Supplementary Material to *Clustering in Complex Latent Variable Models with Many Covariates*

Ya Su[1], Jill Reedy[2] and Raymond J. Carroll[1,3]

[1]*Department of Statistics, Texas A&M University,*

[2]*Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute,*

[3]*School of Mathematical and Physical Sciences, University of Technology Sydney*

Summary

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.

S.1 Theoretical Development

The purpose of this section is to show that there are circumstances where our algorithm of generating realizations of the latent variables will, at least asymptotically, result in the same set of cluster means as if these latent variables were observed. Algorithm 1 is clearly much more general than these contexts. We will illustrate our theoretical results below for two set of models, the classical measurement error model in Section S.1.1 and the model (4) in Section S.1.3.

S.1.1 Nonparametric Deconvolution

Consider nonparametric deconvolution. Following the framework of Section 2, for any function $h \in \mathcal{H}_K$, the expected risk is

$$\mathbb{E}_{F_X}(h) = \int h(y) dF_X(y). \tag{S.1}$$

Assumption 1. Consider nonparametric deconvolution density estimation. Make the following assumptions.
(A). F_X has bounded support.
(B). The characteristic functions of F_X and F_U are both integrable and nonvanishing everywhere.
(C). The function class $\mathcal{H}_K$ is uniformly Glivenko-Cantelli.
(D). The expected risk (loss), $\mathbb{E}_{F_X}(h)$, has a unique minimizer in $\mathcal{H}_K$.

Remark 1.
(i) Assumptions 1(A) and 1(B) are the same as Conditions A3-A5 in Li and Vuong (1998).
(ii) Assumption 1(C) is common in the field of learning theory and empirical processes that is related to and can be characterized by its entropy. Under Assumption 1(A), the function class $\mathcal{H}_K = \{h(\mathbf{z}) = \sum_{k=1}^{K} \|\mathbf{z} - \mathbf{c}_k\|^2 \mathbf{I}(\mathbf{c}_k \text{ is closest to } \mathbf{z}) : \mathbf{c}_1, \ldots, \mathbf{c}_K \in \mathbb{R}^d\}$ is totally bounded for the sup norm, and thus by Proposition 10.5 of Dudley (1999) Assumption 1(C) holds for K-means clustering.
(iii) The validity of Assumption 1(D) is linked to the concept of stability

of an algorithm. Under Assumption 1(D), a clustering algorithm is stable given that it is convergent, see e.g. Ben-David et al. (2006). On the other hand, it has been shown in Ben-David et al. (2006) that instability of a clustering algorithm is inevitable if the underlying probability distribution function F_X maintain a certain symmetry property (there exists a map g under which $F_X(A) = F_X\{g(A)\}$ for all measurable set $A \subset \mathbb{R}^d$) and the expected risk depends only on the distances and F_X. However, the symmetry requirement appears to be something ideal, therefore, from a practical aspect, Assumption 1(D) is not stringent.

The focus next is on the relationship between $\widehat{h}_n$ and $\widetilde{h}_n$. Since both correspond to the minimizers of certain empirical risk function, it turns out to be convenient to look at the almost minimizers set for the expected risk (S.1), namely, for fixed ϵ,

$$Q^{\epsilon}_{F_X} = \{h \in \mathcal{H}_K : \mathbb{E}_{F_X}(h) \leq \inf_{h' \in \mathcal{H}_K} \mathbb{E}_{F_X}(h') + \epsilon\}. \tag{S.2}$$

The notion of diameter is used to characterize the size of the almost minimizers set. Here the diameter is chosen to be the L_1 norm, that is, $\mathrm{diam}(Q^{\epsilon}_{F_X}) = \min_{h_1, h_2 \in Q^{\epsilon}_{F_X}} \|h_1 - h_2\|_{L_1}$.

Proposition 1. Under Assumption 1(D), $\mathrm{diam}(Q^{\epsilon}_{F_X})$ converges to zero as ϵ goes to zero.

Proposition 1 can be shown by contradiction. Refer to Rakhlin and Caponnetto (2006) for details.

Theorem 1. Under Assumptions 1(A), 1(B) and 1(C), for any $\epsilon > 0$ there exists n_ϵ such that for all $n > n_\epsilon$, $\widetilde{h}_n$ and $\widehat{h}_n$ are in the almost minimizers set $Q^{\epsilon}_{F_X}$.

Corollary 1. Under Assumptions 1(A), 1(B), 1(C) and 1(D), $\|\widetilde{h}_n - \widehat{h}_n\|_{L_1} \to 0$ as $n \to \infty$.

Proof of Theorem 1. Let $h^{am} \in Q^{\epsilon/2}_{F_X}$ be one function in the almost minimizers set (here ϵ scaled by 2 is for notational simplicity) then immediately by the definition of $\widetilde{h}_n$ and $\widehat{h}_n$,

$$n^{-1}\textstyle\sum_{i=1}^{n} \widetilde{h}_n(\widetilde{\mathbf{X}}_i) \leq n^{-1}\sum_{i=1}^{n} h^{am}(\widetilde{\mathbf{X}}_i); \tag{S.3}$$

$$n^{-1}\sum_{i=1}^{n}\widehat{h}_n(\mathbf{X}_i) \leq n^{-1}\sum_{i=1}^{n}h^{am}(\mathbf{X}_i). \tag{S.4}$$

Let $\widetilde{F}_{n,m}$ be the empirical distribution function based on $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_n$. In order to compare the expected risk between $\widetilde{h}_n$ and h^{am}, apply the following identical transformation to $\widetilde{h}_n$ and h^{am},

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}h(\widetilde{\mathbf{X}}_i) &= \int h(y)d\widetilde{F}_{n,m}(y) \\ &= \int h(y)d\widetilde{F}_{n,m}(y) - \int h(y)d\widetilde{F}_{X,\mathrm{mes}}(y) \\ &\quad + \int h(y)d\widetilde{F}_{X,\mathrm{mes}}(y) - \int h(y)dF_X(y) + \int h(y)dF_X(y) \end{aligned} \tag{S.5}$$

The last term in (S.5) equals $\mathbb{E}_{F_X}(h)$. Together with (S.3) we are able to provide evidence that with probability approaching 1, $\mathbb{E}_{F_X}(\widetilde{h}_n) \leq \mathbb{E}_{F_X}(h^{am}) + \epsilon/2$, that is, $\widetilde{h}_n \in Q^{\epsilon}_{F_X}$, if one can show

$$\sup_{h\in\mathcal{H}_K}\left|\int h(y)d\widetilde{F}_{n,m}(y) - \int h(y)d\widetilde{F}_{X,\mathrm{mes}}(y)\right| < \epsilon/8; \tag{S.6}$$

$$\sup_{h\in\mathcal{H}_K}\left|\int h(y)d\widetilde{F}_{X,\mathrm{mes}}(y) - \int h(y)dF_X(y)\right| < \epsilon/8, \tag{S.7}$$

both with probability approaching 1.

On the other hand, let $\widehat{F}_X$ be the empirical distribution function based on $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the relationship between $n^{-1}\sum_{i=1}^{n}h(\mathbf{X}_i)$ and $\mathbb{E}_{F_X}(h)$ is

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}h(\mathbf{X}_i) &= \int h(y)d\widehat{F}_X(y) \\ &= \int h(y)d\widehat{F}_X(y) - \int h(y)dF_X(y) + \int h(y)dF_X(y) \end{aligned} \tag{S.8}$$

Now (S.8) holds for $\widehat{h}_n$ and h^{am}. Then a sufficient condition for $\widehat{h}_n \in Q^{\epsilon}_{F_X}$, with probability approaching 1 as $n \to 0$, is that

$$\sup_{h\in\mathcal{H}_K}\left|\int h(y)d\widehat{F}_X(y) - \int h(y)dF_X(y)\right| < \epsilon/4 \tag{S.9}$$

holds with probability approaching 1.

We next confirm the validity of (S.6), (S.7) and (S.9).

Both (S.6) and (S.9) are implied by Assumption 1(C), which says

$$\sup_P P^*\left(\sup_{h\in\mathcal{H}_K} \left| \int h dP_n - \int h dP \right| > \epsilon \right) \to 0, \text{ as } n \to 0. \tag{S.10}$$

Here P^* is an outer measure of the product measure P^n because of the measurability issue of the event inside the probability.

Remark 2. The uniformly Glivenko-Cantelli property is useful in showing that the convergence (in probability) of $\int h dP_n - \int h dP$ is uniform in h when the underlying probability is changing with n and with the sample observed. Under (S.10), (S.9) holds naturally for a single choice of P, F_X. The random quantity in (S.6) lies in the same probability space as that of $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_n$, say, $\widetilde{P}$. Given $\mathbf{W}_1, \ldots, \mathbf{W}_n$, the conditional probability of $\widetilde{P}$ becomes $\widetilde{F}_{X,\text{mes}}$. The uniformly Glivenko-Cantelli property shows that (S.6) holds under all conditional probabilities, so that (S.6) is valid with respect to the marginal distribution $\widetilde{P}$.

To see (S.7), we turn to the literature on the convergence of $\widetilde{F}_{X,\text{mes}}$ to F_X. Li and Vuong (1998) established the uniform convergence of the density of $\widetilde{F}_{X,\text{mes}}$ to that of F_X assuming Assumption 1(A) and 1(B). Given the uniform convergence of the density and bounded support of F_X, the term in (S.7) shrinking to zero follows immediately.

This concludes the proof of Theorem 1. □

S.1.2 Parametric Deconvolution

Here we suppose that we have a parametric model S for the distribution of X and then estimate F_X by maximizing the likelihood function based on marginal distribution of $\mathbf{W}$, i.e.,

$$\widetilde{F}_{X,\text{mes}} = \arg\max_{P_X\in S} \sum_{i=1}^n \log \int f_U(W_i - x) dP_X(x). \tag{S.11}$$

Here, the expected risk is

$$\mathbb{E}_{X,S}(h) = \int h(y) dP_{X,S}(y). \tag{S.12}$$

The function $P_{X,S} \in S$ minimizes the Kullback-Leibler distance between distributions of W under the true model yielded by F_X and the misspecified one in terms of, say $P_X \in S$.

Finally, the almost minimizers set is

$$Q^{\epsilon}_{X,S} = \{h \in \mathcal{H}_K : \mathbb{E}_{X,S}(h) \leq \inf_{h' \in \mathcal{H}_K} \mathbb{E}_{X,S}(h') + \epsilon\}. \tag{S.13}$$

Also, $\text{diam}(Q^{\epsilon}_{X,S}) = \min_{h_1,h_2 \in Q^{\epsilon}_{X,S}} \|h_1 - h_2\|_{L_1}$.

The parametric model yields an estimate $\widetilde{F}_{X,\text{mes}} \in S$. White (1982) has arguments stating the consistency of $\widetilde{F}_{X\text{mes}}$ to $P_{X,S}$ under the following conditions: we have avoided a lot of notation here.

Assumption 2. The following are inherited from the A1, A2 and A3 in White (1982).
(A). The true distribution of **W** has a measurable Radon-Nikodym density.
(B). Under the parametric distribution in S, the distribution of **W** has a measurable Radon-Nikodym density. If the parametric distribution is specified by certain parameter θ, the density is continuous in θ for fixed value of w.
(C). The expectation of the logarithm of the density in Assumption 2(A) exists while the logarithm of the density in (B), say $f(w, \theta)$ is bounded above by some function in w, $m(w)$. Also, m is integrable with respect to the density in Assumption 2(A).
(D). The Kullback-Leibler distance between distributions of W under the true model yielded by F_X and the misspecified one in terms of, say $P_X \in S$, has a unique minimizer, $P_{X,S}$, defined above.

Assumption 3.
(A). The function class $\mathcal{H}_K$ is uniformly Glivenko-Cantelli.
(B). The expected risk (loss), $\mathbb{E}_{X,S}(h)$, has a unique minimizer in $\mathcal{H}_K$.

Proposition 2. Under Assumption 3(B), $\text{diam}(Q^{\epsilon}_{X,S})$ converges to zero as ϵ goes to zero.

The proof of Proposition 2 is the same as Proposition 1 in Section S.1.1.

Theorem 2. Under Assumptions 2(A)-2(D) and Assumption 3(A), for any $\epsilon > 0$ and there exists n_ϵ such that for all $n > n_\epsilon$, $\widetilde{h}_n$ and $\widehat{h}_n$ are in the almost minimizers sets $Q^{\epsilon}_{X,S}$ and $Q^{\epsilon}_{F_X}$.

The steps to prove Theorem 2 are almost identical to that of Theorem 1. The differences lie in the details filled in these steps. The major difference is in (S.5). Instead of the term $\int h(y)dF_X(y)$ there, here under Assumption 2, we have the expected risk term $\int h(y)dP_{X,S}(y)$. As previously mentioned that the distribution $\widetilde{F}_{X,\text{mes}}$ converges to $P_{X,S}$. On the other hand The proof of $\widehat{h}_n$ in $Q^{\epsilon}_{F_X}$ remains the same as Theorem 1.

Remark 3. It is worth pointing out that the parametric model needs to be correctly specified. Theorem 2 indicates that when we fit a parametric model to the distribution of X, it is likely that the clustering function $\widetilde{h}_n$ and the clustering function $\widehat{h}_n$ do not converge to each other unless the true distribution of X falls into the parametric family S.

S.1.3 Classical-Type Measurement Error Model (4) Where the Latent Variable is Associated with Covariates

Under model (4),

$$\mathbf{W}_{ij} = \mathcal{A}\mathbf{Z}_i + \xi_i + \mathbf{U}_{ij}. \tag{S.14}$$

In this set up, $\widetilde{\mathbf{X}}_i = \widehat{\mathcal{A}}\mathbf{Z}_i + \widetilde{\xi}_i$, $\widehat{\mathcal{A}}$ is a consistent estimator of $\mathcal{A}$ and $\widetilde{\xi}_1, \ldots, \widetilde{\xi}_n$ are an independent and identically distributed pseudo-sample from $\widetilde{F}_{\xi,\text{mes}}$.

If $\widetilde{F}_{\xi,\text{mes}}$ is deconvolved nonparametrically, and if, in addition, we assume that $q(z) = \lim_n n^{-1}\sum_i 1_{(\mathbf{Z}_i \le z)}$ and $\mathbf{Z}_i$ for $i = 1, \ldots, n$ are bounded, under the same flow of logic as in Section S.1.1, $\widetilde{h}_n$ will converge to the minimizer of the corresponding expected risk. The expected risk now takes the form

$$\mathbb{E}_{F_{\xi,Z}}(h) = \int h(\mathcal{A}z + y)dF_{\xi,Z}(y,z) \tag{S.15}$$

with $F_{\xi,Z}(y,z) = q(z)F_\xi(y)$.

Under model (4), the distribution of $\mathbf{X}$ depends on $\mathbf{Z}$ and can be expressed as $F_X(x) = \int F_\xi(x - \mathcal{A}z)dq(z)$. Therefore, (S.15) has another variant

$$\mathbb{E}_{F_X}(h) = \int h(x)dF_X(x). \tag{S.16}$$

As in Section S.1.1, we know that $\widehat{h}_n$ converges to the minimizer of (S.16). We conclude that $\widetilde{h}_n$ and $\widehat{h}_n$ will yield the same clusters asymptotically.

Remark 4. In view of model (4), an alternative approach would be to fit a nonparametric model to $\mathbf{X}$ ignoring the covariates entirely. Intuitively the classical nonparametric deconvolution is able to blend the effects of the covariates in the distribution of $\mathbf{X}$ by itself, therefore, we should be able to recover the results by the method in this section when the distribution for ξ is estimated nonparametrically. Guided by Section S.1.2, if one decides to adopt a parametric approach in (4) while completely ignoring the covariates $\mathbf{Z}$ as opposed to the approach in S.1.3 while estimating F_ξ parametrically, in both cases there are likely to be misspecifications, however, we do benefit from the second approach which prevents from a too wild parametric model.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Maximum Possible
Total Fruit	1.57	4.16	3.94	4.28	5
Whole Fruit	1.40	4.62	4.28	4.53	5
Total Grains	4.66	4.57	4.95	4.93	5
Whole Grains	1.05	1.17	2.13	2.34	5
Total Vegetables	3.78	4.14	4.19	4.74	5
DOL	1.29	1.59	1.96	2.77	5
Milk	5.02	5.08	5.75	5.62	10
Meat & Beans	9.85	9.90	7.67	9.88	10
Oil	5.72	6.56	6.24	5.77	10
Saturated Fat	4.51	5.42	8.26	8.34	10
Sodium	1.97	2.59	2.55	1.19	10
SoFAAS	7.39	10.12	11.52	15.44	20
Total Score	48.21	59.92	63.44	69.82	100

Table S.1: Cluster mean scores for the analysis of Section 3.2 with $K = 4$ clusters, and their maximum possible values. The Total Score is the sum of the cluster means. The cluster "sizes", i.e., the sum of the probabilities of being in each cluster, are 81,101, 85,596 and 25,384 and 101,533, respectively.

Subject No.	Cluster 1	Cluster 2	Cluster 3
1	0.00	0.00	1.00
2	0.10	0.75	0.15
3	0.02	0.06	0.92
4	0.05	0.94	0.01
5	0.07	0.02	0.91
6	0.38	0.25	0.37
7	0.46	0.54	0.00
8	0.00	0.26	0.74
9	0.99	0.01	0.00
10	0.00	0.71	0.29

Table S.2: Analysis of the EATS data of Section 3.3. For the first 10 subjects, displayed are the cluster probabilities for 3 clusters. Observe that the cluster assignment for subject 6, for example, is not at all clear.

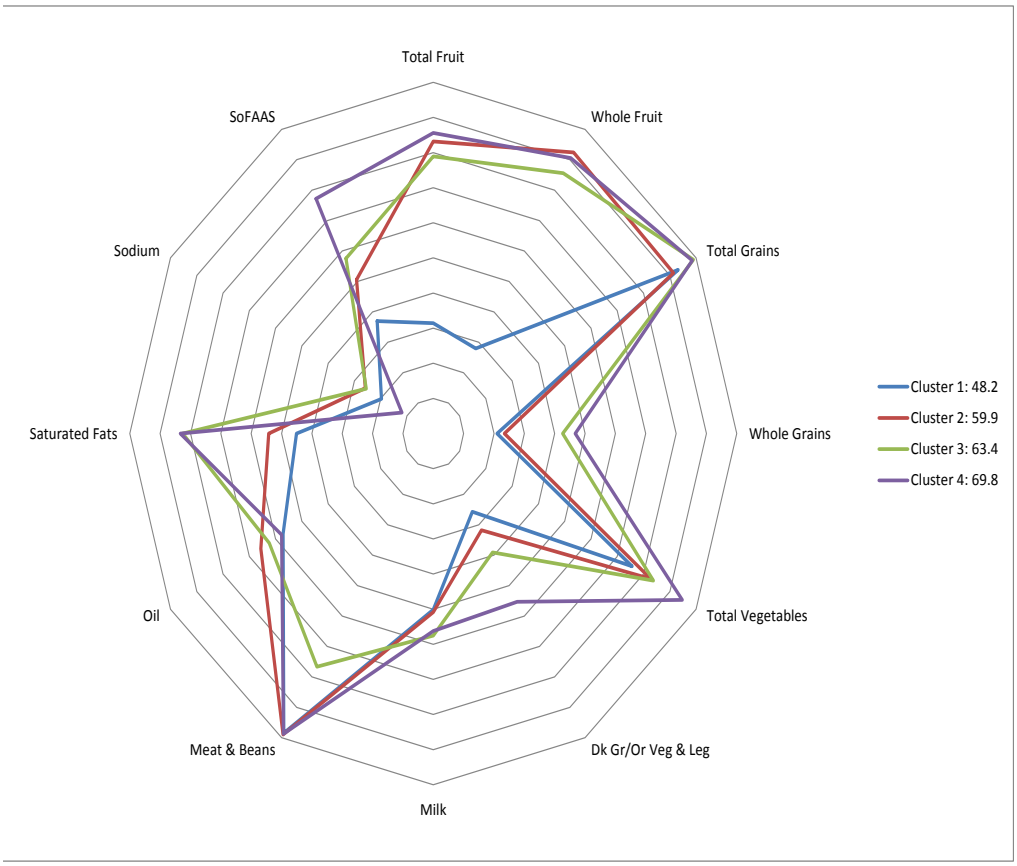

Figure S.1: Data analysis for the HEI-2005 analysis in Section 3.2. Radar plot for the usual, measurement error corrected, intake among men: clustering is based on the HEI-2005 total score. The listed amounts for the clusters are the means of the HEI-2005 total score within the clusters, although the total score was not part of the clustering algorithm. The cluster sizes are 81,101, 85,596, 25,384 and 101,533, respectively.

Bibliography

Ben-David, S., Von Luxburg, U., and Pál, D. (2006). A sober look at clustering stability. In *International Conference on Computational Learning Theory*, pages 5–19. Springer.

Dudley, R. M. (1999). *Uniform Central Limit Theorems.* Cambridge Studies in Advanced Mathematics: 63. Cambridge University Press.

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

Rakhlin, A. and Caponnetto, A. (2006). Stability of k-means clustering. In *Advances in Neural Information Processing Systems*, pages 1121–1128.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.