PARTIAL MEMBERSHIP MODELS FOR HIGH-DIMENSIONAL SPECTROSCOPY DATA

Alessandro Casa[1], Thomas Brendan Murphy[2] and Michael Fop[2]

[1] Faculty of Economics and Management, Free University of Bozen-Bolzano,
(e-mail: alessandro.casa@unibz.it)

[2] School of Mathematics and Statistics, University College Dublin,
(e-mail: michael.fop@ucd.ie, brendan.murphy@ucd.ie)

ABSTRACT: The demand for detecting food adulteration has recently grown, due to its economic and health implications. Infrared spectroscopy provides an efficient method of collecting data for use in food authenticity analyses. Statistical methods are routinely employed to analyze spectroscopy data in order to effectively detect adulterants in different food items and ensure food authenticity. This work presents a novel partial membership model for mid-infrared spectral data. Our approach not only detects the level of adulteration but also provides information on the spectral regions most affected by the adulterant. These insights can be used in combination with subject-matter expertise to characterize the chemical impact of the adulteration.

KEYWORDS: partial membership, latent variable models, food authentication, shrinkage prior

1 Introduction

Expensive foods are often subject to fraud and food adulteration, with some of the original components being removed or replaced by cheaper alternatives, to lower their prices or to increase their bulk. On one hand, this can represent an economic problem for food producers. On the other hand, it might also lead to health issues for the consumers. Therefore, food authenticity studies, which aim to determine if a sample has been adulterated or not, are increasingly important. In this work, we examine Fourier transform mid-infrared (MIR) spectroscopy data, which have been previously used effectively to tackle the aforementioned problem. To the task, we propose a novel partial membership model for spectroscopy data. The model introduces a more sophisticated authentication tool, capable of not only identifying the presence of potential adulterants in food, but also of determining the percentage of contamination. The proposed model also enables the identification of which wavelengths are more impacted by the adulterant, constituting a starting point for further chemical analysis. In Section 2, we introduce this new model and outline the adopted estimation approach. Section 3 reports an application to spectrometry data of Irish honey samples.

2 Model definition and estimation

Individual-level mixture models generalize standard model-based clustering by encompassing situations where units can belong to multiple groups simultaneously, with varying degrees of membership. This idea has been developed in two directions, namely mixed membership and partial membership models (PMM), with the latter being the focus of this work; see Airoldi *et al.* , 2014 for a discussion. Let $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be the observed data. When $\mathbf{y_i} \in \mathbb{R}^p$, a multivariate Gaussian distribution is often assumed for the K component densities. Therefore, according to PMM, $\mathbf{y}_i$ is conditionally distributed as

$$(\mathbf{y}_i|g_i,\Theta) \sim N_p\left(\left(\sum_{k=1}^{K} g_{ik}\Sigma_k^{-1}\right)^{-1}\left(\sum_{k=1}^{K} g_{ik}\Sigma_k^{-1}\mu_k\right), \left(\sum_{k=1}^{K} g_{ik}\Sigma_k^{-1}\right)^{-1}\right) \tag{1}$$

where $\Theta = \{\mu_k, \Sigma_k\}_{k=1}^{K}$ denotes mixture component means and covariance matrices, while $\mathbf{g}_i = (g_{i1}, \ldots, g_{iK})$ is the partial membership vector for the i-th observation with $g_{ik} \in [0,1]$, for $k = 1, \ldots, K$, and $\sum_k g_{ik} = 1$. For food authentication purposes, we consider $K = 2$, with the two components corresponding to the pure food item and the adulterant, respectively. Moreover, we assume that the adulterant has an additive and wavelength-specific effect. As such, we have that

$$\begin{aligned}\mu_1 &= \mu^{pure} = (\mu_1^{pure}, \ldots, \mu_p^{pure})\\ \mu_2 &= \mu^{ad} \;= (\mu_1^{pure} + \delta_1, \ldots, \mu_p^{pure} + \delta_p)\end{aligned}$$

where δ_j, for $j = 1, \ldots, p$, represents the mean-shift induced by the adulterant on the j-th wavelength. Pairing this specification with shrinkage or penalization strategies for δ_j's can lead to the detection of the spectral regions most influenced by the adulterant. Assuming $\Sigma_1 = \Sigma_2 = \Sigma$, model (1) reads as

$$(\mathbf{y}_i|g_i,\Theta) \sim N_p\left(\mu^{pure} + g_{i2}\delta, \Sigma\right) \tag{2}$$

where g_{i2} is the percentage of adulterant in the i-th sample and $\delta = (\delta_1, \ldots, \delta_p)$. When dealing with spectroscopy data, the high number of variables can jeopardize the practical usefulness of model (2). For this reason, simplifying assumptions would consider a factor analytic or a diagonal structure for Σ.

Two alternative ways to estimate model (2) are explored. The first, heuristic, estimation procedure can be used to obtain a naive and fast first model evaluation. More specifically, it aims to maximize iteratively the following quantity

$$SS_f = \sum_{i=1}^{n} (\mathbf{y}_i - \mu^{pure} - g_{i2}\delta)^2$$

with respect to μ^{pure}, δ and g_{i2}. As it is, this procedure does not account for the correlation structure among wavelengths and does not induce shrinkage on δ. Interestingly, it can be used to provide initial values for a Bayesian estimation procedure adopting a Dirichlet prior distribution for the membership vectors $\mathbf{g}_i, i = 1, \ldots, n$ while, for the δ_j's, $j = 1, \ldots, p$, an horseshoe prior (Carvalho *et al.* , 2010) is employed, thus imposing sparsity on the mean-shifts. Lastly, standard conjugate priors are assumed for μ^{pure}, for the diagonal entries of Σ, or for Λ and Ψ, if a factor analytic structure is considered. The model is estimated via MCMC algorithm, by means of the `NIMBLE` software. Note that some degree of supervision can be introduced in the estimation. In particular, for some spectra, g_{i2} can be assumed known, since it is often possible to augment the observed data with experimental data with a controlled amount of adulteration. Unreported analyses showed the beneficial impact of small amount of supervision.

3 Application to honey data

Our proposal is tested on MIR spectral data comprising samples from pure honey and samples contaminated with different adulterants (Kelly *et al.* , 2006). The data have $n = 410$ spectra, $n_H = 290$ from pure honey and $n_B = 120$ adulterated with beet sucrose in different percentages (10%, 20% and 30%). Prior to running the analysis, a data aggregation step has been performed to reduce the overall computational cost. Consequently, the original $p = 285$ wavelengths have been reduced to $p^* = 57$ aggregated ones. Some supervision has been imposed, assuming prior knowledge of the adulteration level for 40 spectra. A diagonal structure for Σ has been considered and the hyperparameters of the horseshoe prior have been selected following suggestions from Piironen & Vehtari, 2017. An excerpt of the results is reported in Figure 1. Here, it is shown how the proposed method is able to precisely estimate the spectrum for the most adulterated samples, with the 95% credible interval always containing the true observed values. A closer inspection for the estimated δ_j's shows

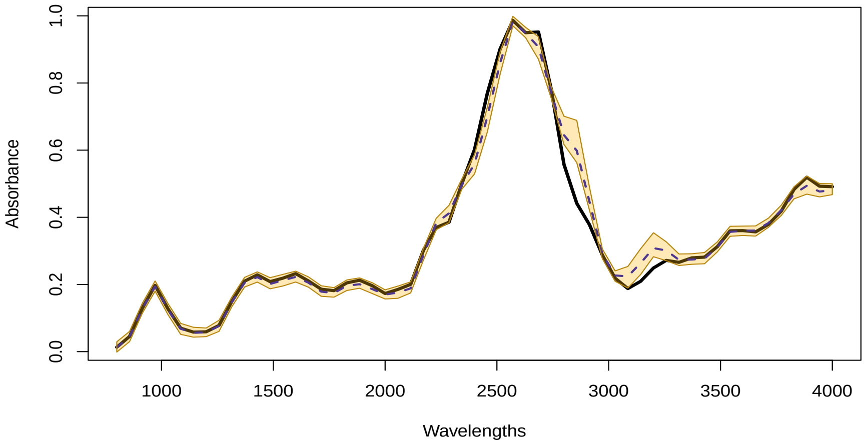

Figure 1. *In black the estimated μ^{pure}. Dashed blue line depicts the observed average spectrum for the most adulterated samples, while the gold shaded area represents the estimated 95% credible interval for the same quantity.*

how beet sucrose seems to have a non negligible impact only on 10 aggregated wavelengths in the region from 2377.46 cm^{-1} to 3166.19 cm^{-1}. These results, if paired with subject-matter knowledge, can shed light on the chemical mechanism underlying the adulteration process.

References

AIROLDI, E.M., BLEI, D., EROSHEVA, E.A., & FIENBERG, S.E. 2014. *Handbook of mixed membership models and their applications*. CRC press.

CARVALHO, C.M., POLSON, N.G., & SCOTT, J.G. 2010. The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

KELLY, J.D., PETISCO, C., & DOWNEY, G. 2006. Application of Fourier transform midinfrared spectroscopy to the discrimination between Irish artisanal honey and such honey adulterated with various sugar syrups. *Journal of Agricultural and Food Chemistry*, **54**(17), 6166–6171.

PIIRONEN, J., & VEHTARI, A. 2017. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *Pages 905–913 of: Artificial Intelligence and Statistics*. PMLR.