Bayesian Analysis (2021) **16**, Number 2, pp. 489–519

A Bayesian Factor Model for Spatial Panel Data with a Separable Covariance Approach

Samantha Leorato* and Maura Mezzetti†,‡

Abstract. A hierarchical Bayesian factor model for multivariate spatially and temporally correlated data is proposed. This method searches factor scores incorporating a dependence within observations due to both a geographical and a temporal structure and it is an extension of a model proposed by Mezzetti (2012) using the results of a separable covariance matrix for the spatial panel data as in Leorato and Mezzetti (2016). A Gibbs sampling algorithm is implemented to sample from the posterior distributions. We illustrate the benefit and the performance of our model by analyzing death rates for different diseases together with some socio-economical and behavioural indicators and by analyzing simulated data.

Keywords: Bayesian inference, correlated factor loadings, factor analysis, autoregressive model, spatial data.

1 Introduction

Factor models are often used to reduce the complexity of data and effectively extract relevant information. The typical assumption of such models is that the observed values of a set of multivariate variables is the result of the action of some unobservable latent variables, or factors, that can also be used to reduce a large number of observables into a smaller number of latent factors. In its original definition, a factor model identifies a set of common factors that are able to reduce the dimension of the response vector for all cross-sectional or time units. A thorough review can be found in Bartholomew et al. (2011).

The first formal definition of a Bayesian approach to factor analysis can be traced back at least to Press (1972) and Kaufman and Press (1973), although a few earlier works did account for subjective information, limitedly to specific problems (see also Martin and McDonald (1975), Press and Shigemasu (1989) and Press and Shigemasu (1997)). These first models were rather restricted, with the same limitations as the classical factor model.

The introduction of iterative Markov chain Monte Carlo (MCMC) tools paved the way to the development of fully Bayesian approaches to factor analysis such as, for example, in Polasek (1997), Arminger and Muthén (1998), Aguilar and West (2000), Lopes and West (2004) and Rowe (2003). These new models allowed for heterogeneities across time of either the latent factors or the factor loadings.

*Department of Economics Management and Quantitative Methods, University of Milan, Italy, samantha.leorato@unimi.it

†Department of Economics and Finance, University of Rome "Tor Vergata", Italy, maura.mezzetti@uniroma2.it

‡Corresponding Author.

© 2021 International Society for Bayesian Analysis https://doi.org/10.1214/20-BA1215

The application of factor models to spatial data, soon led to the introduction of further generalizations, that allowed the so called spatial factor models to cope with specific characteristics of spatial data in terms of heterogeneity and autocorrelation. Wang and Wall (2003) use a common spatial factor model to study cancer risk across counties in Minnesota; Hogan and Tchernis (2004) use a hierarchical Bayesian factor model to account for spatial correlation between locations, where the spatial correlation is introduced on the latent scale variable and can be specified either marginally or conditionally; Lopes and West (2004) and Lopes et al. (2008) consider spatio-temporal models, where temporal and spatial dependence are modeled by latent factors and factor loadings respectively; Mezzetti (2012) generalizes the model by Rowe (2003) and Rowe and Press (1998) by proposing a hierarchical Bayesian factor model for multivariate spatially correlated data. Ren and Banerjee (2013) define a low-rank spatial factor model, to deal with large vectors of outcomes for a large number of locations.

Most of these generalized Bayesian factor models exploit the flexibility of hierarchical Bayesian modeling, that allows to take into account both heterogeneities in factors (loadings and/or latent) and possible correlation between observations. The latter is obtained through the use of proper prior information to simplify the covariance structure.

In this paper we propose a hierarchical Bayesian factor model for multidimensional spatio-temporal data that is able to take into account heterogeneity in latent factors (both in time and space dimensions). This is obtained through a separable covariance structure, assuming a multiplicative effect of the covariance among variables and among regions, combined with a Vector Autoregression (VAR) specification for the latent vectors at every location. Similarly to Lawson et al. (2012), Leorato and Mezzetti (2016) or Chan (2020), the separability of the covariance structure is obtained by the introduction of a Kronecker product. Here the separability is assumed between the N-dimensional spatial covariance matrix and the covariance matrix of the K response variables. Both matrices are given an arbitrary form, thus avoiding one of the major problems with specifications of the covariance structure, which is a sore point in spatial data modeling (see the discussion in Leorato and Mezzetti (2016) and references therein). Further, as noted by Chan (2020), the definition of a Kronecker error covariance structure is also compatible to various commonly used non-Gaussian models.

Most of the literature dealing with factor models for spatio-temporal data does not allow the dynamic factors to be location-specific or does not consider multivariate observations repeated in time on several locations. One of the lines of research where there has been a growing interest in the investigation of the determinants of simultaneous space–time variations of multiple outcomes is health and medicine. For example, Tzala and Best (2008) use a Bayesian hierarchical latent factor model for spatially and temporally correlated multivariate epidemiological data. Their idea is to exploit factor analysis to provide a flexible framework to map the space–time evolution of disease-specific trends in cancer risks. Because of the nature of data, the Poisson model is used to describe the distribution of the observed deaths in all types of cancer analyzed. In these types of applications, the identification of spatio-temporal patterns in a small number of factors that are able to synthesize the variations of a vector of several variables, potentially correlated in a complex way, provides important information and insight on the mechanisms playing a role in the evolution of one or more diseases, reinforcing the evidence

of common underlying risk factors. As pointed out by Tzala and Best (2008), accounting for spatio-temporal heterogeneity in a multivariate factor analysis, is likely to *"lead to improved precision for the estimation of the underlying disease risks by borrowing strength from other diseases as well as close areas or time points"*.

Factor models for multivariate space–time data are also suitable to cope with environmental and ecological applications. For example, Calder (2007) propose a dynamic factor process convolution model. The approach of our paper shares some similarities with the work of Calder (2007), such as the Bayesian estimation approach and the dynamic feature of the factors. However, there are important differences: first, we don't rely on the convolution of an underlying process through a kernel smoother K, because the main purpose of our approach is the reduction of the dimension of the "types of variables" only, allowing for location and time specific factors. Therefore we don't need to address the delicate problem of the choice of tuning parameters such as the resolution of the grid, kernel and bandwidth selection. A second difference consists in the covariance matrix of the factors: our factor component $F(t)$ plays a similar role to Kx_t in Calder (2007), but while the variance of the first include both a spatial (Φ) and a "type-of-variable" component (Ψ), the variance of the latter only accounts for possible heteroskedasticity in the types of variables, through $var(Vec(\epsilon_t)) = I_N \otimes diag(\lambda_\epsilon)$.

Despite the potential application of multivariate factor models for space–time data in other areas of research, such as, for example, in multivariate macroeconomic time series, to the best of our knowledge, limited attention has been drawn to this kind of generalization aside from the aforementioned work by Tzala and Best (2008) and Calder (2007). Few models pay close attention to the spatio-temporal structure of the error terms, typically assuming a strict diagonal structure for the error covariance (as in Fan et al. (2008)).

Therefore, a first motivation of our work is to fill this gap in the literature, by proposing a hierarchical Gaussian-type model, which has the appealing feature of being suitable to a broader range of applications compared to Tzala and Best (2008). Moreover, the use of a Kronecker product in the definition of the covariance between variables and between locations has strong computational advantages, besides those already mentioned relative to the flexibility of the model, and appears to be fairly little used in this framework. A further novelty in our approach is the way we cope with the identification of the factors, which is slightly different from the standard arguments following by Geweke and Zhou (1996): instead of setting a fully informative prior for a subset of the factor loading matrix, we allow for a diffuse conjugate prior and impose the necessary restrictions on the posterior density.

As in Tzala and Best (2008), our motivating application aims at detecting possible common sources that might have a hidden impact on one or more diseases and their eventual dynamics in space and time as well as their mutual interaction. Section 2 describes in more detail the motivating application and the data.

In Section 3, after a brief review of Bayesian factor analysis, we present our framework, we discuss the prior distributions and the identification strategy. Section 4 is dedicated to the computation of the posterior distributions, while Section 5 is dedicated to the estimation of the number of factors. In Section 6 some simulations are

presented to investigate the performance of the proposed model. Section 7 presents the results of the application, with a Subsection 7.1 on sensitivity analysis. Finally, Section 8 concludes.

2 Data

Several authors have tried to use factor analysis to capture hidden relationships between a disease and possible risk factors (Das et al., 2010). Analysis of local variations of disease rates helps identifying latent sources of heterogeneity and uncovers the existence of risk factors that contribute to formulate etiological hypotheses. Considering several diseases simultaneously permits to detect common determinants (Yanai et al., 1978). Moreover, taking advantage of the extra information obtained from different diseases, it allows to improve the precision of the estimates. On the other hand, socio-economical as well as behavioral features tend to have a geographical pattern in Europe: as for example, Scandinavian countries evidence a low unemployment rate and adopt similar welfare programs, while Mediterranean countries share similar dietary intakes. These and other characteristics are potential determinants of one or more diseases but their interrelation is unclear as well as the potential existence of dynamic behaviors such as, for example, local persistence in time or global convergence (Flood et al., 2008).

Following this idea, we perform a factor analysis on the death rates for different diseases, on one hand, and, on the other hand, some socio-economical and behavioural indicators. We aim at highlighting the existence of common spatial patterns in the determinants of morbidity, and its eventual variations through time.

We took the data from the Demographic and Health Surveys (DHS) Program, ICF International and European Health for All databases, World Health Organization (WHO). Data collected refer to the period $1970 - 2015$ and to the following countries: Albania, Armenia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Italy, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Turkmenistan, Ukraine, United Kingdom and Uzbekistan.

Socio economic variables such as: average income measured in dollars and unemployment rate are recorded. For each country few behavioral variables are recorded, in particular: intake of energy from fat, intake of energy from protein and adult per capita consumption of recorded alcohol. Fertility rate is observed for each country. Different health status indicators are also recorded, age standardized death rates (SDR) for 100000 inhabitants for the following diseases: smoking related, alcohol related, malignant neoplastic, female breast cancer, tuberculosis, infection disease, circulatory system disease, ischaemic heart disease, liver disease and all causes mortality rate. Finally country specific life expectancy is recorded.

Due to the high number of missing values, 16 countries and 25 years are dropped and the final dataset regards 27 countries observed during the 10-years period (2000–2009). The remaining number of missing values is quite negligible, being about 0.4%

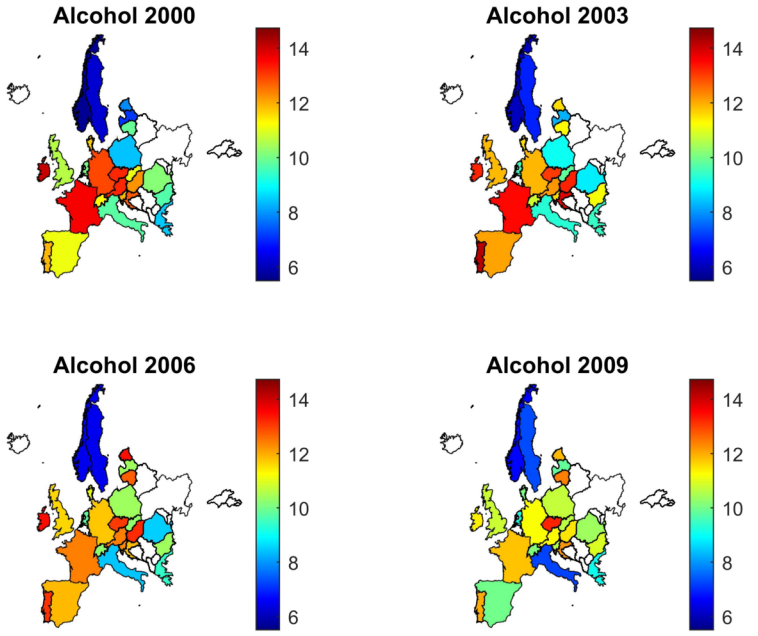

Figure 1: Adult per capita consumption of alcohol (litres per year): Distribution across countries and years.

of the observations; in order to avoid complicated structure, the missing values are just filled with a plausible one, as suggested in Pigott (2001). Both mean for the cases of the observed variable, or the value relative to the previous year provide the same results.

Figures 1–5 show the distribution over time and space of five variables among the seventeen variables considered: alcohol consumption, fertility rate, all cause mortality, annually average income and unemployment rate. A temporal pattern appears to be very strong. Spatial pattern seems to be more difficult to capture. Distribution of standardized mortality rate for all cause of mortality looks like signing a contrast between Western and Eastern countries. On the other hand, a higher value of fertility rate is observed in Northerner countries and France, while France is closer to neighboring countries regarding alcohol consumption distribution. Average income distribution determines three main areas: North, South and Eastern Europe. Unemployment rate shows a weaker spatial pattern with respect to the other variables mapped.

The following variables: fat intake, alcohol related disease SDR, all causes SDR, breast cancer SDR, all cancer SDR, circulatory system SDR, infection disease SDR, ischaemic SDR, liver related disease SDR, smoking related disease SDR, tuberculosis SDR, average income and unemployment rate are considered on a logarithm scale to better be approximated by a Gaussian distribution.

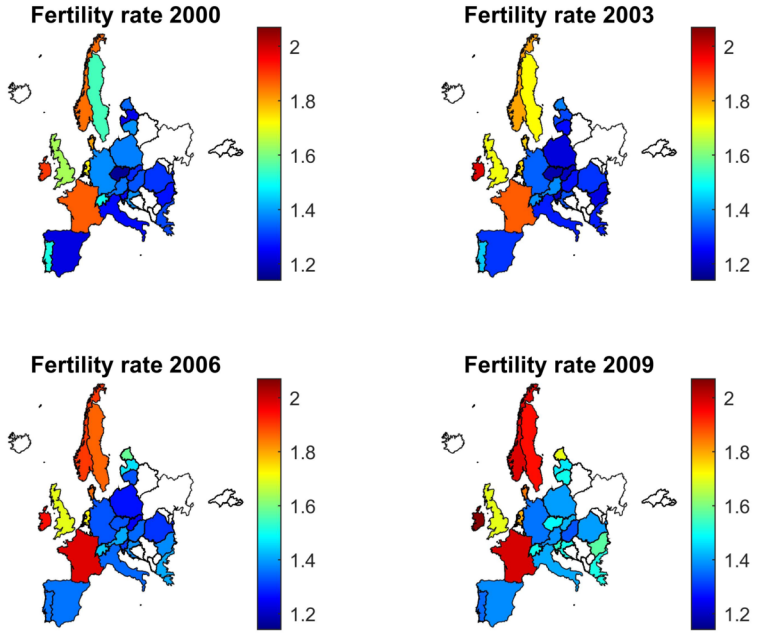

Figure 2: Fertility Rate, as the average number of lives births per 1,000 of the population: Distribution across countries and years.

We can observe three levels of correlation in the data: between time points, between countries and between variables, displayed in Figures 6–8. The three types of correlation are all very strong. We model parametrically only the first one. Overall, all variables show evidence of temporal correlation within variables and countries (see Figure 6). Among the variables considered, the spatial structure assumes different forms, even though some common spatial pattern emerges and some correlation between countries results quite strong (see Figure 7).

Since variables are highly correlated (see Figure 8), it may be interesting to identify a small number of common factors that are able to explain the majority of information within the sample covariance matrix. Each factor will result related to a particular subset of variables, and therefore the definition of factors itself suggests their interpretation. To accomplish this task, the estimation of a factor model is a natural choice. Due to the noticeable heterogeneity of the variables in space and time, it is however necessary to allow factors to vary in both dimensions. Moreover, to take into account for the evident temporal autocorrelation among almost all variables as well as for the possibility that some of the spatial structure is induced by spatial spillovers, we want to allow the country-year-specific factors to be spatially and serially correlated. Further, in the model illustrated in Section 3, the correlation within and between regions are assumed to have a multiplicative effect and separability in the covariance is modelled, while temporal effect on factor scores is assumed additively.

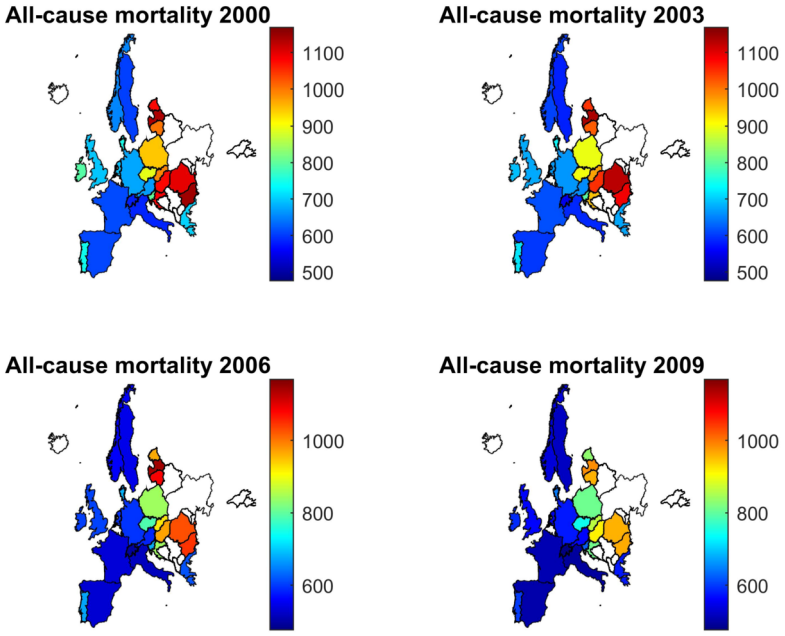

Figure 3: Age-standardized all causes death rate, per 100 000 population: Distribution across countries and years.

3 Model

3.1 Model Proposed

Let X(t) be a matrix $(N \times K)$ of observations on K variables over N countries, at time t and let $X = [X(1)', X(2)', \cdots, X(T)']'$ be the $NT \times K$ dimensional matrix of K observed variables, for the N countries at T time periods. Following Rowe (2003), we write the model for the K observed variables at time t for location n, as

$$\underset{(K\times 1)}{x_n(t)}\,|\mu, \Lambda, f_n, m \;=\; \underset{(K\times 1)}{\mu(t)} + \underset{(K\times m)}{\Lambda}\;\underset{(m\times 1)}{f_n(t)} + \underset{(K\times 1)}{\varepsilon_n(t)}, \tag{3.1}$$

where $\mu(t)$ is a K-dimensional unobserved population mean vector at time t, Λ the $K \times m$ matrix of factor loadings, $f_n(t)$ is the m-dimensional vector of *common* factor scores for the n-th region at time t, and $\varepsilon_n(t)$ is a vector of disturbance terms of the n-th region on the K variables at time t. The parameters $(\mu(t), \Lambda, f(t),\ t \geq 1)$ in the model are unknown and have to be estimated. Usually, the estimation of $\mu(t)$ does not present problems, as typically coincides with the sample mean (see Lawley (1940)), therefore, without lack of generality, we can assume $\mu(t) = 0$. While factor scores are allowed to vary across time and countries, the loadings have dimension $K \times m$. The approach of keeping the loadings fixed across time and regions, also followed in Tzala and Best (2008) or Calder (2007), favors the description of latent factors and their interpretation as well as the identification of local trends or country-specific evolution of the scores.

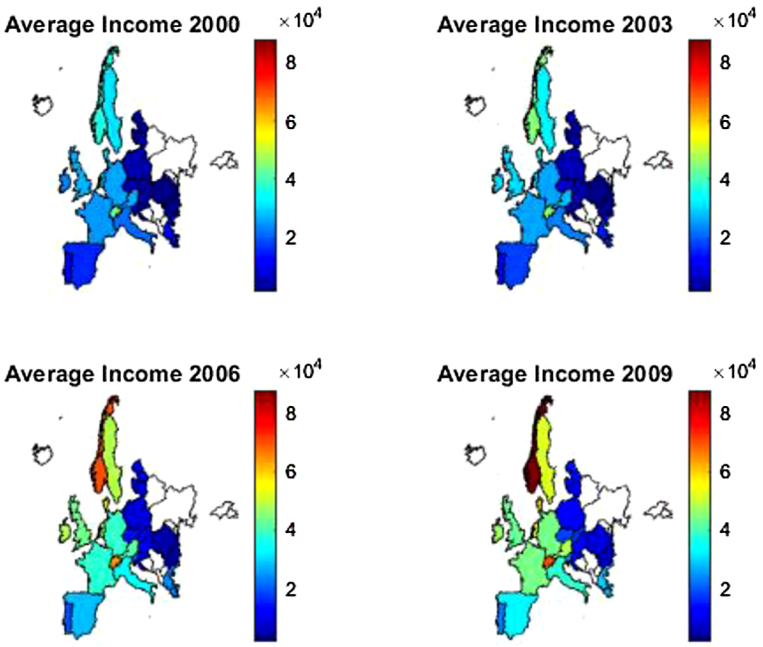

Figure 4: Annually average income measured in dollars: Distribution across countries and years.

In a classical factor analysis model the errors $\varepsilon_n(t)$ are assumed to be normally distributed and independent, conditional on the common factors: thus the common factors are the only source of dependence among the K variables.

Classical and Bayesian factor models assume independent rows of the observation matrix. This assumption turns out to be not reasonable for spatially correlated data. In order to show our proposal to account for possible spatial correlation, let us rewrite the classical factor model as illustrated in (3.1) by rearranging the $N \times K$ terms in $X(t)$ in a $NK \times 1$ vector (where the first K elements are the K-observations relative to first region at time t etc.). We denote by $\text{vec}(X(t))$ the vector obtained by this operation, and get

$$\overset{(NK\times 1)}{\text{vec}(X(t))} \,|\mu, \Lambda, F, m = \begin{pmatrix} \overset{(N\times N)}{I_N} & \otimes & \overset{(K\times m)}{\Lambda} \end{pmatrix} \overset{(Nm\times 1)}{\text{vec}(F(t))} + \overset{(NK\times 1)}{\text{vec}(\varepsilon(t))}, \tag{3.2}$$

where $\otimes$ indicates the Kronecker product between two matrices. By our specification, the residuals are correlated between countries and variables:

$$\begin{aligned} \text{vec}(\varepsilon(t)) &\sim N(0, \Omega), \\ \Omega &= \sigma_\varepsilon^2(\Phi \otimes \Psi), \end{aligned} \tag{3.3}$$

where the two positive definite matrices Φ and Ψ have sizes, $N \times N$ and $K \times K$ and induce spatial correlation and correlation among variables, respectively. Traces are constrained

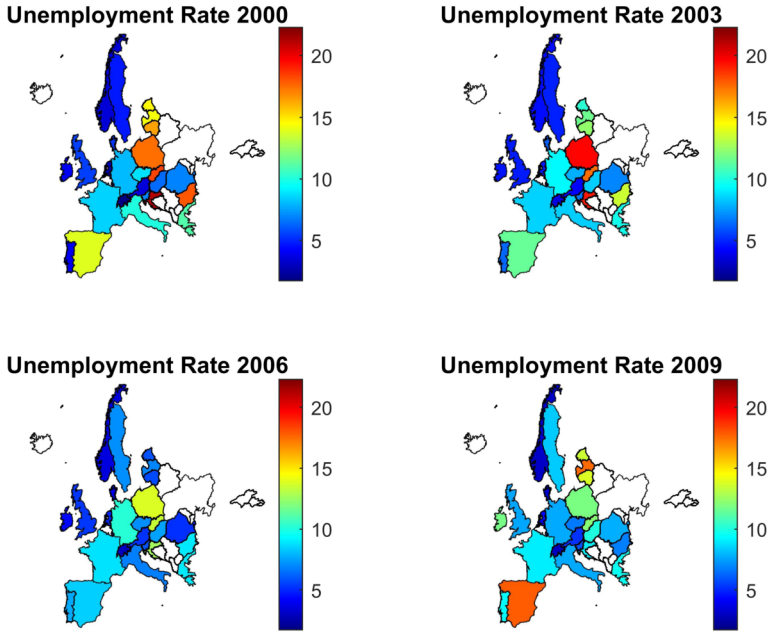

Figure 5: Unemployment rate, computed as the number of unemployed people as a percentage of the labour force: Distribution across countries and years.

to be fixed and proportional to the matrix size, for identification (see also Leorato and Mezzetti (2016)).

The n-th row $x_n(t)$ of the matrix $X(t)$ contains the K observations of variables on region n at time t. Under the separable structure defined by (3.2) and (3.3), $var(x_n(t)|\Phi, \Psi, m, F, \Lambda) = \sigma_\varepsilon^2 \phi_{nn} \Psi$ and the covariance between rows n and n' of X is $\sigma_\varepsilon^2 \phi_{nn'} \Psi$, while the covariance between columns k and l of $X(t)$ (representing observations of variables k and l on all the N regions at time t) is $\sigma_\varepsilon^2 \psi_{kl} \Phi$.

We rewrite (3.2) and (3.3) as:

$$p(X(t)|F(t), \Lambda, \Psi, \Phi, m) \propto$$
$$\sigma_\varepsilon^{-Np} |\Phi|^{-p/2} |\Psi|^{-N/2} \exp\left(-\frac{\sigma_\varepsilon^{-2}}{2} \mathrm{tr}\left(\Psi^{-1}(X(t) - F(t)\Lambda')'\Phi^{-1}(X(t) - F(t)\Lambda')\right)\right),$$
$$t = 1, \ldots, T, \tag{3.4}$$

showing that the distribution of the observations conditional on the factors admits autocorrelation on space and between variables.

Besides that, we want to allow for both a time autocorrelation and a spatial dependence on factors scores F. Specifically, the factor scores $F(t)$ follow an autoregressive model, while we allow spatial autocorrelation in the columns of $F(t)$ (for computational convenience we assume the spatial correlation matrix to be equal to Φ, although in

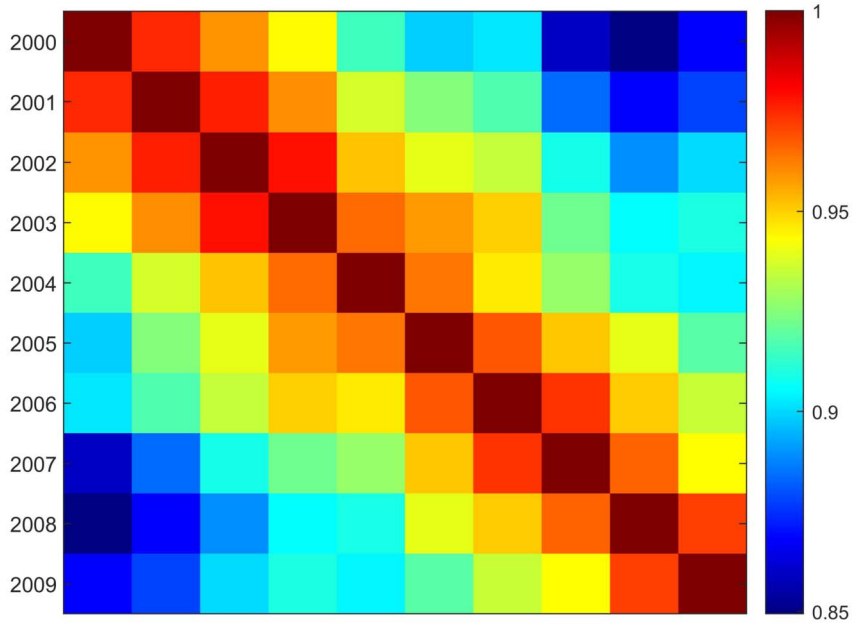

Figure 6: Total correlation of variables between time points, corresponding to years from 2000 and 2009.

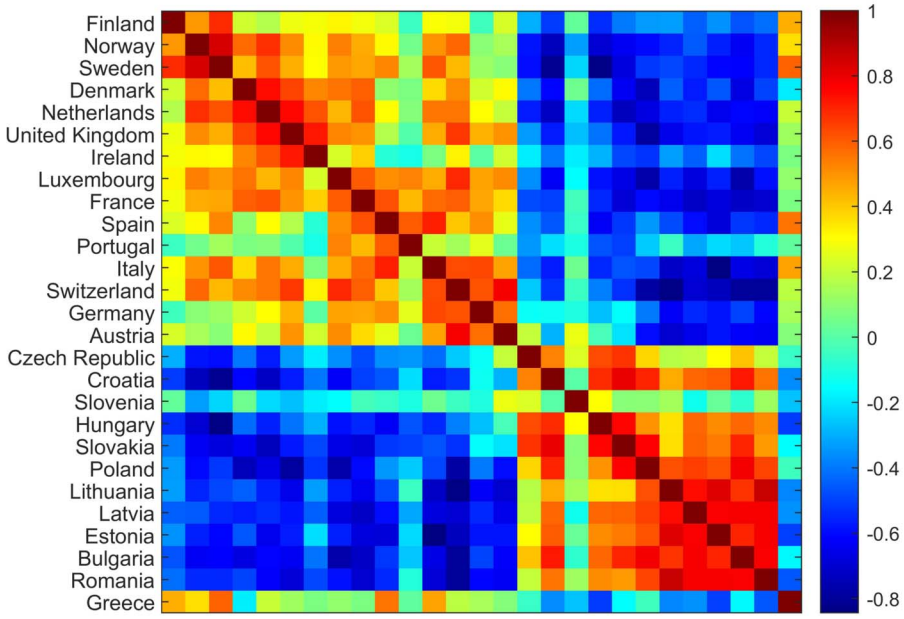

Figure 7: Total correlation of variables between countries. Countries are ordered from North to South and to East.

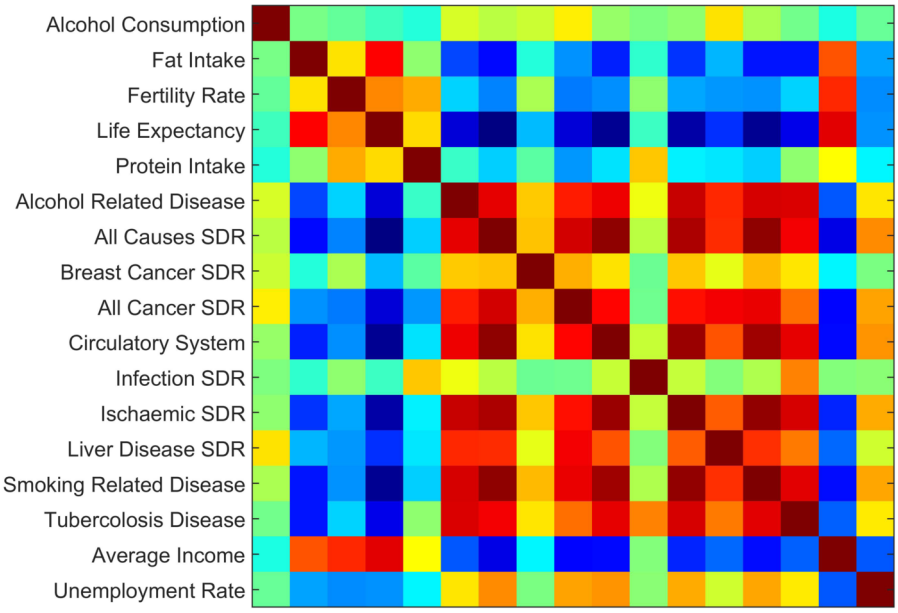

Figure 8: Correlation between variables.

Supplementary Material (Leorato and Mezzetti, 2020) we compute the posterior distribution using an arbitrary positive definite matrix). Further, we permit correlation between factor loadings of different variables through proper definition of prior distributions. Appropriate prior distributions, presented in Section 3.3, are used to account for all these dependencies.

3.2 Identification Strategy

The model in (3.1) presents a well known problem of identification of the parameters. The question of identifiability of factor models has been widely discussed in the literature, and several different proposals have been suggested, each with its specific strengths and drawbacks. This problem is generally solved through the introduction of different sets of constraints on the parameters.

Identification restrictions are necessary due to *rotation invariance* of factor models: model (3.2) cannot be distinguished from a model where both factor loadings and common factors are subject to a rotation. In fact, given a nonsingular $m \times m$ matrix B, transforming $F(t)$ into $F(t)B$ and Λ to ΛB gives the same model as (3.2), for every B:

$$X(t) = F(t)BB'\Lambda' + \varepsilon(t).$$

Practically, this amounts to m^2 redundant parameters (the dimension of B) and the main strategy proposed in the literature has traditionally been the introduction of

the exact same number of restrictions on the parameters. There are several ways to impose those restrictions, and it is not our purpose here to cover all of them (see Bai and Li (2012), Bai (2003) and references therein for a detailed review on this topic). Sometimes all m^2 restrictions are imposed on the factor loadings matrix, leaving the common factors unrestricted. Another possibility is to share the constraints among the sample covariances of the factor loadings and of the common factors.

The question of which one of the set of restrictions is to be chosen is far from being settled. In a Bayesian estimation approach, whenever the computation of the posterior distribution relies on some MCMC tool, the use of one or another identification restriction is more likely to produce relevant differences.

One quite popular choice is to assume the factor loading matrix to be lower triangular (with or without all ones in the main *diagonal*) and to add the remaining restrictions on the covariance matrix of the factor scores.

One of the main drawbacks with this identification strategy is that the posterior distributions, and therefore the interpretation of the results, in general depend heavily on the ordering of the observations. Several authors (see for instance Aßmann et al. (2016), Frühwirth-Schnatter and Lopes (2018) and Conti et al. (2014)) have proposed alternative methods to overcome these liabilities. Nonetheless, this remains one of the most used identification strategies, one of the reasons being the fact that the structure of the loading matrix itself can ease the interpretation of factors.

For this reason, we follow the approach based on a lower triangular structure of Λ:

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \lambda_{2,1} & 1 & 0 & \dots & 0 \\ \lambda_{3,1} & \lambda_{3,2} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & 0 \\ \lambda_{m,1} & \lambda_{m,2} & \lambda_{m,3} & \dots & 1 \\ \lambda_{m+1,1} & \lambda_{m+1,2} & \lambda_{m+1,3} & \dots & \lambda_{m+1,m} \\ \dots & \dots & \dots & \dots & \dots \\ \lambda_{K,1} & \lambda_{K,2} & \lambda_{K,3} & \dots & \lambda_{K,m} \end{bmatrix}. \tag{3.5}$$

However, we distinguish from the standard arguments on one respect: to the best of our knowledge, most of the authors imposing identification restrictions of this form, follow the same ex-ante argument, proposed by Geweke and Zhou (1996) (see for example Aguilar and West (2000), Calder (2007) or Bai and Wang (2015)). This approach consists in the definition of a prior distribution only on the unconstrained elements of the factor loadings, while, for the others, a Dirac distribution (either concentrated in 0 or 1) is assumed. In order to take advantage of the separability of covariance matrices also in the computation of the posterior of Λ, we implement the constraint in a different way. In fact, we impose the constraints by conditioning on the unrestricted posterior distribution of the whole Λ. In this way, we can avoid to resort to a fully informative (deterministic) prior distributions for a subset of the components of Λ. The computation of the constrained posterior distribution of Λ is given in detail in Section 4. In the Supplementary Material we show that the posterior distributions obtained by using our method and the approach by Geweke and Zhou (1996) do not coincide even for very simple factor models.

3.3 Prior Distributions

Following Leorato and Mezzetti (2016), we will use natural conjugate families of prior distributions for the parameters ρ, F, Λ, σ^2_ε, Ψ and Φ. Thus for σ^2_ε, we set an inverse-gamma prior distribution:

$$p(\sigma^2_\varepsilon) \propto \sigma_\varepsilon^{-2\kappa-2} \exp\left(-\frac{\theta}{\sigma^2_\varepsilon}\right). \tag{3.6}$$

As stated in Section 3.1, the factor scores $F(t)$ follow a multivariate autoregressive model of order $h \geq 1$. Let $R_j = \text{diag}(\rho_{j1}, \rho_{j2}, \ldots, \rho_{jm})$, $j \leq h$ and let us define the matrix $F_h(t) = (F(t-1), \ldots, F(t-h))$. Then,

$$\begin{aligned} F(t) &= F_h(t)R + \nu(t), \\ R &= (R_1, \ldots, R_h), \\ \nu(t) &= (\nu^1(t), \nu^2(t), \ldots, \nu^m(t)). \end{aligned} \tag{3.7}$$

The errors $\nu^i(t), \nu^j(t)$ are independent for all $i \neq j$ and are not serially correlated, but they are spatially correlated, through the matrix Φ. In Supplementary Material, we compute the posterior distribution of F, allowing for an arbitrary matrix Γ to govern the spatial correlation matrix, in the particular case when $R = \rho I_m$. However, in the application of the general model, we use $\Gamma = \Phi$ to exploit the computational gain due to simplification of the posterior distribution. A similar remark holds for the prior variance of Λ, given below.

The prior distribution for $F(t)$ is then defined as:

$$p\left(F(t) \mid F_h(t), R, \Phi, m\right) \propto |\Phi|^{-m/2} \exp\left\{-\frac{1}{2}\text{tr}\left(F(t) - F_h(t)R\right)'\Phi^{-1}(F(t) - F_h(t)R)\right\}, \tag{3.8}$$

where we assume the initial conditions $F(-h+1) = \ldots = F(0) = 0$.

We can write the variance of the vector obtained from the matrix F as a $T \times T$ block matrix Σ_F, whose $(t,s) - th$ block is

$$\Sigma_F(t,s) = \Phi \otimes C_{|t-s|}(R),$$

where $C_{|t-s|}(R)$ is a diagonal matrix whose j-th element is the $|t-s|$-th order autocorrelation of an $AR(h)$ process with coefficients $(\rho_{1j}, \ldots, \rho_{hj})$, $j = 1, \ldots, m$.

For example, if $h = 1$, $R = R_1$ and

$$C_{|t-s|}(R) = \left[(I - R_1^2)^{-1} R_1^{|t-s|}\right].$$

As for F, we use a normal prior for Λ, with

$$p(\Lambda|\Psi, m) \propto |\Psi|^{-m/2} \exp\left\{-\frac{1}{2}\text{tr}\left(\Psi^{-1}(\Lambda - \Lambda_0)H(\Lambda - \Lambda_0)'\right)\right\}. \tag{3.9}$$

Therefore, by defining the vectorization $\lambda = \text{vec}(\Lambda)$, the prior distribution of λ is Gaussian, with mean $\lambda_0 = \text{vec}(\Lambda_0)$ and variance $H \otimes \Psi$. The hyperparameter H is fixed, while the matrix Ψ, that allows for correlation between the loadings of different variables, is estimated.

Prior distribution of F reflects our hypothesis of factors with a temporal and a spatial dependence. Note that with our approach to identification, we don't need to specify separate Gaussian prior distributions for all *non null* terms of the rows of Λ, but we can specify Λ, conditionally on Ψ, as a unique Gaussian random matrix with parameters Λ_0 and H.

Both Ψ and Φ follow from *a priori* inverted Wishart distributions with scale matrices A and G respectively:

$$p(\tilde{\Psi}) \propto |\tilde{\Psi}|^{-(\nu+2K+2)/2} \exp\left\{-\frac{\nu}{2}\text{tr}\left(\tilde{\Psi}^{-1}A\right)\right\}, \tag{3.10}$$

$$p(\tilde{\Phi}) \propto |\tilde{\Phi}|^{-(\gamma+2N+2)/2} \exp\left\{-\frac{\gamma}{2}\text{tr}\left(\tilde{\Phi}^{-1}G\right)\right\}. \tag{3.11}$$

Both A and G are symmetric and positive definite matrices, while γ and ν are two scalars. We adopt the following parametrization: $E(\tilde{\Psi}) = A/\nu$ and $E(\tilde{\Phi}) = G/\gamma$.

Since the shape parameters (ν and γ) determine also the level of information of the prior distributions, we add parameters κ_ν and κ_γ such that: $A = \kappa_\nu \nu I_K$ and $G = \kappa_\gamma \gamma I_n$. Thus, the lower the parameters (κ_ν and κ_γ), the less informative the distribution. Furthermore we fix the trace of the two matrices respectively as:

$$\Psi = K \times \frac{\tilde{\Psi}}{\text{tr}(\tilde{\Psi})},$$

$$\Phi = N \times \frac{\tilde{\Phi}}{\text{tr}(\Phi)}.$$

The prior distribution for each ρ_{ij} is a uniform distribution as $p(\rho_{ij}) \propto 1$ for $\rho_{ij} \in (-1, 1)$, $i = 1, \ldots, h$, $j = 1, \ldots, m$.

4 Posterior Distributions

In order to compute the posterior distributions, we rewrite (3.4) as:

$$p(X|F, \Lambda, \rho, \sigma_\varepsilon^2, \Phi, \Psi) \propto \sigma_\varepsilon^{-TNK} |\Phi|^{-KT/2} |\Psi|^{-TN/2} \times \exp\left\{-\frac{\text{tr}\left(\Psi^{-1}\sum_t (X(t) - F(t)\Lambda')'\Phi^{-1}(X(t) - F(t)\Lambda')\right)}{2\sigma_\varepsilon^2}\right\}, \tag{4.1}$$

which can be equivalently written as

$$p(X|F, \Lambda, \rho, \sigma_\varepsilon^2, \Phi, \Psi) \propto \sigma_\varepsilon^{-TNK} |\Phi|^{-KT/2} |\Psi|^{-TN/2} \times \exp\left\{-\frac{\text{tr}\left(\sum_t (X(t) - F(t)\Lambda')\Psi^{-1}(X(t) - F(t)\Lambda')'\Phi^{-1}\right)}{2\sigma_\varepsilon^{-2}}\right\}, \tag{4.2}$$

where $X' = (X(1)', \ldots, X(T)')$ and $F' = (F(1)', \ldots, F(T)')$.

Using the definitions of the likelihood function in (4.1)–(4.2) and the set of prior distributions in (3.6)–(3.11), it is then easy to write down all the conditional posteriors.

The conditional posterior density of the disturbance covariance matrix given the factor scores, the factor loadings and the data, is again an inverted Wishart with updated parameters as:

$$p(\tilde{\Psi}|F,\Lambda,\Phi,X,m) \propto |\tilde{\Psi}|^{-(TN+m+\nu+2K+2)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\tilde{\Psi}^{-1}U\right)\right),$$

where

$$U = \sigma_\varepsilon^{-2}\sum_t (X(t)-F(t)\Lambda')'\Phi^{-1}(X(t)-F(t)\Lambda') + (\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)' + A,$$

$$E(\tilde{\Psi}|F,\Lambda,\Phi,\sigma_\varepsilon^2,X,m) = \frac{NT}{NT+m+\nu}\frac{\sigma_\varepsilon^{-2}\sum_t (X(t)-F(t)\Lambda')'\Phi^{-1}(X(t)-F(t)\Lambda')}{NT} + \\ + \frac{m}{NT+m+\nu}\frac{(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'}{m} + \frac{\nu}{NT+m+\nu}A.$$

Similarly, the posterior distribution for $\tilde{\Phi}$ is an inverted Wishart distribution with updated parameters. The expected value is a weighted sum of empirical covariance, factor scores covariance and prior hypotheses about covariance between observations, more precisely:

$$E(\tilde{\Phi}|F,\Lambda,\Psi,\sigma_\varepsilon^2,X,m) = \frac{KT}{KT+mT+\gamma}\frac{\sigma_\varepsilon^{-2}\sum_t (X(t)-F(t)\Lambda')\Psi^{-1}(X(t)-F(t)\Lambda')'}{KT} + \\ + \frac{mT}{KT+mT+\gamma}\frac{\sum_t F(t)F(t)'}{mT} + \frac{\gamma}{KT+mT+\gamma}G.$$

The posterior distribution for σ_ε^2 is an inverse Gamma, with updated parameters: $\kappa^* = \kappa + \frac{TNK}{2}$ and

$$\theta^* = \theta + \frac{1}{2}\mathrm{tr}\left(\Psi^{-1}\sum_t (X(t)-F(t)\Lambda')'\Phi^{-1}(X(t)-F(t)\Lambda')\right).$$

To compute the posterior distribution of R only (3.8) contributes, as:

$$p(R|X,F,\Psi,\Phi,\Lambda) \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\sum_t (F(t)-F_h(t)R)\Phi^{-1}(F(t)-F_h(t)R)'\right\}.$$

The conditional posterior distribution for the factor scores given the correlation matrix, the disturbance covariance matrix, the factor loadings and the data is multivariate normal as shown in Supplementary Material, and it results as follows:

$$\mathrm{vec}\,(F(1),\ldots,F(T)) \mid X,\Phi,\Psi,R,\ldots \quad \sim \quad N(\tilde{\mu}_F,\tilde{\Sigma}_{F,m}),$$

$$\begin{aligned}\tilde{\Sigma}_{F,m} &= \Phi \otimes \left[\left(M_m^{-1}(R) + \sigma_\varepsilon^{-2} I_T \otimes \Lambda'\Psi^{-1}\Lambda\right)^{-1}\right], \\ \tilde{\mu}_F &= \tilde{\Sigma}_{F,m}(\Phi^{-1} \otimes \sigma_\varepsilon^2 I_T \otimes \Psi^{-1}\Lambda)\text{vec}\left(X(1), \ldots, X(T)\right).\end{aligned}$$

We underline that to compute the posterior distribution of factors in the VAR case, it is convenient to consider $\text{vec}\left(F(1), \ldots, F(T)\right)$, that is the vectorization of a $N \times Tm$ matrix, rather than the vectorization of the matrix F defined above, that has dimension $NT \times m$.

In the particular case when all factor scores have the same autocorrelation coefficient, namely when $R = \rho I_m$, we have instead

$$\begin{aligned}\text{vec}(F) \mid X, \Lambda, \ldots &\sim N(\tilde{\mu}_F, \tilde{\Sigma}_F), \\ \tilde{\mu}_F &= \tilde{\Sigma}_F(I_T \otimes \Phi^{-1} \otimes I_m)\text{vec}(X\Psi^{-1}\Lambda), \\ \tilde{\Sigma}_F &= \left[\left(I_T \otimes \Phi^{-1} \otimes \Lambda'\Psi^{-1}\Lambda\right)\sigma_\varepsilon^{-2} + M(\rho)^{-1} \otimes \Phi^{-1} \otimes I_m\right]^{-1}.\end{aligned}$$

Where $M(\rho)$ is defined as:

$$\mathbf{M}(\rho) = (1-\rho^2)^{-1}\begin{pmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & 1 & \cdots & \rho^{T-2} \\ \vdots & & \ddots & \\ \rho^{T-1} & \rho^{T-2} & \cdots & 1 \end{pmatrix}.$$

Computational details for the posterior distribution of the factors (under these and more general assumptions) are shown in Supplementary Material.

The (unconstrained) posterior distribution of Λ is the following:

$$p(\Lambda|F, \Psi, \Phi, X, m) \propto \exp\left\{-\frac{1}{2}\text{tr}\left(\Psi^{-1}(\Lambda - \tilde{\Lambda})(H + \sum_t F(t)'\Phi^{-1}F(t))(\Lambda - \tilde{\Lambda})'\right)\right\}$$

$$\tilde{\Lambda} = \left(\sum_t X(t)'\Phi^{-1}F(t) + \Lambda_0 H\right)\left(H + \sum_t F(t)'\Phi^{-1}F(t)\right)^{-1}. \quad (4.3)$$

Differently to the standard Bayesian approach, that uses completely informative priors for the elements λ_{kj} subject to the constraints, we impose ex-post the restriction by conditioning the posterior distribution. In this way, the whole posterior distribution of Λ, which enters in all terms of the constrained distribution, is informed by the data.

We consider the $m \times K$ dimensional vector $\text{vec}(\Lambda)$, and define the two sub-vectors λ_u, corresponding to the unconstrained elements in Λ and λ_c, that instead contains the $m(m+1)/2$ terms of Λ that are constrained in (3.5). Then, let $\lambda_{c,0}$ be their corresponding constrained values (either 0 or 1).

Let us denote by

$$\tilde{\Sigma}_\lambda = \left(H + \sum_t F(t)'\Phi^{-1}F(t)\right)^{-1} \otimes \Psi, \quad (4.4)$$

and $\tilde{\lambda} = \text{vec}(\tilde{\Lambda})$, namely, the posterior variance matrix and mean vector of $\text{vec}(\Lambda)$. Then, by appropriately reordering the elements of $\tilde{\Sigma}_\lambda$ and $\tilde{\lambda}$, we can write $\tilde{\Sigma}_\lambda$ and $\tilde{\lambda}$ in blocks, corresponding to λ_u and λ_c respectively:

$$\tilde{\Sigma}_\lambda = \begin{pmatrix} \tilde{\Sigma}_{uu} & \tilde{\Sigma}_{uc} \\ \tilde{\Sigma}_{cu} & \tilde{\Sigma}_{cc} \end{pmatrix} \tag{4.5}$$

and $\tilde{\lambda}' = (\tilde{\lambda}'_u, \tilde{\lambda}'_c)$.

Imposing constraints on the posterior distribution of Λ is equivalent to computing the posterior distribution of λ_u, conditional to $\lambda_c = \lambda_{c,0}$. This distribution is readily computed, from the well known properties of a multivariate Gaussian distribution:

$$\lambda_u \mid X, F, \Phi, \Psi, m, \lambda_c = \lambda_{c,0} \sim \mathcal{N}\left(\tilde{\lambda}_u + \tilde{\Sigma}_{uc}\tilde{\Sigma}_{cc}^{-1}(\tilde{\lambda}_{c,0} - \tilde{\lambda}_c), \tilde{\Sigma}_{uu} - \tilde{\Sigma}_{uc}\tilde{\Sigma}_{cc}^{-1}\tilde{\Sigma}_{cu}\right).$$

5 Number of Factors

An important issue to be faced is the choice of the number of factors. In the previous sections, likelihood, prior and posterior distributions are defined or computed by assuming a fixed number of factors. In this Section we illustrate how their number can be determined. Bayesian and non-Bayesian references that tackle the estimation/choice of the number of common factors are, among many others, Lawley and Maxwell (1962); Bartholomew et al. (2011); Martin and McDonald (1975); Press and Shigemasu (1989); Rowe (2003). The book by Bartholomew (1987) is an excellent overview of the field right before MCMC tools became available.

According to Rowe (2003) the posterior probability of m is given by

$$\begin{aligned} p(m \mid \mu, \Lambda, F, \Psi, \Phi, \sigma_\varepsilon^2, X) \sim\ & p(m)p(\mu)p(\Lambda|F, \Psi, m)p(F|\Phi, m)p(\Psi)p(\Phi)p(\sigma_\varepsilon^2) \\ & \times p(X|\Lambda, F, \Phi, \Psi, \mu, m), \end{aligned}$$

$p(m)$ indicates the prior distribution of m. Since different methods for model selection tend to favour high number of factors, we choose a prior distribution inversely proportional to m.

Due to the computation difficulties to sample from the previous posterior distribution, to choose the optimal number of factors we propose the following algorithm. The main idea is to sample a value $m^\star$ from the prior distribution of m. Given $m^\star$, we estimate all the components of the model according to the method previously exposed: $\hat{F}(t)$, $\hat{\Lambda}$, $\hat{\Psi}$ and $\hat{\Phi}$ as defined in (3.2) and (3.3). Knowing all the components, we can sample a dataset $X^\star$ from the estimated model. We then compute a measure of proximity between observed data and each data set sampled from the estimated distribution, based on either the correlation or a distance. The optimal number of factors is estimated according to one of the following criteria:

- we take the average of the values $m^\star$ weighted by the correlation or the inverse distance,

- all sampled values such that the distance between the observed and sampled data is greater than the $90-th$ percentile of the computed distances are discarded, and then the sample mean of the observed value is the estimated value for m.

6 Simulations

The aims of the simulations are basically twofold: to investigate the ability to reconstruct the data with a reduction of the dimension of the variables and the ability to correctly identify the latent structure.

We need to simulate a data set where the variables are generated by a lower number of factors and investigate whether the method is able to correctly catch the underlying structure. To reach this aim, we first follow the mechanism generating data explained in Ghosh and Dunson (2009), adding an $AR(1)$ structure for the factor scores. By assuming $K = 10$, $N = 20$, $T = 10$ and $m_0 = 3$ we simulate the following three factors model:

$$X = F_0\Lambda_0' + \epsilon, \tag{6.1}$$

$$\epsilon_{n,t} \sim N(0, \Sigma_0), \tag{6.2}$$

$$diag(\Sigma_0) = (0, 2079, 0.1, 0.1525, 0.2, 0.36, 0.1875, 0.1875, 1, 0.2, 0.27), \tag{6.3}$$

$$\Lambda_0' = \begin{bmatrix} 0.89 & 0 & 0.25 & 0 & 0.8 & 0 & 0.5 & 0.1 & 0 & 0 \\ 0 & 0.9 & 0.25 & 0.4 & 0 & 0.5 & 0 & 0 & -0.3 & -0.3 \\ 0 & 0 & 0.85 & 0.8 & 0 & 0.75 & 0.75 & 0 & 0.80 & 0.80 \end{bmatrix}, \tag{6.4}$$

$$F_0(t_0) \sim N(0,\, I_m), \tag{6.5}$$

$$F_0(t) \sim N(\rho_0 F_0(t-1),\, 0.1 \times I_m), \tag{6.6}$$

$$\rho_0 = 0.4. \tag{6.7}$$

We simulate a data set with ten variables generated assuming the structure in (6.1)–(6.7). From each data set simulated, we estimate the factor scores and the factor loadings, with different values of m and, then, reconstruct the data set with $\hat{F}\hat{\Lambda}'$. The number of factors is determined slightly differently from the algorithm illustrated in Subsection 5, because here for each iteration the values of m from 2 to 9 are considered. We are not sampling using the prior distribution of m, since the aim is investigating the ability of the method to reconstruct the simulated data varying the value of m. Once we have estimated the reduced model, we measure the distance between simulated and estimated data through the following three measures: (i) the correlation between the data generated X and the data reconstructed $\hat{F}\hat{\Lambda}'$ (ii) the distance between the mean $\mu_0 = F_0\Lambda_0'$ and (iii) the estimated mean $\hat{\mu} = \hat{F}\hat{\Lambda}'$ and the log condition number of $\Sigma_0^{-1}\hat{\Sigma}$. We repeated these steps 100 times.

In Table 1 we report the results of the simulation. For each value m the tables shows the mean value for the three measures of distance between true and reconstructed data.

From 100 simulations, the number of factors, $\hat{m}$, is estimated to be equal to 3.67 versus the real value of 3. This result again confirms the performance of our method. The average correlation between simulated data and reconstructed data is 0.863.

m	$log(cond(\Sigma_0^{-1}\hat{\Sigma}))$	$\|\|\mu_0 - \hat{\mu}\|\|^2$	$\rho(X, \hat{\mu})$
2	17.56	400.34	0.635
3	**16.09**	**226.01**	**0.75**
4	16.289	735.59	0.697
5	17.28	$> 10^4$	0.578
6	18.08	$> 10^4$	0.421
7	19.07	$> 10^4$	0.274
8	20.92	$> 10^4$	0.259
9	23.83	$> 10^4$	0.123

Table 1: Results for the simulation according to generating mechanism as in (6.1)–(6.7). For each number of factors extracted, the average log condition number of $\Sigma_0^{-1}\hat{\Sigma}$, the average distance between the mean $\mu_0 = F_0\Lambda_0'$ and the estimated one and the average correlation between the data generated and the data reconstructed.

Moreover, we simulate data sets with a random generating structure: p, number of variables, is fixed to be equal to 15, $T = 10$ and ρ_0 is fixed to be equal 0.4. The first step is the random generation of the quantities m_0, F_0, Σ_0 and Λ_0.

$$\begin{aligned} m_0 &\sim UniformDiscrete[2, 11], \\ F_0(t_0) &\sim N(0, I_m), \\ F_0(t) &\sim N(\rho_0 F_0(t-1), 0.1 \times I_m), \\ \Sigma_0 &\sim InverseWishart(I_p, 2+p). \end{aligned}$$

Elements of Λ_0 are randomly sampled from the same elements that appear in the matrix of factor loadings defined in (6.4). At the second step, we simulate a data set X with the sampled generating structure. From matrix X, we extract factor loading and factor scores with number of factors from 2 to 11, and we choose the model that minimize the distance between simulated and reconstructed data. The correlation between the data generated and the model reconstructed has an average of 0.83, ranging from 0.705 and 0.893. Moreover, the correlation between the real value of m and the chosen one is 0.683.

Furthermore, to test the ability of the method proposed to guess the underlying structure, besides reconstructing the data, we simulate data sets with p, number of variables equal to 8, $T = 5$, and $n = 20$. The matrix of factor loading is a 0/1 matrix:

$$\Lambda_0' = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

The rest of the structure remains unchanged:

$$\begin{aligned} X &= F_0\Lambda_0' + \epsilon, \\ \epsilon_{n,t} &\sim N(0, \Sigma_0), \\ F_0(t_0) &\sim N(0, I_m), \\ F_0(t) &\sim N(\rho_0 F_0(t-1), 0.1 \times I_m), \\ \Sigma_0 &\sim InverseWishart(I_p, 2+p). \end{aligned}$$

We simulate 500 dataset following the assumptions above and assuming $\rho = 0$, that corresponds to assume absence of temporal structure. For each dataset, we compute the correlation between matrix Λ_0, factor loadings used to generate the data, and matrix $\hat{\Lambda}$, mean of the samples from the posterior distribution. The average correlation is 0.859, the median is 0.959, with 95% credible interval equal to $[0.277, 0.99]$. Furthermore, the average correlation between matrix F_0, generating the data, and the matrix $\hat{F}$ is 0.821, the median is 0.923, with 95% credible interval equal to $[0.129, 0.975]$.

Finally, we simulate 500 dataset assuming $\rho = 0.4$. For each dataset, we compute the correlation between Λ_0 and $\hat{\Lambda}$, mean of the samples from the posterior distribution, the average value is 0.922, the median is 0.976, with 95% credible interval equal to $[0.212, 0.99]$. Furthermore, the average correlation between F_0 generating the data and the matrix $\hat{F}$ is 0.781, the median is 0.854, with 95% credible interval equal to $[0.067, 0.908]$.

7 Results

The results reported are based on two chains of posterior samples of 200000 iterations per MCMC, with a *burn-in iterations* of 100000. Convergence for each parameter was checked by visual examination of trace plots and by the calculation of the Brooks and Gelman diagnostic (Brooks and Gelman, 1998). Convergence of matrices was checked through convergence of eigenvalues. In terms of real time, the run took about 5 hours. All of our computations performed in Matlab (version 2019b) on a PC with a processor Intel Core $i7-6700T$ and 16Gb of RAM. Note that the estimates are rather stable and the number of iterations can be reduced to 20000 without significant changes.

Applying the method described in Section 5, 1500 values were sampled from the distribution of m, and we obtain a posterior mean of the number of factors equal to 5. In this Section, results obtained by calculating five factors are therefore presented. As illustrated in Section 5, different criteria are suggested to chose the optimal number of factors, all of them leading to the estimate $\hat{m} = 5$:

- the average of the sampled values of m weighted by the correlations between real data and reconstructed data provide 5.45 as the chosen value for the number of factors,
- the average of the sampled values of m weighted by the inverse of distances between real data and reconstructed data provide 5.57 as the chosen value for the number of factors,
- discarding any samples that provide a distance between real data and reconstructed data greater than the $90-th$ percentile, $\hat{m}$ results 5.41.

The choice of hyper-parameters is a fundamental aspect of the analysis and will be discussed in Subsection 7.1. Concerning hyperparameters in the prior distribution of $\tilde{\Psi}$ and $\tilde{\Phi}$, indicated in (3.10) and in (3.11), we assume that the scale parameters, ν and γ, are equal to 1, while κ_ν and κ_γ are fixed to be equal to 0.1.

Matrix H, prior covariance matrix for the distribution of Λ, is assumed to be a diagonal matrix with 0.01 in each cell of the diagonal, and Λ_0 is defined as explained in Section 7.1. The prior distribution for each component ρ_j is a uniform distribution between -1 and 1. Moreover, the parameter σ_ε^2 in (3.3) is assumed to be equal to 1, by standardization of the matrix $X(t)$.

In Figure 9 posterior estimate of matrix Λ is shown. For the first loading, constrained to be correlated with *alcohol consumption*, infection disease is the only variable with a negligible association. The first factor loading results highly correlated with unemployment rate and with SDR of most of the diseases: alcohol and smoking related diseases, all causes disease, circulatory system and ischaemic diseases, tuberculosis and liver disease. Furthermore first factor loading is negatively correlated with protein intake, life expectancy, fertility rate, fat intake and average income, and has a weaker negative association with SDR for breast and all cancer.

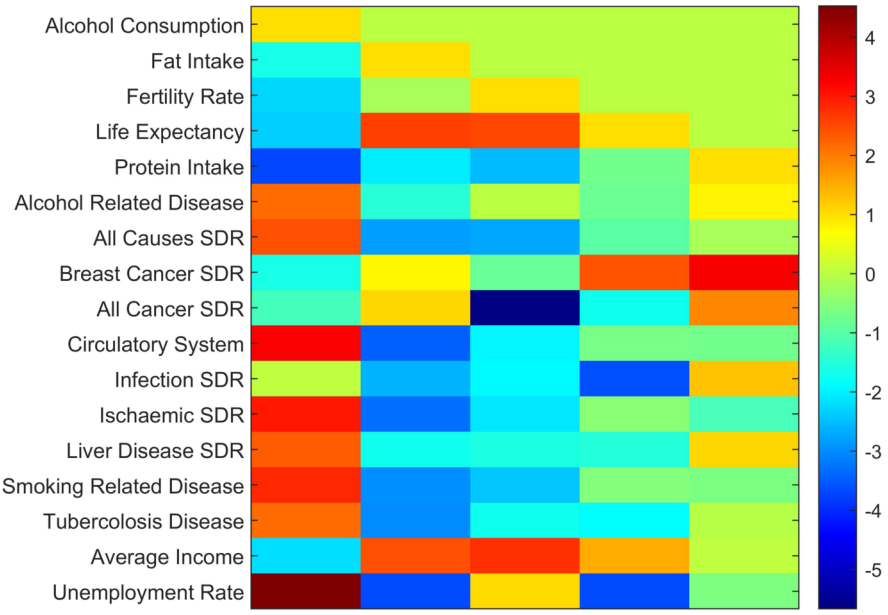

Figure 9: Factor Loadings.

As Figure 10 shows, by plotting the first factor score over Europe, in the years 2000, 2003, 2006 and 2009, it is marked a geographical contrast between Mitteleuropean countries and the rest of Europe. The first factor score clearly marks a contrast between richer versus poorer countries, with the exception of Germany and Austria, whose scores are nearer to those of East European countries. The evolution of the scores through the decade is minimal, and shows a slight divergence in the extreme values of the score. This factor marks differences in lifestyle, that is mostly attributable to heterogeneity in financial resources and impacts on (almost) all health outcomes.

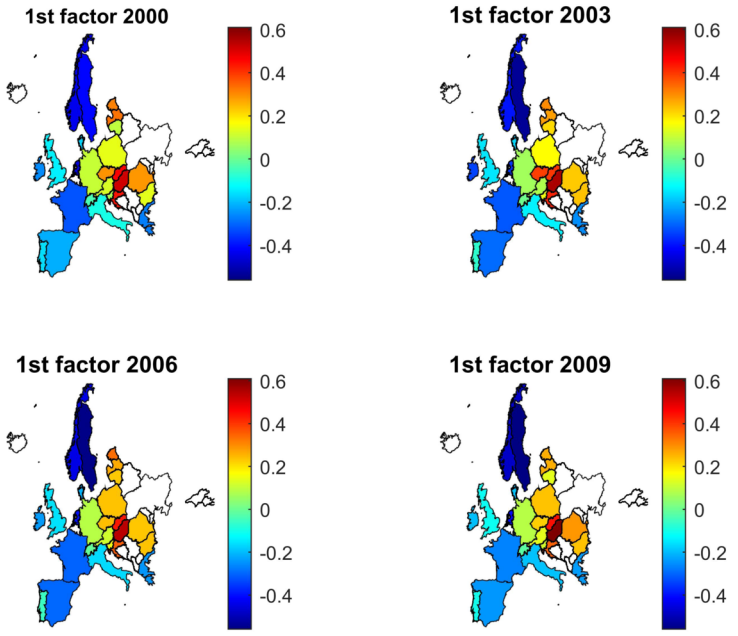

Figure 10: 1^{st} Factor Score.

The second factor loading, constrained to be correlated with *fat intake*, has a high positive association with life expectancy and average income. At the same time it results negatively correlated with unemployment rate, SDRs for tuberculosis, circulatory system and ischaemic diseases, infection disease, all causes disease, smoking related disease and protein intake, and appears also a weaker negative association with liver disease and alcohol related diseases SDR. Among the SDRs, the only positive association is with cancer (all and breast).

As shown in Figure 11, that maps the second factor score over Europe in 2000, 2003, 2006 and 2009, the second factor score signs a contrast between Northern and other countries, but it shows a higher variability in time relative to the first factor. The second factor seems to mark some differences in cancer mortality rates related to heterogeneities both in calories intakes (correlated with income) and dietary habits.

The third factor loading is constrained to be correlated with *fertility rate* and results highly positively correlated to life expectancy and average income. It is worth noting that both the second and third factors have high positive correlation with income and life expectancy, and are negatively correlated with most of the SDRs, but while the second factor shows a positive correlation with cancer SDR, the third factor has a strong negative correlation with all cancer SDRs. In Figure 12, the third factor score over Europe is shown, in 2000, 2003, 2006 and 2009, identifying an evident contrast between eastern and western countries. France is dominating the maps of the third factor scores, probably due to the higher fertility rate compared to other countries with

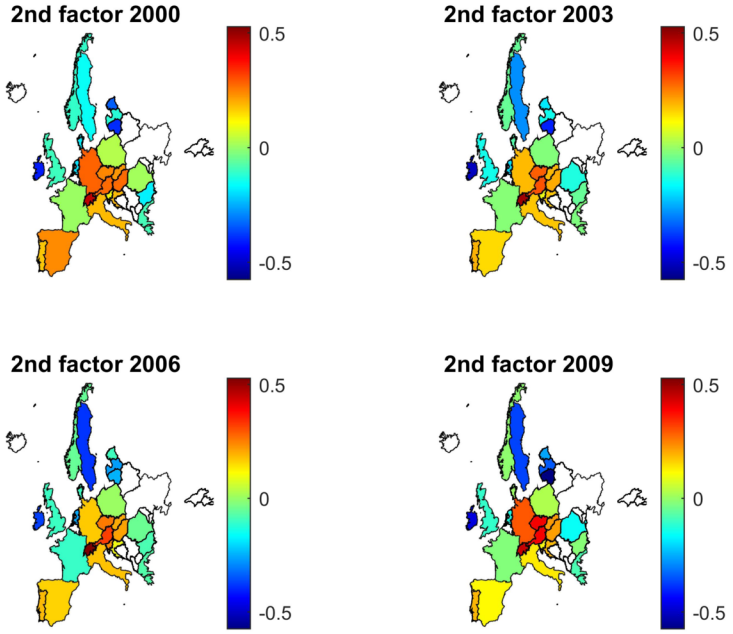

Figure 11: 2^{nd} Factor Score.

similar socio-economic conditions. Here the opposition of richer vs poorer countries is stronger than in the first factor and captures the diffusion of primary prevention and the quality of health care systems.

The fourth factor (maps for the last two factor scores in 2000, 2003, 2006 and 2009 are available upon request), constrained to be correlated with life expectancy, has strong correlation with income (positive) and unemployment rate (negative) and is strongly positively correlated with breast cancer SDR only. It marks a contrast between Mediterranean vs other European countries. This factor is difficult to interpret, it could pattern some hidden social or cultural unobserved variables affecting women participation to prevention programs in rich countries.

The fifth factor is the only one that shows no correlation with the economic variables. It is constrained to be correlated with protein intakes and shows positive correlations with (in descending order) breast cancer, all cancer SDR, infection, liver disease and alcohol related SDR. The lack of correlation with economic variables reflects in the geographical pattern of scores, which does not evidence clusters similar to those of the previous factors (that could be roughly summarized as a South-West/North-East contrast). This factor juxtaposes France and the Baltic states, that dominate the maps in all years, in contrast to Southern-Mediterranean and Scandinavian countries, with highly negative scores. A common feature of the latter group is wide coastal area and the development of fishing industry: this affects the availability of fish protein and might suggest the fifth factor to mark differences in quality of protein intakes.

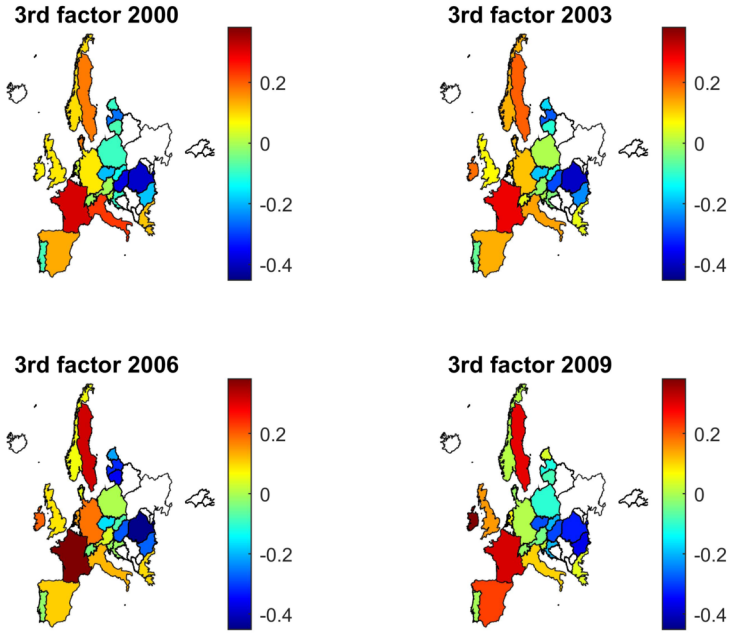

Figure 12: 3^{rd} Factor Score.

Finally, concerning posterior estimates of R we obtained different estimation of auto-correlation coefficients for the different scores, with a final estimate of $diag(R) = (0.313, 0.399, 0.466, 0.445, 0.321)$.

The posterior distributions of each component have a wide area of overlapping, we cannot identify different behaviours regarding temporal dependence of the factors. We then consider the model with $R = \rho I_m$. In Figure 13 the posterior distribution of ρ is shown, posterior estimates of ρ results 0.42 indicating a rather strong temporal correlation between the factor scores. It also results higher than the average autocorrelation of each score.

Overall, the factor analysis seems to suggest the existence of two main geographical blocks: Eastern vs Western countries. Depending on the factor, the latter can be further decomposed into Scandinavian, Mediterranean and Mittel-European countries. These blocks appear to be stable through time, with some slight differences among factors. Occasionally, some countries are associated to different groups: Spain and France, for example, in some cases show factor scores not in line with those of their neighbors.

With the aim of comparing our analysis with the results form a different approach, we implemented, in the WinBugs software, the algorithm proposed by Tzala and Best (2008). The comparison is not straightforward: Tzala and Best (2008) model is applicable to Poisson distributed data, and can extract only one factor. To apply their model to our data, we then limited the attention to variables representing counts of deaths,

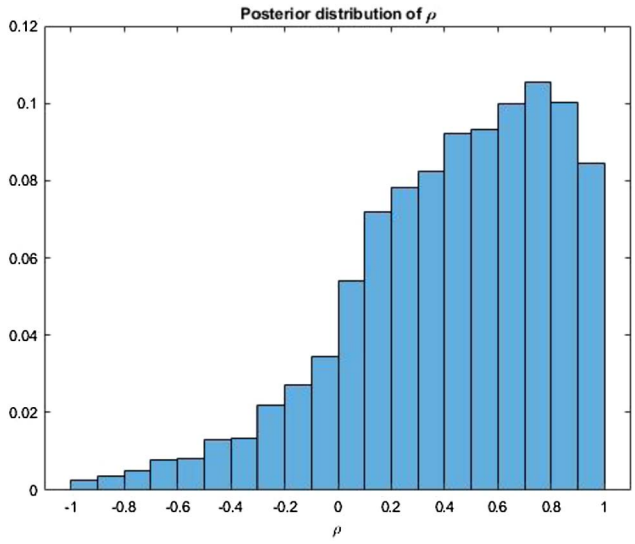

Figure 13: Posterior distribution of parameter ρ.

that lend themselves to be modeled by a Poisson distribution, leaving out the other socio-economical and behavioural variables.

We extracted one factor from a matrix with 10 columns, containing the observed counts of deaths for all the diseases indicated. The results obtained with the WinBugs software are in line with our results, the factor loading estimated contrasts breast and all cancer deaths versus all the other diseases, with again a differentiation between Eastern and Western European countries, quite stable in time. In terms of computing time, the two procedures are pretty similar, with our procedure only slightly faster than Tzala and Best (2008)'s.

7.1 Sensitivity Analysis

Lee and Press (1998) study robustness in Bayesian factor analysis with respect to the choice of hyperparameters. Coherently with their findings, our model confirms that estimation of Λ is very sensitive to the choice of Λ_0. As a consequence, also posterior estimates of F depends heavily on the choice of the elements in Λ_0.

To attenuate the lack of robustness, we use a two-step procedure to model Λ_0. First, we randomly generate elements of a matrix Λ_{00} from a uniform distribution between -2 and 2, according to structure in (3.5). Running the model for 1000 iterations we keep the posterior estimates of Λ as a value of Λ_0 for the next MCMC iterations. The estimates for Λ_0 obtained in the first-step procedure is the prior expected value for the factor loadings and will influence the posterior estimates as shown in (4.3). The two step procedure aims at reducing the dependence on our prior choice for Λ_{00}.

By definition, H is any positive definite matrix, it is assumed here that H is either a diagonal matrix or proportional to the identity. Assuming H scalar matrix with magnitude comparable with the inverse of the sample size looks a reasonable choice. There is no reason to assume a prior correlation between elements in Λ. From (4.4), the posterior estimates of Σ_Λ, covariance matrix of Λ, is the combination of the prior precision matrix, H, and the variance covariance matrix of the column of factor scores F. The posterior estimates of variance covariance matrix of Λ can result far from a diagonal matrix as the covariance matrix of factor scores indicates.

Identifiability of the factors is the most delicate aspect of the factor model. Since we choose a lower triangular structure for Λ, a critical comment to be aware of is that the order of variables impacts the interpretation of factors and might influence the model fit. We investigate whether the posterior estimates of factor loadings and factor scores depend on the order of the variables. We randomly change the order of the variables and compare the posterior estimates of the factor scores and the factor loadings obtained conditionally on the order. The posterior estimates of factor loadings are different, (since posterior estimates are also strongly dependent on Λ_0), but the interpretation of factor scores remains unchanged varying the order of the variables, and furthermore it has no impact on forecasting.

Considering the application to real data illustrated in Section 2, we compare the first three factors extracted from different models, obtained by varying the parameter m, from three to eleven factors. We compute the correlation between the factor extracted and we obtain that the correlation among the first factors in all different models is always greater, in absolute value, than 0.84. Further, all absolute correlations among second factors are greater than 0.73 while the absolute correlation among the third factors is always greater than 0.62.

The choice of the shape hyperparameters in inverse Wishart prior distributions has always been critical, especially when it is desired to achieve arbitrarily high noninformativity (Gelman (2006)). One of the main drawbacks of the inverse Wishart distribution is that the uncertainty of the matrix is determined by the single scalar parameter (ν or γ). The introduction of the parameters $(\kappa_\nu, \kappa_\gamma)$ helps to better control for the uncertainty without changing prior expectation. This limitation of the Wishart prior induced some authors to consider generalizations such as, for instance, the Generalized inverted Wishart distribution (see f.i. Brown et al. (1994)). The less informative choice for γ and ν guaranteeing stability of the estimates, that is $\gamma = 1$, $\nu = 1$, proves to work reasonably well even when sizes N and K are (moderately) unbalanced. However, when the difference between the number of variables and the space dimension increases, the prior precision parameters may differ substantially. In particular, prior precision relative to the covariance matrix with the highest dimension tends to increase and, as a consequence, the weight of the prior expected value on the posterior expected value tends to increase. The introduction of the hyperparameters κ_ν and κ_γ, fixed to 0.1, doesn't appear to impact the posterior distributions of Ψ and Φ: they seem, in fact, to be more sensitive to the order of magnitude of the shape parameters than to the form of the matrices A and G.

8 Conclusion

This paper develops and applies a new hierarchical Bayesian method for factor analysis for multivariate spatially and temporally correlated data. Our method advances the traditional literature on factor analysis in its computational feasibility and interpretation. The main hypothesis underlying the proposed model is that the latent factors share the same between-locations dependence as the observed variables. The underlying structure reflects on one hand our idea of a vector autoregressive pattern in the latent variables relative to adjacent time periods and on the other hand the presence of high correlation between two regions, (or census tracts), that are close to each other.

The main novelty of this work is the development of the factor model for spatially and temporally correlated data motivated by Burkitt (1969) and Burkitt (1970), who stated that the variation in the geographic and temporal pattern of a disease may be related to its cause. This paper applies an innovative approach in trying to analyze the mutual relationship among the geographical distributions of different diseases together with some well known risk factors, both behavioural and socio-economic.

In analyzing high-dimensional, or even moderate-dimensional, multivariate data, one is faced with the problem of estimating a large number of covariance parameters. The method proposed results particularly useful for datasets with a large number of variables where the interest is not in any single variable measured, but in uncovering meaningful common patterns underlying the data.

Assuming an inverted Wishart distribution for matrix Φ, the model proposed does not force the correlation within variables to be necessarily determined by geographic distance, but lets the data suggest the nature of the dependence. Our method borrows information either across geographical areas and different times. Indeed, the resulting multivariate spatio-temporal analysis may also lead to improved precision for the estimation and forecasting of the underlying factors by acquiring strength from other variables as well as close areas or time points.

The flexibility of a separable approach through a Kronecker product can be adapted to specify parametric structures for Φ. For example, the exponential matrix specification is considered in Strauss et al. (2017) and Leorato and Mezzetti (2016), as a way to reduce the number of parameters to be estimated.

A real data example illustrates the relevance of the methodology offering empirical evidence of the model proposed. The application highlights a geographic pattern in the risk factors that mirrors economic differences between Eastern and Western European countries. Furthermore, a simulation study is carried out and provides very encouraging results. The aim of this type of empirical analysis is to verify whether the results contribute to strengthening the hypothesis that environmental and local characteristics can act as a common risk factors for many types of cancers and of other diseases. These environmental factors, are very likely to be shared by adjacent regions, due to spatial spillovers or correlations, thus causing risk factors to show a geographical pattern. The simultaneous variation of factor scores through time and space helps to identify the presence of convergence phenomena in one or more factors or progressive concentrations of some factors in different areas. This may further add evidence to common sources of

influence that reflect underlying shared risk factors, and proves to be particularly useful for diseases with less clear aetiology or rare diseases. In our application, the spatial risk pattern appears to be stable through time, thus rather than reinforcing the hypothesis of spatial convergence of risks, our analysis highlights the existence of persistent regional specific factors. Once the factors are extracted, they can be used for forecasting the variables of interest, by augmenting an observation driven model by the estimated factors.

Supplementary Material

Supplementary Material to "A Bayesian Factor Model for Spatial Panel Data with a Separable Covariance Approach" (DOI: 10.1214/20-BA1215SUPP; .pdf).

References

Aguilar, O. and West, M. (2000). "Bayesian dynamic factor models and portfolio allocation." *Journal of Business and Economic Statistics*, 18: 338–357. 489, 500

Arminger, G. and Muthén, B. (1998). "Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm." *Psychometrika*, 63: 271–300. 489

Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2016). "Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem." *Journal of Econometrics*, 192(1): 190–206. MR3463672. doi: https://doi.org/10.1016/j.jeconom.2015.10.010. 500

Bai, J. (2003). "Inferential theory for factor models of large dimensions." *Econometrica*, 71(1): 135–171. MR1956857. doi: https://doi.org/10.1111/1468-0262.00392. 500

Bai, J. and Li, K. (2012). "Statistical analysis of factor models of high dimension." *The Annals of Statistics*, 40(1): 436–465. MR3014313. doi: https://doi.org/10.1214/11-AOS966. 500

Bai, J. and Wang, P. (2015). "Identification and Bayesian Estimation of Dynamic Factor Models." *Journal of Business and Economic Statistics*, 33(2): 221–240. MR3337059. doi: https://doi.org/10.1080/07350015.2014.941467. 500

Bartholomew, D. J. (1987). *Latent variable models and factors analysis*. Oxford University Press, Inc. 505

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. New York: John Wiley & Sons. MR2849614. doi: https://doi.org/10.1002/9781119970583. 489, 505

Brooks, S. and Gelman, A. (1998). "Alternative methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7: 434–455. MR1665662. doi: https://doi.org/10.2307/1390675. 508

Brown, P., Le, N., and Zideck, J. (1994). "Inference for a covariance matrix." In Lindley, D., Smith, A., and Freeman, P. (eds.), *Aspects of Uncertainty: A Tribute to D. V. Lindley*, Wiley Series in Probability and Mathematical Statistics, 77–92. Wiley. MR1309689. 514

Burkitt, D. (1969). "Related disease-related cause?" *Lancet*, 2: 1229–1231. 515

Burkitt, D. (1970). "Relationship as a clue to causation." *Lancet*, 2: 1237–1240. 515

Calder, C. A. (2007). "Dynamic factor process convolution models for multivariate space–time data with application to air quality assessment." *Environmental and Ecological Statistics*, 14: 229–247. MR2405328. doi: https://doi.org/10.1007/s10651-007-0019-y. 491, 495, 500

Chan, J. C. C. (2020). "Large Bayesian VARs: A flexible Kronecker error covariance structure." *Journal of Business and Economic Statistics*, 38(1): 68–79. MR4044952. doi: https://doi.org/10.1080/07350015.2018.1451336. 490

Conti, G., Frühwirth-Schnatter, A., Heckman, J. J., and Piatek, R. (2014). "Bayesian exploratory factor analysis." *Journal of Econometrics*, 183(1): 31–57. MR3269916. doi: https://doi.org/10.1016/j.jeconom.2014.06.008. 500

Das, M., Pal, S., and Ghosh, A. (2010). "Factor analysis of risk variables associated with metabolic syndrome in adult Asian Indians." *Journal of Cardiovascular Disease Research*, 1(2): 86–91. MR3218231. 492

Fan, J., Fan, Y., and Lv, J. (2008). "High dimensional covariance matrix estimation using a factor model." *Journal of Econometrics*, 147: 186–197. MR2472991. doi: https://doi.org/10.1016/j.jeconom.2008.09.017. 491

Flood, A., Rastogi, T., Wirfält, E., Mitrou, P. N., Reedy, J., Subar, A. F., Kipnis, V., Mouw, T., Hollenbeck, A. R., Leitzmann, M., and Schatzkin, A. (2008). "Dietary patterns as identified by factor analysis and colorectal cancer among middle-aged Americans." *The American Journal of Clinical Nutrition*, 88(1): 176–84. 492

Frühwirth-Schnatter, S. and Lopes, H. F. (2018). "Sparse Bayesian factor analysis when the number of factors is Unknown." Technical report, ArXiV. URL https://arxiv.org/abs/1804.04231 500

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 514

Geweke, J. and Zhou, G. (1996). "Measuring the pricing error of the arbitrage pricing theory." *Review of Financial Studies*, 9(2): 557–587. 491, 500

Ghosh, J. and Dunson, D. (2009). "Default prior distributions and efficient posterior computation in Bayesian factor analysis." *Journal of Computational and Graphical Statistics*, 18: 306–20. MR2749834. doi: https://doi.org/10.1198/jcgs.2009.07145. 506

Hogan, J. and Tchernis, R. (2004). "Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census

data." *Journal of the American Statistical Association*, 99: 314–324. MR2109313. doi: https://doi.org/10.1198/016214504000000296. 490

Kaufman, G. and Press, S. (1973). "Bayesian factor analysis." Technical Report 66273, Sloan School of Management, University of Chicago. URL http://hdl.handle.net/1721.1/1870 489

Lawley, D. (1940). "The estimation of factor loadings by the method of maximum likelihood." *Proceedings of the Royal Society of Edimburgh*, 60: 64–82. MR0002754. 495

Lawley, D. and Maxwell, A. (1962). "Factor analysis as a statistical method." *The Statistician*, 12(3): 209–229. MR0343471. 505

Lawson, A. B., Choi, J., Cai, B., Hossain, M., Kirby, R. S., and Liu, J. (2012). "Bayesian 2-Stage Space–Time Mixture Modeling With Spatial Misalignment of the Exposure in Small Area Health Data." *Journal of Agricultural, Biological, and Environmental Statistics*, 17: 417–441. MR2993274. doi: https://doi.org/10.1007/s13253-012-0100-3. 490

Lee, S. E. and Press, S. J. (1998). "Robustness of Bayesian factor analysis estimates." *Communications in Statistics – Theory and Methods*, 27(8): 1871–1893. MR1647895. doi: https://doi.org/10.1080/03610929808832196. 513

Leorato, S. and Mezzetti, M. (2016). "Spatial panel data model with error dependence: A Bayesian separable covariance approach." *Bayesian Analysis*, 11: 1–35. MR3545473. doi: https://doi.org/10.1214/15-BA979. 489, 490, 497, 501, 515

Leorato, S. and Mezzetti, M. (2020). "Supplementary Material to"A Bayesian Factor Model for Spatial Panel Data with a Separable Covariance Approach"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1215SUPP. 499

Lopes, H. F., Salazar, E., and Gamerman, D. (2008). "Spatial dynamic factor analysis." *Bayesian Analysis*, 3(4): 759–792. MR2469799. doi: https://doi.org/10.1214/08-BA329. 490

Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14: 41–67. MR2036762. 489, 490

Martin, J. and McDonald, R. (1975). "Bayesian estimation in unrestricted factor analysis. A treatment for Heywood cases." *Psychometrika*, 40: 505–517. MR0488503. doi: https://doi.org/10.1007/BF02291561. 489, 505

Mezzetti, M. (2012). "Bayesian factor analysis for spatially correlated data: application to cancer incidence data in Scotland." *Statistical Methods and Applications*, 21(1): 49–74. MR2881629. doi: https://doi.org/10.1007/s10260-011-0177-9. 489, 490

Pigott, T. D. (2001). "A review of methods for missing data." *Educational Research and Evaluation*, 7(4): 353–383. 493

Polasek, W. (1997). "Factor analysis and outliers: a Bayesian approach." *Discussion Paper, University of Basel*. 489

Press, S. (1972). *Applied multivariate analysis*. Holt, Reinhart and Winston. MR0420970. 489

Press, S. and Shigemasu, K. (1989). "Bayesian inference in factor analysis." In L. Gleser, S. P. a. A. S., M. Perlman (ed.), *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, chapter 15. Springer-Verlag, New York. 489, 505

Press, S. and Shigemasu, K. (1997). "Bayesian inference in factor analysis-revised, with an appendix by Rowe, D. B." Technical Report 243, Department of Statistics, University of California, Riverside. 489

Ren, Q. and Banerjee, S. (2013). "Hierarchical factor models for large spatially misaligned data: a low-rank predictive process approach." *Biometrics*, 69(1): 19–30. MR3058048. doi: https://doi.org/10.1111/j.1541-0420.2012.01832.x. 490

Rowe, D. (2003). *Multivariate Bayesian statistics: Models for source separation and signal unmixing*. Chapman & Hall/CRC. MR2000406. 489, 490, 495, 505

Rowe, D. and Press, S. (1998). "Gibbs sampling and hill climbing in Bayesian factor analysis." Technical Report 225, Department of Statistics, University of California, Riverside. 490

Strauss, M. E., Mezzetti, M., and Leorato, S. (2017). "Is a matrix exponential specification suitable for the modeling of spatial correlation structures?" *Spatial Statistics*, 20: 221–243. MR3654011. doi: https://doi.org/10.1016/j.spasta.2017.04.003. 515

Tzala, E. and Best, N. (2008). "Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality." *Statistical Methods in Medical Research*, 17(1): 97–118. MR2420192. doi: https://doi.org/10.1177/0962280207081243. 490, 491, 495, 512, 513

Wang, F. and Wall, M. M. (2003). "Generalized common spatial factor model." *Biostatistics*, 4: 569–582. 490

Yanai, H., Inaba, Y., Takagi, H., Toyokawa, H., and Yamamoto, S. (1978). "An epidemiological study on mortality rates of various cancer sites during 1958 and 1971 by mean of factor analysis." *Behaviormetrika*, 5: 55–74. 492