Statistica Sinica Preprint No: SS-2017-0093.R2	
Title	Clustering in General Measurement Error Models
Manuscript ID	SS-2017-0093.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0093
Complete List of Authors	Raymond Carroll Ya Su and Jill Reedy
Corresponding Author	Raymond Carroll
E-mail	carroll@stat.tamu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Clustering in General Measurement Error Models

Ya Su
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, ysu@stat.tamu.edu

Jill Reedy
Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD 20892, reedyj@mail.nih.gov

Raymond J. Carroll
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, and School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway NSW 2007, Australia, carroll@stat.tamu.edu

Abstract

This paper is dedicated to the memory of Peter G. Hall. It concerns a deceptively simple question: if one observes variables corrupted with measurement error of possibly very complex form, can one recreate asymptotically the clusters that would have been found had there been no measurement error? We show that the answer is yes, and that the solution is surprisingly simple and general. The method itself is to simulate, by computer, realizations with the same distribution as that of the true variables , and then to apply clustering to these realizations. Technically, we show that if one uses K-means clustering or any other risk minimizing clustering, and a multivariate deconvolution device with certain smoothness and convergence properties, then, in the limit, the cluster means based on our method converge to the same cluster means as if there were no measurement error. Along with the method and its technical justification, we analyze two important nutrition data sets, finding patterns that make sense nutritionally.

Some Key Words: Clustering; Deconvolution; K-means; Measurement error; Mixtures of distributions.

Short title: Clustering and Measurement Error

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Dedication to the memory of Peter G. Hall

The last author, Raymond J. Carroll, was very fond of Peter and visited him many times. His facets included brilliance, dedication, kindness, sense of humor, graciousness to young researchers, puzzle solving, madcap driving to take photos of trains, discussions about airplanes, love of cats, and photographic advice. As Peter said in his *Statistical Science* interview Delaigle and Wand (2016), *I always like working with Ray, because I felt I could contribute something from the problem solving side, the theoretical side, whereas he is more an applied person, ... in working with Ray we bring to the table things that don't overlap, and which complemented each other very well.*

In his interview, Peter also mentioned that a lot of his joint work with Raymond grew out of nutrition research, and hence this paper is an appropriate contribution to this special issue. It involves a deceptively simple question: if one observes variables corrupted with measurement error of possibly complex form, such as occurs in nutritional and radiation applications, can one recreate the clusters that would have been found had there been no measurement error? We show that the answer is yes, and that the solution is surprisingly simple and general.

1 Introduction

We consider the question of how to perform a cluster analysis in measurement error problems when the variable of interest is latent, and to do this clustering in such a way that, in large samples, it reproduces the clusters that would have been formed had the latent variable actually been observed. We develop a surprisingly simple, general strategy, to address this goal, and give theoretical evidence that it does have the requisite convergence.

There are many types of measurement error problems, depending on the problem at hand: (a) classical additive homoscedastic error; (b) classical heteroscedastic measurement error; (c) additive Berkson error; (d) multiplicative Berkson error; (e) combinations of (a)-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

(d); (f) multiplicative measurement error with excess zeros Kipnis et al. (2009); Zhang et al. (2011); (g) various multivariate versions of all of these; (h) combinations of misclassified and continuous variables Yi et al. (2015), etc. Full-length books on the topic include Gustafson (2004), Carroll et al. (2006), Buonaccorsi (2010), and Yi (2016).

Whatever the particular situation, measurement error problems have a few commonalities. There are data $\mathbf{X}$, multivariate in our case, which are not observable and the desire is to cluster them. There are observed proxies $\mathbf{W}$ that are related to $\mathbf{X}$, and there are additional error-free data, $\mathbf{Z}$, that can include covariates. In some cases, there may also be responses $\mathbf{Y}$ and a regression model relating these responses to $\mathbf{X}$ and some components of $\mathbf{Z}$; in this paper, we absorb $\mathbf{Y}$ into $\mathbf{W}$ for simplicity of notation. See Section 3 for the application of our ideas to two complex nonlinear measurement error model settings.

The problem is to find clusters for the distribution of $\mathbf{X}$, even though they were not observable. The solution to this problem is surprisingly simple, and consists of the following algorithm.

Algorithm 1.

- *Perform a measurement error analysis, of whatever kind.*
- *Estimate the distribution $F_X(\cdot)$ of $\mathbf{X}$ as $\widetilde{F}_{X,\mathrm{mes}}(\cdot)$, where the "mes" emphasizes that the estimation is based on a measurement error analysis.*
- *Generate realizations of $\mathbf{X}$ from the distribution of $\widetilde{F}_{X,\mathrm{mes}}(\cdot)$. Depending on the initial sample size, it may be advisable to generate multiple realizations for each individual in the study so as to remove simulation variability.*
- *Perform one's favorite cluster analysis on these realizations.*

Clearly, the algorithm is intuitive, since it involves generating data that have, asymptotically, the same distribution as that of $\mathbf{X}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Our main result is this algorithm. In Section 2, we discuss the classical deconvolution setting that estimates $F_X(\cdot)$. We show theoretically, under technical conditions, that if the algorithm is a generalization of K-means clustering, the algorithm converges, as the sample size tends to ∞, to the same cluster solutions as if the true variable were observed. In Section 3, we describe two data analyses where the measurement error models are very different, and describe how the clusters found make scientific sense.

We emphasize that we are not advertising that we can cluster the individual $\mathbf{X}$. We can only estimate the algorithm that would assign an individual $\mathbf{X}$ to a cluster if it were observed. Since these variables are latent, the only thing we can possibly hope to do for an individual is to estimate the *probabilities* that the particular individual $\mathbf{X}$ is in the various clusters, see the discussion in Section 4.

2 The Case of Nonparametric Deconvolution

Algorithm 1 works very generally, as we indicate at the end of this section. However, for specificity, we first consider here the special case of nonparametric deconvoluting density estimation in the classical measurement error model when observations are d-dimensional. The literature on this problem is large with very strong theoretical results; a small sample includes Carroll and Hall (1988), Stefanski and Carroll (1990), Fan (1991), Masry (1991), Li and Vuong (1998), and Comte et al. (2013). In this model, $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{X}$ and $\mathbf{U}$ are independent, have distribution functions F_X and F_U, respectively, and where F_X is unknown but, as is often assumed in the deconvolution literature, F_U is known; there are also papers where this last assumption is weakened.

In this case, Algorithm 1 works as follows. Suppose an independent identically distributed sample $\mathbf{W}_1, \ldots, \mathbf{W}_n$ is observed. Then the distribution function of $\mathbf{X}$ is estimated through deconvoluting density estimation, which is denoted as $\widetilde{F}_{X,\text{mes}}$. The theory that we give is based on the Fourier series estimator in Li and Vuong (1998), and on their theoretical results, but refer to Remark 1 to see why it holds for deconvoluting kernel density estimation.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Following such estimation, a pseudo-sample $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_n$ is generated from $\widetilde{F}_{X,\text{mes}}$, and a clustering procedure is applied to this pseudo-sample.

For specificity, we consider a class of clustering procedures defined as follows. Consider a clustering algorithm characterized by empirical risk (loss) minimization, where the risk function is chosen among a function class $\mathcal{H}_K$. Mathematically, if $\mathbf{X}$ could be observed, we define the clustering result as

$$\widehat{h}_n = \arg\min_{h \in \mathcal{H}_K} n^{-1} \textstyle\sum_{i=1}^n h(\mathbf{X}_i). \tag{1}$$

In K-means clustering, the aim is to find a set of cluster centers $\mathcal{C} = (\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_K)$ corresponding to the optimization problem

$$\mathcal{C} = \arg\min_{\mathbf{c}_1,\ldots,\mathbf{c}_K} n^{-1} \textstyle\sum_{i=1}^n \sum_{k=1}^K \|\mathbf{X}_i - \mathbf{c}_k\|^2 \mathbf{I}(\mathbf{c}_k \text{ is closest to } \mathbf{X}_i). \tag{2}$$

With $\mathcal{H}_K = \{h(\mathbf{z}) = \sum_{k=1}^K \|\mathbf{z} - \mathbf{c}_k\|^2 \mathbf{I}(\mathbf{c}_k \text{ is closest to } \mathbf{z}) : \mathbf{c}_1, \ldots, \mathbf{c}_K \in \mathbb{R}^d\}$, K-means clustering (2) is a special case of (1).

Similarly, since we cannot observe $\mathbf{X}$, with pseudo-observations $\widetilde{\mathbf{X}}$ we do actual data clustering by solving

$$\widetilde{h}_n \;=\; \arg\min_{h \in \mathcal{H}_K} n^{-1} \textstyle\sum_{i=1}^n h(\widetilde{\mathbf{X}}_i). \tag{3}$$

The question of whether Algorithm 1 gives, asymptotically, a solution that converges to the solution if $\mathbf{X}$ were observable, can be rephrased as whether the distance between $\widetilde{h}_n$ and $\widehat{h}_n$ converges to zero as $n \to \infty$. Of course, the empirical risk is a sample version of the expected risk. The underlying measure for the expected risk would vary as the method of constructing $\widetilde{F}_{X,\text{mes}}$ differs.

Remark 1. To see why Algorithm 1 works quite generally, observe that the difference in the empirical risk for any function $h \in \mathcal{H}_K$ between observing $\mathbf{X}$ and instead using $\widetilde{\mathbf{X}}$ is $\Delta(h) = \int h(x) d\{\widehat{F}_X(x) - \widetilde{F}_{n,m}(x)\}$, where $\widehat{F}_X$ and $\widetilde{F}_{n,m}$ are the empirical distribution function of the latent $\mathbf{X}_i$ and the pseudo-sample $\widetilde{\mathbf{X}}_i$. Standard theory has established that $\widehat{F}_X(\cdot)$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

converges to the true $F_X(\cdot)$. Assuming that similar theory in relation to the measurement error analysis is justified, it is also the case that $\widetilde{F}_{n,m}(\cdot)$ converges to $F_X(\cdot)$. It is then a technical matter of showing that $\Delta(h) \to 0$ uniformly for all $h \in \mathcal{H}_K$.

Using this insight, in the Supplementary Material, Section S.1.1 under Assumption 1, we show that within the function class $\mathcal{H}_K$, as the sample size $n \to \infty$, $\widehat{h}_n - \widetilde{h}_n \to 0$, and so asymptotically the clustering done using the $\widetilde{\mathbf{X}}$ and the clustering done using the unmeasured data $\mathbf{X}$ have the same asymptotic risk. For K-means clustering, this means that the cluster centers for the two converge to the same values.

Remark 2. For the classical measurement error model considered in this section, in the Supplementary Material, Section S.1.2, we also consider the scenario that the distribution of $\mathbf{X}$ is estimated parametrically in a misspecified family. Generally, in that case, the clusters thus found do not converge to the clusters based on the unobserved true variable. A semi-parametric generalization is given in Section 3.3 and in the Supplementary Material, Section S.1.3.

3 Examples

3.1 Background

The Dietary Patterns Methods Project is a collaborative project among multiple institutions (Fred Hutchinson Cancer Research Center, National Cancer Institute, University of Hawaii Cancer Center, University of South Carolina) investigating what dietary patterns there are and the relationship of such patterns with mortality and disease George et al. (2014); Harmon et al. (2015); Liese et al. (2015); McCullough (2014); Reedy et al. (2014). One way to define such patterns is through the use of cluster analysis, which is commonly used in this context Freitas-Vilela et al. (2016); Kim et al. (2015); Reedy et al. (2010); Thorpe et al. (2016); Villegas et al. (2004); Wirfält et al. (2009).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

However, it is well-known that usual (long-term average) dietary intakes are impossible to measure accurately, and the instruments used, such as 24 hour recalls and food frequency questionnaires, are subject to bias and measurement error. It is thus of considerable interest to understand dietary clusters based on usual intake, and not based on biased and error-prone measurements. In this section, we report on two data sets, in different contexts, and show the results of what our methodology obtains, and make good nutritional sense.

The examples considered are based on complex parametric multivariate measurement error models, not fitting into the classical measurement error model context, with the estimation of the parameters being done in a Bayesian way using MCMC. There is no asymptotic theory for such complex problems but, because they are in the end parametric models, we are assuming (reasonably) that the necessary convergence described in Remark 1 and the Supplementary Material, Section S.1.2, holds. The methods for both Sections 3.2 and 3.3 have been demonstrated in simulations to have good finite-sample behavior with little bias, so that such an assumption seems reasonable.

3.2 Clustering of Dietary Pattern Scores

We first consider clustering of usual dietary intakes using the National Institutes of Health-AARP Diet and Health Study (NIH-AARP) Reedy et al. (2008); Schatzkin et al. (2001). There are $n = 293,615$ men in our analysis. The clustering is based on 12 components of the Healthy Eating Index-2005 HEI-2005, Guenther et al. (2008), a multi-component index meant to measure adherence to the 2005 U. S. Department of Agriculture (USDA) Dietary Guidelines for Americans. Each food or nutrient is adjusted for energy (caloric) intake. The index components are listed in Table 1, as is the scoring system used, e.g., low amounts of saturated fat intake relative to energy intake produces a maximum component score of 10, while higher amounts of whole grains relative to energy in the diet produces a maximum total score of 5. It is traditional to sum up the scores into a total score and relate it to disease, but there is also great interest in understanding the dietary patterns of the 12 components,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

which is our aim.

The data are described in Section 2.1 of Potgieter, et al. (2016). The data generating mechanism for that data is extremely complex, consisting of two types of dietary data and multiple nutrition variables that have excess zeros, the episodically consumed foods. Consequently, W is extremely complex. One type of dietary variables measured is 24-hour recalls of diets, a measures of what the subject consumed in the previous day. These variables are considered unbiased for long-term dietary intakes $\mathbf{X}$, and if we call them $\mathbf{W}$, then $E(\mathbf{W}|\mathbf{X}) = \mathbf{X}$. Unlike in the classical measurement error model, however, some of the variables $\mathbf{W}$ measured by the 24-hour recall have excess zeros because, for example, a subject might not consume whole fruit on a given day. The observed data are also highly heteroscedastic. The other type of dietary variables measured in the study is food frequency questionnaires that measure the subject's estimate of $\mathbf{X}$ over the past six months, although they are biased for $\mathbf{X}$.

Much more detailed background of how the measurement error is modelled and how the scores are adjusted for measurement error is given in Zhang et al. (2011) and Potgieter et al. (2016). Supplementary material to Zhang et al. (2011) gives Matlab code, and SAS programs being used by many researchers in nutrition are at the web site https://epi.grants.cancer.gov/diet/usualintakes/method.html.

The details of the modeling efforts are quite lengthy, and we take it as a given that the methodology can be applied. Instead, in the interest of brevity, here we denote by $\mathbf{Z}$ covariates that affect the usual intake component scores $\mathbf{X}$. In the NIH-AARP Study, $\mathbf{Z}$ is of dimension 36, with 23 demographic components (age, body mass index category, etc.), and 13 components as measured by a food frequency questionnaire; these components are described in Table 1. Then, for a complex but known nonlinear function $\mathcal{F}$, as described in Zhang et al. (2011) and Potgieter et al. (2016), and for a normally distributed but unobserved random variable $\mathcal{U}$, the HEI component scores based on long-term average intakes are of the form $\mathbf{X} = \mathcal{F}(\mathbf{Z},\mathcal{U})$: in our setting, $\mathbf{X}$ is 12-dimensional. Of course, since we do not observe

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

$\mathcal{U}$, we cannot observe $\mathbf{X}$. We also add here that, for simplicity, we have suppressed notation that indicates that the model has parameters which are estimated.

To implement the method described in Section 1, we use the following procedure. For each individual i, we generated $j = 1, ..., J = 5$ normal random variables $\mathcal{U}_{ij}$, and then formed the realizations $\widetilde{\mathbf{X}}_{ij} = \mathcal{F}(\mathbf{Z}_i, \mathcal{U}_{ij})$ for a total sample size of $n \times J$. We assume the measurement error model is properly specified so that these realizations have the same distribution, asymptotically, as the true but unobserved $\mathbf{X}_i$, and thus fit into the framework in Sections 1-2. We then combined the data across $i = 1, ..., n$ and $j = 1, ..., J$ and applied K-means clustering to the entire data set, setting the number of clusters to $K = 3$. The Supplementary Material, Table S.1 and Figure S.1, gives results for $K = 4$, which are similar. When applying K-means, we first centered and standardized the $\widetilde{\mathbf{X}}_{ij}$, computed the resulting cluster means, and then back-calculated to the original data scale.

Table 2 gives the resulting cluster means, with the following interesting results.

- The cluster means differ largely only in total fruit, whole fruit, saturated fat, and empty calories. For total fruit and whole fruit, Clusters 2 and 3 have much higher cluster means than does Cluster 1. For whole grains and DOL (dark-green and orange vegetables and legumes), Cluster 3 has higher cluster means than Clusters 1 and 2.

- Cluster means are also somewhat higher for Cluster 3 for two other components: whole grains, and DOL (dark-green and orange vegetables and legumes).

- The clusters are ordered by the total of the cluster means, an important finding not guaranteed a priori. Since lower scores mean worse diets, it is also clear that Cluster 1 has the worst diets (48.4 points out of 100 possible points), while Cluster 3 has the best diets (69.9 points), and Cluster 2 has the in-between diets (60.3 points).

- For saturated fat and empty calories, Cluster 3 has higher means than does Cluster 2, and Cluster 2 has cluster means that are much higher than those of Cluster 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Figure 1 provides a different view via a radar plot. Here, because it is of nutritional interest and fairly standard practice, what is plotted is the % of the maximum possible score for each dietary component. This is useful because the HEI-2005 scoring system uses different maximum scores for each component, as described above. We see in Figure 1 that the worst diet group is vastly different from the best diet group as a % of total score for total fruit, whole fruit, saturated fat, and SoFAAS (empty calories), as before. However, visually, we see important differences as well for whole grains and dark green/orange vegetables and legumes.

A striking feature of these results is how the cluster with the best diets, Cluster 3, has higher component scores than the worst diets, Cluster 1, on *every* dietary component except an essential tie for Meat & Beans, and a clear discrepancy for Sodium. Thus, the clustering done here makes a great deal of scientific sense: the scores were based on the USDA Dietary Guidelines, and the scoring system explicitly gives higher scores to those who more closely adhere to these guidelines.

For a recent look at the complexities of the issue of the benefit of low sodium intake on cardiovascular health, see Oparil (2014).

3.3 Clustering of Relative Dietary Amounts

Next, we use data from the Eating at America's Table Study Subar et al. (2001), which consists of 965 subjects who completed four 24 hour dietary recalls over the course of a year. Because absolute nutrient intakes are highly correlated with total energy/calories, it is common to normalize these numbers for the amount of energy consumed, as was done in Section 3.2, see Table 2. Nutritionists call such quantities nutrient or food "densities". The variables considered here are the percentage of kilocalories/energy coming from protein, saturated fat, and fiber. The percentages are computed as protein density = 400 * protein in grams / energy, saturated fat density = 900 * saturated fat in grams / energy and fiber density = 400 * fiber in grams /energy.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

There are covariates $\mathbf{Z}$ that are also predictive of the observed mean nutrient densities, including age, sex, body mass index in 4 categories, ethnicity in 3 categories and education in 4 categories, and it makes sense to include these as predictors of usual intake. Thus, the 24 hour recalls are $\mathbf{W}_{ij}$, and a natural model is

$$\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}, \qquad \mathbf{X}_i = \mathcal{A}\mathbf{Z}_i + \xi_i, \tag{4}$$

where ξ_i is independent of $\mathbf{X}_i$ and has mean zero. This is a far different model than that used in Section 3.2. We thus need to model the joint distribution of $(\xi, \mathbf{U}_i)$ flexibly. To do this, we follow the flexible semiparametric approach of Sarkar et al. (2017), see also Sarkar et al. (2014) for a univariate version. Computation was done using their R program. In the model of Sarkar et al. (2017), the distribution of ξ was modelled by a flexible multivariate mixture of normals. Then the measurement error distribution of $\mathbf{U}_i$ was modeled as conditionally heteroscedastic, so that

$$\mathbf{U}_{ij} = \mathbf{S}(\xi_i)\epsilon_{ij}, \tag{5}$$

where $\mathbf{S}(\xi)$ is a diagonal matrix, with each diagonal function a Bspline. In addition, ϵ_{ij} was also modeled as a flexible multivariate mixture of normals. In the Supplementary Material, Section S.1.3, we show a theoretical justification for Algorithm 1 to cluster the $\widetilde{\mathbf{X}}_i$ under model (4) is given.

In practice what we do is to regress the mean recalls $\overline{\mathbf{W}}_{i\cdot}$ of the dietary variables on the covariates, obtain an estimate $\widehat{\mathcal{A}}$, form the residuals, and then fit the model of Sarkar et al. (2017) to the residuals: computation of the last step was done via an R program included in their Supplementary Material. Sarkar et al. (2017) do the multivariate deconvolution using an MCMC approach. In our example, we took a burn-in of 1000 steps, and then generated a further sample with 4000 steps. Upon convergence of the sampler, the MCMC allows us to use the MCMC steps to generate realizations of the $\widetilde{\xi}_i$, and hence to generate realizations of usual intake $\widetilde{\mathbf{X}}_i$ by adding on $\widehat{\mathcal{A}}\mathbf{Z}_i$. To do this, after the burn-in, we took a realization of $\widetilde{\xi}_i$ for every 100^{th} iteration in the MCMC.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

After creating the realizations of usual intake $\widetilde{\mathbf{X}}_i$, we used K-means clustering based on $K = 3$ clusters, the resulting cluster means are given in Table 3. The results are striking here as well. Cluster 1 has the highest protein intake, the highest fiber intake, and the lowest saturated fat intake. Clusters 2 and 3 have much lower fiber intakes, while differing on protein intakes. These are very different dietary configurations.

4 Discussion

At the end of Section 1, we emphasized that our method, and indeed no method, can actually and precisely place the latent variables $\mathbf{X}$ into a cluster, and that one can only estimate the probabilities that an individual is in a cluster. At least for the discussion of dietary patterns, estimating these individual probabilities is not a major practical or scientific issue. However, it becomes so when the interest is in relating cluster membership to a disease. It is quite easy to estimate the probabilities that an individual is in a cluster. Recall that in Section 3.2, in order to cut down on simulation variability for building the clusters, for $j = 1, ..., J = 5$ we generated realizations $\widetilde{\mathbf{X}}_{ij}$. We then created a pseudo-sample $(\widetilde{\mathbf{X}}_{1j}, ..., \widetilde{\mathbf{X}}_{nj})$ across i and j, and performed the clustering. We did the same thing in Section 3.3, but there $J = 40$, because the sample size in Section 3.3 is much smaller than that in Section 3.2.

Having formed the clusters, we now set J to be a rather large number, and again create pseudo-observations $\widetilde{\mathbf{X}}_{ij}$ for $i = 1, ..., n$ and $j = 1, ..., J$, but this time J is much larger. Then, for an arbitrary person i_*, compute the cluster assignments for $\widetilde{\mathbf{X}}_{i_*j}$ for $j = 1, ..., J$. By the law of large numbers, the fraction of the time that the pseudo-observations are assigned to Cluster 1 (say) is an estimate of the probability that the individual i_* is in Cluster 1. This is done for each cluster and each individual, thus forming cluster probabilities for the entire sample. The procedure itself is entirely general, and of course can be applied in the two examples in Section 3.

In Table 4, for the example in Section 3.2, we give a listing of the cluster membership probabilities for the first 10 subjects in the data file for the example of Section 3.2, which

were formed by taking $J = 1000$. For a few subjects, it is obvious which cluster they are probably in, e.g., Subject 1 has a 98% chance of being in Cluster 1. However, for other subjects, a "cluster call" makes no sense, e.g. Subject 7 is essentially equally likely to be in Cluster 2 or Cluster 3. Table S.3 of the Supplementary Material gives the same information for the example in Section 3.3, with similar results, there we just continued the MCMC and formed pseudo-observations with $J = 1000$. How to use these probabilities efficiently in an analysis of disease risk is a topic for future research.

Supplementary Material

The online Supplementary Material includes proofs, the analysis of Section 3.2 but with $K = 4$ clusters (table and radar plot), cluster membership probabilities for 10 individuals in the analysis of Section 3.3, and additional references.

Acknowledgments

Su and Carroll were supported by a grant from the National Cancer Institute (U01-CA057030). We especially thank the referees for their cogent comments on the paper, and Susan Krebs-Smith and Amy Subar of the National Cancer Institute for their long-term support of this line of research.

References

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods and Applications*. Chapman & Hall.

Carroll, R. J. and P. Hall (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association 83*, 1184–1186.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall.

Comte, F., C. Lacour, et al. (2013). Anisotropic adaptive kernel deconvolution. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 49*, 569–609.

Delaigle, A. and M. P. Wand (2016). A Conversation with Peter Hall. *Statistical Science 31*, 275–304.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics 19*, 1257–1272.

Freitas-Vilela, A. A., A. D. Smith, G. Kac, R. M. Pearson, J. Heron, A. Emond, J. R. Hibbeln, M. B. T. Castro, and P. M. Emmett (2016). Dietary patterns by cluster analysis in pregnant women: relationship with nutrient intakes and dietary patterns in 7-year-old offspring. *Maternal & Child Nutrition 13*, DOI:10.1111/mcn.12353.

George, S. M., R. Ballard-Barbash, J. E. Manson, J. Reedy, J. M. Shikany, A. F. Subar, L. F. Tinker, M. Vitolins, and M. L. Neuhouser (2014). Comparing indices of diet quality with chronic disease mortality risk in postmenopausal women in the Women's Health Initiative Observational Study: evidence to inform national dietary guidance. *American Journal of Epidemiology 180*(6), 616–625.

Guenther, P. M., J. Reedy, S. M. Krebs-Smith, and B. B. Reeve (2008). Evaluation of the Healthy Eating Index-2005. *Journal of the American Dietetic Association 108*, 1854–1864.

Gustafson, P. (2004). *Measurement Error and Misclassication in Statistics and Epidemiology*. Chapman and Hall/CRC.

Harmon, B. E., C. J. Boushey, Y. B. Shvetsov, R. Ettienne, J. Reedy, L. R. Wilkens, L. Le Marchand, B. E. Henderson, and L. N. Kolonel (2015). Associations of key diet-quality indexes with mortality in the Multiethnic Cohort: the Dietary Patterns Methods Project. *The American Journal of Clinical Nutrition 101*(3), 587–597.

Kim, J., A. Yu, B. Y. Choi, J. H. Nam, M. K. Kim, D. H. Oh, and Y. J. Yang (2015). Dietary patterns derived by cluster analysis are associated with cognitive function among Korean older adults. *Nutrients 7*, 4154–4169.

Kipnis, V., D. Midthune, D. W. Buckman, K. W. Dodd, P. M. Guenther, S. M. Krebs-Smith, A. F. Subar, J. A. Tooze, R. J. Carroll, and J. A. Freedman (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics 65*, 1003–1010.

Li, T. and Q. Vuong (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis 65*(2), 139–165.

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

Masry, E. (1991). Multivariate probability density deconvolution for stationary random processes. *IEEE Transactions on Information Theory 37*, 1105–1115.

McCullough, M. L. (2014). Diet patterns and mortality: common threads and consistent results. *The Journal of Nutrition 144*(6), 795–796.

Oparil, S. (2014). Low sodium intake: cardiovascular health benefit or risk? *New England Journal of Medicine 371*, 677–679.

Potgieter, C. J., R. Wei, V. Kipnis, L. S. Freedman, and R. J. Carroll (2016). Moment reconstruction and moment-adjusted imputation when exposure is generated by a complex, nonlinear random effects modeling process. *Biometrics*, DOI: 10.1111/biom.12524.

Reedy, J., S. M. Krebs-Smith, P. E. Miller, A. D. Liese, L. L. Kahle, Y. Park, and A. F. Subar (2014). Higher diet quality is associated with decreased risk of all-cause, cardiovascular disease, and cancer mortality among older adults. *Journal of Nutrition 144*(6), 881–889.

Reedy, J., P. N. Mitrou, S. M. Krebs-Smith, E. Wirfält, A. V. Flood, V. Kipnis, M. Leitzmann, T. Mouwand, A. Hollenbeck, A. Schatzkin, and A. F. Subar (2008). Index-based dietary patterns and risk of colorectal cancer: the NIH-AARP Diet and Health Study. *American Journal of Epidemiology 168*, 38–48.

Reedy, J., E. Wirfält, A. Flood, P. N. Mitrou, S. M. Krebs-Smith, V. Kipnis, D. Midthune, M. Leitzmann, A. Hollenbeck, A. Schatzkin, et al. (2010). Comparing 3 dietary pattern methods-cluster analysis, factor analysis, and index analysis-with colorectal cancer risk in the NIH–AARP Diet and Health Study. *American Journal of Epidemiology 171*(4), 479–487.

Sarkar, A., B. K. Mallick, J. Staudenmayer, D. Pati, and R. J. Carroll (2014). Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics 23*, 1101–1125.

Sarkar, A., D. Pati, A. Chakraborty, B. K. Mallick, and R. J. Carroll (2017). Bayesian semiparametric multivariate density deconvolution. *Journal of the American Statistical Association 112*.

Schatzkin, A., A. F. Subar, F. E. Thompson, et al. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: the national institutes of health-aarp diet and health study. *American Journal of Epidemiology 154*, 1119–1125.

Stefanski, L. and R. J. Carroll (1990). Deconvoluting kernel density estimators. *Statistics 21*, 165–184.

Subar, A. F., F. E. Thompson, V. Kipnis, D. Mithune, P. Hurwitz, S. McNutt, A. McIntosh, and S. Rosenfeld (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology 154*, 1089–1099.

Thorpe, M. G., C. M. Milte, D. Crawford, and S. A. McNaughton (2016). A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. *International Journal of Behavioral Nutrition and Physical Activity 13*, 1.

Villegas, R., A. Salim, M. Collins, A. Flynn, and I. Perry (2004). Dietary patterns in middle-aged Irish men and women defined by cluster analysis. *Public Health Nutrition 7*(08), 1017–1024.

Wirfält, E., D. Midthune, J. Reedy, P. Mitrou, A. Flood, A. F. Subar, M. Leitzmann, T. Mouw, A. R. Hollenbeck, A. Schatzkin, et al. (2009). Associations between food patterns defined by cluster analysis and colorectal cancer incidence in the NIH–AARP Diet and Health Study. *European Journal of Clinical Nutrition 63*(6), 707–717.

Yi, G. Y. (2016). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer.

Yi, G. Y., Y. Ma, D. Spiegelman, and R. J. Carroll (2015). Functional and structural methods with mixed measurement error and misclassification in covariates. *Journal of the American Statistical Association 109*, 681–696.

Zhang, S., D. Midthune, P. M. Guenther, S. M. Krebs-Smith, V. Kipnis, K. W. Dodd, D. W. Buckman, J. A. Tooze, L. Freedman, and R. J. Carroll (2011). A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Annals of Applied Statistics 5*, 1456–1487.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Component	Units	HEI-2005 score calculation
Total Fruit	cups	$\min(5, 5 \times (\text{density}/.8))$
Whole Fruit	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Vegetables	cups	$\min(5, 5 \times (\text{density}/1.1))$
DOL	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Grains	ounces	$\min(5, 5 \times (\text{density}/3))$
Whole Grains	ounces	$\min(5, 5 \times (\text{density}/1.5))$
Milk	cups	$\min(10, 10 \times (\text{density}/1.3))$
Meat and Beans	ounces	$\min(10, 10 \times (\text{density}/2.5))$
Oil	grams	$\min(10, 10 \times (\text{density}/12))$
Saturated Fat	% of energy	if density ≥ 15 score $= 0$ else if density ≤ 7 score $= 10$ else if density > 10 score $= 8 - (8 \times (\text{density} - 10)/5)$ else, score $= 10 - (2 \times (\text{density} - 7)/3)$
Sodium	milligrams	if density ≥ 2000 score=0 else if density ≤ 700 score=10 else if density ≥ 1100 score $= 8 - \{8 \times (\text{density} - 1100)/(2000 - 1100)\}$ else score $= 10 - \{2 \times (\text{density} - 700)/(1100 - 700)\}$
SoFAAS (Empty Calories)	% of energy	if density ≥ 50 score $= 0$ else if density ≤ 20 score=20 else score $= 20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$

Table 1: Description of the HEI-2005 scoring system. Except for saturated fat and SoFAAS, density is obtained by multiplying usual intake by 1000 and dividing by usual intake of kilo-calories. For saturated fat, density is 900 usual saturated fat (grams) divided by usual calories, i.e., the percentage of usual calories coming from usual saturated fat intake. For SoFAAS, the density is the percentage of usual intake that comes from usual intake of calories, i.e., the division of usual intake of SoFAAS by usual intake of calories. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. The total HEI-2005 score is the sum of the individual component scores.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

	Cluster 1	Cluster 2	Cluster 3	Maximum Possible
Total Fruit	1.60	4.18	4.30	5
Whole Fruit	1.45	4.63	4.55	5
Total Grains	4.66	4.63	4.93	5
Whole Grains	1.07	1.24	2.41	5
Total Vegetables	3.80	4.11	4.73	5
DOL	1.30	1.58	2.77	5
Milk	5.05	5.12	5.68	10
Meat & Beans	9.76	9.69	9.62	10
Oil	5.73	6.55	5.79	10
Saturated Fat	4.57	5.79	8.44	10
Sodium	1.97	2.65	1.26	10
Empty Calories	7.42	10.15	15.44	20
Total Score	48.38	60.32	69.92	100

Table 2: Cluster mean scores for the NIH-AARP analysis of Section 3.2, and their maximum possible values. The Total Score is the sum of the cluster means. The four bold-faced dietary components show the greatest difference between Cluster 1, the poor diet group, and Cluster 3, the best diet group. The cluster "sizes", i.e., the sum of the probabilities of being in each cluster, are 85878, 98014 and 109723, respectively.

Cluster	Protein	Saturated Fat	Fiber
1	17.49	8.62	4.67
2	16.85	12.51	2.85
3	13.76	10.51	2.94

Table 3: Analysis of the EATS data in Section 3.3. The table displays the cluster centers using K-means clustering with $K = 3$. The estimated sizes for these clusters are 10,830, 13,326 and 14,404.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

Subject	Cluster 1	Cluster 2	Cluster 3
1	0.980	0.016	0.004
2	0.719	0.265	0.016
3	0.191	0.351	0.458
4	0.468	0.515	0.017
5	0.217	0.478	0.305
6	0.001	0.545	0.454
7	0.045	0.447	0.508
8	0.118	0.735	0.147
9	0.472	0.320	0.208
10	0.194	0.488	0.318

Table 4: Cluster membership probabilities for the first 10 subjects in the HEI-2005 example of Section 3.2. A cluster "call" is difficult to make for subjects 3, 4, 5, 6, 7, 9 and 10.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0093

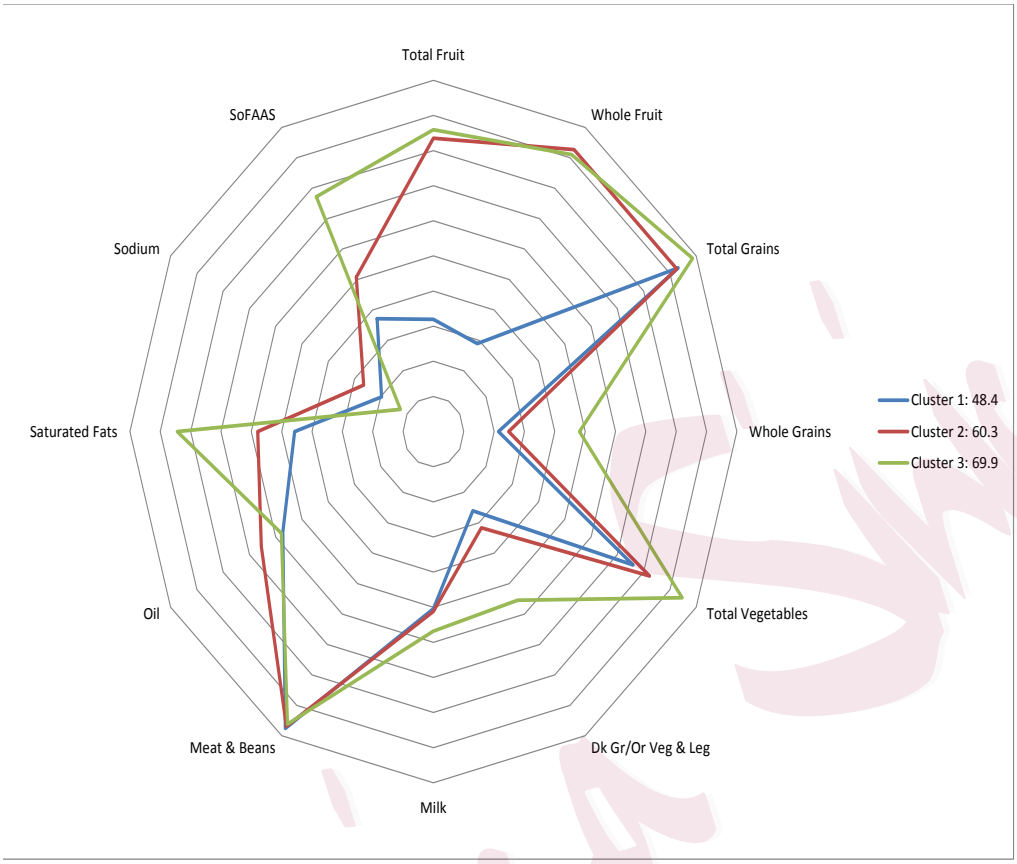

Figure 1: Data analysis for the NIH-AARP HEI-2005 analysis in Section 3.2. Radar plot of usual intake HEI-2005 scores. The listed amounts for the clusters are the means of the HEI-2005 total score within the clusters, although the total score was not part of the clustering algorithm. The cluster sizes were 85,878, 90,104 and 109,723, respectively. The online version of this plot is in color, but the 3 clusters are easily distinguished in the black and white plot.