane perpendicular to one variable taken as reference (for example, $\mathbf{x}_1$):

$$Pr_1\mathbf{x}_p = \mathbf{x}_p - (\mathbf{x}_p^\top \mathbf{x}_1)\mathbf{x}_1,$$

The FPCA displays the vectors $Pr_1\mathbf{x}_p(p = 2, \dots, P)$ in a low-dimensional space, obtaining a $(P-1) \times (P-1)$ covariance matrix:

$$\Sigma_{p,k,1} = (Pr_1\mathbf{x}_p)^\top(Pr_1\mathbf{x}_k),$$

and projects them onto the space spanned by a few of the eigenvectors with the largest eigenvalues. The projected and scaled vectors $\mathbf{x}_p, 2 \leq p \leq P$, are displayed in a scatterplot of the eigenvector component. Each of the P points represents a variable with two components: 1) the correlations between $\mathbf{x}_1$ and the other variables' vectors, 2) the correlations with the other vectors.

The FPCA derives the dietary patterns by the variable of interest in the hypersphere or concentric circles. Circles of smaller radius represent stronger correlations. The center of these circles is the variable of interest. Negative and positive correlations with the variable of interest are differentiated in the graph by use of different colors. Two points close to one another indicate a strong positive correlation between the intakes of the corresponding food groups/nutrients, whereas two diametrically opposed points indicate a strong negative correlation between the intakes of the corresponding food groups/nutrients; two points placed at a similar distance from the origin, parallel to one of the axes, indicate absence of correlation between the intakes of the corresponding food groups/nutrients. Finally, a dashed circle may additional delimit statistical significance at some level [14].

In a study of 1,968 Brazilian adults interviewed with a 26-item FFQ [14], FPCA is applied with three focus variables, age, income, and schooling, to identify the relationship between diet and the variable of interest and the correlation between different foods [42]. These analyses allow to associate socio-economic inequities with dietary patterns and provide reasonable results. For example, whole-wheat foods, fruit, and vegetables are positively correlated with income and schooling, whereas for refined cereals, animal fats (lard), and white bread, the same correlation is negative.

Among major advantages, FPCA provides an immediate graphical representation to answer to a tailored research question. As a major limitation, only one variable of interest at a time can be considered. This opens up the question on how to compare dietary patterns obtained from similar variables of interest in the same study.

12. CROSS-STUDY REPRODUCIBILITY OF DIETARY PATTERNS: NOVEL STATISTICAL APPROACHES

Compared to most a priori dietary patterns – especially indexes of overall diet quality – which can be more easily used across different study populations, the a posteriori approach estimates population-specific dietary patterns. Indeed, if the patterns are derived by explaining the variability among diets of one population, it is unlikely the same patterns would be found in another population. While their reproducibility is necessarily more limited, a posteriori dietary patterns reflect the actual dietary practices in the population under study and provide crucial information [37]. Dietary patterns should reveal those latent characteristics of interest, including socio-demographic and socio-economic factors, ethnic background, religion, and several other environmental factors, like food supply, ability to purchase/prepare foods, advertisements for foods, and the efforts of the government and the nutrition community to foster healthy diets [63], which are at the origin of actual dietary practices. Common latent characteristics may end up in common dietary patterns across studies.

Re-analyses of existing evidence on dietary patterns have been based on the same standardized approach [4, 86, 78, 74] across a few European or US studies. Especially for the a posteriori dietary patterns, this standardized approach of analysis has guaranteed that the 2 to 4 consistently identified dietary patterns do represent common dietary habits across European cohorts representing different countries [4, 86, 78]. On the other hand, differences in the identified dietary patterns have been identified and are more likely to reflect genuine differences in dietary habits than artifacts of statistical analysis [4, 86, 78]. A partial sharing of the a posteriori dietary patterns across countries is therefore supported by existing evidence [4, 86, 78].

In national multi-center studies including groups with different culinary habits, heterogeneous dietary pattern compositions can be estimated based on differences in food group/nutrient intake distributions, likely related to latent population characteristics [64]. This is true in particular for regional and ethnic diversity that can be explored in a multicultural perspective using a posteriori dietary patterns [63].

Although the reproducibility of dietary patterns across populations is becoming crucial for assessing critical aspects of the diet at national and international levels, standard statistical approaches have been used so far [37]. EFA has been initially applied by merging dietary information from different studies in one dataset and forcing the studies to have common dietary factors [39]. Other papers presented separate EFAs by study and provided some ad-hoc standard statistical solutions to improve comparability of food grouping schemes, factor loadings, and factor scores across centers/studies or populations [17, 16, 86].

However, all these attempts have not fully explored the presence and role of population-specific patterns, essential to detect traditional or specific aspects of the diet among subpopulations and their association with disease outcomes, including cancer. Critical in this sense is the lack of a rigorous statistical approach that can simultaneously manage the identification of shared and subpopulation-specific a posteriori dietary patterns, together with an objective selection of the optimal number of shared and subpopulation-specific patterns.

The two novel statistical methods proposed in the next well tackle this issue by extending EFA and CA to discover common and study-specific (or subpopulation-specific) dietary patterns.

The multi-study factor analysis (MSFA) [25] generalizes EFA and has three main goals. First, it combines multiple studies to identify common factors consistent across the studies. Second, it identifies an additional variability component specific to single studies; study-specific latent factors capture that. Third, analyzing study-specific latent factors allows for identifying possible idiosyncratic variations lacking cross-study reproducibility. Indeed, the model allows for a residual component defined for each study and each variable.

The observed variables in study s are decomposed into K factors shared with the other studies and J_s factors reflecting unique sources of variation. Factor loadings relate the observed variables to the latent factors linearly. Let $\mathbf{f}_{is},\ i = 1, \ldots, n_s$ be the common latent factor, and $\mathbf{\Phi}$ with K columns be its loadings; also let $\mathbf{l}_{is},\ i = 1, \ldots, n_s$, be the study-specific latent factor and $\mathbf{\Lambda}_s$ with J_s columns be its loadings. The MSFA assumes that observation of the vector i in the study s, $s = 1, \ldots, S$, $\mathbf{x}_{is}$ is decomposed as:

$$\mathbf{x}_{is} = \mathbf{\Phi}\mathbf{f}_{is} + \mathbf{\Lambda}_s\mathbf{l}_{is} + \mathbf{e}_{is},$$

where $\mathbf{e}_{is}$ is the Gaussian error term, with mean vector equal to zero and variance equal to the diagonal matrix $\mathbf{\Psi}_s = \{\psi_{1s}, \ldots, \psi_{ps}\}$. As a result, the marginal distribution of $\mathbf{x}_{is}$ is multivariate Gaussian with mean vector 0 and covariance matrix:

$$\mathbf{\Sigma}_s = \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s.$$

The estimation is based on the maximum likelihood method, computed via an Expectation Conditional-Maximization (ECM) algorithm [82]. The approach addresses two separate identifiability issues. The first issue deals with the orthogonal rotation indeterminacy, similar to the standard EFA. The MSFA model must be further constrained to avoid orthogonal rotation indeterminacy and obtain an identifiable model. Specifically, if

$$\mathbf{\Phi}^* = \mathbf{\Phi}\mathbf{Q} \quad \text{and } \mathbf{\Lambda}_s^* = \mathbf{\Lambda}_s\mathbf{Q}_s, \quad s = 1, \ldots, S,$$

where $\mathbf{Q}$ and each $\mathbf{Q}_s$ are square orthogonal matrices with K and J_s rows respectively, there could be infinite possible solutions for

$$\mathbf{\Sigma}_s = \mathbf{\Phi}^*\mathbf{\Phi}^{*^\top} + \mathbf{\Lambda}_s^*\mathbf{\Lambda}_s^{*^\top} + \mathbf{\Psi}_s.$$

and thus the covariance matrix $\mathbf{\Sigma}_s$ is not uniquely identified. One popular constraint is to assume a block lower triangular (LT) matrix [76, 15]) for the factor loading matrix. The MSFA adopts this constraint, and $\mathbf{\Phi}$ and all the $\mathbf{\Lambda}_s$s are LT matrices. Similarly to EFA, this solves the orthogonal rotation indeterminacy.

However, in MSFA, a second identifiability issue emerges, since there will be S different equations, and the block LT does not uniquely identify the $S + 1$ solutions $\{\mathbf{\Phi}, \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_s\}$. To address this, further conditions for the MSFA are required. In particular, this paper [25] assumes that the concatenated matrix $\mathbf{\Omega} = \{\mathbf{\Phi}, \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_s\}$ has a full column rank:

$$r(\mathbf{\Omega}) = K + \sum_{i=1}^{s} J_s, \quad \text{with } K + \sum_{i=1}^{s} J_s \leq P.$$

This solves all the identifiability issues that arise in the MSFA.

While the method is focused on genomic applications, it can be applied to other situations where similarities and differences are warranted across multiple data sets. Following this direction, the approach is receiving much attention in the nutritional epidemiological context, as some nutritional epidemiological papers have already adopted it, enhancing the crucial theoretical contributions of the MSFA [24, 37]. A recent paper [24] adopts the MSFA model to derive nutrient-based dietary patterns in the International Head and Neck Cancer Epidemiology (INHANCE) consortium [10] considering 10,668 subjects from 7 different countries (3 from Europe and 4 from the US). The method examines the shared and study-specific patterns in relation to head and neck cancers via multiple logistic regression, taking confounders into account. While this approach does not account for covariates (e.g., energy intake and sex) in dietary pattern identification, it well manages the choice of the number of common and study-specific factors within the statistical model by relying on a combination of information criteria and standard techniques adopted in EFA [24]. It does not provide, however, a way to integrate the selection of the number of factors to be retained within parameter estimation.

The Bayesian MSFA generalizes MSFA in the Bayesian framework [26]. The method explores the multiplicative gamma shrinkage prior, widely used in the Bayesian approach for a single EFA [7], and generalizes it in a multi-study setting. The method develops a fast and efficient Gibbs Sampling, with increasing shrinkage as the column index increases. The prior defined in this approach, called the Sparse Bayesian Infinite factor model, is for the loadings in the $s = 1, \dots, S$ study:

$$\lambda_{pjs} \mid \omega_{pjs}, \tau_{js} \sim N(0, \omega_{pjs}^{-1}\tau_{js}^{-1}),\ p = 1, \dots, P,\ j = 1, \dots, \infty,$$

$$\omega_{pjs} \sim \Gamma\left(\frac{\nu^s}{2}, \frac{\nu^s}{2}\right), \quad \tau_{js} = \prod_{l=1}^{j} \delta_l^s,$$

$$\delta_1^s \sim \Gamma(a_1^s, 1),\ \delta_l^s \sim \Gamma(a_2^s, 1),\ \geq 2,$$

where δ_l^s $(l = 1, 2, \dots)$ are independent, τ_{js} is the global shrinkage parameter for column j, and ω_{pjs} is the local shrinkage for the element p in column j for the study $s = 1, \dots, S$. The idea is rather suitable for high dimensions, where, if more factors are added, it is crucial to consider an increment of the shrinkage.

The Bayesian MSFA is adopted in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) [110], a US multi-site community-based cohort, focusing on health and risk factors of cardiovascular and pulmonary outcomes of Hispanic/Latino adults [27]. The method captures common and subpopulation-specific dietary patterns on 42 nutrients across available combinations of 4 US field sites (Bronx, Chicago, Miami, and San Diego) and 6 Hispanic/Latino ethnic backgrounds (Cuban, Dominican Republic, Mexican, Puerto Rican, Central and South American). This analysis identified four common patterns: "Plant-based foods", "Processed foods", "Dairy products", and "Seafood". Twelve additional study-specific patterns, one for each ethnic background – site category, represent variants of foods from animal sources. These variants can be grouped based on visual inspection of the factor-loading matrices and congruence coefficients between factor loadings into 3 overarching (i.e., more similar) dietary patterns.

A direct comparison with frequentist MSFA and standard EFA (with the principal component method) on the overall sample is also carried out and highlights the Bayesian approach's merits. The introduction of prior distributions (which act like rotations) shows that:

- Bayesian MSFA-based shared dietary patterns are equivalent to their counterparts from standard EFA and frequentist MSFA;
- ethnic background site-specific dietary patterns from Bayesian MSFA are better characterized than those from frequentist MSFA, which oppose vegetable and animal sources of foods in most ethnic background – site categories.

This is the first attempt to use a Bayesian multi-study framework in nutritional epidemiology, which suggests that relevant dietary patterns are shared across different ethnic backgrounds and recruitment sites; additional patterns exist also that are specific to a single category.

The Robust Profile Clustering (RPC) [113] is a generalization of mixture models to handle diverse populations or studies. The RPC identifies robust "global" clusters for each individual $i = 1, \dots, n_s$ and variable $p = 1, \dots, P$ across all subpopulations s_i and locally within a subject's respective subpopulation.

The method performs for the food items, and thus the vector variables $\mathbf{x}_1, \dots, \mathbf{x}_p$ are drawn from a multinomial distribution. The RPC approach focuses on probability models with three different elements: 1) the global clustering C_i, 2) the deviation indicator G_{ip}, and 3) the local clustering membership, L_{ip} with the corresponding probabilities equal to

$$Pr(C_i = h) = \pi_h,$$

$$Pr(G_{ip} = 1 | s_i = s) = \upsilon_p^{(s)}$$

$$Pr(L_{ip} = l | s_i = s) = \lambda_l^{(s)}.$$

The deviation indicator is equal to one, $G_{ip} = 1$, if the variable, in this case, the food item, p is identified in the global cluster C_i for subject i, $G_{ip} = 0$ otherwise. They also adopted a Bayesian approach for parameter estimation with a Gibbs sampler. Then, the model for the subpopulation parameter is a beta-Bernoulli process:

$$G_{ip} \sim Bern(\upsilon_p^{(s)}),\ \upsilon_p^{(s)} \sim Bern(1, \beta^{(s)}),\ \ \beta^{(s)} \sim \Gamma(a, b),$$

with the hyperparameters (a, b) used for varying the overall weight of each local component (deviated food item) of its corresponding subpopulation.

For the global clustering process, the RPC adopts an overfitted finite mixture model [125], with an upper bound equal to 50 for the number of clusters K:

$$Pr(C_i = h) = \pi_h,$$

$$\pi. = (\pi_1, \dots, \pi_K)^\top \sim Dir\left(\frac{1}{K}, \dots, \frac{1}{K}\right).$$

where Dir indicates a Dirichlet distribution. The RPC replicates this scheme to specify the model for the local clustering:

$$Pr(L_{ip} = l | s_i = s) = \lambda_l^{(s)},$$

$$\lambda^{(s)}_{\cdot} = (\lambda_1^{(s)}, \ldots, \lambda_K^{(s)})^\top \sim Dir\left(\frac{1}{K}, \ldots, \frac{1}{K}\right).$$

The method is applied to the National Birth Defects Prevention Study [133], a case-control study of birth defects in the United States focusing on 9,010 control live-born infants without any birth defects and a total of $p = 63$ food items with four consumption levels ($d = 4$). The approach estimates 7 global cluster patterns: “meats and fatty foods”, “fast foods”, “chicken, cheese, and beef”, “Tex-Mex” or Latino style diet, “snack-style foods”, “caffeine products”, “wheat bread, fruit cocktail and low-fat milk”.

As the BMSFA, the RPC is also used in the HCHS/SOL [110] to estimate differences in dietary consumption for study sites and Hispanic/Latino ethnic backgrounds [114].

The RPC method estimates both shared and specific consumption behaviors derived from 132 food groups. Specifically, the method estimates 48 shared consumption behaviors of foods and beverages across all the subpopulations, with some differences in subpopulations for these same foods. Several foods were common within the study site cluster (e.g., chicken, orange juice, milk) and ethnic background (e.g., papayas, plantain, coffee). Different from the BMSFA, which provides an accurate estimate of the factor loadings and scores allowing a direct association with the disease, the RPC is a Bayesian nonparametric approach estimating food item response differences within a region/subpopulation. The RPC allows the number of clusters not to exceed a certain number for better interpretability.

However, this clustering method estimates components likely to describe differences in global behaviors compared to those relevant to specific subpopulations. It becomes crucial to associate these derived clusters with response variables, including socio-demographic or socio-economic factors, or a disease outcome. Specifically for the RPC method, different subjects from two subpopulations could be in the same global cluster and differ in behavioral patterns and/or socio-economic and socio-demographic factors.

A recent method developed the RPC in a supervised setting [115], namely the Supervised RPC. The method is a two-step process: first, it develops the global and local cluster, then it builds the likelihood by associating the cluster to a binary response variable y_i of individual i, adopting a probit regression model.

The global cluster dietary pattern is adjusted for the covariates ξ in the model, i.e., potential confounders, and the subject-specific vector of observed demographic information is represented by W_i. Then the probit model can be written as:

$$\begin{aligned} Pr(y_i\!=\!1|\xi, W_i) &= \Phi(W_i\xi) \\ &= \Phi\left(\sum_{i=1}^{K_0} \mathbb{1}(C_i = h)\xi_h + W_{dem}\xi_{dem}\right), \end{aligned}$$

where K_0 is the number of global clusters.

They jointly derived a one-estimation-step process for handling the two models together. The model is applied to the National Birth Defects Prevention Study [133], finding the same seven global clusters mentioned before and associating those with orofacial clefts among offspring. The method is one attempt to associate clusters with disease outcomes in multi-study settings.

13. CONCLUSIONS

Diet is a complex exposure, which calls for multiple approaches to examine its relationship with non-communicable diseases, including cancer. Evidence on the effect of diet is enhanced when results from multiple study types (i.e., observational studies, randomized trials of intermediate responses, in vitro and in vivo studies) and from multiple forms of dietary exposure (i.e., food items/groups, nutrients, biomarkers, and dietary patterns) are consistent.

Being complementary to the traditional single-component analysis, the dietary pattern approach has been proposed in nutritional epidemiology to exploit the collinearity of nutrients and foods. This approach is not effective if the effect is “caused” by a specific nutrient, because the effect of the nutrient would be diluted. It may be useful when traditional single-component analyses have identified few dietary associations with the disease (e.g., breast cancer). On the other hand, when many dietary associations have been demonstrated for the disease (e.g., coronary heart disease or colorectal cancer), dietary pattern analysis may also prove to be useful because it allows to examine the effects of this overall, but likely well-structured, dietary exposure. In addition, a dietary pattern can be used as a covariate when examining a specific nutrient/food group, to determine whether its effect is independent of the overall dietary pattern. Furthermore, dietary pattern analysis can be useful in evaluating dietary guidelines [56].

As methods to assess dietary patterns have been refined and the evidence base has been strengthened, the advantages that dietary patterns offer as an approach for informing public health recommendations have increasingly been recognized [18].

In the future, like all of nutritional epidemiology, patterns research will be advanced by using methods of dietary capture and analysis that better estimate usual intakes and by considering how they may change over time. In addition, continued clarification of the most useful treatment of input variables for EFA and CA, continued development of best practices for standardizing statistical procedures in a posteriori dietary patterns, refinement of indices, and progress in methods to correct for measurement error would advance the field [70].

Novel statistical methods have been reviewed in this paper, especially those aimed at evaluating cross-study reproducibility of a posteriori or mixed-type dietary patterns.

Among relevant features common to them, it is worth to mention sparsity modeling and related adjustment for additional covariates influencing individuals dietary practices, such as age, sex, socio-cultural factors, and energy intake [65].

In parallel with standardization of statistical procedures for dietary patterns identification, these contributions will further enhance our understanding of the dietary habits – cancer risk association as far as content knowledge of nutritionists and the broad know-how of nutritional epidemiologists meet with statisticians willing to tailor methods to the specific needs of dietary pattern analysis.

ACKNOWLEDGEMENTS

This research is supported by the U.S.A.'s National Institutes of Health, grant NIGMS/NIH COBRE CBHD P20GM109035 (RDV).

Accepted 28 April 2023

REFERENCES

[1] Ascherio, A., Stampfer, M. J., Colditz, G. A., Rimm, E. B., Litin, L. and Willett, W. C. (1992). Correlations of vitamin A and E intakes with the plasma concentrations of carotenoids and tocopherols among American men and women. *J Nutr* **122**(9) 1792–1801.

[2] Assi, N., Moskal, A., Slimani, N., Viallon, V., Chajes, V., Freisling, H., Monni, S., Knueppel, S., Förster, J., Weiderpass, E. and others. (2016). A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Public Health Nutr* **19**(2) 242–254.

[3] Avalos-Pacheco, A., Rossell, D. and Savage, R. S. (2022). Heterogeneous large datasets integration using Bayesian factor regression. *Bayesian Anal* **17**(1) 33–66. https://doi.org/10.1214/20-ba1240. MR4377136

[4] Balder, H. F., Virtanen, M., Brants, H. A., Krogh, V., Dixon, L. B., Tan, F., Mannisto, S., Bellocco, R., Pietinen, P., Wolk, A. and others. (2003). Common and country-specific dietary patterns in four European cohort studies. *J Nutr* **133**(12) 4246–4251.

[5] Bédard, A., Garcia-Aymerich, J., Sanchez, M., Le Moual, N., Clavel-Chapelon, F., Boutron-Ruault, M. -C., Maccario, J. and Varraso, R. (2015). Confirmatory factor analysis compared with principal component analysis to derive dietary patterns: a longitudinal study in adult women. *J Nutr* **145**(7) 1559–1568.

[6] Bertuccio, P., Rosato, V., Andreano, A., Ferraroni, M., Decarli, A., Edefonti, V. and La Vecchia, C. (2013). Dietary patterns and gastric cancer risk: a systematic review and meta-analysis. *Ann Oncol* **24**(6) 1450–1458.

[7] Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98**(2) 291–306. https://doi.org/10.1093/biomet/asr013. MR2806429

[8] Biesbroek, S., van der A, D. L., Brosens, M. C., Beulens, J. W., Verschuren, W. M., van der Schouw, Y. T. and Boer, J. M. (2015). Identifying cardiovascular risk factor-related dietary patterns with reduced rank regression and random forest in the EPIC-NL cohort. *Am J Clin Nutr* **102**(1) 146–154.

[9] Blashfield, R. K. and Aldenderfer, M. S. (1978). The literature on cluster analysis. *Multivariate Behav Res* **13**(3) 271–295.

[10] Bravi, F., Lee, Y. q. C. A., Hashibe, M., Boffetta, P., Conway, D. I., Ferraroni, M., La Vecchia, C., Edefonti, V. and INHANCE Consortium investigators (2021). Lessons learned from the INHANCE consortium: An overview of recent results on head and neck cancer. *Oral Dis* **27**(1) 73–93.

[11] Breiman, L. (2017) *Classification and regression trees*. Routledge.

[12] Cade, J., Thompson, R., Burley, V. and Warm, D. (2002). Development, validation and utilisation of food-frequency questionnaires – a review. *Public Health Nutr* **5**(4) 567–587.

[13] Camp, N. J. and Slattery, M. L. (2002). Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Cause Control* **13**(9) 813–823.

[14] Canuto, R., Camey, S., Gigante, D. P., Menezes, A. and Olinto, M. T. A. (2010). Focused principal component analysis: a graphical method for exploring dietary patterns. *Cadernos de Saúde Pública* **26** 2149–2156.

[15] Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *J Am Stat Assoc* **103**(484) 1438–1456. https://doi.org/10.1198/016214508000000869. MR2655722

[16] Castelló, A., Buijsse, B., Martín, M., Ruiz, A., Casas, A. M., Baena-Cañada, J. M., Pastor-Barriuso, R., Antolín, S., Ramos, M., Muñoz, M. and others. (2016). Evaluating the applicability of data-driven dietary patterns to independent samples with a focus on measurement tools for pattern similarity. *J Acad Nutr Diet* **116**(12) 1914–1924.

[17] Castelló, A., Lope, V., Vioque, J., Santamariña, C., Pedraz-Pingarrón, C., Abad, S., Ederra, M., Salas-Trejo, D., Vidal, C., Sánchez-Contador, C. and others. (2016). Reproducibility of data-driven dietary patterns in two groups of adult Spanish women from different studies. *Brit J Nutr* **116**(4) 734–742.

[18] Cespedes, E. M. and Hu, F. B. (2015). Dietary patterns: from nutritional epidemiologic analysis to national guidelines. *Am J Clin Nutr* **101**(5) 899–900.

[19] Clinton, S. K., Giovannucci, E. L. and Hursting, S. D. (2020). The World Cancer Research Fund/American Institute for Cancer Research third expert report on diet, nutrition, physical activity, and cancer: impact and future directions. *J Nutr* **150**(4) 663–671.

[20] Dalmartello, M., Decarli, A., Ferraroni, M., Bravi, F., Serraino, D., Garavello, W., Negri, E., Vermunt, J. and La Vecchia, C. (2020). Dietary patterns and oral and pharyngeal cancer using latent class analysis. *Int J Cancer* **147**(3) 719–727.

[21] Dalmartello, M., Vermunt, J., Serraino, D., Garavello, W., Negri, E., Levi, F. and La Vecchia, C. (2021). Dietary patterns and oesophageal cancer: a multi-country latent class analysis. *J Epidemiol Community Health* **75** 567–573.

[22] De Keyzer, W., Huybrechts, I., De Vriendt, V., Vandevijvere, S., Slimani, N., Van Oyen, H. and De Henauw, S. (2011). Repeated 24-hour recalls versus dietary records for estimating nutrient intakes in a national food consumption survey. *Food Nutr Res* **55**.

[23] De Stefani, E., Ronco, A. L., Boffetta, P., Deneo-Pellegrini, H., Correa, P., Acosta, G. and Mendilaharsu, M. (2012). Nutrient-derived dietary patterns and risk of colorectal cancer: a factor analysis in Uruguay. *Asian Pac J Cancer Prev* **13**(1) 231–235.

[24] De Vito, R., Lee, Y. C. A., Parpinel, M., Serraino, D., Olshan, A. F., Zevallos, J. P., Levi, F., Zhang, Z. F., Morgenstern, H., Garavello, W. and others. (2019). Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium. *Epidemiology* **30**(1) 93.

[25] De Vito, R., Bellio, R., Trippa, L. and Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics* **75**(1) 337–346. https://

doi.org/10.1111/biom.12974. MR3953734

[26] De Vito, R., Bellio, R., Trippa, L. and Parmigiani, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *Ann Appl Stat* **15**(4) 1723–1741. https://doi.org/10.1214/21-aoas1456. MR4355073

[27] De Vito, R., Stephenson, B., Sotres-Alvarez, D., Siega-Riz, A. M., Mattei, J., Parpinel, M., Peters, B. A., Bainter, S. A., Daviglus, M. L., Van Horn, L. and others. (2022). Shared and ethnic background site-specific dietary patterns in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *medRxiv*.

[28] Devlin, U. M., McNulty, B. A., Nugent, A. P. and Gibney, M. J. (2012). The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting. *Proc Nutr Soc* **71**(4) 599–609.

[29] DiBello, J. R., Kraft, P., McGarvey, S. T., Goldberg, R., Campos, H. and Baylin, A. (2008). Comparison of 3 methods for identifying dietary patterns associated with risk of disease. *Am J Epidemiol* **168**(12) 1433–1443.

[30] Dietary Guidelines Advisory Committee (2015) *Scientific Report of the 2015 Dietary Guidelines Advisory Committee: Advisory Report to the Secretary of Health and Human Services and the Secretary of Agriculture*. U.S. Department of Agriculture, Agricultural Research Service, Washington, DC.

[31] Easton, J. F., Román Sicilia, H. and Stephens, C. R. (2019). Classification of diagnostic subcategories for obesity and diabetes based on eating patterns. *Nutr Diet* **76**(1) 104–109.

[32] Edefonti, V., Randi, G., Decarli, A., La Vecchia, C., Bosetti, C., Franceschi, S., Dal Maso, L. and Ferraroni, M. (2009). Clustering dietary habits and the risk of breast and ovarian cancers. *Ann Oncol* **20**(3) 581–590.

[33] Edefonti, V., Randi, G., La Vecchia, C., Ferraroni, M. and Decarli, A. (2009). Dietary patterns and breast cancer: a review with focus on methodological issues. *Nutr Rev* **67**(6) 297–314.

[34] Edefonti, V., Hashibe, M., Parpinel, M., Turati, F., Serraino, D., Matsuo, K., Olshan, A. F., Zevallos, J. P., Winn, D. M., Moysich, K. et al. (2015). Natural vitamin C intake and the risk of head and neck cancer: A pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Int J Cancer* **137**(2) 448–462.

[35] Edefonti, V., Nicolussi, F., Polesel, J., Bravi, F., Bosetti, C., Garavello, W., La Vecchia, C., Bidoli, E., Decarli, A., Serraino, D. and others. (2015). Nutrient-based dietary patterns and nasopharyngeal cancer: evidence from an exploratory factor analysis. *Br J Cancer* **112**(3) 446–454.

[36] Edefonti, V., De Vito, R., Dalmartello, M., Patel, L., Salvatori, A. and Ferraroni, M. (2020). Reproducibility and Validity of A Posteriori Dietary Patterns: A Systematic Review. *Adv Nutr* **11**(2) 293–326.

[37] Edefonti, V., De Vito, R., Salvatori, A., Bravi, F., Patel, L., Dalmartello, M. and Ferraroni, M. (2020). Reproducibility of A Posteriori Dietary Patterns across Time and Studies: A Scoping Review. *Adv Nutr* **11**(5) 1255–1281.

[38] Edefonti, V., Di Maso, M., Tomaino, L., Parpinel, M., Garavello, W., Serraino, D., Ferraroni, M., Crispo, A., La Vecchia, C. and Bravi, F. (2022). Diet quality as measured by the Healthy Eating Index 2015 and oral and pharyngeal cancer risk. *J Acad Nutr Diet* **122**(9) 1677–1687.

[39] Edefonti, V., Hashibe, M., Ambrogi, F., Parpinel, M., Bravi, F., Talamini, R., Levi, F., Yu, G., Morgenstern, H., Kelsey, K. and others. (2012). Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium. *Ann Oncol* **23**(7) 1869–1880.

[40] Engelhardt, B. E. and Stephens, M. (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* **6**(9) 1001117.

[41] Fahey, M. T., Ferrari, P., Slimani, N., Vermunt, J. K., White, I. R., Hoffmann, K., Wirfält, E., Bamia, C., Touvier, M., Linseisen, J. et al. (2012). Identifying dietary patterns using a normal mixture model: application to the EPIC study. *J Epidemiol Community Health* **66**(1) 89–94.

[42] Falissard, B., Corruble, E., Mallet, L. and Hardy, P. (2001). Focused principal component analysis: a promising approach for confirming findings of exploratory analysis? *Int J Meth Psych Res* **10**(4) 191–195.

[43] Falissard, B. (1999). Focused principal component analysis: looking at a correlation matrix with a particular interest in a given variable. *J Comput Graph Stat* **8**(4) 906–912.

[44] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Ass* **97**(458) 611–631. https://doi.org/10.1198/016214502760047131. MR1951635

[45] Garcia-Larsen, V., Morton, V., Norat, T., Moreira, A., Potts, J. F., Reeves, T. and Bakolis, I. (2019). Dietary patterns derived from principal component analysis (PCA) and risk of colorectal cancer: a systematic review and meta-analysis. *Eur J Clin Nutr* **73**(3) 366–386.

[46] Gianfredi, V., Ferrara, P., Dinu, M., Nardi, M. and Nucci, D. (2022). Diets, dietary patterns, single foods and pancreatic cancer risk: An umbrella review of meta-analyses. *Int J Environ Res Public Health* **19**(22) 14797.

[47] Gleason, P. M., Boushey, C. J., Harris, J. E. and Zoellner, J. (2015). Publishing nutrition research: a review of multivariate techniques part 3: data reduction methods. *J Acad Nutr Diet* **115**(7) 1072–1082.

[48] Gorst-Rasmussen, A., Dahm, C. C., Dethlefsen, C., Scheike, T. and Overvad, K. (2011). Exploring dietary patterns by using the treelet transform. *Am J Epidemiol* **173**(10) 1097–1104.

[49] Greve, B., Pigeot, I., Huybrechts, I., Pala, V. and Börnhorst, C. (2016). A comparison of heuristic and model-based clustering methods for dietary pattern analysis. *Public Health Nutr* **19**(2) 255–264.

[50] Harrington, D. (2009) *Confirmatory factor analysis*. Oxford University Press, Oxford, UK.

[51] Hearty, A. P. and Gibney, M. J. (2008). Analysis of meal patterns with the use of supervised data mining techniques–artificial neural networks and decision trees. *Am J Clin Nutr* **88**(6) 1632–1642.

[52] Hébert, J. R., Hurley, T. G., Steck, S. E., Miller, D. R., Tabung, F. K., Peterson, K. E., Kushi, L. H. and Frongillo, E. A. (2014). Considering the value of dietary assessment data in informing nutrition-related health policy. *Adv Nutr* **5**(4) 447–455.

[53] Hernan, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *J Epidemiol Community Health* **60**(7) 578–586.

[54] Hill, A. B. (1965). The environment and disease: association or causation? *Proc R Soc Med* **58**(5) 295–300.

[55] Hoffmann, K., Schulze, M. B., Schienkiewitz, A., Nöthlings, U. and Boeing, H. (2004). Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol* **159**(10) 935–944.

[56] Hu, F. B. (2002). Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* **13**(1) 3–9.

[57] Imamura, F. and Jacques, P. F. (2011). Invited commentary: dietary pattern analysis. *Am J Epidemiol* **173**(10) 1105–1108.

[58] Jacobs, D. R. (2014). What comes first: the food or the nutrient? Executive summary of a symposium. *J Nutr* **144**(4 Suppl) 543–546.

[59] Jacobs, D. R. and Steffen, L. M. (2003). Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am J Clin Nutr* **78**(3 Suppl) 508–513.

[60] Jacobs, D. R. and Tapsell, L. C. (2007). Food, not nutrients, is the fundamental unit in nutrition. *Nutr Rev* **65**(10) 439–450.

[61] Jacobs, D. R. and Tapsell, L. C. (2013). Food synergy: the

key to a healthy diet. *Proc Nutr Soc* **72**(2) 200–206.

[62] Jacobs, D. R., Gross, M. D. and Tapsell, L. C. (2009). Food synergy: an operational concept for understanding nutrition. *Am J Clin Nutr* **89**(5) 1543–1548.

[63] Jacques, P. F. and Tucker, K. L. (2001). Are dietary patterns useful for understanding the role of diet in chronic disease? *Am J Clin Nutr* **73**(1) 1–2.

[64] Jannasch, F., Kröger, J. and Schulze, M. B. (2017). Dietary patterns and type 2 diabetes: a systematic literature review and meta-analysis of prospective studies. *J Nutr* **147**(6) 1174–1182.

[65] Joo, J., Williamson, S. A., Vazquez, A. I., Fernandez, J. R. and Bray, M. S. (2018). Advanced dietary patterns analysis using sparse latent factor models in young adults. *J Nutr* **148**(12) 1984–1992.

[66] Judd, S. E., Letter, A. J., Shikany, J. M., Roth, D. L. and Newby, P. K. (2015). Dietary patterns derived using exploratory and confirmatory factor analysis are stable and generalizable across race, region, and gender subgroups in the REGARDS study. *Front Nutr* **1** 29.

[67] Kant, A. K. (1996). Indexes of overall diet quality: a review. *J Am Diet Assoc* **96**(8) 785–791.

[68] Kant, A. K. (2004). Dietary patterns and health outcomes. *J Am Diet Assoc* **104**(4) 615–635.

[69] Kipnis, V., Freedman, L. S., Brown, C. C., Hartman, A., Schatzkin, A. and Wacholder, S. (1993). Interpretation of energy adjustment models for nutritional epidemiology. *Am J Epidemiol* **137**(12) 1376–1380.

[70] Krebs-Smith, S. M., Subar, A. F. and Reedy, J. (2015). Examining dietary patterns in relation to chronic disease: matching measures and methods to questions of interest. *Circulation* **132**(9) 790–793.

[71] Lazarou, C., Karaolis, M., Matalas, A. L. and Panagiotakos, D. B. (2012). Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. *Comput Methods Programs Biomed* **108**(2) 706–714.

[72] Lee, A. B., Nadler, B. and Wasserman, L. (2008). Treelets – An adaptive multi-scale basis for sparse unordered data. *Ann Appl Stat* **2**(2) 435–471. https://doi.org/10.1214/07-AOAS137. MR2524336

[73] Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D. and Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med* **26**(3) 172–181.

[74] Liese, A. D., Krebs-Smith, S. M., Subar, A. F., George, S. M., Harmon, B. E., Neuhouser, M. L., Boushey, C. J., Schap, T. E. and Reedy, J. (2015). The Dietary Patterns Methods Project: synthesis of findings across cohorts and relevance to dietary guidance. *J Nutr* **145**(3) 393–402.

[75] Lo Siou, G., Yasui, Y., Csizmadi, I., McGregor, S. E. and Robson, P. J. (2011). Exploring statistical approaches to diminish subjectivity of cluster analysis to derive dietary patterns: The Tomorrow Project. *Am J Epidemiol* **173**(8) 956–967.

[76] Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Stat Sinica* **14** 41–67. MR2036762

[77] Männistö, S., Dixon, L. B., Balder, H. F., Virtanen, M. J., Krogh, V., Khani, B. R., Berrino, F., van den Brandt, P. A., Hartman, A. M., Pietinen, P. and others. (2005). Dietary patterns and breast cancer risk: results from three cohort studies in the DIETSCAN project. *Cancer Causes Control* **16**(6) 725–733.

[78] Männistö, S., Harald, K., Kontto, J., Lahti-Koski, M., Kaartinen, N. E., Saarni, S. E., Kanerva, N. and Jousilahti, P. (2014). Dietary and lifestyle characteristics associated with normal-weight obesity: the National FINRISK 2007 Study. *Brit J Nutr* **111**(5) 887–894.

[79] Martinez, M. E., Marshall, J. R. and Sechrest, L. (1998). Invited commentary: Factor analysis and the search for objectivity. *Am J Epidemiol* **148**(1) 17–19.

[80] McEligot, A. J., Poynor, V., Sharma, R. and Panangadan, A. (2020). Logistic LASSO regression for dietary intakes and breast cancer. *Nutrients* **12**(9) 2652.

[81] Melaku, Y. A., Gill, T. K., Taylor, A. W., Adams, R. and Shi, Z. (2018). A comparison of principal component analysis, partial least-squares and reduced-rank regressions in the identification of dietary patterns associated with bone mass in ageing Australians. *Eur J Nutr* **57**(5) 1969–1983.

[82] Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**(2) 267–278. https://doi.org/10.1093/biomet/80.2.267. MR1243503

[83] Michels, K. B. and Schulze, M. B. (2005). Can dietary patterns help us detect diet–disease associations? *Nutr Res Rev* **18**(2) 241–248.

[84] Milà-Villarroel, R., Bach-Faig, A., Puig, J., Puchal, A., Farran, A., Serra-Majem, L. and Carrasco, J. L. (2011). Comparison and evaluation of the reliability of indexes of adherence to the Mediterranean diet. *Public Health Nutr* **14**(12A) 2338–2345.

[85] Moeller, S. M., Reedy, J., Millen, A. E., Dixon, L. B., Newby, P., Tucker, K. L., Krebs-Smith, S. M. and Guenther, P. M. (2007). Dietary patterns: challenges and opportunities in dietary patterns research. *J Am Diet Assoc* **107**(7) 1233–1239.

[86] Moskal, A., Pisa, P. T., Ferrari, P., Byrnes, G., Freisling, H., Boutron-Ruault, M., Cadeau, C., Nailler, L., Wendt, A., Kühn, T. and others. (2014). Nutrient patterns and their food sources in an International Study Setting: report from the EPIC study. *PLoS One* **9**(6) 98647.

[87] Mozaffarian, D., Katan, M. B., Ascherio, A., Stampfer, M. J. and Willett, W. C. (2006). Trans fatty acids and cardiovascular disease. *N Engl J Med* **354**(15) 1601–1613.

[88] National Institutes of Health, National Cancer Institute (2023). *Dietary Assessment Primer*. https://dietassessmentprimer.cancer.gov/ Accessed 2023-02-09.

[89] Newby, P. K. and Tucker, K. L. (2004). Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr. Rev.* **62**(5) 177–203.

[90] Newby, P. K., Weismayer, C., Akesson, A., Tucker, K. L. and Wolk, A. (2006). Long-term stability of food patterns identified by use of factor analysis among Swedish women. *J Nutr* **136**(3) 626–633.

[91] Ocké, M. C. and Kaaks, R. J. (1997). Biochemical markers as additional measurements in dietary validity studies: application of the method of triads with examples from the European Prospective Investigation into Cancer and Nutrition. *Am J Clin Nutr* **65**(4 Suppl) 1240–1245.

[92] Ocké, M. C. (2013). Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis. *P Nutr Soc* **72**(2) 191–199.

[93] Oluwagbemigun, K., Foerster, J., Watkins, C., Fouhy, F., Stanton, C., Bergmann, M. M., Boeing, H. and Nöthlings, U. (2020). Dietary patterns are associated with serum metabolite patterns and their association is influenced by gut bacteria among older German adults. *J Nutr* **150**(1) 149–158.

[94] Padmadas, S. S., Dias, J. G. and Willekens, F. J. (2006). Disentangling women's responses on complex dietary intake patterns from an Indian cross-sectional survey: a latent class analysis. *Public Health Nutr* **9**(2) 204–211.

[95] Park, T. and Casella, G. (2008). The Bayesian lasso. *J Am Stat Assoc* **103**(482) 681–686. https://doi.org/10.1198/016214508000000337. MR2524001

[96] Patterson, B. H., Dayton, C. M. and Graubard, B. I. (2002). Latent class analysis of complex sample survey data: application to dietary data. *J Am Stat Ass* **97**(459) 721–741. https://doi.org/10.1198/016214502388618465. MR1941406

[97] Preis, S. R., Spiegelman, D., Zhao, B. B., Moshfegh, A., Baer, D. J. and Willett, W. C. (2011). Application of a repeat-measure biomarker measurement error model to 2 valida-

tion studies: examination of the effect of within-person variation in biomarker measurements. *Am J Epidemiol* **173**(6) 683–694.

[98] Rhee, J. J., Cho, E. and Willett, W. C. (2014). Energy adjustment of nutrient intakes is preferable to adjustment using body weight and physical activity in epidemiological analyses. *Public Health Nutr* **17**(5) 1054–1060.

[99] Rita Gaio, A., Costa, J. P., Santos, A. C., Ramos, E. and Lopes, C. (2012). A restricted mixture model for dietary pattern analysis in small samples. *Stat Med* **31**(19) 2137–2150. https://doi.org/10.1002/sim.5336. MR2956067

[100] Ročková, V. and George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J Am Stat Assoc* **111**(516) 1608–1622. https://doi.org/10.1080/01621459.2015.1100620. MR3601721

[101] Romaguera, D., Vergnaud, A. C., Peeters, P. H., van Gils, C. H., Chan, D. S., Ferrari, P., Romieu, I., Jenab, M., Slimani, N., Clavel-Chapelon, F. et al. (2012). Is concordance with World Cancer Research Fund/American Institute for Cancer Research guidelines for cancer prevention related to subsequent risk of cancer? Results from the EPIC study. *Am J Clin Nutr* **96**(1) 150–163.

[102] Rosner, B., Spiegelman, D. and Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol* **132**(4) 734–745.

[103] Ryman, T. K., Boyer, B. B., Hopkins, S., Philip, J., O'brien, D., Thummel, K. and Austin, M. A. (2015). Characterising the reproducibility and reliability of dietary patterns among Yup'ik Alaska Native people. *Brit J Nutr* **113**(4) 634–643.

[104] Satija, A., Yu, E., Willett, W. C. and Hu, F. B. (2015). Understanding nutritional epidemiology and its role in policy. *Adv Nutr* **6**(1) 5–18. MR3337656

[105] Schatzkin, A., Kipnis, V., Carroll, R. J., Midthune, D., Subar, A. F., Bingham, S., Schoeller, D. A., Troiano, R. P. and Freedman, L. S. (2003). A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. *Int J Epidemiol* **32**(6) 1054–1062.

[106] Schoenaker, D. A., Dobson, A. J., Soedamah-Muthu, S. S. and Mishra, G. D. (2013). Factor analysis is more appropriate to identify overall dietary patterns associated with diabetes when compared with Treelet transform analysis. *J Nutr* **143**(3) 392–398.

[107] Schulz, C. A., Oluwagbemigun, K. and Nöthlings, U. (2021). Advances in dietary pattern analysis in nutritional epidemiology. *Eur J Nutr* **60**(8) 4115–4130.

[108] Schulze, M. B., Martínez-González, M. A., Fung, T. T., Lichtenstein, A. H. and Forouhi, N. G. (2018). Food based dietary patterns and chronic disease prevention. *BMJ* **361** 2396.

[109] Slattery, M. L., Boucher, K. M., Caan, B. J., Potter, J. D. and Ma, K. N. (1998). Eating patterns and risk of colon cancer. *Am J Epidemiol* **148**(1) 4–16.

[110] Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L., Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M. et al. (2010). Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol* **20**(8) 629–641.

[111] Sotres-Alvarez, D., Herring, A. H. and Siega-Riz, A. M. (2010). Latent class analysis is useful to classify pregnant women into dietary patterns. *J Nutr* **140**(12) 2253–2259.

[112] Spiegelman, D., Zhao, B. and Kim, J. (2005). Correlated errors in biased surrogates: study designs and methods for measurement error correction. *Stat Med* **24**(11) 1657–1682. https://doi.org/10.1002/sim.2055. MR2137643

[113] Stephenson, B. J., Herring, A. H. and Olshan, A. (2020). Robust clustering with subpopulation-specific deviations. *J Am Stat Assoc* **115**(530) 521–537. https://doi.org/10.1080/01621459.2019.1611583. MR4107655

[114] Stephenson, B. J., Sotres-Alvarez, D., Siega-Riz, A. M., Mossavar-Rahmani, Y., Daviglus, M. L., Van Horn, L., Herring, A. H. and Cai, J. (2020). Empirically derived dietary patterns using robust profile clustering in the Hispanic Community Health Study/Study of Latinos. *J Nutr* **150**(10) 2825–2834.

[115] Stephenson, B. J., Herring, A. H., Olshan, A. F. and others. (2022). Derivation of maternal dietary patterns accounting for regional heterogeneity. *J R Stat Soc C: Appl Stat* **71**(5) 1957–1977. https://doi.org/10.1111/rssc.12604. MR4511136

[116] Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A. and Rosenfeld, S. (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. *Am J Epidemiol* **154**(12) 1089–1099.

[117] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**(3) 209–249.

[118] Tapsell, L. C., Neale, E. P., Satija, A. and Hu, F. B. (2016). Foods, nutrients, and dietary patterns: Interconnections and implications for dietary guidelines. *Adv Nutr* **7**(3) 445–454.

[119] Teng, J. H., Lin, K. C. and Ho, B. S. (2007). Application of classification tree and logistic regression for the management and health intervention plans in a community-based study. *J Eval Clin Pract* **13**(5) 741–748.

[120] Thompson, F. E., Kirkpatrick, S. I., Subar, A. F., Reedy, J., Schap, T. E., Wilson, M. M. and Krebs-Smith, S. M. (2015). The National Cancer Institute's dietary assessment primer: a resource for diet research. *J Acad Nutr Diet* **115**(12) 1986–1995.

[121] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc B* **58**(1) 267–288. MR1379242

[122] Trichopoulou, A., Kouris-Blazos, A., Vassilakou, T., Gnardellis, C., Polychronopoulos, E., Venizelos, M., Lagiou, P., Wahlqvist, M. L. and Trichopoulos, D. (1995). Diet and survival of elderly Greeks: a link to the past. *Am J Clin Nutr* **61**(6 Suppl) 1346–1350.

[123] Tuglus, C. and van der Laan, M. J. (2008). Discussion of: Treelets – An adaptive multi-scale basis for sparse unordered data. *Ann Appl Stat* **2**(2) 489. https://doi.org/10.1214/08-AOAS137F. MR2524342

[124] Turati, F., Edefonti, V., Bravi, F., Ferraroni, M., Talamini, R., Giacosa, A., Montella, M., Parpinel, M., La Vecchia, C. and Decarli, A. (2012). Adherence to the European food safety authority's dietary recommendations and colorectal cancer risk. *Eur J Clin Nutr* **66**(4) 517–522.

[125] Van Havre, Z., White, N., Rousseau, J. and Mengersen, K. (2015). Overfitting Bayesian mixture models with an unknown number of components. *PloS one* **10**(7) 0131739.

[126] Varraso, R., Garcia-Aymerich, J., Monier, F., Le Moual, N., De Batlle, J., Miranda, G., Pison, C., Romieu, I., Kauffmann, F. and Maccario, J. (2012). Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. *Am J Clin Nutr* **96**(5) 1079–1092.

[127] Weikert, C. and Schulze, M. B. (2016). Evaluating dietary patterns: the role of reduced rank regression. *Curr Opin Clin Nutr Metab Care* **19**(5) 341–346.

[128] Westenbrink, S., Roe, M., Oseredczuk, M., Castanheira, I. and Finglas, P. (2016). EuroFIR quality approach for managing food composition data; where are we in 2014? *Food Chem* **193** 69–74.

[129] Willett, W. and Stampfer, M. J. (1986). Total energy intake: implications for epidemiologic analyses. *Am J Epidemiol* **124**(1) 17–27.

[130] Willett, W. C., Stampfer, M. J., Underwood, B. A., Speizer, F. E., Rosner, B. and Hennekens, C. H. (1983). Validation of a dietary questionnaire with plasma carotenoid and alpha-tocopherol levels. *Am J Clin Nutr* **38**(4) 631–639.

[131] Willett, W. (2012) *Nutritional epidemiology* **40**. Oxford University Press.

[132] Williams, C. M. (2022). Mechanistic evidence underpinning dietary policy: bringing the jigsaw pieces together? *Proc Nutr Soc* 1–8.

[133] Yoon, P. W., Rasmussen, S. A., Lynberg, M. C., Moore, C. A., Anderka, M., Carmichael, S. L., Costa, P., Druschel, C., Hobbs, C. A., Romitti, P. A. and others. (2001). The National Birth Defects Prevention Study. *Public Health Rep* **116**(Suppl 1) 32–40.

[134] Zhang, F., Tapera, T. M. and Gou, J. (2018). Application of a new dietary pattern analysis method in nutritional epidemiology. *BMC Med Res Methodol* **18**(1) 119.

[135] Zhao, J., Li, Z., Gao, Q., Zhao, H., Chen, S., Huang, L., Wang, W. and Wang, T. (2021). A review of statistical methods for dietary pattern analysis. *Nutr J* **20**(1) 37.

Valeria Edefonti. Branch of Medical Statistics, Biometry, and Epidemiology "G. A. Maccacaro", Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Italy.
E-mail address: valeria.edefonti@unimi.it

Roberta De Vito. Department of Biostatistics, Data Science Initiative, and Center for Computational Molecular Biology, Brown University, USA.
E-mail address: roberta_devito@brown.edu

Maria Parpinel. Department of Medicine - DAME, Università degli Studi di Udine, Italy.
E-mail address: maria.parpinel@uniud.it

Monica Ferraroni. Branch of Medical Statistics, Biometry and Epidemiology "G. A. Maccacaro", Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Italy.
E-mail address: monica.ferraroni@unimi.it