Discussion of "How to Find an Appropriate Clustering for Mixed Type Variables with Application to Socio-Economic Stratification," by Hennig, C. and Liao, T. F.

de Carvalho, M., and Page, G. L. (2013), Discussion of "How to Find an Appropriate Clustering for Mixed Type Variables with Application to Socio-Economic Stratification," by Hennig, C. & Liao, T. F. In Journal of the Royal Statistical Society, Ser. C, 62, 343–344.

Discussion of "How to Find an Appropriate Clustering for Mixed Type Variables with Application to Socio-Economic Stratification" by Christian Hennig and Tim F. Liao

J. R. Statist. Soc. C, Vol. **62**, Issue 3 (2013, to appear)

Miguel de Carvalho (Pontificia Universidad Católica de Chile, Universidade Nova de Lisboa) **and Garritt L. Page** (Pontificia Universidad Católica de Chile).

We congratulate the authors for a stimulating paper on principles concerning applied statistical modeling for clustering. Interpretation is certainly an important step in our investigations, and we often see it as *the* ultimate step of a data analysis (Cox and Donnelly, 2011, §1.2). This paper encourages our Society to reflect on the problems arising in data-partition analyses (e.g., covariate/cluster method selection), when these are not suitably supplemented with interpretation and subject-matter knowledge.

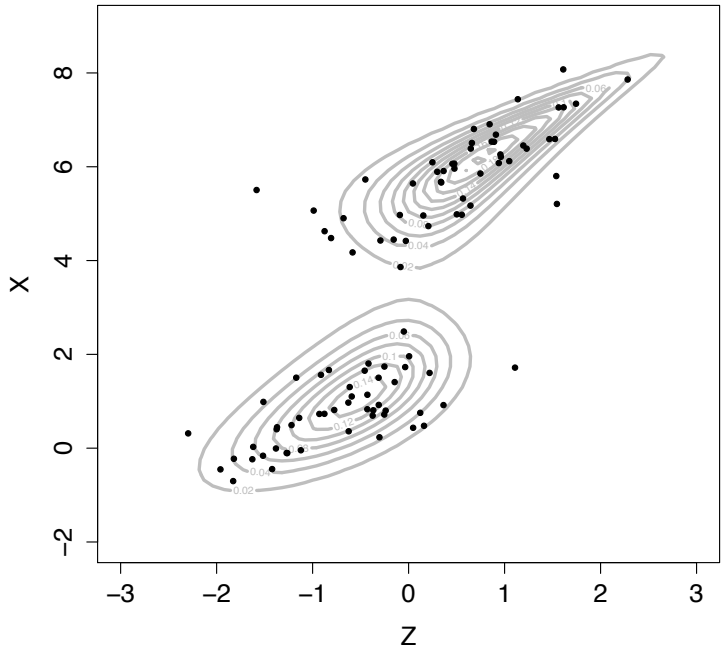

Figure 1: Data generated from a Gumbel copula; the marginal for Z is a standard normal, and the marginal for X is a mixture of $N(1,1)$ and $N(6,1)$ ($\pi_1 = \pi_2 = 1/2$).

We focus on discussing a simple setup related to the appearance of 'spurious' clusters due to (inadequate) data preprocessing, as in Fig. 3 (c) of the paper, with thoughts being illustrated using simulated data. We suppose that there exists a latent variable Z with distribution function

$$F_Z(\cdot) = \sum_{k=1}^{K} \pi_k F(\,\cdot\,; \theta_k), \tag{1}$$

whose mixture components define the 'meaningful' K clusters the researcher expects to see. The challenge is on using the data $\{X_i\}_{i=1}^n \sim F_X$ to learn about Z. Here $\pi_k \in (0,1)$, $\sum_{k=1}^K \pi_k = 1$, and $\{F(\,\cdot\,;\theta) : \theta \in \Theta\}$ denotes a parametric family indexed on a parameter space Θ; more complex sampling schemes could have been used for Z (e.g. Booth *et al.*, 2008, eq. 2), but (1) suffices for our purposes. We assume that the dependence between X and Z is described through an unknown copula function $C\{F_X(u), F_Z(v)\} = F_{X,Z}(u,v)$, for $(u,v) \in [0,1]^2$, where $F_{X,Z}$ denotes the joint distribution function. In practice Z cannot be directly measured and therefore X (which is typically highly correlated with Z) is used as a proxy. However, we often forget that X may not be as informative about Z as one might hope (e.g., when Z is happiness and X income), and preprocessing is used to suitably tilt the distribution of X so that it becomes more similar to that of Z.

In §6.1 the authors provide scientifically relevant arguments why the zero savings group of Fig. 3 (c) fails to be meaningful, and thus motivating the need to employ a somewhat arbitrary $c = 50$. Additionally, a naive application of a pattern recognition technique could lead to spurious clustering—a pattern on X without any correspondent on Z. To illustrate the appearance of such spurious clusters in our setup, consider Fig. 1 which displays 100 points simulated according to a Gumbel copula $C_\psi(p,q) = \exp[-\{(-\log p)^\psi + (-\log q)^\psi\}^{1/\psi}]$, for $(p,q) \in [0,1]^2$, with $\psi = 3$. The marginal for Z is a standard normal, and the marginal for X is a mixture of $N(1,1)$ and $N(6,1)$ ($\pi_1 = \pi_2 = 1/2$). This example is certainly artificial—as in practice only $\{X_i\}_{i=1}^n$ would be observed—but it is interesting to observe that a spurious cluster on X may exist, even when Z is strongly correlated with X (Pearson correlation = 0.79).

From a modeling point of view, the paper clearly puts forward the key role that subject-specific interpretations play in helping link X to Z. Since the authors strongly advocate incorporating researcher intuition in clustering (of which we agree), we wonder whether the Bayesian paradigm should play a more active role in the proposed 'clustering philosophy.' Particularly, product partition models have been recently devised for assessing uncertainty about the configuration of the clusters (Müller *et al.*, 2008). These methods are able to incorporate uncertainty associated with *a priori* 'expected' data partitions via a prior distribution assigned to the cluster configuration. The Bayesian approach would also seem natural for a less debatable choice of c in the preprocessing stage, or for the specification of a prior distribution on the structure of dependence between X and Z.

References

Booth, J. G. and Casella, G. (2008). Clustering using objective functions and stochastic search. *J. R. Statist. Soc.* B, **70**, 119–139.

Cox, D. R. and Donnelly, C. A. (2011). *Principles of Applied Statistics.* Cambridge: Cambridge University Press.

Müller, P., Quintana, F. and Rosner, G. L. (2006). A product partition model with regression covariates. *J. Computnl Graph. Statist.*, **20**, 260–278.