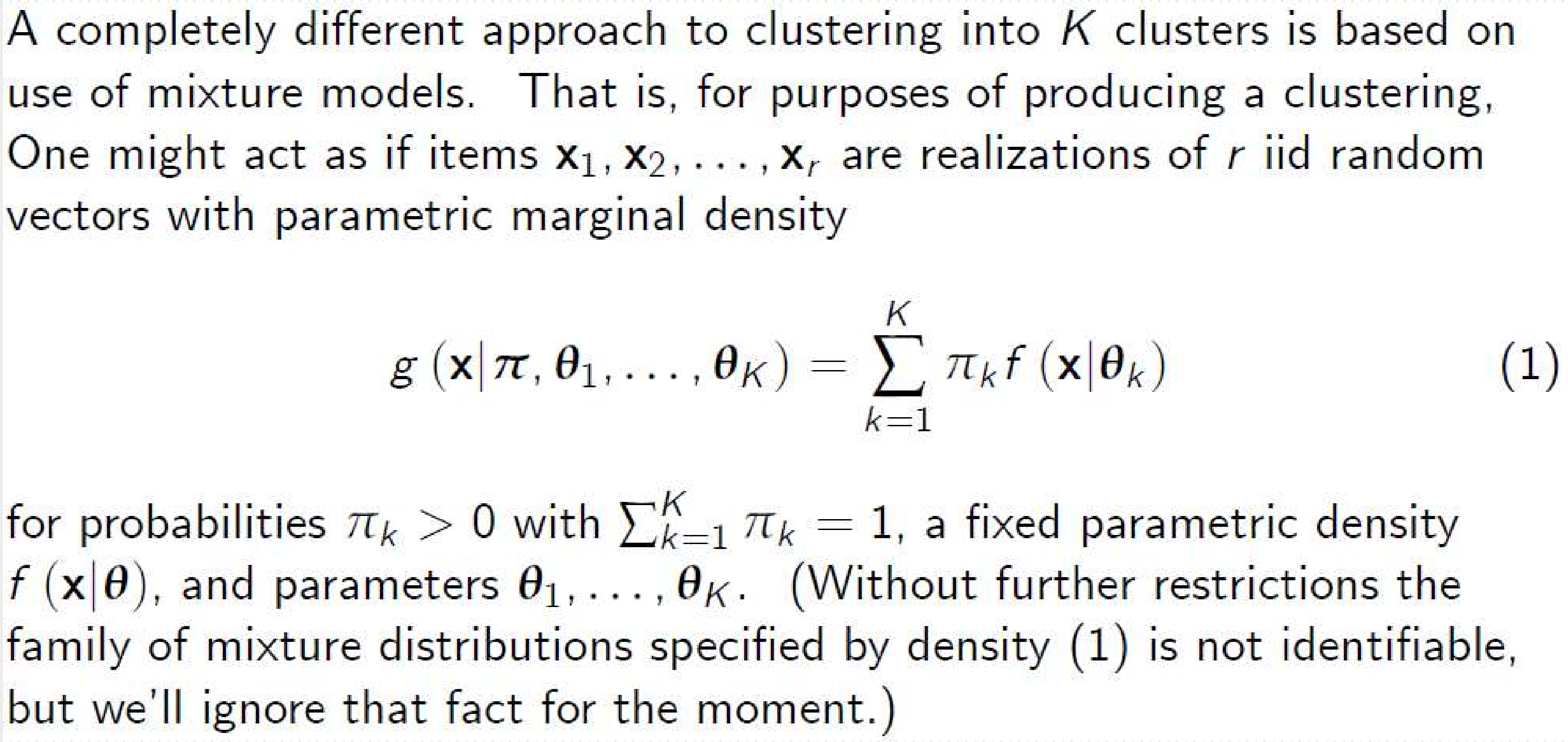
(Mixture) Model-Based Clustering

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Mixture models

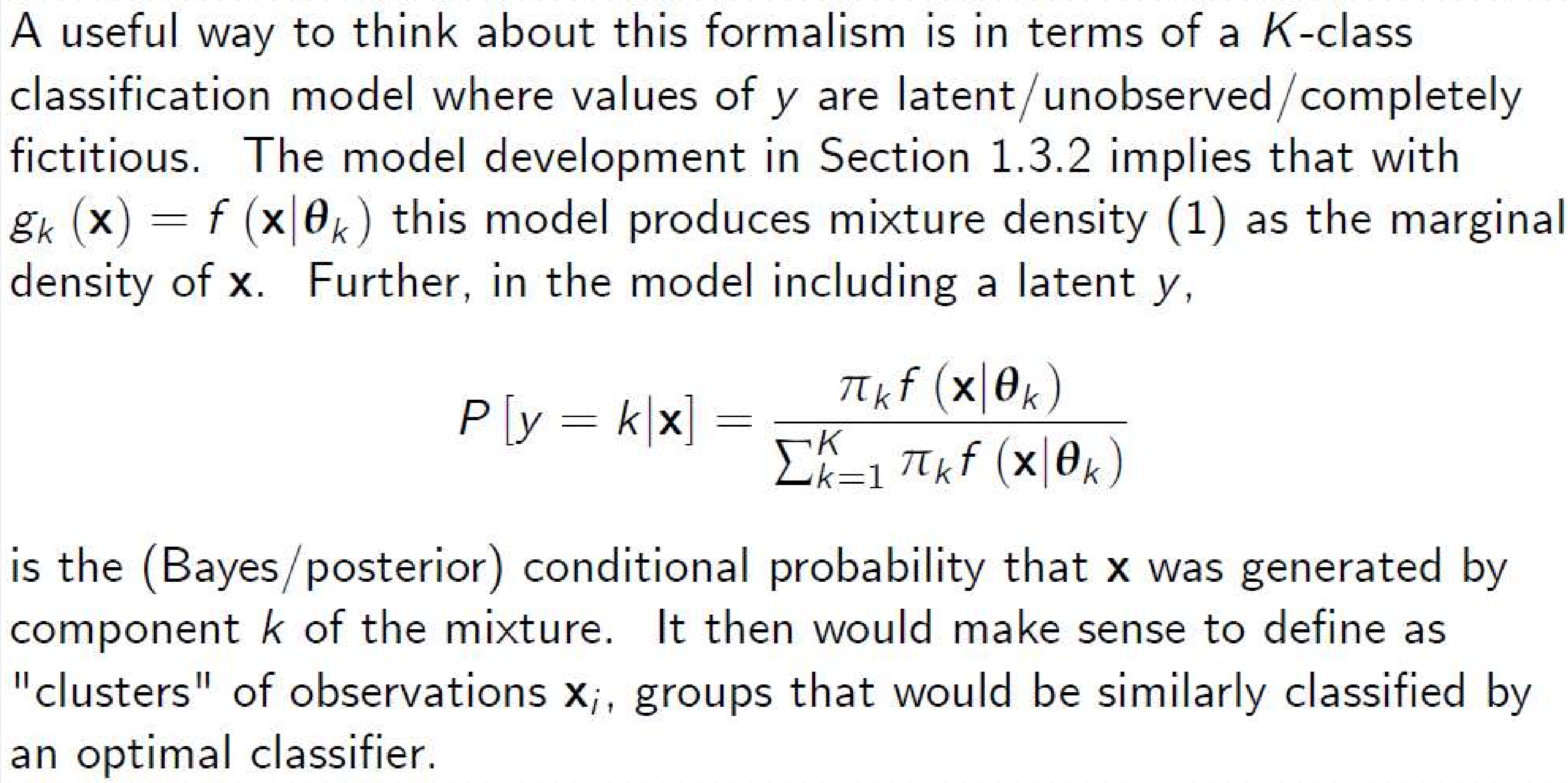


A completely different approach to clustering into K clusters is based on use of mixture models. That is, for purposes of producing a clustering, One might act as if items $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ are realizations of r iid random vectors with parametric marginal density

$$g\left(\mathbf{x} \middle| \boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\right) = \sum_{k=1}^{K} \pi_k f\left(\mathbf{x} \middle| \boldsymbol{\theta}_k\right) \qquad (1)$$

for probabilities $\pi_k > 0$ with $\sum_{k=1}^{K} \pi_k = 1$, a fixed parametric density $f\left(\mathbf{x} \middle| \boldsymbol{\theta}\right)$, and parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. (Without further restrictions the family of mixture distributions specified by density (1) is not identifiable, but we'll ignore that fact for the moment.)

K-class model with latent *y*

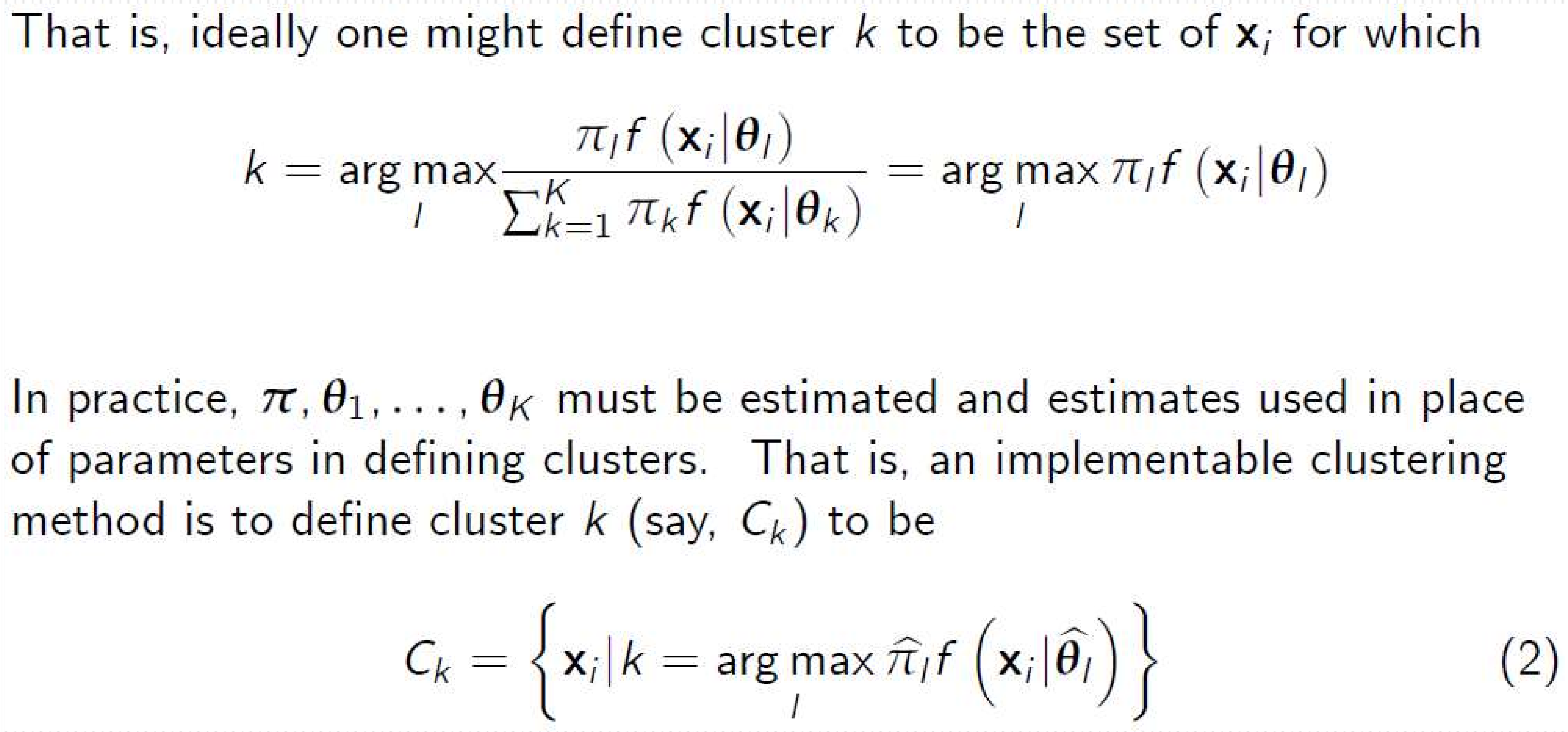


A useful way to think about this formalism is in terms of a K-class classification model where values of y are latent/unobserved/completely fictitious. The model development in Section 1.3.2 implies that with $g_k(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}_k)$ this model produces mixture density (1) as the marginal density of $\mathbf{x}$. Further, in the model including a latent y,

$$P[y = k|\mathbf{x}] = \frac{\pi_k f(\mathbf{x}|\boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k f(\mathbf{x}|\boldsymbol{\theta}_k)}$$

is the (Bayes/posterior) conditional probability that $\mathbf{x}$ was generated by component k of the mixture. It then would make sense to define as "clusters" of observations $\mathbf{x}_i$, groups that would be similarly classified by an optimal classifier.

Clusters from classifiers



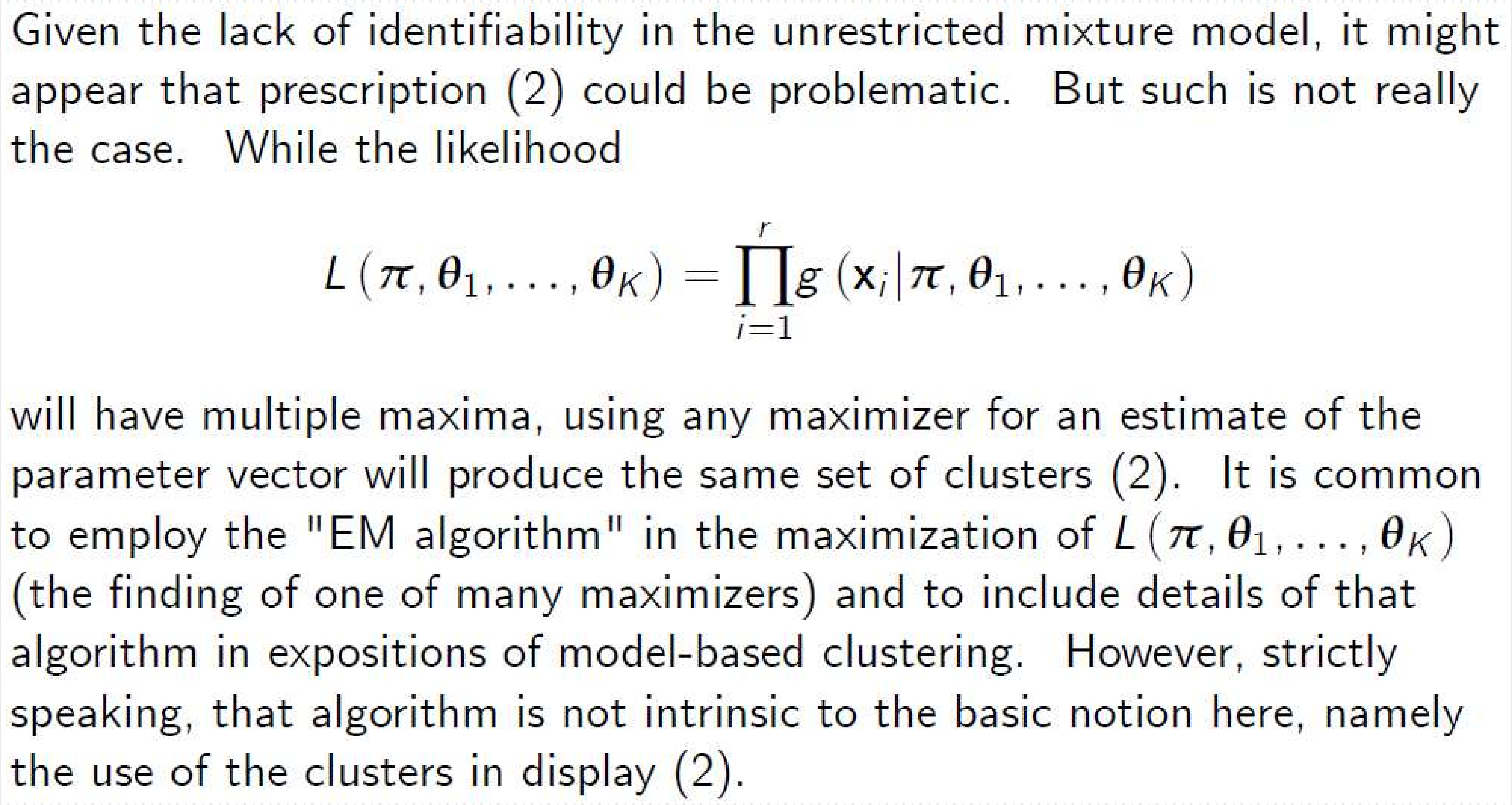
That is, ideally one might define cluster k to be the set of $\mathbf{x}_i$ for which

$$k = \arg\max_{l} \frac{\pi_l f\left(\mathbf{x}_i|\boldsymbol{\theta}_l\right)}{\sum_{k=1}^{K} \pi_k f\left(\mathbf{x}_i|\boldsymbol{\theta}_k\right)} = \arg\max_{l} \pi_l f\left(\mathbf{x}_i|\boldsymbol{\theta}_l\right)$$

In practice, $\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ must be estimated and estimates used in place of parameters in defining clusters. That is, an implementable clustering method is to define cluster k (say, C_k) to be

$$C_k = \left\{\mathbf{x}_i | k = \arg\max_{l} \widehat{\pi}_l f\left(\mathbf{x}_i|\widehat{\boldsymbol{\theta}_l}\right)\right\} \qquad (2)$$

Model fitting



Given the lack of identifiability in the unrestricted mixture model, it might appear that prescription (2) could be problematic. But such is not really the case. While the likelihood

$$L\left(\boldsymbol{\pi},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K\right)=\prod_{i=1}^{r}g\left(\mathbf{x}_i|\boldsymbol{\pi},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K\right)$$

will have multiple maxima, using any maximizer for an estimate of the parameter vector will produce the same set of clusters (2). It is common to employ the "EM algorithm" in the maximization of $L\left(\boldsymbol{\pi},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K\right)$ (the finding of one of many maximizers) and to include details of that algorithm in expositions of model-based clustering. However, strictly speaking, that algorithm is not intrinsic to the basic notion here, namely the use of the clusters in display (2).