PARSIMONIOUS BAYESIAN FACTOR ANALYSIS FOR MODELLING LATENT STRUCTURES IN SPECTROSCOPY DATA

BY ALESSANDRO CASA\textsuperscript{1,3}, TOM F. O’CALLAGHAN\textsuperscript{2} AND THOMAS BRENDAN MURPHY\textsuperscript{1,3}

\textsuperscript{1}School of Mathematics & Statistics, University College Dublin, alessandro.casa@ucd.ie; brendan.murphy@ucd.ie
\textsuperscript{2}School of Food & Nutritional Sciences, University College Cork, tom.ocallaghan@ucc.ie
\textsuperscript{3}Vistamilk SFI Research Centre

In recent years, within the dairy sector, animal diet and management practices have been receiving increased attention, in particular examining the impact of pasture-based feeding strategies on the composition and quality of milk and dairy products, in line with the prevalence of premium grass-fed dairy products appearing on market shelves. To date, methods to thoroughly investigate the more relevant differences induced by the diet on milk chemical features are limited; enhanced statistical tools exploring these differences are required.

Infrared spectroscopy techniques are widely used to collect data on milk samples and to predict milk related traits and characteristics. While these data are routinely used to predict the composition of the macro components of milk, each spectrum also provides a reservoir of unharnessed information about the sample. The accumulation and subsequent interpretation of these data present some challenges due to their high-dimensionality and the relationships amongst the spectral variables.

In this work, directly motivated by a dairy application, we propose a modification of the standard factor analysis to induce a parsimonious summary of spectroscopic data. Our proposal maps the observations into a low-dimensional latent space while simultaneously clustering the observed variables. The method indicates possible redundancies in the data and it helps disentangle the complex relationships among the wavelengths. A flexible Bayesian estimation procedure is proposed for model fitting, providing reasonable values for the number of latent factors and clusters. The method is applied on milk mid-infrared (MIR) spectroscopy data from dairy cows on distinctly different pasture and non-pasture based diets, providing accurate modelling of the correlation, clustering of variables, and information on differences among milk samples from cows on different diets.

1. Introduction. In recent years, the food industry has gone through rapid changes, partially due to ever evolving consumer preferences and increased consumers’ awareness of food and health. We are currently at the forefront of growing demand for detailed and accurate knowledge concerning food quality and security. The sectors potentially most vulnerable to these changing trends are those processing and preparing foodstuffs of animal origin. As a consequence, the dairy sector has been particularly involved in this transition, with an increasing attention towards product quality, traceability and adherence to procedures respectful to animal welfare and the environment.

In this scenario, one aspect which is gaining more attention concerns with cattle feeding regimen. In general, consumers regard pasture-based feeding as more respectful of animal well-being, producing more natural and healthy products (Elgersma, 2012). As such, in many
markets, grass fed dairy products usually demand a premium price, and as often happens with expensive products, milk produced by grass fed cows is susceptible to food adulteration and fraud. Therefore, there has been an increased requirement for milk authentication methods (see Kamal and Karoui, 2015, for a recent review). Such techniques usually rely on some biomarkers (see Capuano et al., 2014, and references therein) and they specifically focus on a proper discrimination between samples coming from animals on different diets. Accordingly, they are not explicitly designed to provide a characterization of the differences implied by distinct feeding regimens, and the impact of the diets on the milk chemical features remains an overlooked problem.

To study if the above-mentioned consumers perceptions have some basis, some papers have demonstrated that outdoor pasture-based feeding produces milk and dairy products having enhanced beneficial nutrients with respect to the ones produced by indoor total mixed ration-based (TMR) regimens (O’Callaghan et al., 2016a,b; O’Callaghan et al., 2017). Furthermore, pasture-based feeding seems to lead to ameliorations in the organoleptic characteristics of dairy products providing signature like traits (Faulkner et al., 2018; Alothman et al., 2019; Garvey et al., 2020) and improved quality. Furthermore, O’Callaghan et al. (2016a, 2018) highlighted the ability of fatty acid profiling coupled with multivariate analysis and Hydrogen-1 nuclear magnetic resonance (1H-NMR) to distinguish between milk from pasture and TMR-based diets. Nonetheless, these approaches rely on expensive and time-consuming techniques since they require laboratory extraction routines to collect the data, compromising their widespread effective utility.

On the other hand vibrational spectroscopy techniques, such as Fourier transform near-infrared (NIR) and mid-infrared (MIR) spectroscopy, are cheap, rapid and non-disruptive alternatives to collect large amounts of data and to analyze different biological materials. When a material is analyzed via MIR spectroscopy, the light is passed through a sample of that material at a sequence of wavelengths in the mid-infrared region (900 to 5000 cm$^{-1}$). The passage of the light activates the sample’s chemical bonds leading to an absorption of energy from the light. The amount of energy absorbed or transmitted by the sample, at different wavelengths, creates the spectrum of that sample that might be subsequently used to analyze its characteristics (see Figure 1 for a graphical illustration of some MIR spectra). Such methods have been already proven useful to analyze a wide range of different foodstuffs (see e.g., Downey, 1996; Reid, O’Donnell and Downey, 2006; Murphy, Dean and Raftery, 2010). Moreover, they have been exploited in the dairy framework to determine milk characteristics such as protein, lactose and casein concentration (De Marchi et al., 2014) as well as to predict milk and animal related traits such as milk fatty acids (Bonfatti et al., 2017) and energy efficiency and intake (McParland et al., 2014; McParland and Berry, 2016). Recently they have been proved useful also to perform authentication, providing good results when considered to discriminate between different dietary treatments (Frizzarin et al., 2021). However, the usefulness of infrared spectroscopy data to obtain insights about the chemical impact of feeding regimens on milk characteristics, has been less widely explored.

Therefore, we need a thorough exploration of the features of spectroscopic data when used to analyze samples coming from differently fed animals. NIR and MIR spectroscopy data introduce statistical challenges that need careful attention. The first one is concerned with their high-dimensionality since, usually, each single observation consists of more than 1000 transmittance or absorbance values over the MIR or NIR regions. Spectral data are highly correlated with underlying chemical processes entailing rather complex correlation structures, as can be seen in Figure 2. While adjacent wavelengths are obviously highly correlated, strong relationships are also observed among distant regions of the spectrum. The information contained in each spectrum is indeed known to be structured in a rather complicated way and possibly spread over different locations.
Statistical methodologies can provide a parsimonious and useful representation of the correlation structures in spectroscopy data. First, by summarizing the information into lower dimensional spaces, they can mitigate high-dimensionality related issues. Second, a proper reconstruction of the relations seen in Figure 2 may help in identifying which spectral regions are carrying similar information, when the aim is to obtain a characterization of the main features of milk samples coming from cows on different feeding strategies. This identification, when coupled with subject-matter knowledge, can highlight which chemical structures are responsible for the structures seen in the milk sample spectra. Moreover, it can be useful to identify the main nutritional differences implied in the milk from different diet regimens, serving as a stepping stone for classification purposes.

In this framework, techniques such as Partial Least Squares (PLS) and principal component analysis (PCA) are widely used both for predictive purposes and to reduce the dimensionality of the data, by summarizing the information in a smaller number of newly built features. In a similar fashion, factor analysis (FA, Everitt, 1984; Bartholomew, Knott and Moustaki, 2011) provides a parsimonious representation of the observed data, by building new variables called factors, while simultaneously explaining the correlation among high-dimensional observations. For this reason, when the aim is to reconstruct structures as the ones in Figure 2, factor analysis represents a suitable strategy to follow. Nonetheless, even if FA effectively reduces the dimensionality of the data, standard FA does not provide information about possible redundancies in the observed features. In this work, we propose a suitable modification of the standard factor analysis model which allows the detection of redundant variables and which produces a partition of the variables themselves, thus possibly gaining useful insights about similarly behaving spectral regions.

In Section 2 we describe the mid-infrared spectroscopy data which motivates our proposal. In Section 3 we outline our proposal with a specific focus on the proposed Bayesian estimation procedure and on the involved model selection steps. Some analyses on synthetic datasets are reported in Section 4, while in Section 5 we present the results obtained on the milk spec-
troscopy data. Finally, in Section 6, we conclude with some final remarks and highlight some advantageous avenues for future research.

2. Dairy diet MIR spectroscopy data. The data we consider in this study have been collected during an experiment designed by O’Callaghan et al. (2016a), which represent the first study of its kind in Ireland, and probably the world; for further information on the experimental setting see O’Callaghan et al. (2016b), O’Callaghan et al. (2017), O’Callaghan et al. (2018) and Faulkner et al. (2018). The dataset consists of a collection of mid-infrared spectra from 4320 milk samples produced from cattle on three dietary treatments over a three year period on the Teagasc Moorepark Dairy research Farm (Fermoy, Co. Cork, Ireland). The data are comprised of spectra extracted from morning (am) and evening (pm) milk samples collected weekly from Holstein-Freisian cows using a ProFoss FT6000 series instruments (FOSS, Ireland) between the period of May and August in 2015, 2016 and 2017. In each year 54 cows were randomly assigned to each of the dietary treatment for the entire lactation period of that year. Treatments included grass (GRS) which consisted of cows maintained outdoors on a perennial ryegrass sward only, clover (CLV) whereby cows were maintained outdoors on a perennial ryegrass with 20% white clover sward only, and total mixed ration (TMR) where cows were maintained indoors year round and nutrients are combined together in a single nutritional mix consisting of grass silage, maize silage and concentrates. More specifically, 2931 samples come from cows being fed with pasture (GRS and CLV), while the remaining 1389 come from cows being fed with TMR. Even if the original milk samples may be grouped into three different classes (GRS, CLV and TMR), in our work the first two classes have been merged together into a general pasture-based diet group, because of their strong similarities from a compositional perspective.

The experiment has 120 cows, with a mean number of 36 samples per cow. The samples have been collected following a yearly balanced scheme and they represent a balance of different parities. The samples considered in this work have been restricted to the ones being collected mainly in the summer months, since they represent a period of milk production with highest prevalence of grass growth. Note that, for each sample, a spectrum consists of 1060 transmittance measurements in the region going from 925 cm$^{-1}$ to 5010 cm$^{-1}$. We also have additional information on fat, protein and lactose content in the available milk samples using channels on the FT6000 calibrated against wet chemistry results.
3. Factor analysis with redundant variables.

3.1. Framework. Standard Factor Analysis (denoted as FA in the following) provides a convenient and parsimonious representation of the dependence structure among high-dimensional observations by mapping them in a low-dimensional latent space. Let \( X = \{x_1, \ldots, x_n\} \), with \( x_i \in \mathbb{R}^p \), the set of the observed data. Factor analysis models each observation \( x_i \) as a linear combination of latent variables, called factors, as follows

\[
x_i = \mu + \Lambda u_i + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \( \mu \in \mathbb{R}^p \) is the mean vector, \( \Lambda = \{\lambda_{jk}\}_{j=1, \ldots, p; k=1, \ldots, K} \) is a \( p \times K \) factor loadings matrix with \( K \) being the number of factors, \( u_i \in \mathbb{R}^K \) denotes the factor scores vector while \( \varepsilon_i \sim \mathcal{N}_p(0, \Psi) \) is an error term where \( \Psi = \text{diag}(\psi_1, \ldots, \psi_p) \), with \( \psi_j \)'s often referred to as the uniquenesses. Without loss of generality, we assume in the following that the data are centered (hence \( \mu = 0 \)) and the latent factors are normally distributed with zero mean and identity covariance matrix. Consequently, we have that \( (x_i|u_i) \sim \mathcal{N}_p(\Lambda u_i, \Psi) \). Marginally \( x_i \) is distributed according to a Gaussian distribution with zero mean and covariance matrix

\[
\Sigma = \Lambda \Lambda^T + \Psi.
\]

In practical applications the number of variables \( p \) is considerably higher than the number of factors \( K \). Therefore the decomposition in (2) introduces a convenient and parsimonious representation of the relationships among the observed features in high-dimensional settings.

From (2), and recalling that \( \Psi \) is a diagonal matrix, it follows, in standard FA, that the correlation between the original variables is modelled via the loading matrix \( \Lambda \). Therefore, in the literature, the attention has been focused on how to model and estimate \( \Lambda \) appropriately. In recent years a lot of different solutions have been proposed, both from a frequentist (see e.g. Hirose and Konishi, 2012; Hirose and Yamamoto, 2015) and from a Bayesian (see e.g. Bhattacharya and Dunson, 2011; Ročková and George, 2016; Legramanti, Durante and Dunson, 2020) standpoint, to obtain sparse estimates of the loading matrix. Setting some values of \( \Lambda \) exactly equal to zero allows an even more parsimonious representation of the original covariance structure. This could be convenient from an interpretative point of view, since relating each factor to a smaller number of observed variables helps to give meaning to the factors themselves. Lastly, note that if all the elements in a row of the loading matrix are equal to zero, the corresponding variable is uncorrelated with all the others and essentially represented as noise.

In recent years FA has been thoroughly studied and several modifications have been proposed to enlarge its applicability to different scenarios. The above-mentioned ideas and sparsity inducing estimation schemes have been exploited to model high-dimensional and complex structured data arising in many distinct contexts (see e.g. Pournara and Wernisch, 2007; Knowles and Ghahramani, 2007; Carvalho et al., 2008; Blum et al., 2010; Palla, Ghahramani and Knowles, 2012; Runcie and Mukherjee, 2013; Gao, Brown and Engelhardt, 2013). With a biomedical related application in mind, De Vito et al. (2018, 2019) proposed multi-study FA methods using both Bayesian and frequentist approaches. Group Factor Analysis (GFA) has been proposed to encompass scenarios where coupled observations for the same sample are available (see e.g. Klami et al., 2014; Zhao et al., 2016). Finally note that Ferrari and Dunson (2021) recently proposed the so called Factor Analysis for Interactions (FIN) to accommodate interactions in a Bayesian latent factor regression context.
3.2. Model specification. Several modifications of standard FA have been proposed to encompass high-dimensional situations to be faced in many different applications. Feature selection is often conducted in the FA framework by inducing sparsity on the elements of $\Lambda$. By contrast, methods for detecting features redundancy are far less common. A variable is defined as redundant when it carries information similar to the one provided by another variable (or variables), usually due to the strong correlation between them. The effective detection of redundancies, which is a challenging task, can lead to more parsimonious modelling. As illustrated in Figure 2, redundancy can be a complex issue when analyzing spectroscopy data. To properly account for it, here we introduce a model in which some of the variables are mapped into the latent space by means of the same loading coefficients thus giving an indication of possible grouping structures in the observed features themselves. The proposed model is then defined as follows

\[ x_i = Z\Lambda_c u_i + \varepsilon_i \]

for $i = 1, \ldots, n$, with $x_i, u_i$ and $\varepsilon_i$ previously defined while $Z = \{z_j\}_{j=1}^{p}$, with $z_j = (z_{j1}, \ldots, z_{jG})$, is a $p \times G$ latent allocation matrix, where $G$ is the number of variable clusters. Here the standard binary partition is adopted for $Z$; therefore $z_{jg} = 1$ if the $j$-th variable belongs to the $g$-th group and 0 otherwise. Lastly, $\Lambda_c = \{\Lambda_{c,g}\}_{g=1}^{G}$, with $\Lambda_{c,g} = (\lambda_{c,g1}, \ldots, \lambda_{c,gK})$, is a $G \times K$ matrix whose $g$-th row contains the unique and representative loading values for the $g$-th variable cluster. As a consequence of the specification of the model (3), $\Lambda$ has duplicate row values. We believe this constitutes a sensible way to account for redundancy in the observed features, by constraining the relations with the latent factors to be equal for those variables belonging to the same cluster.

The distributional properties highlighted above for the standard FA model are still valid. Here we have that $(x_i|u_i, z) \sim N_p(\tilde{\Lambda}u_i, \Psi)$ while $(x_i|z) \sim N_p(0, \tilde{\Sigma})$ where $\tilde{\Sigma} = \tilde{\Lambda}\tilde{\Lambda}^T + \Psi$. Therefore, the proposed model induces an even more parsimonious decomposition of the covariance matrix. In fact, the specification of our model entails a possibly drastic reduction in the total number of covariance parameters to estimate; for model (3) this number is equal to $(G \times k) + p$, for model (1) it is equal to $(p \times k) + p$; due to rotational invariance in FA, the number of identifiable parameters is fewer than this. Clearly, the smaller the number of variable clusters $G$, hence the more redundancy is observed in the data, the greater will be the reduction.

From an interpretative point of view the estimation of the allocation matrix $Z$ allows obtaining a clustering of the variables. This partition gives insights into the redundancy phenomenon by highlighting which variables are strongly correlated, hence providing similar information.

Finally, our proposal may be adapted to detect both redundant and uninformative variables by a priori forcing all the elements in a single specific row of $\Lambda_c$ to be exactly equal to zero. This would imply that all the variables assigned to the corresponding group are modelled as noise and are uncorrelated with all the other variables.

3.3. Likelihood and prior specification. Under the specification of model (3), and recalling that $(x_i|u_i, z) \sim N_p(Z\Lambda_c u_i, \Psi)$, the corresponding likelihood function is given by

\[
\mathcal{L}(X|\Lambda_c, \Psi, Z, U) = \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |\Psi|^{-\frac{1}{2}} \exp \left\{-\frac{1}{2} (x_i - Z\Lambda_c u_i)^T \Psi^{-1} (x_i - Z\Lambda_c u_i) \right\}
\]

\[
\propto |\Psi|^{-\frac{1}{2}} \exp \left\{-\frac{1}{2} \text{tr} \left[ \Psi^{-1} (X - U\Lambda_c^T Z^T)^T (X - U\Lambda_c^T Z^T) \right] \right\}
\]

(4)
where \( U = \{ u_i \}_{i=1,...,n} \), with \( u_i = (u_{i1}, \ldots, u_{iK}) \), is the \( n \times K \) factor scores matrix while \( X, \Lambda_c, Z \) and \( \Psi \) are defined as in the previous section. Lastly \(|A|\) and \( \text{tr}[A] \) denotes respectively the determinant and the trace of a generic matrix \( A \).

Different strategies might be adopted to estimate the parameters involved in (3). From a frequentist perspective, the maximum likelihood estimates are usually obtained via iterative algorithms such as the ones proposed in Jöreskog (1967), Jennrich and Robinson (1969) and Rubin and Thayer (1982). Conversely, in this work we adopt a Bayesian approach to factor analysis estimation (see e.g. Press and Shigemasu, 1989; Arminger and Muthén, 1998; Song and Lee, 2001). More specifically, we assume independent prior distributions for the model parameters as follows

\[
\begin{align*}
(5) & \quad \Lambda_{c,g} \sim \mathcal{N}_K(0, \sigma^2_{\Lambda} I_K) & & \text{for } g = 1, \ldots, G \\
(6) & \quad u_i \sim \mathcal{N}_K(0, I_K) & & \text{for } i = 1, \ldots, n \\
(7) & \quad \psi_j \sim \mathcal{IG}(\alpha, \beta_j) & & \text{for } j = 1, \ldots, p \\
(8) & \quad Z \sim \text{PPM}(\alpha_z)
\end{align*}
\]

The choice of the hyperparameters for the inverse gamma prior on the uniquenesses is guided by the suggestions in Frühwirth-Schnatter and Lopes (2010); here the authors avoid encountering in the Heywood problem by choosing \( \alpha \) and \( \beta_j \) so that \( \psi_j \) tends to be bounded away from zero. More specifically in our analyses we set \( \alpha = 2.5 \) and \( \beta_j = (\alpha - 1)/S_{jj}^{-1} \) where \( S^{-1} \) represents the inverse of the sample covariance matrix. A subjectively large value for \( \sigma^2_{\Lambda} \) might be chosen to consider an uninformative prior for the rows of \( \Lambda_c \).

Some words of caution are required for the prior in (8). Let \( c \) be a clustering of indices \( \{1, \ldots, p\} \); even if different representations might be possible, we consider \( c = \{C_1, \ldots, C_G\} \) as a collection of disjoint subsets such that \( C_g \) contain all the indices of the variables belonging to cluster \( g \)-th. A product partition model (PPM, Hartigan, 1990; Barry and Hartigan, 1992) assumes that the prior probability for \( c \) is expressed as follows

\[
\pi(c = \{C_1, \ldots, C_G\}) \propto \prod_{g=1}^G \rho(C_g)
\]

where \( \rho(\cdot) \) is known as the cohesion function. Since we have a one-to-one correspondence between the representation of the partition \( c \) as a collection of blocks and the one via the allocation matrix \( Z \), with a slight abuse of notation, we specify the prior as in (8) even in our framework. Several different specifications for \( \rho(\cdot) \) have been proposed in literature: we consider \( \pi(c) \propto \alpha_z^G \prod_{g=1}^G (|C_g| - 1)^! \), where \(|C_g|\) denotes the cardinality of the \( g \)-th cluster, sharing strong connections with the Dirichlet process that is widely used in the Bayesian clustering framework (Quintana and Iglesias, 2003). In our analyses we set \( \alpha_z = 1 \) as standard; however, this choice did not seem to be influential. When specifying a prior over the set of the partitions, another reasonable approach to take within our framework would consist of borrowing ideas from the Bayesian spatial clustering literature or to consider some additional information such as distances between the objects to be clustered (see e.g. Blei and Frazier, 2011; Page et al., 2016; Dahl, Day and Tsai, 2017; Wehrhahn et al., 2020, and references therein). In such a way, contiguous groups of wavelengths would be favored, leading to some advantages in the modelling process for some specific applications.

Given the likelihood function in (4) and the specification of the priors outlined above, the posterior distribution is defined as follows

\[
\pi(\Lambda_c, \Psi, Z, U|X) = \mathcal{L}(X|\Lambda_c, \Psi, Z, U)\pi(\Lambda_c|\sigma^2_{\Lambda})\pi(\Psi|\alpha, \beta_j)\pi(Z|\alpha_z)\pi(U)
\]
5. Denote with \( Z \)

4. Select randomly some probability issues related to the rotational invariance property. A common solution considers some reallocation of blocks of variables with reasonable sizes.

The acceptance ratio by considering moves involving similar clusters and by proposing the gabilities involved in the selection of \( Z \) is sampled via a Metropolis-Hastings step. The full conditional distributions are listed below (we refer to the supplementary material for more details on their derivation)

\[
\begin{align*}
\text{vec}(\Lambda_c) \ldots & \sim \mathcal{N}_{G \times K}(\mu_\lambda, \Sigma_\lambda) \\
|u_i| \ldots & \sim \mathcal{N}_K(\mu_u, \Sigma_u) \\
|\psi_j| \ldots & \sim IG(\alpha + n/2, \beta_j^*)
\end{align*}
\]

For the Metropolis-Hastings step to sample the allocation matrix \( Z \), we adapt to our case one of the moves proposed by Nobile and Fearnside (2007) in the so called allocation sampler. Each single move attempts to reallocate to cluster \( g_2 \) a group of variables previously assigned to cluster \( g_1 \); in this way, by possibly reallocating blocks of variables, big moves are proposed so that the space will be explored faster. The detailed steps of the procedure are outlined hereafter:

1. Draw, from the total \( G \) variable clusters, a group \( g_1 \). If \( n_{g_1} = |C_{g_1}| = 0 \) the move fails;
2. Compute \( d(\Lambda_{c,g_1}, \Lambda_{c,g'}) \), the Euclidean distance between \( \Lambda_{c,g_1} \) and \( \Lambda_{c,g'} \), \( \forall g' \in \{1, \ldots, G\} \) with \( g' \neq g_1 \). Afterwards a second group \( g_2 \) is drawn from the set \( \{1, \ldots, G\} \setminus g_1 \) with \( P(g' = g_2) \propto d(\Lambda_{c,g_1}, \Lambda_{c,g'})^{-1} \);
3. Draw \( M \) from the set \( \{1, \ldots, n_{g_1}\} \) with \( P(M = m) \propto 1/m, \forall m = 1, \ldots, n_{g_1} \);
4. Select randomly \( M \) observations among the \( n_{g_1} \) belonging to group \( g_1 \) and reallocate them to group \( g_2 \);
5. Denote with \( Z \) and \( Z' \) respectively the starting latent allocation matrix and the one after the reallocation move. The move is then accepted with probability \( \min\{1, R\} \), where \( R \) is given by

\[
R = \frac{\pi(\Lambda_c, \Psi, Z', U | X) \; P(Z' \to Z)}{\pi(\Lambda_c, \Psi, Z, U | X) \; P(Z \to Z')}
\]

It can be shown that the proposal ratio is

\[
\frac{P(Z' \to Z)}{P(Z \to Z')} = \frac{\sum_{m=1}^{n_{g_1}} \frac{1}{m} \; n_{g_1}!n_{g_2}!}{\sum_{m=1}^{n_{g_2}+M} \frac{1}{m} \; (n_{g_1} - M)!(n_{g_2} + M)!}.
\]

We modify the procedure proposed by Nobile and Fearnside (2007) by changing the probabilities involved in the selection of \( g_2 \) and \( M \). In fact, the modification aims to increase the acceptance ratio by considering moves involving similar clusters and by proposing the reallocation of blocks of variables with reasonable sizes.

The model we are proposing, having a factor analytic structure, inherits standard identifiability issues related to the rotational invariance property. A common solution considers some
constraints on the factor loadings (see e.g., Arminger and Muthén, 1998; Lopes and West, 2004). In our framework, where the factor analytic structure of the model may be seen as a tool to reconstruct $\Sigma$ in a parsimonious way, identification is not strictly necessary. Furthermore, Bhattacharya and Dunson (2011) pointed out that identifiability constraints may lead to order dependence among the variables and general inefficiencies. As a consequence, we decided not to consider such constraints in our modelling strategy.

3.5. Model selection. In the previous sections, the number of factors $K$ has been considered as fixed; in practice inference on $K$ constitutes one of the most challenging issues when considering factor analytic models. A standard approach resorts to information criteria as selection tools; nonetheless they might not be reliable in high-dimensional settings. From a Bayesian standpoint Lopes and West (2004) proposed a reversible jump Markov Chain Monte-Carlo algorithm. In a nonparametric fashion, models with an infinite number of factors have been considered in combination with shrinkage priors on the loadings allowing to automatically select the number of active factors (see e.g. Bhattacharya and Dunson, 2011; Durante, 2017; Schiavon and Canale, 2020).

In the framework we developed, model selection is even more troublesome since it involves the choice of both $K$ and the number of variable clusters $G$. A similar problem, even if arising in the mixture of factor analyzers framework (Ghahramani and Hinton, 1996), has been faced by Fokoué and Titterington (2003) who proposed a stochastic model search to jointly select the number of clusters and factors.

In this work, considering that both $K$ and $G$ might span over a wide range of values, exhaustive searches of the model space are computationally expensive, if not infeasible. Our focus is on models providing good and parsimonious reconstructions of the covariance matrices, jointly with indications about which variables provide similar information through redundancy, rather than on finding the optimal number of factors and groups. For these reasons, we consider an ad hoc initialization strategy which yields a promising configuration $(K_{\text{init}}, G_{\text{init}})$ for the number of factors and variable clusters. The procedure consists in the following steps:

1. Estimate a standard FA model as defined in (1) for $k = 1, \ldots, K_{\text{max}}$, with $K_{\text{max}}$ chosen sufficiently large. This yields the loading matrices $\Lambda_k$, $k = 1, \ldots, K_{\text{max}}$;
2. Use a model-based clustering strategy (see Fraley and Raftery, 2002; Bouveyron et al., 2019, for a recent review) to obtain a partition of the rows of $\Lambda_k$, for $k = 1, \ldots, K_{\text{max}}$, into $G_k$ groups with $G_k$ automatically selected by means of the Bayesian Information Criterion (BIC);
3. Build new loading matrices $\overline{\Lambda}_k$, for $k = 1, \ldots, K_{\text{max}}$, where the rows of $\Lambda_k$ are replaced with the mean of the cluster they belong to; the repeated row values structure of $\overline{\Lambda}$ in (3) is then mimicked;
4. Considering the distributional properties of FA models, compute the BIC for all the models corresponding to different configurations $(k, G_k)$, with $k = 1, \ldots, K_{\text{max}}$. Select as $(K_{\text{init}}, G_{\text{init}})$ the configuration which attains the highest value for the BIC.

This approach allows to find reasonable values that might be used as the starting point of a local search. More specifically, once $(K_{\text{init}}, G_{\text{init}})$ are obtained, we consider a greedy search model selection strategy. Operationally, we fit four different models corresponding to $(K_{\text{init}} \pm 1, G_{\text{init}} \pm 1)$ and we compare them by means of the BIC-MCMC (Frühwirth-Schnatter, 2011). The model with the best value of the information criterion is then selected and its neighboring models are subsequently estimated. These two steps are iterated until no improvements in the BIC-MCMC are found. The best model according to the BIC-MCMC is then selected and used to reconstruct the covariance structure and to obtain a partition of the wavelengths. Other
information criteria, such as the BICM (BIC Monte-Carlo) or the AICM (AIC Monte-Carlo) proposed by Raftery et al. (2007), might be used but, from our experience, BIC-MCMC is more reliable being less influenced by possible fluctuations and jumps in the log-likelihood values across the MCMC draws.

Lastly, some sensitivity analyses reported in the next section confirms that running a global and exhaustive search is not strictly necessary if covariance reconstruction is the final aim.

4. Synthetic data. In this section we investigate the performances of the proposed procedure on some synthetic datasets. The aim of the analyses is twofold. On one hand, we want to quantify the deterioration of the results when a wrong model, having different \((K, G)\) values with respect to the model generating the data, is employed. The possible deterioration is studied in terms of variable partitions quality, that is measured according the Adjusted Rand Index (ARI, Hubert and Arabie, 1985), and of correlation reconstruction. The latter is evaluated considering two different criteria measuring the dissimilarity between the true correlation matrix \(R\) and the estimated one \(\hat{R}\). The first criterion considered is the Mean Squared Error (MSE), defined as

\[
\text{MSE}(R, \hat{R}) = \frac{1}{p(p+1)/2} \sum_{j=1}^{p} \sum_{j \geq j'} (R_{jj'} - \hat{R}_{jj'})^2
\]

where \(R_{jj'}\) represents the \((j, j')\)-th element of the matrix \(R\). The second criterion adopted is the RV coefficient (Abdi, 2007) expressed as

\[
\text{RV}(R, \hat{R}) = \frac{\text{tr}(R^T \hat{R})}{\sqrt{\text{tr}(R^T R) \text{tr}(\hat{R}^T \hat{R})}}
\]

and taking values between 0 and 1 where values closer to 1 denote a greater similarity between the matrices. We consider the correlation matrices and not the covariance ones to have more interpretable values from a MSE perspective.

The second aim of the simulation study consists in the numerical exploration of the quality of the initialization strategy proposed in Section 3.5 to find promising configurations for the number of factors \(K\) and variable clusters \(G\).

A total of \(B = 100\) samples have been drawn with sample size \(n = 4000\) and \(p = 500\) variables. The data are generated according to the probabilistic mechanism underlying model (3) so that the sampled vectors \(x_i\)'s are distributed as a Gaussian random variables with zero mean and covariance matrix \(\Sigma_{\text{true}} = \Lambda_{\text{true}} A_{\text{true}}^T + \Psi_{\text{true}}\). The true number of factors and of variable clusters have been fixed to \(K_{\text{true}} = 4\) and \(G_{\text{true}} = 20\) respectively, to build a simulation setting being coherent with the situation faced when analyzing the milk MIR spectroscopy data. Prior to running the Gibbs sampler outlined in Section 3.4, the parameters are initialized by estimating a standard FA model, where the factor loadings are obtained as the cluster centroids of a \(k\)-means clustering procedure, which also allows to obtain the starting values for the loadings partition. The hyperparameters have been selected according to the approach given in Section 3.3 with \(\sigma_{\lambda} = 5\). All the analyses have been conducted within the \(\text{R}\) environment (R Core Team, 2020) with the aid of the \texttt{mclust} package (Scrucca et al., 2016).

Results are reported in Tables 1, 2, 3 and 4. First of all, note that variable partitions might be erroneously seen as a byproduct of the procedure proposed in Section 3.2, helping to reduce even further the number of free parameters when resorting to factor analysis. In contrast, obtaining variable clusters can represent the final aim of the analyses since it produces relevant insights about the phenomenon that we are studying, as it will be clear for the application in Section 5. For this reason, a proper evaluation of the clustering performance is crucial to validate our procedure. Table 1 shows how the quality of the obtained partitions is
robust to possible under or overestimation of the number of factors and clusters. In fact, even if overestimating \( G \) can be slightly beneficial from a clustering perspective, the ARI values do not show substantial changes across different \((K, G)\) configurations. Generally speaking, as long as \( K \) and \( G \) are in a neighborhood of \( K_{\text{true}} \) and \( G_{\text{true}} \), the obtained variable partitions are close to the true groupings.

In Tables 2 and 3 the results concerning correlation matrix reconstruction are reported. Both the MSE and the RV coefficient tend to provide very similar indications. The underestimation of \( K \) has a detrimental effect on the values of both the indices, especially with \( K = 2 \) when the quality of the estimates of \( R \) shows a clear degradation. Nonetheless, as long as \( K \) is greater or equal to \( K_{\text{true}} \), good reconstructions of the correlation matrices are obtained. The results are more robust to different choices for \( G \), with overestimation of the number of variable clusters appearing to be slightly beneficial.

Finally, in Table 4 the performances of the model selection initialization strategy outlined in Section 3.5 are displayed. Promisingly, the true model generating the data is the one selected more often. The procedure provides extremely good results in terms of the selection of \( K \), as the true number of factors is chosen in more than 95% of the cases. The correct number of clusters seems harder to detect, as it is observed a slight tendency to underestimate it. Nonetheless, most of the time values reasonably close to the true number of groups are selected, with \( G \in \{G_{\text{true}} - 1, G_{\text{true}}, G_{\text{true}} + 1\} \) for more than 85% of the simulated samples.
We believe that the results reported in this section have to be considered as a whole, to obtain useful indications about reasonable paths to take when analyzing real datasets. First of all, note that the strategy outlined in Section 3.5 never selected $K \leq 3$, corresponding to those scenarios which showed a deterioration for the MSE and the RV values. Generally speaking, the results in Table 4 show how the model selection step chooses $(K,G)$ configurations which lead to extremely good results in terms of the quality of the partitions and which provide an accurate reconstruction of the covariance matrix. As a consequence, we believe that the proposed strategy can be fruitfully used as a fast and effective replacement of more intensive and time consuming grid searches over $K$ and $G$ coupled with the reliance to some information criterion that has to be carefully selected.

Some unreported analyses showed that a computationally intensive grid searches based on information criteria, such as the BIC-MCMC, AICM and BICM, produce less satisfactory results with respect to the one reported in Table 4; this serves as confirmation of the soundness of the initialization strategy outlined in Section 3.5.

5. Application to the milk MIR spectroscopy data. In this section, the proposed method is applied to the milk MIR spectroscopy data described in Section 2. The initialization of the parameters and the specification of the hyperparameters have been carried out coherently with what we have done for the synthetic data. Prior to running the proposed methodology, we removed from each single spectrum three wavelength regions reported to be highly noisy (Hewavitharana and van Brakel, 1997) namely the ones from 1592 cm$^{-1}$ to 1720 cm$^{-1}$, from 2996 cm$^{-1}$ to 3698 cm$^{-1}$ and from 3818 cm$^{-1}$ to 5010 cm$^{-1}$. Consequently, we work with a dataset having $n = 4320$ milk samples and $p = 533$ wavelengths.

The initialization procedure outlined in Section 3.5 selects different $(K,G)$ configurations for the Pasture and for the TMR samples. If in the former case it selects a number of factors $K$ equal to 4 and a number of variable clusters $G$ equal to 25, in the latter one $K = 3$ and $G = 19$. This might give a rough indication of the more complex wavelength relationships underlying the samples coming from pasture fed cows, since to capture these relations an higher number of factors and clusters are needed.

In Figure 3 the estimated correlation matrices, obtained running the proposed model with the mentioned $(K,G)$ values, are reported. By comparing these matrices with the sample correlation ones in Figure 2 we can obtain an indication about the capabilities of our methodology to map the data into lower dimensional subspaces while retaining the relevant correlation structures. Denoting with $R$ the sample correlation matrix and with $\hat{R}$ the estimated one we have that, for the pasture samples, $\text{MSE}(R_{\text{Pasture}}, \hat{R}_{\text{Pasture}}) = 0.021$ and $\text{RV}(R_{\text{Pasture}}, \hat{R}_{\text{Pasture}}) = 0.980$ while, in the case of TMR, $\text{MSE}(R_{\text{TMR}}, \hat{R}_{\text{TMR}}) = 0.035$ and $\text{RV}(R_{\text{TMR}}, \hat{R}_{\text{TMR}}) = 0.965$. These results suggest that our method reconstructs in quite a satisfactory way the relations among the wavelengths and that the initialization strategy selects reasonable values for $K$ and $G$.

The graphical inspection of Figure 3, shows how our variable clustering mechanism tends to favour the appearance of blocky structures, thus possibly simplifying the interpretation of
the relations among wavelengths. Despite a rather similar correlation structure, this characteristic of our proposal highlights even more the differences in the correlation among different wavelengths regions between milk samples coming from pasture and TMR fed cows. These differences may serve as an interesting starting point to study how the diet regimens can impact the chemical processes underlying the spectral behaviour.

Another way we consider to assess the performances of our methodology consists of investigating the partitions of the variables. In Table 5, we report the confusion matrix comparing the two clusterings of the wavelengths obtained on the pasture and on the TMR milk samples. Despite having a different number of clusters, the two partitions are similar as the table shows an almost diagonal structure. The quite high value of the ARI, equal to 0.651, confirms their
FIG 4. Wavelengths partitions obtained on samples from pasture fed (on the top) and TMR fed (on the bottom) cows. The grey shaded areas correspond to the removed noisy regions.

similarity. The agreement between the two partitions is expected since we are examining milk samples where the only different experimental condition consists in the different diet regimens. This behaviour can be seen as a strong signal about the presence of a real clustering structure in the measured wavelengths, thus entailing a traceable redundancy in the information they provide. A careful analysis of the results in Table 5 reveals how the different number of clusters among the partitions generally imply that the large TMR variable clusters are split in two pasture variable clusters, as it happens for example for TMR clusters 11 and 17. This might provide some initial indications, possibly deserving further explorations, about how the diets can impact chemical features in the milk, in turn modifying the structures we see in the spectral data. Similar insights can be drawn from Figure 4 where the partitions of the wavelengths for the two different diet regimens are visually represented.

Furthermore, the results obtained from a clustering perspective may be exploited to build variable selection tools possibly useful both for exploratory or graphical analyses and for classification purposes. In fact, the strong redundancy implied by the observed clustering
structures can be used to build new features defined as summaries of the groups themselves possibly highlighting differences among pasture and TMR samples.

We also considered a cluster-specific predictive analysis, to gain some further practical knowledge on the phenomenon that we are studying. In particular, we run different linear regression models, separately for pasture and TMR samples, where the covariates are given by the spectral measurements at the wavelengths belonging to a cluster, while the response variables are the content of fat, protein and lactose in the samples. These analyses, as briefly mentioned in the introduction, are important to understand if spectroscopy data can be used to predict some important features of the milk in a rapid and non-expensive way. Moreover, the predictions are based only on a small subset of variables, namely those assigned to a specific cluster, thus alleviating high-dimensionality induced issues. In Figure 5 we report the obtained results in terms of the Adjusted R-squared index. At first glance, it seems that the content of fat in the milk samples is easy to predict, regardless of the specific spectral region considered and of the diet regimens. The predictive performances are generally higher for the TMR samples in comparison to the pasture ones. A closer inspection of the results, if paired with the suggestions obtained about the wavelength clusters, allows us to indicate how the information carried in some spectral regions is different depending on the diet. For example, from Table 5 we can see how TMR cluster 15 find its correspondence with pasture clusters 16, 17 and 19. The wavelengths in the corresponding regions seem to produce better predictions of the lactose, in terms of the Adjusted R-squared, for the TMR with respect to the pasture milk samples. Similar indications can be found by carefully studying jointly the results shown in Table 5 and Figure 5.

The clustering results obtained, if paired with previously conducted studies, can lead to other relevant insights. For example, the work by Picque et al. (1993) suggests that the measurements in the region spanning from 1515cm\(^{-1}\) to 1593cm\(^{-1}\) are characteristic of the lactate ion. On the other hand, the regions from 1040cm\(^{-1}\) to 1100cm\(^{-1}\) and from 1298cm\(^{-1}\) to 1470cm\(^{-1}\) are related to galactose component of milk. Note that, both lactate and galactose can be seen as indicators of the milk quality. As is visually clear in Figure 4, the wavelengths

---

**Table 5. Adjusted R-squared values of the regression models where only the wavelengths in a specific cluster are used to predict the content of three different milk traits, namely Fat, Protein and Lactose contents. On the left the results for the pasture samples, on the right the ones for the TMR ones.**

**Pasture**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Fat</th>
<th>Protein</th>
<th>Lactose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>0.97</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>5</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
</tr>
</tbody>
</table>

**TMR**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Fat</th>
<th>Protein</th>
<th>Lactose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>0.94</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>3</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>4</td>
<td>0.90</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>0.88</td>
<td>0.88</td>
<td>0.89</td>
</tr>
</tbody>
</table>

---

**Figure 5.** Adjusted R-squared values of the regression models where only the wavelengths in a specific cluster are used to predict the content of three different milk traits, namely Fat, Protein and Lactose contents. On the left the results for the pasture samples, on the right the ones for the TMR ones.
FIG 6. Estimated correlation sub-matrices corresponding to the first 50 wavelengths computed on the milk samples produced by pasture fed cows (on the left) and TMR fed cows (on the right).

pertaining to the lactate ion region mainly belong to pasture cluster 8 and to TMR cluster 7; a closer inspection of Table 5 reveals how these groups are strongly related in the two partitions, giving an additional indication about the coherency of the clustering results. The wavelengths belonging to the galactose regions are split into groups 2, 5 and 7, for the pasture samples, while they are mainly associated with groups 3 and 6 for the TMR samples; again with a strong correspondence among these clusters visible in the confusion matrix. The results obtained from the cluster specific regression analyses show how these groups are among the best ones for predicting the lactose content in the milk samples. Since lactose is a disaccharide molecule formed by the linking of the galactose and glucose monosaccharide modules, these results serve as a confirmation of the practical utility of the variable partitions obtained from the model.

Lastly, the correlation matrices estimated by using the proposed methodology can be used as an exploratory tool to study the relations among wavelengths in a specific spectral region. Focusing on the correlations among the first 50 wavelengths, from Figure 3 we note that these wavelengths seems to relate one to the other differently depending on the diet; this region is shown more closely in Figure 6. Here we can see how the relationships among the spectral values at these wavelengths is highly dependent on the diet in this region. If we compare the two sub-matrices using the metrics considered previously, we have $\text{MSE}(\hat{R}_{\text{Pasture}}^{(1:50)}, \hat{R}_{\text{TMR}}^{(1:50)}) = 0.058$ and $\text{RV}(\hat{R}_{\text{Pasture}}^{(1:50)}, \hat{R}_{\text{TMR}}^{(1:50)}) = 0.931$, thus providing an indication of stronger discrepancies compared to the one observed among the full covariances. Again, considering jointly these indications with the results outlined above, we can hypothesize the reasons behind this difference. More specifically, the initial wavelengths seem to give consistently better performances when predicting the protein content for the TMR samples compared to the pasture samples. Similar analyses can be conducted also for other spectral regions, depending on the specific application interest.

6. Discussion and further work. In this paper we have presented a modification of a standard Factor Analysis model where the factor loadings matrix is reparameterised so that redundancy in the originally observed variables can be detected. We implement a flexible Metropolis-within-Gibbs sampler to estimate the proposed model. Our method yields a parsimonious representation of strongly dependent high-dimensional data with complex correlation structures and leads to many fewer parameters to be estimated compared to a standard FA model. Further, as a direct consequence of the specification itself, the model yields a
grouping of the original variables when mapping them into the lower-dimensional subspace. The subsequent partition throws light on the relations among the observed features and about their possible redundancies. These indications, when supported by subject matter knowledge, can be translated into practical knowledge about the phenomenon under study.

Our proposal was directly motivated by an application to vibrational spectroscopy data analysis and showed good performances on the dairy feed experiment data under investigation, both in terms of correlation reconstruction and interpretability of the results.

Spectroscopy data are high-dimensional with strong correlation structures and complex redundancies, and hence present recurring statistical challenges. The model we introduced has been proven particularly useful in the given context since it has provided a parsimonious characterization of the correlation matrix. Our approach highlighted differences among milk samples, and variable clustering led to useful insights about spectral regions that carry the same information and hence, even if distant, may be influenced by similar chemical processes. Finally, our proposal can be considered as an initial approach for building classification tools to discriminate samples according to the relationships among the observed features thus possibly helping in authenticity assessment and in preventing food adulteration.

We developed our methodology for MIR spectroscopy data but, in principle, it can be applied to other data with similar characteristics. A possible useful extension and research direction consists in the exploration of different possibilities concerning $\pi(Z)$, the prior distribution for the allocation matrix. When accounting for peculiar correlation structures in the data, it can be appropriate, as we briefly mentioned in Section 3.3, to explore prior distributions incorporating information about specific relations and constraints, such spatial or temporal ones, for the variables to be clustered. Another aspect that is worth examining is concerned with the model selection. In Section 3.5 we introduced an initialization strategy that provides good indications about reasonable values for the number of factors $K$ and clusters $G$. Nonetheless, several different approaches may be adopted and a thorough exploration of different model selection tools may be beneficial. As mentioned in Section 3.5, a possible extension consists in allowing $K$ to go toward infinity and in considering shrinkage priors on the factor loadings as proposed in Bhattacharya and Dunson (2011) and Murphy, Viroli and Gormley (2020). This strategy for choosing $G$ and $K$ allows to circumvent the issues related to the selection of $K$, by automating the choice of the active factors, i.e. the ones with non-negligible loading values, in characterizing the covariance structure. Finally, our proposal can be thought as a stepping stone when building new classification tools. A possible straightforward strategy, pointing in this direction, would consist in embedding the model we introduced in a Mixture of Factor Analysis (MFA, Ghahramani and Hinton, 1996) framework thus allowing to perform classification and clustering of high-dimensional data.

Acknowledgements. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and the Department of Agriculture, Food and Marine on behalf of the Government of Ireland under grant number (16/RC/3835) and the SFI Insight Research Centre under grant number (SFI/12/RC/2289_P2).

REFERENCES


